Universidad de Navarra

DOCTORAL THESIS

Interpretable Precision Medicine for Acute Myeloid Leukemia

By

MARIAN GIMENO COMBARRO

TECNUN. UNIVERSIDAD DE NAVARRA

A dissertation submitted to the University of Navarra in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in TECNUN, Faculty of Engineering

DECEMBER 2022

Donostia-San Sebastián

This work was supported by Cancer Research UK [C355/A26819] and FC AECC and AIRC under the Accelerator Award Programme, and Synlethal Project (RETOS Investigacion, Spanish Government).

Agradecimientos

Simplemente, reconocer que el mérito de este trabajo es fruto no sólo mío si no de mucha gente que ha estado detrás durante estos cuatro años de estudio. Agradecer, en primer lugar, a mis directores de tesis Ángel y Fer por animarme a seguir y creer en el proyecto en aquellos momentos en los que se veía muy negro. Por vuestro trabajo para poder sacar adelante cada una de las publicaciones y proyectos que hemos ido realizando estos años, soy consciente de que sin vuestro apoyo no estaría aquí hoy.

Agradecer también a mi familia, a mi marido, a mis padres y a mis hermanos. Durante los momentos de penumbra han sabido dar luz, ánimo y esperanza. Habéis recogido mis incertidumbres, alegrías y llantos siempre animándome a seguir hacia delante. Gran parte de este mérito es sobre todo vuestro. Gracias, especialmente a ti, Ricardo, por tanto.

Gracias, también, a nuestros queridos colaboradores del CIMA y de la CUN, a Xabi, Edurne, Felipe y Sara. Por vuestro interés, ayuda, trabajo y consejo con nuestros proyectos, especialmente con AML. Se nos ha quedado la espinita de las validaciones, pero gracias por vuestra disponibilidad.

Gracias a todas las personas que han formado parte de estos cuatro años de estudio, a mis compañeros de grupo: Katyna, Juan, César, Carlos, Naroa, Luisvi, Iñigo, Francis, Telmo, Francesco, Danel, Sergio, Idoia, Jesús y a esos compis que, aunque no sean del grupo, los queremos igual: Ane, Camila, María, Giorgia, Mireya, Andrés, Nacho, Jacobo. Gracias especialmente al comité de cumpleaños por humanizar y alegrar el día especial de cada uno de nosotros. Gracias a Miriam, por su disponibilidad, interés y atención cada mañana, gracias a mi querido equipo de limpieza, especialmente a Belén y a Amaia.

Gracias a mis asesoradas, especialmente a las de cuarto curso, ha sido muy emocionante compartir estos años con vosotras, gracias por compartir vuestras incertidumbres, aspiraciones e ilusiones, que tanto me han ayudado.

Deo Omnis Gloria

Table of Contents

Abstract

Precision medicine (PM) is a branch of medicine that defines a disease at a higher resolution using genetic and other technologies to enable more specific targeting of its subgroups. Because of its uses in clinical treatment and diagnostics, this field exemplifies the modern era of medicine. PM looks for not just the right drug, but also the right dosage and treatment regimen. PM encounters a variety of challenges, which will be explored in this dissertation.

Large-scale sensitivity screens and whole-exome sequencing experiments (WES) have fostered a new wave of targeted treatments based on finding associations between drug sensitivity and response biomarkers. These experiments with the aid of state-of-the-art artificial intelligence (AI) algorithms are opening new therapeutic opportunities for diseases with unmet clinical needs. It has been proved that AI is capable of predicting novel personalized treatments based on complex genotypic and phenotypic patterns in tumors. The scientific community should make an effort to make these algorithms to be interpretable to humans so that the results could be easily approved by the medical regulators. The purpose of this thesis is to apply AI algorithms for precision oncology that are highly accurate, while guaranteeing that the predictions are interpretable by humans.

This work is divided in three main sections. The first section comprises a new methodology to increase the predictive power of the discovery of novel treatments in large-scale screenings by exploiting that some biomarkers tend to appear in many treatments. This fact is called hub effect in gene essentiality (HUGE). Content of this section was published in [1]. The second section contains a novel interpretable AI method -called multi-dimensional module optimization (MOM)- that associates drug screening with genetic events and proposes a treatment guideline. Content of this section was published in [2]. Finally, the third section includes a detailed comparison of different recently published algorithms that attempt to overcome the barriers proposed by today's precision medicine. This study also includes two novel algorithms specifically designed to solve the challenges of applicability to clinical practice: Optimal Decision Tree (ODT) and Multinomial Lasso.

The characterization of Interpretable Artificial Intelligence as approach with strong potential for use in clinical practice is one of the study's most significant achievements. We present

unique methods for PM that are highly interpretable, and we summarize the needs that could be considered for constructing interpretable AI. We are confident that this method will transform the way PM is addressed, bridging the gap between AI and clinical practice.

List of Figures

List of Tables

Glossary

PM	Precision Medicine
LED(s)	Lethal Dependency
HUGE	Hub-Effect in Gene Essentiality
AML	Acute Myeloid Leukemia
MOM	Multi-dimensional Module Optimization
ODT	Optimal Decision Tree
XAI	Explainable Artificial Intelligence
WES	Whole-Exome Sequencing
ML	Machine Learning
DL	Deep Learning
KRL	Kernelized-ranked learning
GDSC	Genomics of Drug Sensitivity in Cancer

Introduction

Chapter 1. Introduction to Precision Medicine Challenges and Acute Myeloid Leukemia

Precision medicine (PM) is the science that "defines a disease at a higher resolution by genomic and other technologies to enable more precise targeting of its subgroups" [3]. It is an emerging field that epitomes the new era of medicine owing to its applications in clinical treatment and diagnosis [4].

PM tries to find not only the right drug but also the right dosage and the proper treatment schedule. These goals are usually summed up as "targeting the right treatments to the right patients at the right time" [5]. PM faces different challenges that will be described in this introduction.

The challenge of getting the patients' response to drugs.

PM requires the different patients' characteristics to make their predictions [6] such as genomic and transcriptomic data, health records, lifestyle characteristics, etc. (**Figure 1**). With an adequate data policy, they are reasonably easy to obtain; genomic data can be acquired from sequencing techniques, wearable technologies can collect data that provide lifestyle information, EHRs are invaluable sources of information on health status and previous conditions, etc. Its integrative analysis requires complex models and a solid understanding of the interaction of biological systems [7].

However, PM also requires drug sensitivity information which is much more difficult to find, having most likely incomplete information on all patients' response to all available drugs, i.e. each patient is given one or, at most, a few drugs, not all the possible ones[8] (**Figure 1**). Even in these cases, distinguishing between responders and non-responders is not an easy task and requires tailoring methods specific to each disease. In turn, these different criteria for different diseases make it difficult to compare diseases or drugs [9].

Large-scale sensitivity screenings such as PDX (patient-derived xenografts), loss-of-function screens or *ex-vivo* experiments can be used as proxies to estimate the patients' response to several drugs [10]. Ex-vivo experiments in hematological cancers are of great importance since they are performed directly on the patient's living tumor cells[11,12].

All these three approaches have strong limitations. In the case of PDX, the animal models' immune system is usually compromised and, in the case of loss-of-function screens there

is not an interaction of a specific drug but the effect of its target, and, in the case of ex vivo experiments –used mainly in hematologic oncology–, the interaction of the cells and the immune system is not properly modeled. Despite these difficulties, they are reasonable sources of information to predict the response of the patients to different treatments [13].

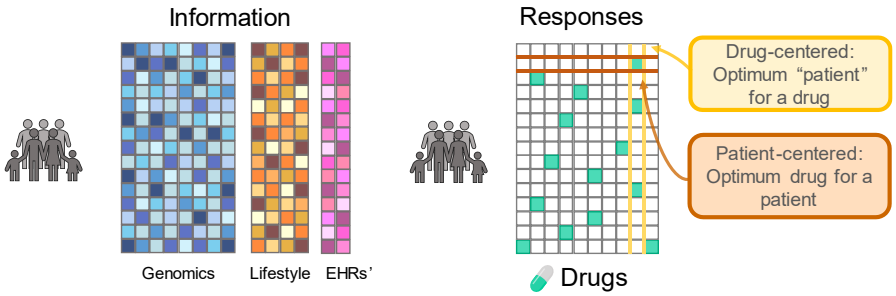


Figure 1. Precision Medicine paradigm. The left-side panel represents patients' data the and right-side panel shows the data available for patients' responses to treatment.

The multiple Hypothesis Problem for biomarker's finding when using large-scale sensitivity screening

The advance of personalized medicine, and in particular precision oncology, is partially based on the development of drug sensitivity studies. These experiments are promoting the discovery of new drugs, biomarkers of sensitivity, and drug repositioning. With increasing frequency, these studies have widened their scope from single drug studies to experiments involving hundreds of drugs −or even combination of drugs− and targets, also known as sensitivity screenings.

In recent years, sensitivity screenings are being carried out on hundreds of cell lines giving rise to large-scale sensitivity screening datasets,– e.g., GDSC, which includes 130 screened drugs in an average of 368 lines per drug [14]–, and large-scale loss-of function sensitivity screens –e.g. the Achilles Project [15,16], or The Project Score [17]. Combining these sensitivity studies with tumor genotypes makes it possible to associate the response to treatment with genetic alterations (biomarkers), thus promoting the search for new personalized therapies[18].

However, the multiple testing problem, related to the large number of gene knockouts or drug tested, and the number of possible biomarkers, limits the statistical power of these studies and, therefore, their potential to find new therapies.

Precision Medicine falls beyond traditional machine learning (ML) problems.

PM can be considered an assignation problem: each patient must be provided a drug (or a set of drugs) given the patient's information. This assignation problem does not perfectly fit in any of the "traditional" fields of machine learning. It is not a unsupervised problem, although, with a proper selection of variables, patients with identical "optimal" drugs should cluster together [19,20].

Regarding supervised machine learning, it is not either a standard regression problem since the aim is not to predict the effectiveness of a drug on each patient but to find *the most effective ones* [21]. Nevertheless, both problems are related and, if the effectiveness of each drug were exactly modeled, the "perfect" drug for a patient would be simply the most effective one predicted by the model. PM assignation could also be treated as a classification problem dividing the drugs for each patient into two classes: the most effective one belongs to one class and the others belong to another class. Again, it solves the problem if the predictions were perfect. However, since this simplistic model only considers misclassifications (the second-best drug is as bad as the worst), it does not work well in practice.

Finally, it can also be considered a reinforced learning problem [22]. For example, [23] includes a review of reinforcement learning applications to oncology. The objective of this field of machine learning is to learn an optimal, or nearly optimal, policy that maximizes the "reward function" –in this case, the patient's response to treatment. Reinforced learning is traditionally applied to teach the computer how to play games (chess, Go, or video games) [20]. In this case, different methods state how use a reinforced learning algorithm to "find a policy that maximizes the patient response to treatment".

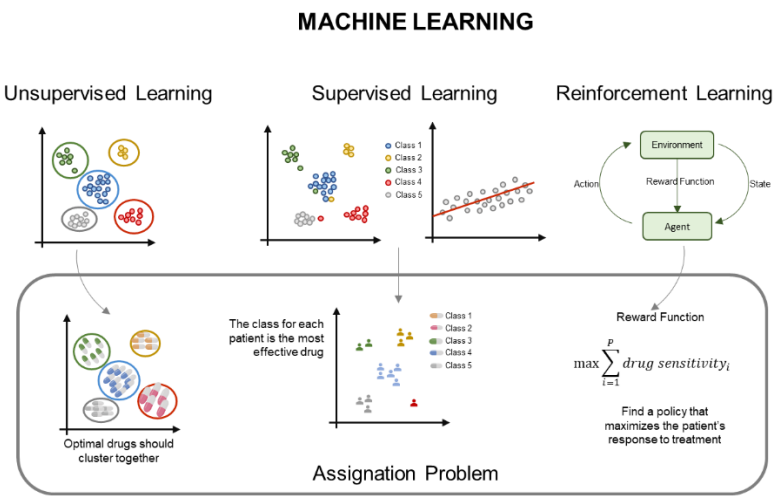


Figure 2: Relationship between Machine Learning and the Assignation Problem. The assignation problem is not a specific machine learning problem but could be addressed from all the machine learning branches.

As a result, precision medicine –assigning the proper drug to each patient– given the patients' data is a problem that shares characteristics of different machine learning fields (**Figure 2**) and can be tackled in many ways.

Different approaches for solving the assignation problem.

There are two main approaches to solve the goal of "targeting the right treatments to the right patients" (**Figure 1**). The first one is to state which is the proper drug for a specific patient. We will name this approach "patient-centered". The other approach consists of finding the patient or patients that are responders for a specific drug, named "drug-centered" in this review. This problem –closely related to finding biomarkers of response– is interesting for the pharma industry.

If the output of the algorithm is a continuous value, it is possible to adapt a drug-based method to solve the patient-based problem and vice-versa. For example, many drug-centered methods return a sensitivity score for each patient when applied to a specific drug. If this score is computed for all the drugs, it can be used to select the drug that maximizes sensitivity for each patient.

The challenge of interpretability.

One common problem in ML is the (lack of) interpretability. Exploring the potential of large-scale sensitivity screenings, artificial intelligence (AI) algorithms for personalized medicine focus on the analysis of such datasets to bridge the gap for drug discovery. Some studies use machine learning algorithms for monotherapy prediction [24,25], other approaches are based on training deep learning (DL) models from patients' omics data [26,27]. These methods create black-box predictors that make agnostic inferences of treatment for a patient based on complex non-linear relationships. The output is, for these cases, an individual therapy for a patient, instead of a general treatment guideline [28]. This approach has the inherent disadvantages of methods based on neural networks: they require a huge amount of data, and are unable to show the criteria that trigger the decision –since neural networks tend to be black-box models–. In many cases the blackbox algorithm gives no clues on why a specific decision is taken [29,30,39–41,31–38]. It is difficult, if not reckless, for a physician to use a treatment guideline with no information on the ultimate reasons that drove this recommendation. These technical challenges are limiting the translation of drug screening experiments to clinical practice.

Explainable AI tries to give a solution to "black-box" algorithms, by analyzing the weights and variables of the different models. There are several differences between Explainable AI and Interpretable AI. The article by Rubin, C., details how these two conceptions differ from one another [42]. The first is the method by which an explanation of the process followed by a "black-box" model is suggested, although frequently these justifications are not totally true or are not enough to apply critical judgment on the algorithm's thinking. The second, though, is the feature that enables the expert using the algorithm to offer his or her personal assessment of the outcomes it produces. It is important to start designing an interpretable ML model with subject-matter specialists, paying particular attention to the logic's clarity [42].

Interpretability focuses on making AI understandable to humans by the usage of "white-box" algorithms [43,44]. Interpretable AI is an active field of research: it justifies the response and ensures that, given the *a priori* knowledge, the recommendation is sensible. It also helps to improve the results since, as they are understandable by physicians, they can provide expert feedback to fine tune the algorithms [45–48]. The importance of using interpretable models in the finding of new personalized treatments is twofold: therapeutic pipelines can be more easily adopted in normal clinical guides (e.g., using a decision tree that does not require a complex model with a high number of variables) [44] and drug

regulators, such as the Food and Drug Administration (FDA), or European Medicines Agency (EMA) will have an easier journey to approve a drug if the companion biomarkers are reasonable and robust [49,50]. Some methods have tried to explain their reasoning to become more explainable but not of them could be defined as interpretable [43,46,51–56]. Consequently, interpretable ML opens the door to bridge the gap between clinical practice and bioinformatics [43,57].

Acute Myeloid Leukemia

We selected Acute Myeloid Leukemia (AML) as a disease model, a highly heterogeneous type of cancer that affects bone marrow cell precursors. In AML, genomic profiling is essential to understand its biology, diagnosis, and treatment [58–60]. Unfortunately, 70% of adult people diagnosed with this disease die within five years of diagnosis [61]. The current ELN (European Leukemia Network) risk stratification is based on the genetic biomarkers of the disease [62].

Current patient stratification guides divide AML patients into three subgroups according to their prognosis, namely favorable-, intermediate- and adverse-risk. Each subgroup is defined by a combination of genetic biomarkers that can be either chromosomal rearrangements, genetic mutations, or allele deletions. Thus, the favorable risk subgroup –a 5-year overall survival (OS) of 45% to 80%– includes 45% of AML patients and is diagnosed mainly through the biomarkers $NPM1^{Mut}$, chromosome 16 inversion (inv(16)), and $CEBPA^{Mut}$. The intermediate-risk subgroup –5 years OS of 30%– comprises 25% of AML cases and is associated with the internal tandem duplications in the *FLT3* gene *(FLT3-ITD)*, and $NPM1^{WT}$. Finally, the adverse risk subgroup –5 years OS of 10%– represents 30% of AML cases and has scattered deletions and complex karyotypes as biomarkers [59].

Although there are big prognosis differences across these genetic groups, the current approach for young and fit patients is a standard induction cytotoxic therapy ("3+7") [59,62] with different dosages and aggressiveness depending on the severity and with the addition of targeted therapies, mainly *FLT3* inhibitors, to a specific group of AML patients [59]. Recently, *FLT3* inhibitors have been incorporated as a treatment directed to *FLT3-ITD* patients, but effective treatments for patients who do not have *this* alteration remain an unmet clinical challenge [63].

Despite eight new drugs have been approved for AML in the last years, its lethality is still very high. In addition, there are no targeted treatments directed to *FLT3*^{WT} patients –70% of all AML cases[63]. An interpretable machine learning approach that identifies the most adequate FLT3 inhibitor as well as the treatment for other AML genotypes, would allow the discovery of new indications for other drugs for the AML. As a result, a new classification guide based on the response to therapy for specific genetic alterations would be beneficial in clinical practice.

In the following sections we will discuss how to address these challenges. Section 1 includes the definition and performance of a method to solve the multiple hypothesis problem. Section 2 comprises a new Machine Learning method designed to be interpretable and it solves the assignation problem. Finally, Section 3 includes a quantitative comparison that defined Interpretability of a model and compares six methods suited for PM.

Hypothesis and objectives

In the context of precision medicine, we propose the development of a model that, using artificial intelligence, solves the assignment problem with an accuracy comparable to other state-of-the-art methods, but also prioritizes its ease clinical implementation. We use Acute Myeloid Leukemia as a test model due to its clinical urgency and genomic heterogeneity.

The following objectives are defined in order to carry out this proposal:

1. Accuray: To improve accuracy, data from large-scale screenings will be used. As a result, the first goal is to increase the predictive power of these experiments, resulting in a larger number of reliable and significant hypotheses when analyzing these screenings.
2. Interpretability: After resolving the multiple hypothesis correction problem, we proposed the definition of an algorithm that apart from accurate, is also interpretable and simple so that it improves clinical traslationality.
3. Benchmarking: Finally, this model will be compared with other similar state-of-the-art models in terms of interpretability −the main feature that enhances clinical traslationality. Furthermore, we will define two refined interpretable artificial intelligence model that will aim to overcome the shortcomings observed in state-of-the-art models. The most essential characteristics of interpretability will be compiled in order to encourage the development of interpretable methods.

This report has been divided into three sections each of them illustrates how the preceding objectives were approached and accomplished. The method that solves the multiple hypothesis problem is explained in Section 1: "A Novel Method to Predict Lethal Dependencies with High Predictive Power." In section 2, "Interpretable Artificial Intelligence for Precision Medicine in Acute Myeloid Leukemia" an accurate, interpretable, and simple artificial intelligence model is defined. Finally, in section 3 "The challenge of interpretability", the algorithm developed in this doctoral thesis is compared with other methods in terms of interpretability, for which a quantitative and qualitative metric is defined, and two new methods are added, which obtained excellent results.

Section 1: A Novel Method to Predict Lethal Dependencies with High Predictive Power

Solving a massive Multiple Hypothesis testing problem

Introduction to Section 1

The traditional concept of synthetic lethality consists of the concurrent loss of functionality of two genes resulting in cellular death. A relevant example is the effectiveness of *PARP* inhibitors in tumors with inactivated *BRCA1* and *BRCA2* [64]. In recent years, the advances in functional genomics triggered by large-scale loss-of-function screening -such as *CRISPR-Cas9* or RNA interference (RNAi) screens- have boosted the discovery of hundreds of novel targets and context-specific lethal dependencies (LEDs) [15–17,65–67], defined as any association between two genes that results in differential viability depending on their genetic context (**Figure 3**).

Several studies have carried out large-scale functional genomic screens to identify genome-wide targets and LEDs [15,17,65,66]. The Project Score [17], the Achilles Project [15,16] and the Project DRIVE [67] are three studies that performed genome-wide gene-knockouts in cancer cells aiming at establishing novel targets and LEDs. The refinement of computational and technical tools have improved the potential of loss-of-function screening to identify cancer vulnerabilities [66,68,69]. However, the multiple-hypothesis problem related to the large number of gene knockouts limits the statistical power of these studies.

In this section we show that previous efforts to predict LEDs from functional screening can be significantly improved by considering the "HUb effect" in Genetic Essentiality (HUGE) of some gene alterations: a few specific sets of gene alterations are statistically associated with large changes in the essentiality of multiple genes. These "hub" aberrations lead to more statistically reliable LEDs than other alterations that do not participate in such hubs. We incorporated the HUGE effect in the statistical analysis of three recent loss-of-function experiments of both The Project Score and The Achilles Project (two datasets) showing that the number of LEDs discovered for a given FDR considerably improves for both *CRISPR-Cas9* and RNAi screens.

Using acute myeloid leukemia (AML), breast cancer (BRCA), lung adenocarcinoma (LUAD) and colon adenocarcinoma (COAD) as disease models, we validated that the predictions are enriched in associations used in the clinic. Finally, we validated *in-vitro* an example of a therapy guideline based in LED selection in AML. The HUGE analysis will help discover novel tumor vulnerabilities in specific genetic contexts, providing valuable candidates -

targets and genetic variants as biomarkers- for further personalized treatments in hematological diseases or other cancer disorders.

The HUGE-based methodology is published in [1].

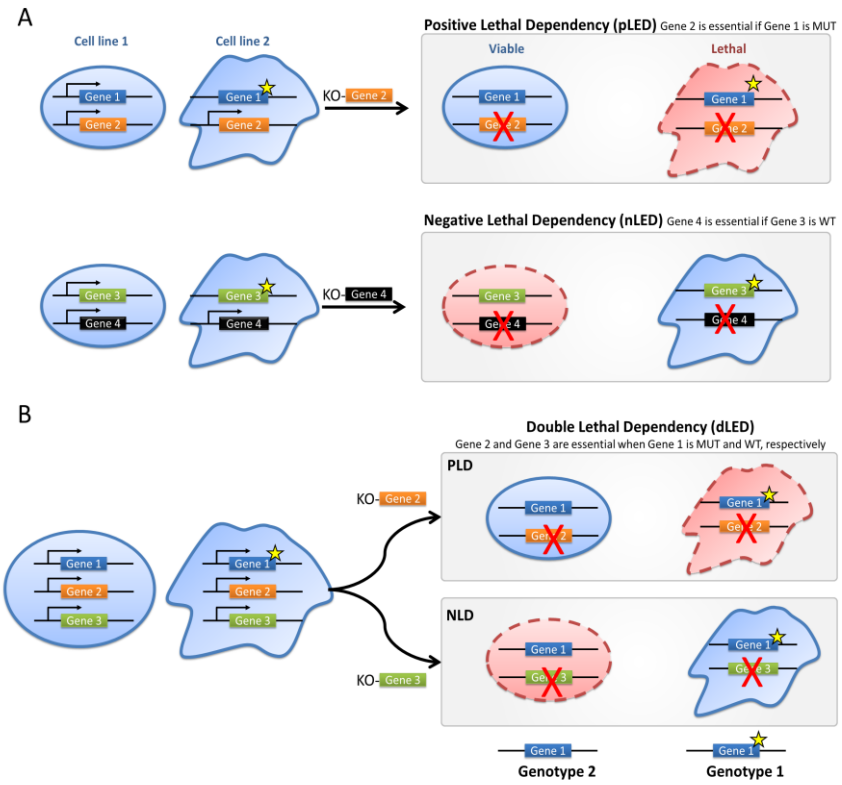


Figure 3. Types of Lethal Dependencies. Lethal dependencies that affect two genes: (A) in a Positive Lethal Dependency (pLED), a gene is essential for tumor survival when another gene is mutated (MUT). This is the traditional concept of synthetic lethality, in which a gene knockout (KO) causes cellular death only for another gene's mutant phenotype. (B) Conversely, in Negative Lethal Dependency (nLED), a gene is essential for tumor survival when another gene is not genetically altered (wild type-WT), here gene variant confers resistance to the inhibition. (C) A lethal dependency that affects three genes: Dual Lethal Dependency (dLED), an altered gene (Gene 1) confers, at the same time, sensitivity to the inhibition of one gene (Gene 2) and resistance to the inhibition of another gene (Gene 3). In this figure, the shape of the cells denotes different cell-types with different genomic characteristics. The color of the cells denote whether the cell survives to the knock down or not. The star shape in a gene denotes a genetic variant. The red crosses denote pharmacological inhibition.

Chapter 2. Methods

Data integration

Data of loss-of-function screens libraries (17,980 knockout genes in 412 cancer cell lines) of the project Achilles [70] were integrated with gene expression and their corresponding gene alteration profiles (gene variants in ~1600 genes; **Figure 4A**) obtained from CCLE and Shao et al. [16]. We gathered gene expression of cells using RNA-seq data to confirm that the genes that were essential for a cohort of cells were expressed before the RNAi library experiment was performed [71]. Gene variant panels were filtered out using the parameters of CCLE's authors to avoid common polymorphisms, low allelic fraction, putative neutral variants, and substitutions located outside of the coding sequence [72].

We used the DEMETER score [15,68] as a measure of gene essentiality of the RNAi libraries of the project Achilles [70]. DEMETER quantizes the competitive proliferation of the cell lines controlling the effect of off-target hybridizations of siRNAs by solving a complex optimization problem. The more negative the DEMETER score is, the more essential the gene is for a cell line. We imputed missing elements of DEMETER using the nearest neighbor averaging algorithm [73]. Besides, we collected gene expression patterns from RNA-seq data [71] to confirm that essential genes are expressed when they are essential. Based on DEMETER data, we first identified genes that were essential for a selected tumor subtype. Essential genes were required to meet several criteria: i) they must be essential for at least 20% samples of the selected cancer subtype, ii) they must be specific to the cancer type under study, i.e. they must be non-essential for other cancer types and iii) they must be expressed before RNAi experiment (>1TPM at least in 75% samples).

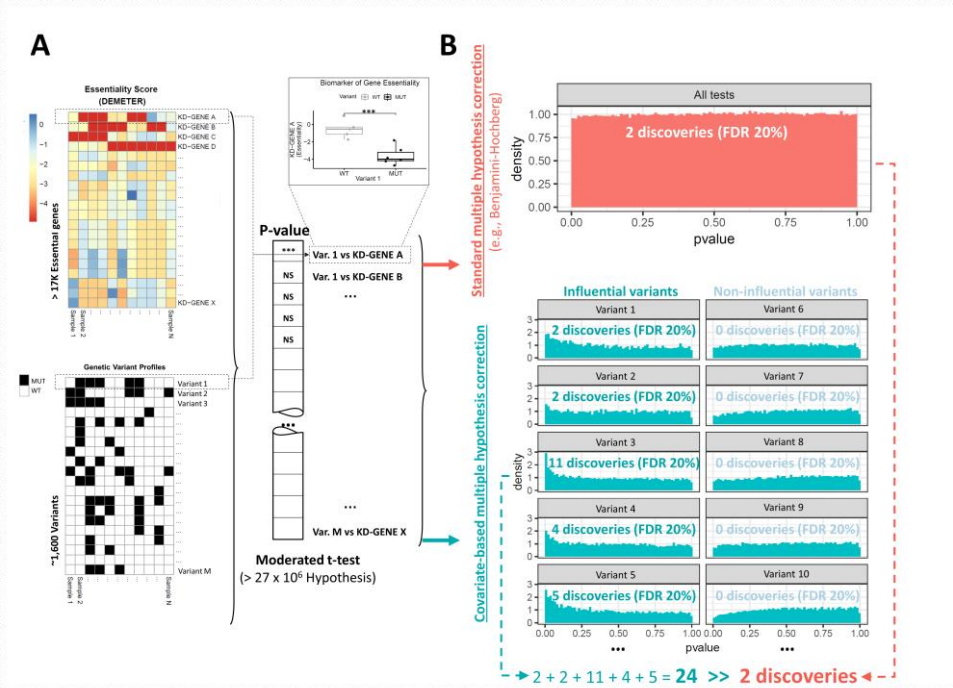


Figure 4. Computational pipeline to find lethal dependencies. (A) Scheme of data integration for N samples (cell lines). RNAi libraries data (gene essentiality; DEMETER score) and gene variant panels are represented as two heatmaps. Each pair of a knock-down gene (KD-gene) and a gene variant defines a p-value, which represents a lethal dependency. Boxplots represent the DEMETER score of a cell line when inhibiting one gene (X-axis) depending on the genetic alteration of another gene (Y-axis). Genes with a DEMETER score < -2 are considered essential for a cell line. (B) Scheme of the histogram of P-values using standard approaches (e.g., Storey-Tibshirani), in red; and using a covariate-based algorithm, in blue.

Statistical model

We developed a statistical algorithm to identify genes whose essentiality is highly associated with the genetic alteration of other genes. Dealing with this statistical issue implies solving a large multiple hypotheses problem (more than one million hypotheses). In similar scenarios, traditional corrections -such as Benjamini-Hochberg (BH), Bonferroni, or Holm- showed very few or no gene-biomarker LEDs for a given FDR [74]. To overcome this problem, we developed a covariate-based statistical approach -similar to the Independent Hypothesis Weighting procedure [74] (**Figure 4**).

Let *e* denote the number of RNAi target genes and *n* denote the number of screened samples. Let **D** be an $e \times n$ matrix of essentiality whose entries d_{ij} represent the

DEMETER score for the RNAi target *i* in sample *j*. Let **m** be a $m \times n$ dichotomized matrix whose entry m_{ij} denotes whether sample *j* is mutant or not according to the previous criteria:

$$m_{ij} = \begin{cases} 1, & if mutant(MUT) \\ 0, & if\ wild-type\ \ (\mathrm{WT})\ ' \end{cases} \quad (1)$$

Let **s** be a subset of n' cell lines that yields an essentiality vector $\boldsymbol{d_s} = \left(d_{e_{s_1}}, \ldots, d_{e_{s_{n'}}}\right)$ for the e[th] RNAi target. Let $\boldsymbol{m_s} = \left(m_{s_1} \ldots, m_{\mathrm{s_n}}\right)$ be the expression vector of a putative gene biomarker. The null hypotheses are defined as:

$$H_0^g : E(\boldsymbol{d_s} | \boldsymbol{m_s} \in MUT) = E(\boldsymbol{d_s} | \boldsymbol{m_s} \in WT) \quad (2)$$

This null hypothesis is, therefore: "the expected essentiality of a gene knock-down is identical in mutant and wild-type cell lines". To test this hypothesis, we used a moderated t-test implemented in *limma*[75]. We applied this test for each RNAi target and all the gene variants to get the corresponding p-values **(Figure 4).** Dealing with these p-values implies correcting for multiple hypotheses.

In our case, we divided the p-values corresponding to all the tests into n groups, where n is the number of altered genes. For each of these groups, we computed the local false discovery rate (local FDR) [76]. The local FDR estimates, for each test, the probability of the null hypothesis to be true, conditioned on the observed p-values. The formula of the local FDR is the following:

$$P(H_0|z) = localFDR(z) = \frac{\pi_0 f_0(z)}{f(z)}, \quad (3)$$

where *z* is the observed p-values, π_0 is the proportion of true null hypotheses –estimated from the data-, $f_0(z)$ the empirical null distribution –usually a uniform (0,1) distribution for well-designed tests- and $f(z)$ the mixture of the densities of the null and alternative hypotheses, which is also estimated from the data.

As stated by B. Efron and R. Tibshirani [76], «the advantage of the local FDR is its specificity: it provides a measure of belief in gene i's 'significance' that depends on its p-

value, not on its inclusion in a larger set of possible values» as it occurs, for example with q-values or the standard FDR. The local FDR and π_0 were estimated using the Bioconductor's R Package *qvalue* [77].

Comparison with the Project Score

To compare our results with Project Score's ones, we selected the same 12 primary cancer tissues shared in both datasets. The comparison followed two steps: 1) using CCLE and DEMETER scores with the Project Score's algorithm, 2) running our approach adapted to Project Score conditions. In the first step, following the code published in their work, an ANOVA test was performed on each tissue to calculate all possible dependent partners. The Storey-Tibshirani correction was then used, using the criteria mentioned in Project Score methods [17]. This enabled us to correct the ANOVA *p*-values and get significant associations. Secondly, the comparison between both methodologies was only possible if the same adjusted p-value is calculated for both datasets. Therefore, we estimated the FDR with our data as it is the q-value selected by the Project Score. The FDR correction was obtained using the Bioconductor R package *IHW* [74], which enables the consideration of covariates-based multiple hypothesis correction, as well as estimating the FDR. Discoveries from both methodologies in DEMETER and CCLE datasets were plotted in different volcano plots, and the number of significant LEDs were counted (FDR<20%).

Integration of the VICC knowledgebase of clinical interpretations of genomic variants

We downloaded 19,551 clinical interpretations of somatic genomic variants in cancer of the Variant Interpretation for Cancer Consortium (VICC) [78,79] (version December 2020). We filtered out incomplete (e.g., entries without annotated drug or biomarker) and redundant associations. We then selected all associations that are annotated to acute myeloid leukemia (AML) and synonyms. From all drugs, we selected those that have an annotated protein target. To do so, we retrieved the data publicly available in the ChEMBL [80] and DrugBank [81] online repositories. In total, 216 out of 19,551 associations matched these criteria. We considered a true positive if either HUGE or ST identifies an LED whose mutation biomarker coincides with a VICC's association and the protein target is included in the same association, or at least in a gene of the same pathway in the STRING database (v.11, STRING score threshold = 400; default value on STRING for "medium" confidence) [82].

We calculated ROC and PR curves considering the two top evidence levels included in VICC [78,79], namely, (i) evidence from professional guidelines or FDA-approved therapies; and (ii) evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus.

Application to acute myeloid leukemia as a disease model

We applied the pipeline to the AML cohort of cell-lines (n=15). In the first step, essential genes were required to be: (i) essential for at least 25% AML samples, (ii) specific for AML cells, and (iii) expressed before the RNAi experiment. The algorithm outputs a ranking of significant Lethal Dependencies (LEDs) that consist of a couple of genes in which the first one is essential depending on the genetic alteration of the other.

For the final ranking for AML, we selected those LEDs that showed a p-value < 0.05 and local FDR ≤ 0.6, |ΔDEMETER| > 2 (default value suggested by DEMETER's authors). Additionally, we interrogated which of these LEDs had direct relationships (co-expressed, annotated in the same pathway database, or contained in a common experiment) in the STRING database [82] to ensure there is an established biological relationship between the essential gene and the subrogate biomarker. This biological double-check is not necessary and can be omitted when the researcher looks for novel relationships.

In vitro validation was performed using siRNAs against *NRAS* and *PTPN11* in four different AML cell lines, two with *NRAS*-genetic variants (HL-60 and OCI-AML3) and two *NRAS*-wt cell lines (MV4-11 and HEL). Finally, the model was compared with 3 standard statistical methods (namely Benjamini-Hochberg (BH), Bonferroni and Holm) known to have suboptimal sensitivity (recall of true positives) in specific scenarios in 19 additional tumor subtypes to define the potential for controlling the FDR. [74] See **Appendix 1** for more details on the cell line culture protocol and the demonstration of the increased statistical power.

Chapter 3. Results

Gene variants associated with multiple essential genes increase the power of loss-of-function screens

One of the main statistical challenges to find LEDs by integrating genome-wide functional screens with -omics datasets is the multiple hypothesis testing problem. Correction for multiple hypothesis reduces the statistical significance of results (meaning a decreased detection rate and an increased false-positive rate). The Project Score presented a large-scale genome-wide *CRISPR-Cas9* screening analysis targeting 18,009 genes in 30 different cancer types, across 14 different tissues [17,83]. They presented a methodology to detect LEDs based on finding differences in genetic essentiality in cell lines associated with the presence of specific gene variants (ANOVA test [84] with the Storey-Tibshirani p-value correction). Following this procedure, the Project Score was able to identify genetic LEDs in 7 out of 14 individual tissues analyzed [17,83].

Analyzing Project Score's data, we noticed that for each tumor type a few specific genetic alterations were significantly associated with the genetic essentiality of a large set of genes. This handful of genetic aberrations shows a hub effect, in which a gene variant is associated with large changes in the essentiality of multiple genes. We termed this behavior as "HUb effect in Genetic Essentiality" (HUGE) (**Figure 5A**; other tumor types can be visualized in https://fcarazo.shinyapps.io/visnetShiny/). From the point of view of statistics, the HUGE effect is defined as improvement of the statistical power by using gene variants as co-variates in a multiple hypothesis problem. Other biological covariates such as gene expression or copy number alterations has also shown to be covariates that increase the statistical power [74]. Using gene variants as statistical covariates provides a larger number of positives for a given FDR, which consequently means an increased specificity and sensitivity, or type I and type II errors, as demonstrated in **Appendix 1**. Interestingly, the analysis shows that HUGE effect is present in all tumors analyzed, significantly improving the predictive power of LEDs.

The presence of the HUGE effect in a cancer type can be also understood as a predictive model in which each mutation has a different capability to define the genetic essentiality of

multiple genes. To show it visually, the histogram of p-values of a gene alteration represents how gene alterations are associated with the genetic essentiality of multiple genes. Histograms of the p-values for alterations that conform to a “hub” show a peak near the origin, which means that cells with these alterations are sensitive to the depletion of a large number of genes (**Figure 5B**). Conversely, if the hubs of alterations are not considered, the relationships of mutations and viability show a flat histogram of p-values. This does not necessarily mean that such relationships are not biologically relevant, but that is difficult to distinguish them from random associations and will be considered as artifacts after multiple testing correction.

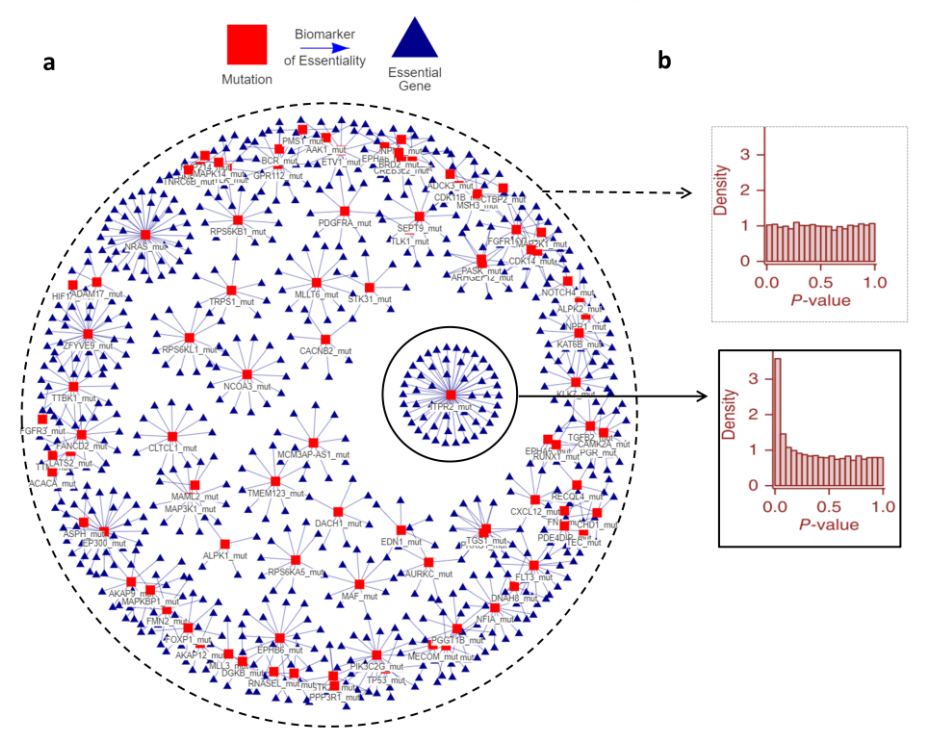


Figure 5 The hub effect in genetic essentiality (HUGE) in Acute Myeloid Leukemia: in a given cell, a small set of gene aberrations is associated with large changes in genetic essentiality. (a) A bipartite graph in which red squares represent gene variants (e.g., mutations), blue triangles represent significant changes in cell viability related to knocked-down genes. Both vertexes are linked by a line if the variations in the essentiality have a statistically significant association with the presence of the gene variant. (b) Implications in p-value histograms of the HUGE effect. Hub associations show a high peak close to zero p-values indicating that the null hypothesis is rejected in more cases and that these genetic variants are associated to a higher response to the inhibition of more gene products. Segregating the statistical analysis according to the alteration provides more statistical power. Essential genes and other tumor types can be visualized in https://fcarazo.shinyapps.io/visnetShiny/

The HUGE effect helps to palliate the multiple hypothesis correction problem. Using the mutation under study as a covariate, multiple hypotheses can be differently treated considering the overall association of gene alteration in the complete set of essential genes (**Figure 4** and **Figure 6**). Using this concept, we developed a statistical model that integrates HUGE information to find LEDs (**Figure 4**).

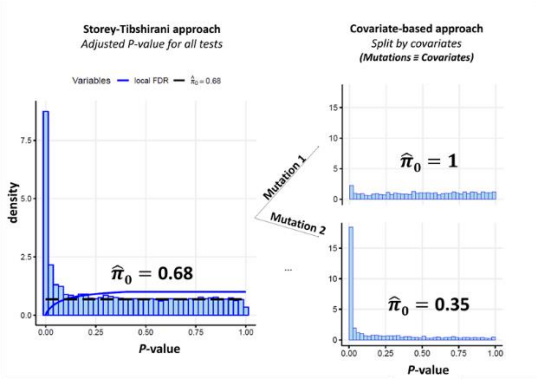


Figure 6. Schematic representation of the covariate-based statistical approach in this context. In this case, the use of genetic variants as covariates of a covariate-based problem allows the reduction of false positive rate, and consequently, the percentage of true null-hypothesis in a statistical test.

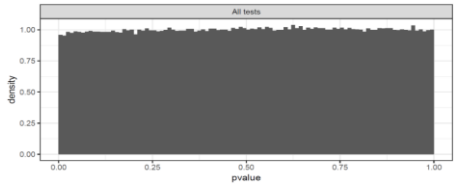


Figure 7. Histogram of P-values of all lethal dependencies in acute myeloid leukemia. Previous efforts to correct multiple testing in this problem consider a single set of tests (all gene aberrations and CRISPR-Cas9 knockouts) and apply a correction that control the FDR, such as Storey-Tibshirani (ST), as done in the Project Score. Interestingly, in this approach histogram of p-values shows flat-shaped histograms.

Previous efforts to correct multiple testing in this problem consider a single set of tests (all gene aberrations and *CRISPR-Cas9* knockouts) and apply a correction that controls the FDR, such as Storey-Tibshirani (ST), as done in the Project Score. Interestingly, in all tumors our approach increases the statistical power of the analysis. From a statistical point view, a flat histogram is compatible with the null hypothesis for all the tests and, therefore, multiple hypothesis correction drives to none or few discoveries (**Figure 7**). Every single tumor shows p-value histograms related to specific gene variants that have a higher zero-

peak than the histogram associated to all tests in such tumor. To test this approach, we compare the results using HUGE with previous LED identification strategies in three genome-wide functional genomic projects: The Project Score [17], the DEMETER score and the CERES score (DEMETER and CERES are included in the Achilles Project [15,16]). First, to test the potential of HUGE to predict LEDs with *CRISPR-Cas9* screens, we analyze the Project Score dataset [17]. Project Score integrates 215 different genetic events across 14 tumor types, including SNVs and CNVs. In the same reference, the authors found at least one LED in 7 out of the 14 tumor types analyzed. 40 out of 215 events were detected to be significant biomarkers of essentiality (FDR ≤ 20%), which correspond to 77 unique LEDs (a single genetic event can be associated with several essential genes). Analyzing Project Score's data using the HUGE-based methodology, we identify 1,438 unique associations with the same FDR (18 times larger than Project Score, **Figure 8A**), corresponding to 80 single genetic events. Besides, using HUGE we detect at least one LED in all the 14 tumors analyzed, finding LEDs in 10 tumors that would have been missed using the original pipeline, affecting around 10-20 genes for each disease type.

We also tested HUGE in the DEMETER score of the Achilles Project to predict LEDs, in this case using RNAi screening. The DEMETER dataset [15,70] is a large-scale genome-wide experiment of RNA interference libraries (17,085 knockdown genes) in 19 tumor types. We integrated the DEMETER data with the corresponding cell line gene alteration profiles (genetic variants in ~1,600 genes) obtained from the Cancer Cell Line Encyclopedia (CCLE) [72] and Shao et al. [16]. This integration turns out to have 27 Million hypotheses, which will hardly impair p-values after multiple hypothesis correction (**Figure 4**). Then, we replicate the Project Score's pipeline with the DEMETER dataset and compare it with the HUGE-based approach to find LEDs, also including in the comparison other two standard p-value corrections used to control the FDR, namely Holm and Bonferroni. Using the standard ST procedure, we find 126 LEDs (FDR ≤ 20%). There are LEDs for 7 out of 19 tumors. The same dataset and FDR threshold using the HUGE-based approach provides 9,535 LEDs (75.7 times larger than using ST). All cancer types (19 out of 19) showed significant LEDs in the HUGE-based analysis (**Figure 8B**). HUGE identifies 1,675 LEDs in 6 tumor types in which other methods recall no LEDs (FDR ≤ 20%); and 9,409 LEDs in 19 tumor types that would have been missed using previous procedures (FDR ≤ 20%; **Figure 8C**). These results show that the HUGE effect is present with different intensities in all tumor types analyzed (https://fcarazo.shinyapps.io/visnetShiny/).

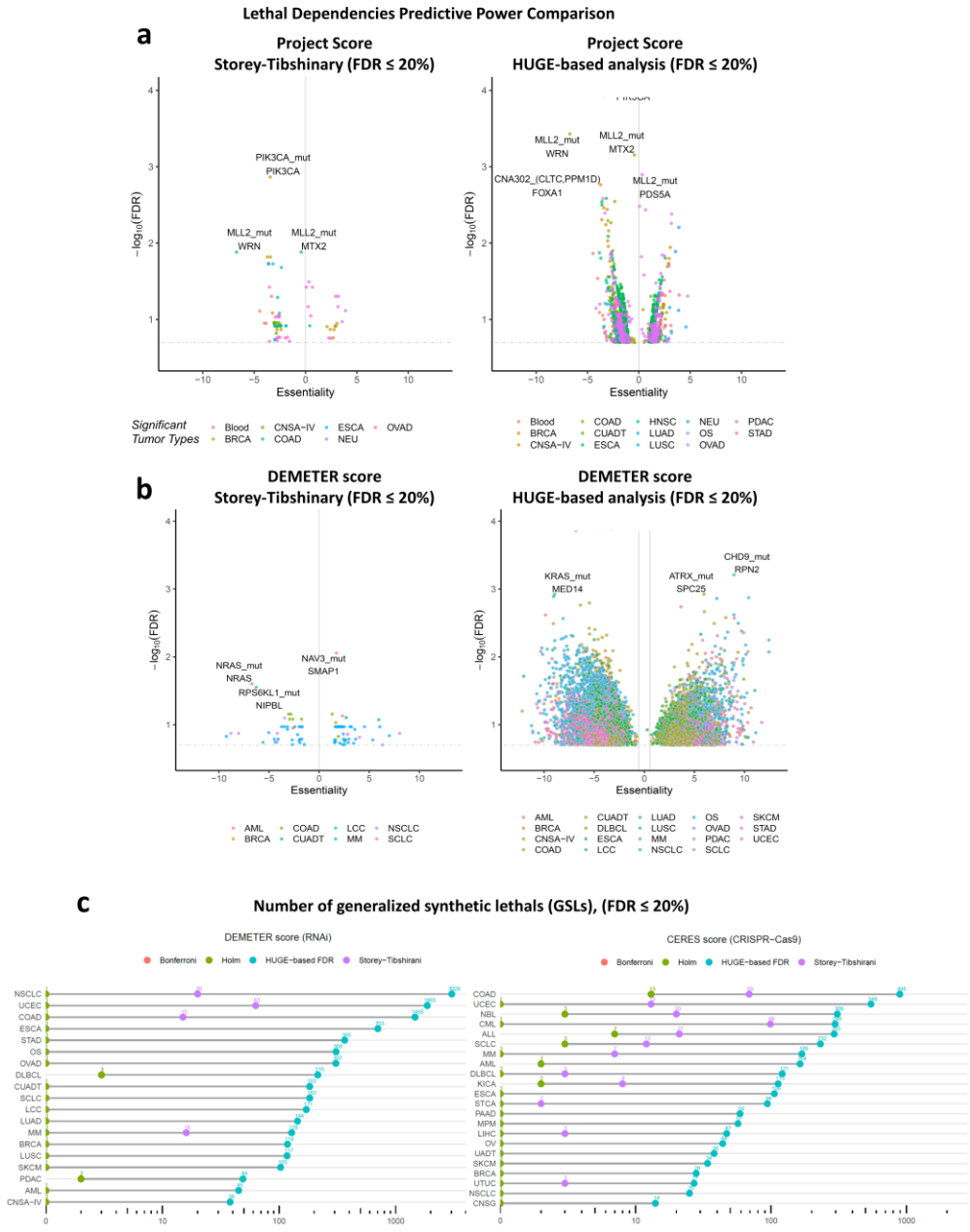


Figure 8. HUGE-based analysis with Project Score and Achilles Project datasets. (a) Volcano-plots of LEDs identified in the Project Score dataset. From left to right: i) result of Project Score, ii) results of analyzing Project Score dataset with HUGE-based methodology. Each dot represents a significant LED (FDR<20%). The X-axis represents the difference in gene essentiality when the event (gene variants) is present. The Y-axis represents the FDR values (-log10) for that change. (b)

Equivalent volcano-plots using Achilles Project. From left to right: i) results of Achilles Project analyzed with the standard procedure, ii) results of analyzing Achilles Project dataset with HUGE-based methodology. (c) The number of LEDs found (FDR ≤ 20%) in 19 tumors of the DEMETER score (RNAi) and 22 tumors of the CERES score (CRISPR-Cas9) using standard statistical pipelines (Storey-Tibshinary, Bonferroni, and Holm) and the HUGE-based algorithm. Bonferroni and Holm return the same number of hypotheses in all cases. LEGEND: ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia; BRCA: breast ductal carcinoma; CNSA-IV: central nervous system astrocytoma grade IV; COAD: colon adenocarcinoma; CUADT: upper aero-digestive tract squamous cell carcinoma; DLBCL: diffuse large B-cell lymphoma; ESCA: esophagus squamous cell carcinoma; KIRC: kidney renal clear cell carcinoma; LCC: lung large cell carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; MM: multiple myeloma; NSCLC: non–small cell lung carcinoma; OS: osteosarcoma; OVAD: ovary adenocarcinoma; PDAC: pancreas ductal carcinoma; SCLC: small cell lung carcinoma; SKCM: skin carcinoma; UCEC: endometrium adenocarcinoma.

As a further test of the increase predictive power of HUGE we carry out a similar analysis using the CERES score, a *CRISPR-Cas9* experiment of 22 tumors also included in the Achilles Project. In this case, the number of significant pairs is enriched 14 times over the standard approaches (FDR ≤ 20%; **Figure 8C-right panel**).

LEDs predicted by HUGE have better validation rates than standard approaches

Validating a ranking of LEDs is not a simple task: it is desirable to have a gold standard of disease-specific list of validated target-biomarker associations. We selected as our gold standard The Variant Interpretation for Cancer Consortium (VICC) Meta-Knowledgebase [78,79]. This database integrates different datasets of clinical associations and includes the level of evidence for each entry: spanning from professional FDA guidelines to preclinical findings.

We tested the enrichment in associations included in VICC in four tumor types, namely acute myeloid leukemia (AML), breast cancer (BRCA), lung adenocarcinoma (LUAD) and colon adenocarcinoma (COAD) for both HUGE and standard statistical methods. The VICC knowledgebase integrates (in September 2021) 19,551 clinical interpretations of somatic genomic variants in cancer of both resistant and sensitive biomarkers. We deleted duplicated and incomplete associations, focused on those related to confirmed mutations and manually selected associations that match each tumor type (including synonyms).

We first run the two procedures (HUGE and Storey-Tibshirani; ST) with AML cell lines to find LEDs and compare how many LEDs predicted by HUGE and by ST are included in the VICC knowledgebase. For instance, if HUGE or the ST procedure predicts *FLT3* mutant AML genotypes to be sensitive to *FLT3* inhibition, it will be considered a true positive LED, as *FLT3* is a well-known target of AML and mutations in *FLT3,* the fms-like receptor-type

tyrosine-protein kinase [85,86], are known to be sensitive biomarkers of the effectiveness of most *FLT3*-inhibitors [87,88].

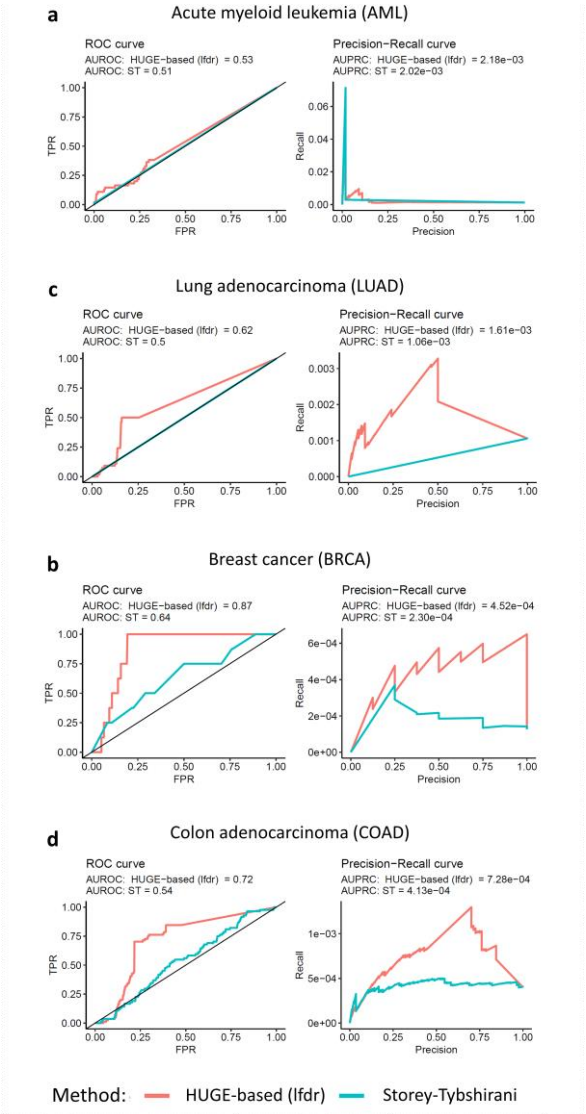


Figure 9. ROC and precision-recall curves of four tumor types. True positives were found somatic genomic variants in the knowledbase of the Variant Interpretation for Cancer Consortium (VICC). a) AML, b) BRCA, c) LUAD and d) COAD. We selected associations indicated for each tumor type that are within the three highest levels of confidence (Level A: Evidence from professional guidelines or FDA-approved therapies relating to a biomarker and disease; Level B: Evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus; and Level C: Evidence for

therapeutic predictive markers from case studies, or other biomarkers from several small studies, or evidence for biomarker therapeutic predictions for established drugs for different indications).

In total, 216 out of 19,551 associations matched these filters. Getting the top 500 LEDs according to the ranking using the HUGE algorithm with AML, we find 17 LEDs that match the VICC knowledgebase of known clinic relationships (**Table 1**; Fisher p-value < 1e-51). An equivalent analysis using the standard pipeline (ANOVA test [84] with the Storey-Tibshirani p-value correction) shows that out of the top 500 LEDs, only 1 is included in the VICC knowledgebase (**Table 1**; Fisher p-value = 6.551e-3). This means that HUGE analysis identifies 16 true positive dependencies not recovered by ST (Fisher p-value = 6.41e-5). The global value of AUROC (0.53) is not too far from the baseline of 0.5 (**Figure 9A**), perhaps because of the scarcity of true positives in our gold standard. We performed the same analysis with LUAD, BRCA and COAD getting AUCROC values of 0.62 (vs 0.5), 0.87 (vs 0.64) and 0.72 (vs 0.54) for HUGE and ST respectively. All cases show better values for HUGE than for ST (**Figure 9A to D**).

Table 1. Associations within the top 500 pairs predicted using the HUGE-based and standard pipeline algorithms in AML that match the knowledgebase of clinical interpretations of somatic genomic variants in cancer of the Variant Interpretation for Cancer Consortium (VICC).

Essential Gene	Biomarker Gene	Difference Essentiality	P-Value	Local Fdr	Method
NRAS	NRAS	-6,83	4,67E-08	1,38E-04	HUGE
FLT3	FLT3	-6,36	2,28E-04	2,00E-01	HUGE
TACR2	NRAS	4,71	9,21E-03	3,07E-01	HUGE
SH2D1A	NRAS	-4,96	9,74E-03	3,14E-01	HUGE
APBB1	FLT3	-2,87	5,54E-03	3,89E-01	HUGE
FGF18	NRAS	2,58	1,62E-02	3,89E-01	HUGE
FLNA	NRAS	4,53	1,85E-02	4,13E-01	HUGE
IL12RB1	NRAS	2,62	1,87E-02	4,15E-01	HUGE
FGF13	NRAS	3,01	2,10E-02	4,37E-01	HUGE
CD1C	FLT3	-3,21	9,22E-03	4,66E-01	HUGE
PPP4C	NPM1	3,55	1,35E-03	4,78E-01	HUGE
FGF13	FLT3	-3,09	1,09E-02	4,96E-01	HUGE
CCR7	FLT3	-3,74	1,12E-02	5,01E-01	HUGE
GATA6	FLT3	-3,43	1,18E-02	5,11E-01	HUGE
TYMS	FLT3	-4,22	1,21E-02	5,15E-01	HUGE
SRSF2	NRAS	3,86	3,23E-02	5,31E-01	HUGE
CCND3	NRAS	-2,62	3,35E-02	5,40E-01	HUGE
NRAS	NRAS	-6,70	1,48E-08	2,49E-02	ST

Applying HUGE methodology to acute myeloid leukemia cell-lines discovers potential therapy biomarkers

AML is a hematologic neoplasm characterized by a remarkable phenotypic and genomic heterogeneity [89], a challenging disease model to test the applicability and impact of HUGE. We run the complete HUGE pipeline with AML and validate *in vitro* two of the predicted LEDs.

As a preliminary step, we identified the potential genes that are essential for AML cell survival. The Achilles Project yield 443 essential genes that are essential and specific for AML cells compared to other tumors. Some of these genes belong to pathways known to be deregulated in AML (e.g., *MYB* [90] or *CEBPA* [91]). Interestingly, 160 of these 443 genes have previously been identified as potential cancer drivers in hematological malignancies according to the Candidate Cancer Gene Database (p-value = 7.76e-05, Fisher exact test) [92].

We then run the HUGE algorithm to identify genomic alterations that could be defined as LED partners of those 443 essential genes. In this pipeline, we required predicted pairs to be biologically related to each other in the STRING database[82] (co-expressed, annotated in the same pathway database or contained in a common experiment). LED associations can be broken down into three groups regarding their dependency type: *positive lethal dependency (pLED)*, when a gene variant marks sensitivity to the inhibition of another gene; *negative lethal dependency (nLED)*, when a gene variant marks resistance to the inhibition of another gene; or *dual lethal dependency (dLED)*, when the same gene variant confers, concurrently, sensitivity to the inhibition of one gene and resistance to the inhibition of another gene (**Figure 3**). In total, we predicted 24 LEDs, (12 *pLEDs* and 12 *nLEDs*, including 2 *dLEDs*; p-value < 0.05, local FDR ≤ 0.6 and |ΔEssentiality| > 2; **Figure 10A, Table 2**). Using the standard multiple hypotheses correction only 1 dependency turns out to be statistically significant. We provided the identified LEDs for the 19 tumors included in the Achilles Project following a similar pipeline.

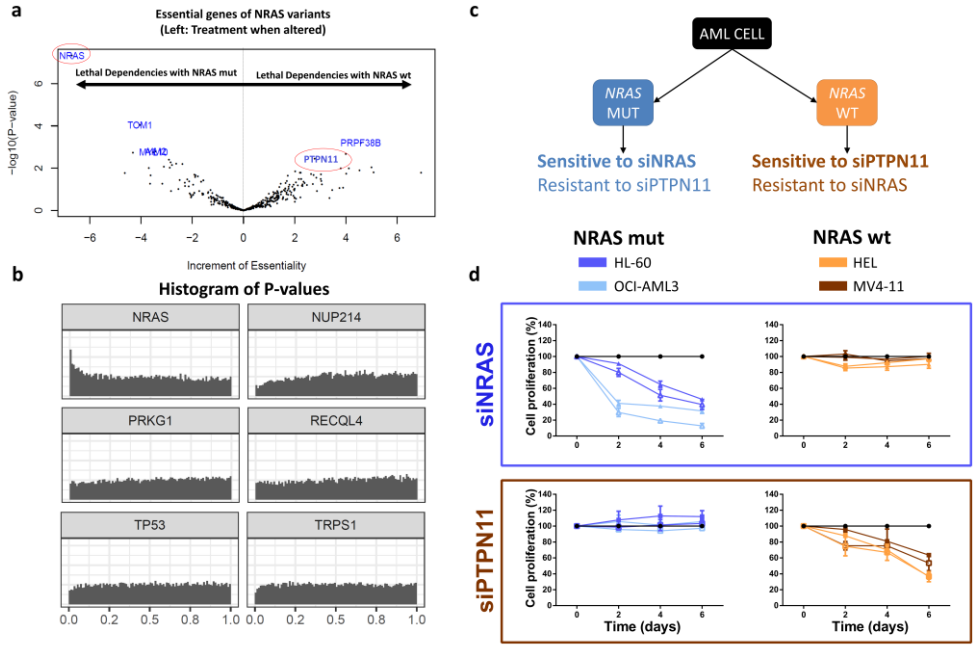


Figure 10. Gene variants-based treatment **guidelines in acute myeloid leukemia (AML).** (a) Volcano-plot of LEDs related to NRAS genetic mutations (left; MUT) and -wildtype (right; WT) phenotypes. Increment of Essentiality and -log10 (p-value) are shown in X-axis and Y-axis, respectively. (b) Histogram of p-values for 6 genetic sequence variants in AML. NRAS-alteration is enriched in close to zero p-values, which is the basic concept of HUGE-based statistical approach. All genetic variants histograms of p-values can be found in the Supplementary Material. (c) Summary of the computational predictions validated: NRAS-altered cells were predicted to be sensitive to siNRAS and resistant to siPTPN11. Conversely, NRAS-wt cells were predicted to be sensitive to siPTPN11 and resistant to siNRAS. (D) Tumor proliferation of the four AML cell lines after inhibiting NRAS (siNRAS) and PTPN11 (siPTPN11) with specific siRNAs. Blue: NRAS-altered AML cell lines (HL-60 and OCI-AML3); Orange: NRAS-wild-type AML cell lines (MV4-11 and HEL).

NRAS mutation ranks first in the analysis. Lethally dependent partners associated with *NRAS* genetic sequence variants show a p-value histogram that peaks at the origin (**Figure 10A and B**), meaning that *NRAS* mutations are associated with more tumor vulnerabilities than other alterations. Interestingly, *NRAS* alteration forms a *Dual Lethal Dependency* with *PTPN11* (**Table 2, Figure 10C**): it confers tumor sensitivity to *NRAS* inhibition and resistance to *PTPN11* inhibition.

To validate our prediction, we first checked that both *NRAS* and *PTPN11* siRNAs efficiently decreased the *NRAS* and *PTPN11* expression, respectively, in four AML cell lines (**Figure 11**). Then, we confirmed the computational hypothesis: the downregulation of *NRAS* significantly decreases cell proliferation only in the *NRAS*-altered AML cell lines, and the inhibition of *PTPN11* expression produces an equivalent effect, but specifically in the *NRAS*-wt AML cell lines (**Figure 10D**), validating the predicted *dLED*. Remarkably, the validated *PTPN11*-*NRAS*-wt pair was not detected using standard methodologies.

Table 2. Ranking of Lethal Dependencies in AML using the HUGE-based statistical approach. The ranking is divided into three groups regarding the typology of the lethal dependency relationship: Positive Lethal Dependency, Negative Lethal Dependency or Dual Lethal Dependency (Figure 1). The Increment of Essentiality column represents the average variation in the DEMETER score between altered and wild-type cells, and its sign is related to the lethal dependency relationship. Lethal dependencies that share the same essential gene and the same Increment of Essentiality sign have been omitted in this table.

Gene variant Biomarker	Essential Gene	Increment of Essentiality	t-score	P-value	Local FDR
Positive Lethal Dependency					
TGS1	SNRPF	-7,87	-4,05	6,69E-04	3,36E-01
CLTCL1	UBR5	-6,66	-3,59	1,99E-03	2,20E-01
FLT3	FLT3	-6,36	-4,53	2,28E-04	2,00E-01
CDK14	CDK2	-3,95	-2,75	1,28E-02	4,30E-01
AURKC	ACTL6A	-3,26	-3,89	9,55E-04	4,99E-01
Negative Lethal Dependency					
NPM1	EEF2	3,81	3,34	3,39E-03	5,96E-01
PIK3C2G	CDK6	3,35	2,95	8,20E-03	3,51E-01
NCOA3	EP300	3,04	2,75	1,25E-02	4,94E-01
CDK14	CCND2	2,97	2,22	3,88E-02	4,99E-01
EPHB6	ZNF266	2,53	2,77	1,22E-02	3,42E-01
ZFYVE9	TOM1L2	2,14	2,35	2,96E-02	5,12E-01
Dual Lethal Dependency					
NRAS	NRAS	-6,83	-8,71	4,67E-08	1,38E-04
NRAS	PTPN11	4,17	2,2	4,05E-02	5,89E-01
EP300	PLK1	-8,11	-4,04	7,01E-04	2,17E-01
EP300	KLF2	3,69	4,08	6,38E-04	2,12E-01

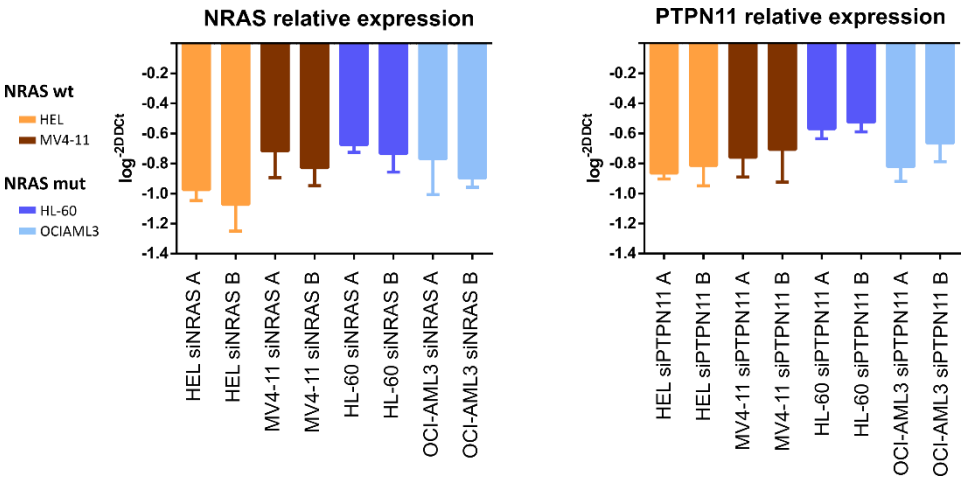


Figure 11. mRNA expression of *NRAS* and *PTPN11* genes after nucleofection with the specific siRNAs. Data are referred to *GUSB* gene and an experimental group nucleofected with negative control siRNA.

The HUGE method was also used in another publication in which several members from our research group proved and validated known LEDs from literature. In addition, they performed an exhaustive study of the LEDs detected by HUGE in SCLC and successfully validated *in-vitro PLK1* essentiality when *CREBBP* was mutant [88] (**Figure 12**).

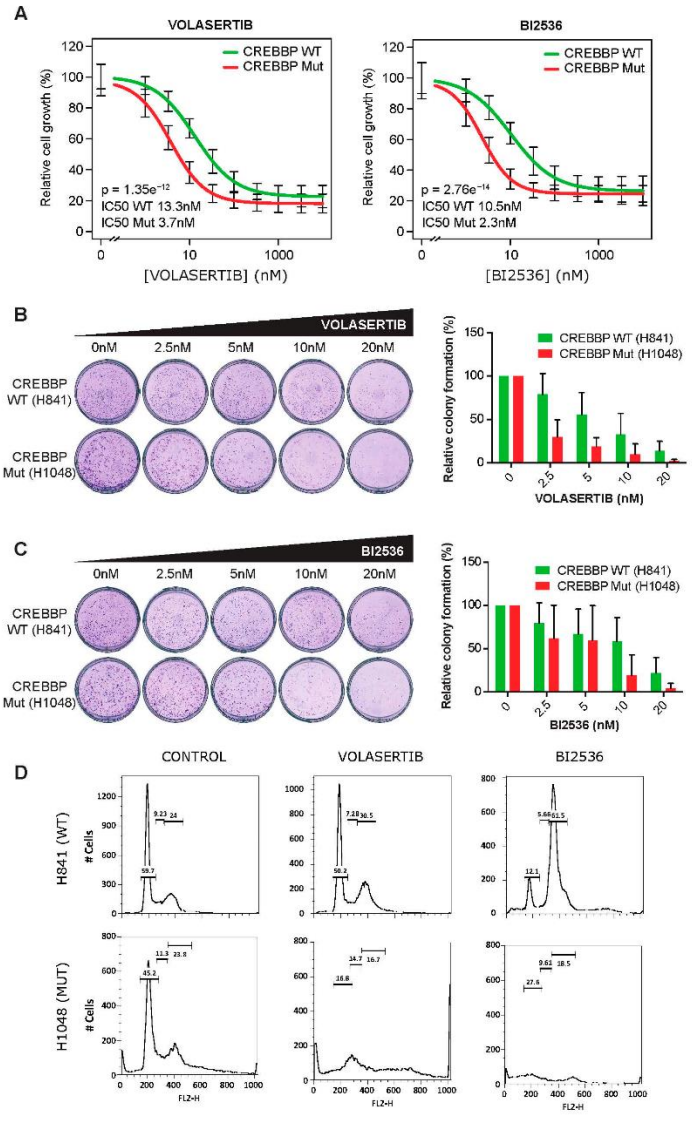


Figure 12: In vitro validation of the sensitivity of CREBBP-mutant SCLC cell lines to two PLK1 inhibitors: Volasertib and BI2536. (A) Dose–response curves showing the effect of Volasertib and BI2536 treatment on the viability of CREBBP-WT NCI-H841, NCI-H889, NCI-H2171, NCI-H146 cells, and CREBBP-MUT NCI-H1048, NCI-H1963, NCI-H211, HCC33 cells. Cells were treated with the indicated doses for 72 h. Cell viability was measured using the cell viability (MTS) assay and the IC50 was calculated for each cell line. (B) Colony formation assays of NCI-H841 and NCI-H1048 cells. Cells were seeded onto a six-well plate and were treated with vehicle (0.1% DMSO) or increasing doses of Volasertib or BI2536 for 72 h. After treatment, cells were incubated in a drug-free culture medium for 14 days, fixed and stained with crystal violet. (C) Quantification of the number of colonies obtained in each condition with Fiji software. (D) FACS cell cycle analysis of NCI-H841 and NCI-H1048 cells conducted upon 5 nM Volasertib and BI2536 treatment for 24 h.

Discussion of Section 1

The advent of large-scale functional genomic screens has allowed the identification of hundreds of novel gene targets and the prediction of genome-wide LEDs [17,93]. This strategy has multiplied treatment strategies, as using LEDs, the drug targets can be decoupled from their corresponding predictive biomarkers. The main statistical limit to find LEDs is the large number of hypotheses that result from integrating gene essentiality and genetic functional events. In this section, we present HUGE, a novel analysis of *CRISPR-Cas9* and RNAi large-scale screens that significantly improves the predictive power to find LEDs from loss-of-function screens in human tumors. It relies on the fact that some gene alterations are statistically related to the essentiality of large sets of genes. Using this characteristic as a prior covariate we significantly improved the predictive power of LEDs.

Notably, the presence of the HUGE effect does not necessarily mean biological causality. HUGE dependencies are more statistically reliable than others, but this does not imply that predicted alterations are the major players in tumor development thus, they are not necessarily driver genes; i.e., they are just genetic biomarkers of gene essentiality. In other words, the Hub-Effect is a statistical association. Since "correlation does not imply causation" is not legitimate to deduce a cause-and-effect relationship between the presence of a mutation and the sensitivity to knocking down a gene. Even more, it cannot be concluded that the HUGE top-ranked genes (either the mutations or the knockdown genes) are driver genes. This would require further experimentation and validation. HUGE simply computes biomarkers of the vulnerability to a knockdown gene, that in turn, could be targeted by a drug. However, the fact that gene alterations co-occur with multiple LEDs in genetic hubs can be exploited to improve the statistical power.

To measure the increased predictive power of HUGE, we carry out three different comparisons within three functional genomic datasets: the Project Score, the DEMETER score and the CERES score. HUGE identifies LEDs with 14 and 75 times larger statistical power than using state-of-the-art methods in *CRISPR-Cas9* and RNAi, respectively. However, it could be argued that this result could be an artifact of the statistical technique and that −lowering the threshold for standard procedures− would provide LEDs with similar reliability. This is not the case. As shown in the results, using the same number of predictions, HUGE's results are more enriched in clinically validated biomarkers than ST's results. Remarkably, one of the 16 LEDs only identified by HUGE is the known interaction

of *FLT3*-mutant genotypes sensitive to *FLT3* inhibitors, such as Midostaurin. This fact is only an example of the key importance of considering the HUGE effect when analyzing LEDs with large-scale functional screens.

A p-value histogram can be modeled as the superposition of two distributions, a uniform distribution (which corresponds to the null hypothesis) and another distribution with a larger proportion of low p-values. A good covariate splits the overall p-value histogram into histograms with different enrichments in small p-values. If all the histograms related to a covariate have similar shapes, it means that the covariate is uninformative. Here, we show that stating which gene is mutated in each test is a good covariate for the LED prediction problem because there is a hub effect of gene aberrations in gene essentiality. The usage of covariates has successfully been incorporated before in other genomics applications (e.g., the abundance of a gene is known to be informative in differential expression analyses; or the proximity of loci in the genome is known to play a role in genome-wide association studies), but it has not yet been exploited in large-scale functional genomic screens.

One main limitation lies in the volume of data required for its execution due to the need for multiple hypotheses to detect the Hub-Effect. Hence, the HUGE-based approach will not obtain such striking results if applied to the analysis of smaller experiments in number, it would perform similarly to current standard methods. Nevertheless, this method has been developed for large-scale screening analyses.

We are confident that the HUGE-based approach to calculating LEDs has great potential if applied to the study of patient data. Nowadays, drug development usually starts from large-scale loss-of-function screenings. Therefore, this work has identified a large number of LEDs across 19 tumor types in 3 different large-scale experiments. Besides, to facilitate the *in-vitro* validation of these LEDs as possible therapeutic targets, we have added information regarding targeted drugs for those essential genes that are drug targets

Predicting true LEDs is especially challenging for tumors with high genetic heterogeneity. In AML, for instance, state-of-the-art approaches only recover 2 LEDs. The HUGE-based approach captured 24 LEDs for the same False Discovery Rate (FDR). Interestingly, *NRASwt-PTPN11* LED, which was only identified by HUGE, has been validated *in vitro*. The validation in AML highlights the potential of the HUGE-based approach to discover and validate new LEDs of biomarkers and drug targets. We pinpoint the *dLED* characteristic of the *NRAS* gene, meaning that if a tumor has *NRAS* mutated a treatment that targets *NRAS*

itself would be the best option to reduce their tumorigenicity, whereas if it is *NRAS* wild-type, a *PTPN11* inhibition would be a better recommendation. This *dLED* discovery confers special relevance to clinically translational therapeutic strategies, as it has been proved effective in AML cell lines, further validation in *ex-vivo* analysis and murine models is required but if resulting effective, it could be suggested as a treatment and it could incentivize drug development targeting *NRAS* and *PTPN11*. This methodology has potential applications both in basic and clinical research.

Section 2: Interpretable Artificial Intelligence for Precision Medicine in Acute Myeloid Leukemia

Novel Interpretable Method for Precision Oncology

Section 2: Interpretable Artificial Intelligence for Precision Medicine in Acute Myeloid Leukemia

Introduction to Section 2

Drug sensitivity studies have helped personalized medicine evolve, particularly in the field of precision oncology. Combining these drug sensitivity studies with tumor genotypes makes it possible to associate the response to treatment with genetic alterations (biomarkers), thus promoting the search for new personalized therapies[18].

Exploring the potential of these experiments, artificial intelligence (AI) algorithms for personalized medicine focus on the analysis of such datasets to bridge the gap for drug discovery. However, the concept of “black box" in AI limits the potential of this approach to be translated into the clinical practice. In contrast, Interpretable AI focuses on making AI results understandable to humans.

In this section we include the development of a new Interpretable AI method, called Multi-dimensional Module Optimization (MOM) algorithm, to predict therapeutic strategies based on large-scale drug screening data. This method systematically associates drugs with combined sets of genetic biomarkers that can be generalized and applied to other cohorts of patients. The therapeutic strategies provided by MOM can easily be understood by humans and are easy to implement in the clinical practice with a process equivalent to a decision tree. The optimization problem considers the effect of drug toxicity focusing on providing drugs that are differentially effective to patients with a specific genotype. MOM’s result is deterministic –this is important to get regulatory approvals– and guaranteed to be optimal, the overall sensitivity of the patients is maximized.

We applied MOM to an AML cohort, the BeatAML project cohort, which carried out WES (Whole Exome Sequencing) and drug screening experiments of 122 drugs with *ex-vivo* AML tumor samples from 319 patients [12]. Ex-vivo experiments in hematological cancers are of great importance since they are performed directly on the patient’s living tumor cells [11,12], allowing to correlate drug sensitivity to the patient’s genotype. The results obtained using MOM are *in-silico* validated using K-fold cross-validation and in three independent large-scale experiments, one based on pan-cancer drug sensitivity and two referred to pan-cancer gene essentiality using siRNA and *CRISPR*-cas9. MOM’s patient indications require only three different biomarkers, which makes them to be easily understood by the clinician.

MOM is published in [2].

Chapter 4. Results

An interpretable artificial intelligence method to predict optimal treatments based on patient genotype

The implementation of a clinical translational interpretable AI model requires the development of a robust method to associate biomarkers to specific targeted treatments. and, thus, relating drug sensitivity and patient genetic events -including SNVs, indels, fusion genes, or even epigenetics. The development of an AI algorithm in this context requires to solve three important challenges: (i) proper modeling of the toxicity of screened drugs (most aggressive drugs are not necessarily better treatments), (ii) dealing with a high number of statistical hypotheses that intrinsically increase false discovery rate (FDR), and (iii) explaining the internal reasoning that the model uses to propose a decision so that it is easy to approve and implement in the clinical practice.

We propose an algorithm named Multi-dimensional Module Optimization (MOM) that addresses each of these challenges by dividing the problem into three main steps (**Figure 13**): preprocessing the input drug sensitivity scores, associating single biomarkers to drugs with an increased statistical power and combining individual treatments to unveil multi-step treatment pipelines to stratify patients based on drug-response.

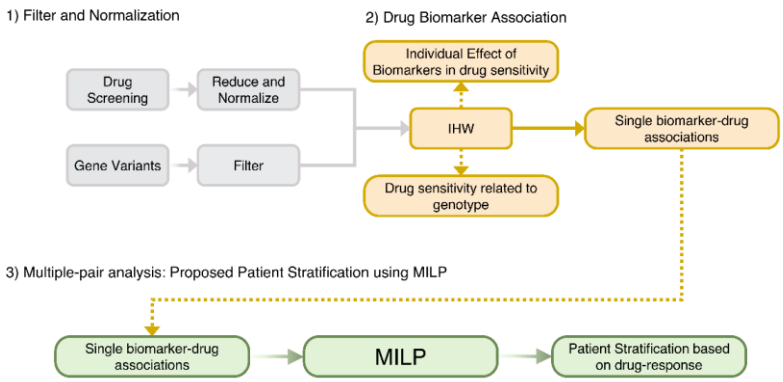


Figure 13. Overview of MOM's pipeline. (1) Filter and Normalization. (2) Generate individual Drug-Biomarker Associations using IHW, (3) Multiple-pair analysis that generates a patient stratification

guideline using a novel MILP model (IHW: Independent Hypothesis Weighting, MILP: Mixed-integer Linear Programming).

MOM is developed to optimally stratify patients following a decision tree based on simple logical rules, in which each step is defined by the presence or absence of a certain biomarker and the recommendation of one drug. In turn, MOM requires genetic variants information and drug sensitivity screenings as input data.

To illustrate the steps of the algorithm, let us consider a toy example with 8 drugs and their corresponding drug-response scores for 6 patients (**Figure 14**). In this case, as in every precision medicine scenario, we want to find robust companion biomarkers that, associated to drugs allow us to maximize patient response with minimized toxicity.

In the first step, MOM preprocesses drug sensitivity scores (**Figure 14.1**). For which, instead of using the standard measure of IC_{50}, we proposed an incremental version of the logarithm of the IC_{50}, named IC50* (See **Chapter 5** for more details). The proposed correction has two main advantages. First, MOM prioritizes drugs that have a differential effect on different patients, which are, in turn, better candidates to develop a personalized treatment based on a companion biomarker. Second, drugs whose effectiveness does not depend on patient genotype are more unspecific and, therefore, more prone to be toxic for different tissues. In the next section, we will illustrate this fact with a real case scenario.

To exemplify this normalization, let us return to the toy example with 6 patients, 8 drugs and their corresponding log(IC_{50}) scores measured in ex-vivo tumors (**Figure 14.1**). Considering raw log(IC_{50}) exclusively (left-hand heatmap), it could be argued that Drug 1 is the most effective drug and, therefore, it should be indicated to all patients regardless their genotype. However, since the dose can be adjusted for each patient, Drugs 1 and 8 will be given at a small and a large dose respectively balancing their effect. Using IC50* (right-hand panel) allows MOM to maximize the genetic dependence of drugs, rather than the absolute cellular death in patient tumors.

In the second step (**Figure 14.2**), MOM provides single biomarker-treatment associations by prioritizing the drugs whose response is associated with patient genotype. The selected statistical analysis to find the biomarker-treatment associations is the Independent Hypothesis Weighting (IHW) algorithm. This algorithm has been proved to increase the power of tests in several biological scenarios [1,94].

This algorithm provides also two interesting “by-products”: i) identifies which biomarkers are related to drug sensitivity, e. g. *TP53* mutation is usually a source of resistance, ii) identifies drugs whose efficacy is related to the genetic profile, Olaparib is effective only for $BRCA^{Mut}$ patients [95].

In the third step (**Figure 14.3**), MOM predicts a sequential treatment guideline that maximizes the drug effect on the group of patients that share the genotype of the selected biomarkers. Using Mixed integer Linear Programing (MILP), MOM gets the optimal treatment guideline (decision tree). MILP is a versatile optimization method that allows the solution of complex mathematical problems using integer variables and assures that the drug assignation is optimal. This solution (i) is interpretable; (ii) eases the translation into clinical practice; and (iii) assures a global and deterministic optimum to the problem.

MOM Pipeline

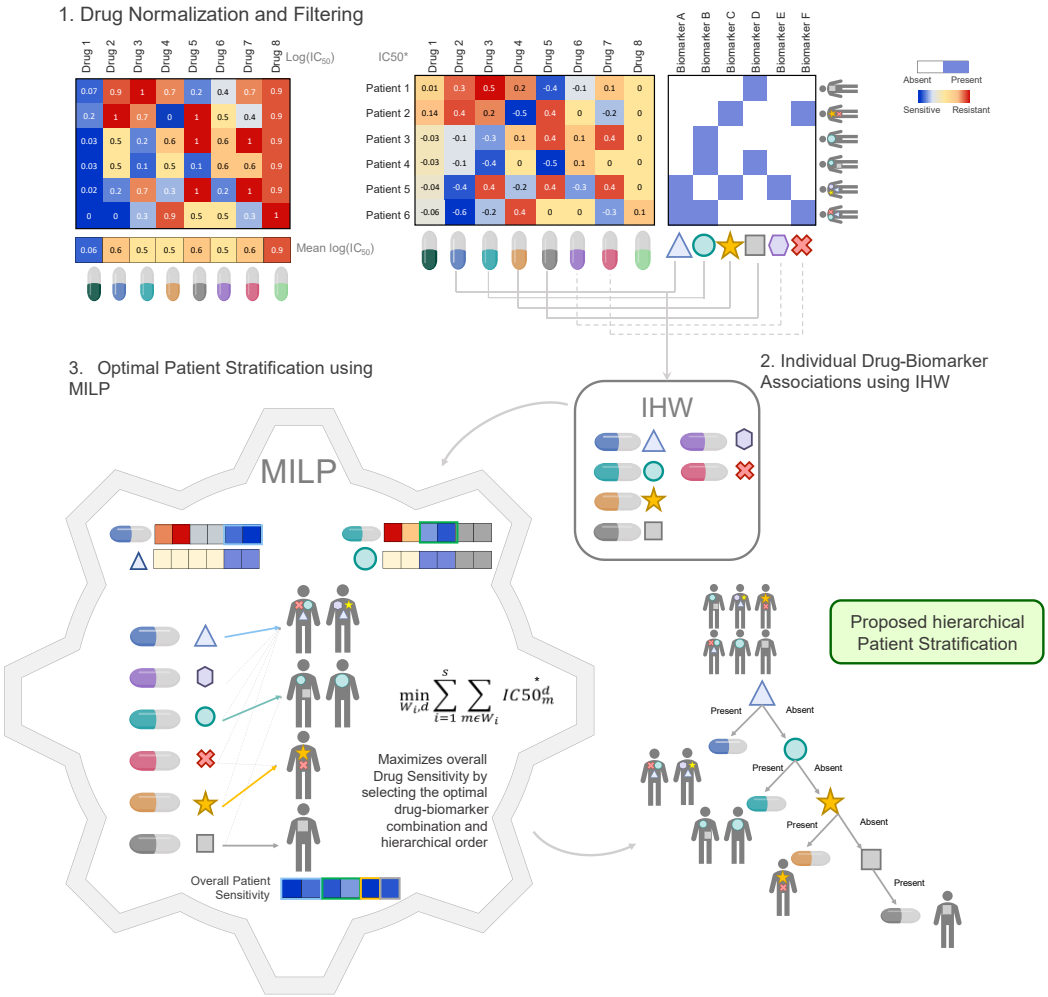


Figure 14. MOM Pipeline: MOM pipeline is defined by 3 major steps: 1) Drug normalization to reduce drug toxicity. It is performed by removing drug mean effect in all patients. The blue color represents drug sensitivity for the sample whereas the red color represents drug resistance in the sample. 2) Individual Drug biomarker associations using IHW. Drugs are matched to biomarkers profile, all individual associations generate a p-value that is corrected using IHW. IHW selects the candidate biomarkers and treatment and are used as input to the MILP problem. 3) Optimal Patient Stratification using MILP. The MILP module receives as input the normalized drug information IC50*

and the candidate individual associations and outputs a decision tree for clinical decision-making guidance. Within this module, the treatment is optimized so that each patient receives the drug for which is more sensitive. (IHW: Independent Hypothesis Weighting, MILP: Mixed-Integer linear Programing).

FLT3, *CBFβ-MYH11*, and *NRAS* variants play a key role in Acute Myeloid Leukemia sensitivity to Quizartinib, Trametinib, and Selumetinib.

We selected the BeatAML cohort to test MOM as it contains *ex-vivo* drug sensitivity screenings of 122 drugs in AML tumors derived from 319 patients [12], and includes both whole-exome sequencing experiments (WES) and drug sensitivity for every patient. Analyzing the WES data, we described the genetic landscape of the cohort shown in **Figure 15** and **Appendix 2**. Patients within this cohort are in different therapeutic stages, e.g., induction, maintenance, consolidation, or palliative care (among others), there also are 32 *de novo* patients (**Figure 16**).

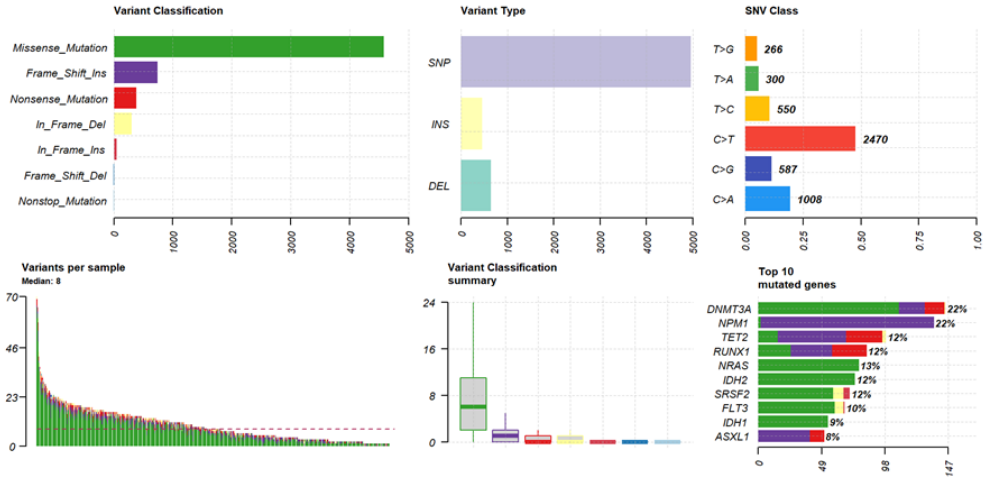


Figure 15. Genetic Variant Type Summary in BeatAML. Variant classification plot represents the different genetic variant types, in terms of functionality and we see that the most common is missense variant, the Variant type plot represents the type of structural variant, whether they are Single Nucleotide Variants (SNVs) or Indels with a clear predominance of SNVs, SNV classification plot shoed the type of mutational signature that is predominant with the signature C>T that is quite frequent in malignant cancer types followed by C>A which is associated with environmental exposure[96]. The Variants per Sample Plot, tells that there is a median of 8 variants per patient. The variant classification summary plot summarises all the prior plots and, finally, the Top 10 mutated genes plot, shows for each of the top 10 genes the type of mutation.

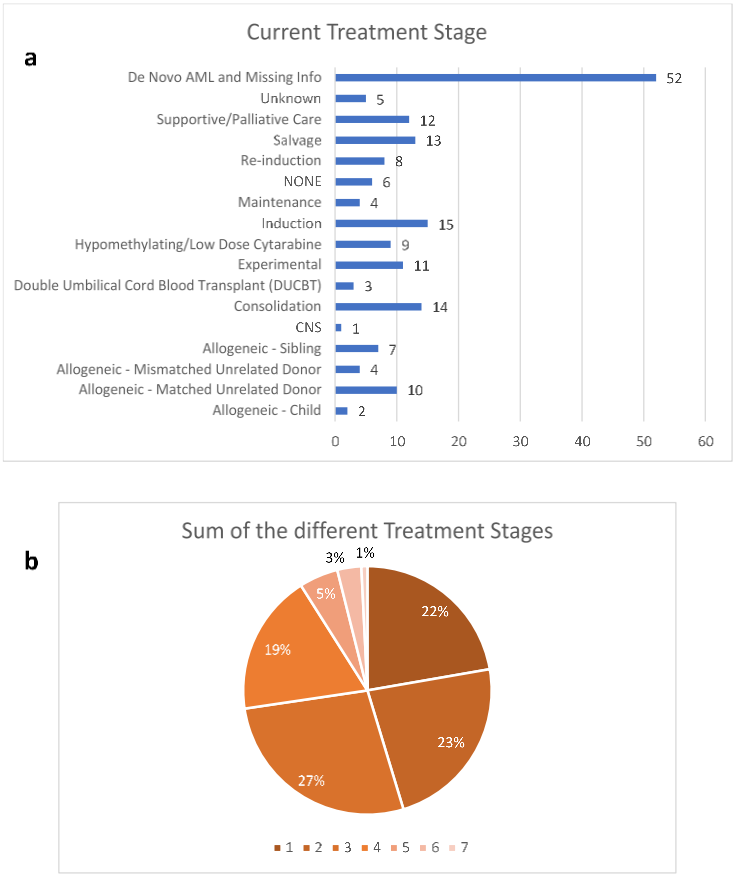


Figure 16. Treatment Stages in the BeatAML cohort. A) From the 319 patients contained in the study coming from the BeatAML cohort, the chemotherapy phase veries considerably. B) Distribution of the patients accounting for the different cumulative chemotherapy phases excluding the patients with missing info or who did not take any chemotherapy.

The drugs studied in the BeatAML cohort cover a wide variety of different cancers and diseases: 24% are indicated for AML, 16% for other leukemias types, 10% for multiple myeloma, and 4% for lymphomas. This means that 54% of the drugs have been studied for hematological malignancies. The rest 46% include drugs used in lung, breast, or renal cancers among other diseases (**Figure 17**). Focusing on AML, the dataset provides a total of 11 AML drugs already in clinical use -e.g. Venetoclax, Quizartinib, or Gilteritinib- and 18 AML experimental drugs -e.g. Panobinostat, Lestaurtinib, or Pazopanib.

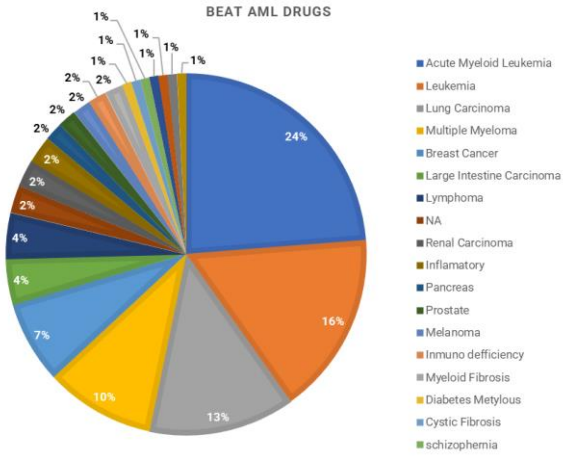


Figure 17. Beat AML Drugs distribution. Drugs studied in the BeatAML cohort cover a wide variety of different cancers and diseases, from which 24% are designed, experimental, or already prescribed for AML, 16% for other leukemias, 10% are shown concerning multiple myeloma, and 4% for different types of lymphomas. This means that 54% of the drugs have been studied for hematological malignancies.

We filtered gene variants to keep the ones that appear in at least 4 out of 319 patients (1%). This process provides 64 potential single biomarkers. We also removed drugs used in less than 20% of the patients, and those without a candidate gene target. After matching samples with *ex-vivo* and WES experiments, we finally get the *ex-vivo* screening of 111 drugs for 319 patients (see **Chapter 5** for more details). We then applied the MOM algorithm to this cohort to unveil groups of AML patients that share genotype and drug sensitivity. In the first step, MOM normalizes the IC_{50} values to define a score that better defines tumor sensitivity, namely IC50*.

Let us illustrate this with a paradigmatic example. In our dataset, the median IC_{50} for Elesclomol is much smaller than the median IC_{50} for Quizartinib (**Figure 18a,** left panel). Consequently, Elesclomol seems a better option to treat patients with AML. **Figure 18b** gives a completely different reading: Elesclomol is more toxic in almost any tissue if compared with the AML lines. On the contrary, Quizartinib is more toxic on AML than in most other tissues. This simple example shows that plain IC_{50} must not be used to select the treatment guideline for the patients. The higher value of IC_{50} for Quizartinib could be corrected by adjusting the dose. In **Figure 18a,** right-panel, after the normalization, the IC_{50}* for Elesclomol appears less effective, whereas Quizartinib preserves its sensitivity profile, which, in this example, it is related to the *FLT3* status of the tumor.

.

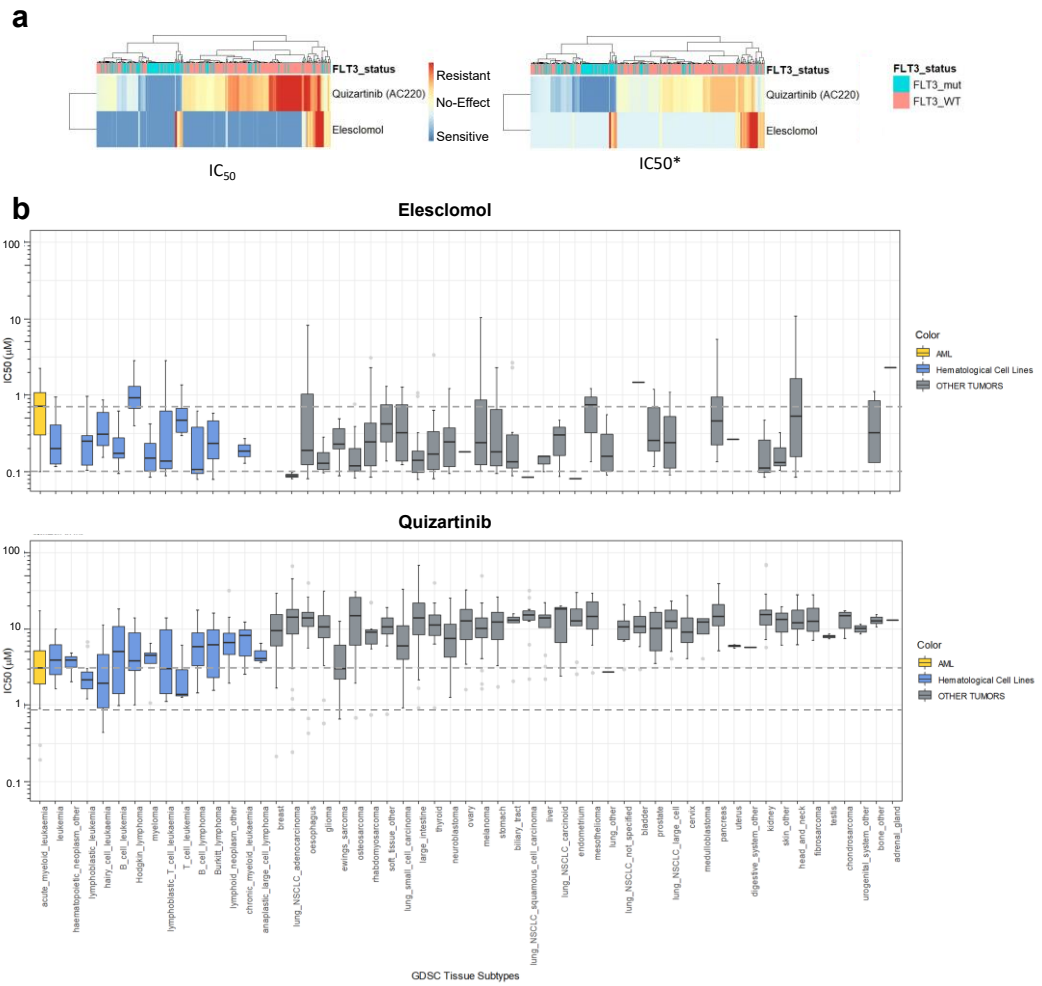


Figure 18. IC_{50} Normalization to Avoid Drug Toxicity. a) Drug Sensitivity Heatmap in BeatAML cohort. The left panel shows the IC_{50} values for AML tumors of BeatAML. Effectiveness of a drug in a patient is plotted in blue color, and resistance is represented in red color. The right panel shows the sensitivity in IC50* score. b) Drug sensitivity of Quizartinib and Elesclomol across different tissue types using GDSC. IC_{50} values relative to different tissues are shown in the graph. In yellow color are plotted the sensitivity values of AML cell lines, in blue color are plotted the drug sensitivity values for the Hematological cell lines, and finally in grey color, are plotted the sensitivity values for the non-hematological tissues from GDSC. Dotted grey lines represent the second IC_{50} quantile for AML cell lines (GDSC: Genomics of Drug Sensitivity in Cancer).

In the second step, MOM calculates individual associations between drugs and genetic alterations using the HUGE approach from IHW package [94]. This approach sheds light

on which drugs can be influenced by patient genotype (**Figure 19a**). IHW also provides a weight for each genetic variant related to the probability of such variant to be a true positive. Non-zero IHW weights represent genetic variants that reduce the FDR and increase the power of tests as demonstrated by IHW authors [94]. IHW estimates that, in our AML cohort, 37 biomarkers have weights greater than zero. IHW weights can be therefore used to state the relevance of each biomarker. We sorted IHW weights confirming that $FLT3^{Mut}$, $NPM1^{Mut}$, $NRAS^{Mut}$, $TP53^{Mut}$, and $KRAS^{Mut}$ are included in the top 5 biomarkers (**Figure 19b**), which have already been described in previous studies [97–102]. IHW also provides an adjusted p-value for each drug-biomarker association. For instance, the pipeline identified the known relation of *FLT3* internal tandem duplications (*FLT3-ITD*) patients being more sensitive to Sorafenib, Quizartinib, or Gilteritinib (**Figure 20**).

Interestingly, an indirect output of this second MOM step is the quantification of the sensitiveness or resistance triggered by a specific genetic variant. Summarizing this score, gene variants can be classified by their effect: either sensitive or resistant to the tested drugs (**Figure 19c**). For example, variants in *FLT3* or *NPM1* are associated with a more sensitive response for the cohort of drugs in this experiment, whereas genetic alterations in *KRAS*, *NRAS*, or *TP53* are more likely resistance-conferring. Other results include *CCND3*, *WDR52*, *CELSR2*, *CBFβ-MYH11*, and *SMC1A* as biomarkers of sensitivity and *STAG2* of resistance. This effect is relative to the studied dataset, Beat AML, and occurs across 66 different drugs studied or prescribed for hematological malignancies.

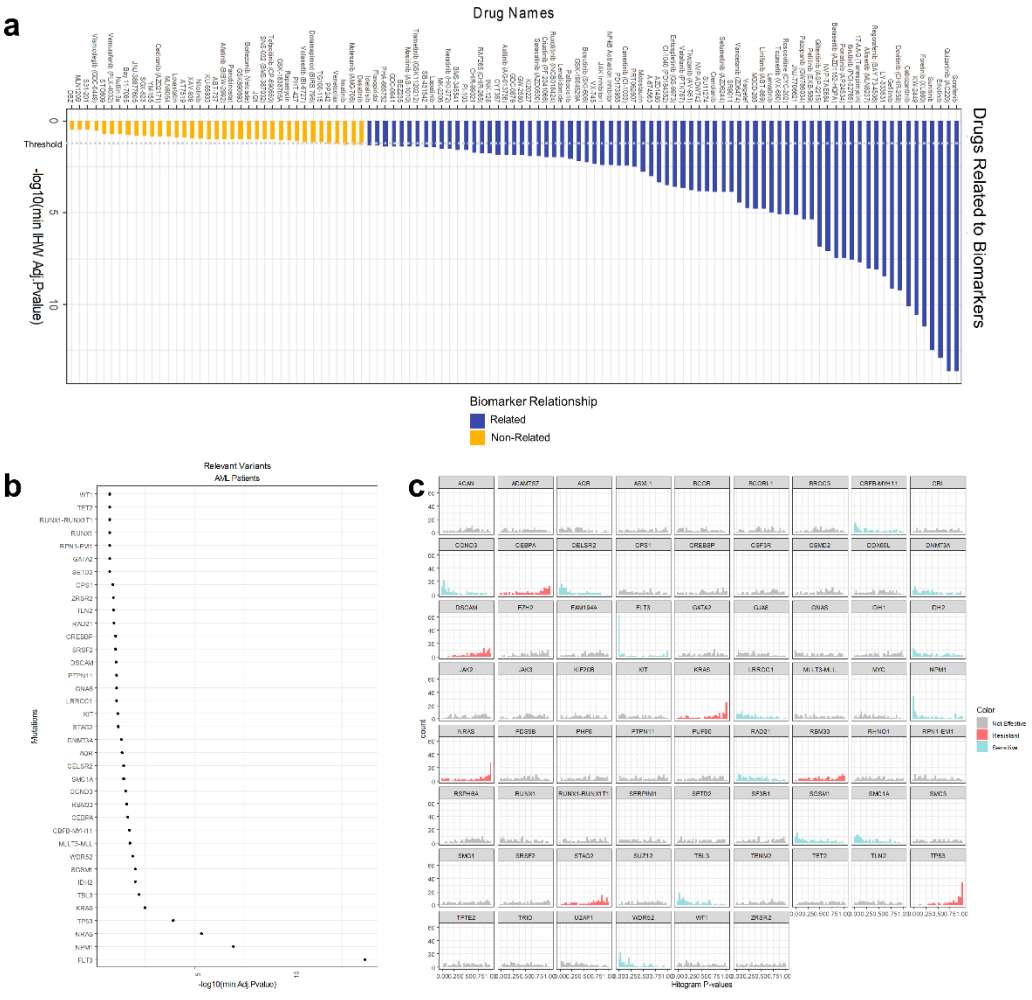


Figure 19. Analysis of single interactions biomarker-drug. a) Overall score of 122 drugs whose IC50* is related or non-related to cell genotype according to our model. A drug is related to a relevant variant (those whose IHW weight is greater than zero) if its adjusted p-value is below 0.05. b) Global effect of AML gene variants in AML drug sensitivity. The x-axis shows the logarithm of the minimum adjusted p-value of the biomarker with any of the drugs. Only those biomarkers whose IHW weight is greater than zero are shown. c) One-tail p-value histograms comparing drug sensitivity according to the biomarker status in AML. If a histogram has a strong peak near zero, patients with the biomarker are sensitive to many drugs. On the contrary, if a histogram has a strong peak near one, patients with the biomarker are resistant to many drugs. A genetic variant is considered to confer sensitiveness if the number of drugs whose p-value <0.2 is twice the number of p-values >0.2. Similarly, a variant

confers resistance if fulfills that the number of p-values>0.8 is twice the number of p-values<0.8. (IHW: Independent Hypothesis Weighting).

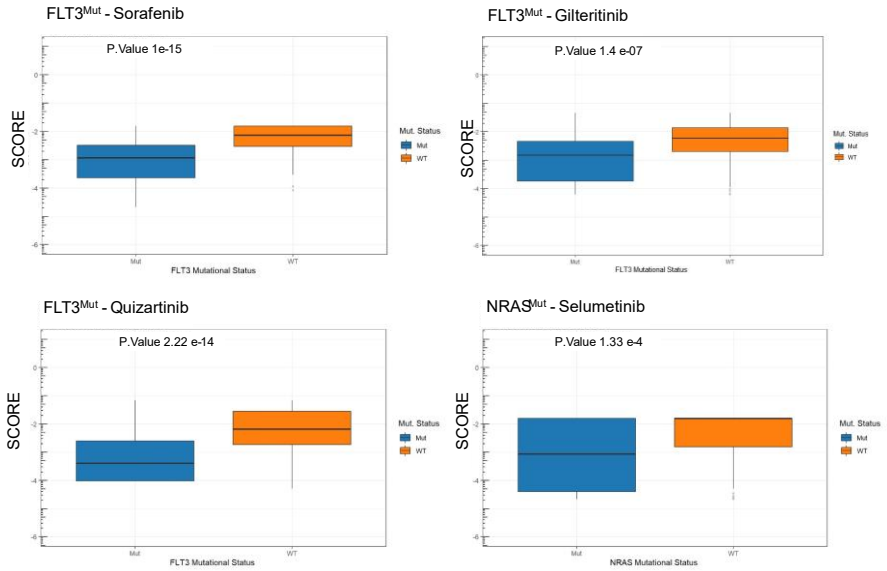


Figure 20. **Relevant Individual Drug-Biomarker Associations.** In blue the patients with the biomarker in orange the patients without the biomarker. P-value association is corrected using IHW. The score shows the differential IC_{50}, more negative means more sensitive.

Finally, in the third step, we solved the MILP problem from MOM using the individual candidate associations. As a result, MOM returns a decision tree that, depending on the presence or absence of several biomarkers, recommends a treatment for each patient. In this case, the patients are divided into four subgroups (one for each level of the tree) denoted by *FLT3*^{Mut}, *NRAS*^{Mut}, and inv(16) biomarkers (**Table 3; Figure 21**).

Table 3. MOM Output: Patient stratification based on drug response to guide clinical decision-making

Name	Biomarkers	Drug	Patients Treated
Subgroup 1	*FLT3*^{Mut}	Quizartinib	103
Subgroup 2	*FLT3*^{WT} & inv(16)	Trametinib	15
Subgroup 3	*FLT3*^{WT} & no inv(16) & *NRAS*^{Mut}	Selumetinib	42
Subgroup 4	FLT3^{WT} & no inv(16) & *NRAS*^{WT}	Crizotinib	159

Following the new therapeutic strategy, the first biomarker is *FLT3^{Mut}*-including *FLT3-ITD*. Patients carrying *FLT3^{Mut}* would be treated with Quizartinib a 2nd generation *FLT3* inhibitor that is currently facing several clinical trials showing an increase in overall survival for AML patients[63]. This group of patients represents 30% of patients [98], in our study, 103 patients out of 319 belong to this group. The second subgroup comprises 15 patients and is characterized by *FLT3^{WT}* and the inv(16), which generates the fusion gene *CBFβ-MYH11*. Patients with these biomarkers are sensitive to Trametinib, a *MAPK* inhibitor that prevents cell replication and has been initiated in phase I clinical trials for hematological malignancies [103]. Interestingly, within this group, the patients with *NRAS^{Mut}* (4 out of 16) are the most sensitive to Trametinib. The third group is defined by the absence of previous biomarkers and *NRAS^{Mut}*. This subgroup poses special interest in the research as *NRAS* is one of the biomarkers most closely related to the general resistance to treatments of this disease [104]. *NRAS* gene variants are mutually exclusive with *FLT3* variants (p-value <0.05; **Figure 22**). Patients within this subgroup are sensitive to Selumetinib, a *MAPK* inhibitor that has started clinical trials for acute lymphoblastic leukemia in the UK [105], it is a mitogen-activated pathway inhibitor, which could inhibit *RAS* pathway functionality [106].

Finally, the fourth subgroup comprises the rest of the patients with none of the above mutational biomarkers but with other possible mutated biomarkers, for which the best treatment is Crizotinib -an *ALK* and *MAPK* inhibitor- approved by the FDA for lung cancer. It has not been enrolled in clinical trials for AML. Nevertheless, it has been used in studies of high-risk AML patients, with *TP53^{Mut}* and obtained very promising results [107].

To further validate the MOM's algorithm, we first run MOM on the BeatAML *ex-vivo* dataset using 10-fold cross-validation and compare the results that MOM outputs with each fold. This analysis shows that the MILP optimization returns robust results as 90% folds share 4 out of 5 biomarkers (**Figure 23**). Specifically, *FLT3^{Mut}* and *NRAS^{Mut}* subgroups appear in 10 out of 10 folds and subgroup with inv(16) in 3 out of 10 folds.

We assessed the sensitivity of MOM using a novel metrics. Since MOM suggests a single drug for each patient, the potential contingency matrix will be very unbalanced: for each patient, only the drug suggested by MOM is a positive and all the other treatments are negatives. Instead, we plotted (**Figure 24**) the sorted ranks of the drug predicted by MOM for each patient. We computed the p-values according to this distribution using thresholds for 1%, 5%, 10% (0.005, 4.58e-11, and 2.24e-23 respectively). A "prediction" algorithm that prescribes a drug by chance would show a curve close to the diagonal in this graph.

We then evaluated the treatment guideline proposed by running MOM with BeatAML within three independent AML datasets: two large-scale loss-of-functionality experiments that used both RNAi (DEMETER 2 [108]) and *CRISPR*-Cas9 (CERES [109,110]), and an additional large-scale cell-drug sensitivity analysis (Genomics of Drug Sensitivity in Cancer, GDSC [14,111,112]). We characterize cell lines using the Cancer Cell Line Encyclopedia's (CCLE [113,114]) genetic variant files, from which we clustered the AML cell lines into the four subgroups predicted by MOM using as input BeatAML. For CERES and DEMETER 2, we identified the main target and model drug effects to be proportional to the depletion of their target, which is the information these databases included.

For each subgroup, we compared each experiment's sensitivity (CERES score, DEMETER 2 score, and GDSC-IC50) dividing patients according to the presence of the biomarkers predicted by MOM in BeatAML and summing their sensitivity scores of the other three databases. We compute the sensitivity scores for the 4 subgroups, and the 3 datasets independently DEMETER2 (n=18 AML cell lines), CERES (n=14 AML cell lines), and GDSC (n=23 AML cell lines) (**Figure 21**). For the GDSC dataset, we compared the IC_{50} value from the cell lines with the selected biomarker and without the biomarker for a given subgroup drug. Finally, we performed an additional validation using DEMETER RNAi dataset (n=15 AML cell lines; **Figure 25**).

The change in sensitivity for the selected treatments is strongly significant using the MOM's predicted biomarkers in the three experiments (p-values of 5.5e-05, 6.8e-06, and 5.5e-04 for CERES, DEMETER2, and GDSC, respectively). Remarkably, inv(16) is difficult to be validated using cell lines, as commercial cell lines mostly lack this alteration. The ME-1 cell line is an exception to that, but GDSC is the only dataset that includes the translocation. Although this comparison is not statistically significant due to the lack of data, the GDSC-IC50 of ME-1 is 30 times lower than the average of cells without inv(16).

Figure 21. **Decision Tree for the Proposed Patient Stratification using MOM.** MILP from MOM obtained a hierarchical clinical guideline for patient stratification consisting of 4 different subgroups.

Each of them is denoted by a biomarker and represented by color (blue, turquoise, orange, and grey). These subgroups were validated in the BeatAML ex-vivo cohort, CERES, DEMETER2, and GDSC. Boxplots show the results of the validation. The y-axis represents the essentiality score from the different experiments and the x-axis represents the biomarker presence-absence of the samples. The validation was performed sequentially, already treated samples from previous subgroups were excluded in the following subgroups i.e. samples with *FLT3*^{Mut} (blue) from the first boxplot are not plotted in the non-biomarker (grey) in the second boxplot. CERES and DEMETER2 do not have experiments with cell lines having inv(16).

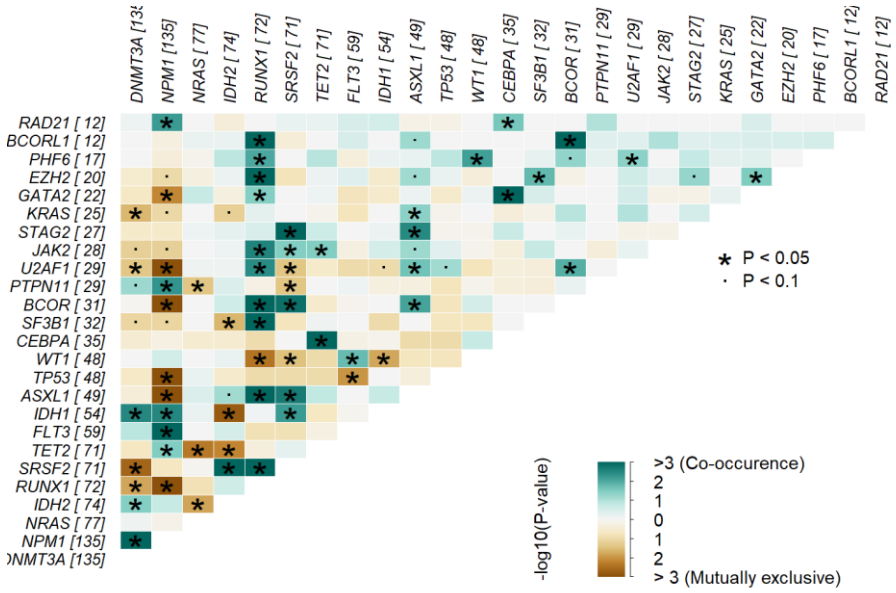


Figure 22. Somatic Interactions. Mutually exclusive or co-occurring set of genes calculated using pair-wise Fisher's Exact test. Associations plotted in green represent Co-occurrence while brown is a sign of mutual exclusivity. Stars are assigned to associations with P< 0.05. We appreciated that FLT3 and NPM1 variants are co-occurrent, and FLT3 and TP53 and NRAS and IDH2 are mutually exclusive respectively.

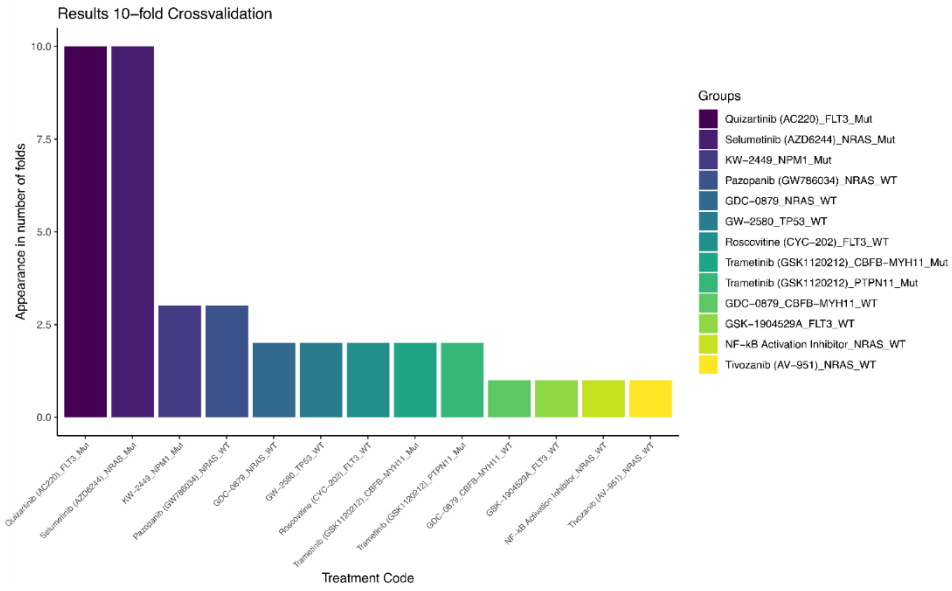


Figure 23. Results 10-fold Cross-validation. In the Y-axis is plotted the number of folds in which each subgroup appeared. In the X-axis all the different subgroups that appeared in the cross-validation.

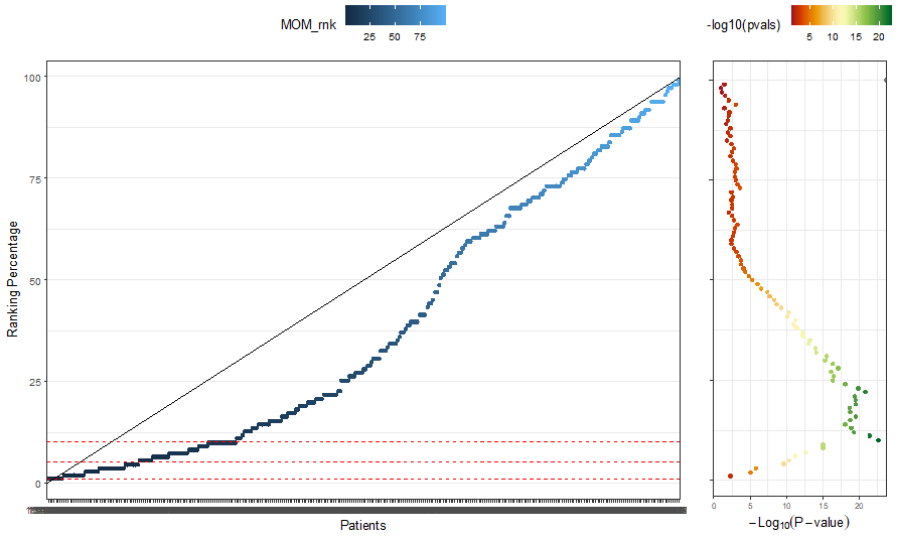


Figure 24. Results Sensitivity. We plotted the sorted ranks of the drug predicted by MOM for each patient. This plot shows that the suggested treatment was the best one in 3% of the cases, within the top 10% in 30% of the cases, within the first quartile in 46% of cases. The statistical significance for each of the thresholds can be stated using a Bernoulli distribution. We also included the p-values

according to this distribution using thresholds for 1% (p-value=0.005), 5%(p-value=4.58e-11), 10%(p-value=2.24e-23) and 25%(p-value= 4.32e-17)

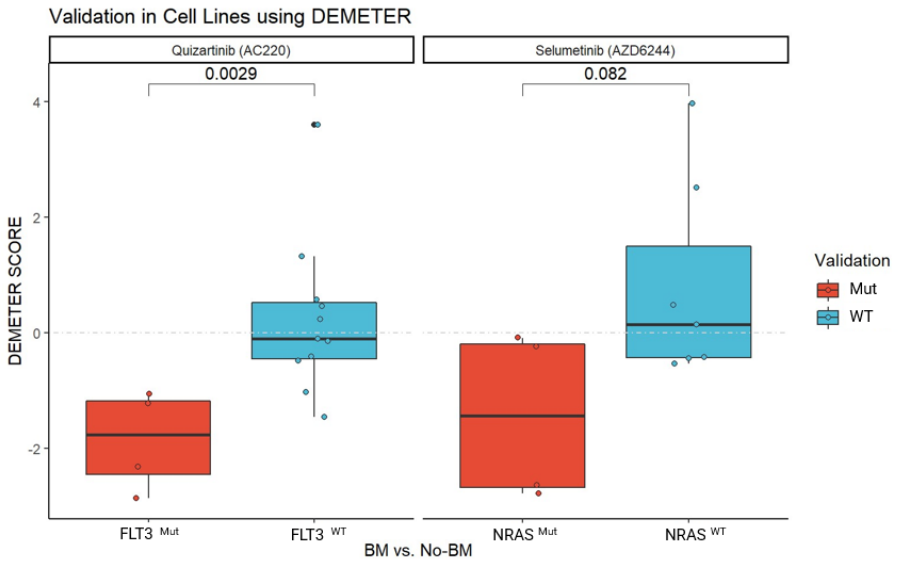


Figure 25. Validation in cell lines using DEMETER Score. In red the cell lines with the biomarker associated with the treatment; in blue the cell lines without the biomarker. On the left, FLT3Mut subgroup, on the right NRASMut subgroup.

We carried out a functional enrichment analysis to unveil the patient genotype according to the stratification proposed by MOM. We calculated the differentially expressed genes that are representative of each subgroup and computed the enriched biological functions of patients that belong to each group. The first subgroup, defined by $FLT3^{Mut}$, is characterized by downregulation in Myeloid Leukocyte Migration (adjusted p-value < 5e-3), this result is present in other functional enrichment studies involving *FLT3* mutated subgroup [115,116]. This subgroup has been repeatedly mentioned in literature and *FLT3* inhibitors are being implemented in the clinic [63]. The second subgroup, defined by samples with inv(16) and $FLT3^{WT}$ shows upregulated cell proliferation (adjusted p-value < 1e-3) including angiogenesis and endothelial cell migration upregulated among others, also described in other studies concerning this genetic aberration [117–119].

We also found that the $NRAS^{Mut}$ subgroup is related to the downregulation of alternative splicing (AS; adjusted p-value < 0.2). This subgroup has an upregulation of the transforming growth factor-beta (TGF-β) signaling pathway (adjusted p-value < 5e-03), which is mentioned in other studies concerning AS, especially in myelodysplastic syndromes

[120,121]. Furthermore, several studies have attempted to address the relationship between AML and AS, with promising results [122–124].

Finally, patients who do not have the previous biomarkers, have a downregulation in the amino acid catabolism process (adjusted p-value < 0.05), i.e. they are less able to metabolize amino acids than the rest of the subgroups [125]. A study demonstrates that for a subpopulation of AML leukemia stem cells the metabolism of amino acids from the medium is essential, and its absence leads to cell death [125]. Further description of the enriched functions for each subgroup, as well as their relationships and statistical significance, can be found in the **Appendix 2**.

Chapter 5. Methods

Filter and normalization

Filtering and Imputation

We used data from *ex-vivo* experiments, WES, and RNA-Seq from 319 Acute Myeloid Leukemia (AML) patients included in the BeatAML cohort [12]. Data was filtered to ensure all samples contained the gene variants and drug sensitivity information, the new dataset containing genomic aberrations and drug IC_{50} for the same patients was used as a starting point for the study. Genetic variant samples were previously pathogenically filtered by Tyner *et al.* [12] and we defined as a biomarker a genetic variant present in more than 1% of the patients (n≥4), leaving a total number of 64 possible biomarkers.

For missing drug sensitivity information in the ex-vivo experiments, we imputed the missing data using the *k*-Nearest Neighbourhood (kNN) Impute method, from Impute R package [126] (version 1.68.0).

*Drug Normalization: From IC50 to IC50**

Initially, we tried to use as drug sensitivity values the half-minimal inhibitory concentration, (IC_{50}) i.e., the concentration of a drug -in micro molar- for which half of the cell from the ex-vivo experiment die. Instead of using the IC_{50}, we propose the usage of an incremental version of the IC_{50}, named IC50*. As described in the results section, the usage of IC50* instead of IC_{50} is a convenient way to deal with the different toxicity of the drugs under study

After imputation, IC_{50} values were taken the $\log_{10}$ logarithm, normalized by subtracting the IC_{50} mean value for each drug, and these scores were made negative by subtracting an offset to the normalized IC_{50} value –the optimization model assumes negative values of drug sensitivity. The obtained drug sensitivity values are named **IC50***. The transformation from IC_{50} to **IC50*** is represented in **equation (4)**. Despite the formidable aspect of the formula, IC50* is simply an incremental and version of the logarithm of IC_{50} with an offset.

Let $\mathbf{IC_{50}}$ be a $T \, x \, P$ matrix, with T the total number of drugs and P the total number of patients, for which each element $\mathbf{ic50}_{t,p}$ is a value contained in (0,10] μM.

$$\boldsymbol{ic50^{*}_{t,p}} = (\log_{10}(ic50_{t,p}) - 1) - \frac{1}{P}\sum_{p=1}^{P}(\log_{10}(ic50_{t,p}) - 1) - \\ - \max\left((log_{10}(ic50_{t,p}) - 1) - \frac{1}{P}\sum_{p=1}^{P}(log_{10}(ic50_{t,p}) - 1)\right) \quad (4)$$

The obtained **IC50*** is a $T x P$ matrix containing the new drug sensitivity values.

Drug-biomarker association

Following with MOM's second step, we implemented a two-tailed Wilcoxon test to assess whether a biomarker influences the sensitivity of each the treatment. Each biomarker is tested against each drug and these associations were ranked according to the p-value. The p-values were adjusted following the methodology described by Gimeno *et al.* [1], using the R package *IHW* [94] (version 1.22.0). The package provides (given the p-values and the covariates –in our study genetic alterations–) a weight for each covariate related to its influence on the p-value significance.

Using these results, we included two consecutive filters. Firstly, we selected the biomarkers whose relative importance (the weight outputted by IHW) is larger than zero. IHW assigns a strictly positive weight to biomarkers relevantly correlated to the potency of a drug. Afterwards, we removed the drugs with no statistically significant relationship to the selected biomarkers (IHW p-value >0.05).

After this analysis, 122 treatments (biomarker-drug associations), with $\Delta\mathbf{IC50}^{*} > 0.2$ (including vs lacking the biomarker) and adjusted p-value<0.05 were considered for therapy.

MOM: MILP MODULE

Finally, in the third step, we proceed with the treatment assignation. We developed a MILP module described in **Chapter 4**. This module receives as input the 122 treatments and solves an optimization problem.

The core of MOM is an integer programming optimization model that predicts the combination of drugs and biomarkers that optimize patient response to treatment (i.e., **IC50***). Let us define a treatment as a combination of a drug and a companion biomarker. The solution to the optimization problem consists of a set of treatments that will be applied sequentially to patients in a defined number of steps (one treatment per step).

Let *S, P,* and *T* be the total number of possible steps, patients, and treatments included in the study respectively. Let **OM** be an input $P\ x\ T$ matrix of essentiality, whose elements $om_{p,t}$ contains the normalized sensitivity score (IC50*) of *ex-vivo* experiments for a patient *p* and a treatment *t*, which fulfills $om_{p,t} \leq 0$. IC50* values are all negative.

Let **K** be a $P\ x\ T$ a binary matrix whose element k_{pt} denotes whether a patient *p* is eligible for the treatment *t*, i.e., the treatment's companion biomarker is present in the patient, as follows:

$$k_{pt} \begin{cases} 1, & if\ the\ patient\ p\ has\ the\ biomarker\ associated\ with\ treatment\ t \\ 0, & otherwise \end{cases}$$

Let **X** be a binary $S\ x\ P\ x\ T$ array whose element x_{spt} states whether a patient *p* is treated with treatment t in step *s*, as follows:

$$x_{spt} \begin{cases} 1, & if\ patient\ p\ is\ given\ the\ treatment\ t\ in\ step\ s \\ 0, & otherwise \end{cases}$$

Let **Y** be a $S\ x\ T$ binary matrix whose element y_{st} represents whether a treatment *t* is used in step *s*, as follows:

$$y_{st} \begin{cases} 1, & if\ the\ treatment\ t\ is\ used\ in\ step\ s \\ 0, & otherwise \end{cases}$$

Given these variables, the MOM algorithm was built as a MILP optimization problem defined by the following equations.

(5) $$\text{minimize} \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{p=1}^{P} om_{pt} \cdot x_{spt}$$

(6) $$s.t. \quad \sum_{t=1}^{T} y_{st} = 1 \qquad s = 1, \dots, S$$

(7) $$\sum_{s=1}^{S}\sum_{t=1}^{T} x_{spt} \leq 1 \qquad p = 1, \dots, P$$

(8) $$x_{spt} \leq y_{st} \cdot k_{pt} \qquad s = 1, \dots, S; t = 1, \dots, T; p = 1, \dots, P$$

(9) $$x_{spt} + \sum_{n=1}^{s-1}\sum_{m=1}^{T} x_{npm} \geq y_{st} \cdot k_{pt} \qquad s = 1, \dots, S; t = 1, \dots, T; p = 1, \dots, P$$

The objective function of the MILP problem, **equation (5)**, is to minimize drug sensitivity scores (**IC50***) for all patients. The sensitivity score will be considered if it is included in the problem solution (x_{spt}). As drug sensitivity scores are negative, the MILP solution will intrinsically maximize the number of treated patients, as each included patient adds a negative term to the objective solution.

The proposed MILP problem has four sets of restrictions, namely **equation (6)** to **equation (9)**. **Equation (6)** is a set of *S* restrictions stating that each step consists of one treatment. **Equation (7)** is a set of *P* restrictions stating that at most one treatment must be used to treat each patient. **Equation (8)** is a set of $S\ x\ T\ x\ P$ restrictions stating that the treatment t can be applied to patient *p* in step *s*, only if (i) the patient *p* is eligible for treatment *t* based on his/her biomarkers ($k_{pt} = 1$), and (ii) the treatment t is used in step *s* ($y_{st} = 1$). Finally, **equation (9)** is set of $S\ x\ T\ x\ P$ restrictions that impose that the treatments included in the solution must be selected hierarchically, i.e., if we have a patient that would be eligible for two treatments, only the first treatment must be considered in the optimal solution.

To solve the model, we used CPLEX™© 12.10.0, Python 3.7.3, and the reticulate R package [127] (version 1.25.0).

MILP results can be directly translated into a decision tree for guiding clinical decision-making. The number of levels of the tree was set to four. Each level of this tree will be defined as one therapeutic AML subgroup and each subgroup is defined by a biomarker and a recommended drug.

Performance of MOM

The application of typical performance measures of machine learning (specificity, accuracy, sensitivity, ROC and PR curves, etc.) to this specific problem is not straightforward. Since MOM suggests a single drug per patient, the potential contingency matrix will be very unbalanced: for each patient, only the drug suggested by MOM is a positive and all the other treatments are negatives. If the suggested drug is the one with the lowest IC50*, the prediction will be a true positive. Otherwise, it will be a false positive. On the other hand, all the drugs that were not selected are true negatives (except the one with the lowest IC50*). The drug with the lowest IC50*, if not selected by MOM, will be a false negative.

Instead of this approach, we plotted the sorted ranks of the drug predicted by MOM for each patient. This plot shows that the suggested treatment was the best one in 2% of the cases, within the top 10% in 30% of the cases and so on. The statistical significance for each of the thresholds can be stated using a Bernoulli distribution. We computed the p-values according to this distribution using thresholds for 1%, 5%, 10%.

External Cohort Validation

For validating the different subgroups, we compared patients that are given a drug in a specific subgroup against the remaining non-treated patients. We validated our results using cell lines, specifically, used 2 different large-scale gene essentiality experiments including RNAi (DEMETER 2 [108]) and CRISPR-Cas9 (CERES [109,110]), and an additional large-scale cell-drug sensitivity analysis (Genomics of Drug Sensitivity in Cancer, GDSC [14,111,112]). We characterized the cell lines using the Cancer Cell Line Encyclopedia (CCLE [113,114]) genetic variants files, from which we were able to divide the cells into different subgroups.

We performed the following test for validation. Cells were divided into two groups. The first group includes cells with the biomarker associated to that subgroup, and the other group, contains the cells without the biomarker that had not been previously treated. This comparison was computed for the 4 subgroups, and the 2 datasets DEMETER 2, and CERES. DEMETER 2 and CERES were compared using the viability score that corresponds to knocking out the corresponding targets for each drug. For the GDSC dataset, we used the IC_{50} value provided in the experiments. All tests were one-tailed Wilcoxon's test to check that the sensitivity increase in the cells with the biomarker.

Functional Analysis of the Subgroups

Functional analysis of the subgroups was performed using gene expression data from the BeatAML [12] cohort. We performed a differential gene expression analysis using limma R package [128] (version 3.50.3). The contrast matrix compared one group against all the others, therefore, there was a different contrast for each group.

Genes differentially expressed were ranked according to its t-statistic, if t >0, genes were considered overexpressed, if t<0, genes were considered underexpressed. For each subgroup, we selected the top 500 over and under expressed genes and performed a Gene Ontology Enrichment Analysis (GEA) using Fisher's Test. We analyzed the biological process ontology. Enriched functions on the overexpressed genes were upregulated, and functions obtained from the underexpressed genes were considered to be downregulated. The statistics were computed using clusterProfiler R package [129] (version 3.10.1). We set an adjusted p-value cutoff of 0.2 for considering a function differentially enriched, adjusted p-values were computed using the Benjamini-Hochberg procedure.

.

Discussion of Section 2

Despite the advances in drug *ex-vivo* screening and computational methods for precision medicine, there are technical issues that limit their translation to clinical practice. Some of these issues are the influence of drug toxicity, the enormous number of statistical hypotheses, the complexity of developing algorithms understandable by the clinician, and the difficulty of proposing an effective treatment guideline that assigns the best drug for each patient. MOM faces and solves each of these challenges.

These statements are not yet covered by current AI strategies, which are focused on increasing accuracy and sensitivity regardless of the complexity of the end model [28,130]. In these AI methods, the absence of interpretability of the feature used for classification prevents further research and downplays the need for clinically defined subgroups [19,131,132]. Indeed, the need of developing interpretable AI algorithms is not only related to easing the diagnosis pipeline in cancer but also to increase and facilitate that the pharma industry brings new drugs and biomarkers to market. Drug regulators -such as the Food and Drug Administration- value that the process to unveil novel biomarkers is robust and transparent [49]. In contrast, the patient stratification guideline provided by MOM has the following characteristics, i) allows treatment assignment by using a simple genetic panel, ii) the results are non-stochastic, they are the same for all possible re-runs of the model, iii) the algorithm outputs a decision tree for treatment guidance.

IC_{50}, EC_{50}, and AUC (used for example in [14,27,133]) are reasonable metrics to determine the efficacy of a drug. None of them, however, considers the overall toxicity of the drug. Using IC50* in the optimization problem, we focus on the differential effectiveness of a drug among different patients, and therefore, drugs that are toxic for most samples will not be included in the solution.

IHW provides us with the ability to increase the power of tests and reduce the FDR. With this strategy, we are also able to identify the direction of the influence of genetic events in drug response, i.e., whether it defines sensitivity or resistance. With this approach, we successfully detected *FLT3* as highly influential in terms of sensitivity to treatment, which is coherent with other studies [98]. *NRAS*, instead, showed as a mutation associated with treatment resistance also coherent with literature [99,104]. One promising conclusion for this study is that we managed to find a drug for which *NRAS* correlates to drug sensitivity.

Interpretable AI defined by MILP ensures that the subgroups obtained are optimal. This feature is not common to other classification methods. However, it also presents two main limitations. The first one is computational resources, which increases exponentially with the number of possible biomarkers, drugs, or patients (on a standard desktop, the presented work required 2.5 hours of computing time). In addition, the incorporation of new non-binary diagnostic markers requires the redefinition of the model. However, once the optimization problem is solved, assigning a treatment to a novel patient is immediate.

Our AML patient stratification includes a subgroup defined by the absence of a genetic mutation, i.e., wild type. It also includes patients who have $TP53^{Mut}$ genotype, which are biomarkers associated with poor prognosis [134]. MOM recommends treating these patients with Crizotinib, a drug used in other studies with $TP53^{Mut}$ AML patients which in fact showed very promising results[107]. In addition, this subgroup shows a deficiency in amino acid metabolism which may lead to alternative treatment therapies based on metabolomics.

The subgroup defined by the *CBFβ-MYH11* fusion gene appears characterized in a very small percentage of AML cell line cohorts but is nevertheless present in 7% of AML patients[135], which enhances the relevance of this biomarker. *CBFβ-MYH11* is a clear indicator of sensitivity to Trametinib, a clinical drug that inhibits cell replication pathway [136], which, in turn, appeared as an upregulated biological process in this subgroup. In the remaining subgroups, $FLT3^{Mut}$ is widely described in the literature [98]. In contrast, $NRAS^{Mut}$ appears as a biomarker of sensitivity for Selumetinib and has downregulated the alternative splicing (AS) process. This subgroup contains, on balance, effective treatment for a resistance-associated mutation and a new line of research linking alternative splicing and AML.

It is remarkable the appearance of three different *MAPK* inhibitors in the proposed therapeutic strategy, which is coherent with the disease behavior. Our biomarker analysis revealed that the RTK-RAS pathway is the most affected in our cohort of AML samples (**Appendix 2-Section BeatAML Cohort**). Of all drugs suggested as treatment, only Quizartinib is clinically approved for AML patients [60]. This study aims to accelerate -once the results are validated in cell lines and murine models- the process of approving these drugs for AML.

The validation of the results is challenging in a real cohort since most patients are treated with standard induction cytotoxic therapy (only 7.5% of AML patients in TCGA are treated with other treatments). We propose a strategy to take advantage of cell lines loss-of-function datasets. Nevertheless, even using cell lines -that are quite different from ex vivo samples- we validated the subgroups and the IC_{50} of the lines with indication was significantly better than the IC_{50} of those without indication. Therefore, in the absence of clinical data for validation, we consider the results using cell lines data to sufficiently support this study.

The concept of MOM is also applicable to other disease types using *ex-vivo* experiments as well as to other sensitivity measurements, leaving an open door for new patient stratifications based either on drug response or even on any other experiment to measure the effectiveness of certain drugs in the future. We believe that Interpretable AI will help physicians and regulators understand AI medical decisions and, therefore, ease the translations of AI analysis of drug screening experiments to clinical practice.

Section 3: The challenge of Interpretability

Assessing the Interpretability of novel and state-of-the-art methods

Introduction to Section 3

As mentioned in the first chapter of this thesis dissertation, interpretability is an essential feature for the applicability of a machine learning model. Interpretability makes an ML model understandable by an expert and allows him/her to critically judge the algorithm output. This characteristic is especially important in Precision Medicine.

In this section, we will compare various methods that solve the assignment problem posed by PM in terms of their interpretability. To do so, we reviewed the current literature to summarize the state-of-the-art and selected different methods that were defined as interpretable and solved the assignment problem.

We have compared 22 different algorithms suited for solving the assignation problem, out of which, 13 algorithms were black-boxes [29,30,39–41,31–38], thus, they are not interpretable and nor suited for the comparison. None of them were defined as explainable [43,46,51–56]. We divided these latter into the two main approaches mentioned in **Chapter 1**, "patient-centered" meaning that the method outputs a drug assignation for each patient [2,137], and "drug-centered" meaning that the method finds which patient or patients that are responders for a specific drug [138–140]. Regarding this last group, "drug-centered" methods can be transformed into "patient-centered" if their output is a continuous variable, as the best drug for each patient will be the drug with the highest sensitivity from the drugs predicted for that patient. We thus, selected from this last group only the methods that are suited for "patient-centered" approach.

Within this last classification, the top-ranked algorithms in the state-of-the-art are Multidimensional Optimization Module (MOM) [2] and Kernelized Rank Learning (KRL) [137] from the "patient-centered" perspective, and BOSO [138] and Lasso Regression [140] in the "drug-centered" approach. MOM uses mixed integer linear programming (MILP) to discover the optimal therapeutic strategy that is returned as a decision tree and was described in the previous section. KRL is a machine learning method based on an optimization problem that applies a kernel approach to circumvent the convexity limitations and also solves the problem using MILP. BOSO and Lasso can be applied to predict the IC_{50} of a drug in different patients. BOSO is a MILP model built up from the Lasso Regression equations that have been modified to predict a numeric variable with the least number of features, improving the reduced interpretability of Lasso Regression.

We also included in this analysis two novel “patient-centered” algorithms developed in this thesis: Optimal Decision Trees (ODT) and an adaptation of the Multinomial Lasso. ODTs are decision trees that recursively optimize the drug recommendation on each branch until a preset group size is reached. Finally, Multinomial Lasso is a modified Lasso regression methodology for which each patient “selects” its best drug using a vote sharing scheme.

Our main challenge in this comparison was to provide a definition of interpretability. There is no immediate quantitative way to compare two methods in terms of their interpretability, so in this section we will define a quantitative metric to evaluate the interpretability of a method. In turn, we also include some qualitative metrics to complement the quantitative ones.

We compared the methods in this section in terms of interpretability, focusing specially on the accuracy, multi-omics capability, explainability, and implementability. Method comparison was performed using the BeatAML [12] dataset and the Genomics of Drug Sensitivity in Cancer (GDSC) [112] dataset for Acute Myeloid Leukemia (AML).

We focused on the BeatAML dataset due to its abundance of patient information –e.g. genomic data, gene expression, clinical data–, and drug sensitivity information which proceeded from ex-vivo experiments performed on patient samples instead of cell lines [12]. Indeed, ex-vivo drug sensitivity provided more information for patient sensitivity than conventional information from clinical data due to the possibility of testing more drugs on the living tumor without injuring the patient and solving possible harmful drug interactions from previous treatments. Although this information could be less reliable, it solves the sparsity issue of drug sensitivity data explained in **Chapter 1** of this dissertation. Furthermore, drug screens performed on ex-vivo experiments improve data reliability if compared to cell line screenings. Nevertheless, further experimental validation is required for clinical applications.

Chapter 6. Methods

Focusing on approaches to address the complex Precision Medicine (PM) problem, we found different methodologies from the "patient-centered" perspective, Multidimensional Optimization Module (MOM) [2] and Kernelized Rank Learning (KRL) [137]. We also included in this group two novel algorithms: Optimal Decision Trees (ODT) and an adaptation of the Multinomial Lasso.

In the "drug-centered" approach, BOSO [138] and Lasso Regression [140] can be applied to predict the IC_{50} of a drug in different patients. Both methods select a small number of variables to make their predictions. Once the predictions are obtained, comparing the predicted IC_{50} for each drug on a patient, the drug with the minimal IC_{50} is selected. The description of the six methods is summarized in **Table 4**. Some of the methods only accept binary data as input. These methods cannot be applied to gene expression unless using a hard threshold.

Optimal Decision Trees (ODT)

In this work, we are introducing a novel algorithm that uses a tree-like method for precision medicine. This method is intrinsically different from classification or regression trees, as will be shown.

In a classification tree, in each step, the tree is split into two subtrees finding the variable (with its corresponding threshold) that best splits the tree according to some figure of merit (Gini index, entropy, information gain, etc.). This figure of merit measures the overall enrichment of the classes in the subtrees.

On the contrary, the ODT algorithm selects for each step the splitting variable (selecting a proper threshold) *and the treatments for each split*. The selection is based on the optimization of an overall measure of the sensitivity of both branches to the selected treatments (**Figure 26**).

Table 4: Precision Medicine Pipelines selected for comparison. This table collects the description of each of the methods. Algorithm shows the method's given name. Type refers to whether the method is patient- or drug-centered. The software column collects all the required software environment programs for the model to be run. Method refers to the pipeline description. Suitable for mutational data has a "Yes" if the method could use genetic variants as input. Suitable for gene expression has a "Yes" if the method could use gene expression data as input and a "No" otherwise. Output explains the raw output of the model. Reference contains the publications in which the method was defined.

Algorithm	*Type*	*Software*	*Method*	*Suitable for mutational data*	*Suitable for gene expression*	*Output*	*Reference*
MOM	Patient	Python 3.7, R 4.2, and CPLEX	Feature Selection and MILP	Yes	No	Drug assignation	[2]
ODT	Patient	R 4.2	Recursive Decision Tree	Yes	Yes	Drug assignation	Novel
Multinomial	Patient	R 4.2	Adapted Lasso	Yes	Yes	Drug assignation	Novel
KRL	Patient	Python 2.7	Kernelized MILP	Yes	No	Drug assignation	[137]
BOSO	Drug	R 4.2 and CPLEX	Lasso regression using MILP	Yes	Yes	Predicted IC50 for a drug	[138]
Lasso	Drug	R 4.2	Traditional Lasso regression	Yes	Yes	Predicted IC50 for a drug	[140]

Specifically, let **Y** be a P x D matrix where P is the number of patients and D is the number of tested drugs. Each of the entries of the matrix quantifies the sensitivity of each patient to a drug, i.e., the matrix Y can be either the IC_{50} or a modified version of it, the area under the concentration-response curve, etc. Let **X** be a P x M matrix where P is the number of patients and M is the number of biomarkers. The matrix **X** can be a matrix of mutations, gene expression, or other characteristics specific to each patient.

In the case of binary variables (mutations for example), for each step in the splits of tree, the following optimization problem is solved (**Equations 4-6**):

$$\max_{m,d_1,d_2} A + B \tag{4}$$

$$A = \sum_{p \in split} y_{pd_1}(x_{pm} == 1) \tag{5}$$

$$B = \sum_{p \in split} y_{pd_2}(x_{pm} == 0) \tag{6}$$

Where split is the set of patients under study (all patients belong to the split in the case of the root node), *m* is the selected mutation or biomarker, and d_1 and d_2 are the selected drugs for the patients that have or do not have the mutation m respectively. The notation "(*condition*)" represents 1 or 0 depending on whether the expression inside the parenthesis is true or false (**Equations 5,6**). This problem can be easily extended to continuous variables, using a threshold (**Equations 7-9**). In this case the optimization problem is:

$$\max_{m,th,d_1,d_2} A + B \tag{7}$$

$$A = \sum_{p \in split} y_{pd_1}(x_{pm} >= th) \tag{8}$$

$$B = \sum_{p \in split} y_{pd_2}(x_{pm} < th) \tag{9}$$

Both optimization problems start by setting all the patients within the studied split. The optimization splits the patients into two groups. For each of these groups, the algorithm is applied recursively until the number of patients in the split is smaller than a given number or until the optimization problem results in the same drug for both splits.

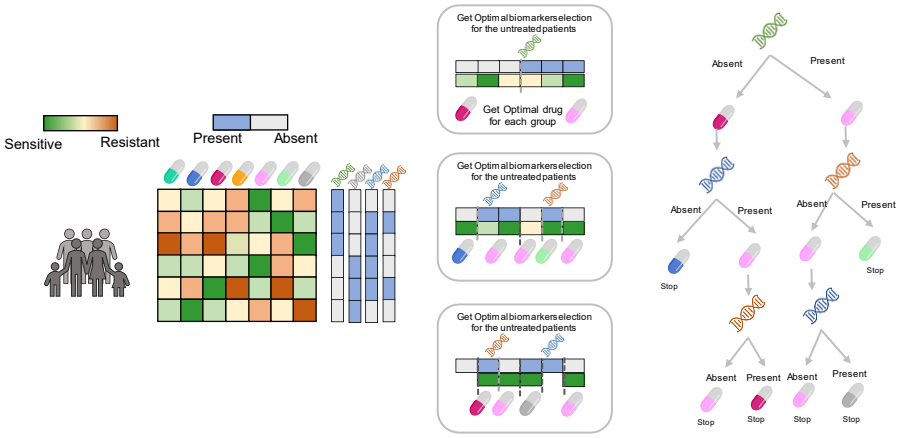


Figure 26. ODT Model Performance. The ODT model uses as input the sensitivity matrix and the biomarker matrix, on each step it splits the patients into two groups according to the presence or absence of a biomarker. This split is optimized so that the drug-assigned is the most sensitive to each of the splits. It recursively splits the different branches until a predefined group size is reached.

The **equations (5,6,8,9)** maximize the sum of the sensitivities of the patients of each of the branches. Using the same algorithm, it is possible to apply any transformation of the sensitivity and include them in the optimization process. In this case, equations (5) and (6) are transformed into:

$$A = \sum_{p \in split} f(y_{pd_1})(x_{pm} >= th)$$

$$B = \sum_{p \in split} f(y_{pd_2})(x_{pm} < th)$$

Equations (5) and (6) can be transformed in an analogous way. To minimize the effect of outliers in the sum, we used the square root function to diminish the dynamical range of the data. The transformation is named ODT Sqrt in this work.

Multinomial logistic Lasso regression

The assignation of the proper drug to each patient problem can be tackled as a multiclass classification problem: the number of classes is the number of drugs and each patient is assigned the most effective drug for him/her. Using this approach, a multinomial regression can be applied to select the proper drug for each patient.

Predicting exclusively the most effective drug can be simplistic, since the penalty for misclassification is identical for the second most effective drug or for the least effective drug. Since the multinomial regression can also be applied to classes defined by continuous variables, it is possible to give a "vote" for each patient that can be shared among all the drugs: the most effective drug will receive more shares of this vote that the least effective drug. Assigning the whole vote to the most effective drug can be seen as a particular case of this approach. Vote sharing can be seen as a transformation of the PM problem to a classification problem using probabilistic labels.

The Lasso regression is also implemented for multinomial regression. The implementation of glmnet (R Package) [140] is fast and convenient and allows for automatic selection of the regularization parameters using cross-validation.

More specifically, the multinomial regression builds the multinomial regression model (**Equation 10**)

$$\boldsymbol{X\beta \sim Z} \qquad (10)$$

where **X** is a P x M matrix where P is the number of patients and M is the number of biomarkers, $\boldsymbol{Z}$ is P x D voting matrix where P is the number of patients and D is the number of tested drugs. All the elements of Z are positive and the sum of its elements by rows is equal to one. Finally, $\boldsymbol{\beta}$ the output of the regression is a M x D coefficient matrix. $\boldsymbol{X\beta}$ are the predicted logits for each drug being the most effective for each patient (**Figure 27**)**.**

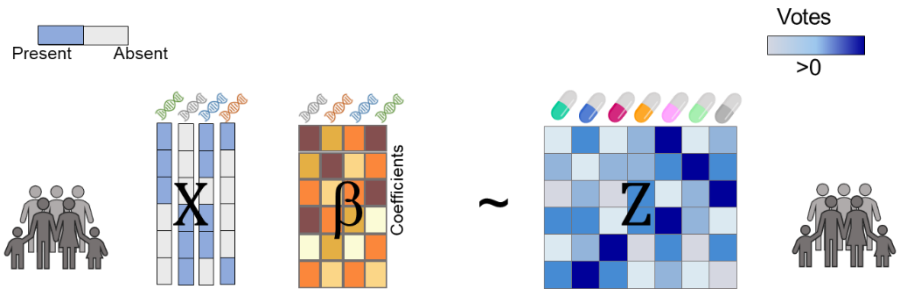


Figure 27: Multinomial Model. The multinomial Model corresponds to a modified Multinomial logistic lasso regression, where the output represents the votes that each patient assigns to each of the drugs.

The specific selection of the entries for the Z matrix is shown in **Equation 11.**

$$z_{pd} = \frac{\exp\left(-K\frac{y_{pd}}{\min_p(y_{pd})}\right)}{\sum_{i=1}^{D}\exp\left(-K\frac{y_{pi}}{\min_p(y_{pi})}\right)} \quad (11)$$

Where y_{pd} are the entries of the **Y** matrix (that measures the sensitivity to a drug) and *K* is a predefined constant. If K>>1, all the exponentials of the summations of the denominator but the min(y_{pi}) vanished and the vote is given to the most effective drug. If K = 0, all the drugs share 1/D votes.

Data for Comparisons

We focused on Acute Myeloid Leukemia (AML) to compare the different methods described above. This disease was selected due to the availability of a wide cohort of patients with genomics data and ex-vivo drug sensitivity screening data. Ex-vivo data is more reliable than drug screenings performed on cell lines due to the similarity of the AML patients' blood to the tumor tissue. Furthermore, AML is a highly heterogeneous disease with not standard PM therapeutic strategy, even though there is a growing field of drug development likely suited for these patients, e.g. Tyrosine Kinase Inhibitors (TKIs)[141].

Consequently, we selected the BeatAML cohort [12] for training the models and predicting different therapeutic strategies. This cohort is publicly available at http://www.vizome.org/ . We normalized the drug sensitivity IC_{50} from the ex-vivo experiments into IC50* described in [2]. To validate the predictions and due to the absence of more large-scale ex-vivo experiments we used as an independent cohort testing set, the GDSC drug screening for AML cell lines [112], which could be found publicly available at https://www.cancerrxgene.org/.

We compared the different algorithms primarily on the basis of four aspects that define interpretability (**Figure 28**): i) *the accuracy* of the method, for which we performed a 5-fold cross-validation in the training set, an independent cohort validation and an intragroup validation with the predicted groups in the training and validation set, ii) *the multi-omics capability*, for which we tested the ability and performance of the methods when training with gene expression and genomic variants, iii) *the explainability*, for which we performed a qualitative comparison of all algorithms, analyzed the number of variables that each algorithm uses for prediction, and iv) *the implementability*, for which apart from qualitative

comparisons based on the method definition, the computing time that each model requires for training becomes essential.

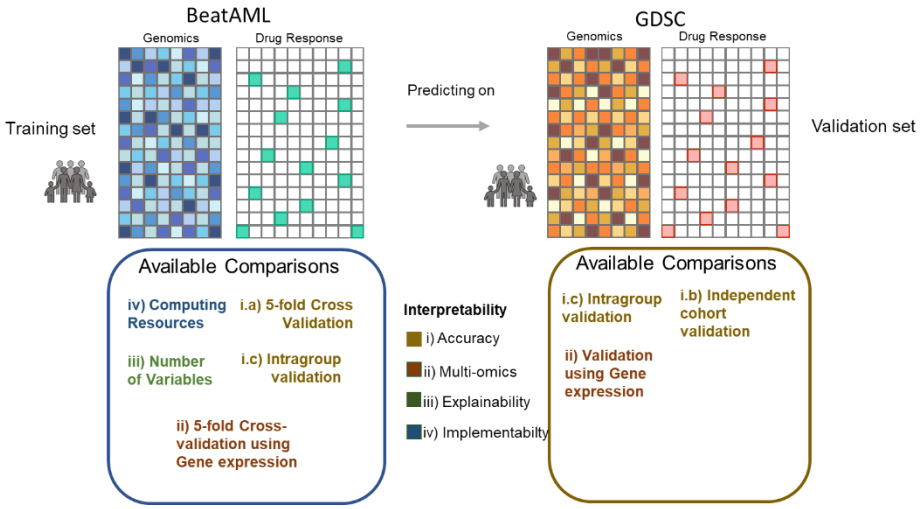


Figure 28: Summary of the available comparisons performed in this study. We trained the different models in BeatAML cohort and tested the predictions predicting over GDSC. From the training step we were able to obtain the training computing time, the number of variables required to make the predictions, a 5-fold cross-validation using mutational and gene expression data, and intragroup validation. Whereas for the testing step we performed and independent cohort prediction validation using mutational and gene expression data, and another intragroup validation

Accuracy

The first "*sine qua non*" characteristic of a PM methods is the accuracy. An "interpretable" method with low accuracy becomes irrelevant. We define the accuracy as the difference of the IC50* for the assigned drug and the drug with maximum IC50* for that patient.

For assessing the accuracy of each of the methods, we performed the following comparisons: 5-fold cross-validation, independent cohort validation, and Intra-group validation.

5-fold cross-validation in BeatAML

We performed a 5-fold cross-validation using the BeatAML dataset. We trained all models with genetic variants data from 319 patients, dividing the cohort between the training samples 4-folds and testing samples the selected 1-fold. Each of the folds were tested, and the predicted IC50* for the 5-fold testing was compared for all the methods and compared

against the Oracle -the drug with the optimum IC50*- (**Figure 29**). We calculated the Oracle as the minimum IC50* value for each patient.

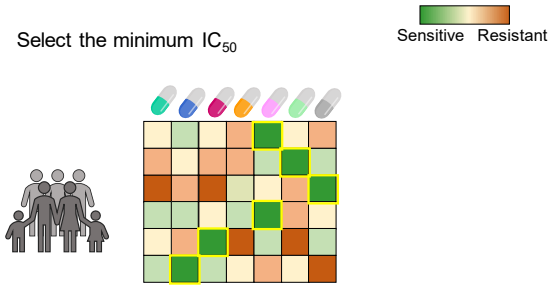


Figure 29: Oracle Method. The Oracle predicts the most sensitive drug for each patient or cell line.

Independent cohort validation

One of the main challenges of Machine Learning, including Precision Medicine, is generalization, i.e. the ability to adapt to new, previously unseen data. All the methods were tested on the GDSC AML dataset to check their generalization ability. The models were trained using the BeatAML dataset and were used to predict the optimal drug for AML cell lines from GDSC using its mutation files. Each of the cell lines was recommended a drug, we compared the all-samples IC_{50} for all the models and against the Oracle (the drug with the minimum IC_{50} for each cell line).

Intra group validation

We compared if the IC50* of a drug in the patients where it was recommended, was lower the IC50* in patients where it was not recommended. Using this information, we compared the sensitivity to a drug for a specific group against the sensitivity to that drug for the rest of the samples by using a 2-tailed Wilcoxon test. This analysis was performed both for the BeatAML dataset (training dataset) and the GDSC AML cell lines cohort (predicted dataset). This analysis was performed using the predicted drug recommendation for the BeatAML dataset (training dataset) and the GDSC AML cell lines cohort (predicted dataset).

Multi-omics suitability

Some of the methods only accept as input binary variables. Although, genomic variants can be transformed into binary variables, gene expression, methylation, or openness of the chromatin are intrinsically continuous variables. We have included a table that checks if the algorithm accepts only binary inputs (only genomic variants) or if it accepts continuous data

as well (gene expression, methylation, etc.). For the methods that accept continuous variables, we assessed the performance of the predictions (5-fold cross-validation) in the BeatAML dataset using both data sources. We state the statistical significance using a 2-tail Wilcoxon's test comparing the IC50* using as input either genetic variants or gene expression.

Explainability

PM is more suited for healthcare if it can be interpreted. A machine learning method is interpretable if it provides the decision criteria that define the pathway that leads to the solution.

Explainability is defined by three different aspects: i) the explainability of the results, which checks if the method provides a ranking of the variables according to their importance for drug recommendation, ii) the capacity to output an easy-to-apply decision criteria, and iii) the understandability of the methods, this category mentions if the process of the algorithm to reach the classification criteria is easy to understand.

For assessing these characteristics, we performed a qualitative analysis based on the method description and execution. Furthermore, we analyzed the number of variables that each model requires to make the predictions. A model with a small number of variables is easier to understand, improves the understanding of the variable ranking, and is easier to for clinical diagnosis. Therefore, we paid special attention to the number of variables.

Implementability

Implementability is the easiness of a method being implemented into clinical research or practice. We measured the implementability of a method by analyzing four main features: i) the feasibility for wet-lab validations, ii) the consideration of the physician's experience, iii) the generation of a clinical guideline, and iv) technical implementation, which refers to the computational burden and software that the method requires. We used qualitative grades for the first characteristics. Regarding the technical implementation, we considered the computational burden. Despite it could be considered less important, some of the algorithms require hours of computing time for the BeatAML of the subset of AML samples in GDSC -that be considered to be small/medium size. Requiring fewer resources makes an algorithm more attractive to be applied for larger datasets. We also analyzed the software environment that each model requires to be run.

Chapter 7. Results

In this work, we compare several aspects of the performance of different interpretable models [131]. These models were classified into two main groups. The first one, named patient-based, are models that return a specific therapeutic strategy for each patient. The second one, named drug-based, are models that provide the patient(s) that are especially sensitive to a specific drug. Patient-based models include Multi-dimensional Module Optimization (MOM) [2], Optimal Decision Trees (ODT), Kernelized Rank Learning (KRL) [137], and Multinomial Lasso. Drug-based models are more suited for physicians and clinical investigation. This group comprises BOSO [138] and Lasso [140]. Patient-based methods rank the effectiveness of the drugs for a specific patient. Drug-based methods rank the effectiveness of a specific drug for each of the patients.

All the methods were developed to predict the drug response or develop a treatment strategy using genetic variants information. Thus, we trained the methods to predict drug efficacy using patients' samples and ex-vivo drug efficacy from the BeatAML [12] dataset. The methods were compared in terms of interpretability, which was defined according to four properties namely accuracy, adaptability, explainability, and easiness of implementation.

Accuracy: all the methods provided good estimates

The first test to assess the accuracy was a 5-fold cross-validation in BeatAML [12]. Results for this analysis can be found in **Figure 30.a.** Multinomial Lasso achieves the lowest median, -the highest sensitivity- although it also entails the highest variability. Lasso's prediction is similar to the former one, but its standard deviation is smaller. Finally, MOM and BOSO achieve almost identical median. ODT –in both versions – has the highest IC50* prediction, i.e. the smallest value for sensitivity. However, the performance of the methods –excluding ODT and ODT Sqrt– was not statistically significant (p-value >0.05). ODT and ODT Sqrt predictions were significantly worse than Multinomial (two-sided Wilcoxon test p-value=0.005921 and p-value=0.004942, respectively).

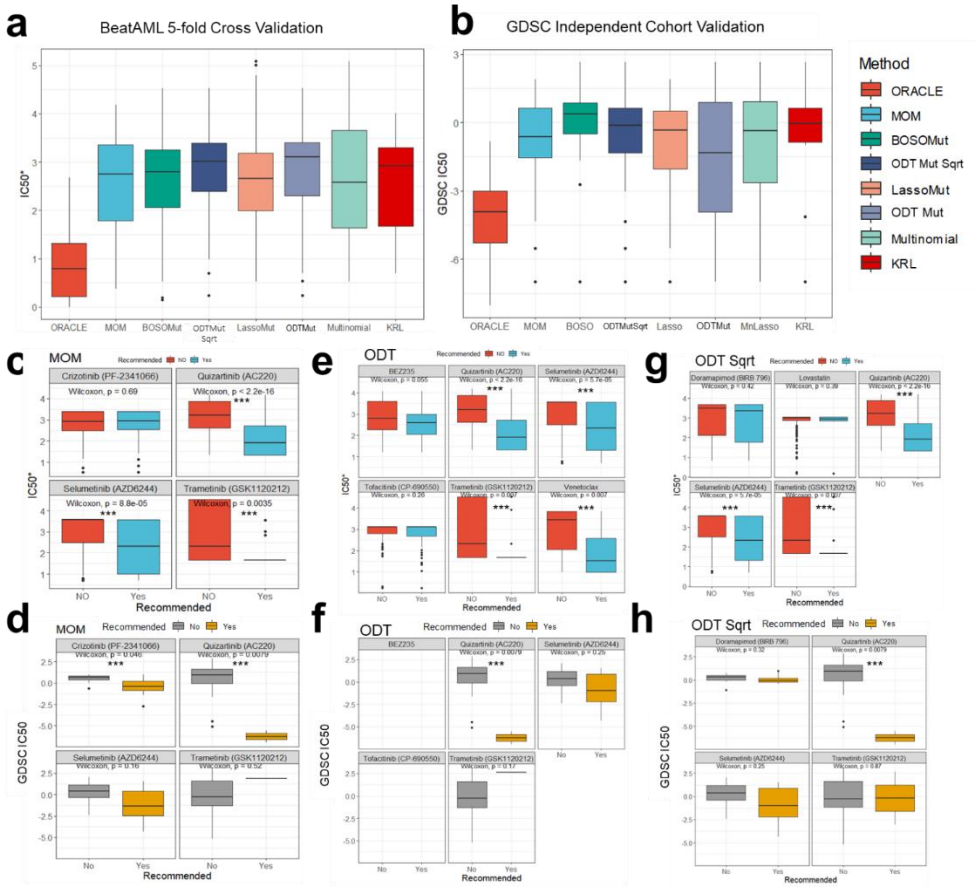


Figure 30: Accuracy comparison. A) **Accuracy in 5-fold cross validation from BeatAML cohort**. The different boxplots show the predicted IC50* of the drugs assigned to each of the patients. The lower the IC50* is, the more sensitive the method is, ORACLE is the control that shows the best possible drug to every patient in the cohort. B) **Accuracy in independent cohort validation**. The different boxplots show the predicted IC50 of the drugs assigned to each of the patients in GDSC. The lower the IC50 is, the more sensitive the method is, ORACLE is the control that shows the best possible drug to every patient in the cohort. Models were trained in BeatAML and predicted over GDSC. C)**Intragroup validation of MOM in BeatAML**. Each of the subplots represents the efficacy of one drug, in blue the patients that were recommended that drug and in red the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05). D) **Intragroup validation of MOM in GDSC**. Each of the subplots represents the efficacy of one drug, in yellow the patients that were recommended that drug and in grey the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05). E) **Intragroup validation of ODT in BeatAML**. Each of the subplots represents the efficacy of one drug, in blue the patients that were recommended that drug and in red the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05). F) **Intragroup validation of ODT in GDSC**. Each of the subplots represents the efficacy of one drug, in yellow the patients that were recommended that drug and in grey the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05). G)

Intragroup validation of ODT Sqrt in BeatAML. Each of the subplots represents the efficacy of one drug, in blue the patients that were recommended that drug and in red the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05). H) **Intragroup validation of ODT Sqrt in GDSC**. Each of the subplots represents the efficacy of one drug, in yellow the patients that were recommended that drug and in grey the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05).

In the second test, the used the models trained on the full BeatAML, and tested them against the Genomics of Drug Sensitivity in Cancer (GDSC) AML dataset. This dataset contains the genetic variants information for each cell line and the IC_{50} values for most of the drugs in the same cell lines. The independent cohort validation showed very different results from the 5-fold cross-validation (**Figure 30.b**), In this case, the ODT standard method achieved the best sensitivity score followed by MOM, Multinomial Lasso, Lasso, ODT square root and BOSO. The IQR for ODT and its standard deviation were much larger than for other methods. Nevertheless, there were no statistical significance in the difference of the predicted GDSC IC_{50} comparing any of the methods.

In the third test, we analyzed the intra-group classification performance. In this test we compared the IC50* of patients that were recommended a drug with the IC50* of the rest of patients using BeatAML and GDSC. The models with the best intragroup performance were MOM and ODT in their standard form followed by ODT Sqrt. MOM showed a significant sensitivity difference in 3 out of 4 groups for the BeatAML dataset (**Figure 30.c**) and 3 out of 4 for the GDSC dataset (**Figure 30.d**). ODT standard achieved a significant intragroup sensitivity in 4 out of 6 groups for BeatAML (**Figure 30.e**) and 2 out of 5 for GDSC (**Figure 30.f**). Finally, ODT Sqrt significantly recommended the usage of 3 drugs out of 5 for BeatAML (**Figure 30.g**), and 1 out of 4 in GDSC (**Figure 30.h**). No statistical significance was found for the rest of the methods. Probably, this is owing to that there are more 10 different suggested treatments and the number of patients is small to get statistical significance (**Appendix 3**:**Figure S6-Figure S13**).

Multi-omics: using gene expression as input provides similar accuracy if compared to genetic variants.

We tested whether using gene expression could improve the method accuracy [142]. We trained all models (except MOM and KRL, since they do not accept continuous inputs) using BeatAML gene expression (GE) data. We performed a 5-fold cross-validation in BeatAML dataset for the models predicting GE and genetic variants data. The results in **Figure 31** show that the predictions do not significantly change when varying the type of

input, except in the Multinomial Lasso, where the use of gene expression significantly increased the precision of the method and in the Lasso, where it significantly decreased the sensitivity of the method. This analysis was also performed training in BeatAML and predicting in GDSC with the mutational and GE models. For which, **Appendix 3: Figure S14** showed no statistical significant difference in model sensitivity for any of the methods.

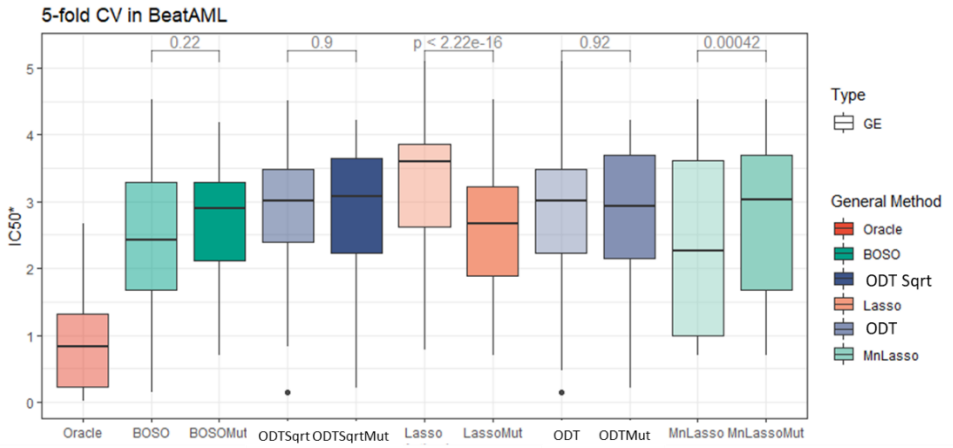


Figure 31: Using GE data over Mutational Data does not improve Method Precision. We compared for each of the algorithms the accuracy in response using Mutational and Gene Expression data as input. Distributions plotted in a lighter color are the responses obtained by each of the methods when using Gene Expression, whereas distributions plotted in darker non-transparent colors are the responses of the methods obtained when using mutational data.

Explainability: tree-like methods (MOM and ODT) require much less variables than any other methods

To measure the explainability of the method, we trained the models with the BeatAML dataset and checked the number of variables that each model required to make the predictions. Results are included in **Figure 32.a**. Remarkably, MOM and ODT use less than 5 variables, almost ten times less than the rest of the methods. BOSO, Multinomial Lasso, and Lasso use more than 30 variables. Among them, BOSO (with 33 variables) is the method that requires less variables. BOSO builds a linear model for each of the drugs. Each of the results (as occurs in Lasso) are sparse: it requires only 5 variables to predict drug response for some drugs. Since these variables are not identical for every drug, in the end, it requires 33 variables to make the predictions. Multinomial Lasso and Lasso were coded to preserve the same variables for predicting over all the drugs. BOSO did not implement this option.

Regarding the KRL method the number of variables it does not provide automatic feature selection but use regularization methods. Thus, all the 69 gene variants are used.

ODT and MOM output the decision criteria it the form of a decision-tree. The main difference between ODT and MOM decision trees is their structure, ODT's tree structure have several branches where drugs for each of them. MOM's tree structure is linear, it is divided into different sequential steps, each of them defined by a biomarker, and there is a drug recommendation on every step. Regression-based methods (BOSO, Lasso and Multinomial Lasso) provide the weights for each of the biomarkers on each of the drugs. Therefore, it is possible to check which are the critical biomarkers for each drug. KRL use kernels to guess the proper treatment. In this case, it is much more complex to understand which are the key genomic variants for the recommendation system.

Implementability: Optimal Decision Trees and MOM are the most prone to clinical practice and ODT the least computing time consuming

We also considered the easiness to implement the methods in wet lab or even clinical practice according to four different points: i) the feasibility for wet-lab validations, ii) the consideration of the physician's experience, iii) the generation of a clinical guideline, and iv) the computational implementation.

Tree-based models require less biomarkers than regression models or KRL. In addition, there are few operations to perform the predictions that can be done directly by hand. On the contrary, regression models and KRL require more genes and a computer-based environment to perform the drug assignation.

Regarding the computational burden of each of the methods, all the methods need to be trained in different software environments such as R or Python. Once trained, the tree-based models directly provide a guideline that do not require the environment anymore. We have timed the training process of the 6 models (**Figure 32.b**) using Mutational data and Gene Expression (where possible). ODT is the fastest method to train (0.05 seconds for training using mutational data and less than 5 seconds using gene expression data). Multinomial requires around 15 second using either mutational data or gene expression data. Lasso lasts 10 and 100 seconds using mutational and gene expression data respectively. Finally, MOM, KRL and BOSO require several hours for training their models. MOM and KRL are not suitable for gene expression data so they have been excluded for the timing analysis with this data. Prediction time is similar (and negligible if compared with

training time) in all the 6 methods. Focusing on the installation, models based on MILP (BOSO, KRL, and MOM) require a complex installation of software (**Table 4**). They are also the most time-consuming methods. ODT, Multinomial, and Lasso, only require of R installation to run. All these conclusions that could lead to rank methods according to Interpretablility have been summed up in Table 5.

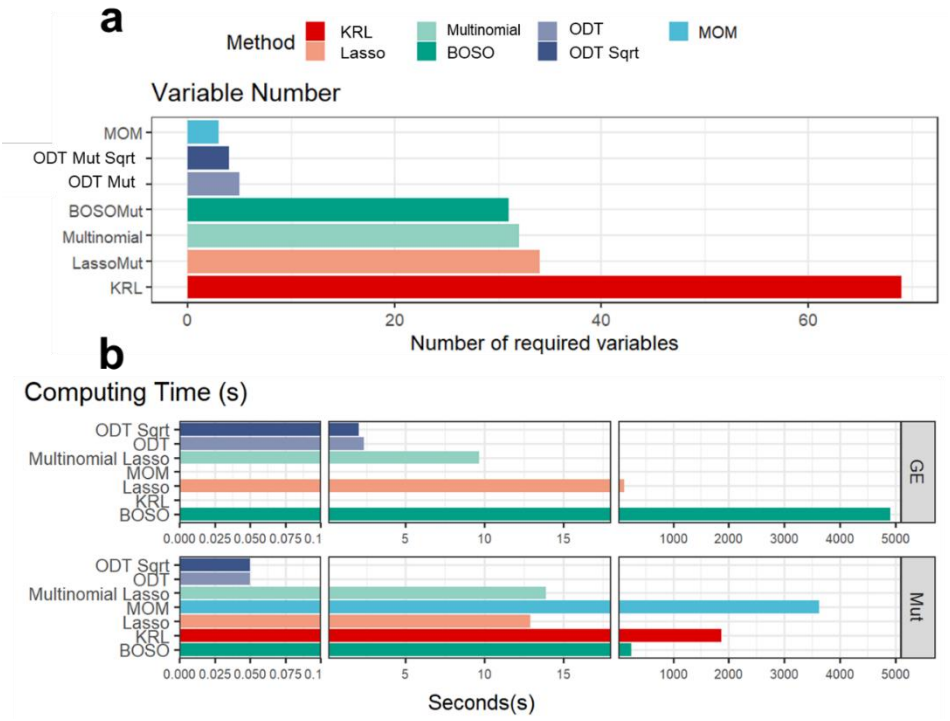


Figure 32: Variable Number and Computer timing performance comparisons. A) **Variable number comparisons.** All methods were trained in BeatAML cohort, after the training process we extracted the number of non-zero weighted input variables that each model requires for making the predictions. The horizontal axis shows the number of variables required by each method. B) **Computer timing comparison**. We measured the training time that each model requires using genetic variants (lower plot) or gene expression (upper plot) as input, time is shown in seconds in the horizontal axis.

Method	*Multi-omics*		*Explainability*			*Implementability*			
	Gene Expression	Genetic Variants	Small number of Variables	Understandable Method	Outputs Decision Criteria	Easy to Validate	Considering Experience	Clinical Guideline	Computational Burden
MOM	∅	*	***	*	*	***	***	***	∅
ODT	*	*	***	*	*	***	***	***	***
Multinomial	*	*	∅	*	∅	*	∅	∅	***
Lasso	*	*	∅	*	∅	*	∅	∅	*
BOSO	*	*	∅	*	∅	*	∅	∅	∅
KRL	∅	*	∅	*	∅	*	∅	∅	∅

Table 5: Table containing the interpretability comparisons for each method. The values No(∅), Yes(*), and Very(***), reflect whether the method does not, fulfills or greatly fulfills -respectively-, the conditions mentioned in the columns. Gene Expression refers to whether the method is suitable for this type of omics data, Genetic Variants refer to whether the method is suitable for this type of omics data. Explainable, refers to whether the method provides a reference of variable importance that could lead to a reasoning of the classification, Understandable Method, is defined as the logical explanation of the method, i.e., it does not rely on random selection, Outputs Decision Criteria, refers to whether the method provides the decision criteria used in the assignation, Easy translation, refers to the ability of the method to be wet-lab validated, to enhance the physician or researcher experience and output a therapeutic strategy for ulterior patients.

Discussion of Section 3

In this work we have selected four precision medicine methods –MOM, BOSO, Lasso, and KRL– and developed two additional ones –Optimal Decision Trees and Multinomial Lasso–, to compare them regarding their interpretability. We performed six quantitative comparisons and four qualitative comparisons. All the methods were similar in terms of accuracy. However, MOM and Optimal Decision Trees were the most interpretable and easy to implement.

PM is a topic that is being widely addressed and there are new algorithm proposals. It may seem surprising that we included only four of them in this comparison and, indeed, we developed two additional ones. A systematic review of all the methods --cited in the Introduction section-- included Machine Learning (ML) methods (using either deep learning, neural networks, support vector machines, random forests, etc). Among the 24 methods that used ML for making their predictions, only 10 were explainable. Of those 10, 5 of them did not solve the "patient-centred" problem: assign the proper drug to each patient. We ended up with MOM, BOSO, KRL, and LOBICO, and added Lasso as a control of a traditional approach in the ML field. LOBICO approach, that was also tested on this dataset elsewhere [139] is drug-centred and, since the output variable is discrete, it cannot be transformed into the patient-centred problem and not suitable for this comparison [139]. We developed two additional methods, both patient-centred, with two different approaches: regression (Multinomial) and tree classification (ODT).

In this work, we have defined Interpretability splitting it into four main concepts: Accuracy, Multi-omics capacity, Explainability, and Implementability. An interpretable PM method should be accurate and understandable by the common researcher or clinician. Accuracy is strictly necessary: if a method is not accurate, it becomes irrelevant despite being easy to understand. Multi-omics capacity, measures the robustness of the method to adapt into different data sources, that could be essential for new lines of research. Explainability is also essential, it should show the reasoning for reaching the results. Finally, the ease of implementation defines the ability of the method to incorporate the clinician experience and provide an easy technical usage.

We focused on a specific sensitivity value named IC50*. This metric was previously described in [2] or in [137] and is a normalization of the logarithm of the IC_{50}. Normalizing the IC_{50} --or other sensitivity value-- is crucial as the best drug is not necessary the drug

with the lowest IC_{50} value. In fact, a drug with a low IC_{50} can be toxic for the patients. Toxic drugs tend to have low IC_{50} values in all tissues, whereas the focus must be set on drugs with differential sensitivity for different tissues. Normalizing the logarithm of the IC_{50} by removing the mean sensitivity value of the drug in all patients, preserves the sensitivity profiles of the drugs and penalizes drugs that are sensitive or resistant in all tissues. The dosage of drugs with higher IC_{50} can be adjusted to obtain drug effectiveness. We trained all the models with the normalized version, IC50*, to avoid the aforementioned problems.

All the methods predict reasonably well in terms of accuracy. The 5-fold cross-validation and the independent cohort validation, showed that the different methods had similar median and the differences were not statistically significant. The intragroup validation, showed that the regression-based models (KRL, BOSO, Lasso, and Multinomial) were not able to distinguish between responders and non-responders to a specific drug. This result is reasonable since these methods do not divide the patients according to responders or not responders to one biomarker, but cherry-picked patients based on weighted combinations of biomarkers. MOM, on the contrary, has a restriction within its model formulation that means that all patients with a biomarker that confers sensitivity to a specific drug should be treated in that current step [2]. Nevertheless, not having a successful intragroup validation does not invalidate the model.

The multi-omics suitability is a “hot-topic” in PM, as there is not a current gold standard based on which type of data is more accurate when predicting drug response. Some models use genetic variants to promote interpretability, whereas other use gene expression or integrated omics for improving accuracy. In this work, we compared the accuracy changes when training and predicting on gene expression and genetic variants separately, and found almost no significant statistical difference in the performance. Drug response is mediated in living beings by complex regulatory and metabolomic processes that are most likely to be solved using an integrated omics input, instead of just one single omics. However, the more complex the model becomes, the less interpretable it is.

Regarding the explainability, we included also a qualitative comparison since focusing only on the number of variables, does not justify that the method is understandable. It is also desirable that the method can provide decision criteria, i.e. a complete process that a clinician can follow and understand. This consideration has paramount importance if it is to be approved by regulators for medicine [49,143]. Consequently, we focused on the ease to understand the output of the methods, and the explainability of the results. We defined

the latter, as the ability of the method to rank the input variables in order of importance for drug assignation. Of course, a smaller number of variables is easier to understand. The tree-based models require less than six variables, and it increases up to five times in the regression-based models. BOSO, however, uses only five variables to predict response of just one drug, but when translated into a patient-centred approach, the total number of variables used for predicting in all drugs is equal to 33. For Lasso and Multinomial, the number of variables has been optimized to predict response in all drugs. KRL, however, did not consider this parameter and uses all variables provided as input to make the predictions, being the less explainable method.

Implementability is a concept easier to understand, as it directly facilitates the clinical translation. Most of the implementability comparisons were qualitative, but we performed a technical comparison of the methods regarding its computational burden. There we showed that MOM, which was leading the accuracy comparisons, is the most time consuming up to 2.5 hours on a normal machine, and it is the model that requires the highest number of software environments: R, Python, and CPLEX need to be installed in the machine (and related to each other). It is the most resource consuming. However, if compared against ODT, which achieved similar accuracy performance, the latter only requires R and the algorithm is trained, even using gene expression, in less than 5 seconds. Besides, ODT is more explainable than MOM, because the method is easier to understand, although it is quite similar to MOM regarding the other explainability and implementability criteria.

Nonetheless, Multinomial and Lasso are also explainable, if not compared against other methods, and there are additional functions -not defined in the methods themselves- that can be applied to extract the algorithm reasoning or decision criteria. Also, linear models can be understandable as the βs reflect the variable importance for prediction.

To summarize, in this work we defined a quantitative method for evaluating the interpretability of a given machine learning method, because, as previously discussed, accuracy is not the only important factor in the complex field of health. The defined criteria can serve as a guide for developing new translational methods aimed at solving precision medicine problems

.

Discussion and Conclusions

The defined objectives have been identified, developed, and achieved within this PhD thesis paper.

The first section provided a computational approach to identify LEDs with increased predictive power and was validated both in silico and in vitro. Predictions of LEDs from functional screens can be dramatically improved by incorporating the “HUb effect in Genetic Essentiality” (HUGE) of gene alterations. We analyzed three recent genome-wide loss-of-function screens -Project Score, CERES score and DEMETER score- identifying LEDs with 75 times larger statistical power than using state-of-the-art methods. Using acute myeloid leukemia, breast cancer, lung adenocarcinoma and colon adenocarcinoma as disease models, we validate that our predictions are enriched in a recent harmonized knowledgebase of clinical interpretations of somatic genomic variants in cancer (AUROC > 0.87). Our approach is effective even in cancers with large genetic heterogeneity such as acute myeloid leukemia, where we identified LEDs not recalled by previous pipelines, including *FLT3*-mutant genotypes sensitive to *FLT3* inhibitors. Interestingly, *in-vitro* validations confirm lethal dependencies of either *NRAS* or *PTPN11* depending on the *NRAS* mutational status.

The second section presented a novel explainable method -called multi-dimensional module optimization (MOM)- that associates drug screening with genetic events, while guaranteeing that predictions are interpretable and robust. We applied MOM to an AML cohort of 319 *ex-vivo* tumor samples with 122 screened drugs and WES. MOM returned a therapeutic strategy based on the *FLT3*, *CBFβ-MYH11*, and *NRAS* status, which predicted AML patient response to Quizartinib, Trametinib, Selumetinib, and Crizotinib. We successfully validated the results in three different large-scale screening experiments.

The third section compared six different machine learning methods to provide guidance for defining interpretability by focusing on: Accuracy, Multi-omics Capability, Explainability, and Implementability. Our selection of algorithms included tree, regression, and kernel-based methods. We also included two novel explainable methods in the comparison. There were no significant differences in accuracy when comparing methods or when using gene expression instead of mutational status as input to these methods. This allowed us to concentrate on the current intriguing challenge: model comprehension, and ease of use. We discovered that the tree-based methods were the most interpretable of those tested.

Thus, the objectives that were stated at the beginning of the writing of this dissertation report have been fulfilled. However, one of the main limitations of the development of this

work is the subjectivity regarding traslationality to clinical practice or research. The issues that exist on the implemention of these models are not only raised by regulatory agencies, e.g. FDA or EMA. They also directly depend on the research centre surroundings, researcher's expertise, the availability of technical resources or funds for data collection.

The primary goal of the interpretability criterion is to not decrease the accuracy of the method and by secondary goal to reduce application costs and staff training time. Accuracy is a must since the patient will bear the brunt of the consequences. However, the criterion we established is debatable. We also believe that the correct quantitative definition of interpretability should be reached through consensus with a specialized committee comprised of physicians and researchers from various backgrounds and research institutions, e.g., regulatory agencies, bioinformaticians, or physicians.

We have defined and validated several methods that solve two major problems in the state-of-the-art. The HUGE method successfully solves the problem of multiple hypothesis correction for treatment search. In turn, this method has been implemented in an algorithm known as MOM, which generates a very simple and accurate treatment pipeline given a tumor type. It is one of the first methods to solve the assignment problem by returning a hierarchical and sequential treatment guideline.

This fact distinguishes it from other explainable methods that return the variable hierarchy but do not explain the algorithm's reasoning, which is critical for regulatory validation. Finally, the ODT algorithm significantly improves the limitations presented in MOM because it achieves a similar result in terms of accuracy using a recursive optimization method rather than MILP, making it much more implementable. Both methods achieve results that are very close to the state of the art in terms of accuracy, and both methods produce a simple and hierarchical therapeutic strategy.

The evolution of ODT has allowed us to demonstrate that accuracy does not always go hand in hand with statistically complex models, but that it is critical to consider the model's objective or implementation requirements before defining it. Regarding the approach to solving the assignation problem, we have seen that the "patient-centred" approach allows for more understandable results. Furthermore, it is the most logical solution to the problem.

As a result, the methods HUGE, MOM, ODT, and Multinomial—which were created alongside other members of the research team and the PhD thesis' co-directors—are introduced into the science knowledge. Although these techniques were applied to AML,

they can be applied to any disease as long as the types of data that each of them needs as input are available. The quantitative metric of interpretability is likewise contributed to the state of the art as a strategy for developing new Precision Medicine methods.

Finally, the following are proposed as future research directions:

1. The incorporation of new omics, including multi-omics specifically, methylation or chromatin access, to increase the versatility of methods in various fields of research. It would be necessary to adapt the methods, particularly HUGE and MOM, because they do not allow for this adaptability.
2. To improve the accuracy of ODT, this method employs a recursive optimization for the generation of the decision tree, but it could be extended to hundreds of trees mimicking the random forest technique. Although the method would be less interpretable, the accuracy can improve and will also return alternative treatments for each patient.
3. We also propose an in vitro validation of the classification obtained by MOM, Multinomial, or ODT in AML, which was not done during the doctoral thesis due to a scarcity of time. This validation, which was completed with HUGE, could result in a fresh approach in disease research since the sensitivity of patients to recommended drugs or the effect of the proposed biomarkers could be studied. Contributing to the much-needed new treatments for this type of cancer.
4. After in vitro testing with AML were successful, the use of these algorithms could be transferred to the study of therapies for other diseases, adapting the methods to the requirements of the researcher or practitioner.

Appendix 1. Extended Information of HUGE method to predict Lethal Dependencies

Appendix 1: Extended Information of HUGE method to predict Lethal Dependencies

Protocol for In-vitro validation

Cell culture

The AML cell lines HL-60, HEL, MV4-11 and OCI-AML3 were maintained in culture in RPMI-1640 medium supplemented with 10% fetal bovine serum (Gibco, Grand Island, NY), penicillin/streptomycin (BioWhitaker, Walkersvill, MD) at 37 °C in a humid atmosphere containing 5% CO2. All cell lines were tested for mycoplasma (MycoAlert Sample Kit, Cambrex) and were authenticated by performing a short tandem repeat allele profile.

Cell transfection

Cells were passaged 24 hours before nucleofection, and cells for nucleofection were in their logarithmic growth phase. The transfection of siRNAs was done with the Nucleofector II device (Amaxa GmbH, Köln, Germany) following the Amaxa guidelines. Briefly, 1×10^6 of HL-60, HEL, MV4-11 and OCI-AML3 cells were resuspended in 100 µL of supplemented culture medium or solution V in the case of HL-60 cells, with 75nM of NRAS or PTPN11 siRNAs or Silencer Select Negative Control-1 siRNA (Ambion, Austin, TX) and nucleofected with the Amaxa nucleofector apparatus using programs A030 (HEL, MV4-11 and OCI-AML3) or T019 (HL-60). We used two different siRNAs against *NRAS* target (siNRAS A: GAACCACUUUGUAGAUGAA; siNRAS B: AAGGACAGTTGATACAAAA) and *PTPN11* (siPTPN11 A: AGAUGUCAUUGAGCUUAAA; siPTPN11 B: GAAAGAAGCAGAGAAAUUA) to demonstrate that the results obtained with siRNA nucleofection are not due to a combination of inconsistent silencing and sequence specific off-target effects. Silencer Select Negative Control-1 siRNA was used to demonstrate that the nucleofection did not induce non-specific effects on gene expression. Nucleofection was performed twice with a 24 hours interval. 48 h after the second nucleofection, the *NRAS* and *PTPN11* mRNA expression was analyzed by Q-PCR (*GUSB* was employed as the reference gene). Cell proliferation was analyzed 0, 2 ,4 and 6 days after two repetitive transfections. Transfection efficiency was determined by flow cytometry using the BLOCK IT Fluorescent Oligo (Invitrogen Life Technologies, Paisley, UK).

Cell proliferation assay

Cell proliferation was analyzed using the CellTiter 96 Aqueous One Solution Cell Proliferation Assay (Promega, Madison, W). This is a colorimetric method for determining the number of viable cells in proliferation. For the assay, 100 µL of nucleofected cells were

plated in 96 wells plates 0, 2, 4 and 6 days after the last nucleofection. Plates with suspension cells were centrifuged at 800 g for 10 minutes and medium was removed. Then, cells were incubated with 100 μL/well of medium and 20 μL/well of CellTiter 96 Aqueous One Solution reagent. The plates were incubated for 1-4 hours, depending on the cell line at 37 ºC in a humidified, 5 % CO2 atmosphere. The absorbance was recorded at 490 nm using 96-well plate readers until absorbance of control cells without treatment was around 0.8. The background absorbance was measured in wells with only cell line medium and solution reagent. First, the average of the absorbance from the control wells was subtracted from all other absorbance values. Data were calculated as the percentage of total absorbance of siRNA transfected cells/absorbance of control cells.

Quantitative-PCR (Q-PCR)

The expression of *NRAS* and *PTPN11* was analyzed by Q-PCR in HL-60, HEL, MV4-11 and OCI-AML3 AML cell lines. First, total mRNA was extracted with Trizol© Reagent 5791 (Life Technologies, Carlsbad, CA, USA) following the manufacturer instructions. RNA concentration was quantified using NanoDrop Specthophotometer (NanoDrop Technologies, USA). cDNA was synthesized from 1 μg of total RNA using the PrimeScript RT reagent kit (Perfect Real Time) (Cat No RR037A, TaKaRa) following the manufacturer's instructions. The quality of cDNA was checked by a multiplex PCR that amplifies *PBGD*, *ABL*, *BCR* and *β2-MG* genes. Q-PCR was performed in a QuantStudio 5 Real-Time PCR System (Applied Biosystems), using 20 ng of cDNA in 2 μL, 1 μL of each primer at 5μM (NRAS F: 5'-CGCACTGACAATCCAGCTAA-3'; NRAS R: 5'-CCAACAAACAGGTTTCACCA-3'; PTPN11 F: 5'-CGGAGCCTGAGCAAGGAG-3'; PTPN11 R: 5'-CTGCCTCCACACCAGTGATA-3'; GUSB F:5' gaaaatatgtggttggagagctcatt-3'; GUSB R: 5'- ccgagtgaagatccccttttta-3'), 6 μL of SYBR Green PCR Master Mix 2X (Cat No 4334973, Applied Biosystems) in 12 μL reaction volume. The following program conditions were applied for Q-RT-PCR running: 50 ºC for 2 min, 95 ºC for 60 s following by 45 cycles at 95 °C for 15 s and 60 °C for 60 s; melting program, one cycle at 95 °C for 15 s, 40 °C for 60 s and 95 °C for 15 s. The relative expression of each gene was quantified by the Log $2^{(-\Delta\Delta Ct)}$ method using the gene *GUSB* as an endogenous control.

Demonstration of the increased statistical power

Lemma: *Let us consider two methods that correct multiple hypothesis test, and let us consider that both methods provide a different number of positives for the same FDR. Then,*

the method that provides a larger number of positives has more statistical power. It is also more specific and sensitive.

The power or sensitivity of a statistical test is the probability that the test correctly rejects the null hypothesis H_0 when the alternative hypothesis H_1 is true. Its value is *TP/(TP+FN)*.

Let's consider that the estimation of the FDR is performed by two tests A and B and both tests have the same False Discovery Rate (20% for example). The FDR will be

$$FDR = \frac{FP_A}{TP_A + FP_A} = 1 - \frac{TP_A}{TP_A + FP_A} = \frac{FP_B}{TP_B + FP_B} = 1 - \frac{TP_B}{TP_B + FP_B} (1a)$$

The power of each test will be

$$PW_A = 1 - \beta_A = \frac{TP_A}{TP_A + FN_A} (1b)$$

$$PW_B = 1 - \beta_B = \frac{TP_B}{TP_B + FN_B} (1c)$$

Since both tests are performed on the same dataset, the number of true null hypothesis H_0 (FP + TN) and true alternative hypothesis H_1 (TP+FN) will be identical, i.e.,

$$FP_A + TN_A = FP_B + TN_B (1d)$$

$$TP_A + FN_A = TP_B + FN_B (1e)$$

Notice that the denominators of the expression of the power (eq (1b) and (1c)) are identical according to (1e).

The total number of positives returned by each test is $TP_A + FP_A$ and $TP_B + FP_B$. Let's assume that method A, returns more positives than method B, i.e.

$$TP_A + FP_A > TP_B + FP_B (2)$$

Using eq. (1a), and (2)

$$TP_A = (1 - FDR)(TP_A + FP_A) (3a)$$

And,

$$TP_B = (1 - FDR)(TP_B + FP_B) (3b)$$

Since (2), the righthand member of equation (3a) is larger than the righthand member of equation (3b) and therefore,

$$TP_A > TP_B (4)$$

As a result,

$$PW_A > PW_B$$

Corolary I. Since $PW_A = 1 - \beta_A$ the type II error using A is smaller than using B.

$$\beta_A < \beta_B$$

Corolary II. The type I error is

$$\alpha_A = \frac{FP_A}{FP_A + TN_A}$$

And the sensitivity is:

$$1 - \alpha_A = \frac{TN_A}{FP_A + TN_A}$$

By (1e) and (4), it is straightforward to conclude that

$$\alpha_A < \alpha_B$$

Therefore, the method that provides a larger number of positives outperforms the other both in terms of specificity and sensitivity (or type I and type II errors).

Appendix 2. Extended Information of MOM performance and pipeline

Appendix 2: Extended Information of MOM performance and pipeline

Extended Information Beat AML Cohort

Biomarker Analysis

We performed an extensive biomarker analysis to characterize the WES from Beat AML[144] cohort using the maftools [145] R package (version 2.10.05). We intended to understand the different processes that are regulating the biomolecular characterization of this cohort. To do it, we plotted several figures that contain the relevant information concerning the genetic profile of the patients in the cohort.

Figure S1 shows that 88.16% of patients have at least one genetic variation, the majority of these variants are missense, and *DNMT3A*, *NPM1* and *NRAS* are the most commonly mutated genes in this cohort. Moreover, from these variants, the majority correspond to single Nucleotide Variants (SNVs) with the signature C>T that is quite frequent in malignant cancer types followed by C>A which is associated with environmental exposure[96] (**Figure 15**). We appreciated that the median of genetic variants per patient is 8 variants and that only *FLT3* and *SRSF2* had insertions i.e. *FLT3-ITD*.

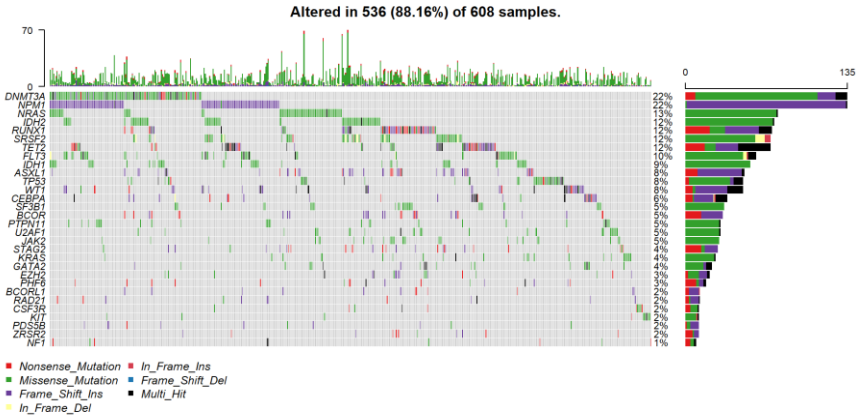


Figure S1: Mutational Status of Beat AML cohort. This plot shows the different types of genetic variants, the lateral barplot shows the sum of all the genetic alterations in all patients and colours the type of variant. In the horizontal axis we have the individual patients' information, showing some patients having up to 70 co-occurring mutations. Of the 608 patients, 538 had at least one genetic variant (88.16%).

We tried to understand more in-depth the SNVs changes and classified them into transitions (two-ring purines or one-ring purines changes) and transversions (changes of

purines for pyrimidines) we discovered that in this cohort the transitions are more frequent with the most common transition being C>T, followed by a transversion C>A (**Figure S2**).

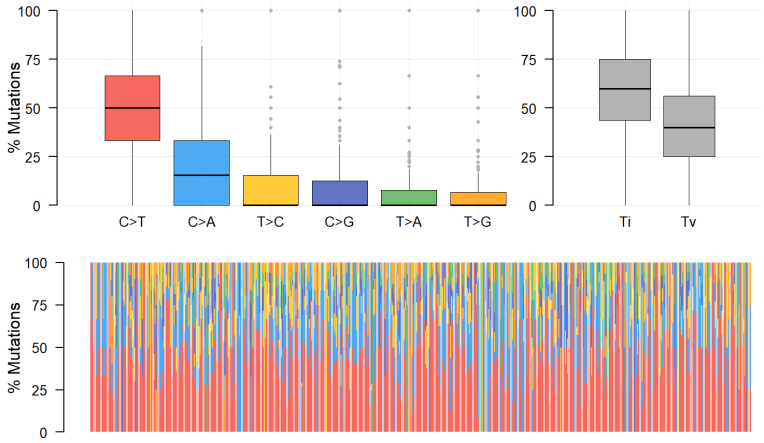


Figure S2: Transitions(Ti) and Transversions (Tv) landscape in Beat AML cohort. Ti vs. Tv plot shows the number of the transitions and transversions showing that even when transversions seem most probable to occur, transitions are more present in this cohort. Boxplot showing overall distribution of six different conversions and as a stacked barplot showing the fraction of conversions in each sample. The most common transition is C>T, followed by a transversion of C>A.

Using the Whole Exome Sequencing (WES) data provided in the Beat AML cohort we analysed co-occurrence and mutual exclusivity of genetic variants at the gene level. We appreciated that *FLT3* and *NPM1* variants are co-occurrent (p-value <0.05), and *FLT3* and *TP53* (p-value <0.05) and NRAS and IDH2 are mutually exclusive respectively (p-value <0.05) (**Figure 22**).

AML is a highly heterogenetic disease and consequently, genetic translocations are included as possible biomarkers. All translocations can be identified by a gene fusion product: inv(16) with *CBFβ-MYH11*, inv(3) with *RPN1-EVI1*, t(9;11) with *MLLT3-MLL*, and t(8;21) with *RUNX1-RUNX1T1*. From these translocations, inv(16) appears in co-occurrence with *FLT3*, *KIF20B*, and *ADAMTS7* variants. Whereas t(9;11) can appear with *NRAS* variants and inv(3) with *KIT* variants (**Figure S3**).

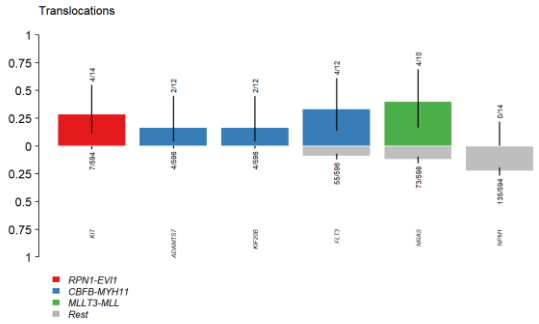


Figure S3: Translocations and SNVs. All translocations can be identified by a gene fusion product: inv(16) with *CBFβ-MYH11*, inv(3) with *RPN1-EVI1*, t(9;11) with *MLLT3-MLL*, and t(8;21) with *RUNX1-RUNX1T1*. From these translocations, inv(16) appears in co-occurrence with *FLT3*, *KIF20B*, and *ADAMTS7* variants. Whereas t(9;11) can appear with *NRAS* variants and inv(3) with *KIT* variants

Finally, we addressed the biological consequences of this mutational landscape by interrogating the alteration of the most common oncogenic pathways (**Figure S4**). We saw that RTK-RAS is the most affected pathway, having an alteration in 31 out of 85 genes and it is present in 237 out of 608 patients. Remarkable alterations include NOTCH, WNT, MYC, TP53 and TGF-β pathways. We included a summary by the patient showing the complete pathway alterations (**Figure S5**).

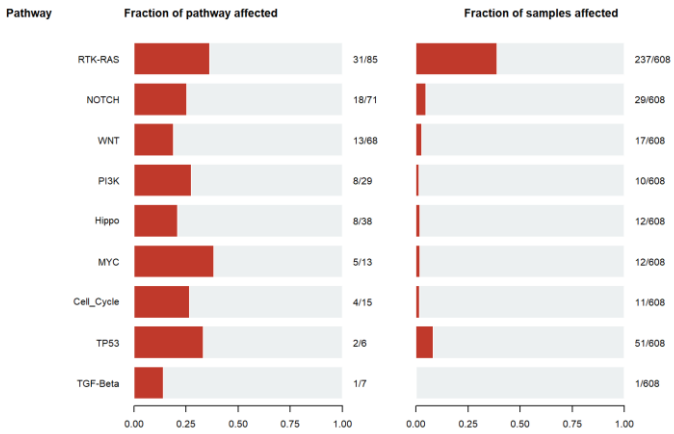


Figure S4: Oncogenic Signalling Pathways altered in Beat AML cohort. The barplot on the left represents the proportion of genes that are altered in the pathway, whereas the barplot on the right, shows the proportion of samples that are having an alteration in that pathways.

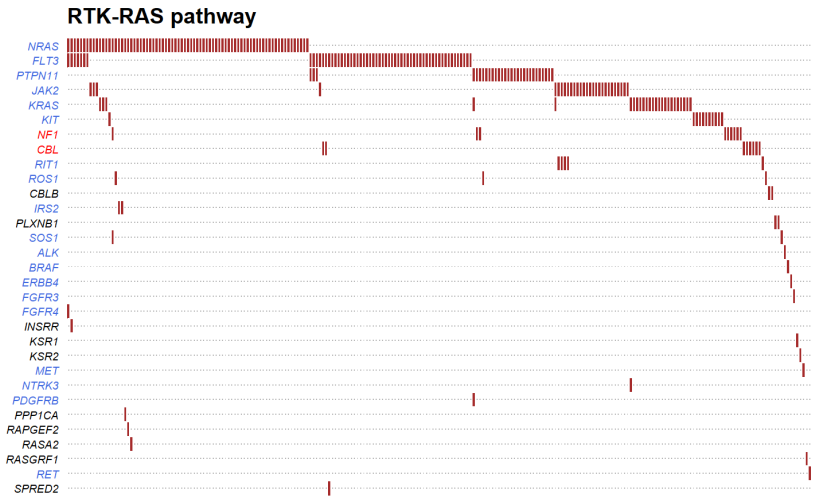


Figure S5: RTK-RAS pathway alterations. In the Y-axis all genes are included in the pathway, in blue the oncogenes and red tumour suppressor genes. In the X-axis all the samples with RTK-RAS altered and the red marks show the pathway genes altered for each sample.

Additional Results on GO analysis

We tried to understand more in-depth each of the subgroups whose treatment according to MOM is different. The methodology is described in the Methods section of the main manuscript. This section includes the results from the enrichment analysis based on gene expression including all the functions that appeared as statistically enriched from the two conditions up and downregulated.

We also included **Table S1-Table S8**, contained detail statistical information of the top 10 significant upregulated and downregulated ontologies for each subgroup.

Appendix 2: Extended Information of MOM performance and pipeline

Table S1:Top 10 GO upregulated FLT3^{Mut}-Quizartinib subgroup

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0007389	pattern specification process	33/304	314/13290	2,15E-13	6,86E-10
GO:0003002	regionalization	27/304	245/13290	1,22E-11	1,95E-08
GO:0044782	cilium organization	30/304	308/13290	1,97E-11	2,10E-08
GO:0048704	embryonic skeletal system morphogenesis	15/304	70/13290	4,05E-11	3,23E-08
GO:0098840	protein transport along microtubule	14/304	61/13290	6,83E-11	3,63E-08
GO:0099118	microtubule-based protein transport	14/304	61/13290	6,83E-11	3,63E-08
GO:0060271	cilium assembly	28/304	294/13290	1,63E-10	7,42E-08
GO:0048598	embryonic morphogenesis	35/304	449/13290	2,05E-10	8,18E-08
GO:0048706	embryonic skeletal system development	16/304	92/13290	2,49E-10	8,86E-08
GO:0042073	intraciliary transport	12/304	46/13290	3,21E-10	1,03E-07

Table S2. Top 10 GO downregulated FLT3^{Mut}-Quizartinib subgroup.

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0042119	neutrophil activation	59/397	471/13290	2,68E-21	9,86E-18
GO:0036230	granulocyte activation	59/397	477/13290	5,09E-21	9,86E-18
GO:0043312	neutrophil degranulation	57/397	459/13290	2,10E-20	2,71E-17
GO:0002283	neutrophil activation involved in immune response	57/397	462/13290	2,88E-20	2,79E-17
GO:0002446	neutrophil mediated immunity	57/397	472/13290	8,12E-20	6,30E-17
GO:0043299	leukocyte degranulation	58/397	499/13290	2,40E-19	1,55E-16
GO:0002430	complement receptor mediated signaling pathway	6/397	11/13290	2,78E-07	0,000154203
GO:0097529	myeloid leukocyte migration	20/397	185/13290	6,75E-07	0,000327246
GO:0060326	cell chemotaxis	24/397	257/13290	7,88E-07	0,000339463

GO:0071621	granulocyte chemotaxis	14/397	104/13290	2,47E-06	0,000957551

Table S3. Top 10 upregulated GO in Inv(16)-Trametinib subgroup

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0001525	angiogenesis	35/369	422/13290	7,10E-09	2,58E-05
GO:0048514	blood vessel morphogenesis	36/369	492/13290	1,08E-07	0,000195429
GO:0043542	endothelial cell migration	19/369	182/13290	7,33E-07	0,000853836
GO:0031589	cell-substrate adhesion	25/369	299/13290	9,40E-07	0,000853836
GO:0001935	endothelial cell proliferation	15/369	132/13290	3,81E-06	0,002400726
GO:0001667	ameboidal-type cell migration	26/369	345/13290	3,97E-06	0,002400726
GO:0002040	sprouting angiogenesis	13/369	104/13290	5,88E-06	0,002540273
GO:0007160	cell-matrix adhesion	18/369	194/13290	7,74E-06	0,002540273
GO:0010631	epithelial cell migration	21/369	254/13290	8,41E-06	0,002540273
GO:0090132	epithelium migration	21/369	254/13290	8,41E-06	0,002540273

Table S4. Top 10 downregulated GO for Inv(16)-Trametinib subgroup

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0042743	hydrogen peroxide metabolic process	8/338	46/13290	1,81E-05	0,061785701
GO:0015669	gas transport	5/338	16/13290	3,58E-05	0,061785701

Table S5. Top 10 Go upregulated for $NRAS^{Mut}$-Selumetinib subgroup

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0036230	granulocyte activation	35/348	477/13290	3,57E-08	6,48E-05
GO:0043312	neutrophil degranulation	34/348	459/13290	4,50E-08	6,48E-05
GO:0002283	neutrophil activation involved in immune response	34/348	462/13290	5,27E-08	6,48E-05

GO:0042119	neutrophil activation	34/348	471/13290	8,40E-08	6,53E-05
GO:0002446	neutrophil mediated immunity	34/348	472/13290	8,84E-08	6,53E-05
GO:0043299	leukocyte degranulation	35/348	499/13290	1,09E-07	6,68E-05
GO:0007179	transforming growth factor beta receptor signaling pathway	18/348	167/13290	3,83E-07	0,000202153
GO:0071560	cellular response to transforming growth factor beta stimulus	18/348	210/13290	1,04E-05	0,004387859
GO:0045765	regulation of angiogenesis	20/348	252/13290	1,07E-05	0,004387859
GO:0030512	negative regulation of transforming growth factor beta receptor signaling pathway	10/348	70/13290	1,31E-05	0,00446212

Table S6. Top 10 GO downregulated for $NRAS^{Mut}$-Selumetinib subgroup

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust
GO:0022613	ribonucleoprotein complex biogenesis	23/328	432/13290	0,000458011	0,539178113
GO:0060571	morphogenesis of an epithelial fold	4/328	18/13290	0,000847547	0,539178113
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	19/328	343/13290	0,000887653	0,539178113
GO:0000398	mRNA splicing, via spliceosome	19/328	343/13290	0,000887653	0,539178113
GO:0000375	RNA splicing, via transesterification reactions	19/328	346/13290	0,000984103	0,539178113
GO:0008380	RNA splicing	22/328	431/13290	0,001055832	0,539178113
GO:0009954	proximal/distal pattern formation	4/328	20/13290	0,00129069	0,564953245
GO:0006397	mRNA processing	23/328	478/13290	0,001779044	0,681374007
GO:0060601	lateral sprouting from an epithelium	3/328	11/13290	0,002121238	0,722163542
GO:0016331	morphogenesis of embryonic epithelium	9/328	121/13290	0,003057585	0,936844029

Table S7. Top 10 upregulated GO for Rest-Crizotinib subgroup

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0050808	synapse organization	24/356	335/13290	1,20E-05	0,041169411
GO:0001906	cell killing	13/356	130/13290	4,55E-05	0,053426939
GO:0099173	postsynapse organization	14/356	149/13290	4,68E-05	0,053426939
GO:0031640	killing of cells of other organism	6/356	35/13290	0,000297916	0,176246865
GO:1900120	regulation of receptor binding	5/356	23/13290	0,000303058	0,176246865
GO:0045216	cell-cell junction organization	13/356	157/13290	0,000309024	0,176246865
GO:0008016	regulation of heart contraction	13/356	166/13290	0,000528978	0,244860946
GO:0060047	heart contraction	14/356	189/13290	0,000572439	0,244860946
GO:0003015	heart process	14/356	198/13290	0,000905207	0,344179825
GO:0030198	extracellular matrix organization	18/356	297/13290	0,001098946	0,355146597

Table S8. Top 10 downregulated GO for Rest-Crizotinib subgroup

ID	Description	GeneRatio	BgRatio	P-value	P-adjust
GO:0009063	cellular amino acid catabolic process	13/307	105/13290	8,60E-07	0,002908716
GO:1901606	alpha-amino acid catabolic process	11/307	86/13290	4,38E-06	0,007405178
GO:1901605	alpha-amino acid metabolic process	13/307	161/13290	9,25E-05	0,104297732
GO:0009081	branched-chain amino acid metabolic process	5/307	24/13290	0,000188586	0,11250965
GO:0006520	cellular amino acid metabolic process	17/307	271/13290	0,000190288	0,11250965
GO:0006790	sulfur compound metabolic process	18/307	299/13290	0,000206601	0,11250965
GO:0072350	tricarboxylic acid metabolic process	4/307	14/13290	0,000232733	0,11250965
GO:0015936	coenzyme A metabolic process	4/307	15/13290	0,000311615	0,131813013
GO:0009953	dorsal/ventral pattern formation	7/307	58/13290	0,000359022	0,13499231
GO:0120031	plasma membrane bounded cell projection assembly	23/307	464/13290	0,000490436	0,143965241

Appendix 2: Extended Information of MOM performance and pipeline

Appendix 3. Extended Information for Precision Medicine Method Comparison

Appendix 3: Extended Information for Precision Medicine Method Comparison

Supplementary Figures

Figure S6: Statistical significance between the different therapeutic strategies using BOSO in BeatAML.

Figure S7: Statistical significance between the different therapeutic strategies using BOSO in GDSC.

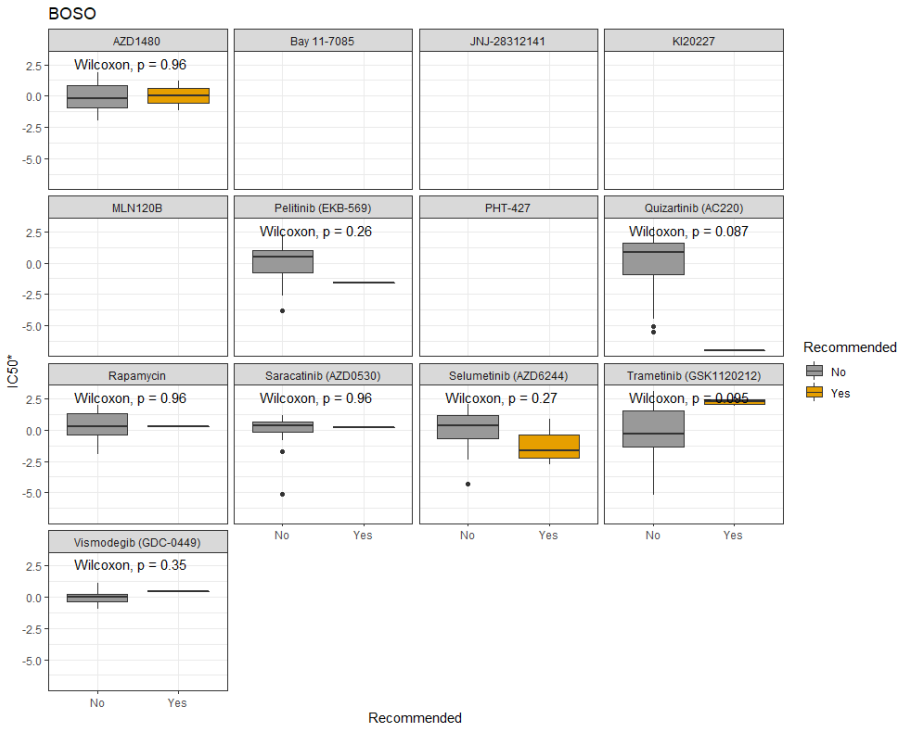


Figure S8: Statistical significance between the different therapeutic strategies using Lasso in BeatAML.

Figure S9: Statistical significance between the different therapeutic strategies using Lasso in GDSC.

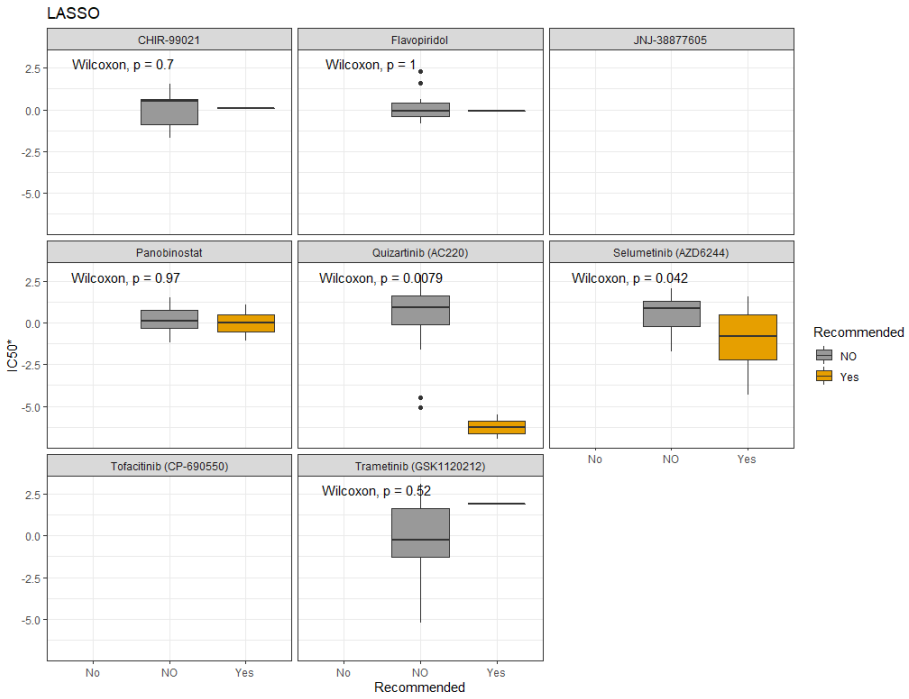


Figure S10: Statistical significance between the different therapeutic strategies using Multinomial in BeatAML.

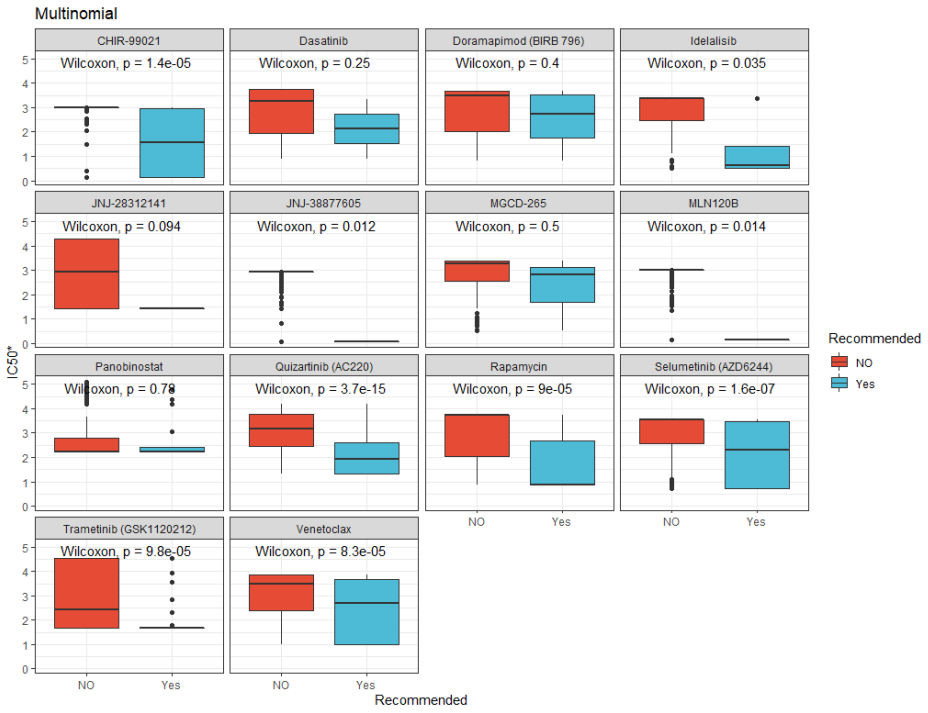


Figure S11: Statistical significance between the different therapeutic strategies using Multinomial in GDSC.

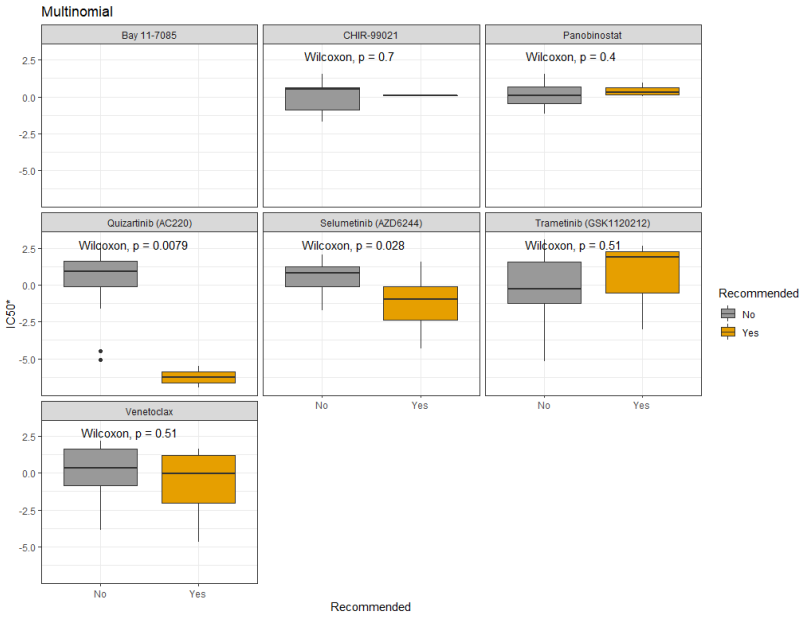


Figure S12: Statistical significance between the different therapeutic strategies using KRL in BeatAML.

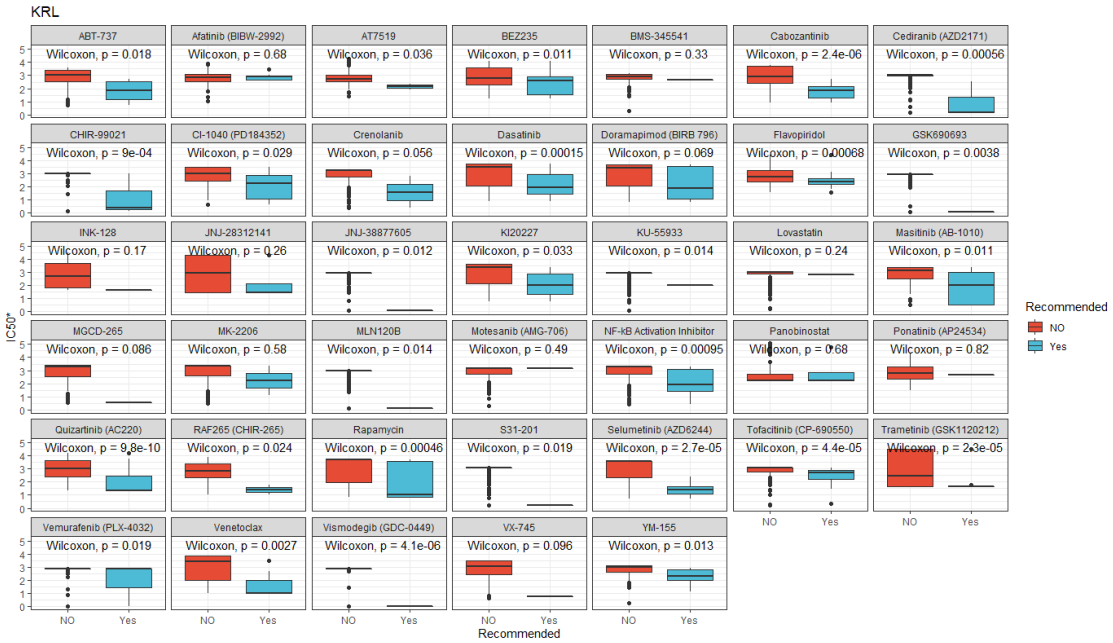


Figure S13: Statistical significance between the different therapeutic strategies using KRL in GDSC.

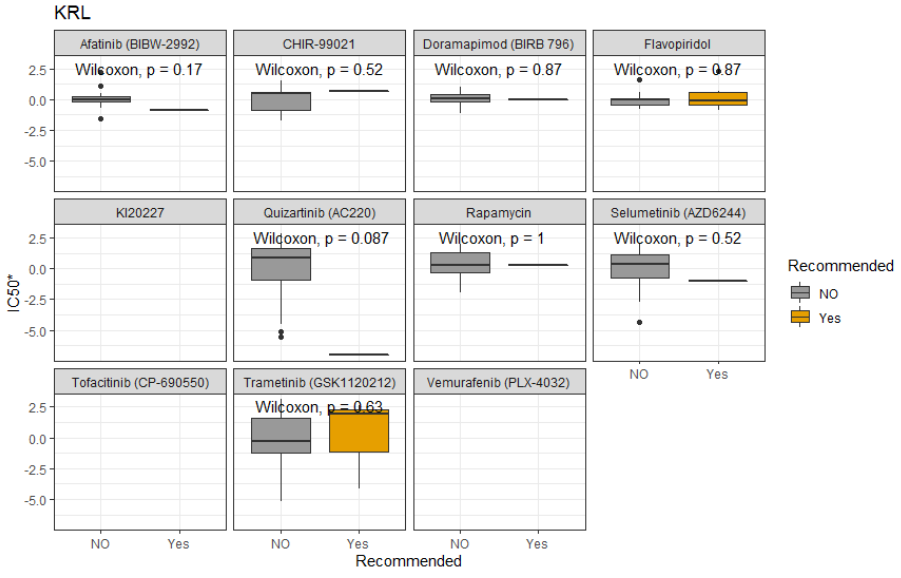
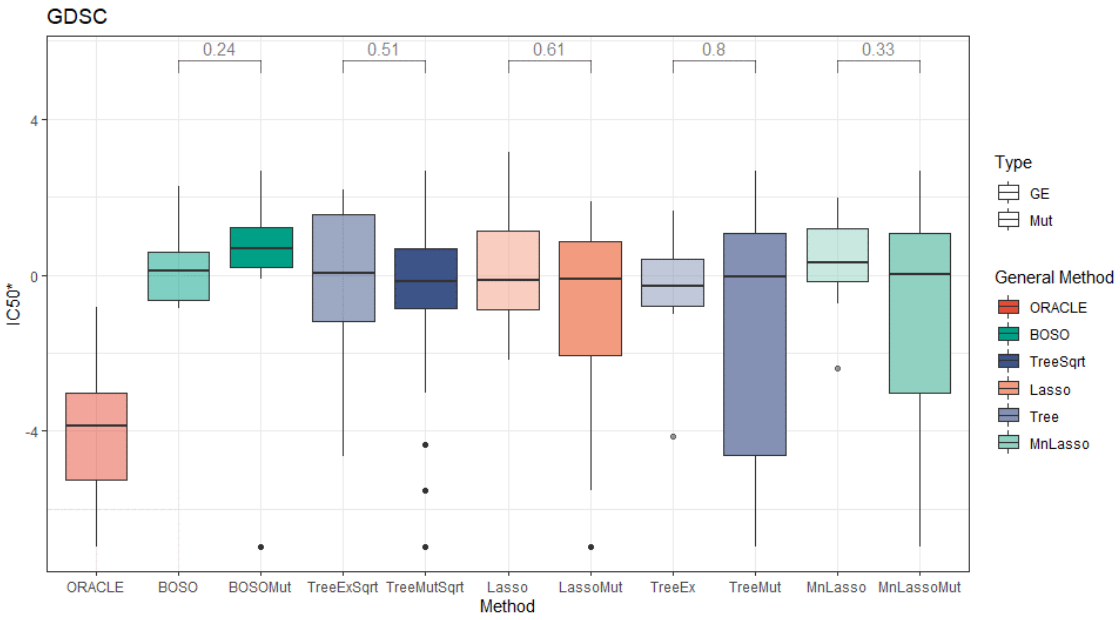


Figure S14: GE vs Mut in GDSC



Author Publications

OBTAINED FROM THIS WORK

GIMENO, MARIAN, EDURNE SAN JOSÉ-ENÉRIZ, ANGEL RUBIO, LEIRE GARATE, ESTÍBALIZ MIRANDA, CARLOS CASTILLA, XABIER AGIRRE, FELIPE PROSPER, AND FERNANDO CARAZO. 2022. "IDENTIFYING LETHAL DEPENDENCIES WITH HUGE PREDICTIVE POWER" *CANCERS* 14, NO. 13: 3251. HTTPS://DOI.ORG/10.3390/CANCERS14133251 IMPACT FACTOR (JCR 2021): 6.575. Q1 (POSITION 60/245)

GIMENO M, SAN JOSÉ-ENÉRIZ E, VILLAR S, AGIRRE X, PROSPER F, RUBIO A AND CARAZO F (2022) EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PRECISION MEDICINE IN ACUTE MYELOID LEUKEMIA. FRONT. IMMUNOL. 13:977358. DOI: 10.3389/FIMMU.2022.977358. IMPACT FACTOR (JCR 2021): 8.787 Q1 (POSITION 35/162)

GIMENO, MARIAN, SADA KATYNA, RUBIO, ANGEL. (UNDER CONSIDERATION). "PRECISION ONCOLOGY: THE CHALLEGE OF INTERPRETABILITY". UNDER REVIEW IN BRIEF. BIOINFORM. IMPACT FACTOR (JCR 2021): 13.994 Q1 (POSITION 1/57)

ADDITIONAL PUBLICATIONS

JUAN A FERRER-BONSOMS, MARIAN GIMENO, DANEL OLAVERRI, PABLO SACRISTAN, CÉSAR LOBATO, CARLOS CASTILLA, FERNANDO CARAZO, ANGEL RUBIO, "EVENTPOINTER 3.0: FLEXIBLE AND ACCURATE SPLICING ANALYSIS THAT INCLUDES STUDYING THE DIFFERENTIAL USAGE OF PROTEIN-DOMAINS", *NAR Genomics and Bioinformatics*, VOLUME 4, ISSUE 3, SEPTEMBER 2022, LQAC067,https://doi.org/10.1093/nargab/lqac067. (NEW JOURNAL NO IF AVAILABLE)

GIMENO, M., LOBATO, C., SAN MARTÍN, A., ANORBE, A., RUBIO, A., FERRER-BONSOMS, J.A. "A SYSTEMATIC IDENTIFICATION OF RBPS DRIVING ABERRANT SPLICING IN CANCER". (UNDER REVIEW) FRONT. MOL. BIOSCI. IMPACT FACTOR (JCR 2021):6.11 Q1 (POSITION 73/297)

CARAZO, FERNANDO, CRISTINA BÉRTOLO, CARLOS CASTILLA, XABIER CENDOYA, LUCÍA CAMPUZANO, DIEGO SERRANO, MARIAN GIMENO, FRANCISCO J.

PLANES, RUBEN PIO, LUIS M. MONTUENGA, AND ANGEL RUBIO. 2020. "DRUGSNIPER, A TOOL TO EXPLOIT LOSS-OF-FUNCTION SCREENS, IDENTIFIES CREBBP AS A PREDICTIVE BIOMARKER OF VOLASERTIB IN SMALL CELL LUNG CARCINOMA (SCLC)" CANCERS 12, NO. 7: 1824. https://doi.org/10.3390/cancers12071824. IMPACT FACTOR (JCR 2019): 6.12 Q1 (POSITION 37/244)

ALDAREGIA, JUNCAL, PEIO ERRARTE, ANE OLAZAGOITIA-GARMENDIA, MARIAN GIMENO, JOSE JAVIER URIZ, TIMOTHY R. GERSHON, IDOIA GARCIA, AND ANDER MATHEU. 2020. "ERBB4 IS REQUIRED FOR CEREBELLAR DEVELOPMENT AND MALIGNANT PHENOTYPE OF MEDULLOBLASTOMA" CANCERS 12, NO. 4: 997. https://doi.org/10.3390/cancers12040997. IMPACT FACTOR (JCR 2019): 6.12 Q1 (POSITION 37/244)

CARAZO, F., GIMENO, M., FERRER-BONSOMS, J.A. ET AL.(2019) “INTEGRATION OF CLIP EXPERIMENTS OF RNA-BINDING PROTEINS: A NOVEL APPROACH TO PREDICT CONTEXT-DEPENDENT SPLICING FACTORS FROM TRANSCRIPTOMIC DATA.” BMC GENOMICS 20, 521. https://doi.org/10.1186/s12864-019-5900-1. IMPACT FACTOR (JCR 2017): 3.73 Q1 (POSITION 40/161)

cancers

MDPI

Article

Identifying Lethal Dependencies with HUGE Predictive Power

Marian Gimeno [1,†], Edurne San José-Enériz [2,3,†], Angel Rubio [1,4], Leire Garate [3,5], Estíbaliz Miranda [2,3], Carlos Castilla [1], Xabier Agirre [2,3,*,‡], Felipe Prosper [2,3,5,*,‡] and Fernando Carazo [1,4,*,‡]

[1] Departamento de Ingeniería Biomédica y Ciencias, TECNUN, Universidad de Navarra, 20009 San Sebastian, Spain; mgimenoc@unav.es (M.G.); arubio@tecnun.es (A.R.); ccastilla.1@tecnun.es (C.C.)
[2] Programa Hemato-Oncología, Centro de Investigación Médica Aplicada, IDISNA, Universidad de Navarra, 31008 Pamplona, Spain; esanjose@alumni.unav.es (E.S.J.-E.); emelizalde@unav.es (E.M.)
[3] Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), 28029 Madrid, Spain; lgarate@unav.es
[4] Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), Universidad de Navarra, 31080 Pamplona, Spain
[5] Departamento de Hematología, Clínica Universidad de Navarra, Universidad de Navarra, 31008 Pamplona, Spain
* Correspondence: xaguirre@unav.es (X.A.); fprosper@unav.es (F.P.); fcarazo@tecnun.es (F.C.)
† These authors contributed equally to this work.
‡ These authors share senior and last authorship.

Simple Summary: This work shows that the predictions of lethal dependencies (LEDs) between genes can be dramatically improved by incorporating the "HUb effect in Genetic Essentiality" (HUGE) of gene alterations. In three genome-wide loss-of-function screens—Project Score, CERES score and DEMETER score—LEDs are identified with 75 times larger statistical power than using state-of-the-art methods. In AML, we identified LEDs not recalled by previous pipelines, including FLT3-mutant genotypes sensitive to FLT3 inhibitors. Interestingly, in-vitro validations confirm lethal de-pendencies of either NRAS or PTPN11 depending on the NRAS mutational status.

Abstract: Recent functional genomic screens—such as *CRISPR*-Cas9 or RNAi screening—have fostered a new wave of targeted treatments based on the concept of synthetic lethality. These approaches identified LEthal Dependencies (LEDs) by estimating the effect of genetic events on cell viability. The multiple-hypothesis problem is related to a large number of gene knockouts limiting the statistical power of these studies. Here, we show that predictions of LEDs from functional screens can be dramatically improved by incorporating the "HUb effect in Genetic Essentiality" (HUGE) of gene alterations. We analyze three recent genome-wide loss-of-function screens—Project Score, CERES score and DEMETER score—identifying LEDs with 75 times larger statistical power than using state-of-the-art methods. Using acute myeloid leukemia, breast cancer, lung adenocarcinoma and colon adenocarcinoma as disease models, we validate that our predictions are enriched in a recent harmonized knowledge base of clinical interpretations of somatic genomic variants in cancer (AUROC > 0.87). Our approach is effective even in tumors with large genetic heterogeneity such as acute myeloid leukemia, where we identified LEDs not recalled by previous pipelines, including *FLT3*-mutant genotypes sensitive to *FLT3* inhibitors. Interestingly, in-vitro validations confirm lethal dependencies of either *NRAS* or *PTPN11* depending on the *NRAS* mutational status. HUGE will hopefully help discover novel genetic dependencies amenable for precision-targeted therapies in cancer. All the graphs showing lethal dependencies for the 19 tumor types analyzed can be visualized in an interactive tool.

Keywords: *CRISPR*-Cas9 screening; precision medicine; synthetic lethality

Citation: Gimeno, M.; San José-Enériz, E.; Rubio, A.; Garate, L.; Miranda, E.; Castilla, C.; Agirre, X.; Prosper, F.; Carazo, F. Identifying Lethal Dependencies with HUGE Predictive Power. *Cancers* **2022**, *14*, 3251. https://doi.org/10.3390/cancers14133251

Academic Editor: Ada Funaro

Received: 16 May 2022
Accepted: 28 June 2022
Published: 1 July 2022

1. Introduction

The traditional concept of synthetic lethality consists of the concurrent loss of functionality of two genes resulting in cellular death. A relevant example is the effectiveness

of *PARP* inhibitors in tumors with inactivated *BRCA1* and *BRCA2* [1]. In recent years, the advances in functional genomics triggered by large-scale loss-of-function screening—such as *CRISPR*-Cas9 or RNA interference (RNAi) screens—have boosted the discovery of hundreds of novel targets and context-specific lethal dependencies (LEDs) [2–7], defined as any association between two genes that results in differential viability depending on their genetic context (Figure S1).

Several studies have carried out large-scale functional genomic screens to identify genome-wide targets and LEDs [2–5]. The Project Score [4], the Achilles Project [5,6] and the Project DRIVE [7] are three studies that performed genome-wide gene-knockouts in cancer cells aiming at establishing novel targets and LEDs. The refinement of computational and technical tools has improved the potential of loss-of-function screening to identify cancer vulnerabilities [3,8,9]. However, the multiple testing problem, related to a large number of gene knockouts, limits the statistical power of these studies and, therefore, their potential to find new targets.

Here, we show that previous efforts to predict LEDs from functional screening can be significantly improved by taking into account the "HUb effect" in Genetic Essentiality (HUGE) of some gene alterations: a few specific sets of gene alterations are statistically associated with large changes in the essentiality of multiple genes. These "hub" aberrations lead to more statistically reliable LEDs than other alterations that do not participate in such hubs. We incorporate the HUGE effect in the statistical analysis of three recent loss-of-function experiments of both The Project Score and The Achilles Project (two datasets) showing that the number of LEDs discovered for a given FDR considerably improves for both *CRISPR*-Cas9 and RNAi screens.

Using acute myeloid leukemia (AML), breast cancer (BRCA), lung adenocarcinoma (LUAD) and colon adenocarcinoma (COAD) as disease models, we validate that the predictions are enriched in associations used in the clinic. Finally, we validated in vitro an example of a therapy guideline based on LED selection in AML. The HUGE analysis will help discover novel tumor vulnerabilities in specific genetic contexts, providing valuable candidates—targets and genetic variants as biomarkers—for further personalized treatments in hematological diseases or other cancer disorders.

2. Materials and Methods

2.1. Data Integration

Data of loss-of-function screens libraries (17,980 knockout genes in 412 cancer cell lines) of the project Achilles [10] were integrated with gene expression and their corresponding gene alteration profiles (gene variants in ~1600 genes) obtained from CCLE and Shao et al. [6]. We gathered gene expression of cells using RNA-seq data to confirm that the genes that were essential for a cohort of cells were expressed before the RNAi library experiment was performed [11]. Gene variant panels were filtered out using the parameters of CCLE's authors to avoid common polymorphisms, low allelic fractions, putative neutral variants, and substitutions located outside of the coding sequence [12].

We used the DEMETER score [5,8] as a measure of gene essentiality of the RNAi libraries of the project Achilles [10]. DEMETER quantizes the competitive proliferation of the cell lines controlling the effect of off-target hybridizations of siRNAs by solving a complex optimization problem. The more negative the DEMETER score is, the more essential the gene is for a cell line. We imputed missing elements of DEMETER using the nearest neighbor averaging algorithm [13]. Moreover, we collected gene expression patterns from RNA-seq data [11] to confirm those essential genes are expressed when they are essential. Based on DEMETER data, we first identified genes that were essential for a selected tumor subtype. Essential genes were required to meet several criteria: (i) they must be essential for at least 20% of samples of the selected cancer subtype, (ii) they must be specific to the cancer type under study, i.e., they must be non-essential for other cancer types, and (iii) they must be expressed before RNAi experiment (>1TPM at least in 75% samples).

2.2. Statistical Model

We developed a statistical algorithm to identify genes whose essentiality is highly associated with the genetic alteration of other genes. Dealing with this statistical issue implies solving a large multiple hypotheses problem (more than one million hypotheses). In similar scenarios, traditional corrections—such as Benjamini-Hochberg (BH), Bonferroni, or Holm—showed very few or no gene-biomarker LEDs for a given FDR [14]. To overcome this problem, we developed a covariate-based statistical approach—similar to the Independent Hypothesis Weighting procedure [14] (Figure S2).

Let e denote the number of RNAi target genes and n denote the number of screened samples. Let $\mathbf{D}$ be an $e \times n$ matrix of essentiality whose entries d_{ij} represent the DEMETER score for the RNAi target i in sample j. Let $\mathbf{m}$ be a $m \times n$ dichotomized matrix whose entry m_{ij} denotes whether sample j is mutant or not according to the previous criteria:

$$m_{ij} = \begin{cases} 1, \; if \; mutant (MUT) \\ 0, \; if \; wild - type \; (WT) \end{cases}, \tag{1}$$

Let $\mathbf{s}$ be a subset of n' cell lines that yield an essentiality vector $\boldsymbol{d_s} = (d_{e_{s_1}}, \ldots, d_{e_{s_{n'}}})$ for the eth RNAi target. Let $\boldsymbol{m_s} = (m_{s_1} \ldots, m_{s_n})$ be the expression vector of a putative gene biomarker. The null hypotheses are defined as:

$$H_0^g : \; E(\boldsymbol{d_s} | \boldsymbol{m_s} \in MUT) = E(\boldsymbol{d_s} | \boldsymbol{m_s} \in WT) \tag{2}$$

This null hypothesis is, therefore: "the expected essentiality of a gene knock-down is identical in mutant and wild-type cell lines". To test this hypothesis, we used a moderated t-test implemented in limma [15]. We applied this test for each RNAi target and all the gene variants to obtain the corresponding p-values (Figure S2). Dealing with these p-values implies correcting for multiple hypotheses.

In our case, we divided the p-values corresponding to all the tests into n groups, where n is the number of altered genes. For each of these groups, we computed the local false discovery rate (local FDR) [16]. The local FDR estimates, for each test, the probability of the null hypothesis to be true, conditioned on the observed p-values. The formula of the local FDR is the following:

$$P(H_0|z) = localFDR(z) = \frac{\pi_0 f_0(z)}{f(z)}, \tag{3}$$

where z is the observed p-values, π_0 is the proportion of true null hypotheses—estimated from the data, $f_0(z)$ the empirical null distribution—usually a uniform (0, 1) distribution for well-designed tests—and $f(z)$ the mixture of the densities of the null and alternative hypotheses, which is also estimated from the data.

As stated by B. Efron and R. Tibshirani [16], "the advantage of the local FDR is its specificity: it provides a measure of belief in gene i's 'significance' that depends on its p-value, not on its inclusion in a larger set of possible values" as it occurs, for example, with q-values or the standard FDR. The local FDR and π_0 were estimated using the Bioconductor's R Package q-value [17].

2.3. Comparison with the Project Score

To compare our results with Project Score's ones, we selected the same 12 primary cancer tissues shared in both datasets. The comparison followed two steps: (1) using CCLE and DEMETER scores with the Project Score's algorithm, (2) running our approach adapted to Project Score conditions. In the first step, following the code published in their work, an ANOVA test was performed on each tissue to calculate all possible dependent partners. The Storey–Tibshirani correction was then used, using the criteria mentioned in Project Score methods [4]. This enabled us to correct the ANOVA p-values and obtain significant associations. Secondly, the comparison between both methodologies was only possible if the same adjusted p-value is calculated for both datasets. Therefore, we estimated the

FDR with our data as it is the q-value selected by the Project Score. The FDR correction was obtained using the Bioconductor R package *IHW* [14], which enables the consideration of covariates-based multiple hypothesis correction, as well as estimating the FDR. Discoveries from both methodologies in DEMETER and CCLE datasets were plotted in different volcano plots, and the number of significant LEDs was counted (FDR < 20%).

2.4. Integration of the VICC Knowledgebase of Clinical Interpretations of Genomic Variants

We downloaded 19,551 clinical interpretations of somatic genomic variants in cancer from the Variant Interpretation for Cancer Consortium (VICC) [18,19] (version December 2020). We filtered out incomplete (e.g., entrees without annotated drug or biomarker) and redundant associations. We then selected all associations that are annotated with acute myeloid leukemia (AML) and synonyms. From all drugs, we selected those that have an annotated protein target. To do so, we retrieved the data publicly available in the ChEMBL [20] and DrugBank [21] online repositories. In total, 216 out of 19,551 associations matched these criteria. We consider a true positive if either HUGE or ST identifies an LED whose mutation biomarker coincides with a VICC's association and the protein target is included in the same association, or at least in a gene of the same pathway in the STRING database (v.11, STRING score threshold = 400; default value on STRING for "medium" confidence) [22].

We calculated ROC and PR curves considering the two top evidence levels included in VICC [18,19], namely, (i) evidence from professional guidelines or FDA-approved therapies; and (ii) evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus.

2.5. Application to Acute Myeloid Leukemia (AML) as a Disease Model

We applied the pipeline to the AML cohort of cell lines (n = 15). In the first step, essential genes were required to be: (i) essential for at least 25% AML samples, (ii) specific for AML cells, and (iii) expressed before the RNAi experiment. The algorithm outputs a ranking of significant gene pairs (LEDs) that consist of a couple of genes in which the first one is essential depending on the genetic alteration of the other.

For the final ranking for AML, we selected those LEDs that showed a p-value < 0.05 and local FDR $\leq$ 0.6, |D DEMETER| > 2 (default value suggested by DEMETER's authors). Additionally, we interrogated which of these LEDs had direct relationships (co-expressed, annotated in the same pathway database, or contained in a common experiment) in the STRING database [22] to ensure there is an established biological relationship between the essential gene and the subrogate biomarker. This biological double-check is not necessary and can be omitted when the researcher looks for novel relationships.

In vitro validation was performed using siRNAs against *NRAS* and *PTPN11* in four different AML cell lines, two with *NRAS*-genetic variants (HL-60 and OCI-AML3) and two *NRAS*-wt cell lines (MV4-11 and HEL). Finally, the model was compared with 3 standard statistical methods (namely Benjamini-Hochberg (BH), Bonferroni and Holm) known to have suboptimal sensitivity (recall of true positives) in specific scenarios in 19 additional tumor subtypes to define the potential for controlling the FDR [14]. See File S1 for more details.

3. Results

3.1. Gene Variants Associated with Multiple Essential Genes Increase the Power of Loss-of-Function Screens

One of the main statistical challenges to finding LEDs by integrating genome-wide functional screens with -omics datasets is the multiple hypothesis testing problem. Correction for multiple hypotheses reduces the statistical significance of results (meaning a decreased detection rate and an increased false-positive rate). The Project Score presented a large-scale genome-wide *CRISPR*-Cas9 screening analysis targeting 18,009 genes in 30 different cancer types, across 14 different tissues [4,23]. They presented a methodology

to detect LEDs based on finding differences in genetic essentiality in cell lines associated with the presence of specific gene variants (ANOVA test [24] with the Storey–Tibshirani *p*-value correction). Following this procedure, the Project Score was able to identify genetic LEDs in 7 out of 14 individual tissues analyzed [4,23].

Analyzing Project Score's data, we observed that for each tumor type, a few specific genetic alterations were significantly associated with the genetic essentiality of a large set of genes. This handful of genetic aberrations shows a hub effect, in which a gene variant is associated with large changes in the essentiality of multiple genes. We termed this behavior the "HUb effect in Genetic Essentiality" (HUGE) (Figure 1A; other tumor types can be visualized in https://fcarazo.shinyapps.io/visnetShiny/ (accessed on 24 June 2022)). From the point of view of statistics, the HUGE effect is defined as an improvement of the statistical power by using gene variants as co-variates in a multiple hypothesis problem. Other biological covariates such as gene expression or copy number alterations have also shown to be covariates that increase the statistical power [14]. Using gene variants as statistical covariates provides a larger number of positives for a given FDR, which consequently means an increased specificity and sensitivity, or type I and type II errors, as demonstrated in File S1, Section S6. Interestingly, the analysis shows that the HUGE effect is present in all tumors analyzed, significantly improving the predictive power of LEDs.

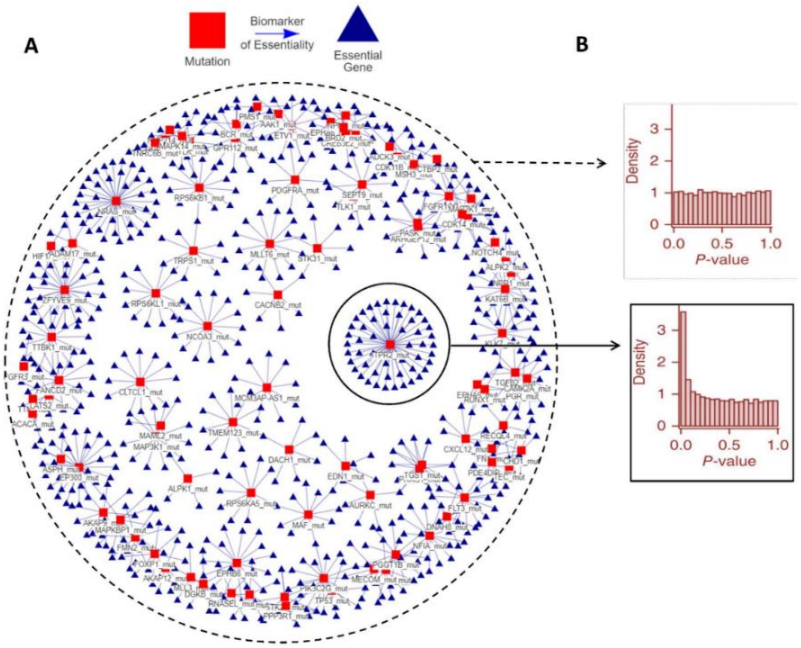


Figure 1. The hub effect in genetic essentiality in Acute Myeloid Leukemia. In each cell, a small set of gene aberrations is associated with large changes in genetic essentiality. (**A**) A bipartite graph in which red squares represent gene variants (e.g., mutations), blue triangles represent significant changes in cell viability related to knocked-down genes. Both nodes are linked by a line if the variations in the essentiality have a statistically significant association with the presence of the gene

variant. (**B**) Implications in *p*-value histograms of the HUGE effect. Hub associations show a high peak close to zero *p*-values indicating that the null hypothesis is rejected in more cases and that these genetic variants are associated with a higher response to the inhibition of more gene products. Segregating the statistical analysis according to the alteration provides more statistical power. Essential genes and other tumor types can be visualized in https://fcarazo.shinyapps.io/visnetShiny/ (accessed on 24 June 2022). Abbreviations. HUGE: The hub effect in genetic essentiality.

The presence of the HUGE effect in a cancer type can be also understood as a predictive model in which each mutation has a different capability to define the genetic essentiality of multiple genes. To show it visually, the histogram of *p*-values of a gene alteration represents how gene alterations are associated with the genetic essentiality of multiple genes. Histograms of the *p*-values for alterations that conform to a "hub" show a peak near the origin, which means that cells with these alterations are sensitive to the depletion of a large number of genes (Figure 1B). Conversely, if the hubs of alterations are not considered, the relationships of mutations and viability show a flat histogram of *p*-values. This does not necessarily mean that such relationships are not biologically relevant, but that it is difficult to distinguish them from random associations and will be considered as artifacts after multiple testing corrections.

The HUGE effect helps palliate the multiple hypothesis correction problem. Using the mutation under study as a covariate, multiple hypotheses can be differently treated considering the overall association of gene alteration in the complete set of essential genes (Figures S2 and S3). Using this concept, we developed a statistical model that integrates HUGE information to find LEDs (Figure S2).

Previous efforts to correct multiple testing in this problem consider a single set of tests (all gene aberrations and *CRISPR*-Cas9 knockouts) and apply a correction that controls the FDR, such as Storey–Tibshirani (ST), as performed in the Project Score. Interestingly, in all tumors, our approach increases the statistical power of the analysis. From a statistical point of view, a flat histogram is compatible with the null hypothesis for all the tests and, therefore, multiple hypothesis correction drives to none or few discoveries (Figure S4). Every single tumor shows *p*-value histograms related to specific gene variants that have a higher zero-peak than the histogram associated with all tests in such tumor (Figures S5–S23). To test this approach, we compare the results using HUGE with previous LED identification strategies in three genome-wide functional genomic projects: The Project Score [4], the DEMETER score and the CERES score (DEMETER and CERES are included in the Achilles Project [5,6]). First, to test the potential of HUGE to predict LEDs with *CRISPR*-Cas9 screens, we analyze the Project Score dataset [4]. Project Score integrates 215 different genetic events across 14 tumor types, including SNVs and CNVs. In the same reference, the authors found at least one LED in 7 out of the 14 tumor types analyzed. A total of 40 out of 215 events were detected to be significant biomarkers of essentiality (FDR $\leq$ 20%), which correspond to 77 unique LEDs (a single genetic event can be associated with several essential genes). Analyzing Project Score's data using the HUGE-based methodology, we identify 1438 unique associations with the same FDR (18 times larger than Project Score, Figure 2A), corresponding to 80 single genetic events. Moreover, using HUGE we detect at least one LED in all the 14 tumors analyzed, finding LEDs in 10 tumors that would have been missed using the original pipeline, affecting around 10–20 genes for each disease type.

We also tested HUGE in the DEMETER score of the Achilles Project to predict LEDs, in this case using RNAi screening. The DEMETER dataset [5,10] is a large-scale genome-wide experiment of RNA interference libraries (17,085 knockdown genes) in 19 tumor types (Table S5). We integrate the DEMETER data with the corresponding cell line gene alteration profiles (genetic variants in ~1600 genes) obtained from the Cancer Cell Line Encyclopedia (CCLE) [12] and Shao et al. [6]. This integration turns out to have 27 Million hypotheses, which will hardly impair *p*-values after multiple hypothesis correction (Figure S2). Then, we replicate the Project Score's pipeline with the DEMETER dataset and compare it with the HUGE-based approach to find LEDs, also including in the comparison other two standard *p*-value corrections used to control the FDR, namely Holm and Bonferroni. Using

the standard ST procedure, we find 126 LEDs (FDR $\leq$ 20%). There are LEDs for 7 out of 19 tumors. The same dataset and FDR threshold using the HUGE-based approach provides 9535 LEDs (75.7 times larger than using ST). All cancer types (19 out of 19) showed significant LEDs in the HUGE-based analysis (Figure 2B). HUGE identifies 1,675 LEDs in six tumor types in which other methods recall no LEDs (FDR $\leq$ 20%); and 9409 LEDs in 19 tumor types that would have been missed using previous procedures (FDR $\leq$ 20%; Figure 2C). These results show that the HUGE effect is present with different intensities in all tumor types analyzed (Figures S5–S23).

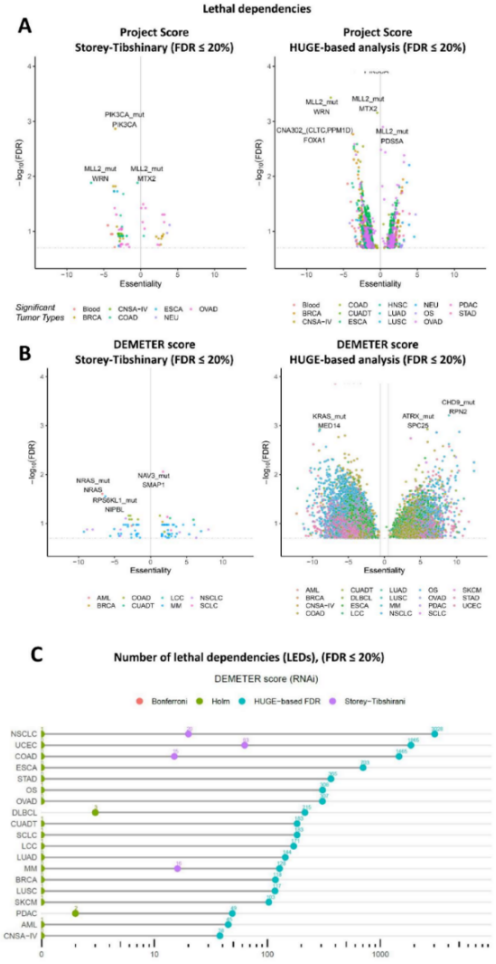


Figure 2. HUGE-based analysis with Project Score and Achilles Project datasets. (**A**) Volcano plots of lethal dependencies, LEDs, identified in the Project Score dataset. From left to right: (i) result of

Project Score, (ii) results of analyzing Project Score dataset with the HUGE-based methodology. Each dot represents a significant LED (FDR < 20%). The X-axis represents the difference in gene essentiality when the event (gene variants) is present. The Y-axis represents the FDR values (−log10) for that change. (**B**) Equivalent volcano plots using Achilles Project. From left to right: (i) results of Achilles Project analyzed with the standard procedure, (ii) results of analyzing Achilles Project dataset with HUGE-based methodology. (**C**) The number of LEDs found (FDR $\leq$ 20%) in 19 tumors of the DEMETER score (RNAi) and 22 tumors of the CERES score (*CRISPR*-Cas9) using standard statistical pipelines (Storey–Tibshirani, Bonferroni, and Holm) and the HUGE-based algorithm. Bonferroni and Holm return the same number of hypotheses in all cases. Abbreviations. LED: lethal dependency; ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia; BRCA: breast ductal carcinoma; CNSA-IV: central nervous system astrocytoma grade IV; COAD: colon adenocarcinoma; CUADT: upper aero-digestive tract squamous cell carcinoma; DLBCL: diffuse large B-cell lymphoma; ESCA: esophagus squamous cell carcinoma; KIRC: kidney renal clear cell carcinoma; LCC: lung large cell carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; MM: multiple myeloma; NSCLC: non-small cell lung carcinoma; OS: osteosarcoma; OVAD: ovary adenocarcinoma; PDAC: pancreas ductal carcinoma; SCLC: small cell lung carcinoma; SKCM: skin carcinoma; UCEC: endometrium adenocarcinoma.

As a further test of the increased predictive power of HUGE, we carry out a similar analysis using the CERES score, a *CRISPR*-Cas9 experiment of 22 tumors also included in the Achilles Project. In this case, the number of significant pairs is enriched 14 times over the standard approaches (FDR $\leq$ 20%; Figure S24).

3.2. LEDs Predicted by HUGE Have Better Validation Rates Than Standard Approaches

Validating a ranking of LEDs is not a simple task: it is desirable to have a gold standard of a disease-specific list of validated target-biomarker associations. We select as our gold standard The Variant Interpretation for Cancer Consortium (VICC) Meta-Knowledgebase [18,19]. This database integrates different datasets of clinical associations and includes the level of evidence for each entry: spanning from professional FDA guidelines to preclinical findings.

We test the enrichment in associations included in VICC in four tumor types, namely acute myeloid leukemia (AML), breast cancer (BRCA), lung adenocarcinoma (LUAD) and colon adenocarcinoma (COAD) for both HUGE and standard statistical methods. The VICC knowledgebase integrates (in September 2021) 19,551 clinical interpretations of somatic genomic variants in cancer of both resistant and sensitive biomarkers. We delete duplicated and incomplete associations, focused on those related to confirmed mutations and manually selected associations that match each tumor type (including synonyms).

We first run the two procedures (HUGE and Storey-Tibshirani; ST) with AML cell lines (Table S5) to find LEDs and compare how many LEDs predicted by HUGE and by ST are included in the VICC knowledgebase. For instance, if HUGE or the ST procedure predicts *FLT3* mutant AML genotypes to be sensitive to *FLT3* inhibition, it will be considered a true positive LED, as *FLT3* is a well-known target of AML and mutations in *FLT3,* the fms-like receptor-type tyrosine-protein kinase [25,26], are known to be sensitive biomarkers of the effectiveness of most *FLT3*-inhibitors [27,28].

In total, 216 out of 19,551 associations matched these filters. Getting the top 500 LEDs according to the ranking using the HUGE algorithm with AML, we find 17 LEDs that match the VICC knowledgebase of known clinic relationships (Table S1; Fisher p-value < 1×10^{-51}). An equivalent analysis using the standard pipeline (ANOVA test [24] with the Storey–Tibshirani p-value correction) shows that out of the top 500 LEDs, only one is included in the VICC knowledgebase (Table S1; Fisher p-value = 6.551×10^{-3}). This means that HUGE analysis identifies 16 true positive dependencies not recovered by ST (Fisher p-value = 6.41×10^{-5}). The global value of AUROC (0.53) is not too far from the baseline of 0.5 (Figure 3A), perhaps because of the scarcity of true positives in our gold

standard. We perform the same analysis with LUAD, BRCA and COAD getting AUCROC values of 0.62 (vs. 0.5), 0.87 (vs. 0.64) and 0.72 (vs. 0.54) for HUGE and ST, respectively. All cases show better values for HUGE than for ST (Figures 3B–D and S25).

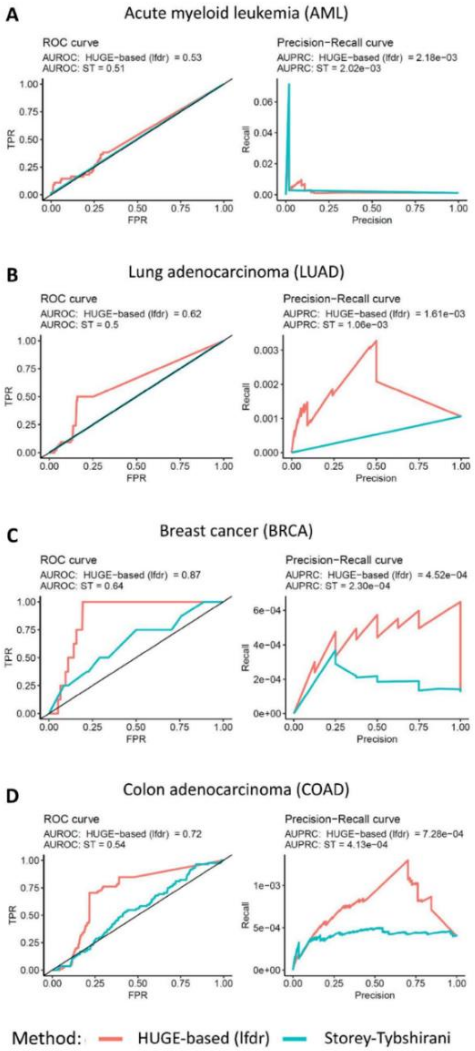


Figure 3. ROC and precision-recall curves of four tumor types. (**A**) Acute myeloid leukemia, (**B**) lung adenocarcinoma, (**C**) breast cancer and (**D**) colon adenocarcinoma. True positives were extracted from the knowledge base of the Variant Interpretation for Cancer Consortium [18,19]. For

each tumor type, we selected only those associations that belong to the three highest levels of confidence (Level A: Evidence from professional guidelines or FDA-approved therapies relating to a biomarker and disease; Level B: Evidence from clinical trials or other well-powered studies in clinical populations, with expert consensus; and Level C: Evidence for therapeutic predictive markers from case studies, or other biomarkers from several small studies, or evidence for biomarker therapeutic predictions for established drugs for different indications).

3.3. Applying HUGE Methodology to Acute Myeloid Leukemia Cell-Lines Discovers Potential Therapy Biomarkers

AML is a hematologic neoplasm characterized by a remarkable phenotypic and genomic heterogeneity [29], a challenging disease model to test the applicability and impact of HUGE. We run the complete HUGE pipeline with AML and validate in vitro two of the predicted LEDs.

As a preliminary step, we identify the potential genes that are essential for AML cell survival. The Achilles Project yielded 443 essential genes that are essential and specific for AML cells compared to other tumors (Table S2). Some of these genes belong to pathways known to be deregulated in AML (e.g., *MYB* [30] or *CEBPA* [31]). Interestingly, 160 of these 443 genes have previously been identified as potential cancer drivers in hematological malignancies according to the Candidate Cancer Gene Database (p-value = 7.76×10^{-5}, Fisher exact test) [32].

We then run the HUGE algorithm to identify genomic alterations that could be defined as LED partners of those 443 essential genes. In this pipeline, we require predicted pairs to be biologically related to each other in the STRING database (see Online Methods). LED associations can be broken down into three groups regarding their dependency type: *positive lethal dependency (pLED)*, when a gene variant marks sensitivity to the inhibition of another gene; *negative lethal dependency (nLED)*, when a gene variant marks resistance to the inhibition of another gene; or *dual lethal dependency (dLED)*, when the same gene variant confers, concurrently, sensitivity to the inhibition of one gene and resistance to the inhibition of another gene (Figure S1). In total, we predict 24 LEDs, (12 *pLEDs* and 12 *nLEDs*, including two *dLEDs*; p-value < 0.05, local FDR $\leq$ 0.6 and $|\Delta\text{Essentiality}| > 2$; Figure 4A, Table 1, Figure S26, and Table S3). Using the standard multiple hypotheses correction only one dependency turns out to be statistically significant. We provide the identified LEDs for the 19 tumors included in the Achilles Project following a similar pipeline (Tables S6–S24).

NRAS mutation ranks first in the analysis. Lethally dependent partners associated with *NRAS* genetic sequence variants show a p-value histogram that peaks at the origin (Figure 4A,B), meaning that *NRAS* mutations are associated with more tumor vulnerabilities than other alterations. Interestingly, *NRAS* alteration forms a *Dual Lethal Dependency* with *PTPN11* (Table 1, Figure 4C): it confers tumor sensitivity to *NRAS* inhibition and resistance to *PTPN11* inhibition.

To validate our prediction, we first check that both *NRAS* and *PTPN11* siRNAs efficiently decreased the *NRAS* and *PTPN11* expression, respectively, in four AML cell lines (Figure S27). Then, we confirm the computational hypothesis: the downregulation of *NRAS* significantly decreases cell proliferation only in the *NRAS*-altered AML cell lines, and the inhibition of *PTPN11* expression produces an equivalent effect, specifically in the *NRAS*-wt AML cell lines (Figure 4D), validating the predicted dLED. Remarkably, the validated *PTPN11*-*NRAS*-wt pair was not detected using standard methodologies.

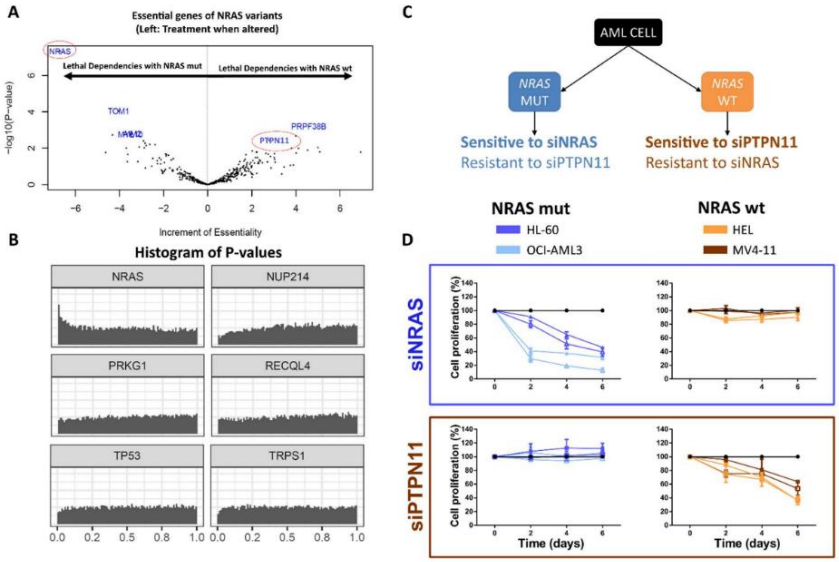


Figure 4. Gene variants-based treatment guidelines in acute myeloid leukemia. (**A**) Volcano-plot of lethal dependencies, LEDs, related to NRAS genetic mutations (left; MUT) and wildtype (right; WT) phenotypes. Increment of Essentiality and −log10 (*p*-value) are shown on X-axis and Y-axis, respectively. (**B**) Histogram of *p*-values for 6 genetic sequence variants in acute myeloid leukemia. NRAS-alteration is enriched in close to zero *p*-values, which is the basic concept of HUGE-based statistical approach. All genetic variants histograms of *p*-values can be found in the Supplementary Material. (**C**) Summary of the computational predictions validated: NRAS-altered cells were predicted to be sensitive to siNRAS and resistant to siPTPN11. Conversely, NRAS-wt cells were predicted to be sensitive to siPTPN11 and resistant to siNRAS. (**D**) Tumor proliferation of the four AML cell lines after inhibiting NRAS (siNRAS) and PTPN11 (siPTPN11) with specific siRNAs. Blue: NRAS-altered AML cell lines (HL-60 and OCI-AML3); Orange: NRAS-wild-type AML cell lines (MV4-11 and HEL).

Table 1. Ranking of lethal dependencies in AML using the covariate-based statistical approach. The ranking is divided into three groups regarding the typology of the lethal dependency relationship: Positive Lethal Dependency (PLD), Negative Lethal Dependency (NLD) or Dual Lethal Dependency (DLD) (Figure S1). The Increment of Essentiality column represents the average variation in the DEMETER score between altered and wild-type cells, and its sign is related to the lethal dependency relationship. Lethal dependencies that share the same essential gene and the same Increment of Essentiality sign were omitted in this table (see complete data in Supplementary Table S3).

Gene Variant Biomarker	Essential Gene	Increment of Essentiality	t-Score	*p*-Value	Local FDR
		Positive Lethal Dependencies			
TGS1	SNRPF	−7.87	−4.05	6.69×10^{-4}	3.36×10^{-1}
CLTCL1	UBR5	−6.66	−3.59	1.99×10^{-3}	2.20×10^{-1}
FLT3	FLT3	−6.36	−4.53	2.28×10^{-4}	2.00×10^{-1}
CDK14	CDK2	−3.95	−2.75	1.28×10^{-2}	4.30×10^{-1}
AURKC	ACTL6A	−3.26	−3.89	9.55×10^{-4}	4.99×10^{-1}

Table 1. *Cont.*

Gene Variant Biomarker	Essential Gene	Increment of Essentiality	t-Score	*p*-Value	Local FDR
		Negative Lethal Dependencies			
NPM1	EEF2	3.81	3.34	3.39×10^{-3}	5.96×10^{-1}
PIK3C2G	CDK6	3.35	2.95	8.20×10^{-3}	3.51×10^{-1}
NCOA3	EP300	3.04	2.75	1.25×10^{-2}	4.94×10^{-1}
CDK14	CCND2	2.97	2.22	3.88×10^{-2}	4.99×10^{-1}
EPHB6	ZNF266	2.53	2.77	1.22×10^{-2}	3.42×10^{-1}
ZFYVE9	TOM1L2	2.14	2.35	2.96×10^{-2}	5.12×10^{-1}
		Dual Lethal Dependencies			
NRAS	NRAS	−6.83	−8.71	4.67×10^{-8}	1.38×10^{-4}
NRAS	PTPN11	4.17	2.2	4.05×10^{-2}	5.89×10^{-1}
EP300	PLK1	−8.11	−4.04	7.01×10^{-4}	2.17×10^{-1}
EP300	KLF2	3.69	4.08	6.38×10^{-4}	2.12×10^{-1}

4. Discussion

The advent of large-scale functional genomic screens has allowed the identification of hundreds of novel gene targets and the prediction of genome-wide LEDs [4,33]. This strategy has multiplied treatment strategies, as using LEDs, the drug targets can be decoupled from their corresponding predictive biomarkers. The main statistical limit to finding LEDs is the large number of hypotheses that result from integrating gene essentiality and genetic functional events. In this work, we present HUGE, a novel analysis of *CRISPR*-Cas9 and RNAi large-scale screens that significantly improves the predictive power to find LEDs from loss-of-function screens in human tumors. It relies on the fact that some gene alterations are statistically related to the essentiality of large sets of genes. Using this characteristic as a prior covariate we significantly improve the predictive power of LEDs.

Notably, the presence of the HUGE effect does not necessarily mean biological causality. HUGE dependencies are more statistically reliable than others, but this does not imply that predicted alterations are the major players in tumor development thus, they are not necessarily driver genes, i.e., they are just genetic biomarkers of gene essentiality. In other words, the Hub-Effect is a statistical association. Since "correlation does not imply causation" is not legitimate to deduce a cause-and-effect relationship between the presence of a mutation and the sensitivity to knocking down a gene. Even more, it cannot be concluded that the HUGE top-ranked genes (either the mutations or the knockdown genes) are driver genes. This would require further experimentation and validation. HUGE simply computes biomarkers of the vulnerability to a knockdown gene, that in turn, could be targeted by a drug. However, the fact that gene alterations co-occur with multiple LEDs in genetic hubs can be exploited to improve the statistical power.

To measure the increased predictive power of HUGE, we carry out three different comparisons within three functional genomic datasets: the Project Score, the DEMETER score and the CERES score. HUGE identifies LEDs with 14 and 75 times larger statistical power than using state-of-the-art methods in *CRISPR*-Cas9 and RNAi, respectively. However, it could be argued that this result could be an artifact of the statistical technique and that lowering the threshold for standard procedures would provide LEDs with similar reliability. This is not the case. As shown in the results, using the same number of predictions, HUGE's results are more enriched in clinically used biomarkers than ST's results. Remarkably, 1 of the 16 LEDs only identified by HUGE is the known interaction of *FLT3*-mutant genotypes sensitive to *FLT3* inhibitors, such as Midostaurin. This fact is only an example of the key importance of considering the HUGE effect when analyzing LEDs with large-scale functional screens.

A *p*-value histogram can be modeled as the superposition of two distributions, a uniform distribution (which corresponds to the null hypothesis) and another distribution with a larger proportion of low *p*-values. A good covariate splits the overall *p*-value

histogram into histograms with different enrichments in small *p*-values. If all the histograms related to a covariate have similar shapes, it means that the covariate is uninformative. Here, we show that stating which gene is mutated in each test is a good covariate for the LED prediction problem because there is a hub effect of gene aberrations in gene essentiality. The usage of covariates has successfully been incorporated before in other genomics applications (e.g., the abundance of a gene is known to be informative in differential expression analyses; or the proximity of loci in the genome is known to play a role in genome-wide association studies), but it has not yet been exploited in large-scale functional genomic screens.

One main limitation lies in the volume of data required for its execution due to the need for multiple hypotheses to detect the Hub-Effect. Hence, the HUGE-based approach will not obtain such striking results if applied to the analysis of smaller experiments in number, it would perform similarly to current standard methods. Nevertheless, this method was developed for large-scale screening analyses.

We are confident that the HUGE-based approach to calculating LEDs has great potential if applied to the study of patient data. Nowadays, drug development usually starts from large-scale loss-of-function screenings. Therefore, this work has identified a large number of LEDs across 19 tumor types in three different large-scale experiments. Moreover, to facilitate the in vitro validation of these LEDs as possible therapeutic targets, we added information regarding targeted drugs for those essential genes that are drug targets.

Predicting true LEDs is especially challenging for tumors with high genetic heterogeneity. In AML, for instance, state-of-the-art approaches only recover two LEDs. The HUGE-based approach captured 24 LEDs for the same False Discovery Rate (FDR). Interestingly, *NRASwt-PTPN11* LED, which was only identified by HUGE, was validated in vitro. The validation in AML highlights the potential of the HUGE-based approach to discover and validate new LEDs of biomarkers and drug targets. We pinpoint the *dLED* characteristic of the *NRAS* gene, meaning that if a tumor has *NRAS* mutated a treatment that targets *NRAS* itself would be the best option to reduce their tumorigenicity, whereas if it is *NRAS* wild-type, a *PTPN11* inhibition would be a better recommendation. This *dLED* discovery confers special relevance to clinically translational therapeutic strategies, as it was proved effective in AML cell lines, further validation in ex vivo analysis and murine models is required but if the result is effective, it could be suggested as a treatment and it could incentivize drug development targeting *NRAS* and *PTPN11*. This methodology has potential applications both in basic and clinical research.

5. Conclusions

In conclusion, this work provides a computational approach to identifying LEDs with increased predictive power. This analysis opens new possibilities for the use of genetic variants as predictive events for precision oncology, by analyzing both previous and future functional genomic screens. Moreover, this analysis enhances current applications in translational oncology, such as drug development or drug repositioning projects.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/cancers14133251/s1, Supplementary Methods: Section S1: Cell Culture, Section S2: Cell Transfection, Section S3: Cell proliferation assay, Section S4: Quantitative-PCR(Q-PCR), Section S5: Statistical pipeline, Section S6: A larger number of positives outperforms specificity and sensitivity. Supplementary Figures: Figure S1: Types of Lethal Dependencies, Figure S2: Computational pipeline to find lethal dependencies, Figure S3: Schematic representation of the covariate-based statistical approach in this context, Figure S4: Histogram of *p*-values of all LEDs in AML, Figure S5: Histogram of *p*-values of all lethal dependencies in acute myeloid leukemia vs. *p*-values associated with each gene variant, Figure S6: Histogram of *p*-values oflethal dependencies in breast cancer vs. *p*-values associated with each gene variant, Figure S7: Histogram of *p*-values of all lethal dependencies in central nervous system astrocytoma grade IV vs. *p*-values associated with each gene variant, Figure S8: Histogram of *p*-values of all lethal dependencies in colon adenocarcinoma vs. *p*-values associated with each gene variant, Figure S9: Histogram of *p*-values of all lethal dependencies in

upper aerodigestive tract squamous cell carcinoma vs. *p*-values associated with each gene variant, Figure S10: Histogram of *p*-values of all lethal dependencies in diffuse large B-cell lymphoma vs. *p*-values associated with each gene variant, Figure S11: Histogram of *p*-values of all lethal dependencies in esophagus squamous cell carcinoma vs. *p*-values associated with each gene variant, Figure S12: Histogram of *p*-values of all lethal dependencies in lung large cell carcinoma vs. *p*-values associated with each gene variant, Figure S13: Histogram of *p*-values of all lethal dependencies in lung adenocarcinoma vs. *p*-values associated with each gene variant, Figure S14: Histogram of *p*-values of all lethal dependencies in lung squamous cell carcinoma vs. *p*-values associated with each gene variant, Figure S15: Histogram of *p*-values of all lethal dependencies in multiple myeloma vs. *p*-values associated with each gene variant, Figure S16: Histogram of *p*-values of all lethal dependencies in non–small cell lung carcinoma vs. *p*-values associated with each gene variant, Figure S17: Histogram of *p*-values of all lethal dependencies in osteosarcoma vs. *p*-values associated with each gene variant, Figure S18: Histogram of *p*-values of all lethal dependencies in ovary adenocarcinoma vs. *p*-values associated with each gene variant, Figure S19: Histogram of *p*-values of all lethal dependencies in pancreas ductal carcinoma vs. *p*-values associated with each gene variant, Figure S20: Histogram of *p*-values of all lethal dependencies in small cell lung carcinoma vs. *p*-values associated with each gene variant, Figure S21: Histogram of *p*-values of all lethal dependencies in skin carcinoma vs. *p*-values associated with each gene variant, Figure S22: Histogram of *p*-values of all lethal dependencies in stomach adenocarcinoma vs. p-values associated with each gene variant, Figure S23: Histogram of *p*-values of all lethal dependencies in uterine corpus endometrial carcinoma vs. *p*-values associated with each gene variant, Figure S24: The number of LEDs found (FDR $\leq$ 20%), Figure S25: ROC and precision-recall curves of four tumor types, Figure S26: Volcano plot of Synthetic lethal genes related to NRAS-mutated (A) and EP300-mutated (B) phenotypes, Figure S27: mRNA expression of NRAS and PTPN11 genes after nucleofection with the specific siRNAs. Supplementary Tables: Table S1: Associations within the top 500 pairs predicted using the HUGE-based and standard pipeline algorithms in AML that match the knowledgebase of clinical interpretations of somatic genomic variants in cancer of the Variant Interpretation for Cancer Consortium (VICC), Table S2: Essential genes for AML. Selected genes meet the following criteria: (i) must be essential in $\geq$25% of AML cell lines (DEMETER essentiality threshold set to -2), Table S3: Complete ranking of lethal dependencies in AML using the HUGE-based statistical approach. The Increment of Essentiality (deltaEs) column represents the average variation in the DEMETER score between altered and wild-type cells, and its sign is related to the lethal dependy relationship, Table S4: Cell lines included in the analysis, Table S5: AML cell lines included in the analysis, Table S6: Ranking of pairs mutation biomarker and essential genes in 19 tumor types using a covariate-based statistical model, Table S7: Ranking of pairs mutation biomarker and essential genes in OS, Table S8: Ranking of pairs mutation biomarker and essential genes in BRCA, Table S9: Ranking of pairs mutation biomarker and essential genes in CNSA-IV, Table S10: Ranking of pairs mutation biomarker and essential genes in UCEC, Table S11: Ranking of pairs mutation biomarker and essential genes in COAD, Table S12: Ranking of pairs mutation biomarker and essential genes in DLBCL, Table S13: Ranking of pairs mutation biomarker and essential genes in MM, Table S14: Ranking of pairs mutation biomarker and essential genes in LUAD, Table S15: Ranking of pairs mutation biomarker and essential genes in LCC, Table S16: Ranking of pairs mutation biomarker and essential genes in NSCLC, Table S17: Ranking of pairs mutation biomarker and essential genes in SCLC, Table S18: Ranking of pairs mutation biomarker and essential genes in LUSC, Table S19: Ranking of pairs mutation biomarker and essential genes in ESCA, Table S20: Ranking of pairs mutation biomarker and essential genes in OVAD, Table S21: Ranking of pairs mutation biomarker and essential genes in PDAC, Table S22: Ranking of pairs mutation biomarker and essential genes in SKCM, Table S23: Ranking of pairs mutation biomarker and essential genes in STAD, Table S24: Ranking of pairs mutation biomarker and essential genes in STAD.

Author Contributions: F.C.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Roles/Writing—original draft, Writing—review and editing. E.S.J.-E.: Conceptualization, Data curation, Formal analysis, Investigation, Validation, Roles/Writing—original draft, Writing—review and editing. M.G.: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Roles/Writing—original draft, Writing—review and editing. L.G.: Investigation, Validation. E.M.: Investigation, Validation. C.C.: Data curation, Formal analysis, Investigation, Software. X.A.: Conceptualization, Formal analysis, Funding ac-

quisition, Investigation, Project administration, Supervision, Validation, Roles/Writing—original draft, Writing—review and editing. A.R.: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Project administration, Supervision, Roles/Writing—original draft, Writing—review and editing. F.P.: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Roles/Writing—original draft, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Provincial Council of Gipuzkoa through the MINEDRUG project, by grants from Instituto de Salud Carlos III (ISCIII) PI14/01867, PI16/02024, PI17/00701, PI19/01352, PI20/01306 and TRANSCAN EPICA AC16/00041 (Co-financed with European Union FEDER funds), Fundació La Marató de TV3, Minister of Economy and Competitiveness of Spain [PID2019-110344RB-I00], PIBA Programme of the Basque Government [PIBA_2020_01_0055], Cancer Research UK [C355/A26819] and FC AECC and AIRC under the Accelerator Award Programme, CIBERONC CB16/12/00489 (Co-financed with European Union FEDER funds), Spanish Ministry of Economy, Industry and Competitivity (RTHALMY SAF2017-92632-EXP), Gobierno de Navarra, Departamento de Salud 40/2016 and Departamento de Industria (Proyecto Estrategico, Reto Genomica, DIANA), Synlethal Project (RETOS Investigacion Referencia PID2019-110344RB-I00, Spanish Government). FC was partially supported by a Basque Government predoctoral Grant [PRE_2016_1_0194].

Institutional Review Board Statement: The AML cell lines used in this study were purchased from ATCC or DSMZ and were authenticated by performing a short tandem repeat allele profile.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the graphs showing lethal dependencies for the 19 tumor types analyzed can be visualized in the following interactive tool https://fcarazo.shinyapps.io/visnetShiny/ (accessed on 24 June 2022). All genes and biomarkers predicted in this study can be downloaded from Tables S7–S24.

Acknowledgments: The authors would like to thank Francisco J. Planes, Luis V. Valcárcel, Xabier Cendoya and Lucía Campuzano for their fruitful comments on the development of the methodology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lord, C.J.; Ashworth, A. PARP inhibitors: The first synthetic lethal targeted therapy. *Science* **2017**, *355*, 1152–1158. [CrossRef] [PubMed]
2. O'Neil, N.J.; Bailey, M.L.; Hieter, P. Synthetic lethality and cancer. *Nat. Rev. Genet.* **2017**, *18*, 613–623. [CrossRef] [PubMed]
3. Huang, A.; Garraway, L.A.; Ashworth, A.; Weber, B. Synthetic lethality as an engine for cancer drug target discovery. *Nat. Rev. Drug Discov.* **2020**, *19*, 23–38. [CrossRef] [PubMed]
4. Behan, F.M.; Iorio, F.; Picco, G.; Gonçalves, E.; Beaver, C.M.; Migliardi, G.; Santos, R.; Rao, Y.; Sassi, F.; Pinnelli, M.; et al. Prioritization of cancer therapeutic targets using *CRISPR*-Cas9 screens. *Nature* **2019**, *568*, 511–516. [CrossRef]
5. Tsherniak, A.; Vazquez, F.; Montgomery, P.G.; Weir, B.A.; Kryukov, G.; Cowley, G.S.; Fill, S.; Harrington, W.F.; Pantel, S.; Krill-Bureger, J.M.; et al. Defining a cancer dependency map. *Cell* **2017**, *170*, 564–576.e16. [CrossRef]
6. Shao, D.D.; Tsherniak, A.; Gopal, S.; Weir, B.A.; Tamayo, P.; Stransky, N.; Schumacher, S.E.; Zack, T.I.; Beroukhim, R.; Garraway, L.A.; et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res.* **2013**, *23*, 665–678. [CrossRef]
7. McDonald, E.R.; De Weck, A.; Schlabach, M.R.; Billy, E.; Mavrakis, K.J.; Hoffman, G.R.; Belur, D.; Castelletti, D.; Frias, E.; Gampa, K.; et al. Project DRIVE: A compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. *Cell* **2017**, *170*, 577–592.e10. [CrossRef]
8. McFarland, J.M.; Ho, Z.V.; Kugener, G.; Dempster, J.M.; Montgomery, P.G.; Bryan, J.G.; Krill-Burger, J.M.; Green, T.M.; Vazquez, F.; Boehm, J.S.; et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.* **2018**, *9*, 4610. [CrossRef]
9. Shalem, O.; Sanjana, N.E.; Zhang, F. High-throughput functional genomics using *CRISPR*-Cas9. *Nat. Rev. Genet.* **2015**, *16*, 299–311. [CrossRef]
10. Lee, A.J.X.; Swanton, C. Tumour heterogeneity and drug resistance: Personalising cancer medicine through functional genomics. *Biochem. Pharmacol.* **2012**, *83*, 1013–1020. [CrossRef]
11. Wilcox, R.R. ANOVA: A paradigm for low power and misleading measures of effect size. *Rev. Educ. Res.* **1995**, *65*, 51–77. [CrossRef]
12. Ignatiadis, N.; Klaus, B.; Zaugg, J.B.; Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **2016**, *13*, 577–580. [CrossRef]

13. Cowley, G.S.; Weir, B.A.; Vazquez, F.; Tamayo, P.; Scott, J.; Rusin, S.; East-Seletsky, A.; Ali, L.D.; Gerath, W.F.J.; Pantel, S.A.; et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* **2014**, *1*, 140035. [CrossRef]
14. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehar, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [CrossRef]
15. Wagner, A.H.; Walsh, B.; Mayfield, G.; Tamborero, D.; Sonkin, D.; Krysiak, K.; Deu-Pons, J.; Duren, R.P.; Gao, J.; McMurry, J.; et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat. Genet.* **2020**, *52*, 448–457. [CrossRef]
16. Alterovitz, G.; Heale, B.; Jones, J.; Kreda, D.; Lin, F.; Liu, L.; Liu, X.; Mandl, K.D.; Poloway, D.W.; Ramoni, R.; et al. FHIR Genomics: Enabling standardization for precision medicine use cases. *NPJ Genomic Med.* **2020**, *5*, 13. [CrossRef]
17. Kazi, J.U.; Rönnstrand, L. FMS-like tyrosine kinase 3/FLT3: From basic science to clinical implications. *Physiol. Rev.* **2019**, *99*, 1433–1466. [CrossRef]
18. López-Zabalza, M.J.; Martínez-Lausín, S.; Bengoechea-Alonso, M.T.; López-Moratalla, N.; González, A.; Santiago, E. Signaling pathway triggered by a short immunomodulating peptide on human monocytes. *Arch. Biochem. Biophys.* **1997**, *338*, 136–142. [CrossRef]
19. Pratz, K.W.; Sato, T.; Murphy, K.M.; Stine, A.; Rajkhowa, T.; Levis, M. FLT3-mutant allelic burden and clinical status are predictive of response to FLT3 inhibitors in AML. *Blood* **2010**, *115*, 1425–1432. [CrossRef]
20. Metzelder, S.; Röllig, C. FLT3 inhibitors for the treatment of acute myeloid leukemia. *Best Pract. Onkol.* **2018**, *13*, 182–190. [CrossRef]
21. Papaemmanuil, E.; Gerstung, M.; Bullinger, L.; Gaidzik, V.I.; Paschka, P.; Roberts, N.D.; Potter, N.E.; Heuser, M.; Thol, F.; Bolli, N.; et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **2016**, *374*, 2209–2221. [CrossRef]
22. Pattabiraman, D.R.; McGirr, C.; Shakhbazov, K.; Barbier, V.; Krishnan, K.; Mukhopadhyay, P.; Hawthorne, P.; Trezise, A.; Ding, J.; Grimmond, S.M.; et al. Interaction of c-Myb with p300 is required for the induction of acute myeloid leukemia (AML) by human AML oncogenes. *Blood* **2014**, *123*, 2682–2690. [CrossRef]
23. Smith, M.L.; Cavenagh, J.D.; Lister, T.A.; Fitzgibbon, J. Mutation of CEBPA in familial acute myeloid leukemia. *N. Engl. J. Med.* **2004**, *351*, 2403–2407. [CrossRef]
24. Abbott, K.L.; Nyre, E.T.; Abrahante, J.; Ho, Y.Y.; Vogel, R.I.; Starr, T.K. The candidate cancer gene database: A database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* **2015**, *43*, D844–D848. [CrossRef]
25. Pacini, C.; Dempster, J.M.; Najgebauer, H.; Mcfarland, J.M.; Tsherniak, A.; Iorio, F. Integrated cross-study datasets of genetic dependencies in cancer. *Nat. Commun.* **2021**, *12*, 1661. [CrossRef]
26. Tatlow, P.J.; Piccolo, S.R. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *Sci. Rep.* **2016**, *6*, 39259. [CrossRef]
27. Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525.
28. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
29. Efron, B.; Tibshirani, R. Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **2002**, *23*, 70–86. [CrossRef]
30. Storey, J.D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2002**, *64*, 479–498. [CrossRef]
31. Gaulton, A.; Hersey, A.; Nowotka, M.L.; Patricia Bento, A.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [CrossRef] [PubMed]
32. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef] [PubMed]
33. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef] [PubMed]

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fimmu.2022.977358

OPEN ACCESS

EDITED BY
Giuseppe Lia,
University of Turin, Italy

REVIEWED BY
Yu Wang,
Shanghai Jiao Tong University, China
Gian Maria Zaccaria,
National Cancer Institute Foundation (IRCCS), Italy

*CORRESPONDENCE
Angel Rubio
arubio@tecnun.es
Fernando Carazo
fernando.carazo@veeva.com

SPECIALTY SECTION
This article was submitted to Alloimmunity and Transplantation, a section of the journal Frontiers in Immunology

RECEIVED 24 June 2022
ACCEPTED 13 September 2022
PUBLISHED 29 September 2022

CITATION
Gimeno M, San José-Enériz E, Villar S, Agirre X, Prosper F, Rubio A and Carazo F (2022) Explainable artificial intelligence for precision medicine in acute myeloid leukemia. *Front. Immunol.* 13:977358. doi: 10.3389/fimmu.2022.977358

Explainable artificial intelligence for precision medicine in acute myeloid leukemia

Marian Gimeno[1], Edurne San José-Enériz[2,3], Sara Villar[4], Xabier Agirre[2,3], Felipe Prosper[2,3,4], Angel Rubio[1,5]* and Fernando Carazo[1,5]*

[1]Departamento de Ingeniería Biomédica y Ciencias, TECNUN, Universidad de Navarra, San Sebastián, Spain, [2]Programa Hemato-Oncología, Centro de Investigación Médica Aplicada, Instituto de Investigación Sanitaria de Navarra (IDISNA), Universidad de Navarra, Pamplona, Spain, [3]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain, [4]Departamento de Hematología and CCUN (Cancer Center University of Navarra), Clínica Universidad de Navarra, Universidad de Navarra, Pamplona, Spain, [5]Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), Universidad de Navarra, Pamplona, Spain

Artificial intelligence (AI) can unveil novel personalized treatments based on drug screening and whole-exome sequencing experiments (WES). However, the concept of "black box" in AI limits the potential of this approach to be translated into the clinical practice. In contrast, explainable AI (XAI) focuses on making AI results understandable to humans. Here, we present a novel XAI method -called multi-dimensional module optimization (MOM)- that associates drug screening with genetic events, while guaranteeing that predictions are interpretable and robust. We applied MOM to an acute myeloid leukemia (AML) cohort of 319 *ex-vivo* tumor samples with 122 screened drugs and WES. MOM returned a therapeutic strategy based on the *FLT3*, *CBFβ-MYH11*, and *NRAS* status, which predicted AML patient response to Quizartinib, Trametinib, Selumetinib, and Crizotinib. We successfully validated the results in three different large-scale screening experiments. We believe that XAI will help healthcare providers and drug regulators better understand AI medical decisions.

KEYWORDS

biomarkers, treatment selection, assignation problem, explainable artificial intelligence, drug repositioning, large-scale screening, ex-vivo experiment, drug sensitivity

1 Introduction

The advance of personalized medicine, and in particular precision oncology, is partially based on the development of drug sensitivity studies. These experiments are promoting the discovery of new drugs, biomarkers of sensitivity, and drug repositioning. With increasing frequency, these studies have widened their scope from single drug studies to experiments involving hundreds of drugs, also known as drug screening. In recent years, drug screenings are being carried out on hundreds of cell lines giving rise to large-scale drug screening datasets, e.g., GDSC, which includes 130 screened drugs in an average of 368 lines per drug (1). Combining these drug sensitivity studies with tumor genotypes makes it possible to associate the response to treatment with genetic alterations (biomarkers), thus promoting the search for new personalized therapies (2).

Exploring the potential of these experiments, artificial intelligence (AI) algorithms for personalized medicine focus on the analysis of such datasets to bridge the gap for drug discovery. Some studies use machine learning algorithms for monotherapy prediction (3, 4), other approaches are based on training deep learning (DL) models from patients' omics data (5, 6). These methods create black-box predictors that make agnostic inferences of treatment for a patient based on complex non-linear relationships. The output is, for these cases, an individual therapy for a patient, instead of a general treatment guideline (7). Despite optimizing patient treatment, this approach has the inherent disadvantages of methods based on neural networks: they require a huge amount of data, and therefore experiments are unable to show the criteria that trigger the decision –since neural networks tend to be black-box models–. These technical challenges are limiting the translation of drug screening experiments to clinical practice.

Explainable Artificial Intelligence (XAI) focuses on making AI understandable to humans by the usage of "white-box" algorithms that allow end-users to understand why the model predicts a certain solution (8, 9). The importance of using XAI models in the finding of new personalized treatments is twofold: therapeutic pipelines can be more easily adopted in normal clinical guides (e.g., using a decision tree that does not require a complex model with a high number of variables) (9) and drug regulators, such as the Food and Drug Administration (FDA), or European Medicines Agency (EMA) will have an easier journey to approve a drug if the companion biomarkers are reasonable and robust (10, 11). Consequently, XAI opens the door to bridge the gap between clinical practice and bioinformatics (8, 12).

In this study we have developed a new XAI method, called multi-dimensional module optimization (MOM) algorithm, to predict therapeutic strategies based on large-scale drug screening data. This method systematically associates drugs with combined sets of genetic biomarkers that can be generalized and applied to other cohorts of patients. The therapeutic strategies provided by MOM can easily be understood by humans and are easy to implement in the clinical practice with a process equivalent to a decision tree. The optimization problem considers the effect of drug toxicity focusing on providing drugs that are differentially effective to patients with a specific genotype. MOM's result is deterministic –this is important to get regulatory approvals– and guaranteed to be optimal, each patient is given the best possible treatment.

We selected Acute Myeloid Leukemia (AML) as a disease model, a highly heterogeneous type of cancer that affects bone marrow cell precursors. In AML, genomic profiling is essential to understand its biology, diagnosis, and treatment (13–15). Unfortunately, 70% of adult people diagnosed with this disease die within five years of diagnosis (16). The current ELN (European Leukemia Network) risk stratification is based on the genetic biomarkers of the disease (17). Although there are big prognosis differences across these genetic groups, the current approach for young and fit patients is a standard induction cytotoxic therapy ("3+7") (14, 17) with the addition of targeted therapies, mainly *FLT3* inhibitors, to a specific group of AML patients (14). Despite 8 new drugs have been approved for AML in the last years, its lethality is still very high. In addition, there are no targeted treatments directed to $FLT3^{WT}$ patients –70% of all AML cases (18). A machine learning approach that identifies the most adequate FLT3 inhibitor as well as the treatment for other AML genotypes, would allow the discovery of new indications for other drugs for the AML. As a result, a new classification guide based on the response to therapy for specific genetic alterations would be beneficial in clinical practice.

We applied MOM to the BeatAML project cohort, which carried out WES (Whole Exome Sequencing) and drug screening experiments of 122 drugs with *ex-vivo* AML tumor samples from 319 patients (19). Ex-vivo experiments in hematological cancers are of great importance since they are performed directly on the patient's living tumor cells (19, 20), allowing to correlate drug sensitivity to the patient's genotype. The results obtained using MOM are *in-silico* validated using K-fold cross-validation and in three independent large-scale experiments, one based on pan-cancer drug sensitivity and two referred to pan-cancer gene essentiality using siRNA and *CRISPR*-cas9. MOM's patient indications require only three different biomarkers, which makes them to be easily understood by the clinician.

2 Results

2.1 An explainable artificial intelligence method to predict optimal treatments based on patient genotype

The implementation of a clinical translational XAI model requires the development of a robust method to associate

10.3389/fimmu.2022.977358

biomarkers to specific targeted treatments. and, thus, relating drug sensitivity and patient genetic events -including SNVs, indels, fusion genes, or even epigenetics. The development of an AI algorithm in this context requires to solve three important challenges: (i) proper modeling of the toxicity of screened drugs (most aggressive drugs are not necessarily better treatments), (ii) dealing with a high number of statistical hypotheses that intrinsically increase false discovery rate (FDR), and (iii) explaining the internal reasoning that the model uses to propose a decision so that it is easy to approve and implement in the clinical practice.

We propose an algorithm named Multi-dimensional Module Optimization (MOM) that addresses each of these challenges by dividing the problem into three main steps (Figure 1): preprocessing the input drug sensitivity scores, associating single biomarkers to drugs with an increased statistical power and combining individual treatments to unveil multi-step treatment pipelines to stratify patients based on drug-response.

MOM is developed to optimally stratify patients following a decision tree based on simple logical rules, in which each step is defined by the presence or absence of a certain biomarker and the recommendation of one drug. In turn, MOM requires genetic variants information and drug sensitivity screenings as input data.

To illustrate the steps of the algorithm, let us consider a toy example with 8 drugs and their corresponding drug-response scores for 6 patients (Figure 2). In this case, as in every precision medicine scenario, we want to find robust companion biomarkers that, associated to drugs allow us to maximize patient response with minimized toxicity.

In the first step, MOM preprocesses drug sensitivity scores (Figure 2.1). For which, instead of using the standard measure of IC_{50}, we proposed an incremental version of the logarithm of the IC_{50}, named IC50* (See Methods for more details). The proposed correction has two main advantages. First, MOM prioritizes drugs that have a differential effect on different patients, which are, in turn, better candidates to develop a personalized treatment based on a companion biomarker. Second, drugs whose effectiveness does not depend on patient genotype are more unspecific and, therefore, more prone to be toxic for different tissues. In the next section, we will illustrate this fact with a real case scenario.

To exemplify this normalization, let us return to the toy example with 6 patients, 8 drugs and their corresponding log (IC_{50}) scores measured in ex-vivo tumors (Figure 2.1). Considering raw log(IC_{50}) exclusively (left-hand heatmap), it could be argued that Drug 1 is the most effective drug and, therefore, it should be indicated to all patients regardless their genotype. However, since the dose can be adjusted for each patient, Drugs 1 and 8 will be given at a small and a large dose respectively balancing their effect. Using IC50* (right-hand panel) allows MOM to maximize the genetic dependence of drugs, rather than the absolute cellular death in patient tumors.

In the second step (Figure 2.2), MOM provides single biomarker-treatment associations by prioritizing the drugs whose response is associated with patient genotype. The selected statistical analysis to find the biomarker-treatment associations is the Independent Hypothesis Weighting (IHW) algorithm. This algorithm has been proved to increase the power of tests in several biological scenarios (21, 22).

This algorithm provides also two interesting "by-products": i) identifies which biomarkers are related to drug sensitivity, e. gr. *TP53* is usually a source of resistance, ii) identifies drugs whose efficacy is related to the genetic profile, Olaparib is effective only for $BRCA^{Mut}$ patients (23).

In the third step (Figure 2.3), MOM predicts a sequential treatment guideline that maximizes the drug effect on the group

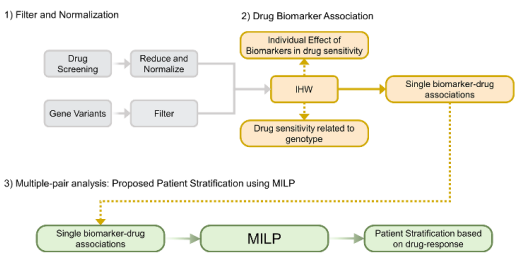


FIGURE 1
Overview of MOM's pipeline. **(1)** Filter and Normalization. **(2)** Generate individual Drug-Biomarker Associations using IHW, **(3)** Multiple-pair analysis that generates a patient stratification guideline using a novel MILP model (IHW: Independent Hypothesis Weighting, MILP: Mixed-integer Linear Programming).

10.3389/fimmu.2022.977358

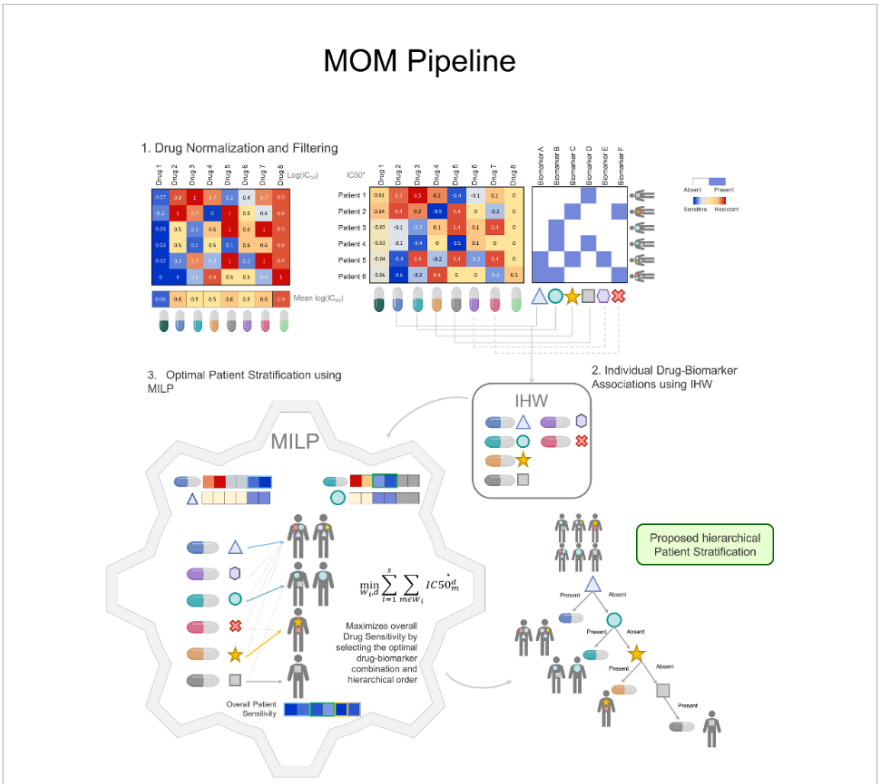


FIGURE 2
MOM Pipeline: MOM pipeline is defined by 3 major steps: 1) Drug normalization to reduce drug toxicity. It is performed by removing drug mean effect in all patients. The blue color represents drug sensitivity for the sample whereas the red color represents drug resistance in the sample. 2) Individual Drug biomarker associations using IHW. Drugs are matched to biomarkers profile, all individual associations generate a p-value that is corrected using IHW. IHW selects the candidate biomarkers and treatment and are used as input to the MILP problem. 3) Optimal Patient Stratification using MILP. The MILP module receives as input the normalized drug information IC50* and the candidate individual associations and outputs a decision tree for clinical decision-making guidance. Within this module, the treatment is optimized so that each patient receives the drug for which is more sensitive. (IHW: Independent Hypothesis Weighting, MILP: Mixed-Integer linear Programing).

of patients that share the genotype of the selected biomarkers. Using Mixed integer Linear Programing (MILP)(see Supplementary Methods), MOM gets the optimal treatment guideline (decision tree). MILP is a versatile optimization method that allows the solution of complex mathematical problems using integer variables and assures that the drug assignation is optimal. This solution (i) is explainable (XAI); (ii) eases the translation into clinical practice; and (iii) assures a global and deterministic optimum to the problem.

2.2 *FLT3*, *CBFβ-MYH11*, and *NRAS* variants play a key role in acute myeloid leukemia sensitivity to quizartinib, trametinib, and selumetinib

We selected the BeatAML cohort to test MOM as it contains *ex-vivo* drug sensitivity screenings of 122 drugs in AML tumors derived from 319 patients (19), and includes both whole-exome sequencing experiments (WES) and drug sensitivity for every

patient. This cohort, allows us to measure the impact of genetic variants on drug sensitivity (Supplementary Figures 13-19). In addition, AML is a good disease model to develop precision treatments, as it is a highly heterogeneous disease in which genomic profiling is essential to understand its biology, diagnosis, and treatment (13–15). Patients within this cohort are in different therapeutic stages, e.g., induction, maintenance, consolidation, or palliative care (among others), there also are 32 *de novo* patients (Supplementary Figure 12).

The drugs studied in the BeatAML cohort cover a wide variety of different cancers and diseases: 24% are indicated for AML, 16% for other leukemias types, 10% for multiple myeloma, and 4% for lymphomas. This means that 54% of the drugs have been studied for hematological malignancies. The rest 46% include drugs used in lung, breast, or renal cancers among other diseases (Supplementary Figure 20). Focusing on AML, the dataset provides a total of 11 AML drugs already in clinical use -e.g. Venetoclax, Quizartinib, or Gilteritinib- and 18 AML experimental drugs -e.g. Panobinostat, Lestaurtinib, or Pazopanib.

We filtered gene variants to keep the ones that appear in at least 4 out of 319 patients (1%). This process provides 64 potential single biomarkers. We also removed drugs used in less than 20% of the patients, and those without a candidate gene target. After matching samples with *ex-vivo* and WES experiments, we finally get the *ex-vivo* screening of 111 drugs for 319 patients (see Methods for more details). We then applied the MOM algorithm to this cohort to unveil groups of AML patients that share genotype and drug sensitivity. In the first step, MOM normalizes the IC_{50} values to define a score that better defines tumor sensitivity, namely IC50*.

Let us illustrate this with a paradigmatic example. In our dataset, the median IC_{50} for Elesclomol is much smaller than the median IC_{50} for Quizartinib (Figure 3A, left panel). Consequently, Elesclomol seems a better option to treat patients with AML. Figure 3B gives a completely different reading: Elesclomol is more toxic in almost any tissue if compared with the AML lines. On the contrary, Quizartinib is more toxic on AML than in most other tissues. This simple example shows that plain IC_{50} must not be used to select the treatment guideline for the patients. The higher value of IC_{50} for Quizartinib could be corrected by adjusting the dose. In Figure 3A, right-panel, after the normalization, the IC_{50}* for Elesclomol appears less effective, whereas Quizartinib preserves its sensitivity profile, which, in this example, it is related to the *FLT3* status of the tumor.

In the second step, MOM calculates individual associations between drugs and genetic alterations using the IHW strategy (21). This approach sheds light on which drugs can be influenced by patient genotype (Figure 4A). IHW also provides a weight for each genetic variant related to the probability of such variant to be a true positive. Non-zero IHW weights represent genetic variants that reduce the FDR and increase the power of tests as demonstrated by IHW authors (21). IHW estimates that, in our AML cohort, 37 biomarkers have weights greater than zero. IHW weights can be therefore used to state the relevance of each biomarker. We sorted IHW weights confirming that $FLT3^{Mut}$, $NPM1^{Mut}$, $NRAS^{Mut}$, $TP53^{Mut}$, and $KRAS^{Mut}$ are included in the top 5 biomarkers (Figure 4B), which have already been described in previous studies (24–29). IHW also provides an adjusted p-value for each drug-biomarker association. For instance, the pipeline identified the known relation of *FLT3* internal tandem duplications (*FLT3-ITD*) patients being more sensitive to Sorafenib, Quizartinib, or Gilteritinib (Supplementary Table 1; Supplementary Figure 21).

Interestingly, an indirect output of this second MOM step is the quantification of the sensitiveness or resistance triggered by a specific genetic variant. Summarizing this score, gene variants can be classified by their effect: either sensitive or resistant to the tested drugs (Figure 4C). For example, variants in *FLT3* or *NPM1* are associated with a more sensitive response for the cohort of drugs in this experiment, whereas genetic alterations in *KRAS*, *NRAS*, or *TP53* are more likely resistance-conferring. Other results include *CCND3*, *WDR52*, *CELSR2*, *CBFβ-MYH11*, and *SMC1A* as biomarkers of sensitivity and *STAG2* of resistance. This effect is relative to the studied dataset, Beat AML, and occurs across 66 different drugs studied or prescribed for hematological malignancies.

Finally, in the third step, we solved the MILP problem from MOM using the individual candidate associations. As a result, MOM returns a decision tree that, depending on the presence or absence of several biomarkers, recommends a treatment for each patient. In this case, the patients are divided into four subgroups (one for each level of the tree) denoted by $FLT3^{Mut}$, $NRAS^{Mut}$, and inv(16) biomarkers (Table 1; Figure 5).

Following the new therapeutic strategy, the first biomarker is $FLT3^{Mut}$-including *FLT3-ITD*. Patients carrying $FLT3^{Mut}$ would be treated with Quizartinib a 2nd generation *FLT3* inhibitor that is currently facing several clinical trials showing an increase in overall survival for AML patients (18). This group of patients represents 30% of patients (25), in our study, 103 patients out of 319 belong to this group. The second subgroup comprises 15 patients and is characterized by $FLT3^{WT}$ and the inv(16), which generates the fusion gene *CBFβ-MYH11*. Patients with these biomarkers are sensitive to Trametinib, a *MAPK* inhibitor that prevents cell replication and has been initiated in phase I clinical trials for hematological malignancies (30). Interestingly, within this group, the patients with $NRAS^{Mut}$ (4 out of 16) are the most sensitive to Trametinib. The third group is defined by the absence of previous biomarkers and $NRAS^{Mut}$. This subgroup poses special interest in the research as *NRAS* is one of the biomarkers most closely related to the general resistance to treatments of this disease (31). *NRAS* gene variants are mutually exclusive with *FLT3* variants (p-value<0.05; Supplementary Figure 16). Patients within this subgroup are sensitive to Selumetinib, a *MAPK* inhibitor that has started clinical trials for acute lymphoblastic leukemia in the UK (32),

10.3389/fimmu.2022.977358

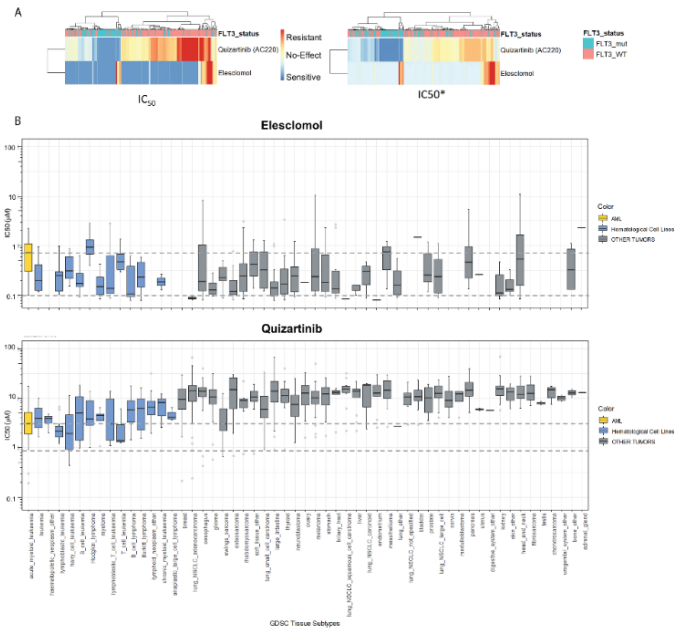


FIGURE 3
IC_{50} Normalization to Avoid Drug Toxicity. **(A)** Drug Sensitivity Heatmap in BeatAML cohort. The left panel shows the IC_{50} values for AML tumors of BeatAML. Effectiveness of a drug in a patient is plotted in blue color, and resistance is represented in red color. The right panel shows the sensitivity in IC50* score. **(B)** Drug sensitivity of Quizartinib and Elesclomol across different tissue types using GDSC. IC_{50} values relative to different tissues are shown in the graph. In yellow color are plotted the sensitivity values of AML cell lines, in blue color are plotted the drug sensitivity values for the Hematological cell lines, and finally in grey color, are plotted the sensitivity values for the non-hematological tissues from GDSC. Dotted grey lines represent the second IC_{50} quantile for AML cell lines (GDSC: Genomics of Drug Sensitivity in Cancer). IC50* is the name of our new sensitivity score.

it is a mitogen-activated pathway inhibitor, which could inhibit *RAS* pathway functionality (33).

Finally, the fourth subgroup comprises the rest of the patients with none of the above mutational biomarkers but with other possible mutated biomarkers, for which the best treatment is Crizotinib -an *ALK* and *MAPK* inhibitor- approved by the FDA for lung cancer. It has not been enrolled in clinical trials for AML. Nevertheless, it has been used in studies of high-risk AML patients, with $TP53^{Mut}$ and obtained very promising results (34).

To validate the MOM's algorithm, we first run MOM on the BeatAML *ex-vivo* dataset using 10-fold cross-validation and compare the results that MOM outputs with each fold. This analysis shows that the MILP optimization returns robust results as 90% folds share 4 out of 5 biomarkers (Supplementary Figure 5). Specifically, $FLT3^{Mut}$ and $NRAS^{Mut}$ subgroups appear in 10 out of 10 folds and subgroup with inv(16) in 3 out of 10 folds.

We then evaluated the treatment guideline proposed by running MOM with BeatAML within three independent AML datasets: two large-scale loss-of-functionality experiments that used both RNAi (DEMETER 2 (35)) and *CRISPR*-Cas9 (CERES (36, 37)), and an additional large-scale cell-drug sensitivity analysis (Genomics of Drug Sensitivity in Cancer, GDSC (1, 38, 39)). We characterize cell lines using the Cancer Cell Line Encyclopedia's (CCLE (40, 41)) genetic variant files, from which we clustered the AML cell lines into the four subgroups predicted by MOM using as input BeatAML. For CERES and DEMETER 2, we identified the main target and model drug effects to be proportional to the depletion of their target, which is the information these databases included.

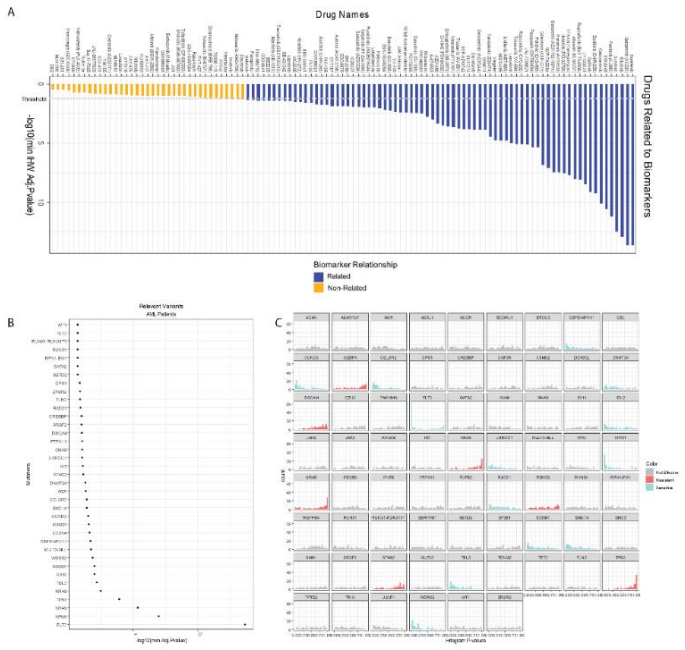


FIGURE 4
Analysis of single interactions biomarker-drug. **(A)** Overall score of 122 drugs whose IC50* is related or non-related to cell genotype according to our model. A drug is related to a relevant variant (those whose IHW weight is greater than zero) if its adjusted p-value is below 0.05. **(B)** Global effect of AML gene variants in AML drug sensitivity. The x-axis shows the logarithm of the minimum adjusted p-value of the biomarker with any of the drugs. Only those biomarkers whose IHW weight is greater than zero are shown. **(C)** One-tail p-value histograms comparing drug sensitivity according to the biomarker status in AML. If a histogram has a strong peak near zero, patients with the biomarker are sensitive to many drugs. On the contrary, if a histogram has a strong peak near one, patients with the biomarker are resistant to many drugs. A genetic variant is considered to confer sensitiveness if the number of drugs whose p-value<0.2 is twice the number of p-values >0.2. Similarly, a variant confers resistance if fulfills that the number of p-values>0.8 is twice the number of p-values<0.8. (IHW: Independent Hypothesis Weighting).

For each subgroup, we compared each experiment's sensitivity (CERES score, DEMETER 2 score, and GDSC-IC50) dividing patients according to the presence of the biomarkers predicted by MOM in BeatAML and summing their sensitivity scores of the other three databases. We compute the sensitivity scores for the 4 subgroups, and the 3 datasets independently DEMETER2 (n=18 AML cell lines), CERES (n=14 AML cell lines), and GDSC (n=23 AML cell lines) (Figure 5). For the GDSC dataset, we compared the

TABLE 1 MOM Output: Patient stratification based on drug response to guide clinical decision-making.

Name	Biomarkers	Drug	Patients Treated
Subgroup 1	$FLT3^{Mut}$	Quizartinib	103
Subgroup 2	$FLT3^{WT}$ & inv(16)	Trametinib	15
Subgroup 3	$FLT3^{WT}$ & no inv(16) & $NRAS^{Mut}$	Selumetinib	42
Subgroup 4	$FLT3^{WT}$ & no inv(16) & $NRAS^{WT}$	Crizotinib	159

10.3389/fimmu.2022.977358

FIGURE 5
Decision Tree for the Proposed Patient Stratification using MOM. MILP from MOM obtained a hierarchical clinical guideline for patient stratification consisting of 4 different subgroups. Each of them is denoted by a biomarker and represented by color (blue, turquoise, orange, and grey). These subgroups were validated in the BeatAML ex-vivo cohort, CERES, DEMETER2, and GDSC. Boxplots show the results of the validation. The y-axis represents the essentiality score from the different experiments and the x-axis represents the biomarker presence-absence of the samples. The validation was performed sequentially, already treated samples from previous subgroups were excluded in the following subgroups i.e. samples with $FLT3^{Mut}$ (blue) from the first boxplot are not plotted in the non-biomarker (grey) in the second boxplot. CERES and DEMETER2 do not have experiments with cell lines having inv(16).

IC_{50} value from the cell lines with the selected biomarker and without the biomarker for a given subgroup drug. Finally, we performed an additional validation using DEMETER RNAi dataset (n=15 AML cell lines; Supplementary Figures 7-8).

The change in sensitivity for the selected treatments is strongly significant using the MOM's predicted biomarkers in the three experiments (p-values of 5.5e-05, 6.8e-06, and 5.5e-04 for CERES, DEMETER2, and GDSC, respectively;

Supplementary Figures 9-11). Remarkably, inv(16) is difficult to be validated using cell lines, as commercial cell lines mostly lack this alteration. The ME-1 cell line is an exception to that, but GDSC is the only dataset that includes the translocation. Although this comparison is not statistically significant due to the lack of data, the GDSC-IC50 of ME-1 is 30 times lower than the average of cells without inv(16).

We carried out a functional enrichment analysis to unveil the patient genotype according to the stratification proposed by MOM. We calculated the differentially expressed genes that are representative of each subgroup (Supplementary Tables 2-5) and computed the enriched biological functions of patients that belong to each group. The first subgroup, defined by $FLT3^{Mut}$, is characterized by downregulation in Myeloid Leukocyte Migration (adjusted p-value< 5e-3; Supplementary Figure 23, Supplementary Table 7), this result is present in other functional enrichment studies involving *FLT3* mutated subgroup (42, 43). This subgroup has been repeatedly mentioned in literature and *FLT3* inhibitors are being implemented in the clinic (18). The second subgroup, defined by samples with inv(16) and $FLT3^{WT}$ shows upregulated cell proliferation (adjusted p-value< 1e-3) including angiogenesis and endothelial cell migration upregulated among others (Supplementary Figure 24, Supplementary Table 8), also described in other studies concerning this genetic aberration (44–46).

We also found that the $NRAS^{Mut}$ subgroup is related to the downregulation of alternative splicing (AS; adjusted p-value< 0.2; Supplementary Figure 27, Supplementary Table 11). This subgroup has an upregulation of the transforming growth factor-beta (TGF-β) signaling pathway (adjusted p-value< 5e-03; Supplementary Figure 26, Supplementary Table 10), which is mentioned in other studies concerning AS, especially in myelodysplastic syndromes (47, 48). Furthermore, several studies have attempted to address the relationship between AML and AS, with promising results (49–51).

Finally, patients who do not have the previous biomarkers, have a downregulation in the amino acid catabolism process (adjusted p-value< 0.05; Supplementary Figure 29, Supplementary Table 13), i.e. they are less able to metabolize amino acids than the rest of the subgroups (52). A study demonstrates that for a subpopulation of AML leukemia stem cells the metabolism of amino acids from the medium is essential, and its absence leads to cell death (52). Further description of the enriched functions for each subgroup, as well as their relationships and statistical significance, can be found in the supplementary material (Supplementary Figures 22-29, Supplementary Tables 6-13).

3 Discussion

Despite the advances in drug *ex-vivo* screening and computational methods for precision medicine, there are technical issues that limit their translation to clinical practice. Some of these issues are the influence of drug toxicity, the enormous number of statistical hypotheses, the complexity of developing algorithms understandable by the clinician, and the difficulty of proposing an effective treatment guideline that assigns the best drug for each patient. MOM faces and solves each of these challenges.

These statements are not yet covered by current AI strategies, which are focused on increasing accuracy and sensitivity regardless of the complexity of the end model (7, 53). In these AI methods, the absence of interpretability of the feature used for classification prevents further research and downplays the need for clinically defined subgroups (54–56). Indeed, the need of developing XAI algorithms is not only related to easing the diagnosis pipeline in cancer but also to increase and facilitate that the pharma industry brings new drugs and biomarkers to market. Drug regulators -such as the Food and Drug Administration- value that the process to unveil novel biomarkers is robust and transparent (10). In contrast, the patient stratification guideline provided by MOM has the following characteristics, i) allows treatment assignment by using a simple genetic panel, ii) the results are non-stochastic, they are the same for all possible re-runs of the model, iii) the algorithm outputs a decision tree for treatment guidance.

IC_{50}, $EC_{50,}$ and AUC (used for example in (1, 6, 38)) are reasonable metrics to determine the efficacy of a drug. None of them, however, considers the overall toxicity of the drug. Using IC50* in the optimization problem, we focus on the differential effectiveness of a drug among different patients, and therefore, drugs that are toxic for most samples will not be included in the solution.

IHW provides us with the ability to increase the power of tests and reduce the FDR. With this strategy, we are also able to identify the direction of the influence of genetic events in drug response, i.e., whether it defines sensitivity or resistance. With this approach, we successfully detected *FLT3* as highly influential in terms of sensitivity to treatment, which is coherent with other studies (25). *NRAS*, instead, showed as a mutation associated with treatment resistance also coherent with literature (26, 31). One promising conclusion for this study is that we managed to find a drug for which *NRAS* correlates to drug sensitivity.

XAI defined by MILP ensures that the subgroups obtained are optimal. This feature is not common to other classification methods. However, it also presents two main limitations. The first one is computational resources, which increases exponentially with the number of possible biomarkers, drugs, or patients (on a standard desktop, the presented work required 2.5 hours of computing time). In addition, the incorporation of new non-binary diagnostic markers requires the redefinition of the model. However, once the optimization problem is solved, assigning a treatment to a novel patient is immediate.

Our AML patient stratification includes a subgroup defined by the absence of a genetic mutation, i.e., wild type. It also

includes patients who have $TP53^{Mut}$ genotype, which are biomarkers associated with poor prognosis (14). MOM recommends treating these patients with Crizotinib, a drug used in other studies with $TP53^{Mut}$ AML patients which in fact showed very promising results (34). In addition, this subgroup shows a deficiency in amino acid metabolism which may lead to alternative treatment therapies based on metabolomics.

The subgroup defined by the *CBFβ-MYH11* fusion gene appears characterized in a very small percentage of AML cell line cohorts but is nevertheless present in 7% of AML patients (57), which enhances the relevance of this biomarker. *CBFβ-MYH11* is a clear indicator of sensitivity to Trametinib, a clinical drug that inhibits cell replication pathway (58), which, in turn, appeared as an upregulated biological process in this subgroup. In the remaining subgroups, $FLT3^{Mut}$ is widely described in the literature (25). In contrast, $NRAS^{Mut}$ appears as a biomarker of sensitivity for Selumetinib and has downregulated the alternative splicing (AS) process. This subgroup contains, on balance, effective treatment for a resistance-associated mutation and a new line of research linking alternative splicing and AML.

It is remarkable the appearance of three different *MAPK* inhibitors in the proposed therapeutic strategy, which is coherent with the disease behavior. Our biomarker analysis revealed that the RTK-RAS pathway is the most affected in our cohort of AML samples (Supplementary Figures 18-19). Of all drugs suggested as treatment, only Quizartinib is clinically approved for AML patients (15). This study aims to accelerate -once the results are validated in cell lines and murine models- the process of approving these drugs for AML.

The validation of the results is challenging in a real cohort since most patients are treated with standard induction cytotoxic therapy (only 7.5% of AML patients in TCGA are treated with other treatments). We propose a strategy to take advantage of cell lines loss-of-function datasets. Nevertheless, even using cell lines -that are quite different from ex vivo samples- we validated the subgroups and the IC_{50} of the lines with indication was significantly better than the IC_{50} of those without indication. Therefore, in the absence of clinical data for validation, we consider the results using cell lines data to sufficiently support this study.

The concept of MOM is also applicable to other disease types using *ex-vivo* experiments as well as to other sensitivity measurements, leaving an open door for new patient stratifications based either on drug response or even on any other experiment to measure the effectiveness of certain drugs in the future. We believe that XAI will help doctors and regulators understand AI medical decisions and, therefore, ease the translations of AI analysis of drug screening experiments to clinical practice.

4 Methods

4.1 Filter and normalization

4.1.1 Filtering and imputation

We used data from *ex-vivo* experiments, WES, and RNA-Seq from 319 Acute Myeloid Leukemia (AML) patients included in the BeatAML cohort (19). Data was filtered to ensure all samples contained the gene variants and drug sensitivity information, the new dataset containing genomic aberrations and drug IC_{50} for the same patients was used as a starting point for the study. Genetic variant samples were previously pathogenically filtered by Tyner et al. (19) and we defined as a biomarker a genetic variant present in more than 1% of the patients (n≥4), leaving a total number of 64 possible biomarkers.

For missing drug sensitivity information in the ex-vivo experiments, we imputed the missing data using the *k*-Nearest Neighbourhood (kNN) Impute method, from Impute R package (59) (version 1.68.0). An analysis of the missing values –both patients and drugs– is included in the supplementary material

4.1.2 Drug normalization: from IC_{50} to IC50*

Initially, we tried to use as drug sensitivity values the half-minimal inhibitory concentration, (IC_{50}) i.e., the concentration of a drug -in micro molar- for which half of the cell from the ex-vivo experiment die. Instead of using the IC_{50}, we propose the usage of an incremental version of the IC_{50}, named IC50*. As described in the results section, the usage of IC50* instead of IC_{50} is a convenient way to deal with the different toxicity of the drugs under study

After imputation, IC_{50} values were taken the $\log_{10}$ logarithm, normalized by subtracting the IC_{50} mean value for each drug, and these scores were made negative by subtracting an offset to the normalized IC_{50} value –the optimization model assumes negative values of drug sensitivity. The obtained drug sensitivity values are named IC50*. The transformation from IC_{50} to IC50* is represented in equation (1). Despite the formidable aspect of the formula, IC50* is simply an incremental and version of the logarithm of IC_{50} with an offset.

Let IC_{50} be a T x P matrix, with T the total number of drugs and P the total number of patients, for which each element $ic50_{t,p}$ is a value contained in (0,10] μM.

$$ic50^{*}_{t,p} = (\log_{10}(ic50_{t,p}) - 1) - \frac{1}{P}\sum_{p=1}^{P}(\log_{10}(ic50_{t,p}) - 1) - \max\left((\log_{10}(ic50_{t,p}) - 1) - \frac{1}{P}\sum_{p=1}^{P}(\log_{10}(ic50_{t,p}) - 1)\right) \quad (1)$$

The obtained **IC50*** is a T x P matrix containing the new drug sensitivity values.

4.2 Drug-biomarker association

Following with MOM's second step, we implemented a two-tailed Wilcoxon test to assess whether a biomarker influences the sensitivity of each the treatment. Each biomarker is tested against each drug and these associations were ranked according to the p-value. The p-values were adjusted following the methodology described by Gimeno *et al.* (22), using the R package *IHW* (21) (version 1.22.0). The package provides (given the p-values and the covariates –in our study genetic alterations–) a weight for each covariate related to its influence on the p-value significance.

Using these results, we included two consecutive filters. Firstly, we selected the biomarkers whose relative importance (the weight outputted by IHW) is larger than zero. IHW assigns a strictly positive weight to biomarkers relevantly correlated to the potency of a drug. Afterwards, we removed the drugs with no statistically significant relationship to the selected biomarkers (IHW p-value >0.05).

After this analysis, 122 treatments (biomarker-drug associations), with $\Delta IC50^{*}>0.2$ (including *vs* lacking the biomarker) and adjusted p-value<0.05 were considered for therapy.

4.3 MOM: MILP Module

Finally, in the third step, we proceed with the treatment assignation. We developed a MILP module described in the Results section. This module receives as input the 122 treatments and solves an optimization problem (described in detail in de Supplementary Material) MILP results can be directly translated into a decision tree for guiding clinical decision-making. The number of levels of the tree was set to four. Each level of this tree will be defined as one therapeutic AML subgroup and each subgroup is defined by a biomarker and a recommended drug.

Additional information regarding the algorithm, its *in-silico* validation, and its performance can be found in Supplementary Material (Section Supplementary Methods).

4.4 External cohort validation

For validating the different subgroups, we compared patients that are given a drug in a specific subgroup against the remaining non-treated patients. We validated our results using cell lines, specifically, used 2 different large-scale gene essentiality experiments including RNAi (DEMETER 2 (35)) and CRISPR-Cas9 (CERES (36, 37)), and an additional large-scale cell-drug sensitivity analysis (Genomics of Drug Sensitivity in Cancer, GDSC (1, 38, 39)). We characterized the cell lines using the Cancer Cell Line Encyclopedia (CCLE (40, 41)) genetic variants files, from which we were able to divide the cells into different subgroups.

We performed the following test for validation. Cells were divided into two groups. The first group includes cells with the biomarker associated to that subgroup, and the other group, contains the cells without the biomarker that had not been previously treated. This comparison was computed for the 4 subgroups, and the 2 datasets DEMETER 2, and CERES. DEMETER 2 and CERES were compared using the viability score that corresponds to knocking out the corresponding targets for each drug. For the GDSC dataset, we used the IC_{50} value provided in the experiments. All tests were one-tailed Wilcoxon's test to check that the sensitivity increase in the cells with the biomarker.

4.5 Functional analysis of the subgroups

Functional analysis of the subgroups was performed using gene expression data from the BeatAML (19) cohort. We performed a differential gene expression analysis using limma R package (60) (version 3.50.3). The contrast matrix compared one group against all the others, therefore, there was a different contrast for each group.

Genes differentially expressed were ranked according to its t-statistic, if t >0, genes were considered overexpressed, if t<0, genes were considered underexpressed. For each subgroup, we selected the top 500 over and under expressed genes and performed a Gene Ontology Enrichment Analysis (GEA) using Fisher's Test. We analyzed the biological process ontology. Enriched functions on the overexpressed genes were upregulated, and functions obtained from the underexpressed genes were considered to be downregulated. The statistics were computed using clusterProfiler R package (61) (version 3.10.1). We set an adjusted p-value cutoff of 0.2 for considering a function differentially enriched, adjusted p-values were computed using the Benjamini-Hochberg procedure.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://vizome.org/additional_figures_BeatAML.html

Author contributions

MG, AR and FC conceived this study. MG, EJ-E, SV, XA, FP, AR and FC designed the MOM requirements and provided

biological insights for the assignation problem. MG and FC developed the pre-processing pipeline. MG, AR and FC carried out the computational implementation and validation. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Cancer Research UK [C355/A26819] and FC AECC and AIRC under the Accelerator Award Programme, and Synlethal Project (RETOS Investigacion, Spanish Government).

Acknowledgments

The authors would like to thank Francisco J. Planes, Iñigo Apaolaza, and Luis V. Valcárcel for the fruitful comments on the development of the methodology. The authors would like to acknowledge Katyna Sada for proof-reading and her suggestions to improve readability.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2022.977358/full#supplementary-material

References

1. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nat* (2012) 483:570–5. doi: 10.1038/nature11005

2. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* (2011) 10:188–95. doi: 10.1038/nrd3368

3. McVeigh TP, Hughes LM, Miller N, Sheehan M, Keane M, Sweeney KJ, et al. The impact of oncotype DX testing on breast cancer management and chemotherapy prescribing patterns in a tertiary referral centre. *Eur J Cancer* (2014) 50:2763–70. doi: 10.1016/j.ejca.2014.08.002

4. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: Another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn* (2009) 9:417–22. doi: 10.1586/erm.09.32

5. Wu L, Yang Y, Guo X, Shu XO, Cai Q, Shu X, et al. An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk. *Nat Commun* (2020) 11:1–11. doi: 10.1038/s41467-020-17673-9

6. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* (2020) 38:672–684.e6. doi: 10.1016/j.ccell.2020.09.014

7. Malani D, Kumar A, Brück O, Kontro M, Yadav B, Hellesøy M, et al. Implementing a functional precision medicine tumor board for acute myeloid leukemia. *Cancer Discov* (2022) 12:388–401. doi: 10.1158/2159-8290.CD-21-0410

8. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intell* (2020) 2:573–84. doi: 10.1038/s42256-020-00236-4

9. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol* (2020) 4:19. doi: 10.1038/s41698-020-0122-1

10. U.S. Food and Drug Administration. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)-Discussion Paper and Request for Feedback*. FDA. (2019) 20. Available from: https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm514737.pdf.

11. European Medicines Agency. *Artificial intelligence in medicine regulation | European medicines agency*. Available at: https://www.ema.europa.eu/en/news/artificial-intelligence-medicine-regulation (Accessed 15th March 2022).

12. Lazar AJ, Demicco EG. Human and machine: Better at pathology together? *Cancer Cell* (2022) 40:806–8. doi: 10.1016/j.ccell.2022.06.004

13. Perry AM, Attar EC. New insights in AML biology from genomic analysis. *Semin Hematol* (2014) 51:282–97. doi: 10.1053/j.seminhematol.2014.08.005

14. Zeisig BB, Kulasekararaj AG, Mufti GJ, Eric So CW. SnapShot: Acute myeloid leukemia. *Cancer Cell* (2012) 22:698–698.e1. doi: 10.1016/j.ccr.2012.10.017

15. Wander SA, Levis MJ, Fathi AT. The evolving role of FLT3 inhibitors in acute myeloid leukemia: quizartinib and beyond. *Ther Adv Hematol* (2014) 5:65–77. doi: 10.1177/2040620714532123

16. NIH N. C. I. G. D. C. Acute Myeloid Leukemia — Cancer Stat Facts.

17. Ragon BK, Odenike O, Baer MR, Stock W, Borthakur G, Patel K, et al. Oral MEK 1/2 inhibitor trametinib in combination with AKT inhibitor GSK2141795 in patients with acute myeloid leukemia with RAS mutations: A phase II study. *Clin Lymphoma Myeloma Leuk* (2019) 19:431–440.e13. doi: 10.1016/j.clml.2019.03.015

18. Sutamtewagul G, Vigil CE. Clinical use of FLT3 inhibitors in acute myeloid leukemia. *Onco Targets Ther* (2018) 11:7041–52. doi: 10.2147/OTT.S171640

19. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* (2018) 562:526–31. doi: 10.1038/s41586-018-0623-z

20. Snijder B, Vladimer GI, Krall N, Miura K, Schmolke AS, Kornauth C, et al. Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. *Lancet Haematol* (2017) 4:e595–606. doi: 10.1016/S2352-3026(17)30208-9

21. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods* (2016) 13:577–80. doi: 10.1038/nmeth.3885

22. Gimeno M, San José-Enériz E, Rubio A, Garate L, Miranda E, Castilla C, et al. Identifying lethal dependencies with HUGE predictive power. *Cancers (Basel)* (2022) 14:3251. doi: 10.3390/cancers14133251

23. Guo XX, Wu HL, Shi HY, Su L, Zhang X. The efficacy and safety of olaparib in the treatment of cancers: a meta-analysis of randomized controlled trials. *Cancer Manage Res* (2018) 10:2553. doi: 10.2147/CMAR.S169558

24. Hill R, Cautain B, De Pedro N, Link W. Targeting nucleocytoplasmic transport in cancer therapy. *Oncotarget* (2014) 5:11–28. doi: 10.18632/oncotarget.1457

25. Daver N, Schlenk RF, Russell NH, Levis MJ. Targeting FLT3 mutations in AML: review of current knowledge and evidence. *Leukemia* (2019) 33:299–312. doi: 10.1038/s41375-018-0357-9

26. Wang S, Wu Z, Li T, Li Y, Wang W, Hao Q, et al. Mutational spectrum and prognosis in NRAS-mutated acute myeloid leukemia. *Sci Rep* (2020) 10:12152. doi: 10.1038/s41598-020-69194-6

27. Hunter AM, Sallman DA. Current status and new treatment approaches in TP53 mutated AML. *Best Pract Res: Clin Haematol* (2019) 32:134–44. doi: 10.1016/j.beha.2019.05.004

28. Thiede C, Koch S, Creutzig E, Steudel C, Illmer T, Schaich M, et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood* (2006) 107:4011–20. doi: 10.1182/blood-2005-08-3167

29. Zhang H, Nakauchi Y, Köhnke T, Stafford M, Bottomly D, Thomas R, et al. Integrated analysis of patient samples identifies biomarkers for venetoclax efficacy and combination strategies in acute myeloid leukemia. *Nat Cancer* (2020) 1:826–39. doi: 10.1038/s43018-020-0103-x

30. Wright CJM, McCormack PL. Trametinib: First global approval. *Drugs* (2013) 73:1245–54. doi: 10.1007/s40265-013-0096-1

31. Gui P, Bivona TG. Stepwise evolution of therapy resistance in AML. *Cancer Cell* (2021) 39:904–6. doi: 10.1016/j.ccell.2021.06.004

32. Markham A, Keam SJ. Selumetinib: First approval. *Drugs* (2020) 80:931–7. doi: 10.1007/s40265-020-01331-x

33. Kiessling M, Rogler G. Targeting the RAS pathway by mitogen-activated protein kinase inhibitors. *Swiss Med Wkly* (2015) 145:w14207. doi: 10.4414/smw.2015.14207

34. Antony ML, Noble-Orcutt K, Ogunsan O, He F, Sachs Z. Cell type-specific effects of crizotinib in human acute myeloid leukemia with TP53 alterations. *Blood* (2019) 134:2563–3. doi: 10.1182/blood-2019-130487

35. McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat Commun.* (2018) 9(1):4610.

36. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* (2017) 49:1779–84. doi: 10.1038/ng.3984

37. Wang T, Yu H, Hughes NW, Liu B, Kendirli A, Klein K, et al. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic ras. *Cell* (2017) 168:890–903.e15. doi: 10.1016/j.cell.2017.01.013

38. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. *Cell* (2016) 166:740–54. doi: 10.1016/j.cell.2016.06.017

39. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* (2013) 41:D955–61. doi: 10.1093/nar/gks1111

40. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* (2012) 483:603–7. doi: 10.1038/nature11003

41. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov G V., Lo CC, McDonald ER, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* (2019) 569:503–8. doi: 10.1038/s41586-019-1186-3

42. Chen S, Chen Y, Zhu Z, Tan H, Lu J, Qin P, et al. Identification of the key genes and microRNAs in adult acute myeloid leukemia with FLT3 mutation by bioinformatics analysis. *Int J Med Sci* (2020) 17:1269. doi: 10.7150/ijms.46441

43. Lucena-Araujo AR, Souza DL, De Oliveira FM, Benicio MTL, Figueiredo-Pontes LL, Santana-Lemos BA, et al. Results of FLT3 mutation screening and correlations with immunophenotyping in 169 Brazilian patients with acute myeloid leukemia. *Ann Hematol* (2010) 89:225–8. doi: 10.1007/s00277-009-0817-4

44. Gutiérrez NC, López-Pérez R, Hernández JM, Isidro I, González B, Delgado M, et al. Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. *Leukemia* (2005) 19:402–9. doi: 10.1038/sj.leu.2403625

45. Zhang L, Nguyen LXT, Chen Y-C, Wu D, Cook GJ, Hoang DH, et al. Targeting miR-126 in inv(16) acute myeloid leukemia inhibits leukemia development and leukemia stem cell maintenance. *Nat Commun* (2021) 12:6154. doi: 10.1038/s41467-021-26420-7

46. Wunderlich M, Krejci O, Wei J, Mulloy JC. Human CD34+ cells expressing the inv(16) fusion protein exhibit a myelomonocytic phenotype with greatly enhanced proliferative ability. *Blood* (2006) 108:1690–7. doi: 10.1182/blood-2005-12-012773

47. Bewersdorf JP, Zeidan AM. Transforming growth factor (TGF)-β pathway as a therapeutic target in lower risk myelodysplastic syndromes. *Leukemia* (2019) 33:1303–12. doi: 10.1038/s41375-019-0448-2

48. Muench DE, Ferchen K, Velu CS, Pradhan K, Chetal K, Chen X, et al. SKI controls MDS-associated chronic TGF-β signaling, aberrant splicing, and stem cell fitness. *Blood* (2018) 132:e24–34. doi: 10.1182/blood-2018-06-860890

49. Bowman TV. Improving AML classification using splicing signatures. *Clin Cancer Res* (2020) 26:3503–4. doi: 10.1158/1078-0432.CCR-20-1021

50. De Necochea-Campion R, Shouse GP, Zhou Q, Mirshahidi S, Chen CS. Aberrant splicing and drug resistance in AML. *J Hematol Oncol* (2016) 9:1–9. doi: 10.1186/s13045-016-0315-9

51. Grinev V V., Barneh F, Ilyushonak IM, Nakjang S, Smink J, van Oort A, et al. RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia. *Nat Commun* (2021) 12:520. doi: 10.1038/s41467-020-20848-z

52. Jones CL, Stevens BM, D'Alessandro A, Reisz JA, Culp-Hill R, Nemkov T, et al. Inhibition of amino acid metabolism selectively targets human leukemia stem cells. *Cancer Cell* (2018) 34:724–740.e4. doi: 10.1016/j.ccell.2018.10.005

53. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet* (2017) 49:332–40. doi: 10.1038/ng.3756

54. Shamout F, Zhu T, Clifton DA. Machine learning for clinical outcome prediction. *IEEE Rev Biomed Eng* (2021) 14:116–26. doi: 10.1109/RBME.2020.3007816

55. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '18)*. New York, NY, USA: Association for Computing Machinery (2018). p. 559–60. doi: 10.1145/3233547.3233667

56. Oh M, Park S, Kim S, Chae H. Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Brief Bioinform* (2021) 22:66–76. doi: 10.1093/bib/bbaa032

57. Surapally S, Tenen DG, Pulikkan JA. Emerging therapies for inv(16) AML. *Blood* (2021) 137:2579–84. doi: 10.1182/blood.2020009933

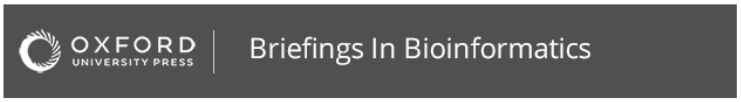
58. Zeiser R, Andrlová H, Meiss F. Trametinib (GSK1120212). In: *Recent results in cancer research*, vol. 211. Springer, Cham (2018). p. 91–100. Available from: https://link.springer.com/chapter/10.1007/978-3-319-91442-8_7

59. Hastie T, Tibshirani R, Narasimhan B, Gilbert C. Impute: Imputation for microarray data. *Bioinformatics* (2001) 17:520–5.

60. Zeiser R, Andrlová H, Meiss F. Trametinib (GSK1120212). In: *Recent results in cancer research*. Springer, Cham (2018). 91–100 p. Available at: https://link.springer.com/chapter/10.1007/978-3-319-91442-8_7

61. Hastie T, Tibshirani R, Narasimhan B, Gilbert C. Impute: Imputation for microarray data. *Bioinformatics* (2001) 17(6):520–5.

Manuscripts submitted to Briefings in Bioinformatics



Precision Oncology: a review to assess interpretability in several explainable methods

Journal:	*Briefings in Bioinformatics*
Manuscript ID	BIB-22-2259
Manuscript Type:	Review
Date Submitted by the Author:	16-Nov-2022
Complete List of Authors:	Gimeno, Marian; Engineering School TECNUN, Biomedical Engineering and Sciences Sada, Katyna; Engineering School TECNUN, Biomedical Engineering and Sciences Angel , Rubio; Engineering School TECNUN, Biomedical Engineering and Sciences; Instituto de Ciencia de Datos e Inteligencia Artificial (DATAI), Universidad de Navarra
Keywords:	Interpretability, Asignation Problem, Precision Medicine, Explainable Artificial Intelligence, Drug Recommendation, Machine Learning

Briefings in Bioinformatics

Precision Oncology: a review to assess interpretability in several explainable methods

Marian Gimeno[1], Katyna Sada del Real[1] and Angel Rubio[1,2]*

[1] Departamento de Ingeniería Biomédica y Ciencias, TECNUN, Universidad de Navarra, 20009, San Sebastián, Spain.
[2] Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), Universidad de Navarra, 31008, Pamplona, Spain.

Corresponding Author
Angel Rubio
Tel.:+34 943 21 98 77; E-mail: arubio@tecnun.es

Abstract

Great efforts have been made to develop Precision Medicine (PM)-based treatments using Machine Learning. In this field, where the goal is to provide the optimal treatment for each patient based on his/her medical history and genomic characteristics, it is not sufficient to make excellent predictions. The challenge is to understand and trust the model's decisions while also being able to easily implement it. However, one of the issues with machine learning algorithms -particularly deep learning- is their lack of interpretability. This review compares six different machine learning methods to provide guidance for defining interpretability by focusing on: Accuracy, Multi-omics Capability, Explainability, and Implementability. Our selection of algorithms includes tree, regression, and kernel based methods, which we selected for their ease of interpretation for the clinician. We also included two novel explainable methods in the comparison. There were no significant differences in accuracy when comparing methods or when using gene expression instead of mutational status as input to those methods. This allowed us to concentrate on the current intriguing challenge: model comprehension, and ease of use. We discovered that the tree-based methods were the most interpretable of those tested.

Keywords:

Interpretability, Precision Medicine, Machine Learning, Explainable Artificial Intelligence, Drug Recommendation, Asignation Problem, Method Comparison.

Briefings in Bioinformatics

Introduction

Precision medicine (PM) is the science that "defines a disease at a higher resolution by genomic and other technologies to enable more precise targeting of its subgroups" [1]. It is an emerging field that epitomes the new era of medicine owing to its applications in clinical treatment and diagnosis [2].

PM tries to find not only the right drug but also the right dosage and the proper treatment schedule. These goals are usually summed up as "targeting the right treatments to the right patients at the right time" [3]. In this review, we will focus on considering the patients' genome, environment, and lifestyles to provide each patient with the "best" treatment according to these characteristics. PM faces different challenges that will be described in this introduction.

The challenge of getting the patients' response to drugs.

PM requires the different patients' characteristics to make their predictions [4] such as genomic and transcriptomic data, health records, lifestyle characteristics, etc. (**Figure 1**). With an adequate data policy, these characteristics are reasonably easy to obtain; genomic data can be acquired from sequencing techniques, wearable technologies can collect data that provide lifestyle information, EHRs are invaluable sources of information on health status and previous conditions, etc. Its integrative analysis requires complex models and a solid understanding of the interaction of biological systems [5].

However, PM also requires drug sensitivity information which is much more difficult to find, having most likely incomplete information on all patients' response to all available drugs, i.e. each patient is given one or, at most, a few drugs, not all the possible ones[6] (**Figure 1**). Even in these cases, distinguishing between responders and non-responders is not an easy task and requires tailoring methods specific to each disease. In turn, these different criteria for different diseases make it difficult to compare diseases or drugs [7].

PDX (patient-derived xenografts) or *ex vivo* experiments can be used as proxies to estimate the patients' response to several drugs [8]. Both approaches have strong limitations. In the case of PDX, the animal models' immune system is usually compromised and, in the case of ex vivo experiments, –used mainly in hematologic oncology– the interaction of the cells and the immune system is not properly modeled. Despite these difficulties, they are reasonable sources of information to predict the response of the patients to different treatments [9].

PM falls beyond traditional machine learning (ML) problems.

Precision Medicine can be considered an assignation problem: each patient must be provided a drug (or a set of drugs) given the patient's information. This assignation problem does not perfectly fit in any of the "traditional" fields of machine learning. It is not a unsupervised problem, although, with a proper selection of variables, patients with identical "optimal" drugs should cluster together[10,11].

Regarding supervised machine learning, it is not either a standard regression problem since the aim is not to predict the effectiveness of a drug on each patient but to find *the most effective ones* [12]. Nevertheless, both problems are related and, if the effectiveness of each drug were exactly modeled, the "perfect" drug for a patient would be simply the most effective one predicted by the model. It could also be treated as a classification problem dividing the drugs for each patient into two classes: the most

Briefings in Bioinformatics

effective one belongs to one class and the others belong to another class. Again, it solves the problem if the predictions were perfect. However, since this simplistic model only considers misclassifications (the second best drug is as bad as the worst), it does not work well in practice.

Finally, it can also be considered a reinforced learning problem [13]. For example, [14] includes a review of reinforcement learning applications to oncology. The objective of this field of machine learning is to learn an optimal, or nearly optimal, policy that maximizes the "reward function" –in this case, the patient's response to treatment. Reinforced learning is traditionally applied to teach the computer how to play games (chess, Go, or video games) [11]. Applied to PM, different methods state how use a reinforced learning algorithm to "find a policy that maximizes the patient response to treatment".

As a result, precision medicine –assigning the proper drug to each patient– given the patients' data is a problem that shares characteristics of different machine learning fields (**Figure 2**) and can be tackled in many ways as will be shown in the different analyzed approaches.

Patient-centered vs. drug centered.

There are two main approaches to solve the goal of "targeting the right treatments to the right patients" (**Figure 1**). The first one is to state which is the proper drug for a specific patient. We will name this approach "patient-centered". The other approach consists of finding the patient or patients that are responders for a specific drug, named "drug-centered" in this review. This problem –closely related to finding biomarkers of response– is interesting for the pharma industry.

If the output of the algorithm is a continuous value, it is possible to adapt a drug-based method to solve the patient-based problem and vice-versa. For example, many drug-centered methods return a sensitivity score for each patient when applied to a specific drug. If this score is computed for all the drugs, it can be used to select the drug that maximizes sensitivity for each patient.

The challenge of interpretability.

One common problem in ML is the (lack of) interpretability. In many cases the algorithm is a blackbox that gives no clues on why a specific decision is taken [15–27]. It is difficult, if not reckless, for a physician to use a treatment guideline with no information on the ultimate reasons that drove this recommendation. Explainable artificial intelligence (XAI) is an active field of research: it justifies the response and ensures that, given the *a priori* knowledge, the recommendation is sensible. XAI also helps to improve the results since, as they are understandable by physicians, they can provide expert feedback to fine tune the algorithms [28–31]. Some methods have tried to explain their reasoning to become more explainable [29,32–38].

Treatment guidelines that require few biomarkers are easier to understand by a human. Therefore, the number of variables is one of the characteristics used to determine how explainable a method is. Some methods [17,18,25–27,34,39,40] automatically select the optimal number of variables to accomplish a task. In other cases, the selection of variables must be done beforehand using either a filter or a wrapper technique depending on whether the result of the predictions is included in the loop to select the variables [19,41–44].

Briefings in Bioinformatics

Method of comparison.

In this review we included interpretable methods amenable to the precision medicine problem, i.e. find the best drug for a patient. We systematically reviewed the current literature to summarize the state-of-the-art and compare different methods that solve the assignation problem.

Some of the methods available in the literature −that are said to be interpretable and output a drug assignation− refer to methods that are currently under the "patient-centered" approach [45,46], others are nested under the "drug-centered" approach [47–49]. From the latter, only the methods that predict a continuous variable for drug sensitivity could be used for patient assignation.

We compared the methods in this work in terms of interpretability, focusing specially on accuracy, multi-omics capacities, and translation into clinical practice. Method comparison was performed using BeatAML [50] dataset and Genomics of Drug Sensitivity in Cancer (GDSC) [51] dataset for Acute Myeloid Leukemia (AML). We focused on the BeatAML dataset due to its abundancy of patient information –e.g. genomic data, gene expression, clinical data–, and drug sensitivity information which proceeded from ex-vivo experiments performed on patient samples instead of cell lines [50]. Indeed, ex-vivo drug sensitivity provided more information for patient sensitivity than conventional information from clinical data due to the possibility of testing more drugs on the living tumor without injuring the patient and solving possible harmful drug interactions from previous treatments. Using this information –although it could be less reliable– solves the sparsity issue of drug sensitivity data. Furthermore, drug screens performed on ex-vivo experiments improve data reliability if compared to cell line screenings. Nevertheless, further experimental validation is required for clinical applications.

AML is an infrequent blood tumor that originates in the bone marrow of the patients and has a very poor progression-free survival. In addition, it is a highly heterogeneous disease for which finding effective treatments is a challenge [30,52–54]. Due to the technical difficulties in finding suitable data from this disease to implement a common ML model, the need to find therapeutic strategies, and the availability of ex-vivo drug screening experiments, we believe that the BeatAML dataset is perfectly suited for this comparison.

Briefings in Bioinformatics

Methods

Focusing on approaches to address the complex Precision Medicine (PM) problem, we found two methodologies from the "patient-centered" perspective, Multidimensional Optimization Module (MOM) [45] and Kernelized Rank Learning (KRL) [55]. We also included in this group two novel algorithms: Optimal Decision Trees (ODT) and an adaptation of the Multinomial Lasso. Both are described in more detail in the Methods section. MOM uses mixed integer linear programming (MILP) to discover the optimal therapeutic strategy that is returned as a decision tree. KRL is a machine learning method based on an optimization problem -minimizing the sensitivity error-, it applies a kernel to solve the convexity limitations and solves the problem also using MILP. ODTs are decision trees that recursively optimize the drug recommendation on each branch until a preset group size is reached. Finally, Multinomial Lasso is a modified Lasso regression methodology for which each patient "selects" its best drug using a vote sharing scheme.

In the "drug-centered" approach, BOSO [47] and Lasso Regression [49] can be applied to predict the IC_{50} of a drug in different patients. Both methods select a small number of variables to make their predictions. Once the predictions are obtained, comparing the predicted IC_{50} for each drug on a patient, the drug with the minimal IC_{50} is selected. BOSO is a MILP model built up from the Lasso Regression equations that have been modified to predict a numeric variable with the least number of features, improving the reduced interpretability of Lasso Regression. The description of the six methods is summarized in **Table 1**. Some of the methods only accept binary data as input. These methods cannot be applied to gene expression unless using a hard threshold.

Optimal Decision Trees (ODT)

In this work, we are introducing a novel algorithm that uses a tree-like method for precision medicine. This method is intrinsically different from classification or regression trees, as will be shown.

In a classification tree, in each step, the tree is split into two subtrees finding the variable (with its corresponding threshold) that best splits the tree according to some figure of merit (Gini index, entropy, information gain, etc.). This figure of merit measures the overall enrichment of the classes in the subtrees.

On the contrary, the ODT algorithm selects for each step the splitting variable (selecting a proper threshold) *and the treatments for each split*. The selection is based on the optimization of an overall measure of the sensitivity of both branches to the selected treatments (**Error! Reference source not found.**).

Specifically, let **Y** be a P x D matrix where P is the number of patients and D is the number of tested drugs. Each of the entries of the matrix quantifies the sensitivity of each patient to a drug, i.e., the matrix Y can be either the IC_{50} or a modified version of it, the area under the concentration-response curve, etc. Let **X** be a P x M matrix where P is the number of patients and M is the number of biomarkers. The matrix **X** can be a matrix of mutations, gene expression, or other characteristics specific to each patient.

In the case of binary variables (mutations for example), for each step in the splits of tree, the following optimization problem is solved (**Equations 1-3**):

$$\max_{m,\, d_1, d_2} \; A + B \qquad (1)$$

Briefings in Bioinformatics

$$A = \sum_{p \in split} y_{pd_1}(x_{pm} == 1) \quad (2)$$

$$B = \sum_{p \in split} y_{pd_2}(x_{pm} == 0) \quad (3)$$

Where split is the set of patients under study (all patients are used in the case of the root node), *m* is the selected mutation or biomarker, and d_1 and d_2 are the selected drugs for the patients that have or do not have the mutation m respectively. The notation "(*condition*)" represents 1 or 0 depending on whether the expression inside the parenthesis is true or false (**Equations 2,3**). This problem can be easily extended to continuous variables, using a threshold (**Equations 4-6**). In this case the optimization problem is:

$$\max_{m, th, d_1, d_2} A + B \quad (4)$$

$$A = \sum_{p \in split} y_{pd_1}(x_{pm} >= th) \quad (5)$$

$$B = \sum_{p \in split} y_{pd_2}(x_{pm} < th) \quad (6)$$

Both optimization problems start by setting all the patients within the studied split. The optimization splits the patients into two groups. For each of these groups, the algorithm is applied recursively until the number of patients in the split is smaller than a given number or until the optimization problem results in the same drug for both splits.

Equations (2,3,5,6) maximize the sum of the sensitivities of the patients of each of the branches. Using the same algorithm, it is possible to use any transformation of the sensitivity and include them in the optimization process. In this case, equations (5) and (6) are transformed into:

$$A = \sum_{p \in split} f(y_{pd_1})(x_{pm} >= th)$$

$$B = \sum_{p \in split} f(y_{pd_2})(x_{pm} < th)$$

Equations (2) and (3) can be transformed in an analogous way. To minimize the effect of outliers in the sum, we used the square root function to diminish the dynamical range of the data. The transformation is named ODT Sqrt in this work.

Multinomial logistic Lasso regression

The assignation of the proper drug to each patient problem can be tackled as a multiclass classification problem: the number of classes is the number of drugs and each patient is assigned the most effective drug for him/her. Using this approach, a multinomial regression can be applied to select the proper drug for each patient.

Predicting exclusively the most effective drug can be simplistic, since the penalty for misclassification is identical for the second most effective drug or for the least effective drug. Since the multinomial regression can also be applied to continuous variables, it is possible to give a "vote" for each patient that can be shared among all the drugs: the

Briefings in Bioinformatics

most effective drug will receive more shares of this vote that the least effective drug. Assigning the whole vote to the most effective drug can be seen as a particular case of this approach.

The Lasso regression is also implemented for multinomial regression. The implementation of glmnet (R Package) [49] is fast and convenient and allows for automatic selection of the regularization parameters using cross-validation.

More specifically, the multinomial regression builds the multinomial regression model (**Equation 8**)

$$\boldsymbol{X\beta} \sim \boldsymbol{Z} \tag{8}$$

where **X** is a P x M matrix where P is the number of patients and M is the number of biomarkers, $\boldsymbol{Z}$ is P x D voting matrix (in fact, probabilistic labels) where P is the number of patients and D is the number of tested drugs. All the elements of Z are positive and the sum of its elements by rows is equal to one. Finally, $\boldsymbol{\beta}$ the output of the regression is a M x D coefficient matrix. $\boldsymbol{X\beta}$ are the predicted logits for each drug being the most effective for each patient (**Figure 4**).

The specific selection of the entries for the Z matrix is shown in **Equation 9.**

$$z_{pd} = \frac{\exp\left(-K\frac{y_{pd}}{\min\limits_{p}(y_{pd})}\right)}{\sum_{i=1}^{D}\exp\left(-K\frac{y_{pi}}{\min\limits_{p}(y_{pi})}\right)} \tag{9}$$

Where y_{pd} are the entries of the **Y** matrix (that measures the sensitivity to a drug) and *K* is a predefined constant. If K>>1, all the exponentials of the summations of the denominator except the y_{pi} that corresponds to min(y_{pi}) vanish and the vote is given to the most effective drug. If K = 0, all the drugs share 1/D votes.

More information regarding the other algorithms already published can be found in the Supplementary Material (Section Supplementary Methods).

Data for Comparisons

We focused on Acute Myeloid Leukemia (AML) to compare the different methods described above. This disease was selected due to the availability of a wide cohort of patients with genomics data and ex-vivo drug sensitivity screening data. Ex-vivo data is more reliable than drug screenings performed on cell lines since the tests are performed directly on the AML patients' blood. Furthermore, AML is a highly heterogeneous disease with not standard PM therapeutic strategy, even though there is a growing field of drug development likely suited for these patients, e.g. Tyrosine Kinase Inhibitors (TKIs)[56].

Consequently, we selected the BeatAML cohort [50] for training the models and predicting different therapeutic strategies. This cohort is publicly available at http://www.vizome.org/ . We normalized the drug sensitivity IC_{50} from the ex-vivo experiments into IC50* described in [45]. This normalization equalizes the toxicity of the different drugs. To validate the predictions and due to the absence of more large-scale ex-vivo experiments, we used as an independent cohort testing set, the GDSC

Briefings in Bioinformatics

drug screening for AML cell lines [51], which could be found publicly available at https://www.cancerrxgene.org/ .

We compared the different algorithms based on four aspects that define interpretability (**Figure 5**): i) *the accuracy* of the method, for which we performed a 5-fold cross-validation in the training set, an independent cohort validation and an intragroup validation with the predicted groups in the training and validation set, ii) *the multi-omics capacity*, for which we tested the ability and performance of the methods when training with gene expression and genomic variants, iii) *the explainability*, for which we performed a qualitative comparison of all algorithms, analyzed the number of variables that each algorithm uses for prediction, and iv) *the implementability*, for which apart from qualitative comparisons based on the method definition, the computing time that each model requires for training becomes essential.

This four categories are explained more in depth in the following paragraphs.

Accuracy

The first "*sine qua non*" characteristic of a PM methods is the accuracy. An "interpretable" method with low accuracy becomes irrelevant. We define the accuracy as the difference of the IC50* for the assigned drug and the drug with maximum IC50* for that patient.

For assessing the accuracy of each of the methods, we performed the following comparisons: 5-fold cross-validation, independent cohort validation, and Intra-group validation.

5-fold cross-validation in BeatAML. We performed a 5-fold cross-validation using the BeatAML dataset. We trained all models with genetic variants data from 319 patients, dividing the cohort between the training samples 4-folds and testing samples the selected 1-fold. Each of the folds were tested, and the predicted IC50* for the 5-fold testing was compared for all the methods and compared against the Oracle -the drug with the optimum IC50*- (**Figure 6**). We calculated the Oracle as the minimum IC50* value for each patient.

Independent cohort validation. One of the main challenges of Machine Learning, including Precision Medicine, is generalization, i.e. the ability to adapt to new, previously unseen data. All the methods were tested on the GDSC AML dataset to check their generalization ability. The models were trained using the BeatAML dataset and were used to predict the optimal drug for AML cell lines from GDSC using its mutation files. Each of the cell lines was recommended a drug, we compared the all-samples IC_{50} for all the models and against the Oracle (the drug with the minimum IC_{50} for each cell line).

Intra group validation. We compared whether the IC50* of a drug in patients in whom it was recommended was lower than the IC50* in patients in whom it was not recommended. Using this information we compared the sensitivity to a drug for a specific group against the sensitivity to that drug for the rest of the samples by using a 2-tailed Wilcoxon test. This analysis was performed both for the BeatAML dataset (training dataset) and the GDSC AML cell lines cohort (predicted dataset).

Multi-omics suitability

Some of the methods only accept as input binary variables. Although, genomic variants can be transformed into binary variables, gene expression, methylation, or openness of the chromatin are intrinsically continuous variables. We have included a table showing

Briefings in Bioinformatics

whether the algorithm accepts only binary inputs (genomic variants only) or whether it also accepts continuous data (gene expression, methylation, etc.). For the methods that accept continuous variables, we assessed the performance of the predictions (5-fold cross-validation) in the BeatAML dataset using both data sources. We state the statistical significance using a 2-tail Wilcoxon's test comparing the IC50* using as input either genetic variants or gene expression.

Explainability

PM is more suited for healthcare if it can be interpreted. A machine learning method is interpretable if it provides the decision criteria that define the pathway that leads to the solution.

Explainability is defined by three different aspects: i) the explainability of the results, which checks if the method provides a ranking of the variables according to their importance for drug recommendation, ii) the capacity to output an easy-to-apply decision criteria, and iii) the understandability of the methods, this category mentions if the process of the algorithm to reach the classification criteria is easy to understand.

For assessing these characteristics, we performed a qualitative analysis based on the method description and execution. Furthermore, we analyzed the number of variables that each model requires to make the predictions. A model with a small number of variables is easier to understand, improves the understanding of the variable ranking, and is easier to for clinical diagnosis. Therefore, we paid special attention to the number of variables.

Implementability

Implementability is the easiness of a method being implemented into clinical research or practice. We measured the implementability of a method by analyzing four main features: i) the feasibility for wet-lab validations, ii) the consideration of the physician's experience, iii) the generation of a clinical guideline, and iv) technical implementation, which refers to the computational burden and software that the method requires. We used qualitative grades for the first characteristics. Regarding the technical implementation, we considered the computational burden. Despite it could be considered less important, some of the algorithms require hours of computing time for the BeatAML of the subset of AML samples in GDSC -that be considered to be small/medium size. Requiring fewer resources makes an algorithm more attractive to be applied for larger datasets. We also analyzed the software environment that each model requires to be run.

Briefings in Bioinformatics

Results

In this work, we compare several aspects of the performance of different interpretable models[57]. These models were classified into two main groups. The first one, named patient-based, are models that return a specific therapeutic strategy for each patient. The second one, named drug-based, are models that provide the patient(s) that are especially sensitive to a specific drug. Patient-based models include Multi-dimensional Module Optimization (MOM) [45], Optimal Decision Trees (ODT), Kernelized Rank Learning (KRL) [46], and Multinomial Lasso. Drug-based models are more suited for physicians and clinical investigation. This group comprises BOSO [47] and Lasso [49]. Patient-based methods rank the effectiveness of the drugs for a specific patient. Drug-based methods rank the effectiveness of a specific drug for each of the patients.

All the methods were developed to predict the drug response or develop a treatment strategy using genetic variants information. Thus, we trained the methods to predict drug efficacy using patients' samples and ex-vivo drug efficacy from the BeatAML [50] dataset. The methods were compared in terms of interpretability, which was defined according to four properties namely accuracy, adaptability, explainability, and easiness of implementation.

Accuracy: all the methods provided good estimates.

The first test to assess the accuracy was a 5-fold cross-validation in BeatAML [50]. Results for this analysis can be found in **Figure 7.a.** Multinomial Lasso achieves the lowest median, -the highest sensitivity- although it also entails the highest variability. Lasso's prediction is similar to the former one, but its standard deviation is smaller. Finally, MOM and BOSO achieve almost identican median. ODT –in both versions – has the highest IC50* prediction, i.e. the smallest value for sensitivity. However, the performance of the methods –excluding ODT and ODT Sqrt– was not statistically significant (p-value >0.05). ODT and ODT Sqrt predicitons were significantly worse than Multinomial (two-sided Wilcoxon test p-value=0.005921 and p-value=0.004942, respectively).

In the second test, we used the models trained on the full BeatAML, and tested them agains the Genomics of Drug Sensitivity in Cancer (GDSC) AML dataset. This dataset contains the genetic variants information for each cell line and the IC_{50} values for most of the drugs in the same cell lines. The independent cohort validation showed very different results from the 5-fold cross-validation (**Figure 7.b**), In this case, the ODT standard method achieved the best sensitivity score followed by MOM, Multinomial Lasso, Lasso, ODT square root and BOSO. The IQR for ODT and its standard deviation were much larger than for other methods. Nevertheless, there were no statistical significance in the difference of the predicted GDSC IC_{50} comparing any of the methods.

In the third test, we analyzed the intra-group classification performance. In this test we compared the IC50* of patients that were recommended a drug with the IC50* of the rest of patients using BeatAML and GDSC. The models with the best intragroup performance were MOM and ODT in their standard form followed by ODT Sqrt. MOM showed a significant sensitivity difference in 3 out of 4 groups for the BeatAML dataset (**Figure 7.c**) and 3 out of 4 for the GDSC dataset (**Figure 7.d**). ODT standard achieved a significant intragroup sensitivity in 4 out of 6 groups for BeatAML (**Figure 7.e**) and 2 out of 5 for GDSC (**Figure 7.f**). Finally, ODT Sqrt significantly recommended the usage of 3 drugs out of 5 for BeatAML (**Figure 7.g**), and 1 out of 4 in GDSC (**Figure 7.h**). No statistical significance was found for the rest of the methods. This is most likely due to

Briefings in Bioinformatics

the fact that there are more than ten different treatments proposed, and the number of patients is insufficient to achieve statistical significance. (**Supplementary Figures 4-11**).

Accuracy: using gene expression as input provides similar accuracy if compared to genetic variants.

We tested whether using gene expression could improve the method accuracy [58]. We trained all models (except MOM and KRL, since they do not accept continuous inputs) using BeatAML gene expression (GE) data. We performed a 5-fold cross-validation in BeatAML dataset for the models predicting GE and genetic variants data. The results in **Figure 8** show that the predictions do not significantly change when varying the type of input, except in the Multinomial Lasso, where the use of gene expression significantly increased the precision of the method and in the Lasso, where it significantly decreased the sensitivity of the method. This analysis was also performed training in BeatAML and predicting in GDSC with the mutational and GE models. For which, **Supplementary Figure 16** showed no statistical significant difference in model sensitivity for any of the methods.

As shown in the preceding paragraphs, the differences are minor, and no method outperforms the others in all cases.

Explainability: tree-like methods (MOM and ODT) require much less variables than any other methods.

To measure the explainability of the method, we trained the models with the BeatAML dataset and checked the number of variables that each model required to make the predictions. Results are included in **Figure 9.a**. Remarkably, MOM and ODT use less than 5 variables, almost ten times less than the rest of the methods. BOSO, Multinomial Lasso, and Lasso use more than 30 variables. Among them, BOSO (with 33 variables) is the method that requires less variables. BOSO builds a linear model for each of the drugs. Each of the results (as occurs in Lasso) are sparse: it requires only 5 variables to predict drug response for some drugs. Since these variables are not identical for every drug, in the end, it requires 33 variables to make the predictions (**Supplementary Figure 14**). Multinomial Lasso and Lasso were coded to preserve the same variables for predicting over all the drugs (**Supplementary Figures 12-13**). BOSO did not implement this option.

Regarding the KRL method the number of variables it does not provide automatic feature selection but use regularization methods. Thus, all the 69 gene variants are used.

ODT and MOM output the decision criteria it the form of a decision-tree. The main difference between ODT and MOM decision trees is their structure, ODT's tree structure have several branches where drugs for each of them. MOM's tree structure is linear, it is divided into different sequential steps, each of them defined by a biomarker, and there is a drug recommendation on every step. Regression-based methods (BOSO, Lasso and Multinomial Lasso) provide the weights for each of the biomarkers on each of the drugs. Therefore, it is possible to check which are the critical biomarkers for each drug. KRL use kernels to guess the proper treatment. In this case, it is much more complex to understand which are the key genomic variants of the recommendation system.

Briefings in Bioinformatics

Implementability: Optimal Decision Trees and MOM are the most prone to clinical practice and ODT the least computing time consuming

We also considered the easeness to implement the methods in wet lab or even clinical practice according to four different points: i) the feasibility for wet-lab validations, ii) the consideration of the physician's experience, iii) the generation of a clinical guideline, and iv) the computational implementation.

Tree-based models require less biomarkers than regression models or KRL. Furthermore, only a few operations are required to perform the predictions, which can be done by hand. On the contrary, regression models and KRL require more genes and a computer-based environment to perform the drug assignation.

Regarding the computational burden of each of the methods, all the methods need to be trained in different software environments such as R or Python. Once trainced, the tree-based models directly provide a guideline that do not require the environment anymore. We have timed the training process of the 6 models (**Figure 9.b**) using Mutational data and Gene Expression (where possible). ODT is the fastest method to train (0.05 seconds for training using mutational data and less than 5 seconds using gene expression data). Multinomial requires around 15 second using either mutational data or gene expression data. Lasso lasts 10 and 100 seconds using mutational and gene expression data respectively. Finally, MOM, KRL and BOSO require several hours for training their models. MOM and KRL are not suitable for gene expression data so they have been excluded for the timing analysis with this data. Prediction time is similar (and negligible if compared with training time) in all the 6 methods. Focusing on the installation, models based on MILP (BOSO, KRL, and MOM) require a complex installation of software (**Table 1**). They are also the most time-consuming methods. ODT, Multinomial, and Lasso, only require of R installation to run.

All these conclusions that could lead to rank methods according to Interpretablility have been summed up in **Error! Reference source not found.2**.

Briefings in Bioinformatics

Discussion

In this work we have selected four precision medicine methods –MOM, BOSO, Lasso, and KRL– and developed two additional ones –Optimal Decision Trees and Multinomial Lasso–, to compare them regarding their interpretability. We performed six quantitative comparisons and four qualitative comparisons. All the methods were similar in terms of accuracy. However, MOM and Optimal Decision Trees were the most interpretable and easy to implement.

PM is a topic that is being widely addressed and there are new algorithm proposals. It may seem surprising that we included only four of them in this comparison and, indeed, we developed two additional ones. A systematic review of all the methods --cited in the Introduction section-- included Machine Learning (ML) methods (using either deep learning, neural networks, support vector machines, random forests, etc). Among the 24 methods that used ML for making their predictions, only 10 were explainable. Of those 10, 5 of them did not solve the "patient-centered" problem: assign the proper drug to each patient. We ended up with MOM, BOSO, KRL, and LOBICO, and added Lasso as a control of a traditional approach in the ML field. LOBICO approach, that was also tested on this dataset elsewhere [48] is drug-centered and, since the output variable is discrete, it cannot be transformed into the patient-centered problem and not suitable for this comparison [48]. We developed two additional methods, both patient-centered, with two different approaches: regression (Multinomial) and tree classification (ODT).

In this work, we have defined Interpretability splitting it into four main concepts: Accuracy, Multi-omics capacity, Explainability, and Implementability. An interpretable PM method should be accurate and understandable by the common researcher or clinician. Accuracy is strictly necessary: if a method is not accurate, it becomes irelevant despite being easy to understand. Multi-omics capacity, measures the robustness of the method to adapt into different data sources, that could be essential for new lines of research. Explainability is also essential, it should show the reasoning for reaching the results. Finally, the ease of implementation defines the ability of the method to incorporate the clinician experience and provide an easy technical usage.

We focused on a specific sensitivity value named IC50*. This metric was previously described in [45] or in [46] and is a normalization of the logarithm of the IC_{50}. Normalizing the IC_{50} --or other sensitivity value-- is crucial as the best drug is not necessary the drug with the lowest IC_{50} value. In fact, a drug with a low IC_{50} can be toxic for the patients. Toxic drugs tend to have low IC_{50} values in all tissues, whereas the focus must be set on drugs with differential sensitivity for different tissues. Normalizing the logarithm of the IC_{50} by removing the mean sensitivity value of the drug in all patients, preserves the sensitivity profiles of the drugs and penalizes drugs that are sensitive or resistant in all tissues. The dosage of drugs with higher IC_{50} can be adjusted to obtain drug effectiveness. We trained all the models with the normalized version, IC50*, to avoid the aforementioned problems.

All the methods predict reasonably well in terms of accuracy. The 5-fold cross-validation and the independent cohort validation, showed that the different methods had similar median and the differences were not statistically significant. The intragroup validation, showed that the regression-based models (KRL, BOSO, Lasso, and Multinomial) were not able to distinguish between responders and non-responders to a specific drug. This result is reasonable since these methods do not divide the patients according to responders or not responders to one biomarker, but cherry-picked patients

Briefings in Bioinformatics

based on weighted combinations of biomarkers. MOM, on the contrary, has a restriction within its model formulation that means that all patients with a biomarker that confers sensitivity to a specific drug should be treated in that current step [45]. Nevertheless, not having a successful intragroup validation does not invalidate the model.

The multi-omics suitability is a "hot-topic" in PM, as there is not a current gold standard based on which type of data is more accurate when predicting drug response. Some models use genetic variants to promote interpretability, whereas other use gene expression or integrated omics for improving accuracy. In this work, we compared the accuracy changes when training and predicting on gene expression and genetic variants separately, and found almost no significant statistical difference in the performance. Drug response is mediated in living beings by complex regulatory and metabolomic processes that are most likely to be solved using an integrated omics input, instead of just one single omics. However, the more complex the model becomes, the less interpretable it is.

Regarding the explainability, we included also a qualitative comparison since focusing only on the number of variables, does not justify that the method is understandable. It is also desirable that the method can provide decision criteria, i.e. a complete process that a clinician can follow and understand. This consideration has paramount importance if it is to be approved by regulators for medicine [59,60]. Consequently, we focused on the ease to understand the output of the methods, and the explainability of the results. We defined the latter, as the ability of the method to rank the input variables in order of importance for drug assignation. Of course, a smaller number of variables is easier to understand. The tree-based models require less than six varables, and it increases up to five times in the regression-based models. BOSO, however, uses only five variables to predict response of just one drug, but when translated into a patient-centered approach, the total number of variables used for predicting in all drugs is equal to 33. For Lasso and Multinomial, the number of variables has been optimized to predict response in all drugs. KRL, however, did not consider this parameter and uses all variables provided as input to make the predictions, being the less explainable method.

Implementability is a concept easier to understand, as it directly facilitates the clinical translation. Most of the implementability comparisons were qualitative, but we performed a technical comparison of the methods regarding its computational burden. There we showed that MOM, which was leading the accuracy comparisons, is the most time consuming up to 2.5 hours on a normal machine, and it is the model that requires the highest number of software environments: R, Python, and CPLEX need to be installed in the machine (and relatd to each other). It is the most resource consuming. However, if compared against ODT, which achieved similar accuracy performance, the latter only requires R and the algorithm is trained, even using gene expression, in less than 5 seconds. Besides, ODT is more explainable than MOM, because the method is easier to understand, although it is quite similar to MOM regarding the other explainability and implementability criteria.

Nonetheless, Multinomial and Lasso are also explainable, if not compared against other methods, and there are additional functions -not defined in the methods themselves- that can be applied to extract the algorithm reasoning or decision criteria. Also, linear models can be understandable as the βs reflect the variable importance for prediction.

Briefings in Bioinformatics

To summarize, in this work we defined a quantitative method for evaluating the interpretability of a given machine learning method, because, as previously discussed, accuracy is not the only important factor in the complex field of health. The defined criteria can serve as a guide for developing new translational methods aimed at solving precision medicine problems.

Key Points

- For a machine learning method to be interpretable, it needs to be accurate, suitable to different multi-omics data, explainable and implementable.
- Traditional Machine Learning does not solve the complete assignation problem, thus, there are many creative methodologies to tackle the drug assignation.
- There are several methods amenable to be interpretable, and can be classified into two main groups: "patient-centered" or "drug-centered". "Drug-centered" methods can be transformed into "patient-centered" methods.
- In terms of explainability and implementability, it is highly important for the method to provide the decision criteria.
- From the methods compared: MOM, ODT, Multinomial, BOSO, KRL, and Lasso. They all achieved similar results in terms of accuracy, ODT and MOM are the most explainable, and ODT the most implementable.

Funding

This work was supported by the Minister of Economy and Competitiveness of Spain [PID2019-110344RB-I00, project Synlethal], PIBA Programme of the Basque Government [PIBA_2020_01_0055, project DeepLethal], Elkartek programme of the Basque Government [KK-2020/00008], Cancer Research UK and AECC under the Accelerator Award Programme [C355/A26819, project Editor]

Acknowledgments

Authors would like to thank Fernando Carazo for the support given when developing the study.

Conflicts-of-interest

The authors declare there are not conficts of interest.

Data Availability Statement

We selected the BeatAML cohort for training the models and predicting different therapeutic strategies. This cohort is publicly available at http://www.vizome.org/ . To validate the predictions we used the GDSC drug screening for AML cell lines, which could be found publicly available at https://www.cancerrxgene.org/ .

Author Contributions

MG and AR conceived this study, and carried out the computational implementation. AR developed ODT and Multinomial. MG, KS, and AR read and wrote the manuscript and approved the submitted version.

Briefings in Bioinformatics

References

1. Ashley EA. Towards precision medicine. Nat. Rev. Genet. 2016; 17:507–522

2. He M, Xia J, Shehab M, et al. The development of precision medicine in clinical practice. Clin. Transl. Med. 2015 41 2015; 4:1–4

3. Collins FS, Varmus H. A New Initiative on Precision Medicine. N. Engl. J. Med. 2015; 372:793–795

4. Gerstung M, Papaemmanuil E, Martincorena I, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat. Genet. 2017; 49:332–340

5. Xu J, Yang P, Xue S, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum. Genet. 2019 1382 2019; 138:109–124

6. Li J, Zheng S, Chen B, et al. A survey of current trends in computational drug repositioning. Brief. Bioinform. 2016; 17:2–12

7. Bhinder B, Gilvary C, Madhukar NS, et al. Artificial Intelligence in Cancer Research and Precision Medicine. Cancer Discov. 2021; 11:900–915

8. Granat LM, Kambhampati O, Klosek S, et al. The promises and challenges of patient-derived tumor organoids in drug development and precision oncology. Anim. Model. Exp. Med. 2019; 2:150–161

9. Roife D, Dai B, Kang Y, et al. Ex vivo testing of patient-derived xenografts mirrors the clinical outcome of patients with pancreatic ductal adenocarcinoma. Clin. Cancer Res. 2016; 22:6021–6030

10. Shamout F, Zhu T, Clifton DA. Machine Learning for Clinical Outcome Prediction. IEEE Rev. Biomed. Eng. 2021; 14:116–126

11. Scott IA, Cook D, Coiera EW, et al. Machine learning in clinical practice: prospects and pitfalls. Med. J. Aust. 2019; 211:203

12. Adlung L, Cohen Y, Mor U, et al. Machine learning in clinical decision making. Med 2021; 2:642–665

13. Oh SH, Lee SJ, Park J. Precision Medicine for Hypertension Patients with Type 2 Diabetes via Reinforcement Learning. J. Pers. Med. 2022, Vol. 12, Page 87 2022; 12:87

14. Eckardt J-N, Wendt K, Bornhäuser M, et al. Reinforcement Learning for Precision Oncology. Cancers 2021, Vol. 13, Page 4624 2021; 13:4624

15. Liu Q, Hu Z, Jiang R, et al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 2020; 36:i911–i918

16. Lee BKB, Tiong KH, Chang JK, et al. DeSigN: Connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics 2017; 18:934

17. Huang C, Clayton EA, Matyunina L V., et al. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. Sci. Rep. 2018; 8:

18. Guo W, Ji Y, Catenacci DVT. A subgroup cluster-based Bayesian adaptive design for precision medicine. Biometrics 2017; 73:367–377

19. Kim Y, Kim D, Cao B, et al. PDXGEM: Patient-derived tumor xenograft-based gene expression model for predicting clinical response to anticancer therapy in cancer patients. BMC Bioinformatics 2020; 21:288

Briefings in Bioinformatics

20. Preuer K, Lewis RPI, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics 2018; 34:1538–1546

21. Robert J, Vekris A, Pourquier P, et al. Predicting drug response based on gene expression. Crit. Rev. Oncol. Hematol. 2004; 51:205–227

22. Seo H, Tkachuk D, Ho C, et al. SYNERGxDB: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. Nucleic Acids Res. 2020; 48:W494–W501

23. Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLoS One 2019; 14:e0219774

24. Boichard A, Richard SB, Kurzrock R. The Crossroads of Precision Medicine and Therapeutic Decision-Making: Use of an Analytical Computational Platform to Predict Response to Cancer Treatments. Cancers (Basel). 2020; 12:166

25. Siah KW, Khozin S, Wong CH, et al. Machine-Learning and Stochastic Tumor Growth Models for Predicting Outcomes in Patients With Advanced Non–Small-Cell Lung Cancer. JCO Clin. Cancer Informatics 2019; 1–11

26. Chang Y, Park H, Yang HJ, et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci. Rep. 2018; 8:

27. Joo M, Park A, Kim K, et al. A Deep Learning Model for Cell Growth Inhibition IC50 Prediction and Its Application for Gastric Cancer Patients. Int. J. Mol. Sci. 2019; 20:6276

28. Matchett K, Lynam-Lennon N, Watson R, et al. Advances in Precision Medicine: Tailoring Individualized Therapies. Cancers (Basel). 2017; 9:146

29. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. npj Precis. Oncol. 2019; 3:1–5

30. Biankin A V. The road to precision oncology. Nat. Genet. 2017; 49:320–321

31. Chua IS, Gaziel-Yablowitz M, Korach ZT, et al. Artificial intelligence in oncology: Path to implementation. Cancer Med. 2021; 10:4138–4149

32. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. 2020; 2:573–584

33. Kuenzi BM, Park J, Fong SH, et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell 2020; 38:672-684.e6

34. Khakabimamaghani S, Kelkar YD, Grande BM, et al. SUBSTRA: Supervised Bayesian Patient Stratification. Bioinformatics 2019; 35:3263–3272

35. Kim Y, Bismeijer T, Zwart W, et al. Genomic data integration by WON-PARAFAC identifies interpretable factors for predicting drug-sensitivity in vivo. Nat. Commun. 2019; 10:1–12

36. Vougas K, Sakellaropoulos T, Kotsinas A, et al. Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining. Pharmacol. Ther. 2019; 203:107395

37. Astras G, Papagiannopoulos CI, Kyritsis KA, et al. Pharmacogenomic Testing to Guide Personalized Cancer Medicine Decisions in Private Oncology Practice: A Case

Briefings in Bioinformatics

Study. Front. Oncol. 2020; 10:521

38. Pai S, Hui S, Isserlin R, et al. netDx: interpretable patient classification using integrated patient similarity networks. Mol. Syst. Biol. 2019; 15:e8497

39. Gerstung M, Papaemmanuil E, Martincorena I, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat. Genet. 2017; 49:332–340

40. Liu H, Zhao R, Fang H, et al. Entropy-based consensus clustering for patient stratification. Bioinformatics 2017; 33:2691–2698

41. Stetson LC, Pearl T, Chen Y, et al. Computational identification of multi-omic correlates of anticancer therapeutic response. BMC Genomics 2014; 15:S2

42. Ingelman-Sundberg M, Mkrtchian S, Zhou Y, et al. Integrating rare genetic variants into pharmacogenetic drug response predictions. Hum. Genomics 2018; 12:26

43. Oberthuer A, Juraeva D, Hero B, et al. Revised risk estimation and treatment stratification of low- and intermediate-risk neuroblastoma patients by integrating clinical and molecular prognostic markers. Clin. Cancer Res. 2015; 21:1904–1915

44. Cheng L, Majumdar A, Stover D, et al. Computational Cancer Cell Models to Guide Precision Breast Cancer Medicine. Genes (Basel). 2020; 11:263

45. Gimeno M, San José-Enériz E, Villar Fernandez S, et al. Explainable Artificial Intelligence for Precision Medicine in Acute Myeloid Leukemia. Front. Immunol. 2022; 0:5805

46. He X, Folkman L, Borgwardt K. Kernelized rank learning for personalized drug recommendation. Bioinformatics 2018; 34:2808–2816

47. Valcárcel L V., San José-Enériz E, Cendoya X, et al. BOSO: A novel feature selection algorithm for linear regression with high-dimensional data. PLOS Comput. Biol. 2022; 18:e1010180

48. Knijnenburg TA, Klau GW, Iorio F, et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. Sci. Reports 2016 61 2016; 6:1–14

49. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 2010; 33:1–22

50. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018; 562:526–531

51. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2012; 41:D955–D961

52. Zeisig BB, Kulasekararaj AG, Mufti GJ, et al. SnapShot: Acute Myeloid Leukemia. Cancer Cell 2012; 22:698-698.e1

53. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood 2017; 129:424–447

54. NIH NCIGDC. Acute Myeloid Leukemia — Cancer Stat Facts.

55. He X, Folkman L, Borgwardt K. Kernelized rank learning for personalized drug recommendation. Bioinformatics 2018; 34:2808–2816

56. Döhner H, Estey EH, Amadori S, et al. Diagnosis and management of acute

Briefings in Bioinformatics

myeloid leukemia in adults: Recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood 2010; 115:453–474

57. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. Proc. 2018 ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics 2018; 559–560

58. Lucena-Araujo AR, Coelho-Silva JL, Pereira-Martins DA, et al. Combining gene mutation with gene expression analysis improves outcome prediction in acute promyelocytic leukemia. Blood 2019; 134:951–959

59. . Artificial intelligence in medicine regulation | European Medicines Agency.

60. U.S. Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)-Discussion Paper and Request for Feedback. FDA 2019; 20

Table 1. Precision Medicine Pipelines selected for comparison. This table collects the description of each of the methods. Algorithm shows the method's given name. Type refers to whether the method is patient- or drug-centered. The software column collects all the required software environment programs for the model to be run. Method refers to the pipeline description. Suitable for mutational data has a "Yes" if the method could use genetic variants as input. Suitable for gene expression has a "Yes" if the method could use gene expression data as input and a "No" otherwise. Output explains the raw output of the model. Reference contains the publications in which the method was defined.

Algorithm	*Type*	*Software*	*Method*	*Suitable for mutational data*	*Suitable for gene expression*	*Output*	*Reference*
MOM	Patient	Python 3.7, R 4.2, and CPLEX	Feature Selection and MILP	Yes	No	Drug assignation	[45]
ODT	Patient	R 4.2	Recursive Decision Tree	Yes	Yes	Drug assignation	Novel
Multinomial	Patient	R 4.2	Multinomial Lasso	Yes	Yes	Drug assignation	Novel
KRL	Patient	Python 2.7	Kernelized MILP	Yes	No	Drug assignation	[55]
BOSO	Drug	R 4.2 and CPLEX	Lasso regression using MILP	Yes	Yes	Predicted IC50 for a drug	[47]
Lasso	Drug	R 4.2	Standard Lasso regression	Yes	Yes	Predicted IC50 for a drug	[49]

Table 2: Table containing the interpretability comparisons for each method. The values No(∅), Yes(*), and Very(***), reflect whether the method does not, fulfills or greatly fulfills -respectively-, the conditions mentioned in the columns. Gene Expression refers to whether the method is suitable for this type of omics data. Genetic Variants refer to whether the method is suitable for this type of omics data. Small number of variables, refers to the number of variables that the method uses for classification, Understandable Method, is defined as the logical explanation of the method, i.e., it does not rely on random selection; Outputs Decision Criteria, refers to whether the method provides the decision criteria used in the assignation, Easy to validate, refers to the ability of the method to be wet-lab validated. Considering Experience shows whether the method could enhance the physician or researcher experience. Clinical Guideline shows methods that output a therapeutic strategy for ulterior patients. Computational Burden summarizes the computational resources consumed by the method, where No(∅) refers to most resources consuming, Yes(*) a reasonably resources consuming and Very(***) refers to the best methods.

	Multi-omics		*Explainability*			*Implementability*			
Method	Gene Expression	Genetic Variants	Small number of Variables	Understandable Method	Outputs Decision Criteria	Easy to Validate	Considering Experience	Clinical Guideline	Computational Burden
MOM	∅	*	***	*	*	***	***	***	∅
ODT	*	*	***	*	*	***	***	***	***
Multinomial	*	*	∅	*	∅	*	∅	∅	**
Lasso	*	*	∅	*	∅	*	∅	∅	**
BOSO	*	*	∅	*	∅	*	∅	∅	∅
KRL	∅	*	∅	*	∅	*	∅	∅	∅

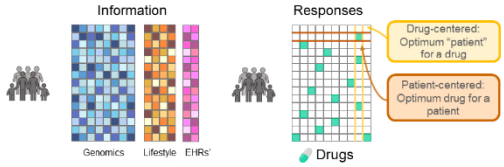


Figure 1: Precision Medicine paradigm. The left-side panel represents patients' data the and right-side panel shows the data available for patients' responses to treatment.

673x227mm (38 x 38 DPI)

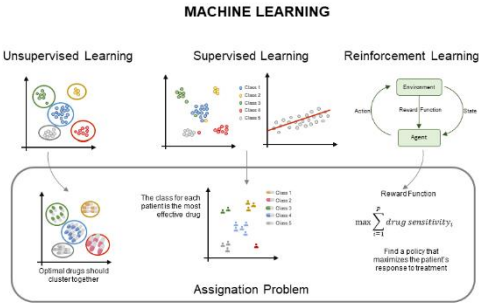


Figure 2: Relationship between Machine Learning and the Assignation Problem. The assignation problem is not a specific machine learning problem but could be addressed from all the machine learning branches.

259x145mm (96 x 96 DPI)

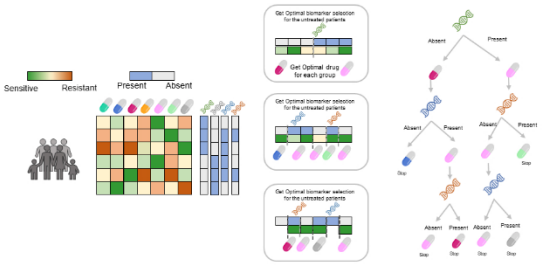


Figure 3. ODT Model Performance. The ODT model uses as input the sensitivity matrix and the biomarker matrix, on each step it splits the patients into two groups according to the presence or absence of a biomarker. This split is optimized so that the drug-assigned is the most sensitive to each of the splits. It recursively splits the different branches until a predefined group size is reached.

260x121mm (96 x 96 DPI)

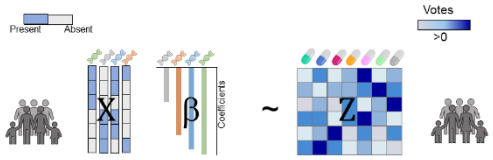


Figure 4. Multinomial Model. The multinomial Model corresponds to a modified Multinomial logistic lasso regression, where the output represents the votes that each patient assigns to each of the drugs.

266x85mm (96 x 96 DPI)

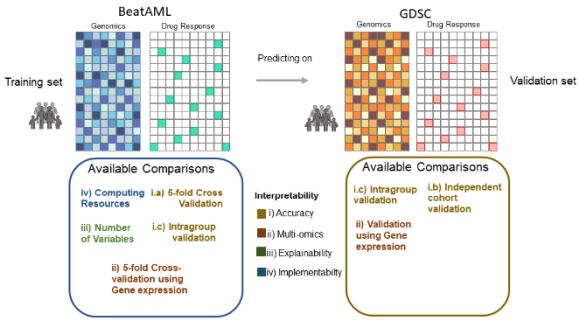


Figure 5. Summary of the available comparisons performed in this study. We trained the different models in BeatAML cohort and tested the predictions predicting over GDSC. From the training step we were able to obtain the training computing time, the number of variables required to make the predictions, a 5-fold cross-validation using mutational and gene expression data, and intragroup validation. Whereas for the testing step we performed and independent cohort prediction validation using mutational and gene expression data, and another intragroup validation

263x146mm (96 x 96 DPI)

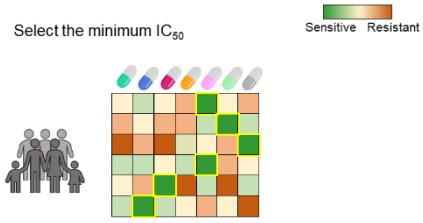


Figure 6. Oracle prediction. The Oracle predicts the most sensitive drug for each patient or cell line.

263x106mm (96 x 96 DPI)

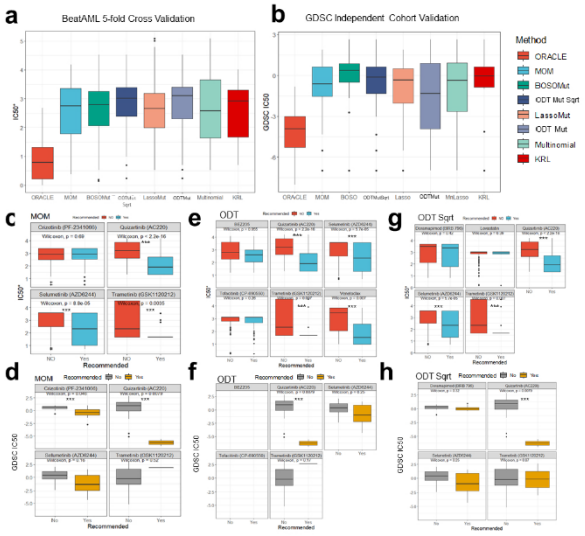


Figure 7. Accuracy comparison. A) Accuracy in 5-fold cross validation from BeatAML cohort. The different boxplots show the predicted IC50* of the drugs assigned to each of the patients. The lower the IC50* is, the more sensitive the method is, ORACLE is the control that shows the best possible drug to every patient in the cohort. B) Accuracy in independent cohort validation. The different boxplots show the predicted IC50 of the drugs assigned to each of the patients in GDSC. The lower the IC50 is, the more sensitive the method is, ORACLE is the control that shows the best possible drug to every patient in the cohort. Models were trained in BeatAML and predicted over GDSC. C-E-G)Intragroup validation of MOM, ODT and ODT Sqrt in BeatAML. Each of the subplots represents the efficacy of one drug, in blue the patients that were recommended that drug and in red the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05). D-F-H) Intragroup validation of MOM, ODT, and ODT Sqrt in GDSC. Each of the subplots represents the efficacy of one drug, in yellow the patients that were recommended that drug and in grey the patients that did not have that drug. Stars show the significance of the two-tailed Wilcoxon test (*** means p-value <0.05).

266x256mm (96 x 96 DPI)

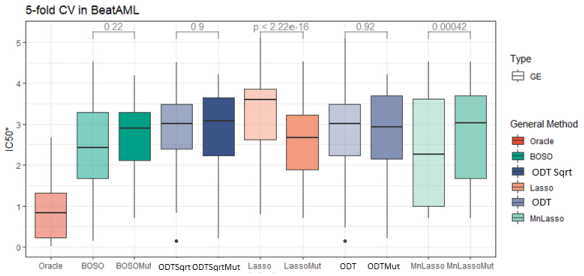


Figure 8. Using GE data over Mutational Data does not improve Method Precision. We compared for each of the algorithms the accuracy in response using Mutational and Gene Expression data as input. Distributions plotted in a lighter color are the responses obtained by each of the methods when using Gene Expression, whereas distributions plotted in darker non-transparent colors are the responses of the methods obtained when using mutational data.

206x96mm (96 x 96 DPI)

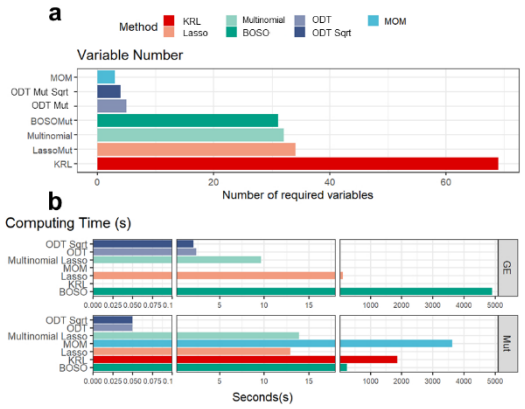


Figure 9. Variable Number and Computer timing performance comparisons. A) Variable number comparisons. All methods were trained in BeatAML cohort, after the training process we extracted the number of non-zero weighted input variables that each model requires for making the predictions. The horizontal axis shows the number of variables required by each method. B) Computer timing comparison. We measured the training time that each model requires using genetic variants (lower plot) or gene expression (upper plot) as input, time is shown in seconds in the horizontal axis.

266x192mm (96 x 96 DPI)

Bibliography

1. Gimeno M, San José-Enériz E, Rubio A, et al. Identifying Lethal Dependencies with HUGE Predictive Power. Cancers (Basel). 2022; 14:3251

2. Gimeno M, San José-Enériz E, Villar Fernandez S, et al. Explainable Artificial Intelligence for Precision Medicine in Acute Myeloid Leukemia. Front. Immunol. 2022; 0:5805

3. Ashley EA. Towards precision medicine. Nat. Rev. Genet. 2016; 17:507–522

4. He M, Xia J, Shehab M, et al. The development of precision medicine in clinical practice. Clin. Transl. Med. 2015 41 2015; 4:1–4

5. Collins FS, Varmus H. A New Initiative on Precision Medicine. N. Engl. J. Med. 2015; 372:793–795

6. Gerstung M, Papaemmanuil E, Martincorena I, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat. Genet. 2017; 49:332–340

7. Xu J, Yang P, Xue S, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum. Genet. 2019 1382 2019; 138:109–124

8. Li J, Zheng S, Chen B, et al. A survey of current trends in computational drug repositioning. Brief. Bioinform. 2016; 17:2–12

9. Bhinder B, Gilvary C, Madhukar NS, et al. Artificial Intelligence in Cancer Research and Precision Medicine. Cancer Discov. 2021; 11:900–915

10. Granat LM, Kambhampati O, Klosek S, et al. The promises and challenges of patient-derived tumor organoids in drug development and precision oncology. Anim. Model. Exp. Med. 2019; 2:150–161

11. Snijder B, Vladimer GI, Krall N, et al. Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. Lancet Haematol. 2017; 4:e595–e606

12. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018; 562:526–531

13. Roife D, Dai B, Kang Y, et al. Ex vivo testing of patient-derived xenografts mirrors the clinical outcome of patients with pancreatic ductal adenocarcinoma. Clin. Cancer Res. 2016; 22:6021–6030

14. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nat. 2012 4837391 2012; 483:570–575

15. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a Cancer Dependency Map. Cell 2017; 170:564-576.e16

16. Shao DD, Tsherniak A, Gopal S, et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res. 2013; 23:665–678

17. Behan FM, Iorio F, Picco G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature 2019; 568:511–516

18. Macarron R, Banks MN, Bojanic D, et al. Impact of high-throughput screening in biomedical research. Nat. Rev. Drug Discov. 2011; 10:188–195

19. Shamout F, Zhu T, Clifton DA. Machine Learning for Clinical Outcome Prediction. IEEE Rev. Biomed. Eng. 2021; 14:116–126

20. Scott IA, Cook D, Coiera EW, et al. Machine learning in clinical practice: prospects and pitfalls. Med. J. Aust. 2019; 211:203

21. Adlung L, Cohen Y, Mor U, et al. Machine learning in clinical decision making. Med 2021; 2:642–665

22. Oh SH, Lee SJ, Park J. Precision Medicine for Hypertension Patients with Type 2 Diabetes via Reinforcement Learning. J. Pers. Med. 2022, Vol. 12, Page 87 2022; 12:87

23. Eckardt J-N, Wendt K, Bornhäuser M, et al. Reinforcement Learning for Precision Oncology. Cancers 2021, Vol. 13, Page 4624 2021; 13:4624

24. McVeigh TP, Hughes LM, Miller N, et al. The impact of Oncotype DX testing on breast cancer management and chemotherapy prescribing patterns in a tertiary referral centre. Eur. J. Cancer 2014; 50:2763–2770

25. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert Rev. Mol. Diagn. 2009; 9:417–422

26. Wu L, Yang Y, Guo X, et al. An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk. Nat. Commun. 2020; 11:1–11

27. Kuenzi BM, Park J, Fong SH, et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell 2020; 38:672-684.e6

28. Malani D, Kumar A, Brück O, et al. Implementing a Functional Precision Medicine Tumor Board for Acute Myeloid Leukemia. Cancer Discov. 2022; 12:388–401

29. Liu Q, Hu Z, Jiang R, et al. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 2020; 36:i911–i918

30. Lee BKB, Tiong KH, Chang JK, et al. DeSigN: Connecting gene expression with therapeutics for drug repurposing and development. BMC Genomics 2017; 18:934

31. Preuer K, Lewis RPI, Hochreiter S, et al. DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics 2018; 34:1538–1546

32. Robert J, Vekris A, Pourquier P, et al. Predicting drug response based on gene expression.

Crit. Rev. Oncol. Hematol. 2004; 51:205–227

33. Seo H, Tkachuk D, Ho C, et al. SYNERGxDB: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. Nucleic Acids Res. 2020; 48:W494–W501

34. Lind AP, Anderson PC. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. PLoS One 2019; 14:e0219774

35. Boichard A, Richard SB, Kurzrock R. The Crossroads of Precision Medicine and Therapeutic Decision-Making: Use of an Analytical Computational Platform to Predict Response to Cancer Treatments. Cancers (Basel). 2020; 12:166

36. Siah KW, Khozin S, Wong CH, et al. Machine-Learning and Stochastic Tumor Growth Models for Predicting Outcomes in Patients With Advanced Non–Small-Cell Lung Cancer. JCO Clin. Cancer Informatics 2019; 1–11

37. Chang Y, Park H, Yang HJ, et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci. Rep. 2018; 8:

38. Joo M, Park A, Kim K, et al. A Deep Learning Model for Cell Growth Inhibition IC50 Prediction and Its Application for Gastric Cancer Patients. Int. J. Mol. Sci. 2019; 20:6276

39. Huang C, Clayton EA, Matyunina L V., et al. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. Sci. Rep. 2018; 8:

40. Guo W, Ji Y, Catenacci DVT. A subgroup cluster-based Bayesian adaptive design for precision medicine. Biometrics 2017; 73:367–377

41. Kim Y, Kim D, Cao B, et al. PDXGEM: Patient-derived tumor xenograft-based gene expression model for predicting clinical response to anticancer therapy in cancer patients. BMC Bioinformatics 2020; 21:288

42. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 2019; 1:206–215

43. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. 2020; 2:573–584

44. Adam G, Rampášek L, Safikhani Z, et al. Machine learning approaches to drug response prediction: challenges and recent progress. npj Precis. Oncol. 2020; 4:19

45. Matchett K, Lynam-Lennon N, Watson R, et al. Advances in Precision Medicine: Tailoring Individualized Therapies. Cancers (Basel). 2017; 9:146

46. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. npj Precis. Oncol. 2019; 3:1–5

47. Biankin A V. The road to precision oncology. Nat. Genet. 2017; 49:320–321

48. Chua IS, Gaziel-Yablowitz M, Korach ZT, et al. Artificial intelligence in oncology: Path to implementation. Cancer Med. 2021; 10:4138–4149

49. U.S. Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)-Discussion Paper and Request for Feedback. FDA 2019; 20

50. European Medicines Agency. Artificial intelligence in medicine regulation | European Medicines Agency.

51. Kuenzi BM, Park J, Fong SH, et al. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell 2020; 38:672-684.e6

52. Khakabimamaghani S, Kelkar YD, Grande BM, et al. SUBSTRA: Supervised Bayesian Patient Stratification. Bioinformatics 2019; 35:3263–3272

53. Kim Y, Bismeijer T, Zwart W, et al. Genomic data integration by WON-PARAFAC identifies interpretable factors for predicting drug-sensitivity in vivo. Nat. Commun. 2019; 10:1–12

54. Vougas K, Sakellaropoulos T, Kotsinas A, et al. Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining. Pharmacol. Ther. 2019; 203:107395

55. Astras G, Papagiannopoulos CI, Kyritsis KA, et al. Pharmacogenomic Testing to Guide Personalized Cancer Medicine Decisions in Private Oncology Practice: A Case Study. Front. Oncol. 2020; 10:521

56. Pai S, Hui S, Isserlin R, et al. netDx: interpretable patient classification using integrated patient similarity networks. Mol. Syst. Biol. 2019; 15:e8497

57. Lazar AJ, Demicco EG. Human and machine: Better at pathology together? Cancer Cell 2022; 40:806–808

58. Perry AM, Attar EC. New Insights in AML Biology From Genomic Analysis. Semin. Hematol. 2014; 51:282–297

59. Zeisig BB, Kulasekararaj AG, Mufti GJ, et al. SnapShot: Acute Myeloid Leukemia. Cancer Cell 2012; 22:698-698.e1

60. Wander SA, Levis MJ, Fathi AT. The evolving role of FLT3 inhibitors in acute myeloid leukemia: quizartinib and beyond. Ther. Adv. Hematol. 2014; 5:65–77

61. NIH NCIGDC. Acute Myeloid Leukemia — Cancer Stat Facts.

62. Ragon BK, Odenike O, Baer MR, et al. Oral MEK 1/2 Inhibitor Trametinib in Combination With AKT Inhibitor GSK2141795 in Patients With Acute Myeloid Leukemia With RAS Mutations: A Phase II Study. Clin. Lymphoma Myeloma Leuk. 2019; 19:431-440.e13

63. Sutamtewagul G, Vigil CE. Clinical use of FLT3 inhibitors in acute myeloid leukemia. Onco. Targets. Ther. 2018; Volume 11:7041–7052

64. Lord CJ, Ashworth A. PARP Inhibitors: The First Synthetic Lethal Targeted Therapy. Science (80-.). 2017; 355:1152–1158

65. O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. Nat. Rev. Genet. 2017; 18:613–623

66. Huang A, Garraway LA, Ashworth A, et al. Synthetic lethality as an engine for cancer drug target discovery. Nat. Rev. Drug Discov. 2020; 19:23–38

67. McDonald ER, de Weck A, Schlabach MR, et al. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. Cell 2017; 170:577-592.e10

68. McFarland JM, Ho Z V., Kugener G, et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat. Commun. 2018; 9:1–13

69. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat. Rev. Genet. 2015; 16:299–311

70. Cowley GS, Weir BA, Vazquez F, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci. data 2014; 1:140035

71. Tatlow PJ, Piccolo SR. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci. Rep. 2016; 6:39259

72. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483:603–307

73. Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17:520–525

74. Ignatiadis N, Klaus B, Zaugg JB, et al. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat. Methods 2016; 13:577–580

75. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47

76. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet. Epidemiol. 2002; 23:70–86

77. Storey JD. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 2002; 64:479–498

78. Wagner AH, Walsh B, Mayfield G, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat. Genet. 2020; 52:448–457

79. Alterovitz G, Heale B, Jones J, et al. FHIR Genomics: enabling standardization for precision medicine use cases. npj Genomic Med. 2020; 5:9–12

80. Gaulton A, Hersey A, Nowotka ML, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017; 45:D945–D954

81. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 2018; 46:D1074–D1082

82. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47:D607–D613

83. Lee AJX, Swanton C. Tumour heterogeneity and drug resistance: Personalising cancer medicine through functional genomics. Biochem. Pharmacol. 2012; 83:1013–1020

84. Wilcox RR. ANOVA: A Paradigm for Low Power and Misleading Measures of Effect Size? Rev. Educ. Res. 1995; 65:51–77

85. Kazi JU, Rönnstrand L. FMS-like tyrosine kinase 3/FLT3: From basic science to clinical implications. Physiol. Rev. 2019; 99:1433–1466

86. López-Zabalza MJ, Martínez-Lausín S, Bengoechea-Alonso MT, et al. Signaling pathway triggered by a short immunomodulating peptide on human monocytes. Arch. Biochem. Biophys. 1997; 338:136–142

87. Pratz KW, Sato T, Murphy KM, et al. FLT3-mutant allelic burden and clinical status are predictive of response to FLT3 inhibitors in AML. Blood 2010; 115:1425–1432

88. Metzelder S, Röllig C. FLT3 inhibitors for the treatment of acute myeloid leukemia. Best Pract. Onkol. 2018; 13:182–190

89. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N. Engl. J. Med. 2016; 374:2209–2221

90. Pattabiraman DR, McGirr C, Shakhbazov K, et al. Interaction of c-Myb with p300 is required for the induction of acute myeloid leukemia (AML) by human AML oncogenes. Blood 2014; 123:2682–2690

91. Matthew L. Smith, M.B., B.S., Jamie D. Cavenagh, M.D., T. Andrew Lister MD, and Jude Fitzgibbon PD. Mutation of CEBPA in Familial Acute Myeloid Leukemia. N. Engl. J. Med. 2004; 351(23):2403–2407

92. Abbott KL, Nyre ET, Abrahante J, et al. The candidate cancer gene database: A database of cancer driver genes from forward genetic screens in mice. Nucleic Acids Res. 2015; 43:D844–D848

93. Pacini C, Dempster JM, Najgebauer H, et al. Integrated cross-study datasets of genetic dependencies in cancer. Nat. Commun. 2021; 12:1–14

94. Ignatiadis N, Klaus B, Zaugg JB, et al. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat. Methods 2016; 13:577–580

95. Guo XX, Wu HL, Shi HY, et al. The efficacy and safety of olaparib in the treatment of cancers: a meta-analysis of randomized controlled trials. Cancer Manag. Res. 2018; 10:2553

96. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. Nature 2020; 578:94–101

97. Hill R, Cautain B, De Pedro N, et al. Targeting nucleocytoplasmic transport in cancer therapy. Oncotarget 2014; 5:11–28

98. Daver N, Schlenk RF, Russell NH, et al. Targeting FLT3 mutations in AML: review of current knowledge and evidence. Leukemia 2019; 33:299–312

99. Wang S, Wu Z, Li T, et al. Mutational spectrum and prognosis in NRAS-mutated acute

myeloid leukemia. Sci. Rep. 2020; 10:12152

100. Hunter AM, Sallman DA. Current status and new treatment approaches in TP53 mutated AML. Best Pract. Res. Clin. Haematol. 2019; 32:134–144

101. Thiede C, Koch S, Creutzig E, et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). Blood 2006; 107:4011–4020

102. Zhang H, Nakauchi Y, Köhnke T, et al. Integrated analysis of patient samples identifies biomarkers for venetoclax efficacy and combination strategies in acute myeloid leukemia. Nat. Cancer 2020; 1:826–839

103. Wright CJM, McCormack PL. Trametinib: First Global Approval. Drugs 2013; 73:1245–1254

104. Gui P, Bivona TG. Stepwise evolution of therapy resistance in AML. Cancer Cell 2021; 39:904–906

105. Markham A, Keam SJ. Selumetinib: First Approval. Drugs 2020; 80:931–937

106. Kiessling M, Rogler G. Targeting the RAS pathway by mitogen-activated protein kinase inhibitors. Swiss Med. Wkly. 2015; 145:w14207

107. Antony ML, Noble-Orcutt K, Ogunsan O, et al. Cell Type-Specific Effects of Crizotinib in Human Acute Myeloid Leukemia with TP53 Alterations. Blood 2019; 134:2563–2563

108. McFarland JM, Ho Z V., Kugener G, et al. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. Nat. Commun. 2018; 9:4610

109. Meyers RM, Bryan JG, McFarland JM, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. 2017; 49:1779–1784

110. Wang T, Yu H, Hughes NW, et al. Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell 2017; 168:890-903.e15

111. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016; 166:740–754

112. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2012; 41:D955–D961

113. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012; 483:603–607

114. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 2019; 569:503–508

115. Chen S, Chen Y, Zhu Z, et al. Identification of the key genes and microRNAs in adult acute myeloid leukemia with FLT3 mutation by bioinformatics analysis. Int. J. Med. Sci. 2020; 17:1269

116. Lucena-Araujo AR, Souza DL, De Oliveira FM, et al. Results of FLT3 mutation screening and correlations with immunophenotyping in 169 Brazilian patients with acute myeloid leukemia. Ann. Hematol. 2010; 89:225–228

117. Gutiérrez NC, López-Pérez R, Hernández JM, et al. Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. Leukemia 2005; 19:402–409

118. Zhang L, Nguyen LXT, Chen Y-C, et al. Targeting miR-126 in inv(16) acute myeloid leukemia inhibits leukemia development and leukemia stem cell maintenance. Nat. Commun. 2021; 12:6154

119. Wunderlich M, Krejci O, Wei J, et al. Human CD34+ cells expressing the inv(16) fusion protein exhibit a myelomonocytic phenotype with greatly enhanced proliferative ability. Blood 2006; 108:1690–1697

120. Bewersdorf JP, Zeidan AM. Transforming growth factor (TGF)-β pathway as a therapeutic target in lower risk myelodysplastic syndromes. Leukemia 2019; 33:1303–1312

121. Muench DE, Ferchen K, Velu CS, et al. SKI controls MDS-associated chronic TGF-β signaling, aberrant splicing, and stem cell fitness. Blood 2018; 132:e24–e34

122. Bowman T V. Improving AML Classification Using Splicing Signatures. Clin. Cancer Res. 2020; 26:3503–3504

123. De Necochea-Campion R, Shouse GP, Zhou Q, et al. Aberrant splicing and drug resistance in AML. J. Hematol. Oncol. 2016; 9:1–9

124. Grinev V V., Barneh F, Ilyushonak IM, et al. RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia. Nat. Commun. 2021; 12:520

125. Jones CL, Stevens BM, D'Alessandro A, et al. Inhibition of Amino Acid Metabolism Selectively Targets Human Leukemia Stem Cells. Cancer Cell 2018; 34:724-740.e4

126. Hastie T, Tibshirani R, Narasimhan B, et al. impute: Imputation for microarray data. Bioinformatics 2001; 17:520–525

127. JJ Allaire, Kevin Ushey, Yuan Tang and DE. reticulate: R Interface to Python. 2017;

128. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47–e47

129. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov. 2021; 2:100141

130. Gerstung M, Papaemmanuil E, Martincorena I, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat. Genet. 2017; 49:332–340

131. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. Proc. 2018 ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics 2018; 559–560

132. Oh M, Park S, Kim S, et al. Machine learning-based analysis of multi-omics data on the

cloud for investigating gene regulations. Brief. Bioinform. 2021; 22:66–76

133. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell 2016; 166:740–754

134. Zeisig BB, Kulasekararaj AG, Mufti GJ, et al. SnapShot: Acute Myeloid Leukemia. Cancer Cell 2012; 22:698-698.e1

135. Surapally S, Tenen DG, Pulikkan JA. Emerging therapies for inv(16) AML. Blood 2021; 137:2579–2584

136. Zeiser R, Andrlová H, Meiss F. Trametinib (GSK1120212). Recent Results Cancer Res. 2018; 211:91–100

137. He X, Folkman L, Borgwardt K. Kernelized rank learning for personalized drug recommendation. Bioinformatics 2018; 34:2808–2816

138. Valcárcel L V., San José-Enériz E, Cendoya X, et al. BOSO: A novel feature selection algorithm for linear regression with high-dimensional data. PLOS Comput. Biol. 2022; 18:e1010180

139. Knijnenburg TA, Klau GW, Iorio F, et al. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. Sci. Reports 2016 61 2016; 6:1–14

140. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 2010; 33:1–22

141. Döhner H, Estey EH, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: Recommendations from an international expert panel, on behalf of the European LeukemiaNet. Blood 2010; 115:453–474

142. Lucena-Araujo AR, Coelho-Silva JL, Pereira-Martins DA, et al. Combining gene mutation with gene expression analysis improves outcome prediction in acute promyelocytic leukemia. Blood 2019; 134:951–959

143. . Artificial intelligence in medicine regulation | European Medicines Agency.

144. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature 2018; 562:526–531

145. Mayakonda A, Lin DC, Assenov Y, et al. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018; 28:1747–1756