

Imperial College London
Department of Bioengineering

**Machine Learning Prediction of
Atezolizumab Treatment Response in
Metastatic Urothelial Carcinoma Using
Gene Expression and Clinical Data**

Thesis by
Chayanit Piyawajanusorn

Supervisor: Dr. Pedro Ballester

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Bioengineering of the Imperial College London
and the Diploma of Imperial College London, December 2024

Declaration of Originality

I (Chayanit Piyawajanusorn) hereby declare that this submission is the result of my own original research, with contributions from colleagues as indicated in the co-authored papers listed in section 1.4. I affirm that none of the content in this work has been previously published or authored by another person, nor has it been substantially used in the pursuit of any other academic qualification, whether a degree or award from this university or any other institute of higher education. Any contributions or ideas from external sources have been properly acknowledged within the text. This declaration verifies the integrity and authenticity of my academic contributions contained within this thesis.

Chayanit Piyawajanusorn

Declaration of Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

"The only way to do great work is to love what you do."

- Steve Jobs

And this is what I love

Abstract

Cancer heterogeneity leads to variability in treatment responses among patients. Precision medicine addresses this by tailoring treatments based on individual tumour molecular profiles. ML offers a promising approach for predicting drug responses, but is challenged by high-dimensional, low-sample datasets, requiring careful data preparation and analysis.

This PhD project demonstrates successful ML application in predicting drug responses across four drug-cancer-type binomials. Since the predictive performance depends on the specific problem, it is important to incorporate diverse molecular profiles tailored to various ML algorithms to enhance the accuracy of drug response predictions. In atezolizumab-treated mUC, CART-OMC trained on the discovery dataset achieved the highest performance, with an MCC of 0.437 in the validation set using just 29 genes. Univariate biomarkers such as TMB, TNB, and PD-L1 were less predictive (MCCs of 0, 0.316, and 0, respectively). When datasets were merged, LGBM-OMC outperformed top modelling approaches like EaSIeR (MCC $\sim$ 0) and JADBio (MCC 0.179), achieving an MCC of 0.252. For taxane-anthracycline-treated breast cancer, LR using mRNAs achieved an MCC of 0.396, with comparable performance on an independent dataset (MCC of 0.351). In doxorubicin-treated breast cancer, CART achieved MCCs of 0.56 and 0.32 using 4 isomiRs and 3 miRNAs, respectively. In gemcitabine-treated pancreatic cancer, RF-OMC with 4 mRNAs achieved an MCC of 0.44, while XGBoost with 12 CpG probes reached an MCC of 0.32.

These findings demonstrate the robustness and generality of our proposed methodology, with strong predictive performance across four case studies. The ML models identified predictive, reproducible, and robust drug response signatures, aiding clinicians in identifying patients most likely to benefit from treatment to propose suitable drugs without delay while minimising the risk of side effects for patients unlikely to respond. We have made our ML models publicly available for prospective validation on patients with the same drug-cancer types.

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Pedro Ballester, for his unwavering support, guidance, and patience throughout this research journey. His expertise and encouragement have not only shaped the direction of this study but have also inspired me to broaden my perspectives and deepen my understanding of the field. His mentorship has been a cornerstone of my growth, both as a researcher and as an individual.

I extend my heartfelt thanks to my colleagues, whose support during the course of my thesis has gone beyond professional camaraderie. In times of doubt and challenge, our shared moments of laughter and encouragement, along with their willingness to listen, share a word of comfort, or simply be there during downtimes, have been invaluable to me.

I am also sincerely grateful to the Bioengineering department at Imperial College London for providing the resources, facilities, and intellectually stimulating environment that enabled the completion of this work. Your support has been instrumental in bringing this research to fruition.

A special note of appreciation goes to the Chulabhorn Royal Academy (Thailand) for their financial support through the Commemoration of Her Royal Highness Princess Chulabhorn's 60th Birthday Anniversary scholarship. This invaluable assistance has made my doctoral studies possible, and I am deeply honoured to have been a recipient of such an award.

To my family, words cannot fully capture my gratitude for your love and support. To my parents and my siblings, thank you for believing in me, cheering me on, and encouraging me to pursue my goals. To my partner, who supported me through all the stages of both this research and my life, including the late nights, the highs and lows.

This accomplishment would not have been possible without the contributions of everyone mentioned here, and I am profoundly grateful for their involvement. Without your support, this journey would have been different, and for that, I am thankful.

List of Abbreviations

10CV	Ten-fold CV
5CV	Five-fold CV
ADASYN	Adaptive Synthetic Sampling
AHPS	Algorithm and hyper-parameter space selection
AI	Artificial intelligence
ANOVA	Analysis of variance
APC	Antigen-presenting cell
AutoML	Automated machine learning
BBC-CV	Bootstrap bias corrected-cross validation
BC	Bladder cancer
BCG	Bacillus calmette–guérin
BH	Benjamin–Hochberg
BLCA	Bladder urothelial carcinoma
BSMOTE	Borderline-synthetic minority over-sampling technique
CART	Classification and regression tree
CCLE	Cancer cell line encyclopedia
CE	Configuration evaluator
CG	Configuration generator
CI	Confidence interval
ClusterC	Cluster centroids
CNN	Condensed nearest neighbour
CT	Computed tomography
CTLA-4	Cytotoxic T-lymphocyte associated antigen 4
CV	Cross-validation
DC	Dendritic cell
DL	Deep learning
DNN	Deep neural network
DREEP	Drug response estimation from single-cell expression profiles
DT	Decision tree
EaSIeR	Estimate systems immune response
ECOG	Eastern cooperative oncology group
ED	Early dropping
EFB	Exclusive feature budling
ELISA	Enzyme-linked immunosorbent assay
EMA	European medicines agency
EMR	Electronic medical record
ENN	Edited nearest neighbours
ER	Enrichment ratio
ES	Early stopping
FDA	Food and drug administration

FDR	False discovery rate
FN	False negative
FP	False positive
FPR	False positive rate
GBM	Gradient boosting machine
GC	Gemcitabine-cisplatin
GDC	Genomic data commons
GDSC	Genomics of drug sensitivity in cancer
GEO	Gene expression omnibus
GO	Gene ontology
GOSS	Gradient-based one-side sampling
hENT1	High human equilibrative nucleoside transporter 1
HR	Hazard ratio
IC50	Half maximal inhibitory concentration
ICI	Immune checkpoint inhibitor
IFN- γ	Interferon-gamma
IgG1	Immunoglobulin G1
IHC	Immunohistochemistry
IHT	Instance hardness threshold
Indels	Insertions and deletions
irAE	Immune-related adverse event
isomiR	miRNA isoform
IUS	Investigational use only
IV	Intravenous
JADBio	Just add data bio
KMSMOTE	KMeans synthetic minority over-sampling technique
KNN	k-nearest neighbors
LASSO	Least absolute shrinkage and selection operator
LGBM	Light gradient boosting machine
LOOCV	Leave one out cross-validation
LR	Linear regression
MCC	Matthew's correlation coefficient
MHC	Major histocompatibility complex
MIBC	Muscle-invasive bladder cancer
ML	Machine learning
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
MSE	Mean squared error
mUC	Metastatic urothelial carcinoma
Mut/Mb	Mutations per megabase
MVAC	Methotrexate, vinblastine, doxorubicin, and cisplatin
NCBI	National centre for biotechnology information
NCI	National cancer institute

NGS	Next-generation sequencing
NK	Natural killer
NMIBC	Non-muscle invasive bladder cancer
NPV	Negative predictive value
NPV	Negative predictive value
NS	Non-significant
NSCLC	Non-small cell lung cancer
OMC	Optimal model complexity
ORR	Objective response rate
OS	Overall survival
OSS	One-sided selection
PAAD	Pancreatic adenocarcinoma
pCR	Pathological complete response
PD-1	Programmed cell death protein 1
PD-L1	Programmed cell death protein ligand 1
PDX	Patient-derived xenograft
PE	Performance estimator
PFS	Progression-free survival
PPI	Protein-protein interaction
PR-AUC	Precision-recall area under the curve
RBF	Radial basis function
RD	Residual disease
RECIST	Response evaluation criteria in solid tumours
RF	Random forest
RMTLR	Elastic net regularized multi-task linear regression
RMTLR	Elastic net regularized multi-task linear regression
ROC-AUC	Area under the receiver operating characteristic curve
RUS	Research use only
SCLC	Small-cell lung cancer
scRNA-seq	Single-cell RNA sequencing
SD	Standard deviation
SES	Statistical equivalent signature
SHAP	Shapley additive explanations
SMOTE	Synthetic minority over-sampling technique
SNP	Single nucleotide polymorphism
SNV	Single-nucleotide variant
SSR	Sum of squared residual
SVM	Support vector machine
SVM SMOTE	Support vector machine synthetic minority over-sampling technique
SVR	Support vector regression
TCC	Transitional cell carcinoma
TCGA	The cancer genome atlas
TF	Transcription factor

TGF- β	Transforming growth factor- β
TIC	Tumour-infiltrating immune cell
TIDE	Tumour immune dysfunction and exclusion
TIL	Tumour-infiltrating lymphocyte
TMB	Tumour mutational burden
TME	Tumour microenvironment
TN	True negative
TNB	Tumour neoantigen burden
TNR	True negative rate
TP	True positive
TPR	True positive rate
Treg	regulatory T-cell
TUR	Transurethral resection
UC	Urothelial carcinoma
WebGestalt	WEB-based Gene SeT AnaLysis Toolkit
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World health organization
XGBoost	eXtreme gradient boosting

Contents

Declaration of Originality	i
Declaration of Copyright	ii
Abstract	iv
Acknowledgements	v
List of Abbreviations	vi
Contents	x
List of Tables	xiii
List of Figures	xiv
Chapter 1	
1. Introduction	
1.1 Motivation	1
1.2 Aims and objectives	3
1.3 Thesis outline	4
1.4 Publications	5
Chapter 2	
2. Background theory	
2.1 The overview of UC	7
2.1.1 UC	7
2.1.2 Environmental and hereditary risk factors for UC	8
2.1.3 UC screening and diagnosis	8
2.1.4 Treatment guidelines for UC and challenges in treatment	9
2.2 ICIs in mUC	10
2.3 Atezolizumab: mechanism of action and clinical significance	12
2.4 Clinical response biomarkers of ICIs and their challenges	14
2.4.1 PD-L1 expression	14
2.4.2 TMB	15
2.4.3 TNB	16
2.5 Gene expression profiles as predictors of response to ICIs in cancer patients	17
2.6 The overview of ML in cancer drug response prediction	18
2.7 ML in predicting the response to ICIs	20
2.8 The state-of-the-art ML in drug response prediction	22
2.8.1 EaSIeR	22
2.8.2 JADBio autoML	24
2.9 ML fundamentals	27
2.9.1 The general architecture of the supervised ML workflow	28
2.9.2 The overview of data pre-processing for ML	30
2.9.2.1 Handling missing values	30
2.9.2.2 Categorical feature encoding	31
2.9.2.3 Data scaling	31
2.9.2.4 Handling data imbalance in ML	32
2.9.2.5 Data splitting	35
2.9.3 Binary classification task by regression algorithms	35

2.9.4	Supervised ML algorithms	36
2.9.4.1	Linear/logistic regression	36
2.9.4.2	Decision tree	37
2.9.4.3	Random forest	37
2.9.4.4	eXtreme gradient boosting	38
2.9.4.5	Light gradient boosting machine	39
2.9.4.6	Support vector machine	40
2.9.4.7	K nearest neighbour	40
2.9.5	Model building and evaluation	41
2.9.5.1	The general procedure of k-fold CV	41
2.9.5.2	Nested k-fold CV with OMC feature selection	42
2.9.6	ML model evaluation metrics	43
2.9.6.1	MCC	44
2.9.6.2	ROC-AUC	45
2.9.6.3	PR-AUC	45
Chapter 3		
3. Materials and methods		
3.1	Gene expression profiles and clinical data acquisition	47
3.2	Data pre-processing for ML	47
3.3	Building OMC regression models	51
3.4	Building models based on clinical data only	51
3.5	Building models based on immunotherapy response biomarkers	52
3.6	Building random class permutation models	52
3.7	Model performance evaluation	52
3.8	Prediction of response to atezolizumab using the EaSIeR	53
3.9	Prediction of response to atezolizumab using the JADBio	54
3.10	Statistical analysis	54
Chapter 4		
4 Results		
4.1	Cohort characteristics of atezolizumab-treated mUC patients	55
4.2	ML model performance in predicting mUC patient response to atezolizumab using gene expression profiles	56
4.3	The impact of clinical data integration with gene expression profiles in predicting mUC patient response to atezolizumab	60
4.4	Comparison of the best-performing models to clinical immunotherapy response biomarkers	67
4.5	Assessing the robustness of the best-performing model in predicting mUC patient response to atezolizumab	67
4.6	Analysis of cluster heat map using the 29 predictive genes	68
4.7	Feature importance of the 29 predictive genes	71
4.8	Differential expression of the 29 predictive genes	74
4.9	PPI network analysis of the 29 predictive genes	75
4.10	Wikipathway cancer and GO pathway enrichment analysis of the 29 predictive genes	78
4.11	Comparison of the best-performing models to EaSIeR models	84

4.12 Enhancing prediction on the merged dataset by integrating TME descriptors, immune response scores, and the EaSleR score into the best-performing models	88
4.13 Comparison of the best-performing models to JADBio autoML	91
Chapter 5	
5 Predictive modelling for other drug responses in cancer patients	
5.1 Case study I: Predictive modelling of breast cancer patient response to taxane-anthracycline chemotherapy using gene expression profiles	94
5.1.1 Cohort characteristics of taxane-anthracycline chemotherapy-treated breast cancer patients	94
5.1.2 Development and validation of ML models	95
5.1.3 Predicting breast cancer patient response to taxane-anthracycline chemotherapy from gene expression profiles	96
5.1.4 External validation of the best-performing model	96
5.2 Case study II: Predictive modelling of breast cancer patient response to doxorubicin based on their tumour molecular profiles	98
5.3 Case study III: Predictive modelling of response to gemcitabine in PAAD patients based on their tumour molecular profiles and clinical data	100
Chapter 6	
6 Conclusion and discussion	
6.1 Summary of thesis achievement	104
6.2 Clinical applications	110
6.3 Limitations of the study	111
6.4 Future works	111
Bibliography	114
Appendix I. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling	147
Appendix II. Interpretable machine learning models to predict the resistance of breast cancer patients to doxorubicin from their microRNA profiles	164
Appendix III. Large-scale machine learning analysis reveals DNA methylation and gene expression response signatures for gemcitabine-treated pancreatic cancer	179

List of Tables

Table 1.	Summary of various studies that employed ML to predict ICI treatment outcome in cancer patients	21
Table 2.	Five descriptive signatures of the tumour microenvironment from the EaSIeR tool	23
Table 3.	10 published signatures of the immune response	23
Table 4.	ML algorithms and optimisation techniques, including preprocessing, feature selection, model selection, and hyperparameter tuning used in JADBio	27
Table 5.	The baseline clinical characteristics of mUC patients from the discovery and validation cohorts	46
Table 6.	Five-repetition evaluation metrics for the validation dataset across eight ML models using gene expression profiles	56
Table 7.	The list of 29 predictive genes derived from the best-performing CART-OMC model	56
Table 8.	Five-repetition evaluation metrics for the validation dataset across eight ML models using integrated gene expression profiles with clinical data	60
Table 9.	The list of 69 predictive features (67 genes, TMB per megabase, and TNB per megabase) derived from the best-performing LR-OMC model	60
Table 10.	The Wikipathway cancer and GO pathway enrichment analysis of 29 predictive genes from the best-performing models (CART-OMC) in predicting atezolizumab response in mUC patients	78
Table 11.	The list of common features between the best-performing models using gene expression profile (GEP) and integrated gene expression profiles and clinical data (GEP + clinical) from validation and merged analyses	82
Table 12.	The predictive performance of the best-performing models vs. JADBio models in predicting atezolizumab response of 320 mUC patients from the merged discovery and validation datasets	89
Table 13.	The evaluation metrics of the best-performing model using gene expression profiles on the independent dataset	95
Table 14.	Overview of drug-cancer type binomials and best-performing models	100

List of Figures

Figure 1.	Mechanism of ICIs in Cancer Therapy	11
Figure 2.	The workflow of the JADBio system	26
Figure 3.	Schematic of ML workflow	29
Figure 4.	The example of k-fold CV with k=10	39
Figure 5.	Nested k-fold CV scheme for feature selection	40
Figure 6.	Confusion matrix for binary classification	41
Figure 7.	ROC-AUC and PR-AUC	43
Figure 8.	ML workflow to predict atezolizumab response in mUC patients	46
Figure 9.	Five-repetition evaluation metrics for the validation dataset across eight ML models using gene expression profiles	55
Figure 10.	Five-repetition evaluation metrics for the validation dataset across eight ML models using integrated gene expression profiles with clinical data	59
Figure 11.	Five-repetition evaluation on the validation dataset of the best-performing models versus those using TMB, TNB, or PD-L1 immunotherapy response biomarkers	63
Figure 12.	CV performance of the best-performing CART-OMC model compared to its permutation model	65
Figure 13.	Unsupervised clustering heatmap of 29 predictive genes in atezolizumab-treated mUC patients	67
Figure 14.	Feature importance plot of 29 predictive genes from the best-performing CART-OMC model for predicting atezolizumab responses in mUC patients	69
Figure 15.	Model interpretation using SHAP	70
Figure 16.	Boxplot showing the expression pattern of 29 predictive genes from the best-performing CART-OMC model between responders and non-responders	72
Figure 17.	PPI network analysis	74
Figure 18.	Wikipathway cancer and GO pathway enrichment analysis of 29 predictive genes for predicting atezolizumab response in mUC patients	76
Figure 19.	Five-repetition MCCs of the merged discovery and validation datasets across eight ML models using either gene expression profiles or integrated gene expression profiles plus clinical data	82
Figure 20.	Performance comparison of the best-performing models with those from the 60 EaSIeR model prediction on 320 mUC patients from the merged discovery and validation datasets (245 non-responders and 75 responders)	84

Figure 21.	Comparison of the predictive performance of the best-performing models using their predictive features to those combining with TME descriptors, immune response scores, and the EaSIeR score for atezolizumab response prediction in 320 mUC patients from the merged discovery and validation datasets	87
Figure 22.	Five-repetition evaluation metrics on the discovery dataset across eight ML models using gene expression profiles	94
Figure 23.	Heatmap showing the mMCC of five 10-CV repetitions for 128 models in predicting response to doxorubicin in breast cancer patients	96
Figure 24.	Heatmap presenting the mMCC of five 10-fold CV repetitions for 128 models predicting gemcitabine response in PAAD patients	98

Chapter 1

Introduction

1.1 Motivation

Urothelial carcinoma (UC) is the ninth most common type of cancer worldwide, affecting the urinary tract system. Its incidence is consistently rising, with about 600,000 new cases and 200,000 global deaths (31.7%) in 2022[1]. 15% of patients are initially diagnosed with locally advanced or metastatic urothelial carcinoma (mUC), and just 5% of patients have a 5-year survival rate in the metastatic setting[2], [3]. For decades, platinum-based chemotherapy has been the standard first-line treatment for mUC patients[4]. While this drug treatment has been shown to improve patient's clinical outcomes, its toxicity has been a significant concern in clinical practice. More than 60% of patients with mUC are not eligible for this treatment due to their poor functional ability, as determined by Eastern Cooperative Oncology Group (ECOG) performance status, or other comorbidities such as advanced age, renal impairment, hearing loss, and heart failure[5]–[8].

To address this clinical need, immune checkpoint inhibitors (ICIs) have become a promising alternative for cancer patients who have either progressed or have been considered ineligible for platinum-based chemotherapy. The U.S. Food and Drug Administration (FDA) has approved three categories of ICIs: two programmed cell death protein 1 (PD-1) inhibitors (nivolumab and pembrolizumab), three programmed cell death protein ligand 1 (PD-L1) inhibitors (atezolizumab, durvalumab, and avelumab), and cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4) inhibitors (Ipilimumab), in various solid tumours, including mUC[8]–[10].

Atezolizumab is an Fc-engineered, humanised immunoglobulin G1 (IgG1) monoclonal antibody that binds to the PD-L1 protein expressed on the surface of tumour cells and other PD-L1-expressing cells, such as tumour-infiltrating immune cells (TICs). By binding to PD-L1, atezolizumab inhibits its interaction with PD-1 receptors, mainly expressed in T-cells, thereby reactivating T-cell-mediated anti-tumour responses[11]. Although atezolizumab has considerably improved patient outcomes in several cancer types, including mUC[11]–[15], there is substantial heterogeneity in response to therapy. Most patients still do not respond to atezolizumab, experience immune-related adverse events (irAEs) at considerable cost and lack robust biomarkers. In the phase II IMvigor210 trial[2], the safety and efficacy of atezolizumab in mUC were first evaluated in patients who had progressed during or following platinum-based chemotherapy. The objective response rate (ORR) of these patients, defined as the proportion who achieved either a complete or partial response to treatment according to the response evaluation criteria in solid tumours (RECIST) version 1.1, was low (14.8%), with only 5% exhibiting a complete response. Treatment-related adverse events occurred, leading to treatment discontinuation. Therefore, it is an urgent clinical need and an essential step to identify predictive biomarkers to select mUC patients

who are most likely to benefit from ICI-based therapy upfront. Doing so will enable the administration of promising drugs to eligible patients without delay while minimising the risk of immunological toxicities for those patients who are unlikely to respond to ICIs[16].

The majority of research on ICI response has focused primarily on univariate markers. Positive PD-L1 expression, high tumour mutational burden (TMB), and high tumour neoantigen burden (TNB) have been proposed as biomarkers that could potentially predict ICI response in multiple cancer types[17]–[22]. However, the results from these studies are often inconsistent. For example, other studies have reported no significant correlation between these biomarkers and the ICI treatment response[23], [24], and some studies have even demonstrated that the effect of ICI therapy is not only restricted to patients with high PD-L1 expression but also to those with low PD-L1 expression, who may also benefit from ICI treatment[25]–[29]. In addition, the cutoffs for PD-L1 positivity to select patients for ICI therapy vary within and between different cancer types, and there is a lack of a standard method for measuring PD-L1[30]. Despite immunohistochemistry (IHC) assay being used as a reference assay to evaluate the membrane expression of PD-L1, variations in PD-L1 testing kits, staining techniques, scoring systems, or components tested (tumour cells, TICs, or both), can lead to inconsistent results. Moreover, the measurement of TMB and TNB is challenging due to the cost and time-consuming of the process, as well as the lack of consensus in the methodologies used to consistently quantify TMB and TNB across different sequencing platforms and laboratories. Furthermore, establishing robust cutoffs for identifying a high burden varies across different cancer types[31], [32]. These present significant limitations to the use of PD-L1, TMB, and TNB as univariate markers in clinical practice[33].

The advances in high-throughput sequencing and microarray technologies characterising biological scales, spanning from whole tumour tissues to single cells, have led to the discovery of a wide range of biomarker strategies from detecting a single biomarker to developing multifactorial synergistic predictive biomarkers. These developments have generated an abundance of data and facilitated the growth of machine learning (ML) in precision oncology to aid in cancer detection and diagnosis, predict patient outcomes, and inform treatment planning[34]. ML, a subfield of artificial intelligence (AI), has been recently used as a promising approach that can efficiently improve the accuracy of clinical predictive models[35]–[37]. The ML-based model can identify the predictive combination of features in high-dimensional data such as genomic, transcriptomic, and imaging, unlike traditional statistical approaches (e.g., Cox regression), which typically focus on a limited set of predictors such as tumour size, patient demographics, and biomarker expression levels[38]–[41]. In recent years, several studies have developed multiple ML models predicting drug response, each tailored to specific cancer types and their treatment data. Our previous research has successfully developed predictive ML models for tumour response to chemotherapy in cell lines[42]–[44], patient-derived xenografts (PDXs)[45], and cancer patients[46] using high-dimensional multi-omics and/or clinicopathological data. We achieved maximum Matthew's correlation coefficients (MCCs) slightly above 0.5 and 0.3 using random forest (RF) with transcriptomic profiles [42], [45] and genomic profiles[43] on

preclinical pharmaco-omics data, respectively. Clinical pharmaco-omics datasets, which are the most relevant for the patient, are much less available to train and evaluate ML models. Thanks to the public availability of large-scale, multi-omics cancer databases (e.g., The Cancer Genome Atlas (TCGA); <https://www.cancer.gov/tcga>), Bomane et al.[46] were focused on predicting paclitaxel response in breast cancer patients, expanding their analysis to include 6 molecular profiles and 10 ML algorithms. Interestingly, decision tree (DT) and, eXtreme gradient boosting (XGBoost), when integrated with feature selection, demonstrated highly predictive models for this drug-cancer-type binomial, based on the miRNA and DNA methylation profiles (with an MCC of 0.25 for XGBoost_CGI methylation, 0.52 for DT_CpG methylation, and 0.43 for DT_miRNA).

Beyond conventional chemotherapy response prediction, several studies have demonstrated the use of ML-based models for assessing the response to ICIs in cancer patients[47]–[52]. Nassar et al.[52] reported a multivariable logistic regression model that combined clinical and genomic features to predict the response to anti-PD1/anti-PD-L1 therapy among 62 mUC patients. However, the results require further validation due to the limited sample size and the lack of an independent validation cohort. Despite ICIs demonstrating meaningful antitumor activity, patient response to ICIs may be influenced by various biological factors, including complex interactions in the tumour microenvironment (TME)[53]. Charoentong et al.[50] focusing on the interactions between tumours and immune cells, developed an immune-related gene signature called the immunophenoscore. This score could be used to stratify patients who are likely to respond to CTLA-4 and PD-1 inhibitors. Lapuente-Santana et al.[48] developed Estimate Systems Immune Response (EaSIeR), a multi-task ML approach for predicting anti-PD-1/anti-PD-L1 response based on system-based signatures of the TME. While EaSIeR marks a significant advancement, it obtained an Area Under the Receiver Operating Characteristic Curve (ROC-AUC) of just 0.65 on the mUC cohort. This highlights the need for more predictive models.

1.2 Aims and objectives

To our knowledge, ML has not yet been applied to predict the response of mUC patients to atezolizumab. With this purpose, we retrospectively collected tumour gene expression profiles and clinical data of mUC patients treated with atezolizumab from two independent studies: the IMVigor210 phase II clinical trial (N = 298)[2] and the study by Snyder et al. (N = 22)[54]. Our objectives can be summarised as follows:

Objective 1: To develop a highly predictive and robust ML model that could identify responders and non-responders to atezolizumab in mUC patients based on their tumour gene expression profiles and clinical data. Since the predictive ability of ML algorithms generally depends on the specific problem, it is important to evaluate a range of ML algorithms for predicting atezolizumab response in mUC patients, to avoid missing the most predictive models due to limited algorithm choice. Eight regression algorithms were applied to build models with optimal model complexity (OMC) feature selection on the

IMvigor210 discovery dataset[2], which were then validated on the Snyder validation dataset[54].

Objective 2: To compare the predictive performance of our best-performing model to univariate immunotherapy biomarkers, including TMB, TNB, and PD-L1 protein expression.

Objective 3: To compare the predictive performance of our best-performing model to state-of-the-art ML drug response prediction models, including EaSIeR[48] and Just Add Data Bio (JADBio)[55].

Objective 4: To demonstrate the robustness and generality of the proposed ML methodology by extending its application to different chemotherapeutic drug-cancer-type binomials: taxane-anthracyclines in breast cancer, doxorubicin in breast cancer, and gemcitabine in pancreatic adenocarcinoma (PAAD). In addition, we aim to demonstrate the advantages of incorporating diverse tumour molecular profiles, including mRNA, DNA methylation, copy number variations (CNVs), miRNAs and their isoform, tailored to various ML algorithms, to enhance the accuracy of drug response predictions.

1.3 Thesis outline

This thesis comprises the following chapters:

Chapter 2: This chapter provides an extensive review of literature relevant to the thesis topic, starting with an overview of mUC, followed by ICIs in mUC, focusing on atezolizumab. The thesis then addresses clinical predictive biomarkers of response to ICIs, discussing PD-L1 protein expression, TMB, and TNB. An analysis of gene expression profiles in predicting ICI responses sets the stage for exploring the use of ML in cancer drug response prediction. This includes a review of ML fundamentals and data pre-processing techniques. A detailed discussion on the use of regression algorithms for classification tasks is provided, covering a range of ML algorithms. Model building and validation techniques such as data splitting and cross-validation are covered, along with evaluation metrics to assess the model's performance. Finally, the thesis also includes state-of-the-art ML drug response prediction models, including the EaSIeR and JADBio tools.

Chapter 3: This chapter describes the materials and methods used in the study, beginning with the acquisition of gene expression profiles and clinical data, followed by data pre-processing steps, and the development and evaluation of our predictive ML models, as well as state-of-the-art-ML models like EaSIeR and JADBio for drug response prediction. The analysis extends to agglomerative hierarchical clustering, the construction of a protein-protein interaction (PPI) network, and the Wikipathway cancer and gene ontology (GO) pathway enrichment analysis to provide insights into molecular interactions and underlying biological processes associated with response to atezolizumab and its mechanism. Furthermore, the statistical tests used to assess the difference in research findings are also described.

Chapter 4: This chapter presents the key findings of the research. First, we identified the best-performing models developed on the discovery dataset, which accurately stratified responders and non-responders to atezolizumab among mUC patients on the validation dataset, based on their tumour gene expression profiles and clinical data. These models were then compared to univariate immunotherapy biomarkers, including TMB, TNB, and PD-L1. Furthermore, we also highlight that the best-performing model could identify the predictive genes associated with the response to atezolizumab and its underlying mechanism. To challenge the model with higher test set diversity, the discovery and validation datasets were merged. This allowed us to compare the predictive performance of our best-performing models with state-of-the-art ML models (e.g., EaSIeR and JADBio). Finally, we explore whether incorporating predictors associated with ICI response from the EaSIeR into our predictive models could improve the accuracy of predicting responses to atezolizumab.

Chapter 5: We explore the application of ML to predict responses to three different chemotherapeutic drug-cancer-type binomials, extending our previous successes with atezolizumab immunotherapy in mUC patients. Since the predictive ability of ML models generally depends on the specific problem, we carried out a large-scale ML analysis employing various ML algorithms across multiple molecular profiles, including mRNA, DNA methylations, CNVs, miRNAs and their isoform, tailored for each drug-cancer-type binomial: taxane-anthracyclines in breast cancer, doxorubicin in breast cancer, and gemcitabine in PAAD. These studies demonstrated the robustness and generality of our ML methodology and highlighted the importance of using multiple algorithms and molecular profiles to enhance the accuracy of drug response predictions.

Chapter 6: This chapter provides a concise summary of the main research findings. It also highlights the clinical application of predictive ML models in medical settings, the limitations of the study, and future directions.

1.4 Publications

1. **C. Piyawajanusorn**, L. C. Nguyen, G. Ghislat, and P. J. Ballester, "A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling," *Brief. Bioinform.*, vol. 22, no. 6, Nov. 2021, doi: 10.1093/BIB/BBAB312.
2. A. Z. Ogunleye, **C. Piyawajanusorn**, A. Gonçalves, G. Ghislat, and P. J. Ballester, "Interpretable Machine Learning Models to Predict the Resistance of Breast Cancer Patients to Doxorubicin from Their microRNA Profiles," *Adv. Sci. (Weinheim, Baden-Wurttemberg, Ger.)*, vol. 9, no. 24, Aug. 2022, doi: 10.1002/ADVS.202201501.
3. A. Ogunleye, **C. Piyawajanusorn**, G. Ghislat, and P. J. Ballester, "Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer," *Heal. Data Sci.*, Jan. 2024, doi: 10.34133/HDS.0108.
4. G. Ghislat, S. Hernandez-Hernandez, **C. Piyawajanusorn**, and P. J. Ballester, "Data-centric challenges with the application and adoption of artificial intelligence for drug

discovery," *Expert Opin. Drug Discov.*, pp. 1–11, Sep. 2024, doi:
10.1080/17460441.2024.2403639.

Chapter 2

Background theory

2.1 The overview of UC

2.1.1 UC

UC, also known as transitional cell carcinoma (TCC), is the most common malignancy affecting the urinary tract system. It is characterised by abnormal cell growth in the urothelium, which is the lining of either the lower (bladder, urethra) or upper (ureters, renal pelvis) urinary tract. It is the most common histologic type of bladder cancer, accounting for approximately 90% of all bladder and only about 7% of all types of kidney cancer, including cancer of the renal pelvis and ureter[56]. UC is categorised into three groups based on the extent of invasion into the bladder wall and the presence of metastasis: (1) non-muscle invasive bladder cancer (NMIBC), which is limited to the urothelium and does not extend beyond the lamina propria; (2) muscle-invasive bladder cancer (MIBC), which grows in the lamina propria and invades the muscle layer; (3) unresectable locally advanced or mUC, when it has spread to regional lymph nodes, abdominal or pelvic wall or distant sites. The most common sites of metastasis include regional lymph nodes, liver, lungs, and bones, less commonly, brain and heart[57]. Each histological variant has distinct characteristics, such as metastatic potential, expression of immunotherapy targets, and susceptibility to radiation or chemotherapy[58].

UC is one of the most aggressive cancers. According to data from the World Health Organization (WHO)[1], there were more than 600,000 newly diagnosed cases and over 200,000 deaths in 2022, making it the ninth most common cancer worldwide, and its prevalence has been on the rise globally. The rates of incidence and mortality vary significantly across different regions of the world, with Western countries, such as those in Europe and North America, reporting the highest incidence rates. In Europe, approximately 52,000 new cases and around 17,000 deaths occurred, while the United States of America reported an estimated 22,000 new cases and 6,000 deaths during the same year. The incidence and mortality rates for UC in males are three to four times higher than in females. Although the majority of cases are diagnosed with early NMIBC, about 30% of patients present with MIBC or are 10-15% initially diagnosed with locally advanced or mUC at the time of diagnosis[57], [59]. This contributes to a significant worldwide burden of morbidity and mortality. Moreover, loco-regional and/or distant disease recurrence following radical cystectomy is common, with local relapse occurring in around 30% of patients and distant metastases reported in up to 50% of UC cases[60]. In addition, UC has a 5-year survival rate of around 70% for localised cancers, 39.2% for regional cancers, and 5% for distant metastatic cancers[61]. The disease is associated with poorer clinical outcomes and significantly worsens prognosis, particularly in the metastatic stage; therefore, ongoing efforts to improve management options are crucial[62].

2.1.2 Environmental and hereditary risk factors for UC

UC develops and progresses through diverse genetic and acquired pathways[63]. Environmental factors are the major risk factors for UC development, with cigarette smoking being the most significant risk factor, accounting for approximately 50% of UC cases in both males and females. Current smokers are four to five times more likely to develop UC than non-smokers[64], [65]. Other environmental risk factors associated with the development of UC include occupational exposure to high-risk carcinogens (e.g., aromatic amines or polycyclic aromatic hydrocarbons), radiation, contaminated drinking water, and chronic bladder infections[66]. The consequences of these risk factors and exposures result in alterations within the urothelial tract, increasing the likelihood of tumour recurrence, a phenomenon known as polychronotropism[67]. While environmental factors contribute significantly to the increased risk, genetic factors are also implicated in the development of UC. Epidemiologic studies indicate that individuals with a family history of cancer doubles the risk of developing the disease, with numerous reports of multiple-case UC kindreds[68], [69].

2.1.3 UC screening and diagnosis

The effectiveness of UC treatment largely depends on early detection. As for many tumours, early detection increases the chances of successful treatment or prolonged survival. Since the bladder serves as the exclusive reservoir of urine, thus, urine is a more specific medium for detecting urologic tumour markers than serum. UC diagnosis involves a combination of medical history assessment, physical examination, imaging investigations, and biopsy of the affected tissue. Microhaematuria testing and urine cytology are the most common and widely used techniques for detecting microscopic haematuria and screening for UC. However, due to their low sensitivity for detecting early-stage UC, cystoscopy is required to confirm the diagnosis. Urine cystoscopy combined with histopathologic examination of biopsied tissue is the gold standard for diagnosis of the disease. However, cystoscopy is a relatively expensive procedure and unsuitable for widespread use in screening programs. These issues have led to the development of other non-invasive molecular screening methods that can detect UC with greater sensitivity and specificity. For example, BTA stat®, NMP22® BladderChek® and Accu-Dx™ are point-of-care molecular screening tests. These tests are enzyme-linked immunosorbent assays (ELISAs), which require specialised setup, instrumentation, and personnel; urine specimens must be sent to reference laboratories. However, most molecular screening tests for UC have median sensitivities less than 90%[70]. Regarding prognostic markers, the scope of combining multiple molecular screening tests to improve detection capability is also being examined. Research addressing their potential for UC screening and surveillance is described elsewhere[70]. In addition, Vriesema et al.[71] reported that 89% of patients prefer urine cystoscopy over molecular screening methods for follow-up if the sensitivity of the latter is less than 90%.

2.1.4 Treatment guidelines for UC and challenges in treatment

Treatment aims to reduce tumour burden, treat cancer-related symptoms, increase overall survival (OS), and prevent recurrence and/or progression. Treatment options are determined by various key factors, including type of cancer, grade, stage, patient's performance status, concurrent comorbidities, and organ function (including cognitive function assessment). Current guidelines describe complete transurethral resection (TUR) as a crucial first step in treating NMIBCs and clinical staging to ensure total tumour removal and to reduce the chance of recurrence and/or progression to invasive forms, resulting in high survival rates[72]. Inadequate excision of the tumour or failure to detect tumours during the first resection might result in disease recurrence or incorrect staging[73], [74].

Patients who undergo TUR may be given additional treatment, such as intravesical Bacillus Calmette–Guérin (BCG) immunotherapy. This therapy is considered a first-line option due to its lower risk of side effects, although its efficacy is only 50%[75]. Intrauterine injection of BCG causes extensive local inflammation in the bladder wall, targeting tumour cells, to reduce the recurrence rate of tumours and to diminish the risk of progression. Nevertheless, intravesical BCG immunotherapy might have short-term immune-stimulating effects[76]. The standard BCG treatment dosage consists of six weekly doses administered as an induction course after TUR[77]. Extra doses of BCG given at regular intervals after the initial induction course, known as maintenance BCG therapy, aim to prevent cancer from recurring and to improve long-term results[78]. Patients with a history of BCG sepsis, immunosuppression, significant haematuria or traumatic catheterisation, or a current urinary tract infection are not suitable candidates for BCG immunotherapy due to the possible hazards[79]. Mild cystitis, low-grade fever, and malaise are common side effects of intravesical BCG immunotherapy. Approximately 90% of individuals may experience these symptoms[80]–[82].

While local therapy with TUR and BCG are commonly used as standard treatments for NMIBCs, patients with locally advanced or mUC are typically recommended to undergo a radical cystectomy. This surgical procedure involves the removal of tumours and surrounding tissues that the cancer might invade[83]. Studies have demonstrated that neoadjuvant chemotherapy for MIBCs has been shown to improve survival rates[84]. In the treatment of mUC, the first-line treatment has been platinum-based chemotherapy, including the combinations of methotrexate, vinblastine, doxorubicin, and cisplatin (MVAC) or gemcitabine-cisplatin (GC). These treatment regimens have achieved a median OS of about 15 months. MVAC demonstrated a significant OS improvement in comparison to cisplatin alone, with a median OS of 12.5 months vs 8.2 months (hazard ratio (HR) not available)[85]. In a phase III trial, the GC combination was directly compared to MVAC. Although the trial was aimed to establish GC's superiority, the two chemotherapeutic regimens demonstrated equivalent long-term OS (14.0 vs 15.2 months; HR 1.09) and progression-free survival (PFS) (7.7 vs 8.3 months; HR 1.09)[86], [87]. However, the clinical application of MVAC has been limited due to a high incidence of severe toxicity, including grade 3 or higher leukopenia, neutropenic fever, mucositis, and gastrointestinal adverse

events, with a drug-related mortality rate of 3%-4%. GC demonstrated a significantly better tolerability profile in terms of adverse events and quality of life than MVAC. These findings resulted in the common use of GC as a first-line chemotherapy regimen for patients with mUC[88]. Although cisplatin-based chemotherapy has response rates in up to 40%–50% of the patients and a 5-year OS rate of 13–15%, durable disease remission is rarely observed, and patients eventually progress within 9 months. Up to 50% of patients are cisplatin-ineligible because of comorbidities, poor patient performance status, or impaired renal function[89]. Cisplatin-ineligibility criteria were defined by Galsky et al.[6] as the presence of any of the following: ECOG performance status of 2 or greater, creatinine clearance less than 60 ml/min, grade >2 neuropathy, grade >2 hearing loss, and New York Heart Association grade III/IV heart failure.

2.2 ICIs in mUC

Under normal physiological conditions, immune checkpoints play a crucial role in maintaining self-tolerance by regulating T-cell activity. These checkpoint proteins send inhibitory signals to T-cells when they bind to partner proteins on other cells, preventing the immune system from attacking healthy tissues. However, in the presence of tumours, the expression of immune-checkpoint proteins was shown to be dysregulated by the tumours. This prevents T-cells to recognise and eliminate tumour cells, serving as a crucial mechanism of immune resistance[90]. Some examples of checkpoint proteins include PD-1 and CTLA-4 on immune cells, and PD-L1 on tumour cells. As shown in Figure 1, PD-1 (CD279) is a membrane receptor expressed by immune cells including T-cells, B-cells, dendritic cells (DCs), and macrophages. It is involved in inhibiting lymphocytes and regulating T-cell-mediated immune responses by interacting with its ligands PD-L1 (CD274), which is expressed by tumour cells[91]. The interaction of PD-1 and PD-L1 inhibits the activation of T-cells and the production of cytotoxic mediators; thus, it promotes tumour progression by increasing tumour cell survival[92]. Some tumours up-regulate the expression of PD-L1 as a coinhibitory ligand to inhibit the activity of PD-1, leading to immune suppression and tumour cells escaping immune surveillance[93]. Similarly, CTLA-4 (CD152) is a transmembrane receptor that competes for the B7 ligand with the co-stimulatory receptor CD28. CTLA-4 is mainly expressed on regulatory T-cells (Tregs) and activated conventional T lymphocytes[94]. Due to its structural similarity to CD28, both receptors can bind to the same B7 ligands, CD80 (B7-1) and CD86 (B7-2); however, CTLA-4 exhibits a significantly higher binding affinity than CD28[95]. CTLA-4 signalling promotes immune tolerance by inhibiting T-cell proliferation and differentiation, inducing cell cycle arrest, and reducing IL-2 production, all of which contribute to exert the inhibitory function in T-cell modulation[96], [97].

By blocking ligand/receptor interactions with ICIs, such as anti-PD-1, anti-PD-L1, and anti-CTLA-4 antibodies, the immune system can overcome the tumour's ability to resist the immune responses, allowing T-cells to kill tumour cells and favouring antitumor cytotoxic activity[98], [99]. The U.S. FDA has approved three categories of ICIs: two PD-1 inhibitors (nivolumab and pembrolizumab), three PD-L1 inhibitors (atezolizumab, durvalumab, and

avelumab), and CTLA-4 inhibitor (Ipilimumab). Recently, ICIs have emerged as a crucial therapeutic strategy in several solid cancers, including non-small cell lung cancer (NSCLC), malignant melanoma, and UC[100]–[106]. To bring forward the use of ICIs to the upfront setting, several therapeutic strategies have been developed. These include the use of ICI as a single treatment, combining chemotherapy with ICIs, and switching to an ICI as a maintenance therapy after non-progression to first-line chemotherapy. ICIs were initially assessed as salvage treatment following progression on platinum-containing chemotherapy and showed favourable efficacy, and safety profiles compared with cytotoxic chemotherapy[107]–[111]. They have now become a standard of care in the first-line treatment for patients who are diagnosed with locally advanced or mUC or are ineligible for platinum-based chemotherapy. Recent treatment guidelines have incorporated atezolizumab and pembrolizumab as first-line therapies for cisplatin-ineligible patients with mUC exhibiting high PD-L1 expression, based on findings from single-arm, phase II studies[2], [112]. In addition, anti-PD-L1 agents like atezolizumab, avelumab, and durvalumab, along with anti-PD-1 agents such as nivolumab and pembrolizumab, are approved for second-line treatment of locally advanced or mUC after platinum-based chemotherapy, regardless of PD-L1 status[108], [110], [113]–[116].

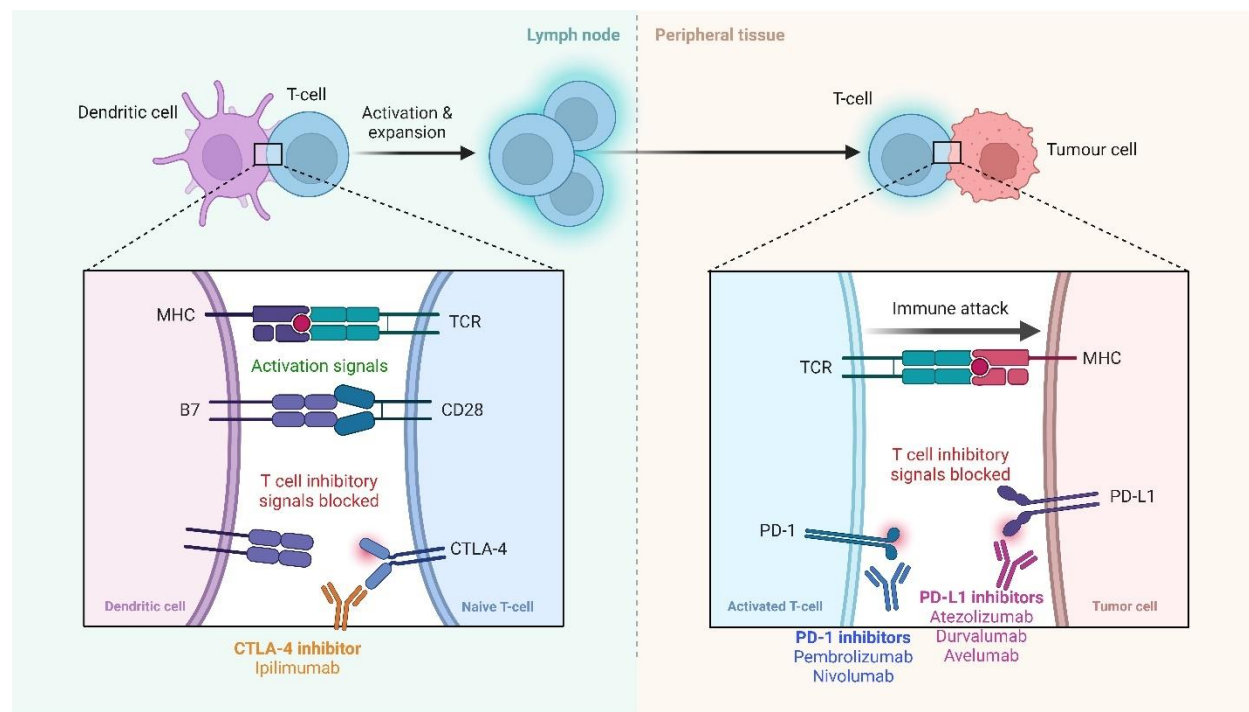


Figure 1. Mechanism of ICIs in Cancer Therapy. The PD-1 and CTLA-4 are inhibitory receptors expressed on T-cell surfaces. When these receptors bind with their corresponding ligands (B7 binding to CTLA-4 and PD-L1 binding to PD-1), T-cell activation is suppressed, allowing tumours to evade immune attack. ICIs (PD-1, PD-L1, and CTLA-4 inhibitors) block these interactions, preventing immune suppression and enabling T-cells to maintain an activated state, thereby enhancing the host's immune response against tumour cells. Created with BioRender.com.

2.3 Atezolizumab: mechanism of action and clinical significance

Atezolizumab (MPDL3280A or Tecentriq) is an Fc-engineered, humanised IgG1 monoclonal antibody developed by Roche Genentech, Inc. It binds to PD-L1, which is expressed on the surface of tumour cells and other PD-L1-expressing cells, such as TICs. By binding to PD-L1, atezolizumab inhibits its interaction with PD-1 receptors, mainly expressed in T-cells, thereby reactivating T-cell-mediated anti-tumour responses [117].

In a phase I trial, atezolizumab was administered intravenously (IV) to mUC patients previously treated with platinum-based chemotherapy or within one year from neoadjuvant chemotherapy, showing a primary endpoint ORR of 26.2% with no dose-limiting toxicity across doses from 0.01 mg/kg to 20 mg/kg. A fixed dose of 1,200 mg every three weeks was administered until disease progression for further clinical investigations. This dosage is equivalent to 17 mg/kg in patients of 70 kg[118], [119]. Following these studies, atezolizumab was granted approval by the U.S. FDA in May 2016 for the treatment of locally advanced or mUC patients who had disease progression during or after platinum-based chemotherapy, or within 12 months of neoadjuvant or adjuvant for such treatment, or are ineligible for platinum-based chemotherapy. This approval was supported by the IMvigor210 single-arm phase II trial (NCT02108652), where 310 platinum-ineligible patients with locally advanced or mUC were treated with 1,200 mg of atezolizumab as first-line therapy every three weeks[2]. This clinical trial reported an ORR of 14.8% (17 (5%) complete responses and 29 (9%) partial responses), as assessed by RECIST version 1.1[120]. Among the responders, 37 patients (11%) showed an ongoing response for a minimum of 6 months, whereas only 6 patients (1%) maintained a response for at least 12 months. The median PFS for all patients was 2.7 months, while the median OS was 15.9 months. Patients were also categorised based on the level of PD-L1 expression on TICs: IC0 (<1% of cells positive for PD-L1, IC1 ($\geq 1\%$ but $< 5\%$), and IC2/3 ($\geq 5\%$). Atezolizumab induced response regardless of PD-L1 expression, including in patients with PD-L1 expression on less than 1% of TICs[118]. The IMvigor130 phase III trial investigated the effectiveness of atezolizumab as the first-line therapy in patients with untreated, locally advanced, or metastatic bladder cancer, including 1,213 patients randomised into three groups: (1) atezolizumab plus platinum-based chemotherapy, (2) atezolizumab as monotherapy, and (3) platinum-based chemotherapy plus placebo. The trial reported a median PFS of 8.2 months in the group receiving atezolizumab with chemotherapy, versus 6.3 months in the group receiving placebo plus chemotherapy group[121]. Recently, atezolizumab has shown promise in treating several types of haematological malignancies and solid tumours, including NSCLC, small-cell lung cancer (SCLC), hepatocellular carcinoma, melanoma, and UC[122], [123]. It is available in two different concentrated solutions of 840 mg/14 mL (60 mg/mL) IV every two weeks and 1200 mg/20 mL (60 mg/mL) IV every three weeks in a single-dose vial, depending on the diagnosis and institution[124].

Although clinical trials evaluating atezolizumab in mUC have demonstrated anti-tumour effectiveness, including prolonged survival rates and durable responses, its efficacy could vary among individuals. Not all eligible mUC patients benefit from atezolizumab, with only

15% of patients achieving a response and just 6% exhibiting a complete response[118]. The factors contributing to the variation in the ICI response among mUC patients are poorly understood. Additionally, a substantial financial burden associated with atezolizumab treatment is costly, exceeding US\$100,000 per patient annually. Moreover, the administration of atezolizumab can be associated with irAEs, which can range from mild to life-threatening and affect different organ systems[125]. The unwanted or unexpected adverse reactions adverse effects depend on the individual drug, combination regimen, and the dose used. Unsurprisingly, concurrent inhibition of immune checkpoints increases the risk of autoimmune toxicities, with higher incidences of high-grade irAEs observed in ICI combination therapy compared to monotherapy[126]. For instance, the irAE incidence in the atezolizumab-containing regimen is about 48%[127], which is higher than the rates observed in atezolizumab monotherapy in the TAIL (10%)[128] and OAK (31%)[129] NSCLC trials. Common reported irAEs are associated with changes in the body's autoimmune system, including the loss of autoimmune tolerance or an increased sensitivity to antigen recognition. This leads to the immune system attacking its own tissues[130]. irAEs are typically organ-specific, with skin-related irAEs being the most prevalent, often presenting as mild itching or rashes, followed by gastrointestinal toxicity, which is usually in the form of diarrhoea and colitis[131], [132]. The third most common are endocrine-related irAEs[133], which include thyroid disorders (both hypothyroidism and hyperthyroidism), pituitary inflammation, and adrenal insufficiency. Commonly reported irAEs also include musculoskeletal toxicity, such as mild joint pain or muscle pain, and ocular toxicity, such as mild dry eye syndrome and uveitis[134]. The majority of patients who experience these irAEs discontinued atezolizumab and have a mortality rate ranging from 10% to 17%. The mortality rate for myocarditis is extremely high, reaching 39.7%. Neurotoxicity is often more severe, with encephalitis and severe myasthenia gravis being the most frequently fatal conditions[135]. As irAEs can affect several organs and systems in the body, careful monitoring and management are needed. Furthermore, most clinical trials did not include cancer patients who had pre-existing autoimmune diseases or chronic infections, such as viral or tuberculosis. Therefore, the incidence of irAEs would be expected to increase further when immunotherapy is administered in real-world clinical settings. In addition, certain symptoms such as cardiovascular or joint toxicity may be overlooked and confused with the cancer-related symptoms of the patients, potentially delaying proper clinical assessment and management of irAEs.

The precise mechanisms underlying irAEs remain unclear, but several hypotheses are proposed. The main concept is that ICIs disrupt immunologic homeostasis and reduce T-cell tolerance, with some cross-reactivity of T-cells between tumour cells and normal tissue[136]. ICIs may increase levels of preexisting autoantibodies, such as antithyroid antibodies, and inflammatory cytokines[137]. Furthermore, ICIs may interfere with the self-protective mechanism in organs, such as in normal pituitary cells or cardiomyocytes, which express CTLA-4 or PD-L1 to protect normal tissue against immune attack[138], [139]. These reasons allow activated T-cells to target healthy tissues, leading to irAEs with autoimmune-like characteristics.

Considering these factors, there is an urgent clinical need to identify predictive biomarkers that could accurately stratify mUC patients who are likely to benefit from atezolizumab. Achieving this objective will enable the administration of promising drugs to eligible patients without delay while also mitigating the risk of irAEs for those who are unlikely to respond to atezolizumab.

2.4 Clinical response biomarkers of ICIs and their challenges

Biomarkers predicting responses to ICIs have been extensively researched and developed. Several biomarkers have been now used in clinical practice, assisting clinicians in selecting patients with suitable therapeutic options. This review highlights FDA-approved biomarkers like PD-L1 expression and TMB, which are commonly used to predict ICI responses and have been extensively reviewed in the literature. In addition, other emerging predictive biomarkers such as TNB and gene expression signatures are also discussed.

2.4.1 PD-L1 expression

As a target of PD-1/PD-L1 antibodies, PD-L1 expression is a logical and commonly examined biomarker for predicting the response to anti-PD-1/PD-L1 therapies. In tumours, PD-L1 expression is induced by interferon-gamma (IFN- γ) secreted from tumour-infiltrating lymphocytes (TILs), suggesting that PD-L1 expression could reflect T-cell infiltration and an anti-tumour immune response. Thus, tumours with elevated PD-L1 expression are associated with a favourable immune response when the PD-1/PD-L1 pathway is blocked. Numerous studies have reported a positive association between PD-L1 expression and the efficacy of PD-1/PD-L1 inhibitors[140]–[143]. In particular, mUC patients with high PD-L1 expression exhibited an increased response rate when treated with pembrolizumab and durvalumab in the second line, or avelumab in the first line of treatment [100], [107], [144].

However, there is currently still insufficient evidence to support the clinical use of PD-L1 expression as a biomarker across all cancer types. The current evidence is limited by inconsistencies observed among different studies. For instance, several studies have found no significant correlation between PD-L1 expression and ICI treatment response[23], [24], and some even indicate that the benefit from ICIs is not only limited to patients with high PD-L1 expression but also to patients with low PD-L1 expression[25]–[29]. Furthermore, the negative predictive value (NPV) of PD-L1 was low, failing to accurately distinguish between responders and non-responders[2], [107], [145]. In addition, the first- and second-line trials with atezolizumab reported contradictory results. While the expression of PD-L1 seems to better predict objective response in a second-line setting with atezolizumab (18% vs. 8%), in a first-line trial there was no difference in efficacy (24% vs. 21%)[2], [109]. Therefore, the role of PD-L1 expression as a biomarker to determine whether patients are likely to respond to ICIs remains controversial.

Indeed, the IHC assay is a simple, cost-effective, and widely available laboratory test that can be easily conducted in hospitals, clinics, and testing labs. However, several challenges are arising. First, there are no standard guidelines for assessing PD-L1 positivity to select

patients for anti-PD-1/PD-L1 therapy. Despite the IHC assay being used as a reference assay to evaluate the membrane expression of PD-L1, there are differences in antibodies for PD-L1 and IHC procedures, including variations in staining techniques, scoring systems, and components tested (tumour cells, TICS, or both), which can lead to inconsistent results[146]. For example, in a study evaluating three different antibodies on NSCLC samples, findings showed only moderate concordance in their positive rates[147]. A systematic review focusing specifically on FDA-approved antibody assays also revealed that, although three of the antibodies were interchangeable, SP142-based assays showed low concordance with the others[148]. Another point of inconsistency among studies on PD-L1 expression is the lack of a standardised threshold to determine whether tumours are classified as PD-L1 positive or negative. A threshold of PD-L1 expression $\geq 1\%$ was associated with a high sensitivity of 73.04% and a low specificity of 39.62%: the reverse was true for a $\geq 5\%$ threshold[146]. Determining the optimal PD-L1 expression cut-off is challenging. It might not be possible to establish a universal PD-L1 threshold across all cancer types, as PD-L1 thresholds not only differ across various cancers but also vary within individual cancer types[149]. Therefore, PD-L1 expression alone remains poorly predictive of ICI response.

2.4.2 TMB

TMB represents the total number of somatic mutations, including single-nucleotide variants (SNVs), nonsense, missense, and splice site mutations, per megabase of genomic sequence analysed from whole genome sequencing (WGS), whole exome sequencing (WES) or next-generation sequencing (NGS) panel in cancer cells. UC is among the most highly mutated cancers[150]. A higher TMB can lead to the increased formation of neoantigens, which are immunogenic peptide sequences recognised by the immune system as foreign, potentially triggering immune-mediated tumour cell death and enhancing anti-tumour responses[151]. Tumours with higher TMB are thus more responsive to immunotherapy and are considered a leading candidate biomarker for predicting immunotherapy response across several cancers, including mUC[2], [100], [152]. Recent studies report that high TMB in tumour tissue of patients with NSCLC, mUC, and melanoma is associated with improved atezolizumab efficacy[2], [153], [154].

Nevertheless, the clinical application and predictive power of TMB need to be further elucidated. There are challenges in using TMB as a biomarker. First, not all mutations produce neoantigens. Only a limited number of mutations can be processed, and expressed on the surface of the major histocompatibility complex (MHC) and recognized by T-cells[155]. Therefore, low TMB is not necessarily associated with not being responsive to immunotherapy, and conversely, high TMB does not always identify those patients most likely to benefit from immunotherapy[156]. Second, there is the availability of many different panels, each varying in targeted genes, number of genes assessed, sequencing depth, and bioinformatics pipeline, which introduces variation and poses a challenge to establishing a unified standardised technique. There are two main methods for assessing TMB: WES and targeted panel-based NGS. Snyder et al.[157] and Rizvi et al.[158] were the first to demonstrate an association between WES-TMB and response to ICIs, suggesting it as a

potential predictive marker for these therapies. The WES-TMB assays usually include nonsynonymous somatic mutations when calculating TMB. This technique is costly and labour-intensive, making it impractical for clinical applications. Instead, targeted gene sequencing is typically used for TMB prediction[159], [160]. The Foundation One CDx assay, developed by Foundation Medicine, Inc., is a U.S. FDA-approved companion diagnostic test to identify patients who may benefit from 28 approved drugs[161]. It is a commercially available targeted NGS panel that analyses 324 genes in cancer patients with solid tumours, considering both synonymous and non-synonymous mutations to enhance assay sensitivity. However, the obvious limitation of this approach is the need to extrapolate the TMB from a limited set of genes. Another challenge is the variability in TMB between cancer types. The U.S. FDA recently approved $\text{TMB} \geq 10$ mutations per megabase (mut/Mb) as a biomarker for administering ICIs across all cancer patients. However, this cutoff has not consistently predicted response to ICIs, raising significant concerns about the FDA's current approval and its clinical utility[162]. Retrospective studies suggest that a cutoff of ≥ 16 mut/Mb is more appropriate than the standard 10 mut/Mb for identifying high TMB associated with atezolizumab response in UC[2].

2.4.3 TNB

TNB is determined by the total number of neoantigens per megabase in the genome region[163], [164]. A patient's TNB can be estimated using bioinformatic algorithms to predict the binding affinity of mutagenic peptides, which are translated from somatic mutations, to MHC class I[165]–[168]. A higher number of TMB is correlated with a higher number of candidate peptides, leading to an increased probability of immunogenic neoantigens being presented by MHCs on the surface of cancer cells, and stimulating host anti-tumour immune response when recognised by T-cells[169]. Thus, TNB has emerged as a biomarker for ICIs, as higher neoantigens are associated with increased responsiveness.[151]. The IMvigor210 trial[2] confirmed the association between high levels of predicted TNB and response to atezolizumab. Compared to TMB, TNB shows a stronger correlation with the response to ICIs and is considered a more effective biomarker, as it directly reflects neoantigen potential[170].

Despite the rapid advancement in immunology and bioinformatics, TNB estimation remains costly and time-consuming. Standard pipelines for predicting neoantigens and determining the cut-off for TNB are yet to be well-defined. In addition, a high intratumoral heterogeneity indicates that the neoantigen might only be present in a subset of tumour cells, thus, the immune response generated against the neoantigen may not be effective against the whole tumour. It is also possible that some patients may lack T-cells capable of recognising the available neoantigen pool, or these T-cells may be present but suppressed by additional or other regulatory mechanisms, leading to inconsistent therapeutic responses among patients. This implies that TNB alone is not sufficient to predict therapeutic outcomes.

In summary, PD-L1 expression, TMB, and TNB are clinical biomarkers that can be used to predict the atezolizumab response and other immunotherapies in several types of cancers,

including UC. However, they each have limitations that need to be considered. In cancers with high PD-L1 expression or a high mutation/neoantigen burden, only a minority of patients achieve clinical responses, with complete responses, where all measurable tumours disappear, are rare. Recent evidence indicates that single biomarkers, such as PD-L1, TMB, and TNB may not be sufficient to accurately predict ICI responses, as some tumours exhibit resistance to ICIs despite the presence of the biomarker and vice versa. Therefore, it is necessary to identify robust biomarkers that can accurately predict the response to ICIs in order to facilitate patient selection for personalised medicine. Ongoing research and progress in biomarker discovery may lead to more precise and reliable predictive biomarkers in the future, ideally with high reproducibility, high accuracy, and cost-effectiveness. The use of PD-L1, TMB, and TNB in combination with other biomarkers, such as gene expression, seems to minimise the limitations of relying on a single biomarker, leading to a more accurate prediction of ICIs for cancer patients.

2.5 Gene expression profiles as predictors of response to ICIs in cancer patients

Cancer is a complex and highly heterogeneous disease, and personalised treatment strategies have become a focal point in the field of oncology. Currently, advancements in high-throughput sequencing and microarray technologies of tumours have been proven as a powerful technique to identify potential ICI response predictors, aiding oncologists in identifying patients who are more likely to benefit from ICIs[171]–[174]. For instance, Ribas et al. [175] identified an IFN- γ gene expression signature associated with improved ORR and longer PFS in metastatic melanoma patients treated with pembrolizumab. IFN- γ is known to have favourable effects on both adaptive and innate anti-cancer immune responses, suggesting that the IFN- γ gene signature has shown promise in predicting ICI response. Furthermore, the Tumour Immune Dysfunction and Exclusion (TIDE) model integrates gene signatures estimating two primary immune evasion mechanisms: T-cell dysfunction and exclusion to predict ICI response, with a high TIDE score indicating non-responder. ROC-AUC shows that TIDE has better predictive performance compared to TMB and PD-L1 for both anti-PD-1 and anti-CTLA4 treatment[176], [177]. Several studies have used data from clinical trials involving mUC to identify gene expression signatures that are associated with either a favourable response or resistance to ICIs[163], [178]–[181]. For example, Chen et al.,[163] revealed 10 prognostic genes, including six OS-favourable genes and four OS-detrimental genes. Among them, six genes were associated with ICI response in mUC patients, with ROC-AUC 0.815 in the training cohort. In addition, Mariathasan et al.,[179] identified Transforming Growth Factor β (TGF- β) pathway gene signatures, including TGFB1 (TGF- β ligand) and TGFBR2 (TGF- β receptor), which were found to have increased mean TGF- β expression in non-responders, and associated with reduced OS and PFS in mUC patients treated with atezolizumab. Furthermore, a recently reported 18-gene tumour inflammation signature (TIS) has shown potential in predicting response to ICIs[182]. Its predictive biomarker was assessed in a prospective cohort of 58 patients with various primary tumours, including 38 lung cancers, 5 melanomas, 10 renal cell carcinomas, 4 UCs, and 1 colon cancer. Across the cohort, patients with high TIS scores had better OS. The TIS has

undergone analytical validation and is currently being investigated in Research Use Only (RUO) and Investigational Use Only (IUO) approaches to further assess its performance as a predictive biomarker.

Overall, gene expression profiling has gained considerable attention with various studies indicating its potential as a predictive biomarker for response to ICIs. Rather than a one-size-fits-all approach, oncologists can tailor treatment plans to individual patients based on the gene expression signature of their tumours. This approach can help better identify patients more likely to benefit from ICIs, maximising the chances of a successful treatment outcome while minimising harmful side effects for those unlikely to respond.

2.6 The overview of ML in cancer drug response prediction

Recent technological advancements in high-throughput sequencing and drug sensitivity screenings have generated an abundance of data yet to be exploited[34], driving the exploration of statistical methods[183] or ML for drug response prediction in cell lines[42]–[44], [184], organoids[185], patient-derived xenografts (PDXs)[45], [186], [187], and even cancer patients[46]. Traditional statistical approaches (e.g., Cox regression) analyse only a small number of predictors, such as tumour size, patient demographics, and biomarker expression levels, within a mathematical framework and make probabilistic assumptions to identify potential predictors of treatment outcomes. These models have generally focused on differences across patient cohorts than focusing on predicting individual patient outcomes. In addition, these methods assume linear relationships, limiting their ability to capture the nonlinear relationships often present in real-world scenarios, which reflects the complexity of biomedical data[188]. In addition, limiting to a few biomarkers are generally insufficient to accurately predict treatment responses for most drug classes[42]; instead, larger sample sizes and multivariate modelling are required for identifying highly predictive and robust biomarkers.

Over the past decade, there has been a significant increase in interest and publications focused on the application of ML techniques in the field of cancer research. ML, a subfield of AI, relies on data-driven, which allows algorithms to learn and identify predictive combinations of features in high dimensional data such as genomic, transcriptomic, and imaging, to accurately predict patient outcomes. ML has been applied to distinguish between myeloid and lymphoblastic acute leukaemia[189], determine the cancer type[190], establish a molecular classification of bladder cancer[191], or define gene expression signatures that can predict survival or metastatic relapse in breast, lung, colorectal, or central nervous system cancers[192]–[195]. In addition, numerous ML have been utilised for drug response prediction, employing diverse inputs, models, and learning strategies. These approaches vary in complexity and interpretability, from well-known linear regression (LR) models to more advanced, sophisticated frameworks like deep neural networks (DNNs)[196]–[198].

Since the predictive performance of ML models relies critically on the availability of high-quality and large sample sizes of training data, creating systematic training datasets for drug response prediction in a clinical trial setting, where numerous drugs are tested across a large

cohort of patients, is both ethically unfeasible and extremely expensive. Therefore, drug response prediction datasets are largely derived from preclinical models like cell lines[42]–[44], [184], organoids[185], and PDXs[45], [186], [187]. The overview cancer pharmacomic modelling using pre-clinical data is described in the original work by[34] (Appendix I). In vitro testing of cancer cell lines across various anti-cancer compounds at different concentrations has provided a good availability of large-scale, systematic datasets with drug sensitivity measurements, (e.g., the half maximal inhibitory concentration (IC50) or the area under the dose-response curve) and paired omics profiles. Large-scaled pharmacogenomic resources, representing approximately 1,000 cell lines across 36 cancer types, such as the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC), have facilitated the development of drug response prediction models leveraging ML. Therefore, there are now numerous studies used cancer cell lines as input for ML-based drug response prediction models[42]–[44], [184]. For instance, Kim et al.[184] explored various ML algorithms to predict pan-cancer cell line responses based on transcriptomic profiles, achieving MCCs under 0.6 in all cases. Other studies also reported predictive performance, with Nguyen et al.[42] reaching MCCs slightly above 0.5 using RF with transcriptomic data, and Naulaerts et al.[43] obtaining MCCs slightly above 0.3 when using RF with genomic profiles. However, while cancer cell lines are easy to culture, they may differ from the primary tumour, and cell culturing may introduce new mutations or alter their characteristics. Unlike cancer cell lines, organoids and PDXs can incorporate interactions with the TME and capture patient-specific tumour heterogeneities in vivo, making them potentially better models of clinical outcomes. Despite these advantages, only few studies have used PDX pharmacogenomic data to train prediction models for drug response[45], [186], [187]. For example, Nguyen et al. used an OMC feature selection with RF to build drug response models for 26 drug-cancer-type binomials using gene expressions, CNVs, and SNVs. They showed that five treatments were predicted with $MCC > 0.4$ in breast cancer, but only two treatments were predicted at this level in colorectal cancer. Tumour biopsies are the best option to observe the characteristics of patient tumours. The TCGA[199] offers a publicly available, large-scale, multi-omics cancer databases, including genomic, epigenomic, transcriptomic, and proteomic data from over 11,000 patients representing 33 types of cancer. Since 2006, TCGA is a joint initiative by the National Cancer Institute (NCI) and the National Human Genome Research Institute, with the objective of advancing understanding of cancer biology and improving cancer treatment and prevention strategies. In TCGA, patient responses to treatments are classified into four categories, complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD), based on tumour progression, assessed through RECIST criteria, an imaging criterion used to evaluate responses in solid tumours[120]. The TCGA dataset is available for download through the Genomic Data Commons (GDC) data portal. In our previous work, we successfully developed ML models to predict chemotherapy response in cancer patients using high-dimensional multi-omics and/or clinicopathological data from TCGA [46]. Bomane et al.[46] focused on predicting paclitaxel response in breast cancer patients, using 6 molecular profiles and 10 ML algorithms. Interestingly, DT and, XGBoost, with OMC feature selection, exhibited highly

predictive models based on the miRNA and DNA methylation profiles. These models could help clinicians to stratify cancer patients likely to respond to treatment in order to propose suitable drugs without delay and avoid harmful side effects for patients unlikely to respond.

2.7 ML in predicting the response to ICIs

ML provides the opportunity for discovering biomarker signatures or ICI prediction scores to predict the response to ICIs using multi-omics data. Typically, these ML approaches involve data collection and selection, feature engineering, model construction, validation, optimisation, and performance evaluation. Due to the limited availability of data, some studies have focused on identifying biomarkers in cancer patients who were not treated with ICIs, a strategy that benefits from the availability of larger sample sizes[48], [50], [177], [200]–[202]. However, these findings require further validation in independent cohorts and might not fully capture ICI-specific biomarkers, potentially limiting the accuracy of predicting responses in ICI-treated cancer patients. Hence, it is crucial to develop robust ML models to identify predictive combinations of biomarkers in ICI-treated patients. Several supervised ML algorithms have been applied in predicting the response of ICIs, including LR, support vector machine (SVM), RF, k-nearest neighbours (KNN), and DNN. For example, Shang et al. developed ML models to predict ICIs response based on DNA methylation immune scores using Lasso regression, support vector machine (SVM), and RF. Among these, the Lasso regression model demonstrated the highest performance. In related work, Sun et al.[203] used an elastic net regularised regression method to develop a computed tomography (CT)-based radiomic model of solid tumours in phase I trial patients treated with anti-PD-1 or anti-PD-L1 immunotherapy. They could identify eight radiomic signatures of CD8+ T-cells, serving as a non-invasive biomarker that accurately predict responses to PD-1 or PD-L1 inhibitors. Park et al.[204] constructed a multivariate LR model built based on radiomics features to predict response to ICIs in mUC. The radiomic-based model could identify high-risk groups associated with poor survival outcomes. Multi-omics-based ML models also show promise for predicting ICI responses by integrating genomics, transcriptomics, epigenomics, proteomics, and radiomics, allowing for a more comprehensive characterisation of the disease. Based on the multi-omics concept, Liu et al.[47] developed an integrative model using clinical, genomic, and transcriptomic data to predict response to anti-PD-1 therapy in metastatic melanoma patients, with validation on smaller independent cohorts. Yang et al.[205] developed a deep learning model based on a multilayer perceptron (MLP) to classify treatment responders and non-responders among advanced NSCLC patients treated with PD-1/PD-L1 inhibitors. By integrating multi-omics data, the model demonstrated a high performance in predicting responses to PD-1/PD-L1 inhibitors. These are just a few examples from growing research that applied ML to predict clinical outcomes of ICIs. Table 1 summarises studies where ML has been used to predict ICI treatment outcomes in various types of cancer. Ongoing research continues to refine treatment strategies and identify potential biomarkers to predict response to ICIs, which in turn further improves clinical outcomes for cancer patients.

Table 1. Summary of various studies that employed ML to predict ICI treatment outcome in cancer patients

Study	Drug	Cancer type	Input feature	Prediction	Algorithm	Ref.
Prat et al., 2017	Nivolumab /pembrolizumab (anti-PD1)	Pancancer	Gene expression profiles	Drug response, PFS	LR	[206]
Charoentong et al., 2017	Anti-PD1/anti-CTLA4	Pancancer	Gene expression profiles	Drug response	RF	[50]
Sun et al., 2018	Anti-PD1/anti-PDL1	Pancancer	CT images and gene expression profiles	Drug response	Elastic-net regularized regression	[203]
Liu et al., 2019	Nivolumab /pembrolizumab (anti-PD1)	metastatic melanoma	Multi-omics profiles	Drug response	LR	[47]
Park et al., 2020	Anti-PD1/anti-PDL1	mUC	CT images	Drug response, OS, PFS	LR	[204]
Johannet et al., 2020	Anti-PD1/anti-CTLA4	Metastatic melanoma	H&E-stained slide images	Drug response, PFS	Deep convolutional neural network	[49]
Hwang et al., 2020	Nivolumab /pembrolizumab (anti-PD1)	Non-small cell lung cancer	Gene expression profiles	Drug response	RF	[207]
Shang et al., 2021	Anti-PD1/anti-PDL1	Non-small cell lung cancer	DNA methylation profiles	Drug response	Lasso regression, RF, SVM	[208]
Benzekry et al., 2021	Anti-PD1/anti-PDL1/anti-CTLA4	Non-small cell lung cancer	Clinical and hematological data	Drug response, OS, PFS	8 ML algorithms	[209]
Yang et al., 2021	Anti-PD1/anti-PDL1	Advanced-stage non-small cell lung cancer	CT images with laboratory and clinical data	Drug response	Deep convolutional neural network	[205]

Kong et al., 2022	Anti-PD1/anti-PDL1	Pancancer	Gene expression profiles	Drug response	Network-based ML approach	[210]
Tabari et al., 2023	Anti-PD1/anti-PDL1/anti-CTLA4	Metastatic melanoma	CT images	Drug response	RF	[211]

2.8 The state-of-the-art ML in drug response prediction

2.8.1 EaSIeR

The TME plays a crucial role in numerous biological processes, particularly in modulating immune response and influencing the effectiveness of ICI therapies[212]. A key aspect is the cellular components of the TME. Different cell types within the TME, especially immune cells, can exert either a pro-tumour or anti-tumour function that regulate cancer progression and response to therapy[213]. Effector T-cells, in particular, play an important role in the anti-tumour immune response, with their phenotype, quantity, and localisation within the TME significantly impact immunotherapy efficacy[214]. In addition, intra- and intercellular regulation of cellular functions also shape the immune response against cancer. External signals received by cells are processed through intracellular signalling pathways, leading to alterations in transcription factor (TF) activity and gene expression. Intracellular regulatory networks of tumour cells are involved in both innate (caused by internal factors such as mutations) and adaptive (caused by external stimuli) mechanisms that enable tumour cells to evade immune surveillance[215]. This evasion can occur through different mechanisms, such as the upregulation in immunological checkpoints[216], a decrease in the production of inflammatory cytokines[217], [218], or impairment of antigen presentation by the MHC[219]. Furthermore, ligand-receptor interactions regulate cell-cell communication within the microenvironment, including interactions among tumour cells, immune cells, and fibroblasts. These interactions are crucial in regulating tumour characteristics and anti-tumour immune responses[218], [220], [221]. Therefore, it is imperative to consider all these aspects to provide a comprehensive description of the TME.

Lapiente-Santana et al. developed the EaSIeR tool[48], which is an R Bioconductor package to assess patients' response to PD-1/PD-L1 inhibitors. By utilising the patients' bulk RNA sequencing data as input, EaSIeR quantifies patient-specific, five real-valued system-based signatures of TME: 11 immune cell types, 14 pathway activities, 118 transcriptional factor activities, 812 ligand-receptor pairs, and 169 cell-cell pairs (Table 2). These signatures comprehensively characterised tumours, illustrating how the complex cellular composition of TMEs, intra- and intercellular mechanisms influence immune response and immunotherapy efficacy. The cancer-specific elastic net regularized multi-task linear regression (RMTLR) models, pretrained on 7,550 TCGA samples across 18 solid cancers, were used to predict 10 tasks of immune response scores: Cytolytic activity (CYT), Ock immune signature (Ock_IS), Roh immune score (Roh_IS), Chemokine signature (Chemokine),

Davoli immune signature (Davoli_IS), IFN γ signature (IFN γ), Expanded immune signature (Ayer_expIS), T-cell inflamed signature (Tcell_inflamed), Repressed immune resistance (RIR), Tertiary Lymphoid structure signature (TLS). These scores have been experimentally validated in the literature (Table 3), reflecting patient response to PD-1/PD-L1 inhibitors, with higher scores indicating a higher likelihood of response.

Table 2. Five descriptive signatures of the tumour microenvironment from the EaSleR tool

System-based signatures of TME	Prior knowledges	Tools	Number of features (N = 1,125)
Immune cell quantification	Finotello et al., 2019	quanTIseq	11 cell types
Pathway activity	Holland, Szalai, and Saez-Rodriguez et al., 2020; Schubert et al., 2018	PROGENy	14 pathways
Transcription Factor activity	Garcia-Alonso et al., 2019	DoRothEA	118 TFs
Ligand-receptor pairs	Ramilowski et al., 2015; Barretina et al., 2012	LR pairs weight (the minimum of the log2(TPM+1) expression of the ligand and receptor)	813 LR pairs
Cell-cell interaction	Ramilowski et al., 2015; Barretina et al., 2012	The number of active LR pairs for each CC pair	169 CC pairs

Table 3. 10 published signatures of the immune response

No.	Signatures of immune response	Studies	Explanation
1.	Cytolytic activity (CYT)	Rooney et al., 2015	The level of two cytolytic effectors, granzyme A and perforin, which are overexpressed upon CD8+ T-cell activation
2.	Ock immune signature (Ock_IS)	Ock et al., 2017	The expression of 105 genes associated with the response to immunotherapy with the MAGE-A3 antigen
3.	Roh immune score (Roh_IS)	Roh et al., 2017	A set of genes involved in immune activation in relation to tumour rejection
4.	Chemokine signature (chemokine)	Messina et al., 2012	A gene set associated with inflammation and immunity, which is

			able to predict host immune reaction and the formation of tumour-localized lymphoid structure.
5.	Davoli immune signature (Davoli_IS)	Davoli et al., 2017	The expression of cytotoxic CD8+ T-cell and natural killer (NK) cell markers
6.	IFN γ signature (IFN γ)	McClanahan et al., 2017	Genes able to separate responders and non-responders in melanoma
7.	Expanded immune signature (Ayer_expIS)	McClanahan et al., 2017	Genes highly correlated with IFN- γ signature genes; this new set included all immune-related genes.
8.	T-cell inflamed signature (Tcell_inflamed)	McClanahan et al., 2017	The joint potential of IFN- γ and T-cell-associated inflammatory genes in predicting response to PD-1 blockade
9.	Repressed immune resistance (RIR)	Jerby-Arnon et al., 2018	Combining a set of gene signatures associated with T-cell exclusion, post-treatment, and functional resistance.
10.	Tertiary lymphoid structure signature (TLS)	Cabrita et al., 2020	Differentially expressed genes in tumours with TLS

2.8.2 JADBio autoML

Automated ML (AutoML) is the process of applying ML models to real-world problems by automating key aspects of ML pipeline, including data preprocessing, feature selection, model selection, and hyperparameter tuning. By minimising the need for manual intervention, AutoML enables users with varying levels of expertise to develop ML models. It benefits researchers by reducing the time and technical barrier to deploying sophisticated algorithms, making advanced data analysis more accessible[222]. Tsamardinos et al.[55] developed JADBio, an AutoML tool, specialising in bioinformatic applications and translational medicine. It designed for analysing high-dimensional omics data with limited sample sizes. By applying various ML algorithms and optimisation techniques, JADBio automates preprocessing, feature selection, model selection, and hyperparameter tuning (see Table 4), selecting the best combinations to provide a predictive model with performance estimates. Users can avoid the time-consuming and complex process of manual model development by simply uploading input data and specifying modelling objectives. Furthermore, JADBio can analyses various types of outcomes, including binary and multi-class classifications, right-censored time-to-event data for survival analysis, and quantitative outcomes for regression tasks. It can process both nominal and continuous features, as well as medical images (e.g., histopathological slides and X-rays) and medical signals (e.g., electrocardiograms and ballistocardiograms).

However, within the scope of our work, we focused only on regression task using tumour gene expression profiles and clinical data. JADBio accepts input as a two-dimensional table, with rows representing patient samples and columns for gene and clinical features, one of

which specifies the outcome of interest (e.g., drug response). JADBio then analyses this dataset fully automatically, applying standardisation and mean imputation to continuous variables, and mode imputation to categorical ones, while removing features with constant values. For feature selection, JADBio used the Statistical Equivalent Signatures (SES), the Least Absolute Shrinkage and Selection Operator (LASSO), and univariate methods. During model development, JADBio employs ridge linear regression, support vector regression (SVR) with linear, polynomial, and Gaussian kernels, DT, and RF. Configurations are obtained by taking all combinations of one modelling algorithm, one feature selection algorithm, and hyperparameter values (imputation is included, if applicable). In brief, JADBio analysis (Figure 2) follows a four-stage process that includes:

- i. **Algorithm and hyper-parameter space selection (AHPS) system:** the process involves selecting the combinations of feature selection techniques, algorithms, and hyperparameter values to generate a configuration space to explore. It also determines the hyperparameter values of the generalised cross-validation (CV) algorithm, i.e., the protocol used to estimate the performance of each configuration. These hyperparameters include the value of R repeated, K -fold, and N -incomplete CV. Once the AHPS determines the configurations to be tested, JADBio uses grid search for hyperparameter tuning and employs early dropping (ED) and early stopping (ES) techniques to reduce computations without compromising quality. ED stops cross-validating configuration that is unlikely to be the winning one, while ES stops repeating CV with different partitions early when there is no improvement in confidence intervals (CIs) across CV repeats.
- ii. **Configuration generator (CG):** it instantiates the configurations to be evaluated.
- iii. **Configuration evaluator protocol (CEP):** it uses the generalised CV to identify the winning configurations. Generalised CV is a stratified, R repeated, K -fold, N -incomplete CV that accepts the values of R , K , and N as hyperparameters, decided by the AHPS system. It repeatedly applies each configuration on different partitions of the data to train a model and estimate its performance on the remaining data. The model's predictions on the held-out fold are stored as the out-of-sample predictions for each configuration. The performance estimate of the configuration is calculated as the mean performance across all folds. Then CEP finds the best ML model with the best performance over all tried configurations and its confidence interval.
- iv. **Performance estimator (PE) and final model generator:** it produces the final model and its performance. The final model is built using all available samples, employing the winning configuration as determined by the CEP, given the final selection of features. In this way, JADBio does not directly assess the performance of the final model. Instead, it evaluates the performance of the configuration used to build the final model by using a generalised CV from the CE, which has already estimated the out-of-sample performance of the models produced by the winning configuration.

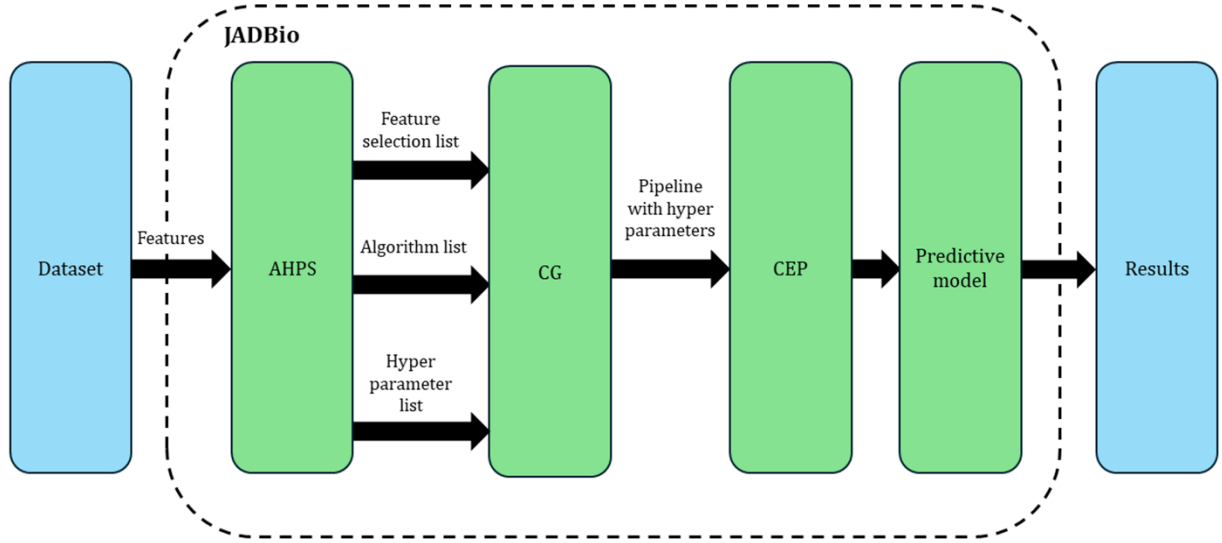


Figure 2. The workflow of the JADBio system. Initially, the dataset are input into the AHPS, which then feeds the CG with a list of feature selection techniques, algorithms, and the range of hyperparameter values. The CEP subsequently identifies the best ML model with the best performance.

To reduce the bias from testing multiple configurations, JADBio applied bootstrap bias corrected-CV (BBC-CV) to adjust prediction performances for each configurations tried. This process uses a matrix of out-of-sample predictions across all configurations during CV, with each row representing an out-of-sample prediction, each column a configuration, along with corresponding true labels. In brief, the procedure of BBC-CV involves the following steps: for each bootstrap iteration, a subset of out-of-sample predictions was randomly sampled, with the replacement, for each configuration. This preserves the sample size of the original dataset, regardless of the potential presence of duplicate samples. The best configuration is selected based on the desired evaluation metric score (e.g., the highest MCC for classification or the lowest mean squared error (MSE) for regression). Then, the model performance is computed on the remaining samples using this best configuration. BBC-CV outputs the probability distribution of the expected performances along with 95% CIs. The final BBC-CV score is estimated as the average of these performance across all bootstrap iterations. Therefore, the performance estimation of BBC-CV can be applied with minimal computation, as it only requires bootstrapping the out-of-sample predictions without training new models.

Table 4. ML algorithms and optimisation techniques, including preprocessing, feature selection, model selection, and hyperparameter tuning used in JADBio

Steps	Methods	Hyperparameters	min	max	
Data-preprocessing	Mean/mode imputation	NA	NA	NA	
	Constant removal	NA	NA	NA	
	Standardisation	NA	NA	NA	
Feature selection	SES	maxK	2	3	
	LASSO	alpha	0.01	0.1	
	Univariate	NA	NA	NA	
Modelling	Ridge linear regression	lamda	0.0001	100	
	Linear support vector regression	epsilon	0.0001	0.5	
		cost	0.001	100	
	Polynomial support vector regression	epsilon	0.0001	0.5	
		cost	0.001	100	
		degree	2	4	
	gamma	0.001	100		
		Gaussian support vector regression	epsilon	0.0001	0.5
			cost	0.001	100
	gamma		0.001	100	
	Random forest	minLeafSize	3	9	
		nTrees	100	1000	
	Decision tree	minLeafSize	3	9	
		alpha	0.01	0.1	
Evaluation protocol	Generalised cross-validation	R repeated	1	20	
		K-fold	1	10	
		N-incomplete CV	1	K	
		Early dropping	0	1	
		Early stopping	0	1	

2.9 ML fundamentals

This section provides guidance for researchers, including non-experts in ML, on how to develop robust ML workflows tailored to their specific applications. It aims to ensure that all important steps are followed, and decisions are made with a solid understanding of fundamental ML principles to avoid common pitfalls. The section begins with an overview of supervised ML workflow architecture, followed by detailed discussions on data preprocessing, and model selection, which includes algorithms and methods for training, validation, and evaluation.

2.9.1 The general architecture of the supervised ML workflow

ML models can be categorised into four types based on the input data used: supervised, unsupervised, semi-supervised, and reinforcement learning. In precision oncology, supervised learning models represent one of the most commonly used ML approaches, applied in cancer diagnosis, treatment response prediction, and cancer survival prediction (e.g., OS or PFS). Supervised learning is where the input data is paired with the corresponding labelled output, allowing the model to learn from the provided examples and be guided by the known correct answers between the input and output to make accurate predictions or classifications on new, unlabelled data. There are two varieties of supervised learning algorithms: classification and regression algorithms. Classification algorithms are probability-based, meaning the outcome is the category to which the algorithm finds the highest probability that the sample belongs to it. Regression algorithms, in contrast, estimate the outcome of samples as continuous values.

A supervised ML workflow typically includes data collection and pre-processing, model training, testing, and evaluation. To ensure that the ML project is successful, it is important to pay attention to these steps carefully. As depicted in Figure 3, an ML workflow begins with the collection of data, including electronic medical records (EMRs), imaging, and omics data. EMR data consists of patient-level information such as patient demographics, medical and social history, medications, diagnoses, procedures, vitals, and laboratory results. These data are standardised across patients within a healthcare system and consistent across systems using the same EMR provider. For patients who continue within a single healthcare system, the EMR provides a comprehensive and longitudinal view on their health progressions. Imaging data, including CT/magnetic resonance imaging (MRI) scans or pathology slides, provides information about a patient's condition and is particularly relevant in cancer diagnosis and treatment. Meanwhile, omics data, including genomics, transcriptomics, epigenomics, and proteomics, has become increasingly available due to the decreasing costs of sequencing technologies and the increased capacity for data storage[223]. Despite these advancements, the availability of high-quality, well-labelled, and sufficiently large datasets remains a significant challenge. Data is like fuel for ML models, no matter how sophisticated the model, it cannot perform well without the reliability and relevance of the data it processes.

Once data is collected, it must be properly pre-processed for it to be useful as part of an ML workflow. This process involves data cleaning (handling missing values or outliers), data normalisation (ensuring a common scale for all features), encoding categorical features (transforming nominal variables into a format that ML algorithms can understand), and feature selection (selecting the most relevant features that facilitate the model prediction). To build a predictive model, it is important to ensure that the model is not overfitting during the training process so that better model generalisation is achieved in the testing phase and future use of this model. Common approaches to prevent model overfitting include CV. The process involves splitting the data into training and testing sets. The training set, where the input data is paired with the corresponding labelled output, is input into ML algorithms for

learning and modelling. The algorithm will identify patterns that best describe the output, which then be used to predict the labels for new, unseen data. Hence, the predicted outcomes are obtained by learning the state of input features. Examples of typical supervised learning include linear/logistic regression, SVC, KNN, DT, RF, XGBoost, and light gradient boosting machine (LGBM).

Following training and testing of the model, model evaluation is an important step that uses various evaluation metrics to understand an ML model's ability to make accurate predictions on new, unseen data. This involves comparing the model's predictions with the true labels. Different evaluation metrics are used depending on the type of model. Classification models are evaluated using metrics derived from the confusion matrix, which categorises outcomes into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), to calculate metrics such as MCC, accuracy, precision, recall, and F1-score. Regression models are assessed using metrics such as mean absolute error, mean squared error, and R-square.

Once we have our model, the final phase in the ML workflow is the model deployment, where the trained *in-silico* model is developed and applied to test new, unseen sample data in real-world settings to verify the learning results.

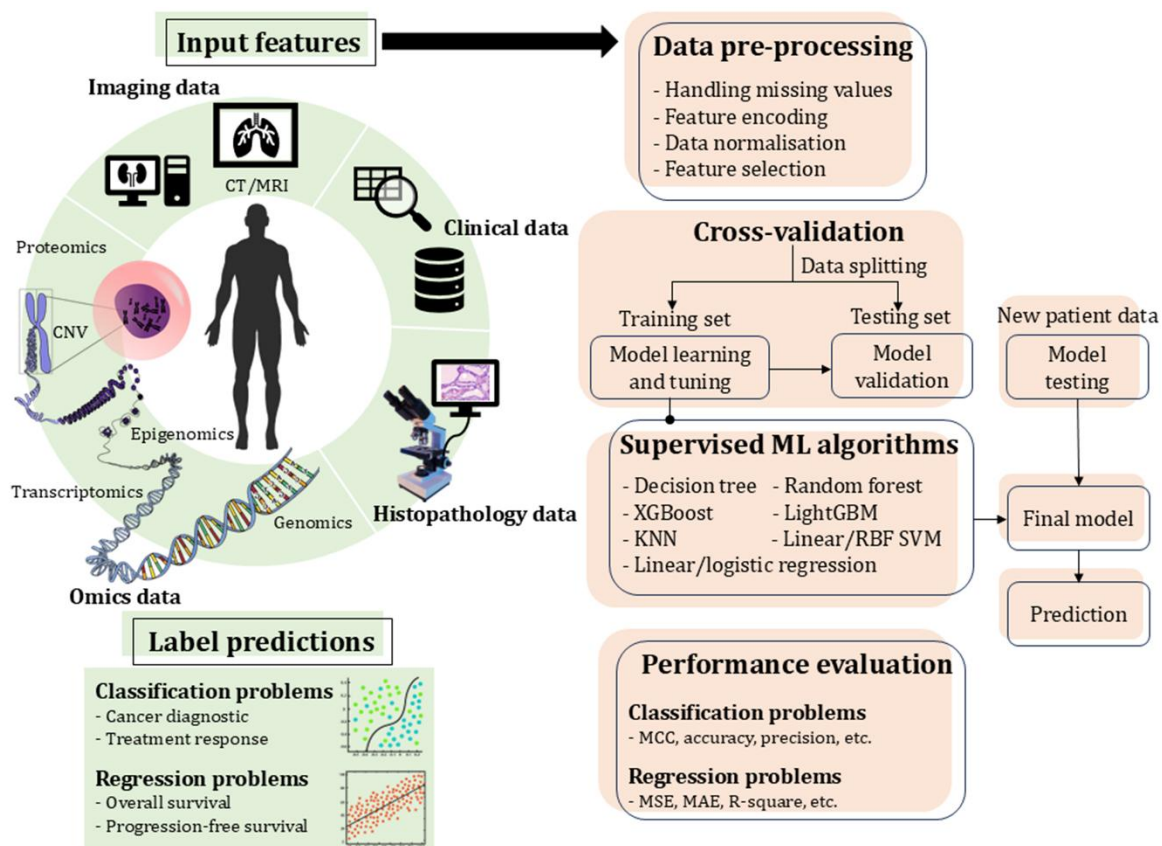


Figure 3. Schematic of ML workflow.

2.9.2 The overview of data pre-processing for ML

In recent years, the application of ML to predict disease outcomes using omics and other medical data has exponentially risen in the field of precision oncology. While ML models generally improve with larger datasets, the quality of the data is equally crucial. High-quality data enhances the ability to identify patterns, extract insights, and make accurate predictions, whereas poor-quality data can degrade model performance. However, real-world datasets used in ML are often messy, containing missing values, inconsistencies, out-of-range values, and duplicates caused by human and measurement errors. Addressing these issues through data pre-processing is essential to build robust ML models. Data preprocessing involves preparing and transforming raw data into a format suitable for training, with steps such as handling missing values, encoding categorical variables, and scaling features. These techniques are key for ensuring data quality and usable format for the ML algorithms. This section outlines guidelines for data pre-processing, including what it is, its benefits, steps, and examples.

2.9.2.1 Handling missing values

The primary issue with omics and medical data involves the presence of missing values, which are typically represented as NaNs or blank. Most ML algorithms require complete data, making it important to address missing values before modelling. Common strategies include removing samples or features with missing values or imputing missing values using a descriptive statistic (e.g., mean, median, or mode) or more advanced methods. Failure to address missing values can lead to inaccurate and flawed conclusions from the data. The following are key approaches to handling missing data:

i. Dropping samples or features

This approach involves discarding samples or features with at least one missing value presented, or with a certain percentage (e.g., 20%, 50%, 70%) of values missing. While this prevents introducing excessive noise into the dataset, can lead to significant data loss. It is recommended only when the dataset has a sufficient number of samples or features.

ii. Statistical imputations

Missing values are replaced with calculated statistics such as the mean, median, or mode. Mean and median for numerical data, and mode for categorical data. In many studies, statistical imputation is commonly used due to its simplicity and ability to minimise information loss. However, they may also introduce bias into the dataset.

iii. Model-based imputations

Earlier approaches for handling missing values do not use the correlation advantage between the feature with the missing value and other features. In contrast, this approach employs predictive models to impute missing values using information from other features in the dataset. Depending on whether the missing feature is categorical or continuous, regression or classification models are applied. While this

method is effective when clear relationships exist among features, it requires advanced expertise and can be computationally intensive.

2.9.2.2 Categorical feature encoding

Since the majority of ML algorithms are primarily designed to operate with numerical data, categorical variables, which can be either nominal or ordinal, need to be converted into numerical formats. This process, known as feature encoding, enables ML models to effectively interpret and utilise categorical information during training.

For nominal variables, one-hot encoding is a widely used method that transforms each category into a binary format. This approach generates new binary columns for each category in the original variable, where a value of 1 indicates the presence of the category and 0 its absence. Unlike assigning numerical values (e.g., male = 0, female = 1), which could imply unintended ranking and bias predictions, one-hot encoding treats categories independently, avoiding such issues. However, it significantly increases dimensionality by adding a column for each category, leading to sparse data with many zero values. This can complicate models, increase training time, and lead to overfitting, especially when the variable has many categories, and the sample size is small. Therefore, one-hot encoding should be applied cautiously, and alternative methods, such as binary encoding, may be more suitable in such cases.

For ordinal variables, label encoding assigns a unique integer to each category based on its meaningful sequential order, typically starting at 0 and increasing by 1 for each category. For instance, a variable like size with categories small, medium, and large can be encoded as 0, 1, and 2, reflecting the ranking small < medium < large. This method keeps the feature space compact, unlike one-hot encoding, which can significantly increase the dataset's dimensionality.

2.9.2.3 Data scaling

Data scaling is a crucial pre-processing step in ML, particularly when datasets include features with varying scales and units. It transforms feature values to a similar scale, ensuring all features contribute equally to the model and often leading to significantly improved performance compared to models trained on unscaled data. In biomedical research, gene expression data from high-throughput sequencing and microarray technologies are widely used. Their workflows typically involve normalisation, which aims to correct systematic variations in measurements before proceeding to data analysis[224]. However, gene expression normalisation and data scaling serve different purposes. While normalisation addresses systematic biases, normalised gene expression data still require additional scaling to optimise suitability for ML model training. Training models on scaled gene expression data consistently yields higher predictive accuracy compared to unscaled data, highlighting the importance of data scaling in ML workflows[225]. Two data scaling algorithms are widely used: the Min-max algorithm and the Z-score algorithm.

i. Min-max algorithm

The min-max algorithm applies a linear transformation to the original data for data scaling. Let x_{min} and x_{max} represents the minimum and the maximum values of a variable within the training datasets. The min-max algorithm maps a given value, x , of this variable to a normalised value, x' , within the range $[0, 1]$ using the following formula:

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

However, if values in testing datasets fall outside of the range $[x_{min}, x_{max}]$ of the training data, the resulting scaled values will be out of the bounds of $[0, 1]$ interval, potentially leading to issues in certain applications.

ii. Z-score algorithm

The Z-score algorithm rescales features to a standard normal distribution. Each feature value is transformed by subtracting the mean (μ) of the feature and dividing by its standard deviation (SD) (σ) in the training samples, resulting in a mean of 0 and an SD of 1. The z-score of a sample i (x'_i) is calculated as follows:

$$x'_i = \frac{x_i - \mu}{\sigma}$$

Z-scores indicate how many SDs a data point is away from the mean. However, this algorithm does not map the original data into an interval and is highly sensitive to outliers, which can skew scaling, especially in small sample sizes. In addition, the mean and SD derived from limited data may not accurately approximate their true values, leading to poorly scaled features and reduced model performance.

2.9.2.4 Handling data imbalance in ML

Data imbalance is a common challenge in medical datasets for diagnosing and predicting malignant tumours[226], heart disease[227], cardiovascular disease[228], and pregnancy-related conditions[229]. It arises when the majority class significantly outnumbers the minority class, leading to biased predictive models. In such cases, models may achieve high overall accuracy while failing to identify minority-class samples [230], [231]. For instance, in a diagnostic dataset of 1,000 cancer patients, only 10 may be diagnosed with cancer, while the rest are classified as healthy. A model that predicts all samples as healthy would achieve 99% accuracy but fail to identify cancer cases. Therefore, addressing this imbalance is crucial for ensuring that ML models perform well across all classes. The following section discusses common data balancing techniques to balance the class distribution and mitigate a skewed class distribution in the classification process.

2.9.2.4.1 Oversampling

Oversampling is a preprocessing technique used to address class imbalances before model training. It increases the representation of the minority class by either duplicating existing instances or generating synthetic data points. This helps prevent algorithms from disregarding underrepresented yet important classes and reduces the cost of FNs. While oversampling is relatively easy to implement, it carries a risk of overfitting, especially if the duplicated or synthetic samples do not accurately reflect the true data distribution. Since some methods simply replicate existing samples, the model may memorize these instances rather than learning generalizable patterns. Moreover, oversampling can be computationally demanding, particularly for large datasets[232]. Common oversampling techniques include:

- i. **Random over-sampling:** This approach mitigates class imbalance by duplicating instances from the minority class at random[233], [234].
- ii. **Synthetic minority over-sampling technique (SMOTE):** New synthetic samples are produced by interpolating between existing minority class instances and their nearest neighbours[235]–[237].
- iii. **Adaptive synthetic sampling (ADASYN):** This method prioritises the generation of synthetic samples for harder-to-learn minority class instances, thereby improving class balance[238], [239].
- iv. **Borderline-SMOTE (BSMOTE):** Synthetic samples are generated from minority class instances located near decision boundaries[240].
- v. **KMeans SMOTE (KMSMOTE):** This technique integrates KMeans clustering with SMOTE to create synthetic samples that better represent the minority class distribution[241].
- vi. **Support vector machine SMOTE (SVMSMOTE):** This method utilizes Support Vector Machines to generate synthetic samples near the minority class decision boundaries[242], [243].

2.9.2.4.2 Undersampling

Undersampling is a resampling technique used to address data imbalance by reducing the number of instances in the majority class to create a more balanced dataset. This prevents the model from being overwhelmed by the majority class and ensures that the minority class receives equal importance in training. While undersampling can simplify the dataset and speed up training, it can also lead to information loss, which may hinder the model's ability to learn crucial patterns. It is particularly useful when dealing with very large datasets, but may not be ideal for small datasets due to potential loss of valuable information[244], [245]. Common undersampling techniques include:

- i. **Random under-sampling:** Addresses class imbalance by randomly discarding samples from the majority class[233], [246].
- ii. **TomekLinks:** Eliminates majority class samples identified as Tomek Links, which are considered to be boundary or noisy instances[247], [248].
- iii. **NearMiss:** Selects majority class samples that are closest to minority class examples, ensuring a more balanced dataset[248], [249].

- iv. **Cluster centroids (ClusterC):** Applies k-means clustering to group majority class samples and balances the dataset using the cluster centroids[250], [251].
- v. **Edited nearest neighbours (ENN):** Utilises the KNN approach to remove misclassified majority class samples[235], [249].
- vi. **Condensed nearest neighbour (CNN):** Reduces the majority class samples by iteratively partitioning the dataset into subsets using the nearest neighbour algorithm[252].
- vii. **Instance hardness threshold (IHT):** Filters out samples that are difficult to classify based on their instance hardness score[253], [254].
- viii. **One-sided selection (OSS):** Combines 1-nearest neighbour (1-NN) with Tomek Links to retain minority class samples and remove boundary majority samples[235], [240].

2.9.2.4.3 Combining resampling

Combining oversampling and undersampling can help mitigate the drawbacks of each method while leveraging their advantages. The goal is to strike a balance between preserving information from both the majority and minority classes and addressing class imbalance. This approach is particularly useful when choosing between oversampling or undersampling is not clear-cut. Depending on the problem, experimenting with different combinations could help identify the most effective strategy. Examples of such techniques include:

- i. **SMOTETomek:** Integrates the SMOTE and TomekLinks techniques by first synthesising new samples for the minority class and subsequently eliminating noisy and borderline samples[247], [248], [255].
- ii. **SMOTEENN:** Combines SMOTE with the ENN method, generating synthetic minority class samples before filtering out misclassified instances[240], [248], [255].

2.9.2.4.4 Algorithm-level techniques

In addition to resampling techniques, algorithm-level methods are another set of strategies used to address data imbalance in ML. These techniques involve adjusting the algorithms themselves to handle imbalanced datasets more effectively. They are particularly useful when working with imbalanced data without modifying the dataset itself, when resampling techniques are not suitable due to dataset size limitations, or concerns about information loss, want to retain the original data distribution while ensuring the model adapts to the imbalance, or if you prefer algorithms that inherently handle class imbalance without additional preprocessing steps, these techniques are a good choice. Here are some key algorithm-level techniques:

- i. **Class weighting:** Many ML algorithms allow assigning different weights to classes, encouraging models to focus more on the minority class by giving it higher weights. This approach avoids altering the dataset size, minimizing the risk of losing valuable information or generating synthetic data[256].
- ii. **Cost-sensitive learning:** Cost-sensitive learning explicitly assigns different misclassification costs to each class. The model is trained to minimize the overall cost,

encouraging better classification of minority class instances, even at the expense of more false positives in the majority class[257].

- iii. **Threshold adjustment:** By modifying the classification threshold, you can balance precision and recall in binary classification. Lowering the threshold increases recall, making the model more sensitive to the minority class, at the expense of precision[258], [259].

2.9.2.5 Data splitting

Proper data splitting is crucial in ML to ensure the reliability and generalisability of a model. The objective is to estimate the performance of ML models when they are used to make predictions on new, unseen data that are not used during the training process, simulating real-world scenarios. Typically, the dataset is divided into two subsets: the training set and the testing set. The training set, containing input features and their corresponding labels, is used to train the model. Through this process, the model identifies patterns in the data and applies these learned patterns to make predictions on the testing set, which is reserved for evaluating the model's performance to avoid any data leakage.

The most common splitting method is random sampling, where the dataset is shuffled and randomly assigned to training or test sets according to predefined ratios. For class-balanced datasets, this approach ensures an unbiased split. However, with class-imbalanced datasets, where certain classes have significantly fewer samples than others, random sampling can introduce bias. In such cases, stratified splitting is preferred as it preserves the original class distribution across the splits. This helps the model recognise patterns and make predictions for all classes, reduces the risks of overfitting or underfitting, and enhances its robustness and reliability[260], [261].

2.9.3 Binary classification by regression algorithms

Binary classification is a type of supervised learning in which an algorithm predicts an outcome that belongs to one of two possible classes based on input features. For instance, a spam filter represents a classic example of binary classification, where incoming emails are categorised as either spam or legitimate. To train such a model, a labelled dataset of emails is provided, where each email is indicated as spam or not. Using this labelled data, the model learns patterns and features associated with spam, such as specific keywords, phrases, or formatting styles. Once trained, the model can classify new incoming emails by analysing their features and predicting their label. A well-trained spam filter enhances user experience by minimising unwanted emails while ensuring legitimate emails are not misclassified.

In oncology, binary classification can potentially be helpful in categorising tumours based on their characteristics, including tissue of origin, histopathology, genetic alterations, or clinical outcomes, such as drug responses [46], [262]–[266]. For instance, a binary classifier can be trained on large datasets to classify whether a breast cancer patient will respond to paclitaxel or not (responder vs. non-responder) using clinical information and molecular profiles[46]. Accurate classification can assist healthcare professionals in guiding treatment decisions,

potentially sparing non-responders from harmful side effects while considering alternative treatments without delay.

Binary classification offers several advantages, including simplicity, computational efficiency, and ease of interpretation. This approach is highly useful in healthcare, where many decisions involve binary outcomes, such as disease presence or drug response. However, its applicability is limited to tasks with only two outcomes. Some cases may have more than two classes. For example, drug response assessed by RECIST criteria includes four categories: CR, PR, SD, and PD. To fit a binary classification, similar categories must be grouped (e.g., CR and PR as responders, SD and PD as non-responders), leading to information-wasting.

An alternative approach is to use regression algorithms, which predict continuous values and can be adapted for classification tasks by applying a decision threshold to assign class labels. Samples with predicted values above the threshold are assigned to one class, while those below the threshold are assigned to another. One key benefit of this approach is that a regression algorithm can learn from all patient response labels (CR, PR, SD, and PD) during training without grouping them into binary categories (CR and PR as responders, SD and PD as non-responders). In addition, applying thresholds to predicted values allows regression models to be evaluated using binary classification metrics such as MCC, precision, and recall, aligning with the objectives of binary classification tasks while mitigating the information loss in binary grouping.

Traditional classification models trained on imbalanced datasets often become biased toward the majority class, leading to high overall accuracy but poor recognition of the minority class. In contrast, regression models are less susceptible to this issue, as they predict a continuous outcome rather than relying solely on majority-class frequency during training. This characteristic reduces the risk of bias and improves the model's ability to learn from all available response labels.

2.9.4 Supervised ML algorithms

2.9.4.1 Linear/logistic regression

LR is a widely used statistical modelling employed to identify the linear relationship between a dependent variable (outcome) and one or more independent variables (features). When there is only one independent variable and one dependent variable, it is known as simple regression. As the number of independent variables increases, it is referred to as multiple regression. LR aims to find the best-fitting straight hyperplane to the training data (a straight line when using only two variables) that minimises the sum of squared residuals (SSR). The LR assumes a parametric linear function for the model as follows:

$$f(X) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

Where β_0 is the y-intercept, $\beta_0, \beta_1, \dots, \beta_n$ represents the coefficients for the respective features $x_1, x_2, \dots, x_n$.

The main advantage of LR lies in its simplicity, making it accessible to a wide range of users. The coefficients within an LR model have clear interpretations, allowing for an easy understanding of the impact of each independent variable on the dependent variable. In addition, it exhibits high computational efficiency as it does not require complicated calculations and runs predictions fast, even with large datasets. However, LR assumes a linear relationship between the dependent and independent variables, which can be overly simplistic. Real-world datasets often exhibit complex and nonlinear relationships that cannot be adequately captured by LR. In such cases, using LR may result in poor predictive performance. Moreover, outliers can have a large effect on the output, as the best-fit line tries to minimise the SSR for the outlier points as well[267].

2.9.4.2 Decision tree

DT is a non-parametric supervised ML algorithm proposed by Leo Breiman[268] as an alternative to linear models for both classification and regression tasks. This algorithm utilises a tree-like structure with nodes and leaf nodes to learn non-linear relationships from multiple input features. The root node initialises the process, branching into internal nodes, each representing decisions based on specific features, and leaf nodes, which indicate the final decisions with a prediction or class label. To build a DT, the feature with the least impurity (measured by MSE in regression, Gini in classification) is selected as the root. Impurity reflects the goodness of a split. The tree grows by repeating this process at each node until reaching a stopping criterion such as a predefined depth of the tree, a minimum number of samples in a leaf node, or other criteria. Predictions are made by traversing the tree from the root to the relevant leaf node. In regression, the final prediction is derived from the mean or weighted average of the values in the leaf node. In classification, however, predictions are based on the majority class within the leaf node.

The strengths of DT include simplicity, making it accessible even to those without a technical background. The tree-like structure provides clear and interpretable rules, with each path from the root to a leaf node indicating a set of conditions leading to a specific outcome. This makes it straightforward to capture and understand non-linear relationships. DT is robust to outliers and efficient in both training and prediction, as it can be built and evaluated quickly, even with large datasets. However, slight changes in the data can produce significantly different trees and DTs tend to overfit, especially with high-dimensional data. As the splitting process progresses with a large number of features, the trees become increasingly complex. This overfitting can limit the tree's ability to generalise to new data. Pruning strategies, such as setting limits on tree depth or minimum samples per leaf node, can be applied to simplify DT and reduce overfitting, though it requires careful tuning[269], [270].

2.9.4.3 Random forest

RF is a supervised learning algorithm that uses an ensemble learning method for classification and regression. It operates by building multiple DTs at training and combines predictions from the individual trees. This could reduce variance and create more robust and accurate final predictions than a single DT. To build an RF model, the method begins with

generating multiple bootstrap samples from the training data, where subsets are randomly selected with replacement, resulting in multiple training datasets of the same sample size as the original, but with different samples. At each node of each tree, a random subset of features is considered for splitting. This adds additional randomness and helps decorrelate the trees, making the final model more generalisable. Each tree is grown until a predefined depth is reached or until further splitting does not significantly improve the model's performance. To predict new data, each of the trees makes its own individual prediction. the mean of all tree outputs is calculated in regression; for classification, majority voting determines the final prediction.

The strengths of RF regression lie in the ensemble of trees. By averaging the predictions of multiple trees, the algorithm tends to be more robust and mitigate overfitting compared to single DT models. The tree-like structure allows RF to naturally learn non-linear relationships. It usually performs great on many problems, including features with non-linear relationships, making it a valuable tool for various predictive tasks in ML. The default settings for hyperparameters controlling the size of the trees (such as the number of trees in the forest, maximum depth of the trees, and the minimum number of samples needed to split an internal node, etc.) result in fully grown, unpruned trees, which can become excessively large and complex, especially with high-dimensional datasets. To minimise memory usage, it is essential to control the complexity and size of the trees by carefully using these hyperparameters. In addition, RF is robust to outliers due to its integration of multiple trees, hence outliers are less likely to significantly affect the overall model performance. Despite its advantages, RF is less interpretable than DT due to its ensemble nature, which involves aggregating multiple trees. Training an RF can be computationally expensive, particularly when dealing with a large number of trees. More trees are required for more accurate predictions, resulting in a slower model[271]–[273].

2.9.4.4 eXtreme gradient boosting

XGBoost, introduced by Chen and Guestrin[274], is an advanced implementation of gradient gradient-boosting framework and decision-tree-based ensemble method designed for enhanced speed and accuracy. It merged weak learners (DTs) to improve the accuracy of predictions. The concept of this algorithm involves sequentially building a series of classification or regression trees, where each subsequent tree aims at correcting errors in the samples misclassified by the preceding one. It does this by assigning more weights to misclassified samples and lower ones to samples that are correctly classified and updating these weights based on the gradient of the loss function relative to current predictions. The process continues, adding DT sequentially in a level-wise manner until a specified stopping criterion is met, such as perfect prediction of the training data or reaching the maximum number of trees. The final prediction is determined by either calculating a weighted average of the predictions from all the DTs for regression or a majority voting for classification.

XGBoost has gained significant attention in recent years for its ability to handle complex non-linear relationships and its high accuracy, making it a popular choice for ML tasks. It is

designed to be fast and efficient by incorporating parallel processing, even for large datasets. It includes regularisation techniques that help to prevent overfitting, which is a common problem in ML. It uses a combination of L1 and L2 regularisation to reduce the complexity of the model, resulting in more robust and accurate predictions. However, XGBoost is complex, requiring technical expertise to effectively implement and optimise. It can be memory-intensive, especially when training with a large dataset. This can make it challenging to run with limited resources, leading to slower performance. While XGBoost includes regularisation techniques to prevent overfitting, it is still possible for the algorithm to overfit the training data when a large number of trees are employed, highlighting the importance of careful hyperparameter tuning. In addition, it lacks transparency, making it difficult to interpret and understand how it makes predictions[275].

2.9.4.5 Light gradient boosting machine

LGBM, developed by Microsoft[276], is a tree-based supervised ML algorithm for both classification and regression tasks. It is a gradient-boosting implementation similar to XGBoost, combining DTs sequentially as weak learners to form a robust predictive model that can accurately classify new data. Unlike level-wise tree growth in the XGBoost algorithm, LGBM adopts a leaf-wise approach, growing the leaf with the maximum loss reduction at each step. This strategy produces deeper, narrower trees, often improving computational speed and reducing loss more effectively. In addition, LGBM addresses XGBoost's limitations in handling high-dimensional and large datasets, where XGBoost requires scanning all the data samples to compute information gain for every possible split point, making it computationally expensive and time-consuming. To overcome this, LGBM incorporates two main techniques: exclusive feature bundling (EFB) and gradient-based one-side sampling (GOSS). EFB is a feature-reduction method by bundling mutually exclusive features (i.e., those that do not take non-zero values simultaneously) into a single feature. It is especially helpful for high-dimensional or sparse datasets, reducing the number of features and accelerating training while minimising memory usage. GOSS prioritises samples contributing most to gradient updates, discarding low-gradient samples while maintaining the model's accuracy. This reduces the data size for split point estimation, further improving efficiency. These techniques make LGBM a fast and effective choice for large-scale datasets and high-dimensional features.

One of its strengths is that LGBM outperforms other algorithms in terms of faster training speed, reduced memory usage, and its ability to handle complex, non-linear datasets with a large number of features, making it a popular choice for real-world applications. It uses techniques like EFB and GOSS to minimise time and computational cost by performing fewer calculations when determining the best split at each tree node. In addition, LGBM is highly customisable, with many different hyperparameters, such as the learning rate, number of leaves, and maximum tree depth, that could be tuned to improve performance. However, its leaf-wise tree growth can increase model complexity, leading to potential overfitting, especially with small datasets or improper hyperparameter tuning. Optimising these

hyperparameters can be time-consuming and challenging, and model complexity can make it harder to interpret[277]–[279].

2.9.4.6 Support vector machine

SVM, developed by Drucker et al.,[280] based on the concept of Vladimir Vapnik[281], is a supervised ML algorithm initially designed for binary classification and later adapted for the prediction of continuous values prediction as SVR. Both approaches aim to minimise prediction error by optimising a hyperplane. In SVM, the hyperplane serves as the classification boundary that best separates classes, with its margin defined as the distance between the closest data points (support vectors) from each class. These support vectors determine the hyperplane's position and orientation. In SVR, the concept of a hyperplane margin is extended to an ε -insensitive tube. This tube allows some deviation of the data points within a distance ε from the hyperplane without being counted as errors. The objective of SVR is to find a hyperplane that best fits the data that fall within the ε -insensitive tube. Points outside this tube (support vectors) contribute to prediction error, quantified by slack variables (ξ), which measure their distance from the tube boundary.

SVM/SVR transform data into a higher-dimensional space to identify a separating decision surface. This process is referred to as a kernel function. One of the reasons for their flexibility and effectiveness is their ability to use different types of kernels. The two most commonly used kernels are the linear kernel and the radial basis function (RBF). The linear kernel is well-suited for large datasets with a linear relationship and provides faster computation compared to other kernels. When dealing with non-linear data, it becomes necessary to consider using non-linear kernels, such as the RBF, which can be more computationally expensive. Thus, selecting the appropriate kernel depends on the type of problem, the computational complexity, and the characteristics of the data. However, a notable drawback is the lack of result transparency, as the algorithm does not provide how predictors are combined to optimise the hyperplane, reducing interpretability[282]–[285].

2.9.4.7 K-nearest neighbour

The KNN algorithm is a non-parametric, supervised learning method widely used for classification and regression tasks in ML. It operates on the principle that similar data points are located near each other, using proximity to make predictions. For classification, KNN assumes that data points with similar input values belong to the same class, while for regression, similar inputs yield similar outputs. To predict the label or value of a new data point, KNN calculates the distance between the query point and all other data points using metrics such as Euclidean, Manhattan, Minkowski, or Hamming distances. It then selects the specified number of data points (K) closest to the query in the feature space. The prediction is made by either assigning the class label most frequently represented among the K nearest neighbours (majority vote) for classification or averaging the labels of the K nearest neighbours for regression. For instance, if $k=1$, the query point is assigned to the same class or value as its single nearest neighbour.

KNN is effective for small to moderately sized datasets, especially when relationships between features are non-linear. Its simplicity, with only one hyperparameter (K) to tune, makes it straightforward to optimise. However, choosing an appropriate K value requires balancing: low K values may result in high variance but low bias (overfitting), while high K values can lead to high bias but low variance (underfitting). Noisy or outlier-prone data typically perform better with higher K values, and CV helps identify the optimal K. Using an odd K for classification avoids ties. Despite its advantages, KNN has limitations. It struggles with high-dimensional data, where distances between points become less meaningful as the number of features increases, and it is computationally intensive during prediction due to the need to calculate distances for all training points. The algorithm is sensitive to noise and outliers, which can significantly affect distance calculations and can skew results. In addition, feature scaling is essential for meaningful distance calculations[286]–[288].

2.9.5 Model building and evaluation

2.9.5.1 The general procedure of k-fold CV

CV is a resampling method used to estimate the performance of ML models by training and testing them on different subsets of the data. It is particularly useful when working with limited data samples. The main objective is determining how well the model generalises to unseen data. Unlike a single train-test split, which evaluates the model only once and risks yielding overly optimistic results by chance, CV performs multiple iterations of training and validation, providing a more robust and reliable evaluation of the model's performance.

A common CV technique is a k-fold CV, where the dataset is divided into k equally sized folds. The model is trained and tested k times, with each fold serving as the testing set once while the remaining folds are used for training. After each iteration, an evaluation metric (e.g., prediction error) is computed, resulting in k performance estimates. The final performance estimate is the average of these values. This process ensures that the model is evaluated on independent subsets of data, reducing bias and providing a more reliable estimate of generalisation error (Figure 4).

The key hyperparameter for k-fold CV is k, the number of folds. Common choices for the value of k are either 5 or 10. For instance, with k = 10, the dataset is split into ten folds, where nine folds are used for training and the remaining one for testing in each iteration. The choice of k depends on the desired balance between bias, variance, and computational cost. Larger k values (e.g., >10) reduce bias in the error estimate but can increase variance, computational time, and may limit the model's ability to capture differences across iterations. Smaller k values (e.g., <5) reduce computational time but may introduce bias into the error estimate[289].

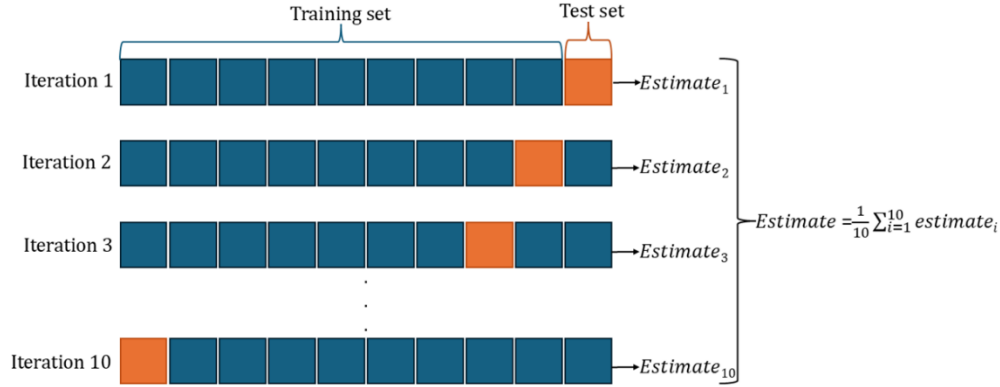


Figure 4. The example of k-fold CV with k=10. The dataset is divided into 10 equally sized folds. In each iteration, one-fold is used as the test set (orange), while the remaining folds are used as the training set (blue). This process is repeated ten times, with each fold serving as the test set exactly once. The model's overall performance is calculated as the average of the estimate across all iterations.

2.9.5.2 Nested k-fold CV with OMC feature selection

High-dimensional omics data, such as genomic, transcriptomic, and epigenomic datasets, has become a more common challenge in biomedical research, where the number of features is larger than the number of samples. Many features may be redundant or irrelevant, which can reduce model accuracy and generalisation. Feature selection is therefore crucial in ML to identify the most relevant features while discarding irrelevant ones, enhancing model performance.

SelectKBest, a widely used filter-based feature selection method in scikit-learn, scores features using statistical tests to select the top k features based on their relevance to the target variable. For regression tasks with numerical inputs and outputs, correlation coefficients like Pearson's (linear relationships) and Spearman's rank (non-linear relationships) are commonly used. When the output is categorical, as in classification tasks, correlation techniques account for the categorical target, with methods like analysis of variance (ANOVA) for linear relationships and Kendall's rank coefficient for non-linear cases (assuming the categorical output is ordinal). For less common cases involving categorical inputs and numerical outputs, the same techniques (e.g., ANOVA, Kendall's rank) are applied in reverse. In classification problems with categorical inputs and outputs, the chi-squared test is the most common approach, while mutual information, derived from information theory, can also be used[290], [291].

Building on this principle, we developed the OMC feature selection strategy to build ML models using only the most relevant features[45]. OMC consists of three key steps: (i) ranking features based on their relevance to the target variable, (ii) training ML models on subsets of features, and (iii) selecting the model with the optimal number of features. Feature relevance is assessed by calculating correlation scores with the target output. We then select the top 2

to $n/2$ features (where n is the sample size) as the most informative, ensuring that the ML model will not be challenged by high-dimensional data. A nested k -fold CV process is used to optimise feature selection and evaluate model performance through a series of train/validation/test set splits. In the inner loop, models are trained on the training set with subsets of features and select the optimal number of features over the validation set. In the outer loop, the final model is built with the selected features and evaluated on the test set, which is not used during feature selection for unbiased performance assessment. This process is repeated across all outer folds, and generalisation error is estimated by averaging test set scores across splits (Figure 5).

Feature selection offers several advantages: it improves model accuracy, reduces dimensionality, and minimises data noise. ML models incorporating feature selection are more interpretable and less prone to overfitting, particularly in high-dimensional datasets. However, nested k -fold CV is computationally intensive due to the large number of models that must be trained in the inner loop. In addition, statistical measures for feature selection evaluate each input feature independently against the target output. While this approach effectively identifies important features, it may lead to the selection of redundant ones and does not account for interactions between them[292].

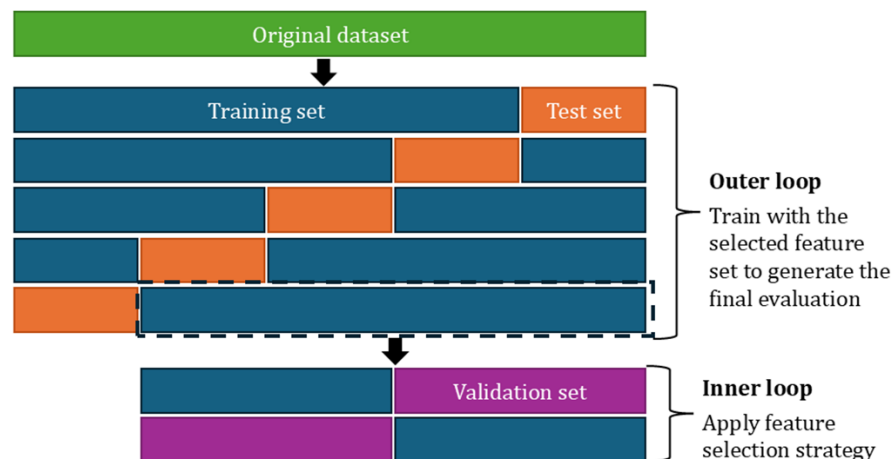


Figure 5. Nested k -fold CV scheme for feature selection.

2.9.6 ML model evaluation metrics

The main objective of predictive modelling is to build a model that has high performance for making accurate predictions on new, unseen data. Achieving this requires robust training on a training set followed by rigorous evaluation on a testing set. Evaluation metrics are quantitative measures used to assess an ML model's performance, offering insights into how well the model performs and facilitating model comparisons. Choosing the right evaluation metrics depends on the problem domain, data type, and desired outcomes. This section provides an overview of key evaluation metrics for binary classification tasks, highlighting their advantages and limitations.

In binary classification, outputs are limited to two classes: positive (1) and negative (0). Model performance is commonly represented using a confusion matrix, a 2×2 table with true classes on the x-axis and predicted classes on the y-axis (Figure 6). The matrix categorises predictions into four groups:

- TP: The model predicts 1, and the actual value is 1.
- TN: The model predicts 0, and the actual value is 0.
- FP: The model predicts 1, but the actual value is 0.
- FN: The model predicts 0, but the actual value is 1

These counts are not directly used for evaluation but are needed for calculating metrics such as MCC, ROC-AUC, and precision-recall area under the curve (PR-AUC)[293], [294].

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

Figure 6. Confusion matrix for binary classification. The target variable consists of two classes: positive or negative. Columns represent true classes, while rows indicate predicted classes. The true and predicted class assigned to cases are compared. The matrix compares true and predicted classifications, presenting TP, TN, FP, and FN to illustrate the model's accuracy and misclassification patterns.

2.9.6.1 MCC

MCC, introduced by Brian Matthews in 1975[295], a widely used metric for evaluating binary classification performance, especially with imbalanced datasets. Despite it was first introduced in biochemistry[295], MCC has gained popularity in several scientific fields, including software defection prediction[296], recommender systems[296], pattern recognition[297], and medicine[298]–[300].

MCC is more robust and reliable on imbalanced datasets than other adopted metrics in binary classification tasks, such as F1 score[301], accuracy[301], balanced accuracy[302], Cohen's Kappa[303], and Brier score[303]. It provides a balanced measure by incorporating TPs, TNs, FPs, and FNs from the confusion matrix, proportionally both to the size of positive and negative samples in the dataset. Thus, it achieves a high score only when most predictions, both positive and negative, are correct. However, MCC relies on a predetermined threshold, typically set at 0.5.

MCC is calculated as follows:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Its values range from -1 to +1, where +1 indicates perfect classification, -1 signifies complete misclassification (all negatives predicted as positives and vice versa), and 0 corresponds to random predictions. The MCC is undefined when the model predicts a single class for all samples, resulting in a zero denominator.

2.9.6.2 ROC-AUC

Using a heuristic threshold for a confusion matrix, such as 0.5, is a common practice in ML but has the flaw of relying on a specific cut-off value to classify samples as either positive or negative classes. In contrast, ROC-AUC is a threshold-independent metric that evaluates a model's performance across all possible thresholds in binary classification. To generate the ROC curve, predicted probabilities are mapped into class labels at varying thresholds. Values above or equal to a threshold are classified as positive, while those below are negative. These labels form a confusion matrix from which the true positive rate (TPR, or recall) and false positive rate (FPR, or 1-specificity) are calculated. The TPR is $\text{TP}/(\text{TP}+\text{FN})$, and the FPR is $\text{FP}/(\text{FP}+\text{TN})$. Plotting TPR (y-axis) against FPR (x-axis) for all thresholds creates the ROC curve, with the ROC-AUC summarising the overall performance the model's ability to distinguish between classes. The ROC-AUC ranges from 0 to 1, where a curve closer to the upper-left corner indicates better discrimination ability. A value of 1 signifies perfect classification, 0 indicates complete misclassification, and 0.5 represents random guessing (a diagonal line). Typically, an ROC-AUC above 0.8 denotes good model performance[293], [294], [304] (Figure 7A).

The ROC-AUC offers several advantages. It is a threshold-independent metric that evaluates model performance across all classification thresholds, providing a comprehensive view without relying on a single threshold. It also visualises the trade-off between TPR and FPR, facilitating the selection of an optimal threshold that maximises the model's performance. However, it has limitations. The ROC-AUC is not suitable for imbalanced datasets, as it can give an overly optimistic assessment of the model's performance. If TNs outnumber TPs, a model's ROC-AUC can be high simply by correctly predicting that the majority of negatives to be negatives. Thus, the ROC-AUC is more suitable for balanced datasets where both classes are equally important or frequent. Using it alone is an insufficient representation of model discrimination and should be interpreted with caution[305].

2.9.6.3 PR-AUC

PR-AUC is a threshold-independent metric used to evaluate binary classification models, particularly in imbalanced datasets. Unlike ROC-AUC, which plots the TPR against the FPR, PR-AUC represents the area under the precision-recall curve. This curve is constructed by varying the classification threshold and plotting precision (proportion of true positives

among all positive predictions: $TP/(TP+FP)$) on the y-axis against recall (proportion of true positives among all actual positives: $TP/(TP+FN)$) on the x-axis. PR-AUC reflect the model's ability to balance high precision (not predicting negative samples as positives) and high recall (correctly predicting positive samples). It ranges from 0 to 1. Unlike ROC-AUC, which maximises in the upper left corner, PR-AUC is optimised in the upper right corner of the plot. A PR-AUC of 1 represents a perfect classifier, while its baseline value equals the proportion of positive samples in the dataset (Figure 7B).

PR-AUC offers advantages over ROC-AUC, particularly in handling class-imbalanced datasets. It measures the model's ability to predict positive classes correctly across all recall levels, making it suitable when the positive class is rare or when FPs are more important than FNs. In addition, it also helps to select a threshold based on the desirable values of precision and recall. However, PR-AUC has limitations, as it does not account for the true negative rate (TNR), focusing only on the performance within the positive class. This can make PR-AUC less informative for assessing overall accuracy in datasets where negative outcomes are also important[306], [307].

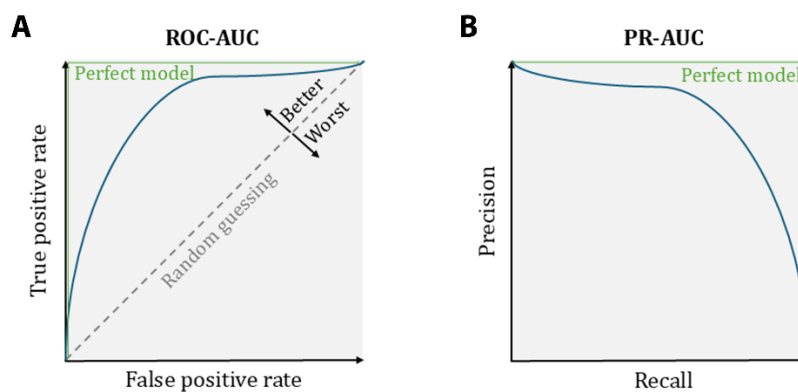


Figure 7. ROC-AUC and PR-AUC. (A) The ROC-AUC demonstrates the trade-off between TPR and FPR across thresholds, where a perfect model (AUC of 1) approaches the top left corner, and a diagonal line indicates random guessing (AUC of 0.5). (B) The PR-AUC illustrates the balance between precision and recall, particularly useful for imbalanced datasets. A perfect model (PR-AUC of 1) stays close to the top-right corner, which means both precision and recall are high across thresholds, while a random model equals to the proportion of positive samples in the dataset.

Chapter 3

Materials and methods

3.1 Gene expression profiles and clinical data acquisition

We retrospectively collected two publicly available datasets for model development and validation: the IMVigor210 phase II clinical trial (registered with ClinicalTrials.gov, number NCT02108652)[2], and the study by Snyder et al.[54].

The IMVigor210 trial served as the discovery dataset for model development[2]. This trial was a multicentre, single-arm that investigated the effectiveness and safety of atezolizumab in 348 locally advanced or mUC patients who were platinum-ineligible or progressed during or following platinum-based chemotherapy. RNA-sequencing data from these patients were generated using TruSeq RNA Access technology (Illumina®). This dataset includes raw counts for 31,286 genes, identified by their Entrez Gene IDs (e.g., 59), along with clinical features. The data can be accessed through the IMvigor210CoreBiologies R package at <http://research-pub.gene.com/IMvigor210CoreBiologies/>). For the validation dataset, we included 29 locally advanced or mUC patients receiving atezolizumab published by Snyder et al.[54]. RNA-sequencing and clinical data for these patients were obtained from <http://doi.org/10.5281/zenodo.546110>. Here, gene expressions of tumour samples were sequenced using an Illumina HiSeq 4000 and were processed to the gene level using tximport, resulting in a total of 39,292 gene expressions with ensemble gene IDs (e.g., ENSG00000139618). Table 5 summarises the baseline clinical features that we used in this study, including gender, ECOG performance status, PD-L1 protein expression status on TICs, TMB per megabase, TNB per megabase, prior receipt of intravesical administration of BCG immunotherapy status, prior receipt of platinum-based chemotherapy status, and pre-platinum sample collection status. Patients in both studies received atezolizumab at 1,200 mg IV every 3 weeks and were evaluated for their response using CT/MRI scans based on RECIST version 1.1[120], which classified patients into four categories: CR, PR, SD, or PD. CR or PR were considered responders, while SD or PD were categorized as non-responders. The discovery dataset showed a response rate of 23% (68 out of 298 patients), while the validation dataset had a rate of 32% (7 out of 22 patients).

3.2 Data pre-processing for ML

Our computational framework, illustrated in Figure 8, outlines the process to develop and validate the predictive ability of gene expression profiles on mUC patients' response to atezolizumab. To facilitate further ML analysis, it's crucial to harmonise the gene nomenclature used between the two datasets. Due to inconsistencies, Ensemble gene IDs in the validation dataset were standardised to Entrez gene IDs based on the gene information provided within the validation dataset itself. This standardisation facilitated the comparison and identification of common genes shared between the two datasets. We addressed missing values by excluding samples with missing values of any features, ensuring that the dataset

remains consistent and that the models are built based on complete information. Consequently, we were left with a dataset comprising 19,941 common gene expressions and 8 clinical features shared between the two datasets with 298 and 22 gene expression profiled atezolizumab-treated mUC patients in the discovery and validation datasets, respectively.

Gene expression profiles were log2 transformed ($\log_2(\text{count} + 1)$), wherein a count of 1 was added before applying the logarithmic operation to prevent taking the logarithm of zero, which is undefined. We assigned unique integer values to categorical variables, including atezolizumab response labels, gender, ECOG performance status, PD-L1 protein expression status on TICs, intravesical administration of BCG immunotherapy status, received platinum-based chemotherapy status, and the sample collected pre-platinum status, as illustrated in Table 5. This encoding process facilitated the utilisation of these categorical variables within the analytical ML framework.

We employed regression ML algorithms to predict the response of mUC patients to atezolizumab. To accomplish this, atezolizumab response labels (CR, PR, SD, and PD) were numerically encoded as 1.0, 2.0, 3.0, and 4.0, respectively. This approach allowed the ML model to learn from each patient's response labels instead of grouping similar response categories together, as is often done in binary classification and can lead to information loss (e.g., grouping CR and PR as responders and SD and PD as non-responders). Furthermore, data standardisation was conducted using StandardScaler (also known as Z-score). This method centres the data by subtracting the mean and scales it by dividing by the SD, resulting in a mean of 0 and an SD of 1, ensuring that all features within the dataset are on a comparison scale.

All analyses were performed using Python version 3.7.3 (available at <https://www.python.org/>) along with packages from scikit-learn version 0.24.2 (accessible at <https://scikit-learn.org>).

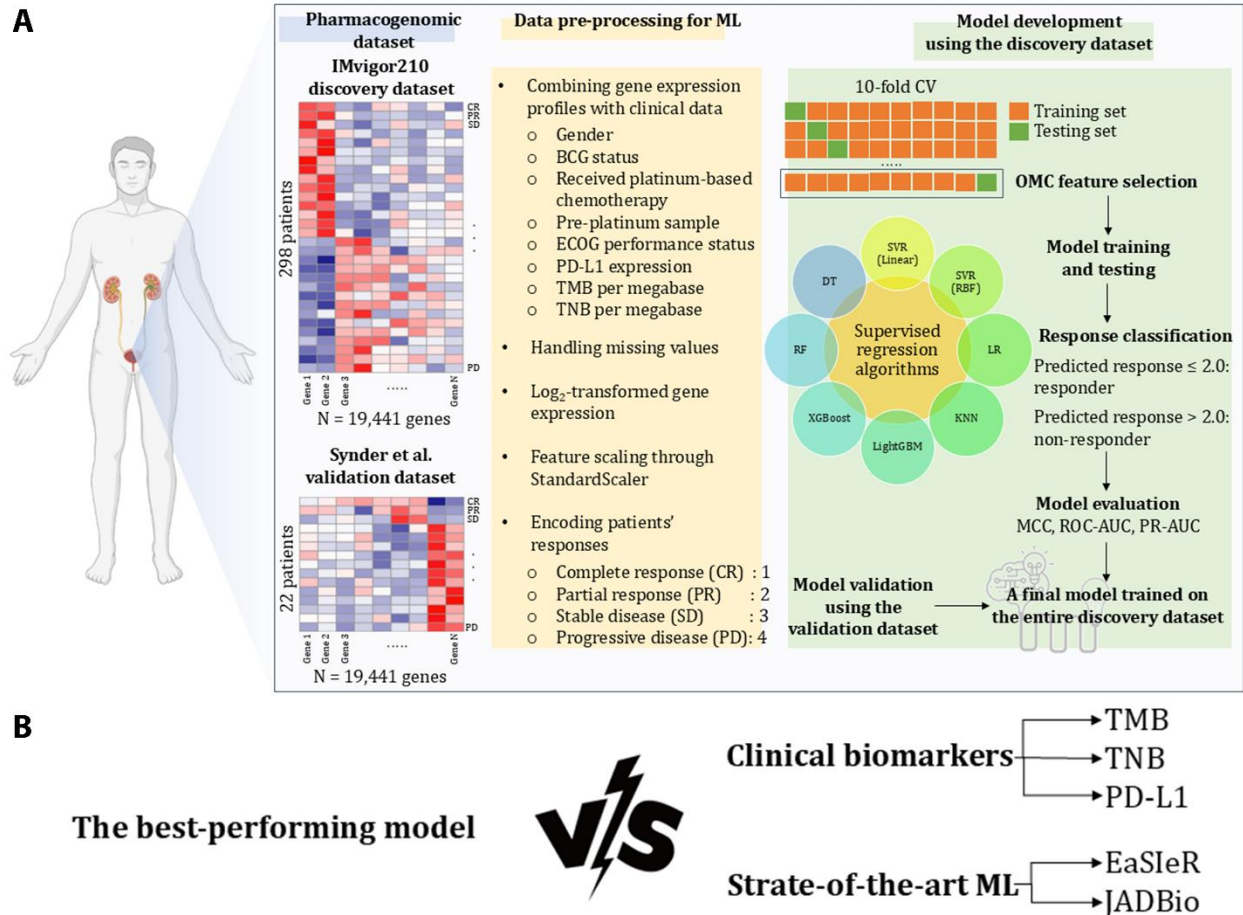


Figure 8. ML workflow to predict atezolizumab response in mUC patients. (A) The workflow involved developing and validating supervised ML models using gene expression profiles from mUC patient tumours treated with atezolizumab, along with clinical data (Table 5). The IMvigor210 phase II clinical trial cohort (298 mUC patients) served as the discovery dataset for model development[2], while the Snyder et al. cohort (22 mUC patients) was the validation dataset for model validation[54]. Both datasets were pre-processed for ML. Eight regression algorithms were employed to develop models with OMC feature selection, reducing the features from the initially considered ~20,000 features. Model performance was assessed using 10-fold CV with five repetitions, each using a unique random seed. Predicted responses were classified as responders or non-responders based on a threshold of 2.0. In this way, the model learned all the information about patients' response labels (CR, PR, SD, and PD), capturing the full granularity of the data rather than pre-grouping similar categories together before modelling, as is often done in binary classification (e.g., grouping CR and PR as responders, and SD and PD as non-responders). The best OMC model, determined by the highest MCC, was selected and applied to the validation dataset. MCC, ROC-AUC, and PR-AUC were reported. (B) The performance of best-performing model was compared against clinical biomarkers (e.g., TMB, TNB, and PD-L1) and state-of-the-art ML models such as EaSIeR and JADBio for atezolizumab response prediction.

Table 5. The baseline clinical characteristics of mUC patients from the discovery and validation cohorts. The following were used as clinical features in ML models: gender, ECOG performance status, PD-L1 expression on TIC status, TMB per megabase, TNB per megabase, received BCG, received platinum-based chemotherapy, sample collected pre-platinum.

Characteristics	ML input	Discovery dataset (N = 298)	Validation dataset (N = 22)
Study cohort		IMVigor210 phase II clinical trial [2]	Snyder et al., 2017[54]
Drug response			
Complete response (CR)	1	25 (8%)	2 (9%)
Partial response (PR)	2	43 (14%)	5 (23%)
Stable disease (SD)	3	63 (21%)	5 (23%)
Progression disease (PD)	4	167 (56%)	10 (45%)
Response rate		23%	32%
Gender			
Male	0	233 (78%)	20 (91%)
Female	1	65 (22%)	2 (9%)
ECOG performance status			
0	0	121 (41%)	1 (5%)
1	1	165 (55%)	21 (95%)
2	2	12 (4%)	0 (0%)
PD-L1 expression on TICs			
IC0	0	83 (28%)	3 (14%)
IC1	1	112 (38%)	11 (50%)
IC2	2	102 (34%)	8 (36%)
NA	NA	1 (<1%)	0 (0%)
TMB per megabase			
Number of patients	Real values	234	21
Range (Median $\pm$ SD)		0-62 (8 $\pm$ 9.38)	0-11 (1 $\pm$ 3.12)
TNB per megabase			
Number of patients	Real values	216	21
Range (Median $\pm$ SD)		0.04 – 11.69 (0.86 $\pm$ 1.60)	0 - 39(3 $\pm$ 9.17)
Received BCG			
Yes	1	231 (78%)	8 (36%)
No	0	67 (22%)	14 (64%)
Received platinum-based chemotherapy			
Yes	1	233 (78%)	19 (86%)
No	0	65 (22%)	3 (14%)
Sample collected pre-platinum			
Yes	1	197 (66%)	10 (45%)
No	0	95 (32%)	9 (41%)
NA	NA	6 (2%)	3 (14%)

3.3 Building OMC regression models

One major challenge in developing predictive ML models is the high dimensionality of data, where the number of features is much larger than the number of patients. Typically, only a small number of tumours have documented responses to the drug, while the molecular profiles of these tumours may include over 20,000 features. In such scenarios, including all available features in the model-building process might negatively affect performance by introducing complexity and noise, leading to overfitting the training data and poor generalisation of new data. To address this, the OMC feature selection strategy[45] was employed to complement the ability of ML algorithms by using only a small subset of informative features while discarding irrelevant ones.

To develop an OMC model, the initial step involves selecting the most informative features associated with atezolizumab responses. As patients' response labels were encoded as 1.0, 2.0, 3.0, and 4.0 for CR, PR, SD, and PD, respectively and the features were real-values of gene expressions, a univariate linear regression test was used to evaluate the significance of each feature in relation to atezolizumab response, returning P -values for each feature. Those with small P -values indicate high discriminatory power to distinguish atezolizumab response. Then, a standard K -fold CV (with a K value of 5, 10, and a number of samples) was used to estimate the predictive performance of each algorithm using only the most informative features of atezolizumab responses. During the CV, each of the eight algorithms (CART, XGBoost, LGBM, KNN, RF, LR, and SVR with linear and RBF kernels) was trained on $K-1$ folds using only the considerate subsets of features (the top 2 to $n/2$ subset of features, where n is the number of samples), and tested on the remaining 1-fold to evaluate model performance. This limit to a subset of $n/2$ features ensures that the ML algorithm will not be challenged by high-dimensional data, as each trained model will have a minimum of two data points of considered feature). The CV process was repeated K times, with each of the K -folds sequentially being used exactly once as the test set. Thus, each patient has exactly one out-of-sample prediction.

In our approach, model performance was not calculated individually on the test set for each fold and then averaged the results because each of the test sets would be too small and thus harm comparison to CV with other K values. As is common with low-sample size, we used instead a merged CV. The resulting out-of-sample predictions from all folds were merged and the MCC were calculated once based on all samples. Among all $n/2$ trained models, the best OMC model, the one with the highest MCC, was trained on all samples of the discovery dataset and subsequently used to predict samples in the validation dataset.

3.4 Building models based on clinical data only

Given the limited number of clinical features (only 8), the OMC feature selection was not applied. Instead, each of the eight regression algorithms was trained on the discovery dataset using the full set of clinical features (Table 5) and subsequently used to predict samples in the validation dataset.

3.5 Building models based on immunotherapy response biomarkers

To compare the performance of the best-performing model with that of individual immunotherapy biomarkers: TMB per megabase, TNB per megabase, or PD-L1 protein expression status on TICs by IHC assay (IC0: <1%, IC1: ≥1% but <5%, and IC2: ≥5%) as provided by the original study[2], [54], we trained an LR algorithm on each feature using the discovery dataset. These models were then used to predict samples in the validation dataset.

3.6 Building random class permutation models

The permutation test assesses whether the model's performance exceeds what would be expected by chance, ensuring that the predictive accuracy reflects true data patterns rather than randomness. To build a permutation model, we randomly generated the atezolizumab response labels (CR, PR, SD, and PD) while keeping the input features unchanged. This creates a scenario where there is no meaningful relationship between features and labels, resulting in a model that predicts labels randomly. Using these permuted labels, we repeated the CV process and evaluated the model following the same procedure as the original version.

3.7 Model performance evaluation

We used regression algorithms for classification by applying thresholding techniques to model-predicted values, converting them into binary class labels (responder or non-responder). Responders were defined as patients with CR or PR, while non-responders were defined as patients with SD or PD. Since atezolizumab responses were encoded as follows: CR as 1, PR as 2, SD as 3, and PD as 4, we classified out-of-sample predictions from K -folds CV into binary categories using a threshold of 2.0, where CR and PR were considered responders. Predictions above or equal to this threshold were classified as non-responders and those below as responders. We then compared actual and predicted classes to compute a confusion matrix, including TP, TN, FP, and FN, which was used to calculate evaluation metrics such as MCC, ROC-AUC, and PR-AUC.

The MCC ranges from -1 to 1, where 1 indicates perfect agreement between actual and predicted classes, and -1 indicates perfect disagreement. A score of 0 suggests performance equivalent to random guessing. An undefined MCC occurs when the model predicts only one class for all instances, resulting in a zero denominator in the MCC formula.

The ROC-AUC, ranging from 0 to 1, measures a model's overall performance to distinguish between two classes. The ROC curve is created by plotting the true positive rate against the false positive rate at various thresholds, each threshold generating a point on the curve. An ROC-AUC of 1 indicates perfect classification, while an ROC-AUC of 0.5 indicates the model performs no better than random chance.

The PR-AUC, ranging between 0 and 1, summarises how well a model can distinguish between two classes, considering both its ability to not mark a negative sample as positive (precision) and its ability to find all the positive samples (recall). PR-AUC plots precision

against recall, where precision is the proportion of TPs among all positive predictions, and recall (or sensitivity) is the proportion of TPs correctly identified by the model. A PR-AUC of 1 indicates perfect precision and recall at all thresholds. The random baseline of the PR-AUC is determined by the proportion of positive samples in the dataset, i.e. positive / (positive + negative).

3.8 Prediction of response to atezolizumab using the EaSIeR

EaSIeR [48] is an R Bioconductor package developed to assess patients' response to PD-1/PD-L1 inhibitors. By using bulk RNA sequencing data from patient tumours as input, EaSIeR quantifies patient-specific, real-valued five types of system-based signatures of TME, including 11 immune cell types, 14 pathway activities, 118 transcriptional factor activities, 812 ligand-receptor pairs, and 169 cell-cell pairs. The cancer-specific elastic net regularized multi-task linear regression (RMTLR) model, built on TCGA samples, was used to predict 10 tasks of immune response scores: Cytolytic activity (CYT), Ock immune signature (Ock_IS), Roh immune score (Roh_IS), Chemokine signature (Chemokine), Davoli immune signature (Davoli_IS), IFNy signature (IFNy), Expanded immune signature (Ayer_expIS), T-cell inflamed signature (Tcell_inflamed), Repressed immune resistance (RIR), Tertiary Lymphoid structure signature (TLS) based on five system-based signatures of TME and their combination. These scores have been experimentally validated in the literature (Table 2) and comprehended key gene signatures for immune activation in the TME in relevant contexts. Higher scores indicate a higher likelihood of a patient responding to PD-1/PD-L1 inhibitors.

We did not train the EaSIeR model on the discovery dataset and evaluate it on the validation dataset, as done with our method because EaSIeR cannot be retrained on new datasets. Instead, the bladder urothelial carcinoma (BLCA)-specific RMTLR model, pre-trained on TCGA-BLCA samples available in the EaSIeR package, was used to predict the likelihood of response to atezolizumab in 320 mUC patients (75 responders and 245 non-responders) from the merged discovery and validation datasets. The model predicted 10 immune response scores based on six sets of features, including five system-based signatures of TME and their combination, producing a total of 60 EaSIeR model predictions.

To compare the predictive performance in this task, we developed OMC models using the same set of 320 mUC patients from the merged discovery and validation datasets. Eight regression algorithms were employed, including CART, RF, XGBoost, LGBM, KNN, LR, and SVR with linear and RBF kernels, to predict the patients' responses to atezolizumab. Each model was built using two sets of features per learning algorithm: gene expression profiles alone and integrated gene expression profiles with clinical data. The model's performance was evaluated using a nested 10-fold CV (the CV is repeated five times changing random seed to assess the robustness and ensure the reliability of the results).

To calculate predictive performance, we first normalised the resulting 320 out-of-sample model's predictions from both our approach and the EaSIeR to a range of 0 to 1 in order to allow for a fair comparison and evaluation across different models, regardless of their scale of prediction output. Once normalised, we then assigned class labels (responder or non-

responder) by applying a threshold value of 0.5 to the normalised predicted values. Subsequently, evaluation metrics (MCC, ROC-AUC, and PR-AUC) were computed.

3.9 Prediction of response to atezolizumab using the JADBio

JADBio version 1.4.94 (www.jadbio.com)[55], an automated ML platform, was used to develop regression models for predicting the response to atezolizumab in 320 mUC patients from the merged discovery and validation datasets (75 responders and 245 non-responders). Two sets of features were considered for model development, either gene expression profiles alone or integrated gene expression profiles and clinical data. Each dataset was input to JADBio as a 2D matrix, where rows corresponded to patient samples and columns corresponded to considered features, with the defined outcome of interest, in this case, being the response to atezolizumab (CR, PR, SD, and PD). To identify the best-performing model, JADBio employs LASSO, SES, and univariate algorithms for feature selection and various regression algorithms, including ridge LR, SVR with linear, polynomial, and Gaussian kernels, RF, and DT, for modelling, along with a range of hyperparameter values (Table 4). This process creates a unique ML pipeline (referred to as a configuration) that takes the dataset, carries out a generalised CV (R repeated, K -fold, and N -incomplete CV) for performance estimation, and searches in the space of admissible configurations to identify the one that yields the highest performance. We opted for the “extensive tuning effort” option to extensively search for the best-performing model, despite the increased computational costs.

To compare the predictive performance in this task, we saved the out-of-sample CV predictions from JADBio’s best-performing model. The evaluation metrics (MCC, ROC-AUC, and PR-AUC) were calculated using normalised predictions (ranging from 0 to 1), following the same way as our approach for comparison purposes. Class labels (responder or non-responder) were assigned based on a threshold of 0.5.

3.10 Statistical analysis

The two-tailed unpaired Welch’s t -test was used to compute P -values in our data analysis. A P -value less than 0.05 was considered statistically significant. For pathway enrichment analysis, WebGestalt uses a hypergeometric test to calculate statistical significance, which further employs the Benjamin–Hochberg (BH) multiple-testing correction for the correct P -value or false discovery rate (FDR). The significantly enriched pathway was identified with FDR less than 0.05.

Chapter 4

Results

4.1 Cohort characteristics of atezolizumab-treated mUC patients

In this study, we retrospectively collected tumour gene expression profiles and clinical data from mUC patients treated with atezolizumab across two independent studies: the IMvigor210 phase II clinical trial[2] and Snyder et al.[54]. Both studies included a total of 320 gene expression-profiled mUC patients who had known responses to atezolizumab. The IMvigor210 discovery dataset included 298 patients, while the Snyder validation dataset comprised 22 patients. Table 5 summarises the baseline clinical characteristics of these patients from both datasets.

In terms of response to atezolizumab, the discovery dataset exhibited varying degrees of response, with 8% achieving a CR (n = 25), 14% a PR (n = 43), 21% SD (n = 63), and 56% PD (167), resulting in a response rate of 23%. Similarly, the validation dataset showed a response rate of 32%, with 9% achieving CR (n = 2), 23% PR (n = 5), 23% SD (n = 5), and 45% PD (n = 10). The gender distribution revealed a predominance of males in both datasets, comprising 78% (n = 233) in the discovery dataset and 91% (n = 20) in the validation dataset. The patient's ECOG performance status skewed towards better performance in the discovery dataset. Specifically, 41% (n = 121) had a status of 0, indicating they were fully active and unrestricted in performing their usual daily activities. Another 55% (n = 165) had a status of 1, meaning they were limited in physically demanding tasks but still capable of light or sedentary work, such as office or housework. Only 4% (n = 12) had a status of 2, where they were able to perform self-care but unable to engage in any work-related activities, though they were active for more than half of their waking hours. While the validation dataset had 5% (n = 1) with a status of 0, 95% (n = 21) 0 with 1, and none with a status of 2. PD-L1 expression on TICs was assessed in both studies, using the SP142 IHC assay, categorising as IC0 (<1%), IC1 ($\geq 1\%$ and <5%), and IC2/3 ($\geq 5\%$) based on the percentage of PD-L1-positive TICs. IC1 was the most prevalent in both datasets, accounting for 38% (n = 112) in the discovery dataset and 50% (n = 11) in the validation dataset. Both the discovery and validation datasets reported TMB and TNB as mut/Mb, normalised to account for differences in gene panel size and the target region. However, detection methodologies, calculation methods, and other bioinformatics test specifications vary across laboratories, contributing to data inconsistencies. In calculating TMB, the discovery dataset considered all somatic coding mutations, including synonymous and non-synonymous, per megabase of the tumour genome examined by the FoundationOne CDx DNA-based commercial panel. The median TMB was 8 mut/Mb, with a range spanning from 0 to 62 mut/Mb across all mUC patients. By contrast, while the validation dataset considered only non-synonymous mutations from the IDT xGen Whole Exome Panel, the validation dataset had a lower median TMB of 1, ranging from 0 to 11 mut/Mb). Despite both studies estimated neoantigens using non-synonymous mutations, they utilised a different bioinformatics tool for neoantigen prediction. The

median TNB in the discovery dataset was 0.86 mut/Mb, with a range from 0.04 to 11.69 mut/Mb, while in the validation dataset, it was 3 mut/Mb, with a range from 0 to 39 mut/Mb. Moreover, prior receipt of intravesical administration of BCG immunotherapy and platinum-based chemotherapy were also considered as features, illustrating variations in treatment patterns among patients and factors influencing treatment. Notably, a higher percentage of patients in the discovery datasets had previously received BCG, with 78% (n = 231) compared to only 36% (n = 8) in the validation dataset. Similarly, a higher percentage of patients across both datasets had previously received platinum-based chemotherapy, with 78% (n = 233) in the discovery dataset and 86% (n = 19) in the validation dataset. Furthermore, sample collection before platinum treatment was more common in the discovery dataset, with 66% (n = 197) of the samples collected pre-platinum, compared to 45% in the validation dataset (n = 10).

4.2 ML model performance in predicting mUC patient response to atezolizumab using gene expression profiles

By using gene expression profiles from 298 atezolizumab-treated mUC patients in the IMvigor210 discovery dataset[2], we develop OMC models with eight regression algorithms to predict the response of atezolizumab. A standard 10-fold CV procedure was conducted on the discovery dataset to select the best OMC model, which was then tested on 22 mUC patients treated with atezolizumab from the Snyder et al. validation dataset[54]. To ensure the reliability and robustness of the models, a 10-fold CV was repeated five times with different random seeds. The computational framework is summarised in Figure 8, fully described in the method section.

The evaluation metrics (MCC, ROC-AUC, and PR-AUC) on the validation dataset across eight ML models using either gene expression profiles (GEP) to predict mUC patient response to atezolizumab are shown in Figure 9 and Table 6. It was observed that four models using GEP showed promising results in distinguishing responders from non-responders, with a median mMCC above 0.2: CART-OMC (mMCC of 0.328), XGBoost-OMC (mMCC of 0.319), SVR(Linear)-OMC (mMCC of 0.319), and LGBM-OMC (mMCC of 0.297). Notably, CART-OMC exhibited the highest MCC of 0.437, with a median MCC of 0.328 ± 0.14 , while also achieving the highest ROC-AUC and PR-AUC values of 0.881 and 0.955, respectively. Hence, it is the recommended model for this challenging problem. Delving deeper into the CART-OMC model, it was revealed that it employed a non-linear combination of only 29 genes (0.15% out of the total 19,441 genes) as predictive features for predicting patient responses to atezolizumab in mUC. The list of these 29 predictive genes is shown in Table 7.

In contrast, RF-OMC, KNN-OMC, LR-OMC, and SVR(RBF)-OMC showed poor predictive performance, each yielding an MCC of 0.0. This indicates that these models failed to correctly classify any instances in the dataset, performing no better than random chance. LR-OMC even exhibited a negative MCC of -0.216, indicating performance worse than randomly guessing.

Note that the KNN, LR, and SVR algorithms produced the same results every time with the same input data and parameters because they are deterministic algorithms that do not

involve random processes, such as stochastic optimization, which would require a random seed. As a result, there is no inherent randomness affecting their computations.

Low MCC and high ROC-AUC were observed in RF-OMC and KNN-OMC models. This typically occurs in an imbalanced dataset when a model performs well on the dominant (positive) class, resulting in a high ROC-AUC but at the cost of a high false positive rate (or a low number of TNs). Unlike ROC-AUC, MCC only generates a high score if the classifier performs well on both positive and negative classes, incorporating all four elements of the confusion matrix (TPs, TNs, FPs, and FNs) in its calculation. The low MCC observed in these models indicates poor performance when both classes are of equal importance. This demonstrates the limitations of ROC-AUC, which can present an overly optimistic view of an algorithm's performance when applied to imbalanced data. Moreover, a high MCC always corresponds to a high ROC AUC, and not vice versa[302]. Therefore, due to the dataset's imbalance, MCC has been selected as the primary metric to select the best-performing model in our study.

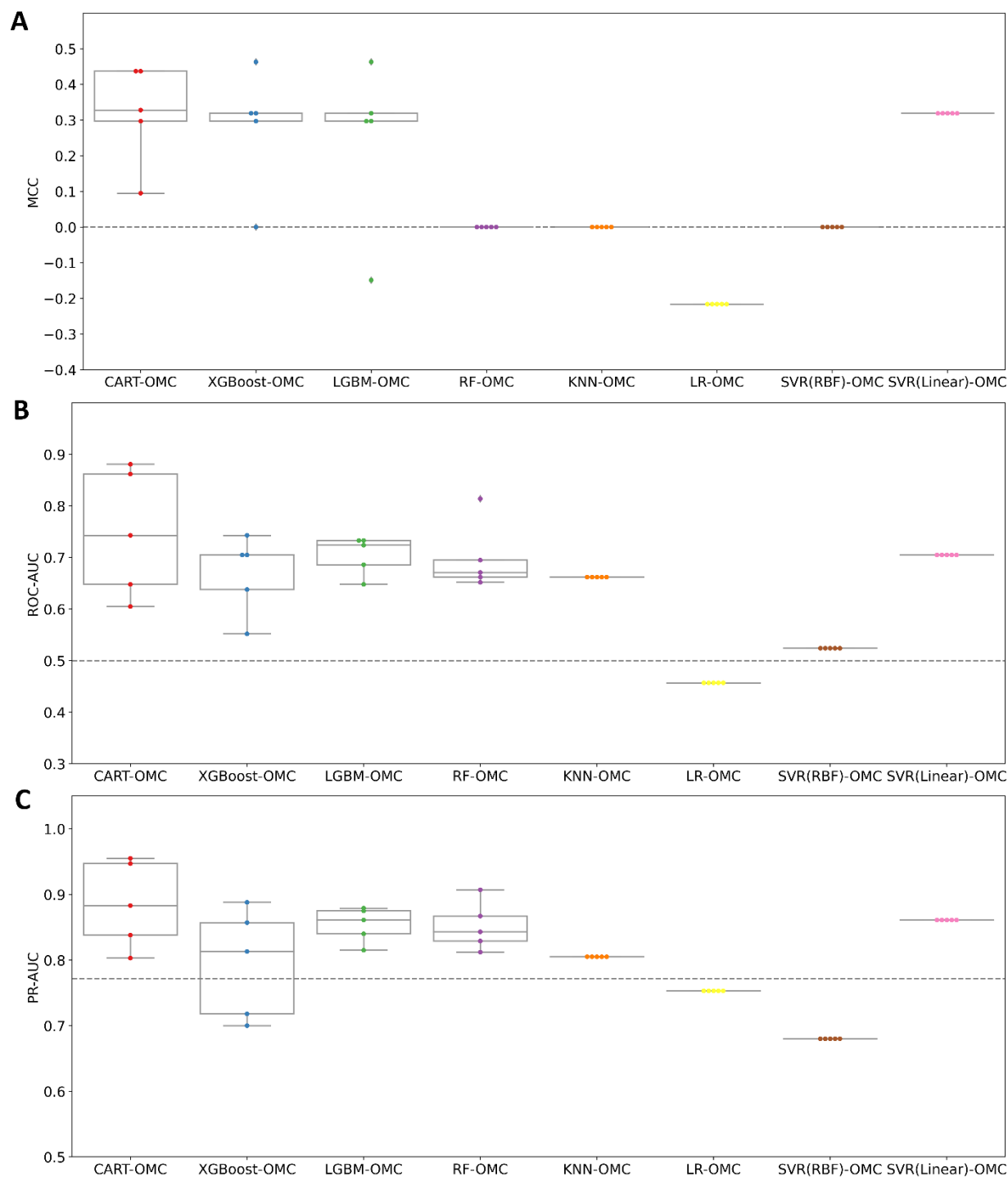


Figure 9 (related to Table 6). Five-repetition evaluation metrics for the validation dataset across eight ML models using gene expression profiles. Boxplots show (A) MCC, (B) ROC-AUC, and (C) PR-AUC, calculated using prediction on the validation dataset of the best OMC model. The best OMC model was selected based on the highest MCC from 10-fold CV on the discovery dataset, retrained on the entire discovery dataset, and tested on the validation dataset. To assess the robustness, five repetitions of CV were performed with

different random seeds. Each regression algorithms with OMC feature selection select only a small subset of informative features, thereby reducing the number of features in the high-dimensional gene expression dataset, facilitating model prediction. Horizontal dashed lines represent random-level performance: 0.0 for MCC, 0.5 for ROC-AUC, and 0.682 for PR-AUC.

Table 6. Five-repetition evaluation metrics for the validation dataset across eight ML models using gene expression profiles. The best OMC for each algorithm, identified via 10-fold CV on the discovery dataset, was tested on the validation dataset (the CV is performed five times changing the random seed). Eight regression algorithms with OMC feature selection select only a small subset of informative features, facilitating model prediction. Random-level performance is as follows: 0.0 for MCC, 0.5 for ROC-AUC, and 0.682 for PR-AUC. The best-performing model was selected based on the highest median MCC as highlighted in bold.

Gene expression profiles			
Algorithms	Highest MCC median MCC $\pm$ SD	Highest ROC-AUC Median ROC-AUC $\pm$ SD	Highest PR-AUC Median PR-AUC $\pm$ SD
CART-OMC	0.437 0.328 $\pm$ 0.14	0.881 0.743 $\pm$ 0.12	0.955 0.883 $\pm$ 0.07
XGB-OMC	0.463 0.319 $\pm$ 0.17	0.743 0.705 $\pm$ 0.08	0.888 0.813 $\pm$ 0.08
LGBM-OMC	0.463 0.297 $\pm$ 0.23	0.733 0.724 $\pm$ 0.04	0.879 0.861 $\pm$ 0.03
RF-OMC	0.0 0.0 $\pm$ 0.0	0.814 0.671 $\pm$ 0.07	0.907 0.843 $\pm$ 0.04
KNN-OMC	0.0 0.0 $\pm$ 0.0	0.662 0.662 $\pm$ 0.0	0.805 0.805 $\pm$ 0.0
LR-OMC	-0.216 -0.216 $\pm$ 0.0	0.457 0.457 $\pm$ 0.0	0.753 0.753 $\pm$ 0.0
SVR(RBF)	0.0 0.0 $\pm$ 0.0	0.524 0.524 $\pm$ 0.0	0.680 0.680 $\pm$ 0.0
SVR(Linear)	0.319 0.319 $\pm$ 0.0	0.705 0.705 $\pm$ 0.0	0.861 0.861 $\pm$ 0.0

Table 7. The list of 29 predictive genes derived from the best-performing CART-OMC model

Entrez gene	Gene symbol	Gene name
2822	GPLD1	glycosylphosphatidylinositol specific phospholipase D1
2835	GPR12	G protein-coupled receptor 12
3458	IFNG	interferon gamma
3627	CXCL10	C-X-C motif chemokine ligand 10
3682	ITGAE	integrin subunit alpha E
3766	KCNJ10	potassium inwardly rectifying channel subfamily J member 10

3805	KIR2DL4	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 4
3822	KLRC2	killer cell lectin like receptor C2
3823	KLRC3	killer cell lectin like receptor C3
3902	LAG3	lymphocyte activating 3
4283	CXCL9	C-X-C motif chemokine ligand 9
4319	MMP10	matrix metalloproteinase 10
5540	NPY4R	neuropeptide Y receptor Y4
6273	S100A2	S100 calcium binding protein A2
8302	KLRC4	killer cell lectin like receptor C4
9437	NCR1	natural cytotoxicity triggering receptor 1
10563	CXCL13	C-X-C motif chemokine ligand 13
10578	GNLY	granulysin
54436	SH3TC1	SH3 domain and tetratricopeptide repeats 1
56896	DPYSL5	dihydropyrimidinase like 5
57149	LYRM1	LYR motif containing 1
59067	IL21	interleukin 21
81691	REXO5	RNA exonuclease 5
83546	RTBDN	retbindin
84630	TTBK1	tau tubulin kinase 1
121643	FOXN4	forkhead box N4
140767	NRSN1	neurensin 1
257101	ZNF683	zinc finger protein 683
340562	SATL1	spermidine/spermine N1-acetyl transferase like 1

4.3 The impact of clinical data integration with gene expression profiles in predicting mUC patient response to atezolizumab

Here, we further looked into the potential improvement of atezolizumab response prediction in mUC patients by integrating both gene expression profiles and clinical data. The baseline clinical features included gender, ECOG performance status, PD-L1 protein expression status on TICs, TMB per megabase, TNB per megabase, prior receipt of intravesical administration of BCG immunotherapy status, prior receipt of platinum-based chemotherapy status, and pre-platinum sample collection status, as illustrated in Table 5. The evaluation metrics (MCC, ROC-AUC, and PR-AUC) on the validation dataset for eight ML models, developed in the discovery dataset using gene expression profiles and clinical data (GEP + clinical), are presented in Figure 10 and Table 8.

Among eight ML models evaluated, only two models demonstrated the predictive ability to distinguish between responders and non-responders, achieving an mMCC of at least 0.2: LR-OMC (with an mMCC of 0.316) and RF-OMC (with an mMCC of 0.289). In terms of ROC-AUC

and PR-AUC, KNN-OMC exhibited the highest ROC-AUC of 0.821 and PR-AUC of 0.867. Given the challenge of class imbalance in the dataset, we prioritized MCC as the primary metric for selecting the best-performing model. This choice was deliberate, as MCC is renowned for its robustness in handling imbalanced datasets. Therefore, LR-OMC was the best-performing model, achieving the highest mMCC of 0.316. This model employed a linear combination of 67 predictive genes alongside TMB per megabase and TNB per megabase (0.35% among 19,449 features) for predicting atezolizumab response, referred to as GEP + TMB + TNB (LR-OMC). The list of these 69 predictive features is shown in Table 9.

However, the prediction based on the combined clinical-gene expression profiles model (GEP + TMB + TNB (LR-OMC)) with mMCC of 0.316 and the baseline clinical features only (clinical (LR-OMC) with mMCC of 0.316, was worse than the model relying solely on gene expression profiles (GEP (CART-OMC) with mMCC of 0.328 (Figure 11). These findings suggest that there is no benefit in integrating patients' clinical data with gene expression profiles in predicting the atezolizumab response of mUC patients.

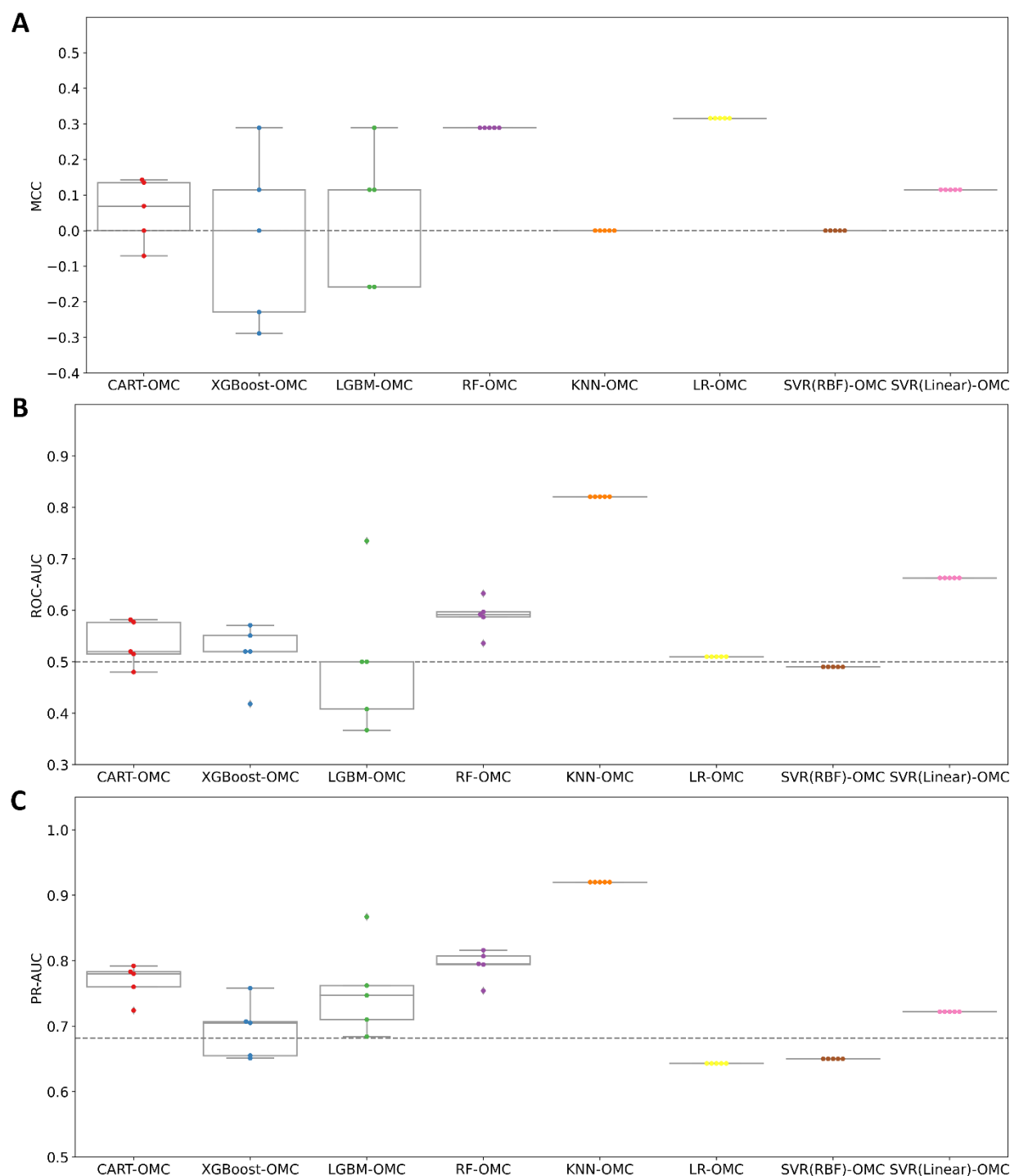


Figure 10 (related to Table 8). Five-repetition evaluation metrics for the validation dataset across eight ML models using integrated gene expression profiles with clinical data. Boxplots show (A) MCC, (B) ROC-AUC, and (C) PR-AUC, calculated using prediction on the validation dataset of the best OMC model. The best OMC model was selected based on the highest MCC from 10-fold CV on the discovery dataset, retrained on the entire discovery dataset, and tested on the validation dataset. To assess the robustness, five repetitions of CV

were performed with different random seeds. Each regression algorithms with OMC feature selection select only a small subset of informative features, facilitating model prediction. Horizontal dashed lines represent random-level performance: 0.0 for MCC, 0.5 for ROC-AUC, and 0.682 for PR-AUC.

Table 8. Five-repetition evaluation metrics for the validation dataset across eight ML models using integrated gene expression profiles with clinical data. The best OMC for each algorithm, identified via 10-fold CV on the discovery dataset, was tested on the validation dataset (the CV is performed five times changing the random seed). Eight regression algorithms with OMC feature selection select only a small subset of informative features, facilitate model prediction. Random-level performance is as follows: 0.0 for MCC, 0.5 for ROC-AUC, and 0.682 for PR-AUC. The best-performing model was selected based on the highest median MCC as highlighted in bold.

Integrated gene expression profiles with clinical data			
Algorithms	Highest MCC median MCC $\pm$ SD	Highest ROC-AUC Median ROC-AUC $\pm$ SD	Highest PR-AUC Median PR-AUC $\pm$ SD
CART-OMC	0.143 0.069 $\pm$ 0.09	0.582 0.520 $\pm$ 0.04	0.792 0.780 $\pm$ 0.03
XGB-OMC	0.289 0.0 $\pm$ 0.24	0.571 0.520 $\pm$ 0.06	0.758 0.705 $\pm$ 0.04
LGBM-OMC	0.289 0.115 $\pm$ 0.19	0.735 0.500 $\pm$ 0.14	0.867 0.747 $\pm$ 0.07
RF-OMC	0.289 0.289 $\pm$ 0.0	0.633 0.592 $\pm$ 0.03	0.816 0.795 $\pm$ 0.02
KNN-OMC	0.0 0.0 $\pm$ 0.0	0.821 0.0 $\pm$ 0.0	0.920 0.0 $\pm$ 0.0
LR-OMC	0.316 0.316 $\pm$ 0.0	0.510 0.510 $\pm$ 0.0	0.643 0.643 $\pm$ 0.0
SVR(RBF)	0.0 0.0 $\pm$ 0.0	0.490 0.490 $\pm$ 0.0	0.650 0.650 $\pm$ 0.0
SVR(Linear)	0.115 0.0 $\pm$ 0.0	0.663 0.663 $\pm$ 0.0	0.722 0.722 $\pm$ 0.0

Table 9. The list of 69 predictive features (67 genes, TMB per megabase, and TNB per megabase) derived from the best-performing LR-OMC model

Entrez gene	Gene symbol	Gene name
356	FASLG	Fas ligand
1521	CTSW	cathepsin W
2012	EMP1	epithelial membrane protein 1
2114	ETS2	ETS proto-oncogene 2, transcription factor
2150	F2RL1	F2R like trypsin receptor 1

2196	FAT2	FAT atypical cadherin 2
2999	GZMH	granzyme H
3458	IFNG	interferon gamma
3627	CXCL10	C-X-C motif chemokine ligand 10
3682	ITGAE	integrin subunit alpha E
3752	KCND3	potassium voltage-gated channel subfamily D member 3
3766	KCNJ10	potassium inwardly rectifying channel subfamily J member 10
3804	KIR2DL3	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 3
3805	KIR2DL4	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 4
3811	KIR3DL1	killer cell immunoglobulin like receptor, three Ig domains and long cytoplasmic tail 1
3812	KIR3DL2	killer cell immunoglobulin like receptor, three Ig domains and long cytoplasmic tail 2
3822	KLRC2	killer cell lectin like receptor C2
3852	KRT5	keratin 5
3902	LAG3	lymphocyte activating 3
4283	CXCL9	C-X-C motif chemokine ligand 9
4319	MMP10	matrix metalloproteinase 10
4624	MYH6	myosin heavy chain 6
4818	NKG7	natural killer cell granule protein 7
5540	NPY4R	neuropeptide Y receptor Y4
6273	S100A2	S100 calcium binding protein A2
6352	CCL5	C-C motif chemokine ligand 5
6846	XCL2	X-C motif chemokine ligand 2
6866	TAC3	tachykinin precursor 3
6947	TCN1	transcobalamin 1
8302	KLRC4	killer cell lectin like receptor C4
8416	ANXA9	annexin A9
9033	PKD2L1	polycystin 2 like 1, transient receptor potential cation channel
9047	SH2D2A	SH2 domain containing 2A
9437	NCR1	natural cytotoxicity triggering receptor 1
9620	CELSR1	cadherin EGF LAG seven-pass G-type receptor 1
10563	CXCL13	C-X-C motif chemokine ligand 13
10578	GNLY	granulysin
26499	PLEK2	pleckstrin 2
30009	TBX21	T-box transcription factor 21
54967	CT55	cancer/testis antigen 55
56097	PCDHGC5	protocadherin gamma subfamily C, 5
56098	PCDHGC4	protocadherin gamma subfamily C, 4

56100	PCDHGB6	protocadherin gamma subfamily B, 6
56105	PCDHGA11	protocadherin gamma subfamily A, 11
56106	PCDHGA10	protocadherin gamma subfamily A, 10
56120	PCDHGB8P	protocadherin gamma subfamily B, 8 pseudogene
56121	PCDHB15	protocadherin beta 15
56122	PCDHB14	protocadherin beta 14
56124	PCDHB12	protocadherin beta 12
56127	PCDHB9	protocadherin beta 9
56895	AGPAT4	1-acylglycerol-3-phosphate O-acyltransferase 4
59067	IL21	interleukin 21
78999	LRFN4	leucine rich repeat and fibronectin type III domain containing 4
79156	PLEKHF1	pleckstrin homology and FYVE domain containing 1
84033	OBSCN	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF
85301	COL27A1	collagen type XXVII alpha 1 chain
121643	FOXN4	forkhead box N4
126520	PLK5	polo like kinase 5 (inactive)
140767	NRSN1	neurensin 1
153769	SH3RF2	SH3 domain containing ring finger 2
162461	TMEM92	transmembrane protein 92
164118	TTC24	tetratricopeptide repeat domain 24
256957	HEATR9	HEAT repeat containing 9
257101	ZNF683	zinc finger protein 683
401427	OR2A7	olfactory receptor family 2 subfamily A member 7
497189	TIFAB	TIFA inhibitor
730249	ACOD1	aconitate decarboxylase 1
TMB per megabase		
TNB per megabase		

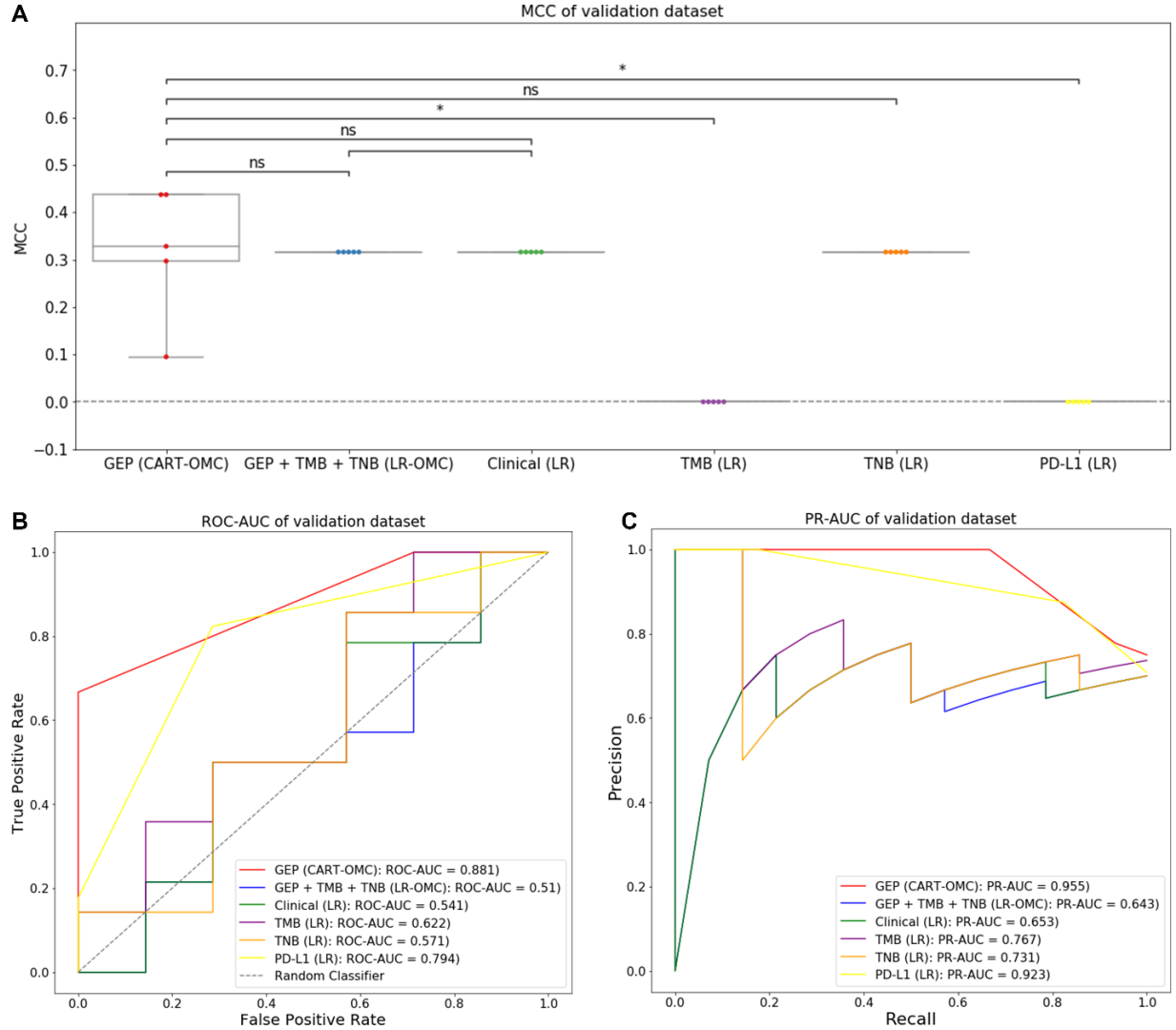


Figure 11. Five-repetition evaluation on the validation dataset of the best-performing models versus those using TMB, TNB, or PD-L1 immunotherapy response biomarkers. (A) MCCs across five model runs (a different random seed per repetition), (B) ROC-AUC for the run with the median MCC, and (C) PR-AUC for the run with the median MCC are shown. Each metric is calculated using predictions on the validation dataset of ML model trained on the discovery dataset with: gene expression profiles (GEP), integrated GEP and clinical data (GEP + clinical), referred to GEP + TMB + TNB, clinical data only (Table 5) (clinical), TMB per megabase (TMB), TNB per megabase (TNB), and PD-L1 protein expression on tumour-infiltrating immune cells (PD-L1). Each dot represents one repetition, visualised with boxplots. Random-level performances are indicated by the horizontal dashed lines: 0.0 for MCC, 0.5 for ROC-AUC, 0.682 for PR-AUC. A two-sided Welch's t-test was used for statistical comparisons between models. P-value significance is denoted as follows: "ns" (not significant, $0.05 < p \leq 1.00$), * ($0.01 < p \leq 0.05$), ** ($0.001 < p \leq 0.01$), *** ($0.0001 < p \leq 0.001$), and **** ($p \leq 0.0001$).

4.4 Comparison of the best-performing models to clinical immunotherapy response biomarkers

Recent studies have reported an association between TMB and TNB with the prediction of immunotherapy response across various cancer types, including mUC. High levels of TMB and TNB have frequently been associated with increased sensitivity to immunotherapeutic drugs[164], [308]–[311]. In addition, PD-L1 protein expression on either tumour cells or TICs is a target for atezolizumab and has emerged as an early biomarker to be evaluated in immunotherapy clinical trials[312], [313]. To compare predictive performance at this task, we trained an LR model using normalised values of TMB per megabase, TNB per megabase, and PD-L1 protein expression status on TICs. We compared its predictive performance with our best-performing models: GEP (CART-OMC) and GEP + TMB + TNB (LR-OMC). The evaluation metrics (MCC, ROC-AUC, and PR-AUC) on the validation dataset for each model, developed in the discovery dataset, are shown in Figure 11.

Our findings reveal that the CART-OMC model, which nonlinearly combines 29 predictive genes, achieved the highest MCC of 0.437, with a corresponding ROC-AUC of 0.881 and PR-AUC of 0.955. This model significantly outperformed the LR models based solely on TMB (MCC of 0.0, ROC-AUC of 0.622, PR-AUC of 0.767), TNB (MCC of 0.316, ROC-AUC of 0.571, PR-AUC of 0.731), and PD-L1 protein expression on TICs (MCC of 0.0, ROC-AUC of 0.794, PR-AUC of 0.923), although the performance difference between GEP (CART-OMC) and TNB (LR-OMC) was not significant. The results suggest that the prediction-based 29 predictive genes are more robust than the use of a single immunotherapy biomarker in predicting atezolizumab response in mUC patients.

4.5 Assessing the robustness of the best-performing model in predicting mUC patient response to atezolizumab

After identifying GEP (CART-OMC) as the best-performing model for predicting atezolizumab response in mUC patients, we examined the impact of training data sizes on its predictive performance. Using the discovery dataset, we carried out five-fold CV (5CV), and leave one out CV (LOOCV) experiments, in addition to the 10CV runs from the previous section, with training data proportions ranging from 80% in 5CV to nearly complete at 99% in LOOCV. In each scenario, CART-OMC models were trained on gene expression profiles, retrained on the full discovery dataset, and evaluated on the validation dataset. Each CV experiment was repeated five times with different random seeds. Figure 12 presents the MCC across five repetitions, showing a trend of performance improvement as the training dataset size increased.

To evaluate whether the model's performance was better than what would be predicted by random chance, we conducted a permutation test comparing the best-performing CART-OMC model to a baseline. This involved replicating all CV runs exactly in the same manner using dataset with permuted class labels, where the labels (CR, PR, SD, and PD) were randomly generated while retaining the original input features. By disrupting the features-label relationship, this approach established a baseline of performance expected by chance.

Results showed that models trained on permuted datasets consistently performed at near-random levels and were worse than CART-OMC models trained on original data in all cases (Figure 12). However, in case of LOOCV, despite using the maximum amount of data for training, no significant difference was observed between the original and permuted versions.

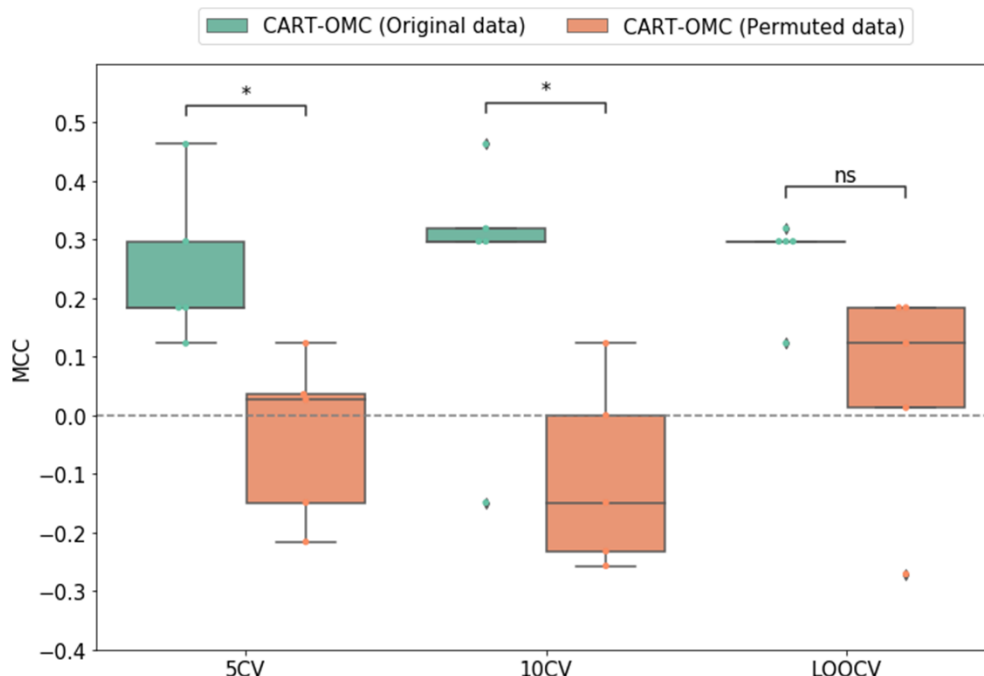


Figure 12. CV performance of the best-performing CART-OMC model compared to its permutation model. Boxplots illustrate the distribution of MCCs across five repetitions of CART-OMC model evaluations on the validation dataset. CART-OMC models were trained on the discovery dataset with either original class labels (green) or permuted class labels (orange), where the labels (CR, PR, SD, and PD) were randomly generated while keeping the original input features unchanged. Each model was evaluated through 5CV, 10CV, and LOOCV, with training data proportions ranging from 80% (5CV) to 99% (LOOCV), and was subsequently tested on the validation dataset. Predictive performance (original vs. permuted) was compared within each CV method. Statistical significance was assessed using two-sided Welch's t-tests, with “ns” indicating no significance ($0.05 < p \leq 1.00$) and “*” indicating significance ($0.01 < p \leq 0.05$). Random-level performances are indicated by the horizontal dashed lines at MCC of 0.0. The CART-OMC shows a trend of improved predictive performance with increasing training data size and consistently outperforms permutation models across all CV methods.

4.6 Analysis of cluster heat map using the 29 predictive genes

Here, we analysed a cohort of 374 mUC patients whose gene expression profiles were available across both discovery ($N = 348$) and validation ($N = 26$) datasets, including those with both known and unknown responses to the atezolizumab. We aimed to determine whether the subset of 298 gene expression-profiled mUC patients with the known response

of the discovery dataset, used during the development phase of a predictive model, could represent all the remaining patients, which could serve as the potential test set. Figure 13 illustrates the clustering of 374 mUC patients based on their expression of 29 predictive genes identified by the best-performing CART-OMC model. We employed agglomerative hierarchical clustering with Ward's linkages and visualized the results in a dendrogram heatmap using Seaborn (version 0.11.2). Samples were colour-coded to denote treatment responses (green for responders, orange for non-responders, and black for missing records) and dataset type (grey for discovery and yellow for validation). The normalised expression levels of 29 predictive genes were visualised in a gradient from green (downregulation) to blue (upregulation). The result shows that patients in the validation dataset and those with unknown atezolizumab response records were evenly distributed across the clusters. This even distribution indicates that the CART-OMC model's predictions were not limited to a specific patient subgroup. Instead, the model's applicability domain was broad and potentially provided accurate predictions across the entire patient cohorts.

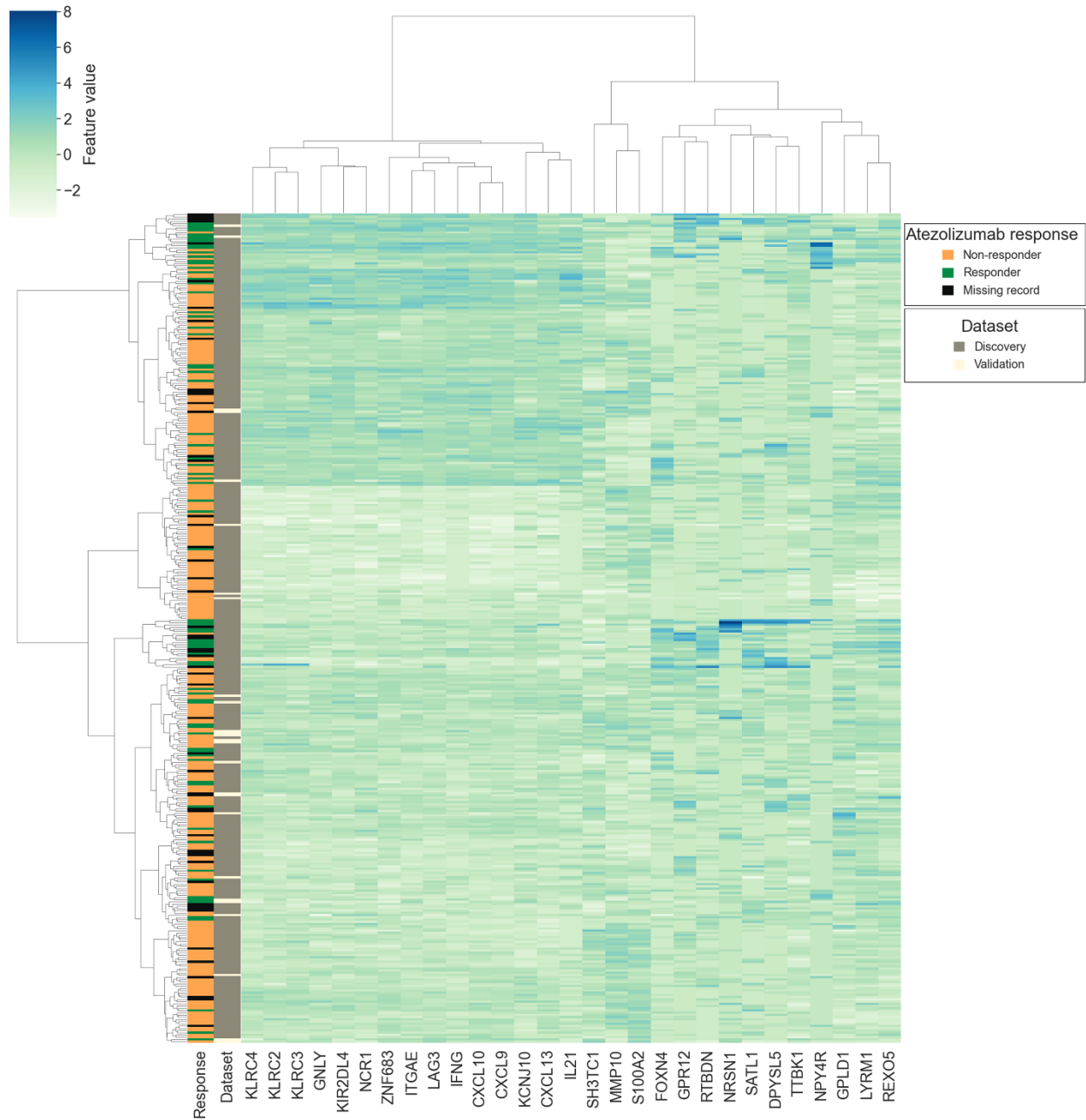


Figure 13. Unsupervised clustering heatmap of 29 predictive genes in atezolizumab-treated mUC patients. The heatmap visualised unsupervised clustering of normalised log2-transformed gene expression data for 29 predictive genes from the GEP (CART-OMC) model across 374 mUC patients in the discovery (N = 348) and validation (N = 26) datasets. Rows represent patients, while columns represent genes, with hierarchically clustered to group similar profiles. On the left, colour bars indicate patient response to atezolizumab: responders (green), non-responders (orange), and missing response records (black). Dataset origins are marked by grey (discovery) and yellow (validation).

4.7 Feature importance of the 29 predictive genes

We next investigated the importance of 29 predictive genes identified by the best-performing CART-OMC model, to estimate the contribution of each gene to model predictive accuracy. In the case of CART-OMC, the feature importance was computed as the reduction of the Gini impurity of each feature in DT algorithms[314]. A higher Gini impurity indicates that a gene is considered to be more important and contributes significantly to partitioning the data into more homogenous groups, which correlates with improved predictive accuracy of the model. The feature importance plot, based on the Gini impurity and inherently built into the CART algorithm in sci-kit learn, was visualised using the bar chart depicted in Figure 14.

Our findings revealed that, at the top of the chart, the CXCL9 gene was the most important feature in predicting the response to atezolizumab, with an importance score close to 0.12. In addition, GPR12 was the second most important gene, showing a level of importance similar to FOXP1. The importance of these genes tapered off further down the chart; however, some other genes had near-zero importance. These genes contributed less to the CART-OMC model prediction compared to top features, but they might still carry valuable information when combined with other features.

However, it is important to note that the feature importance plot solely reflects the contribution of each gene but does not indicate whether a gene has a positive or negative impact on the model predictions. To investigate this question, we employed Shapley Additive Explanations (SHAP) values to relate gene expression levels with their impact on the model's predictions. In brief, this method involved assigning an importance score to each gene, represented as the mean absolute SHAP value calculated across all patients. SHAP values were graphically depicted through colourful horizontal lines composed of dots, where the colour of each dot represented the normalised gene expression levels for that patient, ranging from low (blue) to high (red). In Figure 15, genes were ranked by their mean absolute SHAP values across all patients, listed alongside each gene's name. A gene with a positive SHAP value indicates a higher likelihood of the patient being predicted as a non-responder to atezolizumab. In contrast, a negative SHAP value suggests a greater probability of being predicted as a responder.

Similar to the feature importance plot, the results showed that the CXCL9 gene remained the most important feature in the model's prediction, having the highest mean absolute SHAP values of 0.245. Higher CXCL9 expression was associated with predicted responders. This trend was also observed in some other genes, where higher expression levels generally corresponded with predicted responders.

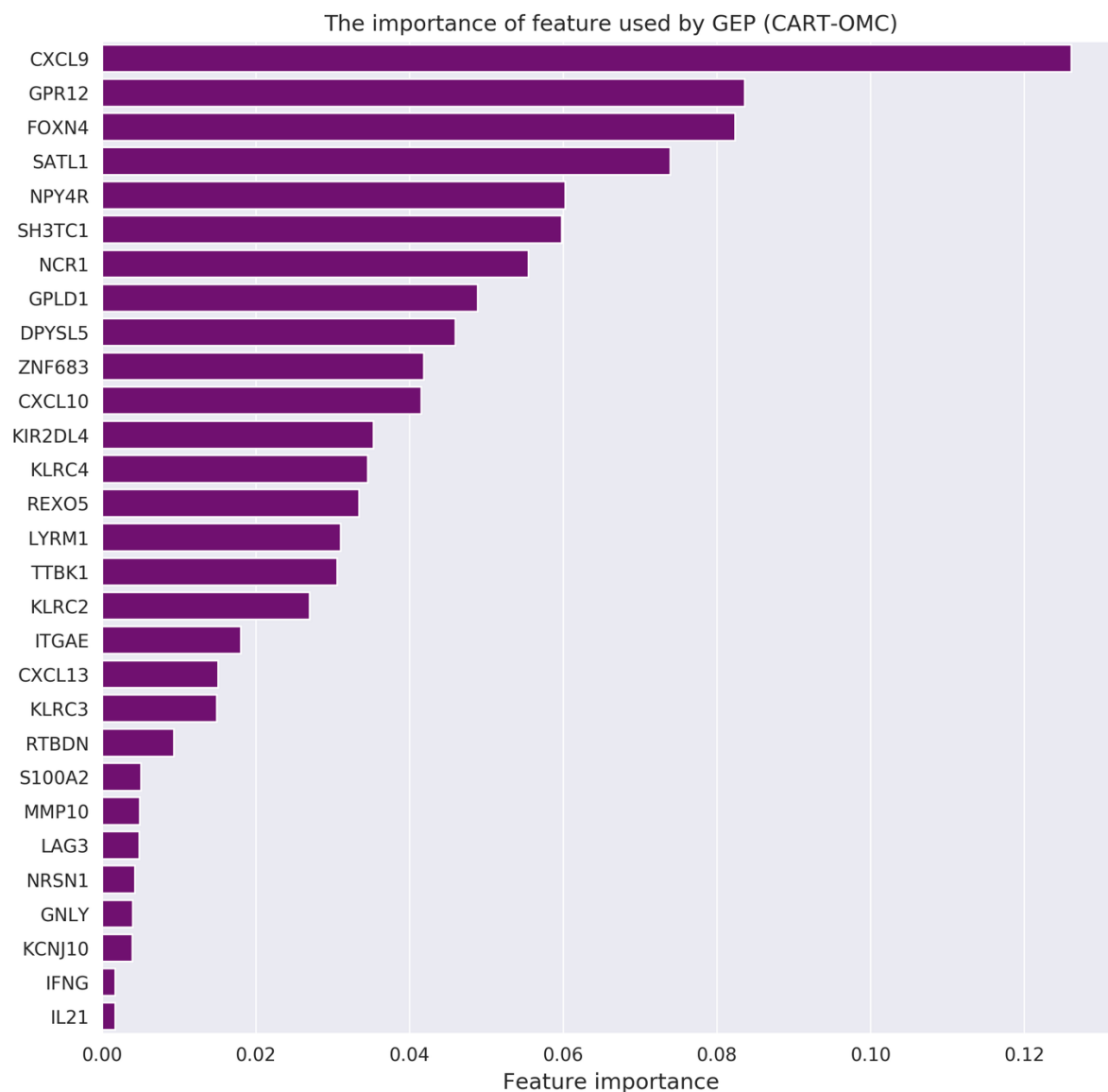


Figure 14. Feature importance plot of 29 predictive genes from the best-performing CART-OMC model for predicting atezolizumab responses in mUC patients. Features are ranked by importance scores, calculated as the reduction of Gini impurity by each feature of the decision tree. Lower important scores indicate less contributions to the model's predictive performance. CXCL9 is the most important feature used by the CART-OMC model.

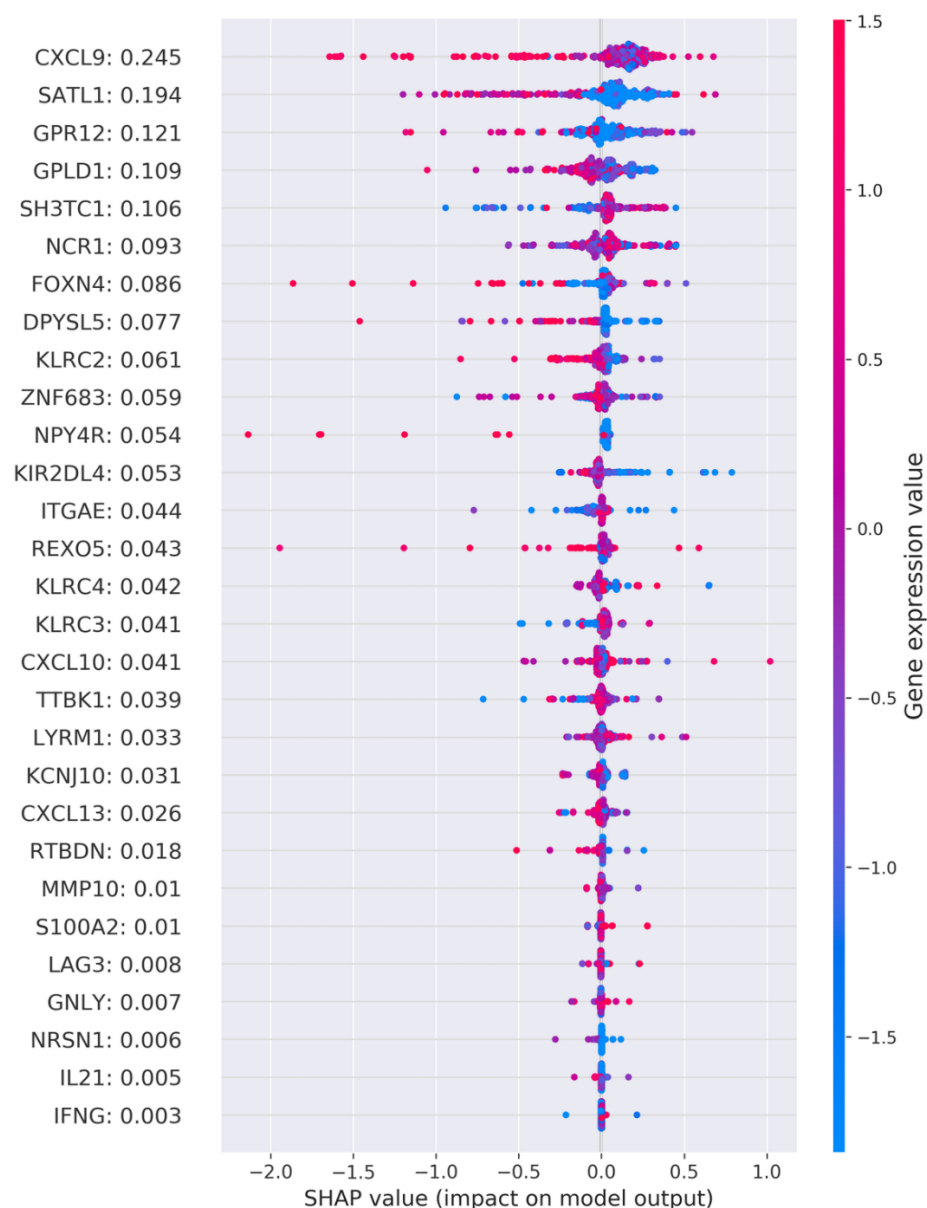


Figure 15. Model interpretation using SHAP. The SHAP summary plot illustrates the importance of 29 predictive genes from the best-performing CART-OMC model in predicting atezolizumab response for 320 mUC patients (75 responders, 245 non-responders) from the merged discovery and validation datasets. Genes are ranked by their mean absolute SHAP values across all patients (displayed next to each gene), with larger values indicating higher important. Each horizontal lines of dots, where the colour of a dot represents the normalised gene expressions (Z-score) for that patient, ranging from low (blue) to high (red). The x-axis shows SHAP values for individual patients. The further a value is away from the vertical centre line at $x = 0$, the larger its impact on the model's prediction. Negative SHAP values to the left increase the chance of being a responder, while positive SHAP values to the right increase the chance of being a non-responder. CXCL9 is the most important feature in the model's prediction, with higher expressions associated with predicted responders.

4.8 Differential expression of the 29 predictive genes

The box plot shown in Figure 16 illustrates the differential expression of 29 predictive genes, identified by the best-performing CART-OMC model, between responders and non-responders across two different datasets: a discovery dataset and a validation dataset. In the discovery dataset (68 responders and 230 non-responders), the comparative analysis of normalised log2-transformed gene expression levels showed that 27 out of 29 genes exhibit statistically significant differences ($P < 0.05$) in expression between responders and non-responders, with higher expression observed in responders. Except for MMP10 and SH3TC1 showed significantly higher expression in non-responders (Figure 16A). A similar trend toward high expression in responders was observed in the validation dataset (7 responders and 15 non-responders), although with fewer statistically significant differences (Figure 16B).

This pattern suggests a robust immune response in responders, characterised by the upregulation of genes involved in the response to atezolizumab and its mechanism such as cellular defence response, immune cell-mediated immunity and cytotoxicity, cell killing, and regulation of immune system process and immune response, chemokine and cytokine receptor binding pathways, as detailed in the subsequent section on Wikipathway cancer and GO pathway enrichment analysis.

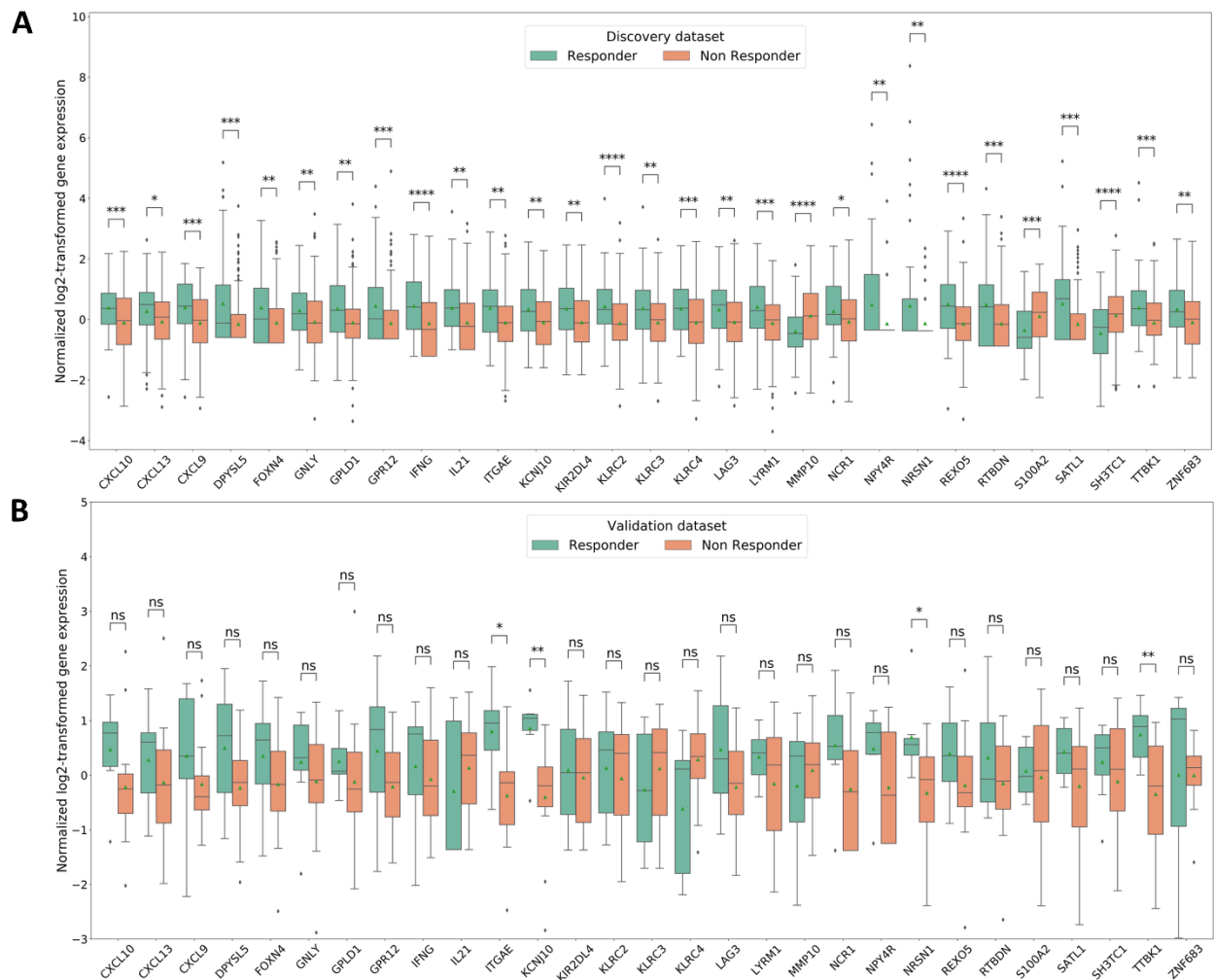


Figure 16. Boxplot showing the expression pattern of 29 predictive genes from the best-performing CART-OMC model between responders and non-responders. The y-axis represents the normalised log₂-transformed gene expression of 29 genes obtained from the GEP (CART-OMC) model between responders and non-responders in the discovery (A) and validation (B) datasets. Each boxplot shows the median (line) and mean (green triangle) expression for a gene, alongside P-values from a two-sided Welch's t-test evaluating mean differences between the two groups. Statistical significance P-value is annotated by stars, where non-significant "ns" means $0.05 < p \leq 1.00$, "*" means $0.01 < p \leq 0.05$, "***" means $0.001 < p \leq 0.01$, "****" means $p \leq 0.0001$, and "*****" means $p \leq 0.0001$. A P-value less than 0.05 is considered as a significant difference in expression levels between the two groups.

4.9 PPI network analysis of the 29 predictive genes

We subsequently analysed both the direct (physical) and indirect (functional) associations among 29 predictive genes and the CD274 gene (encoding PD-L1, the target of atezolizumab), hence could influence atezolizumab response in mUC patients. We performed PPI analysis using the Search Tool for Retrieval of Interacting Genes/Proteins (STRING)

database version 11.5 (<https://string-db.org>)[315], which aggregates known and predicted protein interactions derived from various sources including genomic context, high-throughput experiments, co-expression and the literature.

The resulting PPI network, illustrated in Figure 17, comprises 30 nodes (protein-coding genes) and 54 edges (interactions), with significant PPI enrichment (p-value < 0.0001). This indicates that the input proteins have more interactions among themselves than what would be expected for a randomly drawn set of proteins of the same size and degree distribution from the genome. The network revealed the known and predicted interactions were observed among 15 genes, while the remaining genes appeared as disconnected nodes. Interestingly, eight genes (CXCL13, CXCL10, CXCL9, IFNG, NCR1, GNLY, LAG3, and ITGAE) were directly connected to CD274, suggesting that these gene's activities may be influenced by CD274 signalling upon the binding of atezolizumab and its target, PD-L1. In addition, IFNG interacted with several other genes, including CD274, supporting its potential key role in inflammatory immune responses [316]. Overall, these findings highlight molecular interactions that might contribute to atezolizumab response, suggesting potential targets to improve treatment effectiveness in mUC patients.

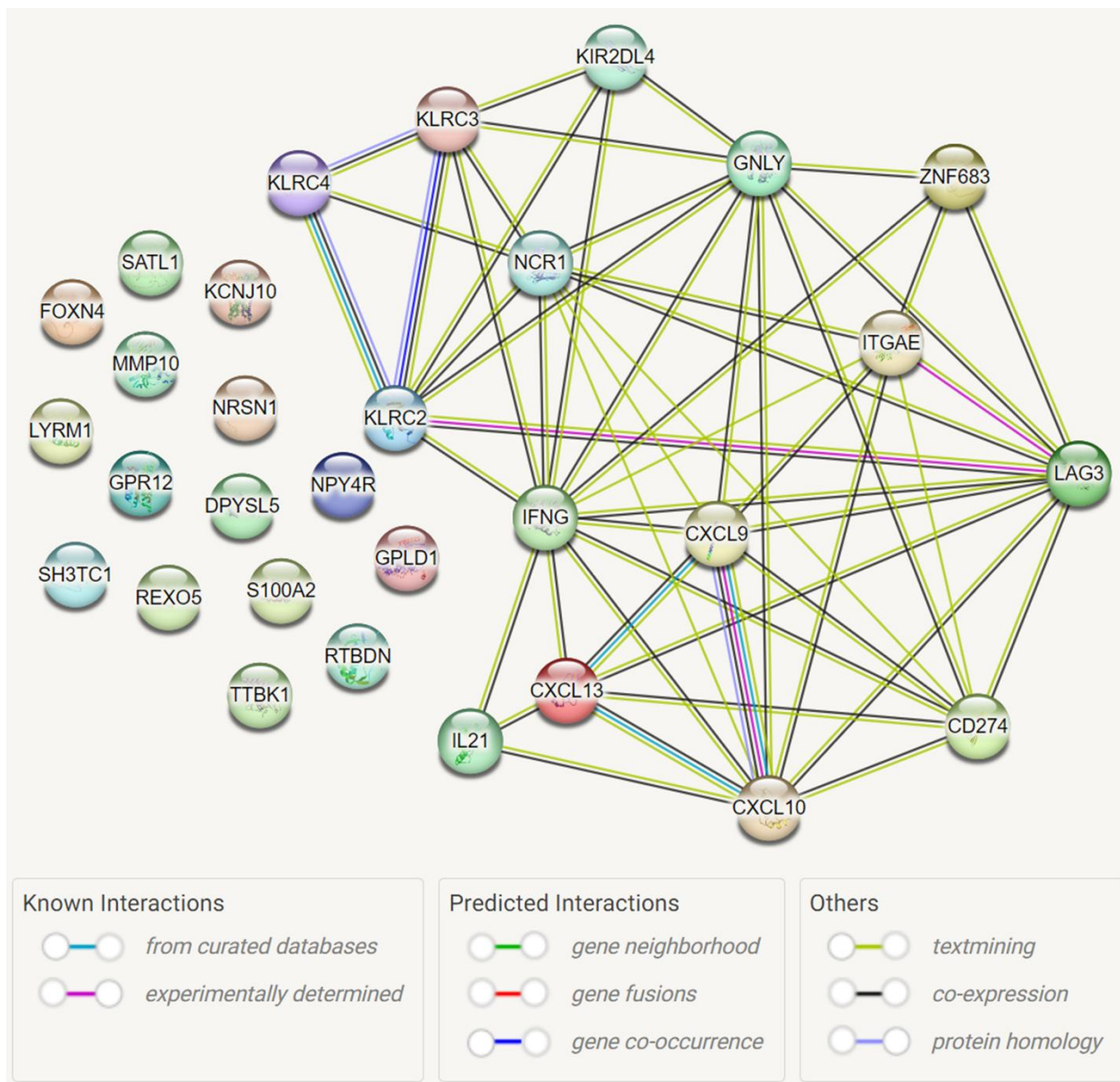


Figure 17. PPI network analysis. The 29 predictive genes identified from the best-performing model (CART-OMC), along with CD274 (encoding PD-L1, atezolizumab target), were analysed using the STRING database (<https://string-db.org/>). The analysis was limited to Homo sapiens with a minimum confidence interaction score > 0.4. This yielded a PPI network comprising 30 nodes and 54 edges, where nodes represent protein-coding genes and edges indicate known or predicted interactions, as shown in the legend. A PPI enrichment p-value of 0.0001 was obtained, suggesting the protein input has more interactions among themselves than what would be expected for a set of proteins of the same size and degree distribution drawn at random from the genome.

4.10 Wikipathway cancer and GO pathway enrichment analysis of the 29 predictive genes

We next identified biological pathways significantly enriched with a subset of 29 predictive genes using the cancer-related repository of WikiPathways (WikiPathway cancer) and GO database in the WEB-based Gene SeT Analysis Toolkit (WebGestalt) 2024 (www.webgestalt.org)[317]. By using the WikiPathway cancer database, one cancer-associated pathway was significantly enrichment (Figure 18 and Table 10). Specifically, three genes (IFNG, CXCL9, and CXCL10) were found to significantly enrich ($FDR \leq 0.05$) in the type II interferon signalling (IFNG) pathway, which is associated with the mechanism of action of atezolizumab. These three genes encode for IFNG cytokine, CXCL9 and CXCL10 chemokines produced by T-cells, natural killer (NK) cells, and macrophages to attach and activate immune cells for eliminating cancer cells when PD1/PD-L1 interaction is blocked by atezolizumab[217], [318]. Therefore, increased intratumoral production of IFNG, CXCL3, and CXCL10 directly suppresses tumour cell proliferation and increases antigen presentation, leading to better recognition and elimination of cancer cells[157]. In addition, IFNG is considered as a crucial biomarker for predicting responses to immunotherapy. Ayers et al.[319] conducted an analysis of gene expression profiles from tumour tissue samples across various cancers, including gastric cancer, metastatic melanoma, and head and neck squamous cell carcinoma, and found that patients who responded favourably to anti-PD1 therapy exhibited higher expression scores for IFNG-related genes compared to non-responders. The authors suggested that the identified IFNG signature (IDO1, CXCL10, CXCL9, HLA-DRA, STAT1, and IFNG) could serve as a predictive biomarker for the clinical response to ICIs.

GO pathways across three terms, including biological processes, cellular components, and molecular functions, significantly enriched ($FDR \leq 0.05$) with a subset of 29 genes were presented in Figure 18 and detailed in Table 10. The enriched biological process GO terms were involved in the immune system's response to cancer, such as cellular defence response, immune cell-mediated immunity and cytotoxicity, cell killing, and regulation of the immune system process and immune response. These pathways play a crucial role in initiating and sustaining anti-tumour activities within the TME, involving the production of cytokines, inflammation, and recruiting T-cells and other immune cells to the TME[320]. Similar to the enriched pathways in molecular function terms, chemokine and cytokine receptor binding pathways lie in their ability to recruit and activate immune cells to target and attack cancer cells, promote the production of inflammatory cytokines, and induce an anti-tumour immune response. The binding of CXCL9, CXCL10, and CXCL11 to their receptor CXCR3 induces the migration of T-cells and NK cells into the TME, where they can recognise and eliminate cancer cells[217]. Furthermore, the cellular component term showed the external side of the plasma membrane was enriched with five genes (CXCL10, CXCL9, ITGAE, LAG3, and RTBDN), indicating the interaction site between PD-L1, which is expressed on the surface of the tumour cells or TICs, and atezolizumab[321].

Altogether, these findings highlight that our best-performing model (CART-OMC) could identify 29 predictive genes that were cancer-related and showed potentially associated with the response to atezolizumab and its mechanism. Changes in their expression levels may contribute to atezolizumab resistance and cancer progression.

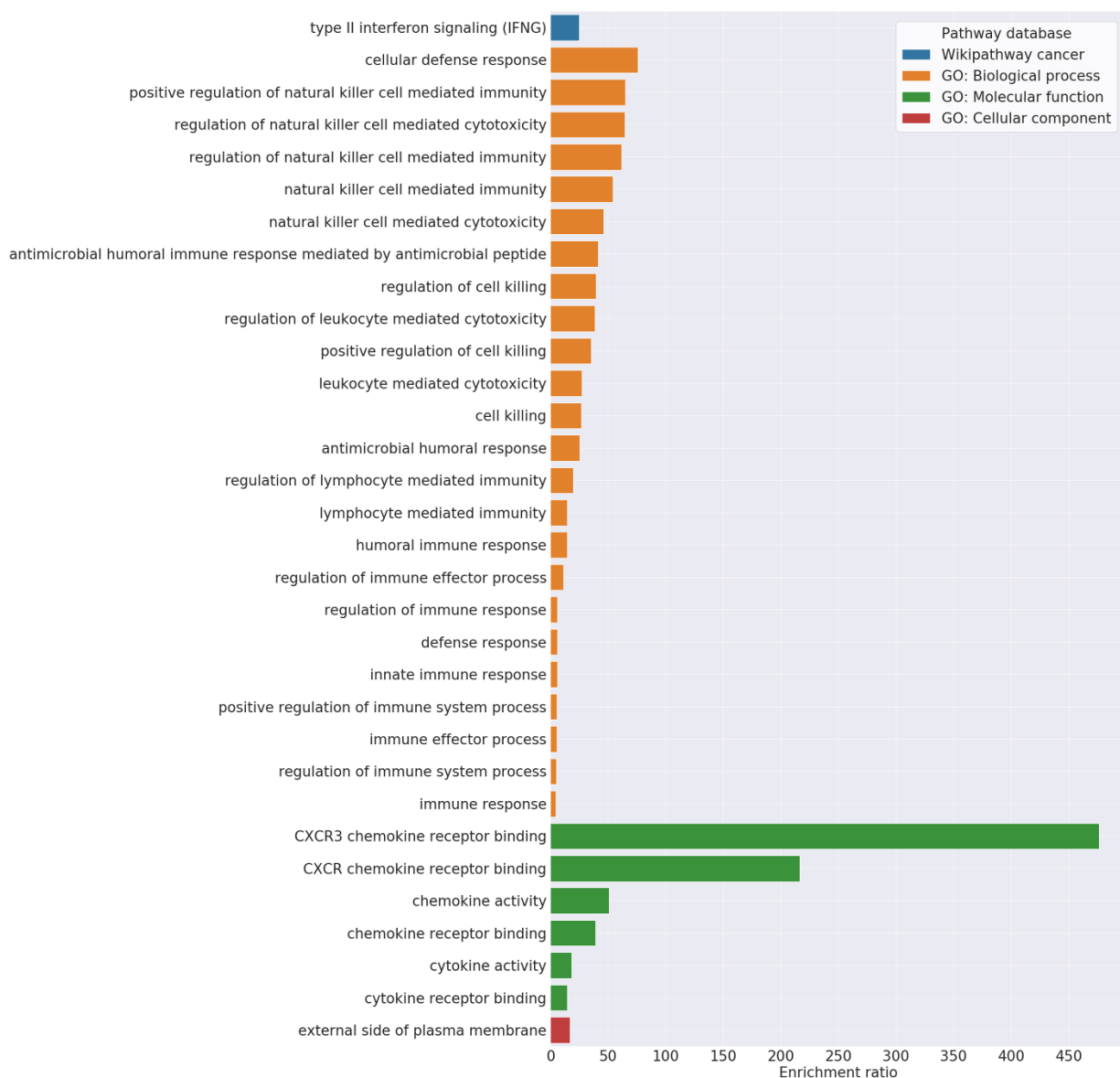


Figure 18 (related to Table 10). Wikipathway cancer and GO pathway enrichment analysis of 29 predictive genes for predicting atezolizumab response in mUC patients. Pathway enrichment analysis of 29 genes associated with predicting atezolizumab response in mUC patients was performed using the cancer-related repository of WikiPathways (WikiPathway cancer) and the GO database in WebGestalt. The bar chart summarises 32 biological pathways that are significantly ($FDR \leq 0.5$) enriched with different subsets of 29 genes in the Wikipathway cancer (blue) and the GO terms, including biological process (orange), molecular function (green), cellular component (red), and presented on the y-axis.

The x-axis represents the enrichment ratio, calculated as the number of genes overlaps with the reference gene set in a given pathway. Our best-performing model (CART-OMC) could identify 29 genes that are primarily cancer-related and known to play a role in the atezolizumab response and its mechanism, including cellular defence response, immune cell-mediated immunity and cytotoxicity, cell killing, and regulation of immune system process and immune response, chemokine and cytokine receptor binding pathways.

Table 10. The Wikipathway cancer and GO pathway enrichment analysis of 29 predictive genes from the best-performing models (CART-OMC) in predicting atezolizumab response in mUC patients (a hypergeometric test corrected with Benjamin-Hochberg)

Reference database	Enriched pathway	#gene in the pathway	#gene overlap	Gene represented in the pathway	Enrichment ratio	P-Value	FDR
Wikipathway cancer	Type II interferon signaling (IFNG)	37	3	CXCL10, CXCL9, IFNG	24.865	0.000	0.011
GO: biological process	cellular defense response	55	6	CXCL9, GNLY, KIR2DL4, KLRC2, KLRC3, NCR1	75.745	0.000	0.000
	positive regulation of natural killer cell mediated immunity	32	3	IL21, KIR2DL4, LAG3	65.094	0.000	0.008
	regulation of natural killer cell mediated cytotoxicity	43	4	IL21, KIR2DL4, LAG3, NCR1	64.589	0.000	0.001
	regulation of natural killer cell mediated immunity	45	4	IL21, KIR2DL4, LAG3, NCR1	61.719	0.000	0.001
	natural killer cell mediated immunity	64	5	IL21, KIR2DL4, KLRC2, LAG3, NCR1	54.245	0.000	0.000
	natural killer cell mediated cytotoxicity	60	4	IL21, KIR2DL4, LAG3, NCR1	46.289	0.000	0.002
	antimicrobial humoral immune response mediated by antimicrobial peptide	67	4	CXCL10, CXCL13, CXCL9, GNLY	41.453	0.000	0.002
	regulation of cell killing	88	5	IFNG, IL21, KIR2DL4, LAG3, NCR1	39.451	0.000	0.000
	regulation of leukocyte mediated cytotoxicity	72	4	IL21, KIR2DL4, LAG3, NCR1	38.574	0.000	0.003
	positive regulation of cell killing	59	3	IFNG, IL21, LAG3	35.305	0.000	0.032

	leukocyte mediated cytotoxicity	102	4	IL21, KIR2DL4, LAG3, NCR1	27.229	0.000	0.008
	cell killing	157	6	GNLY, IFNG, IL21, KIR2DL4, LAG3, NCR1	26.535	0.000	0.000
	antimicrobial humoral response	110	4	CXCL10, CXCL13, CXCL9, GNLY	25.248	0.000	0.009
	regulation of lymphocyte mediated immunity	142	4	IL21, KIR2DL4, LAG3, NCR1	19.559	0.000	0.019
	lymphocyte mediated immunity	238	5	IL21, KIR2DL4, KLRC2, LAG3, NCR1	14.587	0.000	0.010
	humoral immune response	242	5	CXCL10, CXCL13, CXCL9, GNLY, IFNG	14.346	0.000	0.010
	regulation of immune effector process	381	6	IFNG, IL21, KIR2DL4, LAG3, NCR1, ZNF683	10.934	0.000	0.008
	regulation of immune response	909	8	CXCL13, GPLD1, IFNG, IL21, KIR2DL4, LAG3, NCR1, ZNF683	6.111	0.000	0.012
	defense response	1518	13	CXCL10, CXCL13, CXCL9, GNLY, IFNG, IL21, KIR2DL4, KLRC2, KLRC3, LAG3, NCR1, TTBK1, ZNF683	5.946	0.000	0.000
	innate immune response	827	7	IFNG, IL21, KIR2DL4, KLRC2, LAG3, NCR1, ZNF683	5.877	0.000	0.045
	positive regulation of immune system process	979	8	CXCL10, CXCL13, GPLD1, IFNG, IL21, KIR2DL4, LAG3, TTBK1	5.674	0.000	0.019
	immune effector process	1141	9	CXCL10, CXCL9, IFNG, IL21, KIR2DL4, KLRC2, LAG3, NCR1, ZNF683	5.477	0.000	0.009

	regulation of immune system process	1400	10	CXCL10, CXCL13, GPLD1, IFNG, IL21, KIR2DL4, LAG3, NCR1, TTBK1, ZNF683	4.960	0.000	0.008
	immune response	1919	12	CXCL10, CXCL13, CXCL9, GNLY, GPLD1, IFNG, IL21, KIR2DL4, KLRC2, LAG3, NCR1, ZNF683	4.342	0.000	0.003
GO: molecular function	CXCR3 chemokine receptor binding	5	3	CXCL10, CXCL13, CXCL9	476.230	0.000	0.000
	CXCR chemokine receptor binding	11	3	CXCL10, CXCL13, CXCL9	216.470	0.000	0.001
	chemokine activity	47	3	CXCL10, CXCL13, CXCL9	50.663	0.000	0.025
	chemokine receptor binding	61	3	CXCL10, CXCL13, CXCL9	39.035	0.000	0.046
	cytokine activity	217	5	CXCL10, CXCL13, CXCL9, IFNG, IL21	18.288	0.000	0.010
	cytokine receptor binding	274	5	CXCL10, CXCL13, CXCL9, IFNG, IL21	14.484	0.000	0.022
GO: cellular component	external side of plasma membrane	289	5	CXCL10, CXCL9, ITGAE, LAG3, RTBDN	16.910	0.000	0.016

4.11 Comparison of the best-performing models to EaSIeR models

We have so far evaluated models on the validation set, which is robust in that this is an independent patient cohort. However, it is a small set and thus there is the risk that the models with the best validation results may not generalise well to other cohorts. To challenge the model with higher test set diversity, we merged the discovery and validation datasets, resulting in 320 mUC patients treated with atezolizumab (75 responders, 245 non-responders), to obtain out-of-sample predictions for a much more representative set of patients. By using gene expression profiles alone or integrated with clinical data, we built OMC models to predict atezolizumab response from merged discovery and validation datasets. We evaluated eight ML algorithms through five nested 10-fold CV runs (the CV is performed 5 times changing the random seed), with the best-performing model determined by the highest MCC. We then compared the performance of our models with state-of-the-art ML drug response prediction models (e.g., EaSIeR and JADBio).

The five-repetition MCCs on the merged discovery and validation datasets for eight ML models using gene expression profiles (GEP) and clinical data (GEP + clinical) are shown in Figure 19. Using GEP, the LGBM-OMC model achieved the highest MCC of 0.252, which nonlinearly combined 49 genes, referred to as GEP(LGBM). Notably, 19 of these genes overlapped with the 29 genes from the best-performing CART-OMC model in validation dataset (Table 11). The CART-OMC model, which is the best-performing model in the validation dataset, also ranked among the top models in the merge datasets. While CART-OMC on integrated expressions of 65 genes, TMB and TNB features, called GEP + TMB + TNB (CART-OMC), slightly performed better than those using gene expression profiles alone with the highest MCC of 0.253. (Figure 19). 24 of these features overlapped with the 69 features from the best-performing LR-OMC model in the validation dataset (Table 11).

To predict atezolizumab response in the same set of 320 mUC patients using EaSIeR, RNA-sequencing data served as input to quantify five sets of patient-specific, real-valued system-based signatures of TME, including (i) the composition of 11 immune cell types, (ii) the activity of 14 pathways, (iii) the activity of 118 transcriptional factors, (iv) interactions of 812 ligand-receptor pairs, and (v) 169 cell-cell pairs (Table 2). These signatures provide a comprehensive characterisation of the tumours and help to elucidate the complex cellular composition of TMEs, and their inter- and intracellular mechanisms are involved in mediating immune response and affecting the efficacy of immunotherapy. Since mUC often originates from bladder cancer, we employed BLCA-specific RMTLR model from the EaSIeR, which was pre-trained on TCGA-BLCA samples, to predict 10 tasks of immune response scores (CYT, Ock_IS, Roh_IS, Chemokine, Davoli_IS, IFNy, Ayer_expIS, Tcell_inflamed, RIR, TLS; Table 3) in 320 mUC patients from the merged discovery and validation datasets. These 10 score predictions were made based on six sets of features, including five system-based signatures of TME and their combination, producing a total of 60 EaSIeR model predictions (six sets of features x 10 immune response scores).

The evaluation metrics (MCC, ROC-AUC, and PR-AUC) demonstrate the performance of our best-performing models in predicting atezolizumab response for 320 mUC patients, compared to 60 EaSleR model predictions (Figure 20). Overall, the performance of EaSleR models varied across different models, with MCCs at near-random predictive levels (practically MCC ~ 0). Our study introduced two best-performing models, GEP (LGBM-OMC) and GEP + TMB + TNB (CART-OMC), that outperformed all 60 EaSleR model predictions. These results indicate that they are more predictive than TME-based biomarkers from EaSleR models in predicting mUC patients' response to atezolizumab.

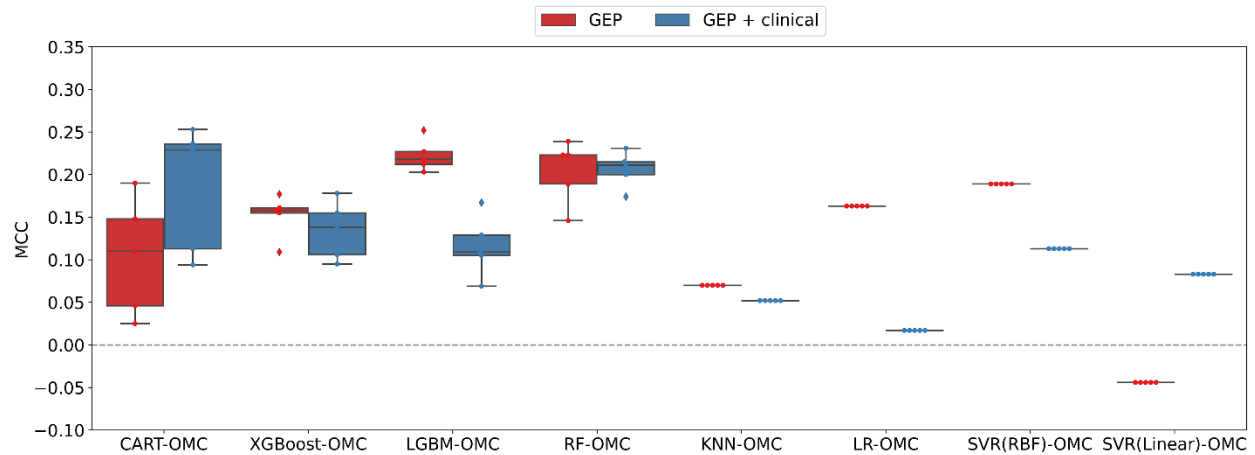


Figure 19. Five-repetition MCCs of the merged discovery and validation datasets across eight ML models using either gene expression profiles or integrated gene expression profiles plus clinical data. MCC is calculated using a merged prediction of a nested 10-fold CV run on the merged discovery and validation datasets (245 non-responders and 75 responders). The CV is repeated five times with different random seeds to assess robustness, and the results are presented in a boxplot. Eight regression algorithms with OMC feature selection select only a small subset of informative features, facilitating model prediction. Two sets of features were employed per learning algorithm: gene expression profiles (GEP) and integrated gene expression profiles with clinical data (GEP + clinical). An MCC of 0.0 indicates random-level performance, as shown by the horizontal dashed line.

Table 11. The list of common features between the best-performing models using gene expression profile (GEP) and integrated gene expression profiles and clinical data (GEP + clinical) from validation and merged analyses

GEP	GEP + clinical
CXCL10	AUNIP
CXCL13	CBX2
CXCL9	CPLX4
FOXN4	CXCL13
GNLY	CXCL9
IFNG	CXCR3
IL21	ESPL1
ITGAE	GABRA3
KCNJ10	GIPC2
KIR2DL4	GNLY
KLRC2	GZMM
KLRC4	IFNG
LAG3	IL21
MMP10	ITGAE
NCR1	KCNJ10
NPY4R	KIR2DL4
NRSN1	NCR1
S100A2	NKG7
ZNF683	NRSN1
	OIP5
	PKD2L1
	TBX21
	TMB per megabase
	TNB per mega base

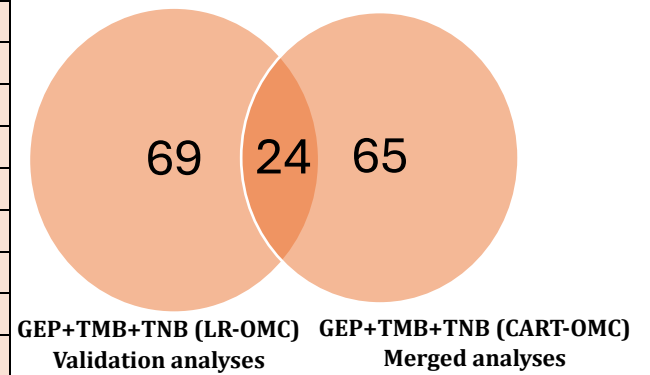
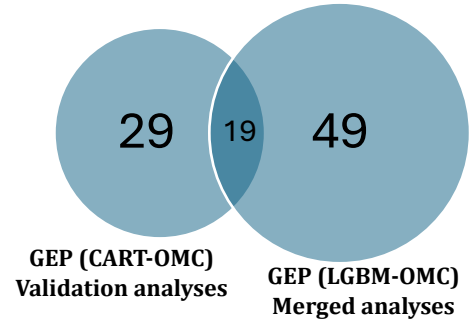


Figure 20. Performance comparison of the best-performing models with those from the 60 EaSIeR model prediction on 320 mUC patients from the merged discovery and validation datasets (245 non-responders and 75 responders). The two leftmost columns are the best-performing models developed in this study using either gene expression profiles, GEP (LGBM-OMC), or integrated gene expression profiles with clinical data, GEP + TMB + TNB (CART-OMC). Both models were evaluated using nested 10-fold CV repeated five times changing random seeds. The one with the highest MCC are shown. Whereas the other 60 columns on the right illustrate the results of running the EaSIeR models on the same dataset. BLCA-specific RMTL models pretrained on TCGA-BLCA samples available in EaSIeR to predict 10 tasks of immune response scores (CYT, Ock_IS, Roh_IS, Chemokine, Davoli_IS, IFNy, Ayer_expIS, Tcell_inflamed, RIR, TLS) based on five system-based signatures of TME and their combinations (ensemble), including cell-cell pairs (ccpairs), immune cell types (immunecells), ligand-receptor pairs (lrpairs), pathway activities (pathways), and transcriptional factor activities (tfs), producing a total of 60 EaSIeR model predictions (named by adding the suffix of immune response scores to the names of system-based signatures of TME e.g., ccpairs_CYT). Random-level performance is delimited by horizontal dashed lines: 0.0 for MCC, 0.5 for ROC-AUC, 0.765 for PR-AUC when using GEP and 60 EaSIeR models, 0.733 for PR-AUC when using GEP + TMB + TNB). Overall, the best-performing models (pink) consistently outperformed all of the 60 EaSIeR models (purple).

4.12 Enhancing prediction on the merged dataset by integrating TME descriptors, immune response scores, and the EaSIeR score into the best-performing models

After identifying the best-performing models from the merged discovery and validation datasets: LGBM-OMC combining 49 genes (GEP (LGBM-OMC)) and CART-OMC combining 63 genes with TMB and TNB (GEP + TMB + TNB (CART-OMC)), we aimed to enhance predictive accuracy by integrating 1,134 additional features computed for each patients using EaSIeR. These include five system-based signatures of the TMEs, referred to as TME descriptors, (1,124 features of 11 immune cell types, 14 pathway activities, 118 transcriptional factor activities, 812 ligand-receptor pairs, and 169 cell-cell pairs; Table 2), nine published immune response scores (CYT, Roh_IS, Chemokine, Davoli_IS, IFNy, Ayer_expIS, Tcell_inflamed, RIR, TLS; Table 3, excluding Ock_IS which is unavailable for estimation in the EaSIeR), and the EaSIeR score. The EaSIeR score is the average of 10 immune response predictions based on ensemble TME descriptors. These features represent predictors of patients' immune response.

The results, illustrated in Figure 21A, show that adding individual sets of features (5 TME descriptors, 9 immune response scores, and the EaSIeR score), or their combinations, to the 49-gene LGBM model improved predictive performance compared to those built solely on 49 genes (mMCC of 0.218). The best performance came from combining 49 genes with 9 immune response scores (mMCC of 0.358), while using only a combination of five TME descriptors, 9 immune response scores, and the EaSIeR score resulted in the lowest performance (mMCC of 0.136).

Similarly, Figure 21B revealed that CART-OMC combining 63 genes, TMB, and TNB, when combined with either the EaSIeR score (mMCC of 0.342), or nine immune response scores (mMCC of 0.394), improved predictive performance compared to those built solely on 63 genes, TMB, and TNB (mMCC of 0.236). CART was not able to cope with the higher number of features involved, and performance declined with the addition of five TME descriptors (mMCC of 0.145), or a combination of five TME descriptors, 9 immune response scores, and the EaSIeR score (mMCC of 0.105). The lowest performance (mMCC of 0.097) was observed when CART was based solely on a combination of five TME descriptors, 9 immune response scores, and the EaSIeR score.

Overall, integrating five TME descriptors, nine immune response scores, and the EaSIeR score, as computed using the EaSIeR software, into a predictive model provides valuable information that can improve the model's predictive ability to predict response to atezolizumab in mUC patients. Specifically, the nine immune response scores significantly contributed to improving the accuracy of both LGBM and CART models. However, including the TME descriptors alone, or in combination with other features, did not consistently lead to performance gains and, in some cases, introduced additional complexity that reduced model effectiveness.

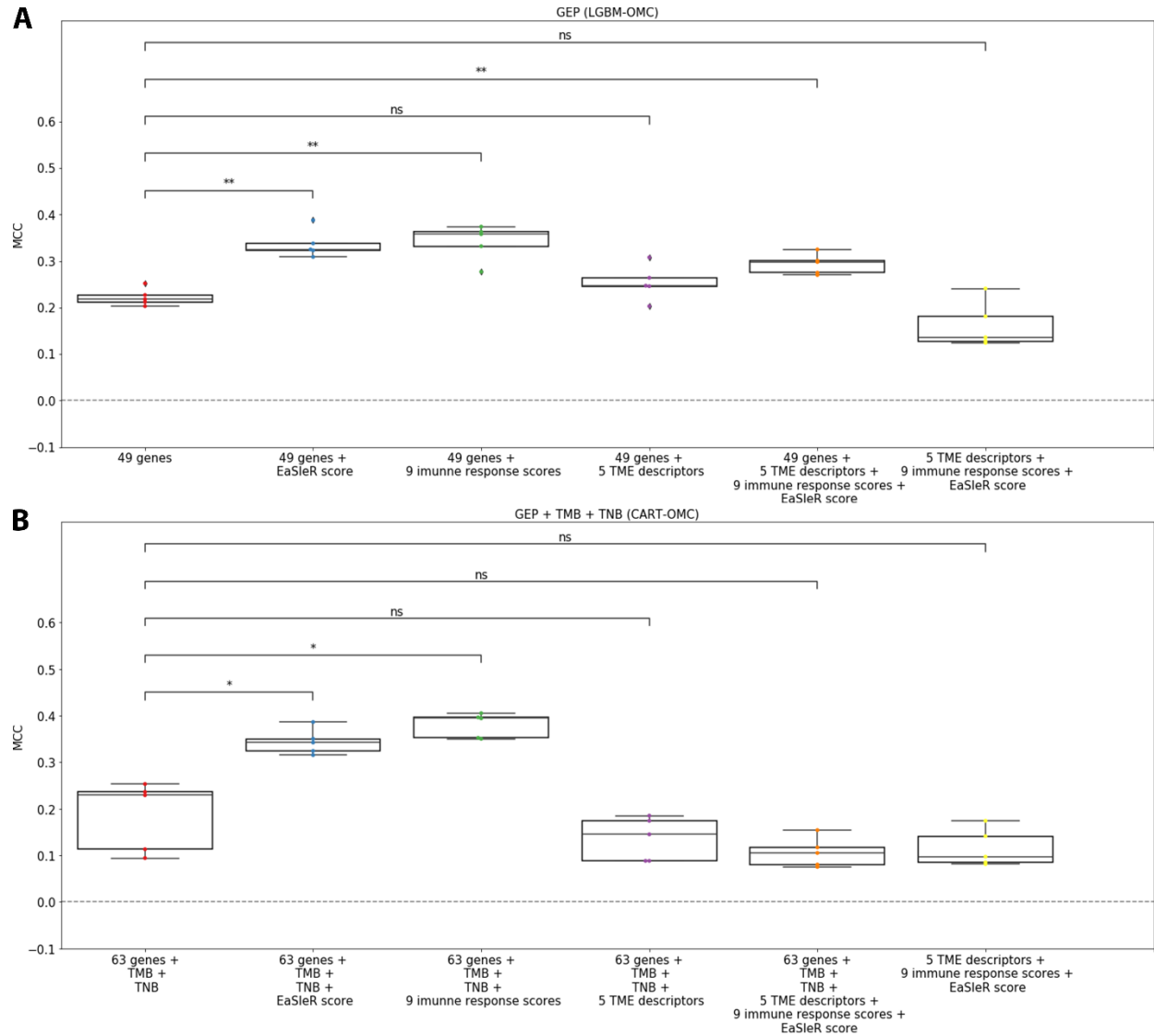


Figure 21. Comparison of the predictive performance of the best-performing models using their predictive features to those combining with TME descriptors, immune response scores, and the EaSIeR score for atezolizumab response prediction in 320 mUC patients from the merged discovery and validation datasets. Boxplots comparing MCCs obtained from five 10-fold CV runs (each dot represents a run) of two best-performing models: LGBM with 49 predictive genes (A), and CART-OMC with 63 genes, TMB and TNB (B). These are compared against models incorporating additional features: five TME descriptors, 9 published immune response scores (Ock_IS is not available in the EaSIeR tool), and EaSIeR score. These features account for the cellular composition of TMEs together with inter- and intracellular communication, offering a comprehensive view of tumours-immune interactions affecting the efficacy of immunotherapy. In addition, immune response scores and the EaSIeR score are proposed as predictors of response to ICIs. Random-level performance is delimited by the horizontal dashed lines at MCC of 0.0. Statistical comparisons between different models were performed using Welch's t-test (two-sided). Stars denote the p-value of the test, where non-significant "ns" means $0.05 < p \leq 1.00$, "*" means $p < 0.05$.

means $0.01 < p \leq 0.05$, “**” means $0.001 < p \leq 0.01$, “***” means $0.0001 < p \leq 0.001$, and “****” means $p \leq 0.0001$. Overall, integrating five TME descriptors, 9 immune response scores, and the EaSleR score into the predictive model improved the accuracy of both LGBM and CART models.

4.13 Comparison of the best-performing models to JADBio autoML

AutoML is the application of ML models to real-world problems by automating key aspects of the ML pipeline, including data preprocessing, feature selection, model selection, and hyperparameter tuning, minimising the need for manual intervention[222]. JADBio[55] is a recently launched AutoML tool tailored to low-sample high-dimensional datasets, as those arising in biomedical or biological problems. It outputs optimal predictive models by automatically training and evaluating numerous ML pipelines, optimising data preprocessing, feature selection, ML algorithms and their hyper-parameters. In this study, we applied JADBio on two sets of features: gene expression profiles (GEP) and integrated gene expression profiles and clinical data (GEP + clinical) to predict atezolizumab response in 320 mUC patients from merged discovery and validation datasets using JADBio’s regression algorithms. The evaluation metrics (MCC, ROC-AUC, and PR-AUC) demonstrate the performance of our best-performing models built in this study on the same set of patients against the JADBio prediction models (Table 12).

When using gene expression profiles as an input, JADBio performs a repeated 10-fold CV (max. repeats = 20) with 4,276 different configurations being tried. The best-performing model was regression RF (with the following hyperparameters: training 100 trees with MSE splitting criterion, minimum leaf size = 9, splits = 1, alpha = 1, and variables to split = number of variables divided by 9.0), along with SES feature selection (with maxK = 2, alpha = 0.1 and budget = 3 * number of variables), achieving an MCC of 0.179. This model generated six equivalent atezolizumab response signatures, each comprising 25 out of 19,440 genes.

Combining gene expression profiles and clinical data slightly improved the model performance in predicting atezolizumab response using JADBio. The best-performing model was regression RF (with training 100 trees with MSE splitting criterion, minimum leaf size = 7, splits = 1, alpha = 1, and variables to split = number of variables divided by 9.0), along with SES feature selection (with maxK = 2, alpha = 0.1 and budget = 3 * number of variables), achieving MCC of 0.198. This model produced 56 equivalent atezolizumab response signatures, each comprising 19 (15 genes, TMB per megabase, and TNB per megabase) out of 19,448 features.

Overall, our best-performing model on the merged dataset, with MCC of 0.252 for GEP (LGBM-OMC) and MCC of 0.253 for GEP + clinical (CART-OMC), outperformed the JADBio best-performing models with MCC of 0.179 for GEP and MCC of 0.198 for GEP + clinical (Table 12).

Table 12. The predictive performance of the best-performing models vs. JADBIO models in predicting atezolizumab response of 320 mUC patients from the merged discovery and validation datasets. The analysis used two feature sets: gene expression profiles (GEP) and combined GEP with clinical data (GEP + clinical). To perform a direct comparison, out-of-sample CV predictions from JADBIO's best-performing model were evaluated alongside our models using the same scripts and metrics for consistency. Random-level performance is 0.0 for MCC, 0.5 for ROC-AUC, 0.765 for PR-AUC when using GEP, and 0.733 for PR-AUC when using GEP + clinical.

Features	The best algorithm	Methods	#Features	MCC	ROC-AUC	PR-AUC
GEP	LGBM-OMC	nested 10-fold CV	49 genes	0.252	0.685	0.860
	JADBIO-RF-SES *	repeated 10-fold CV (maximum repeats = 20)	25 genes	0.179	0.625	0.842
GEP + clinical	CART-OMC	nested 10-fold CV	63 genes + TMB + TNB	0.253	0.664	0.870
	JADBIO-RF-SES **	repeated 10-fold CV (maximum repeats = 20)	15 gene + TMB + TNB	0.198	0.660	0.834

*JADBIO optimal parameters for GEP: Regression RF (with hyper-parameters: training 100 trees with mean squared error (MSE) splitting criterion, minimum leaf size = 9, splits = 1, alpha = 1, and variables to split = number of variables divided by 9.0) with SES feature selection (with hyper-parameters: maxK = 2, alpha = 0.1 and budget = 3 * number of variables)

**JADBIO optimal parameters for GEP + clinical: Regression RF (with hyper-parameters: training 100 trees with MSE splitting criterion, minimum leaf size = 7, splits = 1, alpha = 1, and variables to split = number of variables divided by 9.0) with SES feature selection (with hyper-parameters: maxK = 2, alpha = 0.1 and budget = 3 * number of variables)

Chapter 5

Predictive modelling for other drug responses in cancer patients

ML applications increases in the field of precision oncology, however, many studies still rely on single modelling strategies or a single molecular profile within the identified publication[322]. This approach might lead to missing out on potentially better-suited alternatives that could perform well with specific dataset characteristics. To overcome this limitation, there is a need for diverse algorithms and multiple molecular profiles to be performed, setting a new standard for ML algorithm selection and development, and is recommended as an additional criterion to assess study quality among ML-based modelling.

Thanks to the U.S. NCI GDC, we had access to harmonised datasets across several cancer genome programs, offering a comprehensive set of clinical drug response data paired with omics profiles[199]. TCGA, one of the most ambitious and successful cancer genomics programs to date, has generated, analysed, and made available genomic sequence, expression, methylation, and CNV data for over 11,000 patients across 33 cancer types. Considerable efforts have leveraged TCGA's multi-omics data to predict drug treatment responses in various cancer types[46], [323]–[326]. For instance, Bomane et al.[46] employed TCGA data to predict breast cancer patients' responses to paclitaxel, using six tumour molecular profiles and 10 ML algorithms. Their study identified DT and, XGBoost, with OMC feature selection, exhibited highly predictive models based on the miRNA and DNA methylation profiles. Similarly, Ding et al.[323] developed an elastic net regression with bootstrap and RF model to predict drug responses for six drug-cancer-type binomials, each employing four molecular profiles. These approaches highlight the potential of omics-driven precision medicine to enhance prediction accuracy.

In this chapter, we extend our proposed ML methodology for drug response prediction to three different chemotherapeutic drug-cancer-type binomials, beyond the previous successes with atezolizumab immunotherapy in mUC patients, including taxane-anthracyclines in breast cancer, doxorubicin in breast cancer, and gemcitabine in PAAD. Since the predictive ability of ML models generally depends on the specific problem, we performed ML analysis tailored for each drug-cancer-type binomial, considering various ML algorithms across multiple molecular profiles, including mRNA, DNA methylations, CNVs, miRNAs and their isoforms. In brief, we evaluated the performance of eight ML algorithms to predict response to taxane-anthracycline treatment in breast cancer using tumour gene expression profiles, with and without the OMC feature selection, generating 16 ML models. Additionally, we expanded our analysis to include multiple molecular profiles beyond gene expression to conduct a large-scale ML analysis on matched treatment response and TCGA multi-omics for doxorubicin for breast cancer[298] and gemcitabine for PAAD[299]. We systematically evaluated eight ML algorithms, with and without an OMC feature selection, across eight

different molecular profiles, resulting in 128 ML models (16 algorithms x 8 molecular profiles).

Our results demonstrated the robustness and generality of our proposed ML methodology by showing its consistent performance and adaptability across diverse datasets and highlighted the importance of using multiple algorithms and molecular profiles, leading to the discovery of predictive, reproducible, and robust drug response signatures and enhancing the overall prediction accuracy of drug response.

5.1 Case study I: Predictive modelling of breast cancer patient response to taxane-anthracycline chemotherapy using gene expression profiles

Currently, taxane-anthracycline chemotherapy is the most widely used agent in neo-adjuvant chemotherapy for breast cancer[327]. Since clinical or pathological response can only be evaluated after a period of chemotherapy, we are unable to predict in advance which patient will be more sensitive or more resistant to treatment. Although primary drug resistance will occur in some patients leading to treatment failure, all patients eligible for taxane-anthracycline chemotherapy empirically receive the same treatment in clinical practice. Therefore, identifying robust predictors would be of value to reducing ineffective treatment and related harmful side effects for those unlikely to respond. To our knowledge, a broad range of ML algorithms to classify breast cancer patients' responses to taxane-anthracycline chemotherapy has not yet been investigated so far. With this purpose, we retrospectively collected tumour gene expression profiles from 732 taxane-anthracycline chemotherapy-treated breast cancer patients across two independent studies: Hatzis et al. (488 patients)[328] and Magbanua et al. (244 patients)[329]. We developed ML models with eight classification algorithms to predict the response to taxane-anthracycline chemotherapy using the Hatzis discovery dataset[328]. The best-performing model was then externally validated on an independent dataset from the Magbanua study[329] to assess the model's generalisability. The goal of this study is to identify the robust ML model to classify taxane-anthracycline chemotherapy responders and non-responders in breast cancer patients.

5.1.1 Cohort characteristics of taxane-anthracycline chemotherapy-treated breast cancer patients

The Hatzis discovery dataset[328] included 488 breast cancer patients who received taxane-anthracycline chemotherapy, while the Magbanua independent dataset[329] included 244 patients with the same cancer type and treatment. We retrieved the gene expression profiles and clinical data for both studies from the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database, with GSE25066 for the discovery dataset and GSE32603 for the independent dataset. Gene expression profiling for the discovery dataset was conducted using the Affymetrix Human Genome U133A microarray (HG-U133A), with 22,284 probe IDs, while the independent dataset used the UCSF/HAQQLAB_Human_40986_ISPY microarray, with 5,069 probe IDs. The gene

expression data is a two-dimensional matrix where rows represent patients and columns represent genes, with gene expression levels quantified as normalised log₂-transformed probe intensities in the discovery dataset and normalised log₂ of the Cy5 (tumour)/Cy3 (reference) intensities in the independent dataset.

Taxane-anthracycline chemotherapy response was evaluated by pathologists through a microscopic examination using haematoxylin and eosin staining[330]. Patients achieving a pathologic complete response (pCR), defined as no detectable cancer in post-chemotherapy tissue samples, were classified as responders. While patients with residual disease (RD), which showed persistent cancer cells after treatment, were defined as non-responders. Thus, the response rate was 20% (99 responders and 389 non-responders) in the discovery dataset and 21% (51 responders and 193 non-responders) in the independent dataset.

5.1.2 Development and validation of ML models

To facilitate ML analysis, both datasets were standardised by converting probe IDs to valid gene symbols, averaging expression values for genes with multiple probes. We could identify 8,992 protein-coding genes that overlapped between the discovery and independent datasets. Genes with missing values were replaced with each gene's mean expression, and MinMaxScaler were used to scale the gene expression data. The patients' response labels were encoded as 0 for non-responders and 1 for responders. Due to a class imbalance, where 80% of the data are non-responders, class weights were applied during model training to assign higher weights to the minority class, allowing the model to pay more attention to its patterns and reducing bias towards the majority class. However, class weights could not be applied to the KNN algorithm because it does not support class weighting.

The discovery dataset was divided into ten equal-sized folds as training and testing sets using stratified k-fold ($k = 10$) to preserve the same class distribution in each fold as the original dataset. We developed all-feature and OMC models using eight supervised ML classification algorithms to predict patients' responses to taxane-anthracycline chemotherapy: classification and regression tree (CART), RF, XGBoost, LGBM, KNN, LR, SVC with linear and RBF kernel, resulting in a total of 16 ML models. The all-feature model's predictive performance was evaluated with a standard 10-fold CV using all features. While the OMC model was assessed via nested 10-fold CV using only a subset of informative features (the top 2 to $n/2$ subset of features with the smallest P -values from an ANOVA test, where n is the number of samples). The CV is performed 5 times changing the random seed to assess robustness.

Out-of-sample predictions from the 10-fold CV were merged to calculate the evaluation metrics (MCC, ROC-AUC, and PR-AUC). The classification threshold was set to 0.5, where patients with predicted probabilities above this value were categorised as responders, and those below as non-responders. The best-performing model was selected based on the highest mMCC.

5.1.3 Predicting breast cancer patient response to taxane-anthracycline chemotherapy from gene expression profiles

The evaluation metrics (MCC, ROC-AUC, and PR-AUC) for five 10-fold CV runs on the discovery dataset using gene expression profiles across 16 ML models (all-feature and OMC version) is shown in Figure 22. Overall, five models exhibited promising results in distinguishing responders from non-responders with an mMCC above 0.3: LR (mMCC of 0.396), LR-OMC (mMCC of 0.343), SVC(Linear) (mMCC of 0.332), SVC(Linear)-OMC (mMCC of 0.302), and SVC(RBF)-OMC (mMCC of 0.329). The all-feature LR model achieved the highest mMCC of 0.396, along with the highest ROC-AUC of 0.785, and PR-AUC of 0.473. Thus, it is recommended for this challenging problem.

Notably, six OMC models (CART-OMC, LGBM-OMC, RF-OMC, XGBoost-OMC, KNN-OMC, and SVC(RBF)-OMC) showed higher mMCC compared to their corresponding all-feature models, although some improvements were not statistically significant. This suggests that ML models benefit from feature selection, which aims to extract only the most informative features and remove noisy non-informative features, leading to more accurate prediction.

5.1.4 External validation of the best-performing model

To investigate whether the best-performing model (all-feature LR) could generalise to new, unseen patient data, we externally validated its predictive ability on an independent dataset. To achieve this, the all-feature LR model was retrained using all available features from the entire discovery dataset and was then applied to predict taxane-anthracycline chemotherapy response in breast cancer patients from the Magbanua study[329].

The evaluation metrics (MCC, ROC-AUC, and PR-AUC) of the best-performing model in predicting response to taxane-anthracycline chemotherapy on the independent dataset are shown in Table 13. The model exhibited a slight decrease in performance on the independent dataset (MCC of 0.351) compared to the discovery dataset (MCC of 0.396). These results suggest that, despite being trained on the discovery dataset, the model was robust and accurately predicted the response on a new, unseen dataset, demonstrating its applicability was not limited to a single study.

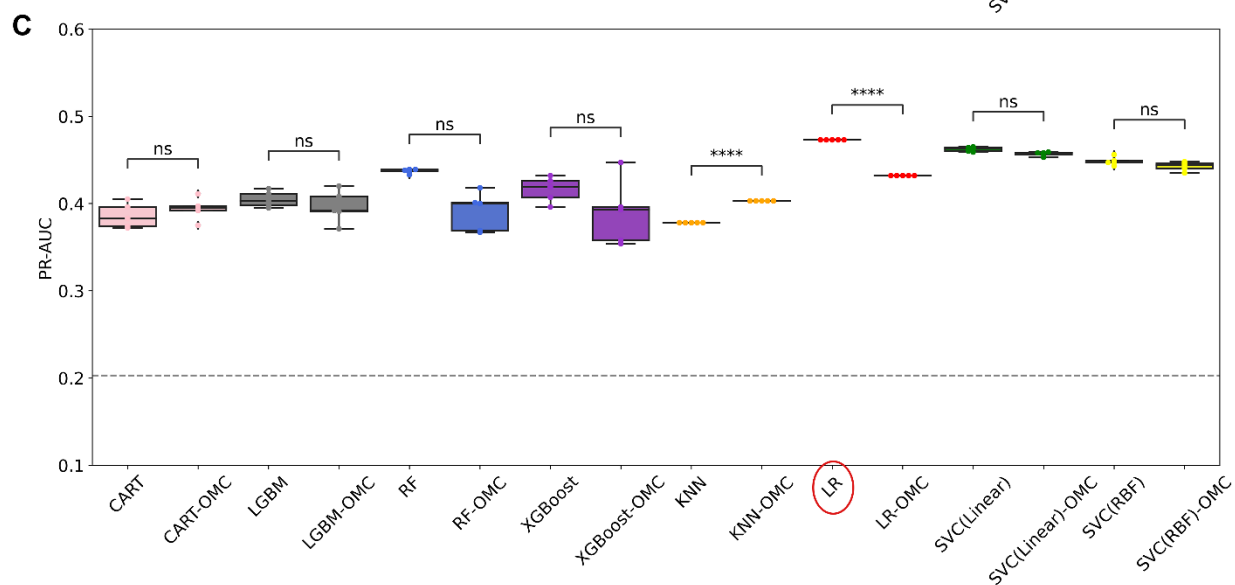
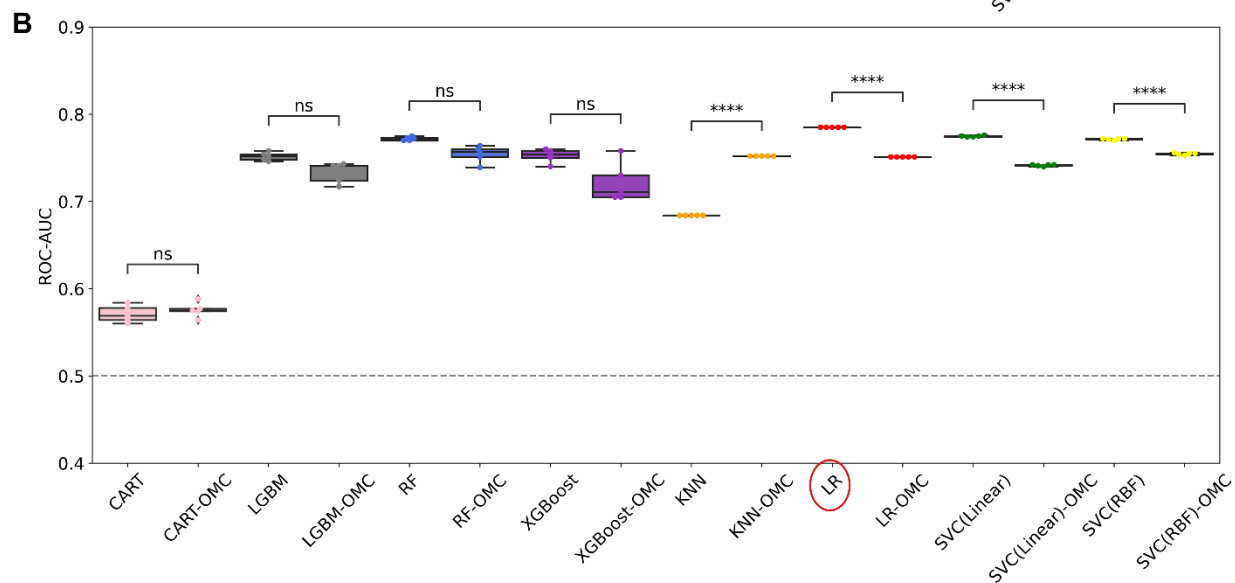
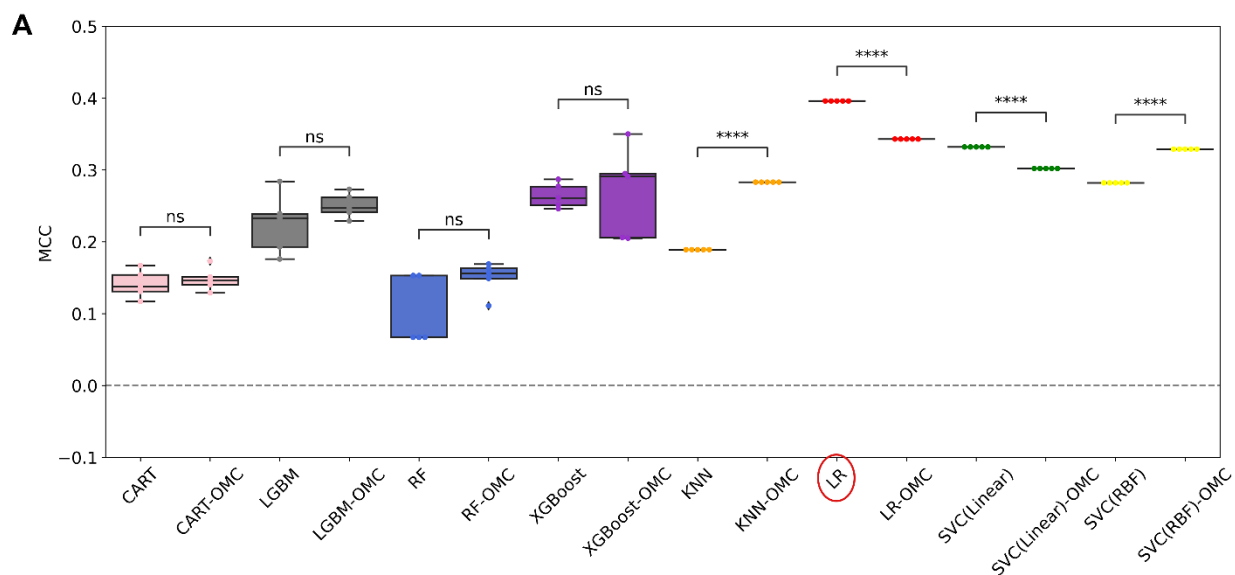


Figure 22. Five-repetition evaluation metrics on the discovery dataset across eight ML models using gene expression profiles. (A) MCC, (B) ROC-AUC, and (C) PR-AUC was calculated from merged 10-fold CV runs on the discovery dataset (the CV is performed five times changing the random seed to assess the robustness). Gene expression profiles were standardised using MinmaxScaler and analysed per learning algorithm either with all features or a small subset of informative features selected via OMC feature selection (the suffix “OMC” was added to the algorithm name). Class imbalances were mitigated by applying class weights during model training, except for KNN, which does not support class weighting. The random-level performance is delimited by the horizontal dashed lines: 0.0 for MCC, 0.5 for ROC-AUC, and 0.203 for PR-AUC. Statistical comparisons between models were performed using Welch’s t-test (two-sided). Stars denote the p-value of the test, where non-significant “ns” means $0.05 < p \leq 1.00$, “*” means $0.01 < p \leq 0.05$, “**” means $0.001 < p \leq 0.01$, “***” means $0.0001 < p \leq 0.001$, and “****” means $p \leq 0.0001$.

Table 13. The evaluation metrics of the best-performing model using gene expression profiles on the independent dataset. The evaluation metrics (MCC, ROC-AUC, and PR-AUC) were calculated using prediction on the independent dataset of the best-performing model (LR). This LR model was selected based on the highest MCC from a 10-fold CV on the discovery dataset, retrained on the full discovery dataset, and tested on the independent dataset. Random-level performance is as follows: MCC of 0.0, a ROC-AUC of 0.5, a PR-AUC of 0.203 for the discovery dataset, and a PR-AUC of 0.209 for the independent dataset.

Dataset	The best-performing model	MCC	ROC-AUC	PR-AUC
Discovery dataset	LR with MinmaxScaler and class weight applied	0.396	0.785	0.473
Independent dataset		0.351	0.586	0.624

5.2 Case study II: Predictive modelling of breast cancer patient response to doxorubicin based on their tumour molecular profiles

Doxorubicin is a commonly used chemotherapeutic agent for various solid tumours[331], [332]. However, not all patients respond effectively to this drug, and its use can sometimes lead to severe, potentially life-threatening side effects[333]. Therefore, accurately identifying doxorubicin-resistant patients is crucial to unnecessary risks and offer alternative treatments without delay. Traditional oncology approaches often rely on single predictors (e.g., HER2 expression)[334]–[336]; however, the complex intra- and intertumoral heterogeneity in breast cancer, which poses diagnostic and therapeutic challenges, leads to insufficient clinical outcomes and responses prediction to current therapies, necessitating a more comprehensive approach[337]. This study leveraged ML applied to multi-omics and clinical data to predict doxorubicin response in breast cancer patients.

Using data from 96 breast cancer patients (87 responders and 9 non-responders) in the publicly available TCGA-BRCA dataset, we analysed clinical and molecular profiles, including transcriptomics (e.g., gene expression, miRNA, and isomiR expression), CNVs, and DNA methylation. These profiles contained high-dimensional features, ranging from expression levels of 927 isomiR to methylation levels of 450,000 DNA probes. We developed ML models using eight binary classification algorithms, with and without OMC feature selection, applied across eight molecular datasets, resulting in 128 ML models. Models were evaluated using 10-fold CV (the CV is repeated five times changing random seeds). Class imbalances were mitigated by applying class weights during model training, except for KNN, which does not support class weighting. For algorithms outputting class probabilities, a threshold of 0.5 was used to classify patients: those with probabilities above 0.5 were classified as responders, and those below as non-responders. The full details of methodology, including data collection and preprocessing and the OMC feature selection strategy, is available in Appendix II[298].

Our large-scale analysis (8 molecular profiles $\times$ 8 algorithms $\times$ 2 feature strategies) identified two best-performing models for doxorubicin response among 128 ML models. The CART algorithm, using the expression levels of four isomiRs (hsa-miR-450a-1, hsa-miR-19b-2, hsa-miR-92a-2, and hsa-miR-92a-1), achieved the highest performance (mMCC of 0.56; Figure 23). In addition, CART incorporating three miRNAs (hsa-miR-4421, hsa-miR-4680, and hsa-miR-514a-1) yielded an mMCC of 0.32. These findings highlight that considering fewer algorithms and/or fewer profiles could easily miss predictive models for doxorubicin response.

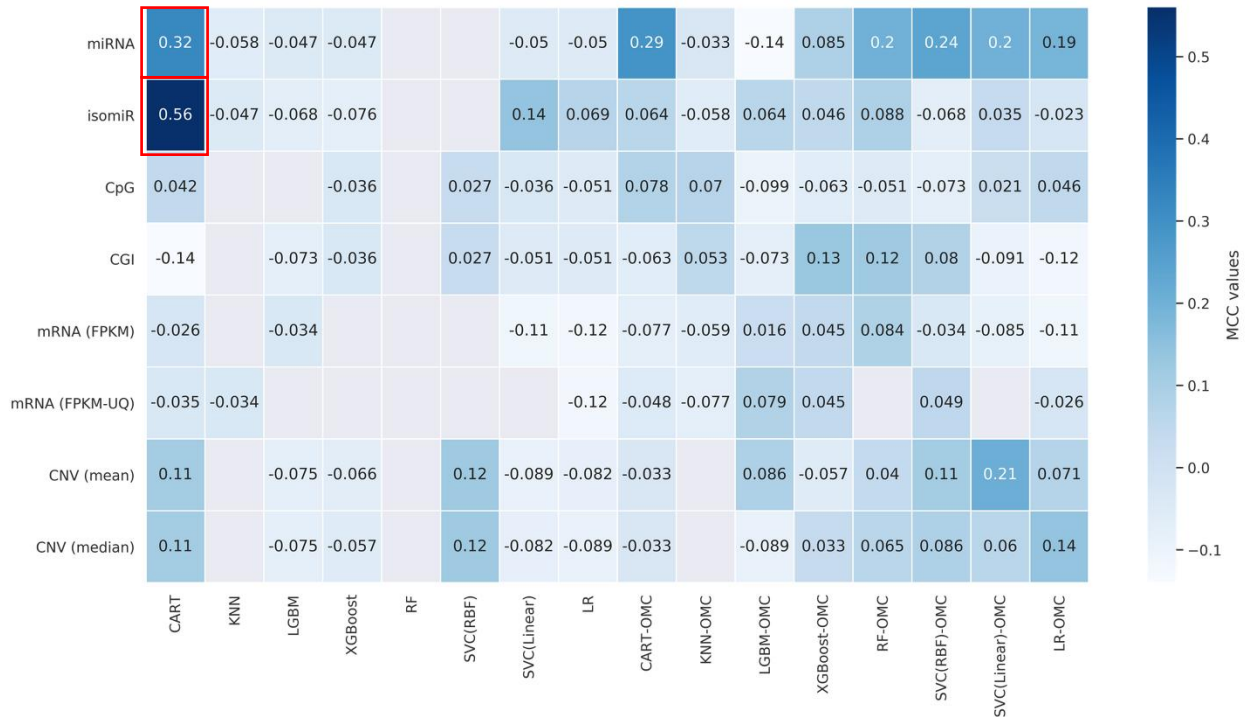


Figure 23. Heatmap showing the mMCC of five 10-CV repetitions for 128 models in predicting response to doxorubicin in breast cancer patients. Rows represent molecular

profiles, while columns correspond to classification algorithms. The first eight algorithms on the left utilised all available features, while the rightmost eight (labelled with the “OMC” suffix) employed OMC to identify a small subset of features, facilitating model prediction. 10-by-10 nested-CV runs were carried out five times, each time with a different random seed. MCCs were computed from merged CV predictions for each run, producing five MCC values per algorithm-profile pair. The heatmap displays the mMCC, with darker blue shades indicating higher predictive performance. Random-level performance corresponds to an MCC of 0.0, while undefined mMCC values (e.g., when the model predicts a single class for all samples, resulting in a zero denominator) are shown as blank grey boxes. The full details of the methodology are fully described in Appendix II[298].

Previous studies identified HER2 status as a marker of doxorubicin sensitivity in breast cancer, with higher HER2 expression linked to increased drug sensitivity[338]–[340]. However, this study found that CART models using three miRNAs (mMCC = 0.32) or four isomiRs (mMCC = 0.56) outperformed the HER2-based model (mMCC = 0.14) in predicting doxorubicin response.

We next performed pathway enrichment analysis using overrepresentation analysis to identify biological pathways enriched with 13,898 miRNA targeted genes from the WikiPathways cancer repository. The analysis revealed four cancer-associated pathways significantly enriched ($FDR \leq 0.05$): DNA damage response, ErbB signaling, endometrial cancer, and chromosomal/microsatellite instability in colorectal cancer. These pathways, particularly DNA damage response and cell signalling, are linked to doxorubicin's mechanism of action and may contribute to resistance[338]–[340]. The ErbB signaling pathway, in particular, was associated with worsened breast tumorigenesis and doxorubicin resistance[341], [342]. In addition, overrepresentation analysis revealed that 23 breast cancer-specific gene sets were enriched in 10 cancer-associated pathways. These genes were also part of drug resistance pathways, like platinum drug resistance and endocrine resistance. The analysis showed that the BC-specific genes regulate key processes in doxorubicin's mechanism of action, including apoptosis and DNA damage response. Alterations in these processes can contribute to doxorubicin resistance.

5.3 Case study III: Predictive modelling of response to gemcitabine in PAAD patients based on their tumour molecular profiles and clinical data

Gemcitabine has been the standard first-line therapy for patients with unresectable locally advanced or metastatic pancreatic cancer. However, only about 23% of patients with advanced disease respond to gemcitabine, with a median survival of approximately six months[343], [344]. Single-gene biomarkers, particularly those related to the gemcitabine mechanism, have shown an association with treatment response, as demonstrated by studies on protein expression[345], [346] and genetic polymorphisms[347], [348]. Human equilibrative nucleoside transporter 1 (hENT1) has been widely studied, with high expression associated with increased gemcitabine sensitivity in PAAD[346], [349]–[351]. Yet, some studies did not observe these benefits due to variability in antibodies and

methodologies used across different laboratories[352], [353]. Multigene signatures have been explored as alternatives, but their predictive power remains limited[354].

Our study leverages ML to address these challenges by identifying robust drug response biosignatures from diverse tumour molecular profiles. Using data from 70 PAAD patients (41 non-responders, 29 responders) available through the TCGA-PAAD project, we evaluated eight molecular profiles using eight ML algorithms. Models were built with full feature sets and only a small subset of informative features identified through OMC feature selection, resulting in 128 ML models. Model performance was assessed via five repetitions of a 10-fold CV, each with a different random seed. Class imbalances were mitigated by applying class weights during model training, except for KNN, which does not support class weighting. A class probability threshold of 0.5 was applied to classify patients, with those predicted to have a response probability above the threshold classified as responders, and those below classified as non-responders. Detailed methodology is provided in Appendix III[299].

Out of 128 models, only seven demonstrated predictive power ($mMCC > 0.3$), highlighting the importance of extensive model exploration across diverse omics profiles and ML algorithms. These seven predictive models were all OMC models, which improve performance by training only on the most informative features, reducing noise from irrelevant features in high-dimensional datasets compared to all-feature models. The RF-OMC integrating four mRNAs (0.007% of 60,483; SEPW1P, RP11-179A10.1, ATF4P4, and CTC-429L19.3) was the best-performing model, achieving an MCC of 0.44. Similar predictive performances were observed for the LR-OMC incorporating five mRNAs (0.008% of 60,483; SEPW1P, RP11-179A10.1, ATF4P4, CTC-429L19.3, and SPCS2P1) with an $mMCC$ of 0.4. In addition, the XGBoost-OMC combining 12 CpG probes (0.003% of 450,000) from DNA methylation profiles also showed relatively strong performance with an MCC of 0.32. These probes are associated with 14 genes: cg03208198 (COL18A1 and MIR6815), cg07239938 (ELANE), cg11528307 (C14orf80 and CRIP1), cg15200796 (TMEM191C), cg16049691 (AHRR), cg16379910(B2M), cg17804635 (ZNF703), cg23854567 (PXN), cg24128434(DNAH2), cg25653341 (PLOD3), cg27152190 (RP11-429P3.3), and ch.2.184672693R (not linked to any gene). These findings highlight the potential of omics-driven ML approaches to refine gemcitabine response predictions and support personalised treatment strategies in PAAD.

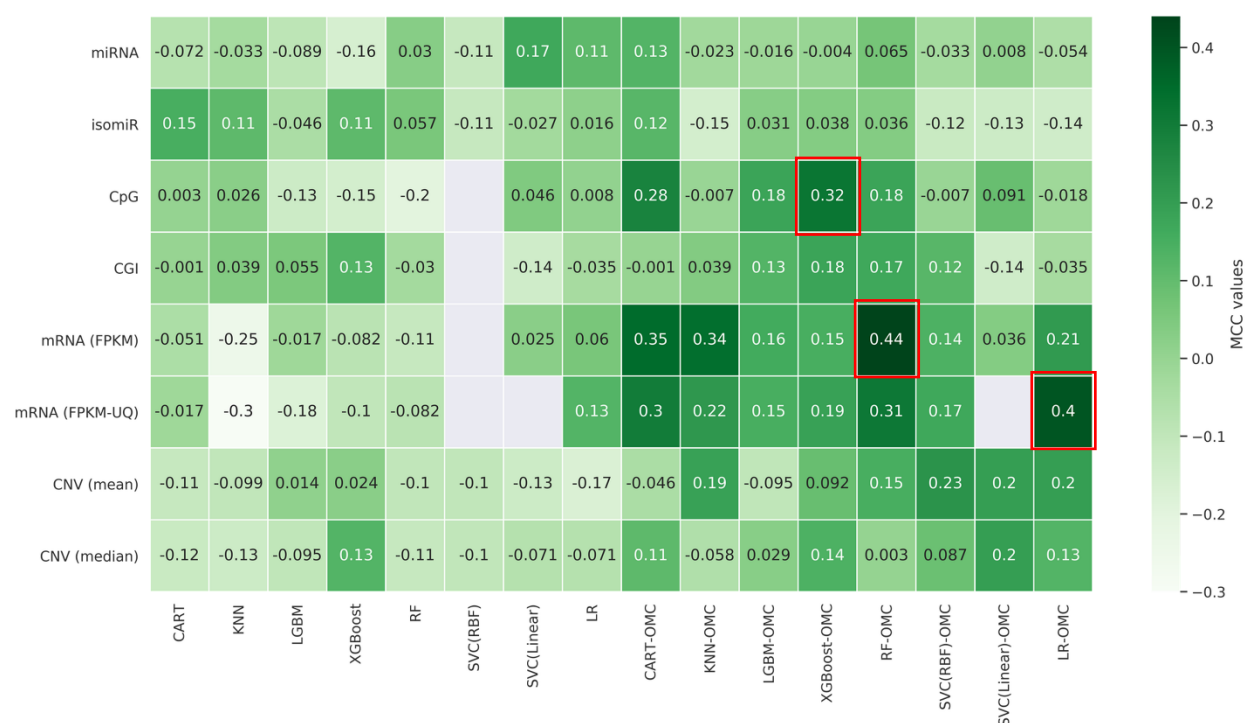


Figure 24. Heatmap presenting the mMCC of five 10-fold CV repetitions for 128 models predicting gemcitabine response in PAAD patients. Rows represent molecular profiles, while columns correspond to classification algorithms. The first eight algorithms on the left utilised all available features, while the rightmost eight (labelled with the “OMC” suffix) employed OMC to identify a small subset of features, facilitating model prediction. 10-by-10 nested-CV runs were carried out five times, each time with a different random seed. MCCs were computed from merged CV predictions for each run, producing five MCC values per algorithm-profile pair. The heatmap displays the mMCC, with darker green shades indicating higher predictive performance. Random-level performance corresponds to an MCC of 0.0, while undefined mMCC values (e.g., when the model predicts a single class for all samples, resulting in a zero denominator) are shown as blank grey boxes. The full details of the methodology are fully described in Appendix III[299].

hENT1 has been identified as a predictive biomarker for gemcitabine response in pancreatic cancer, with high expression linked to greater sensitivity and improved survival[355]–[357]. However, ML models using multiple biomarkers performed significantly better than hENT1 alone. The RF-OMC model combining four mRNAs (mMCC = 0.44) and the XGBoost-OMC model using 12 CpG probes (mMCC = 0.32) both outperformed models based solely on hENT1 expression, which had negative MCC values. These findings suggest that multi-gene and epigenetic markers provide more accurate predictions of gemcitabine response than a single-gene biomarker.

The RF-OMC model combined four predictive mRNA features (SEPW1P, RP11-179A10.1, ATF4P4, and CTC-429L19.3), with some implicated in cancer progression and drug resistance through alterations in their expression. Other studies have already examined the

roles of these genes in other types of cancer. For example, depletion of SEPW1P enhances tumour suppressor activities in breast cancer.[358] The single-nucleotide SNPs located in long noncoding RNA RP11-179A10.1 were reportedly associated with poorer outcomes in ovarian cancer[359], but their association with PAAD has not been documented. DNA methylation profiles with XGBoost-OMC identified 12 CpG sites linked to 14 genes as predictive biomarkers. Increased expression of four mRNAs and methylation at 11 CpG sites correlated with better treatment response, while cg03208198 hypermethylation predicted resistance. Some predictive genes were previously associated with poor prognosis in various cancers. Elevated COL18A1 expression has been associated with unfavourable clinical outcomes in pancreatic cancer[360], gastric cancer[361], and oesophageal squamous cell carcinoma[362]. In addition, reduced CRIP1 expression has been observed in invasive pancreatic cancer[363]. ZNF703, a transcriptional regulator with oncogenic properties, plays a role in cell adhesion, migration, and proliferation. Its overexpression has also been connected to resistance to tamoxifen treatment[364], [365]. Pathway analysis from GO and KEGG pathway database revealed enrichment in cancer-related and drug resistance pathways, supporting their potential mechanistic role in treatment response.

Chapter 6

Conclusion and discussion

6.1 Summary of thesis achievement

Cancer is a highly complex and heterogeneous disease, with numerous possible molecular pathways contributing to the development of even the same cancer type. This complexity explains why the response to any given drug can be highly variable across patients[366]. Precision oncology aims to address this challenge by tailoring treatments based on the molecular profiles of individual patient tumours[367]. The rapid growth of publicly available omics data, doubling approximately every 6–7 months and projected to reach 40 million gigabytes annually within a decade [368], has facilitated the development of ML applications in cancer pharmacogenomics[369]. ML ideally leverage molecular cause-and-effect relationships altered in cancer to predict treatment outcomes and guide personalised therapies. These models aim to stratify patients likely to benefit from specific drugs in order to propose suitable drugs without delay and minimise harmful side effects for those unlikely to respond. Despite its potential, ML is still rarely applied to this type of problem, as they are challenged by high-dimensional low-sample data, which require particularly careful data preparation and analysis.

This PhD project is hence innovative in that it provides a proof of concept that this promising approach is viable. We reported here on the successful application of ML models using tumour gene expression profiles, as well as other molecular data types (such as miRNAs, isomiRs, DNA methylations, and CNVs), to predict patient responses across four drug-cancer-type binomials: atezolizumab in 320 mUC patients [2], [54], taxane-anthracycline regimens in 732 breast cancer patients [328], [329], doxorubicin in 96 TCGA-BRCA patients, and gemcitabine in 70 TCGA-PAAD patients. These large-scale ML analyses considering various molecular data types and ML algorithms led to the discovery of predictive, reproducible, and robust drug response signatures, facilitating the prediction of treatment responses in cancer patients. These retrospective proof-of-concept studies are necessary before initiating further prospective clinical applications in collaboration with healthcare providers.

The summary in Table 14 highlights the robustness and generality of our proposed methodology across these four case studies. The results consistently demonstrate strong predictive performance, showcasing the adaptability of our approach to various scenarios. Notably, the best-performing models differ for each case (different in combinations of molecular profiles and algorithms), highlighting the importance of incorporating diverse molecular profiles (e.g., mRNA, DNA methylation, CNVs, miRNAs and their isoform) tailored to various ML algorithms, to enhance the accuracy of drug response predictions.

Table 14. Overview of drug-cancer type binomials and best-performing models. This table summarises the data sources, the number and proportion of responder (R) and non-responder (NR), the number of profiles analysed and algorithms considered, and evaluation metrics (MCC, PR-AUC, and ROC-AUC) for the best-performing models for each drug-cancer type binomials. For atezolizumab-mUC, metrics were calculated using prediction on the validation dataset of the best OMC model, which was selected based on the highest MCC from the 10-fold CV on the discovery dataset. For other cases, metrics were calculated using merged predictions of a 10-fold CV run. In cases with the independent dataset (e.g., taxane-anthracycline-BC), metrics were calculated on the independent dataset using predictions from the best-performing model trained on the full discovery dataset. Random-level performance is as follows: MCC of 0.0, ROC-AUC of 0.5, and PR-AUCs dataset positive sample proportions: 0.682 (Snyder validation dataset), 0.766 (merged discovery and validation datasets), 0.203 (Hatzis discovery dataset), 0.318 (Magbanua independent dataset), 0.906 (doxorubicin-BC), and 0.414 (gemcitabine-PAAD).

Drug-cancer type	Dataset	#Patient	#Profile	#Algorithm	Best model	MCC	PR-AUC	ROC-AUC
Atezolizumab-mUC	IMVigor210 discovery dataset[2]	298 (R: 68, NR: 230)	1	8 (each with OMC)	mRNA (CART-OMC)	0.437	0.955	0.811
	Snyder validation dataset[54]	22 (R: 7, NR: 15)						
	Merged discovery and validation datasets[2], [54]	320 (R: 75, NR: 245)	1	8 (each with OMC)	mRNA (LGBM-OMC)	0.252	0.860	0.685
Taxane-anthracycline-BC	Hatzis discovery dataset[328]	488 (R: 99, NR: 389)	1	16 (8 algorithms each with and without OMC)	mRNA (LR)	0.396	0.473	0.785
	Magbanua independent dataset[329]	244 (R: 51, NR: 193)	The best-performing model from the Hatzis discovery dataset: mRNA (LR)			0.351	0.624	0.586
Doxorubicin-BC	TCGA-BRCA	96 (R: 87, NR: 9)	8	16 (8 algorithms each with and without OMC)	isomiR (CART)	0.557	0.979	0.804
					miRNA (CART)	0.32	0.643	0.964
Gemcitabine-PAAD	TCGA-PAAD	70 (R: 29, NR: 41)	8	16 (8 algorithms each with and without OMC)	mRNA (RF-OMC)	0.440	0.719	0.785
					mRNA (LR-OMC)	0.404	0.775	0.771
					CpG (XGBoost-OMC)	0.32	0.598	0.697

In our study on predicting atezolizumab response in mUC patients using tumour gene expression profiles and clinical data from two independent cohorts [2], [54], we demonstrated the use of regression algorithms for the classification task by applying a threshold to classify the predicted values into class labels (responder or non-responder). This approach allows the model to learn from all the patients' response labels (CR, PR, SD, and PD) rather than the more common information-wasting binary classes (CR and PR as responders and SD and PD as non-responders). By conducting five 10-fold CV runs on the discovery dataset to select the best OMC model, which was then tested on the validation dataset, this study identified CART-OMC as the best-performing model for predicting atezolizumab response in mUC patients, achieving the highest MCC of 0.437 (Figure 9 and Table 6). The CART-OMC nonlinearly combined only 29 (0.15%) out of the 19,441 genes considered (Table 7), with the feature importance generated by the model shown in Figures 14 and 15. Thus, in a prospective clinical trial, each mUC patient would only need to be profiled for these 29 genes before treatment, which would substantially reduce time and cost compared to determining the full profile. In addition, integrating gene expression profiles and clinical data, GEP + TMB + TNB (LR-OMC), did not improve predictive accuracy (Figure 10 and Table 8). This lack of improvement could be attributed to heterogeneity in TMB and TNB detection and quantification methods across different laboratories, which may reduce the predictive performance of models that incorporate TMB and TNB into their predictions.

The predictive performance of CART-OMC is robust in this context. First, CART-OMC outperformed immunotherapy response biomarkers, including TMB, TNB, and PD-L1 protein expression (Figure 11). Second, it demonstrated greater predictive performance than its permuted version, with accuracy improving as the training dataset size increased (Figure 12). Third, clustering 374 mUC patients from the discovery ($N = 348$) and validation ($N = 26$) datasets, including those with known and unknown atezolizumab response records, based on the 29 predictive genes identified by CART-OMC (Figure 13), revealed that the 298 mUC patients in the training dataset were well-represented across all clusters. These clusters also included the 26 mUC patients in the validation dataset and others with unknown responses, indicating that they fall within the model's applicability domain and are likely to achieve similar predictive accuracy for atezolizumab response. Fourth, CART-OMC could identify 29 genes potentially associated with atezolizumab response and its mechanism. PPI network analysis revealed that several of these genes were physically or functionally associated with CD274, indicating their activities may be influenced upon atezolizumab binding to PD-L1 (Figure 17). Wikicancer and GO pathway enrichment analysis revealed that a subset of these 29 genes was significantly enriched ($FDR < 0.05$) in 32 biological pathways related to cancer-associated and immunotherapy response mechanisms (Figure 18, Table 10). These 29 genes may contribute to atezolizumab resistance and cancer progression through alterations in their expressions as indicated by the differential expression pattern between responders and non-responders (Figure 16).

Lastly, we merged the discovery and validation datasets to create a more challenging test set. With a total of 320 mUC patients, we were able to generate predictions for those patients and

compare our model's performance with state-of-the-art ML drug response prediction models (e.g., EaSIeR and JADBio). Our best-performing models (LGBM-OMC with 49 genes, and CART with 63 genes with TMB and TNB) outperformed EaSIeR (Figure 20) and JADBio (Table 12) in predicting response to atezolizumab in 320 mUC patients, despite implementing a fully autoML approach and extensive tuning effort in JADBio. Furthermore, integrating predictive genes from the best-performing models with five TME descriptors, 9 published immune response scores, and the EaSIeR score, as computed for each patient using EaSIeR, improved the prediction of atezolizumab response (Figure 21). This integration provides a comprehensive approach that incorporates various dimensions of tumour-immune interactions, leading to improved predictive accuracy. Despite merging more features computed using EaSIeR providing a performance boost, we should consider the time and cost overhead of determining the full gene expression profiles in every patient for EaSIeR RNA-sequencing input to compute these features.

Due to the challenge problem in this study associated with predicting atezolizumab response in mUC patients, the EaSIeR model employed multi-task ML built on TCGA cancer patients who were not treated with anti-PD-1/anti PD-L1 inhibitors. Despite benefiting from the availability of many samples, a major limitation is that the model may not accurately predict atezolizumab response from non-immunotherapy-treated TCGA patients. Unlike our approach, which involves training the models on gene expression profiles and clinical data from mUC patients treated with atezolizumab. In addition, JADBio has a limited number of algorithms to build the model. Thus, it may need to consider including additional algorithms such as LGBM, which was the best-performing model in our study, to handle a wider range of datasets and achieve higher accuracy in predictions.

Atezolizumab response assessment in both studies relied on CT/MRI scans evaluated using RECIST version 1.1, which categorises patients as CR, PR, SD, or PD, with CR and PR grouped as responders, and SD and PD as non-responders. Although RECIST 1.1 provides standardised criteria, subtle differences in imaging resolution, contrast enhancement, sensitivity to tumour morphology, and interpretation may lead to inconsistencies in response categorisation. Beyond imaging-based assessment, heterogeneity in biopsy sampling and sequencing methodologies also contribute to data noise. Tumour biopsies often capture only a subset of the tumour, which may not fully represent the molecular heterogeneity present across different regions. The spatial distribution of these subclones influences gene expression patterns, potentially introducing noise into the data that impact the reliability of predictions. In addition, differences in sequencing platforms, protocols, and multi-centre sample integration introduce batch effects. Differences in sequencing depth can impact transcriptomic data accuracy, limiting the detection of low-abundance transcripts and potentially omitting biologically relevant signals. These technical discrepancies are evident in our study datasets. The IMvigor210 dataset employed the TruSeq RNA Access protocol (Illumina®), generating raw counts for 31,286 genes with Entrez Gene IDs, while the validation dataset from Snyder et al. utilized Illumina HiSeq 4000 with tximport processing, reporting 39,292 gene expressions with Ensembl Gene IDs. These technical discrepancies

necessitate rigorous normalisation strategies to mitigate batch effects and ensure cross-study comparability. Given these sources of noise, it is crucial to account for the inherent biological heterogeneity during model development and to consider how variations in biopsy sampling methods (e.g., core needle biopsy versus surgical resection) might affect the consistency and quality of the molecular data used in training predictive models. Addressing these challenges requires robust normalisation techniques, careful model training to account for biological variability, and an awareness of how sampling methods influence molecular profiles.

Building on our success with atezolizumab immunotherapy response prediction in mUC patients using gene expression profiles, we extended our work to develop drug response prediction models for three additional drug-cancer-type binomials involving different drug types (e.g., cytotoxic chemotherapies). For taxane-anthracycline chemotherapy in breast cancer patients, we could identify LR employing gene expression profiles as the best-performing model from five 10-fold CV runs on the discovery dataset (Figure 22). Notably, the LR model showed robustness and generalisability, achieving similar predictive performance on an independent dataset (Table 13). This suggests its applicability is not limited to a single cohort. Interestingly, the fact that the best-performing models are linear, despite testing non-linear models too, hints at the nonlinearity of the data.

Furthermore, we demonstrate the advantages of including more tumour molecular profiles, beyond gene expression alone, to enhance predictive accuracy. Owing to the wealth of curated profiling data from the GDC, we conducted a large-scale ML analysis of 128 ML models built on 8 molecular profiles per patient and 8 algorithms with and without OMC feature selection strategy to predict response to doxorubicin in breast cancer patients ($N = 96$) and gemcitabine in PAAD patients ($N = 70$). Small sample sizes are not uncommon in cancer-specific drug response prediction studies. Here, out-of-sample predictions were carried out using a 10-fold CV, which has been shown to effectively estimate model performance in small datasets. Interestingly, only 2 out of 128 models for doxorubicin-BC (Figure 23) and 7 models for gemcitabine-PAAD (Figure 24) achieved predictive performance with $mMCC > 0.3$, showing that such models could be easily missed by a standard-scale analysis limited to a few ML algorithms or single molecular profiles. For doxorubicin response in breast cancer, the two most predictive models built with CART: $mMCC$ of 0.56 from CART using four isomiRs and $mMCC$ of 0.32 from CART using three miRNAs. For gemcitabine response in PAAD, the two most predictive profiles were mRNA and DNA methylation: $mMCC$ of 0.44 from RF-OMC using four mRNAs, $mMCC$ of 0.4 from LR-OMC using five mRNAs, and $mMCC$ of 0.32 from XGBoost using 12 CpG probes. These findings highlight that using various ML algorithms tailored to molecular profiles beyond gene expression also demonstrates promising predictive power, which underlines the importance of conducting such a large-scale analysis to avoid missing the most predictive models.

Yet, the pharmacogenomics data are often high dimensional, as the number of features significantly outnumbers the available samples, which is challenging for many ML algorithms in biomedical research. In our study, the dataset included high-dimensional profiles,

including expression levels of 927 isomiRs to methylation levels of 450,000 DNA probes across only a few hundred patient samples. This necessitates efficient feature selection methods to identify the most informative predictors for a given outcome. Our group developed a scheme called OMC, which identifies a small subset of informative features from the much larger original feature space before model training and integrated OMC with various ML algorithms[44]–[46]. This approach increases the chance of making accurate predictions of ML models. For instance, the CART-OMC model selected only 29 genes (0.15% out of the total 19,441 genes) to predict atezolizumab response in mUC patients. Similarly, for gemcitabine response in breast cancer patients, the RF-OMC integrating four mRNAs (0.007% of 60,483), the LR-OMC incorporating five mRNAs (0.008% of 60,483), and the XGBoost-OMC combining 12 CpG probes (0.003% of 450,000).

While OMC can be highly predictive in certain cases, embedded feature selection methods within algorithms, such as CART, perform feature selection directly within the model itself and can be just as effective, or sometimes even more effective, in specific contexts. For example, in the doxorubicin-BC cases, CART models using either miRNAs or isoforms achieved the best predictive performance, selecting only three or four features out of approximately 1–2 thousand considered. However, since OMC works by ranking the features based on p-values derived from univariate statistical tests that assess the correlation between each feature and the corresponding label, it may lead to the selection of redundant features and does not account for multicollinearity, when two or more features are correlated.

In addition, hierarchical Bayesian models and causal ML techniques offer valuable approaches for handling high-dimensional data with low sample sizes. Hierarchical Bayesian models enhance predictive accuracy and generalisation by incorporating prior knowledge and sharing information across different levels (e.g., genes, patients), leading to more robust predictions and improved performance on unseen data[370], [371]. While ML algorithms are designed to identify patterns and correlations in data, they do not make explicit assumptions about causal relationships. For example, a RF model may successfully predict drug response in cancer patients, but it does not clarify how specific features influence the response. Causal inference, on the other hand, focus on identifying cause-and-effect relationships between molecular features and drug response[372], [373]. However, both hierarchical Bayesian models and causal ML techniques have limitations. One key challenge of hierarchical Bayesian models is the reliance on prior distributions, which can be difficult to specify accurately, particularly when prior knowledge is limited or uncertain. Causal ML techniques aim to uncover causal relationships rather than optimise predictive performance, while this study focuses on predictive modelling rather than causal discovery. Moreover, these models are computationally intensive and may struggle with scalability in large datasets.

Regarding the use of deep learning (DL) for drug response prediction, it often does not yield significant improvements[374], [375]. This is primarily because DL models typically require large amounts of data to perform well. However, in many biomedical datasets, including those used for drug response prediction, the available data is often limited in size due to the

high costs associated with experiments, making it difficult for DL models to outperform classical ML algorithms. In addition, DL models may not always generalise well on small datasets due to their complexity and the large number of parameters that require optimisation during the training process, making them prone to overfitting easily and leading to poor performance on new, unseen data. DL models also require more computational resources compared to ML methods, including high-performance computing units, powerful GPUs, and substantial storage capacity. Moreover, their "black boxes" nature makes it difficult to interpret how they make decisions. This lack of model interpretability can be a major drawback in fields of drug response prediction, where understanding the biological significance of predictors is crucial.

While multi-omics technology has significantly advanced the development of immunotherapy for mUC, it also encounters several challenges. Firstly, current sequencing technologies impose limitations on sequencing depth, which can impact the accuracy of transcriptomic data. This limitation may hinder the identification of low-frequency variations and result in the omission of crucial biological information. Additionally, significant batch effects often arise during cross-platform sequencing and multi-sample integration. There are substantial discrepancies in the standards for batch processing across various studies, which can reduce the reproducibility of bioinformatics analyses.

6.2 Clinical applications

This ML analysis has led to the discovery of predictive, reproducible, and robust drug response signatures, facilitating the prediction of treatment responses in cancer patients. These signatures outperformed single biomarkers currently used in clinical practice, which are often insufficient to stratify patient outcomes. Integrating ML into clinical practice has the potential to improve decision-making and enhance care efficiency. These models could assist clinicians in identifying patients most likely to benefit from treatment in order to propose suitable drugs without delay while minimising unnecessary exposure to ineffective therapies and reducing the risk of life-threatening side effects for patients unlikely to respond. In such a prospective clinical trial, each patient would only need to be profiled for specific signatures before treatment, which would reduce both the time and cost associated with determining the full profile. Furthermore, this study provides proof of concept and serves as a starting point for future research into the mechanisms of drug resistance, offering insights that could guide the development of novel therapeutic strategies.

We have made our ML models and processed datasets publicly available for prospective validation on patients with the same drug-cancer types. The taxane-anthracycline-BC model will be released upon publication:

- **Atezolizumab-mUC:** <https://codeocean.com/capsule/4702686/tree>
- **Doxorubicin-BC:** https://github.com/adeolu1/BRCA_CART_model
- **Gemcitabine-PAAD:** https://github.com/chayanitpiy/gemcitabine_PAAD

However, when evaluating the performance of predictive models for cancer drug response, it is essential to consider the clinical implications of false positives and false negatives. A false positive occurs when the model incorrectly predicts a patient will respond to a drug when, in reality, they do not. Clinically, this can lead to unnecessary treatment with a potentially ineffective drug, exposing the patient to side effects and delaying the administration of an alternative, more effective treatment. This scenario can be especially problematic in cases where treatment toxicity is a concern, such as with certain chemotherapy drugs or immunotherapies, which may have severe adverse effects.

On the other hand, a false negative happens when the model fails to predict a response in a patient who would have actually benefited from the drug. This is often considered more damaging, as it might result in the patient being deprived of a potentially life-saving treatment. In oncology, where treatment options are often limited and the prognosis can be poor, withholding a beneficial drug can significantly impact the patient's survival or quality of life.

In the context of cancer drug response prediction, minimising false negatives is typically prioritized to ensure that patients who could benefit from a treatment are not wrongly excluded from it. However, the trade-off between false positives and false negatives should be carefully considered in light of the clinical context. For example, in the case of drugs with low toxicity and high potential benefit, a higher rate of false positives may be acceptable. Conversely, for high-risk treatments with severe side effects, the consequences of a false positive may need to be weighed more carefully. Thus, the clinical utility of a model depends not only on traditional performance metrics e.g., MCC but also on the balance between false positives and false negatives, which should be tailored to the specific drug, cancer type, and patient population.

6.3 Limitations of the study

Although the best-performing models have shown predictive and robust performance, there are some limitations to be considered in this study. The analysis was based only on *in silico* and retrospective data from publicly available sources, which included a limited number of patients. A larger curated dataset should be further utilised to further improve model performance. In addition, an independent cohort would further test the generalisability of the best-performing model, i.e., whether this can stratify patients with different characteristics and account for technical variations (e.g., sample preparation, sequencing procedures, and bioinformatics pipelines). Furthermore, experimental validation (e.g., *in vitro*, *in vivo* experiments and clinical trials) is required to confirm the bioinformatics findings and to investigate the functional roles of predictive signatures in drug resistance and cancer progression.

6.4 Future works

The predictive model developed in this thesis for assessing drug response in cancer patients represents a significant step forward in personalised cancer therapy. Building upon these

promising results, this section outlines future research directions to enhance the model's accuracy, reliability, and clinical applicability.

- **Prospective clinical validation**

Although the current model has been validated retrospectively using publicly available large-scale, multi-omics cancer datasets, prospective validation through clinical trials is crucial to confirm its predictive power in real-world settings. Future studies should aim to implement the model in clinical workflows, enrolling patients prospectively to assess its utility in guiding treatment decisions. Collaborating with multiple institutions and research centres would facilitate access to larger, more diverse patient populations, which would enable the collection of heterogeneous data across different demographics, tumour types, and healthcare settings, thereby further refining the model's predictive ability, and enhancing the model's generalizability and robustness.

- **Integration of multi-omics data**

To achieve a more holistic understanding of cancer, future research should consider integrating multi-omics data sources. While gene expression data is the most widely used for predicting drug response, other omics, such as proteomics, transcriptomics, methylomics, histone post-translational modifications, and microbiomics also shape host responses to cancers. The multi-omics integration approach provides deeper insights into disease aetiology, where data from a single-omics cannot capture the complexity of all factors associated with disease[376]. Integrative models for predicting drug sensitivity have become a cornerstone of cancer research, with multi-omics ML and DL methods being increasingly proposed[377]–[380]. For instance, Sharifi-Noghabi et al.[380] developed the multi-omics with DNN model that integrates SNVs, CNVs and gene expression profiles for drug response prediction. However, molecular data, such as methylation profiles, are often high-dimensional (e.g., 450,000 features) while sample sizes remain limited. This can lead to overfitting and challenges in integrating multi-omics data. Moreover, the limited interpretability of DNNs poses challenges to their clinical applicability. Addressing the complexity and scale of such datasets requires advanced computational tools. Thus, future efforts should focus on developing integrative bioinformatics pipelines and ML algorithms capable of handling such high-dimensional data for accurate and meaningful interpretations.

- **Exploration of drug combinations**

A major challenge in cancer research and treatment is the ability of cancer cells to adapt, evolve, and develop resistance to drugs. It is expected that the future of cancer therapy will involve combining treatments to overcome this resistance. However, current knowledge of drug combinations in cancer patients remains limited due to the significant cost and time required to test all possible drug combinations. In this

context, ML methods offer a promising approach to addressing drug resistance and improving therapeutic efficacy. In recent years, various research initiatives have produced extensive databases on combinatorial drug therapy, such as DrugCombDB, SYNERGxDB, and NCI-ALMANAC, which facilitate the discovery of personalised, multitargeted drug combinations. These advancements present new opportunities for the application of large-scale ML methods in predicting effective drug combinations. Numerous studies from both academic and industry sectors have explored the prediction of combination therapies, achieving high predictive performance, as reviewed elsewhere[381], [382].

- **Utilisation of single-cell RNA sequencing technologies**

Single-cell RNA sequencing (scRNA-seq) enables the high-throughput characterisation of thousands of cells at the transcriptome level, making it a powerful tool for studying cellular components and their interactions within the TME. This technology has been applied to uncover associations between TME patterns and clinical outcomes and to investigate cell-specific drug responses in complex tissues. Recent advancements, such as scDRUG[383], [384], scDEAL[385], beyondCell[385], chemCPA[386], DREEP[387], and NCI-DOE[388], leverage scRNA-seq data and publicly available drug sensitivity profiles to predict drug response at the single-cell level. Despite its ability to address cellular heterogeneity, scRNA-seq faces challenges, including high cost, and technical complexity of large, high-dimensional datasets; the copious number of cells increases the dimensionality and scale of the data, highlighting the need for advanced computational methods to effectively process and interpret the data, ensuring its utility in clinical and research applications.

In conclusion, pursuing these research directions brings the field closer to personalised cancer treatment. The integration of prospective clinical validation, multi-omics data, drug combination strategies, and single-cell analyses holds significant promise for improving patient outcomes and quality of life. This framework not only expands the utility of current models but also encourages a broader exploration of integrated, patient-specific cancer treatment approaches. It also highlights the importance of interdisciplinary collaboration, technological innovation, and a patient-centred approach in driving progress in cancer therapy.

Bibliography

- [1] "Cancer Today." <https://gco.iarc.fr/today/en> (accessed Mar. 07, 2024).
- [2] A. V. Balar *et al.*, "Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial," *Lancet (London, England)*, vol. 389, no. 10064, pp. 67–76, Jan. 2017, doi: 10.1016/S0140-6736(16)32455-2.
- [3] C. Frick *et al.*, "Quantitative estimates of preventable and treatable deaths from 36 cancers worldwide: a population-based study," *Lancet Glob. Heal.*, vol. 11, no. 11, pp. e1700–e1712, Nov. 2023, doi: 10.1016/S2214-109X(23)00406-0.
- [4] N. Bukhari, H. O. Al-Shamsi, and F. Azam, "Update on the Treatment of Metastatic Urothelial Carcinoma," *ScientificWorldJournal.*, vol. 2018, 2018, doi: 10.1155/2018/5682078.
- [5] A. Dash *et al.*, "Impact of renal impairment on eligibility for adjuvant cisplatin-based chemotherapy in patients with urothelial carcinoma of the bladder," *Cancer*, vol. 107, no. 3, pp. 506–513, Aug. 2006, doi: 10.1002/CNCR.22031.
- [6] M. D. Galsky *et al.*, "Treatment of patients with metastatic urothelial cancer 'unfit' for Cisplatin-based chemotherapy," *J. Clin. Oncol.*, vol. 29, no. 17, pp. 2432–2438, Jun. 2011, doi: 10.1200/JCO.2011.34.8433.
- [7] M. D. Galsky *et al.*, "Cisplatin Ineligibility for Patients With Metastatic Urothelial Carcinoma: A Survey of Clinical Practice Perspectives Among US Oncologists," *Bl. Cancer*, vol. 5, no. 4, pp. 281–288, Jan. 2019, doi: 10.3233/BLC-190235.
- [8] K. S. Hanna, "Updates and novel treatments in urothelial carcinoma," *J. Oncol. Pharm. Pract.*, vol. 25, no. 3, pp. 648–656, Apr. 2019, doi: 10.1177/1078155218805141.
- [9] Y. Shiravand *et al.*, "Immune Checkpoint Inhibitors in Cancer Therapy," *Curr. Oncol. 2022, Vol. 29, Pages 3044-3060*, vol. 29, no. 5, pp. 3044–3060, Apr. 2022, doi: 10.3390/CURRONCOL29050247.
- [10] H. Houssiau and E. Seront, "Improving the role of immune checkpoint inhibitors in the management of advanced urothelial carcinoma, where do we stand?," *Transl. Oncol.*, vol. 19, p. 101387, May 2022, doi: 10.1016/J.TRANON.2022.101387.
- [11] A. Krishnamurthy and A. Jimeno, "Atezolizumab: A novel PD-L1 inhibitor in cancer therapy with a focus in bladder and non-small cell lung cancers," *Drugs Today (Barc.)*, vol. 53, no. 4, pp. 217–237, Apr. 2017, doi: 10.1358/DOT.2017.53.4.2589163.
- [12] P. C. Barata *et al.*, "Atezolizumab in Metastatic Urothelial Carcinoma Outside Clinical Trials: Focus on Efficacy, Safety, and Response to Subsequent Therapies," *Target. Oncol.*, vol. 13, no. 3, pp. 353–361, Jun. 2018, doi: 10.1007/S11523-018-0561-6.
- [13] M. A. Socinski *et al.*, "Atezolizumab for First-Line Treatment of Metastatic Nonsquamous NSCLC," *N. Engl. J. Med.*, vol. 378, no. 24, pp. 2288–2301, Jun. 2018, doi: 10.1056/NEJMOA1716948/SUPPL_FILE/NEJMOA1716948_DISCLOSURES.PDF.

- [14] R. S. Herbst *et al.*, “Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC,” *N. Engl. J. Med.*, vol. 383, no. 14, pp. 1328–1339, Oct. 2020, doi: 10.1056/NEJMOA1917346/SUPPL_FILE/NEJMOA1917346_DATA-SHARING.PDF.
- [15] S. M. Reddy, E. Carroll, and R. Nanda, “Atezolizumab for the treatment of breast cancer,” *Expert Rev. Anticancer Ther.*, vol. 20, no. 3, pp. 151–158, Mar. 2020, doi: 10.1080/14737140.2020.1732211.
- [16] J. Tang, L. Pearce, J. O’donnell-Tormey, and V. M. Hubbard-Lucey, “Erratum: Trends in the global immuno-oncology landscape (Nature Reviews Drug Discovery, (2018) 17, (783–784), 10.1038/nrd.2018.202),” *Nat. Rev. Drug Discov.*, vol. 17, no. 12, pp. 783–784, Nov. 2018, doi: 10.1038/NRD.2018.167.
- [17] E. B. Garon *et al.*, “Pembrolizumab for the treatment of non-small-cell lung cancer,” *N. Engl. J. Med.*, vol. 372, no. 21, pp. 2018–2028, May 2015, doi: 10.1056/NEJMOA1501824.
- [18] N. McGranahan *et al.*, “Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade,” *Science*, vol. 351, no. 6280, pp. 1463–1469, Mar. 2016, doi: 10.1126/SCIENCE.AAF1490.
- [19] A. Snyder *et al.*, “Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma,” *N. Engl. J. Med.*, vol. 371, no. 23, pp. 2189–2199, Dec. 2014, doi: 10.1056/NEJMOA1406498/SUPPL_FILE/NEJMOA1406498_DISCLOSURES.PDF.
- [20] A. Osipov *et al.*, “Tumor Mutational Burden, Toxicity, and Response of Immune Checkpoint Inhibitors Targeting PD(L)1, CTLA-4, and Combination: A Meta-regression Analysis,” *Clin. Cancer Res.*, vol. 26, no. 18, pp. 4842–4851, Sep. 2020, doi: 10.1158/1078-0432.CCR-20-0458.
- [21] X. Ma, Y. Zhang, S. Wang, and J. Yu, “Predictive value of tumor mutation burden (TMB) with targeted next-generation sequencing in immunocheckpoint inhibitors for non-small cell lung cancer (NSCLC),” *J. Cancer*, vol. 12, no. 2, p. 584, 2021, doi: 10.7150/JCA.48105.
- [22] R. Bai, Z. Lv, D. Xu, and J. Cui, “Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors,” *Biomark. Res. 2020 81*, vol. 8, no. 1, pp. 1–17, Aug. 2020, doi: 10.1186/S40364-020-00209-0.
- [23] M. K. Labriola *et al.*, “Characterization of tumor mutation burden, PD-L1 and DNA repair genes to assess relationship to immune checkpoint inhibitors response in metastatic renal cell carcinoma,” *J. Immunother. Cancer*, vol. 8, no. 1, p. e000319, Mar. 2020, doi: 10.1136/JITC-2019-000319.
- [24] J. M. Placke *et al.*, “Correlation of tumor PD-L1 expression in different tissue types and outcome of PD-1-based immunotherapy in metastatic melanoma – analysis of the DeCOG prospective multicenter cohort study ADOREG/TRIM,” *eBioMedicine*, vol. 96, Oct. 2023, doi: 10.1016/J.EBIOM.2023.104774.
- [25] T. Powles, J. Walker, J. Andrew Williams, and J. Bellmunt, “The evolving role of PD-L1

- testing in patients with metastatic urothelial carcinoma," *Cancer Treat. Rev.*, vol. 82, Jan. 2020, doi: 10.1016/J.CTRV.2019.101925.
- [26] M. A. Socinski *et al.*, "Atezolizumab for First-Line Treatment of Metastatic Nonsquamous NSCLC," *N. Engl. J. Med.*, vol. 378, no. 24, pp. 2288–2301, Jun. 2018, doi: 10.1056/NEJMOA1716948.
 - [27] T. Morihito *et al.*, "PD-L1 expression combined with microsatellite instability/CD8+ tumor infiltrating lymphocytes as a useful prognostic biomarker in gastric cancer," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/S41598-019-41177-2.
 - [28] T. S. K. Mok *et al.*, "Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial," *Lancet (London, England)*, vol. 393, no. 10183, pp. 1819–1830, May 2019, doi: 10.1016/S0140-6736(18)32409-7.
 - [29] L. Horn *et al.*, "Nivolumab Versus Docetaxel in Previously Treated Patients With Advanced Non-Small-Cell Lung Cancer: Two-Year Outcomes From Two Randomized, Open-Label, Phase III Trials (CheckMate 017 and CheckMate 057)," *J. Clin. Oncol.*, vol. 35, no. 35, pp. 3924–3933, Dec. 2017, doi: 10.1200/JCO.2017.74.3062.
 - [30] A. A. Davis and V. G. Patel, "The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors," *J. Immunother. cancer*, vol. 7, no. 1, Oct. 2019, doi: 10.1186/S40425-019-0768-9.
 - [31] L. M. Sholl *et al.*, "The Promises and Challenges of Tumor Mutation Burden as an Immunotherapy Biomarker: A Perspective from the International Association for the Study of Lung Cancer Pathology Committee," *J. Thorac. Oncol.*, vol. 15, no. 9, pp. 1409–1424, Sep. 2020, doi: 10.1016/J.JTHO.2020.05.019.
 - [32] N. I. Vokes *et al.*, "Harmonization of Tumor Mutational Burden Quantification and Association With Response to Immune Checkpoint Blockade in Non-Small-Cell Lung Cancer," *JCO Precis. Oncol.*, vol. 3, no. 3, pp. 1–12, Dec. 2019, doi: 10.1200/PO.19.00171.
 - [33] L. Fancello, S. Gandini, P. G. Pelicci, and L. Mazzearella, "Tumor mutational burden quantification from targeted gene panels: major advancements and challenges," *J. Immunother. cancer*, vol. 7, no. 1, Jul. 2019, doi: 10.1186/S40425-019-0647-4.
 - [34] C. Piyawajanusorn, L. C. Nguyen, G. Ghislat, and P. J. Ballester, "A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling," *Brief. Bioinform.*, vol. 22, no. 6, Nov. 2021, doi: 10.1093/BIB/BBAB312.
 - [35] A. Bomane, A. Gonçalves, and P. J. Ballester, "Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data," *Front. Genet.*, vol. 10, p. 1041, Oct. 2019, doi: 10.3389/FGENE.2019.01041/FULL.

- [36] A. Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat, and P. J. Ballester, "Interpretable Machine Learning Models to Predict the Resistance of Breast Cancer Patients to Doxorubicin from Their microRNA Profiles," *Adv. Sci. (Weinheim, Baden-Wurttemberg, Ger.)*, vol. 9, no. 24, Aug. 2022, doi: 10.1002/ADVS.202201501.
- [37] A. Ogunleye, C. Piyawajanusorn, G. Ghislat, and P. J. Ballester, "Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer," *Heal. Data Sci.*, Jan. 2024, doi: 10.34133/HDS.0108.
- [38] A. Y. Abuhelwa, G. Kichenadasse, R. A. McKinnon, A. Rowland, A. M. Hopkins, and M. J. Sorich, "Machine Learning for Prediction of Survival Outcomes with Immune-Checkpoint Inhibitors in Urothelial Cancer," *Cancers (Basel)*, vol. 13, no. 9, May 2021, doi: 10.3390/CANCERS13092001.
- [39] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015, doi: 10.1016/J.CSBJ.2014.11.005.
- [40] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform.*, vol. 2, pp. 59–77, Jan. 2006, doi: 10.1177/117693510600200030/ASSET/IMAGES/LARGE/10.1177_117693510600200030-FIG5.JPEG.
- [41] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment," *Medicina (B. Aires)*, vol. 56, no. 9, pp. 1–10, Sep. 2020, doi: 10.3390/MEDICINA56090455.
- [42] L. Nguyen, C. C. Dang, and P. J. Ballester, "Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data," *F1000Research*, vol. 5, p. 2927, Mar. 2016, doi: 10.12688/F1000RESEARCH.10529.2.
- [43] S. Naulaerts, C. C. Dang, and P. J. Ballester, "Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours," *Oncotarget*, vol. 8, no. 57, p. 97025, Nov. 2017, doi: 10.18632/ONCOTARGET.20923.
- [44] S. Naulaerts, M. P. Menden, and P. J. Ballester, "Concise Polygenic Models for Cancer-Specific Identification of Drug-Sensitive Tumors from Their Multi-Omics Profiles," *Biomolecules*, vol. 10, no. 6, pp. 1–21, Jun. 2020, doi: 10.3390/BIOM10060963.
- [45] L. C. Nguyen, S. Naulaerts, A. Bruna, G. Ghislat, and P. J. Ballester, "Predicting cancer drug response in vivo by learning an optimal feature selection of tumour molecular profiles," *Biomedicines*, vol. 9, no. 10, Oct. 2021, doi: 10.3390/BIOMEDICINES9101319.
- [46] A. Bomane, A. Gonçalves, and P. J. Ballester, "Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data," *Front. Genet.*, vol. 10, Oct. 2019, doi: 10.3389/FGENE.2019.01041.

- [47] D. Liu *et al.*, “Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma,” *Nat. Med.*, vol. 25, no. 12, p. 1916, Dec. 2019, doi: 10.1038/S41591-019-0654-5.
- [48] Ó. Lapuente-Santana, M. van Genderen, P. A. J. Hilbers, F. Finotello, and F. Eduati, “Interpretable systems biomarkers predict response to immune-checkpoint inhibitors,” *Patterns (New York, N.Y.)*, vol. 2, no. 8, Aug. 2021, doi: 10.1016/J.PATTERN.2021.100293.
- [49] P. Johannet *et al.*, “Using Machine Learning Algorithms to Predict Immunotherapy Response in Patients with Advanced Melanoma,” *Clin. Cancer Res.*, vol. 27, no. 1, pp. 131–140, Jan. 2021, doi: 10.1158/1078-0432.CCR-20-2415.
- [50] P. Charoentong *et al.*, “Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade,” *Cell Rep.*, vol. 18, no. 1, pp. 248–262, Jan. 2017, doi: 10.1016/J.CELREP.2016.12.019.
- [51] C. W. Abbott *et al.*, “Prediction of Immunotherapy Response in Melanoma through Combined Modeling of Neoantigen Burden and Immune-Related Resistance Mechanisms,” *Clin. Cancer Res.*, vol. 27, no. 15, pp. 4265–4276, Aug. 2021, doi: 10.1158/1078-0432.CCR-20-4314.
- [52] A. H. Nassar *et al.*, “A model combining clinical and genomic factors to predict response to PD-1/PD-L1 blockade in advanced urothelial carcinoma,” *Br. J. Cancer* 2019 1224, vol. 122, no. 4, pp. 555–563, Dec. 2019, doi: 10.1038/s41416-019-0686-0.
- [53] O. Lapuente-Santana and F. Eduati, “Toward Systems Biomarkers of Response to Immune Checkpoint Blockers,” *Front. Oncol.*, vol. 10, p. 530889, Jun. 2020, doi: 10.3389/FONC.2020.01027/BIBTEX.
- [54] A. Snyder *et al.*, “Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis,” *PLoS Med.*, vol. 14, no. 5, May 2017, doi: 10.1371/JOURNAL.PMED.1002309.
- [55] I. Tsamardinos *et al.*, “Just Add Data: automated predictive modeling for knowledge discovery and feature selection,” *NPJ Precis. Oncol.*, vol. 6, no. 1, Dec. 2022, doi: 10.1038/S41698-022-00274-8.
- [56] J. P. Yaxley, “Urinary tract cancers: An overview for general practice,” *J. Fam. Med. Prim. Care*, vol. 5, no. 3, p. 533, 2016, doi: 10.4103/2249-4863.197258.
- [57] J. Bellmunt *et al.*, “Recent therapeutic advances in urothelial carcinoma: A paradigm shift in disease management,” *Crit. Rev. Oncol. Hematol.*, vol. 174, p. 103683, Jun. 2022, doi: 10.1016/J.CRITREVONC.2022.103683.
- [58] O. Ogbue, A. Haddad, N. Almassi, J. Lapinski, and H. Daw, “Overview of histologic variants of urothelial carcinoma: current trends and narrative review on treatment outcomes,” *Transl. Androl. Urol.*, vol. 11, no. 6, pp. 877–901, Jun. 2022, doi:

10.21037/TAU-22-43/COIF.

- [59] L. Cardenas, P. Dibajnia, and A. K. Lalani, "The emerging treatment landscape of advanced urothelial carcinoma," *Curr. Opin. Support. Palliat. Care*, vol. 15, no. 4, pp. 247–252, Dec. 2021, doi: 10.1097/SPC.0000000000000573.
- [60] A. Mari *et al.*, "Patterns and predictors of recurrence after open radical cystectomy for bladder cancer: a comprehensive review of the literature," *World J. Urol.*, vol. 36, no. 2, pp. 157–170, Feb. 2018, doi: 10.1007/S00345-017-2115-4.
- [61] "Bladder Cancer — Cancer Stat Facts."
<https://seer.cancer.gov/statfacts/html/urinb.html> (accessed Mar. 07, 2024).
- [62] T. J. Okobi *et al.*, "Immune Checkpoint Inhibitors as a Treatment Option for Bladder Cancer: Current Evidence," *Cureus*, vol. 15, no. 6, Jun. 2023, doi: 10.7759/CUREUS.40031.
- [63] B. Rani, J. J. Ignatz-Hoover, P. S. Rana, and J. J. Driscoll, "Current and Emerging Strategies to Treat Urothelial Carcinoma," *Cancers (Basel)*, vol. 15, no. 19, Oct. 2023, doi: 10.3390/CANCERS15194886.
- [64] N. D. Freedman, D. T. Silverman, A. R. Hollenbeck, A. Schatzkin, and C. C. Abnet, "Association between smoking and risk of bladder cancer among men and women," *JAMA*, vol. 306, no. 7, pp. 737–745, Aug. 2011, doi: 10.1001/JAMA.2011.1142.
- [65] M. G. Cumberbatch, M. Rota, J. W. F. Catto, and C. La Vecchia, "The Role of Tobacco Smoke in Bladder and Kidney Carcinogenesis: A Comparison of Exposures and Meta-analysis of Incidence and Mortality Risks," *Eur. Urol.*, vol. 70, no. 3, pp. 458–466, Sep. 2016, doi: 10.1016/J.EURURO.2015.06.042.
- [66] S. Letaiová *et al.*, "Bladder cancer, a review of the environmental risk factors," *Environ. Health*, vol. 11 Suppl 1, no. Suppl 1, 2012, doi: 10.1186/1476-069X-11-S1-S11.
- [67] D. Theodorescu and B. Ehdaie, "Bladder Cancer," *Encycl. Cancer*, pp. 1–8, 2014, doi: 10.1007/978-3-642-27841-9_662-4.
- [68] L. A. Mucci *et al.*, "Familial Risk and Heritability of Cancer Among Twins in Nordic Countries," *JAMA*, vol. 315, no. 1, pp. 68–76, Jan. 2016, doi: 10.1001/JAMA.2015.17703.
- [69] M. I. Carlo *et al.*, "Cancer Susceptibility Mutations in Patients With Urothelial Malignancies," *J. Clin. Oncol.*, vol. 38, no. 5, p. 406, Feb. 2020, doi: 10.1200/JCO.19.01395.
- [70] A. P. Mitra, R. J. Cote, A. P. Mitra, R. J. Cote, A. P. & Mitra, and R. Cote, "Molecular screening for bladder cancer: progress and potential," *Nat. Rev. Urol.* 2010 71, vol. 7, no. 1, pp. 11–20, Jan. 2010, doi: 10.1038/nrurol.2009.236.
- [71] J. L. J. Vriesema, M. H. Poucki, L. A. L. M. Kiemeney, and J. A. Witjes, "Patient opinion of urinary tests versus flexible urethroscopy in follow-up examination for

- superficial bladder cancer: A utility analysis," *Urology*, vol. 56, no. 5, pp. 793–797, Nov. 2000, doi: 10.1016/S0090-4295(00)00777-9.
- [72] M. Jurewicz and M. S. Soloway, "Approaching the optimal transurethral resection of a bladder tumor," *Turkish J. Urol.*, vol. 40, no. 2, p. 73, 2014, doi: 10.5152/TUD.2014.94715.
 - [73] H. W. Herr and S. M. Donat, "Quality control in transurethral resection of bladder tumours," *BJU Int.*, vol. 102, no. 9 Pt B, pp. 1242–1246, Nov. 2008, doi: 10.1111/J.1464-410X.2008.07966.X.
 - [74] M. Brausi *et al.*, "Variability in the recurrence rate at first follow-up cystoscopy after TUR in stage Ta T1 transitional cell carcinoma of the bladder: A combined analysis of seven EORTC studies," *Eur. Urol.*, vol. 41, no. 5, pp. 523–531, May 2002, doi: 10.1016/S0302-2838(02)00068-4.
 - [75] V. T. Packiam *et al.*, "The association of salvage intravesical therapy following BCG with pathologic outcomes and survival after radical cystectomy for patients with high-grade non-muscle invasive bladder cancer: A multi-institution analysis," *Urol. Oncol.*, vol. 39, no. 7, pp. 436.e1-436.e8, Jul. 2021, doi: 10.1016/J.UROLONC.2021.01.004.
 - [76] R. F. M. Bevers, K. H. Kurth, and D. H. J. Schamhart, "Role of urothelial cells in BCG immunotherapy for superficial bladder cancer," *Br. J. Cancer*, vol. 91, no. 4, pp. 607–612, Aug. 2004, doi: 10.1038/SJ.BJC.6602026.
 - [77] R. J. Sylvester, A. P. M. Van der Meijden, and D. L. Lamm, "Intravesical bacillus Calmette-Guerin reduces the risk of progression in patients with superficial bladder cancer: a meta-analysis of the published results of randomized clinical trials," *J. Urol.*, vol. 168, no. 5, pp. 1964–1970, Nov. 2002, doi: 10.1016/S0022-5347(05)64273-5.
 - [78] A. Böhle and P. R. Bock, "Intravesical bacille Calmette-Guérin versus mitomycin C in superficial bladder cancer: Formal meta-analysis of comparative studies on tumor progression," *Urology*, vol. 63, no. 4, pp. 682–686, Apr. 2004, doi: 10.1016/j.urology.2003.11.049.
 - [79] A. Böhle, D. Jocham, and P. R. Bock, "Intravesical bacillus Calmette-Guerin versus mitomycin C for superficial bladder cancer: a formal meta-analysis of comparative studies on recurrence and toxicity," *J. Urol.*, vol. 169, no. 1, pp. 90–95, Jan. 2003, doi: 10.1016/S0022-5347(05)64043-8.
 - [80] R. Waked, J. Choucair, N. Chehata, E. Haddad, and G. Saliba, "Intravesical Bacillus Calmette-Guérin (BCG) treatment's severe complications: A single institution review of incidence, presentation and treatment outcome," *J. Clin. Tuberc. other Mycobact. Dis.*, vol. 19, May 2020, doi: 10.1016/J.JCTUBE.2020.100149.
 - [81] Y. B. Thyaviahally *et al.*, "Intravesical bacillus Calmette-Guerin (BCG) in treating non-muscle invasive bladder cancer-analysis of adverse effects and effectiveness of two strains of BCG (Danish 1331 and Moscow-I)," *Asian J. Urol.*, vol. 9, no. 2, pp. 157–164, Apr. 2022, doi: 10.1016/J.AJUR.2021.05.002.

- [82] Y. Liu, J. Lu, Y. Huang, and L. Ma, "Clinical Spectrum of Complications Induced by Intravesical Immunotherapy of Bacillus Calmette-Guérin for Bladder Cancer," *J. Oncol.*, vol. 2019, 2019, doi: 10.1155/2019/6230409.
- [83] W. Jäger *et al.*, "Early vs delayed radical cystectomy for 'high-risk' carcinoma not invading bladder muscle: delay of cystectomy reduces cancer-specific survival," *BJU Int.*, vol. 108, no. 8 Pt 2, Oct. 2011, doi: 10.1111/J.1464-410X.2010.09980.X.
- [84] C. Vale, "Neoadjuvant chemotherapy in invasive bladder cancer: A systematic review and meta-analysis," *Lancet*, vol. 361, no. 9373, pp. 1927–1934, Jun. 2003, doi: 10.1016/S0140-6736(03)13580-5.
- [85] P. J. Loehrer *et al.*, "A randomized comparison of cisplatin alone or in combination with methotrexate, vinblastine, and doxorubicin in patients with metastatic urothelial carcinoma: a cooperative group study," *J. Clin. Oncol.*, vol. 10, no. 7, pp. 1066–1073, 1992, doi: 10.1200/JCO.1992.10.7.1066.
- [86] H. Von der Maase *et al.*, "Gemcitabine and cisplatin versus methotrexate, vinblastine, doxorubicin, and cisplatin in advanced or metastatic bladder cancer: results of a large, randomized, multinational, multicenter, phase III study," *J. Clin. Oncol.*, vol. 18, no. 17, pp. 3068–3077, 2000, doi: 10.1200/JCO.2000.18.17.3068.
- [87] H. Von Der Maase *et al.*, "Long-term survival results of a randomized trial comparing gemcitabine plus cisplatin, with methotrexate, vinblastine, doxorubicin, plus cisplatin in patients with bladder cancer," *J. Clin. Oncol.*, vol. 23, no. 21, pp. 4602–4608, 2005, doi: 10.1200/JCO.2005.07.757.
- [88] R. Iacovelli *et al.*, "Management of metastatic urothelial carcinoma: Current approach, emerging agents, and future perspectives," <https://doi.org/10.1177/03915603221139907>, vol. 90, no. 1, pp. 3–10, Dec. 2022, doi: 10.1177/03915603221139907.
- [89] A. Bamias *et al.*, "Impact of contemporary patterns of chemotherapy utilization on survival in patients with advanced cancer of the urinary tract: a Retrospective International Study of Invasive/Advanced Cancer of the Urothelium (RISC)," *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 29, no. 2, pp. 361–369, Feb. 2018, doi: 10.1093/ANNONC/MDX692.
- [90] A. G. Ramsay, "Immune checkpoint blockade immunotherapy to activate anti-tumour T-cell immunity," *Br. J. Haematol.*, vol. 162, no. 3, pp. 313–325, Aug. 2013, doi: 10.1111/BJH.12380.
- [91] L. V. Riella, A. M. Paterson, A. H. Sharpe, and A. Chandraker, "Role of the PD-1 pathway in the immune response," *Am. J. Transplant*, vol. 12, no. 10, pp. 2575–2587, Oct. 2012, doi: 10.1111/J.1600-6143.2012.04224.X.
- [92] A. Akinleye and Z. Rasool, "Immune checkpoint inhibitors of PD-L1 as cancer therapeutics," *J. Hematol. Oncol.*, vol. 12, no. 1, Sep. 2019, doi: 10.1186/S13045-019-0779-5.

- [93] D. M. Pardoll, "The blockade of immune checkpoints in cancer immunotherapy," *Nat. Rev. Cancer*, vol. 12, no. 4, p. 252, Apr. 2012, doi: 10.1038/NRC3239.
- [94] H. Lingel and M. C. Brunner-Weinzierl, "CTLA-4 (CD152): A versatile receptor for immune-based therapy," *Semin. Immunol.*, vol. 42, Apr. 2019, doi: 10.1016/J.SMIM.2019.101298.
- [95] S. Van Coillie, B. Wiernicki, and J. Xu, "Molecular and Cellular Functions of CTLA-4," *Adv. Exp. Med. Biol.*, vol. 1248, pp. 7–32, 2020, doi: 10.1007/978-981-15-3266-5_2.
- [96] E. I. Buchbinder and A. Desai, "CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition," *Am. J. Clin. Oncol.*, vol. 39, no. 1, pp. 98–106, 2016, doi: 10.1097/COC.0000000000000239.
- [97] O. MA, G. RJ, M. DA, L. RB, and S. AH, "CTLA-4 and T cell activation," *Curr. Opin. Immunol.*, vol. 11, no. 3, pp. 294–300, 1999, doi: 10.1016/S0952-7915(99)80047-8.
- [98] D. M. Pardoll, "The blockade of immune checkpoints in cancer immunotherapy," *Nat. Rev. Cancer*, vol. 12, no. 4, p. 252, Apr. 2012, doi: 10.1038/NRC3239.
- [99] M. A. Postow, M. K. Callahan, and J. D. Wolchok, "Immune Checkpoint Blockade in Cancer Therapy," *J. Clin. Oncol.*, vol. 33, no. 17, pp. 1974–1982, Jun. 2015, doi: 10.1200/JCO.2014.59.4358.
- [100] J. Bellmunt *et al.*, "Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma," *N. Engl. J. Med.*, vol. 376, no. 11, pp. 1015–1026, Mar. 2017, doi: 10.1056/NEJMOA1613683.
- [101] T. Powles *et al.*, "Efficacy and Safety of Durvalumab in Locally Advanced or Metastatic Urothelial Carcinoma: Updated Results From a Phase 1/2 Open-label Study," *JAMA Oncol.*, vol. 3, no. 9, Sep. 2017, doi: 10.1001/JAMAONCOL.2017.2411.
- [102] J. Bellmunt, T. Powles, and N. J. Vogelzang, "A review on the evolution of PD-1/PD-L1 immunotherapy for bladder cancer: The future is now," *Cancer Treat. Rev.*, vol. 54, pp. 58–67, Mar. 2017, doi: 10.1016/J.CTRV.2017.01.007.
- [103] M. Reck *et al.*, "Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer," *N. Engl. J. Med.*, vol. 375, no. 19, pp. 1823–1833, Nov. 2016, doi: 10.1056/NEJMOA1606774.
- [104] J. Bellmunt *et al.*, "Association of PD-L1 expression on tumor-infiltrating mononuclear cells and overall survival in patients with urothelial carcinoma," *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 26, no. 4, pp. 812–817, Apr. 2015, doi: 10.1093/ANNONC/MDV009.
- [105] R. L. Ferris *et al.*, "Nivolumab for Recurrent Squamous-Cell Carcinoma of the Head and Neck," *N. Engl. J. Med.*, vol. 375, no. 19, pp. 1856–1867, Nov. 2016, doi: 10.1056/NEJMOA1602252.
- [106] F. Pierantoni *et al.*, "Immunotherapy and urothelial carcinoma: An overview and future perspectives," *Crit. Rev. Oncol. Hematol.*, vol. 143, pp. 46–55, Nov. 2019, doi:

10.1016/J.CRITREVONC.2019.08.005.

- [107] E. R. Plimack *et al.*, "Safety and activity of pembrolizumab in patients with locally advanced or metastatic urothelial cancer (KEYNOTE-012): a non-randomised, open-label, phase 1b study," *Lancet Oncol.*, vol. 18, no. 2, pp. 212–220, Feb. 2017, doi: 10.1016/S1470-2045(17)30007-4.
- [108] P. Sharma *et al.*, "Nivolumab in metastatic urothelial carcinoma after platinum therapy (CheckMate 275): a multicentre, single-arm, phase 2 trial," *Lancet Oncol.*, vol. 18, no. 3, pp. 312–322, Mar. 2017, doi: 10.1016/S1470-2045(17)30065-7.
- [109] J. E. Rosenberg *et al.*, "Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial," *Lancet*, vol. 387, no. 10031, pp. 1909–1920, May 2016, doi: 10.1016/S0140-6736(16)00561-4.
- [110] M. R. Patel *et al.*, "Avelumab in metastatic urothelial carcinoma after platinum failure (JAVELIN Solid Tumor): pooled results from two expansion cohorts of an open-label, phase 1 trial," *Lancet Oncol.*, vol. 19, no. 1, pp. 51–64, Jan. 2018, doi: 10.1016/S1470-2045(17)30900-2.
- [111] C. Massard *et al.*, "Safety and efficacy of durvalumab (MEDI4736), an anti-programmed cell death ligand-1 immune checkpoint inhibitor, in patients with advanced urothelial bladder cancer," *J. Clin. Oncol.*, vol. 34, no. 26, pp. 3119–3125, Sep. 2016, doi: 10.1200/JCO.2016.67.9761/ASSET/IMAGES/LARGE/JCO679761F3.JPEG.
- [112] A. V. Balar *et al.*, "First-line pembrolizumab in cisplatin-ineligible patients with locally advanced and unresectable or metastatic urothelial cancer (KEYNOTE-052): a multicentre, single-arm, phase 2 study," *Lancet Oncol.*, vol. 18, no. 11, pp. 1483–1492, Nov. 2017, doi: 10.1016/S1470-2045(17)30616-2.
- [113] J. Bellmunt *et al.*, "Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma," *N. Engl. J. Med.*, vol. 376, no. 11, pp. 1015–1026, Mar. 2017, doi: 10.1056/NEJMOA1613683.
- [114] T. Powles *et al.*, "Efficacy and Safety of Durvalumab in Locally Advanced or Metastatic Urothelial Carcinoma: Updated Results From a Phase 1/2 Open-label Study," *JAMA Oncol.*, vol. 3, no. 9, pp. e172411–e172411, Sep. 2017, doi: 10.1001/JAMAONCOL.2017.2411.
- [115] T. Powles *et al.*, "Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial," *Lancet*, vol. 391, no. 10122, pp. 748–757, Feb. 2018, doi: 10.1016/S0140-6736(17)33297-X.
- [116] A. B. Apolo *et al.*, "Avelumab, an anti-programmed death-ligand 1 antibody, in patients with refractory metastatic urothelial carcinoma: Results from a multicenter, Phase Ib study," *J. Clin. Oncol.*, vol. 35, no. 19, pp. 2117–2124, Jul. 2017, doi: 10.1200/JCO.2016.71.6795.

- [117] B. A. Inman, T. A. Longo, S. Ramalingam, and M. R. Harrison, "Atezolizumab: A PD-L1-Blocking Antibody for Bladder Cancer," *Clin. Cancer Res.*, vol. 23, no. 8, pp. 1886–1890, Apr. 2017, doi: 10.1158/1078-0432.CCR-16-1417.
- [118] Y. Ning *et al.*, "FDA Approval Summary: Atezolizumab for the Treatment of Patients with Progressive Advanced Urothelial Carcinoma after Platinum-Containing Chemotherapy," *Oncologist*, vol. 22, no. 6, p. 743, Jun. 2017, doi: 10.1634/THEONCOLOGIST.2017-0087.
- [119] T. Powles *et al.*, "MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer," *Nat. 2014 5157528*, vol. 515, no. 7528, pp. 558–562, Nov. 2014, doi: 10.1038/nature13904.
- [120] E. A. Eisenhauer *et al.*, "New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)," *Eur. J. Cancer*, vol. 45, pp. 228–247, doi: 10.1016/j.ejca.2008.10.026.
- [121] M. D. Galsky *et al.*, "Atezolizumab with or without chemotherapy in metastatic urothelial cancer (IMvigor130): a multicentre, randomised, placebo-controlled phase 3 trial," *Lancet*, vol. 395, no. 10236, pp. 1547–1557, May 2020, doi: 10.1016/S0140-6736(20)30230-0.
- [122] T. Powles *et al.*, "MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer," *Nature*, vol. 515, no. 7528, pp. 558–562, Nov. 2014, doi: 10.1038/NATURE13904.
- [123] E. Cha, J. Wallin, and M. Kowanetz, "PD-L1 inhibition with MPDL3280A for solid tumors," *Semin. Oncol.*, vol. 42, no. 3, pp. 484–487, Jun. 2015, doi: 10.1053/J.SEMINONCOL.2015.02.002.
- [124] A. Markham, "Atezolizumab: First Global Approval," *Drugs*, vol. 76, no. 12, pp. 1227–1232, Aug. 2016, doi: 10.1007/S40265-016-0618-8/METRICS.
- [125] Q. Yin *et al.*, "Immune-related adverse events of immune checkpoint inhibitors: a review," *Front. Immunol.*, vol. 14, p. 1167975, May 2023, doi: 10.3389/FIMMU.2023.1167975/BIBTEX.
- [126] D. B. Johnson, C. A. Nebhan, J. J. Moslehi, and J. M. Balko, "Immune-checkpoint inhibitors: long-term implications of toxicity," *Nat. Rev. Clin. Oncol.*, vol. 19, no. 4, p. 254, Apr. 2022, doi: 10.1038/S41571-022-00600-W.
- [127] M. A. Socinski *et al.*, "Association of Immune-Related Adverse Events With Efficacy of Atezolizumab in Patients With Non-Small Cell Lung Cancer: Pooled Analyses of the Phase 3 IMpower130, IMpower132, and IMpower150 Randomized Clinical Trials," *JAMA Oncol.*, vol. 9, no. 4, p. 527, Apr. 2023, doi: 10.1001/JAMAONCOL.2022.7711.
- [128] S. Baxi *et al.*, "Immune-related adverse events for anti-PD-1 and anti-PD-L1 drugs: systematic review and meta-analysis," *BMJ*, vol. 360, 2018, doi: 10.1136/BMJ.K793.
- [129] K. Komiya *et al.*, "Discontinuation due to immune-related adverse events is a possible predictive factor for immune checkpoint inhibitors in patients with non-small cell

- lung cancer," *Thorac. cancer*, vol. 10, no. 9, pp. 1798–1804, Sep. 2019, doi: 10.1111/1759-7714.13149.
- [130] K. Esfahani *et al.*, "Moving towards personalized treatments of immune-related adverse events," *Nat. Rev. Clin. Oncol.*, vol. 17, no. 8, pp. 504–515, Aug. 2020, doi: 10.1038/S41571-020-0352-8.
- [131] I. Puzanov *et al.*, "Managing toxicities associated with immune checkpoint inhibitors: consensus recommendations from the Society for Immunotherapy of Cancer (SITC) Toxicity Management Working Group," *J. Immunother. cancer*, vol. 5, no. 1, Nov. 2017, doi: 10.1186/S40425-017-0300-Z.
- [132] G. De Velasco *et al.*, "Comprehensive Meta-analysis of Key Immune-Related Adverse Events from CTLA-4 and PD-1/PD-L1 Inhibitors in Cancer Patients," *Cancer Immunol. Res.*, vol. 5, no. 4, pp. 312–318, Apr. 2017, doi: 10.1158/2326-6066.CIR-16-0237.
- [133] R. Barroso-Sousa *et al.*, "Incidence of Endocrine Dysfunction Following the Use of Different Immune Checkpoint Inhibitor Regimens: A Systematic Review and Meta-analysis," *JAMA Oncol.*, vol. 4, no. 2, pp. 173–182, Feb. 2018, doi: 10.1001/JAMAONCOL.2017.3064.
- [134] D. J. Lee, H. J. Lee, J. R. Farmer, and K. L. Reynolds, "Mechanisms Driving Immune-Related Adverse Events in Cancer Patients Treated with Immune Checkpoint Inhibitors," *Curr. Cardiol. Rep.*, vol. 23, no. 8, Aug. 2021, doi: 10.1007/S11886-021-01530-2.
- [135] D. Y. Wang *et al.*, "Fatal Toxic Effects Associated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis," *JAMA Oncol.*, vol. 4, no. 12, pp. 1721–1728, Dec. 2018, doi: 10.1001/JAMAONCOL.2018.3923.
- [136] M. A. Postow, R. Sidlow, and M. D. Hellmann, "Immune-Related Adverse Events Associated with Immune Checkpoint Blockade," *N. Engl. J. Med.*, vol. 378, no. 2, pp. 158–168, Jan. 2018, doi: 10.1056/NEJMRA1703481.
- [137] S. Kimbara *et al.*, "Association of antithyroglobulin antibodies with the development of thyroid dysfunction induced by nivolumab," *Cancer Sci.*, vol. 109, no. 11, pp. 3583–3590, Nov. 2018, doi: 10.1111/CAS.13800.
- [138] S. Iwama, A. De Remigis, M. K. Callahan, S. F. Slovin, J. D. Wolchok, and P. Caturegli, "Pituitary expression of CTLA-4 mediates hypophysitis secondary to administration of CTLA-4 blocking antibody," *Sci. Transl. Med.*, vol. 6, no. 230, Apr. 2014, doi: 10.1126/SCITRANSLMED.3008002.
- [139] B. Baban, J. Y. Liu, X. Qin, N. L. Weintraub, and M. S. Mozaffari, "Upregulation of Programmed Death-1 and Its Ligand in Cardiac Injury Models: Interaction with GADD153," *PLoS One*, vol. 10, no. 4, Apr. 2015, doi: 10.1371/JOURNAL.PONE.0124059.
- [140] J. M. Taube *et al.*, "Association of PD-1, PD-1 ligands, and other features of the tumor immune microenvironment with response to anti-PD-1 therapy," *Clin. Cancer Res.*,

- vol. 20, no. 19, pp. 5064–5074, Oct. 2014, doi: 10.1158/1078-0432.CCR-13-3271.
- [141] S. L. Topalian *et al.*, “Safety, activity, and immune correlates of anti-PD-1 antibody in cancer,” *N. Engl. J. Med.*, vol. 366, no. 26, pp. 2443–2454, Jun. 2012, doi: 10.1056/NEJMOA1200690.
 - [142] J. R. Brahmer *et al.*, “Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates,” *J. Clin. Oncol.*, vol. 28, no. 19, pp. 3167–3175, Jul. 2010, doi: 10.1200/JCO.2009.26.7609.
 - [143] H. Li, P. A. van der Merwe, and S. Sivakumar, “Biomarkers of response to PD-1 pathway blockade,” *Br. J. Cancer* 2022 12612, vol. 126, no. 12, pp. 1663–1675, Feb. 2022, doi: 10.1038/s41416-022-01743-4.
 - [144] C. Massard *et al.*, “Safety and Efficacy of Durvalumab (MEDI4736), an Anti-Programmed Cell Death Ligand-1 Immune Checkpoint Inhibitor, in Patients With Advanced Urothelial Bladder Cancer,” *J. Clin. Oncol.*, vol. 34, no. 26, pp. 3119–3125, Sep. 2016, doi: 10.1200/JCO.2016.67.9761.
 - [145] T. Powles *et al.*, “Efficacy and Safety of Durvalumab in Locally Advanced or Metastatic Urothelial Carcinoma: Updated Results From a Phase 1/2 Open-label Study,” *JAMA Oncol.*, vol. 3, no. 9, pp. e172411–e172411, Sep. 2017, doi: 10.1001/JAMAONCOL.2017.2411.
 - [146] J. Huang and X. Teng, “Expression of PD-L1 for predicting response to immune checkpoint inhibitors in metastatic urothelial carcinoma: a systematic review and meta-analysis,” *Curr. Oncol.*, vol. 27, no. 6, p. e656, 2020, doi: 10.3747/CO.27.6437.
 - [147] C. Pang *et al.*, “Assessment of programmed cell death ligand-1 expression with multiple immunohistochemistry antibody clones in non-small cell lung cancer,” *J. Thorac. Dis.*, vol. 10, no. 2, pp. 816–824, Feb. 2018, doi: 10.21037/JTD.2018.01.124.
 - [148] E. A. Prince, J. K. Sanzari, D. Pandya, D. Huron, and R. Edwards, “Analytical Concordance of PD-L1 Assays Utilizing Antibodies From FDA-Approved Diagnostics in Advanced Cancers: A Systematic Literature Review,” *JCO Precis. Oncol.*, vol. 5, no. 5, pp. 953–973, Nov. 2021, doi: 10.1200/PO.20.00412.
 - [149] A. A. Davis and V. G. Patel, “The role of PD-L1 expression as a predictive biomarker: an analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors,” *J. Immunother. cancer*, vol. 7, no. 1, Oct. 2019, doi: 10.1186/S40425-019-0768-9.
 - [150] L. B. Alexandrov *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, pp. 415–421, 2013, doi: 10.1038/NATURE12477.
 - [151] T. N. Schumacher and R. D. Schreiber, “Neoantigens in cancer immunotherapy,” *Science*, vol. 348, no. 6230, pp. 69–74, Apr. 2015, doi: 10.1126/SCIENCE.AAA4971.
 - [152] S. J. Klempner *et al.*, “Tumor Mutational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence,”

- Oncologist*, vol. 25, no. 1, pp. e147–e159, Jan. 2020, doi: 10.1634/THEONCOLOGIST.2019-0244.
- [153] C. S. Shemesh *et al.*, “Pan-cancer population pharmacokinetics and exposure-safety and -efficacy analyses of atezolizumab in patients with high tumor mutational burden,” *Pharmacol. Res. Perspect.*, vol. 8, no. 6, Dec. 2020, doi: 10.1002/PRP2.685.
 - [154] M. Kowanetz *et al.*, “OA20.01 Tumor Mutation Burden (TMB) is Associated with Improved Efficacy of Atezolizumab in 1L and 2L+ NSCLC Patients Neoantigen Targeting in NSCLC Patients with Complete Response to Anti-PD-1 Immunotherapy,” 2017, doi: 10.1016/j.jtho.2016.11.343.
 - [155] P. G. Coulie, B. J. Van Den Eynde, P. Van Der Bruggen, and T. Boon, “Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy,” *Nat. Rev. Cancer*, vol. 14, no. 2, pp. 135–146, Feb. 2014, doi: 10.1038/NRC3670.
 - [156] M. D. M. Leiserson *et al.*, “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes,” *Nat. Genet.* 2014 472, vol. 47, no. 2, pp. 106–114, Dec. 2014, doi: 10.1038/ng.3168.
 - [157] L. B. Ivashkiv, “IFN γ : signalling, epigenetics and roles in immunity, metabolism, disease and cancer immunotherapy,” *Nat. Rev. Immunol.*, vol. 18, no. 9, pp. 545–558, Sep. 2018, doi: 10.1038/S41577-018-0029-Z.
 - [158] N. A. Rizvi *et al.*, “Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer,” *Science*, vol. 348, no. 6230, pp. 124–128, Apr. 2015, doi: 10.1126/SCIENCE.AAA1348.
 - [159] A. M. Goodman *et al.*, “Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers,” *Mol. Cancer Ther.*, vol. 16, no. 11, pp. 2598–2608, Nov. 2017, doi: 10.1158/1535-7163.MCT-17-0386.
 - [160] H. Rizvi *et al.*, “Molecular Determinants of Response to Anti-Programmed Cell Death (PD)-1 and Anti-Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With Non-Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing,” *J. Clin. Oncol.*, vol. 36, no. 7, pp. 633–641, Mar. 2018, doi: 10.1200/JCO.2017.75.3384.
 - [161] R. Woodhouse *et al.*, “Clinical and analytical validation of FoundationOne Liquid CDx, a novel 324-Gene cfDNA-based comprehensive genomic profiling assay for cancers of solid tumor origin,” *PLoS One*, vol. 15, no. 9, Sep. 2020, doi: 10.1371/JOURNAL.PONE.0237802.
 - [162] V. Tateo, P. V. Marchese, V. Mollica, F. Massari, R. Kurzrock, and J. J. Adashek, “Agnostic Approvals in Oncology: Getting the Right Drug to the Right Patient with the Right Genomics,” *Pharm. 2023, Vol. 16, Page 614*, vol. 16, no. 4, p. 614, Apr. 2023, doi: 10.3390/PH16040614.
 - [163] S. Chen, N. Zhang, T. Wang, E. Zhang, X. Wang, and J. Zheng, “Biomarkers of the Response to Immune Checkpoint Inhibitors in Metastatic Urothelial Carcinoma,” *Front. Immunol.*, vol. 11, Aug. 2020, doi: 10.3389/FIMMU.2020.01900/FULL.

- [164] P. Wang, Y. Chen, and C. Wang, "Beyond Tumor Mutation Burden: Tumor Neoantigen Burden as a Biomarker for Immunotherapy and Other Types of Therapy," *Front. Oncol.*, vol. 11, Apr. 2021, doi: 10.3389/FONC.2021.672677.
- [165] S. A. Shukla *et al.*, "Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes," *Nat. Biotechnol.* 2015 3311, vol. 33, no. 11, pp. 1152–1158, Sep. 2015, doi: 10.1038/nbt.3344.
- [166] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015, doi: 10.1093/BIOINFORMATICS/BTU638.
- [167] M. Nielsen *et al.*, "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations," *Protein Sci.*, vol. 12, no. 5, pp. 1007–1017, May 2003, doi: 10.1110/PS.0239403.
- [168] S. Kim *et al.*, "Neoepesee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information," *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 29, no. 4, pp. 1030–1036, Apr. 2018, doi: 10.1093/ANNONC/MDY022.
- [169] Z. R. Chalmers *et al.*, "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden," *Genome Med.*, vol. 9, no. 1, Apr. 2017, doi: 10.1186/S13073-017-0424-2.
- [170] G. Roviello, M. Catalano, S. Nobili, R. Santi, E. Mini, and G. Nesi, "Molecular Sciences Review Focus on Biochemical and Clinical Predictors of Response to Immune Checkpoint Inhibitors in Metastatic Urothelial Carcinoma: Where Do We Stand?," 2016, doi: 10.3390/ijms21217935.
- [171] T. Gevaert *et al.*, "Predictive biomarkers for immunotherapy in the treatment of advanced urothelial carcinoma: where we stand and where we go," *Future Oncol.*, vol. 15, no. 19, pp. 2199–2202, Jul. 2019, doi: 10.2217/FON-2019-0217.
- [172] J. B. Baker *et al.*, "Tumour gene expression predicts response to cetuximab in patients with KRAS wild-type metastatic colorectal cancer," *Br. J. Cancer* 2011 1043, vol. 104, no. 3, pp. 488–495, Jan. 2011, doi: 10.1038/sj.bjc.6606054.
- [173] H. K. Kim *et al.*, "Three-gene predictor of clinical outcome for gastric cancer patients treated with chemotherapy," *Pharmacogenomics J.* 2012 122, vol. 12, no. 2, pp. 119–127, Dec. 2010, doi: 10.1038/tpj.2010.87.
- [174] N. B. Jamieson and A. V. Maker, "Gene-expression profiling to predict responsiveness to immunotherapy," *Cancer Gene Ther.*, vol. 24, no. 3, pp. 134–140, Mar. 2017, doi: 10.1038/CGT.2016.63.
- [175] A. Ribas *et al.*, "Association of response to programmed death receptor 1 (PD-1) blockade with pembrolizumab (MK-3475) with an interferon-inflammatory immune gene signature.," https://doi.org/10.1200/jco.2015.33.15_suppl.3001, vol. 33, no. 15_suppl, pp. 3001–3001, May 2015, doi: 10.1200/JCO.2015.33.15_SUPPL.3001.

- [176] J. Fu *et al.*, “Large-scale public data reuse to model immunotherapy response and resistance,” *Genome Med.*, vol. 12, no. 1, Feb. 2020, doi: 10.1186/S13073-020-0721-Z.
- [177] P. Jiang *et al.*, “Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response,” *Nat. Med.*, vol. 24, no. 10, pp. 1550–1558, Oct. 2018, doi: 10.1038/S41591-018-0136-1.
- [178] P. H. O’Donnell *et al.*, “Biomarker findings and mature clinical results from KEYNOTE-052: First-line pembrolizumab (pembro) in cisplatin-ineligible advanced urothelial cancer (UC).,” https://doi.org/10.1200/JCO.2017.35.15_suppl.4502, vol. 35, no. 15_suppl, pp. 4502–4502, May 2017, doi: 10.1200/JCO.2017.35.15_SUPPL.4502.
- [179] S. Mariathasan *et al.*, “TGF β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells,” *Nature*, vol. 554, no. 7693, pp. 544–548, Feb. 2018, doi: 10.1038/NATURE25501.
- [180] T. B. Powles *et al.*, “6990 Avelumab first-line (1L) maintenance + best supportive care (BSC) vs BSC alone for advanced urothelial carcinoma (UC): Association between clinical outcomes and exploratory biomarkers,” *Ann. Oncol.*, vol. 31, pp. S552–S553, Sep. 2020, doi: 10.1016/J.ANNONC.2020.08.771.
- [181] P. Sharma *et al.*, “Nivolumab in metastatic urothelial carcinoma after platinum therapy (CheckMate 275): a multicentre, single-arm, phase 2 trial,” *Lancet. Oncol.*, vol. 18, no. 3, pp. 312–322, Mar. 2017, doi: 10.1016/S1470-2045(17)30065-7.
- [182] D. Damotte *et al.*, “The tumor inflammation signature (TIS) is associated with anti-PD-1 treatment benefit in the CERTIM pan-cancer cohort,” *J. Transl. Med.*, vol. 17, no. 1, p. 357, Nov. 2019, doi: 10.1186/s12967-019-2100-3.
- [183] C. C. Dang, A. Peón, and P. J. Ballester, “Unearthing new genomic markers of drug response by improved measurement of discriminative power,” *BMC Med. Genomics*, vol. 11, no. 1, pp. 1–14, Feb. 2018, doi: 10.1186/S12920-018-0336-Z/FIGURES/8.
- [184] S. Kim, V. Sundaresan, L. Zhou, and T. Kahveci, “Integrating Domain Specific Knowledge and Network Analysis to Predict Drug Sensitivity of Cancer Cell Lines,” *PLoS One*, vol. 11, no. 9, p. e0162173, Sep. 2016, doi: 10.1371/JOURNAL.PONE.0162173.
- [185] F. Weeber, S. N. Ooft, K. K. Dijkstra, and E. E. Voest, “Tumor Organoids as a Pre-clinical Cancer Model for Drug Discovery,” *Cell Chem. Biol.*, vol. 24, no. 9, pp. 1092–1100, Sep. 2017, doi: 10.1016/J.CHEMBIOL.2017.06.012.
- [186] Y. Kim, D. Kim, B. Cao, R. Carvajal, and M. Kim, “PDXGEM: patient-derived tumor xenograft-based gene expression model for predicting clinical response to anticancer therapy in cancer patients,” *BMC Bioinformatics*, vol. 21, no. 1, Jul. 2020, doi: 10.1186/S12859-020-03633-Z.
- [187] A. Partin *et al.*, “Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images,” *Front. Med.*, vol. 10, p. 1058919, Mar. 2023, doi:

10.3389/FMED.2023.1058919/BIBTEX.

- [188] K. Matsuo *et al.*, "Survival outcome prediction in cervical cancer: Cox models vs deep-learning model," *Am. J. Obstet. Gynecol.*, vol. 220, no. 4, p. 381.e1, Apr. 2018, doi: 10.1016/J.AJOG.2018.12.030.
- [189] T. R. Golub *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, Oct. 1999, doi: 10.1126/SCIENCE.286.5439.531.
- [190] S. Ramaswamy *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 26, pp. 15149–15154, Dec. 2001, doi: 10.1073/PNAS.211566398.
- [191] C. M. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.
- [192] L. J. Van't Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002, doi: 10.1038/415530A.
- [193] K. Shedden *et al.*, "Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nat. Med.*, vol. 14, no. 8, pp. 822–827, Aug. 2008, doi: 10.1038/NM.1790.
- [194] J. Guinney *et al.*, "The consensus molecular subtypes of colorectal cancer," *Nat. Med.*, vol. 21, no. 11, pp. 1350–1356, Nov. 2015, doi: 10.1038/NM.3967.
- [195] D. Capper *et al.*, "DNA methylation-based classification of central nervous system tumours," *Nature*, vol. 555, no. 7697, pp. 469–474, Mar. 2018, doi: 10.1038/NATURE26000.
- [196] E. Papaemmanuil *et al.*, "Genomic Classification and Prognosis in Acute Myeloid Leukemia," *N. Engl. J. Med.*, vol. 374, no. 23, pp. 2209–2221, Jun. 2016, doi: 10.1056/NEJMOA1516192.
- [197] B. B. Campbell *et al.*, "Comprehensive Analysis of Hypermutation in Human Cancer," *Cell*, vol. 171, no. 5, pp. 1042–1056.e10, Nov. 2017, doi: 10.1016/J.CELL.2017.09.048.
- [198] K. A. Hoadley *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, Aug. 2014, doi: 10.1016/J.CELL.2014.06.049.
- [199] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/NG.2764.
- [200] J. S. Lee *et al.*, "Synthetic lethality-mediated precision oncology via the tumor transcriptome," *Cell*, vol. 184, no. 9, pp. 2487–2502.e13, Apr. 2021, doi: 10.1016/J.CELL.2021.03.030.
- [201] N. Auslander *et al.*, "Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma," *Nat. Med.*, vol. 24, no. 10, pp. 1545–1549, Oct. 2018, doi: 10.1038/S41591-018-0157-9.

- [202] A. Bagaev *et al.*, “Conserved pan-cancer microenvironment subtypes predict response to immunotherapy,” *Cancer Cell*, vol. 39, no. 6, pp. 845–865.e7, Jun. 2021, doi: 10.1016/J.CCELL.2021.04.014.
- [203] R. Sun *et al.*, “A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study,” *Lancet. Oncol.*, vol. 19, no. 9, pp. 1180–1191, Sep. 2018, doi: 10.1016/S1470-2045(18)30413-3.
- [204] K. J. Park, J. L. Lee, S. K. Yoon, C. Heo, B. W. Park, and J. K. Kim, “Radiomics-based prediction model for outcomes of PD-1/PD-L1 immunotherapy in metastatic urothelial carcinoma,” *Eur. Radiol.*, vol. 30, no. 10, pp. 5392–5403, Oct. 2020, doi: 10.1007/S00330-020-06847-0.
- [205] Y. Yang *et al.*, “A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer,” *Am. J. Transl. Res.*, vol. 13, no. 2, p. 743, 2021, Accessed: Mar. 13, 2024. [Online]. Available: [/pmc/articles/PMC7868825/](https://pmc/articles/PMC7868825/)
- [206] A. Prat *et al.*, “Immune-Related Gene Expression Profiling After PD-1 Blockade in Non-Small Cell Lung Carcinoma, Head and Neck Squamous Cell Carcinoma, and Melanoma,” *Cancer Res.*, vol. 77, no. 13, pp. 3540–3550, Jul. 2017, doi: 10.1158/0008-5472.CAN-16-3556.
- [207] S. Hwang *et al.*, “Immune gene signatures for predicting durable clinical benefit of anti-PD-1 immunotherapy in patients with non-small cell lung cancer,” *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/S41598-019-57218-9.
- [208] S. Shang *et al.*, “MeImmS: Predict Clinical Benefit of Anti-PD-1/PD-L1 Treatments Based on DNA Methylation in Non-small Cell Lung Cancer,” *Front. Genet.*, vol. 12, p. 676449, May 2021, doi: 10.3389/FGENE.2021.676449/FULL.
- [209] S. Benzekry *et al.*, “Machine Learning for Prediction of Immunotherapy Efficacy in Non-Small Cell Lung Cancer from Simple Clinical and Biological Data,” *Cancers (Basel)*, vol. 13, no. 24, Dec. 2021, doi: 10.3390/CANCERS13246210.
- [210] J. H. Kong *et al.*, “Network-based machine learning approach to predict immunotherapy response in cancer patients,” *Nat. Commun.* 2022 131, vol. 13, no. 1, pp. 1–15, Jun. 2022, doi: 10.1038/s41467-022-31535-6.
- [211] A. Tabari *et al.*, “Machine Learning Improves the Prediction of Responses to Immune Checkpoint Inhibitors in Metastatic Melanoma,” *Cancers (Basel)*, vol. 15, no. 10, May 2023, doi: 10.3390/CANCERS15102700.
- [212] O. Lapuente-Santana and F. Eduati, “Toward Systems Biomarkers of Response to Immune Checkpoint Blockers,” *Front. Oncol.*, vol. 10, Jun. 2020, doi: 10.3389/FONC.2020.01027.
- [213] W. H. Fridman, L. Zitvogel, C. Sautès-Fridman, and G. Kroemer, “The immune contexture in cancer prognosis and treatment,” *Nat. Rev. Clin. Oncol.*, vol. 14, no. 12,

- pp. 717–734, Dec. 2017, doi: 10.1038/NRCLINONC.2017.101.
- [214] J. Galon and D. Bruni, “Approaches to treat immune hot, altered and cold tumours with combination immunotherapies,” *Nat. Rev. Drug Discov.*, vol. 18, no. 3, pp. 197–218, Mar. 2019, doi: 10.1038/S41573-018-0007-Y.
 - [215] S. Spranger and T. F. Gajewski, “Impact of oncogenic pathways on evasion of antitumour immune responses,” *Nat. Rev. Cancer*, vol. 18, no. 3, pp. 139–147, Mar. 2018, doi: 10.1038/NRC.2017.117.
 - [216] I. Zeldes, A. Matikas, J. Bergh, G. Z. Rassidakis, and T. Foukakis, “Genetic, transcriptional and post-translational regulation of the programmed death protein ligand 1 in cancer: biology and clinical correlations,” *Oncogene*, vol. 37, no. 34, pp. 4639–4661, Aug. 2018, doi: 10.1038/S41388-018-0303-3.
 - [217] R. Tokunaga *et al.*, “CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation - A target for novel cancer therapy,” *Cancer Treat. Rev.*, vol. 63, pp. 40–47, Feb. 2018, doi: 10.1016/J.CTRV.2017.11.007.
 - [218] N. Nagarsheth, M. S. Wicha, and W. Zou, “Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy,” *Nat. Rev. Immunol.*, vol. 17, no. 9, pp. 559–572, Sep. 2017, doi: 10.1038/NRI.2017.49.
 - [219] A. M. Cornel, I. L. Mimpfen, and S. Nierkens, “MHC Class I Downregulation in Cancer: Underlying Mechanisms and Potential Targets for Cancer Immunotherapy,” *Cancers (Basel)*, vol. 12, no. 7, pp. 1–33, Jul. 2020, doi: 10.3390/CANCERS12071760.
 - [220] J. A. Ramilowski *et al.*, “A draft network of ligand-receptor-mediated multicellular signalling in human,” *Nat. Commun.*, vol. 6, Jul. 2015, doi: 10.1038/NCOMMS8866.
 - [221] M. P. Kumar *et al.*, “Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics,” *Cell Rep.*, vol. 25, no. 6, pp. 1458–1468.e4, Nov. 2018, doi: 10.1016/J.CELREP.2018.10.047.
 - [222] A. Mustafa and M. Rahimi Azghadi, “Automated Machine Learning for Healthcare and Clinical Notes Analysis,” *Comput. 2021, Vol. 10, Page 24*, vol. 10, no. 2, p. 24, Feb. 2021, doi: 10.3390/COMPUTERS10020024.
 - [223] Z. D. Stephens *et al.*, “Big Data: Astronomical or Genomical?,” *PLoS Biol.*, vol. 13, no. 7, 2015, doi: 10.1371/JOURNAL.PBIO.1002195.
 - [224] M. D. Robinson and A. Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biol.*, vol. 11, no. 3, p. R25, Mar. 2010, doi: 10.1186/GB-2010-11-3-R25.
 - [225] X. H. Cao, I. Stojkovic, and Z. Obradovic, “A robust data scaling algorithm to improve classification accuracies in biomedical data,” *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–10, Sep. 2016, doi: 10.1186/S12859-016-1236-X/TABLES/3.
 - [226] S. Fotouhi, S. Asadi, and M. W. Kattan, “A comprehensive data level analysis for cancer diagnosis on imbalanced data,” *J. Biomed. Inform.*, vol. 90, Feb. 2019, doi:

10.1016/J.JBI.2018.12.003.

- [227] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artif. Intell. Med.*, vol. 128, Jun. 2022, doi: 10.1016/J.ARTMED.2022.102289.
- [228] G. Drouard *et al.*, "Exploring machine learning strategies for predicting cardiovascular disease risk factors from multi-omic data," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/S12911-024-02521-3.
- [229] Y. Ren, D. Wu, Y. Tong, A. López-DeFede, and S. Gareau, "Issue of Data Imbalance on Low Birthweight Baby Outcomes Prediction and Associated Risk Factors Identification: Establishment of Benchmarking Key Machine Learning Models With Data Rebalancing Strategies," *J. Med. Internet Res.*, vol. 25, 2023, doi: 10.2196/44081.
- [230] J. Zhu *et al.*, "Processing imbalanced medical data at the data level with assisted-reproduction data as an example," *BioData Min.*, vol. 17, no. 1, p. 29, Dec. 2024, doi: 10.1186/S13040-024-00384-Y.
- [231] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Appl. Soft Comput.*, vol. 143, p. 110415, Aug. 2023, doi: 10.1016/J.ASOC.2023.110415.
- [232] Y. Yang, H. A. Khorshidi, and U. Aickelin, "A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems," *Front. Digit. Heal.*, vol. 6, 2024, doi: 10.3389/FDGTH.2024.1430245.
- [233] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," *J. Big Data*, vol. 8, no. 1, pp. 1–41, Dec. 2021, doi: 10.1186/S40537-020-00390-X/TABLES/49.
- [234] Z. Zheng, Y. Cai, Y. Li, Z. Zheng, Y. Cai, and Y. Li, "Oversampling Method for Imbalanced Classification," *Comput. Informatics*, vol. 34, no. 5, pp. 1017–1037, 2015, Accessed: Mar. 12, 2025. [Online]. Available: <https://www.cai.sk/ojs/index.php/cai/article/view/1277>
- [235] M. S. Kraiem, F. Sánchez-Hernández, and M. N. Moreno-García, "Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models," *Appl. Sci. 2021, Vol. 11, Page 8546*, vol. 11, no. 18, p. 8546, Sep. 2021, doi: 10.3390/APP11188546.
- [236] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/JAIR.953.
- [237] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, p. 106, Mar. 2013, doi: 10.1186/1471-2105-14-106.
- [238] X. Gu, P. P. Angelov, and E. A. Soares, "A self-adaptive synthetic over-sampling technique for imbalanced classification," *Int. J. Intell. Syst.*, vol. 35, no. 6, pp. 923–943, Jun. 2020, doi: 10.1002/INT.22230.

- [239] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proc. Int. Jt. Conf. Neural Networks*, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.
- [240] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Inform.*, vol. 90, Feb. 2019, doi: 10.1016/J.JBI.2018.12.003.
- [241] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Inf. Sci. (Ny)*, vol. 572, pp. 574–589, Sep. 2021, doi: 10.1016/J.INS.2021.02.056.
- [242] Y. Yang and G. Mirzaei, "Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification," *PLoS One*, vol. 19, no. 2, Feb. 2024, doi: 10.1371/JOURNAL.PONE.0293607.
- [243] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.
- [244] A. Anand, G. Pugalenth, G. B. Fogel, and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino Acids*, vol. 39, no. 5, pp. 1385–1391, Nov. 2010, doi: 10.1007/S00726-010-0595-2.
- [245] A. Kim and I. Jung, "Optimal selection of resampling methods for imbalanced data with high complexity," *PLoS One*, vol. 18, no. 7, p. e0288540, Jul. 2023, doi: 10.1371/JOURNAL.PONE.0288540.
- [246] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, p. 122778, Jun. 2024, doi: 10.1016/J.ESWA.2023.122778.
- [247] E. F. ; Swana *et al.*, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors 2022, Vol. 22, Page 3246*, vol. 22, no. 9, p. 3246, Apr. 2022, doi: 10.3390/S22093246.
- [248] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
- [249] L. Bao, C. Juan, J. Li, and Y. Zhang, "Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets," *Neurocomputing*, vol. 172, pp. 198–206, Jan. 2016, doi: 10.1016/J.NEUCOM.2014.05.096.
- [250] A. S. Palli, J. Jaafar, M. A. Hashmani, H. M. Gomes, and A. R. Gilal, "A Hybrid Sampling Approach for Imbalanced Binary and Multi-Class Data Using Clustering Analysis," *IEEE Access*, vol. 10, pp. 118639–118653, 2022, doi: 10.1109/ACCESS.2022.3218463.
- [251] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-

- imbalanced data," *Inf. Sci. (Ny)*, vol. 409–410, pp. 17–26, Oct. 2017, doi: 10.1016/J.INS.2017.05.008.
- [252] B. Satish and K. P. Supreethi, "An independent condensed nearest neighbor classification technique for medical image retrieval," *J. Ambient Intell. Humaniz. Comput.*, vol. 1, pp. 1–11, Mar. 2021, doi: 10.1007/S12652-021-03028-9/TABLES/4.
- [253] H. Liu, Z. Liu, W. Jia, D. Zhang, and J. Tan, "A Novel Imbalanced Data Classification Method Based on Weakly Supervised Learning for Fault Diagnosis," *IEEE Trans. Ind. Informatics*, vol. 18, no. 3, pp. 1583–1593, Mar. 2022, doi: 10.1109/TII.2021.3084132.
- [254] R. Walsh and M. Tardy, "A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer," *Diagnostics 2023, Vol. 13, Page 67*, vol. 13, no. 1, p. 67, Dec. 2022, doi: 10.3390/DIAGNOSTICS13010067.
- [255] V. Kumar *et al.*, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthc. 2022, Vol. 10, Page 1293*, vol. 10, no. 7, p. 1293, Jul. 2022, doi: 10.3390/HEALTHCARE10071293.
- [256] B. BAKIRARAR and A. H. ELHAN, "Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research," *Turkiye Klin. J. Biostat.*, vol. 15, no. 1, pp. 19–29, 2023, doi: 10.5336/BIOSTATIC.2022-93961.
- [257] I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: a review," *Artif. Intell. Rev. 2024 574*, vol. 57, no. 4, pp. 1–72, Mar. 2024, doi: 10.1007/S10462-023-10652-8.
- [258] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *J. Big Data*, vol. 10, no. 1, pp. 1–22, Dec. 2023, doi: 10.1186/S40537-023-00738-Z/TABLES/37.
- [259] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning," *J. Chem. Inf. Model.*, vol. 61, no. 6, pp. 2623–2640, Jun. 2021, doi: 10.1021/ACS.JCIM.1C00160/ASSET/IMAGES/LARGE/CI1C00160_0015.JPEG.
- [260] Z. Xu and A. De, "Assessing model accuracy using random data split: a simulation study," *J. Biopharm. Stat.*, vol. 33, no. 2, pp. 131–139, 2023, doi: 10.1080/10543406.2022.2089158.
- [261] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, and J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl. Inf. Syst.*, vol. 65, no. 1, pp. 31–57, Jan. 2023, doi: 10.1007/S10115-022-01772-8/TABLES/7.
- [262] S. Huang, P. Hu, and T. M. Lakowski, "Predicting breast cancer drug response using a multiple-layer cell line drug response network model," *BMC Cancer*, vol. 21, no. 1, pp. 1–15, Dec. 2021, doi: 10.1186/S12885-021-08359-6/TABLES/3.
- [263] L. C. Hermida, E. M. Gertz, and E. Ruppín, "Predicting cancer prognosis and drug

- response from the tumor microbiome," *Nat. Commun.* 2022 131, vol. 13, no. 1, pp. 1–15, May 2022, doi: 10.1038/s41467-022-30512-3.
- [264] S. N. Dorman *et al.*, "Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning," *Mol. Oncol.*, vol. 10, no. 1, pp. 85–100, Jan. 2016, doi: 10.1016/J.MOLONC.2015.07.006.
- [265] M. Frejno *et al.*, "Pharmacoproteomic characterisation of human colon and rectal cancer," *Mol. Syst. Biol.*, vol. 13, no. 11, Nov. 2017, doi: 10.15252/MSB.20177701.
- [266] S. Chawla *et al.*, "Gene expression based inference of cancer drug sensitivity," *Nat. Commun.* 2022 131, vol. 13, no. 1, pp. 1–15, Sep. 2022, doi: 10.1038/s41467-022-33291-z.
- [267] T. R. Vetter and P. Schober, "Regression: The apple does not fall far from the tree," *Anesth. Analg.*, vol. 127, no. 1, pp. 277–283, Jul. 2018, doi: 10.1213/ANE.00000000000003424.
- [268] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," *Classif. Regres. Trees*, pp. 1–358, Jan. 2017, doi: 10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-OLSHEN-CHARLES-STONE.
- [269] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch. psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/J.ISSN.1002-0829.215044.
- [270] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: an overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, Oct. 2002, doi: 10.1023/A:1016409317640.
- [271] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.
- [272] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, Jun. 2012, doi: 10.1016/J.YGENO.2012.04.003.
- [273] R. Gonzalez, A. Saha, C. J. V. Campbell, P. Nejat, C. Lokker, and A. P. Norgan, "Seeing the random forest through the decision trees. Supporting learning health systems from histopathology with machine learning models: Challenges and opportunities," *J. Pathol. Inform.*, vol. 15, Dec. 2023, doi: 10.1016/J.JPI.2023.100347.
- [274] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.
- [275] M. Teixeira, F. Silva, R. M. Ferreira, T. Pereira, C. Figueiredo, and H. P. Oliveira, "A review of machine learning methods for cancer characterization from microbiome data," *npj Precis. Oncol.* 2024 81, vol. 8, no. 1, pp. 1–16, May 2024, doi: 10.1038/s41698-024-00617-7.

- [276] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, Accessed: Mar. 18, 2024. [Online]. Available: <https://github.com/Microsoft/LightGBM>.
- [277] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [278] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/S10462-020-09896-5/TABLES/12.
- [279] R. Schwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Inf. Fusion*, vol. 81, pp. 84–90, May 2022, doi: 10.1016/J.INFFUS.2021.11.011.
- [280] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support Vector Regression Machines”.
- [281] C. Cortes, “Support-Vector Networks,” vol. 20, pp. 273–297, 1995.
- [282] W. S. Noble, “What is a support vector machine?,” *Nat. Biotechnol.* 2006 2412, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.
- [283] A. Ben-Hur and J. Weston, “A user’s guide to support vector machines,” *Methods Mol. Biol.*, vol. 609, pp. 223–239, 2010, doi: 10.1007/978-1-60327-241-4_13.
- [284] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics,” *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, Jan. 2018, doi: 10.21873/CGP.20063.
- [285] R. Rodríguez-Pérez and J. Bajorath, “Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery,” *J. Comput. Aided. Mol. Des.*, vol. 36, no. 5, pp. 355–362, May 2022, doi: 10.1007/S10822-022-00442-9.
- [286] Z. Zhang, “Introduction to machine learning: k-nearest neighbors,” *Ann. Transl. Med.*, vol. 4, no. 11, Jun. 2016, doi: 10.21037/ATM.2016.03.37.
- [287] T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.
- [288] H. Raeisi Shahraki, S. Pourahmad, and N. Zare, “K Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data,” *Biomed Res. Int.*, vol. 2017, p. 7560807, 2017, doi: 10.1155/2017/7560807.
- [289] Y. Jung and J. Hu, “A K-fold Averaging Cross-validation Procedure,” *J. Nonparametr. Stat.*, vol. 27, no. 2, pp. 167–179, Apr. 2015, doi: 10.1080/10485252.2015.1010532.
- [290] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, “Benchmark of filter methods for feature selection in high-dimensional gene expression survival data,” *Brief. Bioinform.*, vol. 23, no. 1, Jan. 2022, doi: 10.1093/BIB/BBAB354.
- [291] C. Lazar *et al.*, “A survey on filter techniques for feature selection in gene expression microarray analysis,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 9, no. 4, pp.

- 1106–1119, 2012, doi: 10.1109/TCBB.2012.33.
- [292] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Comput. Biol. Med.*, vol. 112, p. 103375, Sep. 2019, doi: 10.1016/J.COMPBIOMED.2019.103375.
- [293] J. H. Cabot and E. G. Ross, “Evaluating Prediction Model Performance,” *Surgery*, vol. 174, no. 3, p. 723, Sep. 2023, doi: 10.1016/J.SURG.2023.05.023.
- [294] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, no. 1, p. 6086, Dec. 2024, doi: 10.1038/S41598-024-56706-X.
- [295] B. W. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochim. Biophys. Acta*, vol. 405, no. 2, pp. 442–451, Oct. 1975, doi: 10.1016/0005-2795(75)90109-9.
- [296] J. Yao and M. Shepperd, “Assessing software defection prediction performance: Why using the Matthews correlation coefficient matters,” *ACM Int. Conf. Proceeding Ser.*, pp. 120–129, Apr. 2020, doi: 10.1145/3383219.3383232.
- [297] Q. Zhu, “On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset,” *Pattern Recognit. Lett.*, vol. 136, pp. 71–80, Aug. 2020, doi: 10.1016/J.PATREC.2020.03.030.
- [298] A. Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat, and P. J. Ballester, “Interpretable Machine Learning Models to Predict the Resistance of Breast Cancer Patients to Doxorubicin from Their microRNA Profiles,” *Adv. Sci.*, vol. 9, no. 24, Aug. 2022, doi: 10.1002/ADVS.202201501.
- [299] A. Ogunleye, C. Piyawajanusorn, G. Ghislat, and P. J. Ballester, “Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer,” *Heal. Data Sci.*, vol. 4, Jan. 2024, doi: 10.34133/HDS.0108.
- [300] S. M. Saqlain *et al.*, “Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines,” *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, Jan. 2019, doi: 10.1007/S10115-018-1185-Y/FIGURES/12.
- [301] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/S12864-019-6413-7.
- [302] D. Chicco, N. Tötsch, and G. Jurman, “The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation,” *BioData Min.*, vol. 14, no. 1, pp. 1–22, 2021, doi: 10.1186/S13040-021-00244-Z.
- [303] D. Chicco, M. J. Warrens, and G. Jurman, “The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen’s Kappa and Brier Score in Binary Classification

- Assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021, doi: 10.1109/ACCESS.2021.3084050.
- [304] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Oct. 2020, Accessed: Nov. 22, 2024. [Online]. Available: <https://arxiv.org/abs/2010.16061v1>
- [305] F. Movahedi, R. Padman, and J. F. Antaki, "Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores," *J. Thorac. Cardiovasc. Surg.*, vol. 165, no. 4, pp. 1433–1442.e2, Apr. 2023, doi: 10.1016/J.JTCVS.2021.07.041.
- [306] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/JOURNAL.PONE.0118432.
- [307] B. Ozenne, F. Subtil, and D. Maucourt-Boulch, "The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *J. Clin. Epidemiol.*, vol. 68, no. 8, pp. 855–859, Aug. 2015, doi: 10.1016/J.JCLINEPI.2015.02.010.
- [308] R. Bai, Z. Lv, D. Xu, and J. Cui, "Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors," *Biomark. Res.*, vol. 8, no. 1, Aug. 2020, doi: 10.1186/S40364-020-00209-0.
- [309] M. D. Galsky *et al.*, "Nivolumab in Patients with Advanced Platinum-resistant Urothelial Carcinoma: Efficacy, Safety, and Biomarker Analyses with Extended Follow-up from CheckMate 275," *Clin. Cancer Res.*, vol. 26, no. 19, pp. 5120–5128, Oct. 2020, doi: 10.1158/1078-0432.CCR-19-4162.
- [310] D. Sha, Z. Jin, J. Budczies, K. Kluck, A. Stenzinger, and F. A. Sinicrope, "Tumor Mutational Burden as a Predictive Biomarker in Solid Tumors," *Cancer Discov.*, vol. 10, no. 12, pp. 1808–1825, Dec. 2020, doi: 10.1158/2159-8290.CD-20-0522.
- [311] R. Cristescu *et al.*, "Tumor mutational burden predicts the efficacy of pembrolizumab monotherapy: a pan-tumor retrospective analysis of participants with advanced solid tumors," *J. Immunother. Cancer*, vol. 10, no. 1, p. e003091, Jan. 2022, doi: 10.1136/JITC-2021-003091.
- [312] J. Huang and X. Teng, "Expression of PD-L1 for predicting response to immune checkpoint inhibitors in metastatic urothelial carcinoma: a systematic review and meta-analysis," *Curr. Oncol.*, vol. 27, no. 6, pp. e656–e663, 2020, doi: 10.3747/CO.27.6437.
- [313] J. Zouein, C. Kesrouani, and H. R. Kourie, "PD-L1 expression as a predictive biomarker for immune checkpoint inhibitors: between a dream and a nightmare," *Immunotherapy*, vol. 13, no. 12, pp. 1053–1065, Aug. 2021, doi: 10.2217/IMT-2020-0336.
- [314] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and

- prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, Apr. 2015, doi: 10.11919/J.ISSN.1002-0829.215044.
- [315] D. Szklarczyk *et al.*, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. Database issue, p. D607, Jan. 2019, doi: 10.1093/NAR/GKY1131.
 - [316] G. Ghislat *et al.*, "NF- κ B-dependent IRF1 activation programs cDC1 dendritic cells to drive antitumor immunity," *Sci. Immunol.*, vol. 6, no. 61, Jul. 2021, doi: 10.1126/SCIIMMUNOL.ABG3570.
 - [317] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W199–W205, Jul. 2019, doi: 10.1093/NAR/GKZ401.
 - [318] N. Karachaliou *et al.*, "Interferon gamma, an important marker of response to immune checkpoint blockade in non-small cell lung cancer and melanoma patients," *Ther. Adv. Med. Oncol.*, vol. 10, Jan. 2018, doi: 10.1177/1758834017749748.
 - [319] M. Ayers *et al.*, "IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade," *J. Clin. Invest.*, vol. 127, no. 8, pp. 2930–2940, Aug. 2017, doi: 10.1172/JCI91190.
 - [320] A. Amedei, D. Prisco, and M. M. D'Elios, "The Use of Cytokines and Chemokines in the Cancer Immunotherapy," *Recent Pat. Anticancer. Drug Discov.*, vol. 8, no. 2, pp. 126–142, Apr. 2013, doi: 10.2174/1574892811308020002.
 - [321] Y. Wu, W. Chen, Z. P. Xu, and W. Gu, "PD-L1 distribution and perspective for cancer immunotherapy— blockade, knockdown, or inhibition," *Front. Immunol.*, vol. 10, no. AUG, p. 437710, Aug. 2019, doi: 10.3389/FIMMU.2019.02022/BIBTEX.
 - [322] A. Brnabic and L. M. Hess, "Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–19, Dec. 2021, doi: 10.1186/S12911-021-01403-2/FIGURES/2.
 - [323] Z. Ding, S. Zu, and J. Gu, "Evaluating the molecule-based prediction of clinical drug responses in cancer," *Bioinformatics*, vol. 32, no. 19, pp. 2891–2895, Oct. 2016, doi: 10.1093/BIOINFORMATICS/BTW344.
 - [324] X. Y. Liu and X. Y. Mei, "Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches," *Front. Bioeng. Biotechnol.*, vol. 11, 2023, doi: 10.3389/FBIOE.2023.1156372.
 - [325] M. Olivier, R. Asmis, G. A. Hawkins, T. D. Howard, and L. A. Cox, "The Need for Multi-Omics Biomarker Signatures in Precision Medicine," *Int. J. Mol. Sci.*, vol. 20, no. 19, Oct. 2019, doi: 10.3390/IJMS20194781.
 - [326] C. Wang, X. Lye, R. Kaalia, P. Kumar, and J. C. Rajapakse, "Deep learning and multi-omics approach to predict drug responses in cancer," *BMC Bioinformatics*, vol. 22, no.

- 10, pp. 1–15, May 2021, doi: 10.1186/S12859-022-04964-9/FIGURES/2.
- [327] M. Zaheed, N. Wilcken, M. L. Willson, D. L. O’connell, and A. Goodwin, “Sequencing of anthracyclines and taxanes in neoadjuvant and adjuvant therapy for early breast cancer,” *Cochrane database Syst. Rev.*, vol. 2, no. 2, Feb. 2019, doi: 10.1002/14651858.CD012873.PUB2.
- [328] C. Hatzis *et al.*, “GENOMIC PREDICTOR OF RESPONSE AND SURVIVAL FOLLOWING TAXANE-ANTHRACYCLINE CHEMOTHERAPY FOR INVASIVE BREAST CANCER,” *JAMA*, vol. 305, no. 18, p. 1873, May 2011, doi: 10.1001/JAMA.2011.593.
- [329] M. J. M. Magbanua *et al.*, “Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response,” *Breast Cancer Res.*, vol. 17, no. 1, May 2015, doi: 10.1186/S13058-015-0582-3.
- [330] J. K. Litton *et al.*, “Standardized Definitions for Efficacy End Points in Neoadjuvant Breast Cancer Clinical Trials: NeoSTEEP,” *J. Clin. Oncol.*, vol. 41, no. 27, p. 4433, Sep. 2023, doi: 10.1200/JCO.23.00435.
- [331] Y. L. Franco, T. R. Vaidya, and S. Ait-Oudhia, “Anticancer and cardio-protective effects of liposomal doxorubicin in the treatment of breast cancer,” *Breast cancer (Dove Med. Press.)*, vol. 10, pp. 131–141, 2018, doi: 10.2147/BCTT.S170239.
- [332] R. Paridaens *et al.*, “Paclitaxel versus doxorubicin as first-line single-agent chemotherapy for metastatic breast cancer: a European Organization for Research and Treatment of Cancer Randomized Study with cross-over,” *J. Clin. Oncol.*, vol. 18, no. 4, pp. 724–733, 2000, doi: 10.1200/JCO.2000.18.4.724.
- [333] K. Johnson-Arbor and R. Dubey, “Doxorubicin,” *xPharm Compr. Pharmacol. Ref.*, pp. 1–5, Aug. 2023, doi: 10.1016/B978-008055232-3.61650-2.
- [334] K. R. Hess *et al.*, “Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer,” *J. Clin. Oncol.*, vol. 24, no. 26, pp. 4236–4244, Sep. 2006, doi: 10.1200/JCO.2006.05.6861.
- [335] S. Demir *et al.*, “TWIST1 Gene expression as a biomarker for predicting primary doxorubicin resistance in breast cancer,” *Balk. J. Med. Genet.*, vol. 22, no. 2, pp. 25–30, Dec. 2019, doi: 10.2478/BJMG-2019-0025.
- [336] C. Gutierrez and R. Schiff, “HER 2: Biology, Detection, and Clinical Implications,” *Arch. Pathol. Lab. Med.*, vol. 135, no. 1, p. 55, Jan. 2011, doi: 10.1043/2010-0454-RAR.1.
- [337] N. Tuasha and B. Petros, “Heterogeneity of Tumors in Breast Cancer: Implications and Prospects for Prognosis and Therapeutics,” *Scientifica (Cairo.)*, vol. 2020, no. 1, p. 4736091, Jan. 2020, doi: 10.1155/2020/4736091.
- [338] J. Zhang and Y. Liu, “HER2 over-expression and response to different chemotherapy regimens in breast cancer,” *J. Zhejiang Univ. Sci. B*, vol. 9, no. 1, pp. 5–9, Jan. 2008, doi: 10.1631/JZUS.B073003/METRICS.

- [339] A. Gennari *et al.*, “HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials,” *J. Natl. Cancer Inst.*, vol. 100, no. 1, pp. 14–20, Jan. 2008, doi: 10.1093/JNCI/DJM252.
- [340] A. Rody *et al.*, “Gene expression profiling of breast cancer patients treated with docetaxel, doxorubicin, and cyclophosphamide within the GEPARTRIO trial: HER-2, but not topoisomerase II alpha and microtubule-associated protein tau, is highly predictive of tumor response,” *Breast*, vol. 16, no. 1, pp. 86–93, 2007, doi: 10.1016/J.BREAST.2006.06.008.
- [341] H. Masuda, D. Zhang, C. Bartholomeusz, H. Doihara, G. N. Hortobagyi, and N. T. Ueno, “Role of epidermal growth factor receptor in breast cancer,” *Breast Cancer Res. Treat.*, vol. 136, no. 2, pp. 331–345, Nov. 2012, doi: 10.1007/S10549-012-2289-9.
- [342] T. A. H. Järvinen *et al.*, “Amplification and deletion of topoisomerase IIalpha associate with ErbB-2 amplification and affect sensitivity to topoisomerase II inhibitor doxorubicin in breast cancer,” *Am. J. Pathol.*, vol. 156, no. 3, pp. 839–847, 2000, doi: 10.1016/S0002-9440(10)64952-8.
- [343] H. A. Burris *et al.*, “Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial,” *J. Clin. Oncol.*, vol. 15, no. 6, pp. 2403–2413, 1997, doi: 10.1200/JCO.1997.15.6.2403.
- [344] T. Conroy *et al.*, “FOLFIRINOX versus Gemcitabine for Metastatic Pancreatic Cancer,” *N. Engl. J. Med.*, vol. 364, no. 19, pp. 1817–1825, May 2011, doi: 10.1056/NEJMOA1011923/SUPPL_FILE/NEJMOA1011923_DISCLOSURES.PDF.
- [345] P. Yu, S. Luo, J. Cai, J. Li, and C. Peng, “ERAP2 as a potential biomarker for predicting gemcitabine response in patients with pancreatic cancer,” *Aging (Albany NY)*, vol. 14, no. 19, p. 7941, 2022, doi: 10.18632/AGING.204324.
- [346] R. Maréchal *et al.*, “Levels of Gemcitabine Transport and Metabolism Proteins Predict Survival Times of Patients Treated With Gemcitabine for Pancreatic Adenocarcinoma,” *Gastroenterology*, vol. 143, no. 3, pp. 664–674.e6, Sep. 2012, doi: 10.1053/J.GASTRO.2012.06.006.
- [347] S. Y. Lee *et al.*, “Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy,” *Eur. J. Cancer*, vol. 50, no. 4, pp. 698–705, Mar. 2014, doi: 10.1016/J.EJCA.2013.11.028.
- [348] M. Amrutkar and I. P. Gladhaug, “Pancreatic Cancer Chemoresistance to Gemcitabine,” *Cancers 2017, Vol. 9, Page 157*, vol. 9, no. 11, p. 157, Nov. 2017, doi: 10.3390/CANCERS9110157.
- [349] S. Nordh, D. Ansari, and R. Andersson, “hENT1 expression is predictive of gemcitabine outcome in pancreatic cancer: A systematic review,” *World J. Gastroenterol.*, vol. 20, no. 26, p. 8482, Jul. 2014, doi: 10.3748/WJG.V20.I26.8482.
- [350] W. Greenhalf *et al.*, “Pancreatic cancer hENT1 expression and survival from

- gemcitabine in patients from the ESPAC-3 trial," *J. Natl. Cancer Inst.*, vol. 106, no. 1, Jan. 2014, doi: 10.1093/JNCI/DJT347.
- [351] N. T. E. Bird *et al.*, "Immunohistochemical hENT1 expression as a prognostic biomarker in patients with resected pancreatic ductal adenocarcinoma undergoing adjuvant gemcitabine-based chemotherapy," *Br. J. Surg.*, vol. 104, no. 4, pp. 328–336, Mar. 2017, doi: 10.1002/BJS.10482.
- [352] N. Kawada *et al.*, "Human equilibrative nucleoside transporter 1 level does not predict prognosis in pancreatic cancer patients treated with neoadjuvant chemoradiation including gemcitabine," *J. Hepatobiliary. Pancreat. Sci.*, vol. 19, no. 6, pp. 717–722, 2012, doi: 10.1007/S00534-012-0514-X.
- [353] J. Raffenne *et al.*, "hENT1 Testing in Pancreatic Ductal Adenocarcinoma: Are We Ready? A Multimodal Evaluation of hENT1 Status," *Cancers* 2019, Vol. 11, Page 1808, vol. 11, no. 11, p. 1808, Nov. 2019, doi: 10.3390/CANCERS11111808.
- [354] H. Tiriach *et al.*, "Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer," *Cancer Discov.*, vol. 8, no. 9, pp. 1112–1129, Sep. 2018, doi: 10.1158/2159-8290.CD-18-0349.
- [355] J. Spratlin *et al.*, "The absence of human equilibrative nucleoside transporter 1 is associated with reduced survival in patients with gemcitabine-treated pancreas adenocarcinoma," *Clin. Cancer Res.*, vol. 10, no. 20, pp. 6956–6961, Oct. 2004, doi: 10.1158/1078-0432.CCR-04-0224.
- [356] J. J. Farrell *et al.*, "Human equilibrative nucleoside transporter 1 levels predict response to gemcitabine in patients with pancreatic cancer," *Gastroenterology*, vol. 136, no. 1, pp. 187–195, 2009, doi: 10.1053/J.GASTRO.2008.09.067.
- [357] "Nucleoside Transporter Profiles in Human Pancreatic Cancer Cells | Clinical Cancer Research | American Association for Cancer Research."
<https://aacrjournals.org/clincancerres/article/9/13/5000/202242/Nucleoside-Transporter-Profiles-in-Human> (accessed Apr. 02, 2025).
- [358] L. Tan *et al.*, "PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA," *Mol. Cancer*, vol. 18, no. 1, pp. 1–15, Jan. 2019, doi: 10.1186/S12943-019-0940-3/FIGURES/7.
- [359] S. E. Johnatty *et al.*, "Genome-wide analysis identifies novel loci associated with ovarian cancer outcomes: findings from the Ovarian Cancer Association Consortium," *Clin. Cancer Res.*, vol. 21, no. 23, p. 5264, Dec. 2015, doi: 10.1158/1078-0432.CCR-15-0632.
- [360] K. Nakajima *et al.*, "Neoadjuvant therapy alters the collagen architecture of pancreatic cancer tissue via Ephrin-A5," *Br. J. Cancer*, vol. 126, no. 4, pp. 628–639, Mar. 2022, doi: 10.1038/S41416-021-01639-9.
- [361] K. Weng *et al.*, "Collagen family genes and related genes might be associated with prognosis of patients with gastric cancer: an integrated bioinformatics analysis and

- experimental validation," *Transl. Cancer Res.*, vol. 9, no. 10, pp. 6246–6262, Oct. 2020, doi: 10.21037/TCR-20-1726.
- [362] J. Li *et al.*, "The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma," *PeerJ*, vol. 2019, no. 10, p. e7705, Oct. 2019, doi: 10.7717/PEERJ.7705/SUPP-14.
- [363] B. Terris *et al.*, "Characterization of Gene Expression Profiles in Intraductal Papillary-Mucinous Tumors of the Pancreas," *Am. J. Pathol.*, vol. 160, no. 5, p. 1745, 2002, doi: 10.1016/S0002-9440(10)61121-2.
- [364] F. Yang, S. S. Teves, C. J. Kemp, and S. Henikoff, "Doxorubicin, DNA torsion, and chromatin dynamics," *Biochim. Biophys. Acta - Rev. Cancer*, vol. 1845, no. 1, pp. 84–89, Jan. 2014, doi: 10.1016/J.BBCAN.2013.12.002.
- [365] X. Zhang *et al.*, "Luminal Breast Cancer Cell Lines Overexpressing ZNF703 Are Resistant to Tamoxifen through Activation of Akt/mTOR Signaling," *PLoS One*, vol. 8, no. 8, p. e72053, Aug. 2013, doi: 10.1371/JOURNAL.PONE.0072053.
- [366] M. M. Gottesman, O. Lavi, M. D. Hall, and J. P. Gillet, "Toward a Better Understanding of the Complexity of Cancer Drug Resistance," *Annu. Rev. Pharmacol. Toxicol.*, vol. 56, pp. 85–102, Jan. 2016, doi: 10.1146/ANNUREV-PHARMTOX-010715-103111.
- [367] V. Prasad, T. Fojo, and M. Brada, "Precision oncology: origins, optimism, and potential," *Lancet. Oncol.*, vol. 17, no. 2, pp. e81–e86, Feb. 2016, doi: 10.1016/S1470-2045(15)00620-8.
- [368] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, Apr. 2011, doi: 10.1126/SCIENCE.1200970.
- [369] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haike-Kains, and A. Goldenberg, "Machine learning approaches to drug response prediction: challenges and recent progress," *NPJ Precis. Oncol.*, vol. 4, no. 1, Dec. 2020, doi: 10.1038/S41698-020-0122-1.
- [370] F. Liu, "A Bayesian hierarchical model for high-dimensional meta-analysis," *Methods Mol. Biol.*, vol. 620, pp. 538–546, 2010, doi: 10.1007/978-1-60761-580-4_20.
- [371] R. Moriconi, M. P. Deisenroth, and K. S. Sesh Kumar, "High-dimensional Bayesian optimization using low-dimensional feature spaces," *Mach. Learn.*, vol. 109, no. 9–10, pp. 1925–1943, Sep. 2020, doi: 10.1007/S10994-020-05899-Z/FIGURES/5.
- [372] R. Aoki and M. Ester, "Causal Inference from Small High-dimensional Datasets," May 2022, Accessed: Mar. 30, 2025. [Online]. Available: <https://arxiv.org/abs/2205.09281v1>
- [373] L. Zhang and X. Lin, "Some considerations of classification for high dimension low-sample size data," *Stat. Methods Med. Res.*, vol. 22, no. 5, pp. 537–550, Oct. 2013, doi: 10.1177/0962280211428387.

- [374] A. Park, Y. Lee, and S. Nam, “A performance evaluation of drug response prediction models for individual drugs,” *Sci. Reports* 2023 131, vol. 13, no. 1, pp. 1–10, Jul. 2023, doi: 10.1038/s41598-023-39179-2.
- [375] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?”.
- [376] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, “Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities,” *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019, doi: 10.1016/J.INFFUS.2018.09.012.
- [377] M. Olivier, R. Asmis, G. A. Hawkins, T. D. Howard, and L. A. Cox, “The Need for Multi-Omics Biomarker Signatures in Precision Medicine,” *Int. J. Mol. Sci.*, vol. 20, no. 19, Oct. 2019, doi: 10.3390/IJMS20194781.
- [378] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, “Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer,” *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1248–1259, Mar. 2018, doi: 10.1158/1078-0432.CCR-17-0853.
- [379] X. Y. Liu and X. Y. Mei, “Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches,” *Front. Bioeng. Biotechnol.*, vol. 11, p. 1156372, 2023, doi: 10.3389/FBIOE.2023.1156372.
- [380] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, “MOLI: multi-omics late integration with deep neural networks for drug response prediction,” *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, Jul. 2019, doi: 10.1093/BIOINFORMATICS/BTZ318.
- [381] B. Güvenç Paltun, S. Kaski, and H. Mamitsuka, “Machine learning approaches for drug combination therapies,” *Brief. Bioinform.*, vol. 22, no. 6, p. bbab293, Nov. 2021, doi: 10.1093/BIB/BBAB293.
- [382] L. Wu *et al.*, “Machine learning methods, databases and tools for drug combination prediction,” *Brief. Bioinform.*, vol. 23, no. 1, Jan. 2022, doi: 10.1093/BIB/BBAB355.
- [383] C. Y. Hsieh *et al.*, “scDrug: From single-cell RNA-seq to drug response prediction,” *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 150–157, Jan. 2022, doi: 10.1016/J.CSBJ.2022.11.055.
- [384] C. Suphavitai *et al.*, “Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures,” *Genome Med.*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/S13073-021-01000-Y.
- [385] J. Chen *et al.*, “Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data,” *Nat. Commun.*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/S41467-022-34277-7.
- [386] L. Hetzel, S. Boehm, N. Kilbertus, S. Günemann, mohammad lotfollahi, and F. Theis, “Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 26711–26722, Dec. 2022.

- [387] S. Pellecchia, G. Viscido, M. Franchini, and G. Gambardella, "Predicting drug response from single-cell expression profiles of tumours," *BMC Med.*, vol. 21, no. 1, pp. 1–15, Dec. 2023, doi: 10.1186/S12916-023-03182-1/FIGURES/4.
- [388] F. Xia *et al.*, "A cross-study analysis of drug response prediction in cancer cell lines," *Brief. Bioinform.*, vol. 23, no. 1, Jan. 2022, doi: 10.1093/BIB/BBAB356.

Appendix I

A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling

Chayanit Piyawajanusorn, Linh C. Nguyen, Ghita Ghislat and Pedro J. Ballester 

Corresponding authors: Ghita Ghislat, U1104, CNRS UMR7280, Centre d'Immunologie de Marseille-Luminy, Inserm, Marseille, France. E-mail: gghislat1@gmail.com; Pedro J. Ballester, Cancer Research Center of Marseille, INSERM U1068, F-13009 Marseille, France. E-mail: pedro.ballester@inserm.fr

Abstract

A central goal of precision oncology is to administer an optimal drug treatment to each cancer patient. A common preclinical approach to tackle this problem has been to characterize the tumors of patients at the molecular and drug response levels, and employ the resulting datasets for predictive *in silico* modeling (mostly using machine learning). Understanding how and why the different variants of these datasets are generated is an important component of this process. This review focuses on providing such introduction aimed at scientists with little previous exposure to this research area.

Key words: pharmacogenomic modeling; precision oncology; phenotypic screening; molecular profiling; machine learning drug response

Introduction

Cancer is a group of diseases characterized by abnormal cell growth initiated in a tissue and with the potential of spreading to other organs. Patients with the same cancer type can respond very differently to the same drug treatment. Part of the reason for this interpatient variability is the heterogeneity of tumors across patients, which has been extensively documented [1]. This finding gave rise to cancer pharmacogenomics, a research field intended to predict how the response of cancer patients to a given drug treatment is affected by the genomic characteristics of their tumors. With the application of further omics technologies to profile tumors at the molecular level, this field is today more accurately referred to as cancer pharmaco-omics.

Although cancer pharmaco-omic data sharing has been scarce at the clinical level [2, 3], this has not been the case at

the preclinical level. Figure 1 illustrates how these preclinical datasets are commonly generated. A particularly promising research area is the application of machine learning (ML) to infer how drugs affect tumors from their molecular profiles, as this has distinct advantages over classical approaches [4, 5].

A key prerequisite for those investigating the application of ML to this problem is acquiring a minimum understanding of these preclinical datasets. This is not only important to integrate such heterogeneous datasets, but also to get a notion of data uncertainties and facilitate communication in collaborations between ML and domain experts, or simply inducing research trainees to this problem. However, we are not aware of any entry-level introduction to understanding these datasets. This review aims at filling this gap. Thus, concomitant topics such as the prior bioinformatics analysis of tumor profiles [6, 7] or posterior predictive modeling [5, 8] are out of the scope.

Chayanit Piyawajanusorn is a PhD student in bioinformatics at Aix-Marseille University (France). She graduated with a BSc in medical technology and an MSc in medical biochemistry and molecular biology from Khon Kaen University (Thailand).

Linh Nguyen is a Lecturer at the USTH (Vietnam). She graduated with a PhD in bioinformatics at Aix-Marseille University (France), an MSc in biotechnology-pharmacology at USTH (Vietnam) and BSc in Biology at HUS (Vietnam).

Ghita Ghislat is a Marie-Curie Fellow in immuno-oncology at the CIML (France) after a postdoctoral research in autophagy at Cambridge University (UK) and a PhD in cell biology at Prince Felipe Research Institute (Spain).

Pedro J. Ballester is a Group leader in machine learning for healthcare at the CRCM (France); postdoctoral research at EMBL-EBI, Cambridge University and Oxford University (UK); PhD in computational geophysics at Imperial College London (UK).

Submitted: 23 April 2021; Received (in revised form): 25 June 2021

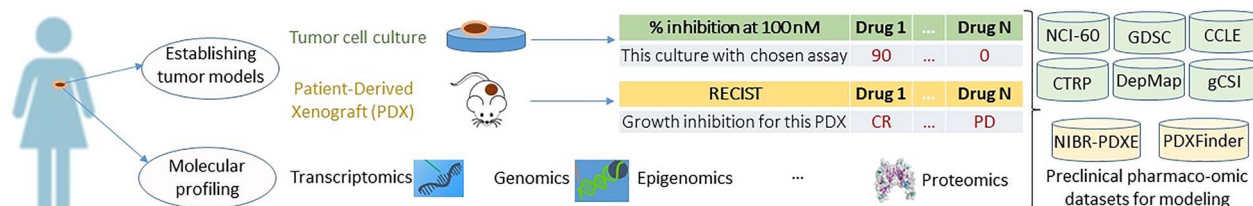


Figure 1. Sketch of the generation of preclinical pharmaco-omic data from a patient's tumor. Several databases sharing this type of datasets are publicly available. Due to space constraints, only the largest databases releasing data from tumor response to monotherapy are shown. Some databases report the molecular profile of the tumor model rather than that of the tumor itself (e.g. CCLE determined the somatic mutations of each cell line, not those in the tumor from which the cell line was established). This has the advantage of growing more tumor for profiling at the cost of representing the human tumor less accurately. It is noteworthy that not all types of molecular profiles are determined in each database.

Experimental models of cancer tumors

To mimic some aspects of human cancer tumors, tumor models are usually employed for experimentation. Here, we provide a brief essential introduction to commonly used preclinical cancer models.

Tumor samples

Tumor biopsies freshly extracted are the best option to observe the characteristics of patient tumors. When it is not viable, storage of these samples for further analyses such as molecular profiling is the norm. The most common way is to store the biobanked biopsies as a fresh-frozen (FF) or formalin-fixed paraffin-embedded (FFPE) tissue. Compared with FFPE, FF tissues are more suitable for keeping the genomic, proteomic and transcriptomic properties unaltered. Therefore, FF tissues are preferred for molecular profiling techniques such as mass spectrometry [9], quantitative polymerase chain reaction (PCR) [10], next-generation sequencing (NGS) [11] or western blot [12].

On the other hand, FFPE tissues are more practical than FF tissues because they can be stored at room temperature. However, FFPE processing is more time consuming, requiring many steps. In addition, some macromolecules can be denatured due to the acidic conditions of unbuffered formaldehyde solutions and formaldehyde cross-linking [13], a process required for immunostaining and histology studies for which FFPE is particularly suitable [14–16]. Therefore, FFPE tissues may not always be suitable for molecular analysis. Despite these difficulties, many studies have successfully generated the genomic [17–20], transcriptomic [21, 22] and epigenomic profiles [23–25] from nucleic acids isolated from FFPE tissues with comparable results with those extracted from FF tissues. Until recently, bulk analysis of tumor samples has been the norm, which measures average properties of the sample (e.g. the average expression of a gene across cells in the sample). Single-cell analysis of tumor samples, i.e. analyzing each cell in the sample individually at the considered omic level, is becoming increasingly more common, as a powerful approach to study tumor heterogeneity with applications to precision medicine and target therapy [26]. For analyses at single-cell resolution heeding intratumoral heterogeneity, protocols on frozen samples are still to be optimized when working on fresh samples is not possible [27]. Other types of samples can be single-cell analyzed, for instance, those containing circulating tumor cells [28].

In vitro tumor models

Cancer cell lines are established from primary cultures by transforming them into cells able to continuously grow *in vitro* via a

process called immortalization [29]. Alternatively, noncancerous cell lines, which acquire some cancer hallmarks [30], such as cells that underwent the epithelial to mesenchymal transition process, can also be used. One of their limitations is the genotypic and phenotypic variation during passages leading to differences between the cell line and the original tumor [31]. These issues are not present in primary cell cultures, which are derived directly from the patient tumor closely representing its origin (i.e. without any genetic manipulation to immortalize them). However, gene editing is more difficult to succeed in primary cell cultures than in immortalized ones. In addition, primary cell cultures have limited lifespan and undergo senescence, special mediums or growth factors are required and culture conditions need to be optimized [32].

Slight variations of the same tumor model have been systematically generated to investigate the impact of each variation on a phenotype. Here the variation is blocking the expression of the targeted gene and the phenotype is the response to the considered drug. Two approaches to such gene silencing have been applied so far: RNA interference (RNAi) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based loss-of-function manipulations in a large number of cell lines. RNAi uses pooled short-hairpin RNA to knockdown the intended RNA target in the cell line (off-target effects is the major challenge of this approach [33]). Instead of generating knockdowns at the messenger RNA (mRNA) level, CRISPR silences genes by generating knockouts at the DNA level. CRISPR-Cas9 gene editing system combined with single-guide RNAs (sgRNAs) are hence designed to introduce a frameshift mutation in the targeted gene. Compared with RNAi, CRISPR has higher screening sensitivity across cell lines, lower noise, minimal off-target effects and higher data reproducibility [34]. RNAi and CRISPR-based screening from the Achilles project with 74 661 sgRNAs targeting 18 119 genes in cancer cell lines are available on the Cancer Dependency Map (DepMap) portal [35–37].

Similar to cell lines, primary cell cultures lack tumor microenvironment and a three-dimensional (3D) structure [38]. In contrast, organoids are 3D *in vitro* cultures of primary cancer cells derived from tumor samples, which have also been used to assess treatment response [39–43]. Organoids can be generated via mechanical digestion of human tumor biopsies followed by *in vitro* growth, or via directed differentiation of embryonic stem cells or pluripotent stem cells to form bodies in 3D culture [44, 45]. Limitations include that organoids typically contain only epithelial layers without tumor microenvironment, the extracellular matrix Matrigel may disrupt drug penetration, as well as the costly enriched growth factors in culture medium may affect drug sensitivity [46, 47].

In vitro tumor models have the advantage of being amenable to high-throughput perturbation studies (currently, this means hundreds of drug molecules tested on over a thousand cell lines). We will later see that these large datasets permit building ML models able to predict the response of unseen cell lines and molecules (i.e. those data instances that were not employed for training these models).

In vivo tumor models

Animals bearing tumors, as cancer models, can better mimic the molecular diversity, cellular heterogeneity and histology of patient tumors, and thus capture to a better extent the tumor pathology, growth, metastasis and disease outcomes [48]. Several types of animals have been used as models of human tumors such as fly, worm, mouse, rat, rabbit, dog, cat and zebrafish [49, 50]. Mice have several advantages over other animal models such as size, inexpensive maintenance, short gestation period with numerous offspring and easy-to-manipulate gene expression [51, 52]. Patient-derived xenograft (PDX) models are the most common model for tumorigenicity testing, metastatic ability and drug testing [53–55]. PDXs are developed through patient primary tumor inoculations into immunocompromised mice. Xenografted tumors return the human histopathological and genetic characteristics and provide an important support for tumor expansion and drug treatment studies [48, 56]. Tumors can be injected in mice by several routes including subcutaneous (under the skin), intraperitoneal (into the abdomen cavity), intracranial (within the skull), intrasplenic (into the spleen), renal subcapsule (into the kidney) and site-specific organ (orthotopic) inoculations. Each inoculation site sets an environment that determines the way the tumor will grow and how much of the unaltered drug dose will reach tumor cells. The latter is also determined by the chosen route of drug administration. Consequently, drug response must be affected by these factors. For example, tumor response to misonidazole has been reported to change with inoculation site [57]. Modeling PDX data without taking into account these differences is likely to lead to analogous inconsistencies as those observed when modeling *in vitro* data coming from different types of assays [58], which will be later discussed in this review.

Limitations of PDX models are the lack of human stroma-tumor interactions, compromised immune system of the mice, slower tumor growth compared with *in vitro* cell lines and shorter cancer doubling time than the original growth in humans [59]. Thus, it does not fully represent the behavior of cancer in patients. In addition, not all cancer types can be easily engrafted in an animal model. Although the success rate of colorectal cancer models is over 75%, melanoma models are harder to establish (about 28% success rate). The successful engraftment rate varies depending on type of cancer cells, tumor microenvironment and the location of transplantation [54, 55, 60, 61]. Metastatic tumor, high stage and histological grade tend to be successfully transplanted in PDXs. Tumor size and treatment status should be taken into consideration to improve engraftment rate [62]. A further drawback of PDX models is their lower-throughput for perturbation studies than *in vitro* models due to ethical, technical and financial reasons (they are performed in animals, grow slower and are more expensive, respectively). However, for the same drug, PDX response is widely regarded as more relevant of patient response than in

vitro response, despite criticisms [63, 64] and *in vitro* success stories [65].

Multi-omics tumor profiles

A tumor sample giving rise to a cancer model can be molecularly profiled at different omics levels (e.g. genomic or transcriptomic). The suffix omic refers to large-scale profiling generating data that represents an entire set of some kind (e.g. transcriptomic refers to large-scale characterization of the types and numbers of RNA molecules in a sample). Multi-omics tumor profiles refer to the generation of multiple large-scale profiles per sample. Omics profiling has largely replaced techniques profiling at a smaller scale [e.g. fluorescent *in situ* hybridization (FISH), Sanger sequencing or blotting] to characterize tumors. Therefore, this review focuses on omics profiling technologies, which have been mostly based on microarrays and NGS (Figure 2). Microarrays require hybridization probes or antibodies that need to be previously determined; thus, they do not allow the identification of new variants of molecules (e.g. a mutated RNA molecule). In contrast, NGS is able to detect new variants, but may bring difficulties in interpreting the effects of those modifications on the phenotype. NGS-based profiling is more recent and of widespread use nowadays. NGS also requires more bioinformatics resources, faster data processing and larger data storage, which can be costly (Table 1). The primary analysis of these molecular profiles has been reviewed elsewhere [6, 7]. This section overviews molecular profiles typically used as tumor features in ML models to predict cancer drug response.

Genomic profiling

Genomic profiling provides the DNA sequence of a tumor sample at single-base-pair resolution. Somatic mutations can be identified by comparing this DNA sequence with that in a reference genome. These mutations include single-nucleotide polymorphisms (SNPs) in sequences of gene and regulatory regions, changes in the number of copies for each gene (copy number variation) or chromosomal rearrangements such as translocation or inversion.

Hybridization-based techniques for genomic profiling include FISH, comparative genomic hybridization (CGH) and DNA microarray [66, 67]. FISH uses fluorescently labeled probes to detect and locate the position of a specific DNA sequence in the genome. Similarly, CGH is used to detect unbalanced copy number changes at the whole-genome level, allowing the identification of new genes that are up- or down-regulated in cancer. Nevertheless, the resolution might be increased using CGH compared with FISH, which is limited to the resolution of the used microscopy [68–71]. In contrast, the millions of probes in a microarray detect thousands of genes and transcripts in the sample. Whole-genome expression arrays (e.g. hg-u133 plus 2.0 microarrays) are useful for the analysis of gene expression profile and copy number variations. In addition, SNP arrays (e.g. Affymetrix SNP 6.0) are used to identify the copy number of variants or the concordance between cell lines [72–74].

Following the emergence of NGS technologies, whole-genome sequencing (WGS) and whole-exome sequencing (WES) have become predominant. WGS sequences the entire human genome, both coding and noncoding regions. Then, comparing the obtained sequences to external DNA references or comparing cancer to healthy cells within the individual can reveal putative somatic mutations and gene fusions [75–77]. WGS has been used to collect large amounts of genomic

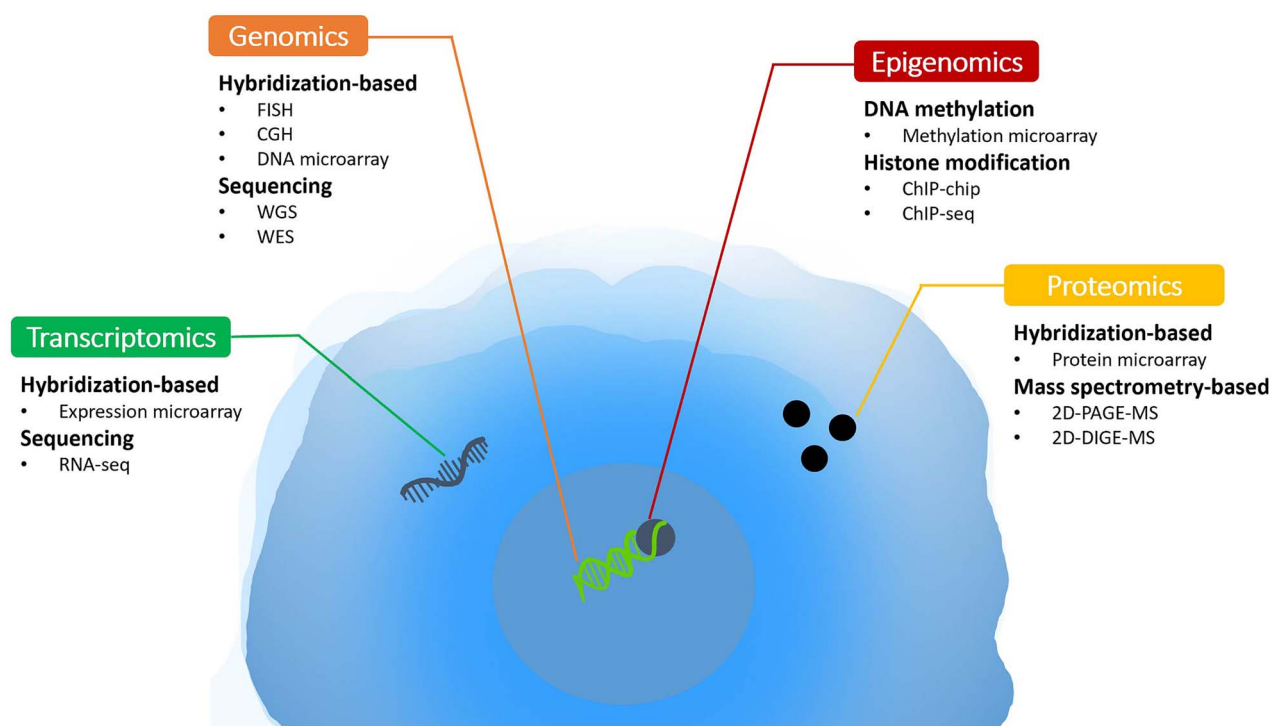


Figure 2. Sketch of most common technologies for molecular profiling of tumor cells for pharmaco-omic modeling purposes.

information across cancer types, which improved knowledge in cancer research overtime including breast cancer [78, 79], neuroblastoma [80], prostate cancer [75] and pediatric medulloblastoma [81]. On the other hand, WES only sequences the coding regions of the sample, which has been useful in detecting both germline and somatic variants, especially identifying the somatic driver mutations within the exomes of various cancers. The cost of WES is cheaper than WGS as WES sequences only about 1% of the genome, but WGS has been increasingly used owing to the continuous decrease in price difference. Several studies applied WES to characterize the genomic aberrations of both cancer cell lines [82] and PDX models [83]. Although most functional regions associated with the disease have so far been found in the coding sequence of genes, identifying noncoding regions of the DNA have the potential to expand our knowledge of genotype–phenotype relationships, which is a very attractive prospect.

Transcriptomic profiling

Transcriptomics is the large-scale study of RNA molecules using high-throughput techniques [84, 85]. Although microarrays detect genes of interest by employing probes previously designed to target them, RNA sequencing (RNA-seq) is a more recent technology that sequences all transcripts from a biological sample, which allows analyzing the expression level of protein-coding and nonprotein-coding genes across the entire genome [86]. This broader coverage of gene expression also adds to the cost. With sufficient sequencing depth, this technique is used to detect and quantify fusion transcripts or alternative splicing, which can be related to various phenotypes including treatment response [87–89]. Furthermore, the combination of genomic and transcriptomic profiling can help drug response studies. For instance, a clinical trial employed WGS, WES

and RNA-seq on gastric cancer patient samples to show that C10orf71 mutations and MYC amplification were correlated with neoadjuvant chemotherapy resistance and sensitivity, respectively [90].

To focus on coding regions, WES requires exome-enrichment steps before the genomic DNA can be sequenced, thereby adding substantially to WES cost. To further optimize the cost effectiveness of variant detection within coding regions, RNA-seq has been considered as a potential substitute for WES [91]. Because the transcriptome from a tissue or group of cells represents a subset of the exome, only the variants in the expressed genes or transcripts can be detected by RNA-seq. Unlike WES, where the major application is variant detection, RNA-seq is useful in determining the global expression profile of a specific cell or tissue, especially under two different conditions (e.g. drug sensitive versus drug resistant). The resulting differentially expressed genes across the two conditions can then be used for a range of purposes. Thus, the choice of which technology to adopt largely depends on the research question [92].

Epigenomic profiling

Epigenetic alterations modify the chemical structure of DNA, leading to transcriptional activity modulation. These alterations include the methylation of CpG islands (CGIs) in promoter regions, histone modifications, chromatin remodeling and noncoding RNA molecules (e.g. microRNA), which can be potential markers of cancer development and progression [93]. Several technologies are used to detect epigenetic alterations including methylation microarray methods as well as chromatin immunoprecipitation-based microarray and sequencing methods [94].

At the DNA level, DNA hypermethylation is arguably the most studied epigenetic alteration in cancer [95]. Methylated

Table 1. Summary of multi-omics tumor profiling techniques and their application in pharmaco-omics studies

Technologies	Scale	Advantages	Limitations	Examples of application	Ref.
FISH	Whole genome	Able to detect and locate the position of specific DNA sequence in the genome, examining chromosome abnormalities and gene mutation	Limited to the resolution of the used microscopy	The impact of EML4-ALK gene rearrangement on the response to tyrosine kinase inhibitors	[169]
CGH	Whole genome	Able to detect unbalanced copy number changes at the whole-genome level	The detection limits decrease with a smaller aberration size (in bases), smaller aberration amplitude (copy number)	Analyze targeted genomic regions associated with chronic lymphocytic leukemia to facilitate the determinations of the relationship between therapeutic options and genomic aberrations	[170]
DNA microarray	Whole genome	Analysis of gene expression profile and able to detect mutations, CNVs or polymorphisms in DNA sequences	Not able to identify new variants due to hybridization probes need to be previously determined, high background level due to cross-hybridization	Monitor the expression profiles of human breast cancer MCF-7 cells that are either transiently treated with doxorubicin or doxorubicin resistance	[171]
WGS	Whole genome	Able to identify SNPs, INDELs, chromosomal aberration and CNVs in the coding and noncoding region of the genome	Cost, large volume of data and complex to interpret the data, equipment and labor intensive	Identify gene variants associated with clopidogrel resistance and positive response to lipid-lowering therapy. Identify rare variants that may be very important in the genomic contribution to treatment response and toxicity	[172, 173]
WES	Whole exome	Able to identify SNPs, INDELs, chromosomal aberration and CNVs in the coding region of the genome, quicker sequencing and data analysis compared with WGS, cost effective	Unable to detect noncoding region, equipment and labor intensive	CCLE have provided genetic variants to improve understanding of the molecular features that contribute to cancer phenotypes, including drug responses	[174]
Expression microarray	Transcribed region of genome	High sample throughput for analysis of known genes and transcripts relies on hybridization	Inability to detect novel transcripts, required previously designed probes to target transcript of interest, high variance for low-expressed genes, cross-hybridization	Differential expression of genes that were associated with drug response and treatment outcome in B-lineage ALL	[85]
RNA-seq	Transcribed region of genome	Quantify the amount of transcripts and detect novel transcripts, protein-coding and nonprotein-coding genes with low level of expression, alternative splicing detected if aligned to genome	Sensitivity limited to transcribed genome, high cost compared with expression microarray, and required high-power computer facilities	Characterized molecular signatures for therapy guidance that can arise from gene expression investigation with RNA-seq	[175]
Methylation microarray	DNA methylations	Able to measure 5-methylcytosine at individual CpG locus in the genome, cost and time effectiveness	Required previously designed probes, cross reactive probes	Predicting drug sensitivity of cancer cells based on DNA methylation levels	[176]

Continued

Table 1. Continued

Technologies	Scale	Advantages	Limitations	Examples of application	Ref.
ChIP-chip	Histone modifica-tions	Map the entire spectrum of the DNA sequences occupied by specific DNA-binding protein across the whole genome on microarray, quantify the specific DNA sequences where proteins bind and detect where epigenetic modification is	The quantification depends on the hybridization efficiency of the ChIP DNA molecules to the probes on the array, coverage limited by probe on array, multiple arrays may be needed for large genomes, and cross-hybridization	Identify histone modification markers in gastric cancer and determine whether these markers were able to discriminate between normal can gastric cancer cells	[177]
ChIP-seq	Histone modifica-tions	Map the entire spectrum of the DNA sequences occupied by specific DNA-binding protein across the whole genome by sequencing, quantify the specific DNA sequences where proteins bind and detect where epigenetic modification is, ChIP-seq does not require preselection of genomic regions	Limited by align ability of reads to the genome, complex data analysis	Prediction drug response-related epigenetic phenotypes in chronic lymphoblastic leukemia	[178]
Protein microarray	Proteome	Measure the expression of protein on array, detection of phosphoproteins possible	Only possible to detect known target, limited by availability of antibody, cross reactivity	Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays	[179]
Mass spectrometry-based techniques	Proteome	Able to quantify protein expression difference, using small amount of sample (1 µg), detection of novel protein possible	Cost, require combination mass spectrometry with other techniques, such as chromatographic techniques, equipment and labor intensive	Measuring quantity of drug target in cancer to improves drug sensitivity prediction	[180]

Notes: FISH, fluorescent in situ hybridization; CGH, comparative genomic hybridization; CNVs, copy number variants; SNPs, single-nucleotide polymorphisms; WGS, whole-genome sequencing; INDELs, insertions and deletions; WES, whole-exome sequencing; ALL, acute lymphocytic leukemia; RNA-seq, RNA sequencing; ChIP-ChIP, chromatin immunoprecipitation with DNA microarray; ChIP-seq, chromatin immunoprecipitation with sequencing.

cytosines at CpG sites can change gene expression patterns affecting embryonic development, cellular proliferation, differentiation and chromosome stability [96]. DNA methylation profiling can be detected by converting unmethylated cytosines to uracil using bisulfite, whereas 5-methylcytosines remain unchanged. This distinguishes between methylated and unmethylated cytosines [97]. Methylation microarrays are designed for the hybridization of bisulfite-converted DNAs with high numbers of predesigned probes providing quantitative methylation measurement of a large number of CGIs [98]. Illumina’s Infinium Human Methylation 450 (HM450K) BeadChip targeting over 450 000 methylation sites that cover 96% of the CGIs has been used to profile cancer methylomes over 8000 samples of 33 cancer types in The Cancer Genomic Atlas through the US National Cancer Institute (NCI) Genomic Data Commons public resource [99].

At the chromatin level, histone modification patterns have been used for the prognosis prediction of various types of cancer [100]. Histone modification profiles in cancer cells can be determined by chromatin immunoprecipitation (ChIP) that uses specific antibodies to identify proteins bound to a DNA region of interest. Two of the most common approaches for ChIPed DNA

analysis are based sequencing (ChIP-seq) and microarray (ChIP-chip) [101]. ChIP-seq has become the gold standard to study histone modifications. ChIP-seq produces profiles with higher resolution and genomic coverage, allowing higher sensitivity and specificity than ChIP-chip, whose resolution is limited by probes on the array. Furthermore, ChIP-seq improves signal-to-noise ratios and is less expensive [102, 103]. The ChIP technologies were applied to analyzed histone modification of H3K4me3 (considered as a transcriptional activator) and H3K27me3 (a transcriptional silencer) in breast cancer [104, 105], glioblastoma [106], prostate cancer [107] and ovarian cancer [108]. The use of ChIP technologies was also fruitful in pharmaco-omics modeling. For instance, a recent study applied multi-omic ML approach on omics-data including ChIP-seq to reveal the mode of action of drugs in Huntingtin disease models [109]. Another study combined ChIP-seq-based profiling with high-throughput drug screening to identify resistance mechanism to ibrutinib in mantle cell lymphoma [110].

Proteomic profiling

Genomic and transcriptomic profiling provide abundant information about genes and transcripts in cancer cells. However,

they cannot fully represent the physiological state of the cells, because they do not capture the abundance of each protein directly participating in cellular processes. Indeed, several studies have reported a discordance between the transcriptomic and proteomic profiles of a cell, due to the posttranscriptional and posttranslational modifications, protein half-lives and other regulatory factors [111]. Therefore, proteomic profiling permits the downstream analysis of alterations at genomic, transcriptomic and/or epigenetic levels. Conversely, there are many functional RNA molecules beyond mRNAs, which are therefore not captured by proteomic profiles and evidence their complementarity with upstream omics profiles [112].

Mass spectrometry-based sequencing is an exploratory tool to study the entire proteome and to relate it to cancer progression as well as drug response [113, 114]. The proteomic analysis generally consists of two steps: separation of proteins followed by their identification. During this approach, the protein mixture from cancer cell lysate is separated by protein feature using polyacrylamide gel electrophoresis (PAGE) and isoelectric focusing (IEF), which is the basis for two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) according to mass and isoelectric point. A modified version of 2D-PAGE, the 2D difference gel electrophoresis, uses different fluorescent dyes labeling 2–3 samples on the same 2D gel. This minimizes spot pattern variability and the number of gels in an experiment, providing higher sensitivity, reproducibility and accurate quantification of protein expression differences. Then, each protein is identified by mass spectrometry-based technologies [114].

Another approach of proteomic profiling is the use of protein microarrays, which can measure the expression of thousands of proteins to detect protein biomarkers that related to cancer progression. Nishizuka *et al.* [115] analyzed the proteomic profiles of 60 cancer cell lines of the NCI panel using the reverse-phase protein array (RPPA). Ma *et al.* [116] integrated these datasets with drug sensitivity to predict drug response on cell lines.

Bulk versus single-cell profiling

So far, this section has only addressed molecular profiles carried out on the bulk of the tumor (i.e. all cells in the sample together). This is because we are not aware of any cancer pharmacomics database including single-cell molecular profiles. Despite their higher cost per tumor, large-scale single-cell profiles are likely to become publicly available in the future, given their distinct advantages for pharmaco-omic modeling. For example, exploiting single-cell RNA-seq data with ML has enabled the systematic identification of patient-customized drug combinations [117]. This has the additional advantage of revealing whether the combination also inhibits the patient's own nonmalignant cells as a proxy for treatment toxicity.

Drug-sensitivity profiling of tumors

High-throughput phenotypic screens evaluating drug sensitivity for drug design and precision oncology purposes have boomed in recent years. Drug-sensitivity profiles are typically performed on cancer cell lines. Each cell line is tested against a panel of molecules to assess their sensitivity or resistance profile *in vitro*. Furthermore, medium-throughput screens with *in vivo* assessment of drugs in appropriate animal models are starting to be released. Several methods are available to measure drug sensitivity, such as *in vitro* cellular enzymes-based assays, nucleic and protein assays, adenosine triphosphate (ATP) assays, the Profiling Relative

Inhibition Simultaneously in Mixtures (PRISM) assays, and *in vivo* tumor growth and survival assays in PDX models (Figure 3). Here, we overview *in vitro* and *in vivo* drug testing methods used to establish drug-sensitivity profile of tumors. The advantage and disadvantages of each assay are discussed and planning strategies of drug testing are also included.

In vitro and ex vivo drug testing

Drug sensitivity testing has been used to investigate biological effects related to drug response and toxicity. Cultured cells are treated with a molecule, or combination of molecules followed, by determining how the treatment has changed the number of viable cells after typically 3 days. To generate reproducible and useful drug response data and minimize experimental variability, biological and technical factors should be optimized. Indeed, biological factors, such as cell type, cell density, incubation time, culture media (e.g. media type and contents), can affect cell proliferation rate. The culture media should be constant throughout the study and the standard culture condition is also provided by American Type Culture Condition or European Collection of Cell Cultures. Cell density should be optimized to avoid slow proliferation due to contact inhibition and nutrient limitation. In addition, cell growth rate varies in different cell types. This issue can be improved by determining time-dependent growth rate inhibition [118]. There are also technical factors, such as the concentration, incubation time and sensitivity assay, employed for measuring the activity of the molecule on the cell culture phenotype. For example, to measure drug sensitivity accurately, the drug concentration range has to be chosen to overlap with the dose-response curve of the drug-cell line pair, at least for the most sensitive cell lines to the drug (this can be hard as the drug concentration at which a response is observed is not known before hand). In cell lines where the dose-response curve is outside the range of tested drug concentrations, the half-maximal inhibitory concentration of the drug (IC_{50}) will be extrapolated instead of interpolated. Hence, such IC_{50} value will not be a good summary of the true dose-response curve. IC_{50} values typically vary from pM to μ M range depending on the drug, cell line and experimental conditions. Testing with a 10-fold drug concentration range is desirable [119, 120].

In addition to IC_{50} , there are others metrics available to represent the measured response of cell lines to a given molecule, including the half-maximal effective concentration (EC_{50}) [121], maximal effect level (A_{max}) [121], the area under the fitted dose-response curve [122] or drug sensitivity scoring [117]. The role of such metrics in improving replicability and reproducibility of cancer drug sensitivity screens has been discussed elsewhere [119]. Any of these metrics can be used as the dependent variable for pharmaco-omic regression models. Alternatively, to build binary classification models, each cell line can be assigned either a resistant or sensitive label by applying a meaningful threshold on the metric values.

Cell cultures are commonly used in these screens, mostly in 2D, where cells grow as a monolayer in a culture flask or in flat petri dish attached to a plastic surface [123]. However, the 2D cell culture has several limitations overcome by 3D culture systems such as organoids, as explained in experimental models section. 2D or 3D testing can be carried out on *in vitro* immortalized cancer cell lines or *ex vivo* cell cultures generated from primary tumor cells of the patients (i.e. without any genetic modification). Drug testing in cancer cell lines can be classified into

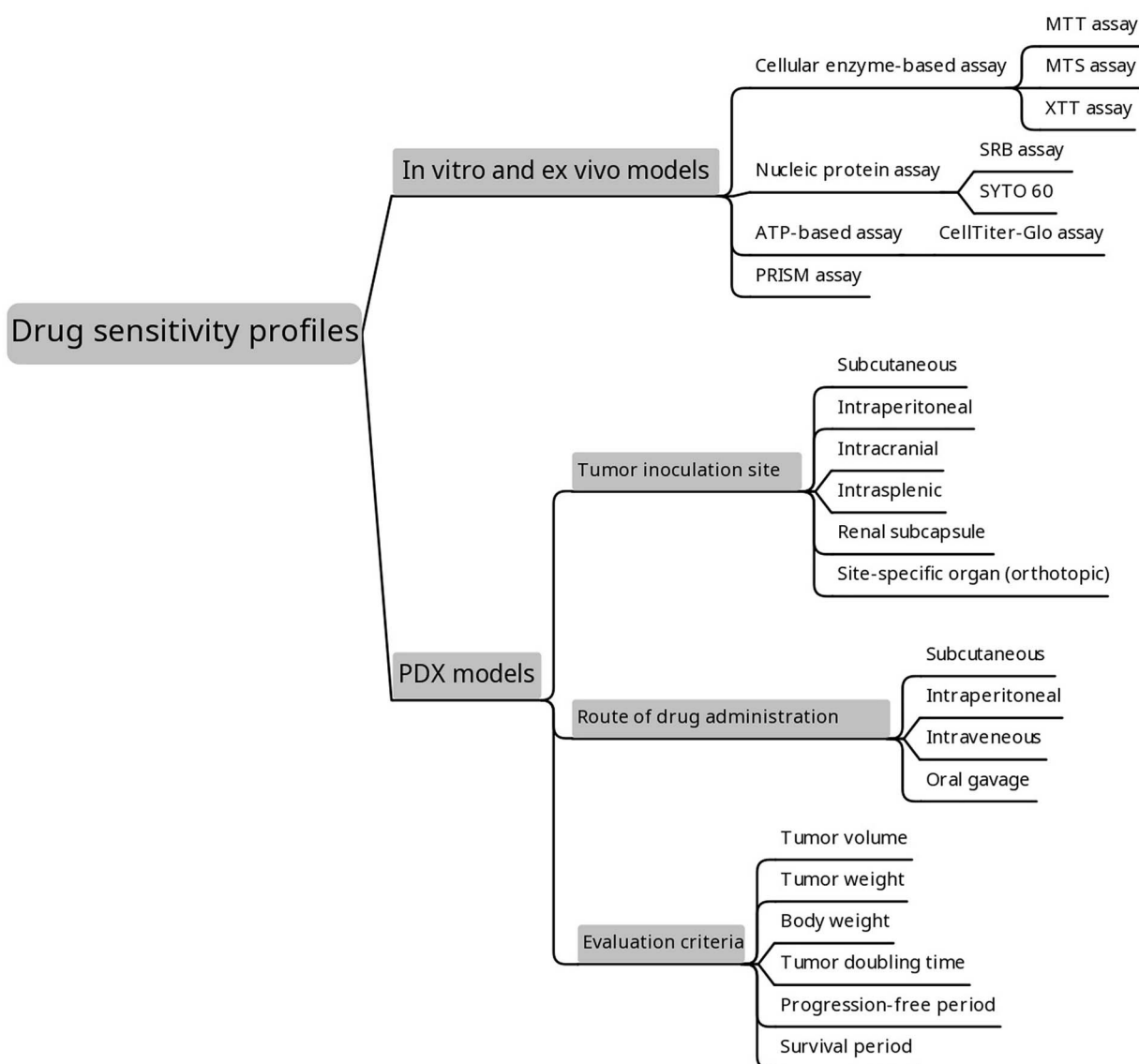


Figure 3. Commonly used cancer models for drug sensitivity profiling. The assays used for ex vivo models are the same as those used for the in vitro models. PDXs are by far the most common type in vivo model in cancer research, although there are exceptions [163].

four groups: cellular enzymes-based assays, nucleic and protein assays, ATP assay and the PRISM assay. These are concisely introduced in the next subsections.

Cellular enzyme-based assay

In most colorimetric cell viability assays, substrates are converted to a colored product by enzymes present in viable cells only, in a proportional way to the number of viable cells. Colorimetric or fluorometric cell viability assays are used in cancer drug sensitivity. The 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl-2H-tetrazolium bromide (MTT) assays are known of being fast, reproducible, simple and homogeneous [124]. This method can underestimate the antiproliferating effect of certain chemical compounds, and its requirements of multiple washing steps may cause loss of less-adherent cells [125]. Other factors, like metabolic rate and number of mitochondria,

may also affect the results [125]. MTT is converted to purple-colored formazan by oxidoreductase or dehydrogenases in viable cells, which can then be then detected by spectrophotometer [126]. Other tetrazolium compounds, such as MTS (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium)) and XTT (2,3-Bis-(2-Methoxy-4-Nitro-5-Sulfophenyl)-2H-Tetrazolium-5-Carboxanilide) can also produce formazan in viable cells [127].

Nucleic acid and protein assay

Another common colorimetric assay is the Sulforhodamine B (SRB) assay. The SRB is a pink aminoxanthine dye with two sulfonic groups that bind to basic amino acid residues in proteins of viable cells under mild acidic conditions [128]. This assay has also been proved to be suitable for large-scale screening and highly cost effective [129]. The intracellular SRB level is

Table 2. Single-drug response data on *in vitro* cancer cell lines (CLs) and *in vivo* patient-derived xenografts (PDXs)

Datasets (assay)	#drugs	#models	Assay	#types	Early modeling studies/reviews
NCI-60 [141]	51 060 molecules with GI ₅₀ s (~3000 drugs)	60 CLs	SRB assay	9	Shoemaker et al. [141], Ma et al. [116]
GDSC [181]	265 molecules with AUC and IC ₅₀	1001 CLs	Syto 60, Resazurin	29	Menden et al. [182], Ammad-ud-din et al. [183]
CCLE [121]	24 molecules with AUC and IC ₅₀	1061 CLs	CellTiter-Glo	22	Barretina et al. [121]
CTRP v2 [184]	481 molecules with AUC and IC ₅₀	823 CLs	CellTiter-Glo	25	Rampásek et al. [185]
gCSI [144]	16 molecules with AUC and IC ₅₀	754 CLs	CellTiter-Glo	23	Kurilov et al. [186]
DepMap [142]	4518 (1448 at eight concentrations)	578 CLs	PRISM	24	Gerdes et al. [187]
NIBR-PDXE [83]	62	277 PDXs	% tumor volume change	6	Nguyen et al. [188], Kurilov et al. [186]
PDXFinder [145]	36	378 PDXs	% tumor volume change	9	None so far

Notes: NCI-60 refers to IC₅₀s as GI₅₀s. Also note that a posterior version of GDSC released at <https://www.cancerrxgene.org/>, called GDSC2, employs the CellTiter-Glo assay.

proportional to the number of viable cells. The SRB assay has been used to test drug sensitivity in cancer cell lines [116, 130]. Another compound, SYTO 60, is a red fluorescent dye that binds to nucleic acids of viable cells. Thus, fluorescence intensity is proportional to the number of viable cells [131].

ATP-based assay

An important metabolic factor for cell viability assays is the level of ATP production. Cancer cells produce low ATP after treatment, leading to cell death. The CellTiter-Glo assay is a luminescence cell viability assay that measures ATP with the luciferase enzyme, which is able to oxidize luciferin in the presence of ATP and Mg²⁺ to release light in a proportional way to the number of viable cells [132]. This class of assays is known for their short measurement time, thus providing rapid results, and is highly sensitive. However, it is more expensive than other assays such as MTT [133]. Besides, this type of assays often required fixation and is not sensitive for nonadherent cells [124]. Other limitations include ATP measurement not being directly linked to cell viability and possibly leading to false positives when the drugs affects cell metabolism. Heiser et al. [134] investigated the viability of different breast cancer cell lines under the effect of 77 anticancer compounds using CellTiter-Glo assay. CellTiter-Glo assay can also be used as drug testing assay in human cancer organoid cultures to screen drug sensitivity [135]. Sachs et al. [136] generated organoids derived from resection of 155 breast cancer patients. Next, drug sensitivity was performed by using CellTiter-Glo 3D assay. The treatment responses were compared in the breast cancer organoids and the corresponding patients. These authors found a high correlation between sensitivity to tamoxifen in the breast cancer organoids and in the original patients with metastatic breast cancer.

The PRISM assay

PRISM is a multiplex technology to measure the viability of hundreds of cancer cell lines concurrently. Each cell line is labeled with 24 base pair unique DNA barcode, allowing cell lines to be pooled together according to doubling time. The pooled cells are then treated with each drug separately. Following drug treatment, cell lines are lysed and mRNAs are isolated and amplified

by PCR and detected using Luminex scanner. The quantity of each barcode remaining after treatment measures the cell line sensitivity to each compound [137]. PRISM is able to scale up drug sensitivity measurements on cell lines at the cost of requiring a complex protocol and being labor intensive.

In vivo drug testing

PDXs are arguably the most commonly used *in vivo* cancer model to assess drug sensitivity. The drug effect is evaluated by measuring the change in tumor volume, tumor weight and body weight following drug treatment [138]. In particular, the tumor can be measured by its length (the largest diameter: *L*) and its width (the smallest diameter: *l*), then the estimated volume (*V*) will be calculated according to standard geometrical formulas such as $V = (L \cdot l^2)/2$. Cancer progression under treatment also can be accessed by measuring the time for the tumor to double its size, called the tumor's doubling time [83]. Moreover, progression-free survival and overall survival are also considered when assessing the sensitivity of PDXs against anticancer drugs, especially for the types of cancer where the tumor volume is difficult to measure, e.g. blood cancers [83, 139]. After a defined period of treatment or until the tumor volume reaches a certain level, mice are sacrificed and tumors are collected for subsequent analyses. An example of the application of PDXs to drug response study is by Garrido-Laguna et al. [140], who employed 69 PDX pancreatic cancer models to predict the survival and response of patients to gemcitabine from their corresponding PDXs.

Primary resources of cancer pharmaco-omic data

The generation of cancer pharmaco-omic datasets have played a vital role in broadening our understanding of cancer mechanisms and cellular process. The NCI-60 project, with 60 human cancer cell lines and from the US NCI, is that with the largest number of tested molecules [141]. More than 130 000 molecules have been screened on this panel, over 50 000 of these molecules being publicly available, each tested against 60 cancer cell lines [141]. Recently, several projects have employed panels with a much larger number of cell lines at the cost of screening a much smaller number of molecules (Table 2). For example, the Cancer

Table 3. Two-drug synergy data on *in vitro* cancer cell lines (CLs) and *in vivo* patient-derived xenografts (PDXs)

Datasets (assay)	#drugs	#unique drug combinations	#models	Assay	#types	Early modeling studies/reviews
AstraZeneca–Sanger data [162]	119	910	85 CLs	Sytox green	8	Li et al. [189], Menden et al. [162]
NCI-ALMANAC [163]	104	5232	60 CLs	CellTiter-Glo	9	Xia et al. [190], Sidorov et al. [191]
NIBR-PDXE [83]	62	26	277 PDXs	% tumor volume change	6	Nguyen et al. [188], Kurilov et al. [186]
PDXFinder [145]	36	9	25 PDXs	% tumor volume change	5	None so far

Table 4. Available omics data for each resource

Datasets	RNA expression	DNA copy number	DNA mutations	DNA methylation	Protein abundances
NCI-60 [141]	cDNA microarray, Affymetrix Hu6800, Affymetrix HGU133	Affymetrix 125 K SNP array, OncoBAC DNA microarray	WES	Infinium Human-Methylation450k BeadChip array	Reverse-phase protein lysate arrays
GDSC [181]	Affymetrix HG-U219A	Affymetrix SNP 6.0	Illumina Hiseq 2000	Illumina array 450 K	–
CCLE [121]	Affymetrix HG-U133 Plus 2.0, Illumina RNA-seq	Affymetrix SNP 6.0	OncoMap, Hybrid Capture Sequencing	Reduce representation bisulfite sequencing	Reverse-phase protein lysate microarray
CTRP v2 [184]	Affymetrix HG-U133 Plus 2.0	Affymetrix SNP 6.0	OncoMap, Hybrid Capture Sequencing	–	–
gCSI [144]	Illumina RNA-seq	Illumina 2.5 M SNP	Sanger	–	–
DepMap [142]	CCLE (Affymetrix HG-U133 Plus 2.0)	CCLE (Affymetrix SNP 6.0)	CCLE (OncoMap, Hybrid Capture Sequencing)	–	Reverse-phase protein lysate microarray
NIBR-PDXE [83]	Affymetrix HG-U133PLUS 2.0, Illumina RNA-seq	Affymetrix SNP 6.0	Illumina Hiseq 2000/2500	–	–
PDXFinder [145]	/ (platform depends on providers)	/ (platform depends on providers)	/ (platform depends on providers)	–	–
AstraZeneca–Sanger data [162]	Affymetrix Human Genome U219	Affymetrix SNP 6.0	Illumina Hiseq 2000	Illumina array 450 K	–
NCI-ALMANAC [163]	cDNA microarray, Affymetrix Hu6800, Affymetrix HGU133	Affymetrix 125 K SNP array, OncoBAC DNA microarray	WES	Infinium Human-Methylation450k BeadChip array	Reverse-phase protein lysate arrays

Cell Line Encyclopedia (CCLE) project [121], where the sensitivity of 479 cancer cell lines against 24 drug molecules was tested. The exception is DepMap [142], a more recent screen midway between these opposite numbers of cell lines and tested drugs. Indeed, 4518 compounds against 578 cell lines are available in the DepMap portal [35] and were originally screened to reveal the inhibitory effect of nononcology drugs in cancer cell lines [142]. These resources have enabled research on many more cancer types, while providing higher statistical power to identify single-gene markers of drug response for a range of cancer types [143].

However, special care must be taken to avoid inconsistencies when integrating drug sensitivity data from different resources. These inconsistencies may originate from different drug testing assay, drug concentration and seeding density across resources [144]. Another recent trend has been to characterize each cell line with a range of molecular profiles. These *in vitro* pharmacomic datasets result from integrating drug response and molecular profiles across cell lines. Furthermore, large collections of PDXs

have also been established to characterize patient tumor heterogeneities *in vivo* [83, 145]. For example, Gao et al. [83] generated PDXs for the Novartis Institute for Biomedical Research PDX encyclopedia and tested them with numerous drugs. Subsequently, Georgopoulou et al. developed the extensive collection of breast cancer PDXs, in which molecular characteristics of original tumor are preserved [146]. These *in vivo* pharmacomic datasets cover a much smaller number of drugs and tumor models than their *in vitro* counterparts due to incurring in higher financial and time costs.

Either way, there are still many cancer types for which no pharmacomic datasets are available. Integrating data from a range of cancer types for transfer learning constitutes an effective way to be able to predict drug response in data-scarce cancer types as long as all tumor models are profiled in a common way [147, 148]. Integrating various multi-omics profiles has been proposed as another way to gain predictive accuracy in general [149]. However, a model integrating multiple profiles would increase time and cost expenses for each patient in a clinical

setting. Furthermore, a multi-omic integration has been typically reported to result in at best incremental benefits [150]. These reviews provide further information about pharmaco-omic data integration [151, 152].

Table 2 compiles major primary data resources that can be used to build predictive models of response of individual drugs. The number of drugs, cell lines and cancer types increase with new releases (the ones provided are indicative and correct at the time of writing). Table 3 compiles major primary data resources that can be used to build predictive models of response or synergy of drug pairs. Such drug combinations can provide many benefits over single-drug therapies: higher efficacy, lower toxicity and at least delayed onset of acquired treatment resistance [153–155]. Systematic screening efforts to measure their synergies on preclinical cancer models have been pursued [156–161], including that of the AstraZeneca–Sanger drug combination prediction DREAM challenge consortium [162] and NCI-ALMANAC [163]. In the case of PDXs, tumor growth inhibition of the combined drugs is compared with each of the drugs individually to determine greater-than-additive effects [83, 164]. Given space constraints, we cannot review databases aggregating and adding value to primary datasets such as PharmacDB [165] or DrugComb [166]. For the same reasons, we do not review studies exploiting these datasets using ML, but provide some references to the earliest of these studies/reviews in Tables 2 and 3 as a starting point. Further information is available in excellent reviews dedicated to this topic [5, 8, 167].

Table 4 compiles molecular profiles available for the resources listed in Tables 2 and 3. Some resources also release molecular profiles, whereas other only provide drug-response profiles with molecular profiles coming from independent resources (e.g. the profiles of NCI60 cell lines are provided in CellMinerCDB [168]). DepMap also provides molecular characterization of cancer cell line data from the CCLE project [121], which comprises the log copy number of gene from WES, protein expression measured by RPPA, quantitative proteomic profile by mass spectrometry, transcript per million-processed RNA-seq and gene mutation data.

Key points

- Cancer is a complex and multifactorial family of genetic diseases involving alterations in different types of macromolecules, which are captured by a range of ‘omics’ technologies.
- Preclinical cancer studies are based on a variety of disease models, including *in vitro*, *ex vivo* and *in vivo* models, generating from a tumor sample of a cancer patient.
- A wide range of omics technologies have been applied for molecular profiling of tumor samples, e.g. at the genomic, transcriptomic, epigenomic and proteomic levels.
- Cancer models are also profiled for their response (sensitivity or resistance) to a panel of drugs with high-throughput experimental approaches, thus also generating a large amount of data.
- The integration of the molecular profiles of cancer models with their corresponding drug-response profiles gives rise to pharmaco-omic datasets, hosted by a rising number of databases.

Funding

The HRH Princess Chulabhorn College of Medical Science (to C.P.) and Marie-Curie actions IEF-Horizon 2020 (under REA grant agreement no. 702933) (to G.G.).

References

1. Hutter C, Zenklusen JC. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* 2018;173:283–5.
2. Clinical Cancer Genome Task Team of the Global Alliance for Genomics and Health, Lawler M, Haussler D, et al. Sharing clinical and genomic data on cancer — the need for global solutions. *N Engl J Med* 2017;376:2006–9.
3. Bomane A, Gonçalves A, Ballester P. Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. *Front Genet* 2019;10:1041.
4. Naclaerts S, Dang C, Ballester PJ. Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget* 2017;8:97025–40.
5. Adam G, Rampásek L, Safikhani Z, et al. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis Oncologia* 2020;4:1–10.
6. Valencia A, Hidalgo M. Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. *Genome Med* 2012;4:1–17.
7. Singer J, Irmisch A, Ruscheweyh H-J, et al. Bioinformatics for precision oncology. *Brief Bioinform* 2019;20:778–88.
8. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev* 2019;11:31–9.
9. Dapic I, Uwugiaren N, Jansen PJ, et al. Fast and simple protocols for mass spectrometry-based proteomics of small fresh frozen uterine tissue sections. *Anal Chem* 2017;89:10769–75.
10. Sánchez-Navarro I, Gámez-Pozo A, González-Barón M, et al. Comparison of gene expression profiling by reverse transcription quantitative PCR between fresh frozen and formalin-fixed, paraffin-embedded breast cancer tissues. *Biotechniques* 2010;48:389–97.
11. Gao XH, Li J, Gong HF, et al. Comparison of fresh frozen tissue with formalin-fixed paraffin-embedded tissue for mutation analysis using a multi-gene panel in patients with colorectal cancer. *Front Oncol* 2020;10:310.
12. Gustafsson OJR, Arentz G, Hoffmann P. Proteomic developments in the analysis of formalin-fixed tissue. *Biochim Biophys Acta Proteins Proteomics* 1854;2015:559–80.
13. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *Am J Pathol* 2002;161:1961–71.
14. Li P, Conley A, Zhang H, et al. Whole-transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq. *BMC Genomics* 2014;15:1087.
15. April C, Klotzle B, Royce T, et al. Whole-genome gene expression profiling of formalin-fixed, paraffin-embedded tissue samples. *PLoS One* 2009;4:e8162.
16. Han Y, Li J. Sample types applied for molecular diagnosis of therapeutic management of advanced non-small cell lung cancer in the precision medicine. *Clin Chem Lab Med* 2017;55:1817–33.
17. Wood HM, Belvedere O, Conway C, et al. Using next-generation sequencing for high resolution multiplex anal-

- ysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res* 2010;**38**:e151.
18. Kerick M, Isau M, Timmermann B, et al. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med Genomics* 2011;**4**:68.
 19. Tuononen K, Mäki-Nevala S, Sarhadi VK, et al. Comparison of targeted next-generation sequencing (NGS) and real-time PCR in the detection of EGFR, KRAS, and BRAF mutations on formalin-fixed, paraffin-embedded tumor material of non-small cell lung carcinoma-superiority of NGS. *Genes Chromosomes Cancer* 2013;**52**:503–11.
 20. Gillespie JW, Best CJM, Bichsel VE, et al. Evaluation of non-formalin tissue fixation for molecular profiling studies. *Am J Pathol* 2002;**160**:449–57.
 21. Weng L, Wu X, Gao H, et al. MicroRNA profiling of clear cell renal cell carcinoma by whole-genome small RNA deep sequencing of paired frozen and formalin-fixed, paraffin-embedded tissue specimens. *J Pathol* 2010;**222**:41–51.
 22. Meng W, McElroy JP, Volinia S, et al. Comparison of microRNA deep sequencing of matched formalin-fixed paraffin-embedded and fresh frozen Cancer tissues. *PLoS One* 2013;**8**:e64393.
 23. Spencer DH, Sehn JK, Abel HJ, et al. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *J Mol Diagnostics* 2013;**15**:623–33.
 24. Fanelli M, Amatori S, Barozzi I, et al. Pathology tissue-chromatin immunoprecipitation, coupled with high-throughput sequencing, allows the epigenetic profiling of patient samples. *Proc Natl Acad Sci U S A* 2010;**107**:21535–40.
 25. Fanelli M, Amatori S, Barozzi I, et al. Chromatin immunoprecipitation and high-throughput sequencing from paraffin-embedded pathology tissue. *Nat Protoc* 2011;**6**:1905–19.
 26. Saliba AE, Westermann AJ, Gorski SA, et al. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 2014;**42**:8845–60.
 27. Moreno P, Huang N, Manning JR, et al. User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. *Nat Methods* 2021;**18**:327–8.
 28. Hu P, Zhang W, Xin H, et al. Single cell isolation and analysis. *Front Cell Dev Biol* 2016;**4**:116.
 29. Dove A. The art of culture: developing cell lines. *Science* 2014;**346**:1013–5.
 30. Mather JP. Concise review: cancer stem cells: in vitro models. *Stem Cells* 2012;**30**:95–9.
 31. Kaur G, Dufour JM. Cell lines. *Spermatogenesis* 2012;**2**:1–5.
 32. Lerescu L, Tucureanu C, Caraş I, et al. Primary cell culture of human adenocarcinomas—practical considerations. *Roum Arch Microbiol Immunol* 2008;**67**:55–66.
 33. Mohr SE, Perrimon N. RNAi screening: new approaches, understandings, and organisms. *Wiley Interdiscip Rev RNA* 2012;**3**:145–58.
 34. Wei L, Lee D, Law CT, et al. Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC. *Nat Commun* 2019;**10**:1–13.
 35. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a Cancer Dependency Map. *Cell* 2017;**170**:564–76.e16.
 36. 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**:1779–84.
 37. Dempster JM, Rossen J, Kazachkova M, et al. Extracting biological insights from the project Achilles genome-scale CRISPR screens in cancer cell lines. *bioRxiv* 2019;720243.
 38. Segeritz CP, Vallier L. Cell culture: growing cells as model systems in vitro. *Basic Sci Methods Clin Res* 2017;**151**–72.
 39. Eder A, Vollert I, Hansen A, et al. Human engineered heart tissue as a model system for drug testing. *Adv Drug Deliv Rev* 2016;**96**:214–24.
 40. Jabs J, Zickgraf FM, Park J, et al. Screening drug effects in patient-derived cancer cells links organoid responses to genome alterations. *Mol Syst Biol* 2017;**13**:955.
 41. Abbasi J. Patient-derived organoids predict cancer treatment response. *JAMA* 2018;**319**:1427.
 42. Seidlitz T, Merker SR, Rothe A, et al. Human gastric cancer modelling using organoids. *Gut* 2019;**68**:207–17.
 43. Vlachogiannis G, Hedayat S, Vatsiou A, et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science* 2018;**359**:920–6.
 44. Xu H, Jiao Y, Qin S, et al. Organoid technology in disease modelling, drug development, personalized treatment and regeneration medicine. *Exp Hematol Oncol* 2018;**7**: 30.
 45. Neal JT, Li X, Zhu J, et al. Organoid modeling of the tumor immune microenvironment. *Cell* 2018;**175**:1972–88.e16.
 46. Xu H, Lyu X, Yi M, et al. Organoid technology and applications in cancer research 11 medical and health sciences 1112 oncology and carcinogenesis. *J Hematol Oncol* 2018;**11**:116.
 47. Corrò C, Novellademunt L, Li VSW. A brief history of organoids. *Am J Physiol Cell Physiol* 2020;**319**:C151–65.
 48. Deroe YS, Wang G, Lin YC, et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat Med* 2011;**17**:1514–20.
 49. Cekanova M, Rathore K. Animal models and therapeutic molecular targets of cancer: utility and limitations. *Drug Des Devel Ther* 2014;**8**:1911–22.
 50. Doke SK, Dhawale SC. Alternatives to animal testing: a review. *Saudi Pharm J* 2015;**23**:223–9.
 51. Perlman RL. Mouse models of human disease: an evolutionary perspective. *Evol Med Public Heal* 2016;**2016**:eow014.
 52. De Jong M, Maina T. Of mice and humans: are they the same? - Implications in cancer translational research. *J Nucl Med* 2010;**51**:501–4.
 53. Mattern J, Bak M, Hahn EW, et al. Human tumor xenografts as model for drug testing. *Cancer Metastasis Rev* 1988;**7**:263–84.
 54. Hidalgo M, Amant F, Biankin AV, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov* 2014;**4**:998–1013.
 55. Aparicio S, Hidalgo M, Kung AL. Examining the utility of patient-derived xenograft mouse models. *Nat Rev Cancer* 2015;**15**:311–6.
 56. Tentler JJ, Tan AC, Weekes CD, et al. Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol* 2012;**9**:338–50.
 57. Hill SA, Denekamp J. Site dependent response of tumours to combined heat and radiation. *Br J Radiol* 1982;**55**:905–12.
 58. Haibe-Kains B, El-Hachem N, Birkbak NJ, et al. Inconsistency in large pharmacogenomic studies. *Nature* 2013;**504**: 389–93.
 59. Dey N, Sun Y, Leyland-Jones B, et al. Evolution of tumor model: from animal model of tumor to tumor model in animal. *J Cancer Ther* 2013;**4**:1411–25.

60. Li G. Patient-derived xenograft models for oncology drug discovery. *J Cancer Metastasis Treat* 2015;1:8–15.
61. Chen Y, Zhang R, Wang L, et al. Tumor characteristics associated with engraftment of patient-derived non-small cell lung cancer xenografts in immunocompromised mice. *Cancer* 2019;125:3738–48.
62. Chen C, Lin W, Huang Y, et al. The essential factors of establishing patient-derived tumor model. *J Cancer* 2021;12:28–37.
63. Ledford H. Cancer-genome study challenges mouse 'avatars'. *Nature* 2017.
64. Willyard C. The mice with human tumours: growing pains for a popular cancer model. *Nature* 2018;560:156–7.
65. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol* 2014;15:R47.
66. Rack K, Vidrequin S, Dargent JL. Genomic profiling of myeloma: the best approach, a comparison of cytogenetics, FISH and array-CGH of 112 myeloma cases. *J Clin Pathol* 2016;69:82–6.
67. Arcellana-Panlilio M, Robbins SM. I. Global gene expression profiling using DNA microarrays. *Am J Physiol Liver Physiol* 2002;282:G397–402.
68. Bishop R. Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Biosci Horizons* 2010;3:85–95.
69. Ratan ZA, Bin ZS, Mehta V, et al. Application of fluorescence in situ hybridization (FISH) technique for the detection of genetic aberration in medical science. *Cureus* 2017;9:e1325.
70. Weiss MM, Hermsen MAJA, Meijer GA, et al. Demystified comparative genomic hybridisation. *J Clin Pathol Mol Pathol* 1999;52:243–51.
71. James LA, Mitchell ELD, Menasce L, et al. Comparative genomic hybridisation of ductal carcinoma in situ of the breast: identification of regions of DNA amplification and deletion in common with invasive breast carcinoma. *Oncogene* 1997;14:1059–65.
72. Nakatsu N, Yoshida Y, Yamazaki K, et al. Chemosensitivity profile of cancer cell lines and identification of genes determining chemosensitivity by an integrated bioinformatical approach using cDNA arrays. *Mol Cancer Ther* 2005;4:399–412.
73. Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol* 2015;33:306–12.
74. Kao J, Salari K, Bocanegra M, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One* 2009;4:e6146.
75. Berger MF, Lawrence MS, Demichelis F, et al. The genomic complexity of primary human prostate cancer. *Nature* 2011;470:214–20.
76. Moorcraft SY, Gonzalez D, Walker BA. Understanding next generation sequencing in oncology: a guide for oncologists. *Crit Rev Oncol Hematol* 2015;96:463–74.
77. Malone ER, Oliva M, Sabatini PJB, et al. Molecular profiling for precision cancer therapies. *Genome Med* 2020;12:1–19.
78. Nik-Zainal S, Alexandrov LB, Wedge DC, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012;149:979–93.
79. Stephens PJ, McBride DJ, Lin ML, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009;462:1005–10.
80. Molenaar JJ, Koster J, Zwijnenburg DA, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* 2012;483:589–93.
81. Rausch T, Jones DTW, Zapatka M, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* 2012;148:59–71.
82. Samorodnitsky E, Jewell BM, Hagopian R, et al. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum Mutat* 2015;36:903–14.
83. Gao H, Korn JM, Ferretti S, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med* 2015;21:1318–25.
84. Jansen MPH, Knijnenburg T, Reijm EA, et al. Hallmarks of aromatase inhibitor drug resistance revealed by epigenetic profiling in breast cancer. *Cancer Res* 2013;73:6632–41.
85. Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000;24:236–44.
86. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57–63.
87. Fernandez-Cuesta L, Sun R, Menon R, et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biol* 2015;16:7.
88. Kim J, Kim S, Ko S, et al. Recurrent fusion transcripts detected by whole-transcriptome sequencing of 120 primary breast cancer samples. *Genes Chromosomes Cancer* 2015;54:681–91.
89. Ju YS, Lee WC, Shin JY, et al. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome Res* 2012;22:436–45.
90. Li Z, Gao X, Peng X, et al. Multi-omics characterization of molecular features of gastric cancer correlated with response to neoadjuvant chemotherapy. *Sci Adv* 2020;6:eaay4211.
91. Ku CS, Wu M, Cooper DN, et al. Exome versus transcriptome sequencing in identifying coding region variants. *Expert Rev Mol Diagn* 2012;12:241–51.
92. Han Y, Gao S, Muegge K, et al. Advanced applications of RNA sequencing and challenges. *Bioinform Biol Insights* 2015;9:29–46.
93. Nebbioso A, Tambaro FP, Dell'Aversana C, et al. Cancer epigenetics: moving forward. *PLoS Genet* 2018;14:e1007362.
94. Nagarajan RP, Fouse SD, Bell RJA, et al. Methods for cancer epigenome analysis. *Adv Exp Med Biol* 2013;754:313–38.
95. Kulis M, Esteller M. DNA methylation and cancer. *Adv Genet* 2010;70:27–56.
96. Robertson KD, Jones PA. DNA methylation: past, present and future directions. *Carcinogenesis* 2000;21:461–7.
97. Frommer M, McDonald LE, Millar DS, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A* 1992;89:1827–31.
98. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics* 2011;98:288–95.
99. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;17:208.
100. Kurdiani SK. Histone modifications in cancer biology and prognosis. *Prog Drug Res* 2011;67:91–106.

101. Gargiulo G, Minucci S. Epigenomic profiling of cancer cells. *Int J Biochem Cell Biol* 2009;**41**:127–35.
102. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**:669–80.
103. Pellegrini M, Ferrari R. Epigenetic analysis: ChIP-chip and ChIP-seq. *Methods Mol Biol* 2012;**802**:377–87.
104. Kaukonen D, Kaukonen R, Polit L, et al. Analysis of H3K4me3 and H3K27me3 bivalent promoters in HER2+ breast cancer cell lines reveals variations depending on estrogen receptor status and significantly correlates with gene expression. *BMC Med Genomics* 2020;**13**:92.
105. Nevedomskaya E, Wessels L, Zwart W. Genome-wide epigenetic profiling of breast cancer tumors treated with aromatase inhibitors. *Genomics Data* 2014;**2**:195–8.
106. Lin B, Lee H, Yoon JG, et al. Global analysis of H3K4me3 and H3K27me3 profiles in glioblastoma stem cells and identification of SLC17A7 as a bivalent tumor suppressor gene. *Oncotarget* 2015;**6**:5369–81.
107. Ke XS, Qu Y, Cheng Y, et al. Global profiling of histone and DNA methylation reveals epigenetic-based regulation of gene expression during epithelial to mesenchymal transition in prostate cells. *BMC Genomics* 2010;**11**:669.
108. Curry E, Zeller C, Masroun N, et al. Genes predisposed to DNA hypermethylation during acquired resistance to chemotherapy are identified in ovarian tumors by bivalent chromatin domains at initial diagnosis. *Cancer Res* 2018;**78**:1383–91.
109. Patel-Murray NL, Adam M, Huynh N, et al. A multi-omics interpretable machine learning model reveals modes of action of small molecules. *Sci Rep* 2020;**10**:954.
110. Zhao X, Wang MY, Jiang H, et al. Transcriptional programming drives Ibrutinib-resistance evolution in mantle cell lymphoma. *Cell Rep* 2021;**34**:108870.
111. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* 2012;**13**:227–32.
112. Cai Y, Yu X, Hu S, et al. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* 2009;**7**:147–154.
113. Carr SA, Abbatiello SE, Ackermann BL, et al. Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol Cell Proteomics* 2014;**13**:907–17.
114. Sallam RM. Proteomics in cancer biomarkers discovery: challenges and applications. *Dis Markers* 2015;**2015**:1–12.
115. Nishizuka S, Charboneau L, Young L, et al. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc Natl Acad Sci U S A* 2003;**100**:14229–34.
116. Ma Y, Ding Z, Qian Y, et al. Predicting cancer drug response by proteomic profiling. *Clin Cancer Res* 2006;**12**:4583–9.
117. Ianevski A, Lahtela J, Javarappa KK, et al. Patient-tailored design for selective co-inhibition of leukemic cell subpopulations. *Sci Adv* 2021;**7**:eabe4038.
118. Hafner M, Niepel M, Chung M, et al. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods* 2016;**13**:521–7.
119. Larsson P, Engqvist H, Biermann J, et al. Optimization of cell viability assays to improve replicability and reproducibility of cancer drug sensitivity screens. *Sci Rep* 2020;**10**:5798.
120. Niepel M, Hafner M, Chung M, et al. Measuring cancer drug sensitivity and resistance in cultured cells. *Curr Protoc Chem Biol* 2017;**9**:55–74.
121. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anti-cancer drug sensitivity. *Nature* 2012;**483**:603–7.
122. Seashore-Ludlow B, Rees MG, Cheah JH, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015;**5**:1210–23.
123. Breslin S, O'Driscoll L. Three-dimensional cell culture: the missing link in drug discovery. *Drug Discov Today* 2013;**18**:240–9.
124. van Tonder A, Joubert AM, Cromarty A. Limitations of the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl-2H-tetrazolium bromide (MTT) assay when compared to three commonly used cell enumeration assays. *BMC Res Notes* 2015;**8**:47.
125. Wang P, Henning SM, Heber D. Limitations of MTT and MTS-based assays for measurement of antiproliferative activity of green tea polyphenols. *PLoS One* 2010;**5**:e10202.
126. Berridge MV, Tan AS. Characterization of the cellular reduction of 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT): subcellular localization, substrate dependence, and involvement of mitochondrial electron transport in MTT reduction. *Arch Biochem Biophys* 1993;**303**:474–82.
127. Ediriweera MK, Tennekoon KH, Samarakoon SR. In vitro assays and techniques utilized in anticancer drug discovery. *J Appl Toxicol* 2019;**39**:38–71.
128. Skehan P, Storeng R, Scudiero D, et al. New colorimetric cytotoxicity assay for anticancer-drug screening. *J Natl Cancer Inst* 1990;**82**:1107–12.
129. Vichai V, Kirtikara K. Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nat Protoc* 2006:1112–1116.
130. Shen L, Kondo Y, Ahmed S, et al. Drug sensitivity prediction by CpG island methylation profile in the NCI-60 cancer cell line panel. *Cancer Res* 2007;**67**:11335–43.
131. Wlodkowic D, Skommer J, Faley S, et al. Dynamic analysis of apoptosis using cyanine SYTO probes: from classical to microfluidic cytometry. *Exp Cell Res* 2009;**315**:1706–14.
132. Aslantürk ÖS. In vitro cytotoxicity and cell viability assays: principles, advantages, and disadvantages. In: *Genotoxicity - A Predictable Risk to Our Actual World*, 2018.
133. Mueller H, Kassack MU, Wiese M. Comparison of the usefulness of the MTT, ATP, and calcein assays to predict the potency of cytotoxic agents in various human cancer cell lines. *J Biomol Screen* 2004;**9**:506–15.
134. Heiser LM, Sadanandam A, Kuo WL, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci U S A* 2012;**109**:2724–9.
135. Francies HE, Barthorpe A, McLaren-Douglas A, et al. Drug sensitivity assays of human cancer organoid cultures. *Methods Mol Biol* 2019;**1576**:339–51.
136. Sachs N, de Ligt J, Kopper O, et al. A living biobank of breast cancer organoids captures disease heterogeneity. *Cell* 2018;**172**:373–86.e10.
137. Yu C, Mannan AM, Yvone GM, et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol* 2016;**34**:419–23.
138. Némati F, Sastre-Garau X, Laurent C, et al. Establishment and characterization of a panel of human uveal melanoma xenografts derived from primary and/or metastatic tumors. *Clin Cancer Res* 2010;**16**:2352–62.
139. Townsend EC, Murakami MA, Christodoulou A, et al. The public repository of xenografts enables discovery

- and randomized phase II-like trials in mice. *Cancer Cell* 2016;**29**:574–86.
140. Garrido-Laguna I, Uson M, Rajeshkumar NV, et al. Tumor engraftment in nude mice and enrichment in stroma-related gene pathways predict poor survival and resistance to gemcitabine in patients with pancreatic cancer. *Clin Cancer Res* 2011;**17**:5793–800.
 141. Shoemaker RH. The NCI60 human tumour cell line anti-cancer drug screen. *Nat Rev Cancer* 2006;**6**:813–23.
 142. Corsello SM, Nagari RT, Spangler RD, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* 2020;**1**:235–48.
 143. Dang CC, Peón A, Ballester PJ. Unearthing new genomic markers of drug response by improved measurement of discriminative power. *BMC Med Genomics* 2018;**11**:10.
 144. Haverty PM, Lin E, Tan J, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* 2016;**533**:333–7.
 145. Conte N, Mason JC, Halmagyi C, et al. PDX Finder: a portal for patient-derived tumor xenograft model discovery. *Nucleic Acids Res* 2019;**47**:D1073–9.
 146. Georgopoulou D, Callari M, Rueda OM, et al. Landscapes of cellular phenotypic diversity in breast cancer xenografts and their impact on drug response. *Nat Commun* 2021;**12**:1998.
 147. Zhu Y, Brettin T, Evrard YA, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. *Sci Rep* 2020;**10**:1–11.
 148. Kim Y, Zheng S, Tang J, et al. Anticancer drug synergy prediction in understudied tissues using transfer learning. *J Am Med Informatics Assoc* 2020;**28**:42–51.
 149. Oh M, Park S, Lee S, et al. DRIM: a web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration. *Front Genet* 2020;**11**:564792.
 150. Costello JC, Heiser LM, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**:1202–12.
 151. Mirza B, Wang W, Wang J, et al. Machine learning and integrative analysis of biomedical big data. *Genes (Basel)* 2019;**10**:87.
 152. Caroli J, Dori M, Bicciato S. Computational methods for the integrative analysis of genomics and pharmacological data. *Front Oncol* 2020;**10**:185.
 153. Sugahara KN, Teesalu T, Karmali PP, et al. Coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs. *Science* 2010;**328**:1031–5.
 154. Holohan C, Van Schaeybroeck S, Longley DB, et al. Cancer drug resistance: an evolving paradigm. *Nat Rev Cancer* 2013;**13**:714–26.
 155. Crystal AS, Shaw AT, Sequist LV, et al. Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science* 2014;**346**:1480–6.
 156. Lieu CH, Tan A-C, Leong S, et al. From bench to bedside: lessons learned in translating preclinical studies in cancer drug development. *J Natl Cancer Inst* 2013;**105**:1441–56.
 157. Yu D, Kahen E, Cubitt CL, et al. Identification of synergistic, clinically achievable, combination therapies for osteosarcoma. *Sci Rep* 2015;**5**:16991.
 158. Budman DR, Calabro A, Rosen L, et al. Identification of unique synergistic drug combinations associated with downexpression of survivin in a preclinical breast cancer model system. *Anticancer Drugs* 2012;**23**:272–9.
 159. Kischkel FC, Eich J, Meyer CI, et al. New in vitro system to predict chemotherapeutic efficacy of drug combinations in fresh tumor samples. *PeerJ* 2017;**5**:e3030.
 160. Kashif M, Andersson C, Hassan S, et al. In vitro discovery of promising anti-cancer drug combinations using iterative maximisation of a therapeutic index. *Sci Rep* 2015;**5**:14118.
 161. Zoli W, Ricotti L, Tesei A, et al. In vitro preclinical models for a rational design of chemotherapy combinations in human tumors. *Crit Rev Oncol Hematol* 2001;**37**:69–82.
 162. Menden MP, Wang D, Mason MJ, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun* 2019;**10**:2674.
 163. Holbeck SL, Camalier R, Crowell JA, et al. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res* 2017;**77**:3564–76.
 164. Ice RJ, Chen M, Sidorov M, et al. Drug responses are conserved across patient-derived xenograft models of melanoma leading to identification of novel drug combination therapies. *Br J Cancer* 2020;**122**:648–57.
 165. Smirnov P, Kofia V, Maru A, et al. PharmacODB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res* 2018;**46**:D994–1002.
 166. Zagidullin B, Aldahdooh J, Zheng S, et al. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res* 2019;**47**:W43–51.
 167. De Niz C, Rahman R, Zhao X, et al. Algorithms for drug sensitivity prediction. *Algorithms* 2016;**9**:77.
 168. Luna A, Elloumi F, Varma S, et al. CellMiner Cross-Database (CellMinerCDB) version 1.2: exploration of patient-derived cancer cell line pharmacogenomics. *Nucleic Acids Res* 2021;**49**:D1083–93.
 169. Thunnissen E, Bubendorf L, Dietel M, et al. EML4-ALK testing in non-small cell carcinomas of the lung: A review with recommendations. *Virchows Arch* 2012;**461**:245–57.
 170. Commo F, Ferte C, Soria JC, et al. Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology. *Ann Oncol* 2015;**26**:582–8.
 171. Kudoh K, Ramanna M, Ravatn R, et al. Monitoring the expression profiles of doxorubicin-induced and doxorubicin-resistant cancer cells by cDNA microarray. *undefined* 2000;**60**:4161–6.
 172. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. *Lancet* 2010;**375**:1525–35.
 173. Gillis NK, Patel JN, Innocenti F. Clinical implementation of germ line cancer pharmacogenetic variants during the next-generation sequencing era. *Clin Pharmacol Ther* 2014;**95**:269–80.
 174. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 2019;**569**:503–8.
 175. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**:570–5.
 176. Miranda SP, Baião FA, Maçaira PM, et al. Predicting drug sensitivity of cancer cells based on DNA methylation levels. *bioRxiv*. 2020; 2020.08.25.266049.
 177. Zhu X, Liu J, Xu X, et al. Genome-wide analysis of histone modifications by ChIP-chip to identify silenced genes in gastric cancer. *Oncol Rep* 2015;**33**:2567–74.

178. Schmidl C, Vladimer GI, Rendeiro AF, et al. Combined chemosensitivity and chromatin profiling prioritizes drug combinations in CLL. *Nat Chem Biol* 2019;**15**:232–40.
179. Nishizuka S, Charboneau L, Young L, et al. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc Natl Acad Sci U S A* 2003;**100**:14229–34.
180. Ali M, Khan SA, Wennerberg K, et al. Global proteomics profiling improves drug sensitivity prediction: Results from a multi-omics, pan-cancer modeling approach. *Bioinformatics* 2018;**34**:1353–62.
181. Iorio F, Knijnenburg TA, Vis DJ, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* 2016;**166**:740–54.
182. Menden MP, Iorio F, Garnett M, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS One* 2013;**8**:e61318.
183. Ammad-Ud-Din M, Georgii E, Gönen M, et al. Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J Chem Inf Model* 2014;**54**:2347–59.
184. Rees MG, Seashore-Ludlow B, Cheah JH, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2016;**12**:109–16.
185. Rampášek L, Hidru D, Smirnov P, et al. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* 2019;**35**:3743–51.
186. Kurilov R, Haibe-Kains B, Brors B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci Rep* 2020;**10**:1–11.
187. Gerdes H, Casado P, Dokal A, et al. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nat Commun* 2021;**12**:1–15.
188. Nguyen L, Naulaerts S, Bomane A, et al. Machine learning models to predict in vivo drug response via optimal dimensionality reduction of tumour molecular profiles. *bioRxiv*. 2018;277772.
189. Li H, Li T, Quang D, et al. Network Propagation Predicts Drug Synergy in Cancers. *Cancer Res* 2018;**78**:5446–57.
190. Xia F, Shukla M, Brettin T, et al. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics* 2018;**19**:71–9.
191. Sidorov P, Naulaerts S, Arieu-Bonnet J, et al. Predicting Synergism of Cancer Drug Combinations Using NCI-ALMANAC Data. *Front Chem* 2019;**7**:509.

Appendix II

Interpretable Machine Learning Models to Predict the Resistance of Breast Cancer Patients to Doxorubicin from Their microRNA Profiles

Adeolu Z. Ogunleye, Chayanit Piyawajanusorn, Anthony Gonçalves, Ghita Ghislat, and Pedro J. Ballester*

Doxorubicin is a common treatment for breast cancer. However, not all patients respond to this drug, which sometimes causes life-threatening side effects. Accurately anticipating doxorubicin-resistant patients would therefore permit to spare them this risk while considering alternative treatments without delay. Stratifying patients based on molecular markers in their pretreatment tumors is a promising approach to advance toward this ambitious goal, but single-gene markers such as HER2 expression have not shown to be sufficiently predictive. The recent availability of matched doxorubicin-response and diverse molecular profiles across breast cancer patients permits now analysis at a much larger scale. 16 machine learning algorithms and 8 molecular profiles are systematically evaluated on the same cohort of patients. Only 2 of the 128 resulting models are substantially predictive, showing that they can be easily missed by a standard-scale analysis. The best model is classification and regression tree (CART) nonlinearly combining 4 selected miRNA isoforms to predict doxorubicin response (median Matthew correlation coefficient (MCC) and area under the curve (AUC) of 0.56 and 0.80, respectively). By contrast, HER2 expression is significantly less predictive (median MCC and AUC of 0.14 and 0.57, respectively). As the predictive accuracy of this CART model increases with larger training sets, its update with future data should result in even better accuracy.

1. Introduction

Breast cancer (BC) has the highest global incidence and mortality rate amongst all cancer types affecting women.^[1,2] In 2020, BC rose above lung cancer to become the most frequently diagnosed cancer worldwide,^[1] with over 2 million new cases recorded (11.7% of all reported new cases) and over 600 000 deaths (6.9% of the overall cancer deaths recorded). Doxorubicin is an intravenous chemotherapy, within the anthracycline class of drugs, used in both the early and the advanced setting of various cancer types, including BC.^[3,4] Doxorubicin works essentially by intercalating between neighboring DNA base pairs.^[5,6] The resulting doxorubicin–DNA complex inhibits topoisomerase II activity, which subsequently disrupts DNA replication and transcription, leading to both cytotoxic and apoptotic cell death.^[7,8] Unfortunately, primary resistance to treatment is common across cancer types and drugs.^[9–12] Such *de novo* resistance is also commonly observed with doxorubicin-containing treatments in BC patients.^[13]

A. Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat,
P. J. Ballester
Cancer Research Center of Marseille (CRCM)
INSERM U1068
Marseille F-13009, France
E-mail: pedro.ballester@inserm.fr

A. Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat,
P. J. Ballester
Cancer Research Center of Marseille (CRCM)
Institut Paoli-Calmettes
Marseille F-13009, France

A. Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat,
P. J. Ballester
Cancer Research Center of Marseille (CRCM)
Aix-Marseille Université
Marseille F-13284, France

A. Z. Ogunleye, C. Piyawajanusorn, A. Gonçalves, G. Ghislat,
P. J. Ballester
Cancer Research Center of Marseille (CRCM)
CNRS UMR7258
Marseille F-13009, France

P. J. Ballester
Department of Bioengineering
Imperial College London
London SW7 2AZ, UK

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/advs.202201501>

© 2022 The Authors. Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202201501

A central goal of precision oncology is to anticipate which patients will be resistant to a given drug.^[14,15] This anticipation would result in the identified patients receiving without delay alternative treatments more likely to stop cancer progression. In the case of doxorubicin, which may cause cardiotoxicity,^[7,16,17] such predictors would also be helpful to avoid resistant-predicted patients taking unnecessary risks. Historically, this goal has been approached by searching for a single-gene marker, which is a molecular feature able to distinguish between sensitive and resistant tumors to the treatment. While many efforts have employed preclinical data with this purpose,^[18–21] much less abundant clinical datasets are by definition the most relevant for patients.^[14] An example of the latter clinical studies is one where high HER2 expression in BC patient tumors was reported to be predictive of patient response to anthracyclines, including doxorubicin.^[22] A major limitation, however, is that many single-gene markers only provide modest predictive accuracy restricted to a very small fraction of patients.^[15] As an example, 16% of EGFR-mutant non-small-cell lung cancer (NSCLC) patients were found to respond to Erlotinib, but the prevalence of this mutation was low and 7% of EGFR-WT NSCLC patients also responded to this drug.^[23] Thus, at least in this cohort, the Matthew correlation coefficient (MCC) of this FDA-approved single-gene marker is just 0.11,^[24] i.e., slightly more predictive than random guessing. These practical limitations of single-gene markers may be due to biological reasons (e.g., patient response to a drug is the outcome of a complex multifactorial process, which is often poorly anticipated by the status of a single gene), but also technical reasons (e.g., using metrics that can strongly overestimate predictive performance in the common scenario where there is a class imbalance in the data).^[25,26] Multigene expression signatures of drug response are also a common approach, but these are limited by neglecting epistatic effects and profiles other than messenger RNA (mRNA) expression.^[27–30]

These limitations are now being overcome by machine learning (ML), which can generate computational models exploiting multiple pretreatment features of patients to predict their response to that drug treatment. There are many of such computational studies using pharmaco-omic data from preclinical models,^[31] especially data from cancer cell lines^[32–35] but also data from primary tumor cultures^[36,37] or from patient-derived xenografts.^[24,38] When systematically and directly compared to single-gene markers, multigene ML models mostly offer higher MCC and almost always much higher recall.^[39,40] However, while some progress on leveraging preclinical data with this purpose has been made,^[24,41–47] preclinical models still tend to struggle to predict drug response in patients with a useful level of accuracy.^[48,49]

ML models exploiting clinical data are hence attractive in this context. However, such studies also have their challenges, e.g., the scarcity of suitable datasets and the need for substantial curation before these datasets can be used for this purpose. For instance, a patient tumor's molecular features are likely to be altered upon drug treatment.^[50,51] Therefore, as these are unlikely to represent the pretreatment molecular features of the tumor faithfully, one aim of curation is to discard tumors that were profiled after the drug is administered (otherwise, these noisy data

instances could degrade model performance). Following such curation, Ding et al. built binary classifiers for six drug-cancer type binomials,^[52] each exploiting four molecular profiles: copy number alteration, DNA methylation, mRNA, and microRNA (miRNA). An elastic net with bootstrapping was used for each profile to select the most predictive molecular features, and a final ensemble classifier was built based on these features. Using fivefold cross-validation (CV) predictions of drug treatment response, the area under the curve (AUC) of the receiver operating characteristic (ROC), hereafter AUC for short, was above random-guess level in 4 of the 6 binomials in at least one of the considered profiles. In Bomane et al.,^[53] our lab focused on a single drug-cancer type binomial, paclitaxel-BC, but expanded the analysis to 6 profiles and 10 ML algorithms. Interestingly, miRNA and DNA methylation profiles were revealed to be highly predictive for this binomial. However, the latter only occurred with two algorithms: classification and regression tree (CART) and, to a lesser extent, extreme gradient boosting (XGBoost) integrated with feature selection. Overall, there are still few studies on the application of ML to predict drug treatment response, in part due to the scarcity of suitable clinical samples for a given drug-cancer type binomial, although its application to predict other patient outcomes requiring less curation has received more attention (e.g., prognosis).^[54]

To our knowledge, ML is yet to be applied to predict BC patient response to doxorubicin treatments. Here we will evaluate a range of ML algorithms, which generally results in much better prediction than restricting to a single algorithm.^[24,53] This is particularly true when considering algorithms that incorporate some form of feature selection to mitigate the impact of high-dimensionality in the training data. The latter may have the additional advantage of providing interpretable molecular hypotheses for interpatient doxorubicin response variability. Integrating optimal model complexity (OMC) strategies with ML algorithms has resulted in improved predictions with problems spanning a range of drugs and cancer types.^[24,53,55] Another novel aspect of this study is that we will also investigate for the first time multivariate predictors based on patient response and profiles other than mRNA expression for this drug. The US National Cancer Institute (NCI) Genomic Data Commons (GDC) datasets have been harmonized across different cancer genome programs,^[56] providing a range of clinical drug response data and omics profiles. Thus, the data used in this study were obtained from the GDC repository (<https://portal.gdc.cancer.gov>). Beyond single-gene and/or multigene expression predictors with modest performance,^[15,23] we considered data from multiple molecular profiles arising from the application of various next-generation sequencing technologies to patient samples and processed via various GDC workflows with their annotated treatment and biospecimen information. Drug response data of BC patients required the most intensive curation (**Figure 1A**), as detailed in the Experimental Section. From a total of 64 drugs administered across BC patients, we could identify 96 of these patients (Table S1, Supporting Information) with both molecular profiling of their tumors and annotated responses to a doxorubicin-containing treatment. **Figure 1B** presents the scheme of our methodology.

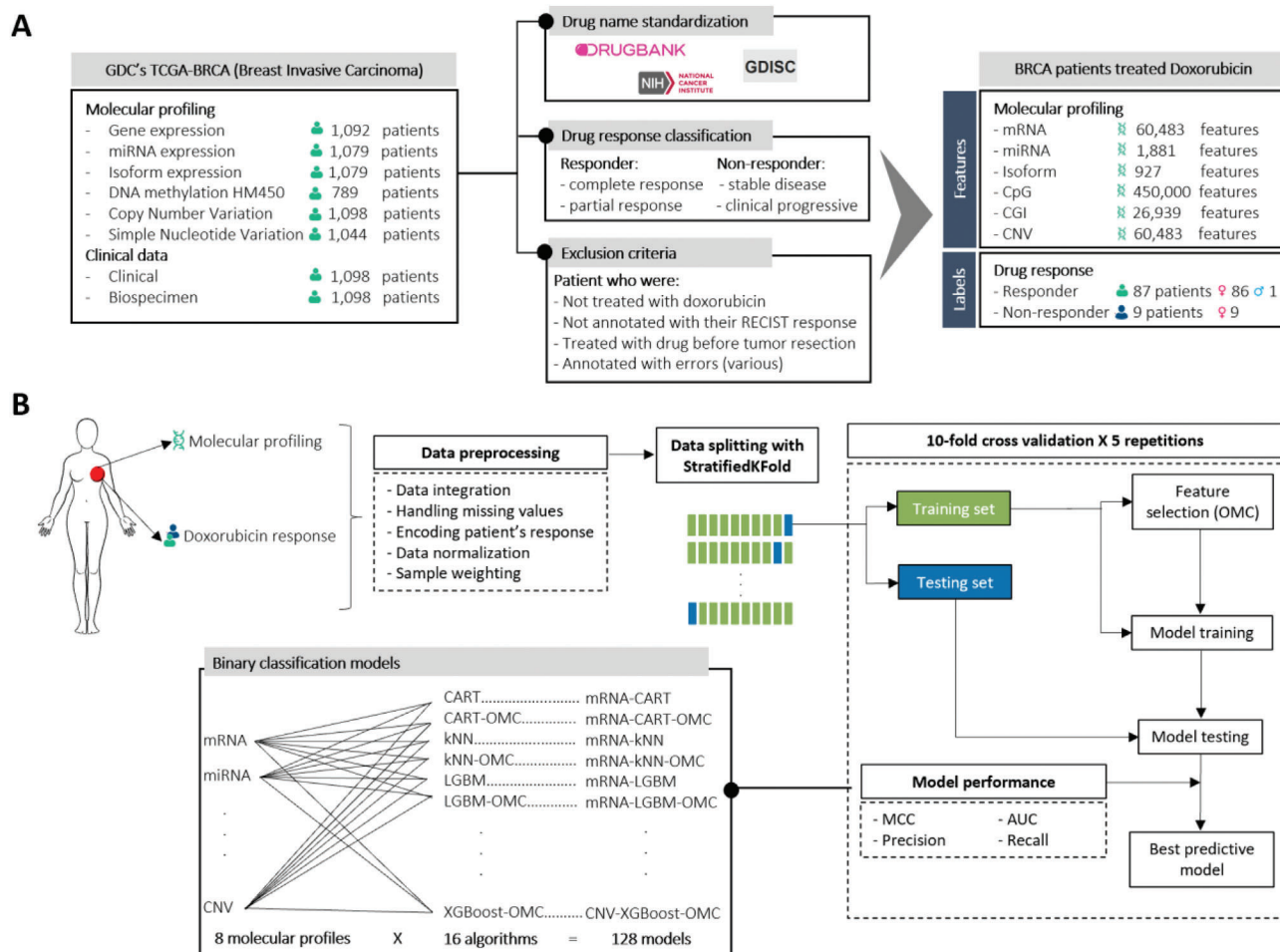


Figure 1. Schematic representation of the development of supervised learning models to predict patient response to doxorubicin. A) Multiomics datasets, including molecular profiles of patient tumors generated with high-throughput technologies such as RNA-Sequencing, miRNA-Sequencing or DNA methylation array, were retrieved from the NCI GDC data repository. The corresponding biospecimen and clinical datasets were also retrieved from the GDC and curated to retain valid records only. All valid records came from the GDC-enriched The Cancer Genome Atlas – Breast Invasive Carcinoma (TCGA-BRCA) project. B) These datasets were subsequently preprocessed and used to build and evaluate a range of supervised learning models by tenfold CV with five repetitions. Because these datasets have high dimensionality, OMC models were also built to identify and retain the most important features.^[24,53] 16 binary classification algorithms, 8 algorithms either with or without the OMC strategy, were applied to each of the 8 molecular profiles, resulting in the generated 128 models. Diverse binary classification metrics were used to evaluate the predictive performance of each developed model.

2. Results

Using data from the 96 BC patients (Table S1, Supporting Information), we employed 16 binary classification algorithms to generate and evaluate ML models to predict the responses to doxorubicin of these patients from the molecular profiles of their tumors. Figure 1 summarizes this process, which is fully specified in the Experimental Section. Figure S1 of the Supporting Information shows the high dimensionality of each profile, ranging from the expression of 927 miRNA isoforms (isomiR) to the methylation levels of 450 000 DNA probes.

2.1. Identifying the Most Predictive ML Algorithms and Molecular Profiles for Doxorubicin-Response Prediction in BC Patients

The accuracy of each set of tenfold CV (10CV) predictions for the patients by each model is quantified by its MCC. The median

MCC (mMCC) of five 10CV repetitions, each with a different initial random seed, is presented in Figure 2. Two models were able to distinguish between responders and nonresponders with an mMCC of at least 0.3, including mMCC of 0.56 from CART using isomiR features and mMCC of 0.32 from CART using miRNA features. Note that most models obtain a near-random predictive level ($MCC \approx 0$), with perfect prediction ($MCC = 1$) being still far away from the best models (this was expected, as we will discuss later). Figure S2 of the Supporting Information presents the results using other evaluation metrics, which evidences that the more common AUC is a less demanding metric than MCC. Indeed, AUC tends to overestimate predictive performance in imbalanced classification problems,^[57,58] which is not the case of MCC.^[59] This is our main reason to use MCC as the primary performance metric.

Looking at the seven models with an mMCC of at least 0.2 (Figure S3, Supporting Information), all were built with

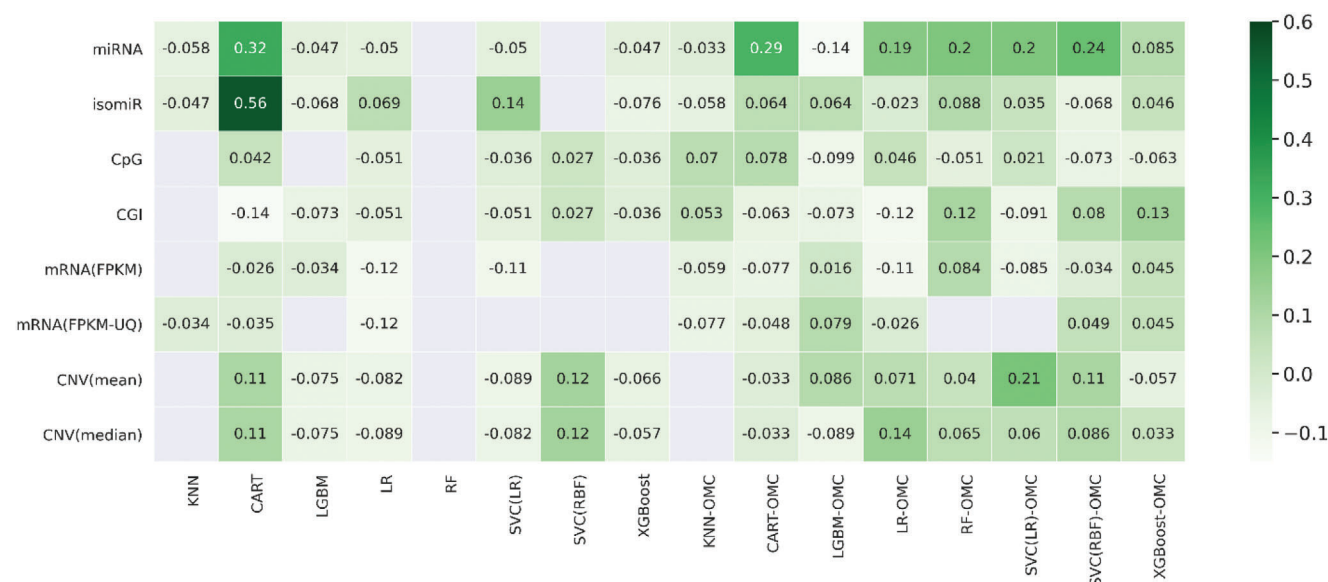


Figure 2. Heatmap showing the median MCC (mMCC) of five tenfold CV runs for each of the 128 models. Each row corresponds to a given molecular profile, each column refers to the employed classification algorithm. The first 8 algorithms on the left were ran without OMC and hence each lead to an all-features model considering all available features from the processed datasets during model building. The rest of algorithms (the other 8 on the right) were ran with OMC to search for a small subset of features facilitating the classification of patients, thus the suffix “OMC” was added to the algorithm name. 10-by-10 nested-CV runs were carried out five times, each time with a different random seed. For each run, CV predictions were merged to calculate the evaluation metrics. In this way, five MCC scores are obtained for each algorithm and molecular profile binomial, with its mMCC being shown in the heatmap. The two most predictive models were built with decision trees: mMCC of 0.56 from CART using isomiR features and mMCC of 0.32 from CART using miRNA features. However, the general trend is the OMC model (right) having higher MCC than its corresponding all-features model (left). Models resulting in undefined mMCC scores are indicated as blank grey boxes. Here MCC is undefined because the model predicts the same class for all instances (e.g., if all are predicted positive, then the sum of true negatives and false negatives is by definition zero, which in turn makes the MCC denominator zero as well).

algorithms inducing feature selection. Two-thirds were OMC models, whereas the rest were all-features CART models. Six of these models employed either isomiR or miRNA features. An inspection of the best results in terms of median AUC (mAUC) and their corresponding mMCC (Figure S4, Supporting Information) shows that these models also have high mAUC (CART_isomiR with an mAUC of 0.80 and CART_miRNA with an mAUC of 0.64). Noticeably, isomiR gives a better prediction of nonresponders when compared to miRNA (Table S2, Supporting Information). By contrast, Figure S4 of the Supporting Information also reveals many models with good mAUC but poor mMCC such as miRNA_LGBM (Light Gradient Boosting Machine). The latter model has 14 false negatives and 9 false positives, whereas CART_isomiR only has 4 false negatives and 3 false positives (Table S2, Supporting Information). This further supports the use of MCC for this type of class-imbalanced problems.

2.2. Determining the Robustness of the Best ML Models

After identifying isomiR_CART and miRNA_CART as the best ML models, here we evaluate how their predictive accuracies vary with different training set sizes and random seeds. With this purpose, we also conducted five runs of threefold CV (3CV) and five-fold CV (5CV) experiments, in addition to the five 10CV runs from the previous section, using different random seeds. Both models were robust to this type of variability given the similar MCC values returned within each set of experiments summa-

rized by a boxplot (Figure 3). Higher MCC values were observed as larger training sets were employed (from training with 67% of the data in 3CV to training with 90% of the data in 10CV). To find out which part of the predictive accuracy comes from signal in the data, we repeated all the CV runs exactly in the same manner from class-permuted versions of the datasets. As a result, all the latter runs obtained near random-level MCC values, which were significantly worse than those arising from models trained on the original data in all cases (Figure 3).

2.2.1. Impacts of Data Integration on Doxorubicin Response Prediction

We next look at whether adding or considering other sets features to describe patients improves prediction further. CART using clinical data was barely predictive (Figure 4; Table S3, Supporting Information). These clinical features comprised sex, age, tumor stage, histological type, menopause status and the status of the estrogen, progesterone, and HER2 receptors (Table S1, Supporting Information). CART with merged clinical and isomiR features did not result in an improvement in MCC. However, integrating clinical and miRNA data improved MCC, although the difference with using miRNA features alone was not significant. We also evaluated CART with merged isomiR and miRNA features, which led to worse MCC values than using either profile alone. Lastly, CART using all the profiles as features led to the worse MCC overall.

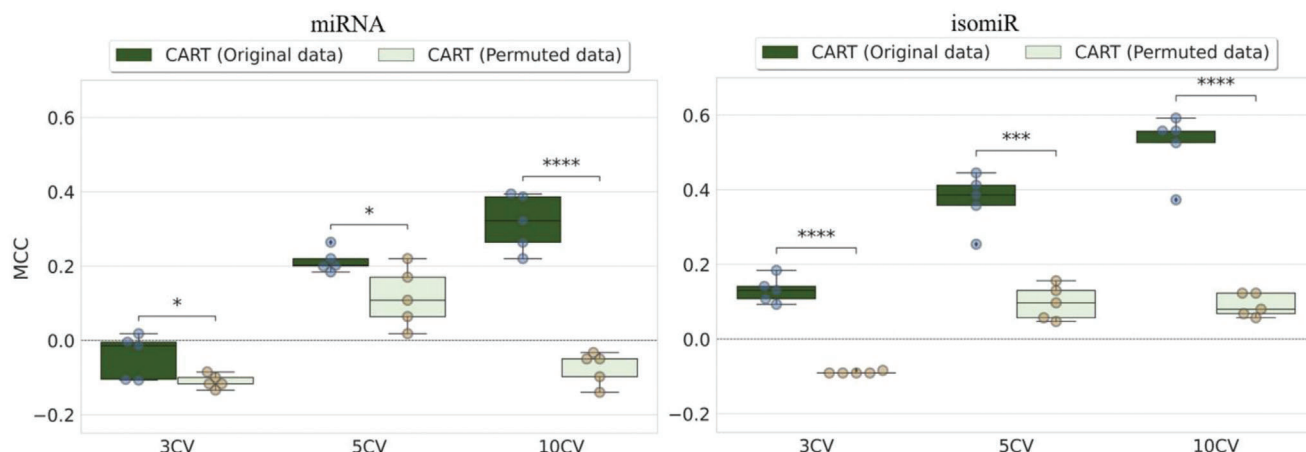


Figure 3. CV performance of the two most predictive models compared to those from training with permuted data. The boxplots present the distributions of MCC scores obtained across five iterations of CART models evaluation implemented on miRNA (left) and isomiR (right) datasets with 3CV, 5CV, and 10CV. CART models trained on the original dataset, i.e., all-features data (deep green), and CART model trained on the class permuted dataset (light green). Each model's predictive performances (original and permuted) are compared within each of the CVs implemented for isomiR and miRNA. The horizontal bars above the boxplots indicate the significance levels between these distributions: “*” means $0.01 < p < 0.05$, “***” means $0.001 < p < 0.01$, “****” means $0.0001 < p < 0.001$, and “*****” means $p < 1.00 \times 10^{-4}$. These p -values were calculated using two-sided Welch's t -tests. CART models obtain significantly better MCC than the permuted models in all the CVs, both with miRNA and isomiR features. Each dot represents a repetition.

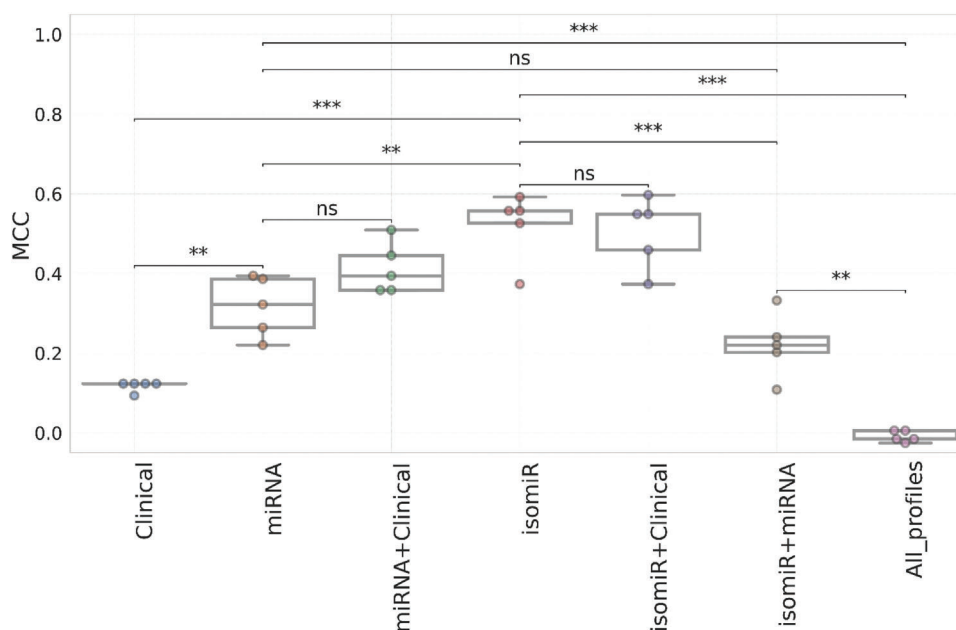


Figure 4. Comparison of the predictive performances of the model trained on the predictive molecular profile to those combining other datasets: Boxplots comparing the MCCs obtained from five runs of tenfold CV for the most predictive profiles (miRNA and isomiR) with those combining clinical data, merged both profiles (isomiR + miRNA) and those merging all the profiles considered in this study (clinical + 8 molecular profiles). The predictive models combining clinical and demographic information (Table S1, Supporting Information) with each of miRNA and isomiR profiles slightly performed better than those trained on miRNA and isomiR individually, but the differences were not statistically significant ($p > 0.05$). However, when the two predictive profiles were combined (miRNA + isomiR), a significant decrease in performance was observed compared with individual predictive profiles. Finally, the model combining all the 8 molecular profiles with clinical data has the least predictive performances; this could be due to the curse of dimensionality. It was observed that the higher the dimension of our dataset, the lower the performance of the models on the molecular profiles. Statistical comparisons between different models were performed using Welch's t -test (two-sided). Stars denote p -value of the test, where; nonsignificant “ns” means $0.05 < p < 1.00$, “*” means $0.01 < p < 0.05$, “**” means $0.001 < p < 0.01$, and “***” means $0.0001 < p < 0.001$. Each dot represents a repeat.

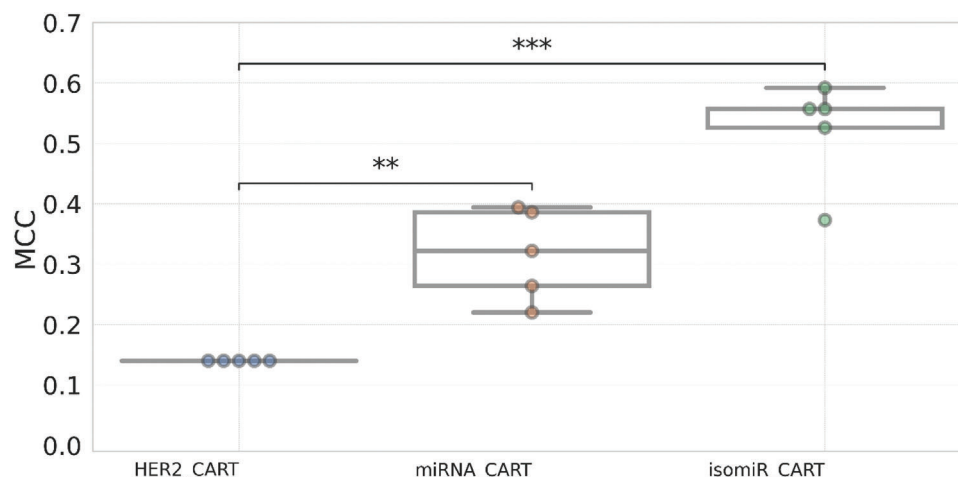


Figure 5. Comparison of the HER2-based model with the best ML models to predict doxorubicin response in BC patients. Boxplots comparing the MCCs of CART models using miRNAs and their isoforms with those obtained using HER2 expression values only. HER2 expression (left) gave a significantly lower ($p < 0.01$) predictive performance than the two best ML models from this study (right). Legend: “***” means $0.001 < p \leq 0.01$ and “**” means $0.0001 < p \leq 0.001$. The p -values were calculated using two-sided Welch’s t -test. Each dot represents a repeated run.

2.2.2. Comparing the ML Models to the Existing Single-Gene Marker of Doxorubicin Response

Gennari et al.,^[60] Rody et al.,^[22] Zhang and Liu^[61] showed independently that HER2 status is a marker of doxorubicin sensitivity in BC, with higher HER2 expression indicating higher sensitivity to the drug. **Figure 5** shows that the mMCC obtained from the HER2 model is just 0.14, which is significantly lower than our predictive models with 3 miRNAs (mMCC = 0.32) or with 4 isomiRs (mMCC = 0.56). Table S4 of the Supporting Information also displays a sharp difference in terms of AUC. These results suggest that these ML models are able to predict doxorubicin response much better than the HER2 marker.

2.3. Assessing the Applicability Domain of the Best ML Models

We are not aware of publicly-available BC datasets with other miRNA-profiled doxorubicin-treated patients. However, there is a total of 1078 miRNA-profiled patients in the GDC TCGA-BRCA project. Therefore, we can investigate whether the subset of 95 miRNA-profiled doxorubicin-treated patients constitutes a representative sample of all these patients. **Figure 6A** shows how these 1078 patients cluster by their similarity in the expression of predictive miRNA features. (Figure 6B shows the clustering of the same patients with respect to predictive isomiR features.) In both cases, the 95 doxorubicin-treated patients are evenly distributed across clusters, which means that all major clusters are represented in the training set of each of the best ML models we identified. This lack of a strong distribution shift between the training and potential test sets suggests that these models should also be predictive on the rest of miRNA-profiled patients.

2.4. The Best ML Models Are Also Highly Interpretable

Out of the 16 models evaluating miRNA features, CART was found to be the most predictive. This model selected 3 (hsa-miR-

4421, hsa-miR-4680, and hsa-miR-514a-1) out of the 1881 considered miRNAs (**Figure 7A**; Table S2, Supporting Information). On the other hand, among the models employing isomiR features, only CART was substantially predictive by selecting 4 (hsa-miR-450a-1, hsa-miR-19b-2, hsa-miR-92a-2, and hsa-miR-92a-1) out of 927 analyzed isomiR features (Figure 7B; Table S2, Supporting Information).

2.4.1. Predicted miRNA Target Genes

A miRNA is a small single-stranded noncoding RNA molecule able to bind to a specific set of mRNA molecules. Such binding results in either degrading the bound mRNA molecule or suppressing its translation into a protein.^[62] That is, an miRNA influences the post-transcriptional regulation of a specific set of protein-coding genes (its targeted genes). A total of 13898 genes are predicted to be targeted by the 7 predictive miRNAs (**Table 1**) according to TargetScan.^[63]

To unveil BC-associated processes, we first determined the overlap between TargetScan genes and the 89 and 37 genes that are reported to be BC-driving genes in Integrative OncoGenomics (IntOGen)^[64] and Catalogue of Somatic Mutation in Cancer (COSMIC)^[65,66] databases, respectively. To enhance the BC-specificity of the analysis, we focus on the 23 genes that we found to be common to these three gene lists (**Figure 8A**): AKT1, ARID1A, ARID1B, BAP1, BRCA1, BRCA2, CASP8, CDH1, CDKN1B, CTCF, ERBB2, ESR1, FOXA1, GATA3, MAP2K4, MAP3K1, NCOR1, PIK3CA, RB1, SALL4, SMARCD1, TBX3, TP53 (Table S6, Supporting Information).

2.4.2. Enrichment Analysis (EA) of the Genes Targeted by the Predictive miRNAs

We next carried out EA using the over representation analysis (ORA) method implemented in the WEB-based Gene SeT

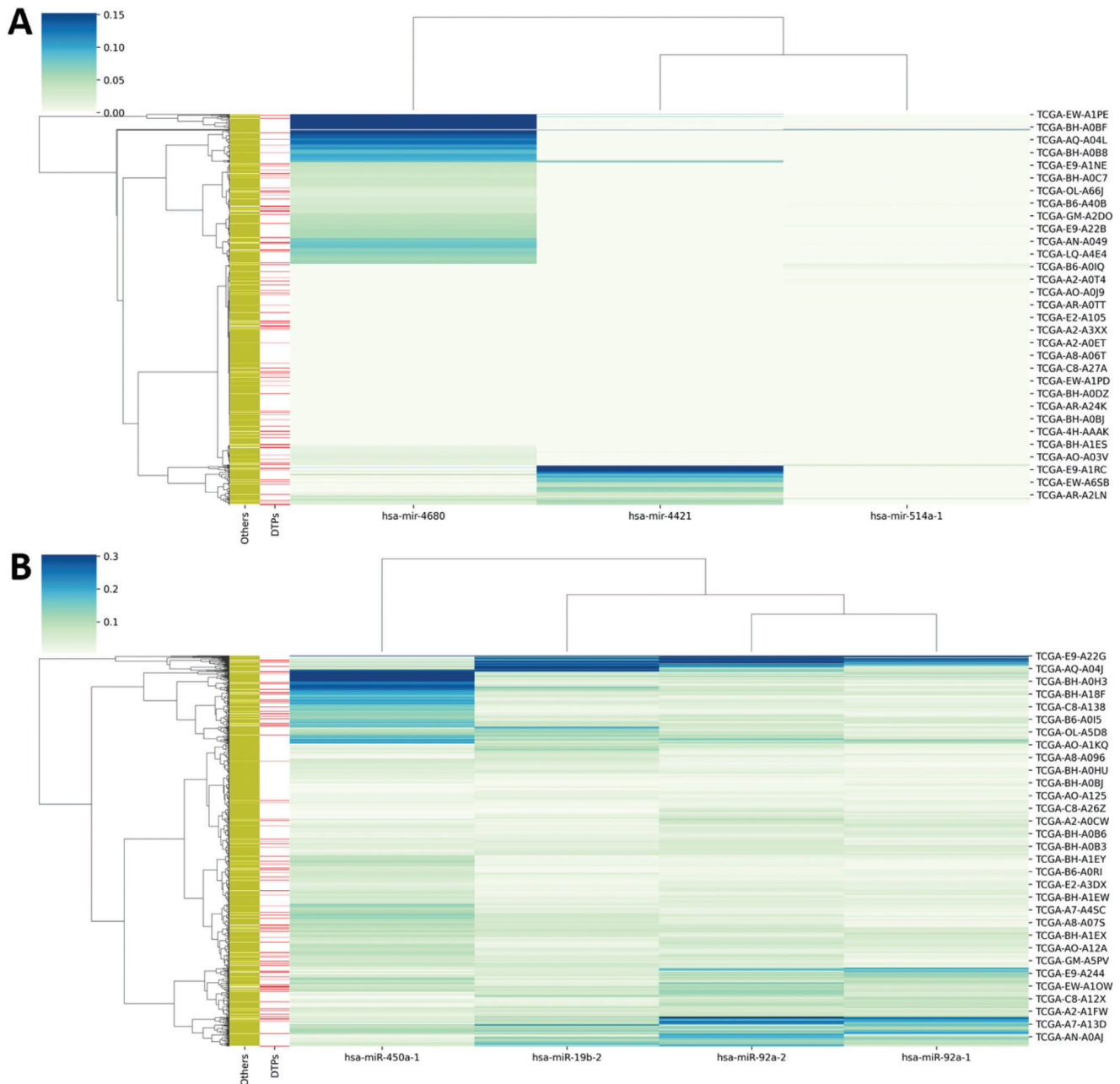


Figure 6. Expression of the selected miRNAs and isomiRs in 1078 TCGA-BRCA cases. Clustering on the RPM-normalized expression for the A) 3 selected miRNAs and B) 4 selected isomiRs with all the 1078 TCGA-BRCA patients having both miRNA and isomiR data (with and without treatments records). The dendrogram on the left shows the clustering of patients, while the one on the top shows the clustering of the selected features (3 miRNAs and 4 isomiRs). The colors on the top-left of the heat map represent the scale of the normalized expression values, while the colors in the heat map represent the expression intensities of the selected subsets of features. For the selected miRNAs (A), hsa-miR-4680 has more cases with the highest expression values among the 3 selected across all the cases considered, has-miR-4421 followed this and hsa-miR-514a-1 has the least expression values as indicated by color partitions in the heatmap. Similarly, the column dendrogram showed that the expression of has-miR-4421 and hsa-miR-514a-1 are more similar (i.e., both were in the same cluster) while that of hsa-miR-4680 are less similar to the other 2 miRNAs as indicated by the distance observed between the clusters. For the selected isomiRs (B), hsa-miR-450a-1 has more cases with higher expression values when compared with others, while the remaining isomiRs (hsa-miR-19b-1, hsa-miR-92a-2, and hsa-miR-92a-1) have very similar expressions across all the cases. Despite their similarity, hsa-miR-19b-2 still has slight dissimilarity from other 2 and as such, it belongs to another cluster, whereas hsa-miR-92a-2 and hsa-miR-92a-1 form a cluster. Patients belonging to the same cluster/group are more similar and are less similar to patients in other clusters. For readability, the labels to the right of the y-axis only present 36 patient IDs from different clusters out of 1078 included in each plot (this was done by minimizing the figure size). On the left of the dendrogram, the DTPs distribution is shown in red against other cohorts in yellow. There is no clear separation between the DTPs and others cohorts in both plots, which shows that the 95 patients are representative of the full 1078-patient cohort.

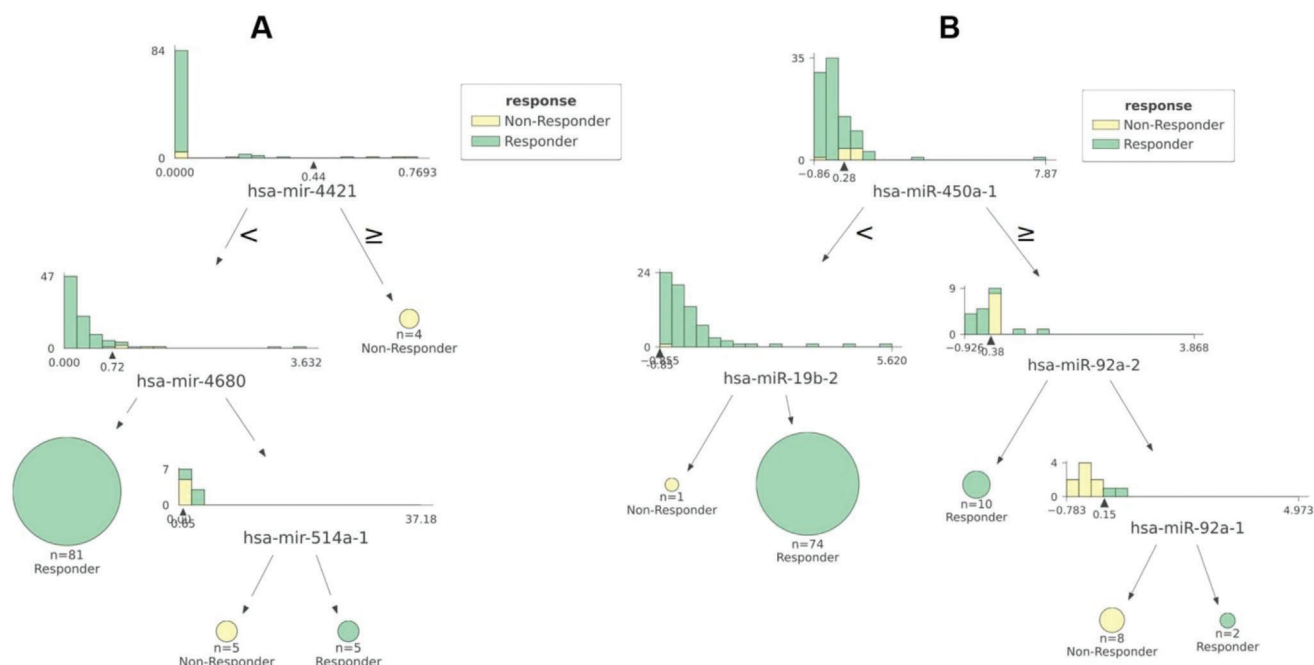


Figure 7. Interpreting the most predictive models based on miRNA and isomiR expression data. A) CART with the 3 selected miRNA features (hsa-miR-4421, hsa-miR-4680, and hsa-miR-514a-1) trained on all 95 patients with miRNA profile. The histogram in each tree node shows the distribution of patients at that node against the feature employed to split the patients into nonresponders and responders. The triangle under each histogram indicates the value of the best split for that feature, whose name can be found beneath the histogram. Each node has two leaves: to the right (patients with a feature value greater than or equal to the best split) and the left (the rest of the patients). Terminal nodes appear as circles and the decision path to get to each of them constitute an explanation of why a patient has been assigned the associated response class. For instance, there are two molecular subtypes in miRNA space associated to doxorubicin resistance: i) patients whose tumors express hsa-miR-4421 ≥ 0.44 are predicted to be nonresponders (4 out of the 9 nonresponders have tumors verifying this decision rule in the trained model), and ii) patients whose tumors express hsa-miR-4421 < 0.44 , hsa-miR-4680 ≥ 0.72 and hsa-miR-514a-1 < 0.65 are also predicted to be nonresponders (the remaining 5 nonresponders follow this decision rule). Otherwise, the patient is predicted to be a responder. B) CART with the 4 selected isomiR features (hsa-miR-450a-1, hsa-miR-19b-2, hsa-miR-92a-2, and hsa-miR-92a-1) trained on all 95 patients. In the isomiR space, there are also two molecular subtypes associated to doxorubicin resistance: i) patients with tumors expressing hsa-miR-450a-1 ≥ 0.28 , hsa-miR-92a-2 ≥ -0.38 , and hsa-miR-92a-1 < 0.15 are predicted to be nonresponders (8 out of the 9 nonresponders have tumors verifying this decision rule in the trained model), and ii) patients whose tumors express hsa-miR-450a-1 < 0.28 , and hsa-miR-19b-2 < -0.85 are also predicted to be nonresponders (the remaining nonresponder follow this decision rule). Conversely, the patient is predicted responder.

Table 1. 23 BC-driving genes are likely to be targeted by the 7 predictive miRNAs.

miRNA	TargetScan	TargetScan $\cap$ IntOGen	TargetScan $\cap$ COSMIC	TargetScan $\cap$ COSMIC $\cap$ IntOGen
hsa-mir-4421	3284	22	11	7
hsa-miR-4680-3p	4870	43	16	9
hsa-miR-4680-5p	3864	24	5	4
hsa-miR-514a-3p	3520	28	9	7
hsa-miR-514a-5p	3851	31	12	10
hsa-miR-450a-1-3p	6030	34	14	9
hsa-miR-19-5p	4906	37	12	8
hsa-miR-92a-2-5p	5873	36	16	10
hsa-miR-92a-1-5p	3287	26	7	4
# of unique genes	13 898	89	37	23

Analysys Toolkit (WebGestalt).^[67] ORA was used to identify the biological pathways significantly enriched with the 13898 targeted genes from the cancer-related repository of WikiPathways (WikiPathway cancer). The four cancer-associated pathways that were found (false discovery rate (FDR) ≤ 0.05) are DNA dam-

age response, ErbB (epidermal growth factor receptor family) signaling pathway, endometrial cancer, and chromosomal and microsatellite instability in colorectal cancer (Table S5, Supporting Information). Figure S5A of the Supporting Information complements this information with the gene ontology (GO) terms,

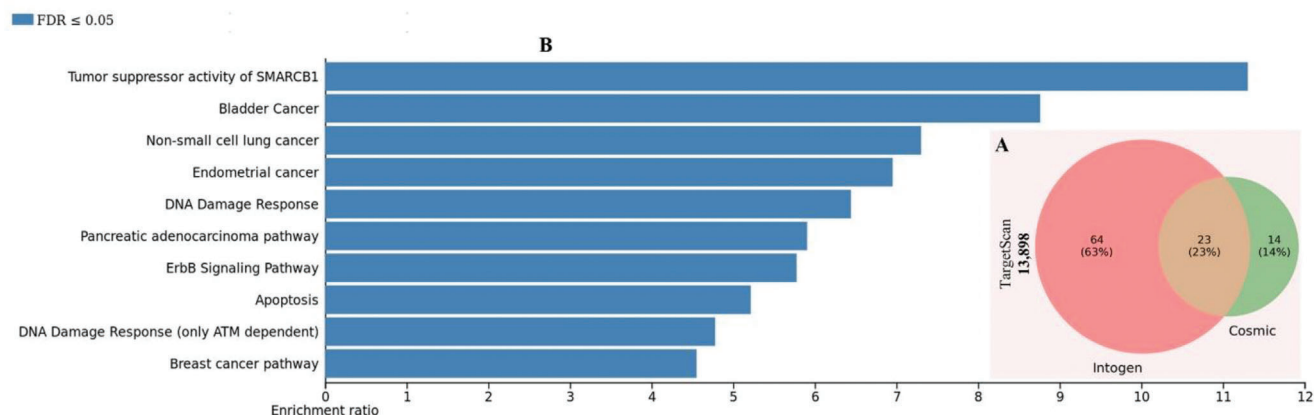


Figure 8. A) Venn diagram presenting the number and percentage of genes TargetScan-mapped to be targeted by the 7 predictive miRNAs in IntOGen and COSMIC databases (23 of these genes overlapped between the three gene lists). B) EA of the 23 overlapping target genes using WikiPathway cancer. The bar chart summarizes 10 biological pathways that are significantly ($FDR < 0.05$) enriched with different subsets of the 23 selected target genes for the predictive miRNAs. The bars represent the enrichment ratio (ER, as the number of gene overlaps over its expected value) obtained by comparing the gene list in our study with the reference gene set from genome protein-coding. FDA: false discovery rate.

including biological process, cellular component, and molecular function, that are enriched with these 13898 genes. Pathways associated with doxorubicin's mechanism of action such as DNA damage response^[68,69] and cell signaling,^[70] were found to be enriched. This suggests that dysregulation of the predicted target genes involved in these pathways could promote doxorubicin resistance. Also, alterations in ErbB signaling pathways could exacerbate breast tumorigenesis,^[71] which would contribute to doxorubicin seeming less effective, and have been linked to doxorubicin resistance.^[72]

Aiming at revealing the most predominant BC-specific biological processes underpinning primary resistance to doxorubicin, EA was conducted for the 23 genes obtained from leveraging BC-driver knowledge consensus and miRNA target prediction (Table 1). ORA was now used to identify biological pathways enriched with these 23 genes from both repositories WikiPathway cancer and Kyoto Encyclopedia of Genes and Genomes (KEGG). This gene list was significantly ($FDR \leq 0.05$) enriched in 10 cancer-associated biological pathways each, as indicated in both WikiPathway cancer (Figure 8B; Table S6, Supporting Information) and KEGG (Figure S5C, Supporting Information), three of these pathways (breast cancer, pancreatic cancer, and endometrial cancer) are common to both WikiPathway cancer and KEGG and appeared enriched with the same sets of genes. Other pathways also emerged significantly enriched ($FDR \leq 0.05$) in WikiPathway of cancer, including tumor suppressor activity, DNA damage response, apoptosis, cell signaling pathway, bladder cancer, and nonsmall cell lung cancer pathway. Of note, 8 of these 23 genes (ERBB2, ESR1, AKT1, PIK3CA, RB1, TP53, BRCA2, BRCA1) specifically overlapped with the BC annotated gene set (Figure S6 and Table S6, Supporting Information) in both WikiPathway cancer and KEGG. The GO terms enriched with these 23 genes are presented in Figure S5B of the Supporting Information. Due to doxorubicin-mediated DNA damage, apoptotic cell death can be induced via the activation of tumor suppressors in the cell cycle control and apoptosis.^[73] However, defect in these regulators can lead to doxorubicin failing to induce cell cycle arrest and apoptosis.^[68,69] The 23 BC-specific genes reg-

ulate processes associated with doxorubicin's mechanism of action and the breast cancer pathways in WikiPathway cancer (Figure 8). These genes are also part of drug resistance KEGG pathways, such as second-ranked platinum drug resistance and third-ranked endocrine resistance pathways (Figure S5C, Supporting Information), which are not found in enrichment analysis of the 13898 genes (Table S5, Supporting Information).

3. Discussion and Conclusions

Many studies^[15,74] have focused on identifying mutation-based single-gene markers and/or gene expression signatures for precision oncology. With fast-growing clinical pharmaco-omic datasets being available in the public domain, ML has become a highly promising approach to discover how these molecular factors could collectively explain and predict drug response.^[14,53,54,75]

Owing to the wealth of curated profiling data from the GDC, we could carry out an unusually broad analysis covering eight tumor profiles per patient. In line with previous studies analyzing other drugs,^[52,53] this large-scale analysis across multiple profiles and algorithms has resulted in the identification of the first ML models able to predict patient response to doxorubicin-containing treatments (CART allied with either miRNAs or their isoforms). Owing to CART's embedded feature selection, these CART models only employ either three or four features of the about 1–2 thousand considered. Another sign of the importance of feature selection when training on high-dimensional datasets is that OMC models were generally more predictive than their all-features counterparts. Overall, only 2 of the 128 models in this large-scale analysis were substantially predictive (Figure 2), showing that they can be easily missed by a standard-scale analysis. Interestingly, the fact that the best models are nonlinear, despite testing linear models too, hints the nonlinearity of the data.

Merging features of potentially complementary nature yielded mixed results (Figure 4). While merging clinical and miRNA features improved the accuracy of miRNA features alone, the difference was not significant and this improvement was not observed with isomiRs. Furthermore, clinical features alone were barely

predictive and combining them with all the molecular profiles led to an almost complete loss of predictive accuracy. We attribute the latter to CART not being able to cope with the far higher number of features involved (over half a million). In fact, higher dimensionality is detrimental well before 3000 features (i.e., merging miRNAs and their isoforms, as this results in CART models with lower MCC values than any of the two profiles in isolation). Thus, as no improvement is achieved by combining the most predictive profiles, determining multiple profiles per tumor is not recommended due to also being much more expensive and time-consuming.

The predictive accuracy of the isomiR-based CART model is high in the context of this problem despite the strong class imbalance. First, the mMCC of its tenfold CV predictions across five independent repetitions is 0.56, which corresponds to a mAUC of 0.80. Importantly, while all models with a substantial mMCC also have a substantial mAUC, the opposite is not true. For instance, the miRNA-based LGBM model returned an mAUC of 0.71, but its mMCC is practically zero. In class-imbalanced problems like this, we recommend MCC as a more appropriate alternative to AUC, as the latter has here overestimated the ability of some models to discriminate between responders and nonresponders. Second, the best ML models are much more predictive than their respective permuted versions (Figure 3). Third, HER2 expression was identified as a doxorubicin response marker, as also found by other studies supporting the use of this single-gene marker,^[22,60,61] but its mMCC is just 0.14 (Figure 5), which is four times lower than the mMCC of the best ML model. Lastly, our best predictive model with mMCC of 0.56 compares well to the very few existing *in vivo* treatment response ML models for other drugs and cancer types (MCC ranging from 0.36 to 0.54^[24,76,77]).

To go beyond these retrospective validations, we performed clustering on an additional 1078 BC patients profiled for both miRNAs and isomiRs (Figure 6). Here we showed that, in either of these profiles, the 95 doxorubicin-treated patients in the training set were well represented in all the clusters in which the 1078 BC patients are partitioned. This means that these patients are within the applicability domain of these models, which are therefore expected to have similarly high accuracy when predicting their response to doxorubicin. Furthermore, the predictive accuracies of these models were robust to using different data partitions and algorithm initializations (Figure 3). This further supports that similarly small fluctuations in accuracy should be observed on other BC patients.

An additional advantage of the identified models is that they are interpretable at the patient level. For example, a patient is predicted to be a nonresponder because her/his expression levels for the 3 selected miRNAs, hsa-miR-4421, hsa-miR-4680, and hsa-miR-514a-1, are nonlinearly combined in a way that has been accurately associated to nonresponders by the CART algorithm (Figure 7A). This 3-miRNA model is further supported by the relevant individual roles of each of their constituting miRNAs. For instance, hsa-miR-4421 and hsa-miR-4680 are overexpressed in luminal A BC^[78] and hereditary BRCA2 BC,^[79] respectively. Also, hsa-miR-514a-1 negatively correlates with BC recurrence.^[80] On the other hand, the second CART model presents predictive combinations of 4 isomiRs: hsa-miR-450a-1, hsa-miR-19b-2, hsa-miR-92a-2, and hsa-miR-92a-1. This model is also supported by the relevant individual roles of these four molecules. hsa-miR-

450a-1 is upregulated in BC.^[81] Also, overexpression of hsa-miR-19b family members are candidate prognostic biomarkers of BC and their involvement in the tumor progression through the PI3K/AKT pathway has also been documented.^[82,83] Downregulation of hsa-miR-92a family members is also associated with aggressive BC and high tumor macrophage infiltration.^[84] Hsa-miR-92a family members are also involved in the formation of blood vessels and the development of some mammalian organs,^[85] with their aberrant expressions being highly associated with different malignant human tumors.^[86–88] Therefore, they have been reported as a potential therapeutic target and novel diagnostic biomarker of human tumors.^[89] In addition, hsa-miR-92a is involved in apoptosis, cell proliferation, and doxorubicin chemosensitivity in gastric carcinoma cells, with the suppression of hsa-miR-92a leading to DNA damage foci and thus sensitivity to doxorubicin treatment.^[90]

In addition to doxorubicin response, individual miRNAs are often associated to other cancer patient outcomes. They play significant roles in RNA silencing as well as post-transcriptional regulation of gene expression.^[91,92] As miRNAs play critical roles in gene regulation and drug resistance, their dysregulation could promote cancer development, recurrence, and chemoresistance.^[93,94] MiRNAs can be used as tools or targets for the treatment of different cancers^[95] because of their essential roles as gene regulators for various human cancers.^[96] Hundreds of genes, including oncogenes and tumor suppressor genes, can be regulated by a single miRNA binding to their mRNA transcripts.^[97,98] Conversely, a single mRNA transcript can bind different miRNAs. That is, one single miRNA usually targets many genes and different miRNAs might regulate the same gene.^[91]

To reveal the most predominant BC-specific biological processes underpinning primary resistance to doxorubicin, EA for the 23 predicted BC-associated genes targeted by our predicted miRNAs revealed ten significantly (FDR < 0.05) enriched cancer-associated biological pathways (Figure 8B), including tumor suppressor activity, apoptosis, DNA damage response, bladder cancer, nonsmall cell lung cancer, endometrial cancer, BC pathway among others (Table S6, Supporting Information). Eight of these genes were enriched in the BC pathway (Figure S6, Supporting Information), five in DNA damage response, five in apoptosis, and so on (Table S6, Supporting Information). Details of the enriched genes targeted by each miRNA are reported in Table S8 of the Supporting Information. Deregulation of any of the components of the identified pathways listed above leads to aberrant expression of several associated genes, consequently resulting in different disorders or diseases, including cancer.^[99] For example, the deregulation of our predicted target genes in DNA damage and apoptotic pathways could be a mechanism that promotes patients resistance to doxorubicin, whose mechanism of action include inducing DNA damage to trigger apoptotic cell death.^[7] This EA provides starting points for studies investigating the molecular mechanisms of primary resistance to doxorubicin in BC patients.

Our study has some limitations to point out. While clustering analysis has shown that the best ML models are likely predictive on an additional cohort with over a thousand BC patients whose tumors were profiled for miRNAs and isomiRs, this is still to be confirmed prospectively. In such prospective clinical trial,

each BC patient would only have to be profiled pretreatment for the 4 isomiRs selected by CART, which would represent a large saving in time and cost with respect to determining the full profile. Another limitation is that the TCGA BRCA project collected treatment response data on a voluntary basis from hundreds of submitting institutions. Thus, the level of curation and harmonization of these datasets is likely to be much lower than that of the molecular profiles. With that said, there is a strong signal in treatment response data, at least for the 96 patients used for training, given that models trained on permuted response data were significantly less predictive.

Overall, this large-scale ML analysis has led to the discovery of highly predictive, robust and even interpretable predictors of BC patient response to doxorubicin. These are CART models nonlinearly combining selected miRNAs and their isoforms in a predictive manner (EA of the genes potentially regulated by these molecules provide starting points for mechanistic studies). These CART models achieved median MCC values that are at least four times higher than those based on HER2 expression and response-permuted data. Importantly, the MCC of these models increased with larger training sets, therefore these should become even more predictive as more data are available in the future.

4. Experimental Section

Acquisition, Retrieval, and Preprocessing of Drug Response Data: The open-access clinical and biospecimen files of the primary tumor samples in the TCGA-BRCA project were downloaded from the GDC portal (version 25.0, July 22nd, 2020 release). Note that this TCGA acronym stands for Breast Invasive Carcinoma, not the BRCA gene. To avoid confusion, BC was employed to refer to this cancer type, for which TCGA contains data for 1098 patients. To curate these datasets, misspellings, formulations, and synonyms (e.g., adriamycin, doxil, doxorubicinum, liposomal doxorubicin, and doxorubicin liposome) of the drug as annotated in the Drug-Bank database^[100] to its generic name (doxorubicin), were first standardized. Next, the 332 BC patients who received doxorubicin were identified. Among these patients, those without annotated doxorubicin responses, those missing sample collection and treatment start dates were excluded. Thus, 236 (71.1%) of patients were excluded from the study for these reasons. After these filtering steps, 96 BC patients treated with doxorubicin were retained (these patients did not receive chemotherapy prior to tumor sampling, as indicated by the times of tumor sample procurement and the start of treatment). 78 (81.3%) of whom had a tumor biopsy taken, while the remaining 18 (18.7%) had their tumors surgically resected instead. Patient responses to doxorubicin-containing treatments, complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD), were provided by the TCGA-BRCA project. As it is common practice,^[26,101] such responses were further categorized into two classes: responder (CR or PR) and nonresponder (SD or PD). This process resulted in a total of 87 responders and 9 nonresponders (Table S1, Supporting Information).

Acquisition, Retrieval, and Preprocessing of Molecular Profiling Data: The open-access molecular profiles for the 1098 TCGA-BRCA patients were also downloaded from the GDC portal. This was restricted to baseline primary tumor samples (i.e., those annotated with “01” in the 14th and 15th positions of the TCGA sample code). A few patients had multiple tumor samples sequenced. In this case, the first labeled sample was chosen, which corresponds to the A-level sample (e.g., TCGA-02-0001-01A). To reduce technical variability, the primary analysis of profiling data was harmonized by the GDC (i.e., carried out according to the same standardized GDC workflows). All the nonrestricted-access profiles were considered for these patients. Thus, eight molecular profiles per patient were considered: mRNA(FPKM) are the messenger RNAs as Fragment Per Kilobase of tran-

script per Million, mRNA(FPKM-UQ) are the messenger RNAs as Upper-Quantile-normalized FPKM, miRNA are the log2-transformed Reads Per Million mapped-normalized microRNAs, isomiR are the log2-transformed Reads Per Million mapped-normalized miRNA isoforms expression at a given locus that are distinguished by their location within the locus, CpG are the DNA methylation beta values of 450k probe at known CpG sites, CGI are the averaged beta values of all the probes at the CpG sites of the corresponding CpG Island, CNV(mean) are the Copy Number Variants calculated using CNTools R packages as the mean of DNA copy number across the segments of the considered gene, and CNV(median) is calculated in the same as CNV(mean) except using the median instead of the mean. Each molecular profiling dataset was used as a set of features for building models using a range of ML algorithms.

Preparing Datasets for ML: Only some doxorubicin-treated BC patients have both treatment responses and molecularly profiles annotated. The number of doxorubicin treated BC patients along with the number of features of each dataset was reported in Figure S1 of the Supporting Information. Each individual dataset was split into training and testing set using stratified K-fold^[102] CV, with K values being 3, 5, and 10. The predictions from each left-out CV fold were merged prior to calculating a given evaluation metric (e.g., MCC), instead of calculating the metric for each left-out fold and average them. This provides a more robust estimation of the metric, while ensuring that the prediction of each instance from the CV was not used in any way for the training or selection of its corresponding model.

The median results from the 5 seeds are reported for each of the 128 ML models in Figure 1 (16 ML models $\times$ 8 molecular profiles). Because of the class imbalance of the processed datasets, the class weights were applied inversely proportional to class frequencies during model fitting. The process was repeated five times with different random seeds. The median of performance metrics of five repetitions was reported. All analyses were performed using python package version 3.7.3 (<https://www.python.org/>) with packages from scikit-learn (<https://scikit-learn.org/>).

Building All-Features Classification Models: 8 ML algorithms were employed: CART,^[103] random forest,^[104] XGBoost,^[105] LGBM,^[106] Logistic regression,^[107] Linear Support Vector Machine and Radius Support Vector Machine,^[108] and K-Nearest Neighbors.^[109] The model hyperparameters were set to their default values.

Standard stratified K-fold CVs were performed to measure the model performance for each algorithm-profile pair using all the available features for that profile. This is called an all-features model (however, as some algorithms such as CART possess embedded feature selection, the resulting model will only employ a fraction of the features to calculate its predictions). During CV, K-1 folds were used for model training, whereas the remaining partition was used to test the trained model. Thus, each fold was used exactly once as a test set. Therefore, for any given model, the response of a patient was predicted using a model trained with data from other patients.

Building OMC Classification Models: Due to the high dimensionality of the datasets (i.e., the number of features is much larger than the number of patients), OMC^[24,53] variants of each of the 8 ML algorithms were implemented to build models using only the most relevant features of the doxorubicin response. This can remove noise from the training data and improve the model performance. To select models with the OMC while estimating their performance, nested CV was carried out for each algorithm-profile pair. In brief, OMC is comprised of three steps. First, features were ranked according to their relevance to doxorubicin response by increasing *p*-values from Analysis of Variance (ANOVA). The *p*-value, one per each feature, indicates the discriminative power to distinguish between responders and nonresponders, the informative features associated with small *p*-values. Then, an ML model was trained with considered subset of features (the top 2 to $n/2$ subset of features, where n is the number of samples). Finally, the best-performing model was selected among all $n/2$ trained model as the one with the highest MCC in the inner loop of the nested CV and evaluated that model performance in its outer loop.

Comparing the Best ML Models to Permutation and HER2 Expression Models: As a baseline, CVs for a top model was also run with the same algorithm and features, but after randomly shuffling doxorubicin response labels across patients. This was called a permutation model.

On the other hand, HER2 status has been identified as a single-gene marker for BC patient sensitivity to doxorubicin in several studies.^[22,60,61] To compare the accuracy of the ML models at this task with that of using HER2 only, an HER2-CART model was built using standard tenfold CV with HER2(ERBB2) expression data derived from the mRNA(FPKM) profile.

Model Performance Evaluation: For MCC, as it is customary, the operating threshold was set to 0.5. Patients with class probabilities above this threshold were predicted to be responders, otherwise they were predicted to be nonresponders. To estimate predictive performance, the true and predicted classes were compared. The numbers of true positive, true negative, false positive, and false negative instances were used to calculate performance metrics. The AUC of the ROC, AUC for short, was also calculated for comparison purposes.

Clustering and Pathway Analysis of BC Patients from Selected miRNA Features: BC patients were clustered using the predictive miRNA features (either miRNAs or isomiRs) with an agglomerative clustering algorithm (Ward's linkage with Euclidean distance to calculate similarity between the clusters of patients). A dendrogram heatmap was used to visualize the results of this clustering analysis, marking the 95 patients with high-quality response and profiling data. Both clustering and its visualization were carried out with clustermap in the seaborn package (version 0.11.2).

Data Processing for EA: To reveal the biological pathways related to predictive miRNAs, the target genes of predictive miRNAs were predicted using the TargetScan database (version 7.0).^[63] This online miRNA target prediction tool identify mRNA by matching the desired miRNA seed region to the conserved complementary sites (sequences) of their mRNA targets.^[110–113] It retrieves the highest number of target genes of any database and ranks the predicted targets of each miRNA based on these matching algorithms.^[63] Then, the list of BC-driving genes was download from IntOGen,^[64] and COSMIC.^[66] Of those, the subset of genes in common of TargetScan, IntOGen, and COSMIC were considered as potential BC miRNA targets. They were used as input data of EA to explore the biological relevance of predictive miRNAs. WEB-based Gene Set Analysis Toolkit (WebGestalt)^[67] (www.webgestalt.org) was utilized for EA for WikiPathway cancer, KEGG, and GO database. WebGestalt uses a hypergeometric test for statistical significance, which further employed the Benjamin–Hochberg FDR multiple-testing correction. The enriched pathways were identified based on FDR threshold which was set to 0.05. For each pathway, input genes that are part of the pathway are counted and enrichment ratio was also calculated.

Statistical Analysis: Preprocessing of data: See subsections entitled “Acquisition, Retrieval, and Preprocessing of Drug Response Data”, “Acquisition, Retrieval, and Preprocessing of Molecular Profiling Data”, and “Preparing Datasets for ML” at the start this section.

Data presentation: Whenever relevant, Figures 2–8 present the distributions of performance metrics (e.g., Figure 3). On the other hand, Table S1 of the Supporting Information presents all the clinical data.

Sample sizes for each statistical analysis: Each sample is formed by 5 runs whose results are summarized by boxplot (Figures 3–5). Each dot represents the model performance from that run (each using a different random seed). The provenance of each sample is specified in the figure captions.

Statistical methods: A two-tailed Welch's *t*-tests for each considered pair of samples. A performance difference was considered to be significant if *p*-value < 0.05.

Software used for statistical analysis: Statistical analysis was carried out using the statannot package in the python 3.7.3 software.

Code availability: The python codes and processed datasets are provided to build the best classifiers and the HER2 baseline, evaluate them and facilitate their application to other cohorts of miRNA-profiled BC patients: https://github.com/adeolu1/BRCA_CART_model.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by grant funding from the Indo-French Centre for the Promotion of Advanced Research – CEFIPRA and the Petroleum Technology Development Fund (PTDF), Nigeria.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

P.J.B. conceived the idea and designed the experiments. A.Z.O. collected and analyzed the patients' data, developed the ML process and interpreted the results. A.Z.O. and P.J.B. wrote the manuscript with the assistance of C. Piyawajanusorn, A. Gonçalves, and G. Ghislat. All authors contributed to the discussion.

Data Availability Statement

The data and the corresponding cancer information were downloaded from the Genomic Data Commons portal (<https://portal.gdc.cancer.gov/>) and were, in whole, based upon open access data generated from the TCGA-BRCA project. Thus, these data were publicly available without restriction, authentication, or authorization.

Keywords

artificial intelligence, machine learning, multiomics, precision oncology, tumor profiling

Received: March 15, 2022

Revised: June 2, 2022

Published online: July 3, 2022

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, *Ca-Cancer J. Clin.* **2021**, *71*, 209.
- [2] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, D. Forman, *Ca-Cancer J. Clin.* **2011**, *61*, 69.
- [3] Y. L. Franco, T. R. Vaidya, S. Ait-Oudhia, *Breast Cancer: Targets Ther.* **2018**, *10*, 131.
- [4] R. Paridaens, L. Biganzoli, P. Bruning, J. G. M. Klijn, T. Gamucci, S. Houston, R. Coleman, J. Schachter, A. Van Vreckem, R. Sylvester, A. Awada, J. Wildiers, M. Piccart, *J. Clin. Oncol.* **2000**, *18*, 724.
- [5] S. M. Cutts, A. Nudelman, A. Rephaeli, D. R. Phillips, *IUBMB Life* **2005**, *57*, 73.
- [6] F. Yang, S. S. Teves, C. J. Kemp, S. Henikoff, *Biochim. Biophys. Acta, Rev. Cancer* **2014**, *1845*, 84.
- [7] O. Tacar, P. Sriamornsak, C. R. Dass, *J. Pharm. Pharmacol.* **2013**, *65*, 157.
- [8] R. Thirumaran, G. C. Prendergast, P. B. Gilman, *Cancer Immunother.* **2007**, 101.
- [9] R. Brown, U. Böger-Brown, *Cytotoxic Drug Resistance Mechanisms*, Humana Press, New Jersey **1999**.
- [10] F. Cardoso, A. Di Leo, C. Lohrisch, C. Bernard, F. Ferreira, M. J. Piccart, *Ann. Oncol.* **2002**, *13*, 197.
- [11] G. Housman, S. Byler, S. Heerboth, K. Lapinska, M. Longacre, N. Snyder, S. Sarkar, *Cancers* **2014**, *6*, 1769.
- [12] H. S. D. C. Ribeiro, P. R. Stevanato-Filho, W. L. Da Costa, A. L. Diniz, P. Herman, F. J. F. Coimbra, *Arq. Gastroenterol.* **2012**, *49*, 266.

- [13] T. R. J. Evans, A. Yellowlees, E. Foster, H. Earl, D. A. Cameron, A. W. Hutcheon, R. E. Coleman, T. Perren, C. J. Gallagher, M. Quigley, J. Crown, A. L. Jones, M. Highley, R. C. F. Leonard, J. L. Mansi, *J. Clin. Oncol.* **2005**, *23*, 2988.
- [14] P. J. Ballester, J. Carmona, *npj Precis. Oncol.* **2021**, *5*, 1.
- [15] M. Huang, A. Shen, J. Ding, M. Geng, *Trends Pharmacol. Sci.* **2014**, *35*, 41.
- [16] C. F. Thorn, C. Oshiro, S. Marsh, T. Hernandez-Boussard, H. McLeod, T. E. Klein, R. B. Altman, *Pharmacogenet. Genomics* **2011**, *21*, 440.
- [17] L. Zhao, B. Zhang, *Sci. Rep.* **2017**, *7*, 44735.
- [18] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Das-tur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, et al., *Nature* **2012**, *483*, 570.
- [19] C. Dang, A. Peón, P. J. Ballester, *BMC Med. Genomics* **2018**, *11*, 10.
- [20] C. Piyaawanusorn, L. C. Nguyen, G. Ghislis, P. J. Ballester, *Briefings Bioinf.* **2021**, *22*, 312.
- [21] S. M. Corsello, R. T. Nagari, R. D. Spangler, J. Rossen, M. Kocak, J. G. Bryan, R. Humeidi, D. Peck, X. Wu, A. A. Tang, V. M. Wang, S. A. Bender, E. Lemire, R. Narayan, P. Montgomery, U. Ben-David, C. W. Garvie, Y. Chen, M. G. Rees, N. J. Lyons, J. M. McFarland, B. T. Wong, L. Wang, N. Dumont, P. J. O'Hearn, E. Stefan, J. G. Doench, C. N. Harrington, H. Greulich, et al., *Nat. Cancer* **2020**, *1*, 235.
- [22] A. Rody, T. Karn, R. Gätje, A. Ahr, C. Solbach, K. Kourtis, M. Munnes, S. Loibl, S. Kissler, E. Ruckhäberle, U. Holtrich, G. von Minckwitz, M. Kaufmann, *Breast* **2007**, *16*, 86.
- [23] M.-S. Tsao, A. Sakurada, J.-C. Cutz, C.-Q. Zhu, S. Kamel-Reid, J. Squire, I. Lorimer, T. Zhang, N. Liu, M. Daneshmand, P. Marrano, G. da Cunha Santos, A. Lagarde, F. Richardson, L. Seymour, M. White-head, K. Ding, J. Pater, F. A. Shepherd, *N. Engl. J. Med.* **2005**, *353*, 133.
- [24] L. C. Nguyen, S. Naulaerts, A. Bruna, G. Ghislis, P. J. Ballester, *Biomedicine* **2021**, *9*, 1319.
- [25] J. J. Li, X. Tong, *Patterns* **2020**, *1*, 100115.
- [26] T. Saito, M. Rehmsmeier, *PLoS One* **2015**, *10*, 0118432.
- [27] C. Cybulski, B. Gliniewicz, A. Sikorski, J. Kładny, T. Huzarski, J. Gron-wald, T. Byrski, T. Dębniak, B. Gorski, A. Jakubowska, D. Wokolor-czyk, S. A. Narod, J. Lubiński, *Cancer Epidemiol., Biomarkers Prev.* **2007**, *16*, 572.
- [28] B. Weigelt, J. S. Reis-Filho, *J. Pathol.* **2014**, *232*, 255.
- [29] M. A. A. Koike Folgueda, D. M. Carraro, H. Brentani, D. F. Da Costa Patrão, E. Mantovani Barbosa, M. Mourão Netto, J. R. Figaro Caldeira, M. L. Hirata Katayama, F. A. Soares, C. Tosello Oliveira, L. F. Lima Reis, J. H. Lima Kaiano, L. P. Camargo, R. Z. Nicolliello Vêncio, I. M. Longo Snitcovsky, F. B. Alves Makdissi, P. J. Da Silva E Silva, J. C. G. Sampaio Góes, M. M. Brentani, *Clin. Cancer Res.* **2005**, *11*, 7434.
- [30] Y. Chen, H. Cai, W. Chen, Q. Guan, J. He, Z. Guo, J. Li, *Front. Mol. Biosci.* **2020**, *7*.
- [31] P. J. Ballester, R. Stevens, B. Haibe-Kains, R. S. Huang, T. Aittokallio, *Briefings Bioinf.* **2021**, *23*, 450.
- [32] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, J. Saez-Rodriguez, *PLoS One* **2013**, *8*, 61318.
- [33] M. Ammad-ud-din, S. A. Khan, K. Wennerberg, T. Aittokallio, *Bioin-formatics* **2017**, *33*, i359.
- [34] L. Parca, G. Pepe, M. Pietrosanto, G. Galvan, L. Galli, A. Palmeri, M. Sciarone, F. Ferrè, G. Ausiello, M. Helmer-Citterich, *Sci. Rep.* **2019**, *9*, 15222.
- [35] S. Fasola, G. Cilluffo, L. Montalbano, V. Malizia, G. Ferrante, S. La Grutta, *Genes* **2021**, *12*, 1.
- [36] L. He, J. Tang, E. I. Andersson, S. Timonen, S. Koschmieder, K. Wen-nerberg, S. Mustjoki, T. Aittokallio, *Cancer Res.* **2018**, *78*, 2407.
- [37] B. Majumder, U. Baraneedharan, S. Thiyagarajan, P. Radhakrish-nan, H. Narasimhan, M. Dhandapani, N. Brijwani, D. D. Pinto, A. Prasath, B. U. Shanthappa, A. Thayakumar, R. Surendran, G. K. Babu, A. M. Shenoy, M. A. Kuriakose, G. Berghthold, P. Horowitz, M. Loda, R. Beroukhir, S. Agarwal, S. Sengupta, M. Sundaram, P. K. Majumder, *Nat. Commun.* **2015**, *6*, 6169.
- [38] R. Kurilov, B. Haibe-Kains, B. Brors, *Sci. Rep.* **2020**, *10*, 1.
- [39] S. Naulaerts, C. C. Dang, P. J. Ballester, *Oncotarget* **2017**, *8*, 97025.
- [40] L. Nguyen, C. C. Dang, P. J. Ballester, *F1000Research* **2017**, *5*, 2927.
- [41] J. A. Williams, *J. Clin. Med.* **2018**, *7*, 41.
- [42] R. Peres Da Silva, C. Suphavitai, N. Nagarajan, *Bioinformatics* **2021**, *37*, i76.
- [43] E. Izumchenko, K. Paz, D. Ciznadija, I. Sloma, A. Katz, D. Vasquez-Dunddel, I. Ben-Zvi, J. Stebbing, W. McGuire, W. Harris, R. Maki, A. Gaya, A. Bedi, S. Zacharoulis, R. Ravi, L. H. Wexler, M. O. Hoque, C. Rodriguez-Galindo, H. Pass, N. Peled, A. Davies, R. Morris, M. Hidalgo, D. Sidransky, *Ann. Oncol.* **2017**, *28*, 2595.
- [44] S. N. Dorman, K. Baranova, J. H. M. Knoll, B. L. Urquhart, G. Mari-ani, M. L. Carcangiu, P. K. Rogan, *Mol. Oncol.* **2016**, *10*, 85.
- [45] P. Geleher, Z. Zhang, F. Wang, R. F. Gruener, A. Nath, G. Morrison, S. Bhutra, R. L. Grossman, R. S. Huang, *Genome Res.* **2017**, *27*, 1743.
- [46] S. Sinha, R. Vegesna, S. Rahman Dhruba, W. Wu, D. Lucas Kerr, O. V. Stroganov, I. Grishagin, K. D. Aldape, C. M. Blakely, P. Jiang, C. J. Thomas, T. G. Bivona, A. A. Schäffer, E. Rupp, *bioRxiv* **2022**.
- [47] B. Györfy, V. Serra, K. Jürchott, R. Abdul-Ghani, M. Garber, U. Stein, I. Petersen, H. Lage, M. Dietel, R. Schäfer, *Oncogene* **2005**, *24*, 7542.
- [48] S. P. Miranda, F. A. Baião, P. M. Maçaira, J. L. Fleck, S. R. Piccolo, *bioRxiv* **2020**.
- [49] C. Willyard, *Nature* **2018**, *560*, 156.
- [50] S. Y. Cindy Yang, S. C. Lien, B. X. Wang, D. L. Clouthier, Y. Hanna, I. Cirlan, K. Zhu, J. P. Bruce, S. El Ghamrasni, M. A. J. Iafolla, M. Oliva, A. R. Hansen, A. Spreafico, P. L. Bedard, S. Lheureux, A. Razak, V. Speers, H. K. Berman, A. Aleshin, B. Haibe-Kains, D. G. Brooks, T. L. McGaha, M. O. Butler, S. V. Bratman, P. S. Ohashi, L. L. Siu, T. J. Pugh, *Nat. Commun.* **2021**, *12*, 1.
- [51] N. Schäfer, G. H. Gielen, L. Rauschenbach, S. Kebir, A. Till, R. Reinartz, M. Simon, P. Niehusmann, C. Kleinschnitz, U. Herrlinger, T. Pietsch, B. Scheffler, M. Glas, *J. Transl. Med.* **2019**, *17*, 96.
- [52] Z. Ding, S. Zu, J. Gu, *Bioinformatics* **2016**, *32*, 2891.
- [53] A. Bomane, A. Gonçalves, P. J. Ballester, *Front. Genet.* **2019**, *10*, 1041.
- [54] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, N. Waddell, *Genome Med.* **2021**, *13*, 1.
- [55] S. Naulaerts, M. P. Menden, P. J. Ballester, *Biomolecules* **2020**, *10*, 1.
- [56] J. N. Weinstein, E. a. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, *Nat. Genet.* **2013**, *45*, 1113.
- [57] F. Movahedi, R. Padman, J. F. Antaki, **2020**, 1.
- [58] C. G. Weng, J. Poon, *Conf. Res. Pract. Inf. Technol. Ser.* **2008**, *87*, 27.
- [59] D. Chicco, G. Jurman, *BMC Genomics* **2020**, *21*, 6.
- [60] A. Gennari, M. P. Sormani, P. Pronzato, M. Puntoni, M. Colozza, U. Pfeffer, P. Bruzzi, *J. Natl. Cancer Inst.* **2008**, *100*, 14.
- [61] J. Zhang, Y. Liu, *J. Zhejiang Univ., Sci., B* **2008**, *9*, 5.
- [62] I. G. Cannell, Y. W. Kong, M. Bushell, *Biochem. Soc. Trans.* **2008**, *36*, 1224.
- [63] V. Agarwal, G. W. Bell, J. W. Nam, D. P. Bartel, *eLife* **2015**, *4*, 1.
- [64] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, N. Lopez-Bigas, *Nat. Rev. Cancer* **2020**, *20*, 555.
- [65] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Har-sha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, S. A. Forbes, *Nucleic Acids Res.* **2018**, *47*, D941.

- [66] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, S. A. Forbes, *Nat. Rev. Cancer* **2018**, 18, 696.
- [67] J. Wang, S. Vasaikar, Z. Shi, M. Greer, B. Zhang, *Nucleic Acids Res.* **2017**, 45, W130.
- [68] L. Y. Li, Y. D. Guan, X. S. Chen, J. M. Yang, Y. Cheng, *Front. Pharmacol.* **2021**, 11, 2520.
- [69] Z. Torki, D. Ghavi, S. Hashemi, Y. Rahmati, D. Rahmanpour, M. Pornour, M. R. Alivand, *Cancer Chemother. Pharmacol.* **2021**, 88, 771.
- [70] P. C. Marinello, C. Panis, T. N. X. Silva, R. Binato, E. Abdelhay, J. A. Rodrigues, A. L. Menciaha, N. M. D. Lopes, R. C. Luiz, R. Cecchini, A. L. Cecchini, *Sci. Rep.* **2019**, 9, 1.
- [71] H. Masuda, D. Zhang, C. Bartholomeusz, H. Doihara, G. N. Hortobagyi, N. T. Ueno, *Breast Cancer Res. Treat.* **2012**, 136, 331.
- [72] T. A. H. Järvinen, M. Tanner, V. Rantanen, M. Bärlund, Å. Borg, S. Grénman, J. Isola, *Am. J. Pathol.* **2000**, 156, 839.
- [73] R. Lüpertz, W. Wätjen, R. Kahl, Y. Chovolou, *Toxicology* **2010**, 271, 115.
- [74] C. Rodríguez-Antona, M. Taron, *J. Intern. Med.* **2015**, 277, 201.
- [75] P. Johannot, N. Coudray, D. M. Donnelly, G. Jour, I. Illa-Bochaca, Y. Xia, D. B. Johnson, L. Wheless, J. R. Patrinely, S. Nomikou, D. L. Rimm, A. C. Pavlick, J. S. Weber, J. Zhong, A. Tsigos, I. Osman, *Clin. Cancer Res.* **2021**, 27, 131.
- [76] X. Xu, H. Gu, Y. Wang, J. Wang, P. Qin, *Front. Genet.* **2019**, 10, 233.
- [77] Y. Chen, R. Zhang, L. Wang, A. M. Correa, A. Pataer, Y. Xu, X. Zhang, C. Ren, S. Wu, Q. H. Meng, J. Fujimoto, V. B. Jensen, M. B. Antonoff, W. L. Hofstetter, R. J. Mehran, G. Pisimisis, D. C. Rice, B. Sepesi, A. A. Vaporciyan, G. L. Walsh, S. G. Swisher, J. A. Roth, J. V. Heymach, B. Fang, *Cancer* **2019**, 125, 3738.
- [78] N. M. Zamarró, *Ph.D. Thesis*, Autonomous University of Madrid, **2015**.
- [79] R. Murria Estal, S. Palanca Suela, I. De Juan Jiménez, C. Egoavil Rojas, Z. García-Casado, M. J. Juan Fita, A. B. Sánchez Heras, Á. Segura Huerta, I. Chirivella González, D. Sánchez-Izquierdo, M. Llop García, E. Barragán González, P. Bolufer Gilabert, *Breast Cancer Res. Treat.* **2013**, 142, 19.
- [80] J. Tang, W. Ma, Q. Zeng, J. Tan, K. Cao, L. Luo, *Dis. Markers* **2019**, 2019, 1.
- [81] R. Baffa, M. Fassan, S. Volinia, B. O'Hara, C. G. Liu, J. P. Palazzo, M. Gardiman, M. Rugge, L. G. Gomella, C. M. Croce, A. Rosenberg, *J. Pathol.* **2009**, 219, 214.
- [82] H. Li, X. Jin, B. Liu, P. Zhang, W. Chen, Q. Li, *BMC Cancer* **2019**, 19, 1.
- [83] L. Yang, W. Zhao, P. Wei, W. Zuo, S. Zhu, *Am. J. Transl. Res.* **2017**, 9, 683.
- [84] S. Nilsson, C. Möller, K. Jirstrom, A. Lee, S. Busch, R. Lamb, G. Landberg, *PLoS One* **2012**, 7, 36051.
- [85] A. Ventura, A. G. Young, M. M. Winslow, L. Lintault, A. Meissner, S. J. Erkeland, J. Newman, R. T. Bronson, D. Crowley, J. R. Stone, R. Jaenisch, P. A. Sharp, T. Jacks, *Cell* **2008**, 132, 875.
- [86] A. R. Elhamamsy, M. S. El Sharkawy, A. F. Zanaty, M. A. Mahrous, A. E. Mohamed, E. A. Abushaaban, *Int. J. Mol. Cell. Med.* **2017**, 6, 1.
- [87] H. Si, X. Sun, Y. Chen, Y. Cao, S. Chen, H. Wang, C. Hu, *J. Cancer Res. Clin. Oncol.* **2013**, 139, 223.
- [88] C. Zhu, C. Ren, J. Han, Y. Ding, J. Du, N. Dai, J. Dai, H. Ma, Z. Hu, H. Shen, Y. Xu, G. Jin, *Br. J. Cancer* **2014**, 110, 2291.
- [89] M. Jiang, X. Li, X. Quan, X. Li, B. Zhou, *Front. Mol. Biosci.* **2019**, 6, 1.
- [90] X.-C. Tao, X.-Y. Zhang, S.-B. Sun, D.-Q. Wu, *Oncol. Rep.* **2019**, 42, 313.
- [91] P. Xu, Q. Wu, J. Yu, Y. Rao, Z. Kou, G. Fang, X. Shi, W. Liu, H. Han, *Front. Genet.* **2020**, 11, 1.
- [92] D. P. Bartel, *Cell* **2018**, 173, 20.
- [93] W. Si, J. Shen, H. Zheng, W. Fan, *Clin. Epigenet.* **2019**, 11, 1.
- [94] I. Koturbash, W. H. Tolleson, L. Guo, D. Yu, S. Chen, H. Hong, W. Mattes, B. Ning, *Biomarkers Med.* **2015**, 9, 1153.
- [95] N. Hosseini, M. Aghapour, P. H. G. Duijff, B. Baradaran, *J. Cell. Physiol.* **2018**, 233, 5574.
- [96] X. Zhang, W. Li, *Mol. Med. Rep.* **2012**, 6, 303.
- [97] B. Zhang, E. J. Stellwag, X. Pan, *Gene* **2009**, 443, 100.
- [98] B. Zhang, X. Pan, G. P. Cobb, T. Anderson, *Dev. Biol.* **2007**, 302, 1.
- [99] A. Das, S. Bhattacharya, *BAOJ Bioinf.* **2017**, 1, 1.
- [100] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, 34, D668.
- [101] A. L. Tarca, V. J. Carey, X. wen Chen, R. Romero, S. Drăghici, *PLoS Comput. Biol.* **2007**, 3, e116.
- [102] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, 12, 2825.
- [103] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA **1984**.
- [104] L. Breiman, *Mach. Learn.* **2001**, 45, 5.
- [105] T. Chen, C. Guestrin, *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining – KDD '16*, ACM, **2016**, pp. 785–794.
- [106] G. Ke, Q. Meng, T. Wang, W. Chen, W. Ma, T.-Y. Liu, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, *Adv. Neural Inf. Process. Syst.* **2017**.
- [107] J. Tolles, W. J. Meurer, *JAMA, J. Am. Med. Assoc.* **2016**, 316, 533.
- [108] V. Vapnik, in *Nonlinear Modeling: Advanced Black-Box Techniques* (Eds: J. A. K. Suykens, J. Vandewalle), Springer, Boston, MA **1998**, pp. 55–85.
- [109] T. M. Cover, P. E. Hart, *IEEE Trans. Inf. Theory* **1967**, 13, 21.
- [110] G. Riolo, S. Cantara, C. Marzocchi, C. Ricci, *Methods Protoc.* **2020**, 4, 1.
- [111] M. T. Witkos, E. Koscińska, J. W. Krzyżosiak, *Curr. Mol. Med.* **2011**, 11, 93.
- [112] S. M. Peterson, J. A. Thompson, M. L. Ufkin, P. Sathyanarayana, L. Liaw, C. B. Congdon, *Front. Genet.* **2014**, 5.
- [113] B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **2005**, 120, 15.

Appendix III

RESEARCH ARTICLE

Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer

Adeolu Ogunleye^{1†}, Chayanit Piyawajanusorn^{2†}, Ghita Ghislat^{3*}, and Pedro J. Ballester^{2*}

¹Department of Organismal Biology, Uppsala University, Uppsala, Sweden. ²Department of Bioengineering, Imperial College London, London, UK. ³Department of Life Sciences, Imperial College London, London, UK.

*Address correspondence to: ghita.ghislat@imperial.ac.uk (G.G.); p.ballester@imperial.ac.uk (P.J.B.)

†Joint first authors.

Background: Gemcitabine is a first-line chemotherapy for pancreatic adenocarcinoma (PAAD), but many PAAD patients do not respond to gemcitabine-containing treatments. Being able to predict such nonresponders would hence permit the undelayed administration of more promising treatments while sparing gemcitabine life-threatening side effects for those patients. Unfortunately, the few predictors of PAAD patient response to this drug are weak, none of them exploiting yet the power of machine learning (ML). **Methods:** Here, we applied ML to predict the response of PAAD patients to gemcitabine from the molecular profiles of their tumors. More concretely, we collected diverse molecular profiles of PAAD patient tumors along with the corresponding clinical data (gemcitabine responses and clinical features) from the Genomic Data Commons resource. From systematically combining 8 tumor profiles with 16 classification algorithms, each of the resulting 128 ML models was evaluated by multiple 10-fold cross-validations. **Results:** Only 7 of these 128 models were predictive, which underlines the importance of carrying out such a large-scale analysis to avoid missing the most predictive models. These were here random forest using 4 selected mRNAs [0.44 Matthews correlation coefficient (MCC), 0.785 receiver operating characteristic–area under the curve (ROC-AUC)] and XGBoost combining 12 DNA methylation probes (0.32 MCC, 0.697 ROC-AUC). By contrast, the hENT1 marker obtained much worse random-level performance (practically 0 MCC, 0.5 ROC-AUC). Despite not being trained to predict prognosis (overall and progression-free survival), these ML models were also able to anticipate this patient outcome. **Conclusions:** We release these promising ML models so that they can be evaluated prospectively on other gemcitabine-treated PAAD patients.

Introduction

Pancreatic cancer (PC) is the seventh leading cause of death among all cancers (495,773 new cases and 466,003 deaths worldwide in 2020 [1]). Surgical resection remains the only potential therapeutic option for patients with early-stage PC [2]. However, PC is difficult to diagnose due to nonspecific initial symptoms. Consequently, most patients are at the advanced or metastatic stage when they are diagnosed, where the 5-year survival rate is about 3% [3]. In addition, PC also lacks effective early detection methods, predictive biomarkers, and risk prediction models.

While several target therapies, chemotherapies, and immunotherapies have been approved for PC, gemcitabine has been widely used as the standard first-line treatment for PC. Gemcitabine is administered in combination with other cytotoxic agents, such as nab-paclitaxel, carboplatin, cisplatin, or oxaliplatin. Such gemcitabine-containing treatments generally

offer enhanced patient survival and clinical benefit [4–10]. The anti-proliferative properties of gemcitabine come from inhibiting DNA synthesis [11]. Gemcitabine requires intracellular phosphorylation to be converted into its active form, difluoro-deoxycytidine triphosphate (dFdCTP), which competes with deoxycytidine triphosphate (dCTP) for incorporation into DNA during DNA elongation, resulting in inhibition of DNA synthesis and inducing apoptosis of cancer cells [12]. However, 54% of gemcitabine-treated PC patients can exhibit primary resistance [13]. It is therefore important to find an effective way to anticipate which patients will benefit from gemcitabine. In this way, a more promising drug treatment could be administered without delay while avoiding gemcitabine's adverse effects for patients who are unlikely to respond.

Several studies examined the gemcitabine-related biomarkers involved in its metabolism [14,15]. Most studies about predictive markers of gemcitabine treatment response focus on

Citation: Ogunleye A, Piyawajanusorn C, Ghislat G, Ballester PJ. Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer. *Health Data Sci.* 2024;4:Article 0108. <https://doi.org/10.34133/hds.0108>

Submitted 15 April 2023

Accepted 8 December 2023

Published 8 January 2024

Copyright © 2024 Adeolu Ogunleye et al. Exclusive licensee Peking University Health Science Center. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY 4.0).

human equilibrative nucleoside transporter 1 (hENT1) expression, with high hENT1 expression being reported to show clinical benefit in selected treatment settings and survival endpoint [16–19]. However, some studies did not observe these benefits [20,21]. This might be due to study differences such as the way hENT1 expression is measured, different clinical endpoints, or different gemcitabine combos. This study will propose promising alternatives for this marker.

Computational models exploiting multi-omics and/or clinico-pathological data to predict cancer patient outcomes have shown their value [22–27]. When the outcome is patient response to drugs, there is abundant preclinical pharmaco-omics data available [28]. These datasets have enabled drug response prediction via univariate markers [29,30] or gene expression signatures [31–33]. Machine learning (ML) algorithms, such as logistic regression, random forest, and deep neural network, have more recently been implemented to analyze such datasets and offer opportunities in personalized therapy to better understand mechanisms of drug resistance [34]. There are now a plethora of studies predicting drug response using ML trained on preclinical pharmaco-omics data [31,35–42]. Clinical pharmaco-omics datasets, despite being the most relevant for the patient, are much less available to train and evaluate ML models [43]. This is the reason for the small number [44–49].

ML techniques have been proven to be powerful tools to generate predictive computational models from the typically high-dimensional pharmaco-omics datasets. In particular, large-scale ML analysis, including algorithms enhancing supervised learning with feature selection such as optimal model complexity (OMC), has been shown to be advantageous [44,45,50]. To our knowledge, despite its promise, no study has yet applied a broad panel of ML algorithms to classify PC patients' response to gemcitabine using various molecular profiles and clinical datasets. With this purpose, we retrospectively collected molecular profiles of The Cancer Genomic Atlas (TCGA)-Pancreatic Adenocarcinoma (PAAD) tumors from the U.S. National Cancer Institute (NCI) Genomic Data Commons (GDC) database (<https://gdc.cancer.gov/>). One of the objectives of GDC is generating standardized molecular profiles and clinical datasets across different cancer genome programs [51]. We span 8 molecular profiles ranking from 1,339 miRNA isoforms (isomiR) to the 450,000 DNA methylation of CpG probes and 8 ML classification algorithms that employed all features and OMC feature selection, producing a total of 128 diverse models to predict PAAD patients' responses to gemcitabine. The goal of this study is to identify the robust ML model to classify gemcitabine responders and nonresponders in PAAD patients and to compare the predictive performances to gemcitabine single-gene marker.

Methods

Clinical data acquisition and preprocessing

The most common type of PC is PAAD [52]. The open access molecular profiling and clinical data of 185 primary pancreatic tumor samples in TCGA-PAAD project was downloaded from GDC application programming interface (API) (version 25.0). Neuroendocrine tumors and other noncancerous samples were thus not considered. To ensure the consistency of drug name, misspelling and synonyms (e.g., gemzar and gemcitabine HCl) of drugs were standardized according to the NCI drug dictionary and DrugBank to gemcitabine. Then, drug response information for each patient was obtained by querying the clinical

records. The patients with missing gemcitabine response or responding inconsistently to a drug were excluded to retain only valid records. The patients who received gemcitabine prior to tumor resection were removed as indicated by the time of tumor procurement and the start of treatment. After these curation steps, 70 gemcitabine-treated PAAD patients remained. The gemcitabine responses provided by TCGA-PAAD were binarized into 2 classes: responder and nonresponder. We defined a responder as a patient who had complete response (CR) or partial response (PR), and a nonresponder as a patient who had stable disease (SD) or progressive disease (PD). TCGA-PAAD reports best response over the entire treatment period, starting from the initiation of the treatment until the end of treatment, i.e., not only the response observed at last follow-up.

Molecular data acquisition and preprocessing

The curated gemcitabine responses were combined with the corresponding molecular profiles via TCGA patient IDs. Eight molecular profiles provided by GDC were considered in this analysis. The mRNA and miRNA expressions were generated from next-generation sequencing (NGS). mRNA(FPKM) was defined as mRNA in fragment per kilobase of exon per million mapped fragments (FPKM) unit, while mRNA(FPKM-UQ) was defined as mRNA in upper-quartile normalized FPKM unit. miRNA and isomiR were log₂-transformed reads per million mapped reads (RPM). CpG (5'-cytosine-phosphate-guanine-3') is beta values of DNA methylation level at known CpG sites using about 450,000 probes from Illumina Human Methylation 450 BeadChip. CGI is the average methylation beta of all probes' values at known CpG sites corresponding to their CpG island. CNV(mean) is the average of copy number variations (CNVs) across segmented DNA corresponding to their gene using CNTools (version 3.8) of R package (version 3.5.1). CNV(median) is determined in the same way as CNV(mean), except the median is calculated. Those features missing any value across patients were removed. Thus, 8 datasets were generated, where each dataset was served as a set of features and subjected to ML classification algorithms to make the predictions. Note that we consider slightly different profile normalization to elucidate the impact of these differences on ML model performance.

Data preparation for ML

A computational framework was developed to estimate the predictive ability of molecular profiles on PAAD patients' response to gemcitabine as shown in Fig. 1. In brief, each dataset was randomly split into a training and testing set using stratified *K*-fold, for *K* values of 5, 10, and 70. Eight ML classification algorithms, including classification and regression tree (CART), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), logistic regression (LR), support vector classifier (SVC) with linear kernel, SVC with the radial basis function (RBF) kernel, and *k*-nearest neighbors (KNN) algorithms, employed all features and OMC feature selection, producing a total of 128 models that were used to predict patients' responses to gemcitabine. We performed 5 *K*-fold cross-validation (CV) repetitions, each with different random seed to assess the variability of resulting ML models. The median of evaluation metrics across 5 repetitions was reported. All analyses of this study were performed using Python version 3.7.3 (<https://www.python.org/>) and scikit-learn version 0.24.2 (<https://scikit-learn.org>).

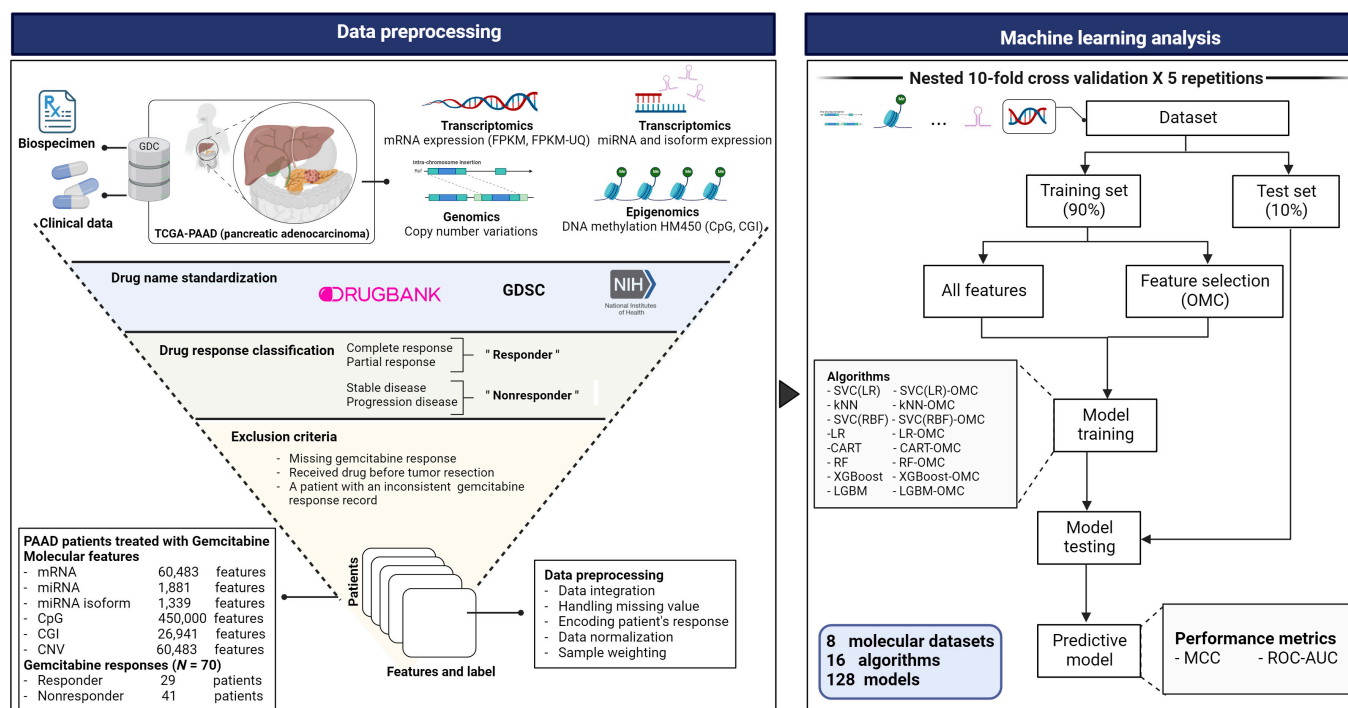


Fig. 1. ML workflow to predict PAAD patient response to gemcitabine. The TCGA-PAAD clinical, biospecimen, and molecular profiles comprising mRNA in FPKM, and FPKM-UQ unit, miRNAs and isoforms, copy number variation, DNA methylation at CpGs, and CGI profiles from high-throughput technologies of patient tumors were retrieved from the NCI GDC. These molecular profiles were matched with corresponding clinical information using the patient's TCGA barcode. The samples meeting the exclusion criteria were filtered out to retain only valid high-quality records. Each molecular profile of the tumors was preprocessed and subsequently submitted by each of the 8 supervised learning algorithm analyses. The performances of the resulting all-features ML models were evaluated by 10-fold CV (5 repetitions per model, each using a different random seed). Thus, 5 MCC and ROC-AUC determinations were carried out for each model. To mitigate the high dimensionality of the datasets, each of the 8 algorithms was coupled with the OMC strategy, with the resulting models undergoing identical performance evaluation. Overall, a total of 128 ML models (16 binary classification models for each of the 8 molecular profiles) were evaluated on the same cohort of patients. Further details can be found in the Methods section.

Building all-features classification models

Stratified K -fold standard CV runs were performed to estimate the predictive performance of ML algorithms for classification using all available features in each molecular dataset. After partitioning of the samples into K -folds, one of the K -folds was used as the test set for testing the model trained on the remaining $K - 1$ folds. The CV process was then repeated K times with each one of the K -folds used sequentially once in the test set. CV ensures that each patient has exactly one out-of-sample prediction regardless of the value of K . Model performance was not calculated on the test set of every fold and then averaged because each of the test sets would be too small and thus harm comparison to CV with other K values. As it is common with low-sample scenarios, we used instead a merged CV, where the 70 out-of-sample predictions were merged from all folds and Matthews correlation coefficient (MCC) was calculated once with all 70 samples.

Building OMC classification models

The OMC models [45] were implemented in this study to overcome the high dimensionality problem, in this case, the number of features is larger than the number of patients. Stratified K -fold nested CV was performed to estimate how well classification ML algorithms perform using only the most relevant features to gemcitabine responses. In brief, the analysis of variance (ANOVA) test was used to calculate the P value of each feature, with a low P value indicating high discriminative power

in terms of gemcitabine response. Then, classification algorithms were trained on only the considered subset of features (the top 2 to $n/2$ subset of features, where n is the number of samples). Among all $n/2$ trained models in the inner loop of the nested CV, the best model that achieved the highest MCC was selected and used to predict samples in its outer loop. The resulting 70 out-of-sample predictions were also merged from all folds, and each model performance metric was calculated once with all 70 samples.

Building a model based on hENT1 expression

hENT1 has been identified as a prognostic molecular marker in PC patients treated with gemcitabine [16,17]. To compare the performance of the best models to that of using a single hENT1 gene only, hENT1 expression derived from the mRNA(FPKM) profile was trained on the same algorithms as the 2 most predictive models (RF and XGBoost) using standard 10-fold CV from merging the out-of-sample predictions from the 10 folds.

Model performance evaluation

When a model estimated real-valued class probabilities, these were transformed to binary classes by setting a threshold to 0.5. The probabilities above this threshold were classed as responders; otherwise, they were classed as nonresponders. The 4 confusion matrix categories, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN),

were counted and used to calculate performance metrics including MCC and receiver operating characteristic–area under the curve (ROC-AUC).

Construction of protein–protein interaction (PPI) network

To unravel the relationships among the genes corresponding to 12 predictive CpG probes, these genes were mapped to STRING database version 11.5 (<https://string-db.org/>) [53] to retrieve both known and predicted protein–protein interactions (PPIs). We limited the species to “Homo sapiens” and required at least the minimum confidence interaction score (combine score > 0.15). The text mining, experiments, database, co-expression, co-occurrence, neighborhood, and gene fusion evidences were included to construct PPI.

The pathway enrichment analysis

In order to understand the roles of the genes found to be associated with the predictive CpG probes in PAAD treated with gemcitabine, the genes corresponding to 12 CpG probes were subjected to the Database for Annotation, Visualization and Integrated Discovery (DAVID) web server [54] for Gene Ontology (GO) pathway enrichment analysis. In addition, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was also investigated using the KEGG Orthology-Based Annotation System (KOBAS) web server (<http://kobas.cbi.pku.edu.cn/>) [55]. A statistically significant *P* value was calculated using a Fisher exact test, which further employed the Benjamini–Hochberg multiple-testing correction for the correct *P* value. The significant enriched pathways were identified based on a *P* value less than 0.05. For each pathway, input genes that are part of the pathway are counted, and the enrichment ratio was also calculated.

Statistical analysis

The two-tailed unpaired Welch’s *t* test was used to generate *P* values in data analysis. The median overall survival (OS) and progression-free survival (PFS) were calculated for 4 groups, including actual and predicted responders and nonresponders, representing the time point on the Kaplan–Meier plot where 50% of the patients in each group have survived (for OS) or have not experienced disease progression (for PFS). Log-rank test was used to compute statistical differences in OS and PFS. The univariate and multivariate Cox regression analyses were performed to assess hazard ratio (HR) and 95% confidence intervals (CIs) of gemcitabine response predictors (4 mRNAs and 12 CpG probes). A *P* value less than 0.05 was considered as statistically significant.

Results

A total of 70 molecularly profiled PAAD patients treated with gemcitabine (41 nonresponders and 29 responders) were derived from TCGA-PAAD project. The patient’s distribution and available features of each molecular profile, spanning from the 1,339 miRNA isoform expressions to the 450,000 methylation levels of DNA probes, were presented in Fig. S1. We trained each of 8 molecular profiles from PAAD patients’ tumor on 8 classification ML algorithms with and without OMC feature selection, producing a total of 128 gemcitabine response prediction models. Computational pipeline is described in the Methods section and is represented in Fig. 1.

ML model performances at predicting PAAD patient response to gemcitabine

For each of the algorithm–profile pair, the model performance was evaluated by stratified 10-fold CV with 5 repetitions, each repetition with a different random seed. Figure 2 shows the out-of-sample median MCC (mMCC) of the 5 repetitions (the boxplot with these 5 MCC determinations can be seen in Fig. S2). OMC models, which are enhanced with further feature selection, tend to perform better than all-features models. Indeed, only 31 (48.4%) of the 64 OMC models yield mMCC values below 0.1, compared to 56 (87.5%) of the all-features models. Models exploiting mRNA and DNA methylation profiles distinguish responders from nonresponders with mMCCs of over 0.3. Both mRNA(FPKM) and mRNA(FPKM-UQ) exhibited the similar predictive performances regardless of normalized gene expression metrics used. mRNA(FPKM) achieved the highest MCC; hence, it is recommended for this challenging problem. RF-OMC and LR-OMC incorporating mRNAs are the 2 best ML models, achieving mMCC of 0.44, and 0.4, respectively. In addition, XGBoost incorporating DNA methylation of CpG probes obtained mMCC of 0.32. Such poor performance was also observed when using SVC for these problem instances. All nonresponders were misclassified using some molecular profiles [either mRNA(FPKM), mRNA(FPKM-UQ), CpG, or CGI], leading to undefined MCCs. Median ROC-AUCs range from 0.407 to 0.674 (Fig. 2).

OMC models are generally more predictive while retaining a concise subset of predictive features

OMC is the process of selecting the most informative features that considerably reduces the number of features for model building. As shown in Fig. 2, the OMC models tend to be more predictive, whereas the all-features models obtained a near-random predictive level. The most predictive models with an mMCC of at least 0.3 were OMC models that employed either mRNAs or CpG probes (Fig. S3). These 7 predictive OMC models also obtained the highest median ROC-AUCs in this challenging problem, ranking from 0.654 with CART-OMC employing mRNA(FPKM) to 0.785 with RF-OMC employing mRNA(FPKM) (Fig. 3). We employ MCC as the main evaluation metric, as it is less flattering than ROC-AUC.

The differential expression of the gemcitabine predictors

mRNA expression is the most predictive profile for stratifying PAAD patients treated with gemcitabine, giving the highest MCC. Four (SEPWI1P, RP11-179A10.1, ATF4P4, and CTC-429L19.3) of 60,483 mRNAs are the most predictive features selected by RF-OMC that employed mRNA(FPKM) profile (mMCC of 0.44). Figure 4 and Table S2 show the differential expression of these 4 predictive genes between the responders and nonresponders using two-sided Welch’s *t* test. All these 4 predictive genes are significantly (*P* < 0.01) up-regulated in responders compared to nonresponders. CTC-429L19.3 is the most expressed genes, while ATF4P4, RP11-179A10.1, and SEPWI1P are the least expressed genes. Interestingly, these 4 predictive genes were commonly selected by other 5 most predictive OMC models with an mMCC of at least 0.3 employing mRNA profiles (Table S3). In particular, CART-OMC, the second most predictive model that employed mRNA(FPKM), also

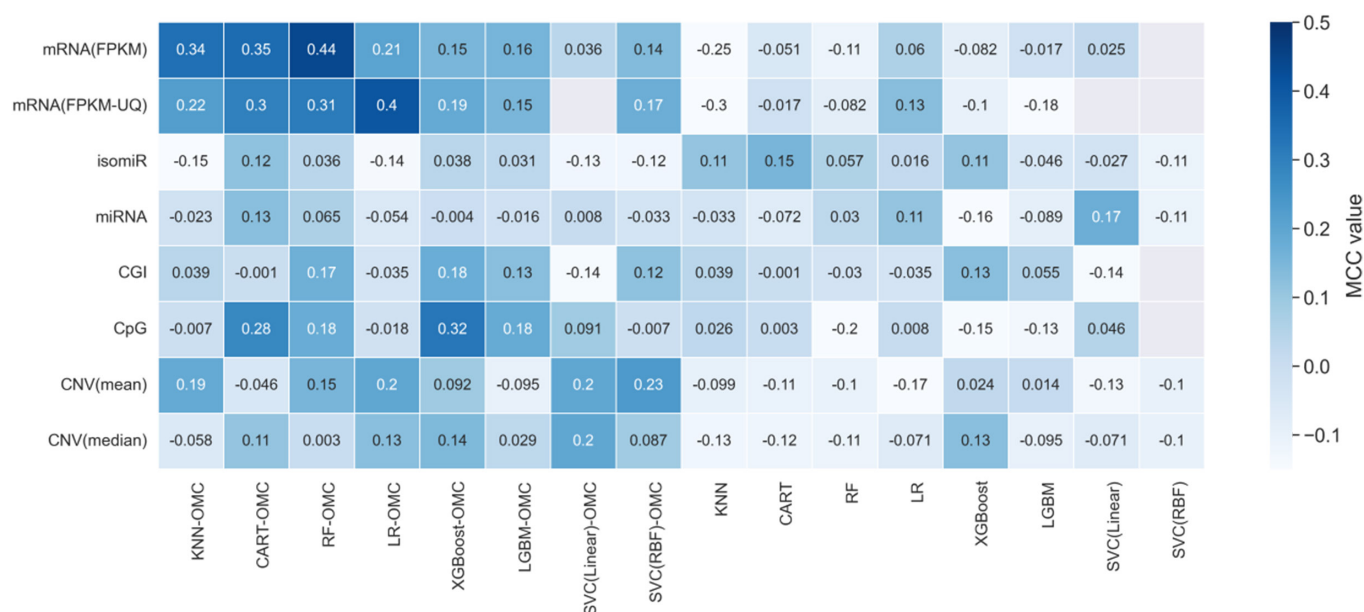


Fig. 2. Heatmap presenting the mMCC of the five 10-fold CV repetitions for each of the 128 models. The vertical axis shows the considered molecular profiles, while the horizontal axis shows the ML algorithms used. The first 8 columns on the left show supervised ML algorithms implemented with OMC feature selection considering only the subset of informative features during the model training (the suffix “OMC” was added to the algorithm name), while the remaining 8 columns on the right show the corresponding algorithms that used all features available in datasets. Five MCCs were obtained from five 10-fold CV runs ranking from -1 to 1 . The mMCC was shown in the heatmap. A positive MCC means that the model accurately predicts gemcitabine responses better than a random classifier. MCC of 0.0 indicates a prediction no better than a random. A negative MCC means that the model performs worse than a random. A model that obtained undefined MCC, predicting the same class for all instances where the denominator of MCC is zero, is indicated as blank boxes. The 2 most predictive molecular profiles were mRNA expression and DNA methylation of CpG: mMCC of 0.44 from RF-OMC using mRNA expression profiles and mMCC of 0.32 from XGBoost-OMC using DNA methylation of CpG profile. In addition, the OMC models tend to perform better than the corresponding all-features models.

combined the same 4 predictive mRNAs to build classification tree (mMCC of 0.35) (Fig. S4).

While 12 of 450,000 CpG probes [cg03208198 (COL18A1 and MIR6815), cg07239938 (ELANE), cg11528307 (C14orf80 and CRIP1), cg15200796 (TMEM191C), cg16049691 (AHRR), cg16379910 (B2M), cg17804635 (ZNF703), cg23854567 (PXN), cg24128434 (DNAH2), cg25653341 (PLOD3), cg27152190 (RP11-429P3.3), and ch.2.184672693R] were selected as predictive features by XGBoost-OMC that employed DNA methylation of CpG profile, the genes found to be related to each of these probes are between the brackets except for ch.2.184672693R for which GDC did not report any associated genes. Two-sided Welch's t tests show that the responders had significantly ($P < 0.01$) higher DNA methylation of these predictive CpG probes than nonresponders, except cg03208198, which showed lower DNA methylation in responders (Fig. 4 and Table S2). In addition, the feature importance for the top 2 best models was shown in Fig. S10. It has been found that ENSG00000267197.1 was selected by RF-OMC model as the most important feature in predicting gemcitabine response (Fig. S10A), while cg27152190 is the most importance feature for XGBoost-OMC model prediction (Fig. S10B).

PPI network analysis of genes corresponding to the 12 predictive CpG probes

The interactions among the genes corresponding to 12 predictive CpG probes were assessed using STRING web server [53], which integrates both known and predicted PPIs. Deciphering the protein interaction network could help in understanding the physical and functional associations of these protein that

could influence chemo-susceptibility. Eleven of 13 genes corresponding to predictive CpG probes were mapped in the STRING database, while the remaining 2 non-protein-coding genes were not mapped (MIR6815 is miRNA, and RP11-429P3.3 is antisense). Figure 5 shows the network with 11 nodes and 14 edges of PPI. The nodes are proteins, while edges are their interactions (PPI enrichment P value of 0.035). The known and predicted interactions were observed among 6 proteins (CRIP1, PXN, COL18A1, B2M, ELANE, and PLOD3). TMEM191B and DNAH2 were linked with text mining of scientific literature evidence. The remaining 3 genes (C14orf80, AHRR, and ZNF703) were disconnected nodes in the network. These 13 genes corresponding to predictive CpG probes were also enriched in cancer-associated pathways in GO and KEGG pathway enrichment analysis (see the next section). Overall, the results show that these gene interactions could contribute to the gemcitabine resistance in PAAD patients.

An analogous analysis could not be performed with the 4 predictive mRNAs, as these do not code for proteins and hence are not included in the STRING database. Two of them (ATF4P4 and SEPW1P) are annotated as pseudogenes, while the other 2 (RP11-179A10.1 and CTC-429L19.3) are long noncoding RNAs.

GO and KEGG pathway enrichment analysis of genes corresponding to the 12 predictive CpG probes

The GO enrichment analysis of genes corresponding to 12 predictive CpG probes revealed a total of 43 significantly ($P < 0.05$) enriched GO pathways (Fig. S5 and Table S4) across 3 GO categories: biological process (BP), cellular component (CC), and molecular function (MF). By analyzing BP, we found that

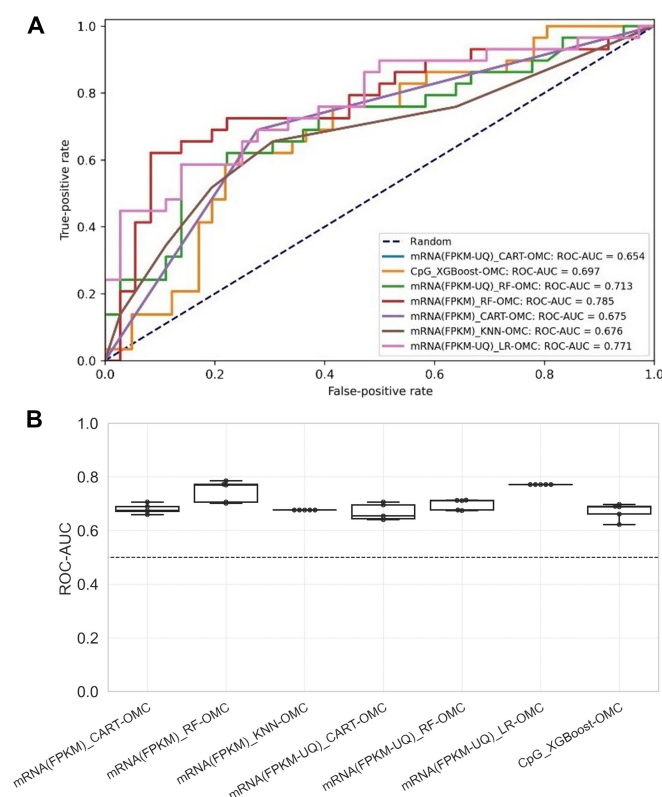


Fig. 3. ROC-AUC across five 10-fold CV runs of the 7 most predictive models with the highest median ROC-AUC. (A) Highest ROC-AUC among five 10-fold CV runs for each model. A ROC curve depicts the trade-off between the true-positive rates and the false-positive rates of a classifier by varying the threshold of the probability of response. A ROC-AUC of 1 is the best model performance at distinguishing between responders and nonresponders, while a ROC-AUC of 0.5 corresponds to the random guessing denoted by a dashed line. RF-OMC that employed the mRNA model has the highest ROC-AUC of 0.785. (B) ROC-AUC variability across the five 10-fold CV runs for each of the 7 most predictive models with the highest median ROC-AUC. The horizontal dashed line at 0.5 represents random-level performance.

genes were enriched in 39 GO BP pathways. Most of these pathways are cancer-related pathways and contributed to cancer progression, for example, cell differentiation and development, cell proliferation, regulation of cell adhesion, and cell motility. Cancer cells undergo rapid proliferation and migration [56]. Cellular adhesion is also an important process in cancer progression, as it allows cancer cells to interact with and invade surrounding tissue [57]. Additionally, the extracellular matrix and extracellular structures are important components of the tumor microenvironment and play a critical role in cancer progression [58]. These pathways provide insights into the underlying biological mechanisms that drive cancer development. Interestingly, 8 genes (COL18A1, PXN, ZNF703, AHRR, PLOD3, CRIP1, B2M, and ELANE) were predominantly enriched in response to chemical and response to stimulus pathways. Dysregulation of these 8 enriched genes in response to chemical or stimulus pathways could promote carcinogenesis and drug resistance [59–61]. Six of these 8 enriched genes were also linked with known and predicted interaction evidences in PPI network, suggesting that these genes may function in coordinated manner and contributed to drug resistance (Fig. 5). Furthermore, GO enrichment through CC showed that the terms endoplasmic reticulum lumen, macromolecular complex, and specific granule

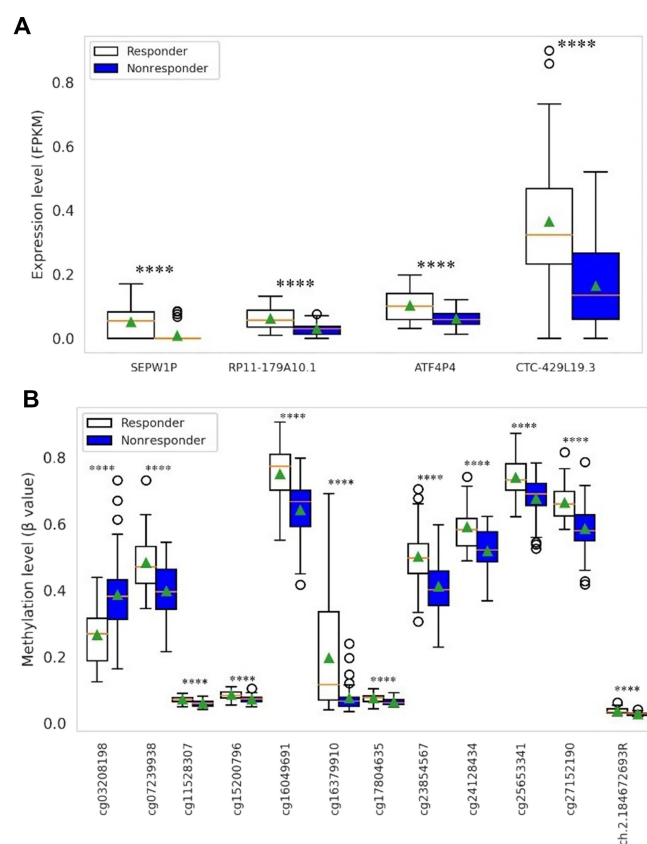


Fig. 4. Boxplot showing the differential expression of the predictive genes and DNA methylation of CpG probes between responders and nonresponders. The y axis represents the normalized expression level for 4 predictive genes from RF-OMC employing mRNA(FPKM) profile (A) and 12 predictive DNA methylation CpG probes from XGBoost-OMC employing CpG profile (B) between responders and nonresponders. The corresponding gene(s) for each CpG probe can be found in Table S6. The orange line inside each boxplot represents the median expression across patients for a given feature, while the green triangle represents the mean expression. The boxplot for each feature contains the *P* value of mean differential expression between responders and nonresponders using a two-sided Welch's *t* test. *****P* ≤ 0.0001, significant difference in expression levels between 2 groups.

lumen indicate an intracellular localization where they are likely to interact. By analyzing MF, 2 genes were enriched in transcriptional repressor activity and RNA polymerase II transcription factor binding pathways. This pathway was related to gemcitabine mechanism, which was involved in negative regulation of transcription from an RNA polymerase II promoter that led to apoptosis [62].

The KEGG pathway enrichment analysis (Table S5) showed that genes corresponding to CpG probes were significantly (*P* < 0.05) enriched in 10 KEGG pathways, with the metabolism of xenobiotics by cytochrome P450 pathway being particularly noteworthy due to its involvement in drug metabolism and patient response to treatment. Overexpression of cytochrome P450 causes rapid drug elimination before it reaches the target site [63]. Moreover, antigen processing and presentation pathway is critical for the recognition of tumor cells by the immune system. Tumor cells can present antigens on their surface, which can be recognized by T cells and trigger an immune response. Immunotherapy approaches such as immune checkpoint inhibitors target this pathway to enhance T cell activity and improve antitumor immunity [64].

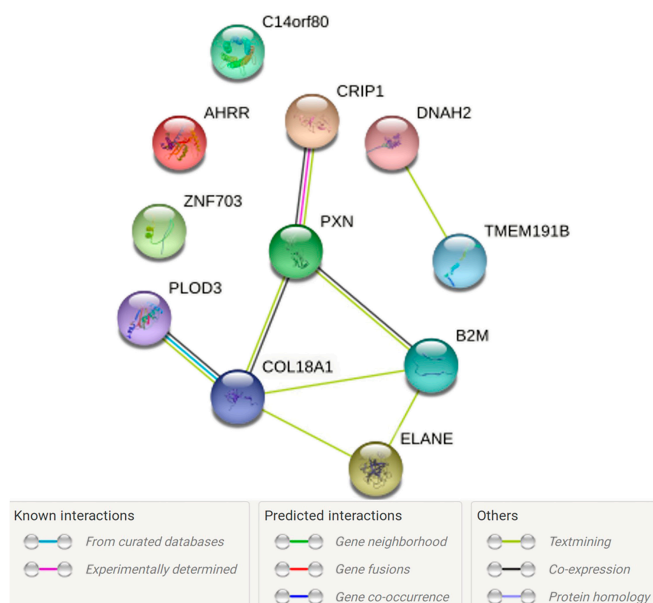


Fig. 5. The PPI network analysis. The genes associated with the 12 predictive CpG probes (Table S8), as annotated by the GDC, were input into STRING database for PPI network analysis. Out of 13 genes corresponding to predictive CpG probes, 11 protein-coding genes were found in the STRING database, while the remaining 2 genes (MIR6815 and RP11-429P3) were not found as they are non-protein-coding genes (MIR6815 is miRNA, and RP11-429P3.3 is antisense). These 11 protein-coding genes were connected by 14 edges representing their known and predicted interactions as shown in the legend. A PPI enrichment P value of 0.035 was obtained, indicating that the protein input has more interactions among themselves than what would be expected for a set of proteins of the same size and degree distribution drawn at random from the genome. As a result, the proteins were connected as a group. The nodes in the PPI network represent proteins, and the edges represent their known and predicted interactions.

Altogether, our findings show that dysregulation of DNA methylation of predictive CpGs in this study could be associated to the alteration of their corresponding genes and pathways in which they are involved, leading to cancer development and gemcitabine resistance in PAAD patients. The pathway enrichment analysis provides insight into the underlying biological mechanisms of genes corresponding to predictive CpG that could be targeted for cancer therapy. However, functional validations are needed to test our bioinformatics findings.

Assessing the robustness of the best model in predicting PAAD patients' gemcitabine responses

We next investigated the effect of training dataset sizes on the predictive performance of the best model. Leave-one-out CV (LOOCV), 5-fold CV (5CV), and 10-fold CV (10CV) were used to evaluate the RF-OMC that employed mRNA (FPKM) profile model performance. All the CVs were repeated exactly in the same manner for all-features models. The MCCs and ROC-AUCs of 5 repetitions, each with different random seeds, are presented in Fig. 6. The results show the improvement in predictive performances as we increase the training dataset size from 80% of the data in 5CV to 99% of the data in LOOCV, while all-features models obtained a near-random predictive level and were significantly worse than those obtained from OMC models in all cases (Fig. 6). We also investigated whether the performances obtained from the signal in the dataset was better than what would be predicted by chance from

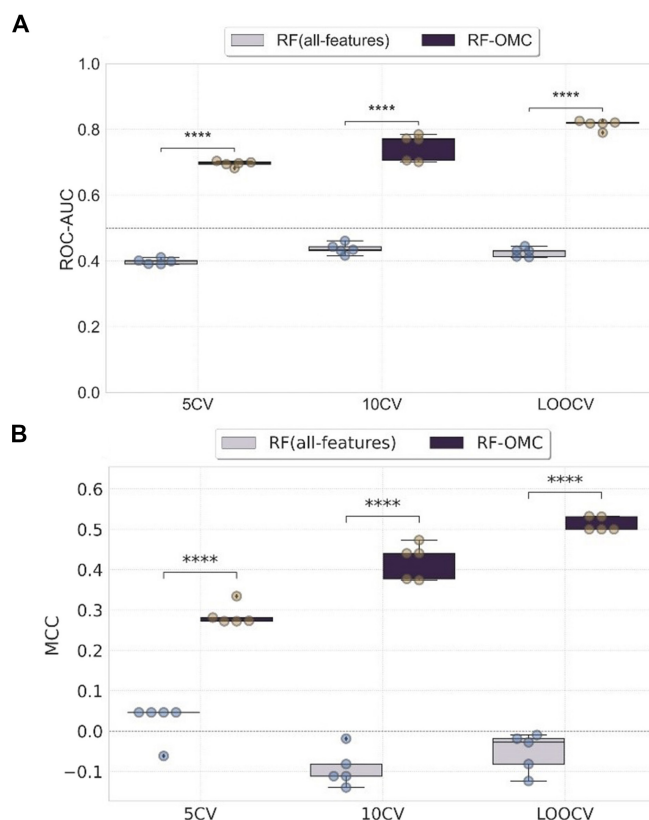


Fig. 6. The impact of training dataset size on the performance of best model versus its corresponding all-features model. The boxplots present the distributions of MCC (A) and ROC-AUC (B) obtained across five 10-fold CV runs of RF-OMC employing mRNA (FPKM) profile with 5CV, 10CV, and LOOCV (i.e., $K = 5, 10$, and 70 , respectively). RF incorporated OMC feature selection (RF-OMC) was represented in deep purple, while RF trained on all available 60,483 mRNA features [RF(all-features)] was represented in light purple. Each boxplot contains the P value of mean differential expression between 2 groups with a two-sided Welch's t test. **** $P \leq 0.0001$. The model performances improved as the training dataset size increased. In addition, the OMC models significantly outperformed the corresponding all-features models in all CVs. The dashed line represents the random classifier at MCC of 0 and ROC-AUC of 0.5. MCC or ROC-AUC of 1 indicates a perfectly accurate prediction.

class-permuted version. As a result, all permutation models were significantly worse than those arising from OMC models in all cases (Fig. S6). The similar result was obtained when using XGBoost-OMC, which employed DNA methylation of CpG profile and showed improved predictive performance as the training dataset size increased (Fig. S9).

CpG-based XGBoost-OMC can predict survival of gemcitabine-treated PAAD patients

We evaluated the OS (defined as the time interval from the date of diagnosis to death or the last known follow-up date) and PFS (defined as the time interval from the date of diagnosis to disease progression or the last known follow-up date) time difference between actual and predicted responders and nonresponders by the XGBoost-OMC combined 12 CpG probes. Among 70 patients with survival data and gemcitabine responses, 41 (58.6%) nonresponders and 29 (41.4%) responders (Fig. 7A), XGBoost-OMC incorporating 12 CpG probes classified these patients into 43 predicted nonresponders and 27 predicted responders. The confusion matrix was shown in Table S3.

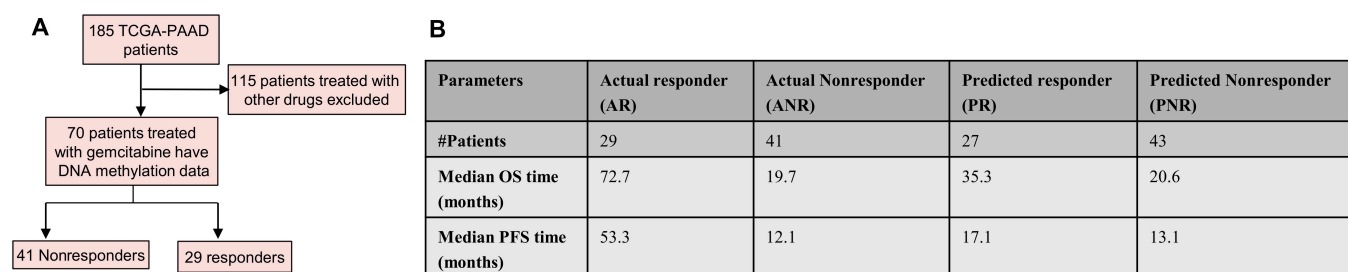


Fig. 7. Comparison between the median OS and PFS time of actual patient response to gemcitabine and those predicted by the CpG-based XGBoost-OMC model. (A) Flowchart of patient selection. (B) Median OS (defined as the time interval from the date of diagnosis to death or the last known follow-up date) and PFS (defined as the time interval from the date of diagnosis to disease progression or the last known follow-up date) time for 4 groups, including actual and predicted responders and nonresponders by the XGBoost-OMC model. The median OS and PFS are defined as the time point on the Kaplan–Meier plot, where 50% of the patients in each group have survived (for OS) or have not experienced disease progression (for PFS). The figure highlights the ability of the CpG-based XGBoost-OMC model to accurately predict patient gemcitabine response, as demonstrated by the similar trend between the predicted and actual outcomes. The actual and predicted responders had longer OS and PFS than actual and predicted nonresponders.

The median OS and PFS were calculated for 4 groups, including actual and predicted responders and nonresponders by the XGBoost-OMC model. The median OS and PFS were not significantly different between the actual and the predicted responders (log-rank test P value of 0.19 for OS and 0.14 for PFS) and nonresponders (log-rank P value of 0.33 for OS and 0.11 for PFS). The actual responders had significantly longer OS (median OS was 72.7 months) and PFS (median PFS was 53.3 months) than actual nonresponders (median OS was 19.7 months; median PFS was 12.1) with log-rank test P value < 0.005 . Similarly, the predicted responders had longer OS (median OS was 35.3 months) and PFS (median PFS was 17.1 months) compared to the predicted nonresponders (median OS was 20.6 months; median PFS was 13.1). However, the difference was not statistically significant with log-rank test P value of 0.19 and 0.09 in OS and PFS, respectively (Fig. 7B).

mRNA-based RF-OMC can also predict survival of gemcitabine-treated PAAD patients

Among 65 mRNA profiled patients, 36 nonresponders and 29 responders (Fig. 8A), RF-OMC incorporating 4 mRNAs classified these patients into 35 predicted nonresponders and 30 predicted responders. The confusion matrix was shown in Table S3. The median OS and PFS were not significantly different between the actual and the predicted responders (log-rank test P value of 0.21 for OS and 0.15 for PFS) and nonresponders (log-rank P value of 0.45 for OS and 0.20 for PFS). The OS and the PFS were significantly longer in the actual responders (median OS was 72.7 months; median PFS was 53.3 months) compared to the actual nonresponders (median OS was 17.7 months; median PFS was 10.2 months) with log-rank test P value < 0.005 . Similarly, the OS and PFS were longer in the predicted responders (median OS was 35.3; median PFS was 16.4) compared to the predicted nonresponders (median OS was 21.1; median PFS was 12.2) with log-rank test P value of 0.19 for OS and 0.04 for PFS (Fig. 8B).

Notably, the median OS and PFS time of actual response to gemcitabine and those predicted by CpG-based XGBoost and mRNA-based RF-OMC models exhibited the similar results. The median PFS of predicted responders and predicted nonresponders by the mRNA-based RF-OMC model shows a statistically significant difference (log-rank test P value of 0.04), despite the small sample size (Fig. 8B). This suggests that RF-OMC combining 4 mRNAs could be a prognostic predictor of gemcitabine-treated PC patients.

Identification of the predictive 4 mRNAs and the 12 CpG probes as the independent survival predictors in gemcitabine-treated PAAD patients

We next investigated whether gemcitabine response predictors (predictive 4 mRNAs and 12 CpG probes) could be survival predictors in PAAD patients. The univariate Cox regression analysis of 4 mRNAs from 65 mRNA profiled patients revealed that 3 genes could be independent biomarkers for prediction of OS and PFS ($P < 0.05$) (Fig. S7A). The high expression of SEPW1P (HR = 0.59 for OS; HR = 0.72 for PFS), RP11-179A10.1 (HR = 0.63 for OS; HR = 0.58 for PFS), and CTC-429L19.3 (HR = 0.64 for OS; HR = 0.58 for PFS) was significantly associated with improved OS and PFS, but it was not significant in multivariate Cox regression analysis (Fig. S7, A and C). These findings were consistent with the differential expression of these 4 mRNAs, which were shown to be highly expressed in responders (Fig. 4).

Furthermore, the univariate analysis of the 12 CpG probes from 70 DNA methylation profiled patients revealed that 3 CpG probes could be used as independent biomarkers for OS prediction ($P < 0.05$) (Fig. S7B). The high DNA methylation of cg15200796 (HR = 0.68) and cg16049691 (HR = 0.7) was significantly associated with improved OS. In contrast, high DNA methylation of cg03208198 (HR = 1.4) was significantly associated with poor OS. In addition, 4 CpG probes could be used as independent biomarkers to predict PFS ($P < 0.05$) (Fig. S7B). The high DNA methylation of cg07239938 (HR = 0.71), cg15200796 (HR = 0.77), and ch.2.184672693R (HR = 0.61) was associated with improved PFS, whereas high DNA methylation of cg03208198 (HR = 1.47) was negatively associated with PFS. The remaining CpG probes had no significant association with patient survival. In multivariate Cox regression analysis, only cg03208198 (HR = 1.56 for OS; HR = 1.41 for PFS) was shown to be significantly associated with OS and PFS (Fig. S7D). These results were consistent with the differential expression of these 12 CpG probes, where high DNA methylation of cg03208198 was associated with nonresponders, while the remaining CpG probes showed high DNA methylation in responders (Fig. 4). These suggest that gemcitabine response predictors could be used as independent survival predictors of PAAD patients receiving gemcitabine.

Subsequently, we also performed univariate Cox regression analysis on hENT1 gene. The result in Fig. S8A shows that hENT1 gene was not significantly associated with both OS

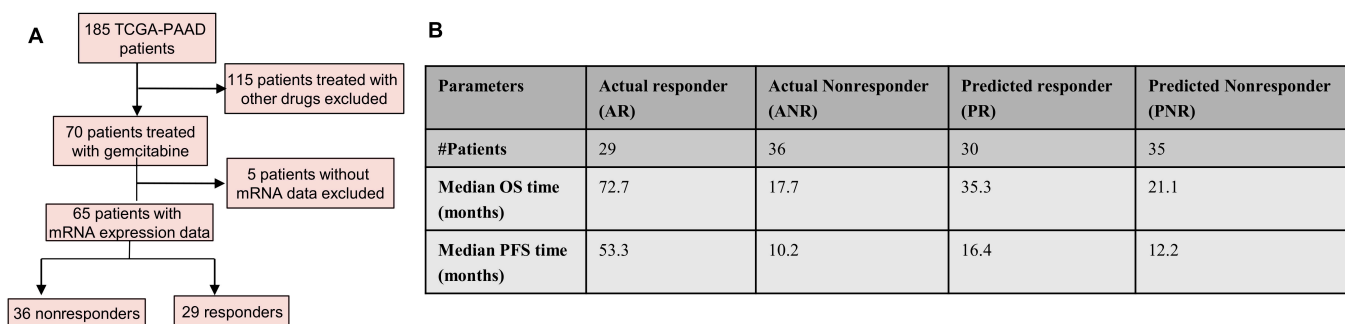


Fig. 8. Comparison between the median OS and PFS time of actual patient response to gemcitabine and those predicted by the mRNA-based RF-OMC model. (A) Flowchart of patient selection. (B) Median OS (defined as the time interval from the date of diagnosis to death or the last known follow-up date) and PFS (defined as the time interval from the date of diagnosis to disease progression or the last known follow-up date) time for 4 groups, including actual and predicted responders and nonresponders by the RF-OMC model. The median OS and PFS are defined as the time point on the Kaplan–Meier plot where 50% of the patients in each group have survived (for OS) or have not experienced disease progression (for PFS). The figure highlights the ability of the mRNA-based RF-OMC model to accurately predict patient gemcitabine response, as demonstrated by the similar trend between the predicted and actual outcomes. The actual and predicted responders had longer OS and PFS than actual and predicted nonresponders.

(HR = 1.01; $P = 0.63$) and PFS (HR = 1.0; $P = 0.76$) in the univariate model, suggesting that hENT1 gene was not prognostic in PAAD patient receiving gemcitabine. In addition, Fig. S8B demonstrates poor patient stratification by using hENT1 gene trained on RF model, in which there was a trend toward better survival in predicted nonresponders.

The effect of data integration on the gemcitabine response prediction

We next investigated the effect of data integration on the predictive performance of the best model. The RF-OMC was trained on mRNA profile, clinical features (Table S1), and integrated mRNA and clinical features. Figure 9A shows the MCCs across five 10-fold CV repetitions. Combining mRNA and clinical features (FPKM + clinical) did not improve the model performance (mMCC of 0.44) compared to those built on only mRNA features (FPKM) (mMCC of 0.44). RF-OMC prediction based on only clinical data (clinical) led to the lowest performance (mMCC of 0.014). If merging the 2 best molecular profiles provided a performance boost, it would justify the time and financial overhead of determining both profiles in every patient. However, RF-OMC integrating the 2 most predictive CpG methylation and mRNA profile performed worse than those by individual profile (mMCC of 0.349). Last, RF-OMC employed all the molecular profiles and clinical features (All_profiles) did not improve the prediction (mMCC of 0.253). Overall, there is no benefit in integrating patients' clinical data with molecular profiles on the predictive performance of the best model. Importantly, increasing number of features could reduce the model performances as a curse of dimensionality. We did not carry out systematic ablation experiments combining more than 2 profiles mostly due to unfavorable cost–benefit reasons.

Comparing the model prediction based on mRNA and CpG methylation profiles to the hENT1, a well-established gemcitabine response biomarker

Many studies reported that hENT1 has shown to be a predictive biomarker of gemcitabine response in PC. The high expression of hENT1 is associated with sensitivity to gemcitabine and increased overall and disease-free survival [16,17,65]. Figure 9B shows that RF-OMC combining 4 mRNAs (mMCC of 0.44) significantly outperformed RF prediction using only hENT1

gene expression (mMCC of -0.066). Similarly, XGBoost-OMC combining 12 CpG probes (mMCC of 0.32) obtained a significantly higher MCC compared to XGBoost that employed only hENT1 gene expression (mMCC of -0.045). Table S6 also presents the predictive performances in terms of MCC and ROC-AUC. The results suggest that the prediction based on either 4 predictive mRNAs or 12 CpG probes outperforms those by hENT1 single-gene biomarker in predicting gemcitabine responses.

Discussion

Omics-based precision oncology, also known as targeted therapies or genomic medicine, has improved cancer patient management. The focus has been on either single-gene marker [17,66,67], a single molecular profile [67–70], or the discovery of rare actionable mutations [71,72]. However, predicting drug responses is a challenging task in oncology. Using a single-gene marker or a few molecular profiles is often insufficient to correctly predict drug responses [50,73,74]. The previous researches [44,50,75] demonstrated that the multiple molecular profiles provide new opportunities to identify robust biosignatures that can discriminate drug responses and aid to tailor treatments for individual cancer patients [44,50,75].

Supervised ML algorithms are increasingly being applied to the multiple molecular profiles as they can predict the clinical outcomes for new data based on what they learned from the previous data. Owing to well-curated molecular profiling data of GDC, we present a large-scale analysis of supervised 128 ML models built on multiple molecular profiles and algorithms to predict gemcitabine responses in 70 PAAD patients. Smaller sample sizes are not uncommon for cancer-specific drug response prediction problems [76]. Here, out-of-sample predictions were carried out with rigorous CV procedures, which have been shown to anticipate held-out test set performance in smaller datasets than the one used in our study [77].

Using five 10-fold CV runs, 7 of 128 (5%) models are predictive with mMCC > 0.3. These 7 predictive models were all OMC models. Training only the most informative features in OMC models increases the chance of making accurate prediction by removing the irrelevant features from high dimensionality dataset early prior to model training, and model performance improves compared to all-features models. The

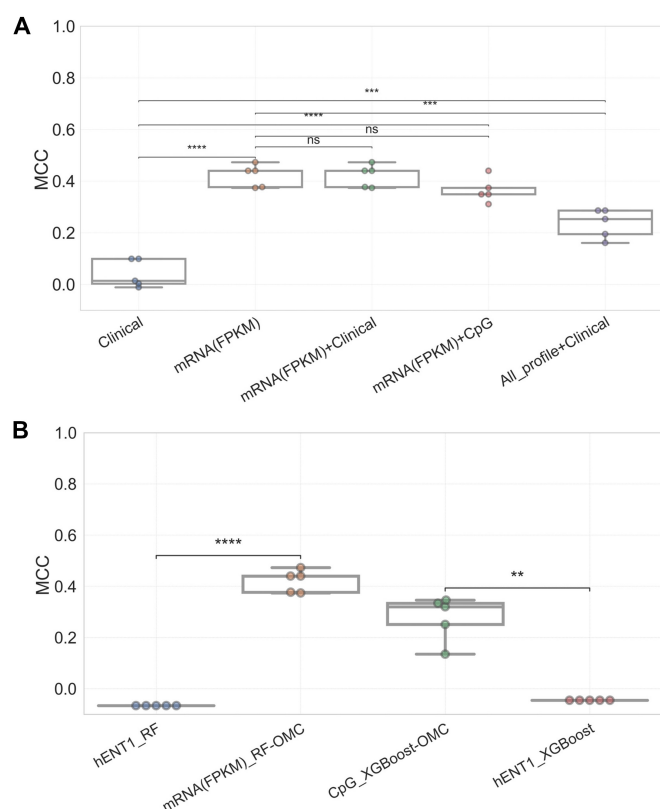


Fig. 9. Comparison of the predictive performances of the best model to those combining other datasets and those using hENT1 single gene only in predicting gemcitabine response in PAAD patients. (A) Boxplots comparing the MCCs obtained from five 10-fold CV runs (each dot represents a run) of the RF-OMC prediction based on mRNA profile [mRNA(FPKM)], combining mRNA and clinical dataset [mRNA(FPKM) + Clinical], combining mRNA and CpG profile [mRNA(FPKM) + CpG] and clinical data only (Clinical) (Table S1), and combining all of 8 molecular profiles and clinical data (All_profile + Clinical). MCC ranges between -1 (misclassification) and 1 (perfect classification), with 0 corresponding to no better than a random. Altogether, integrated patients' clinical data have no effect on the predictive performance of the best model. Importantly, increasing number of features could reduce the model performances. (B) To compare the predictive performances of the 2 most predictive models with those using hENT1 only, hENT1 gene expression from mRNA(FPKM) profile was trained on the same algorithms as the 2 most predictive models (RF and XGBoost). Boxplots comparing the MCCs of the 2 most predictive models [mRNA(FPKM)_RF-OMC and CpG_XGBoost-OMC] to those using the expression of the hENT1 gene only (hENT1_RF and hENT1_XGBoost) in gemcitabine response prediction of PAAD patients. The 2 most predictive models obtained in this study significantly outperformed the prediction based on hENT1 expression as a single predictor. Statistical comparisons between different models were performed using two-sided Welch's *t* test. Stars denote the *P* value, where nonsignificant "ns" means $0.05 < P \leq 1.00$, $*0.01 < P \leq 0.05$, $**0.001 < P \leq 0.01$, $***0.0001 < P \leq 0.001$, and $****P \leq 0.0001$.

2 most predictive molecular profiles in this study revealed 4 of 60,483 mRNAs (0.066%) from RF-OMC model and 12 of 450,000 CpG probes (0.027%) from XGBoost-OMC model as the most informative biosignatures for predicting gemcitabine responses in PAAD patients. These models did not only exhibit high out-of-sample mMCC values robust to different partitions and repetitions, but these values were also significantly higher than those from all-features models (Fig. 6) and class-permuted models (Figs. S6 and S9). Moreover, the performance of these 2 OMC models monotonically increased with larger training set sizes (Fig. 6). Last, note that, in general,

other combinations of molecular profiles and algorithms will be predictive in different drugs and cancer types [44,45,50].

The RF-OMC model nonlinearly combined 4 predictive mRNA features including SEPW1P, RP11-179A10.1, ATF4P4, and CTC-429L19.3 that have accurately predicted PAAD patients' responses to gemcitabine. Two of these predictive mRNAs (ATF4P4 and SEPW1P) are pseudogene, while the remaining mRNAs are RNA genes (RP11-179A10.1 and CTC-429L19.3) (Table S7). These genes might be involved in drug resistance and cancer progression through alterations in their expression. Other studies have already examined the roles of these genes in other types of cancer. For example, depletion of SEPW1P enhances tumor suppressor activities in breast cancer [78]. The single-nucleotide polymorphisms (SNPs) located in long noncoding RNA RP11-179A10.1 were reportedly associated with poorer outcomes in ovarian cancer [79], but their association with PAAD has not been documented.

Previous studies have shown that potential cancer biomarkers can be identified using DNA methylation profiles [80–82]. In this study, we could identify 12 DNA methylation of CpG probes affecting 14 genes as the predictive biomarkers of gemcitabine responses in PAAD patients by XGBoost-OMC model. The genes found to be associated with these CpG probes are presented in bracket: cg03208198 (COL18A1 and MIR6815), cg07239938 (ELANE), cg11528307 (C14orf80 and CRIP1), cg15200796 (TMEM191C), cg16049691 (AHRR), cg16379910 (B2M), cg17804635 (ZNF703), cg23854567 (PXN), cg24128434 (DNAH2), cg25653341 (PLOD3), cg27152190 (RP11-429P3.3), and ch.2.184672693R (not linked to any gene). The 2 CpG probes (cg11528307 and ch.2.184672693R) were commonly selected by other models that employed CpG profile, including CART-OMC, LGBM-OMC, and RF-OMC, with mMCC of 0.28, 0.18, and 0.18, respectively (Table S8). In addition, we also investigated the differential expression of gemcitabine response predictors and found that increased 4 mRNA expressions and DNA methylation of 11 CpG sites in pretreatment biopsies from gemcitabine-treated PAAD patients are correlated with better response to therapy, while increased DNA methylation of cg03208198 predicts resistance to therapy (Fig. 4). In fact, dysregulation of DNA methylation (hypo- or hypermethylation) is related to affected gene activation or associated with gene silencing including known tumor suppressor genes and loss of gene functions in cancer [83]. Some of the genes corresponding to the predictive CpG probes have been previously reported to be associated with the prognosis and clinical outcomes in various cancers. The high expression of COL18A1 in PC [84], gastric cancer [85], and esophageal squamous cell carcinoma [86] showed poor clinical outcomes. Down-regulation of CRIP1 has also been identified in invasive PC [87]. ZNF703 is an oncogenic transcriptional regulator involved in cell adhesion, movement, and proliferation [88]. Overexpression of ZNF703 is linked to tamoxifen resistance [89]. From this subset of predictive genes, we carried out PPI and pathway analyses to further support follow-up mechanistic studies. Here, we found that the genes corresponding to predictive CpG probes were significantly enriched in a number of cancer-associated pathways and drug resistance from GO and KEGG pathway database (Fig. S5 and Tables S4 and S5).

We next analyzed the survival difference observed between predicted responders and nonresponders by using the best models among gemcitabine-treated PAAD patients. The Kaplan–Meier survival analysis shows that the median OS and PFS were not significantly different between the actual and the predicted

responders and nonresponders from the best models. Both OS and PFS were increased in responders compared to nonresponders (Figs. 7 and 8). The univariate analysis suggests that gemcitabine response predictors could be used as independent features for survival prediction of PAAD patients receiving gemcitabine (Fig. S7). Some studies have reported these genes as predictors of PAAD survival outcomes [68,90]. Note that building a better predictor of PAAD survival would require considering all PAAD patients regardless of the administered drug, which is out of the scope of this study.

Additionally, merging features did not improve the model performances (Fig. 9A). Clinical features alone were barely predictive, and combining them with predictive mRNA profile did not improve the model performance. Furthermore, there is no improvement by integrating 2 most predictive molecular profiles. In addition, loss in predictive accuracy was observed when combining all molecular profiles and clinical features. These suggest that RF-OMC was not able to capture the hidden informative features when the dimensionality dataset exceeded 500,000 features.

Furthermore, several studies have reported hENT1 expression as a gemcitabine response biomarker in PC [16,17], but the mMCC across five 10-fold CV runs was barely predictive (mMCC ~ 0), while the prediction based on 4 predictive mRNAs (mMCC of 0.44) and 12 predictive CpGs (mMCC of 0.32) achieved a significantly higher performance (Fig. 9B). This could suggest the attractiveness of the identified biosignatures. However, the improvement of these ML models over the hENT1 marker will have to be more accurately quantified in a prospective clinical trial.

Although the best models, based on mRNAs and CpG profiles, have predictive and robust performances, there are several limitations to be considered. First, this is a retrospective study using publicly available data from TCGA-PAAD project. While the number of patients is large for a cancer-specific drug response study (diagnosis or prognosis datasets are not restricted to a single administered drug and hence are larger), a larger cohort would further improve model performance. Indeed, as shown in Fig. 6, the more data used to train the model, the more accurate the model. Second, an independent validation cohort would further test the generalizability of the best model, i.e., whether this is able to stratify patients with different characteristics or technical validations (e.g., sample preparation, sequencing procedure, and bioinformatics pipeline). With that said, this is not critical here as it is when using clinical trial data, as TCGA data come from clinical practice in different hospitals. Third, the effect of drug combinations could be further evaluated along with its individual effect, as a few patients were given other drugs in combination with gemcitabine.

Conclusions

In summary, this systematic ML analysis revealed the discovery of predictive, reproducible, and robust gene expression and DNA methylation computational models for gemcitabine-treated PAAD patients. These models can be employed as response signatures to guide clinicians to stratify PAAD patients who will respond to gemcitabine in order to propose other suitable drugs without delay and avoid side effects to patients who are unlikely to respond. In addition, these signatures could be a starting point to explore the gemcitabine resistance mechanism.

Acknowledgments

We thank H. Du (Imperial College) for his feedback on this study. The study reuses public data, and thus, no further ethical approval is required.

Funding: This work was supported by the Petroleum Technology Development Fund (Nigeria, PTDF/ED/PHD/APO/27/18), the Commemoration of Her Royal Highness Princess Chulabhorn's 60th Birthday Anniversary scholarship from Chulabhorn Royal Academy (Thailand), the Marie-Curie actions IEF-Horizon (EU, LCII_PA4887), and the Wolfson Foundation and the Royal Society for a Royal Society Wolfson Fellowship (UK, RSWF\R1\221005).

Author contributions: P.J.B. and G.G. conceived the idea and designed the experiments. A.O. and C.P. collected and analyzed the patients' data, developed the ML process, and interpreted the results. A.O., C.P., G.G., and P.J.B. wrote the manuscript.

Competing interests: The authors declare that they have no competing interests.

Data Availability

The datasets supporting the results of this article are open access data downloaded from the TCGA-PAAD project within the GDC resource (<https://portal.gdc.cancer.gov/>). The python code and processed mRNA(FPKM) dataset are provided in order to build the best model to predict and evaluate gemcitabine response in PAAD patients and facilitate their application to other cohorts of mRNA-profiled PAAD patients: https://github.com/chayanitpiy/gemcitabine_PAAD

Supplementary Materials

Figs. S1 to 10
Tables S1 to S8

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–249.
2. Zhu H, Li T, Du Y, Li M. Pancreatic cancer: Challenges and opportunities. *BMC Med.* 2018;16(1):214.
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin.* 2022;72(1):7–33.
4. Jia Y, Xie J. Promising molecular mechanisms responsible for gemcitabine resistance in cancer. *Genes Dis.* 2015;2(4):299–306.
5. Zhang XW, Ma YX, Sun Y, Cao YB, Li Q, Xu CA. Gemcitabine in combination with a second cytotoxic agent in the first-line treatment of locally advanced or metastatic pancreatic cancer: A systematic review and meta-analysis. *Target Oncol.* 2017;12(3):309–321.
6. Sarabi M, Mais L, Oussaid N, Desseigne F, Guibert P, De La Fouchardiere C. Use of gemcitabine as a second-line treatment following chemotherapy with folfirinox for metastatic pancreatic adenocarcinoma. *Oncol Lett.* 2017;13(6):4917–4924.
7. Nagourney RA, Flam M, Link J, Hager S, Blitzer J, Lyons W, Sommers BL, Evans S. Carboplatin plus gemcitabine repeating doublet therapy in recurrent breast cancer. *Clin Breast Cancer.* 2008;8(5):432–435.

8. Pfisterer J, Plante M, Vergote I, du Bois A, Hirte H, Lacave Uwe Wagner AJ, Stähle A, Stuart G, Kimmig R, Olbricht S, et al. Gemcitabine plus carboplatin compared with carboplatin in patients with platinum-sensitive recurrent ovarian cancer: An intergroup trial of the AGO-OVAR, the NCIC CTG, and the EORTC GCG. *J Clin Oncol*. 2006;24(29):4699–4707.
9. von der Maase H, Hansen SW, Roberts JT, Dogliotti L, Oliver T, Moore MJ, Bodrogi I, Albers P, Knuth A, Lippert CM, et al. Gemcitabine and cisplatin versus methotrexate, vinblastine, doxorubicin, and cisplatin in advanced or metastatic bladder cancer: Results of a large, randomized, multinational, multicenter, phase III study. *J Clin Oncol*. 2000;18(17):3068–3077.
10. Sandler AB, Nemunaitis J, Denham C, von Pawel J, Cormier Y, Gatzemeier U, Mattson K, Manegold C, Palmer MC, Gregor A, et al. Phase III trial of gemcitabine plus cisplatin versus cisplatin alone in patients with locally advanced or metastatic non-small-cell lung cancer. *J Clin Oncol*. 2000;18(1):122–130.
11. Amrutkar M, Gladhaug IP. Pancreatic cancer chemoresistance to gemcitabine. *Cancers (Basel)*. 2017;9(11):157.
12. Berdis AJ. Inhibiting DNA polymerases as a therapeutic intervention against cancer. *Front Mol Biosci*. 2017;4:78.
13. Von Hoff DD, Ervin T, Arena FP, Gabriela Chiorean E, Infante J, Moore M, Seay T, Tjulandin SA, Ma WW, Saleh MN. Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N Engl J Med*. 2013;369(18):1691–1703.
14. Bird NTE, Elmasry M, Jones R, Psarelli E, Dodd J, Malik H, Greenhalf W, Kitteringham N, Ghaneh P, Neoptolemos JP, et al. Immunohistochemical hENT1 expression as a prognostic biomarker in patients with resected pancreatic ductal adenocarcinoma undergoing adjuvant gemcitabine-based chemotherapy. *Br J Surg*. 2017;104(4):328–336.
15. Wei CH, Gorgan TR, Elashoff DA, Hines OJ, Farrell JJ, Donahue TR. A meta-analysis of gemcitabine biomarkers in patients with pancreato-biliary cancers. *Pancreas*. 2013;42(8):1303–1310.
16. Spratlin J, Sangha R, Glubrecht D, Dabbagh L, Young JD, Dumontet C, Cass C, Lai R, Mackey JR. The absence of human equilibrative nucleoside transporter 1 is associated with reduced survival in patients with gemcitabine-treated pancreas adenocarcinoma. *Clin Cancer Res*. 2004;10(20):6956–6961.
17. Farrell JJ, Elsaleh H, Garcia M, Lai R, Ammar A, Regine WF, Abrams R, Benson AB, Macdonald J, Cass CE, et al. Human equilibrative nucleoside transporter 1 levels predict response to gemcitabine in patients with pancreatic cancer. *Gastroenterology*. 2009;136(1):187–195.
18. Giovannetti E, Tacca MD, Mey V, Funel N, Nannizzi S, Ricci S, Orlandini C, Boggi U, Campani D, Chiaro MD, et al. Transcription analysis of human equilibrative nucleoside transporter-1 predicts survival in pancreas cancer patients treated with gemcitabine. *Cancer Res*. 2006;66(7):3928–3935.
19. Wang W, Yu X, Li H, Yang C, Jin C, Huang X. hENT1's role in adjuvant intra-arterial gemcitabine-based chemotherapy for resectable pancreatic cancer patients. *BMC Gastroenterol*. 2023;23(1):35.
20. Kawada N, Uehara H, Katayama K, Nakamura S, Takahashi H, Ohigashi H, Ishikawa O, Nagata S, Tomita Y. Human equilibrative nucleoside transporter 1 level does not predict prognosis in pancreatic cancer patients treated with neoadjuvant chemoradiation including gemcitabine. *J Hepatobiliary Pancreat Sci*. 2012;19(6):717–722.
21. Yabushita Y, Mori R, Taniguchi K, Matsuyama R, Kumamoto T, Sakamaki K, Kubota K, Endo I. Combined analyses of hENT1, TS, and DPD predict outcomes of borderline-resectable pancreatic cancer. *Anticancer Res*. 2017;37(5):2465–2476.
22. Liu X, Xiao C, Yue K, Chen M, Zhou H, Yan X. Identification of multi-omics biomarkers and construction of the novel prognostic model for hepatocellular carcinoma. *Sci Rep*. 2022;12(1):12084.
23. Kong L, Liu P, Zheng M, Xue B, Liang K, Tan X. Multi-omics analysis based on integrated genomics, epigenomics and transcriptomics in pancreatic cancer. *Epigenomics*. 2020;12(6):507–524.
24. Guinney J, Wang T, Laajala TD, Winner KK, Bare JC, Neto EC, Khan SA, Peddinti G, Airola A, Pahikkala T, et al. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: Development of a prognostic model through a crowdsourced challenge with open clinical trial data. *Lancet Oncol*. 2017;18(1):132–142.
25. Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, Fuchs C, Gofer E, Kumar L, Wolfinger RD, et al. Prostate Cancer DREAM Challenge Community. A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. *JCO Clin Cancer Inform*. 2017;1:1–15.
26. Ghislat G, Cheema AS, Baudoin E, Verthuy C, Ballester PJ, Crozat K, Attaf N, Dong C, Milpied P, Malissen B, et al. NF- κ B-dependent IRF1 activation programs cDC1 dendritic cells to drive antitumor immunity. *Sci Immunol*. 2021;6(61):Article eabg3570.
27. Menden MP, Wang D, Mason MJ, Szalai B, Bulusu KC, Guan Y, Yu T, Kang J, Jeon M, Wolfinger R, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun*. 2019;10(1):Article 2674.
28. Piyawajanusorn C, Nguyen LC, Ghislat G, Ballester PJ. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. *Brief Bioinform*. 2021;22(6):Article bbab312.
29. Dang CC, Peón A, Ballester PJ. Unearthing new genomic markers of drug response by improved measurement of discriminative power. *BMC Med Genomics*. 2018;11(1):10.
30. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012;483(7391):570–575.
31. Frejno M, Chiozzi RZ, Wilhelm M, Koch H, Zheng R, Klaeger S, Ruprecht B, Meng C, Kramer K, Jarzab A, et al. Pharmacoproteomic characterisation of human colon and rectal cancer. *Mol Syst Biol*. 2017;13(11):951.
32. Chibon F. Cancer gene expression signatures—The rise and fall? *Eur J Cancer*. 2013;49(8):2000–2009.
33. Liu K, Geng Y, Wang L, Xu H, Zou M, Li Y, Zhao Z, Chen T, Xu F, Sun L, et al. Systematic exploration of the underlying mechanism of gemcitabine resistance in pancreatic adenocarcinoma. *Mol Oncol*. 2022;16(16):3034–3051.
34. Ballester PJ, Stevens R, Haibe-Kains B, Huang RS, Aittokallio T. Artificial intelligence for drug response prediction in disease models. *Brief Bioinform*. 2022;23(1):Article bbab450.
35. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics*. 2021;22(1):Article 214.
36. Yu L, Zhou D, Gao L, Zha Y. Prediction of drug response in multilayer networks based on fusion of multiomics data. *Methods*. 2021;192:85–92.

37. Dorman SN, Baranova K, Knoll JHM, Urquhart BL, Mariani G, Carcangiu ML, Rogan PK. Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. *Mol Oncol*. 2016;10(1):85–100.
38. Menden MP, Iorio F, Garnett M, Dermott UM, Benes CH, Ballester PJ, Saez-Rodriguez J. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLOS ONE*. 2013;8(4):Article e61318.
39. Naulaerts S, Dang CC, Ballester PJ. Precision and recall oncology: Combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget*. 2017;8(57):97025–97040.
40. Nguyen L, Dang CC, Ballester PJ. Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *F1000Res*. 2016;5.
41. Naulaerts S, Menden MP, Ballester PJ. Concise polygenic models for cancer-specific identification of drug-sensitive tumors from their multi-omics profiles. *Biomolecules*. 2020;10(6):Article 963.
42. Ammad-Ud-Din M, Khan SA, Malani D, Murumägi A, Kallioniemi O, Aittokallio T, Kaski S. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*. 2016;32(17):i455–i463.
43. Ballester PJ, Carmona J. Artificial intelligence for the next generation of precision oncology. *NPJ Precis Oncol*. 2021;5(1):Article 79.
44. Bomane A, Gonçalves A, Ballester PJ. Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. *Front Genet*. 2019;10:Article 1041.
45. Ogunleye AZ, Piyawajanusorn C, Gonçalves A, Ghislat G, Ballester PJ. Interpretable machine learning models to predict the resistance of breast cancer patients to doxorubicin from their microRNA profiles. *Adv Sci (Weinh)*. 2022;9(24):Article e2201501.
46. Xu Y, Dong Q, Li F, Xu Y, Hu C, Wang J, Shang D, Zheng X, Yang H, Zhang C, et al. Identifying subpathway signatures for individualized anticancer drug response by integrating multi-omics data. *J Transl Med*. 2019;17(1):Article 227.
47. Krishna V, Nimgaonkar V, Tiu E, Krishna V, Bhambhani H, Cook S, Miller D, Vrabac D, Joshi A, Singhi AD, et al. Gemcitabine response prediction in the adjuvant treatment of resected pancreatic ductal adenocarcinoma using an AI histopathology platform. *J Clin Oncol*. 2022;40(16):e16295–e16295.
48. Clayton EA, Pujol TA, McDonald JF, Qiu P. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinformatics*. 2020;21(Suppl 14):364.
49. Foersch S, Glasner C, Woerl A-C, Eckstein M, Wagner D-C, Schulz S, Kellers F, Fernandez A, Tserea K, Kloth M, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med*. 2023;29(2):430–439.
50. Nguyen LC, Naulaerts S, Bruna A, Ghislat G, Ballester PJ. Predicting cancer drug response in vivo by learning an optimal feature selection of tumour molecular profiles. *Biomedicine*. 2021;9(10):1319.
51. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113–1120.
52. Kleeff J, Korc M, Apte M, Vecchia CL, Johnson CD, Biankin AV, Neale RE, Tempero M, Tuveson DA, Hruban RH, et al. Pancreatic cancer. *Nat Rev Dis Primers*. 2016;2:–16022.
53. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605–D612.
54. Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID-WS: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics*. 2012;28(13):1805–1806.
55. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, Wu Y, Zhao L, Liu J, Guo J, et al. KOBAS-i: Intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res*. 2021;49(W1):W317–W325.
56. Cooper GM. *The development and causes of cancer*. Sunderland (MA): Sinauer Associates; 2000.
57. Wittekind C, Neid M. Cancer invasion and metastasis. *Oncology*. 2005;69(Suppl 1):14–16.
58. Xiong G-F, Xu R. Function of cancer cell-derived extracellular matrix in tumor progression. *J Cancer Metastasis Treat*. 2016;2(9):357–364.
59. Stewart BW. Mechanisms of carcinogenesis: From initiation and promotion to the hallmarks. In: Baan RA, Stewart BW, Straif K, editors. *Tumour site concordance and mechanisms of carcinogenesis*. Lyon (France): International Agency for Research on Cancer; 2019.
60. Sever R, Brugge JS. Signal transduction in cancer. *Cold Spring Harb Perspect Med*. 2015;5(4):a006098.
61. Andl CD. The misregulation of cell adhesion components during tumorigenesis: Overview and commentary. *J Oncol*. 2010;2010:Article 174175.
62. Linzer N, Trumbull A, Nar R, Gibbons MD, Yu DT, Strouboulis J, Bungert J. Regulation of RNA polymerase II transcription initiation and elongation by transcription factor TFII-I. *Front Mol Biosci*. 2021;8:Article 681550.
63. Doehmer J, Goeptar AR, Vermeulen NP. Cytochromes P450 and drug resistance. *Cytotechnology*. 1993;12(1–3):357–366.
64. Mpakali A, Stratikos E. The role of antigen processing and presentation in cancer and the efficacy of immune checkpoint inhibitor immunotherapy. *Cancers (Basel)*. 2021;13(1):134.
65. García-Manteiga J, Molina-Arcas M, Casado FJ, Mazo A, Pastor-Anglada M. Nucleoside transporter profiles in human pancreatic cancer cells: Role of hCNT1 in 2,2-difluorodeoxycytidine-induced cytotoxicity. *Clin Cancer Res*. 2003;9(13):5000–5008.
66. Nakano Y, Tanno S, Koizumi K, Nishikawa T, Nakamura K, Minoguchi M, Izawa T, Mizukami Y, Okumura T, Kohgo Y. Gemcitabine chemoresistance and molecular markers associated with gemcitabine transport and metabolism in human pancreatic cancer cells. *Br J Cancer*. 2007;96(3):457–463.
67. Voutsadakis IA. Molecular predictors of gemcitabine response in pancreatic cancer. *World J Gastrointest Oncol*. 2011;3(11):153–164.
68. Wei X, Zhou X, Zhao Y, He Y, Weng Z, Xu C. A 14-gene gemcitabine resistance gene signature is significantly associated with the prognosis of pancreatic cancer patients. *Sci Rep*. 2021;11(1):6087.
69. Nicolle R, Gayet O, Duconseil P, Vanbrugghe C, Roques J, Bigonnet M, Blum Y, Elarouci N, Armenoult L, Ayadi M, et al.

- A transcriptomic signature to predict adjuvant gemcitabine sensitivity in pancreatic adenocarcinoma. *Ann Oncol*. 2021;32(2):250–260.
70. Mehrmohamadi M, Jeong SH, Locasale JW. Molecular features that predict the response to antimetabolite chemotherapies. *Cancer Metab*. 2017;5:8.
 71. Schuh A, Dreau H, Knight SJL, Ridout K, Mizani T, Vavoulis D, Colling R, Antoniou P, Kvikstad EM, Pentony MM, et al. Clinically actionable mutation profiles in patients with cancer identified by whole-genome sequencing. *Cold Spring Harb Mol Case Stud*. 2018;4(2):Article a002279.
 72. Tanaka M, Javle M, Dong X, Eng C, Abbruzzese JL, Li D. Gemcitabine metabolic and transporter gene polymorphisms are associated with drug toxicity and efficacy in patients with locally advanced pancreatic cancer. *Cancer*. 2010;116(22):5325–5335.
 73. Rodríguez-Antona C, Taron M. Pharmacogenomic biomarkers for personalized cancer treatment. *J Intern Med*. 2015;277(2):201–217.
 74. Huang M, Shen A, Ding J, Geng M. Molecularly targeted cancer therapy: Some lessons from the past decade. *Trends Pharmacol Sci*. 2014;35(1):41–50.
 75. He D-X, Gu F, Gao F, Hao J-J, Gong D, Gu X-T, Mao A-Q, Jin J, Fu L, Ma X. Genome-wide profiles of methylation, microRNAs, and gene expression in chemoresistant breast cancer. *Sci Rep*. 2016;6:Article 24706.
 76. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol*. 2014;32(12):1202–1212.
 77. Tsamardinos I, Charonyktakis P, Papoutsoglou G, Borboudakis G, Lakiotaki K, Zenklusen JC, Juhl H, Chatzaki E, Lagani V. Just add data: Automated predictive modeling for knowledge discovery and feature selection. *NPJ Precis Oncol*. 2022;6(1):38.
 78. Tan L, Mai D, Zhang B, Jiang X, Zhang J, Bai R, Ye Y, Li M, Pan L, Su J, et al. PIWI-interacting RNA-36712 restrains breast cancer progression and chemoresistance by interaction with SEPW1 pseudogene SEPW1P RNA. *Mol Cancer*. 2019;18(1):9.
 79. Johnatty SE, Tyrer JP, Kar S, Beesley J, Lu Y, Gao B, Fasching PA, Hein A, Ekici AB, Beckmann MW, et al. Genome-wide analysis identifies novel loci associated with ovarian cancer outcomes: Findings from the Ovarian Cancer Association Consortium. *Clin Cancer Res*. 2015;21(23):5264–5276.
 80. Stirzaker C, Taberlay PC, Statham AL, Clark SJ. Mining cancer methylomes: Prospects and challenges. *Trends Genet*. 2014;30(2):75–84.
 81. Mikeska T, Craig JM. DNA methylation biomarkers: Cancer and beyond. *Genes (Basel)*. 2014;5(3):821–864.
 82. Levenson VV. DNA methylation as a universal biomarker. *Expert Rev Mol Diagn*. 2010;10(4):481–488.
 83. Ehrlich M. DNA methylation in cancer: Too much, but also too little. *Oncogene*. 2002;21(35):5400–5413.
 84. Nakajima K, Ino Y, Naito C, Nara S, Shimasaki M, Ishimoto U, Iwasaki T, Doi N, Esaki M, Kishi Y, et al. Neoadjuvant therapy alters the collagen architecture of pancreatic cancer tissue via Ephrin-A5. *Br J Cancer*. 2022;126(4):628–639.
 85. Weng K, Huang Y, Deng H, Wang R, Luo S, Wu H, Chen J, Long M, Hao W. Collagen family genes and related genes might be associated with prognosis of patients with gastric cancer: An integrated bioinformatics analysis and experimental validation. *Transl Cancer Res*. 2020;9(10):6246–6262.
 86. Li J, Wang X, Zheng K, Liu Y, Li J, Wang S, Liu K, Song X, Li N, Xie S, et al. The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma. *PeerJ*. 2019;7:e7705.
 87. Terris B, Blaveri E, Crnogorac-Jurcevic T, Jones M, Missiaglia E, Ruszniewski P, Sauvanet A, Lemoine NR. Characterization of gene expression profiles in intraductal papillary-mucinous tumors of the pancreas. *Am J Pathol*. 2002;160(5):1745–1754.
 88. Yang F, Teves SS, Kemp CJ, Henikoff S. Doxorubicin, DNA torsion, and chromatin dynamics. *Biochim Biophys Acta*. 2014;1845(1):84–89.
 89. Zhang X, Mu X, Huang O, Xie Z, Jiang M, Geng M, Shen K. Luminal breast cancer cell lines overexpressing ZNF703 are resistant to tamoxifen through activation of Akt/mTOR signaling. *PLOS ONE*. 2013;8(8):e72053.
 90. Wu M, Li X, Zhang T, Liu Z, Zhao Y. Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. *Front Oncol*. 2019;9:9996.