LARGE-SCALE AND PAN-CANCER MULTI-OMIC ANALYSES WITH MACHINE LEARNING

Zhaoxiang Cai

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Children's Medical Research Institute

Faculty of Medicine and Health

The University of Sydney

Jan 2023

Keywords

Cancer; Multi-omic data integration; Drug response, CRISPR-Cas9; Cancer vulnerabilities; Machine learning; Deep learning.

Thesis Abstract

Multi-omic data analysis has been foundational in many fields of molecular biology, including cancer research. Investigation of the relationship between different omic data types reveals patterns that cannot otherwise be found in a single data type alone. With recent technological advancements in mass spectrometry (MS), MS-based proteomics has enabled the quantification of thousands of proteins in hundreds of cell lines and human tissue samples. The proteome of these lines and samples have provided additional insights into disease biology beyond the genome and transcriptome. This thesis presents several machine learning-based methods that facilitate the integrative analysis of multi-omic data.

First, we reviewed five existing multi-omic data integration methods and performed a benchmarking analysis, using a large-scale multi-omic cancer cell line dataset. We evaluated the performance of these machine learning methods for drug response prediction and cancer type classification. Our result provides recommendations to researchers regarding optimal machine learning method selection for their applications.

Second, we generated a pan-cancer proteomic map of 949 cancer cell lines across 40 cancer types and developed a machine learning method DeeProM to analyse the multi-omic information of these lines. DeeProM identifies 8,498 proteins with evidence of cell types, protein-protein interaction, and broad post-transcriptional regulation. It also discovers protein-specific biomarkers of drug response and gene essentiality. The predictive performance of this dataset using machine learning was comparable with the RNA-seq and an external proteomic dataset. Further analysis demonstrated that a random subset of 1,500 proteins had a limited impact on predictive

performance, consistent with protein networks being highly connected and co-regulated. This pan-cancer proteomic map (ProCan-DepMapSanger) is now publicly available and represents a major resource for the scientific community, for biomarker discovery and for the study of fundamental aspects of protein regulation.

Third, we focused on publicly available multi-omic datasets of both cancer cell lines and human tissue samples and developed a Transformer-based deep learning method, DeePathNet, which integrates human knowledge with machine intelligence. DeePathNet incorporates cancer pathway knowledge into its network design by grouping omic data. A Transformer encoder was utilised to dynamically model the interdependency between cancer pathways, further improving the predictive performance. We applied DeePathNet on three evaluation tasks, namely drug response prediction, cancer type classification and breast cancer subtype classification. For a wide range of experiments, DeePathNet achieved better predictive performance than other methods that do not incorporate knowledge of cancer pathways. We used SHapley Additive exPlanations (SHAP) and Layer-wise Relevance Propagation (LRP) for model explanation and identify several key omic features and pathways that were related to breast cancer subtype classification.

Taken together, our analyses and methods allowed more accurate cancer diagnosis and prognosis.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature: _____

Date: _____

Acknowledgements

I would like to thank my supervisor, Dr. Qing Zhong, for his guidance and support throughout my PhD candidature. I am incredibly grateful for his generosity of time and valuable advice, without which this thesis would not have been possible. I would like also to thank my co-supervisors Dr. Rebecca Poulos and Professor Roger Reddel, as well as Dr Erdahl Teber (who co-supervised the first year of my candidature). I am extremely grateful to each of them for their guidance, helpful discussions and feedback on my work.

In addition, I am very grateful to my wife Aofei, my daughter Yueling and my parents – only through their ongoing support and encouragement have I reached this point.

Authorship Attribution Statement

Chapter 2 of this thesis is published as: “Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience* , 103798.”.

I designed the study, analysed the data and wrote and edited the manuscript.

Chapter 3 of this thesis is published as: “Gonçalves, E.*, Poulos, R. C.*, Cai, Z.* [* joint first authors], Barthorpe, S., Manda, S. S., Lucas, N., ... & Reddel, R. R. (2022). Pan-cancer proteomic map of 949 human cell lines. Cancer cell, 40(8), 835-849.”.

I co-designed the analysis plan of the study and co-analysed the data. Specifically, I performed data quality control, unsupervised landscape analysis using dimensionality reduction, and differential analysis. I designed, developed, and performed all machine learning analyses. I co-wrote the manuscript.

Chapter 4 of this thesis is published as a preprint: “Cai, Z., Poulos, R. C., Aref, A., Robinson, P. J., Reddel, R. R., & Zhong, Q. (2022). Transformer-based deep learning integrates multi-omic data with cancer pathways. In *bioRxiv* (p. 2022.10.27.514141). https://doi.org/10.1101/2022.10.27.514141.”.

I designed the study, analysed the data and wrote and edited the manuscript.

Signature:

Date:

Attesting Your Authorship Attribution Statement

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Signature:

Date:

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Signature:

Date:

Table of Contents

Publications Arising from this PhD Research

Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience*, 103798.

Gonçalves, E.*, Poulos, R. C.*, Cai, Z.* [* joint first authors], Barthorpe, S., Manda, S. S., Lucas, N., ... & Reddel, R. R. (2022). Pan-cancer proteomic map of 949 human cell lines. *Cancer cell*, 40(8), 835-849.

Poulos, R. C., Cai, Z., Robinson, P. J., Reddel, R. R., & Zhong, Q. (2022). Opportunities for pharmacoproteomics in biomarker discovery. *Proteomics*, 2200031.

Cai, Z., Poulos, R. C., Aref, A., Robinson, P. J., Reddel, R. R., & Zhong, Q. (2022). Transformer-based deep learning integrates multi-omic data with cancer pathways. In *bioRxiv* (p. 2022.10.27.514141). https://doi.org/10.1101/2022.10.27.514141.

Presentations Arising from this PhD Research

Human Proteome Organisation (HUPO) Congress 2022 | Mexico (2022)

Poster presentation: Transformer-based deep learning integrates multi-omic data with regulatory pathways in cancer

Multi-Omics ONLINE Webinar by Front Line Genomics | Virtual (2022)

Invited oral presentation: Machine Learning for multi-omics data integration in cancer

Sydney Bioinformatics Research Symposium 2022 | Australia (2022)

Poster presentation: Transformer-based deep learning integrates multi-omic data with regulatory pathways in cancer

70th ASMS Conference on Mass Spectrometry and Allied Topics 2022 | USA (2022)

Oral presentation: Pan-cancer proteomic map of 949 human cell lines reveals principles of cancer vulnerabilities

The Combined ABACBS and Phylomania 2021 Hybrid Conference | Virtual/Australia (2021)

Oral presentation: Integrating multi-omic data with biological knowledge by Transformer-based deep learning

KCA Precision Medicine for Childhood Cancer Symposium 2021 | Australia (2021)

Invited oral presentation: Pan-cancer proteomic map of 949 human cell lines reveals principles of cancer vulnerabilities

List of Figures

In Chapter 2

In Chapter 3

In Chapter 4

In Appendix A (relating to Chapter 3)

In Appendix B (relating to Chapter 4)

List of Tables

In Chapter 2

In Appendix A (relating to Chapter 3)

In Appendix B (relating to Chapter 4)

List of Abbreviations

3C	Chromosome conformation capture
4C	Circular Chromosome Conformation Capture
5C	Carbon-Copy Chromosome Conformation Capture
AUC	Area under the curve
CCA	Canonical correlation analysis
CCLE	Cancer Cell Line Encyclopedia
ChiA-PET	Chromatin Interaction Analysis with Paired-End Tag
ChIP-seq	Chromatin immunoprecipitation sequencing
CNV	copy number variation
COSMIC	Catalog of Somatic Mutations In Cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CUP	Cancers of unknown primary
DepMap	The Cancer Dependency Map
DIA	Data-independent acquisition
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPBS	Dulbecco's phosphate buffered saline
EMT	Epithelial-to-mesenchymal transition
FDR	False discovery rate
GPU	Graphics processing unit
H&E	Hematoxylin and eosin
IC50	Half-maximal inhibitory concentration
ICGC	International Cancer Genome Consortium
INDELs	Insertions and deletions
LC	Liquid chromatography
MAE	Mean absolute error
miRNA	MicroRNA
MKL	Multiple kernel learning
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
MSE	Mean squared error
PCA	Principal component analysis
PDX	Patient-derived xenografts
PPI	Protein-protein interaction
RNA-seq	RNA sequencing
scRNA-seq	Single cell RNA sequencing
SDC	Sodium deoxycholate
SNV	Single nucleotide variant
STR	Short tandem repeat
SV	Structural variation

SWATHTM-MS	Sequential Window Acquisition of All Theoretical Mass Spectra
TCGA	The Cancer Genome Atlas
TMT	Tandem mass tag
UMAP	Uniform manifold approximation and projection
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Chapter 1: Introduction

This chapter presents the background (Section 1.1), the aims (Section 1.2) and the outline (section 1.3) of the thesis.

1.1 BACKGROUND

Cancer has been one of the most richly studied diseases for decades. However, technological advances have only recently enabled the mass spectrometry (MS)-based mapping of entire proteomes of human cancers in a high-throughput manner (Aebersold and Mann 2003; Ong and Mann 2005; Wilhelm et al. 2014). Investigating protein dysregulation in cancer has a fundamental role in both understanding cancer mechanisms and developing new therapeutic approaches. Although the behaviour of a cell is largely defined by the proteins it produces, most large-scale and pan-cancer studies to-date (Iorio et al. 2016; Rohart, Gautier, Singh, and Cao 2017; Hoadley et al. 2018) have focused on either a small number of proteins or completely excluded proteomic analyses due to inherent technical challenges. The ACRF International Centre for the Proteome of Human Cancer (ProCan®) is located at the Children's Medical Research Institute in Westmead, Australia (Tully et al. 2019). ProCan endeavours to create a public knowledgebase of cancer proteomes from tens of thousands of cancer samples using data-independent acquisition MS (DIA-MS) (Gillet et al. 2012).

Multi-omic data analysis enables researchers to gain an increased understanding of tumour biology and the identification of more robust therapeutic targets (Rodriguez et al. 2021). A variety of multi-omic studies have led to the improved detection of intra-tumour heterogeneity, identification of novel therapeutic targets, as well as more

robust diagnostic, prognostic and predictive markers (Reel et al. 2021; Picard et al. 2021; Rohart, Gautier, Singh, and Lê Cao 2017; I. Subramanian et al. 2020). Many of these discoveries would not have been possible by analysing any single omic data type alone. However, performing multi-omic analysis presents computational challenges due to the large amount of data generated by high-throughput instruments and the limitations of existing multi-omic data integration methods (Tarazona, Arzalluz-Luque, and Conesa 2021; Cai et al. 2022).

The bedrock of multi-omic data analysis is machine learning, based upon which many tools have been developed (Argelaguet et al. 2020; Mo et al. 2018; Sharifi-Noghabi et al. 2019). Machine learning algorithms are trained to model complex patterns that cannot be accurately captured by traditional mathematical models in high dimensional data (Russell, Russell, and Norvig 2020). Publications of existing methods often emphasize the computational aspects of the proposed models, but lack a thorough introduction to the characteristics of individual omics. Recent reviews on multi-omic data integration focus on either biological applications or machine learning algorithms (Nicora et al. 2020; Picard et al. 2021; Reel et al. 2021; I. Subramanian et al. 2020), or both (Cai et al. 2022).

Focusing on one of the largest ProCan datasets (ProCan-SangerDepMap) (Gonçalves et al. 2022) and other publicly available multi-omic information, the primary aim of my PhD is to perform large-scale multi-omic data analyses using novel machine learning methods on both cancer cell lines and human tissue samples.

1.2 AIMS

Aim 1 (Chapter 2): Conduct a benchmarking analysis to assess the accuracy and runtime efficiency of the existing machine learning methods for multi-omic data integration.

Aim 2 (Chapter 3): Perform an integrative analysis using data from a collection of over 1,000 cancer cell lines (Iorio et al. 2016), comprising genomic and transcriptomic data, proteomic data (ProCan-SangerDepMap) generated in ProCan (Gonçalves et al. 2022), and data from drug response and gene essentiality (Pacini et al. 2021).

Aim 3 (Chapter 4): Design an explainable deep learning model that integrates existing cancer specific domain knowledge with quantitative multi-omic measurements to predict multiple cancer phenotypes, including drug response, cancer type and cancer subtype.

1.3 THESIS OUTLINE

This thesis comprises of five chapters. **Chapter 2** reviews existing studies on multi-omic data integration by machine learning (literature review), and then incorporates an original benchmark analysis that summarises models' performance (Aim 1). **Chapter 3** describes analysis of the ProCan-SangerDepMap dataset (Aim 2) and the development of a novel deep learning model for proteomic biomarker discovery. **Chapter 4** introduces the novel transformer-based deep learning model DeePathNet, which integrates multi-omic data with cancer pathway knowledge to predict several cancer phenotypes (Aim 3).

Chapter 2: Machine learning for multi-omics data integration in cancer (Literature Review)

Text and figures included in this chapter are adapted from the following publication:

Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience*, 103798.

Statement of Contribution

The PhD Candidate completed all data analyses presented in this chapter, under the supervision of Dr. Rebecca C Poulos and Dr. Qing Zhong. The PhD Candidate was also responsible for writing this chapter and the preparation of all figures. Jia Liu also contributed to the writing of the introduction of this work.

2.1 SUMMARY

Multi-omics data analysis is an important aspect of cancer molecular biology studies and has led to ground-breaking discoveries. Many efforts have been made to develop machine learning methods that automatically integrate omics data. Here, we review machine learning tools categorized as either general-purpose or task-specific, covering both supervised and unsupervised learning for integrative analysis of multi-omics data. We benchmark the performance of five machine learning approaches using data from the Cancer Cell Line Encyclopedia, reporting accuracy on cancer type classification and mean absolute error on drug response prediction, and evaluating runtime efficiency. This review provides recommendations to researchers regarding suitable machine learning method selection for their specific applications. It should also promote the development of novel machine learning methodologies for data integration, which will be essential for drug discovery, clinical trial design, and personalized treatments.

2.2 INTRODUCTION

The discipline in molecular biology that aims for the collective characterization and quantification of the genome, transcriptome, and proteome, to influence the structure, function, and dynamics of a biological sample is termed omics (López de Maturana et al. 2019). Biotechnological advancements have enabled researchers to generate molecular datasets and perform individual or integrative analyses across various fields, such as genomics, transcriptomics, and proteomics (O'Donnell, Ross, and Stanton 2019).

In human cancers, there are complex rearrangements at the genetic, transcriptional, and proteomic levels that drive oncogenesis. This process evolves through clonal selection and over time, contributing to resistance to treatment. Single-omics datasets such as those derived from the Human Genome Project (Lander et al. 2001) and initial genomic profiling from The Cancer Genome Atlas (TCGA) projects (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020), have failed to produce the revolution in cancer treatment that was expected for the vast majority of common cancer types (Tannock and Hickman 2016). Next-generation sequencing of tumour genomes has been able to propose targeted treatments in only a small percentage of patients (Bohan et al. 2020), and no improvements in outcome have been found in randomised trials of targeted therapies (Le Tourneau et al. 2015). Consequently, developing a holistic view of cancer behaviour and identification of new therapeutic vulnerabilities may only be possible through multi-omics analysis, which has become an area of increasing interest in biological research over the last decades (I. Subramanian et al. 2020; B. Lee et al. 2019; Sathyanarayanan et al. 2020; Oh et al. 2021) (**Figure 1**). This has been exemplified by the addition of epigenomic, transcriptomic, proteomic, phosphoproteomic, and metabolomic data to the TCGA for

many solid tumour subtypes in recent years (Clark et al. 2019; L.-B. Wang et al. 2021). These large-scale integrative analyses on multi-omics data from various tumour cohorts have shed light on the complex systemic dysregulation associated with specific cancer phenotypes, producing essential insights that cannot be attained by examining only a single omics dataset. For example, a proteogenomic analysis of colon and rectal cancer showed moderate correlation between messenger RNA (mRNA) expression and protein abundance, and identified four cancer subtypes at the proteomic level to enable better prioritization of mutated (affecting DNA) or dysregulated (affecting RNA) cancer driver genes (B. Zhang et al. 2014). Another multi-omics study in an ethnically diverse lung adenocarcinoma cohort used machine learning to reveal four subgroups defined by mRNA transcripts, proteins, phosphoproteins, and acetylated proteins, with multiple potential therapeutic vulnerabilities to targeted therapy as well as immunotherapy resistance (Gillette et al. 2020).

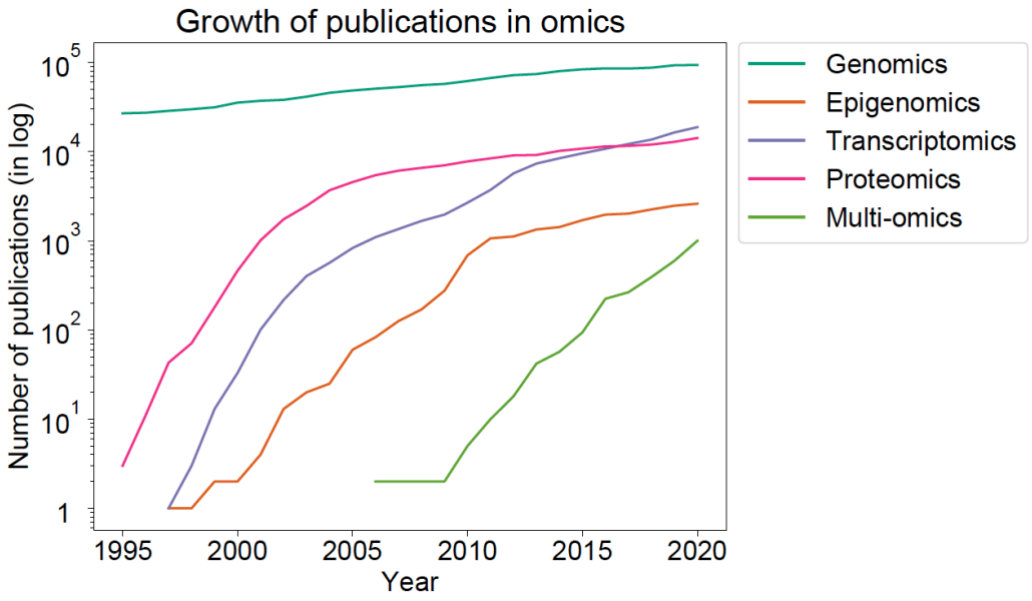


Figure 1: Growth of publications in omics. Line charts showing the number of articles published in each year from 1995 to 2020 in PubMed, coloured by different omics. The y-axis is plotted in log scale. Search terms used are "genomics", "epigenomics", "transcriptomics", "proteomics" and "multi-omics".

Given the immense complexity of data integration across multiple omics, the computational algorithms required to tease out signals from noise become more

complex. Therefore, strategies are required to systematically integrate heterogeneous multi-omics datasets to deliver actionable results that may advance biological sciences and eventually translate into clinical practice. There are three common strategies for multi-omics data integration: early, middle, and late integration (Rappoport and Shamir 2018). Early integration, also known as early concatenation, is a simple concatenation of features from each omics layer into one single matrix. In late integration, modelling and analysis are performed at each omics layer separately, and the results are merged at the end. The difference between early, middle, and late integration is also summarized in **Figure 2**. Because both early and late integration do not involve additional statistical processing or modelling by machine learning, all methods reviewed in this article fall under middle integration, which focuses on using machine learning models to consolidate data without concatenating features or merging results.

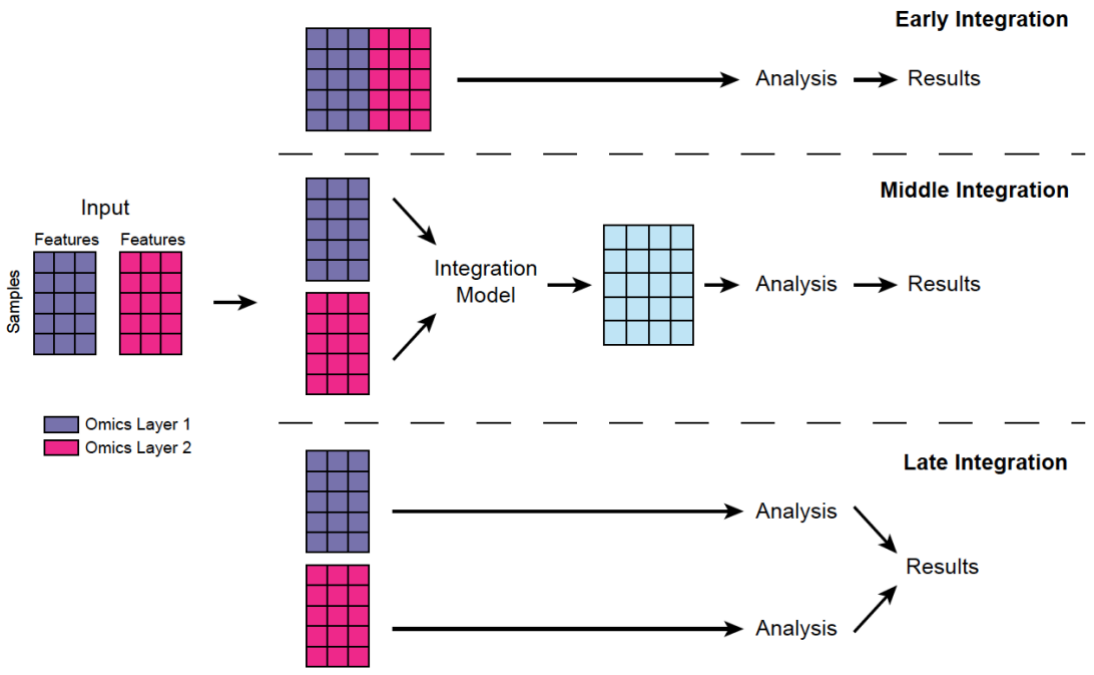


Figure 2: Illustration of early, middle and late integration for merging data matrices generated by different omics. In early integration, features from different data matrices are concatenated. Middle integration uses machine learning models to consolidate data without concatenating features or merging results. In late integration, each omics layer is analysed independently, and results are combined at the end.

The bedrock of multi-omics data analysis is machine learning, based upon which many tools have been developed (Argelaguet et al. 2020; Mo et al. 2018; Sharifi-Noghabi et al. 2019). Machine learning algorithms are trained to model complex patterns that cannot be accurately captured by traditional mathematical models in high dimensional data (Russell, Russell, and Norvig 2020). Publications of existing methods often emphasize the computational aspects of the proposed models, but lack a thorough introduction to the characteristics of individual omics. Recent reviews on multi-omics data integration focus on either biological applications or machine learning algorithms, rather than the combination of both (Nicora et al. 2020; Picard et al. 2021; Reel et al. 2021; I. Subramanian et al. 2020).

Published multi-omics data are usually stored in online portals for public access, serving as resources for both discovery and validation (**Table 1**). Among them is the TCGA project (Campbell et al., 2020) initiated by the National Cancer Institute in 2006, which generated multi-omics data for more than 20,000 tumors spanning 33 cancer types. The International Cancer Genome Consortium (ICGC) was initiated by multiple countries as a collaborative program, which incorporates some projects from TCGA and features a user-friendly online analysis interface (International Cancer Genome Consortium et al. 2010). The Catalog of Somatic Mutations In Cancer (COSMIC) (Iorio et al. 2016; Tate et al. 2019) database is led by the Wellcome Sanger Institute and curates multi-omics data for both cancer cell lines and tumors. The Cancer Dependency Map (DepMap) (Broad 2020) is a platform similar to COSMIC developed by the Broad Institute, which provides genome-wide CRISPR-Cas9 knockout screens with comprehensive multi-omics molecular characterization of cell lines and the corresponding drug screens.

Name	URL	Data types	Notes
TCGA	https://portal.gdc.cancer.gov/	• Genomics • Epigenomics • Transcriptomics	• Tumour data • Large coverage of tumours
ICGC	https://dcc.icgc.org/	• Genomics • Transcriptomics	• Tumour data • Powerful online analytics tools
CPTAC	https://cptac-data-portal.georgetown.edu/cptacPublic/	• Proteomics	• Tumour data • The largest proteomic data portal
COSMIC Cell Lines	https://cancer.sanger.ac.uk/cell_lines	• Genomics • Epigenomics • Transcriptomics • Drug response • CRISPR-Cas9 screen	• Cancer cell line data • Manually curated • Large coverage of cell lines
DepMap	https://depmap.org/portal/	• Genomics • Epigenomics • Transcriptomics • Proteomics • Drug response • CRISPR-Cas9 screen	• Cancer cell line data • Large coverage of omic types • Powerful online tools
COSMIC	https://cancer.sanger.ac.uk/cosmic	• Genomics • Epigenomics • Transcriptomics	• Tumour data • Manually curated • Focus on genomics • Overlap with other portals

Table 1. Key portals for accessing publicly available multi-omics datasets

The unique contribution of this review is three-fold (**Figure 3**). First, this review features a balance of both biological and technical content, so that readers from a range

of backgrounds can benefit from the presented information and guidance when they seek multi-omics integration tools for cancer research. Other similar reviews primarily focus on only one of these aspects or lack comprehensiveness. Second, we propose a new classification that categorizes the reviewed tools into general-purpose and task-specific. This allows researchers to quickly determine which tools are the most applicable for their research questions. In addition, researchers who do not have a strong computational background may be not aware that general-purpose methods can also be applied to their research projects. Third, unlike most review articles, we perform an independent benchmarking analysis using a publicly available dataset called Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al. 2019; Nusinow et al. 2020). The benchmarking exercise enables researchers to choose the most suitable tools for their research question and computational environment.

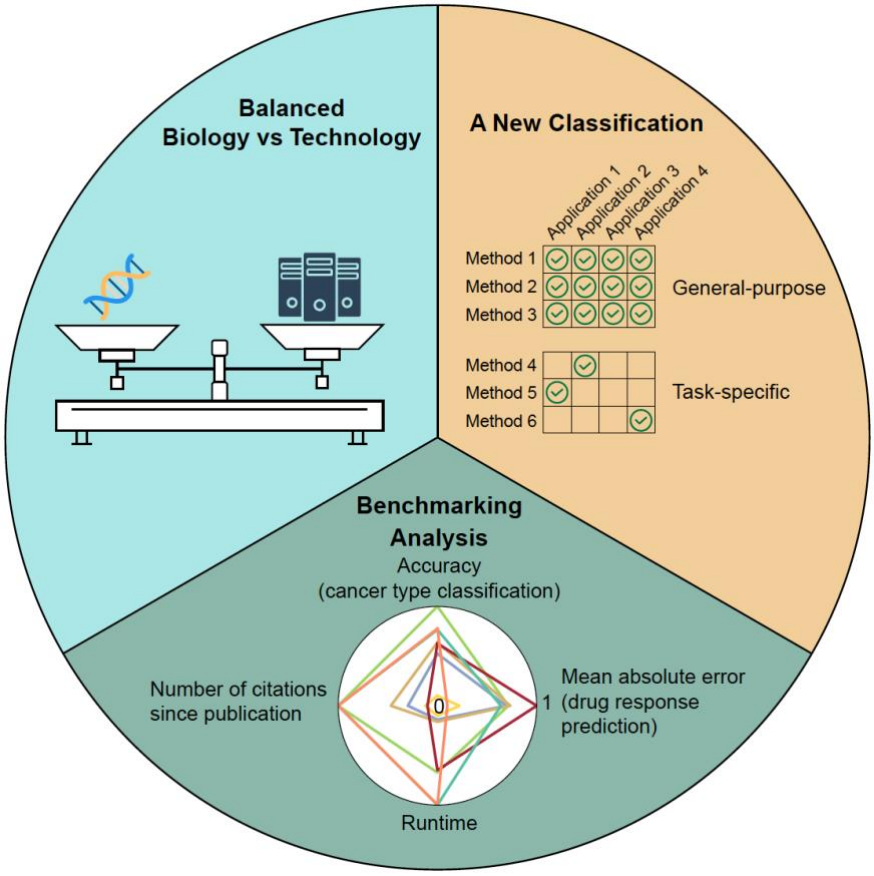


Figure 3: Unique contribution of this review. First, we describe a balance of both biological and technical content covering topics from genomics to proteomics and from machine learning to multi-

omics integration tools. Second, we propose a new classification that categorises the reviewed tools into two categories, namely general-purpose and task-specific, and then review these tools for four types of applications in biomedical sciences. Third, we provide an independent benchmarking analysis to compare integration methods for cancer type classification and drug response prediction.

2.3 OMICS OVERVIEW

Understanding the biological underpinnings of the data in each omics layer and the data formats is crucial to method development and fully utilizing the available tools. For instance, genomic and epigenomic variants influence gene regulation and the quantities of transcribed mRNA (Haraksingh and Snyder 2013). Splicing mechanisms and posttranslational modifications then impact the downstream measurements of the proteome. All of these mechanisms ultimately determine the cellular phenotype (Niklas et al. 2015). In this section, we review several common omics, describe data formats, and discuss corresponding analytical strategies.

2.3.1 Genomics

Genomics examines DNA sequences and seeks to understand the associations between diseases and genomic alterations (Stratton, Campbell, and Futreal 2009). Whole-exome sequencing (WES) and whole-genome sequencing (WGS) (Schwarze et al. 2018) are two popular technologies utilized in genomic studies. WES mostly examines the exonic (mRNA-coding) portion of the genome, whereas WGS aims to examine all nucleotides in the genome including the gene regulatory regions (Nakagawa and Fujita 2018). WES usually involves a lower cost than WGS because it only covers the coding regions, although key regulatory and splice-site mutations that are not in coding regions could be missed by WES (Y. Yang et al. 2013). Genomic analysis focuses on single nucleotide variants (SNVs), insertions and deletions (INDELs), structural variation (SV), and copy number variation (CNV). SNVs are variants of only a single nucleotide that occurs at a specific genomic position. INDELs are small genetic variations with lengths usually shorter than 10,000 nucleotides. SV

covers large variations in the chromosome, including deletions, duplications, insertions, inversions, and translocations of long nucleotide sequences. CNV is a particular form of SV and usually involves the amplification or deletion of a large region of a chromosome. The genome is the most fundamental layer of genetic information and is very well characterized because of the development of advanced sequencing technologies. Genomic analyses have revealed many significantly mutated genes in cancer, known as cancer driver genes, such as *TP53* and *KRAS* (Stratton, Campbell, and Futreal 2009). Novel treatment strategies that have greatly improved outcome for subsets of cancers have been discovered by analysing mutations at the genomic level (Behan et al. 2019). For example, EGFR tyrosine kinase inhibitors are used in the treatment of EGFR-mutant non-small cell lung cancer, whereas HER2-amplified breast cancers are treated with HER2 monoclonal antibodies (Cohen, Cross, and Jänne 2021).

For computational analysis, the data matrices for SNVs, INDELs, and SVs can be summarized as binary values that indicate whether a gene is mutated or wild type. Genes are often filtered so that only those with mutations in sufficient numbers of samples are included to avoid a highly sparse matrix, and mutation frequencies are typically normalized against background mutation rates for that genomic locus (M. S. Lawrence et al. 2013). Filtering may also be required for specific mutation types, depending on the research question. For example, only missense mutations may be considered as mutants in certain situations. By contrast, CNV data are typically presented as a matrix of either counts or continuous values for each gene. The data matrix of CNV is sometimes in the format of log fold-change for cancer studies, reflecting changes of copy numbers compared with the normal ploidy. Outside of coding regions, genomic information can be used to understand elements of gene

regulation and dysregulation in cancer (Andersson and Sandelin 2020). Somatic, germline, and epigenetic variation affecting these regions can have profound effects on gene expression in cancer (Poulos and Wong 2017). Although genomic technology is relatively more mature than other omics, discovering causal relationships in addition to associations remains as one of the biggest challenges (McGuire et al. 2020).

2.3.2 Epigenomics

The epigenome encompasses the set of indirect chemical modifications of nucleotides and proteins that regulate how genes are expressed, without changing the actual nucleotide sequence itself (K. C. Wang and Chang 2018). The study of the epigenome is called epigenomics, which involves investigating DNA methylation (Jost and Saluz 2013) and histone modification (Seligson et al. 2005), as well as understanding the three-dimensional structure of DNA, which is influenced by topologically-associating domains (Szabo, Bantignies, and Cavalli 2019). This three-dimensional structure is examined via sequencing technologies, such as ChiA-PET, 3C, 4C, 5C, and Hi-C (van Berkum et al. 2010), whereas DNA methylation can be measured by a range of methods, such as bisulfite sequencing (Krueger et al. 2012; Wreczycka et al. 2017). Chromatin immunoprecipitation sequencing (ChIP-seq) (Valouev et al. 2008) experiments are often used for high-throughput measurement of histone modifications.

Methylomics is one of the best characterized aspects of epigenomics. It focuses primarily on the effects of promoter DNA methylation on silencing gene expression, but also commonly examines the effects of gene body methylation in cancer (Wong et al. 2014). Studies of DNA methylation play a pivotal role in biomedical research. For example, the promoter hypermethylation of MLH1 was found to result in hereditary nonpolyposis colon cancer (Cunningham et al. 1998). The processed methylation data

matrix often contains continuous values ranging from 0 to 1, representing the proportion of cells in which the relevant nucleotide is found to be methylated. Methylation data usually require normalization and correction according to cancer types (Iorio et al. 2016). However, current high-throughput technologies for measuring methylomic data use probes, which may not properly cover the promoter regions of specific genes. This issue may be especially important if the research question is about one specific gene.

2.3.3 Transcriptomics

Genes are mostly transcribed into mRNA and introns are spliced out, leaving only exons in the mature mRNA, which consists of 5′ and 3' untranslated regions, and a protein-encoding open reading frame (Brouwer and Lenstra 2019). The result is a large pool of cellular mRNA used by ribosomes for translation to proteins. This provides an indirect indicator of the protein expression in a cell, or the activity of the genome at a particular point in time. Transcriptomic analyses measure the abundance of the complete set of mRNA transcripts of each gene, which is also referred to as the gene expression level. Several methods are available to quantify the transcriptome, with the most popular approaches being microarrays and RNA-seq (Malone and Oliver 2011). RNA-seq is now more commonly employed than microarrays, as it provides better performance and data consistency (X. Xu et al. 2013). Biomarkers found at the transcriptomic level can be used for identifying patient subtypes (Nielsen et al. 2010) and developing new cancer treatments. For example, the transcriptome can be more predictive of anticancer drug response than genomic mutations and DNA methylation data (Iorio et al. 2016).

2.3.4 Proteomics

mRNAs are translated to produce proteins, which are sequences of amino acids (Kaeberlein and Kennedy 2007). Analogous to the transcriptome, the proteome encompasses the entire set of expressed proteins in a cell or organism at a particular point in time. Proteins are the functional units that interact with other molecules like metabolites, lipids, or nucleotides (Spirin and Mirny 2003). For this reason, along with the metabolome and lipidome, the proteome is more closely related to cellular phenotypes than the genome or transcriptome (Crick 1970). Protein abundance often differs from gene expression levels because of a number of factors, including posttranslational modifications and protein stability or degradation (Hegde, White, and Debouck 2003). For example, in clear cell renal cell carcinomas, genes related to oxidative phosphorylation-related metabolism, protein translation processes, and some phospho-signaling modules were found to be dysregulated only at the protein level (Clark et al. 2019).

Proteomics refers to the large-scale analysis of proteomes. Recent technological advances in mass spectrometry have enabled proteomics to become high-throughput and reproducible for large-scale cancer analyses (Tully et al. 2019; Poulos et al. 2020). Normalized and imputed proteomic data matrices typically contain continuous values that are usually in the logarithmic scale. However, the high prevalence of missing values of protein abundance in proteomic data presents unique analytical challenges (Poulos et al. 2020). The missing values need to be handled by applying domain-specific knowledge, statistical methods, or machine learning algorithms (Poulos et al. 2020; Välikangas, Suomi, and Elo 2018). Normalized and imputed proteomic data matrices typically contain continuous values that are usually in the logarithmic scale.

2.3.5 Single-cell sequencing

Most omics studies are based on data generated from bulk samples. Therefore, the omics data are averaged measurements of multiple cells from samples. However, tumor samples exhibit a great degree of intertumoral and intratumoral heterogeneity, making analyses extremely challenging (Guo et al. 2018). The advancement of single-cell technologies has enabled researchers to investigate omics profiles at a single-cell level by tagging each cell from a sample (Nam, Chaligne, and Landau 2021), although single-cell analyses face a number of challenges that are distinct from analyses of bulk samples because of the difference of data resolution. Among the different omics data types, single-cell RNA sequencing (scRNA-seq) is relatively more mature and has yielded a number of discoveries (Argelaguet et al. 2019; Yijie Zhang et al. 2021). In the meantime, computational tools for integrating single-cell multi-omics data have been emerging. For example, Seurat 4.0 is able to integrate data from multiple single-cell technologies, including single-cell epigenomic, transcriptomic, and proteomic data (Hao et al. 2021). In this review, we focus on multi-omics data integration for bulk samples.

2.4 MACHINE LEARNING FOR MULTI-OMICS INTEGRATION

Having reviewed different omics, here we discuss core machine learning concepts that are involved in multi-omics data integration. Machine learning is the study of computational algorithms that make predictions and decisions based on experience and data, without having been explicitly programmed to do so (Koza et al. 1996). Machine learning models are generally divided into three categories: supervised, unsupervised, and reinforcement learning. Supervised learning methods use label information from samples for model training, whereas unsupervised learning infers patterns directly from unlabelled data. Reinforcement learning trains an agent to

`choose the best next action when the environment changes. Supervised and unsupervised learning are commonly applied to multi-omics data integration, but to the best of our knowledge, there have been no attempts to use reinforcement learning for this type of task.

As mentioned above, there are three approaches for multi-omics data integration: early, middle, and late. The most straightforward strategy is early concatenation. However, having a vast number of features while the number of available data points is low, known as the "curse of dimensionality" (Bellman 1966), is a particular challenge for the use of early concatenation in multi-omics integration. For example, with the human genome containing more than 20,000 protein-coding genes, multi-omics datasets can easily comprise more than 50,000 features when the genome, transcriptome, and proteome are combined. By contrast, the number of available tumor samples in a dataset is often relatively small, with cancer cohorts typically comprising no more than a few hundred patients. Late integration only involves manual combination of the results from each omics layer, hence out of scope for this review. Therefore, we focus on middle integration that aims to overcome the challenge by using machine learning integration methods that are categorized either as general-purpose or task-specific (**Figure 3**). General-purpose methods couple dimensionality reduction with different downstream algorithms for a variety of applications, whereas task-specific methods are end-to-end models designed for one specific task.

In this section, we first provide a brief introduction to related machine learning concepts, followed by a more comprehensive review on existing general-purpose and task-specific methods for multi-omics data integration.

2.4.1 Basics of related machine learning concepts

Unsupervised learning

Unsupervised learning discovers patterns in multi-omics data without mapping the input data to output data. Most dimensionality reduction techniques are unsupervised methods. Principal component analysis (PCA) (Wold, Esbensen, and Geladi 1987) projects each data point onto a lower-dimensional space by creating orthogonal principal components that are eigenvectors of the data's covariance matrix. Factor analysis (A. P. Singh and Gordon 2008) can also be used to reduce dimensionality, assuming the existence of latent (unobserved) variables that are not limited to linear combinations of features. Factor analysis algorithms then seek such latent variables that can capture the common variance of the whole dataset. Joint latent variable models (Everett 2013) extend factor analysis by allowing more assumptions and configurations on the statistical models. Canonical correlation analysis (CCA) (Hotelling 1992) calculates how well different data matrices are correlated and derives a set of variables such that the correlations between data matrices are maximized. CCA can be applied to multi-omics data integration under the assumption that the correlation between different omics layers is to be maximized. Multiple kernel learning (MKL) (Y.-Y. Lin, Liu, and Fuh 2011) can be used with either supervised or unsupervised learning. Kernels allow the data to be transformed into a higher-dimensional space via kernel tricks (Aizerman 1964). Multiple kernels are used in MKL so that data from different omics layers can be appropriately modelled.

Supervised learning

Supervised learning is usually used to make predictions. Given the input data and the output labels, supervised learning finds a mapping function that maps the input data to the label information. Label information in cancer research can be any phenotype of interest. For example, cancer types can be considered as label

information. If the label information is discrete, then it is called classification. If the label is continuous, then it is regression. Linear regression and logistic regression are the two basic supervised learning models that only use linear predictor functions (Freedman 2009). Elastic net (Zou and Hastie 2005) adds both L_1 (absolute-value norm) and L_2 (Euclidean norm) regularization terms to the basic linear models. L_1 regularization encourages non-informative features to have zero coefficients, and L_2 regularization works well with correlated features by allocating roughly equivalent weights to strongly correlated features. Because of these characteristics, elastic net has been used widely in multi-omics analyses such as drug response studies (Iorio et al. 2016), where simple markers are preferred and interpretability is important. Random forest (Breiman 2001) uses a set of decision trees (Rokach and Maimon 2006) to make predictions based on votes over all the trees in the forest. Random forest is a nonlinear machine learning model that captures more complexity in the data than linear and logistic regression. The feature importance given by random forest represents how well each feature performs in terms of prediction, allowing researchers to prioritize the most important features for their studies. Neural networks are the root of deep learning algorithms that have attracted increasing attention recently and shown better predictive power than other traditional machine learning models in research areas such as natural language processing, computer vision, and biomedical sciences (K. G. Kim 2016). Neural network models are versatile because they can be used for various purposes, including classification, regression, dimensionality reduction, and missing value imputation. Despite its superior predictive performance, deep learning is often criticized for its poor model interpretability. To overcome this limitation, deep learning algorithms that focus on model explanations have emerged in recent years (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016; Ullah et al. 2020).

2.4.2 Multi-omics data integration tools and their applications

In this section, we review machine learning tools that have been developed specifically for biomedical sciences and discuss how these tools are applied in three typical real-world applications, namely cancer type classification, drug response prediction, and patient stratification. Specifically, we used keywords "machine learning" and "multi-omics integration" to broadly search the methods using PubMed and Google Scholar. We then searched for existing reviews of similar topics. For each tool, we indicate whether it is general-purpose or task-specific, and whether it provides an easy-to-use interface for custom datasets (**Table 2**). Although no multi-omics dataset covers all omics layers, most integration tools are flexible enough to support the data analysis of any combination of available omics layers (**Figure 3**).

Cancer type classification

Knowing the cancer subtype is crucial for disease classification, assessing prognosis, and planning treatment. For instance, breast carcinoma can be classified into five transcriptome-based subtypes that lead to different treatment responses and outcomes (Dai et al. 2015). Non-small cell lung cancer can also now be classified into a large number of subtypes depending on both histological appearance (squamous versus adenocarcinoma) and the presence/absence of particular driver mutations (Thomas et al. 2015). It would also be valuable to predict the cell of origin, and therefore the likely therapeutic sensitivity, for cancers of unknown primary (CUP) (M. Y. Lu et al. 2021; Pavlidis and Pentheroudakis 2012), as treatments differ significantly depending on the patient's primary cancer type. Traditional methods for classifying cancer types mainly involve visual inspection by trained anatomical pathologists of cancer sections stained by H&E or by immunohistochemistry. Machine learning tools that integrate multi-omics data have provided more efficient diagnoses for patients

with CUP, allowing the most effective treatment options to be identified (Bavafaye Haghighi et al. 2019). Relevant machine learning packages include the following.

mixOmics (general-purpose) (Rohart, Gautier, Singh, and Cao 2017) is a well-implemented package enabling a set of supervised and unsupervised machine learning models based on linear discriminant analysis (LDA) (R.O. Duda, P.E. Hart, and D. Stork 2000), PCA, and CCA. mixOmics has been used extensively in studies to characterize cancer subtypes. It supports three modes: single omics analysis, data integration across different omics layers, and data integration across different samples. One specific multi-omics integration method proposed in the mixOmics package is DIABLO (A. Singh et al. 2019). PAM50 breast cancer subtype prediction was selected for the classification task, and colon cancer data from TCGA were used for clustering comparisons. The authors tested DIABLO via two model configurations. DIABLO_full represents correlations across all omics layers, which are modeled via CCA, whereas DIABLO_null does not consider correlations between omics layers by simply applying LDA on the dataset as early concatenation. Notably, DIABLO_null performed better than DIABLO_full in both classification and clustering tasks, suggesting that integration may not be effective because the connection of omics layers actually worsens the predictive accuracy. Apart from cancer type classification, mixOmics has been used for tasks such as analysing the association between gut microbial composition and the risk of asthma in childhood (Stokholm et al. 2018).

MOFA/MOFA2 (**M**ulti-**O**mics **F**actor **A**nalysis, **general-purpose**) (Argelaguet et al. 2018, 2020) seeks common factors that can explain the greatest variance of all data from different omics layers. MOFA (Argelaguet et al. 2018) uses factor analysis as the method of dimensionality reduction. It supports different data distributions and optimizes computational runtime for better performance in multi-omics integration.

Various types of regularizations are supported in MOFA for better model interpretability. MOFA also handles missing values automatically and supports partial datasets. In their study, MOFA was able to classify four subtypes of chronic lymphocytic leukemia (Argelaguet et al. 2018). (Alcala et al. 2019) used MOFA on DNA methylation and gene expression data, and identified two latent factors that were associated with survival in large-cell neuroendocrine carcinomas. Apart from survival analysis, MOFA showed satisfactory results in other tasks, such as drug response prediction and patient stratification (Argelaguet et al. 2018). MOFA+/MOFA2 (Argelaguet et al. 2020) is a subsequent enhancement of MOFA that supports single-cell datasets and GPU-accelerated model inference, which is over twenty times faster than the original inference algorithm in MOFA. Although MOFA2 mostly highlighted its usage in analyzing single-cell datasets, bulk sample analyses also benefited from the faster inference of MOFA2 without any loss of functionalities (Argelaguet et al. 2020).

sCCA (sparse CCA, **general-purpose**) is a variation of CCA that imposes additional penalties in modeling so that the number of latent variables can be kept low for better interpretation (Rodosthenous, Shahrezaei, and Evangelou 2020). Several conventional CCA methods were compared with a customized sCCA variation that allowed data from more than two datasets to be used at the same time. The proposed sCCA variation was applied to TCGA gene expression, miRNA, and methylation data, and yielded the highest accuracy in a task of classifying three cancer types, namely breast, kidney, and lung cancer (Rodosthenous, Shahrezaei, and Evangelou 2020).

Drug response prediction

Using multi-omics profiles to predict drug responses allows researchers to discover new treatment opportunities and to provide recommendations on the design

of early-phase clinical trials for either novel drugs or the repurposing of existing drugs for different cancers (Iorio et al. 2016). In this scenario, the problem can be either formulated as a regression task, where a model is trained to predict the half-maximal inhibitory concentration (IC_{50}) value and area under the curve (AUC), or as a simplified classification task where the model predicts whether a given input is sensitive or resistant to a particular drug. Computational researchers have been focusing on designing machine learning models that are able to uncover explainable biomarkers with better prediction, facilitating personalized treatment.

Name	Model	Programming Language	API for custom data	Can handle missing values	Publication year	Citations to date	Source Code
				General-purpose			
MOFA2/MOFA	Matrix factorisation	R	Yes	Yes	2020/2018	77 / 295	https://github.com/bioFAM/MOFA2 https://github.com/bioFAM/MOFA
sCCA	CCA	R	No	No	2020	5	https://github.com/theorod93/sCCA
DIABLO	CCA/LDA	R	Yes	Yes	2019	140	http://mixomics.org/
web-rMKL	Multi-kernel	Web-interface	Yes	No	2019	1	https://web-rmkl.org/home/upload/
iClusterBayes / iClusterPlus / iCluster	Bayesian model	R	Yes	No	2018/ 2013/ 2009	76 / NA / 206	https://www.bioconductor.org/packages/devel/bioc/html/iClusterPlus.html
moCluster	Bayesian model	R	Yes	No	2016	49	https://www.bioconductor.org/packages/release/bioc/html/mogsa.html
sGCCA	CCA	R	Yes	No	2014	134	https://cran.r-project.org/web/packages/RGCCA/index.html
JIVE	Matrix factorisation	R	Yes	No	2013	331	https://cran.r-project.org/src/contrib/Archive/r.jive/
DeepCCA	Deep learning + CCA	Python	No	No	2013	73	https://github.com/VahidooX/DeepCCA

Task-specific							
Clustering							
NEMO	Affinity clustering	R	Yes	No	2019	50	https://github.com/Shamir-Lab/NEMO
Similarity Network Fusion (SNF)	Network-based	R/MATLAB	Yes	No	2014	980	http://compbio.cs.toronto.edu/SNF/SNF/Software.html
Drug response prediction							
MOLI	Deep learning	Python	No	No	2019	78	https://github.com/hosseinshn/MOLI
CaDRReS	Recommender System	Python 2	No	No	2018	49	https://github.com/CSB5/CaDRReS
HNMDRP	Network-based	R/MATLAB	No	No	2018	47	https://github.com/USTC-HIlab/HNMDRP
DRLP	Network-based	MATLAB	No	No	2017	52	http://compbio.case.edu/omics/software/drlp/index.html

Table 2 Machine learning tools for multi-omics data integration

MOLI (**M**ulti-**O**mics **L**ate **I**ntegration, **task-specific**) (Sharifi-Noghabi et al. 2019) is based on deep learning and predicts drug response as a classification task. MOLI uses a deep neural network (Rumelhart, Hinton, and Williams 1986) as a feature extractor for each omics layer, and concatenates the last hidden layers with a triplet loss (Schroff, Kalenichenko, and Philbin 2015) at the end to train the model. Triplet loss facilitates training by minimizing distances between similar samples and maximizing distances between different samples. Despite the "late integration" in its name, MOLI would be better classified as middle integration because MOLI integrates all omics layers using machine learning instead of merging results at the end. MOLI used gene mutation, expression, and copy number to predict cancer drug response by classifying patients as responders or nonresponders to cancer drugs. Various datasets were utilized, including GDSC (Iorio et al. 2016), PDX (Gao et al. 2015), and TCGA (Ding, Zu, and Gu 2016) drug response data. The number of samples included in each dataset for MOLI varied depending on the drugs, ranging from 16 to 856 samples. MOLI was only compared with early concatenation in its related publication (Schroff, Kalenichenko, and Philbin 2015), without being thoroughly compared with similar integration tools.

CaDRReS (**C**ancer **D**rug **R**esponse prediction using a **Re**commender **S**ystem, **general-purpose**) (Suphavilai, Bertrand, and Nagarajan 2018) is a matrix-factorization-based model that treats drug response prediction as a regression task. CaDRReS was developed based on recommender systems, where matrix factorization has shown excellent performance. The fundamental intuition in the recommender system is that if a set of users rate a set of products similarly, they are also likely to give similar ratings on other products (Xue et al. 2017) This idea has been transformed into multi-omics data integration, where the molecular profile of a gene is considered

to share similarities across different omics layers. CaDRReS was used to predict continuous drug responses IC_{50}, using both GDSC (Iorio et al. 2016) and CCLE datasets (Barretina et al. 2012). CaDRReS has been benchmarked against two other matrix-factorization-based methods, as well as some basic models such as regularized linear regression (Suphavilai, Bertrand, and Nagarajan 2018). Although CaDRReS is a general-purpose method with dimensionality reduction, the tool does not allow flexible usages for purposes other than drug response prediction.

HNMDRP (**H**eterogeneous **N**etwork-based **M**ethod for **D**rug **R**esponse **P**rediction, **task-specific**) (F. Zhang et al. 2018) focuses on constructing similarity networks among cell lines, drug structures, and drug target genes to predict drug responses. This method assumes that when a similar cell line is treated with a similar drug, then the drug response should be similar. Application of HNMDRP showed that protein-protein interactions and drug-target interactions were useful to improve prediction results, but drug chemical structure data were not widely accessible in many scenarios.

pairwiseMKL (multiple pairwise kernels for drug bioactivity prediction, **general-purpose**) (Cichonska et al. 2018) is an enhanced extension to MKL. It improves upon both runtime and memory efficiency of MKL and enables its application to drug response prediction, where the term "pairwise" refers to sample and drug pairs. The original study showed that pairwiseMKL ran approximately six times faster and consumed 95% less memory than KronRLS-MKL (Nascimento, Prudêncio, and Costa 2016), when predicting continuous IC_{50} values of drugs in the GDSC (Iorio et al. 2016) dataset.

Patient stratification

Another key task for which researchers use multi-omics data is to cluster tumours and identify potential new cancer subtypes to facilitate better patient stratification (Li and Wong 2019). These new subtypes could show different characteristics from existing known cancer subtypes, and new therapies may need to be developed for them. Computational methods for patient stratification are usually evaluated using simulated data, and methods that are able to clearly separate known phenotypes are generally considered high-performance methods (Argelaguet et al. 2018; Shen, Olshen, and Ladanyi 2009).

iCluster/iClusterPlus/iClusterBayes (**general-purpose**), are a family of machine learning methods based on joint latent variables. iCluster formulates the latent cancer subtypes as a joint variable, which results in a much smaller number of dimensions than early concatenation. The latent variables are modelled to capture common information from different omics layers. The iCluster study (Shen, Olshen, and Ladanyi 2009) shows that the best performance is achieved with less than ten dimensions in the latent variable. Based on iCluster, iClusterPlus (Mo et al. 2013) focuses on modeling different statistical distributions for discrete data types. iClusterBayes (Mo et al. 2018) is the latest version in the series and implements a fully Bayesian inference algorithm, which runs six times faster than iClusterPlus. Bass et al. used iCluster and discovered four subtypes of gastric cancer, enabling better patient stratification and treatment planning (Cancer Genome Atlas Research Network 2014).

moCluster (**general-purpose**) (Meng et al. 2016) is also a joint latent variable-based machine learning model, and was benchmarked against iCluster and iClusterPlus. The main difference between moCluster and the iCluster series is the method of finding latent variables. Instead of using an expectation-maximization

algorithm (Moon 1996), moCluster uses consensus PCA (CPCA) (Westerhuis, Kourti, and MacGregor 1998) to estimate the latent variables. CPCA is a variation of the typical PCA algorithm and allows modelling data from different groups, which can be naturally mapped to different omics layers. moCluster was shown to run 100 to 1,000 times faster than iCluster/iClusterPlus on a simulated dataset with better clustering performance (Meng et al. 2016). On a multi-omics dataset for the NCI-60 cancer cell lines (Gholami et al. 2013), moCluster was able to separate melanoma cell lines from the remaining cell lines, whereas iClusterPlus could not.

SNF (**S**imilarity **N**etwork **F**usion, **task-specific**) (B. Wang et al. 2014) is a network-based machine learning model developed for patient stratification as well as survival analysis. SNF focuses on patient similarity networks and uses a specific network fusion algorithm to iteratively update similarity networks for each omics layer, with information from other omics layers so that the similarity networks become more consistent. The fused network contains information from all omics layers, thus enabling multi-omics data integration. SNF is similar to HNMDRP, as both methods are based on similarity networks (Heckerman 1990). However, SNF only uses molecular profile information, whereas HNMDRP requires drug chemical structure data and drug target information. Despite the similarity between the two methods, no comparison was made for these two methods. SNF revealed two clusters in pancreatic ductal adenocarcinoma using epigenomics and transcriptomics data, demonstrating potential personalized treatment opportunities (Cancer Genome Atlas Research Network. and Cancer Genome Atlas Research Network 2017).

NEMO (**NE**ighborhood based **M**ulti-**O**mics clustering, **task-specific**) (Rappoport and Shamir 2019) provides enhancements for other similarity-based methods, including SNF and HNMDRP. NEMO optimized the similarity inference

algorithms to enable faster runtime and added support for partial datasets (i.e., datasets in which a sample may not have the same coverage across all the omics). In the study, gene expression, DNA methylation, and miRNA expression data from ten cancer types in the TCGA dataset (Ding, Zu, and Gu 2016) were included as the input. NEMO was able to identify patient subgroups that showed significant difference in terms of survival, and achieved superior performance compared with another nine clustering algorithms (Rappoport and Shamir 2019). NEMO not only achieved comparable performance to several other multi-omics integration methods, but its interface is also more user-friendly. Evaluated on ten different cancer datasets, NEMO runs approximately 400 times faster than iClusterBayes and 20% faster than SNF (B. Wang et al. 2014).

2.5 BENCHMARKING

Previous sections about various omics data, fundamental machine learning concepts, and integration tools provide basic knowledge for researchers to perform their multi-omics data integration analysis. To facilitate the decision on which machine learning tools are suitable for specific applications, we performed a comparative analysis for two baseline methods (early concatenation and PCA) and five multi-omics data integration tools (MOFA2 v1.1, DIABLO v6.17.15, iClusterPlus v1.22.0, iClusterBayes v1.22.0, and moCluster v1.20.0) by using a common CCLE dataset. These five methods were selected from Table 2 because they satisfied the following criteria. First, the method provides a software package that allows users to apply the analysis to custom datasets. Second, the method has at least 10 citations. Third, the method was published within the past five years (**Figure 4A**). Because patient stratification aims at discovering new molecular subtypes, which are defined a priori,

we only included cancer type classification and drug response prediction as the two specific applications (**Figure 4B**).

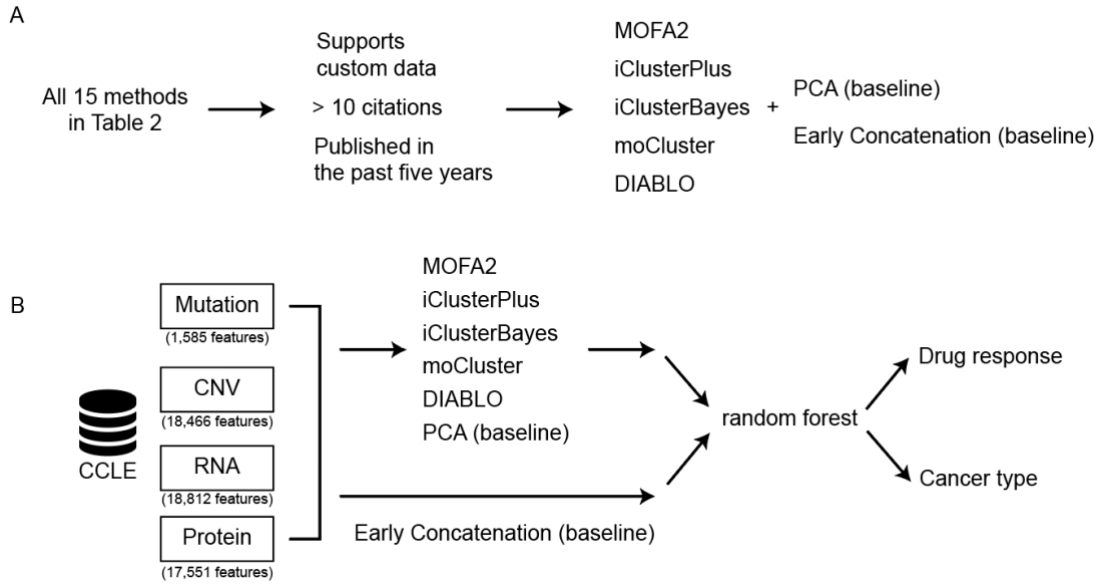


Figure 4: Details of the benchmarking analysis. A, The process of determining the scope of the benchmarking analysis. B, An overview of the steps included in the benchmarking analysis.

CCLE multi-omics data were retrieved from the DepMap portal (version 20Q2), which consists of information from four omics layers (Ghandi et al. 2019; Nusinow et al. 2020), including WES mutation, CNV, gene expression (RNA), and protein abundance. To compare the five general-purpose tools with each other on an equal footing and report both prediction accuracy and runtime efficiency, we selected random forest as the downstream algorithm, because general-purpose tools do not directly predict drug responses and cancer types (**Figure 4B**). Specifically, random forests with 500 trees were fitted using five-fold cross-validation, which was repeated 100 times by the Monte Carlo method to obtain a robust evaluation. Source code for this benchmarking analysis is provided in GitHub https://github.com/CMRI-ProCan/MODIBenchmark.

2.5.1 Cancer type classification

In the CCLE dataset, the number of features for WES mutation, CNV, gene expression (RNA), and protein abundance are 1,585, 18,466, 18,812, and 17,551, respectively (**Figure 4B**). The WES data were sequenced with 76-bp pair-end reads by Illumina HiSeq 2000 or Illumina GAIIX. The tool Picard was used to process the data, and mutation calling was done by MuTect. The copy number information was derived from WES data by ABSOLUTE. The methylation data were generated using reduced representation bisulfite sequencing and aligned by Bismark. For RNA-seq data, STAR was used for alignment and the read counts were normalized to transcripts per million bases. The protein data were generated by TMT10plex (Ghandi et al. 2019; Nusinow et al. 2020). No additional feature selection was performed to ensure unbiased downstream analysis. Only cell lines measured with all four omics layers were considered, because not all methods support partial datasets. A total of 217 cell lines from nine different cancer types were included in this benchmark comparison and CCLE metadata were used as the ground truth. Accuracy is chosen to be the evaluation metric, which is defined as the number of samples that are correctly predicted divided by the total number of samples.

DIABLO showed the highest accuracy in cancer type classification, with an average of 78.7% (**Figure 5A**). DIABLO is the only supervised integration tool, which uses tissue types as label information. Notably, iCluster series, moCluster, and MOFA2 did not perform better than PCA, which was not specifically developed for biomedical applications. MOFA2 and iClusterBayes had lower accuracy than early concatenation, suggesting no real improvement was gained from the statistical integration. In summary, our benchmarking study indicates that DIABLO is the most appropriate tool when label information such as tissue type is available, and general

dimensionality reduction methods are most appropriate for cancer type prediction when unsupervised integration is intended.

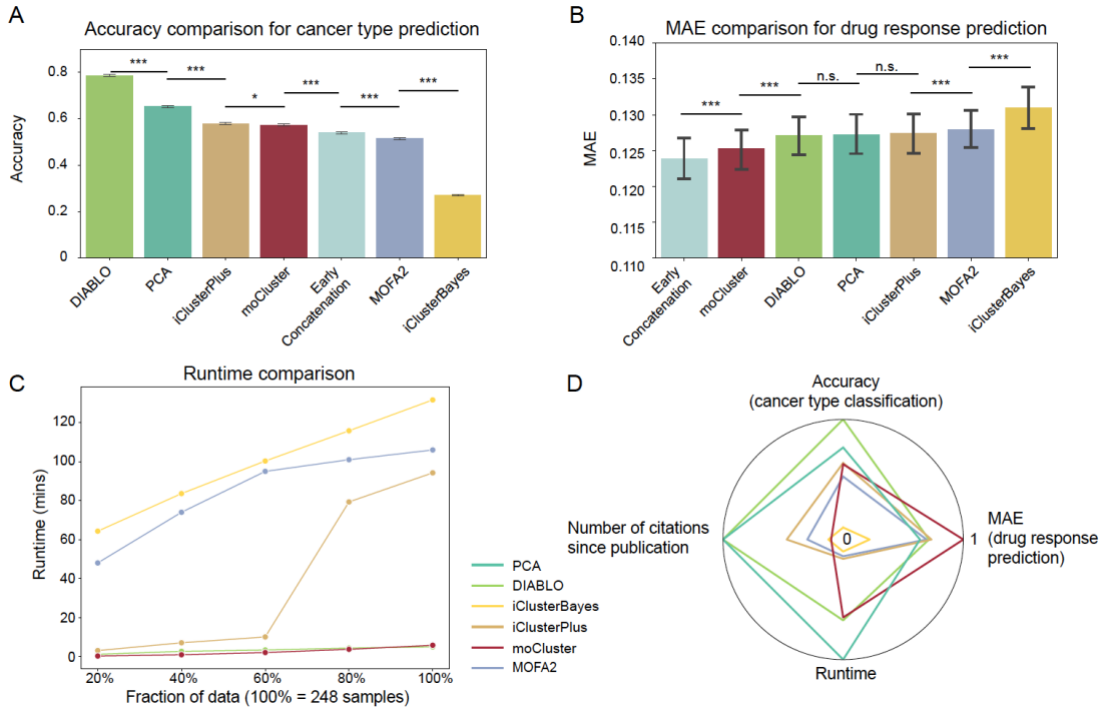


Figure 5: Benchmarking of machine learning-based integration tools using the CCLE multi-omics data. A. Accuracy of each method for cancer type prediction, showing standard errors of the mean derived from 100 runs of five-fold cross-validation, totalling 500 experiments (* signifies p-value < 0.05 and *** signifies p-value < 0.001 by an unpaired two-tailed Student's t-test). B. MAE comparison for drug response prediction across 1,448 drugs, error bars representing standard errors of the mean (*** signifies p-value < 0.001 and n.s. stands for not significant by an unpaired two-tailed Student's t-test). C. Runtime comparison. PCA is omitted as the runtime was negligible compared with the five multi-omics integration methods. D. A summary of the benchmarking study, derived from the results of cancer type prediction, drug response prediction (MAE between the measured AUC and predicted AUC), runtime comparison and the number of citations since publication. The number of citations for PCA was set to the maximum for better visualisation and because of its widespread use. The inverse of the runtime and drug response prediction MAE values are plotted so that higher values indicate better performance in all dimensions, and all values are plotted in the range of 0 to 1 in the radar plot.

2.5.2 Drug response prediction

For drug response prediction, we used the PRISM repurposing secondary screening dataset as the target, covering 1,448 drugs across 248 cell lines (Corsello et al. 2020). The area under the curve (AUC) of plasma concentration of a drug versus time after dosage (Scheff et al. 2011) was used as the drug response measurement. Mean absolute error (MAE) between the measured AUC and predicted AUC was

chosen as the evaluation metric, where MAE measures the arithmetic average of the absolute errors across all the samples.

The performance difference between models was relatively small for drug response prediction when compared with cancer type prediction (**Figure 5B**). Early concatenation yielded the best MAE of 0.1239, followed by moCluster with an MAE of 0.1252. This is similar to a previous study (A. Singh et al. 2019), where early concatenation also outperformed DIABLO. No evident difference was observed between DIABLO, PCA, MOFA2, and iClusterPlus. iClusterBayes produced the worst prediction, with an average MAE of 0.1309 and 0.1311, respectively. Therefore, for drug response prediction, early concatenation should be prioritized if computational resources suffice.

2.5.3 Runtime comparison

Finally, we reported runtime for modelling CCLE data with the two baseline methods and the five integration tools. Early concatenation was not included because it does not involve modelling. The computation was completed on a CPU of Intel® Core™ i9-9880H @ 2.30GH and with default settings.

We set five sample sizes (20%/40%/60%/80%/100% of the data) and compared the time consumed to run each of the methods (**Figure 5C**). PCA took less than one second even for the full dataset. Among the five multi-omics integration tools, DIABLO, and moCluster took the least runtime across all sizes with a linear runtime complexity. Although iClusterBayes also has a linear runtime complexity, it consistently ran slower than other tools for all sizes. MOFA2 has better scalability than iClusterBayes with a logarithmic runtime complexity. For sizes from 20% to 60%, iClusterPlus took less time than iClusterBayes and MOFA2, with a similar runtime compared to moCluster and DIABLO. However, the runtime of iClusterPlus surged

dramatically from 10 minutes to 78 minutes when the number of samples was increased from 60% to 80%. Thus, iClusterPlus indicates a nonlinear growth, and is suboptimal for large datasets.

2.5.4 Recommendation

No single best method across all aspects can be found when analysing the CCLE multi-omics dataset (**Figure 5D**). PCA was robust and scalable among unsupervised methods when compared with MOFA2, iCluster series, and moCluster. In terms of average ranking, moCluster showed satisfactory performance across all three types of comparisons. It was ranked as the fourth for cancer type prediction and the second for drug response prediction with a linear runtime complexity.

2.6 CONCLUSIONS

We comprehensively reviewed current machine learning tools for the integration of multi-omics data across various research tasks in the field of cancer research. Multi-omics data analysis is key to understanding the complex dysregulation that is associated with different cancer phenotypes. Despite the exponential growth of the number of multi-omics experiments and the amount of data available for analysis, limited efforts have been made to develop machine learning tools that automatically integrate these multi-omics datasets. Existing reviews on this topic have predominantly focused on computational approaches, leaving a gap in the literature for a review of multi-omics technologies and data formats. All tools reviewed in this article can only be used via a command-line interface, which is not user friendly for non-computational researchers. Therefore, future tools with a graphical user interface will facilitate wider adoption in cancer research. More importantly, here we only reviewed methods that support custom datasets, and we conducted an independent benchmarking with a range of evaluations, including cancer type prediction, drug

response prediction, and runtime efficiency. We concluded that many multi-omics data integration tools did not show significantly enhanced performance than PCA on a common dataset.

One challenge for multi-omics data integration is to account for the inconsistency between data generated from multiple sites. A meta-analysis across 12 laboratories has shown that consistency for copy number and transcriptomic data is relatively high, whereas methylation and proteomic data only showed moderate to low consistency (Jaiswal et al. 2021). Other attempts at the genomics and proteomics levels have been made to partially mitigate this inconsistency (Collins et al. 2017; Zhong et al. 2018). We expect future machine learning approaches can resolve issues such as batch effects and normalization within the integration analysis.

Certain omics data types are not covered in detail in this review because significant challenges regarding data integration are yet to be addressed. For example, metabolomic data record the levels of small molecules that are involved in cell metabolism, and metabolomics has shown a significant impact on cancer research to date (L.-B. Wang et al. 2021). However, metabolomic data are not stored in a gene-level format, making it difficult for machine learning to integrate metabolomic data with other omics data. Another challenge for most machine learning tools is to incorporate biological knowledge into modelling approaches. Gene regulation is a fundamental biological mechanism that describes a hidden link between layers of multi-omics data, but this association is often inappropriately modelled. Most statistical methods focus on explaining the greatest amount of variation in a dataset by using a small number of surrogate variables. This approach can miss the detail of true biological relationships. We thus hypothesize that dynamical modelling of the regulation between genes and subsequent omics layers might be a promising direction

for the future development of machine learning-based multi-omics data integration tools. Causal models are likely to be applied to modelling gene regulations.

Chapter 3: Pan-cancer proteomic map of 949 human cell lines

Text and figures included in this chapter are adapted from the following publication:

Gonçalves, E.*, Poulos, R. C.*, Cai, Z.* [* joint first authors], Barthorpe, S., Manda, S. S., Lucas, N., ... & Reddel, R. R. (2022). Pan-cancer proteomic map of 949 human cell lines. *Cancer cell*, 40(8), 835-849.

Statement of Contribution

This chapter was co-authored by the PhD Candidate, Dr. Emanuel Gonçalves and Rebecca C Poulos. The PhD candidate completed landscape proteomic analyses, including data quality assurance analysis, unsupervised landscape analysis using dimensionality reduction and differential analysis in the study. The PhD Candidate designed, developed and performed all machine learning analyses. The PhD Candidate was also highly involved in the writing of this chapter and the preparation of figures.

3.1 SUMMARY

The proteome provides unique insights into disease biology beyond the genome and transcriptome. Lack of large proteomic datasets has restricted identification of new cancer biomarkers. Here, proteomes of 949 cancer cell lines across 28 tissue types are analyzed by mass spectrometry. Deploying a workflow to quantify 8,498 proteins, these data capture evidence of cell type and post-transcriptional modifications. Integrating multi-omics, drug response and CRISPR-Cas9 gene essentiality screens with a deep learning-based pipeline reveals thousands of protein biomarkers of cancer vulnerabilities that are not significant at the transcript level. The power of the proteome to predict drug response is very similar to that of the transcriptome. Further, random downsampling to only 1,500 proteins has limited impact on predictive power, consistent with protein networks being highly connected and co-regulated. This pan-cancer proteomic map (ProCan-DepMapSanger) is a comprehensive resource available at https://cellmodelpassports.sanger.ac.uk.

3.2 INTRODUCTION

Precision medicine relies on the identification of specific molecular alterations that can stratify patients and guide the choice of effective therapeutic options. Cancer vulnerabilities, such as synthetic lethalities, can be systematically studied using functional genetic and small molecule screens. To circumvent the limitations of using patient tissue samples for this type of approach, biomarkers of cancer vulnerabilities have been analysed using cancer cell lines, together with deep molecular characterization, functional genetic and pharmacological screens (Iorio et al. 2016; Tsherniak et al. 2017; Ghandi et al. 2019; Behan et al. 2019; Frejno et al. 2020). The direct measurement of proteins provides insights into the dynamic molecular behaviour of cells and can improve our understanding of genotype-to-phenotype

relationships (Y. Liu, Beyer, and Aebersold 2016). Despite the development of precision oncology therapeutics, the complexity of cancer and the inability of genomics to accurately predict the proteome indicate that genomics alone is often insufficient to inform and guide the clinical care of many patients. Measurement of the proteome has the potential to expand our understanding of cancer phenotypes and to improve diagnosis and treatment choices.

Technological and methodological advances have enabled the standardized quantification of thousands of proteins across dozens to hundreds of cell lines (Gholami et al. 2013; R. T. Lawrence et al. 2015; Coscia et al. 2016; Roumeliotis et al. 2017; Nusinow et al. 2020) and the profiling of clinical samples derived from minute tissue biopsies (B. Zhang et al. 2014; Edwards et al. 2015; Pozniak et al. 2016; H. Zhang et al. 2016; Mertins et al. 2016; Frejno et al. 2017; Vasaikar et al. 2019; Clark et al. 2019; Tully 2020). Using a data-independent acquisition (DIA)-mass spectrometry (MS) approach (Gillet et al. 2012; Guo et al. 2015; Ludwig et al. 2018; Lucas et al. 2019), together with a sample processing workflow with novel data processing methods, it is now possible for proteomes to be acquired reproducibly at scale (Tully et al. 2019; Poulos et al. 2020). The generation and distribution of large-scale proteomic datasets have the potential to drive new computational approaches, including deep learning-based algorithms, to investigate the impact of molecular changes on cancer vulnerabilities. This will enable proteomics to contribute important clinical advances for cancer therapeutic applications.

Cell lines have been invaluable models for our understanding of cellular processes and the molecular drivers of carcinogenesis (Garnett et al. 2012; Iorio et al. 2016; Barretina et al. 2012; Tsherniak et al. 2017; Meyers et al. 2017; Behan et al. 2019; Ghandi et al. 2019; Picco et al. 2019), and for identifying cancer cell

vulnerabilities to both genetic (McDonald et al. 2017; Tsherniak et al. 2017; Hart et al. 2015; Meyers et al. 2017; Behan et al. 2019) and pharmacological (Garnett et al. 2012; Iorio et al. 2016; Corsello et al. 2020; Barretina et al. 2012; Seashore-Ludlow et al. 2015; Rees et al. 2016) perturbations. However, proteomic quantifications for cancer cell lines are either limited in the range of cancer types or number of samples analysed, or are largely unavailable (Nusinow et al. 2020; Tsherniak et al. 2017; Ghandi et al. 2019) . Consequently, little is known about the contribution of the proteome to cancer vulnerabilities or how the cancer proteome is regulated in diverse tissues and genetic contexts.

This study reports a pan-cancer cell line proteomic map quantifying 8,498 proteins across 949 cell lines. The generation and analysis of this rich resource involved the development of a workflow with rapid sample processing and minimal complexity, followed by the application of a deep neural network-based computational pipeline to uncover cancer targets. Integration of our proteomic data (referred to as the ProCan-DepMapSanger dataset) with existing molecular and phenotypic datasets from the Cancer Dependency Map (Boehm et al. 2021), showed that protein networks are more strongly co-regulated than are transcriptomics and functional genomics. Our approach identified biomarkers of well-established cancer vulnerabilities and, more important, highlighted those that cannot be identified with genomics or transcriptomics alone. The proteome measured in our study had an equivalent performance to the total measured transcriptome in predicting cancer phenotypes. Furthermore, random subsets of 1,500 proteins downsampled from the complete dataset achieved 88% of the power to predict drug responses. These results have broad implications for the design of future studies, ranging from basic research to clinical applications.

3.3 RESULTS

3.3.1 A resource of 949 cancer cell line proteomes

To construct a pan-cancer proteomic map, proteomes of 949 human cancer cell lines from 28 tissues and more than 40 genetically and histologically diverse cancer types were quantified (**Figures 1A** and **S1A**, **Table S1**). The proteome for each cell line was acquired by DIA-MS from six replicates using a workflow that enables high throughput and minimal instrument downtime (see STAR Methods, **Figure S1B**). The resulting dataset was derived from 6,864 DIA-MS runs acquired over 10,000 MS h (**Table S1**), including peptide preparations derived from the human embryonic kidney cell line HEK293T that were used throughout all data acquisition periods and instruments for quality control. These data, together with the spectral library, were deposited in the Proteomics Identification Database (Perez-Riverol et al. 2019) with dataset identifier PXD030304. Raw DIA-MS data were processed with DIA-NN (Demichev et al. 2020), using retention time-dependent normalization and with a spectral library generated by DIA-NN. For full details of data processing steps and parameters, see STAR Methods and **Table S1**. MaxLFQ (Cox et al. 2014) was then used to quantify a total of 8,498 proteins (**Table S2**, **Figure S1C**), with a median of 5,237 proteins (min-max range: 2,523–6,251) quantified per cell line (**Table S1**, **Figure 1A**). The ProCan-DepMapSanger dataset significantly expands the existing molecular characterizations of this broad range of cancer cell line models (**Figure 1B**). Pharmacological screens of anti-cancer drugs tested against this cell line panel were also expanded in this study, increasing the number of unique drugs tested by 48% over our prior work (n = 625 drugs and investigational compounds; **Figure 1C**), with a total of 578,238 half-maximal inhibitory concentrations (IC_{50}) experimentally measured.

High correlations were observed between replicates of each cell line, yielding a sample-wise median Pearson's correlation coefficient (Pearson's r) of 0.92 (**Figures 1D** and **S1A**). Correlations between unmatched samples from the same instrument or batch were similar to random (median Pearson's r = 0.75, **Figure 1D**). Although integration of outputs from different proteomics platforms is acknowledged to be an unsolved challenge, we have compared our data with previously published proteomic datasets comprising smaller subsets of the same cell lines (R. T. Lawrence et al. 2015; Roumeliotis et al. 2017; Guo et al. 2019; Nusinow et al. 2020; Frejno et al. 2020) and have shown comparable levels of correlation among all datasets (**Figure S1D**). Nonlinear dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018) showed no evidence of instrument or batch effects (**Figure S1E**). Proteins that were detected had higher mean RNA expression across the cell line panel than proteins that were not detected in this study ($p < 0.0001$ by the Mann-Whitney U test) (**Figure S1F**), suggesting some bias toward abundant proteins. Overall, this study generated a high-quality and biologically reproducible pan-cancer proteome map of human cancer cell lines.

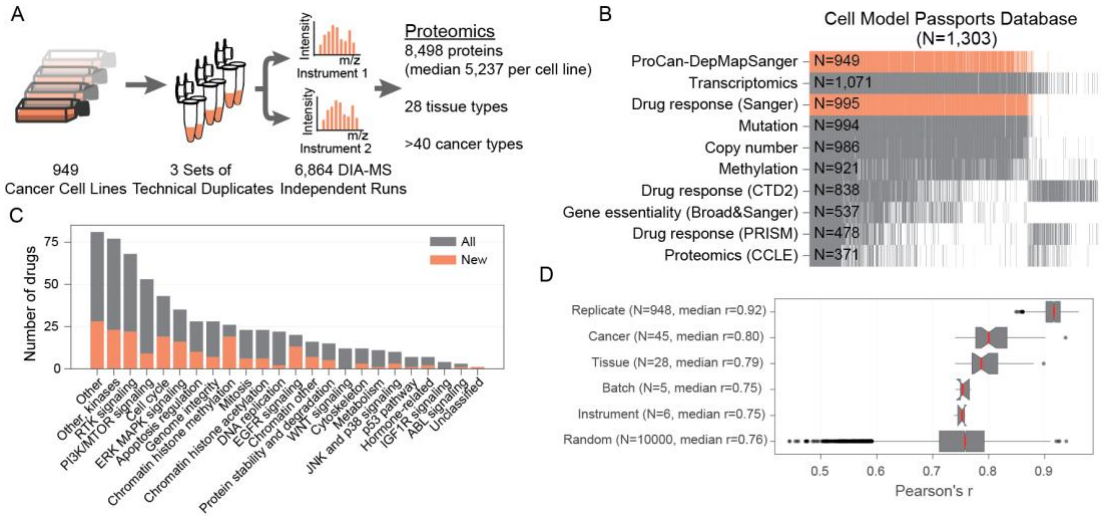


Figure 1. A pan-cancer proteomic map of 949 human cancer cell lines. A, Methodology overview for pan-cancer characterization of 949 human cell lines using a DIA-MS workflow. **B,** Proteomic measurements were integrated with independent molecular and phenotypic datasets spanning 1,303

cancer cell lines as part of the Cell Model Passports Database. Data include proteomics (ProCan-DepMapSanger) presented here, transcriptomics, drug response (Sanger), mutation, copy number, methylation, drug response (CTD2), CRISPR-Cas9 gene essentiality (Broad&Sanger), drug response (PRISM), and proteomics (CCLE). Each grey slice denotes a unique cell line, and the total number of cell lines per dataset is indicated. The proteomic data (ProCan-DepMapSanger) generated in this study are shown in orange, as well as the expanded drug response (Sanger) dataset. **C,** Number of drugs included in the drug response (Sanger) screen, with the orange bar highlighting the additional number of unique drugs presented in this study compared to previous studies. Drugs are grouped by the pathway of their canonical targets. **D,** Pearson's correlations of the proteomes for each set of six technical replicates, as well as each cancer type, tissue type, batch and instrument. Random indicates the correlation between random unmatched sets of replicates. Median Pearson's *r* for each group is reported. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. See also ***Figure S1***, ***Table S1*** and ***Table S2***.

3.3.2 Proteomic profiles reveal cell type of origin

Next, we defined a stringent set of protein quantifications that were supported by measuring more than one peptide (n = 6,692 human proteins) (**Table S2**). Visualization of these protein intensities via UMAP showed groupings by cell type of origin, such as distinct clusters of hematopoietic and lymphoid cells and skin cells (**Figure 2A**). Hematopoietic and lymphoid cells appeared to exhibit further subgroups, and we found that this cell type could be segregated into different cell lineages (**Figure 2B**). This high-level dimensionality reduction suggested a profile of protein expression that relates to cell type of origin. To investigate this further, a set of 279 proteins that are enriched in certain cell types was selected (**Table S3**). These cell type-enriched proteins were defined as any protein quantified in 50% or more of cell lines from no more than two tissue types and 35% or less of cell lines from all remaining tissues, considering only tissues represented by at least 10 cell lines (**Figure 2C**). Cell lines from hematopoietic and lymphoid, peripheral nervous system, and skin cell types showed the greatest numbers of these proteins (**Figure 2D**). Proteins encoded by genes representing gene ontology terms for lymphocyte activation, neuron projection, and pigmentation were identified in each of these cell types, respectively. Further, the cell type-enriched proteins had a higher correlation between the transcriptome and proteome than did other proteins, suggesting that these represent cell type-specific

processes that are more highly conserved between transcription and translation (**Figure 2E**). Overall, this analysis demonstrates a general alignment of the proteomic data with cell lineage, revealing patterns of protein expression that are consistent with some cancer cell types of origin.

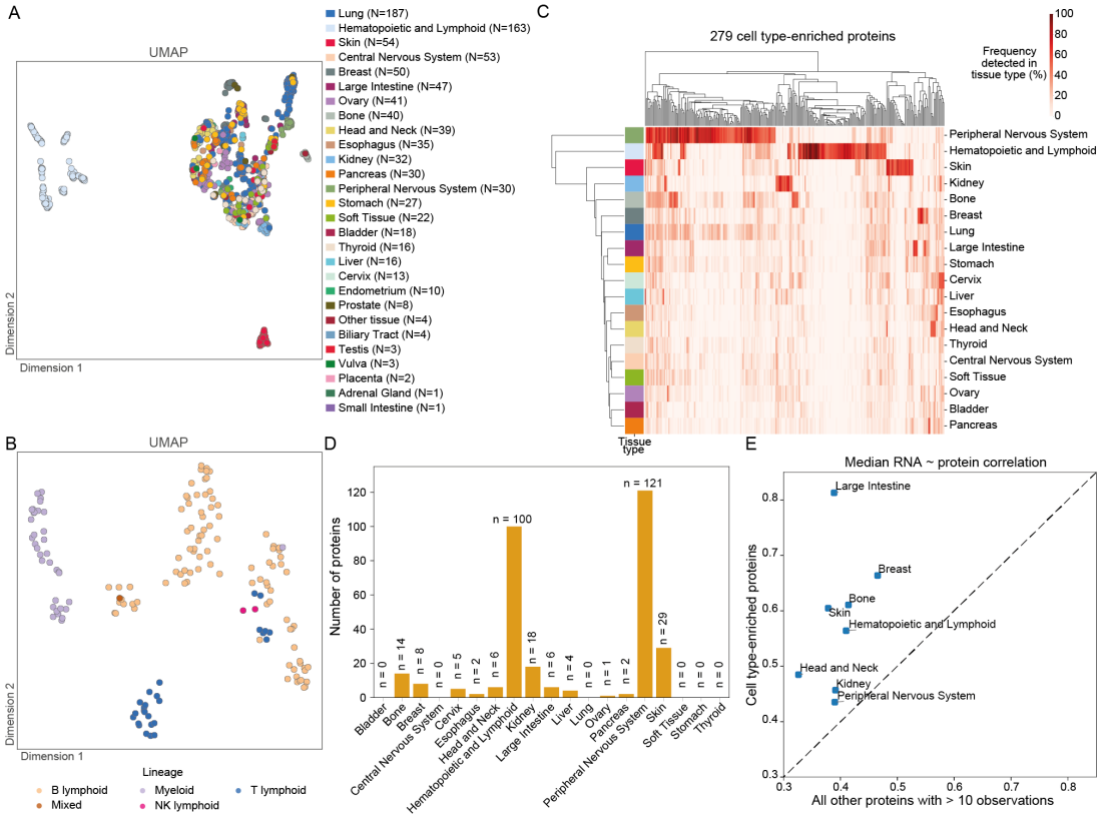


Figure 2. Distinct proteomic profiles according to cell type. A, Proteomic data dimensionality reduction by UMAP, with cell lines colored by tissue. **B,** UMAP of hematopoietic and lymphoid cell lines colored by cell lineage. **C,** Heatmap of the frequency of cell type-enriched proteins observed within each tissue. Tissues and proteins are clustered on the vertical and horizontal axes, respectively. **D,** Number of cell type-enriched proteins identified in each tissue type represented by more than 10 cell lines. **E,** Median RNA-protein correlation of cell type-enriched proteins against all other proteins with more than 10 observations in that tissue type. Only tissues with at least 5 cell type-specific proteins are shown. See also ***Table S3***.

3.3.3 Post-transcriptional regulation in diverse cancer cell types

We next sought to identify the key drivers of the distinct protein expression patterns observed across the cell line panel and to investigate how these integrate with other molecular and phenotypic measurements. Multi-omics factor analysis (MOFA) (Argelaguet et al. 2018, 2020) was used to integrate the proteomic measurements with a range of molecular (promoter methylation, gene expression and protein abundance)

and phenotypic (drug response) datasets available for most cancer cell lines (**Table S4** and **Figure 3A**). MOFA uses a Bayesian group factor analysis framework to enable unsupervised integration of datasets to infer a set of factors (latent variables) that account for biological and technical variability in the data (Argelaguet et al. 2018, 2020). Epithelial-to-mesenchymal transition (EMT) canonical markers, vimentin and E-cadherin, and EMT gene set enrichment analysis enrichment scores (A. Subramanian et al. 2005; Liberzon et al. 2011) were found to be associated with the first two factors (F1 and F2) corresponding with large portions of the variability across all datasets (**Figure 3A**). Technical aspects like cell line media and growth conditions explained little or no variability, while cell size and growth rate were moderately related with some factors (**Figure S2A**). Cancer cell lines from the same tissue of origin showed gradients of EMT markers (**Figure 3B**), which may indicate that the cell lines were established from cancers derived from different epithelial or mesenchymal lineages, or that the cancers had undergone EMT. EMT markers are known to be associated with different stages of cancer progression including initiation, metastasis, and the development of therapy resistance (Brabletz et al. 2018). We observed that some factors capture tissue-specific processes, with their loadings being enriched toward the cell type-enriched proteins defined earlier (**Figure 3A**). For example, a MOFA analysis highlighted an association between factor 12 and skin-derived cell lines (**Figure S2B**), which in this study are primarily from melanomas. Factor 12 also related to phenotypic measurements that are typical of melanomas, correlating with CRISPR-Cas9 gene essentiality scores for BRAF (**Figure 3C**) and with its inhibitor, dabrafenib (**Figure 3D**), both of which are strongly associated with cell lines harbouring BRAF mutations that are very common in cutaneous melanomas. While EMT markers were largely concordant between transcriptomics and proteomics,

more broadly, a modest and variable association was observed between protein and transcript measurements (median protein-wise Pearson's r = 0.42, **Figure S2C**). This is consistent with the expectation that proteomic data capture variability explained by post-transcriptional regulation and the proteostasis network. Focusing on the impact of genomic alterations on proteomic profiles, gene copy number was more weakly correlated with protein levels than with gene expression, indicating attenuation of copy number effects between the transcriptome and the proteome (**Figures 3E** and **Table S4**). This was particularly evident among subunits of protein complexes such as those involved in ribosomes (**Figure S2D**), which can co-regulate their abundance post-transcriptionally to maintain complex stability and stoichiometry (Gonçalves et al. 2017; Roumeliotis et al. 2017; Ryan et al. 2017; Sousa et al. 2019). Proteins involved in protein synthesis and degradation had some of the strongest post-transcriptional regulation, with several proteasome and ribosome subunits showing strong attenuation (**Figure S2D**). Together, this reflects an active proteostasis network, revealed primarily via direct measurement of the proteome (Gumeni et al. 2017).

Using somatic mutation data to stratify the full set of protein quantifications according to mutation status (**Table S4**) revealed 478 proteins with significant differential protein intensities between cell lines that were wild type versus those that had protein-coding mutations (false discovery rate [FDR]-adjusted p value < 0.05) (**Figure 3F**). When mutations were present in the gene encoding a given protein, the majority of proteins had decreased abundance (n = 354 proteins). In contrast, mutations in some genes, such as TP53, were associated with significantly higher protein intensities than wild-type cell lines (TP53: FDR-adjusted p value < 0.0001). This is consistent with the known increase in stability of many mutant P53 proteins, which

results from a decreased rate of proteasome-mediated degradation (Vijayakumaran et al. 2015).

These results indicate that, while variability in protein expression is associated with other molecular and phenotypic layers, there is only partial correlation between transcript and protein abundance, consistent with the effects of post-transcriptional regulation. Thus, the ProCan-DepMapSanger dataset captures additional protein-specific information that can augment our understanding of the impact of genomic alterations affecting, among others, well-established cancer genes.

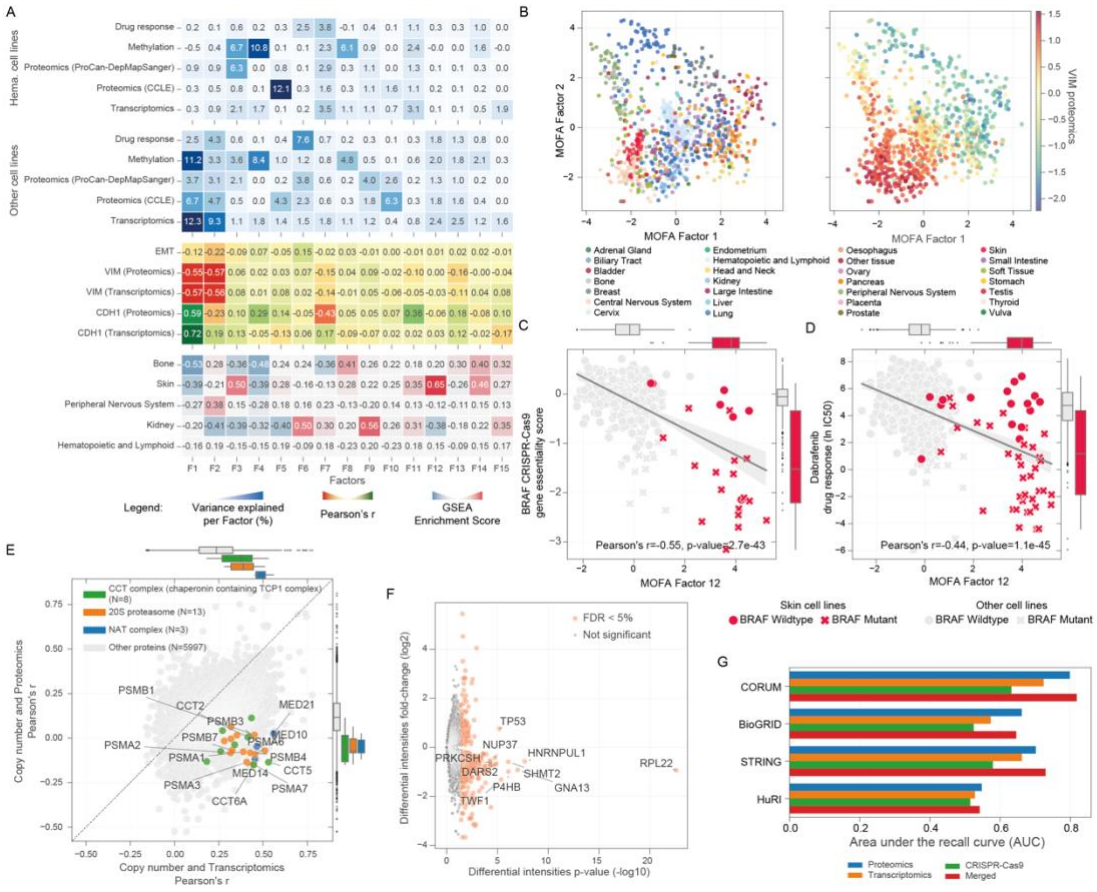


Figure 3. Post-transcriptional regulatory mechanisms of cancer cell lines. A, Identification of shared variability (factors) from MOFA across multiple molecular and phenotypic cancer cell line datasets. Hematopoietic and lymphoid cells are grouped and trained separately from the other cell lines. The upper two heatmaps (blue) report the portion of variance explained by each factor (columns) in each dataset. The central (yellow) heatmap reports Pearson's *r* between each learned factor and various molecular characteristics of the cancer cell lines. The lower heatmap shows GSEA enrichment scores of each factor to cell type-specific proteins. **B,** Separation of cancer cell lines by MOFA Factors 1 and 2, colored by tissue of origin (left) and by EMT canonical marker VIM protein intensities (right). **C,** Scatter plot with linear regression between MOFA Factor 12 and BRAF

CRISPR-Cas9 gene essentiality scores. Skin cancer cell lines are highlighted in red, and BRAF mutant cell lines are marked with a cross. **D,** Similar to **C**, but instead the vertical axis indicates the dabrafenib drug response (IC_{50}) measurements. **E,** Pearson's *r* between gene absolute copy number profiles with transcriptomics (horizontal axis) and with protein intensities from the ProCan-DepMapSanger dataset (vertical axis). Representative CORUM protein complexes with the highest differences between the Pearson's *r* are shown, and the top 15 most attenuated proteins from these complexes are labeled. N indicates the number of proteins in each protein complex. Box-and-whisker plots represent the Pearson's *r* distributions of proteins involved in each highlighted gene ontology term compared to all proteins (gray). **F,** Volcano plot showing differential protein intensities between cell lines that are wild-type versus mutant for each protein in the ProCan-DepMapSanger dataset that is mutated in at least 1% of the cohort. The top 10 proteins by p-value are annotated. The horizontal axis shows the $-\log_{10}$ of the empirical Bayes moderated t-test p-value, and proteins with FDR < 5% are colored in red. **G,** Recall of protein-protein interactions (PPIs), i.e. ability to detect known PPIs, from resources CORUM, STRING, BioGRID and HuRI. All possible protein pairwise correlations (Pearson's p-value) were ranked, using proteomics, transcriptomics and CRISPR-Cas9 gene essentiality. The merged score was defined as the product of the p-values of the different correlations. In **C**, **D** and **E**, Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. See also ***Figure S2*** and ***Table S4***.

3.3.4 Co-regulatory protein networks of cancer cells

Having observed post-transcriptional co-regulation of protein complex abundance (**Figure 3E**), we next investigated whether the abundance of co-regulated proteins could be used to predict putative protein-protein interactions (PPIs). We assessed all possible pairwise protein correlations (n = 16,580,952) and, as a comparator, used corresponding gene expression and CRISPR-Cas9 gene essentiality profiles, where available. As expected, paralogs and protein complex subunits had some of the strongest correlations (**Table S4**; absolute Pearson's r > 0.5 and FDR adjusted p value < 0.05). We systematically assessed this enrichment using multiple resources for protein interactions: protein complex interactions from the Comprehensive Resource of Mammalian Protein Complexes (CORUM) (Ruepp et al. 2010); functional interactions from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (Szklarczyk et al. 2017); physical protein interactions from the Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al. 2015); and the Human Protein Interactome (HuRI) (Luck et al. 2020). For all resources, proteomic measurements had greater ability to detect known PPIs (area under the recall curve [AUC] = 0.55–0.80) than

transcriptomics (AUC = 0.53–0.72) and CRISPR-Cas9 gene essentiality (AUC = 0.51–0.63) (**Figure 3G**), indicating that PPIs and co-regulation are best captured by proteomics.

The correlation p value score (Fisher's combined probability test) for BioGRID and HuRI databases was improved slightly by merging different datasets (AUC = 0.54–0.82), suggesting that different types of interactions are captured across multi-omic layers. Proteins with higher numbers of positive protein-protein correlations were more essential for cancer cell survival, as observed in the CRISPR-Cas9 gene essentiality dataset (**Figure S2E**). This is likely linked to their increased transcript and protein expression levels, and the number of pathways in which they are involved. Paralogs were an exception, as they had largely non-essential profiles independent of their gene expression (**Figure S2F**), consistent with their functional redundancy attenuating the impact of loss (Dandage and Landry 2019).

Considering the high correlations that were observed with known interactions, these pairwise protein correlations could be used to identify novel putative PPIs, such as between protein subunits. Consistent with this, we identified 1,182 putative PPIs with a Pearson's r greater than 0.8 that are not reported in any of the protein network resources analysed here. For example, there were strong correlations between protein profiles for EEF2-EIF3I, RPSA-SERBP1 and CCT6A-EEF2 (**Table S4**). These proteins are not reported to interact directly but are also closely related in the high confidence STRING protein interaction network (Szklarczyk et al. 2017). Overall, this analysis highlights the ability for protein measurements to detect known interactions, and suggests the utility of this dataset for predicting co-regulated interactions.

3.3.5 Identifying biomarkers of cancer vulnerabilities

Next, we considered the application of proteomics to identify biomarkers by harnessing drug (Garnett et al. 2012; Iorio et al. 2016; Picco et al. 2019; Gonçalves et al. 2020) and CRISPR-Cas9 gene essentiality (Behan et al. 2019; Meyers et al. 2017; Pacini et al. 2021) screens. The previous Sanger pharmacological screens were expanded to include a total of 578,238 IC_{50} values (**Figure 1B**). These included a total of 625 unique anti-cancer drugs (48% increase in unique drugs over previously published datasets (Iorio et al. 2016; Picco et al. 2019; Gonçalves et al. 2020)) that were screened across 947 of the 949 cancer cell lines, including U.S. Food and Drug Administration-approved drugs, drugs in clinical development, and investigational compounds. To identify protein biomarkers predictive of cancer cell line response to these drugs or CRISPR-Cas9 gene essentialities (n = 17,486), we applied linear regression to test all pairwise associations between proteins, drug sensitivity, and CRISPR-Cas9 gene dependencies, while considering potentially confounding effects, such as cell line growth rate, culture media, and the average replicate correlation (see STAR Methods for more details) (**Figure 4A**, **Table S5**). Among the strongest significant associations (FDR < 5%), we observed that 57 drugs were associated with the protein abundance of their canonical target(s), including negative associations between EGFR protein abundance and its inhibitor gefitinib, and MET protein abundance and its inhibitor drug response (**Figure 4A**). We observed a significant negative association between ERBB2 (also known as HER2) and lapatinib, a tyrosine kinase inhibitor that targets EGFR and HER2 (**Figure 4A**). This association has been observed in proteomic studies using other preclinical models (breast cancer patient-derived xenografts) (Huang et al. 2017) and in human cancers, with lapatinib already an approved drug used in the treatment of HER2-positive breast cancers (Z.-Q. Xu et al. 2017). For another 132 drugs, significant associations were identified with proteins

functionally related with their targets (i.e., one step away in the STRING PPI network). The majority of the significant drug-protein target associations showed a negative effect size (**Figure 4A**), indicating greater sensitivity of a cell line to a drug when the target of the drug is more abundant. Last, the identification of associations with reported gene copy number alterations, such as amplification of MET and ERBB2, are also observed at the protein level (**Figure 4B**). Non-self interactions were also observed, such as PPA1-PPA2 paralog synthetic-lethal interaction, where cell lines with lower PPA2 abundance are more sensitive to PPA1 knockout (**Figure S3A**).

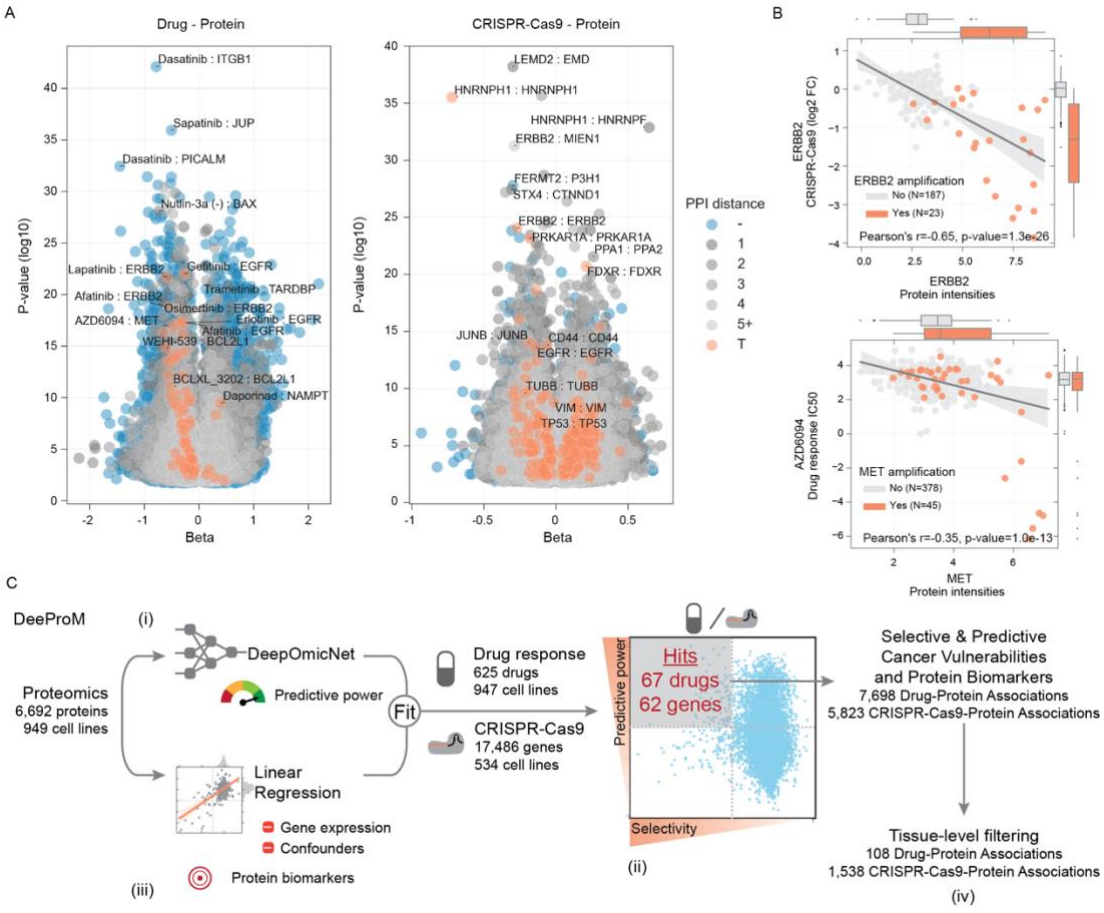


Figure 4. Biomarkers for cancer vulnerabilities. A, Significant linear regression associations (FDR < 5%) between protein measurements and drug responses (left panel) and protein measurements and CRISPR-Cas9 gene essentiality scores (right panel). Each association is represented using the linear regression effect size (beta) and its statistical significance (log-ratio test), and colored according to the distance between the target of the drug or CRISPR-Cas9 and the associated protein in a protein-protein interaction network assembled from STRING. T denotes the associated protein is either a canonical target of the drug or the CRISPR-Cas9 reagents; numbers represent the minimal number of interactions separating the drug or CRISPR-Cas9 targets to the associated proteins; and the symbol '-' denotes associations for which no path was found. Representative examples are labeled. **B,**

Representative top-ranked CRISPR-Cas9-protein and drug-protein associations. Upper panel shows ERBB2 protein intensities associated with CRISPR-Cas9 gene essentiality, where cell lines with ERBB2 amplifications are highlighted in orange. Lower panel shows the association between AZD6094 MET inhibitor and MET protein intensities, where MET amplified cell lines are highlighted in orange. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. **C,** Overview of the DeeProM workflow: (i) deep learning models of DeepOmicNet were trained to predict drug responses and CRISPR-Cas9 gene essentialities, prioritizing those that are best predicted by proteomic profiles; and (ii) Fisher-Pearson coefficient of skewness was calculated to identify drug responses and CRISPR-Cas9 gene essentialities that selectively occur in subsets of cancer cell lines. The selected candidates from (i) and (ii) are illustrated by the gray box. (iii) Linear regression models were fitted to identify significant associations between protein biomarkers, drug responses and CRISPR-Cas9 gene essentialities. (iv) Filtering algorithms were applied to further identify tissue-specific cancer vulnerabilities. See also ***Figure S3*** and ***Table S5***.

To identify biomarker associations that are unique to the proteome and could not be predicted by gene expression measurements alone, we developed a deep learning-based computational pipeline called Deep Proteomic Marker (DeeProM) (**Figure 4C**). DeeProM is powered by DeepOmicNet (see **Figure S3B** and STAR Methods for more details), a deep neural network architecture designed to prioritize drug responses and CRISPR-Cas9 gene essentialities that are highly predictive and specific to subsets of cancer cell lines. In addition, DeeProM incorporates results from linear models described above to highlight biomarkers that are only evident at the proteomic level. As a benchmark, we found that the accuracy of DeepOmicNet, evaluated using Pearson's r between the observed and predicted IC_{50} values, consistently outperformed other machine learning approaches such as elastic net and Random Forest, across a range of multi-omic datasets used in previous studies (Iorio et al. 2016) (**Figure S3C**). DeeProM assessed all possible drug-protein (n = 4,218,788) and CRISPR-protein (n = 86,584,537) associations to identify cancer vulnerabilities that are simultaneously well predicted and selective in subsets of cell lines (**Figure 5A**). These two selection criteria yielded 67 drug responses and 62 gene essentialities, with a total of 7,698 drug-protein and 5,823 CRISPR-Cas9-protein associations, that had significantly improved predictions when compared to models that considered gene expression measurements alone (**Figure 4C** and **Table S5**).

Promising targeted therapeutics are often developed for specific cancer types and can display tissue-specific responses. For this reason, DeeProM was used to interrogate associations at the tissue type level by applying a filtering strategy (**Figure 4C** and STAR Methods). Using the DeeProM workflow, we identified 1,538 tissue type-level CRISPR-Cas9-protein associations (**Table S5**). Among the strongest was the dependency on FOXA1 transcription factor knockout and protein levels of basigin (BSG; also known as CD147), a plasma membrane protein expressed in breast cancer cells (**Figures 5B** and **5C**). This association was not observed at the gene expression level (**Figure S3D**). FOXA1-BSG association occurred in luminal (luminal A and B) and HER2-positive (non-basal) breast cancer cell lines, in which BSG protein abundance is low relative to basal cell lines (**Figure 5C**). BSG has been implicated in breast cancer progression (Landras et al. 2019), and is a marker of the aggressive basal-like and triple-negative subtypes, as well as being associated with poor overall survival within these patients (M. Liu et al. 2018). These data support a model where BSG protein expression is associated with basal-like breast cancer cells, whereas luminal and HER2-positive breast cancer cells with low BSG expression have an increased dependency on estrogen receptor-driven FOXA1 transcriptional activity. Further work that expands the number of samples, and confirmatory studies, would be required to validate this observation.

DeeProM also identified 108 tissue type level drug-protein associations (**Table S5**). Filtering by the effect size, the strongest association identified was between sensitivity to Aurora kinase B/C selective inhibitor GSK1070916 and the protein abundance of peptidyl-prolyl cis-trans isomerase H (PPIH) in cell lines derived from bone (**Figure 5D**). This association was significant at the protein level, but was not significant in the transcriptome (**Figure S4A**). The association was further supported

by examination of the Cancer Cell Line Encyclopedia (CCLE) proteomic dataset (**Figures S4B** and **S4C**) (Nusinow et al. 2020) , the PRISM drug response dataset (**Figure S4D**) (Corsello et al. 2020), and using an independent screening of GSK1070916 in the Sanger drug sensitivity dataset (**Figures S4E** and **S4F**), in which there is a suggestive association that does not reach statistical significance due to the smaller sample size. Furthermore, there was a strong association between PPIH protein levels and Alisertib, a second Aurora kinase inhibitor (**Figures 5E**, **S4G** and **S4H**). PPIH and Aurora kinase A are both regulated by the p53-p21-DREAM-CDE/CHR signalling pathway (Fischer et al. 2016), supporting the identified link between Aurora kinase inhibitor sensitivity and PPIH protein levels. Elucidating the precise mechanism underlying this association will require further research.

Taken together, these results demonstrate the added value of proteomic measurements for the discovery of cancer biomarkers. We identified both established and potential cancer related biomarkers, including protein biomarkers for selective cancer vulnerabilities that were not found using gene expression measurements alone.

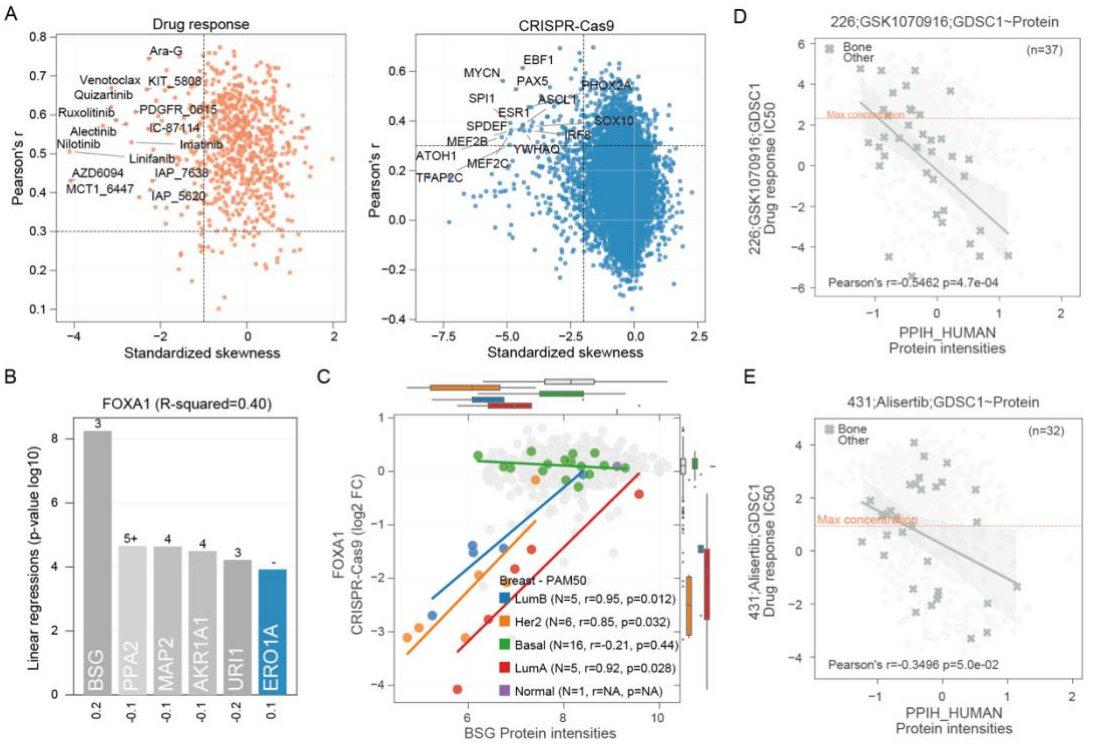


Figure 5. Protein biomarkers identified by DeeProM. A, Predictive performance and selectivity of all drug responses (left) and CRISPR-Cas9 gene essentialities (right) across 947 and 534 cancer cell lines, respectively. Data points toward the top left corner of each plot indicate drug responses or gene essentialities that are both selective and well predicted. Top selective drugs and CRISPR-Cas9 gene essentialities are labeled. **B,** Top significant protein associations with FOXA1 CRISPR-Cas9 gene essentiality scores, each bar representing the statistical significance of the linear regression, and below the effect size (beta). The minimal distance of PPIs in the STRING network between FOXA1 and each protein is annotated in each respective bar and color coded according to the description in **Figure 4A**. **C**, Association between FOXA1 CRISPR-Cas9 gene essentiality scores and BSG protein intensities. Breast cancer cell lines are highlighted and sub-classified using the PAM50 gene expression signature (Parker et al. 2009). Box-and-whisker plots indicate the PAM50 subtypes of breast cancers. Pearson's *r* (r), p-value (p), and number of observations/cell lines (N) within each PAM50 type is provided; for "Normal" subtype no correlation was performed considering N is 1. These plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. **D-E**, Representative examples of tissue-specific associations between drug responses and protein markers for cell lines derived from bone (green; all other cell lines are shown in gray). The number of cell lines and Pearson's *r* from the highlighted tissue type are annotated at the top right and bottom left corners, respectively. The dashed line represents the maximum concentration used in the drug response screens. **D,** The GSK1070916-PPIH association in bone supported by the ProCan-DepMapSanger proteomic dataset. **E,** Similar to **D**, instead showing data for the drug Alisertib. See also ***Figure S3, Figure S4*** and ***Table S5***.

3.3.6 Predictive power of protein sub-networks on cancer cell phenotypes

We have established the utility of proteomics to identify specific biomarkers for cancer vulnerabilities. Using an independent cell line proteomic dataset from the CCLE (Nusinow et al. 2020), we observed comparable performance to the ProCan-DepMapSanger dataset when predicting three independent drug response datasets

(**Figures 6A** and **S5A**) and CRISPR-Cas9 gene essentiality profiles (**Figures 6B** and S5B). We next compared the predictive power of proteomic and transcriptomic data for modelling drug responses and CRISPR-Cas9 gene essentialities. The predictive power of our models was highly similar when trained using either the ProCan-DepMapSanger or the transcriptomics dataset (**Figure 6C**). This was recapitulated by machine learning methods such as elastic net and Random Forest (**Figure 6D**). Notably, the predictive performance of protein measurements and transcript measurements were highly similar, and protein measurements alone outperformed the corresponding overlapping subset of the transcriptome (p value < 0.0001, two-tailed paired Student t test) (**Figure S5C**). Proteomic measurements further showed overall stronger protein pairwise correlations than transcriptomics or CRISPR-Cas9 gene essentialities (**Figure 6E**). Taking these observations together, this demonstrates that proteomics and transcriptomics share comparable predictive power and suggests that proteomics may provide additional relevant information that is not captured by transcriptomics.

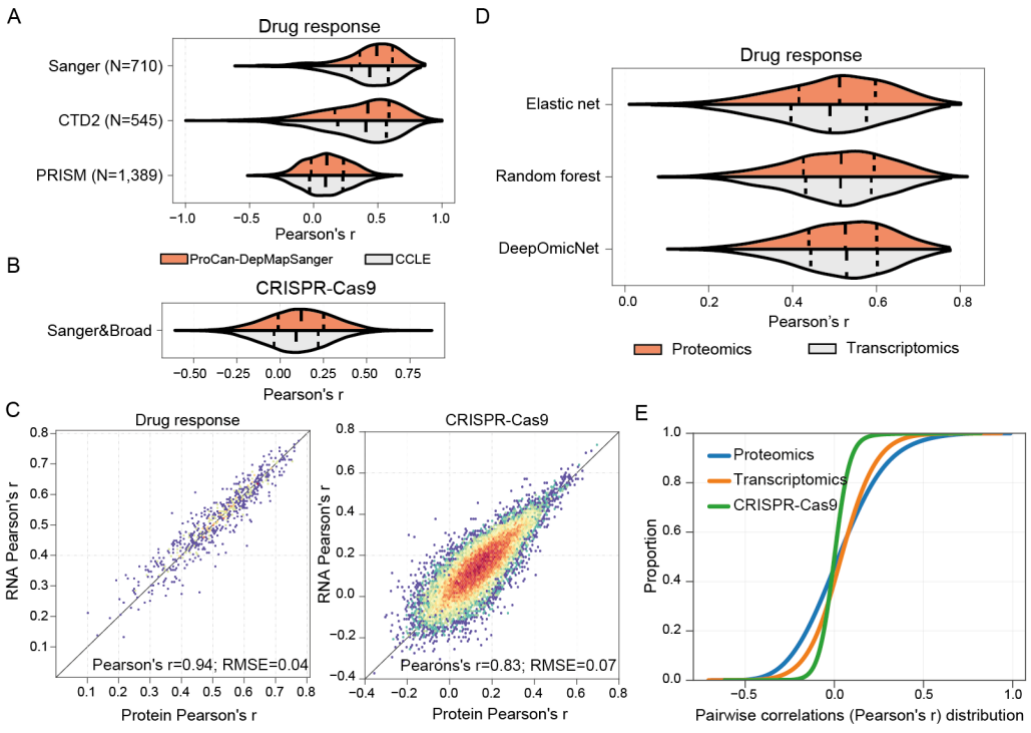


Figure 6. Evaluation of predictive power of DeepOmicNet for multi-omic datasets. A-B, Distribution of the predictive power (mean Pearson's *r* between predicted and observed IC_{50} values) of DeepOmicNet, comparing ProCan-DepMapSanger to an independent proteomic dataset (CCLE), using cell lines in common between the two datasets. Plots show prediction of **A,** drug responses (N represents the total number of drugs tested; n = 290 cell lines) and **B,** CRISPR-Cas9 gene essentialities (n = 234 cell lines). **C**, Two-dimensional density plots showing the predictive power of DeepOmicNet in predicting drug responses (left) and CRISPR-Cas9 gene essentiality profiles (right) using protein (horizontal axis) and transcript (vertical axis) measurements. Each data point denotes the Pearson's *r* between predicted and observed measurements for each drug or CRISPR-Cas9 gene essentialities. **D**, Similar to **A**, distribution of the predictive power of three machine learning models using either proteomic or transcriptomic measurements to train and predict drug responses (Sanger dataset). **E,** Cumulative distribution function of the Pearson's *r* of all pairwise protein-protein correlations compared with transcriptomics and CRISPR-Cas9 gene essentiality measurements. See also ***Figure S5***.

To determine how predictive power is influenced by the number of proteins used in the modelling, a random downsampling analysis was performed to predict drug responses, with a decrease of 500 proteins in each step (see STAR Methods). This showed that a randomly selected subset of 1,500 proteins was able to provide 88% of the predictive power of the full dataset (mean Pearson's r = 0.43 at n = 1,500 proteins versus mean Pearson's r = 0.49 at n = 8,498 proteins) (**Figure 7A**). This implies that a random fraction of quantifiable proteins is sufficient to represent fundamental elements of the proteome involved in mediating key cellular phenotypes. We propose this is in part because proteins are organized into complexes and pathways with connected and co-regulated subunits (**Figure 7B**).

To investigate protein networks, DeeProM analyses of drug-protein and CRISPR-Cas9-protein associations were examined in the context of protein networks (**Figures 7C** and **7D**). The strongest overall associations were observed between the drug and CRISPR-Cas9 targets and their protein intensities (**Figures 7C** and **7D**). Additionally, CRISPR-Cas9-protein associations showed that proteins closer in the PPI network to the targeted proteins had stronger associations than those further apart (**Figure 7D**). This enrichment for targets and functionally closer proteins remains even when the contribution of transcriptomic measurements is removed for the drug-protein and CRISPR-Cas9-protein associations (**Figures S5D** and **S5E**). However, many

seemingly functionally distant proteins (more than two steps away from the perturbation target in PPIs) also exhibit significant drug-protein and CRISPR-Cas9-protein associations (**Figures 7C** and **7D**).

We next explored the relationship between predictive performance and sub-networks comprising proteins of differing frequencies in the full dataset, defining categories of proteins as those found in 90% or more (category A; n = 2,944 proteins), 20%–90% (category B; n = 3,939) or less than 20% (category C; n = 1,615) of the cell lines, respectively (**Figure S5F**). Downsampling these protein sets at random, with a decrease of 250 proteins in each step (see STAR Methods), showed that proteins that are frequently observed in the dataset (category A), provided the highest predictive performance (e.g., mean Pearson's r = 0.463 for 1,500 proteins, or r = 0.435 for 250 proteins) when compared against less frequently observed proteins (category B, mean r = 0.441; and category C, mean r = 0.432, for 1,500 proteins) (**Figure 7E**). Category A proteins had a significantly higher degree of connectivity (mean of 8 degrees) than category B and C proteins (mean of 2 degrees and 1 degree, respectively) in the STRING protein interaction network (Szklarczyk et al. 2017) (**Figure 7F**). It is possible that category A proteins, as a consequence of being more frequently observed, are better studied and therefore have greater annotation in the STRING database. However, together these results suggest that the quantification of small subsets of commonly expressed proteins within highly interconnected networks can be used for predictive modelling of cellular phenotypes.

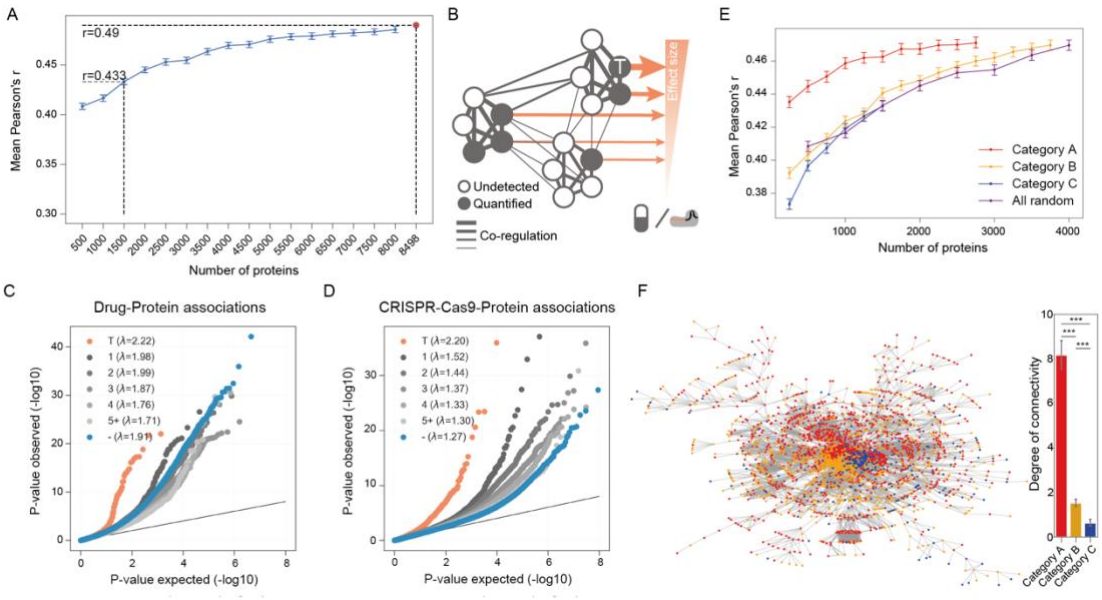


Figure 7. Proteomic support for a network pleiotropy model. A, Comparison of the predictive power of DeepOmicNet trained with randomly downsampled sets of proteins. The dots indicate the means and vertical lines represent 95% confidence intervals derived from 10 iterations of random downsampling. The red point represents the full predictive power using all of 8,498 quantified proteins. **B,** Schematic diagram depicting protein network pleiotropy with widespread protein associations with responses to either drugs or CRISPR-Cas9, and demonstrating the strongly co-regulated nature of protein networks. Nodes represent proteins that could be either quantified or are undetected, where 'T' represents a protein target of a drug or CRISPR-Cas9 gene essentialities. Edges showcase putative interactions, with high correlation coefficients between proteins depicted by thicker edges. Orange arrows represent the variability explained by that protein for the cancer cell line's response to a drug or CRISPR-Cas9 gene perturbation. The size of the arrow is proportional to the variance explained. **C-D,** Quantile-quantile plots of protein associations with **C,** drug responses and **D,** CRISPR-Cas9 gene essentiality profiles. Protein associations are grouped and colored by their distances from the drugs or CRISPR-Cas9 targets using the STRING protein interaction network, where '-' and the blue circles denote associations for which no link in the protein network could be found between the protein and the drug or CRISPR target. P-values were calculated in likelihood ratio tests on all parameters of the linear regression models. Annotation is as described in **Figure 4A**. For each group the p-value distribution inflation factor lambda, 'λ', using the median method (Aulchenko et al. 2007). **E,** Comparison of the predictive power of DeepOmicNet models trained with subsets of Category A, B and C proteins (per ***Figure S5F***) comprising randomly downsampled sets of proteins. The dots indicate the means and vertical lines represent 95% confidence intervals derived from 10 iterations of random samplings. **F,** The STRING protein interaction network diagram (left), with proteins colored according to Category. The bar chart (right) shows the network connectivity for these proteins, where degree represents the number of other proteins connected to a given protein according to the STRING PPI network. *** denotes significant at $P < 0.001$ by unpaired t-test. Error bars represent 95% confidence intervals. See also ***Figure S5***.

3.4 DISCUSSION

The ProCan-DepMapSanger data resource is a large pan-cancer proteomic map that provides multiple insights beyond existing molecular datasets. This map quantifies 8,498 proteins across 949 human cancer cell lines, representing 28 tissues and more than 40 histologically diverse cancer types and a wide range of genotypes, significantly

expanding the molecular characterization of cancer models as part of a Cancer Dependency Map (Boehm et al. 2021). All data are publicly available along with other molecular and phenotypic datasets at http://cellmodelpassports.sanger.ac.uk (van der Meer et al. 2019). This proteomic dataset is a high-quality resource for mechanistic investigation of network organization and regulatory principles of the proteome, as well as for translational discoveries.

This study demonstrated protein expression patterns reflecting cell lineage and, potentially, EMT. The data also revealed widespread protein regulatory events, such as post-transcriptional attenuation of gene copy number effects. ProCan-DepMapSanger allowed the comprehensive characterization of protein expression patterns that could not be captured by the transcriptome, exposing the benefits of directly measuring protein abundance. Furthermore, we developed a deep learning-based pipeline DeeProM, with a deep neural network architecture, which consistently outperformed other machine learning approaches. DeeProM enabled the full integration of proteomic data with drug responses and CRISPR-Cas9 gene essentiality screens to build a comprehensive map of protein-specific biomarkers of cancer vulnerabilities that are essential for cancer cell survival and growth. Notably, this demonstrates that proteomic data spanning a broad range of cancer cell types and molecular backgrounds has significant utility for predicting cancer cell vulnerabilities.

The proteomic workflow used in this study was devised to be clinically relevant, so that our methods could be readily applied for use in human cancer tissue samples. To do so, we used shortened preparation times, low peptide loads, and a short liquid chromatography (LC)/MS run time, enabling the analysis of large numbers of very small cancer samples, with high throughput and minimal instrument downtime. This will facilitate future validation of proteomic predictions from cell line data in clinical

sample cohorts for which outcome of treatment is documented and proteomic data are obtained. The CCLE proteomic dataset was generated using higher peptide loads and longer LC/MS run time to measure more proteins (12,755 proteins). Despite the different depths of protein coverage, the CCLE and ProCan-DepMapSanger proteomic datasets had equivalent power for predicting cancer dependencies. Similarly, the ProCan-DepMapSanger proteomic dataset had similar predictive power to cancer cell line transcriptomic data. Taken together, this demonstrates that a high-throughput sample workflow, as used in this study, produces data with power to inform predictions of cancer dependencies and indicates the potential of proteomics for clinical applications using small biopsies of human cancer tissue in diverse molecular contexts. Subsequent application of this proteomics sample workflow, and integration of this cell line dataset with proteomic data from cancer tissue samples, is likely to provide numerous potential clinical applications, such as the proteomic molecular identification and stratification of cancer subtypes.

Measuring even a fraction of the proteome, as small as 1,500 randomly selected proteins, provided power to predict drug responses that were similar to the full proteome that we report. This suggests that random subsets of protein data comprising a relatively small number of proteins would be sufficient to represent many fundamental cellular processes. This is consistent with an omnigenic model (Boyle, Li, and Pritchard 2017), whereby large numbers of genes are related to many different disease traits in an interconnected manner. This is related to the proteostasis network model of sustaining proteome balance via coordinated protein synthesis, folding, conformation and degradation (Gumeni et al. 2017). In the context of the cancer proteome, we propose that pleiotropic networks of highly connected and co-regulated proteins contribute toward establishing cellular phenotypes. This includes a small

number of core protein modules that are proximal to the phenotype and have the strongest effect, and a much larger set of more distal proteins that together explain a significant portion of total variation.

In conclusion, this dataset represents a major resource for the scientific community, for biomarker discovery and for the study of fundamental aspects of protein regulation that are not evident from existing molecular datasets. This will enable the identification of targets (including cell surface proteins) and treatments for validation in cancer tissue cohorts, with applications in precision oncology.

3.5 STAR METHODS

3.5.1 EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

Cell line authentication. The 949 human cancer cell lines used in this study have been obtained from public repositories and private collections and are cultured in DMEM/F12 or RPMI 1640 (**Table S1**). More detailed information about each cell line can be found at the Cell Model Passports portal (https://cellmodelpassports.sanger.ac.uk/) (van der Meer et al. 2019). All cell line stocks were tested for mycoplasma contamination prior to banking using both a polymerase chain reaction (EZ-PCR Mycoplasma Detection Kit, Biological Industries) and a biochemical test (MycoAlert, Lonza). Cultures testing positive using either method were removed from the collection.

To prevent cross-contamination or misidentification, all banked cryovials of cell lines were analysed using a panel of 94 single nucleotide polymorphisms (SNPs) (Garnett et al. 2012) (Fluidigm, 96.96 Dynamic Array IFC). The data obtained were compared against a set of reference SNP profiles that have been authenticated by short tandem repeat (STR) back to a published reference (typically the supplying

repository). Where a published reference STR profile is not available, the reference SNP profile is required to be unique within the collection/dataset. A minimum of 75% of SNPs is required to match the reference profile for a sample to be positively authenticated. Additionally, one of the replicate cell pellets generated from each cell line for this study underwent authentication via STR profiling at CellBank Australia (Westmead, Australia). To do so, STR loci were amplified using the PowerPlex® 16HS System (Promega) and the data were analysed using GeneMapper™ ID software (ThermoFisher). Only cell lines that passed this quality control metric were retained for analysis (n = 949).

Cell culture and harvesting

For each cell line, distinct cell pellets from a single cell culture were produced as technical replicates. Cells were cultured to semi-confluence at 37°C and 5% CO2 in the appropriate medium and then harvested. Suspension cells were centrifuged at 200 g for 5 min at 4°C and the supernatant was removed. The cells were then washed twice by resuspension in ice-cold Dulbecco's phosphate buffered saline containing no calcium or magnesium (DPBS) and centrifugation at 200 g for 5 min at 4°C. For adherent cells, the culture medium was removed before washing with ice-cold DPBS and the cells were removed by mechanical scraping into fresh ice-cold DPBS. The harvested cells were then centrifuged as before, washed twice in ice-cold DPBS, transferred to 1.5 mL centrifuge tubes (Protein LoBind Tubes, Eppendorf) and centrifuged at 600 g for 5 min at 4°C. The DPBS was removed and the tubes containing the cell pellets were snap frozen on dry ice, then stored at −80°C.

3.5.2 Method details

Cell lysis and digestion

Three cell pellets were analysed for each of the 949 cell lines. The cell pellets were processed using Accelerated Barocycler Lysis and Extraction (ABLE) protocol with minor modifications (Lucas et al. 2019). In brief, all cell pellets were centrifuged to remove residual DPBS, then resuspended in a volume of 1% (w/v) sodium deoxycholate (SDC) that was appropriate for the cell count (between 50 and 400 µL). To this, 1 unit of benzonase was added to digest the DNA/RNA in the samples for 5 min at 37°C and mixed with shaking at 1000 rpm. After incubation, a 50 µL aliquot was taken and further processed, with peptide digestion carried out as previously published (Lucas et al. 2019).

Data independent acquisition (DIA)-MS

We used a workflow that enables high throughput and minimal instrument downtime; 2 µg of peptide was loaded for each replicate with 90-min acquisitions. Three technical replicates of peptide preparations were generated. Each replicate was injected on two of six different SCIEX™ 6600 TripleTOF® mass spectrometers coupled to Eksigent nanoLC 425 high-performance liquid chromatography (HPLC) systems, housed in a single laboratory, ProCan in Westmead, Australia (**Figure S1B**). In each case, an Eksigent nanoLC 425 HPLC system (Sciex) operating in microflow mode was coupled online to a 6600 TripleTOF® system (Sciex) run in sequential windowed acquisition of all theoretical fragment ion spectra (SWATH™) mode using 100 variable isolation windows (**Table S1**). The parameters were set as follows: lower m/z limit 350; upper m/z limit 1250; window overlap (Da) 1.0; collision energy spread was set at 5 for the smaller windows, then 8 for larger windows; and 10 for the largest windows. MS/MS spectra were collected in the range of m/z 100 to 2000 for 30 ms in high resolution mode and the resulting total cycle time was 3.2 s.

The peptide digests (2 µg) were spiked with retention time standards and injected onto a C18 trap column (SGE TRAPCOL C18 G203 300 µm × 100 mm) and desalted for 5 min at 10 µL/min with solvent A (0.1% [v/v] formic acid). The trap column was switched in-line with a reversed-phase capillary column (SGE C18 G203 250 mm × 300 µm ID 3 µm 200 Å), maintained at a temperature of 40°C. The flow rate was 5 µL/min. The gradient started at 2% solvent B (99.9% [v/v] acetonitrile, 0.1% [v/v] formic acid) and increased to 10% over 5 min. This was followed by an increase of solvent B to 25% over 60 min, then a further increase to 40% for 5 min. The column was washed with a 4 min linear gradient to 95% solvent B held for 5 min, followed by a 9 min column equilibration step with 98% solvent A. The TripleTOF® 6600 system was equipped with a DuoSpray source and 50 µm internal diameter electrode and controlled by Analyst 1.7.1 software. The following parameters were used: 5500 V ion spray voltage; 25 nitrogen curtain gas; 100°C TEM, 20 source gas 1, 20 source gas 2.

Spectral library and DIA-MS data processing

An in silico spectral library was created using DIA-NN (version 1.8) (Demichev et al. 2020) for the canonical human proteome (Uniprot Release, 2021_03; 20,612 sequences), along with retention time peptides and commonly occurring microbial and viral sequences. DIA-MS data in wiff file format were collected for 6,981 MS runs (**Table S1**), and all of these MS runs were used to create a spectral library in DIA-NN (Demichev et al. 2020). To reduce the search space, the library was confined to precursors identified in the in silico library only. The following settings were used for library generation. Precursor mass ranges were set between 400 and 1250 m/z and fragment mass ranges were set between 100 and 2000 m/z. Mass accuracies of 40 ppm were set for both MS1 and MS2, with the scan window set to 9. Precursors of charges 2-4 and of length between 7 and 30 were retained. Only Carbamidomethylation at

Cysteine residues was allowed as a fixed modification. Interfering precursor peaks were removed, the robust LC (high accuracy) quantification strategy was used, and precursors were filtered at a q-value of 0.01. Protein grouping was done at the canonical protein sequence level rather than gene level. The final spectral library contained a total of 12,487 proteins and 144,578 precursors. DIA-NN (version 1.8) (Demichev et al. 2020) was used to process the MS data using this spectral library, implemented using RT-dependent normalization. See **Table S1** for the full code and parameters used to run DIA-NN. All MS runs, as well as the FASTA and spectral library files, have been deposited in the Proteomics Identification Database (PRIDE) (Perez-Riverol et al. 2019) with identifier PXD030304.

DIA-NN output data were filtered to retain only precursors from proteotypic peptides with Global.Q.Value $\leqslant$ 0.01. These precursors were then used for protein quantification by maxLFQ (Cox et al. 2014), implemented using the DiaNN R Package (https://github.com/vdemichev/diann-rpackage) and with default parameters. Data were then log2 - transformed. 117 files were discarded from downstream analyses (**Table S1**), as follows: one MS run recorded no peptides, six replicates of SW900 were removed because the cell line failed STR profiling; 32 files from the earliest pilot batch were removed, as these were repeated later in the experiment; 39 files that quantified fewer than 2,000 proteins were removed; 39 files were removed because they had a poor correlation across replicates. Cell lines with a poor replicate correlation were identified using two methods. First, the minimum correlation between replicates was calculated for each cell line. The 10% of cell lines with the lowest correlation across the cohort were then examined to identify whether any MS run had a correlation with an MS run from another cell line that was above the 75% percentile of correlations (n = 11 cell lines). MS runs were then discarded for each cell line if manual

examination of replicate correlations indicated that a sample mix up could have occurred (n = 21 MS runs discarded). Second, any cell line was selected that had a minimum correlation in at least one replicate of < 0.8 or a coefficient of variation, from proteins observed in > 80% of the cohort, across replicates of > 30% (n = 9 cell lines). These MS runs were then also manually examined for each cell line and MS runs that were discordant with the remainder of replicates were removed (n = 18 MS runs discarded). The final dataset, termed ProCan-DepMapSanger, was derived from 6,864 mass spectrometry runs covering 949 cell lines (**Table S1**) and quantifying a total of 8,498 proteins (**Table S2**). A filtering was applied to identify protein quantifications derived from more than one supporting peptide (n = 6,692 human proteins; **Table S2**. MS runs across replicates of each cell line were combined by calculating the geometric mean. Protein quantifications and number of peptides identified per protein in each MS run are available in figshare https://doi.org/10.6084/m9.figshare.19345397.

Assembly of multi-omics cell line datasets

Drug response measurements were assembled from multiple studies (Garnett et al. 2012; Iorio et al. 2016; Picco et al. 2019; Gonçalves et al. 2020) and 204 new compounds were screened and dose-response curves fitted as previously described in detail (Iorio et al. 2016; Vis et al. 2016). A total of 625 unique drugs were included in our drug response dataset. All data and respective details can be accessed at www.cancerRxgene.org (W. Yang et al. 2013). Cell line growth rates were represented as the ratio between the mean of the untreated negative controls measured at day one (time of drug treatment) and the mean of the dimethyl sulfoxide (DMSO) treated negative controls at day four (72 h post drug treatment). Data acquisition and processing was performed as previously described (https://www.cancerrxgene.org/) to systematically fit drug response curves and derive half-maximal inhibitory

concentration (IC_{50}) measurements for each drug across the cell lines measured (Garnett et al. 2012; W. Yang et al. 2013; Iorio et al. 2016; Picco et al. 2019; Gonçalves et al. 2020). The dataset comprises two different screening approaches (W. Yang et al. 2013), and for drugs screened with both modalities, these were kept as separate entries for the downstream analyses by constructing a unique identifier (drug_id) with the pattern of <drug_code>_<drug_name>_<GDSC_version>, resulting in 819 drug_ids. A threshold of a minimum of 300 cell lines was applied to exclude drugs that were screened without enough cell lines for DeeProM analysis, resulting in a total of 710 drug_ids. The natural log of the raw IC50 was used for all computations.

RNA sequencing (RNA-seq) transcriptomics and Infinium HumanMethylation450 methylation measurements for the same set of cancer cell lines were assembled from previous analyses, for which the acquisition and processing are described in detail (Garcia-Alonso et al. 2018; Iorio et al. 2016). Mutation and copy-number calls were inferred from whole-exome sequencing and Affymetrix SNP6 arrays, respectively, as described previously (Iorio et al. 2016).

Genome-wide essentiality measurements for 17,486 genes were assembled for 534 cancer cell lines that overlap with those analyzed in the ProCan-DepMapSanger dataset, using CRISPR-Cas9 screens (Pacini et al. 2021). This is an integrated CRISPR-Cas9 dataset derived from two projects (Behan et al. 2019; Meyers et al. 2017) that removes library biases and represents gene essentiality as log2 fold-changes corrected for copy number bias (Iorio et al. 2018).

Dimensionality reduction and visualization

Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018) was calculated using Python package umap-learn (v.0.4.2) with the default setting of 15 nearest neighbours and the first 50 principal components derived

from the protein matrix. Missing values were replaced with a value representing the first percentile of the input data matrix to calculate the principal components with the Python package scikit-learn (v.0.22.1) (Pedregosa et al. 2012). The first two dimensions were used for visualization.

Multi-omics factor analysis

Multi-omics decomposition by factor analysis was performed using the mofapy2 Python module (v0.5.6) (Argelaguet et al. 2018, 2020). Datasets with continuous measurements were selected for this analysis, i.e., drug responses, methylation, proteomics, and transcriptomics. For proteomic measurements, we used both the ProCan-DepMapSanger dataset and an independently acquired dataset (Nusinow et al. 2020) measuring an overlapping set of 290 cancer cell lines. Considering the strong separation of hematopoietic and lymphoid cell lines from the rest of the cell lines (**Figure 2A**), these were treated as a separate group in the analysis. Different numbers of factors were tested and n = 15 was chosen as it represented a trade-off between the total variance explained and the correlation between factors. Higher numbers of factors increased the correlation between factors and only marginally increased the variance explained, indicating that some factors were unnecessary. For the ProCan-DepMapSanger dataset, the mean sample intensity was regressed out prior to the factor analysis, thereby avoiding it being captured by any factor and artifactually increasing the total variance explained. Mofapy2 was run with convergence mode set to 'slow'. Scale views and groups were set to 'True' to have a unit variance.

Pairwise protein-protein correlations

We considered proteins with corresponding data also measured in the transcriptomics and CRISPR-Cas9 datasets (n = 6,347). For all pairwise protein combinations, we calculated Pearson's r correlations between their protein, gene

expression and essentiality measurements. A minimum of 15 complete observations was required to calculate the correlation, yielding 16,580,952 pairwise combinations. Protein-protein correlations were annotated using multiple sources of protein interactions: Comprehensive Resource of Mammalian Protein Complexes (CORUM) (Ruepp et al. 2010); Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk et al. 2017); Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al. 2015) ; and Human Protein Interactome (HuRI) (Luck et al. 2020). For BioGRID, only physical interactions between proteins from humans were considered. For STRING, the most stringent threshold of the confidence score was chosen, and only interactions with a score $\geqslant$ 900 were considered. The average path length of the STRING PPI networks was 3.9. Protein-protein Pearson's correlations were then used to estimate the capacity to recover interactions from the different resources by ranking in ascending order all correlations according to their p value (x axis) and drawing the cumulative distribution curve of the interactions found in the resource (y axis). The area under the recall curve (AUC) was estimated using the corresponding function from the Python package scikit-learn (v0.24.2) (Pedregosa et al. 2012).

DeeProM (Deep Proteomic Marker) – Overview

We developed a multistep computational workflow, Deep Proteomic Marker (DeeProM), to identify protein biomarkers of cancer vulnerabilities. The analysis steps in DeeProM are fourfold. First, it prioritizes drug responses and CRISPR-Cas9 gene essentialities that can be confidently predicted using proteomic profiles. Second, it prioritizes strong drug responses and gene essentialities that are specific to small subsets of cancer cell lines. Third, it prioritizes protein biomarkers that show

significant associations with drug responses or gene essentialities. Fourth, it prioritizes protein biomarkers that are present in specific tissues.

DeeProM - DeepOmicNet model

DeeProM is powered by a deep neural network architecture, DeepOmicNet, to predict drug responses and CRISPR-Cas9 gene essentialities. DeepOmicNet ranks drugs and gene essentialities based on predicted cellular responses using the ProCan-DepMapSanger dataset as the input. Both the proteomic and the drug response datasets contain missing values, while the gene essentiality dataset provides a complete data matrix. DeepOmicNet models these missing values accordingly. Multilayer perceptron (MLP) is a classic neural network architecture that has been used by default for deep learning in numerous biomedical studies (Z. Zhang et al. 2019). To enhance the predictive performance of MLP with proteomic data, we modified its network architecture and developed DeepOmicNet with the following three major improvements:

Grouped bottleneck. DeepOmicNet uses grouped bottlenecks to avoid fully connected layers, which involves a large number of parameters being optimized. Compared with a fully-connected layer, breaking the connections into smaller groups allows the network to be more memory efficient, thus enabling wider or deeper layers. A weight matrix $W \in R(k\ x\ k)$ containing k^2 parameters with grouped bottlenecks can not only reduce the number of parameters, but also provide better predictive performance. Instead of connecting all pairs of neurons, neurons can be broken into groups, and only neurons within the same group are connected between layers (**Figure S3B)**. The group size g can be set as any number that is divisible by the hidden layer width k. When $g = k$, all neurons are treated as one group, which reduces to a normal fully-connected layer. Multiple configurations were tested and the optimal group size

g was set to 2. The number of parameters for one layer with grouped bottlenecks is significantly reduced from $k^2$to $\frac{k}{2} \times 2^2 = 2k$. The number of parameters with grouped bottlenecks is calculated as $k/g \times g^2 = g \times k$, thus the run-time complexity is decreased from quadratic to linear.

Skip connections. The complete network architecture is visualized in **Figure S3B**. Neurons between every two consecutive layers are connected in MLP, which is computationally intensive and suboptimal for model training. To mitigate this problem, DeepOmicNet utilizes skip connections (He et al. 2016) to connect alternate layers. Let $x \in R^k$ be the vector of the ith hidden layer of a real coordinate space of dimension k (corresponds to the number of neurons in a layer, also known as the layer width), the value of xi is calculated with skip connections as:

$$x_i = f(\, W_{i-1}\, x_{i-1} + b_{i-1} + x_{i-2})$$

where *f* is the activation function, $W_{i-1} \in R^{(kxk)}$is the weight matrix, $b_{i-1} \in R^k$ is the bias vector and $x_{i-2} \in R^k$is the hidden layer ahead of the hidden layer x_{i-1}. The value of the hidden layer *i-2* is fed into the hidden layer *i* by skipping the hidden layer *i-1*, resulting in skip connections (**Figure S3B**). Each hidden layer is set with the same width *k*, which is a hyperparameter for model tuning, and usually is chosen to be slightly smaller than the input feature dimension. In DeepOmicNet, a sigmoid function was chosen to be the activation function *f*, because it outperformed the rectified linear activation function and the hyperbolic tangent function.

Loss function. DeepOmicNet is trained with mini-batches using a customized mean squared error (MSE) as the loss function. DeepOmicNet is applied to predict both drug responses and CRISPR-Cas9 gene essentialities. For the cell line *m*, the loss for the target variable *n* is defined as:

$$L_{m,n} = \begin{cases} (y_{m,n} - \hat{y}_{m,n})^2 \; if \; ground \; truth \; is \; present \; for \; n \\ 0 \; otherwise \end{cases}$$

where $y_{m,n}$ is the ground truth label of the target n (either drug response or gene essentiality) and $\hat{y}_{m,n}$ is the predicted value of the target n. The label $y_{m,n}$ is missing if a particular cell line m was not screened with drug n.

In addition to the three major improvements, other characteristics of DeepOmicNet include the following:

Missing values. One specific challenge of DIA-MS based proteomics is the missing values in the data matrix (Webb-Robertson et al. 2015; Poulos et al. 2020). Imputation is widely used but often leads to distortion to some extent (Runmin Wei et al. 2018). For DeepOmicNet, missing values were replaced by zeros, thus allowing the neural network to ignore the weight update for these inputs.

Hyperparameter tuning. Hyperparameters including model width, depth, learning rate and batch size were tuned to achieve the highest predictive performance. Pearson's r between true and predicted values is used as the evaluation metric. Hyperparameters that resulted in the highest performance in five-fold cross-validation of the 80% training data were chosen for the final evaluation on the 20% independent test set. The chosen hyperparameters can be found in the configuration files in the source code (see Data and code availability).

Thresholding. The thresholds are set to Pearson's r > 0.4 and Pearson's r > 0.3 to prioritize highly predictive drug responses and CRISPR-Cas9 gene essentialities, respectively. This yielded 67 drug responses and 62 gene essentialities.

DeeProM - Prioritising selective associations

DeeProM prioritizes drug responses and CRISPR-Cas9 gene essentialities that are likely to be non-toxic to normal cells. Since the cell lines used in this study were

derived from cancer (the majority) or viral transformation, we approximated this task by finding drug responses and gene essentialities that were selective for only a small fraction of the cell lines, which indicates that the drug response or gene essentiality is less likely to be toxic to normal cells. Ranking of strongly selective drug responses and gene essentialities was performed using Python package scipy (v1.5.2) skew function, which calculates the Fisher-Pearson's coefficient of skewness. Skewness values of −1 and −2 were used for drug responses and gene essentialities, respectively.

DeeProM - Linear regression models

Associations between protein and phenotypic measurements, drug responses and gene essentialities were performed using linear regression models (sklearn v0.24.2 class LinearRegression). Several technical and biological covariates were added to the model to remove potentially spurious associations. First, we built the following technical covariates into the model: (i) the growth rate of the cell lines; (ii) cell culture medium, D/F12 (DMEM/F12: 10% FBS, 1% PenStrep) or R (RPMI 1640: 10% FBS, 1% PenStrep, 4.5 mg/mL Glucose, 1 mM Sodium Pyruvate); (iii) cell line growth properties (i.e., adherent, semi-adherent or suspension), ploidy, and if they are hematopoietic and lymphoid cell lines; (iv) sample mean protein replicates Pearson's correlation; (v) for the CRISPR-Cas9 gene essentiality only, we considered the institute of origin of the CRISPR-Cas9 screen, i.e., Wellcome Sanger or Broad Institute; and (vi) for the drug response models only, we considered the cell line mean IC50 across all drugs. Discrete covariates were represented as dummy binary variables. Second, to identify associations that were exclusively found at the protein level, we added the following gene expression covariates to the model: (i) the first ten gene expression principal components using the Python package scikit-learn (v0.24.2) (Pedregosa et al. 2012) ; and (ii) the corresponding transcript level of the protein being

tested. Formally, we fit the following linear regression model for each drug response/gene essentiality–protein pair:

$$d = M\beta_0 + E\,\beta_1 + e\,\beta_2 + p\,\beta_3 + \varepsilon$$

where, d represents a n x 1 vector of the drug response IC_{50} (n = 710 drugs) for 947 cell lines or CRISPR-Cas9 gene essentiality $\log_2$ fold changes (n = 17,486 genes) for 534 cell lines; M is the n x k matrix of covariates (k = 11 covariates); E is a n x m matrix containing the first (m = 10) principal components of the gene expression dataset; e is a vector of size n x 1 containing the transcriptomics measurements of the corresponding protein p; p is a vector of size n x 1 containing the protein measurements; and ε is the error vector of size n x 1. For each protein, cell lines with missing values were dropped from the modelling. For drug response, missing values were replaced by the drug mean IC_{50}. The number of cell lines with complete information per fit was provided. The model was fitted by minimizing the residual sum of squares to estimate the parameters β_n of each variable. In total, there were 710 x 6,692 = 4,751,320 drug-protein pairs and 17,486 x 6,692 = 117,016,312 possible CRISPR-Cas9 gene essentiality-protein pairs, however, both drug response and CRISPR-Cas9 data covered various subsets of cell lines. We required a minimum number of 60 cell lines to test the association. As a result, a total of 4,218,788 drug-protein and 86,584,537 CRISPR-Cas9-protein tests were performed.

Statistical assessment of the improvement of adding protein measurements to the linear regression was performed using a likelihood ratio test between the full mode and the null model, which excludes the protein measurement and its parameter β_2. Likelihood ratio test's p-value was estimated using a chi-square distribution with one degree of freedom. Adjustment for multiple testing was performed per drug or CRISPR-Cas9 gene essentiality using the Benjamini-Hochberg procedure to control

the false discovery rate (FDR). Associations with FDR < 0.1 for models with covariates, or FDR < 0.001 for models that do not use covariates, were considered as significant associations.

Statistical assessment of the improvement of adding protein measurements to the linear regression was performed using a likelihood ratio test between the full model and the null model, which excludes the protein measurement and its parameter β_2. Likelihood ratio test's *p*-value was estimated using a chi-square distribution with one degree of freedom. Adjustment for multiple testing was performed per drug or CRISPR-Cas9 gene essentiality using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). Associations with FDR < 0.1 for models with covariates, or FDR < 0.001 for models that do not use covariates, were considered as significant associations.

DeeProM - Tissue type level filtering

To investigate drug-protein and CRISPR-Cas9-protein associations for a given tissue type, we used metrics derived from DeepOmicNet, Fisher Pearson's coefficient of skewness and linear regression to prioritize drug responses and CRISPR-Cas9 gene essentialities according to the thresholds described above. The overlap of these three methods was used as the final result, which included 7,698 drug-protein and 5,823 CRISPR-Cas9-protein associations. Finally, we applied additional filters to further prioritize associations that are unique to certain tissue types, yielding the final list of 108 drug-protein associations for 18 drugs and 1,538 CRISPR-Cas9-protein associations for 38 genes (**Figure 4C**). The filtering steps are described below:

Step 1. Tissue types with < 20 protein measurements were filtered out.

Step 2. For each significant association identified from 7,698 drug-protein and 5,823 CRISPR-Cas9-protein associations, Pearson's *r* was calculated. Using protein

data, the significance level was set to 0.1 and Pearson's *r* was defined as $r_{(target, protein)}^{tissue}$ for a target-protein association in a given tissue type. Here, protein indicates a protein of interest, target indicates either a drug response or CRISPR-Cas9 gene essentiality and tissue refers to the tissue type under investigation. We then calculated the Pearson's *r* for the gene that encodes each protein, and we defined this value as $r_{(target, RNA)}^{tissue}$. To prioritize associations that were uniquely identified at the protein level, the difference *d* for a given tissue type between target-protein and target-RNA associations was set to be larger than 0.15, where $d = |r_{(target, protein)}^{tissue}| - |r_{(target, RNA)}^{tissue}|$. This prioritizes associations that have either strong positive or strong negative correlations at the protein level but have weak correlations around zero at the RNA level. Rare cases where $r_{(target, protein)}^{tissue}$ and $r_{(target, RNA)}^{tissue}$ have opposite signs and *d* is close to 0 were not considered.

Step 3. For drug-protein associations, a large value of *d* alone is insufficient to select candidate associations, because a drug may be entirely ineffective for all the cell lines in a particular tissue type. Therefore, we applied an additional filter to ensure that a drug is effective on the cell lines for which the protein abundance is high. That is, for a drug-protein association to be included for a given tissue type, the median IC_{50} of the 20% of cell lines with the highest corresponding protein abundance must be lower than the maximal concentration for that drug.

Step 4. The remaining associations were ranked in descending order according to *d*.

Comparing DeepOmicNet and other models

DeepOmicNet was compared against traditional machine learning models, including elastic net and Random Forest. A total of 947 cell lines were randomly separated into a training set comprising 80% of the cell lines, and a test set with the

remaining cell lines for unbiased evaluation. Grid search was used to find the best hyperparameters for elastic net and Random Forest in the training set by cross-validation. Hyperparameter tuning for DeepOmicNet was performed manually due to the limit of a graphics processing unit (GPU) memory. For each model, missing values were imputed using the method that gave the best prediction based on cross-validation. Specifically, four imputation methods were considered, including imputation by minimum, first percentile of the whole input matrix, mean and zero. Based on the predictive performance of models in cross-validation, imputation by one percentile of the whole matrix was chosen for the elastic net and Random Forest, and imputation by zero was used for DeepOmicNet. This strategy yielded the best prediction accuracy in comparison with other imputation and normalization methods, such as imputation with k-nearest-neighbor, mean values of proteins and zeros. A cut-off was set at a minimum of 300 screened cell lines for the drug response dataset to filter out drugs without sufficient data. A simplified version of DeepOmicNet without grouped bottlenecks was used for omics data other than proteomic data due to the large input dimension. The Python package scikit-learn (v.0.22.1) (Pedregosa et al. 2012) was used to train elastic net and Random Forest models. DeepOmicNet was implemented and trained using PyTorch (v.1.4.0) (Paszke et al. 2019).

Machine learning for lethality prediction

Due to the limit of GPU memory, elastic net, Random Forest and DeepOmicNet were applied only to transcriptomic and proteomic data to predict CRISPR-Cas9 gene essentialities. The same computational strategy for drug response prediction was used to predict CRISPR-Cas9 gene essentialities.

Predictive power comparison with CCLE

The predictive power of machine learning models for two proteomic datasets (ProCan-DepMapSanger and the Cancer Cell Line Encyclopedia (CCLE) (Nusinow et al. 2020)) were compared independently on three drug response datasets (Sanger, CTD2 and PRISM) and the CRISPR-Cas9 gene essentiality dataset (Behan et al. 2019; Meyers et al. 2017; Pacini et al. 2021). The analysis was performed using the 290 overlapping cell lines to ensure a fair comparison. The proteomic (Nusinow et al. 2020) and drug response (CTD2 and PRISM) (Corsello et al. 2020; Seashore-Ludlow et al. 2015; Rees et al. 2016) datasets were retrieved from the DepMap portal (https://depmap.org/portal/). DeepOmicNet was used to compare the predictive power of models for CRISPR-Cas9 gene essentialities, and Random Forest was used for drug response prediction due to the limited number of cell lines for certain drugs. AUC instead of IC_{50} was used for the drug response (PRISM) dataset due to a large proportion of drugs having no IC_{50} values provided.

Downsampling for drug response prediction

For downsampling analysis, the full set of 8,498 proteins were randomly downsampled using a step decrease of 500 proteins (**Figure 7A**). Each step was repeated ten times and for each iteration, results from five-fold cross-validations and an unbiased test were included in evaluating predictive power. Therefore, each downsampling step used ten different random subsets of proteins for six distinct experiments (the five-fold cross-validation and one unbiased test). The predictive power of each DeepOmicNet model was evaluated for each protein set and each iteration, with confidence intervals summarizing the results across the ten iterations. This random downsampling procedure was also performed with step sizes of 250 proteins for proteins in Categories A, B and C (**Figure 7E**).

3.5.3 Quantification and statistical analysis

DIA-NN (version 1.8) (Demichev et al. 2020) was used to build the peptide spectral library and process raw MS data. MaxLFQ (Cox et al. 2014) was used to quantify relative protein intensities. Sigmoid drug response curves were fitted to estimate IC50s (Vis et al. 2016). Associations between pairs of continuous variables were tested by Peason's correlation coefficient r. Statistical tests were adjusted for multiple hypotheses correction using the Benjamini-Hochberg False Discovery Rate (FDR), and statistical significance was considered when FDR < 5%, except when otherwise specified (such as when multiple thresholds were compared). Quantification methods and statistical analyses for the proteomics, drug response and multi-omics datasets are described in the respective sections of the STAR Methods. Unless otherwise stated, relevant statistical parameters are reported in the legend of each figure.

3.6 DATA AND CODE AVAILABILITY

All the source codes are available at the GitHub repository: https://github.com/EmanuelGoncalves/cancer_proteomics

Chapter 4: Transformer-based deep learning integrates multi-omic data with regulatory pathways in cancer

Text and figures included in this chapter are adapted from the following publication:

Cai, Z., Poulos, R. C., Aref, A., Robinson, P. J., Reddel, R. R., & Zhong, Q. (2022). Transformer-based deep learning integrates multi-omic data with cancer pathways. In *bioRxiv* (p. 2022.10.27.514141). https://doi.org/10.1101/2022.10.27.514141.

Statement of Contribution

The PhD Candidate completed all data analyses presented in this chapter, under the supervision of Dr. Rebecca C Poulos, Dr. Qing Zhong, Prof. Phillip J Robinson and Prof. Roger Reddel. The PhD Candidate was also responsible for writing this chapter and the preparation of all figures. Dr. Adel Aref contributed to the writing of the introduction of this work.

4.1 ABSTRACT

Multi-omics data analysis powered by machine learning has significantly improved cancer diagnosis and prognosis. However, traditional machine learning methods only consider omics measurements, failing to integrate domain knowledge such as biological networks that link different omic layers via regulatory pathways. We develop a Transformer-based deep learning model DeePathNet, integrating cancer pathway information, to analyse multi-omics data with model explanation. DeePathNet robustly outperforms traditional methods for the prediction of drug response as well as cancer type and subtype using a variety of large datasets including GDSC, CCLE, TCGA and CPTAC. Combining biomedical knowledge and the power of deep learning, DeePathNet enables reliable biomarker discovery at the pathway level, paving the road to data-driven cancer research and precision medicine.

4.2 INTRODUCTION

Multi-omic analysis of diverse data types enables researchers to gain insights into tumour biology and to identify new and robust therapeutic targets (Mani et al. 2022). One major goal of multi-omic analysis by machine learning is to predict the cancer treatment strategies that are best suited to individuals in the context of precision medicine. A variety of multi-omic studies have led to the improved detection of intra-tumour heterogeneity, identification of novel therapeutic targets, as well as more robust diagnostic and predictive markers (Reel et al. 2021; Picard et al. 2021; Rohart, Gautier, Singh, and Lê Cao 2017; I. Subramanian et al. 2020). Many of these discoveries would not have been possible by analysing any single omic data type alone. However, performing multi-omic analysis presents computational challenges due to the large number of data generated by high-throughput instruments and the limitations of existing multi-omic data integrative methods (Tarazona, Arzalluz-Luque, and Conesa 2021; Cai et al. 2022).

To address this, a plethora of machine learning methods have been developed for integrating large-scale multi-omic data (Cai et al. 2022; Reel et al. 2021; Picard et al. 2021; Rohart, Gautier, Singh, and Lê Cao 2017; Meng et al. 2016; Shen, Olshen, and Ladanyi 2009; Mo et al. 2018; A. Singh et al. 2019). For example, moCluster (Meng et al. 2016) integrates multi-omic data based on joint latent variable models, showing performance superior to previous methods such as iCluster (Shen, Olshen, and Ladanyi 2009) and iCluster Bayes (Mo et al. 2018). Likewise, mixOmics (A. Singh et al. 2019) provides various options for multi-omic data integration, aiming to find common information between different omic data types. These models solely take omic measurements as the input and do not consider existing biomedical knowledge that links different omic data types together, such as the regulatory networks.

Regulatory networks exist in cells to control the expression levels of different gene products, through collections of functionally interacting protein or RNA macromolecules (Karlebach and Shamir 2008). However, models that incorporate existing biomedical knowledge in addition to computational inference have the potential to better capture the interactions that drive biomarker associations, and to increase the predictive power and modelling capacity of these algorithms (Cai et al. 2022).

Several studies have attempted to incorporate existing biomedical knowledge into multi-omic models using deep learning (Kang, Ko, and Mersha 2022). DCell (Ma et al. 2018) and DrugCell (Kuenzi et al. 2020) combine the neural network architecture with known gene ontology information, but they only support the use of gene deletions or mutation as the input. EMOGI (Schulte-Sasse et al. 2021) was designed based on graph neural networks (Wu et al. 2021) and integrates protein-protein interaction (PPI) networks with multi-omic data to predict cancer genes, but its network architecture cannot be easily generalised to other tasks. Besides, gene ontology information and PPI networks used in these models do not precisely reflect cancer-specific information. Therefore, integrating cancer pathways (Kuenzi and Ideker 2020) into multi-omic data analysis by deep learning for general tasks, such as drug response prediction and cancer type or subtype classification, remains an open research topic.

To address this gap, we developed DeePathNet, a Transformer-based (Vaswani et al. 2017) explainable deep learning method that inputs multi-omic data alongside knowledge of cancer pathways. The Transformer is used primarily in the fields of natural language processing and computer vision, by adopting the mechanism of self-attention (Vaswani et al. 2017). In molecular biology, the Transformer-based model AlphaFold has successfully predicted protein structures based on amino acid

sequences (Jumper et al. 2021). However, the Transformer has not yet been incorporated into multi-omic cancer data analysis. Here, we apply DeePathNet by using a Transformer module to integrate several large-scale multi-omic datasets with cancer pathways, allowing more complex patterns to be learned. By comprehensively evaluating multiple datasets with three prediction tasks and a range of metrics, we demonstrate that the predictive power of DeePathNet is superior to that of traditional machine learning methods.

4.3 RESULTS

4.3.1 Overview of DeePathNet

DeePathNet was developed to model biological pathways using a Transformer-based deep learning architecture with both multi-omic data and cancer pathway information as the input **(Fig. 1a)**. The performance of DeePathNet was evaluated on drug response prediction, and cancer type and subtype classification.

DeePathNet consists of three major steps. It starts with a pathway encoder to summarise features from an arbitrary number of omic data types into cancer pathways (Step 1; **Fig. 1b**), and then uses a Transformer encoder to model the interactions between these pathways (Step 2). This is followed by a multi-layer perceptron (MLP) that can be adapted to different prediction tasks (Step 3).

In Step 1, the neural network architecture is constructed based on the LCPathways dataset (Kuenzi and Ideker 2020), which contains 241 literature-curated pathways encompassing 3,164 cancer genes. The LCPathways dataset was selected since it is one of the most recent and comprehensive pathway databases that are specifically curated for cancer research. As such, it is particularly suitable for the applications of DeePathNet. The pathway encoder then uses a fully connected layer to project the multi-omic data (Omics 1–*m*) from genes (Gene 1–*n*) onto a 512-dimension

pathway vector that represents one of the cancer pathways (Pathway 1–*p*; **Supplementary Fig. 1a**, see **Methods**). With this architecture, the pathway encoder allows DeePathNet to capture interactions across different omic data types.

In Step 2, an enhanced version of the Transformer module is developed to encode the interactions between cancer pathways (**Supplementary Fig. 1b,** see **Methods**). First, a dropout layer is used to train only half of the pathways at each iteration, which prevents the model from focusing on specific pathways that may not generalise well to a test dataset. Then, two blocks of the original Transformer module (Vaswani et al. 2017) are used, which contains a list of recurring layers with each layer comprising a sequence of layer normalisation, multi-head self-attention, and a MLP. The Transformer also enables dynamic modelling of the complex relationship between cancer pathways, thus avoiding the generation of fixed weights for the different input, as is the case in traditional machine learning.

In Step 3, a MLP is used to map the encoded pathway vectors to output neurons, which allows the knowledge learned by the Transformer module to be adapted to general prediction tasks.

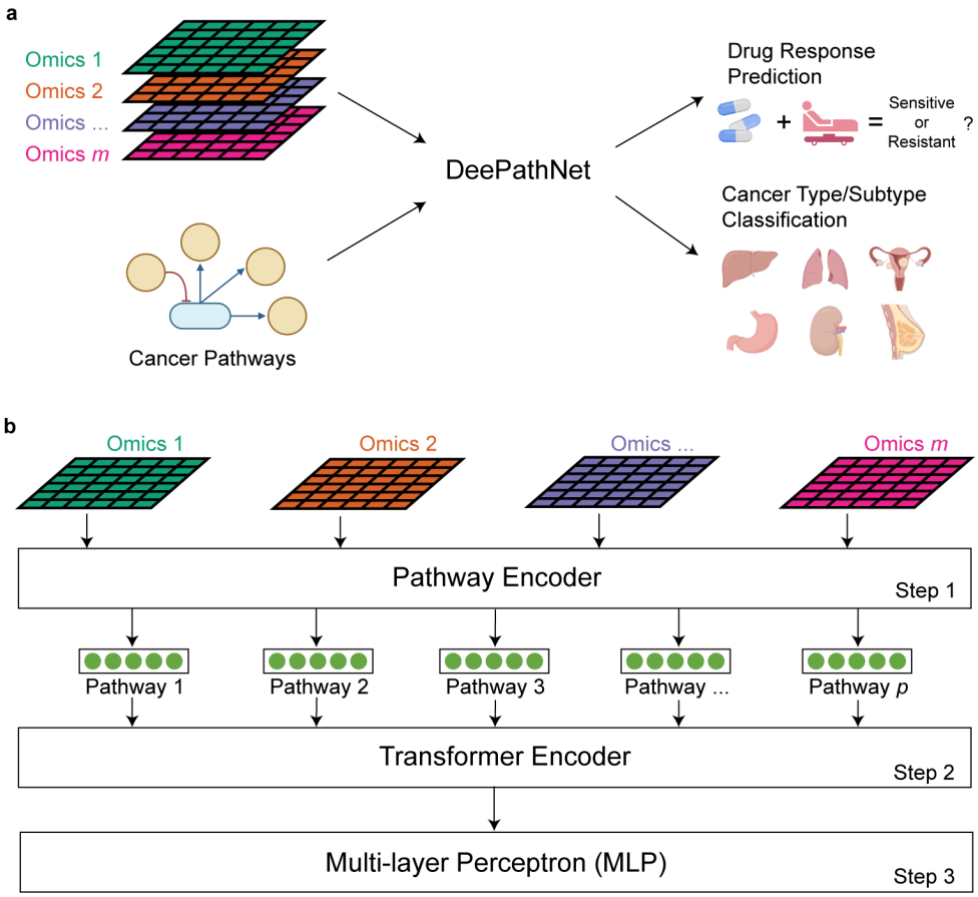


Fig. 1 | Overview of DeePathNet. a, DeePathNet has its network architecture built using the LCPathway dataset and takes multi-omic data as the input to model pathway interactions and predict drug responses or classify cancer types and subtypes. **b,** DeePathNet architecture supports any number of omic data types as the input. Step 1: DeePathNet encodes multi-omic information into an arbitrary number of cancer pathways. Step 2: DeePathNet uses a Transformer encoder to learn the interactions between these pathways. Step 3: The encoded pathway vector is passed into a MLP for the prediction. Circles represent neurons in a neural network. Arrows represent the direction of information flow.

4.3.2 DeePathNet predicts drug response

We first assessed the predictive performance of DeePathNet on a regression task by benchmarking it against random forest (Ho 2002), elastic net (Zou and Hastie 2005), principal component analysis (PCA), mixOmics(Rohart, Gautier, Singh, and Lê Cao 2017) and moCluster (Meng et al. 2016) to predict the responses of anti-cancer drugs to cancer cell lines. These six methods were evaluated using data from the Cell Lines Project (CLP) (Iorio et al. 2016) and the Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al. 2019), the two largest publicly available multi-omic cancer cell line datasets (**Supplementary Table 1,** see **Methods**). Gene mutation, copy number

variation (CNV) and gene expression data from the two datasets were used as the input. For drug response data, we retrieved the half-maximal inhibitory concentration (IC_{50}) from the Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al. 2016) database. For each method, six experimental setups were assessed, comprising two datasets and three evaluation metrics, namely coefficient of determination (R^2), mean absolute error (MAE) and Pearson correlation coefficient (Pearson's *r*) between predicted and actual IC_{50} values.

In DeePathNet, 241 pathway encoders were constructed (**Supplementary Fig. 1a**) to summarise the omic data into pathway vectors defined by the LCPathways (Kuenzi and Ideker 2020). These vectors were then fed into the Transformer module to model the interactions between cancer pathways (**Supplementary Fig. 1b**). Default hyperparameters were used for all six methods (see **Methods**). Omic data were combined using early integration (Cai et al. 2022) for random forest and elastic net. Middle integration (Cai et al. 2022) was used for PCA, moCluster and mixOmics. PCA and moCluster were coupled with random forest for predictions (Cai et al. 2022) (see **Methods**).

To quantitatively and reliably compare the six methods, five-fold cross-validation was repeated five times at random, yielding 25 error measures for each of the R^2, MAE and Pearson's *r* metrics. The mean and 95% confidence interval (CI) of the evaluation metrics was reported, serving as an estimate of the generalisation error. We observed that DeePathNet had significant and consistently better performance in drug response prediction than the other five methods that do not incorporate cancer pathway information (**Fig. 2a-f**, p-value < 0.0001, two-tail paired Student's t-test, **Supplementary Table 2**). By ranking the methods according to the mean measures for each setup, we found that random forest was the second-best performing method

(**Fig. 2g**). To investigate whether drug responses that had relatively lower predictive accuracy by DeePathNet were also challenging for other methods, the correlations between DeePathNet and the other five methods were calculated. We found that the predictive performance of the paired methods was highly concordant (Pearson's r > 0.9), with DeePathNet consistently outperforming the other five methods (**Supplementary Fig. 2**).

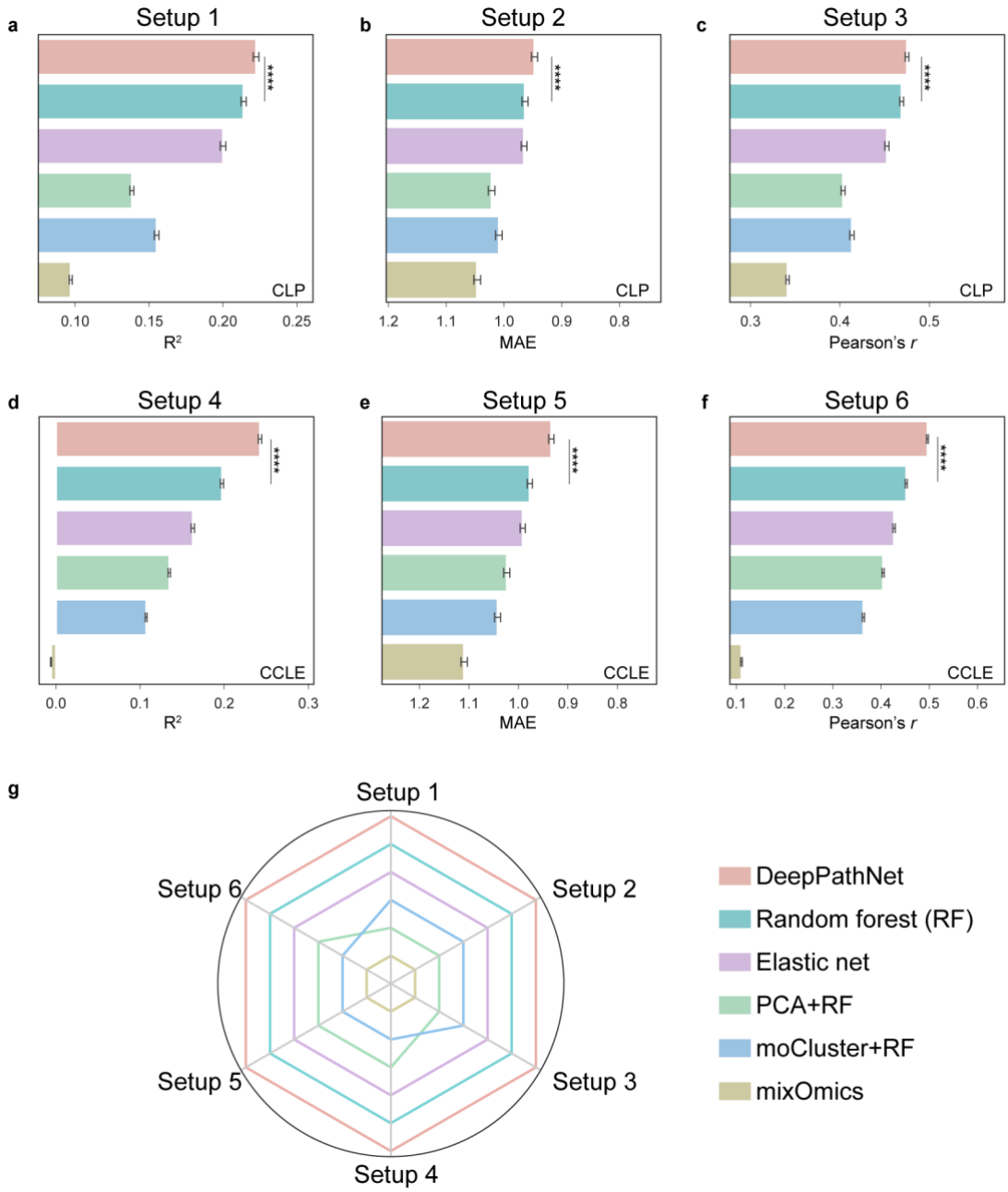


Fig. 2 | Performance evaluation of drug response prediction by cross-validation. a-f, Bar plots showing predictive performances across six experimental setups on the CLP and CCLE datasets by three evaluation metrics: R^2, MAE (inverted on the horizontal axis), and Pearson's *r*. A higher value represents better performance. Error bars are derived from cross-validation, representing 95% confidence intervals of the mean. **** indicates *p*-value < 0.0001 by two-tail paired Student's *t*-test, only showing significance between the first- and second-best performing methods. **g,** Radar plot showing the ranks of each model across the six experimental setups. A larger enclosed area represents better performance. **a-g**, The six methods are colour coded as in **g**.

To evaluate the generalisation error using an independent test set, we trained DeePathNet on the CLP dataset (**Supplementary Table 1**) and tested the final model by predicting drug responses in the CCLE dataset (**Supplementary Table 1**). Cancer pathway information was integrated in the same way as described above and random forest was trained as a baseline model. The test performance for all 549 GDSC anti-cancer drugs was summarised for both DeePathNet and random forest. DeePathNet achieved a statistically significant higher predictive performance than random forest across all three metrics (**Fig. 3a**, *p*-value < 0.0001, two-tail paired Student's *t*-test, **Supplementary Table 3**).

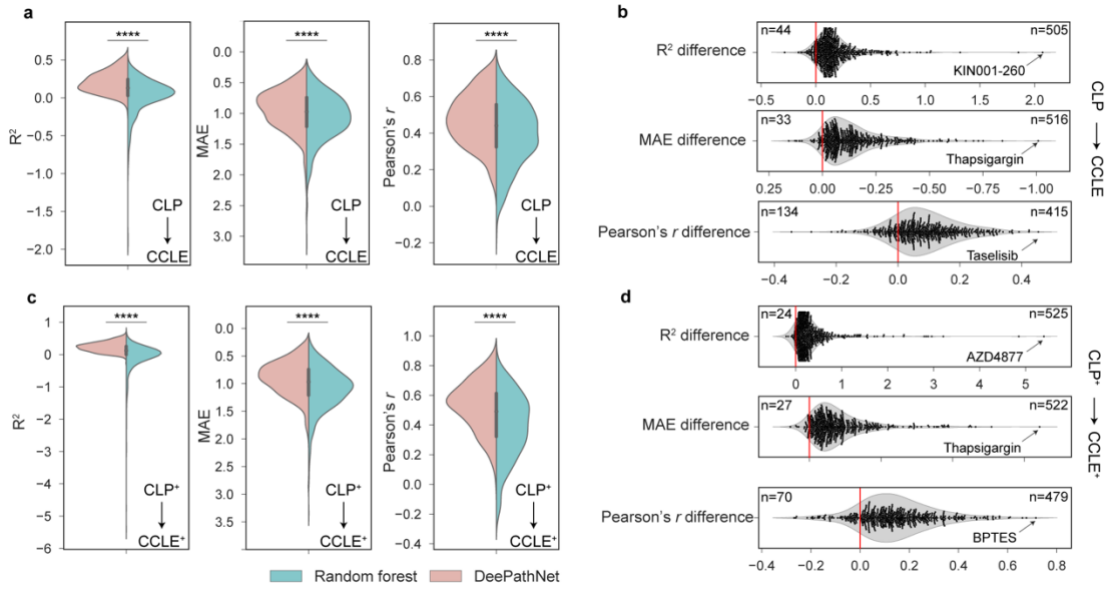


Fig. 3 | Generalisation error of DeePathNet and random forest for drug response prediction. a, Violin plots showing predictive performances of DeePathNet and random forest using CLP as the training set and evaluated on the independent CCLE test set across the 549 GDSC drugs.The vertical axis is inverted for MAE. **** indicates *p*-value < 0.0001 by two-tail paired Student's *t*-test. **b,** Violin and swarm plots showing the performance difference in R^2 (upper), MAE (middle) and Pearson's *r* (lower) between DeePathNet and random forest for each drug. A drug is more accurately predicted by DeePathNet when it exhibits a positive value for the R^2 or Pearson's *r* difference, or a negative value of the MAE difference (the horizontal axis is inverted for MAE). The numbers of drugs that are more

accurately predicted by DeePathNet or random forest are annotated on the upper right and left of the plot, respectively. The name of the drug that achieved the largest improvement with DeePathNet is annotated for each metric. **c** and **d**, Similar to **a** and **b**, but using CLP^+ as the training set and $CCLE^+$ as the independent test set.

To compare the predictive performance of DeePathNet with random forest for each drug, the difference of R^2 between DeePathNet and random forest was measured. Here, 92% (505/549) of drugs had positive values, indicating superior predictive performance from DeePathNet over random forest. Similarly, 94% (516/549) drugs and 76% (415/549) of drugs exhibited improved results by MAE and Pearson's *r*, respectively **(Fig. 3b).** This demonstrates that DeePathNet consistently achieved better predictive performance than random forest for most anti-cancer drugs. The drug that obtained the largest R^2 improvement by DeePathNet was KIN001-260 (**Fig. 3b**), which was poorly predicted by random forest and caused the long tail in the distribution of values (**Fig. 3a**). Drugs that had the largest improvement in MAE and Pearson's *r* with DeePathNet were thapsigargin and taselisib (**Fig. 3b**).

Next, we extended our analysis by including two proteomic cell line datasets from ProCan-DepMapSanger(Gonçalves et al. 2022) and CCLE (Nusinow et al. 2020). ProCan-DepMapSanger is a recently published pan-cancer proteomic dataset of 949 human cell lines generated by our team, supplementing the CLP with proteomic information. DeePathNet and random forest were trained on the combined CLP and ProCan-DepMapSanger datasets (CLP^+; **Supplementary Table 1**), with the final model tested on the expanded CCLE dataset that includes additional proteomic measurements ($CCLE^+$; **Supplementary Table 1**). Pathway information was integrated in DeePathNet as described above. DeePathNet yielded significantly higher test performance than random forest across all three metrics when predicting the 549 GDSC anti-cancer drugs (**Fig. 3c**, **Supplementary Table 3**). Analysing the predictive performance for each drug, DeePathNet also provided significant improvement for the

majority of anti-cancer drugs compared with random forest (**Fig. 3d**). The drugs that had the largest improvement by DeePathNet were AZD4877, thapsigargin and BPTES, measured by the differences of R^2, MAE and Pearson's *r*, respectively (**Fig. 3d**).

To investigate which types of drugs were most accurately predicted by DeePathNet, we grouped the 549 drugs by their canonical target cellular pathways. Drugs targeting ABL signalling and ERK MAPK signalling pathway had the highest mean Pearson's *r* between predicted and actual IC_{50} values (**Supplementary Fig. 3a**). The top 20 most accurately predicted drugs and their pathways are reported in **Supplementary Fig. 3b**.

Taking these observations together, we demonstrated DeePathNet increased predictive performance through several benchmarking analyses in predicting responses to several drugs targeting various signalling pathways.

4.3.3 DeePathNet classifies cancer types

To evaluate DeePathNet on a classification task, we used publicly available data from The Cancer Genome Atlas (TCGA) (Alexandrov et al. 2020) to classify primary cancer types. Gene mutation, CNV and gene expression features were used as the omic data input to train DeePathNet models to classify each of the 6,356 samples into one of 23 cancer types (see **Methods**). A total of seven metrics were used across the analysis to ensure reliable evaluation. The metrics are accuracy, macro-average F_1-score, precision, recall (sensitivity), area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC) and stability (see **Methods**). LCPathways was integrated in the same way as described for the drug response prediction. For benchmarking, elastic net was replaced with *k*-nearest neighbours (*k*-NN) (Fix and Hodges 1989) because elastic net does not support

classification. For all six methods, feature integration and hyperparameter settings were identical to the drug response prediction.

In the absence of an independent dataset comprising the 23 cancer types, cross-validation was performed for the six methods on the TCGA dataset, and the mean and 95% CI of the evaluation metrics were reported as an estimate of the generalisation error. DeePathNet consistently outperformed the other five machine learning methods by accuracy, macro-average F_1-score (**Fig. 4a**, **Supplementary Table 4**). In contrast, other methods such as mixOmics only performed well in one metric, indicating that these methods may be suitable for certain scenarios but can not generalise well across datasets (**Fig. 4a**). Assessing the performance of each method using a set of four metrics including accuracy, macro-average F_1-score, AUROC and stability, showed that DeePathNet was consistently top ranked, followed by random forest (**Fig. 4b, Supplementary Table 4**).

To further investigate DeePathNet's performance for each cancer type, the predicted and actual cancer type for each sample was visualised using a confusion matrix, with the number of samples, precision and recall annotated (**Fig. 4c**). DeePathNet achieved a recall of over 0.95 for most cancer types, with acute myeloid leukemia (LAML), pancreatic adenocarcinoma (PRAD), and thyroid carcinoma (THCA) as the top three most accurately classified cancer types. Rectum adenocarcinoma (READ) was the cancer type with the lowest recall, having 46% of the samples incorrectly classified as colon adenocarcinoma (COAD). The latter outcome is unsurprising because the colon and rectum are adjacent tissue types that share highly similar features, with these two cancer types often grouped together (Cancer Genome Atlas Network 2012) and are treated with similar chemotherapeutic regimens (Cancer Genome Atlas Network 2012). The cancer type exhibiting the

second-lowest recall was stomach adenocarcinomas (STAD), with 19% of STAD samples incorrectly classified as esophageal carcinoma (ESCA). This can be explained by their similar histopathology and the anatomical proximity of STAD and ESCA (Akiyama et al. 1997). Next, AUROC and AUPRC were examined for each cancer type, both displaying high performances for all cancer types, with the exception of AUPRC for READ, due to the tissue proximity of READ to COAD. (**Supplementary Fig. 4a** and **Supplementary Fig. 4b**).

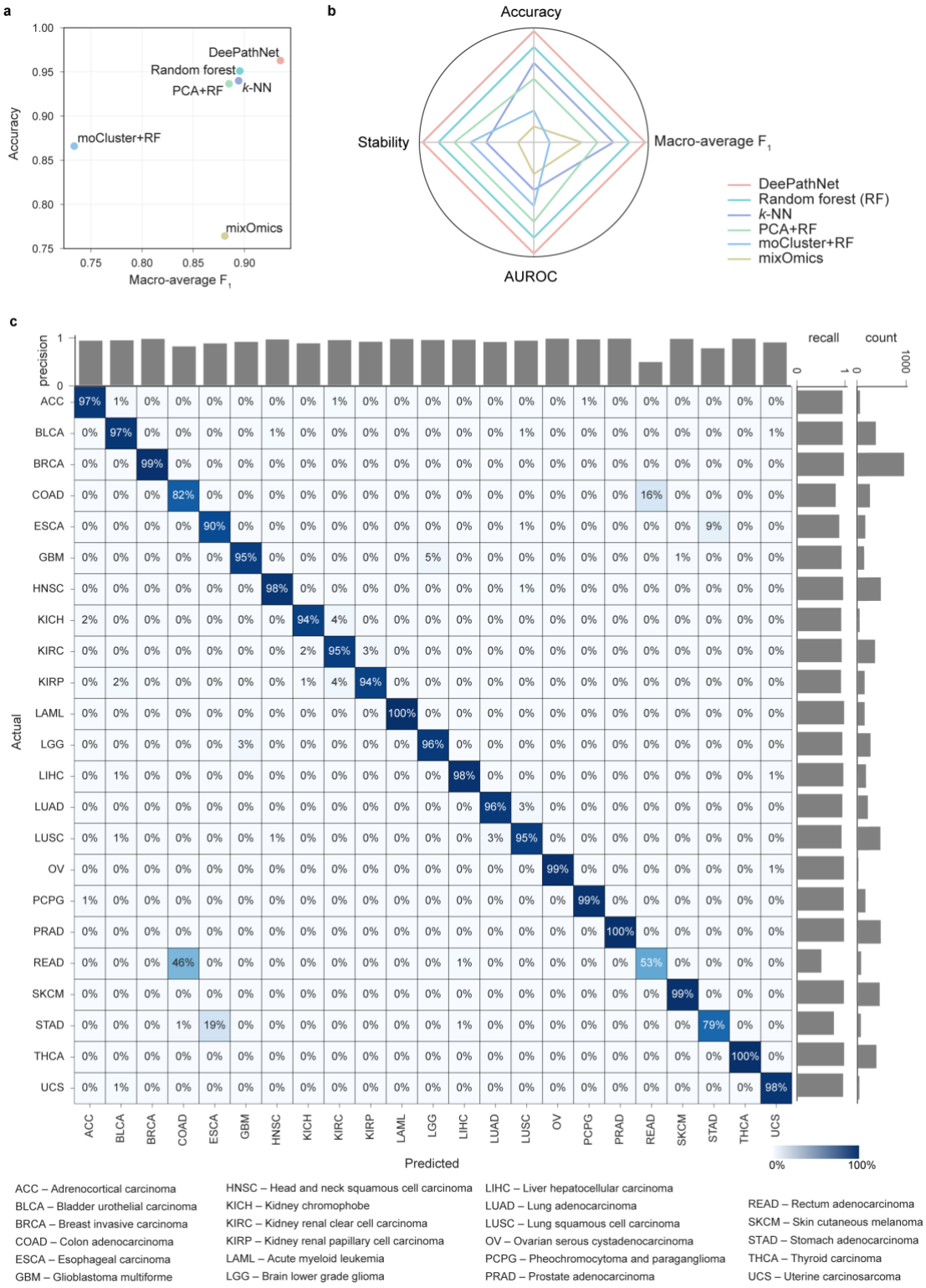


Fig. 4 | Performance evaluation of cancer type classification. a, Model comparison using cross-validation on the TCGA dataset. The x-axis represents macro-average F_1-score, and the y-axis denotes accuracy. **b,** Radar chart showing the model ranks across the set of four metrics. A larger enclosed area indicates better predictive performance. **c,** Confusion matrix for the classification of 23 cancer types. Columns denote predicted labels, and rows represent actual labels. The percentage shown represents the proportion of predictions made for the corresponding cancer type, with each row summing to 1. The diagonal represents correct predictions for each cancer type, with the percentage indicating the recall. Bar plots show precision (horizontal axis), recall (vertical axis, leftmost) and number of samples (vertical axis, rightmost) per cancer type.

4.3.4 DeePathNet classifies breast cancer subtypes

Gene mutation, CNV and gene expression features were used to train DeePathNet models for the classification of five breast cancer subtypes (Luminal A, Luminal B, HER2+, Basal, Normal-like) according to the Prediction Analysis of Microarray 50 (PAM50) (Parker et al. 2009). A total of 974 breast cancer samples from the TCGA dataset were used for training, and a breast cancer cohort of 122 samples from Clinical Proteomic Tumor Analysis Consortium (CPTAC) was included as an independent dataset to evaluate the generalisation error.

Cross-validation for all six methods was first performed on the TCGA dataset, reporting the mean and 95% CI of the evaluation metrics as an estimate of the generalisation error. DeePathNet provided a substantial improvement over the other methods in terms of accuracy and macro average F_1-score (**Fig. 5a**, **Supplementary Table 5**). The performance gain in AUROC was relatively minor but statistically significant (Student's *t*-test *p*-value $< 5 \times 10^{-4}$) (**Supplementary Table 5**). The methods were then ranked according to the same set of four metrics as in cancer type classification. DeePathNet achieved the best performance in all four metrics, with random forest ranked as the second best overall (**Fig. 5b**). Other methods showed inconsistent performance rankings across different metrics, demonstrating the necessity of using multiple evaluation metrics for a comprehensive evaluation.

To evaluate the generalisation error, a DeePathNet model was trained on the TCGA breast cancer cohort, with the final model tested on the independent CPTAC breast cancer cohort. Benchmarked against random forest, DeePathNet yielded a much lower generalisation error on the independent test set (**Fig. 5c**, **Supplementary Table 6)**. Next, the generalisation error of DeePathNet was assessed for each subtype by a confusion matrix. DeePathNet achieved the highest precision and recall in classifying

the Basal subtype (96.6%, **Fig. 5d**), with most tumours in this subtype being high-grade with a poor prognosis (Dai et al. 2015). The most difficult subtype to classify was Normal-like, where three out of the five Normal-like samples were incorrectly classified as Luminal A (**Fig. 5d**). Luminal A and Normal-like subtypes are traditionally difficult to distinguish as they share the same immunohistochemistry markers (Dai et al. 2015). The Normal-like subtype is less frequently used in clinics (Raj-Kumar et al. 2019). Further analyses by AUROC (**Supplementary Fig. 5a**) and AUPRC (**Supplementary Fig. 5b**) demonstrated DeePathNet's high predictive performance for each subtype.

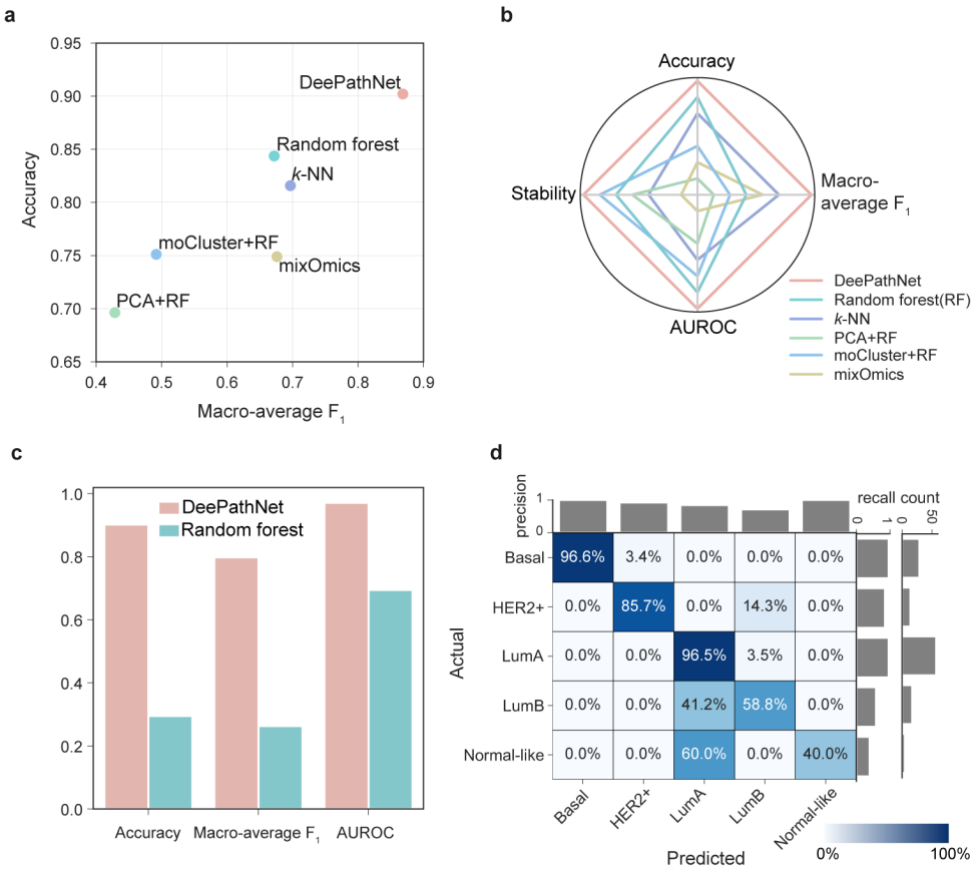


Fig. 5 | Performance evaluation of breast cancer subtype classification. a, Model evaluation by cross-validation. The x-axis represents macro-average F_1-score, and the y-axis represents accuracy. **b,** Radar chart showing the model ranks across the set of four metrics. A larger enclosed area represents better classification performance. **c,** Performance metrics showing generalisation errors for DeePathNet and random forest when using CPTAC data as the independent test set. **d,** Confusion matrix showing generalisation errors when using CPTAC data as the independent test set. Statistics are annotated in the same way as described in **Fig. 4c**.

4.3.5 DeePathNet provides model explanation

The DeePathNet model is explainable at both omic and pathway levels by using feature importance derived from SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) and Layer-wise Relevance Propagation (LRP) (Bach et al. 2015). SHAP attributes the prediction to all features and assigns each feature an importance value, while LRP assumes that the classifier can be decomposed into several layers of computation, with these layers being parts of the feature extraction. Thus, both SHAP and LRP are post-hoc model explanation approaches that establish relationships between feature values and the predictions after DeePathNet is trained. Breast cancer subtype classification was used to demonstrate model explanation.

To explain the model at the omic level, SHAP was used to calculate feature importance. Specifically, feature importance was computed and visualised for the top five genes as stack bar plots comprising each omic data type for each breast subtype (**Fig. 6a**). DeePathNet was able to identify known biomarker genes as top features, such as ESR1, ERBB2 and KRT17, whose gene expression is routinely used to determine the PAM50 subtypes in the clinic(Parker et al. 2009) (**Fig. 6a**). Most genes had their high feature importance attributable to transcriptomic data (**Fig. 6a**), consistent with the fact that PAM50 classifications are RNA-based subtypes(Parker et al. 2009).

To explain the models at the pathway level, LRP was used to calculate feature importance. Since the cancer pathways are represented as an encoded vector that summarises multi-omic information, feature importance of a cancer pathway is computed for all omic data types jointly. For each cancer subtype, the top five pathways with the highest feature importance values were ranked (**Fig. 6b**). DeePathNet identified the FOXM1 transcription factor network as the most important

pathway for predicting all PAM50 subtypes (**Fig. 6b**). FOXM1 shows distinct patterns of expression in different breast cancer subtypes and is seen as a promising candidate target in breast cancer treatment (X.-F. Lu et al. 2018). FOXM1 is also an adverse prognostic factor of survival in Luminal A and B subtypes (J.-J. Lee et al. 2016). The ARF6 pathway was shown to be overexpressed in triple negative breast cancer and to be associated with breast cancer invasion and metastasis (Marchesin et al. 2015). Similarly, Notch Signalling pathways are involved in cell proliferation, apoptosis, hypoxia and epithelial to mesenchymal transition and were found to be over-expressed in HER2+ positive and triple-negative breast cancer (Acar et al. 2016).

Taken together, these findings suggest that DeePathNet provides reliable model explanation with a strong biological basis by providing feature importance at both the omic and pathway level.

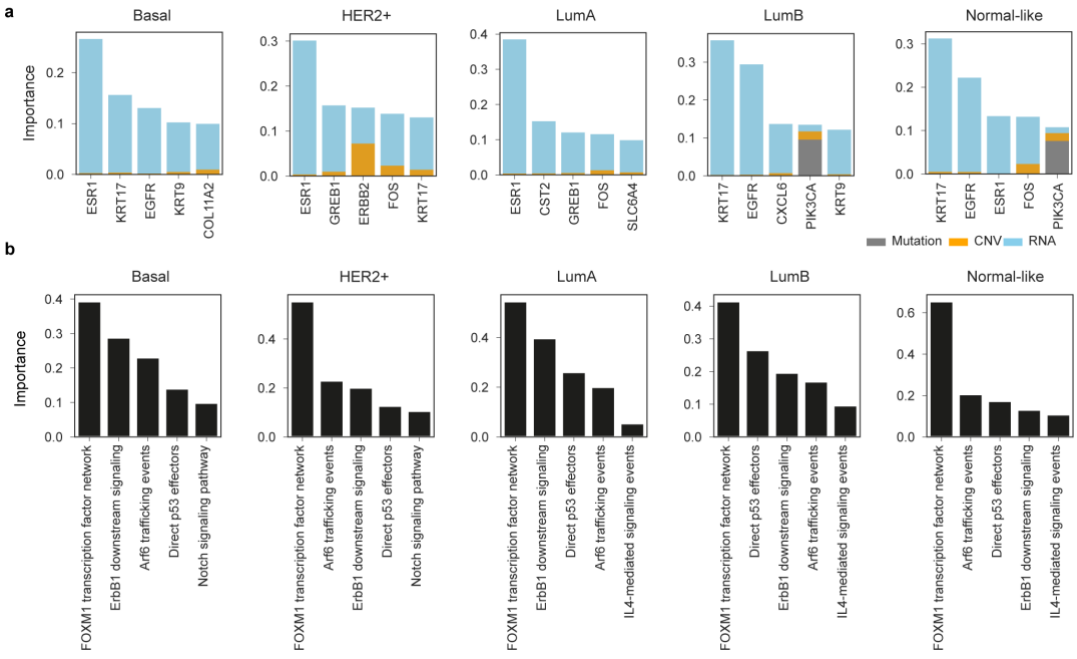


Fig. 6 | DeePathNet model explanation by omic level and pathway level feature importance. a, Stacked bar plots showing the omic level feature importance of the top five genes for each omic data type (indicated by grey, yellow and blue colour). **b,** Bar plots showing the DeePathNet pathway level feature importance of the top five pathways.

4.4 DISCUSSION

DeePathNet is a Transformer-based deep learning model that overcomes the limitation of existing machine learning approaches that do not consider known cancer biology. DeePathNet integrates multi-omic data with cancer pathway knowledge to accurately predict drug responses and classify cancer types and subtypes. The self-attention mechanism of the Transformer module dynamically models the interdependency between pathways, thus capturing regulatory effects across different biological processes and the effects of dysregulation.

The predictive performance of DeePathNet was evaluated by one regression and two classification tasks. The evaluation was conducted on a larger scale than previous similar studies (Meng et al. 2016; A. Singh et al. 2019), using multiple big datasets and a range of metrics with both cross-validation and independent testing. Incorporating cancer pathway information, DeePathNet outperformed other machine learning methods that only use omic data as input features. A low generalisation error when validating DeePathNet models on independent datasets suggests that DeePathNet work well even when different experimental protocols were implemented between these independent datasets. DeePathNet provides model explanations at the pathway level, which has not yet been accomplished by other multi-omic integration tools for the prediction of drug response and classification of cancer type and subtype. DeePathNet was able to highlight known biomarkers when predicting breast cancer subtypes, including ESR1, ERBB2 and the FOXM1 network pathways. This suggests that other top-ranked genes and pathways may provide novel insights into cancer biology and drug discovery.

Despite these comprehensive evaluations, this study only concentrated on a limited number of omic data types because large-scale studies of some omic data

types are still in their infancy (Satpathy et al. 2021; Nusinow et al. 2020; Yiqun Zhang et al. 2022). As large proteomic and metabolomic datasets become increasingly available, the predictive power of DeePathNet can improve further, because deep learning is likely to obtain performance boost with increased amount of data (Esteva et al. 2019).

In conclusion, DeePathNet combines multi-omics, deep learning and existing biological knowledge to predict cancer phenotypes accurately with a model explanation. The application of DeePathNet may lead to more accurate diagnosis and prognosis, and will facilitate researchers to understand unknown cancer mechanisms and prioritise putative drug targets.

4.5 METHODS

Multi-omic and drug response data collection. For drug response prediction, multi-omic data were retrieved from 941 CLP (Iorio et al. 2016) and 696 CCLE cell lines (Ghandi et al. 2019). In total, 19,099 gene mutation, 19,116 CNV and 15,320 gene expression features are in the CLP, and 18,103 gene mutation, 27,562 CNV and 19,177 gene expression features are in the CCLE.

For drug response prediction analysis with proteomic data, the ProCan-DepMapSanger dataset(Gonçalves et al. 2022) was added to the CLP (CLP + ProCan-DepMapSanger = CLP^+) and the CCLE's proteomic dataset(Nusinow et al. 2020) was also used (CCLE + CCLE proteomic data = $CCLE^+$). The ProCan-DepMapSanger and CCLE proteomic datasets contain 8,498 and 12,755 protein features, respectively. The combined datasets have 910 and 292 cell lines for CLP^+ and $CCLE^+$, respectively. No additional processing was performed on both the omics and drug response datasets (**Supplementary Table 1)**.

For cancer type and subtype classification, multi-omic data from TCGA cohorts were retrieved using TCGA-assembler 2 (L. Wei et al. 2018). In total, 6,356 samples were collected, containing 31,949 features from gene mutation, 23,529 features from CNV and 20,435 features from gene expression. In addition, multi-omic data from 122 breast cancer samples were retrieved from a CPTAC breast cancer cohort (Krug

et al. 2020), containing 11,877 features from gene mutation, 23,692 features from CNV and 23,121 features from gene expression. For breast cancer subtype classification, the PAM50 classification (Luminal A, Luminal B, HER2+, Basal and Normal-like) was retrieved from the TCGA and CPTAC datasets (**Supplementary Table 1)**.

Overview of DeePathNet. DeePathNet has a pathway encoder (Step 1), a Transformer encoder (Step 2) and a MLP (Step 3).

In Step 1, DeePathNet encodes multi-omic information into cancer pathways, defined by the 241 cancer pathways in LCpathways (Kuenzi and Ideker 2020). Let $g_{mutation} \in \{0,1\}$ represent the mutation, $g_{CNV} \in \mathbb{R}$ the CNV, $g_{RNA} \in \mathbb{R}$ the gene expression, and $g_{prot} \in \mathbb{R}$ the protein intensity of a gene g. Then the vector that contains omic features for a pathway that contains n genes with four omic data types, is defined as:

$$\boldsymbol{a}_{omics} = \left[g^1_{mutation}, g^1_{CNV}, g^1_{RNA}, g^1_{prot}, \cdots, g^n_{mutation}, g^n_{CNV}, g^n_{RNA}, g^n_{prot}\right]$$

Next, the vector $\boldsymbol{a}_{omics}$ is encoded into the pathway vector $\boldsymbol{a}_{encoded}$ via a MLP. Here, the notation is converted into the matrix form to include the number of samples. Thus, for N samples, the total features from the four omic data types for a pathway can be represented as a matrix A_{omics} of dimension $N \times 4n$. DeePathNet then uses a fully connected layer to encode these omic features into an encoded pathway matrix $A_{encoded}$, calculated as:

$$A_{encoded} = A_{omics} W^T + B$$

where W and B represent the learnable weights matrix and bias term in the fully connected layer. The dimension of the weight matrix W is set as $512 \times 4n$. The dimension of both bias B and $A_{encoded}$ is $N \times 512$. In total, 241 cancer pathways were used and 241 matrices $A^1_{encoded}, A^2_{encoded}, \cdots, A^{241}_{encoded}$ are combined as a tensor $\boldsymbol{A}_{encoded}$ with a dimensionality of $N \times 512 \times 241$. $\boldsymbol{A}_{encoded}$ is used as the input into the Transformer encoder (**Supplementary Fig. 1**).

In Step 2, DeePathNet uses a Transformer encoder to learn the interdependence between regulatory pathways in cancer. In contrast to the general attention mechanism that models the interdependence between the input and target, self-

attention is used by the Transformer module to model interdependence within the input (T. Lin et al. 2021) (i.e., features from the multi-omic data). The Transformer encoder starts with a dropout layer with a probability of 0.5 on the 241 cancer pathways, ensuring that on average half of the pathways are dropped out during training to prevent potential overfitting. The set of selected pathways is sampled independently for each training batch, allowing different pathways to be used. The Transformer block was configured the same way as the original version (Vaswani et al. 2017), denoted as $Transformer$ below. Since the Transformer encoder contains recurrent layers, we use a superscript with parenthesis to represent the $\boldsymbol{A}_{encoded}$ at different layers, where $\boldsymbol{A}_{encoded}^{(0)}$ represents the data before entering the first layer. After the first layer of the Transformer block, $\boldsymbol{A}_{encoded}^{(0)}$ becomes $\boldsymbol{A}_{encoded}^{(1)}$ as follows:

$$\boldsymbol{A}_{encoded}^{(1)} = Transformer\left(\boldsymbol{A}_{encoded}^{(0)}\right)$$

DeePathNet contains two layers of Transformer block, therefore:

$$\boldsymbol{A}_{encoded}^{(2)} = Transformer\left(\boldsymbol{A}_{encoded}^{(1)}\right)$$

Finally, in Step 3, DeePathNet uses a MLP to map $\boldsymbol{A}_{encoded}^{(2)}$ to the final prediction. The output dimension of a MLP depends on the prediction task. For drug response prediction, the number of output dimensions is equal to the number of drugs, and for cancer type and subtype classification, the number of output dimensions is equal to the number of cancer types and subtypes.

Model training. All methods were trained with default hyperparameters for both regression and classification tasks. The default hyperparameters of DeePathNet and optimiser used can be found in the GitHub repository. Default hyperparameters were used for random forest, elastic net, PCA (top 200 PCs) and *k*-NN (*k* = 5) and details can be found in the official API of scikit-learn (v1.0.2). Default hyperparameters were also set for mixOmics and moCluster, and details can be found in their original publications. To train DeePathNet for regression, mean squared error (MSE) loss was computed between the predicted and actual IC_{50}. For classification, we computed the cross-entropy (CE) loss to train DeePathNet.

Evaluation metrics. For regression, R^2, MAE and Pearson's *r* were used to evaluate the performance and they are defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$MAE = \frac{\sum_{i=1}^{n} |(y_i - \hat{y}_i)|}{n}$$

$$Pearson's\ r = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}}$$

For a given drug, y_i represents the actual IC_{50} of cell line i, $\hat{y}_i$ represents the predicted IC_{50} value of cell line i, $\bar{y}$ represents the mean value of all actual IC_{50} values, $\bar{\hat{y}}$ represents the mean value of all predicted IC_{50} values, and n represents the total number of cell lines. For classification, multiple metrics were used to evaluate the predictive performance of DeePathNet and other models, including accuracy, macro-average F_1-score, precision, recall, AUROC, AUPRC and stability. Let TP, TN, FP, FN represent true positive, true negative, false positive and false negative prediction. Accuracy is defined as $\frac{TP+TN}{TP+TN+FP+FN}$. Precision is defined as $\frac{TP}{TP+FP}$. Recall is defined as $\frac{TP}{TP+FN}$,. Then the F_1-score is calculated as the harmonic mean of the precision and recall and defined as $\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$. The macro-average F_1-score is calculated by computing the arithmetic mean of F_1-scores from all the cancer types or subtypes. The ROC curve is created by plotting the recall and false positive rate ($\frac{FP}{FP+TN}$) at various thresholds. AUROC is calculated as the area under the ROC curve. The precision and recall (PR) curve is created by plotting the precision and recall at various thresholds, and the AUPRC is calculated as the area under the PR curve. The stability is measured by the standard deviation.

Chapter 5: Discussions and Conclusions

Some text included in this chapter is adapted from the following publications:

Poulos, R. C., Cai, Z., Robinson, P. J., Reddel, R. R., & Zhong, Q. (2022). Opportunities for pharmacoproteomics in biomarker discovery. *Proteomics*, 2200031

5.1 SUMMARY OF THE RESEARCH PRESENTED IN THIS THESIS

Multi-omic data analysis has transformed cancer diagnosis, prognosis, and therapeutical development in a variety of ways. However, large-scale multi-omic analysis, especially when including proteomic data, is still in its infancy. Both multi-omic data and related machine learning tools are crucially needed to further improve our understanding of the mechanism of cancer. This thesis presents analyses of both cancer cell lines and human tumour tissue samples using multi-omic data combined with novel machine learning approaches. The overarching aim of this thesis has been to develop methods to integrate proteomic data with other omic data types, and to implement these methods to achieve better predictions of various cancer phenotypes.

Chapter 2 presents both a comprehensive review and a benchmarking analysis of current machine learning tools for the integration of multi-omic data. The review and analysis are based on several key tasks in cancer research, including cancer type prediction and drug response prediction. Multi-omic data analysis is key to understanding the nature and result of complex dysregulation events that are associated with different cancer phenotypes. Despite exponential growth in the number of multi-omic experiments being performed in the research community, and therefore in the amount of data available to researchers, limited efforts have been made to develop

machine learning tools that automatically integrate these multi-omic datasets. This chapter further incorporates a benchmarking analysis, comparing several recently published machine learning-based multi-omic data integration tools. The results of this benchmarking study allow researchers to select the most appropriate tools for their applications.

Chapter 3 introduces a large pan-cancer proteomic map and the use of a novel neural network to gain insight into cancer phenotype beyond what could be obtained from existing molecular datasets. Specifically, this map quantifies 8,498 proteins across 949 human cancer cell lines, representing 28 tissues and more than 40 histologically diverse cancer types and a wide range of genotypes. To analyse these data, a deep learning-based pipeline was designed to find biomarkers of drug response and gene essentiality. Further, a random downsampling analysis was conducted that reveals highly connected and co-regulated protein networks.

Chapter 4 presents DeePathNet, a machine learning algorithm that combines multi-omic, deep learning and existing biological knowledge to predict cancer phenotypes accurately with a model explanation. Through the analysis of biological datasets, the application of DeePathNet is shown to predict cancer type, subtype and drug response accurately with meaningful model explanations. This approach may lead to more accurate diagnosis and prognosis, and will facilitate researchers to understand unknown cancer mechanisms and prioritise putative drug targets.

5.2 NEW RESEARCH QUESTIONS ARISING FROM THE FINDINGS PRESENTED IN THIS THESIS

5.2.1 Arising from Chapter 2: How do sample size and number of omic data types affect predictive performance of multi-omic integration tools?

In the benchmarking analysis, we only estimated the run-time of the multi-omic integration but didn't evaluate the predictive performance when varying sample sizes. Machine learning methods benefit greatly from a large amount of data to avoid overfitting (Zhou et al. 2017), and complex models such as deep learning require more data than basic machine learning methods such as a logistic regression. However, the sample size is usually small when training machine learning models in multi-omic data analysis. To evaluate which multi-omic integration methods can work well with small cohorts, our future work will involve assessing the performance of a range of machine learning methods with varying numbers of samples, on the tasks of drug response prediction and cancer type classification.

A second question arising from this work is whether adding more omic data types of the same samples will lead to improved prediction. In multi-omic data analysis, the total number of features is significantly larger than the number of samples, known as the curse of dimensionality (Mirza et al. 2019). It has been shown that more omic data types may not necessarily improve predictive performance (T. Wang et al. 2021). This result is expected because by introducing more features, it further worsens the curse of dimensionality. What remains to be investigated is the trade-off between information gained from the extra omic data types and the increased feature space.

5.2.2 Arising from Chapter 3: Can DeeProM trained with cancer cell lines data be validated in tumour tissues data?

In **Chapter 3**, we developed DeeProM to predict drug response using the proteomic data (ProCan-SangerDepMap) of 949 human cancer cell lines. A follow-up project could be investigating the applicability of DeeProM in other model systems, such as organoids and patient-derived xenografts (PDXs) rather than human cancer tissues. This is because we can test multiple drugs simultaneously on the same cell line, but it is intractable for human tissue samples.

Compared to cell line models, pharmacogenomic biomarker studies have been performed at a smaller scale in organoids (van de Wetering et al. 2015; Pauli et al. 2017) and PDXs (Woo et al. 2019; Conte et al. 2019). Organoids, which contain more than one cell type and are capable of mimicking tissue organisation, and tumoroids, which are a three-dimensional cell culture, offer an effective way to additionally model organ structure and function and are now possible at large scale (Durens et al. 2020; Daniszewski et al. 2018). PDXs are derived from human tissue excised from a patient's tumour and transplanted into an immunodeficient mouse. All of these pre-clinical models bear many resemblances to their tissues of origin, but they are unable to sufficiently represent all elements of a human tissue sample (Salvadores, Fuster-Tormo, and Supek 2020; Goodspeed et al. 2016; Mirabelli, Coppola, and Salvatore 2019; J. Kim, Koo, and Knoblich 2020; Jung, Seol, and Chang 2018; Trastulla et al. 2022). For example, none of the model systems incorporate a fully functioning immune system. Cell lines lack the complexity of other pre-clinical models including three-dimensional structure. In general, organoids and PDXs more closely resemble their human tissue counterparts than cell lines, but they are more costly to establish and maintain. In the case of PDX samples, an additional challenge is to differentiate human from mouse proteins, which share considerable sequence homology.

Therefore, to test the applicability of DeeProM beyond cell line models, future works can start with organoids and PDXs samples, but ultimately human tumour tissues should be investigated due to their unique characteristics that pre-clinical models do not provide.

5.2.3 Arising from Chapter 4: Does the choice of human knowledge database affect the predictive performance of multi-omic integrative models?

In **Chapter 4**, we developed DeePathNet as a novel tool to incorporate human knowledge into multi-omic data integration. One highlight of DeePathNet is that the model utilises cancer specific pathways as the input of human knowledge. However, whether DeePathNet can utilise databases in different structures such as GeneOntology (GO) (Ashburner et al. 2000; Gene Ontology Consortium 2021) and STRING (Szklarczyk et al. 2021) has yet to be investigated.

The architecture of DeePathNet would need to be modified to utilise GO and STRING, and each database would confer different challenges. To use GO with DeePathNet, one open question would be how to handle the hierarchies between different GO entities. GO has been utilised in other deep learning methods with multiple layers of neurons with predefined connections between neurons (Elmarakeby et al. 2021). However, DeePathNet does not support using hierarchies in the pathway dataset due to technical constraints from the transformer module. To overcome this limitation, the pathway encoder and transformer module would need to be redesigned to reflect the hierarchy of information in the GO database.

To use DeePathNet with STRING, modifications to DeePathNet are also required because STRING contains graphs that are connected at a very large scale and does not have any sub-groups. DeePathNet requires omic-level features to be grouped into middle level sub-groups such as cancer pathways. Since the STRING database is

essentially a computational graph, the most appropriate method is to utilise graph neural networks (Wu et al. 2021) to perform data integration. Graph transformer network (GTN) (Yun et al. 2022) has been proposed recently and is an appropriate model to utilise in this scenario. DeePathNet could be modified with a variation of using GTN to support the use of human knowledge in the form of graphs.

5.3 NEW TECHNOLOGIES AND FUTURE DIRECTIONS

5.3.1 Single-cell omic approaches create additional opportunities for machine learning

Although omic studies in cancer research primarily focus on data obtained from bulk samples, recent advancements in single-cell technology has enabled researchers to analyse tumour samples at higher resolutions (Peng et al. 2020). Since each cell is a sample by itself, large-scale single-cell datasets will provide more data, allowing machine learning to train complex models that are currently impossible with bulk data due to the limited numbers of samples. More importantly, intra-tumour heterogeneity, meaning a tumour may contain different sub-populations that contain different molecular profiles and phenotypes, plays a key role in developing cancer therapies (Marusyk, Almendro, and Polyak 2012). Single-cell omic measurements enables us to study the relationship and interactions between different cells within the tumour (Tellez-Gabriel et al. 2016), Machine learning approaches are expected to further facilitate the single-cell data analysis.

5.3.2 Few-shot deep learning in biomedical research

New technologies are currently emerging in the field of deep learning that can enable models to be well trained with only small datasets. This is called few-shot

learning (FSL) (Y. Wang et al. 2021). FSL can quickly generalise to new tasks comprising only a few samples with supervised information by utilising past knowledge. In biomedical research, FSL has been applied to text analysing tasks (Hofer et al. 2018) and histopathological images (Medela et al. 2019). It is anticipated that FSL may also be applied to omic dataset successfully. In cancer, this may enable the use of deep learning to study cohorts where it can be difficult to obtain large sample sizes, such as paediatric cancers or rare adult cancer types. Apart from FSL, data synthesis algorithms using generative models such as variational autoencoder (Ruoqi Wei and Mahmood 2021) may also play a key role in overcoming the shortage of data, specifically when analysing rare cancer types.

5.3.3 More accessible computational power for small labs to use deep learning

Many deep learning models, especially complex ones, require the usage of at least one high-end GPU to accelerate computation. However, powerful GPUs are not equipped in standard office desktop or laptop computers. Therefore, small labs who do not have access to such powerful computational resources may not be able to benefit from the latest tools. However, products such as Google Colab are now being developed, allowing more deep learning platforms with free GPU usage to become available for individual researchers. This will enable a wider adoption of deep learning methods in cancer research, with the potential to facilitate the discoveries that can lead to new cancer treatments.

5.3.4 Explainable machine learning to aid biological discovery

As the availability of multi-omic data expands, more complex machine learning models, the application of increasingly complex machine learning models, including advanced deep learning architectures, has the potential to revolutionise cancer diagnosis and prognostic predictions. However, the interpretability of such data

presents a substantial challenge. Therefore, the role of explainable machine learning becomes instrumental (Mathews 2019; Lötsch, Kringel, and Ultsch 2021). We envisage a future where machine learning systems not only predict outcomes with high precision but also provide interpretable reasoning behind these predictions, thereby promoting trust in their utility. The development of such machine learning models would enable researchers to explain intricate biological pathways and mechanisms underlying cancer pathogenesis, progression, and response to therapy.

Bibliography

Acar, Ahmet, Bruno M. Simões, Robert B. Clarke, and Keith Brennan. 2016. "A Role for Notch Signalling in Breast Cancer and Endocrine Resistance." *Stem Cells International* 2016 (January): 2498764.

Aebersold, Ruedi, and Matthias Mann. 2003. "Mass Spectrometry-Based Proteomics." *Nature* 422 (6928): 198–207.

Aizerman, M. A. 1964. "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning." *Automation and Remote Control* 25: 821–37.

Akiyama, H., M. Tsurumaru, H. Udagawa, and Y. Kajiyama. 1997. "Esophageal Cancer." *Current Problems in Surgery* 34 (10): 765–834.

Alcala, N., N. Leblay, A. A. G. Gabriel, L. Mangiante, D. Hervas, T. Giffon, A. S. Sertier, et al. 2019. "Integrative and Comparative Genomic Analyses Identify Clinically Relevant Pulmonary Carcinoid Groups and Unveil the Supra-Carcinoids." *Nature Communications* 10 (1): 3407.

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. "The Repertoire of Mutational Signatures in Human Cancer." *Nature* 578 (7793): 94–101.

Andersson, Robin, and Albin Sandelin. 2020. "Determinants of Enhancer and Promoter Activities of Regulatory Elements." *Nature Reviews. Genetics* 21 (2): 71–87.

Argelaguet, Ricard, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, and Oliver Stegle. 2020. "MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data." *Genome Biology* 21 (1): 111.

Argelaguet, Ricard, Stephen J. Clark, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, et al. 2019. "Multi-Omics Profiling of Mouse Gastrulation at Single-Cell Resolution." *Nature* 576 (7787): 487–91.

Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2018. "Multi-Omics Factor Analysis-a Framework for Unsupervised Integration of Multi-Omics Data Sets." *Molecular Systems Biology* 14 (6): e8124.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nature Genetics* 25 (1): 25–29.

Aulchenko, Yurii S., Stephan Ripke, Aaron Isaacs, and Cornelia M. van Duijn. 2007. "GenABEL: An R Library for Genome-Wide Association Analysis." *Bioinformatics* 23 (10): 1294–96.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." *PloS One* 10 (7): e0130140.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483 (7391): 603–7.

Bavafaye Haghighi, Elham, Michael Knudsen, Britt Elmedal Laursen, and Søren Besenbacher. 2019. “Hierarchical Classification of Cancers of Unknown Primary Using Multi-Omics Data.” *Cancer Informatics* 18 (August): 1176935119872163.

Behan, Fiona M., Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M. Beaver, Giorgia Migliardi, Rita Santos, et al. 2019. “Prioritization of Cancer Therapeutic Targets Using CRISPR-Cas9 Screens.” *Nature* 568 (7753): 511–16.

Bellman, R. 1966. “Dynamic Programming.” *Science* 153 (3731): 34–37.

Berkum, Nynke L. van, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A. Mirny, Job Dekker, and Eric S. Lander. 2010. “Hi-C: A Method to Study the Three-Dimensional Architecture of Genomes.” *Journal of Visualized Experiments: JoVE* 39 (39): e1869.

Boehm, Jesse S., Mathew J. Garnett, David J. Adams, Hayley E. Francies, Todd R. Golub, William C. Hahn, Francesco Iorio, James M. McFarland, Leopold Parts, and Francisca Vazquez. 2021. “Cancer Research Needs a Better Map.” *Nature* 589 (7843): 514–16.

Bohan, Sandy S., Jason K. Sicklick, Shumei Kato, Ryosuke Okamura, Vincent A. Miller, Brian Leyland-Jones, Scott M. Lippman, and Razelle Kurzrock. 2020. “Attrition of Patients on a Precision Oncology Trial: Analysis of the I-PREDICT Experience.” *The Oncologist* 25 (11): e1803–6.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” *Cell* 169 (7): 1177–86.

Brabletz, Thomas, Raghu Kalluri, M. Angela Nieto, and Robert A. Weinberg. 2018. “EMT in Cancer.” *Nature Reviews. Cancer* 18 (2): 128–34.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Broad. 2020. “DepMap.” 2020.

Brouwer, Ineke, and Tineke L. Lenstra. 2019. “Visualizing Transcription: Key to Understanding Gene Expression Dynamics.” *Current Opinion in Chemical Biology* 51 (August): 122–29.

Cai, Zhaoxiang, Rebecca C. Poulos, Jia Liu, and Qing Zhong. 2022. “Machine Learning for Multi-Omics Data Integration in Cancer.” *IScience* 25 (2): 103798.

Cancer Genome Atlas Network. 2012. “Comprehensive Molecular Characterization of Human Colon and Rectal Cancer.” *Nature* 487 (7407): 330–37.

Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Characterization of Gastric Adenocarcinoma.” *Nature* 513 (7517): 202–9.

Cancer Genome Atlas Research Network., and Cancer Genome Atlas Research Network. 2017. “Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma.” *Cancer Cell* 32 (2): 185-203.e13.

Chatr-Aryamontri, Andrew, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, et al. 2015. “The BioGRID Interaction Database: 2015 Update.” *Nucleic Acids Research* 43 (Database issue): D470-8.

Cichonska, Anna, Tapio Pahikkala, Sandor Szedmak, Heli Julkunen, Antti Airola, Markus Heinonen, Tero Aittokallio, and Juho Rousu. 2018. “Learning with Multiple Pairwise Kernels for Drug Bioactivity Prediction.” *Bioinformatics* 34 (13): i509–18.

Clark, David J., Saravana M. Dhanasekaran, Francesca Petralia, Jianbo Pan, Xiaoyu Song, Yingwei Hu, Felipe da Veiga Leprevost, et al. 2019. “Integrated

Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma." *Cell* 179 (4): 964-983.e31.
Cohen, Philip, Darren Cross, and Pasi A. Jänne. 2021. "Kinase Drug Discovery 20 Years after Imatinib: Progress and Future Directions." *Nature Reviews. Drug Discovery* 20 (7): 551–69.
Collins, Ben C., Christie L. Hunter, Yansheng Liu, Birgit Schilling, George Rosenberger, Samuel L. Bader, Daniel W. Chan, et al. 2017. "Multi-Laboratory Assessment of Reproducibility, Qualitative and Quantitative Performance of SWATH-Mass Spectrometry." *Nature Communications* 8 (1): 291.
Conte, Nathalie, Jeremy C. Mason, Csaba Halmagyi, Steven Neuhauser, Abayomi Mosaku, Galabina Yordanova, Aikaterini Chatzipli, et al. 2019. "PDX Finder: A Portal for Patient-Derived Tumor Xenograft Model Discovery." *Nucleic Acids Research* 47 (D1): D1073–79.
Corsello, Steven M., Rohith T. Nagari, Ryan D. Spangler, Jordan Rossen, Mustafa Kocak, Jordan G. Bryan, Ranad Humeidi, et al. 2020. "Discovering the Anti-Cancer Potential of Non-Oncology Drugs by Systematic Viability Profiling." *Nature Cancer* 1 (2): 235–48.
Coscia, F., K. M. Watters, M. Curtis, M. A. Eckert, C. Y. Chiang, S. Tyanova, A. Montag, R. R. Lastra, E. Lengyel, and M. Mann. 2016. "Integrative Proteomic Profiling of Ovarian Cancer Cell Lines Reveals Precursor Cell Associated Proteins and Functional Status." *Nature Communications* 7 (1): 12645.
Cox, Jürgen, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. "Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ." *Molecular & Cellular Proteomics: MCP* 13 (9): 2513–26.
Crick, F. 1970. "Central Dogma of Molecular Biology." *Nature* 227 (5258): 561–63.
Cunningham, Julie M., Eric R. Christensen, David J. Tester, Cheong-Yong Kim, Patrick C. Roche, Lawrence J. Burgart, and Stephen N. Thibodeau. 1998. "Hypermethylation of the HMLH1 Promoter in Colon Cancer with Microsatellite Instability." *Cancer Research* 58 (15): 3455–60.
Dai, Xiaofeng, Ting Li, Zhonghu Bai, Yankun Yang, Xiuxia Liu, Jinling Zhan, and Bozhi Shi. 2015. "Breast Cancer Intrinsic Subtype Classification, Clinical Use and Future Trends." *American Journal of Cancer Research* 5 (10): 2929–43.
Dandage, Rohan, and Christian R. Landry. 2019. "Paralog Dependency Indirectly Affects the Robustness of Human Cells." *Molecular Systems Biology* 15 (9): e8871.
Daniszewski, Maciej, Duncan E. Crombie, Rachael Henderson, Helena H. Liang, Raymond C. B. Wong, Alex W. Hewitt, and Alice Pébay. 2018. "Automated Cell Culture Systems and Their Applications to Human Pluripotent Stem Cell Studies." *SLAS Technology* 23 (4): 315–25.
Demichev, Vadim, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. 2020. "DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput." *Nature Methods* 17 (1): 41–44.
Ding, Zijian, Songpeng Zu, and Jin Gu. 2016. "Evaluating the Molecule-Based Prediction of Clinical Drug Responses in Cancer." *Bioinformatics* 32 (19): 2891–95.

Durens, Madel, Jonathan Nestor, Madeline Williams, Kevin Herold, Robert F. Niescier, Jason W. Lunden, Andre W. Phillips, Yu-Chih Lin, Derek M. Dykxhoorn, and Michael W. Nestor. 2020. “High-Throughput Screening of Human Induced Pluripotent Stem Cell-Derived Brain Organoids.” *Journal of Neuroscience Methods* 335 (108627): 108627.

Edwards, Nathan J., Mauricio Oberti, Ratna R. Thangudu, Shuang Cai, Peter B. McGarvey, Shine Jacob, Subha Madhavan, and Karen A. Ketchum. 2015. “The CPTAC Data Portal: A Resource for Cancer Proteomics Research.” *Journal of Proteome Research* 14 (6): 2707–13.

Elmarakeby, Haitham A., Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, et al. 2021. “Biologically Informed Deep Neural Network for Prostate Cancer Discovery.” *Nature* 598 (7880): 348–52.

Esteva, Andre, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. “A Guide to Deep Learning in Healthcare.” *Nature Medicine* 25 (1): 24–29.

Everett, B. 2013. *An Introduction to Latent Variable Models*. Springer Science & Business Media.

Fischer, Martin, Marianne Quaas, Lydia Steiner, and Kurt Engeland. 2016. “The P53-P21-DREAM-CDE/CHR Pathway Regulates G2/M Cell Cycle Genes.” *Nucleic Acids Research* 44 (1): 164–74.

Fix, Evelyn, and Joseph Lawson Hodges. 1989. “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties.” *Revue Internationale de Statistique [International Statistical Review]* 57 (3): 238.

Freedman, David A. 2009. *Statistical Models: Theory and Practice*. 2nd ed. Cambridge University Press.

Frejno, Martin, Chen Meng, Benjamin Ruprecht, Thomas Oellerich, Sebastian Scheich, Karin Kleigrewe, Enken Drecoll, et al. 2020. “Proteome Activity Landscapes of Tumor Cell Lines Determine Drug Responses.” *Nature Communications* 11 (1): 3639.

Frejno, Martin, Riccardo Zenezini Chiozzi, Mathias Wilhelm, Heiner Koch, Runsheng Zheng, Susan Klaeger, Benjamin Ruprecht, et al. 2017. “Pharmacoproteomic Characterisation of Human Colon and Rectal Cancer.” *Molecular Systems Biology* 13 (11): 951.

Gao, Hui, Joshua M. Korn, Stéphane Ferretti, John E. Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, et al. 2015. “High-Throughput Screening Using Patient-Derived Tumor Xenografts to Predict Clinical Trial Drug Response.” *Nature Medicine* 21 (11): 1318–25.

Garcia-Alonso, Luz, Francesco Iorio, Angela Matchan, Nuno Fonseca, Patricia Jaaks, Gareth Peat, Miguel Pignatelli, et al. 2018. “Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer.” *Cancer Research* 78 (3): 769–80.

Garnett, Mathew J., Elena J. Edelman, Sonja J. Heidorn, Chris D. Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, et al. 2012. “Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells.” *Nature* 483 (7391): 570–75.

Gene Ontology Consortium. 2021. “The Gene Ontology Resource: Enriching a GOld Mine.” *Nucleic Acids Research* 49 (D1): D325–34.

Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald 3rd, Jordi Barretina, et al. 2019.

"Next-Generation Characterization of the Cancer Cell Line Encyclopedia." *Nature* 569 (7757): 503–8.

Gholami, Amin Moghaddas, Hannes Hahne, Zhixiang Wu, Florian Johann Auer, Chen Meng, Mathias Wilhelm, and Bernhard Kuster. 2013. "Global Proteome Analysis of the NCI-60 Cell Line Panel." *Cell Reports* 4 (3): 609–20.

Gillet, Ludovic C., Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. 2012. "Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis." *Molecular & Cellular Proteomics: MCP* 11 (6): O111.016717.

Gillette, Michael A., Shankha Satpathy, Song Cao, Saravana M. Dhanasekaran, Suhas V. Vasaikar, Karsten Krug, Francesca Petralia, et al. 2020. "Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma." *Cell* 182 (1): 200-225.e35.

Gonçalves, Emanuel, Athanassios Fragoulis, Luz Garcia-Alonso, Thorsten Cramer, Julio Saez-Rodriguez, and Pedro Beltrao. 2017. "Widespread Post-Transcriptional Attenuation of Genomic Copy-Number Variation in Cancer." *Cell Systems* 5 (4): 386-398.e4.

Gonçalves, Emanuel, Rebecca C. Poulos, Zhaoxiang Cai, Syd Barthorpe, Srikanth S. Manda, Natasha Lucas, Alexandra Beck, et al. 2022. "Pan-Cancer Proteomic Map of 949 Human Cell Lines." *Cancer Cell* 40 (8): 835-849.e8.

Gonçalves, Emanuel, Aldo Segura-Cabrera, Clare Pacini, Gabriele Picco, Fiona M. Behan, Patricia Jaaks, Elizabeth A. Coker, et al. 2020. "Drug Mechanism-of-Action Discovery through the Integration of Pharmacological and CRISPR Screens." *Molecular Systems Biology* 16 (7): e9405.

Goodspeed, Andrew, Laura M. Heiser, Joe W. Gray, and James C. Costello. 2016. "Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics." *Molecular Cancer Research: MCR* 14 (1): 3–13.

Gumeni, Sentiljana, Zoi Evangelakou, Vassilis Gorgoulis, and Ioannis Trougakos. 2017. "Proteome Stability as a Key Factor of Genome Integrity." *International Journal of Molecular Sciences* 18 (10): 2036.

Guo, Tiannan, Petri Kouvonen, Ching Chiek Koh, Ludovic C. Gillet, Witold E. Wolski, Hannes L. Röst, George Rosenberger, et al. 2015. "Rapid Mass Spectrometric Conversion of Tissue Biopsy Samples into Permanent Quantitative Digital Proteome Maps." *Nature Medicine* 21 (4): 407–13.

Guo, Tiannan, Li Li, Qing Zhong, Niels J. Rupp, Konstantina Charmpi, Christine E. Wong, Ulrich Wagner, et al. 2018. "Multi-Region Proteome Analysis Quantifies Spatial Heterogeneity of Prostate Tissue Biomarkers." *Life Science Alliance* 1 (2). https://doi.org/10.26508/lsa.201800042.

Guo, Tiannan, Augustin Luna, Vinodh N. Rajapakse, Ching Chiek Koh, Zhicheng Wu, Wei Liu, Yaoting Sun, et al. 2019. "Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines." *IScience* 21 (November): 664–80.

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck 3rd, Shiwei Zheng, Andrew Butler, Maddie J. Lee, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184 (13): 3573-3587.e29.

Haraksingh, Rajini R., and Michael P. Snyder. 2013. "Impacts of Variation in the Human Genome on Gene Regulation." *Journal of Molecular Biology*, Understanding Molecular Effects of Naturally Occurring Genetic Differences, 425 (21): 3970–77.

Hart, Traver, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R. Brown, Graham MacLeod, Monika Mis, et al. 2015. “High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.” *Cell* 163 (6): 1515–26.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–78. IEEE.

Heckerman, David. 1990. “Probabilistic Similarity Networks.” *Networks. An International Journal* 20 (5): 607–36.

Hegde, Priti S., Ian R. White, and Christine Debouck. 2003. “Interplay of Transcriptomics and Proteomics.” *Current Opinion in Biotechnology* 14 (6): 647–51.

Ho, Tin Kam. 2002. “Random Decision Forests.” In *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press. https://doi.org/10.1109/icdar.1995.598994.

Hoadley, Katherine A., Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, et al. 2018. “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.” *Cell* 173 (2): 291-304.e6.

Hofer, Maximilian, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. 2018. “Few-Shot Learning for Named Entity Recognition in Medical Text.” *ArXiv [Cs.CL]*. arXiv. http://arxiv.org/abs/1811.05468.

Hotelling, Harold. 1992. “Relations Between Two Sets of Variates.” In *Breakthroughs in Statistics: Methodology and Distribution*, edited by Samuel Kotz and Norman L. Johnson, 162–90. New York, NY: Springer New York.

Huang, Kuan-Lin, Shunqiang Li, Philipp Mertins, Song Cao, Harsha P. Gunawardena, Kelly V. Ruggles, D. R. Mani, et al. 2017. “Proteogenomic Integration Reveals Therapeutic Targets in Breast Cancer Xenografts.” *Nature Communications* 8 (1): 14864.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. “Pan-Cancer Analysis of Whole Genomes.” *Nature* 578 (7793): 82–93.

International Cancer Genome Consortium, Thomas J. Hudson, Warwick Anderson, Axel Artez, Anna D. Barker, Cindy Bell, Rosa R. Bernabé, et al. 2010. “International Network of Cancer Genome Projects.” *Nature* 464 (7291): 993–98.

Iorio, Francesco, Fiona M. Behan, Emanuel Gonçalves, Shriram G. Bhosle, Elisabeth Chen, Rebecca Shepherd, Charlotte Beaver, et al. 2018. “Unsupervised Correction of Gene-Independent Cell Responses to CRISPR-Cas9 Targeting.” *BMC Genomics* 19 (1): 604.

Iorio, Francesco, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, et al. 2016. “A Landscape of Pharmacogenomic Interactions in Cancer.” *Cell* 166 (3): 740–54.

Jaiswal, Alok, Prson Gautam, Elina A. Pietilä, Sanna Timonen, Nora Nordström, Yevhen Akimov, Nina Sipari, et al. 2021. “Multi-Modal Meta-Analysis of Cancer Cell Line Omics Profiles Identifies ECHDC1 as a Novel Breast Tumor Suppressor.” *Molecular Systems Biology* 17 (3): e9526.

Jost, J., and H. Saluz. 2013. *DNA Methylation: Molecular Biology and Biological Significance*. Vol. 64. Birkhäuser.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Jung, Jaeyun, Hyang Sook Seol, and Suhwan Chang. 2018. “The Generation and Application of Patient-Derived Xenograft Model for Cancer Research.” *Cancer Research and Treatment: Official Journal of Korean Cancer Association* 50 (1): 1–10.

Kaeberlein, Matt, and Brian K. Kennedy. 2007. “Protein Translation, 2007.” *Aging Cell* 6 (6): 731–34.

Kang, Mingon, Euiseong Ko, and Tesfaye B. Mersha. 2022. “A Roadmap for Multi-Omics Data Integration Using Deep Learning.” *Briefings in Bioinformatics* 23 (1). https://doi.org/10.1093/bib/bbab454.

Karlebach, Guy, and Ron Shamir. 2008. “Modelling and Analysis of Gene Regulatory Networks.” *Nature Reviews. Molecular Cell Biology* 9 (10): 770–80.

Kim, Jihoon, Bon-Kyoung Koo, and Juergen A. Knoblich. 2020. “Human Organoids: Model Systems for Human Biology and Medicine.” *Nature Reviews. Molecular Cell Biology* 21 (10): 571–84.

Kim, Kwang Gi. 2016. “Book Review: Deep Learning.” *Healthcare Informatics Research* 22 (4): 351.

Koza, John R., Forrest H. Bennett, David Andre, and Martin A. Keane. 1996. “Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming.” In *Artificial Intelligence in Design ’96*, edited by John S. Gero and Fay Sudweeks, 151–70. Dordrecht: Springer Netherlands.

Krueger, Felix, Benjamin Kreck, Andre Franke, and Simon R. Andrews. 2012. “DNA Methylome Analysis Using Short Bisulfite Sequencing Data.” *Nature Methods* 9 (2): 145–51.

Krug, Karsten, Eric J. Jaehnig, Shankha Satpathy, Lili Blumenberg, Alla Karpova, Meenakshi Anurag, George Miles, et al. 2020. “Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy.” *Cell* 183 (5): 1436-1456.e31.

Kuenzi, Brent M., and Trey Ideker. 2020. “A Census of Pathway Maps in Cancer Systems Biology.” *Nature Reviews. Cancer* 20 (4): 233–46.

Kuenzi, Brent M., Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. 2020. “Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells.” *Cancer Cell* 38 (5): 672-684.e6.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (6822): 860–921.

Landras, Alexandra, Coralie Reger de Moura, Fanelie Jouenne, Celeste Lebbe, Suzanne Menashi, and Samia Mourah. 2019. “CD147 Is a Promising Target of Tumor Progression and a Prognostic Biomarker.” *Cancers* 11 (11): 1803.

Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* 499 (7457): 214–18.

Lawrence, Robert T., Elizabeth M. Perez, Daniel Hernández, Chris P. Miller, Kelsey M. Haas, Hanna Y. Irie, Su-In Lee, C. Anthony Blau, and Judit Villén. 2015.

"The Proteomic Landscape of Triple-Negative Breast Cancer." *Cell Reports* 11 (4): 630–44.

Le Tourneau, Christophe, Jean-Pierre Delord, Anthony Gonçalves, Céline Gavoille, Coraline Dubot, Nicolas Isambert, Mario Campone, et al. 2015. "Molecularly Targeted Therapy Based on Tumour Molecular Profiling versus Conventional Therapy for Advanced Cancer (SHIVA): A Multicentre, Open-Label, Proof-of-Concept, Randomised, Controlled Phase 2 Trial." *The Lancet Oncology* 16 (13): 1324–34.

Lee, Bohyun, Shuo Zhang, Aleksandar Poleksic, and Lei Xie. 2019. "Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis." *Frontiers in Genetics* 10: 1381.

Lee, Jeong-Ju, Hee Jin Lee, Byung-Ho Son, Sung-Bae Kim, Jin-Hee Ahn, Seung Do Ahn, Eun Yoon Cho, and Gyungyub Gong. 2016. "Expression of FOXM1 and Related Proteins in Breast Cancer Molecular Subtypes." *International Journal of Experimental Pathology* 97 (2): 170–77.

Li, Xiangtao, and Ka-Chun Wong. 2019. "Evolutionary Multiobjective Clustering and Its Applications to Patient Stratification." *IEEE Transactions on Cybernetics* 49 (5): 1680–93.

Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. "Molecular Signatures Database (MSigDB) 3.0." *Bioinformatics (Oxford, England)* 27 (12): 1739–40.

Lin, Tianyang, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. "A Survey of Transformers." *ArXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/2106.04554.

Lin, Yen-Yu, Tyng-Luh Liu, and Chiou-Shann Fuh. 2011. "Multiple Kernel Learning for Dimensionality Reduction." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (6): 1147–60.

Liu, Ming, Julia Y. S. Tsang, Michelle Lee, Yun-Bi Ni, Siu-Ki Chan, Sai-Yin Cheung, Jintao Hu, Hong Hu, and Gary M. K. Tse. 2018. "CD147 Expression Is Associated with Poor Overall Survival in Chemotherapy Treated Triple-Negative Breast Cancer." *Journal of Clinical Pathology* 71 (11): 1007–14.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. "On the Dependency of Cellular Protein Levels on MRNA Abundance." *Cell* 165 (3): 535–50.

López de Maturana, Evangelina, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano, Silvia Pineda, Lucas Piorno, M. Luz Calle, and Núria Malats. 2019. "Challenges in the Integration of Omics and Non-Omics Data." *Genes* 10 (3): 238.

Lötsch, Jörn, Dario Kringel, and Alfred Ultsch. 2021. "Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients." *BioMedInformatics* 2 (1): 1–17.

Lu, Ming Y., Tiffany Y. Chen, Drew F. K. Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. 2021. "AI-Based Pathology Predicts Origins for Cancers of Unknown Primary." *Nature* 594 (7861): 106–10.

Lu, Xiao-Feng, De Zeng, Wei-Quan Liang, Chun-Fa Chen, Shu-Ming Sun, and Hao-Yu Lin. 2018. "FoxM1 Is a Promising Candidate Target in the Treatment of Breast Cancer." *Oncotarget* 9 (1): 842–52.

Lucas, Natasha, Andrew B. Robinson, Maiken Marcker Espersen, Sadia Mahboob, Dylan Xavier, Jing Xue, Rosemary L. Balleine, Anna deFazio, Peter G. Hains, and Phillip J. Robinson. 2019. "Accelerated Barocycler Lysis and Extraction Sample Preparation for Clinical Proteomics by Mass Spectrometry." *Journal of Proteome Research* 18 (1): 399–405.

Luck, Katja, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, et al. 2020. “A Reference Map of the Human Binary Protein Interactome.” *Nature* 580 (7803): 402–8.

Ludwig, Christina, Ludovic Gillet, George Rosenberger, Sabine Amon, Ben C. Collins, and Ruedi Aebersold. 2018. “Data-Independent Acquisition-Based SWATH-MS for Quantitative Proteomics: A Tutorial.” *Molecular Systems Biology* 14 (8): e8126.

Lundberg, Scott M., and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc.

Ma, Jianzhu, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. 2018. “Using Deep Learning to Model the Hierarchical Structure and Function of a Cell.” *Nature Methods* 15 (4): 290–98.

Malone, John H., and Brian Oliver. 2011. “Microarrays, Deep Sequencing and the True Measure of the Transcriptome.” *BMC Biology* 9 (1): 34.

Mani, D. R., Karsten Krug, Bing Zhang, Shankha Satpathy, Karl R. Clauser, Li Ding, Matthew Ellis, Michael A. Gillette, and Steven A. Carr. 2022. “Cancer Proteogenomics: Current Impact and Future Prospects.” *Nature Reviews. Cancer* 22 (5): 298–313.

Marchesin, Valentina, Antonio Castro-Castro, Catalina Lodillinsky, Alessia Castagnino, Joanna Cyrta, Hélène Bonsang-Kitzis, Laetitia Fuhrmann, et al. 2015. “ARF6-JIP3/4 Regulate Endosomal Tubules for MT1-MMP Exocytosis in Cancer Invasion.” *The Journal of Cell Biology* 211 (2): 339–58.

Marusyk, Andriy, Vanessa Almendro, and Kornelia Polyak. 2012. “Intra-Tumour Heterogeneity: A Looking Glass for Cancer?” *Nature Reviews. Cancer* 12 (5): 323–34.

Mathews, Sherin Mary. 2019. “Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review.” In *Intelligent Computing*, 1269–92. Springer International Publishing.

McDonald, E. Robert, 3rd, Antoine de Weck, Michael R. Schlabach, Eric Billy, Konstantinos J. Mavrakis, Gregory R. Hoffman, Dhiren Belur, et al. 2017. “Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening.” *Cell* 170 (3): 577-592.e10.

McGuire, Amy L., Stacey Gabriel, Sarah A. Tishkoff, Ambroise Wonkam, Aravinda Chakravarti, Eileen E. M. Furlong, Barbara Treutlein, et al. 2020. “The Road Ahead in Genetics and Genomics.” *Nature Reviews. Genetics* 21 (10): 581–96.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *ArXiv [Stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Medela, Alfonso, Artzai Picon, Cristina L. Saratxaga, Oihana Belar, Virginia Cabezon, Riccardo Cicchi, Roberto Bilbao, and Ben Glover. 2019. “Few Shot Learning in Histopathological Images:Reducing the Need of Labeled Data on Biological Datasets.” In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. https://doi.org/10.1109/isbi.2019.8759182.

Meer, Dieudonne van der, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E. Francies, and Mathew J. Garnett. 2019. “Cell Model Passports-a Hub for Clinical, Genetic and Functional Datasets of Preclinical Cancer Models.” *Nucleic Acids Research* 47 (D1): D923–29.

Meng, Chen, Dominic Helm, Martin Frejno, and Bernhard Kuster. 2016. “MoCluster: Identifying Joint Patterns Across Multiple Omics Data Sets.” *Journal of Proteome Research* 15 (3): 755–65.

Mertins, Philipp, NCI CPTAC, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, et al. 2016. “Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer.” *Nature* 534 (7605): 55–62.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. “Computational Correction of Copy Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells.” *Nature Genetics* 49 (12): 1779–84.

Mirabelli, Peppino, Luigi Coppola, and Marco Salvatore. 2019. “Cancer Cell Lines Are Useful Model Systems for Medical Research.” *Cancers* 11 (8): 1098.

Mirza, Bilal, Wei Wang, Jie Wang, Howard Choi, Neo Christopher Chung, and Peipei Ping. 2019. “Machine Learning and Integrative Analysis of Biomedical Big Data.” *Genes* 10 (2). https://doi.org/10.3390/genes10020087.

Mo, Qianxing, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S. Chan, and Susan G. Hilsenbeck. 2018. “A Fully Bayesian Latent Variable Model for Integrative Clustering Analysis of Multi-Type Omics Data.” *Biostatistics* 19 (1): 71–86.

Mo, Qianxing, Sijian Wang, Venkatraman E. Seshan, Adam B. Olshen, Nikolaus Schultz, Chris Sander, R. Scott Powers, Marc Ladanyi, and Ronglai Shen. 2013. “Pattern Discovery and Cancer Gene Identification in Integrated Cancer Genomic Data.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (11): 4245–50.

Moon, T. K. 1996. “The Expectation-Maximization Algorithm.” *IEEE Signal Processing Magazine* 13 (6): 47–60.

Nakagawa, Hidewaki, and Masashi Fujita. 2018. “Whole Genome Sequencing Analysis for Cancer Genomics and Precision Medicine.” *Cancer Science* 109 (3): 513–22.

Nam, Anna S., Ronan Chaligne, and Dan A. Landau. 2021. “Integrating Genetic and Non-Genetic Determinants of Cancer Evolution by Single-Cell Multi-Omics.” *Nature Reviews. Genetics* 22 (1): 3–18.

Nascimento, André C. A., Ricardo B. C. Prudêncio, and Ivan G. Costa. 2016. “A Multiple Kernel Learning Algorithm for Drug-Target Interaction Prediction.” *BMC Bioinformatics* 17 (January): 46.

Nicora, Giovanna, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. 2020. “Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools.” *Frontiers in Oncology* 10 (June): 1030.

Nielsen, Torsten O., Joel S. Parker, Samuel Leung, David Voduc, Mark Ebbert, Tammi Vickery, Sherri R. Davies, et al. 2010. “A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor-Positive Breast Cancer.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 16 (21): 5222–32.

Niklas, Karl J., Sarah E. Bondos, A. Keith Dunker, and Stuart A. Newman. 2015. "Rethinking Gene Regulatory Networks in Light of Alternative Splicing, Intrinsically Disordered Protein Domains, and Post-Translational Modifications." *Frontiers in Cell and Developmental Biology* 3 (February): 8.

Nusinow, David P., John Szpyt, Mahmoud Ghandi, Christopher M. Rose, E. Robert McDonald 3rd, Marian Kalocsay, Judit Jané-Valbuena, et al. 2020. "Quantitative Proteomics of the Cancer Cell Line Encyclopedia." *Cell* 180 (2): 387-402.e16.

O'Donnell, Shane Thomas, R. Paul Ross, and Catherine Stanton. 2019. "The Progress of Multi-Omics Technologies: Determining Function in Lactic Acid Bacteria Using a Systems Level Approach." *Frontiers in Microbiology* 10: 3084.

Oh, Minsik, Sungjoon Park, Sun Kim, and Heejoon Chae. 2021. "Machine Learning-Based Analysis of Multi-Omics Data on the Cloud for Investigating Gene Regulations." *Briefings in Bioinformatics* 22 (1): 66–76.

Ong, Shao-En, and Matthias Mann. 2005. "Mass Spectrometry–Based Proteomics Turns Quantitative." *Nature Chemical Biology* 1 (5): 252–62.

Pacini, Clare, Joshua M. Dempster, Isabella Boyle, Emanuel Gonçalves, Hanna Najgebauer, Emre Karakoc, Dieudonne van der Meer, et al. 2021. "Integrated Cross-Study Datasets of Genetic Dependencies in Cancer." *Nature Communications* 12 (1): 1661.

Parker, Joel S., Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, et al. 2009. "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 27 (8): 1160–67.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. D\textquotesingle Alché-Buc, E. Fox, and R. Garnett, 8026–37. Curran Associates, Inc.

Pauli, Chantal, Benjamin D. Hopkins, Davide Prandi, Reid Shaw, Tarcisio Fedrizzi, Andrea Sboner, Verena Sailer, et al. 2017. "Personalized in Vitro and in Vivo Cancer Models to Guide Precision Medicine." *Cancer Discovery* 7 (5): 462–77.

Pavlidis, Nicholas, and George Pentheroudakis. 2012. "Cancer of Unknown Primary Site." *The Lancet* 379 (9824): 1428–35.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2012. "Scikit-Learn: Machine Learning in Python." *ArXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/1201.0490.

Peng, Anghui, Xiying Mao, Jiawei Zhong, Shuxin Fan, and Youjin Hu. 2020. "Single-Cell Multi-Omics and Its Prospective Application in Cancer Biology." *Proteomics* 20 (13): e1900271.

Perez-Riverol, Yasset, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J. Kundu, Avinash Inuganti, et al. 2019. "The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data." *Nucleic Acids Research* 47 (D1): D442–50.

Picard, Milan, Marie-Pier Scott-Boyer, Antoine Bodein, Olivier Périn, and Arnaud Droit. 2021. "Integration Strategies of Multi-Omics Data for Machine

Learning Analysis." *Computational and Structural Biotechnology Journal* 19 (June): 3735–46.
Picco, Gabriele, Elisabeth D. Chen, Luz Garcia Alonso, Fiona M. Behan, Emanuel Gonçalves, Graham Bignell, Angela Matchan, et al. 2019. "Functional Linkage of Gene Fusions to Cancer Cell Fitness Assessed by Pharmacological and CRISPR-Cas9 Screening." *Nature Communications* 10 (1): 2198.
Poulos, Rebecca C., Peter G. Hains, Rohan Shah, Natasha Lucas, Dylan Xavier, Srikanth S. Manda, Asim Anees, et al. 2020. "Strategies to Enable Large-Scale Proteomics for Reproducible Research." *Nature Communications* 11 (1): 3793.
Poulos, Rebecca C., and Jason W. H. Wong. 2017. "Cis-Regulatory Driver Mutations in Cancer Genomes." In *ELS*, 1–10. John Wiley & Sons, Ltd.
Pozniak, Yair, Nora Balint-Lahat, Jan Daniel Rudolph, Cecilia Lindskog, Rotem Katzir, Camilla Avivi, Fredrik Pontén, Eytan Ruppin, Iris Barshack, and Tamar Geiger. 2016. "System-Wide Clinical Proteomics of Breast Cancer Reveals Global Remodeling of Tissue Homeostasis." *Cell Systems* 2 (3): 172–84.
Raj-Kumar, Praveen-Kumar, Jianfang Liu, Jeffrey A. Hooke, Albert J. Kovatich, Leonid Kvecher, Craig D. Shriver, and Hai Hu. 2019. "PCA-PAM50 Improves Consistency between Breast Cancer Intrinsic and Clinical Subtyping Reclassifying a Subset of Luminal A Tumors as Luminal B." *Scientific Reports* 9 (1): 7956.
Rappoport, Nimrod, and Ron Shamir. 2018. "Multi-Omic and Multi-View Clustering Algorithms: Review and Cancer Benchmark." *Nucleic Acids Research* 46 (20): 10546–62.
———. 2019. "NEMO: Cancer Subtyping by Integration of Partial Multi-Omic Data." *Bioinformatics* 35 (18): 3348–56.
Reel, Parminder S., Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. 2021. "Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review." *Biotechnology Advances* 49 (July): 107739.
Rees, Matthew G., Brinton Seashore-Ludlow, Jaime H. Cheah, Drew J. Adams, Edmund V. Price, Shubhroz Gill, Sarah Javaid, et al. 2016. "Correlating Chemical Sensitivity and Basal Gene Expression Reveals Mechanism of Action." *Nature Chemical Biology* 12 (2): 109–16.
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *ArXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/1602.04938.
R.O. Duda, P.E. Hart, and D. Stork. 2000. *Pattern Classification*. Wiley.
Rodosthenous, Theodoulos, Vahid Shahrezaei, and Marina Evangelou. 2020. "Integrating Multi-OMICS Data through Sparse Canonical Correlation Analysis for the Prediction of Complex Traits: A Comparison Study." Edited by Jonathan Wren. *Bioinformatics* 36 (17): 4616–25.
Rodriguez, Henry, Jean Claude Zenklusen, Louis M. Staudt, James H. Doroshow, and Douglas R. Lowy. 2021. "The next Horizon in Precision Oncology: Proteogenomics to Inform Cancer Diagnosis and Treatment." *Cell* 184 (7): 1661–70.
Rohart, Florian, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. 2017. "MixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration." *PLoS Computational Biology* 13 (11): e1005752.

Rohart, Florian, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. 2017. "MixOmics: An R Package for 'omics Feature Selection and Multiple Data Integration." *PLoS Computational Biology* 13 (11): e1005752.

Rokach, Lior, and Oded Maimon. 2006. "Decision Trees." In *Data Mining and Knowledge Discovery Handbook*, 165–92. New York: Springer-Verlag.

Roumeliotis, Theodoros I., Steven P. Williams, Emanuel Gonçalves, Clara Alsinet, Martin Del Castillo Velasco-Herrera, Nanne Aben, Fatemeh Zamanzad Ghavidel, et al. 2017. "Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells." *Cell Reports* 20 (9): 2201–14.

Ruepp, Andreas, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. 2010. "CORUM: The Comprehensive Resource of Mammalian Protein Complexes--2009." *Nucleic Acids Research* 38 (Database issue): D497-501.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–36.

Russell, Stuart Jonathan, Stuart Russell, and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach*. 4th ed. Pearson.

Ryan, Colm J., Susan Kennedy, Ilirjana Bajrami, David Matallanas, and Christopher J. Lord. 2017. "A Compendium of Co-Regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events." *Cell Systems* 5 (4): 399-409.e5.

Salvadores, Marina, Francisco Fuster-Tormo, and Fran Supek. 2020. "Matching Cell Lines with Cancer Type and Subtype of Origin via Mutational, Epigenomic, and Transcriptomic Patterns." *Science Advances* 6 (27). https://doi.org/10.1126/sciadv.aba1862.

Sathyanarayanan, Anita, Rohit Gupta, Erik W. Thompson, Dale R. Nyholt, Denis C. Bauer, and Shivashankar H. Nagaraj. 2020. "A Comparative Study of Multi-Omics Integration Tools for Cancer Driver Gene Identification and Tumour Subtyping." *Briefings in Bioinformatics* 21 (6): 1920–36.

Satpathy, Shankha, Karsten Krug, Pierre M. Jean Beltran, Sara R. Savage, Francesca Petralia, Chandan Kumar-Sinha, Yongchao Dou, et al. 2021. "A Proteogenomic Portrait of Lung Squamous Cell Carcinoma." *Cell* 184 (16): 4348-4371.e40.

Scheff, Jeremy D., Richard R. Almon, Debra C. Dubois, William J. Jusko, and Ioannis P. Androulakis. 2011. "Assessment of Pharmacologic Area under the Curve When Baselines Are Variable." *Pharmaceutical Research* 28 (5): 1081–89.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. 2015. "FaceNet: A Unified Embedding for Face Recognition and Clustering." In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–23. IEEE.

Schulte-Sasse, Roman, Stefan Budach, Denes Hnisz, and Annalisa Marsico. 2021. "Integration of Multiomics Data with Graph Convolutional Networks to Identify New Cancer Genes and Their Associated Molecular Mechanisms." *Nature Machine Intelligence* 3 (6): 513–26.

Schwarze, Katharina, James Buchanan, Jenny C. Taylor, and Sarah Wordsworth. 2018. "Are Whole-Exome and Whole-Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature." *Genetics in Medicine:*

Official Journal of the American College of Medical Genetics 20 (10): 1122–30.

Seashore-Ludlow, Brinton, Matthew G. Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E. Coletti, Victor Jones, et al. 2015. "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset." *Cancer Discovery* 5 (11): 1210–23.

Seligson, David B., Steve Horvath, Tao Shi, Hong Yu, Sheila Tze, Michael Grunstein, and Siavash K. Kurdistani. 2005. "Global Histone Modification Patterns Predict Risk of Prostate Cancer Recurrence." *Nature* 435 (7046): 1262–66.

Sharifi-Noghabi, Hossein, Olga Zolotareva, Colin C. Collins, and Martin Ester. 2019. "MOLI: Multi-Omics Late Integration with Deep Neural Networks for Drug Response Prediction." *Bioinformatics* 35 (14): i501–9.

Shen, Ronglai, Adam B. Olshen, and Marc Ladanyi. 2009. "Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis." *Bioinformatics* 25 (22): 2906–12.

Singh, Ajit P., and Geoffrey J. Gordon. 2008. "A Unified View of Matrix Factorization Models." In *Machine Learning and Knowledge Discovery in Databases*, edited by Walter Daelemans, Bart Goethals, and Katharina Morik, 5212:358–73. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.

Singh, Amrit, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, and Kim-Anh Lê Cao. 2019. "DIABLO: An Integrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays." *Bioinformatics* 35 (17): 3055–62.

Sousa, Abel, Emanuel Gonçalves, Bogdan Mirauta, David Ochoa, Oliver Stegle, and Pedro Beltrao. 2019. "Multi-Omics Characterization of Interaction-Mediated Control of Human Protein Abundance Levels." *Molecular & Cellular Proteomics: MCP* 18 (8 suppl 1): S114–25.

Spirin, Victor, and Leonid A. Mirny. 2003. "Protein Complexes and Functional Modules in Molecular Networks." *Proceedings of the National Academy of Sciences of the United States of America* 100 (21): 12123–28.

Stokholm, Jakob, Martin J. Blaser, Jonathan Thorsen, Morten A. Rasmussen, Johannes Waage, Rebecca K. Vinding, Ann-Marie M. Schoos, et al. 2018. "Maturation of the Gut Microbiome and Risk of Asthma in Childhood." *Nature Communications* 9 (1): 141.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. "The Cancer Genome." *Nature* 458 (7239): 719–24.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Subramanian, Indhupriya, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. 2020. "Multi-Omics Data Integration, Interpretation, and Its Application." *Bioinformatics and Biology Insights* 14 (January): 1177932219899051.

Suphavilai, Chayaporn, Denis Bertrand, and Niranjan Nagarajan. 2018. “Predicting Cancer Drug Response Using a Recommender System.” *Bioinformatics* 34 (22): 3907–14.

Szabo, Quentin, Frédéric Bantignies, and Giacomo Cavalli. 2019. “Principles of Genome Folding into Topologically Associating Domains.” *Science Advances* 5 (4): eaaw1668.

Szklarczyk, Damian, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, et al. 2021. “The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets.” *Nucleic Acids Research* 49 (D1): D605–12.

Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, et al. 2017. “The STRING Database in 2017: Quality-Controlled Protein–Protein Association Networks, Made Broadly Accessible.” *Nucleic Acids Research* 45 (D1): D362–68.

Tannock, Ian F., and John A. Hickman. 2016. “Limits to Personalized Cancer Medicine.” *The New England Journal of Medicine* 375 (13): 1289–94.

Tarazona, Sonia, Angeles Arzalluz-Luque, and Ana Conesa. 2021. “Undisclosed, Unmet and Neglected Challenges in Multi-Omics Studies.” *Nature Computational Science* 1 (6): 395–402.

Tate, John G., Sally Bamford, Harry C. Jubb, Zbyslaw Sondka, David M. Beare, Nidhi Bindal, Harry Boutselakis, et al. 2019. “COSMIC: The Catalogue Of Somatic Mutations In Cancer.” *Nucleic Acids Research* 47 (D1): D941–47.

Tellez-Gabriel, Marta, Benjamin Ory, Francois Lamoureux, Marie-Francoise Heymann, and Dominique Heymann. 2016. “Tumour Heterogeneity: The Key Advantages of Single-Cell Analysis.” *International Journal of Molecular Sciences* 17 (12). https://doi.org/10.3390/ijms17122142.

Thomas, Anish, Stephen V. Liu, Deepa S. Subramaniam, and Giuseppe Giaccone. 2015. “Refining the Treatment of NSCLC According to Histological and Molecular Subtypes.” *Nature Reviews. Clinical Oncology* 12 (9): 511–26.

Trastulla, Lucia, Javad Noorbakhsh, Francisca Vazquez, James McFarland, and Francesco Iorio. 2022. “Computational Estimation of Quality and Clinical Relevance of Cancer Cell Lines.” *Molecular Systems Biology* 18 (7): e11017.

Tsherniak, Aviad, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, et al. 2017. “Defining a Cancer Dependency Map.” *Cell* 170 (3): 564-576.e16.

Tully, Brett. 2020. “Toffee - a Highly Efficient, Lossless File Format for DIA-MS.” *Scientific Reports* 10 (1): 8939.

Tully, Brett, Rosemary L. Balleine, Peter G. Hains, Qing Zhong, Roger R. Reddel, and Phillip J. Robinson. 2019. “Addressing the Challenges of High-Throughput Cancer Tissue Proteomics for Clinical Application: ProCan.” *Proteomics* 19 (21–22): e1900109.

Ullah, Hsan, Andre Rios, Vaibhav Gala, and Susan Mckeever. 2020. “Explaining Deep Learning Models for Structured Data Using Layer-Wise Relevance Propagation.” *ArXiv [Cs.LG]*. arXiv. http://arxiv.org/abs/2011.13429.

Välikangas, Tommi, Tomi Suomi, and Laura L. Elo. 2018. “A Comprehensive Evaluation of Popular Proteomics Software Workflows for Label-Free Proteome Quantification and Imputation.” *Briefings in Bioinformatics* 19 (6): 1344–55.

Valouev, Anton, David S. Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M. Myers, and Arend Sidow. 2008. "Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data." *Nature Methods* 5 (9): 829–34.

Vasaikar, Suhas, Chen Huang, Xiaojing Wang, Vladislav A. Petyuk, Sara R. Savage, Bo Wen, Yongchao Dou, et al. 2019. "Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities." *Cell* 177 (4): 1035-1049.e19.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need.

Vijayakumaran, Reshma, Kah Hin Tan, Panimaya Jeffreena Miranda, Sue Haupt, and Ygal Haupt. 2015. "Regulation of Mutant P53 Protein Expression." *Frontiers in Oncology* 5 (December): 284.

Vis, Daniel J., Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J. Garnett, and Lodewyk Fa Wessels. 2016. "Multilevel Models Improve Precision and Speed of IC50 Estimates." *Pharmacogenomics* 17 (7): 691–700.

Wang, Bo, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. 2014. "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale." *Nature Methods* 11 (3): 333–37.

Wang, Kevin C., and Howard Y. Chang. 2018. "Epigenomics: Technologies and Applications." *Circulation Research* 122 (9): 1191–99.

Wang, Liang-Bo, Alla Karpova, Marina A. Gritsenko, Jennifer E. Kyle, Song Cao, Yize Li, Dmitry Rykunov, et al. 2021. "Proteogenomic and Metabolomic Characterization of Human Glioblastoma." *Cancer Cell* 39 (4): 509-528.e20.

Wang, Tongxin, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. 2021. "MOGONET Integrates Multi-Omics Data Using Graph Convolutional Networks Allowing Patient Classification and Biomarker Identification." *Nature Communications* 12 (1): 3445.

Wang, Yaqing, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2021. "Generalizing from a Few Examples." *ACM Computing Surveys* 53 (3): 1–34.

Webb-Robertson, Bobbie-Jo M., Holli K. Wiberg, Melissa M. Matzke, Joseph N. Brown, Jing Wang, Jason E. McDermott, Richard D. Smith, et al. 2015. "Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics." *Journal of Proteome Research* 14 (5): 1993–2001.

Wei, Lin, Zhilin Jin, Shengjie Yang, Yanxun Xu, Yitan Zhu, and Yuan Ji. 2018. "TCGA-Assembler 2: Software Pipeline for Retrieval and Processing of TCGA/CPTAC Data." *Bioinformatics* 34 (9): 1615–17.

Wei, Runmin, Jingye Wang, Mingming Su, Erik Jia, Shaoqiu Chen, Tianlu Chen, and Yan Ni. 2018. "Missing Value Imputation Approach for Mass Spectrometry-Based Metabolomics Data." *Scientific Reports* 8 (1): 663.

Wei, Ruoqi, and Ausif Mahmood. 2021. "Recent Advances in Variational Autoencoders with Representation Learning for Biomedical Informatics: A Survey." *IEEE Access: Practical Innovations, Open Solutions* 9: 4939–56.

Westerhuis, Johan A., Theodora Kourti, and John F. MacGregor. 1998. “Analysis of Multiblock and Hierarchical PCA and PLS Models.” *Journal of Chemometrics* 12 (5): 301–21.

Wetering, Marc van de, Hayley E. Francies, Joshua M. Francis, Gergana Bounova, Francesco Iorio, Apollo Pronk, Winan van Houdt, et al. 2015. “Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients.” *Cell* 161 (4): 933–45.

Wilhelm, Mathias, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, et al. 2014. “Mass-Spectrometry-Based Draft of the Human Proteome.” *Nature* 509 (7502): 582–87.

Wold, Svante, Kim Esbensen, and Paul Geladi. 1987. “Principal Component Analysis.” *Chemometrics and Intelligent Laboratory Systems* 2 (1–3): 37–52.

Wong, C. C. Y., E. L. Meaburn, A. Ronald, T. S. Price, A. R. Jeffries, L. C. Schalkwyk, R. Plomin, and J. Mill. 2014. “Methylomic Analysis of Monozygotic Twins Discordant for Autism Spectrum Disorder and Related Behavioural Traits.” *Molecular Psychiatry* 19 (4): 495–503.

Woo, Xing Yi, Anuj Srivastava, Joel H. Graber, Vinod Yadav, Vishal Kumar Sarsani, Al Simons, Glen Beane, et al. 2019. “Genomic Data Analysis Workflows for Tumors from Patient-Derived Xenografts (PDXs): Challenges and Guidelines.” *BMC Medical Genomics* 12 (1): 92.

Wreczycka, Katarzyna, Alexander Gosdschan, Dilmurat Yusuf, Björn Grüning, Yassen Assenov, and Altuna Akalin. 2017. “Strategies for Analyzing Bisulfite Sequencing Data.” *Journal of Biotechnology* 261 (November): 105–15.

Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. “A Comprehensive Survey on Graph Neural Networks.” *IEEE Transactions on Neural Networks and Learning Systems* 32 (1): 4–24.

Xu, Xiao, Yuanhao Zhang, Jennie Williams, Eric Antoniou, W. Richard McCombie, Song Wu, Wei Zhu, Nicholas O. Davidson, Paula Denoya, and Ellen Li. 2013. “Parallel Comparison of Illumina RNA-Seq and Affymetrix Microarray Platforms on Transcriptomic Profiles Generated from 5-Aza-Deoxy-Cytidine Treated HT-29 Colon Cancer Cells and Simulated Datasets.” *BMC Bioinformatics* 14 Suppl 9 (9): S1.

Xu, Zhi-Qiao, Yan Zhang, Ning Li, Pei-Jie Liu, Ling Gao, Xin Gao, and Xiao-Jing Tie. 2017. “Efficacy and Safety of Lapatinib and Trastuzumab for HER2-Positive Breast Cancer: A Systematic Review and Meta-Analysis of Randomised Controlled Trials.” *BMJ Open* 7 (3): e013053.

Xue, Hong-Jian, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. “Deep Matrix Factorization Models for Recommender Systems.” In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 17:3203–9. California: International Joint Conferences on Artificial Intelligence Organization.

Yang, Wanjuan, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, et al. 2013. “Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells.” *Nucleic Acids Research* 41 (Database issue): D955-61.

Yang, Yaping, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, et al. 2013. “Clinical Whole-Exome

Sequencing for the Diagnosis of Mendelian Disorders." *The New England Journal of Medicine* 369 (16): 1502–11.
Yun, Seongjun, Minbyul Jeong, Sungdong Yoo, Seunghun Lee, Sean S. Yi, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. 2022. "Graph Transformer Networks: Learning Meta-Path Graphs to Improve GNNs." *Neural Networks: The Official Journal of the International Neural Network Society* 153 (September): 104–19.
Zhang, Bing, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C. Chambers, et al. 2014. "Proteogenomic Characterization of Human Colon and Rectal Cancer." *Nature* 513 (7518): 382–87.
Zhang, Fei, Minghui Wang, Jianing Xi, Jianghong Yang, and Ao Li. 2018. "A Novel Heterogeneous Network-Based Method for Drug Response Prediction in Cancer Cell Lines." *Scientific Reports* 8 (1): 3355.
Zhang, Hui, Tao Liu, Zhen Zhang, Samuel H. Payne, Bai Zhang, Jason E. McDermott, Jian-Ying Zhou, et al. 2016. "Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer." *Cell* 166 (3): 755–65.
Zhang, Yijie, Dan Wang, Miao Peng, Le Tang, Jiawei Ouyang, Fang Xiong, Can Guo, et al. 2021. "Single‐cell RNA Sequencing in Cancer Research." *Journal of Experimental & Clinical Cancer Research: CR* 40 (1): 81.
Zhang, Yiqun, Fengju Chen, Darshan S. Chandrashekar, Sooryanarayana Varambally, and Chad J. Creighton. 2022. "Proteogenomic Characterization of 2002 Human Cancers Reveals Pan-Cancer Molecular Subtypes and Associated Pathways." *Nature Communications* 13 (1): 2669.
Zhang, Zhiqiang, Yi Zhao, Xiangke Liao, Wenqiang Shi, Kenli Li, Quan Zou, and Shaoliang Peng. 2019. "Deep Learning in Omics: A Survey and Guideline." *Briefings in Functional Genomics* 18 (1): 41–57.
Zhong, Qing, Ulrich Wagner, Henriette Kurt, Francesca Molinari, Gieri Cathomas, Paul Komminoth, Jasmin Barman-Aksözen, et al. 2018. "Multi-Laboratory Proficiency Testing of Clinical Cancer Genomic Profiling by next-Generation Sequencing." *Pathology, Research and Practice* 214 (7): 957–63.
Zhou, Lina, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. 2017. "Machine Learning on Big Data: Opportunities and Challenges." *Neurocomputing* 237 (May): 350–61.
Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 67 (2): 301–20.

Appendices

Appendix A – Supplementary Data relating to Chapter 3

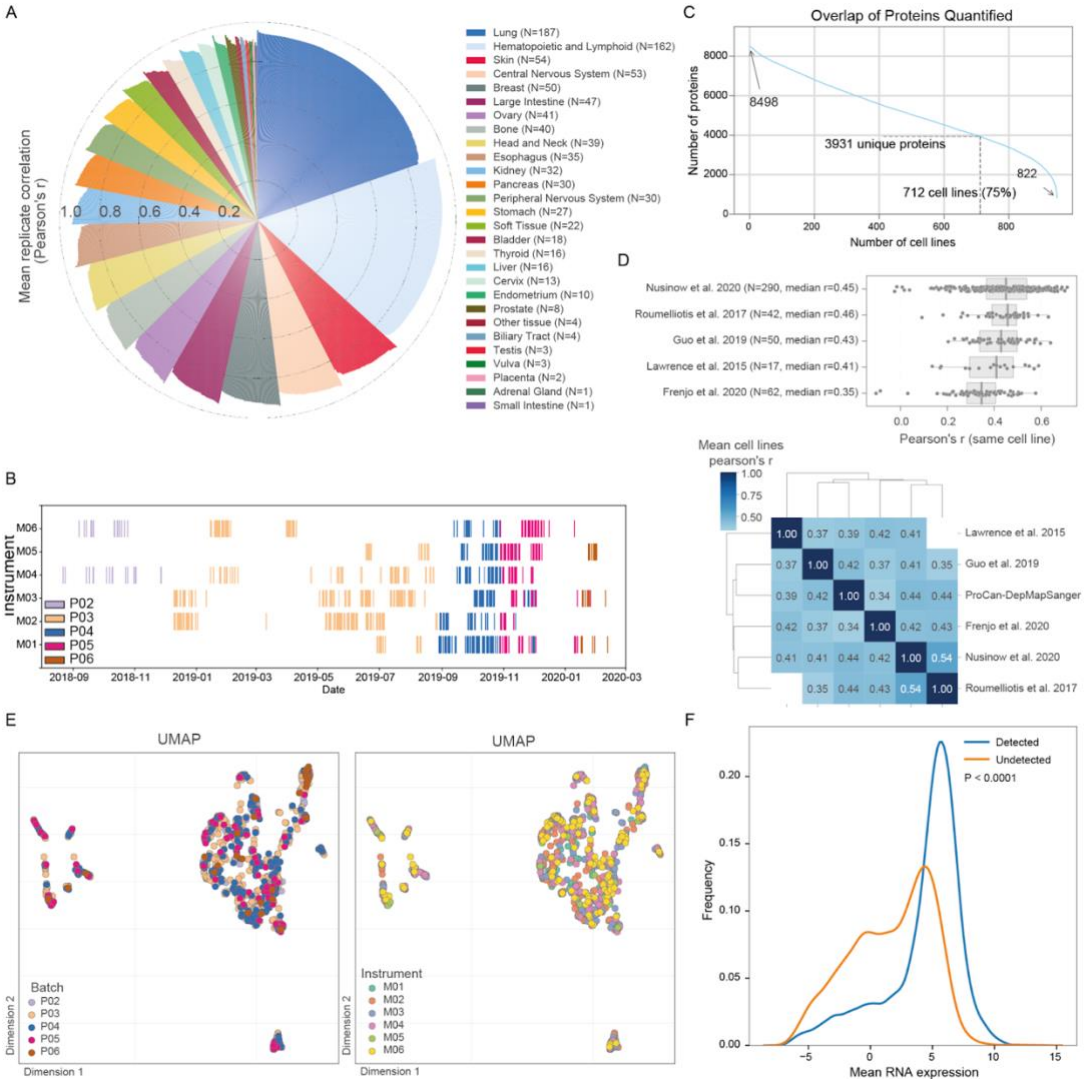


Figure S1. A pan-cancer proteomic map of 949 human cancer cell lines by Data Independent Acquisition Mass Spectrometry (DIA-MS), Related to Figure 1. A, Mean Pearson's r for replicates of each cancer cell line, coloured by tissue of origin. **B,** Timeline of MS data acquisition across mass spectrometers, coloured according to processing batches (P02 - P06). **C,** Frequency of proteins identified across the 949 cancer cell lines. **D,** Upper panel, correlation by Pearson's r of ProCan-DepMapSanger dataset against independent proteomic datasets that comprise subsets of the same cell lines. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. Lower panel, heatmap and dendrogram (average method with euclidean metric) of the mean pairwise correlations between the same cell lines in different studies. **E,** Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction of cell line proteomes coloured by processing batches (left) and mass spectrometer (right). **F,** Distribution of mean RNA-seq expression for genes corresponding to proteins that were detected (blue) or undetected (orange) in the ProCan-DepMapSanger dataset. Significance is indicated by Mann-Whitney U test.

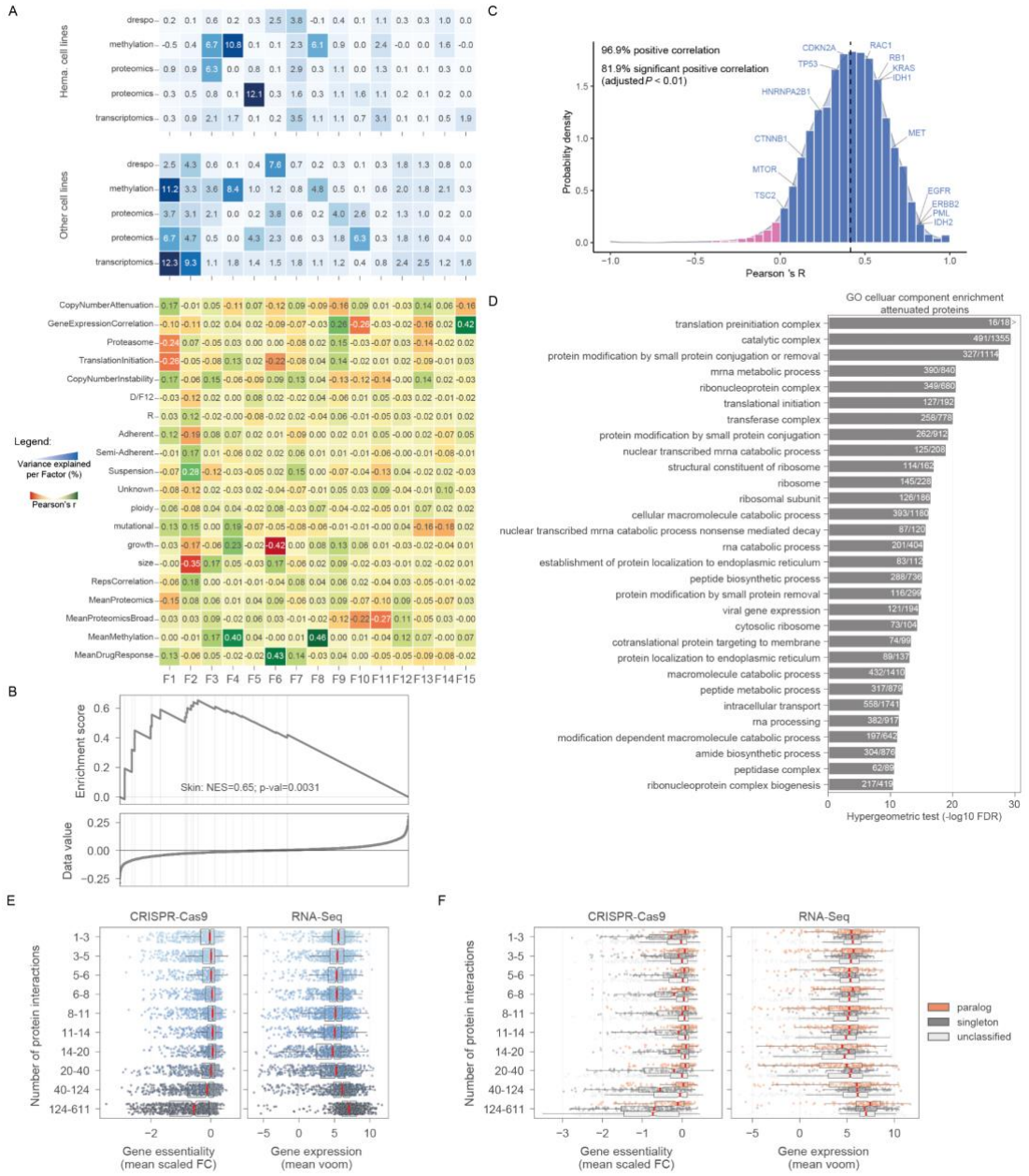


Figure S2. Multi-Omics Factor Analysis (MOFA) and post-transcriptional regulation, Related to Figure 3. A, Similar to **Figure 3A**, MOFA factors across molecular and phenotypic cancer cell line datasets, including ProCan-DepMapSanger. Hematopoietic and lymphoid cells are grouped and trained separately from the other cell lines corresponding to each factor (column). The upper two heatmaps (blue) report the portion of variance explained by each factor (columns) in each dataset. The lower heatmap reports Pearson's r between each learned factor and various molecular characteristics of the cancer cell lines. **B,** Gene Set Enrichment Analysis (GSEA) demonstrating enrichment of skin cell type-enriched proteins in MOFA Factor 12. **C,** Per-gene Pearson's r between protein and RNA expression for all proteins quantified. Mean correlation (r = 0.42) is indicated by a dashed line. The locations of several cancer-related genes are shown. **D,** Enrichment analysis of proteins that were highly attenuated (n = 1,215), as determined by correlations between protein and copy number, and gene expression and copy number, in terms of Pearson's r. The top most significantly enriched sets are annotated. **E, F,** Proteins were grouped by their number of significant positive correlations by Pearson's r for putative protein interactions (FDR < 5%, r > 0.5). For each protein, the respective mean scaled CRISPR-Cas9 gene essentiality fold-change (FC) and gene expression (RNA-seq voom) measurement are calculated. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. **E,** The distribution

across all proteins is represented and **F,** proteins are subgrouped into paralog, singleton and unclassified, as previously determined (Dandage and Landry 2019).

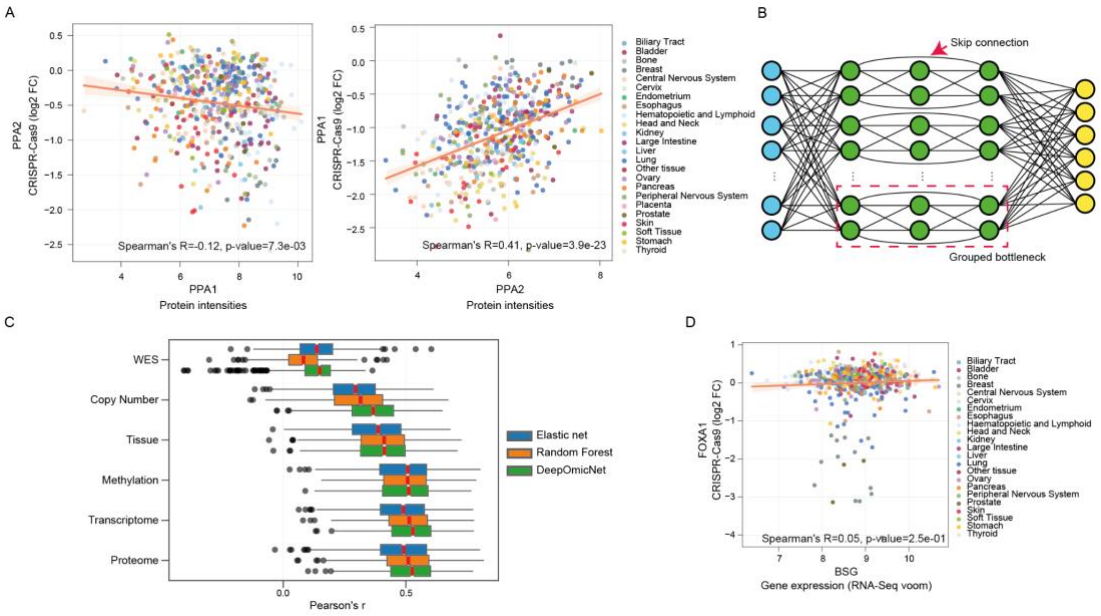


Figure S3. Drug-protein and CRISPR-Cas9-protein associations and Deep Proteomic Marker (DeepProM) analysis pipeline, Related to Figure 4 and Figure 5. A, Synthetic lethal association between PPA2 and PPA1. Left panel, scatter plot between protein intensities of PPA1 and PPA2 CRISPR-Cas9 gene essentiality scores. Right panel, scatter plot between protein intensities of PPA2 and CRISPR-Cas9 gene essentiality scores of PPA1. Cell lines are coloured by tissue types. **B,** Neural network architecture of DeepOmicNet. In addition to the basic multilayer perceptron (MLP) architecture, skip connections with grouped bottlenecks were added to provide a deeper and wider network. Blue circles represent input neurons, green represents hidden layer neurons and yellow represents output neurons. **C,** Comparison of observed drug responses and predicted drug responses by DeepProM, elastic net and Random Forest. WES, mutation data from whole exome sequencing; Copy number, copy number profiles; Tissue, categorical variable representing the cell line's tissue of origin; Methylation, promoter region methylation level; Transcriptome, RNA-seq data; Proteome, the ProCan-DepMapSanger dataset. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. **D,** Scatter plot between FOXA1 CRISPR-Cas9 gene essentiality scores and BSG gene expression measurements. Data points are coloured according to tissue type.

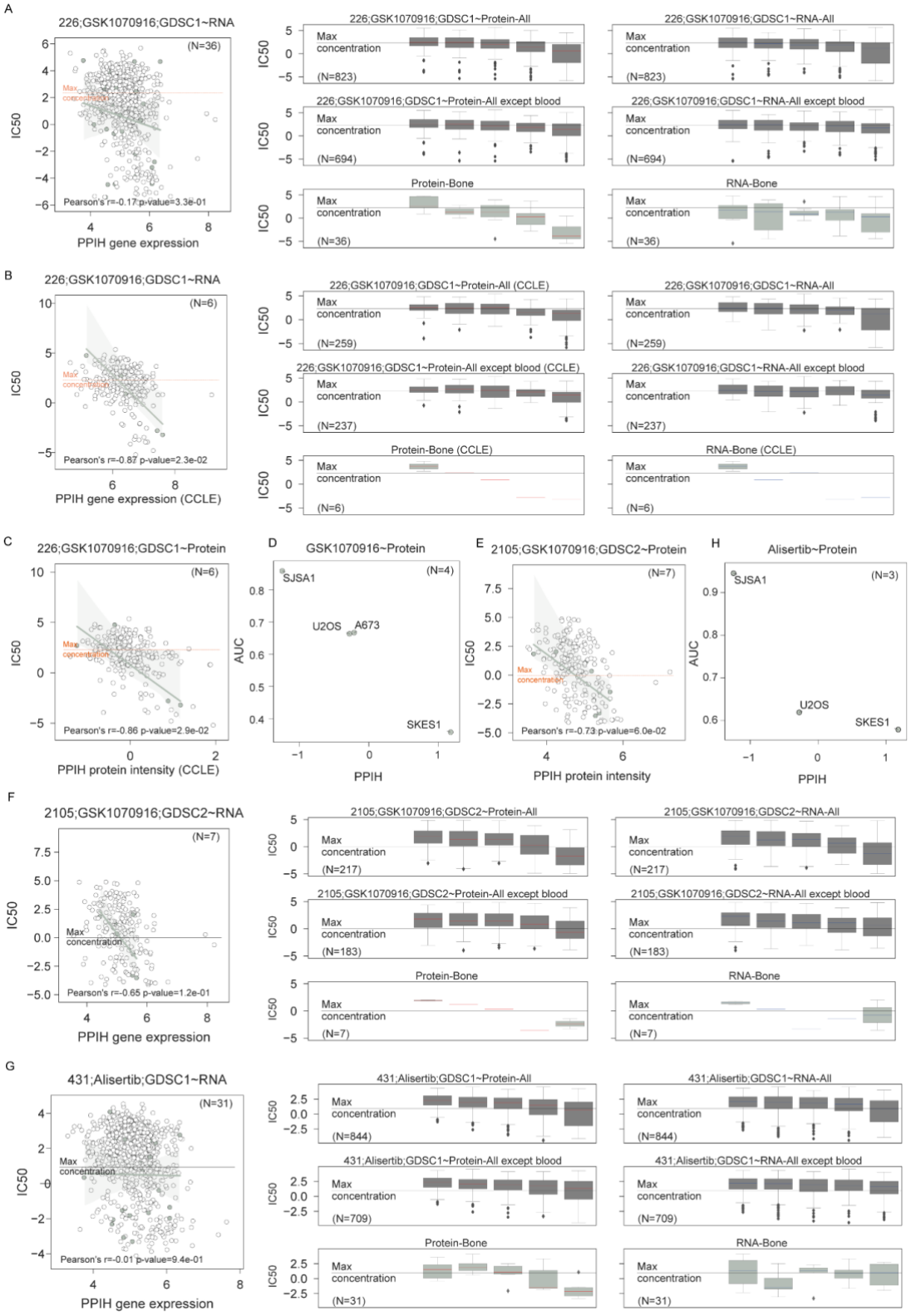


Figure S4. Tissue-level protein biomarkers for GSK1070916 across datasets, Related to Figure 5. A-C, E-G Scatter plots show the relationship between the drug and the protein biomarker, using either protein abundance or underlying RNA expression in cell lines from bone (green; all other cell lines are shown in gray). The number of cell lines and Pearson's r from the highlighted tissue type are annotated at the top right and bottom left corners, respectively. The dashed line represents the maximum concentration used in the drug response screens. In **A**, **B**, **F** and **G**, Box-and-whisker plots summarize the information from the scatter plots of protein (left) and RNA (right). The protein

intensity is divided into five equally spaced quantiles from low to high, and the corresponding IC_{50} values in the natural logarithmic scale are shown for each quantile. The first row of plots shows the relationship for all cell lines, and non-hematopoietic cell lines are shown in the second row. The relationship for only cell lines from bone is shown in the third row. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. **A,** The association between PPIH gene expression and GSK1070916 drug response in cell lines from bone. **B,** Similar to **A,** instead using CCLE gene expression data. **C,** Similar to **A,** instead using the CCLE proteomic dataset. **D,** The association between PPIH protein abundance and response to GSK1070916 using the CCLE proteomic dataset and PRISM drug response data, with the figure obtained from the DepMap Portal. **E,** Similar to **Figure 5D**, instead using the GDSC2 drug response dataset. **F,** Similar to **E,** instead using gene expression data. **G,** The association between PPIH gene expression and Alisertib drug response from GDSC data in cell lines from bone. **H,** Similar to **D**, instead showing response to Alisertib.

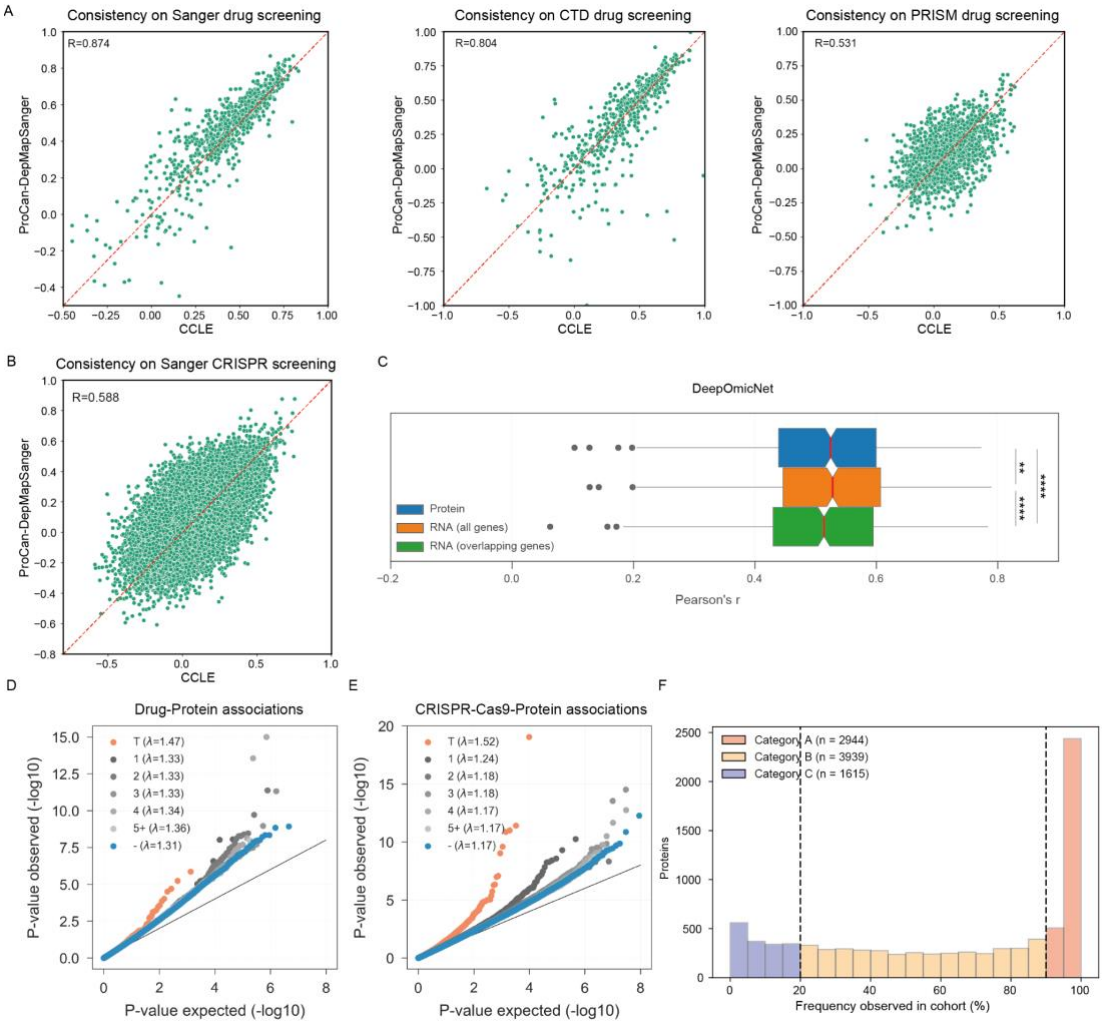
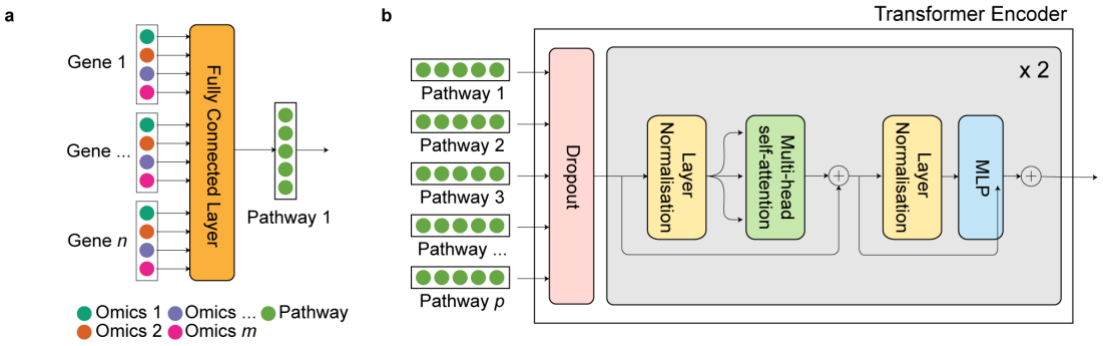


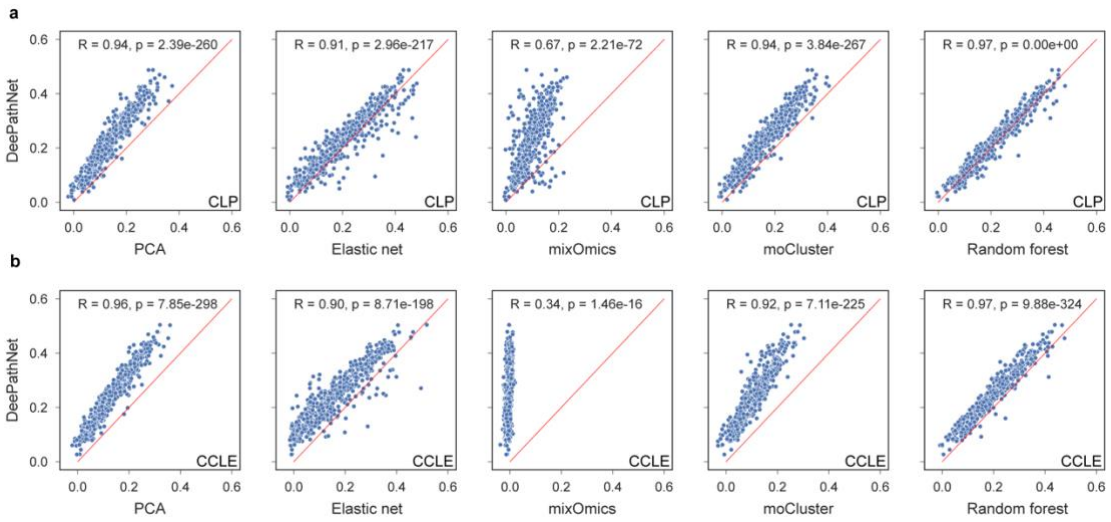
Figure S5. Predictive power benchmarks and comparisons, Related to Figure 6 and Figure 7. A, Comparison of DeepOmicNet mean predictive power across three independent drug response datasets, trained using the ProCan-DepMapSanger and the CCLE proteomic dataset**. B,** Similar to **A**, models trained to predict CRISPR-Cas9 gene essentiality profiles. **C,** Comparison of the predictive power of the DeepOmicNet model trained using the ProCan-DepMapSanger dataset and RNA expression data, both for all genes or for only genes that have their corresponding proteins quantified (overlapping genes). **** p-value < 0.0001, ** p-value < 0.01. Significance is by two-tailed paired Student's t-test. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 x IQRs. **D,** Drug responses (drug-protein) and **E,** CRISPR-Cas9 gene essentiality (CRISPR-Cas9-protein) associations, identified with linear regression models without taking gene expression as covariates. See **Figure 7C-D** for a general description of the plot. **F,** Histogram showing the distribution of the frequency of proteins detected in each of the cell lines in the cohort, with Category A, B and C proteins indicated. The dashed vertical lines indicate the frequency thresholds for defining the categories.

Please see the soft copy contents accompanying this thesis for a copy of **Table S1-S5**.

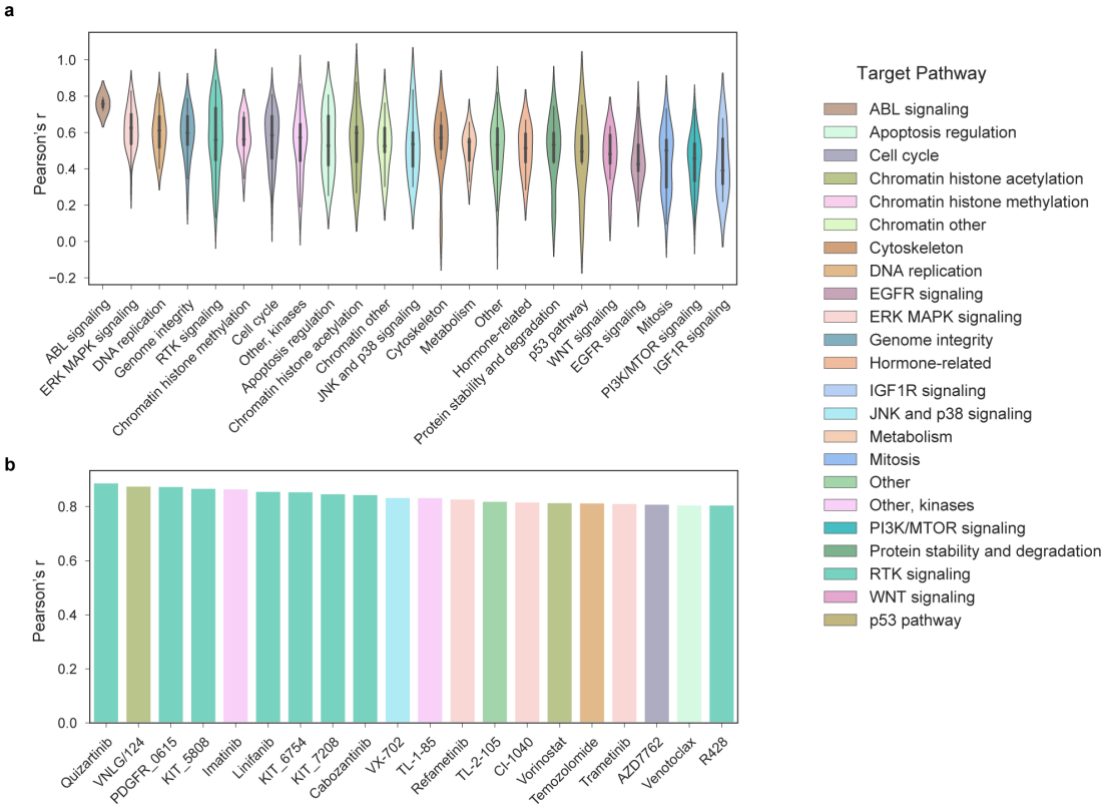
Appendix B – Supplementary Data relating to Chapter 4



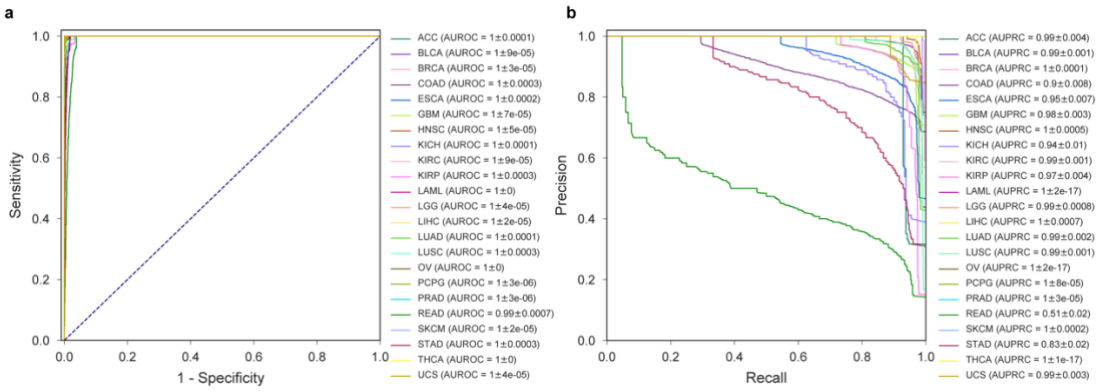
Supplementary Figure 1 | Details of pathway encoder and Transformer encoder. a, Detailed illustration of the pathway encoder. A fully connected layer encodes multi-omic data from genes into a pathway vector. **b,** Detailed illustration of the Transformer encoder. Pathway vectors are first fed into a dropout layer, followed by a recurrent sequence (grey box) of layer normalisation, multi-head self-attention and multi-layer perceptron (MLP). The components in the grey box recur twice in DeePathNet. Arrows represent the direction of information flow and ⊕ represents matrix addition.



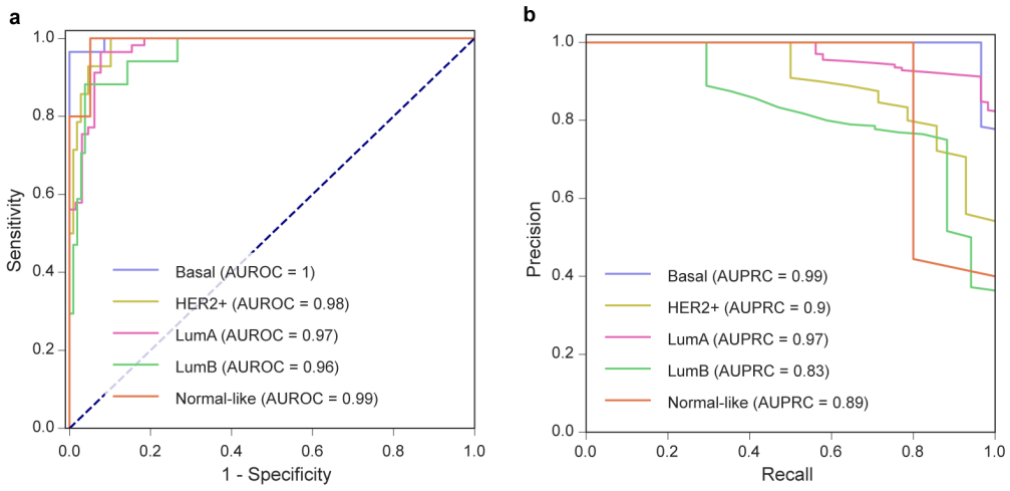
Supplementary Figure 2 | Consistency of drug response predictions from different machine learning models. a, Scatter plots showing R^2 of drugs from DeePathNet (vertical axis) against each of the remaining five machine learning models (horizontal axis) evaluated on the CLP dataset. The diagonal red line indicates equal performance between the two models. Points above the red line represent drugs that are more accurately predicted by DeePathNet. Pearson's *r* (R) and *p*-value (p) are annotated. **b,** Similar to **a**, but evaluated on the CCLE dataset.



Supplementary Figure 3 | Analysis of performance of drug response prediction by target pathways. a, A violin plot showing the predictive performance grouped by drug canonical target pathways, ranked by the mean Pearson's *r* of the group. **b,** Top 20 drugs ranked by Pearson's *r*. Drugs are coloured by their canonical target pathways.



Supplementary Figure 4 | ROC curves and precision-recall curves for TCGA cancer type classification. a, ROC curves for DeePathNet classification of TCGA cancer types. Mean AUROC and standard error of the mean are annotated. **b,** Precision-recall curves for DeePathNet classification of TCGA cancer types. Mean AUPRC and standard error of the mean are annotated. Full terms of the abbreviations in **a** and **b** are listed in **Fig. 4c**.



Supplementary Figure 5 | ROC curves and precision-recall curves for breast cancer subtype classification. a, ROC curves for DeePathNet classification of breast cancer subtypes using TCGA as the training data and CPTAC as the test data. **b,** Similar to **a**, but showing AUPRC.

Data sets	**Number of features**			
(n = sample size)	**Gene mutation**	**CNV**	**Gene expression**	**Protein**
Drug response prediction (n = 549 GDSC drugs)				
CLP (n = 941)	19,099	19,116	15,320	-
CLP^+ (n = 910)	19,099	19,116	15,320	8,498
CCLE (n = 696)	18,103	27,562	19,117	-
$CCLE^+$ (n = 292)	18,103	27,562	19,117	12,755
Cancer type classification				
TCGA (n = 6,356)	31,949	23,529	20,435	-
Breast cancer subtype classification				
TCGA (n = 974)	31,949	23,529	20,435	-
CPTAC (n = 122)	11,877	23,692	23,121	-

Supplementary Table 1. Overview of datasets used in this study. CLP^+ = CLP + ProCan-DepMapSanger and $CCLE^+$= CCLE + CCLE proteomic dataset.

	R2 mean ± 95%CI	MAE mean ± 95%CI	Pearson's r mean ± 95%CI
CLP			
DeePathNet	**0.222±0.0020**	**0.947±0.0021**	**0.475±00.0053**
Random forest (RF)	0.214±0.0018	0.964±0.0021	0.469±0.0054
Elastic net	0.200±0.0020	0.965±0.0023	0.452±0.0051
moCluster+RF	0.155±0.0015	1.009±0.0022	0.413±0.0057
PCA+RF	0.138±0.0014	1.021±0.0022	0.403±0.0058
mixOmics	0.097±0.0009	1.047±0.0019	0.342±0.0061
CCLE			
DeePathNet	**0.242±0.0020**	**0.934±0.0020**	**0.496±0.0052**
Random forest (RF)	0.197±0.0018	0.977±0.0022	0.452±0.0055
Elastic net	0.163±0.0019	0.991±0.0025	0.427±0.0053
PCA+RF	0.135±0.0014	1.024±0.0023	0.404±0.0059
moCluster+RF	0.107±0.0012	1.042±0.0023	0.363±0.0060
mixOmics	-0.006±0.0006	1.110±0.0016	0.110±0.0066

Supplementary Table 2. Benchmarking six methods to predict drug responses by reporting mean cross-validation performance with 95% confidence interval (CI).

	R^2 mean ± 95%CI	MAE mean ± 95%CI	Pearson's r mean ± 95%CI
CLP > CCLE			
DeePathNet	**0.208±0.0118**	**0.933±0.0274**	**0.476±0.0119**
Random forest	0.027±0.0200	1.07±0.0315	0.390±0.0143
CLP^+ > $CCLE^+$			
DeePathNet	**0.233±0.0126**	**0.899±0.0262**	**0.532±0.0139**
Random forest	-0.106±0.0441	1.107±0.0330	0.388±0.0183

Supplementary Table 3. Comparing two methods by evaluating mean generalisation errors with 95% CI of drug response prediction.

	Accuracy mean ± 95%CI	Macro-average F1-score mean ± 95%CI	AUROC mean ± 95%CI	Stability
DeePathNet	**0.963±0.0015**	**0.935±0.0030**	**0.998±0.0001**	**0.004**
Random forest (RF)	0.951±0.0018	0.895±0.0038	0.997±0.0003	0.005
k-NN	0.940±0.0018	0.894±0.0045	0.982±0.0016	0.007
PCA+RF	0.937±0.0023	0.885±0.0039	0.996±0.0003	0.005
moCluster	0.866±0.0031	0.734±0.0034	0.987±0.0010	0.006
mixOmics+RF	0.764±0.0040	0.881±0.0091	0.919±0.0041	0.015

Supplementary Table 4. Benchmarking six methods to predict cancer types by reporting cross-validation performance.

	Accuracy mean ± 95%CI	Macro-average F1-score mean ± 95%CI	AUROC mean ± 95%CI	Stability
DeePathNet	**0.902±0.0081**	**0.868±0.0115**	**0.980±0.0030**	**0.019**
Random forest (RF)	0.844±0.0095	0.672±0.0178	0.969±0.0038	0.026
k-NN	0.816±0.0097	0.697±0.0200	0.897±0.0094	0.033
PCA+RF	0.696±0.0109	0.429±0.0111	0.880±0.0094	0.027
mixOmics	0.749±0.0525	0.676±0.0369	0.731±0.0169	0.090
moCluster+RF	0.751±0.0090	0.492±0.0152	0.917±0.0051	0.025

Supplementary Table 5. Benchmarking six methods to predict breast cancer subtypes by reporting cross-validation performance.

	Accuracy	Macro-average F1-score	AUROC
DeePathNet	**0.902**	**0.798**	**0.971**
Random forest	0.295	0.263	0.694

Supplementary Table 6. Comparing two methods by evaluating generalisation errors of breast cancer subtype prediction.