Cancer molecular subtyping: a machine learning approach

Brian Karlberg

Ph.D., Oregon Health and Science University, 2025

A DISSERTATION

Presented to the Department of Biomedical Engineering of the Oregon Health & Science University School of Medicine in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computational Biology

School of Medicine
Oregon Health and Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of
Brian Karlberg
has been approved, June 2025

_____
Advisor, Kyle Ellrott, Ph.D.

_____
Member and Chair, Emek Demir, Ph.D.

_____
Member, Stuart Ibsen, Ph.D.

_____
Member, Young Hwan Chang, Ph.D.

_____
Member, Andrew Emili, Ph.D.

_____
Member, Sara Gosline, Ph.D

Table of contents

Acknowledgments

Deep gratitude to my mentor Kyle Ellrott for teaching me how to write scientifically and the opportunity for learning and practicing rigorous experimental methodology. I am grateful to each member of the Ellrott Lab: Jordan Lee, Brian Walsh, Adam Struck, Liam Beckman, Matthew Peterkort, Nasim Santini, and Quinn Wai Wong. Ein außergewöhnliches Gefühl tiefer Dankbarkeit gilt meinem Kooperationspartner und Freund Raphael Kirchgässner; Ich wünsche ihm alles Gute. And to my dissertation committee: Emek Demir, Stuart Ibsen, Young Hwan Chang, Andrew Emili, Sara Gosline – thank you each – your time and experience are greatly appreciated. To the PMedIC team: Laura Erker, Liz Sturgill, and Beth Habacker – your conscientious work was instrumental in navigating me to Pacific Northwest National Laboratory in Seattle. Thank you to Oregon Health and Science University, the School of Medicine and Biomedical Engineering Department, the Registrar's Office, OHSU Biomedical Communications Center, the Advanced Computing Center, Holly Chung, Alyssa Yadao, and Sandra Rugonyi. I would also like to thank Jeremy Geocks.

List of Abbreviations

ACC .....Adrenocortical Carcinoma

AJCC .....American Joint Committee on Cancer

AKLIMATE.....Algorithm for Kernel Learning with Integrative Modules of Approximating Tree Ensembles

ARC .....Advanced Research Cluster (OHSU's HPC, formerly Exacloud)

AUC.....area under the curve

AURORA.....Aiming to Understand the Molecular Aberrations in Metastatic Breast Cancer

BeatAML .....Beat Acute Myeloid Leukemia

BLCA.....Bladder Urothelial Carcinoma

BRCA.....Breast Invasive Carcinoma

CCG .....Center for Cancer Genomics

CESC .....Cervical and Endocervical Cancer

COADREAD .....Colorectal Adenocarcinoma

CoderData .....Cancer Omics Drug Experiment Response Dataset

COMP .....composite (type of mutation)

COSMIC .....Catalogue Of Somatic Mutations in Cancer

CpG.....cytosine-phosphate-guanine

CPTAC.....Clinical Proteomic Tumor Analysis Consortium

DL .....deep learning

DNA.....deoxyribonucleic acid

DT .....decision tree

ESCC .....Esophageal Squamous Cell Carcinoma

FB-ED.....forward backward early dropping

FDA .....Food and Drug Administration

FFPE .....formalin-fixed paraffin embedded

GAN.....generative adversarial networks

GDAN.....Genomic Data Analysis Network

GEA .....Gastroesophageal Adenocarcinoma

GEXP .....gene expression (referred to as type of molecular profile)
HCMI .....Human Cancers Model Initiative
HER2 .....a molecular subtype of BRCA
HLVS .....Hybrid Latent Variable Samples
HNSC.....Head and Neck Squamous Cell Carcinoma
HOTS.....hot-spot (type of mutation)
HPC .....high performance computing
HUGO.....Human Genome Organization
JADBio .....Just Add Data Bio, an auto-ML platform
KDx .....OHSU Knight Diagnostic Lab oncogenes
KIRCKICH .....Pan-Kidney Clear Cell and Chromophobe Carcinoma
KIRP .....Kidney Renal Papillary Cell Carcinoma
LGGGBM .....Glioblastoma and Lower Grade Glioma
LIHCCHOL .....Pan-Hepatic Cholangiocarcinoma & Hepatocellular Carcinoma
LOF.....loss-function (type of mutation)
LR .....logistic regression
LUAD .....Lung Adenocarcinoma
LUSC .....Lung Squamous Cell Carcinoma
MAD .....mean absolute deviation
MC3 .....Multi-Center Mutation Calling in Multiple Cancers
MESO .....Mesothelioma
METABRIC .....Molecular Taxonomy of Breast Cancer International Consortium
METH .....methylation (referred to as type of molecular profile)
ML .....machine learning
MMD .....maximum mean discrepancy
MPNST .....malignant peripheral nerve sheath tumor
mRNA.....messenger ribonucleic acid
MUTA.....mutation (referred to as type of molecular profile)
MXM .....Mens eX Machina (feature selection software)
NCI .....National Cancer Institute
nons.....non-synonymous (type of mutation)

OHSU .....Oregon Health and Science University

OV.....Ovarian Serous Cystadenocarcinoma

PAAD.....Pancreatic Adenocarcinoma

PAM50.....Prediction Analysis of Microarray 50

PCPG .....Pheochromocytoma and Paraganglioma

PNNL.....Pacific Northwest National Laboratory

PRAD.....Prostate Adenocarcinoma

SARC.....Sarcoma

SHAP .....SHapley Additive exPlanations

SKCM.....Skin Cutaneous Melanoma

SLURM .....Simple Linux Utility for Resource Management

SMOTE.....Synthetic Minority Over-sampling Technique

SVM..... support vector machine

SynTRUST .....Synthesis Study Trustworthy Test

RF .....random forest

RFE.....recursive feature elimination

RMSD.....root mean squared deviation

RNLVS.....Random Noise Latent Variable Samples

RNN.....recurrent neural network

TCGA.....The Cancer Genome Atlas

TGCT.....Testicular Germ Cell Tumors

THCA.....Thyroid Carcinoma

THYM .....Thymoma

TMP .....Tumor Molecular Pathology

TNM .....tumor, lymph node, metastasis

UCEC.....Uterine Corpus Endometrial Carcinoma

UMAP.....Uniform Manifold Approximation and Projection

UVM.....Uveal Melanoma

VAE.....variational auto encoder

WES.....whole exome sequencing

WGS .....whole genome sequencing

List of Figures, Tables, and Algorithms

Abstract

Molecular phenotypes, or subtypes, can describe cancer as distinct diseases within primary tissues-of-origin. Machine learning (ML) can be applied to this molecular taxonomy of cancer for classifying newly diagnosed samples in supporting clinical decision making and informing development of molecular therapeutics. The Cancer Genome Atlas (TCGA) provides high-dimensional genomic profiles for solid-tumor cancers where each sample is labeled with an intrinsic subtype. ML classifiers can be trained across tumor types to reveal cancer type-specific biology and models for use in trials and studies. Although gene-expression signatures are frequently used to delineate cancer subtypes, they are the downstream transcriptional effect of proteogenomic alterations, including somatic mutations in the exome. To facilitate analysis of the relationship between cancer transcriptomic states and their underlying mutation profiles, a feature selection method within a sub-sampling framework was developed to identify corresponding sets of mutated and expressed genes at subtype resolution. Rare subtypes within cancer molecular profile data present challenges due to limited statistical power. To address this, a variational autoencoder-based sample generation method was developed and evaluated to produce synthetic gene expression data with properties similar to the real training samples. Cancer model systems, such as cell lines and organoids, provide a means to obtain empirical data on responses of different cancer types to anti-cancer compounds. However, data from these model systems suffer from nested batch effects and have substantial differences from in-vivo conditions. We propose a phenotype-preserving latent feature representation to remove these effects and glean insight into cancer-specific biology. In sum, this dissertation demonstrates rationale for ML in cancer genomics and transcriptomics, application of ML in cancer subtyping, interpretation of cancer biology with ML, improving ML in cancer molecular profiling, and transferring knowledge between domains of cancer genomics and transcriptomics with ML.

Chapter 1— Introduction to multi-omic cancer subtyping applications of ML

Brian Karlberg

Publishing/permissions: NA

Thesis

Machine learning (ML) applied to human and model system cancer molecular profiles, in particular molecular subtyping, can advance precision medicine toward the clinic (Chapter 2), be used to interpret the molecular biology of cancer (Chapter 3), improve modeling performance via synthetic data generation (Chapter 4), and transfer knowledge between data platforms or model systems (Chapter 5).

Cancer incidence and epidemiology

Cancer incurs the highest burden of all human diseases as measured by Daily Adjusted Life Years[1]. This corresponds to a global incidence probability of 20% prior to age 75 with 10% chance of death[2]. Cancer is more prevalent in developed countries; this socioeconomic aspect of cancer can be quantified with a measure of Human Development Index[3]. Lung, breast, and prostate are among the most common cancers and incidence rate comparisons can be standardized by age and are typically reported in terms of sex-specificty[4]. In the United States, incidence has decreased for males while remaining constant for females while mortality has decreased for both males and females over the 15-year period from 1999 to 2014 [5]. These trends are attributed to advancements in earlier detection and more effective treatment.

Machine learning in precision oncology

In precision oncology, subtyping of cancer informs prognosis and therapeutic development; for example, mutation and expression profiles of individual tumors can be used to develop tailored molecular therapeutics[6]. Early work in the systematic prediction of phenotypic class based on molecular features was done for acute leukemias based on gene expression data[7]. Determination of minimum gene sets in prediction tasks subsequently emerged as a research goal to facilitate the development of diagnostic tools[8]. High-throughput sequencing technologies have led to the omics datasets that make ML applications possible across detection and diagnosis tasks[9]. Identification of parsimonious feature sets, that are gene-centric, advances clinical implementation of precision oncology[10]. Tissue-of-origin can be predicted based on mutation

profiles however driver mutations have been shown to not be the most performant feature type[11]. Additional considerations in the application of ML to cancer genomics are data-specificity of models, interpretability of features, and limited sample sizes[12]. Generative models, such as variational autoencoders (VAEs) can be adapted to produce synthetic data in augmenting traditional, interpretable ML[13].

Cancer as a disease of the genome, epigenome, transcriptome, and proteome

The American Joint Committee on Cancer (AJCC) has indicated precision molecular oncology as a complementary approach with which to evolve the utility of the traditional tumor, lymph node, metastasis (TNM) cancer patient classification[14]. High-throughput sequencing technologies have enabled comprehensive molecular characterization of cancer with The Cancer Genome Atlas (TCGA) project seeking to uncover the genomic roots of cancer via delineation in finer categorizations of molecular subtypes[15]. The TCGA culminated over ten years of pilot and production data generation to deliver the Pan-Cancer Atlas of molecular profiles for ~11,000 patient samples, 7 data types, and resulted in a new taxonomy of cancer that included subtypes defined with primary tissues of origin[16]. TCGA provides a basis for aggregated reports of cancer-type specific gene alterations and biological process signalling patterns[17] that can be used for comparison in subsequent studies. In sum, the TCGA data make it possible to interrogate the degree to which molecular profiles can inform clinical decision making, investigate interactions between genomic alterations and predictive transcriptomic features, and characterize the statistical power effects of limited sample sizes[18]. Additionally, programs such as the Clinical Proteomic Tumor Analysis Consortium[19] and the Human Cancer Models Initiative[20] provide further molecular data coverage in both humans and cancer model systems.

While somatic mutations affecting protein-coding regions of oncogenes and tumor suppressor genes are established drivers of tumorigenesis, they represent only one dimension of a complex landscape of molecular alterations. The oncogenic phenotype is frequently shaped by a confluence of events that extend beyond the exome. Epigenetic dysregulation, for instance, including aberrant DNA methylation patterns and histone modifications, can profoundly alter gene expression programs, leading to the silencing of critical tumor suppressor genes or the

ectopic activation of oncogenic pathways independent of direct sequence mutation. Furthermore, the functional non-coding genome plays a significant role; mutations in regulatory elements such as enhancers and promoters, or alterations in the expression of non-coding RNAs (e.g., miRNAs, lncRNAs), can disrupt entire gene networks and contribute to malignant transformation. These processes, coupled with post-translational modifications that modulate protein function and stability, create a multi-layered system of oncogenic inputs that collectively define the cellular state.

The transcriptome, as measured by mRNA abundance, serves as a critical nexus that integrates these diverse molecular inputs. A cell's transcriptional state is not a direct reflection of its genomic sequence alone but is rather the net output of its underlying genetic lesions, its dynamic epigenetic landscape, the status of intracellular signaling cascades, and post-transcriptional regulatory mechanisms. Consequently, a gene expression signature provides a functionally coherent and highly informative snapshot of the tumor's biological state. While the analysis of the mutational landscape is indispensable for identifying initiating events and potential targets for therapy, the transcriptional profile often yields a more powerful and robust signal for molecular subtyping. It effectively captures the integrated downstream consequences of myriad oncogenic processes, providing a more comprehensive basis for classifying tumors into clinically and biologically distinct subgroups.

Cancer molecular profiling technologies and challenges

Molecular omics technologies have been developed to measure mutations and methylation to DNA, gene expression levels, and other data types typically using bulk tissue samples[21]. Definition of molecular subtypes within primary tissue-of-origin cancers is made possible by the outputs of these data generating technologies sometimes in combination with traditional attributes of cancer such as histological features[22]. Technical artifacts may exist within these data due to different platforms used to measure the same biology as with gene expression measured by both RNA microarray[23] and the more recently developed RNA-seq[24]. Rescaling data between RNA-seq and microarray distributions is essential when re-applying trained models across studies or samples. Within sequencing technologies, variations in read length, reference genome

alignment, variant calling, and other technical challenges can result in non-biological noise in the data[25]. Challenges in the analysis of genomics and transcriptomics data include limitations of clustering techniques for prediction of new samples, risk of overfitting models, and the need for interpretation of biological function[26]. The curse of dimensionality is where the number features substantially outnumbers the number of samples and is common in biomedical datasets due to the diversity of data sources. In cancer molecular profiles and cell line experiments, there may be cancer types with rare subtypes or observations with 10 or less samples and tens of thousands of features. Overcoming this curse of dimensionality to locate biological signals within high dimensional genomic, epigenomic, and transcriptomic molecular profile data is a common theme in the development of clinical ML models.

Information theory in feature selection

Feature selection is a set of methods that aim to reduce the noisy and uninterpretable aspects of high-dimensional datasets such as those generated with genomic assays[27]. Feature selection addresses the curse of dimensionality by reducing the ratio of features to samples. Feature selection methods are generally binned into four categories: filter methods, wrapper methods, embedded methods, and hybrid methods[28,29]. Feature selection can address specific data challenges such as noise reduction, highlighting biological signals, and enhancing interpretability by removing irrelevant and redundant features such as co-expressed genes. In cancer, identification of mutation patterns of mutual exclusivity and finding subtype-specific mutation features is of interest in the context of characterizing functional relationships between genes[30]. Domain knowledge can integrate with computational feature selection techniques, for example in cancer molecular subtyping this could be running feature selection across a mixture of different data types or or within individual data types.

Entropy and mutual information are important information theory concepts relevant to feature selection and feature engineering[31]. Entropy is the minimum descriptive complexity of a random variable. In the context of cancer subtype feature selection, we aim to identify gene-centric features that minimize the entropy of the target variable, thereby maximizing the information content. Mutual information measures the amount of information shared between two random

variables. By maximizing mutual information between features and the target variable, highly predictive features can be identified.

Machine learning overview

Learning from data includes supervised and unsupervised methods[32]. Supervised learning is where a series of training cases, such as molecular profiles, with corresponding response measurements, such as tissue-of-origin or molecular subtype, are used to learn, or fit, a predictive model. The intention is to then apply the fitted model to make the same type of predictions on similar data that the model has not previously seen, a process termed generalization. Importantly, there is an implicit ground truth in supervised learning that the category labels have been assigned a priori as attributes of the training samples. In unsupervised learning, clusters, such as groups of tumor samples with similar genomic profiles, are inferred via patterns in the values of unlabeled molecular profiles. These clusters of samples can then be used to assign labels to samples for subsequent training of supervised models. Unsupervised methods include association rules, principal components, and clustering. Clustering methods can be combined i.e. clusters-of-clusters; this approach has been applied in designating molecular subtypes within breast cancer[33].

The diverse array of machine learning methods that have been developed for various types of data differ greatly in their underlying algorithms. To find a particular model effective for a given data set can thus require empirical search over model types, not just optimizing within a given model, when developing a particular ML task[34]. Types of ML methods can be divided into neural network and non-neural network-based methods with the former including canonical model types such as decision trees (DT), random forests (RF), logistic regression (LR), and support vector machines (SVM), among others[35]. Neural network ML models work by updating weight parameters on a set of interconnected nodes during training and are equivalently diverse in their various implementations. Canonical neural network architectures included recurrent neural networks (RNNs) for time series modeling and convolutional neural networks with image modeling capabilities[36]. Most neural networks consist of more than one hidden layer of nodes thus are considered to be a deep learning (DL) architecture. DL models can be categorized as

discriminative or generative and within generative DL models are a class known as autoencoders that possess the genomically-useful property of learning dimensionally reduced, or latent, representations of data. These latent factors capture the essence of data distributions by means of a representative minimal set of composite features that capture most of the variability while reducing the effect of unnecessary or redundant features[37].

In ML, the class-imbalance problem in labeled data is an important problem with implications for how machine learning results are evaluated[38]. When the ratio of samples in the majority class substantially outnumbers the number of samples in minority classes (or minority class in the case of binary data where there are only two class labels), a given model may inadvertently mis-predict those minority samples disproportionately with respect to the majority class in optimizing its accuracy during training. F1 score, or the harmonic mean of precision and recall, is a model evaluation metric that protects minority classes from being exploited to boost overall accuracy[39]. F1 is preferred over other metrics, such as balanced accuracy, from a clinical perspective of molecular cancer subtyping due to more stringent treatment of false positive diagnosis and false negative diagnosis of rare variants — both outcomes that result in substantial consequences for patients.

Patient impact

The intention of this research is to translate these computational advancements into tangible benefits for patients. The frameworks developed here provide a foundation for a more personalized approach to cancer care. By providing the means for molecular subtype classifications, this work aids clinicians in selecting the most effective existing therapies and offers patients a clearer prognosis and understanding of their tumor's likely progression. Looking forward, the ability to correct for differences between preclinical models and human tumors is an essential step toward predicting which drugs will be effective for an individual patient. Furthermore, by enabling the generation of synthetic data for rare cancer variants, these methods provide the potential to develop better predictive models for patients. In sum, this dissertation provides foundational tools and insights that can be used for guiding therapy, improving prognostication, and ultimately prolonging patients' lives.

This introduction chapter has reviewed the salient genomic and transcriptomic concepts of cancer as well as machine learning concepts in the context of molecular subtype prediction with an emphasis on the inherent challenges of molecular profiling data. Associated concepts in information theory, feature selection, and model evaluation have also been introduced providing the context for the succeeding chapters.

Concept map

Chapter 2, Implement machine learning — predict TCGA subtype of novel tumors

- Label assignment methods not suitable for class prediction; clinical motivation
- Classifier development process, 26 TCGA tumor types with 106 subtypes
- Transferability of trained models, validate TCGA-trained models with external cohorts
- Inform further collection of data, disentangle biological signal from statistical power
- Library of classifiers and feature sets, resources for trials and studies

Chapter 3, Interpret machine learning — somatic mutations in oncogenes vs mRNA signatures

- Apply TMP findings, sub-sampling threshold and feature selection methods
- Benchmark feature selection, importance, and onco-screening methods
- Subtype-specific mutation-expression interaction networks
- Data type predictive signal comparison: mutation, methylation, and gene expression
- Expression signature performance on cancer samples stratified by detected onco-status

Chapter 4, Improve machine learning — synthetic sample generation

- Review ML applications in biology, genomics, cancer subtyping
- Review DL, transfer learning, latent representations of data
- Review synthetic data concepts and methods
- VAE transfer learning; pretrain and finetune model to prepare encoder and decoder
- Latent variable recombination methods for synthetic sample generation
- Evaluation methods to quantify the quality of synthetic data

Chapter 5, Extend machine learning — cancer model systems batch correction

- Cancer model systems background
- Concepts of nested batch effects, structuring data to be machine learning ready
- Develop and test evaluation frameworks and batch correction methods

Chapter 6, Conclusions, overall scientific contributions, and individual contributions

- Conclusions specific to each of the results chapters — 2 through 5
- Summary-level description of contributions to science of these experiments
- Chapter-level breakdown of individual contributions to presented herein

Through these concept reviews and experimental results, this dissertation demonstrates how machine learning can address fundamental challenges in cancer omics including clinical implementation of subtyping for diagnosis and guidance of treatment, generalization of findings to other studies, building subtype-specific gene interaction networks, demonstrating the utility of expression signatures in the context of mutational state, synthetic data generation of molecular profile samples, and approaches to model system correction and evaluation pipelines. The interdisciplinary nature of modern scientific publishing afforded the opportunity to practice simultaneous collaboration with individual initiative and contribution. Cancer molecular subtyping is important because it provides patients and clinicians with both a more-accurate prognosis and better-informed decisions in treatment. Additionally, ML applied to subtyping can help with connecting molecular features that are specific to different types of cancer with other data such as gene-interaction knowledge to advance understanding of the underlying biology of the disease.

Chapter 2 — Classification of tumors to TCGA molecular subtypes, model validation, and guidance of further data collection

Kyle Ellrott, Christopher K. Wong, Christina Yau, Mauro A.A. Castro, Jordan A. Lee, Brian J. Karlberg, Jasleen K. Grewal, Vincenzo Lagani, Bahar Tercan, Verena Friedl, Toshinori Hinoue, Vladislav Uzunangelov, Lindsay Westlake, Xavier Loinaz, Ina Felau, Peggy I. Wang, Anab Kemal, Samantha J. Caesar-Johnson, Ilya Shmulevich, Alexander J. Lazar, Ioannis Tsamardinos, Katherine A. Hoadley, The Cancer Genome Atlas Analysis Network, A. Gordon Robertson, Theo A. Knijnenburg, Christopher C. Benz, Joshua M. Stuart, Jean C. Zenklusen, Andrew D. Cherniack, Peter W. Laird

Content adapted from: “Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets”, Cancer Cell, 2025

Abstract

Molecular subtypes have been defined in the The Cancer Genome Atlas (TCGA) and delineate a cancer's underlying biology. Subtype discovery methods such as unsupervised clustering and histology are not sufficient for classification of new undocumented samples in a clinical setting. To address these challenges, five machine learning methods were explored to identify classifiers and compact feature sets specific to primary tumor types. These feature sets were derived from 5 data types of gene-centric genomic, epigenomic, and transcriptomic molecular profiles and were often not the same type of features used to define the subtype categories. The most performant models frequently selected expression features over the other four data types. External validation of classifier subtype prediction concordance was conducted for one cancer type, breast invasive carcinoma (BRCA). Biological distinctions between subtypes can be determined via comparison of classifier-selected features with signaling pathways. Sample-size effect modeling allows for determination of ultimate cohort prediction performance with additional data of the same type. The models are accessible in Docker containers as a public resource for non-TCGA patient sample classification. In sum, this work is an interdisciplinary approach toward ML-driven clinical translation of 'omics data.

Introduction

The traditional basis for classifying cancers has been anatomic site or organ of origination along with AJCC/UICC TNM stage, morphologic grade, and histological features[40]. Subtyping of cancer informs prognosis and guides therapeutic decision making. A classification scheme of cancer based upon tissue-of-origin is substantiated by genomic studies where distinct genomic and transcriptomic biology is associated with anatomic site[41–43]. The Cancer Genome Atlas (TCGA) provides a database of cancer genome and transcriptome profiles; data to support molecular subtyping was an originating goal of the data collection and aggregation project. The project originally stemmed from a collaboration between the National Cancer Institute and the National Human Genome Research Institute. The concept of distinct transcriptomics, epigenomics, and genomics in defining cancer subtypes is supported by previous studies[44–50]. However, methods that were used for assigning labels to training samples are not sufficient for

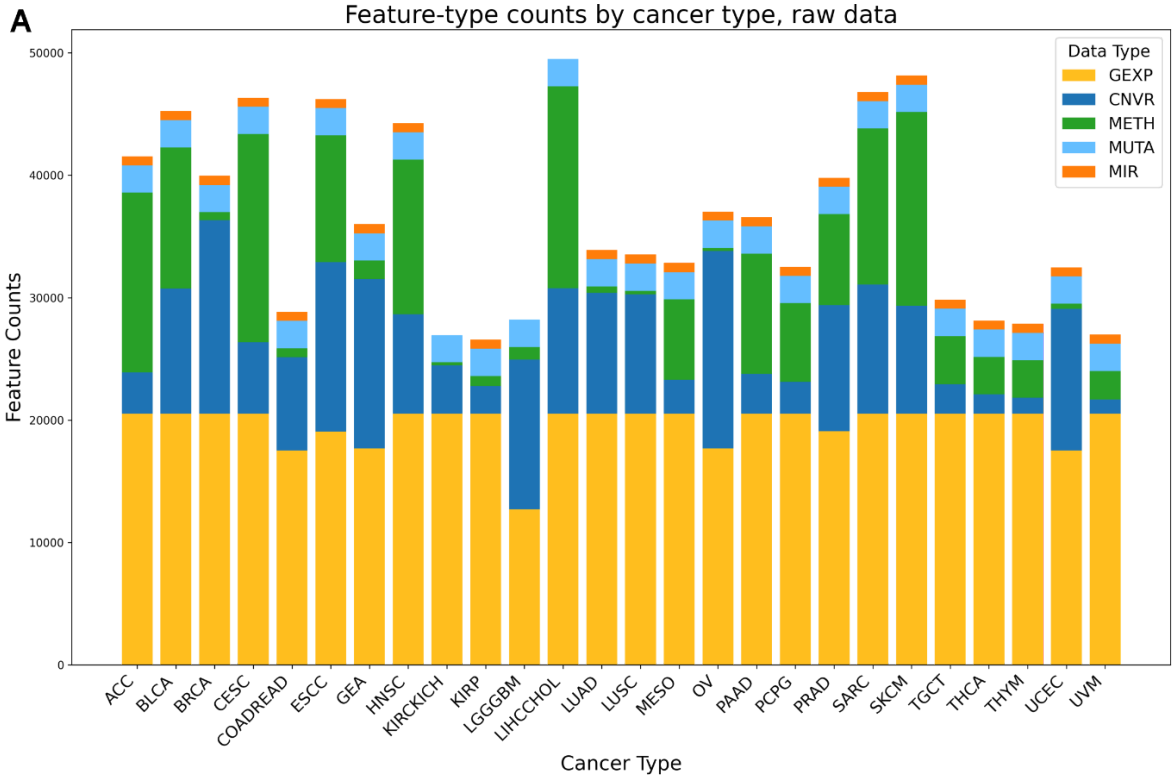
predicting the subtype category for a newly diagnosed sample in a clinical implementation. For genome studies and clinical trials to leverage TCGA cancer types, models must be capable of prediction of newly diagnosed samples using features amenable to a pre-trained classifier that leverages a previously defined classification scheme. In this study, a gene-centric TCGA dataset for 8,791 tumors across 26 primary tumor types and 106 subtypes with 5 data types of features was prepared and machine learning classifiers spanning 5 ML methods were trained and cross-validated with multiple feature selection methods to produce a library of 412,585 models. From these results, a set of 737 top models by cancer type and data type were identified and made publicly available as a resource. Future work to incorporate the availability of additional 'omics may result in refinement of the subtype label assignments to existing and new samples and new approaches to clinical screening panel development whereas the underlying direction of the field of precision medicine becoming more molecularly driven will remain constant. Ultimately, each patient's tumor is unique to that person and represents an n-of-1 subtype with a unique response to therapy. This work of extending coarse definitions of cancer types from tissue-of origin to molecular bridges toward that ideal of precision medicine. Lessons learned from the present work include data type performance, transfer learning capabilities, guidance of further data collection via sample size effect analysis, and pathway biology distinctions of subtypes.

Results

Tumor subtypes defined and classifier models developed

Molecular subtypes were defined within primary tumor types following the cancer-type-centric approach to patient care based on histopathology and anatomic location. Some TCGA cohorts like COADREAD and KIRCKICH were merged based on inter-cancer subtype overlap. Dataset preparation included retrieval of subtypes from the PanCancer Atlas resources and assembly of the molecular profiles. The five data types comprising the aggregated molecular profile within each cancer were DNA mutation, RNA-seq, DNA methylation, micro-RNA, and copy number variation. Feature and sample characteristics of the data are shown in Fig 2-1, Panels A and B, respectively. The result of preparing ML-ready subtype prediction data was 26 uniformly

formatted molecular profiles where each mapped to an anatomical location on the human body and included a corresponding cross validation file.


A
Feature-type counts by cancer type, raw data
Data Type
GEXP
CNVR
METH
MUTA
MIR
Feature Counts
0
10000
20000
30000
40000
50000
ACC
BLCA
BRCA
CESC
COADREAD
ESCC
GEA
HNSC
KIRCKICH
KIRP
LGGGBM
LIHCCHOL
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
SARC
SKCM
TGCT
THCA
THYM
UCEC
UVM
Cancer Type

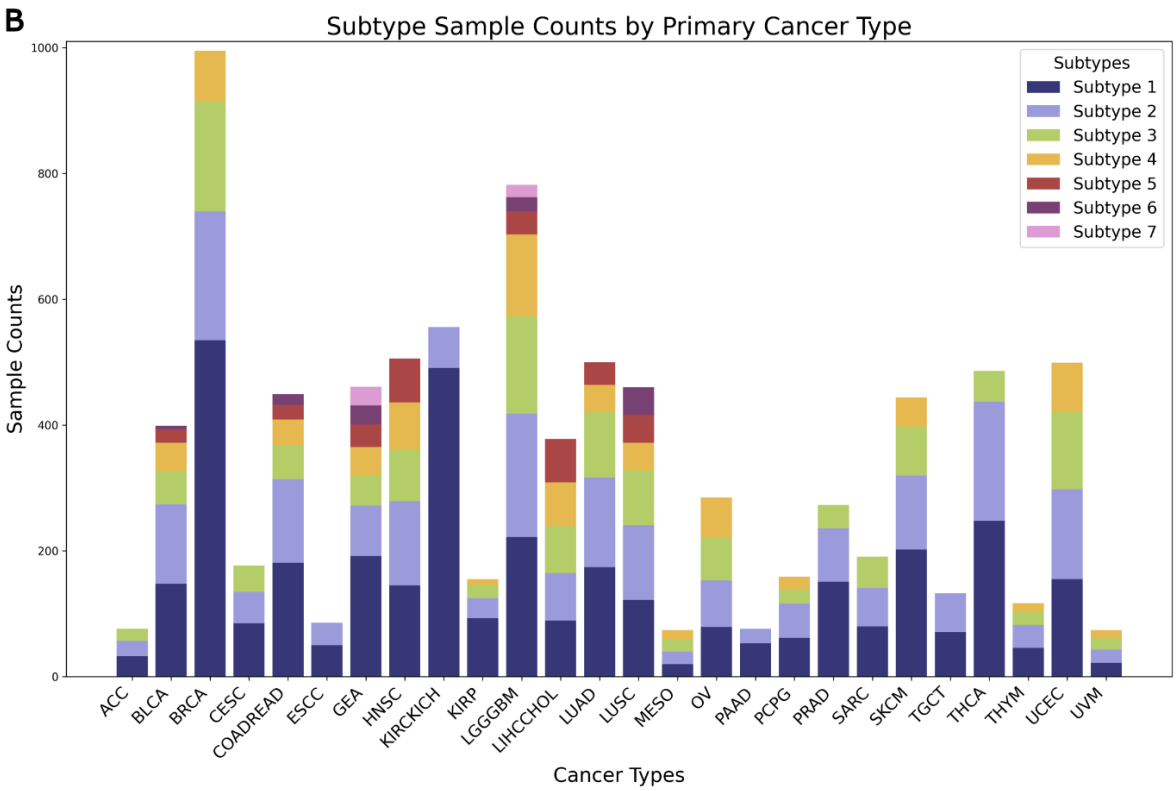

B
Subtype Sample Counts by Primary Cancer Type
Subtypes
Subtype 1
Subtype 2
Subtype 3
Subtype 4
Subtype 5
Subtype 6
Subtype 7
Sample Counts
0
200
400
600
800
1000
ACC
BLCA
BRCA
CESC
COADREAD
ESCC
GEA
HNSC
KIRCKICH
KIRP
LGGGBM
LIHCCHOL
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
SARC
SKCM
TGCT
THCA
THYM
UCEC
UVM
Cancer Types

Fig. 2-1 TMP subtype feature distributions and subtype-level sample distributions. **A** Feature counts by data type within each primary tumor type. Total raw feature counts on the order of tens of thousands per primary tumor type. Expression features comprised approximately half of the total features per tumor type. **B** Sample counts by subtype within primary tumor types. The number of subtype classes varies from two to seven per primary tumor type with the number of samples per class asymmetrically distributed on the order of tens to hundreds of samples per class

The five classifier methods utilized for subtype prediction model development were Subscope[51], Scikit Grid[52], AKLIMATE[53], Cloud Forest[54], and JADBio[55]. Subscope was the neural network-based model, AKLIMATE incorporated prior information on biological pathways, JADBio was an auto-ML technique, Cloud Forest was a modified random forest method, and Scikit Grid searched over ML models in the scikit-learn library. Training of each of the five classifier methods was conducted over each of the 26 cancer types and using individual and combined datatypes within each cancer type. These combinatorial effects in conjunction with Scikit Grid and JADBio both containing embedded method versions resulted in the large number, 412,585, of models as the aggregate experimental output. To account for the class-imbalance nature of the data, all of the raw predictions produced by this developed model set were scored with the F1 measure[38,56] for subtype prediction performance. These F1 scoring results were aggregated into a unified results matrix to facilitate inter-cancer comparisons.

Four primary observations were identified in the analysis of prediction performance, the data used to originally define the subtypes, and the feature sets selected by the models. First, in cohorts where subtypes were originally defined by multiomics or histology generally yielded performant classifiers. Second, the mRNA datatype for model-selected features predominated among the higher performing models. Third, for cancers defined by mutation or methylation datatypes, such as SKCM and LGGGBM, model-selected data types tended to match the defining data type. Fourth, subtypes defined with genome-wide features i.e. mutation load, chromosome instability, or CpG island methylator phenotype such as GEA and COADREAD - were difficult for gene-centric the classification approach and lower performance was observed in these cases. These four observations are summarized with a comparative plot of F1prediction performance and datatypes as shown in Fig. 3.

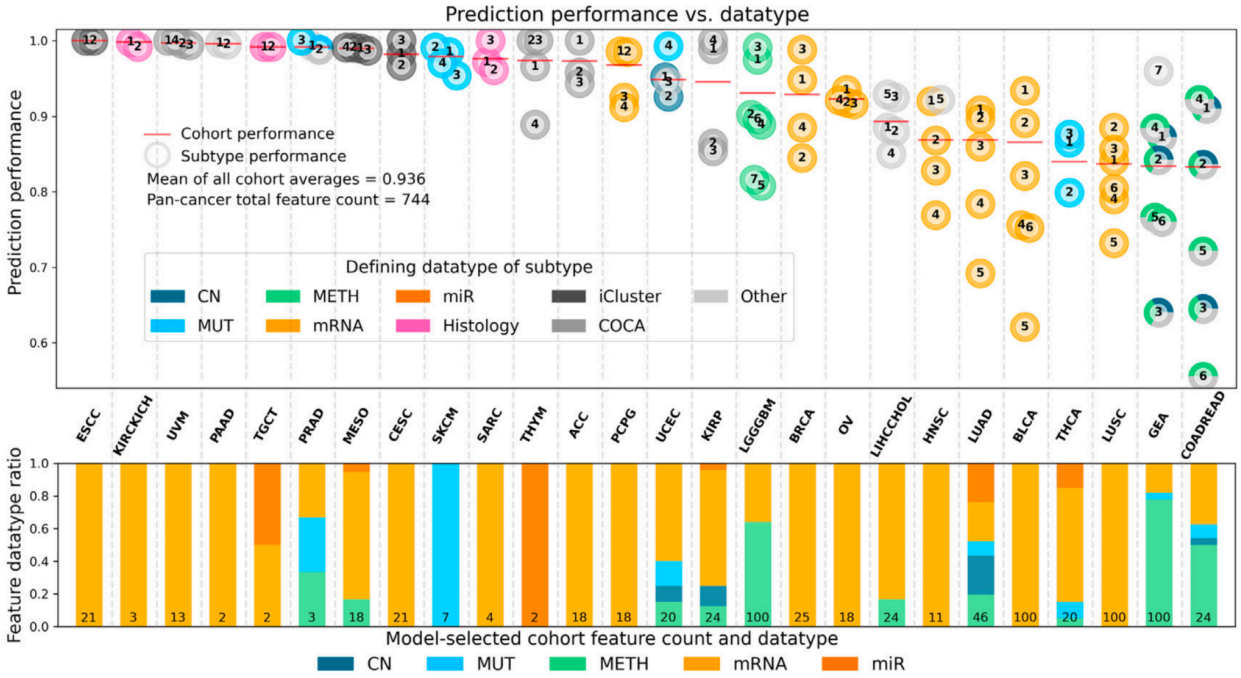


Fig. 2-2 Classifier performance metrics, defining datatypes, and model-selected features. Classifier subtype prediction performance, quantified with F1 scores, for highest performing identified model within each of the 26 tumor types is shown at the top of the figure. Mean overall weighted F1 score for each primary tumor type is shown with a horizontal red bar. Individual subtype performance within each primary tumor type is plotted as round markers, numbered by subtype, and colored by the data type(s) used originally to define that subtype. The stacked bars at the base of the figure shows the proportion of model-selected feature-set data types for the top model in each cohort. Printed at the base of each cancer's stacked bar is the number of gene-based features comprising the set identified by that cohort's most-performant subtype classifier

PAM50 predictions recapitulated in external validation

To assess the generalizability of our classifier development method, we tested whether two of our TCGA-trained BRCA gene expression classifiers could recapitulate PAM50 subtype predictions in external data sets. The AKLIMATE and Scikit-grid methods were each tested in two external data sets - METABRIC[57] and AURORA[58]. The METABRIC cohort used microarray gene expression measurements whereas our TCGA training data was based on RNA-seq data.

The AURORA dataset samples were prepared using an alternate FFPE (formalin-fixed paraffin-embedded) method compared to our fresh frozen TCGA sample prep.

In the METABRIC validation, the BRCA subtype prediction performance was conducted in the context of sample-level silhouette scores[59]. Briefly, silhouette scores are the similarity of each sample to its own subtype distribution vs the next-closest subtype distribution. Positive silhouette scores indicate samples that are more similar to their own subtype class and negative for more similar to another class. Both the AKLIMATE and Scikit-grid models were first trained on the full TCGA BRCA sample set then used to predict each sample in the METABRIC set; prediction results are shown as aggregate bars in the Sankey diagram shown in Fig. 2-3. Highly concordant calls were observed for both models with the majority of calls being concordant within each model-subtype combination except for HER2 with Scikit-grid. Both models showed similar behavior in discordant calls for LumB as LubA with similar proportions of samples. Scikit-grid miscalled HER2 as Basal more frequently than the AKLIMATE model. The central horizontal bars in Fig. 2-3 depict the silhouette scores for each sample. Samples with lower Silhouette similarities were more likely to be called discordantly.

The greater GDAN-TMP experiment included a multi-class pairs analysis and sub-experiments on sub-setting of samples by silhouette, that was work beyond the scope covered here and those results are detailed in our team's publication[10]; Ellrott (2025).

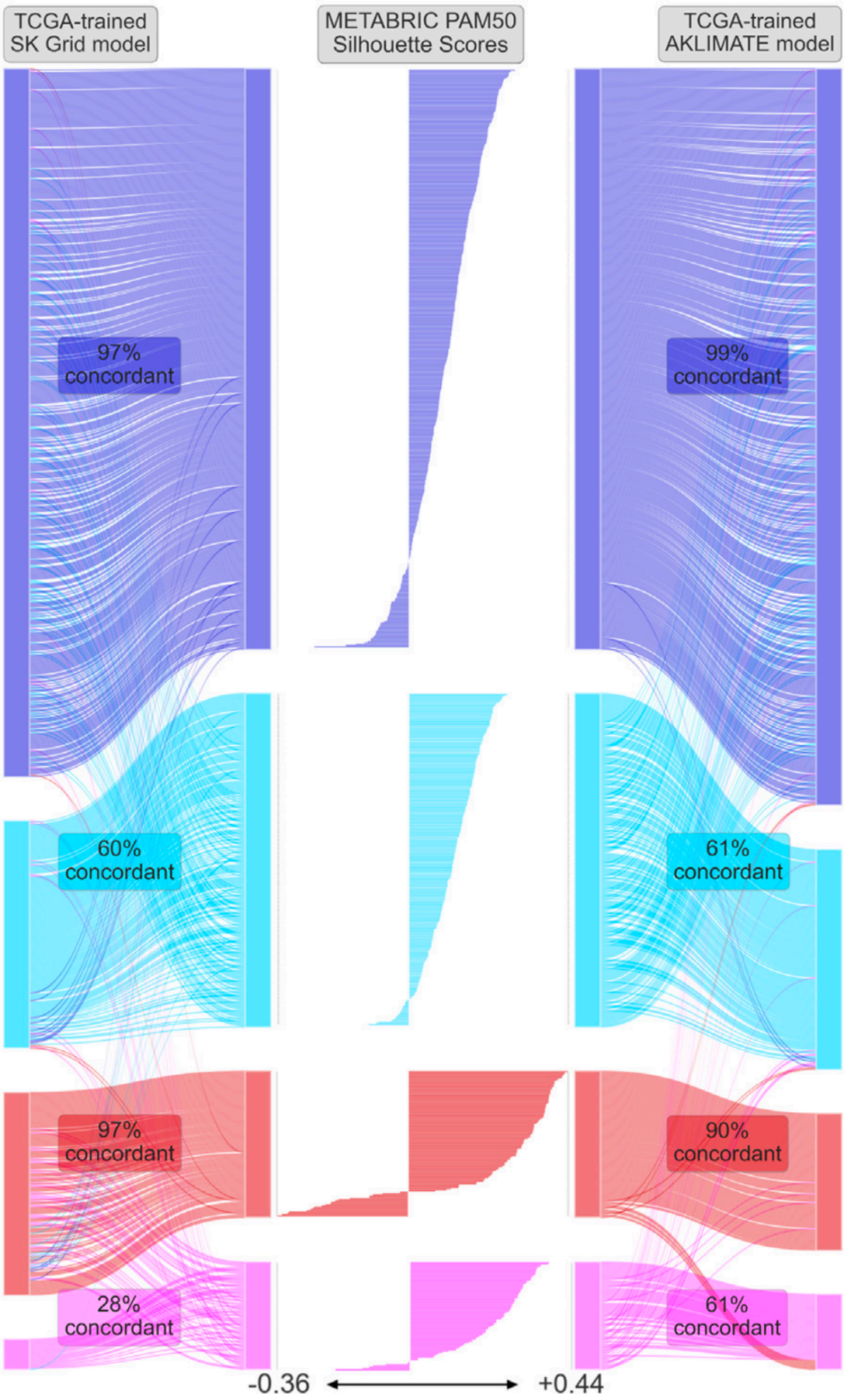


Fig. 2-3 Sample-level concordance analysis of model validation on external cohort for BRCA. Original METABRIC PAM50 calls to SK Grid (left) and AKLIMATE (right) classifications. Center horizontal bars

represent silhouette scores — the similarity of each sample to its own subtype distribution vs next-closest subtype distribution. Positive silhouette scores indicate samples more similar to their own subtype class and negative for samples more similar to another class

The results of the METABRIC external validation are summarized in a Venn diagram with Fig. 2-4.

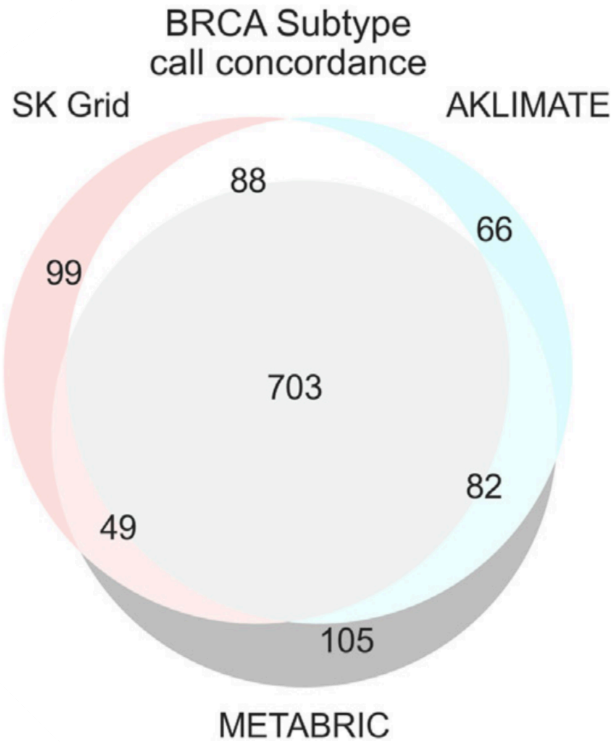


Fig. 2-4 Venn diagram illustrating agreement and disagreement of the classifiers on the METABRIC cohort. Intersection and relative compliments of concordant and discordant predictions for METABRIC as ground truth labels against calls of TCGA-trained classification models

A correlation between classifier confidence and Silhouette score was hypothesised. To test this, we leveraged the confidence scores emitted from the AKLIMATE model. The calls that are more confident would be expected to have a greater difference in confidence between the first and second calls. Conversely, the less confident calls would have a relatively lower difference in confidence between first and second calls. We observed this correlation to hold to a degree that varied by subtype as shown in Fig 2-5. Luminal A (Spearman Rho = 0.69, p =53 1067) and basallike (Spearman Rho = 0.60, p =1012) subtypes exemplified this relationship.

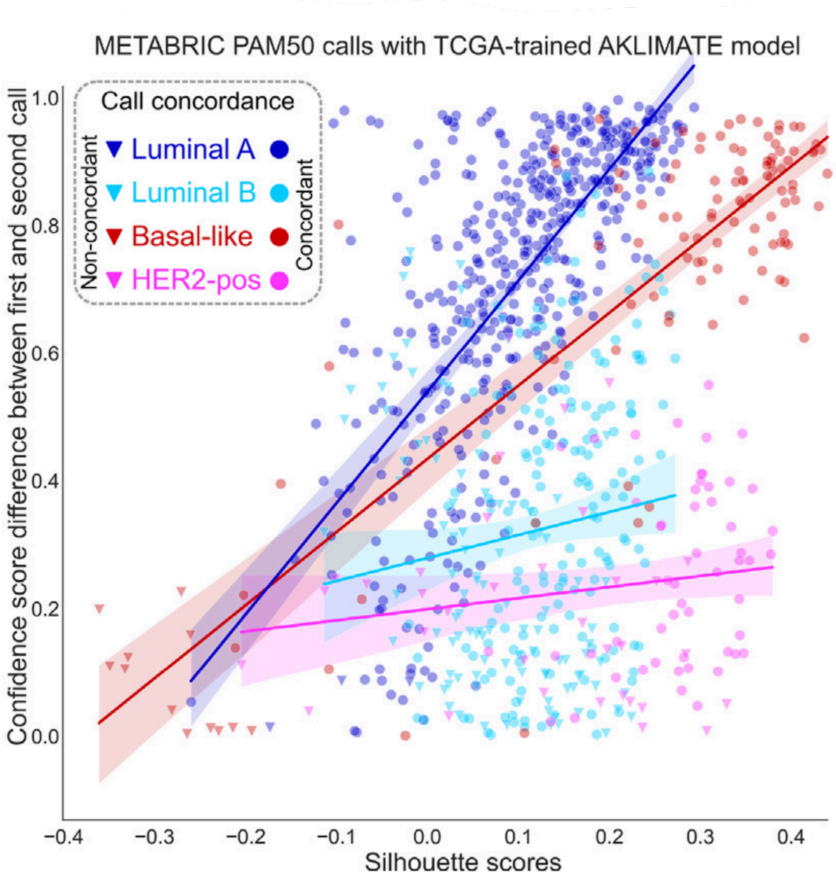


Fig. 2-5 Sample silhouette scores vs. classifier confidence. The confidence score is defined by the difference in confidence between the best and second-best sample prediction confidence calls for AKLIMATE. Circles indicate samples with concordant calls and triangles indicate samples with discordant calls. Linear regression trend lines for each subtype with 95% confidence intervals

Similar validation results were obtained for our TCGA-trained classifiers on the external Aurora cohort where a quantile rescaling to correct for distributional differences between FFPE and fresh-frozen[60] TCGA training samples was critical.

Number of samples needed to train classifiers

We applied our feature sets and classifiers to predict the number samples needed to achieve adequate prediction performance in a given tumor type. This extrapolation of performance can be affected by the type of classifier used, the feature sets used, and the fidelity of the original labels used for training[61–63]. In alignment with Figure 2-2, the main performance and data type comparison figure, discrepancies in individual cohort performances can be attributed to differences in the recapitulation nature of this work where supervised models here are attempting

to predict subtype labels based on specifically curated gene-centric features whereas the labels may have been assigned in the previous project with other gene-centric features or genome-wide feature not present in these data. A power-law function was fit to prediction scores made with subsets of samples within cancer types. The same general trend in power law curves captures the behavior of learning curves held across cancer types, Fig 2-6. Predictive performance was observed to plateau at around 150 samples for cohorts with at least that many samples.

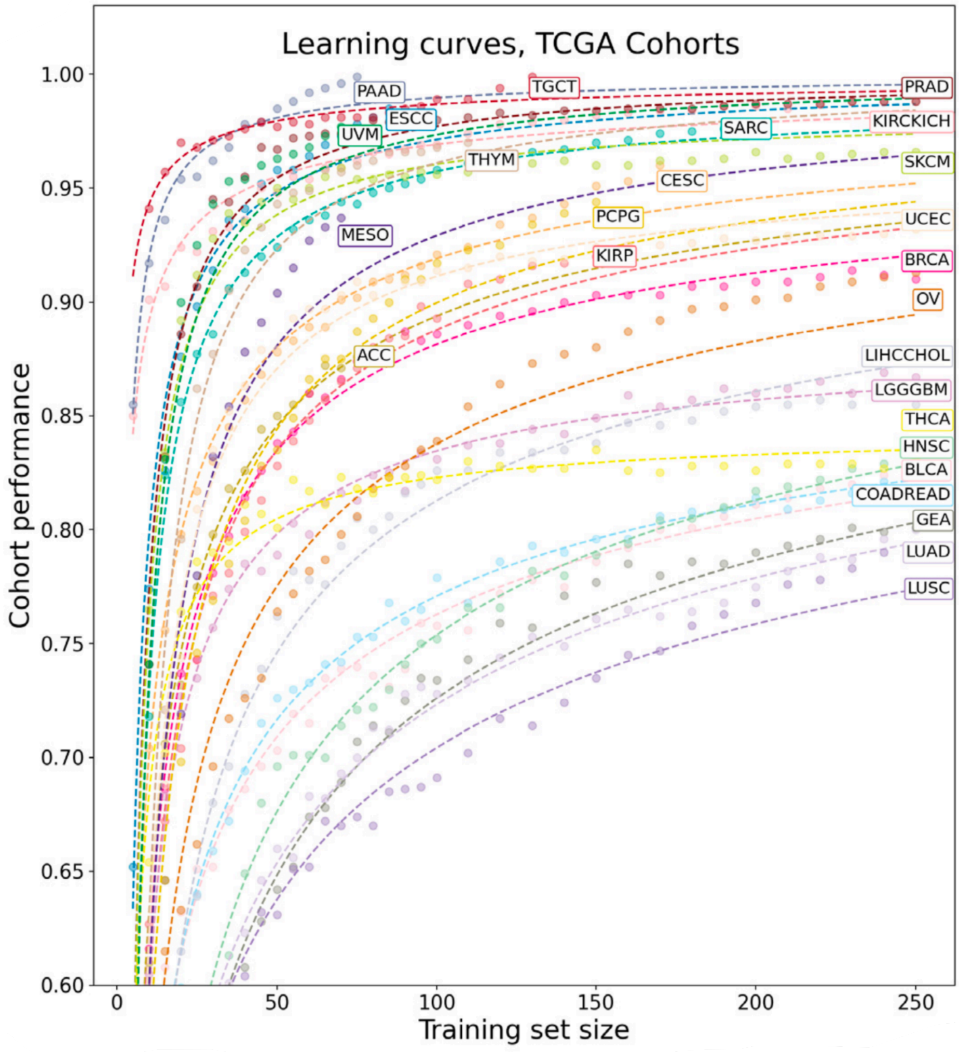


Fig. 2-6 Power law curves fit to subtype prediction performance as a function of sample size. Sub-sampling was repeated 100 times with corresponding performance averaged at each sample size increment

A method was developed to fit a Burr statistical distribution to a set of 100 power law function projections each fit to a sub-sampling derived learning curve in the range of 35 to 70 samples.

The method was developed on the cancer types with at least 250 samples by optimizing against the mean squared error over the 15 cohorts. These results are presented as actual vs. projected performance in Fig. 2-7.

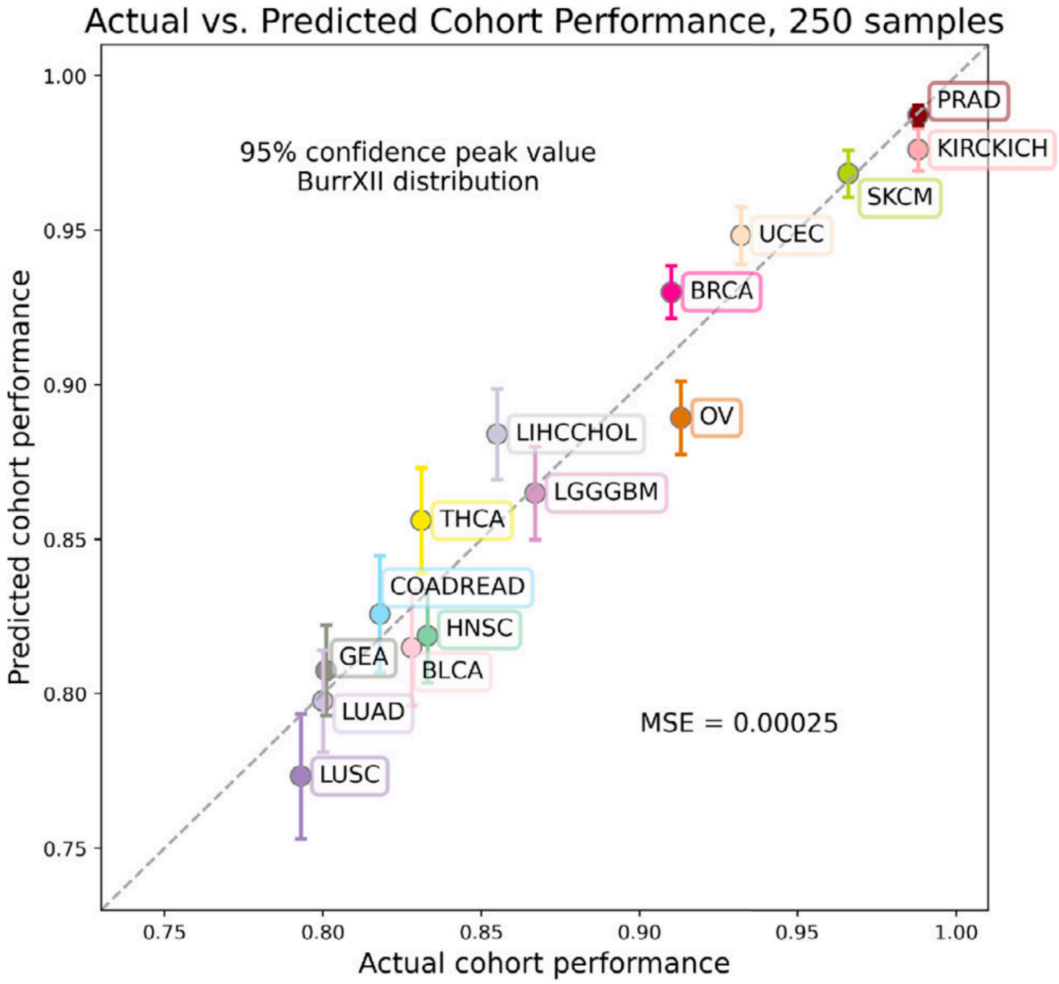


Fig. 2-7 Predicted vs actual subtype classification score for 15 cancer cohorts with at least 250 samples. Performance projections based on fitting power law curve to sub-sampling range of 35 to 70 samples in increments of 5

The performance extrapolation method was applied to adrenocortical carcinoma to model an approximate doubling of the cohort's sample count to 150 samples. This prediction performance extrapolation for adrenocortical carcinoma is shown in Fig. 2-8.

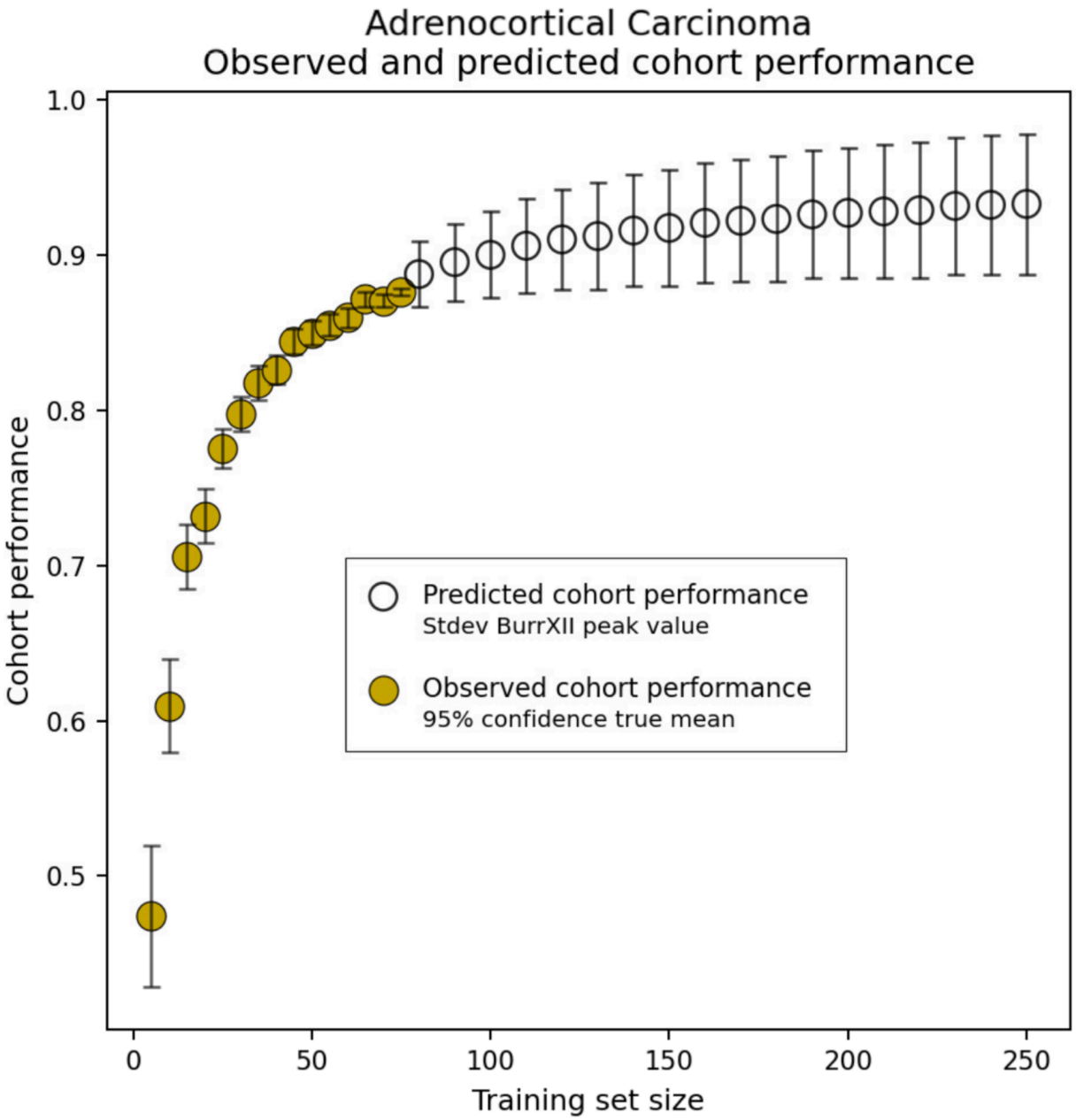


Fig. 2-8 Adrenocortical carcinoma performance projection to 250 samples. Projections with error based on repeated fitting of power law curve within 35 to 70 sample range sub-sampling framework

Common features, tumor biology, and pathways

Reproducibility in feature set selection can be difficult within a given ML method[64] especially in the case of redundant features such as coregulated genes. Given this, a set of core genes based on

features that were selected by two or more methods were identified within cancer type. Biological themes were evident in these gene-sets such as BRCA oncogenes ESR1 and FOXC1 and COADREAD and SKCM feature data types matching their respective defining data types, methylation and mutation.

The presence of equivalently predictive genes in the feature sets of our subtype prediction ML models potentially resulted from co-membership in biological signal pathways in a comparison with Pathway Commons[65]. The oncogene status of the identified feature sets contained genes both associated and unassociated with the COSMIC oncogene database.

Discussion

Model interpretability was an intent of the gene-centric dataset design. The core concept of this study was that certain methods for label assignment, such as unsupervised clustering on genome-wide features like chromosome or microsatellite instability, are not sufficient for predicting the subtype label of a newly diagnosed sample. By identifying cancer type-specific gene-centric feature sets with corresponding classifiers, this study provides a basis for further interpretation of biological distinctions between cancer in other studies as well as the basis for development of clinical screening panels that would rely on cancer type-specific genes that probabilistically define molecular subtypes within those primary tumor types.

Classifier prediction performance was evaluated across 26 TCGA primary types comprising 106 tumor subtypes. With a model defined as the combination of a classifier and a feature set, a model specific to each of the 26 primary tumor types was identified. For more than half the tumor types, model performance exceeded 90% and for the remaining tumor types, model performance exceeded 80%. Measuring performance with the F1 score provided a more realistic assessment of the classifiers given the class-imbalances of the data.

The relationship between the features used to define the subtypes and the features selected by the classifiers as predictive of subtype, as shown in Figure 3, was one of the main overall learnings of the study. The most frequent data type selected was gene expression. This indicates that for

gene-centric measurements, transcriptomic features are generally the most information-efficient data type with which to construct minimal feature sets.

Generalizability in machine learning is a fundamental concern. An external validation was conducted with the METABRIC cohort. Two types of models - Scikit Grid and AKLIMATE - from our study recapitulated the four main BRCA subtype labels with remarkable concordance. The silhouette score of how well a given training sample fits within its assigned class can be an indicator of how well that sample will be predicted. The validation experiment on METABRIC shows a relatively high degree of concordance in the transferability of our models. Future work could investigate the poor outcome observed in HER2 with the SK Grid model via inclusion of additional feature sets, additional confidence and distribution comparison metrics, and direct inclusion of the original BRCA subtype definition methods to evaluate the fidelity of the label assignments taken as the ground truth.

In considering how our library of models could be implemented in another study or in a clinical trial whether data is pre-existing and the type of data available drives the decision flow.

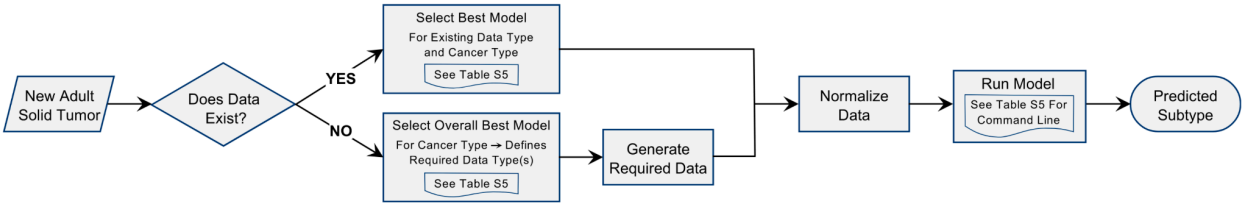


Fig 2-9 Workflow to select model for predicting subtype status of a new sample. The decision criteria is 'omics data existence and type of new sample. TMP project Table S5 available via the supplemental information of that resource[10]

In a study setting where an understanding of statistical power vs. predictive signal is required, a learning curve analysis on the existing sample set can yield insight in the relationship between sample size and predictive performance. Secondarily, a projection of prediction performance improvement can be cast from the learning curve to inform decisions on data collection.

In the full GDAN TMP project results, it was observed that the redundancy of biological pathway membership of classifier-selected features yields confidence that these features that were identified from multiple methods are of biological relevance to their respective cancer types. Interchangeability of features within different feature sets of similar predictive power was explored; these results are available for review in the full project manuscript.

Conclusions

By determining a combination of feature set and classifier specific to primary tumor types, ML models can be developed that rely on minimally-sized feature sets for efficient prediction of the molecular subtype of a given sample. This capability scaled across TCGA solid tumors brings molecular medicine closer to clinical implementation and provides a basis for further inquiry into the molecular biology of cancer subtypes. Standardized cross validation and performance scoring appropriate for multi-class predictions can enable comparisons across diverse feature selection and classification methods to support machine learning model development programs. Transfer learning for models and features is possible between studies if care is taken in rescaling data as seen with the microarray to RNA-seq BRCA experiment. The constraint of limited sample sizes in cancer genomics data can be characterized with learning curves. These learning curves can also be used to predict ultimate predictive performance upon addition of more samples with the same molecular profiling.

A potential limitation of our method is in the case of a new or undocumented subtype, the capability of a trained classifier is constrained by the validity of the original subtype definitions. Compared with ensemble methods, which would require feature sets with potentially little or no overlap for each of the constituent models, our method of model building via classifier-feature set search preserves the parsimonious feature set goal. Although our dataset, with less than 10,000 samples, is far too small for training a large language model, the results of our study could potentially be used for fine-tuning of large models.

Methods

This work on model development and evaluation was part of a larger project conducted by the Tumor Molecular Pathology (TMP) Analysis Working Group of the Genomic Data Analysis Network (GDAN). Informed consent of samples used was covered under the local Institutional Review Boards of the TCGA Research Network.

Gene-centric molecular profiles for 26 primary tumor types spanning 106 subtypes were obtained from the project's controlled Synapse repository[66]. Five data types comprised these profiles: mRNA-seq, DNA mutations, DNA methylation, copy number variation, and micro-RNA. These data were tabular in format and contained no missing values. So that cross validation was applied consistently across the five machine learning modeling methods, each cancer had a pre-designated 100-repeat, 5-fold cross validation file in addition to its molecular profile file. One cancer type, LIHCHOL, required a rebuild of its molecular profile and corresponding cross-validation file due to post-hoc changes in the included samples. To rebuild these project files, raw files were obtained from each of the five data sources and processed into a unified format then concatenated to the final .tsv project file format. To remove missing values, an iterative process of removing 20% of samples with missing values alternated with removing 20% of features with missing values was applied. The 100-repeat, 5-fold cross validation file creation method was reverse engineered from the other cancer types and implemented with scikit-learn train-test split. The completed set of subtype-balanced repeated cross-validation folds utilized as training and testing sets for model development.

Of the 5 total ML methods utilized in this study, Scikit Grid was the one developed at OHSU by the Ellrott Lab and is the focus here. The Scikit Grid models that were developed for this study consisted of a feature set and corresponding scikit-learn classifier. The feature sets for this work were prepared by Jordan Lee of the Ellrott Lab with two methods - RFE within Scikit-learn and the forward-backward early dropping algorithm with the R package MXM. The classifiers utilized were: Adaboost, Bernoulli Naive Bayes, Decision Tree, Extra Trees, Gaussian Naive Bayes, Gaussian Process, K Nearest Neighbors, Logistic Regression, Multi-layer Perceptron, Multinomial Naive Bayes, Passive Aggressive, Random Forest, Stochastic Gradient Descent, and Support Vector Machine.

To develop each grid of hyperparameter configurations for the Scikit-learn classifiers, common ranges of hyperparameter settings were determined via online resources for each specific hyperparameter, such as learning rate or number of estimators, within each specific model. A survey of common hyperparameter settings specific to each of the scikit-learn classifiers and from this a config file was created to support deployment of models and hyperparameter combinations over the feature sets via OHSU's cluster, Exacloud. The Scikit-grid component of the experiment was deployed in conjunction with GDAN-TMP co-author Kyle Ellrott.

Crisp, or individual sample predictions within each repeat cross-fold, prediction results were scored with Scikit-learn's function F1 score:

$$sklearn.metrics.f1\_score()$$

This ML-performance metric is defined as the harmonic mean of precision and recall and functions to address the class imbalance characteristic of these data as outlined in the introduction. Prediction results aggregated with classifier-selected features were sorted by highest prediction score, lowest prediction standard deviation, then lowest feature count. These results were aggregated with subtype-defining data integrated to produce Fig. 3.

For the external validation on BRCA, raw instrument microarray gene expression data were obtained from the METABRIC experiments[57]. Probe values were averaged by gene and distributionally-scaled to the TCGA mRNA data. Sample silhouette scores[59] were first calculated for the METABRIC cohort. Then AKLIMATE and Scikit-Grid models were pre-trained on the full set of TCGA samples and used to predict each METABRIC sample to produce the results visualized in Fig 2-2. Crucially, a quantile rescaling of the expression values in the METABRIC microarray data was performed to align the distribution of expression values with those of the TCGA training set to facilitate the transferability of classifiers. Overcoming an FFPE sample preparation artifact was the focal challenge for the AURORA experiment. The METABRIC and AURORA external validations were conducted in collaboration with Jordan Lee of the Ellrott Lab at OHSU and Chris Wong of the Stuart Lab at University of California, Santa Cruz.

Python packages in addition to Scikit-learn utilized for analysis and plotting: Scipy, Numpy, Matplotlib, and Seaborn.

Table 1 TCGA demographics[41]

Statistic	**Value**
Age at Diagnosis	
Median (years)	60
Range (years)	10 - 90
Sex	
Female	52%
Male	48%
AJCC pathologic tumor stages for TCGA cancers, source DOI: https://doi.org/10.1038/s41416-018-0140-8	**Table of TCGA stages:** https://www.nature.com/articles/s41416-018-0140-8/tables/1

Chapter 3 — mRNA signatures incorporate multiple initiating processes of cancer: a comparison with mutational landscapes

Brian Karlberg, Kyle Ellrott

Publishing/permissions: NA

Abstract

Gene expression signatures, such as the PAM50 breast cancer signature, are frequently used as molecular definitions of cancer subtypes. While these transcriptomic measurements of protein-coding genes provide the ability to differentiate between different cancer subtypes, they do not necessarily correspond to the putatively initiating mutational landscapes. We hypothesized that some of the mRNA signatures instead represent the integrated effects of many layers of regulation and their predictive utility would hold over the 26 GDAN-TMP primary TCGA tumor types encompassing 106 molecular subtypes. To decipher these complex relationships, we first utilized a memo-sort algorithm - based on the fundamental information theory concept of memoization - to identify mutated genes at subtype resolution, with a characterization of mutual exclusivity. Next, feature selection within a repeated sub-sampling framework was applied to mRNA-seq gene abundance measurements to identify corresponding expression signatures. Gene networks were built by mapping these oncogene and mRNA feature sets using gene-to-gene interactions reported in Pathway Commons with widely variable results over both subtypes and primary tumor types. In testing four onco-screening approaches, cancer samples were stratified by whether each set of oncogenes would have detected their cancer or not; gene expression signatures are shown to perform equivalently well between these sample cohorts. Gene expression incorporates multiple initiating molecular processes of cancer of which a tumor's mutational landscape is one component — the results of these experiments comprise coherent evidence in support of the utility of gene expression signatures toward clinical panel development and biological interpretation.

Background

The coding region of the genome - sections of DNA directly associated with production of functional proteins - has generally been the focus in studies aiming to determine onco-driver genes[67,68]. This focus resulted from the lower cost of whole exome sequencing (WES) compared with whole genome sequencing (WGS) in the effort to translate mutational signatures to the clinic[69,70]. Next-generation sequencing technologies have since emerged and include single-cell sequencing, immunophenotyping, epigenetic profiling, and transcriptomics[71]. While somatic

mutations, typically instantiated with WES oncogene measurements, are widely understood as one of the major drivers of tumor differentiation[72], our research, Ellrott et al. (2025), has shown that alterations in the coding transcriptome provide the best source of information for defining molecular subtypes[10]. Machine learning (ML) methods consistently identify gene expression measurements over mutational profiles as more performant in delineating cancer subtypes. This reinforces the concept that mRNA portraits of the transcriptome are of utility in capturing how cancer changes complex proteogenomic interactions[73]. The set of experiments in this work focused on the GDAN-TMP MUTA (MC3 oncogene filtered mutation) and GEXP (mRNA-seq expression) data with one experiment here including the datatype METH (DNA methylation) as a control.

Feature selection algorithms are used to overcome the curse of data dimensionality - where the number of features is extensively more than the number of samples. Feature importance methods score the values of individual features relative to other features[74] for the purpose of interpreting what factors drive model performance. Feature importance in ML can be conceptualized similarly to the weights in a regression model - the sign and magnitude of the coefficients in an equation indicate how independent variables affect the dependent variable. For analysis of high dimensional genome-transcriptome data, these feature interpretation capabilities are essential to building ML pipelines of clinical utility[75].

Several findings from the GDAN-TMP project were relevant to the design of this study. First, different ML methods applied to a given cancer type can yield divergent results such as different sets of features selected that lead to both variability in biological interpretations and differences in prediction performance. We applied this lesson of methodological benchmarking in this work in comparing multiple onco-screening methods over the 8,791 TCGA cancer samples included in these data. We also applied a benchmarking approach in the comparisons of mRNA-seq feature selection methods and feature importance methods. Next, a threshold of 70 samples had been observed in the TMP study as the sample size where predictive performance markedly improved model performance across the cohorts; this observation was used to determine the sub-sampling threshold in the feature selection framework in this work. In another example - the external METABRIC cohort validation in the TMP project - a silhouette score was utilized to characterize

goodness-of-fit of individual samples to their respective classes; we extended that concept here with feature importance scores applied to the selected mRNA features as a means to inform feature engineering and feature set analysis. Additionally, the GDAN-TMP mutation profiles utilized in these experiments included four mutation types: LOF (loss-function[76,77]), nons (non-synonymous[78]), HOTS (hot-spot[79,80]), and COMP (composite[81]). Briefly, LOF mutations result in partial or complete loss of protein function, nons includes gain-of-function and neutral mutations such as nonsense and read-through with missense as the most-common type of nons mutation, HOTs represent unusually high mutation frequency across different tumors or patients, and COMP indicates more than one single nucleotide variant within a single gene.

Gene interaction networks have been cataloged and indicate what other genes a given gene may interact with[82,83]. Oncogene lists have been established that can be used to identify the mutational status of individual samples[84,85]. Investigation of expressed gene sets that correspond with mutated gene states can be done at both the primary tumor and molecular subtype levels. These data on gene interactions may also include directionality and the type interaction such as "controls expression of". Mutual exclusivity and co-occurrence of somatic driver mutations have been previously characterized in the MC3 project by Ellrott et al. (2018) for TCGA samples[86] in the context of signal transduction pathways.

Results

To facilitate direct gene-to-gene comparisons, the mutation feature values were first aggregated by gene - each unique gene in the mutation data was given an overall designation of mutated if that gene was indicated as mutated for any of the four mutation types within any feature. The mutation and expression molecular profiles were mapped within each primary tumor type via the Pathway Commons gene-interaction .sif file. This was done by taking the three-way intersection of the mutation genes, known interactions, and expression genes. This gene-list filtering was done to control for equivalent probability in support of permutation tests. Corresponding mutation and expression feature sets within each primary tumor type were then identified from these mapped gene sets. For the mutation data, a de novo implementation of a memo-sort mutual exclusivity algorithm[87,88], based on the concept of memoization, was developed and applied at

subtype resolution. For the expression data, the MXM[89] forward-backward early-dropping (FB-ED) feature selection was run within a repeated subsampling framework, with replacement, which produced a compendium feature set that was then sorted by frequency to reveal a ranked set of the most commonly selected expression features for each cancer type. Directed gene-network interaction graphs were then constructed from these feature sets; a representative example of these analytical processes is shown in the flow chart in Fig. 3-1.

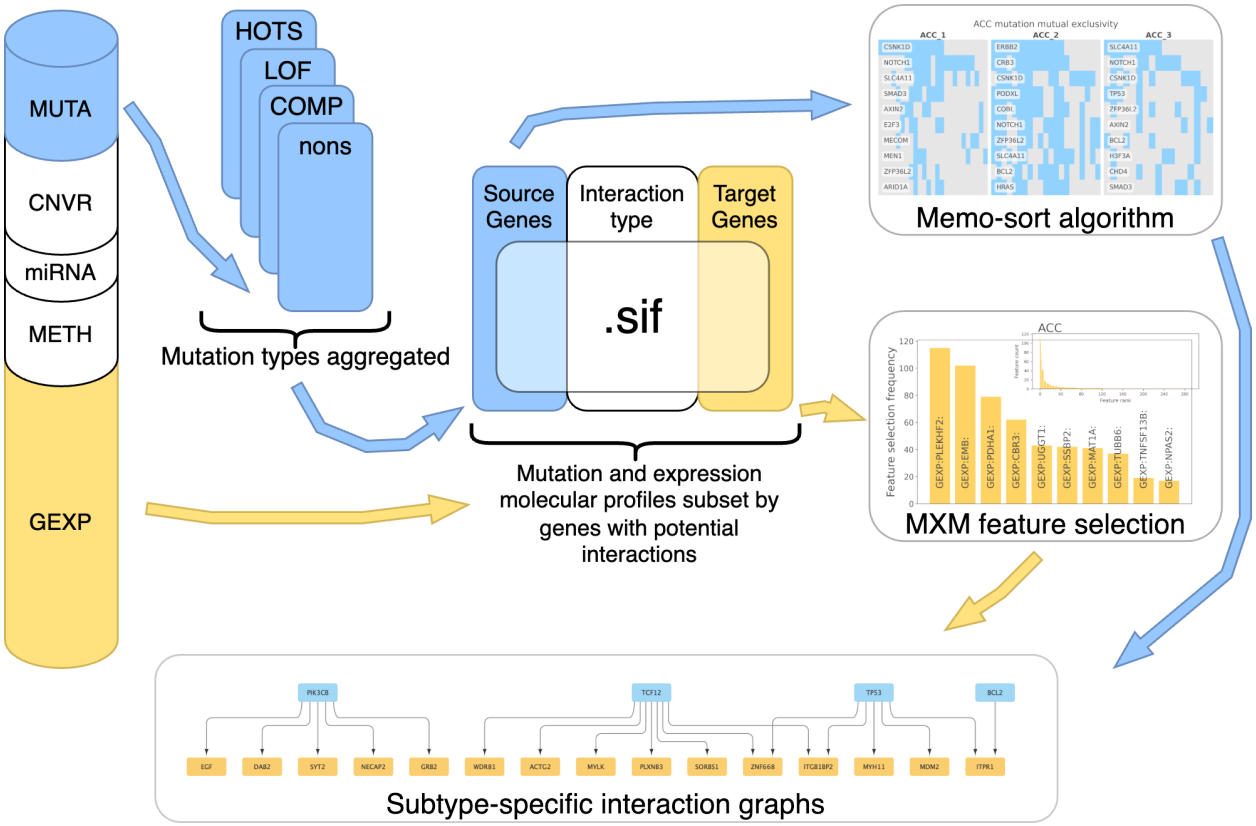


Fig. 3-1 Pipeline overview for mRNA and mutational feature set identification and interaction analysis. Within each primary tumor type, mutation status was aggregated over mutation types at the gene level. After taking the intersection with known interactions, our memo-sort algorithm was applied to the mutation features whereas for the expression profiles, the MXM forward-backward early-dropping selection algorithm was applied within a repeated sub-sampling framework. From this, gene-interaction graphs with primary, subtype, or mixed resolution can be built from the engineered feature sets

Memo-sort algorithm applied to mutation profiles

The memo-sort algorithm revealed the most-frequently mutated genes within each cancer type and the degree to which these mutations co-occurred, Fig. 3-2. For most subtypes, mutations in only two or three genes cover the majority of the samples within that subtype. For example, in ACC subtype 2 (ACC_2), mutations in ERBB2, the first most-frequently mutated feature, or C5NK1D, the third most-frequently occurring gene, would be indicative of this subtype for almost all the samples.

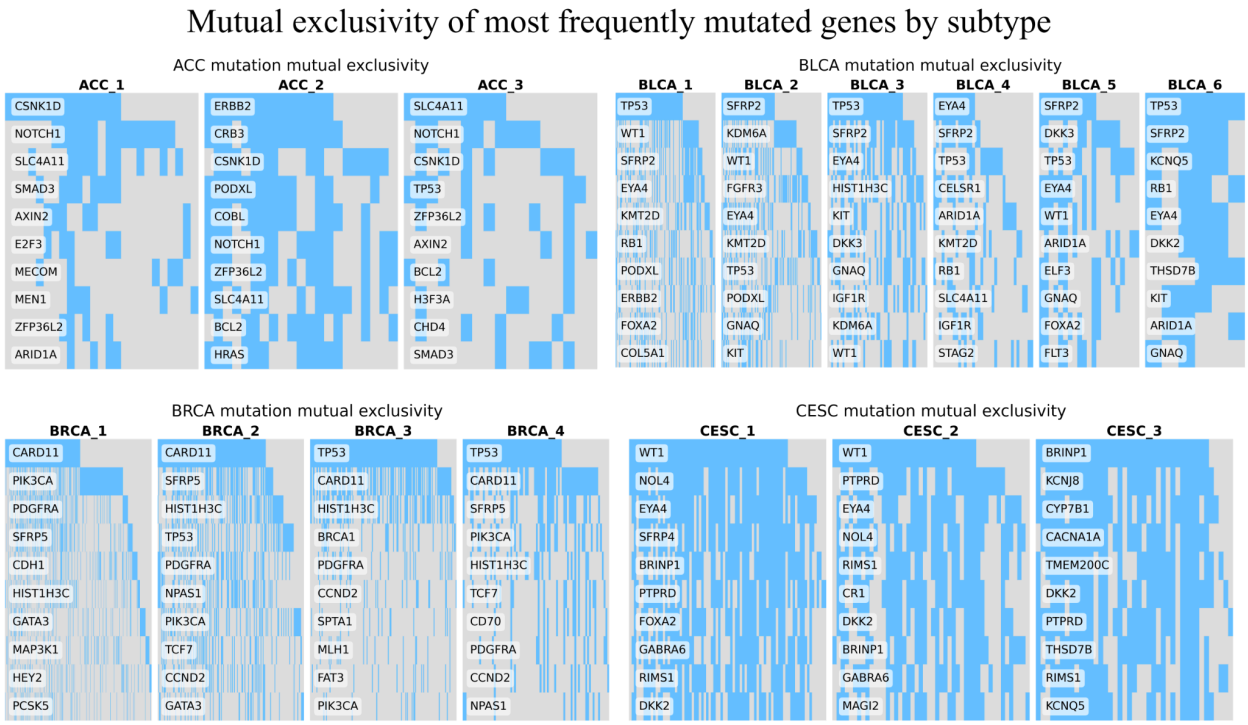


Fig. 3-2 Mutual exclusivity and co-occurance of mutations at subtype resolution. Waterfall plots of the memo-sorted of mutation genes characterized by the memo-sort algorithm within TCGA molecular subtypes showing the general patterns of mutual exclusivity within each subtype with simultaneous overlap of genes between subtypes. These observed patterns held for most of the 26 GDAN-TMP primary tumor types

We next validated the application of the memo-sort algorithm to the raw data of mutated genes by first running the MXM forward-backward early-dropping feature selection method within a sub-sampling framework and then applying the memo-sort. The results of this sequential process reveal both similarities and differences to the direct application of the memo-sort, as shown in Fig. 3-3. For example, the gene ERBB2 was still the predominant feature for ACC_2 whereas a

divergence from the raw data was observed with BCL2 emerging as a potentially predictive gene in selected genes for ACC subtypes 1 and 2.

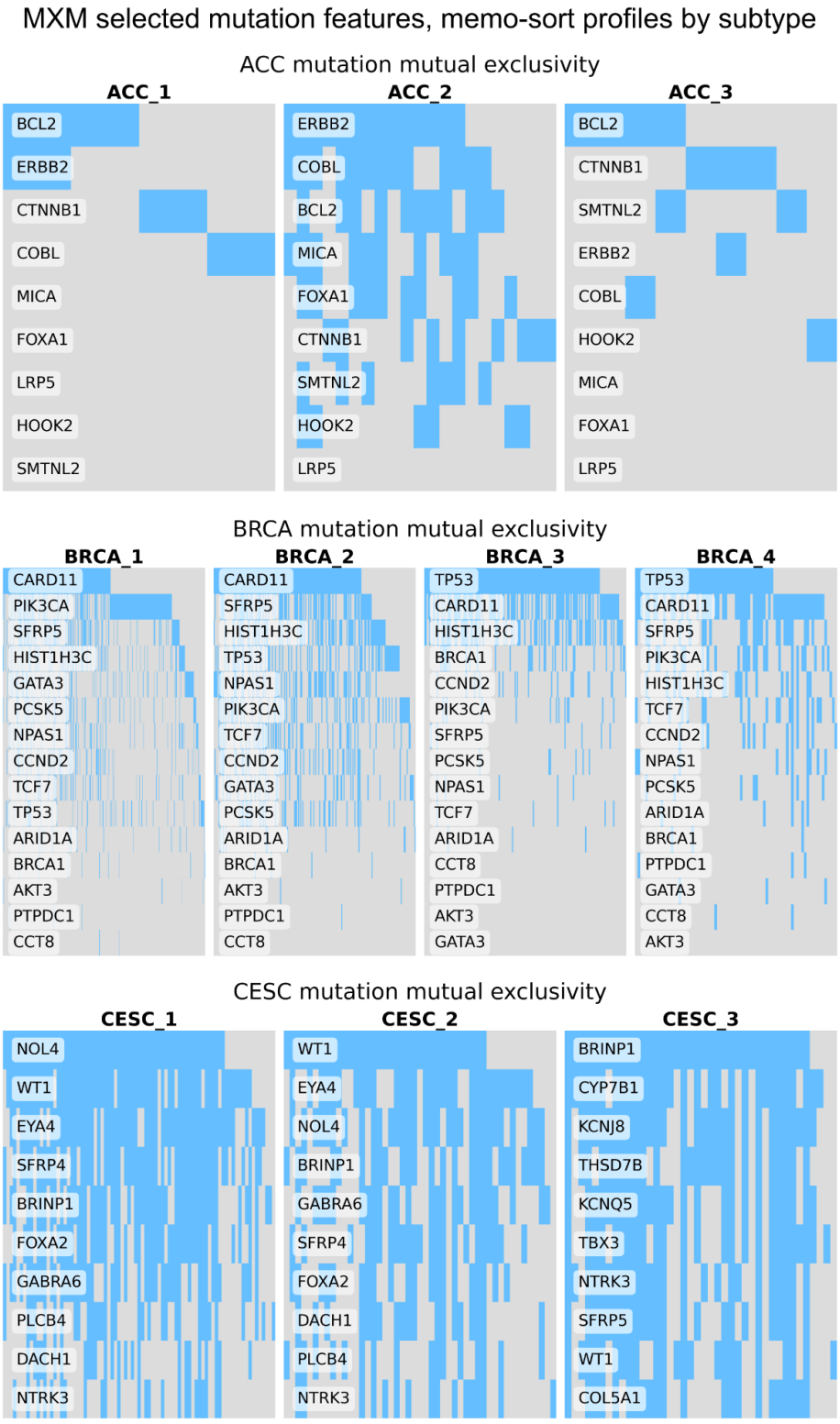


Fig. 3-3 Benchmarking of mutation memo-sort with prior feature selection. Statistical feature was run to down-select predictive features prior to application of the memo-sort for comparison with the direct

application of the memo-sort algorithm shown in Fig. 3-2 Both similarities and differences in mutational patterns across cancer types are revealed

mRNA feature selection within a subsampling framework

To identify mRNA features relevant to primary tumor types, we focused on the MXM forward-backward early-dropping feature selection algorithm that was identified as relatively performant in the GDAN-TMP project. To produce a quantification of relative feature importance and counteract over-fitting, we utilized a sub-sampling-with-replacement strategy where the feature selection was repeated 250 times for each primary tumor-type at a threshold of 70 samples. This sub-sampling threshold was previously determined in an experiment on statistical power-analysis in the GDAN-TMP project[10], Ellrott et al. (2025). The selection frequency across the 250 repeats of the top-10 most-often selected expression features at primary tumor resolution are shown in the main panels of Fig. 3-4 with inset panels showing the full distributions of selected mRNA-seq features over the 250 selection replicates.

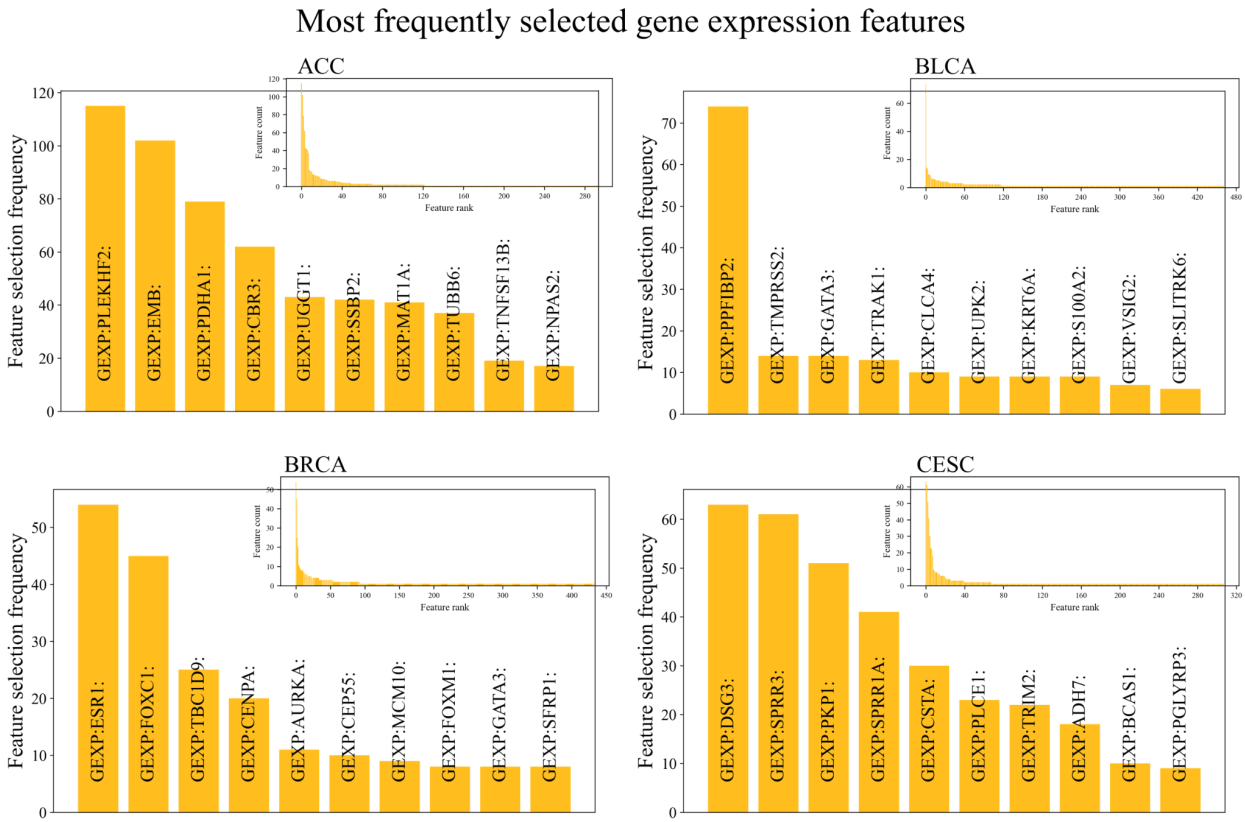


Fig. 3-4 The 10 most-frequently selected expression features at primary tumor resolution. Forward-backward early-dropping statistical feature selection was run in a repeated sub-sampling framework to identify predictive expression features. Main panels show the top-10 features selected for the first-four primary tumors; inset panels show the full set of selected features over the 250 selection replicates. The observed patterns generalized over all 26 tumor types

To benchmark the performance of the MXM feature selection algorithm within the sub-sampling framework, recursive feature elimination (RFE) feature was also run on the expression features in the same 250-repeat 70-sample threshold framework. Again interpreting the frequency of selection as a measure of biological relevance of individual genes to specific cancer types, the RFE method was observed to return a lesser degree of separation in rate of selection for individual features, Fig. 3-5.

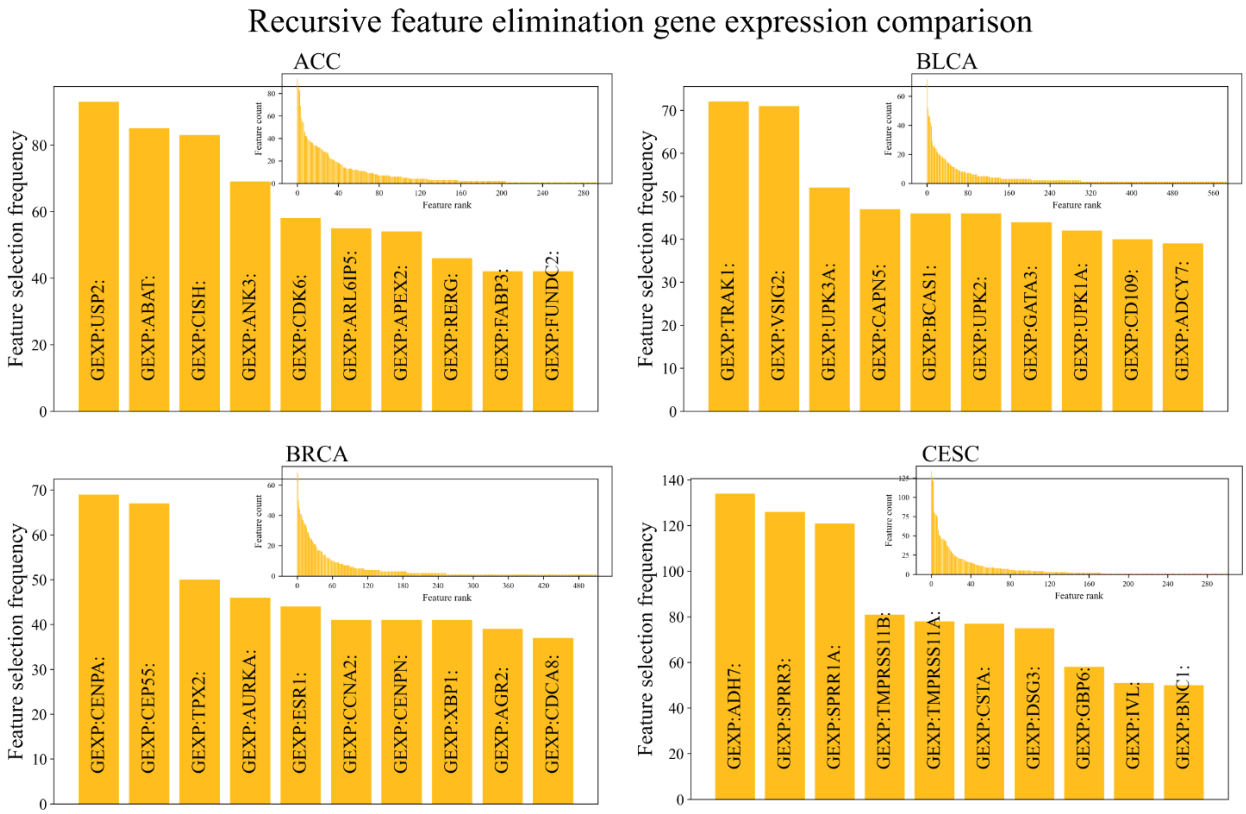


Fig. 3-5 Recursive feature elimination (RFE) feature selection method benchmark for the MXM-based selection shown in Fig. 3-4

The relative compactness of the most-frequently-selected features MXM sets was quantified with an area-under-the-curve calculation for the first four cancers, Fig. 3-6. The MXM forward-backward early-dropping method selects features with significantly (Mann-Whitney test, p-value 0.029) sharper drop-off in frequency of less-often selected features over the selection replicates. This pattern held across the 26 TCGA cancers under investigation.

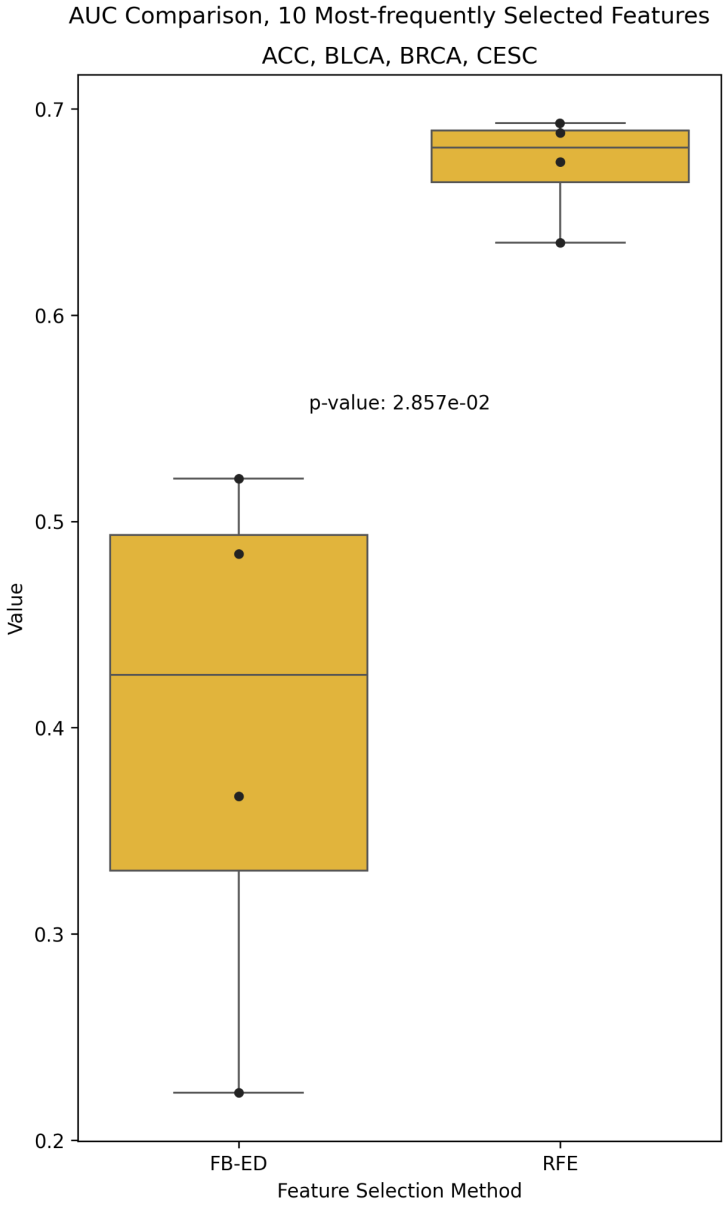


Fig. 3-6 Area-under-curve (AUC) comparison of feature selection rates by method. For the aggregated sets of expression genes for MXM, as shown in Fig 3.4 and RFE shown in Fig 3.5, MXM converges on a narrower set of frequently-selected features as seen in the sharper spikes of selection frequency at the taller end on the left-hand side of the mRNA (expression) histograms for the MXM and RFE feature

selection methods. This effect is quantified here with an area-under-the-curve (AUC) calculation showing how MXM selects significantly (Mann-Whitney test, p-value 0.029) more-parsimonious feature sets for the first four TCGA cancers; this pattern holds in general over the remaining cancer types

mRNA feature importance benchmarking

These comparatively-performant MXM feature sets were carried forward for downstream interaction analysis of mutated genes specific to subtypes and analysis of mRNA signature-based onco-status. Interpretability of these mRNA feature sets was quantified with two feature importance methods at subtype resolution - scikit-learn decision tree (DT) Gini and SHAP. Better separation of individual features within subtypes was observed for the scikit-learn feature importance method; feature importance quantifications with error bars for the standard deviation over 30 replicates of the importance calculations are shown in Fig. 3-7 and Fig. 3-8 Supplemental. This comparison of subtype-specific feature importances can inform development of subtype-specific clinical panels as well as elucidate subtype-specific cancer biology. For example, for the first-ranked feature for ACC_1, both methods identify PDHA1 as the most important feature; however only the scikit-learn method does so with statistical significance as determined by the error bars of standard deviation. For ACC_2, TNSF13B is identified by scikit-learn as the most important by a wide margin; although SHAP identifies this feature in the top three, it does so with no significance across the entire set of 10 features. This observation reinforces the theme of applying multiple methods for ML applications in molecular subtype analysis for validation and robust interpretation of results.

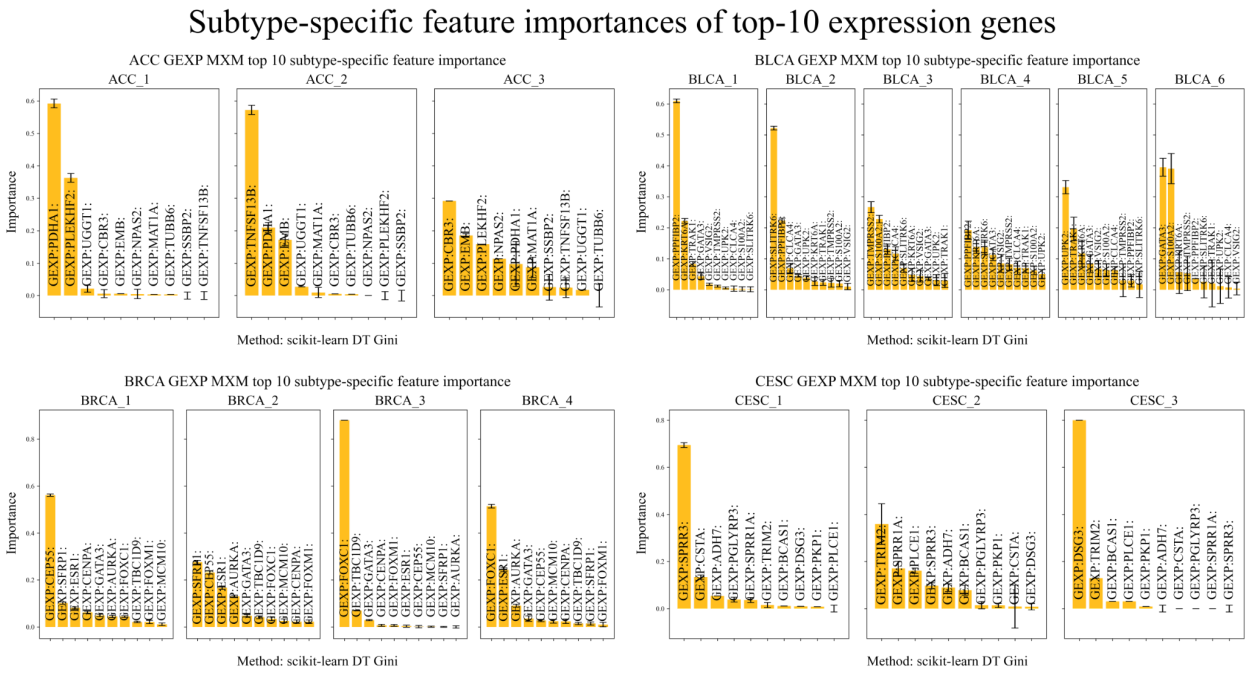


Fig. 3-7 Subtype-specific feature importance calculations with scikit-learn Gini method. Error bars are standard deviation over 30 importance calculation replicates

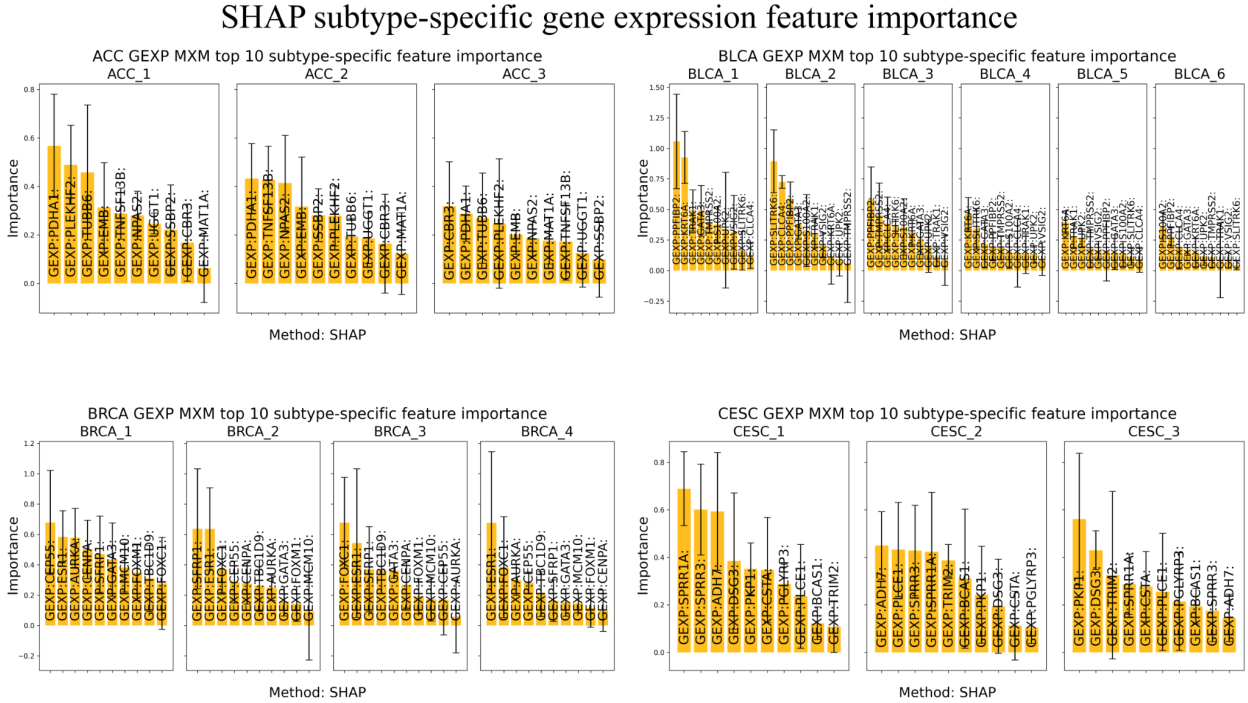


Fig. 3-8 SHAP implementation of class-specific feature importances. Error bars standard deviation; for comparison with scikit-learn Gini importances shown in Fig. 3-7

Mutation-expression interaction graph

Next, network interaction graphs were constructed for mutation and expression genes within each primary tumor type. The two most-mutated genes identified in the mutual-exclusivity analysis for each subtype within each tumor type were selected and the union of these mutation features composed the upstream component of the graphs. Any expression genes occurring more frequently than the p-value 0.01 (binomial test, alternative = greater) threshold were then added and interactions reported in Pathway Commons with the mutation genes as the source and the expression genes as the target were used to construct the graphs. Examples of increasingly complex graphs are shown in Fig. 3-9.

The same MC3 oncogene was among the two most-mutated for more than one subtype within 22 out of the 26 primary cancer types. Increasing the threshold to the strictest possible level of only the single most-mutated gene for each subtype within each primary tumor type still resulted in overlap of the same gene being the most-frequently mutated for more than one subtype within 18 out of the 26 primary cancers. The average number of mRNA features occurring more frequently than expected at the 0.01 p-value selection significance for each cancer was 105 with a standard deviation of 41. This was out of an average 445, standard deviation 253, total unique mRNA features per cancer selected over the 250 MXM sub-sampling runs. The redundancy in common mutations at subtype resolution shows potential overlap in biological processes between cancer subtypes. Combined with the cross-relationships between subtype-specific mutations and expression genes of importance, this highlights the need for minimal feature sets that capture the interplay of initiating processes and are capable of robust probabilistic prediction of subtypes.

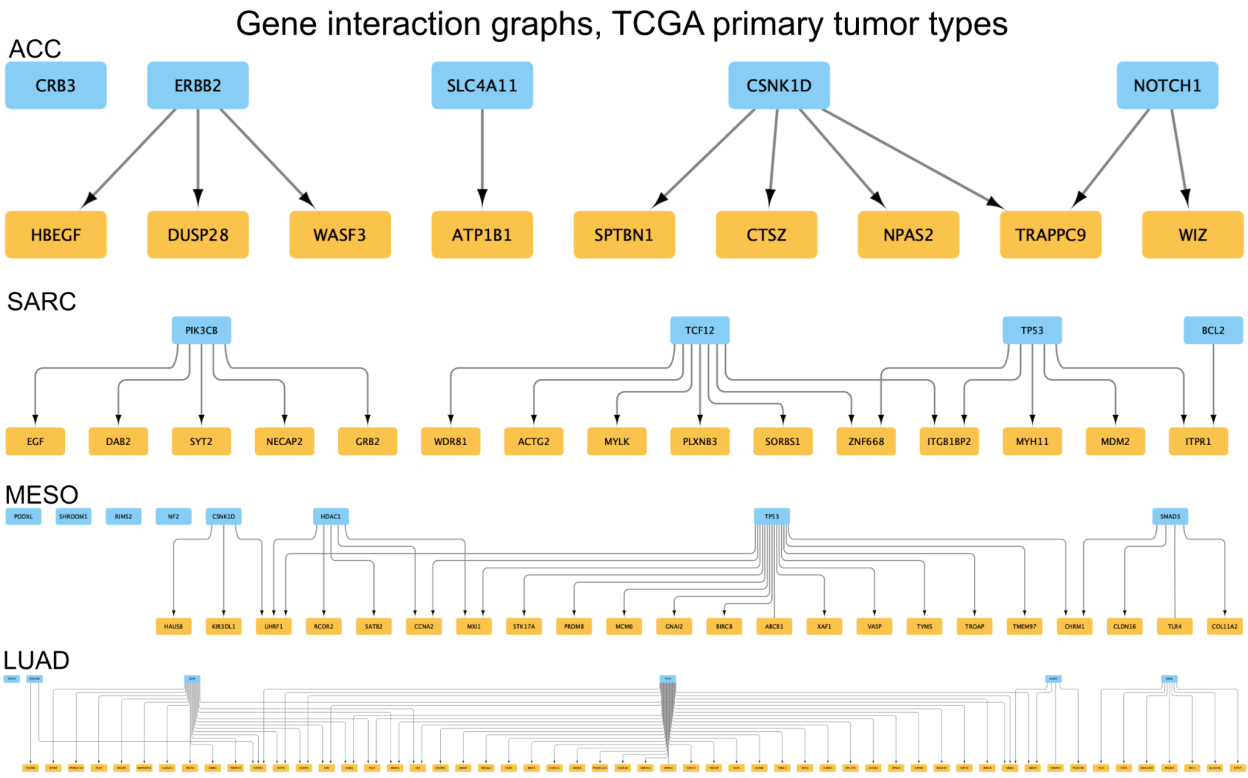


Fig. 3-9 Gene interaction graphs at primary tumor resolution. Union of the top 2 most-frequently mutated subtype-specific MC3 oncogenes mapped to mRNA genes significant at the primary tumor level; binomial test, p-value 0.01

Annotated versions of these interaction graphs could be built with the subtype-specific ranking of the mutation genes, the selection frequency and corresponding p-value of the expression genes, and the interaction type. An example of this annotation is shown in Fig. 3-10.

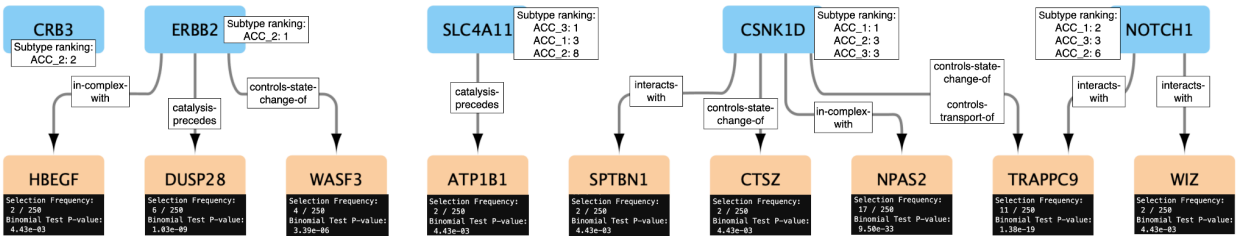


Fig. 3-10 Annotated interaction graph for adrenocortical carcinoma. The rank order of the ten-most frequently mutated genes within each subtype, the interaction type, and the selection frequency and corresponding p-value for the expression genes are reported

Datatype prediction performance benchmarking

To compare the differences in predictive signal over data types, a random forest (RF) classifier was trained using the top feature sets identified in the GDAN-TMP project for three data types within each cancer type. DNA methylation (METH) was included with the mutation (MUTA) and gene expression (GEXP) feature sets as a control. Gene-expression signatures and DNA methylation signatures differentiate cancer subtypes with ML models generally better than mutation signatures as shown in Fig. 3-11.

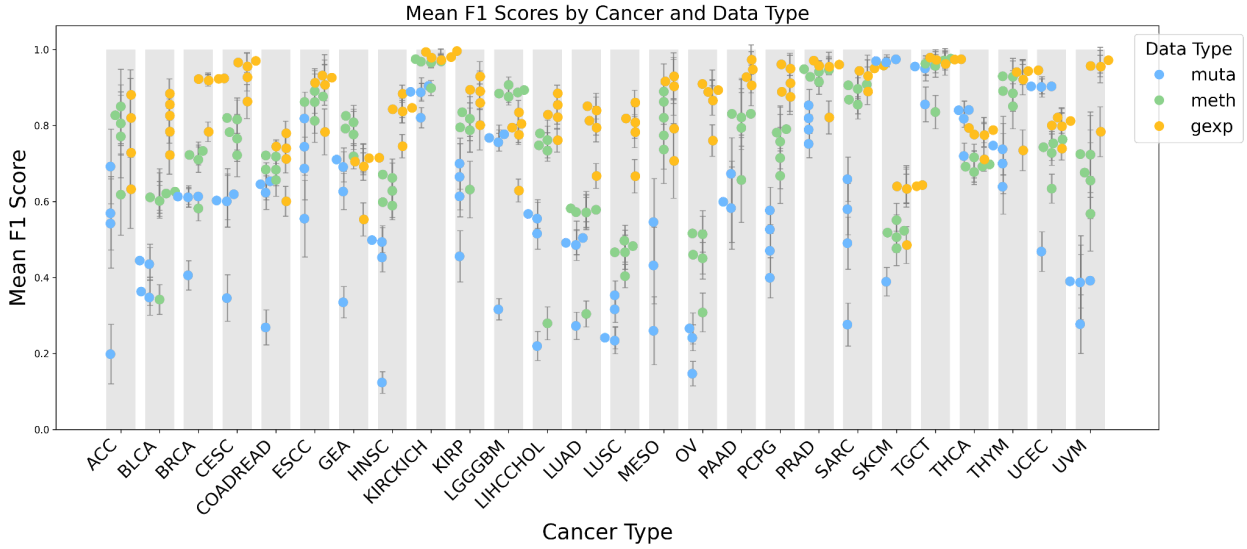


Fig. 3-11 Comparison of data types in predicting subtypes within primary tumor types. The same random forest (RF) classifier was trained and tested over 30 data-splits using the GDAN-TMP top-model feature sets for the MUTA, METH, and GEXP data types. Gene expression (mRNA) signatures frequently outperform other data types across TCGA cancers

Transcriptional signature performance stratified by oncogene screenings

We then hypothesized that mRNA signatures could differentiate molecular subtypes equivalently well for these TCGA samples that either would have been detected as cancerous or not according to known oncogene mutations and clinically-implemented onco-screening gene sets. To test this, four oncogene lists were used to screen the mutational status for these 8,791 TCGA cancer samples within each primary tumor type. These oncogene screening lists were the COSMIC

oncogenes, the OHSU Knight Diagnostic Lab (KDx) oncogenes, the Ampliseq50 genes, and primary tumor-specific onco-lists derived from our memoization-based sorting algorithm[87]: Our memo-screen gene sets were composed of the union of the top-10 most-frequently mutated genes in each subtype for that primary tumor type. Similar but not concurrent onco-status was determined with these four methods across the 26 GDAN-TMP cancer types as shown in Fig. 3-12. Proportions of onco-negative samples are calculated relative to total sample count within each cancer type according to each corresponding onco-screening method.

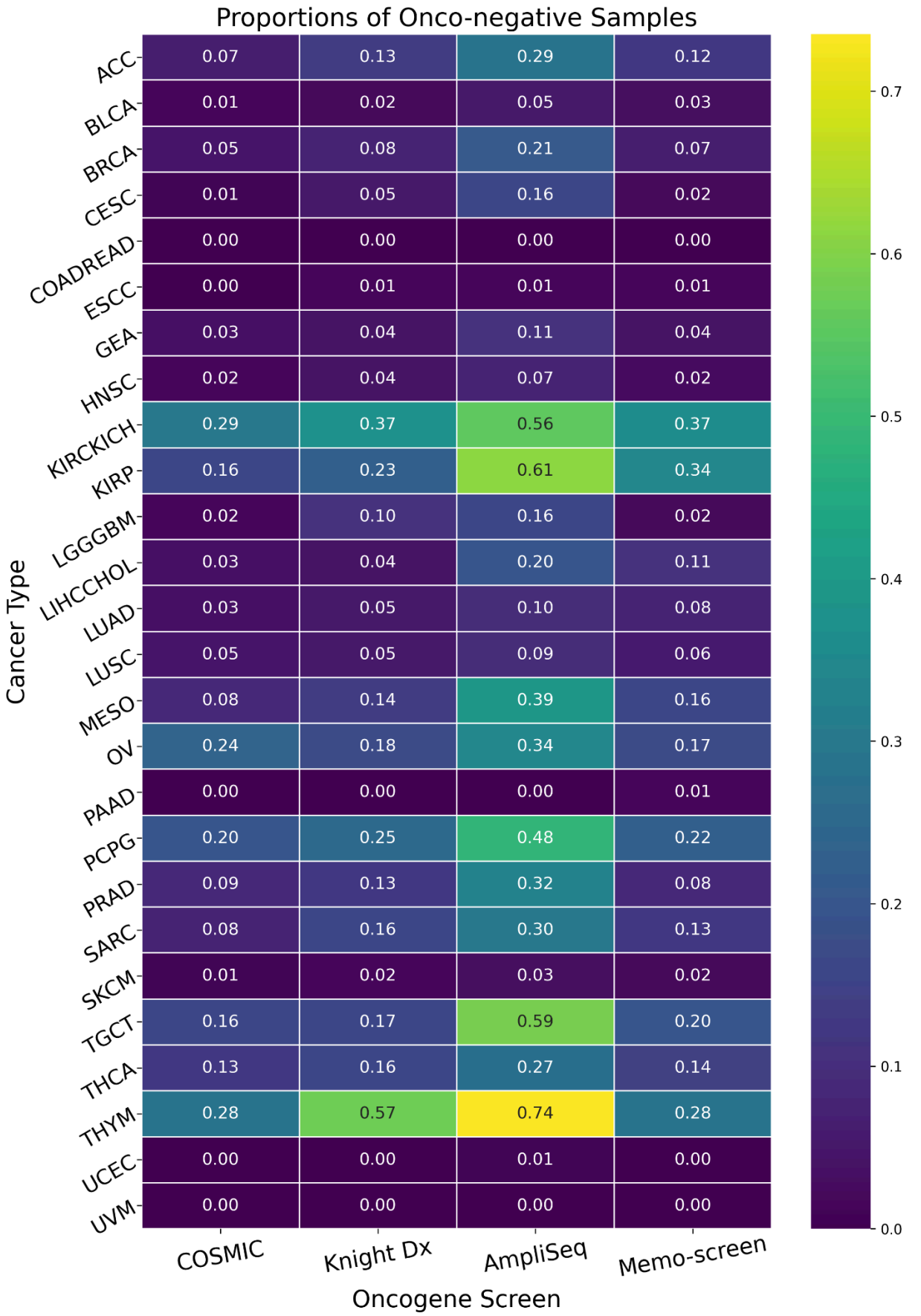

Proportions of Onco-negative Samples
Cancer Type
ACC 0.07 0.13 0.29 0.12
BLCA 0.01 0.02 0.05 0.03
BRCA 0.05 0.08 0.21 0.07
CESC 0.01 0.05 0.16 0.02
COADREAD 0.00 0.00 0.00 0.00
ESCC 0.00 0.01 0.01 0.01
GEA 0.03 0.04 0.11 0.04
HNSC 0.02 0.04 0.07 0.02
KIRCKICH 0.29 0.37 0.56 0.37
KIRP 0.16 0.23 0.61 0.34
LGGGBM 0.02 0.10 0.16 0.02
LIHCCHOL 0.03 0.04 0.20 0.11
LUAD 0.03 0.05 0.10 0.08
LUSC 0.05 0.05 0.09 0.06
MESO 0.08 0.14 0.39 0.16
OV 0.24 0.18 0.34 0.17
PAAD 0.00 0.00 0.00 0.01
PCPG 0.20 0.25 0.48 0.22
PRAD 0.09 0.13 0.32 0.08
SARC 0.08 0.16 0.30 0.13
SKCM 0.01 0.02 0.03 0.02
TGCT 0.16 0.17 0.59 0.20
THCA 0.13 0.16 0.27 0.14
THYM 0.28 0.57 0.74 0.28
UCEC 0.00 0.00 0.01 0.00
UVM 0.00 0.00 0.00 0.00
COSMIC
Knight Dx
AmpliSeq
Memo-screen
Oncogene Screen
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Fig. 3-12 Proportions of TCGA cancer samples designated as onco-negative via mutation screening. Sample mutation screening comparison of four oncogene lists: COSMIC, OHSU Knight Diagnostics (Knight Dx), Ampliseq50, and the Memo-sort-identified frequently mutated genes (Memo-screen). Onco-screens that return a mixture of cancer samples designated onco-positive and onco-negative are input to subsequent F1 score and confidence comparisons. *Note: all samples screened are in fact cancer-positive as derived from the TCGA regardless of status assigned by screening method*

An alternate visualization of these onco-screening results for TCGA samples known to be all cancerous, in the form of an upset plot of designated onco-status intersections comparing the subtype-specific onco-screening results, is shown in Fig. 3-13. For 60 of the GDAN-TMP cancer subtypes, more than half of the 106 total, at least one TCGA cancer sample was not detected as cancerous according to all four onco-screening methods tested.

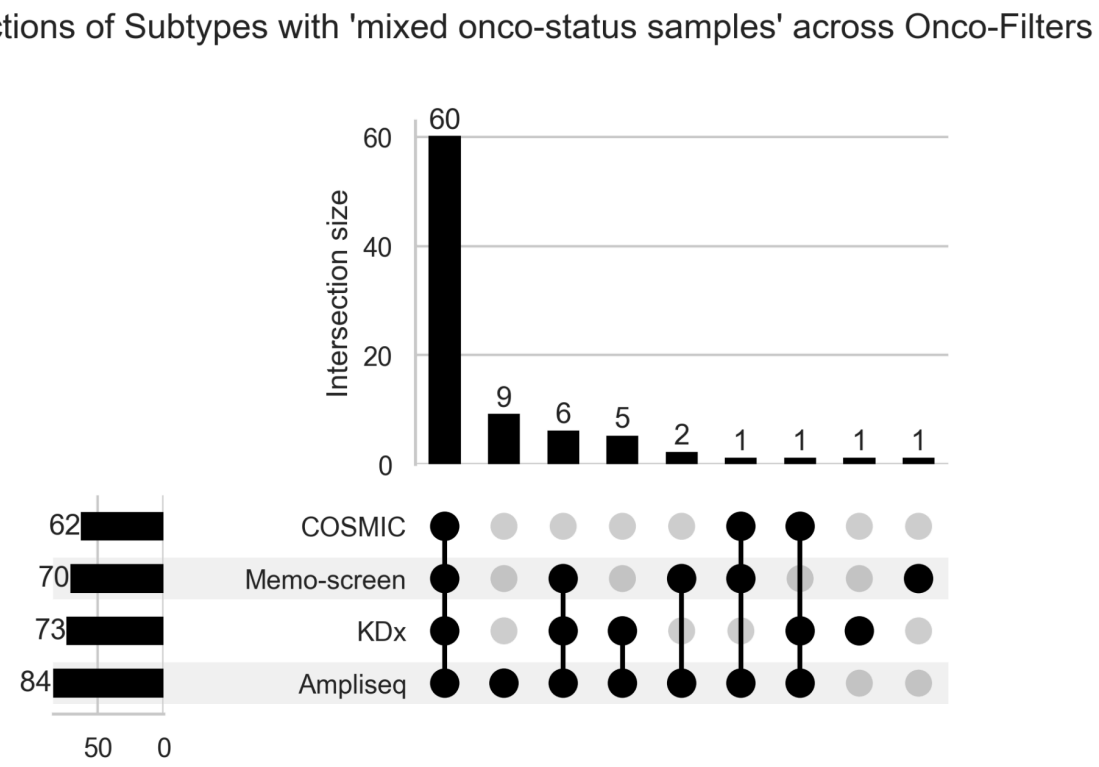


Fig. 3-13 Subtype-resolution of cancer sample mutation screening. Intersection pattern of subtype onco-status according to four onco-screen gene panels for the 106 cancer subtypes defined by the GDAN-TMP project. Onco-screens returning a mixture of samples designated as onco-positive and onco-negative for at least one subtype within a given primary tumor type are the source of input to

subsequent subtype prediction and confidence comparisons by onco-status. 60 subtypes contained one or more TCGA diagnosed cancer samples that would have been flagged as non-cancerous according to all four screens

Our memo-screen mutation gene lists were comparatively efficient, as measured by the average number of genes comprising each list. This analysis was inline with a primary theme of the GDAN-TMP project – finding compact feature sets. Despite the comparable cancer detection rates, the memo-sort algorithm utilized 23.6 genes per primary tumor type compared with 753 for the COSMIC, 281 for the KDx, and 50 for the Ampliseq screens, Fig. 3-14.

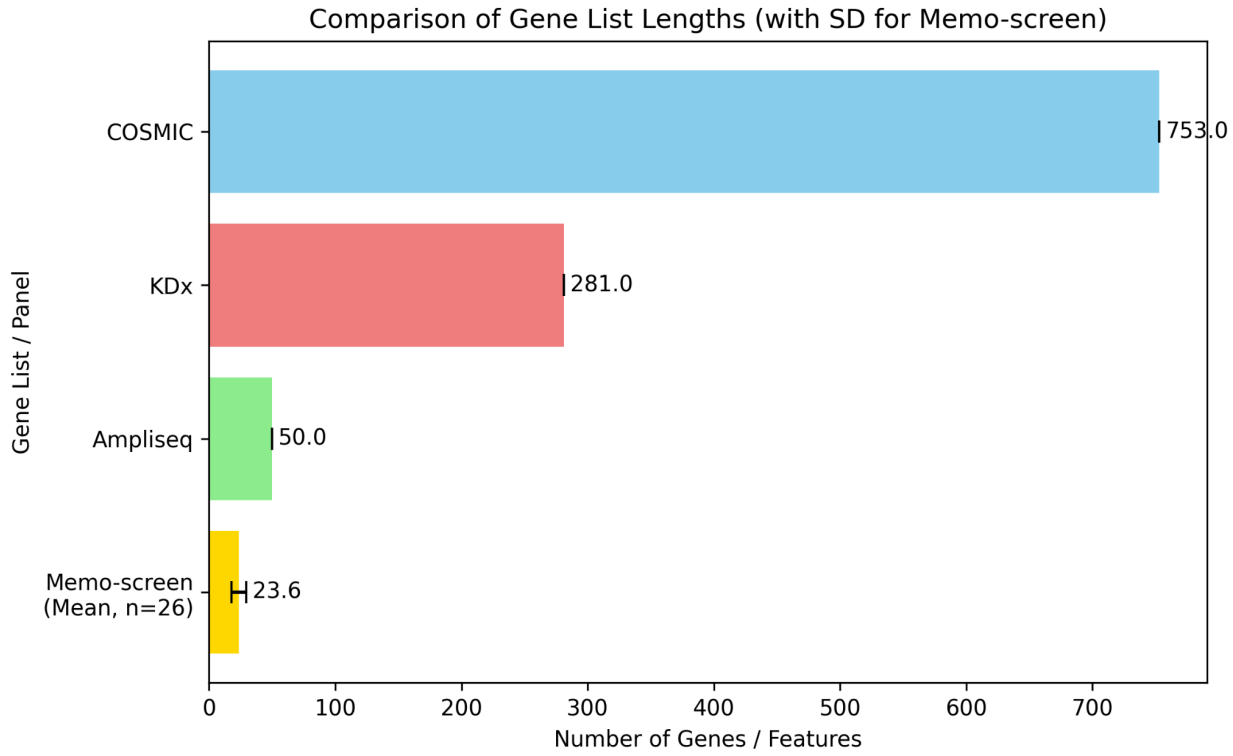


Fig. 3-14 Onco-screen gene count comparisons. The yellow bar shows the average gene count of our memo-sort-derived, cancer-specific lists compared with the COSMIC gene list, Knight Diagnostic Labs (KDx), and the Ampliseq50 genes. With respect to potential clinical implementation, this shows the relative efficiency of memo-sort derived oncogene lists for mutation-based cancer screening

To test the consistency of gene-expression-based feature sets in making subtype predictions on cohorts of cancer samples stratified by designated onco-status, predictions were made over the 26 cohorts – each using 157 different GEXP feature sets – where the designated onco-status was

tracked and used to bin each of the 8,971 cancer samples by putative onco-status. This test was replicated for each of the four onco-screening methods. Expression mRNA signatures were found to consistently return similar performance, as measured by F1 score and confidence, for cancer samples designated as either onco-positive or onco-negative across cancer types, specific GEXP feature sets, and screening methods as shown in Fig 3-15.

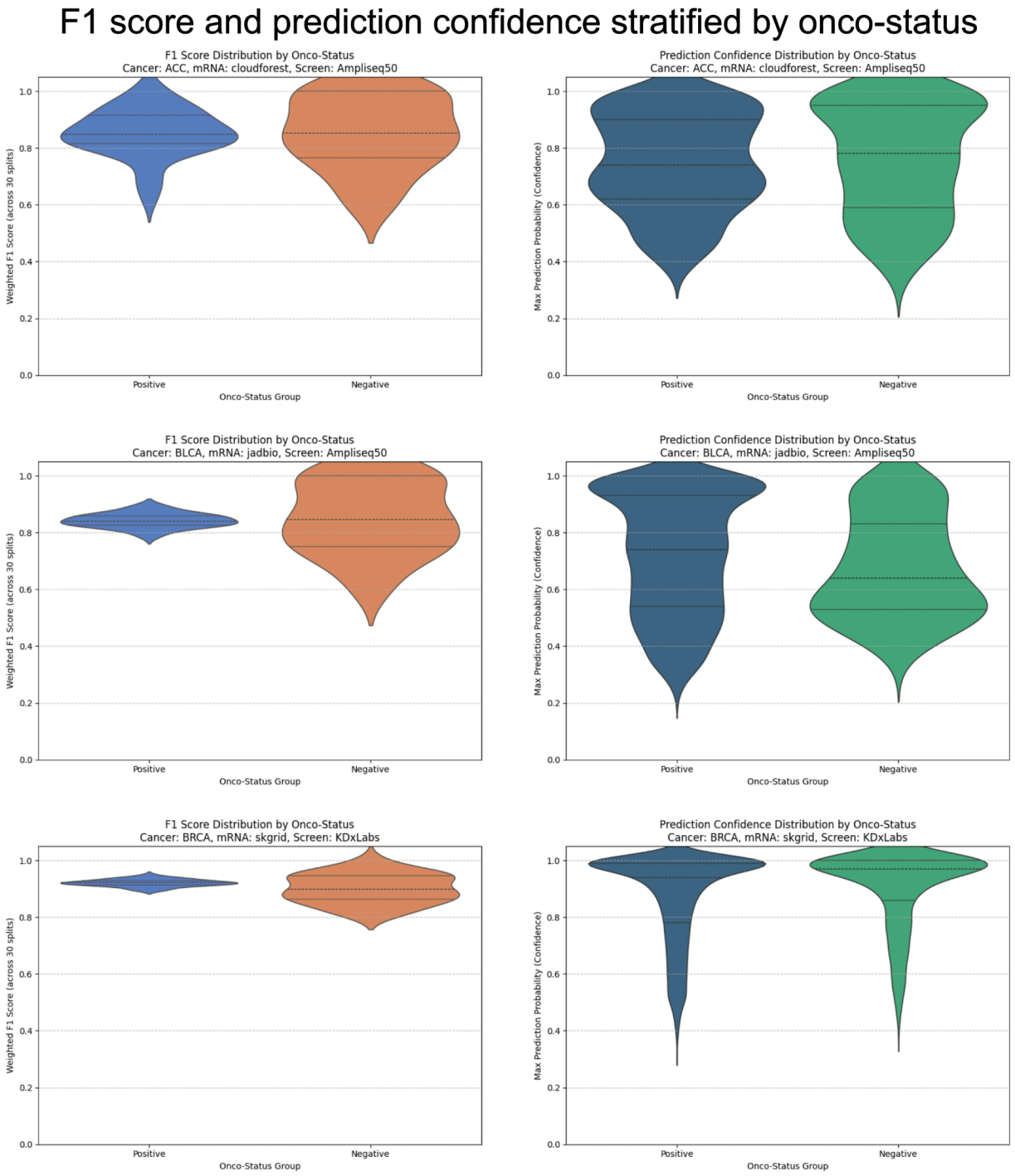


Fig. 3-15 mRNA-seq signatures predict subtypes in cancer samples regardless of mutation-based screening results. Primary tumor-level F1 scores and sample-level confidence scores are concordant for cancer samples designated either onco-positive or onco-negative using gene expression feature sets for the prediction of subtypes within each primary tumor type. This trend holds across cancer types, onco-screens

and expression feature sets with exceptions that align with the findings of the GDAN-TMP project, such as skin cutaneous melanoma (SKCM)

Discussion

This study sought to demonstrate how feature engineering, supervised categorical prediction, and feature importance methods can combine to interpret the utility of different data types in the context of molecular subtype prediction. The results show the clinical utility of mRNA-based transcriptomic signatures in leveraging integrated information from the various initiating processes of cancer to delineate subtypes. Our approach focused on two of the five datatypes from the GDAN-TMP project – the mutation and expression profiles. We applied methodologic benchmarking across these experiments in identifying cancer-specific gene-sets for both data types, quantifying the importance of individual expression features, and characterizing the interactions of these identified mutation and expression genes to interpret the utility of different TCGA data types in molecular subtype prediction. This rigor was further applied in a summary recapitulation of the GDAN-TMP results where multiple feature sets within each of the MUTA, METH and GEXP datatypes were used for comparing the difference in predictive signal between datatypes with gene expression generally most performant. Noted exceptions were reproduced with GEA and LGGGBM where DNA methylation was used in determining the ground truth label assignment in the training data and SKCM which was originally defined by mutations. In the final experiments in this work, testing multiple screening methods for determination of cancer detectability rates of these cancer samples combined with the subsequent subtype prediction by stratification of detection or not demonstrated the unique utility of gene expression data. The overall result of these analyses show how mRNA-seq captures the interplay of cancer-initiating processes, of which mutational profiles are a component, in determining TCGA molecular subtypes.

The memo-sorted mutation profiles, aggregated from feature level to gene level, showed that the most-frequently mutated genes often occur in more than one subtype for most primary tumor types. This presents a limitation of genomic (DNA-based) measurements for delineating cancer subtypes with ML. At the same time, the memo-sort waterfall plots showed that mutations in

typically two or three genes often occurred with a pattern of mutual exclusivity and covered the majority of samples within subtypes – this sub-analysis yields a lens of interpretation toward the specific molecular biology of individual subtypes based on mutation data. However, the memoization-based analysis that is possible because of the binary format of the mutation profile values could not be directly compared with the continuous values comprising the expression data necessitating alternate integrative approaches.

Our previous research in the GDAN-TMP project had identified the performant forward-backward early-dropping algorithm within the R package MXM. By running this feature selection algorithm in a subsampling replicate framework, compendium feature sets of expression genes could be built yielding feature selection frequencies analogous to the mutation rates by genes used in the mutation analysis. This allowed for ranking the expression genes by selection frequency to identify the most relevant genes for predicting subtypes, building the expression-mutation interaction plots, and conducting the subtype-specific expression feature importance analysis. Each of these contributed toward interpretability of datatype utility and subtype-specific biology. Limitations, however, of the MXM feature selection algorithm included its non-python implementation and substantial computational expense compared with RFE. The early-dropping component of the algorithm halts further search upon diminished returns in prediction performance is intended to improve computational efficiency. However, when input with raw feature sets on the order of tens of thousands of features as were these gene expression data, deployment with parallelization on a compute cluster is necessary which, when combined with its implementation in R, adds overhead to the tractability of the experiment. Testing python-ported versions of the FB-ED algorithm, implemented by means of recently emerged versions of advanced AI-based coding tools, could improve the iterative capacity of these types of feature engineering analyses.

The two main themes of the preceding GDAN-TMP project explored here were parsimonious feature sets and benchmarking of multiple ML methods. For the practical consideration of cost of implementation in the clinic, a minimum set of genes to screen for in determining onco-status and predicting subtypes is desirable. Both the memo-sorted mutation data and the ranked frequency selection-based expression feature sets provided a means to derive compact feature

sets via designation of arbitrary cutoff feature count values at both the primary tumor and subtype levels. While not at the consortia-project scale of GDAN-TMP, this study extended the theme of method comparisons throughout. The inclusion of the RFE selection method as a baseline for the MXM method with both methods run in the identical subsampling framework yielded a quantifiable comparison via the area-under-curve comparison. Testing the memo-sort algorithm with and without prior feature selection on the mutation data showed how combining methods can reveal both similarities and differences in feature identification and provide a route to interpretability of machine learning. The SHAP importance calculation modified for subtype specificity allowed for a characterization of performance for the scikit-learn feature importance method. For the onco-status detection analysis of the TCGA cancer samples, utilization of every TMP top-performing expression feature set in addition the the MXM sub-sampling-derived mRNA sets developed in this work resulted in a robust foundation for the subsequent conclusion of equivalent predictability in determined onco-status cohort comparison experiments. The robustness of the onco-detection observations for the TCGA cancer samples was reinforced by inclusion of four onco-screening methods. Overall, the observation in the TMP project where ML models frequently selected transcriptomic features was explored in further depth in these experiments illuminating some of the reasons why machine learning models prefer expression signatures to differentiate molecular subtypes.

Future directions for ML-based explorations of omics data

The field is moving toward industrial-scale interconnected foundation models[90,91]. These systems can now exceed one trillion model parameters, can integrate data across all modalities, are capable of inter-species transfer learning, and can design de novo proteins with diffusion techniques. While these new approaches will take time to come to fruition, our work in multimodal TCGA data integration and modeling, Ellrott et al. (2025), represents an immediate step in the direction of tangible ML-based tools for the clinic. The sub-sampling with replacement framework developed in this study that utilizes repeated feature selection is potentially a strategy to reduce overfitting of feature sets because the resulting aggregated feature set within each primary tumor type is yielded from a composite of selection runs over many random combinations of minimally-sized samples across subtypes; a future study could be

designed to test this. The gene interaction analysis could be extended to include relationships within the mutation genes and within the expression genes. Inclusion of the specific MC3 mutation type i.e. LOF, COMP, etc. could also be a facet of future analysis. In constructing the gene interaction network plots, future work could control for network complexity by dynamically varying the n=2 top-subtype MUTA gene and .01 p-value GEXP significance thresholds over the cancer types. A multi-label framework may be of utility in future subtyping taxonomies where a sample could have labels of tissue-of-origin, TCGA molecular subtype, mutation subtype, expression subtype, and a biological process-type such as immune-type[92].

Conclusions

Multi-modal feature selection can be applied to identify biological relationships between genomic alterations (DNA mutations) and transcriptomic state (gene expression) in a cancer molecular subtype context. Other genomic data, in this case methylation, can be used as a control in these comparisons. Repeated selection of feature sets from the same molecular profile for a given cancer type can yield statistical quantification of feature relevance to primary tumor type Feature importance quantification methods can highlight which subtype that specific features may be associated with. Selection methods and frameworks can be combined with feature importance quantifications to build subtype-specific gene interaction graphs in the context of tumor-initiating mutation profiles and resulting gene expression states. The significance of these interactions can be quantified to reveal subtype-specific genomic dependencies and inform more-precise targeting of therapeutics and a more complete characterization of tumor biology. These results again demonstrate that drawing conclusions in ML interpretation efforts requires methodological rigor; the arbitrary selection of unique or a limited set of experimental variable combinations can lead to incomplete conclusions. This trend suggests that heterogeneous mutation profiles in cancer can result in similar transcriptomic phenotype patterns.

Methods

Data provenance and processing, feature engineering

Molecular profile data for 26 TCGA tumors were obtained from the NCI Tumor Molecular Pathology (TMP) working group's publication data page[10]: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022. Demographics for these data are shown in Table 2-1. Gene interaction data were downloaded from: PathwayCommons12.All.hgnc.sif.gz. COSMIC oncogenes were retrieved from: https://cancer.sanger.ac.uk/cosmic/download/cosmic/v101/cancergenecensus. The Ampliseq50 oncogene list was obtained from Gemini Advanced 2.5 (experimental) and confirmed with GPT-Pro o4-mini-high: https://g.co/gemini/share/9a4b2176e776
https://chatgpt.com/share/68151846-a7dc-8010-8753-a01a5f134702
The Knight Diagnostics Lab oncogene list was derived from the GeneTrails listings at https://knightdxlabs.ohsu.edu/

For each of the 26 primary tumor types: 1) mutation and expression features were first extracted from the full set of five TMP project datatypes. 2) the mutation features were aggregated into a single binary indicator for each gene by the feature-code-embedded HUGO gene identifiers over the four MC3 mutation types: composite (COMP), non-silent (nons), loss-of-function (LOF), and hotspot (HOTS). A gene was only considered not mutated if zero of its constituent features were indicated as not mutated as defined by a value of zero for all features measuring any of the four mutation types for that HUGO. 3) the mutation and expression profiles were sub-set to the intersection of interacting genes via mapping with the Pathway Commons .sif file as visually depicted in the initial steps of Fig. 3-1. Specifically, within each cancer type, the mutation genes were mapped to the "Source" column of the interaction (.sif) file whereas the expression genes were mapped to the "Target" column.

Next, the mutation profiles of these intersection sub-set genes were sorted in a two-stage process termed memo-sorting. Here, the tabular data were subset by subtype and transposed to genes as rows and samples as columns. In stage one, genes with the most instances of mutated samples were sorted to the top and in stage two, mutated samples were grouped as chunks of columns left to right beginning with the top row, most-mutated gene, and descending row by row. The caching of columns left to right within each chunking iteration is the connection with the canonical memo-sort concept in computer science, hence the name. The result of plotting these

subtype-specific organizations of mutation patterns jointly reveals the patterns of mutual exclusivity of mutated genes across samples within subtypes and the co-occurrence of frequently mutated genes between subtypes. Taking the union of the 10-most frequently mutated genes over all the subtypes with each primary tumor type yielded the mutation feature set component of each Pathway Commons-based interaction graph.

Gene expression feature sets were obtained for each primary cancer type via application of a statistical feature selection algorithm called forward-backward early-dropping. This algorithm was implemented with the R package MXM[89]. The algorithm was applied to the expression profiles of the intersection sub-set genes again visually depicted in Figure 3-1. To control for overfitting and attain a diverse set of features representative of various sample combinations, the MXM feature selection method was run on sub-sets, with replacement, of 70 samples at a time. The compendium set of selected features was derived from a run of 250 repetitions within each cancer cohort. As a control, recursive feature elimination[93] (RFE) using the scikit-learn implementation was run within the same sub-sampling framework of 250 repeats at the 70-sample threshold. The feature selection was deployed on OHSU's Advanced Research Cluster, ARC using a shell script loop passing args to a SLURM script the argument array option to call the respective R and Python feature selection scripts.

Scikit-learn feature importance values were then calculated for each subtype over the ten most significant expression features within each cancer type. These feature values were evaluated against a custom class-specific implementation of the SHAP[94] feature importance algorithm as a control. The SHAP feature importance values were calculated with a Catboost classifier using 300 interactions and a learning rate of 0.01.

Interaction graph construction

The top-two mutated genes for each subtype within each primary tumor type were extracted for inclusion as the upstream genes into each interaction plot. For the expression features, a p-value was calculated against random probability with a binomial test for each gene and genes occurring at a frequency above a significance threshold of 0.01 added to the graph as downstream

interactors from the mutation genes. This process was instantiated by writing .graphml objects to disk via the NetworkX Python package for export to plotting with Cytoscape[82].

Predictive signal by GDAN-TMP datatype comparison

To prepare data ready for machine learning in the GDAN-TMP predictive signal comparison of the MUTA, METH, and GEXP datatypes, the code repository was cloned from:
https://github.com/NCICCGPO/gdan-tmp-models.git
Combinations of models, datatypes, and cancers contained in the results file

>> model_info.json

file were retrieved from: https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022
per repo instructions. Feature sets were retrieved for all available model types within each cancer type via the

>> tools_ml.get_model_info()

function for the datatypes MUTA, METH, and GEXP. Molecular profiles for each TMP model-specific feature set were constructed from mapping the feature lists to the original raw TMP feature profiles. A scikit-learn random forest (RF) model with default hyperparameters was trained and tested on 30 data-splits for each TMP model's feature set within each of the three data types within each of the 26 cancer types. Each of these predictions was scored with the F1 measure using scikit-learn's *metrics* library to account for class imbalances of the sample distributions inherent in most of these TCGA cancer types. The error for these prediction results is reported as standard deviation in Fig 3-10.

Onco-screen predictive signal comparison

The memo-screening gene lists used for comparison in determining onco-status against the COSMIC, KDxL, and Ampliseq50 oncogenes were derived from the GDAN-TMP mutation

genes ranked by frequency of mutation using the memo-sort algorithm across the samples within each subtype; the output of the mutation mutual exclusivity analysis. The union of the 10 most-frequently mutated genes in each subtype were aggregated within each primary tumor type to build the primary tumor-type-specific memo-screen gene lists.

To determine sample-level onco-status, the four screening methods - COSMIC, Ampliseq-50, KDxL, and GDAN-TMP memo-screen oncogene list filters were run against the 157 GEXP feature sets composed of the 126 top-performing GDAN-TMP mRNA sets along with 26 de novo expression sets derived from the top-10 genes identified with our MXM sub-sampling framework for predicting F1 score and sample-level prediction confidence. The predict_proba method within scikit-learn was used to calculate the confidence scores.

Chapter 4 — SyntheVAEiser: augmenting traditional machine learning methods with VAE-based gene expression sample generation for improved cancer subtype predictions

Brian Karlberg, Raphael Kirchgaessner, Jordan Lee, Matthew Peterkort, Liam Beckman, Jeremy Goecks, Kyle Ellrott

Published in: Genome Biology, 2025

Abstract

The accuracy of machine learning methods is often limited by the amount of training data that is available. We proposed to improve machine learning training regimes by augmenting datasets with synthetically generated samples. We present a method for synthesizing gene expression samples and test the system's capabilities for improving the accuracy of categorical prediction of cancer subtypes. We developed SyntheVAEiser, a variational autoencoder based tool that was trained and tested on over 8000 cancer samples. We have shown that this technique can be used to augment machine learning tasks and increase performance of recognition of underrepresented cohorts.

Background

Machine learning (ML) has become common in genomics as a means of modeling with complex biological data[95,96]. Across numerous publications from The Cancer Genome Atlas (TCGA)[97], bulk RNA-sequencing has been shown as a robust way for defining cancer subtypes[98–102]. Bulk RNA-seq based signatures have been translated from basic research into FDA approved diagnosis used in the clinic[103,104]. While this technique has found use in more common cancers, issues begin to arise with more rare cancer variants. Small sample counts within genomics datasets can impede model performance because of the high dimensionality of the feature space and imbalanced classes. In training performance analysis, we have found that about 120 samples are often needed before a machine learning recognizer can achieve best possible performance. For rare cancers, the resulting low sample counts of these omics datasets limit the capability of machine learning to improve patient outcomes. In this paper, we show that synthetic sample generation is one possible mechanism to mitigate these issues.

Synthetic data have been shown to improve the sample efficiency of learning across diverse domains such as image processing, physics modeling, and neuroscience[105]. We propose to apply data synthesis methods to augmenting transcriptomic data sets and improve the performance of a variety of prediction tasks. Neural networks with multiple hidden layers known as deep learning (DL) models combined with transfer learning techniques have demonstrated utility across a wide

range of modeling applications within the rapidly evolving field of ML[106]. Generative deep modeling has emerged as a route to generate new samples and works by creating representations of complicated, high-dimensional probability distributions[107].

A variational autoencoder (VAE) is a feed-forward neural network that approximates a function for mapping high dimensional variables into representative, or latent, variables of a reduced dimension[108–110]. Continuous normalizing flows and generative adversarial networks (GANs) are similar generative models to VAEs[111]. VAE training is an unsupervised machine learning technique, and is unaware of any outside labels, such as cancer subtype, and is only concerned with organizing a low dimensional latent space based on the sample data. The defining characteristic of a VAE is stochastic backpropagation[108] which allows the model to overcome the accuracy and scalability challenges of modeling high-dimensional data.

The aims of this study were to **1** build a generative model for creating synthetic gene expression samples, **2** develop an algorithm for creating synthetic samples based on combining these latent representations of multiple parent samples with a labeled dataset, and **3** integrate this generative modeling framework with a traditional ML classifier to robustly quantify the improvement in predictive power from the addition of synthetic samples. This will demonstrate that VAEs can be trained on pan-cancer data and use that information to extrapolate into new tissue types. In these new cohorts, a minimal set of examples can be used to extrapolate a larger training set, and that extended training set can help to improve the performance of machine learning methods.

Traditional reasons for developing synthetic data sets for genomics and imaging include insufficient sample sizes, too many or too few features, disproportionate feature to sample size ratio, and the class imbalance problem[112]. Methods used to deal with class imbalance can be seen as analogous to synthetic sample generation methods. SMOTE[113] is the canonical method addressing the class imbalance problem. This method seeks to improve classifier performance by undersampling the majority class and oversampling the minority class. The minority samples are not directly sampled with replacement, rather the feature values of two or more samples are recombined with the feature value differences multiplied by a random number between zero and one to generate novel samples. However, in cases of high feature dimensionality and low

signal-to-noise such as gene expression applications, the performance of SMOTE has been shown to both lack robust performance and be classifier dependent[114]. In cancer imaging, synthetic data have advanced to the point where a Synthesis Study Trustworthy Test (SynTRUST) has been proposed as a meta-analysis framework to address specific challenges across research and clinical care[115]. For computer vision tasks, there are a multitude of techniques for data augmentation[116] including skin lesion image synthesis[117]. Generative methods have been shown to be robust across multiple data types, and as our research shows, this trend continues with transcriptomic data.

In the field of transcriptomic sample generation, there are previous publications outlining the use of GANs to create synthetic mRNA samples and improve prediction tasks[118]. These methods utilize noise or alternate omics inputs to generate new synthetic samples. Our method differs from these approaches in how the basis for new samples are seeded. Rather than utilizing random noise for permuting existing models, our model mixes features of multiple samples in latent space before reconstructing a new synthetic sample. Importantly, the mixing of features in the low dimensional latent space occurs between samples of the same target label. This ensures that each synthetic sample is effectively a high dimensional average of similar elements and avoids mixing samples from different classes.

When compared to other machine learning methods, deep learning methods are viewed as “black boxes” that produce predictions based on uninterpretable methods. Many times, especially when thinking about clinically oriented tasks, non-DL machine learning methods can provide interpretable models that can be connected to specific biological elements. These more interpretable models may be seen favorably for translational use cases, but may lack the ability to extract additional information from large sample populations in the same way that deep learning methods are able. For this study, we demonstrated that traditional ML can benefit from adding synthetic data generated by a VAE. By combining the pan-cancer training set, the VAE model is able to learn common patterns seen across multiple cancer types, and use that information to enrich a traditional machine learning task, even if that problem is only specific to a single cancer type. Because these performance gains are seen in methods, such as random forest (RF) based

models, that are commonly viewed as being interpretable, the results of this technique can be interrogated.

Results

Generative model overview

A new method combining a VAE with a RF classifier and a corresponding software tool for sample synthesis was developed for applications in ML applied to gene expression data. Our dataset, based on samples from the TCGA, was structured for supervised categorical prediction where each sample was labeled with a cancer subtype within 25 primary tumor types based on gene expression profiles. In total, the 25 different tumor types were segmented into 99 molecular subtypes. For example, breast cancer (TCGA code BRCA), is subdivided into luminal A, luminal B, basal, and HER2[119]. A transfer learning framework was applied for training the VAE on a sample set composed of all TCGA samples using a tumor sample holdout strategy, Fig. 1A. This involved a sequence of training and fine-tuning a VAE and using a RF classifier to compare the predictive accuracy of the data modes. The VAE was never trained on or received any information about tissue type or cancer subtype. So in the case of the BRCA cohort, the trained VAE was not presented with any BRCA samples, but rather learned the patterns from all other available cancer types. Thus in that experiment, BRCA could be viewed as a rare cancer that had never been encountered. A VAE model is trained to compress gene expression data into a latent space and then decompress a faithful copy of the original signal. This encoder/decoder pair is then used to translate data into a "latent space" where values can be altered and decompressed back into "normal space" to create new samples. For our cross fold experiment, we produced 25 separate encoder/decoder pairs that each ignored a single cancer type. The sample generation pipeline was built around the Tybalt VAE[120], Fig. 1B. The corresponding feature engineering pipeline takes the intersection of genes across cancer types and reduces the feature counts with mean absolute deviation. Original gene counts varied by primary tumor type are shown in Supplemental Table 1.

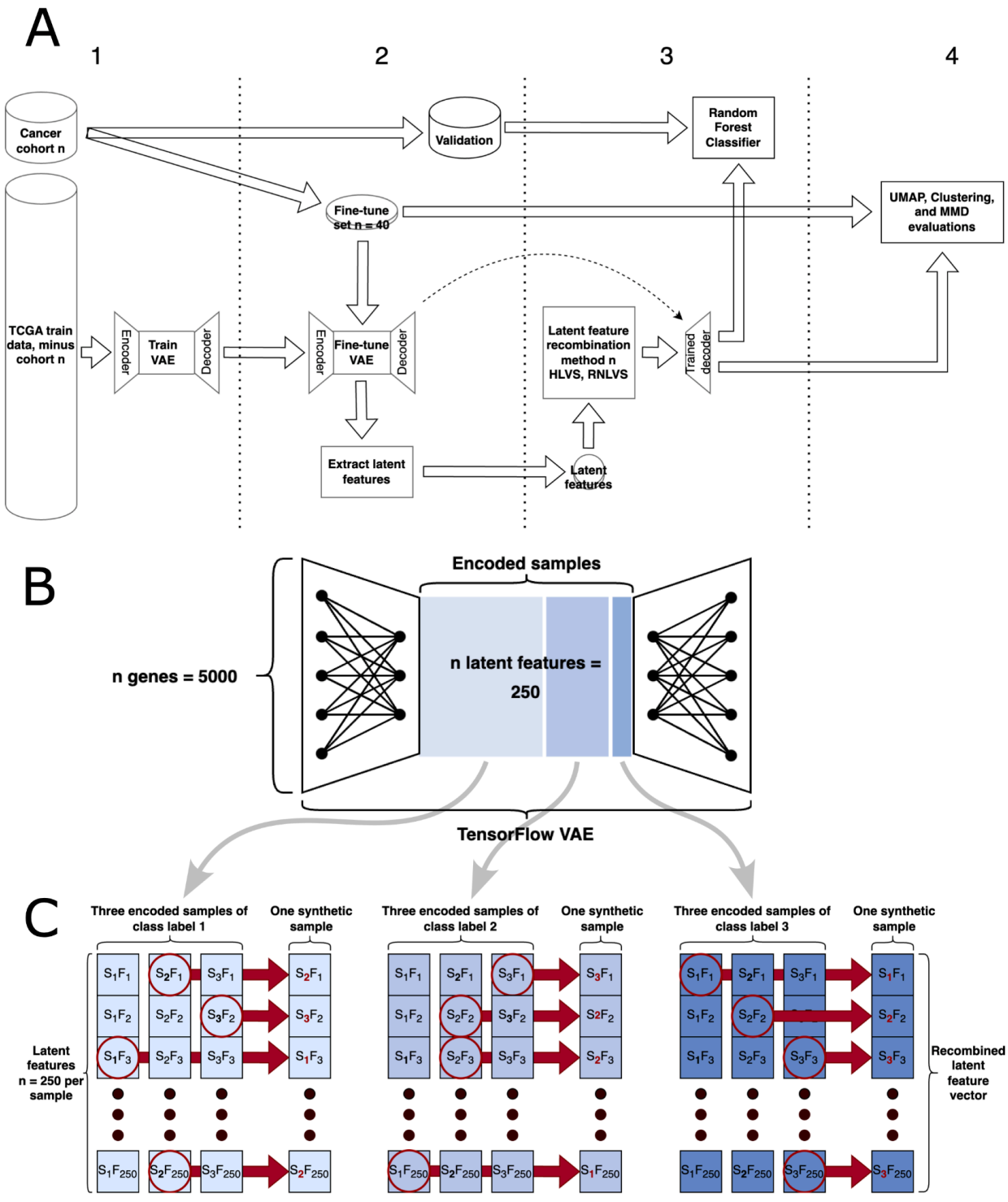


Fig. 4-1 Overview of the synthetic TCGA gene expression sample generation pipeline. **A** One cancer cohort at a time is designated for sample generation and removed from the TCGA sample set. The Tybalt VAE adapted from Way and Greene[120] is trained on these TCGA samples and then fine-tuned on 40 samples from the designated cohort_n. The remaining samples from cohort_n are used as validation. The latent feature values of three randomly selected samples from within each subtype are randomly recombined to form a latent sample feature vector which is then decoded with the trained decoder to

generate a synthetic sample with feature dimensionality restored to that of the 5000 input genes. This latent feature value recombination and decoding process is repeated to generate 200 samples per subtype per validation split. The random forest classifier is trained five times, each time predicting on the entire held-out validation set to return a subtype prediction accuracy with quantified error. The train-validation split point at cohort_n and ensuing processes comprise a single experimental replicate which is repeated 25 times per cancer cohort. **B** Input gene expression features and latent dimension of the Tybalt VAE component of the pipeline. **C** Depiction of the three-sample version of the HLVS algorithm operating within each labeled class

Using our hybrid DL/traditional ML synthesis and analysis pipeline, we analyzed the effect on subtype prediction performance with the RF classifier for 25 cancer types, using the cohort holdout strategy, where specific cancers were limited to 40 samples for training the RF classifier with all other samples from that cancer type used for performance validation. Effectively, our protocol simulated 25 separate rare cancer cases by restricting the RF training set to 40 samples. This process was repeated across these 25 cancer types, generating 200 additional samples per subtype to augment the 40 original samples. Thus, the number of synthetic samples generated varied for each primary tumor type, varying from 400 for the binary cancers up to 1400 for gastroesophageal (GEA) with seven subtypes. Using the validation sets, we measured F1 score performance improvement on the prediction of held out samples by a mean of 6.85% and a maximum improvement of 13.2% in lung squamous cell carcinoma (LUSC).

The transfer learning strategy involved first training the VAE on the gene expression data for approximately 8000 samples from the TCGA dataset, holding out one specific cancer type for testing. After the initial training, the VAE was fine-tuned on a subset of 40 randomly selected samples from the testing cancer type. The rationale for using this threshold of 40 samples for fine-tuning and sample generation across the 25 cancers was to balance a simulated reduced sample set with diminished accuracy while still having enough samples with which to generate quality synthetic samples. Reducing the batch size parameter of the VAE when transferring the model from training on a relatively large dataset to fine-tuning on a smaller dataset was identified as an important factor in learning a model capable of generating samples that improved predictive accuracy.

The effect of the quantity of training varied by cancer and could be inferred by the shape of the learning curves. In these data, the ratio of sample sizes in the training sets to fine-tuning sets was

approximately two orders of magnitude and the number of epochs utilized in the training phase was observed to be a primary parameter in controlling the performance results of the generated synthetic data. This can be approached in absolute terms of training and fine-tuning epoch counts as well as from a ratio perspective. To investigate these effects, the quantity of TCGA training epochs was varied while holding the fine-tuning epochs constant at 150. The proportion of pan-TCGA training epochs to fine-tuning epochs on the cohort targeted for sample generation was observed to affect model performance asymmetrically across cohorts thus is a key point of consideration for generalizing this model to data with other distributional characteristics.

Synthetic sample generation

We tested two methods for synthetic sample generation: Random Noise Latent Variable Samples (RNLVS) and Hybrid Latent Variable Samples (HLVS). For a baseline, we deployed RNLVS which modulates samples with random noise in the latent space to create synthetic samples that are slightly perturbed from their original parent sample. We contrasted that method against HLVS which is designed to generate a synthetic sample of a specific subtype. It does this by randomly recombining the latent feature values of two or three samples from the same subtype into a novel latent feature vector (Fig. 1C). Both two- and three-sample versions of HLVS were tested. The rationale for using three samples was to balance a generalized subtype representation based on a greater number of samples with the fact that for cancers with many subtypes, random samplings would begin to return one or zero samples of the rare subtypes as test set sizes decreased which negated the possibility of latent feature recombination. The decoder component of the VAE was then used to project each HLVS vector back into gene expression space. To validate the performance of RNLVS vs. HLVS derived synthetic samples, we tested machine learning models derived from cohorts generated using the two methods. We noted a marked improvement in performance using HLVS derived samples, as shown in Fig. 2.

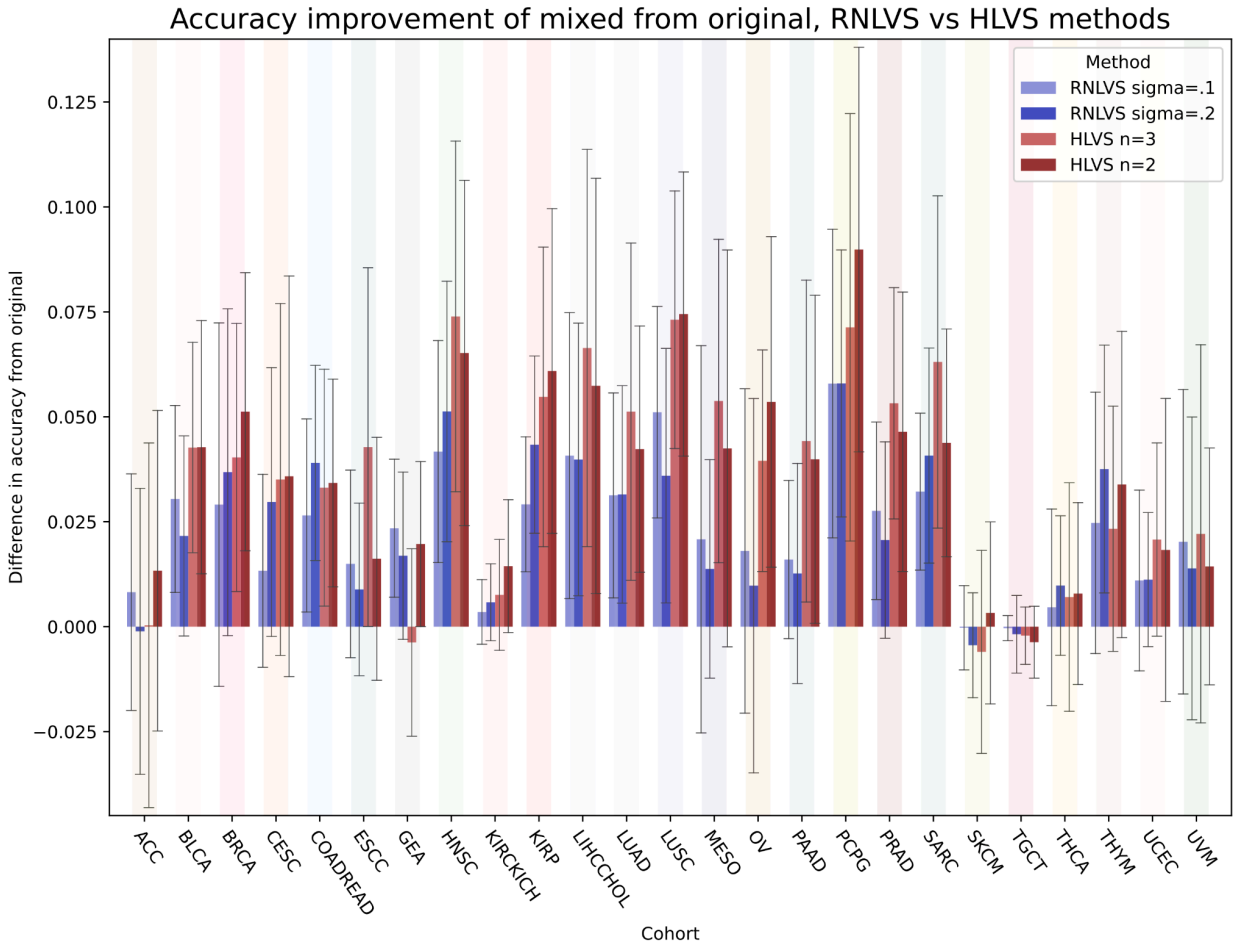


Fig. 4-2 Comparison of cancer subtype prediction accuracy improvement between the two RNLVS methods and two HLVS methods tested. With feature sets and model parameters fixed across primary cancer types, the HLVS methods return synthetic samples that result in greater accuracy improvement for 21 out of 25 cancer types

For both the RNLVS and the HLVS sample generation methods and for each set of the experimental replicates, 200 samples were generated within each subtype for each of 25 replicates of 40 randomly selected training samples for a total of 5000 synthetic samples per subtype per replicate set. The trained decoder contained both pan-TCGA information as well as information from all subtypes via the 40 samples selected from within the cohort designated for sample generation. This was the result of the transfer learning design of the experiment in leveraging the combined learned representation of what a molecular cancer subtype is in general, with how molecular subtypes within a primary tumor cohort differed from each other.

After the synthetic samples were generated, they were mixed with the original training samples and then used to train a traditional ML RF classifier to predict on a validation set to assess performance of the sample generation. Across the 25 cancer subtype learning tasks, this resulted in improved classification accuracy for the majority of cancers. In our testing, 16 out of 25 cancers returned a statistically significant improved subtype prediction raw accuracy at a p value threshold of at least 0.05 as a result of mixing with the original 40 samples all of the 200 synthetic samples per subtype across the 25 experimental replicates.

Synthetic sample assessment

To quantify and compare the quality of the sample embeddings and generated synthetic samples, a Scikit-Learn RF classifier was selected based on its observed performance as a traditional ML method[121,122]. The default hyperparameters of the RF classifier were used. Within each cohort and experimental replicate, the RF was first trained on the 40 original samples then used to predict on the validation set. This training of the RF was repeated on the VAE reconstruction of the same 40 samples once they had been encoded then re-coded back to gene expression space at the end of the fine-tuning epochs. The RF trained on these re-coded samples was then used to predict on the same validation as was used to evaluate the original 40 samples. Finally, this RF training and validation scheme was repeated on the pure synthetic and the mixture of the 40 original samples with the 200 synthetic samples per subtype. Raw prediction accuracy [Scikit-Learn metrics] was utilized for these comparisons. For each of these four data phases, the RF model was trained on the test set five times and used to predict on the validation each time to control for stochasticity in the RF model. The results of these five runs were averaged. A comparison of the performance results for two configurations within both the HLVS and RNLVS latent feature modification methods across the 25 TCGA cancers is shown in Fig. 2. The error shown is standard deviation and the magnitude relates to subsampling effects of low sample sizes. This illustrates heterogeneity within cohorts and number of subtypes within cohorts.

Once establishing this baseline configuration of the VAE training to attain predictive accuracy improvement for the majority of cohorts, learning curves were generated. The original and mixed datasets were subsampled in incremental steps with the random forest again repeated five times

and averaged on each subsample set at each increment size. Learning curves for four selected cancers that returned increased raw accuracy from the addition of synthetic samples are shown in Fig. 3 with learning curves for the other 21 cohorts in Supplemental Fig. 1.

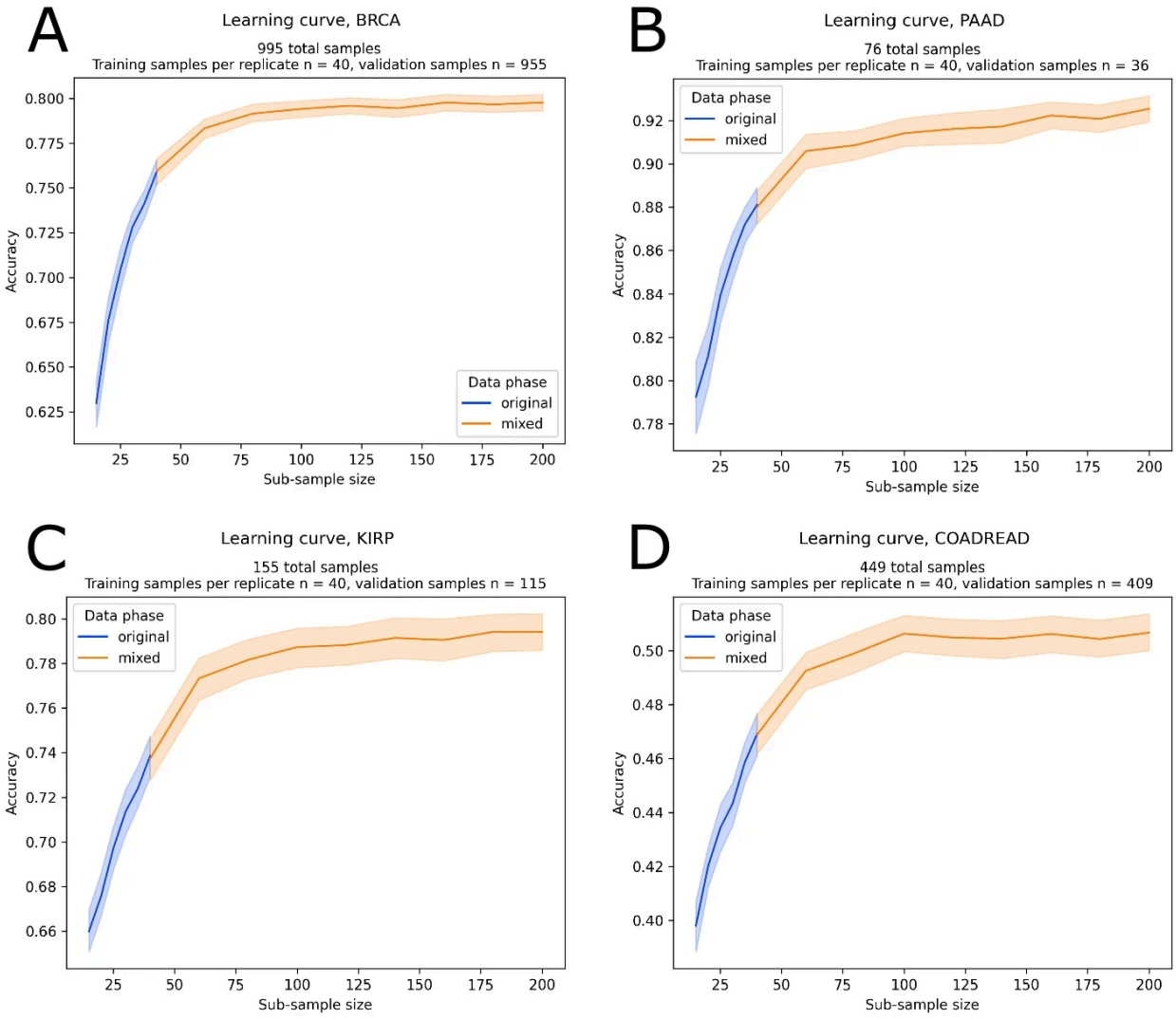


Fig. 4-3 Learning curve comparisons of individual cancers; predictive accuracy as a function of sample size aggregated across 25 experimental replicates. Original sample sets in blue showing subsampled accuracy growth up the 40 sample training threshold. Continuation of learning curves at larger sample counts with subsampling mixed original/synthetic sample sets in orange. **A** Breast invasive carcinoma learning curve, relatively smooth improvement in predictive accuracy with addition of synthetic samples up to a peak at approximately 150 samples. **B** Pancreatic adenocarcinoma, with 76 original samples shows a gradual improvement in predictive accuracy observed past 100 samples. **C** Performance improvement behavior of adding synthetic samples for kidney renal papillary cell carcinoma with 76 original samples,

third smallest cohort. **D** Learning curve for colorectal adenocarcinoma, with more challenging to predict subtypes showing plateau in improved performance at around 50% accuracy

To characterize the similarity of the gene expression value distributions within the respective subtype label categories for the synthetic samples with the original samples from which they were generated, maximum mean discrepancy (MMD) was calculated for each pairwise combination of samples within three cancer types representing a range of subtype counts shown in Fig. 4A. A scatter plot of 2D UMAP dimensionality reduction was applied to visualize clustering of samples by subtype with mixing of original and synthetic data, Fig. 4B. If the distance between the expression value distributions of the original and synthetic samples is minimal, it would be expected that original and synthetic samples would cluster randomly within each subtype, with subtype status driving the clustering. Affirmingly, when applied to a mixed set of the original and synthetic samples, this clustering shows general separation of samples consistent by subtype as illustrated in Fig. 4C. Clustering of synthetic samples within a given subtype may be driven by the synthetic gene expression vectors being based on combinations of latent values from real samples resulting in synthetic samples being a non-linear interpolation of real samples. Although some degree of clustering by synthetic and original sample status is observed, despite this limitation, there is still an improvement in subtype predictive accuracy with either the pure synthetic or mixed data sets. A full survey covering another 22 TCGA cancer types can be found in Supplemental Fig. 2.

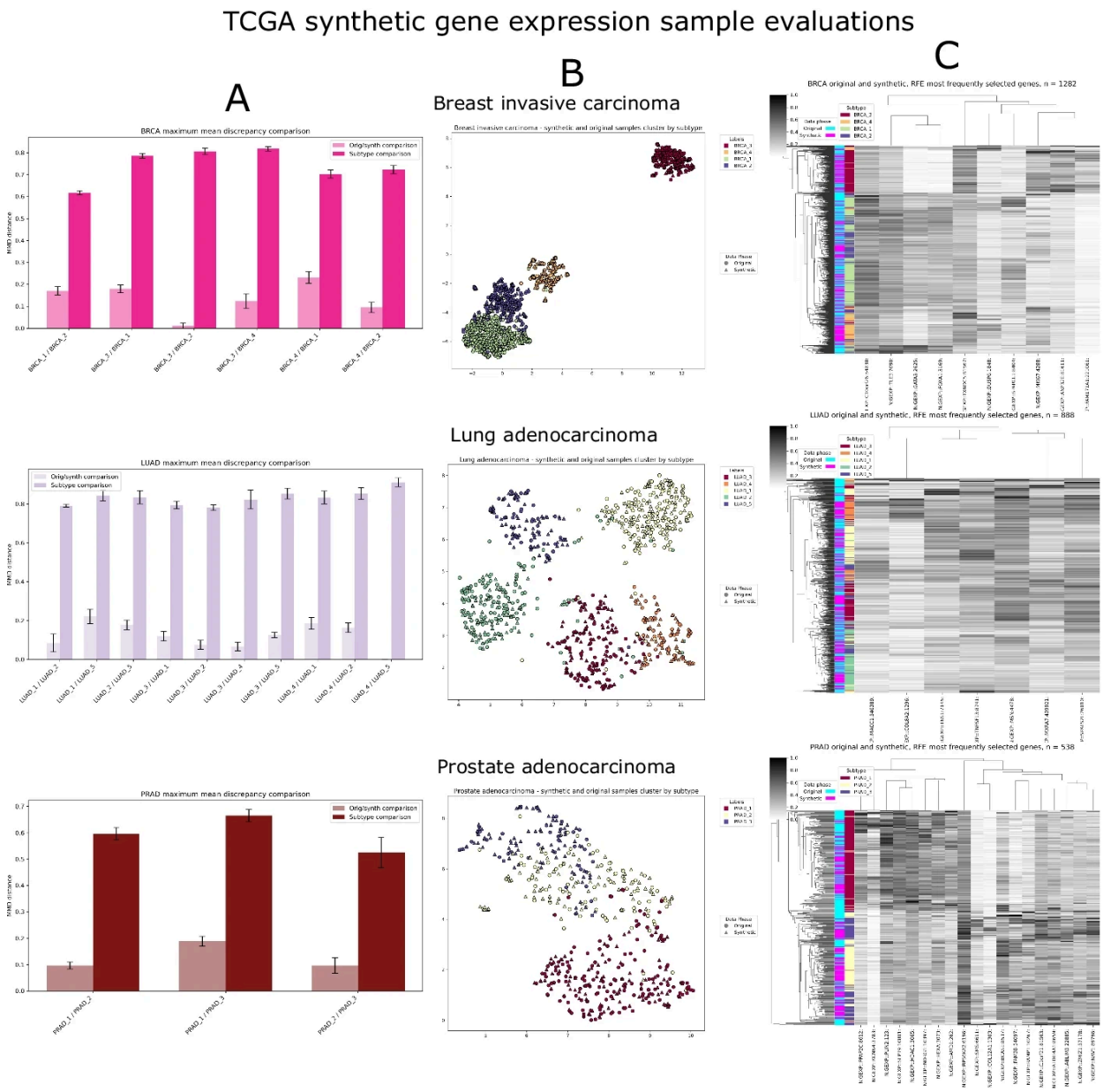


Fig. 4-4 Comparison of evaluation methods for synthetic vs original samples. **A** MMD statistics for each pair of cancer subtypes within each primary cancer type comparing the difference of gene distributions with samples split by subtype vs. samples split by original/synthetic. **B** Scatter plots of 2D UMAP projections showing interspersed clustering of original and synthetic samples separated by cancer type. **C** Cluster maps showing propensity of samples to cluster by subtype with interspersion of synthetic and original samples within each subtype. Color bars on left in pink and light blue show original or synthetic sample status and saturated color bars on right show subtype sample status

An additional quantitative inspection of the original and re-coded gene expression values was conducted with a root mean squared deviation (RMSD) comparison.

$$rmsd = \sqrt{mean((predictions - targets)^2)}.$$

For each of the 40 samples in each experimental replicate, RMSD was calculated across the 5000 genes for the original and re-coded versions of the values. One thousand RMSD values, 40 samples times 25 replicates, for each cohort are shown in Fig. 5.

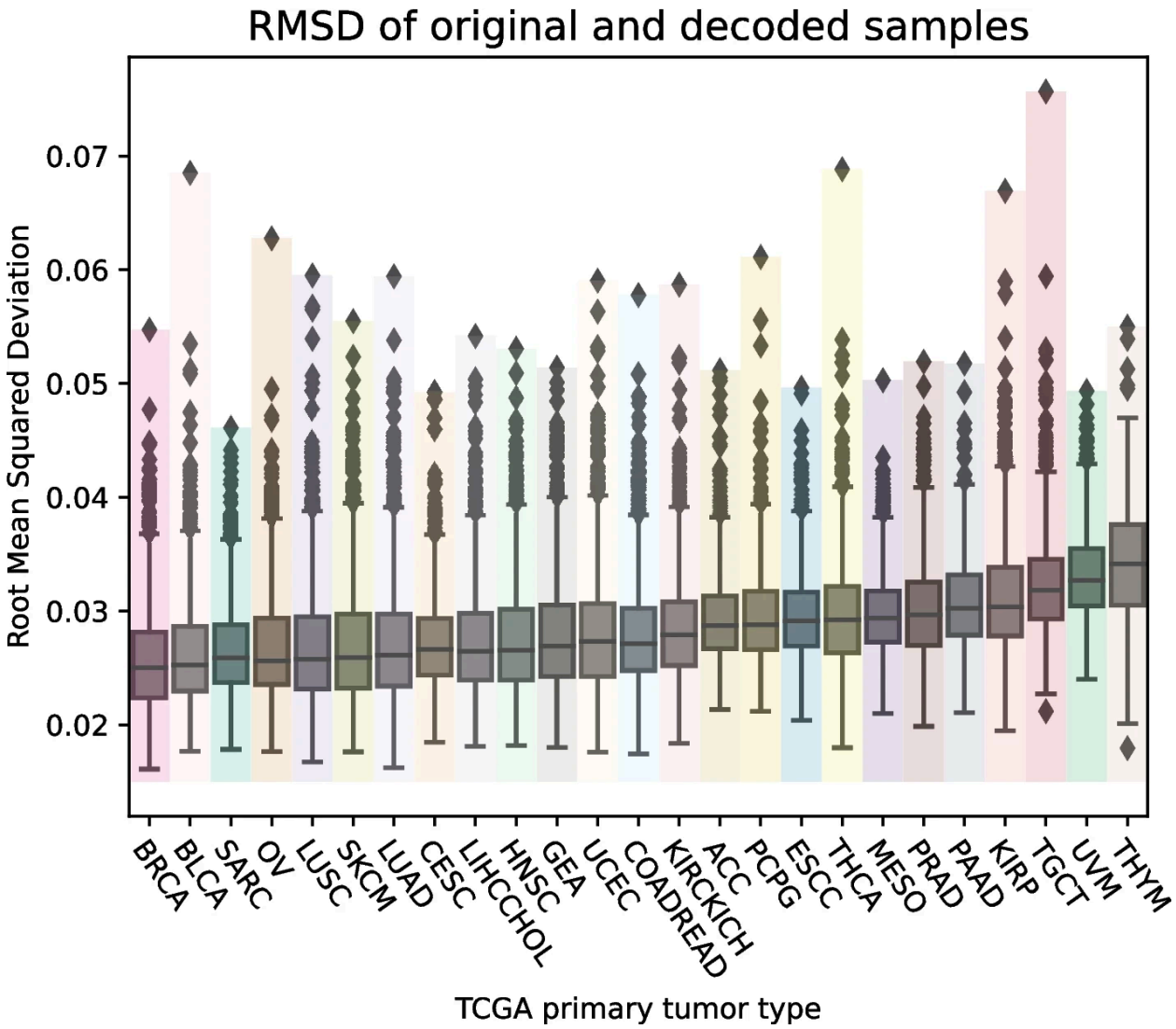


Fig. 4-5 Correlation of gene expression RMSD with the difference in prediction accuracy by primary cancer cohort. The gene expression RMSD is the average root mean squared deviation across each sample's 5000 gene expression values input to the VAE with the corresponding re-coded values of encoding and decoding these input values. The y-axis, delta accuracy is the change in average subtype

predictive accuracy across the 25 replicates of 40 input samples vs. the average of the predictions at 140 and 160 sample size mixed sample sets of the original 40 samples and synthetic samples within each experimental replicate

Recursive feature elimination, a statistical feature selection algorithm, was applied to identify specific gene features of importance within the original, re-coded, and synthetic samples. For three selected primary tumor types, BRCA, LUAD, and PRAD, the intersections of features selected across the three data phases are presented in Fig. 6A. Consistency in the specific features selected from each phase of the data would be expected in the case of consistency in the gene expression values across the data phases. For these three cancers, this pattern of consistency was observed—in BRCA, 71 features were commonly selected across all three phases of the data compared with 39, 17, and 16 features commonly selected across the pairwise combinations of the data phases. Eighty-four and 66 features were commonly selected across all data phases for LUAD and PRAD, respectively, with lower numbers again observed for any pairwise combinations of data phases. This observation indicates biological consistency of the synthetic data with the original samples. Permutation-based feature importance scores were calculated within each of the three data phases for each of these three cancers for these selected features shown in Fig. 6B. The gene FOXA1 scored in the top three of the most important features for BRCA across all data phases and SEPT9 scored in the top three across all phases for PRAD.

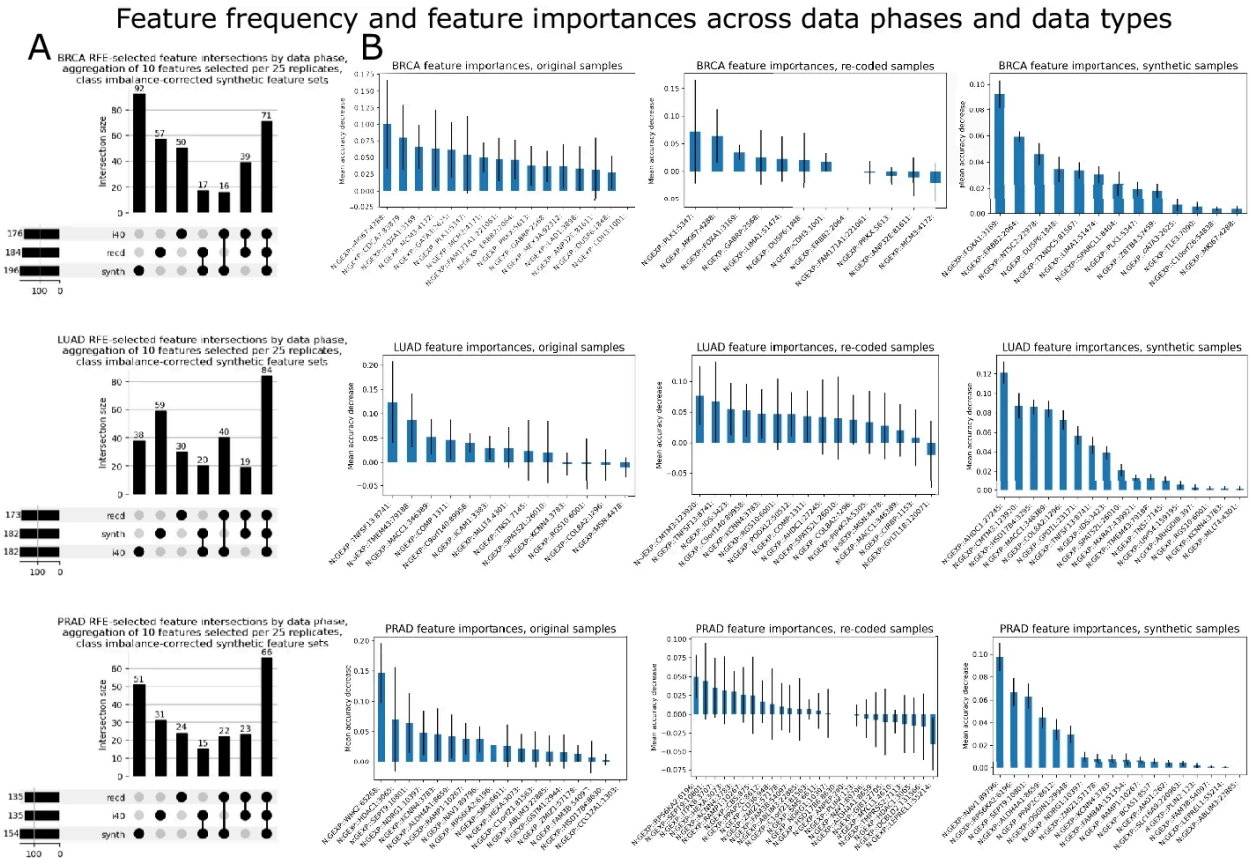


Fig. 4-6 Selection frequency and importance comparisons of feature sets. **A** Intersections of features across original, re-coded, and synthetic samples. **B** Feature importance scores calculated with Scikit-Learn Permutation Importance algorithm for features selected three or more times across the 25 experimental replicates

For further validation of the VAE-based genomic samples, we tested the algorithm on single-cell data, by using oligodendroglioma intra-tumor heterogeneity gene expression data obtained from the Broad Single Cell Portal[123]. To create two distinct cohorts, this data was filtered for malignant and Microglia/Macrophage cell labels which were the analog to the cancer subtype labels in the original experiments. The Microglia/Macrophage class was down-sampled to 250 samples to approximately match the 235 samples in the malignant class. Filtering samples with missing expression values from this set yielded a prepared set of 418 samples with 235 samples of the Microglia/Macrophage class and 183 samples of the malignant class. The 23,686 raw gene features were reduced to the 5000 gene features with the same greatest mean absolute deviation method utilized in the original experiments. The data was randomly split into a pre-training set of

268 samples and a fine-tuning set of 150 samples for input to the VAE sample generation tool in its same configuration from the original experiments. The generated data were evaluated against the original data with UMAP clustering (Supplemental Fig. 3), showing synthetic and original single-cell samples clustering by cell type and not clustering by real or synthetic status.

Discussion

In order to test the robustness of our method, we benchmarked the recognition of cancer subtypes as defined by the TCGA cohort. Because each tissue type has extremely different dynamics, and the subtypes within each of these cancers are defined by different rules, this allowed us to perform robust benchmarking in translation, by removing entire cancer types from the original training set. Additionally, the dataset has cohorts of extremely different sample sizes, with groups with 995, such as the case of breast invasive carcinoma (BRCA) and as few as 74 in the case of mesothelioma (MESO) and uveal melanoma (UVM). In our tests with the TCGA dataset, the sample size limitation is most pronounced in cancers with rare subtypes such as bladder urothelial carcinoma (BLCA) or kidney renal papillary cell carcinoma (KIRP), primary tumors with subtypes containing less than 10 samples. Using a leave-tissue out cross fold strategy, every cancer type was tested as if it was a rare cancer type. Our method to increase sample sizes of rare, molecularly defined subtypes to solve the class imbalance problem could be of particular utility for feature sets reduced to the number of samples required to train accurate models.

Augmenting datasets with synthetic samples created with the HLVS methods outperformed the RNLVS derived samples in 20 out of 25 of the specific machine learning tasks tested. The three-sample and two-sample variations of the HLVS method performed comparably well with average predictive improvement over the original samples of $3.64\% \pm 0.04\%$ and $3.67 \pm 0.04\%$ percentage points, respectively. Although random noise methods combined with generative modeling improved performance for the majority of tested cancers, the performance gains were greater across most cancers with the combination of generative modeling and HLVS methods.

This study sought to leverage the representation learning capabilities of generative modeling with the interpretability of traditional ML to develop a method for transcriptomic sample

generation. The software tool developed can be directly applied to supervised categorical prediction tasks with gene expression data sets and potentially adapted to other transcriptomic based ML tasks including regression. This was an improvement on previous methods in robustness for this type of genomics prediction task characterized by a large ratio of features to samples. By using transfer learning techniques to train a model on data related to the fine-tuning data and final prediction domain, the model is less prone to overfitting. The training method utilized in this study was to include all of the TCGA cohorts, except the cancer type designated for testing, to prepare the VAE for fine-tuning. The RMSD statistics characterizing the reconstruction values between the best-fit cancer, BRCA, and poorest-fit cancer, THYM, showed that the mean of every tested cancer was within the error of every tested cancer. This demonstrates generalizability of a transfer learning strategy where fewer epochs are used for training than fine-tuning and the batch size is reduced in the fine-tuning from the training.

For the benchmarking observed in Fig. 2, issues beyond sample generation likely prevented SKCM and TGCT from improving their performance. The subtypes in skin cutaneous melanoma (SKCM) were originally defined using mutation markers. Training ML models on gene expression fails to capture that original information used for defining the subtyping, and instead relies on gene expression values that happen to be correlated with the subtype, rather than elements with direct biological implications. Similarly testicular germ cell cancer (TGCT) subtypes are largely defined by DNA methylation and miRNA[124]. In these cases, boosting the population of gene expression data will do little to better illuminate the underlying biology.

To quantify the similarity of the synthetic and original data, maximum mean discrepancy (MMD), a nonparametric distance statistic that is robust in comparing sample groups comprising different distributions[125], was calculated for each subclass pair within three primary cancer types of differing numbers of subtypes. For all subclass pair comparisons, the distance between subclasses was significantly greater than the distance between the original and synthetic samples as shown in Fig. 4A. This observation is reinforced with UMAP clustering behavior shown in Fig. 4B, where original and synthetic samples cluster uniformly within each cancer subtype. The sample cluster map of gene expression value experiments, seen in Fig. 4C, also showed aggregation of samples within subtypes of mixed synthetic and original data.

The feature selection experiments reveal a greater intersection of features across the original, re-coded, and synthetic samples than within any pairwise combination of these three phases as shown in Fig. 6A. This observation is validating of both the model encoding and the synthetic data.

The feature importance scores indicate reduced error associated with the synthetic data compared with the original and re-coded feature importance scores as shown in Fig. 6B. This effect is driven by improved statistical power of synthetic data sets and the solving of the class imbalance problem with 200 synthetic samples per cancer subtype vs. 40 total original samples within each replicate. This demonstrates the potential utility of the method to improve confidence in biomarker target identification for rare cancer subtypes.

Conclusions

This work demonstrates that generative models based on neural networks can be combined with traditional ML as an effective means to generate synthetic gene expression samples. This allows for information from other tissue and cancer types to provide priors for learning patterns in a new cohort. Rare cancers, which traditionally see much lower rates of collection and sequencing, can benefit from augmenting their dataset. Additionally, non-DL machine learning methods, traditionally seen as more trustworthy or easier to interpret than DL models, can still benefit from these methods.

Methods

Data provenance and feature engineering

The data utilized for developing this sample generation method and software tools were derived from a TCGA-based curated dataset from the Tumor Molecular Pathology working group and can be downloaded from the NCI's Genomic Data Commons[10] [https://gdc.cancer.gov/about-data/publications/CCG-TMP-2022]. These data files were tabular

comprising 8009 samples across 25 primary tumor types and 99 subtypes. The gene expression features utilized in this study were down-selected via mean absolute deviation to the 5000 most differentially expressed features per the original Tybalt method[120]. The raw expression values were normalized with the Scikit-Learn MinMaxScaler function within each cohort and within each feature. Four of the cancers utilized in this study have only two subtypes making them a binary supervised classification problem whereas the remaining cancers are multiclass with three to seven subtypes per primary tumor type.

Generative modeling framework

The sample generation model shown in Fig. 1 was built around a variational autoencoder (VAE) adapted from[120]. A latent feature dimension of 250 was used for all experiments and all experiments used 150 epochs for model fine-tuning. One cohort at a time was designated for generating synthetic samples and removed from the combined TCGA set. The VAE was then trained on all of the remaining TCGA samples for 1, 2, 3, 4, 10, 20, or 30 epochs. The batch size was set at 50 for each of these initial TCGA trainings. From the cohort selected for sample generation, a training set of 40 samples was randomly selected without replacement. The remaining samples were used as a validation set of size $nv = n - 40$. The various epoch-count and feature set versions of the TCGA-trained VAE were then each fine-tuned for 150 epochs at batch size of 10 on the 40 samples within each replicate. A learning rate of 0.0005 was used for both the TCGA training and fine-tuning steps. This framework is represented symbolically in Algorithm 1.

Given N samples within Γ cohorts such that for each cohort $\Psi \subset \Gamma$, contains $\curlywedge$ number of classes ψ_i:

ϱ = an experimental replicate
$\rho = 25$, n replicates to repeat for each Ψ
λ = a subset of Ψ for training VAE
$n \leftarrow 40$, n samples for fine-tuning and synthesis within Ψ
$\gamma \leftarrow 50$, random sample repeats to attain minimum ν
Θ = Latent sample vectors of Ψ
θ = Latent sample vectors of ψ_i
$\nu \leftarrow 3$, n samples within θ from which to generate a synthetic sample
$\vartheta = \nu$ samples from within θ
dim $\vartheta = (n, 250)$
$\underline{\omega}$ = a synthetic latent feature vector
ω = a synthetic latent feature value
Ω = Synthetic latent feature vector set from Θ
$\varphi \leftarrow 200$, n samples to generate for each ψ_i

---

for ϱ **to** ρ **do**
| λ = sample(Ψ, n)
| **while** min(ψ_i *of* λ) $\geq \nu$ **do**
| | λ = sample(Ψ, n)
| | ι += 1
| | **if** $\iota = \gamma$ **then**
| | | continue
| **end**
| train VAE on Γ - Ψ and fine-tune on n
end

---

for VAE trained on Γ - Ψ and fine-tuned on n **do**
| D $\leftarrow$ decoder extracted from VAE
| **for** θ *in* Θ **to** $\curlywedge$ **do**
| | ϑ = sample(θ, ν)
| | **for** $\underline{\omega}$ in Θ **to** $\curlywedge$ **do**
| | | ω = sample($\underline{\omega}$, 1)
| | | Ω.append(ω)
| | **end**
| **end**
| D(Ω)
end

Algorithm 1 Generation of categorically labeled synthetic samples from the latent feature vectors of a variational autoencoder

The initial validation split of 40 fine-tuning samples within the cohort designated for sample generation defined each experimental replicate. Within each replicate, the samples not selected into the set of 40 for fine-tuning are designated as the validation set such that the number of validation samples varies by cohort because each cancer cohort contains a different number of total samples. Results for 25 replicates were produced for each cohort. Replicates returning less than three (or two in the alternate HLVS version) samples for any subtype within the random 40 cohort samples were rejected because this was the sampling threshold for the latent feature recombination algorithm, described below.

The training/validation split constituted an experimental replicate and was repeated 25 times for each cohort. If a training set contained less than three samples within a subtype, the sampling was repeated up to 50 times attempting to obtain at least three samples per subtype. The replicate was omitted if three (or two) samples were not obtained over these 50 repeats. The latent feature object was subset by subtype. Three samples at a time were chosen without replacement and sent to a function where the latent feature values from these three samples were randomly recombined into a novel latent feature vector. Two hundred synthetic samples were generated within each subtype for each primary tumor type. This 200 synthetic subtype sample by 150 synthetic latent feature object was returned to the original 5000 dimension feature space using the trained VAE decoder.

To evaluate the HLVS results, a set of experimental control results were generated with RNLVS derived from Gaussian noise injection. The effectiveness of Gaussian noise injection has been mathematically described for multi-layer perceptron neural networks in terms of the heat kernel and Taylor expansions [32][126]. This form of noise injection was implemented in the present study with sigma values of 0.1 and 0.2 for the Gaussian function applied to corresponding sets of latent feature values with a zero-floor or rectification operation to prevent negative expression values.

Within each experimental replicate, the 40 training samples were used to train a Scikit-Learn random forest model with default hyperparameters. This random forest was trained on the original training samples of the data then was used to predict on the validation set as to establish

a baseline accuracy score with which to compare with the synthetic samples. The process of training the random forest and predicting on the validation set was repeated for the re-coded, synthetic, and mixed sample sets denoted by the green, red, and orange arrows, respectively, in Fig. 1. The mixed sample set was the generated synthetic sample set blended with the original 40 training samples.

The imbalanced class problem was eliminated by adding 200 synthetic samples to each class. The result was that subtypes with relatively few samples were augmented with proportionally more synthetic samples.

For the comparisons of the distributions of the original and synthetic samples within the cancer subtype class pairs shown in Fig. 4A, the MMD formula utilized is given in Algorithm 2.

Given two sets of samples, *gexp_1* and *gexp_2*, and a kernel parameter γ:

$K\_XX \leftarrow$ rbf_kernel(*gexp_1*, *gexp_1*, γ)
$K\_XY \leftarrow$ rbf_kernel(*gexp_1*, *gexp_2*, γ)
$K\_YY \leftarrow$ rbf_kernel(*gexp_2*, *gexp_2*, γ)
$m \leftarrow$ number of rows in *gexp_1*
$n \leftarrow$ number of rows in *gexp_2*
$mmd \leftarrow (\sum(K\_XX) - trace(K\_XX))/(m*(m-1))$
$mmd \mathrel{+}= (\sum(K\_YY) - trace(K\_YY))/(n*(n-1))$
$mmd \mathrel{-}= 2*\sum(K\_XY)/(m*n)$
$mmd \leftarrow \max(mmd, 0)$
$mmd \leftarrow \sqrt{mmd}$

return *mmd*

Algorithm 2 Compute MMD

The UMAP clusterings of original with synthetic samples within each intended cancer subclass shown in Fig. 4B were done by subsampling the pool of generated samples within each subtype the same number of synthetic samples as unique original samples in the aggregated input across

the 25 experimental replicates. This unified set of balanced counts of original and synthetic samples within each subtype for each primary tumor type was input to the UMAP dimensionality reduction algorithm for subsequent scatter plotting. The clustering algorithm was the default "average" method implemented in the Scipy dependency of the Seaborn Clustermap function [33][127].

The feature importance algorithm utilized was Scikit-Learn Permutation Importance and was run on each of the 25 experimental replicates within the original gene expression data, the reconstructed expression data, and the synthetic sample expression data. Ten features were selected from each replicate within each data phase. The intersections of every combination of selected features were identified and binned for plotting in the UpSet plot.

Software tool requirements:

- TensorFlow 2.10
- Python 3.9
- Scikit-Learn 1.1.3

Data Availability

The software tool, SyntheVAEiser, is available at https://github.com/ohsu-comp-bio/syntheVAEiser [34][128] and https://doi.org/10.5281/zenodo.13948571 [35][129] under the Apache 2.0 license.

Chapter 5 — Preclinical cancer model systems: methods and evaluations for biological and technical data artifact correction

Brian Karlberg, Raphael Kirchgaessner, Jordan Led, Jeremy Jacobson, Sara J Gosline, Kyle Ellrott

Abstract Published: American Association for Cancer Research, 2024, conference abstract submission 7393 and poster presented

Abstract

Despite large-scale efforts to measure the effect of drug screens in cancer cell lines, mapping the effects of drugs to patient samples has been a challenge. Biological differences between cell lines and patients, such as lack of immune system or microbiome, in-vitro survival adaptations, and biases in measurement technologies create differences across sample modalities that can confound analysis including prediction with machine learning. In this work, we propose a multiway batch correction strategy to enable algorithmic prediction of tumor drug response across model systems and patient data. Recent advances in batch correction algorithms have been motivated by the need to correct for batch effects in single-cell omics and include diverse approaches such as variational autoencoders (VAEs) and generative adversarial networks (GANs). Given the successes of these generative deep learning methods in single cell sequencing analysis, we worked to employ similar approaches to correct large omics measurements across various cancer datasets. Here, we describe mapping of datasets from diverse data sources and model systems to the same space, so that a predictive model of drug response built in a system such as cell lines can be used in biologically relevant models such as organoids, patient-derived xenografts, and tumor data. Specifically, we introduce a modified loss function in a VAE using cosine similarity distance to minimize the effect of different cancer model systems in predicting cancer types. We evaluate the method on standard data types for drug response prediction - gene expression, copy number variation, and protein abundance. For this method, the cosine similarity is added as an additional term to the VAE reconstruction and Kullback-Leibler divergence loss terms. This injects a quantification of the dissimilarity between the tumor and tumor model distributions into the backpropagation and gradient descent for updating the model parameters resulting in an encoded representation of the data where the effect of data source has been attenuated while preserving the phenotypic signal. We evaluate our approach for biological signal preservation while reducing model system-specific noise with logistic regression and Euclidean distance. Our results show that the proposed VAE can effectively correct for platform effects and improve the accuracy of downstream integrative analyses. This study has the potential to improve the accuracy and translatability of proteogenomic drug response studies. The proposed modified VAE could be used to correct for platform effects in a variety of datasets, including those from different studies, different platforms, and different cancer types. This could

lead to new insights into cancer biology, calibration of cancer patient digital twins, and the development of new diagnostic and therapeutic strategies.

Introduction

The complexity of biological processes in cancer and the ethics of testing new therapeutics necessitate the use of models[130]. Cancer model systems include cell lines, organoids, and patient-derived xenografts[131]. These systems provide a means with which to screen putative therapeutic compounds against various types of cancers[132–134]. However, nested batch effects - unwanted noise - arise in data derived from these model systems due to biological artifacts such as cell lines having no immune system and technical artifacts such as differences in assays[135]. Additionally, non-linear relationships arise in data from biological model systems due to biochemically encoded information flow between system components such as gene regulation and metabolism[136]. The motivation for correction of batch effects in cancer model systems is to predict drug response in humans based on observations in model systems. A successful model system batch correction method could also inform future data collection in terms of what additional biological signals to target for collection or what noise signals to design against in the design of data collection instrumentation. Methods for batch correction in single-cell data have been developed and systematically benchmarked[137]. Variational autoencoders (VAEs) are versatile, neural network models that generate tunable latent representations of data thus are amenable to modification of data distributions[120].

VAEs are deployed in python whereas single-cell batch correction packages are written primarily in R presenting unique installation and intermediate data structuring challenges in evaluating method performance. Thus, package usability instead of performance emerges as a primary factor in the design and comparison of different correction methods. We hypothesize that neural network methods can both improve both the tractability of constructing system-specific correction pipelines and correct for batch effects as well as or better than existing methods while also improving the usability of batch correction tooling. Our observations are based on experiments, datasets, methods, and evaluations that were undergoing simultaneous development resulting in discrepancies over the lifetime of the project. However, the two cancer model system

batch correction themes of aligning the biological signal component of omics data distributions while retaining separation of the phenotypic signals remain consistent across all combinations of data modalities, correction methods, and evaluation methods.

Results

CoderData platform and batch correction strategy

Cancer model systems data were obtained using Pacific Northwest National Laboratory's CoderData (Cancer Omics Drug Experiment Response Data) integration platform[138]. This package integrates five data sources: Broad Sanger, CPTAC (Clinical Proteomic Tumor Analysis Consortium[19]), HCMI (Human Cancer Models Initiative[20]), BeatAML[139], and MPNST (malignant peripheral nerve sheath tumor[140]). Each of these datasets contain samples with molecular profiles of various combinations of data types including transcriptomics, proteomics, mutations, and copy number variation. For example, CPTAC contained samples with molecular profiles covering all four data types with 10 cancer types whereas the HCMI organoids contained no proteomics profiles and a different combination of 10 cancer types. Understanding the intersections of these data attributes is necessary for controlling the design of machine-learnable data structures. This includes controlling for statistical power and class imbalance. The strategy to evaluate the model system correction methods was to predict cancer type and model system source before and after the correction and to cluster by cancer type and model system labels before and after the correction. The model system batch correction workflow of data acquisition, structuring, and comparing DL with linear correction methods is shown in Fig. 5-1.

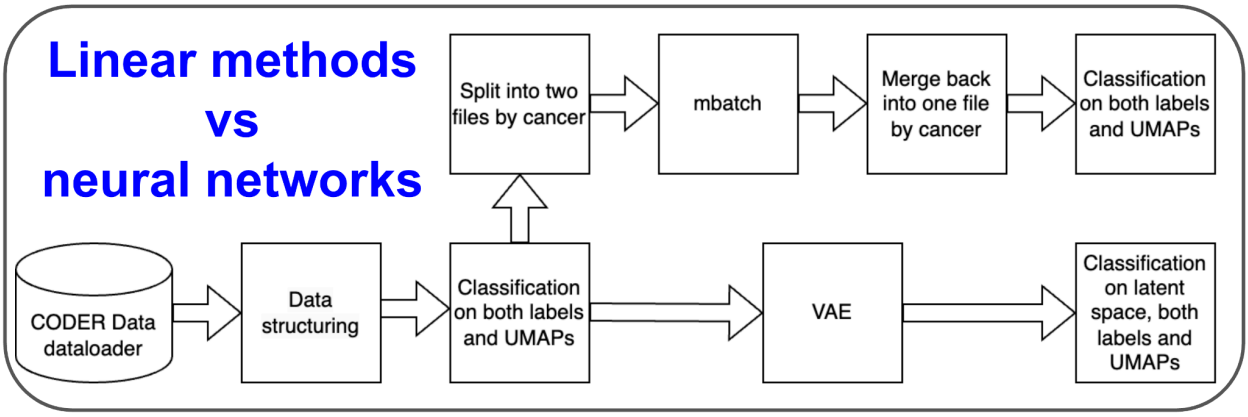


Fig. 5-1 Model systems batch correction process flow diagram. Corresponding data across model systems, data types, and cancer types were sourced from the CoderData platform at Pacific Northwest National Laboratory. Classification on both the cancer type labels and the data source labels in addition to UMAP clustering on the independent variables by both label types established baseline results on the pre-corrected data. The pipeline then bifurcated for the linear method vs neural network comparison with additional unique data structuring on each track to correct for model system batch effects and then reformatted the data for direct post-correction comparisons

Machine learnable data preparation

The batch-correction evaluation was implemented with a dual-label tabular data structure. Label set one was the model system or data source and label set two was the cancer type; these two label sets were the y targets for the scikit-learn classifiers and the UMAP dimensionality reduction. One or more data types of biomolecular measurements i.e. transcriptomics, proteomics, etc. comprised the independent variable portion of the data. The result was dual-labeled, mono- or multi-modal machine learnable and dimensionally reducible data. A generic tabular representation of this structure for one biological modality is shown in Fig. 5-2. This representation shows two model systems and two cancer types however data structures with more than two model systems per cancer and/or more than two cancers are possible and were tested in conjunction with these binary dual-label structures.

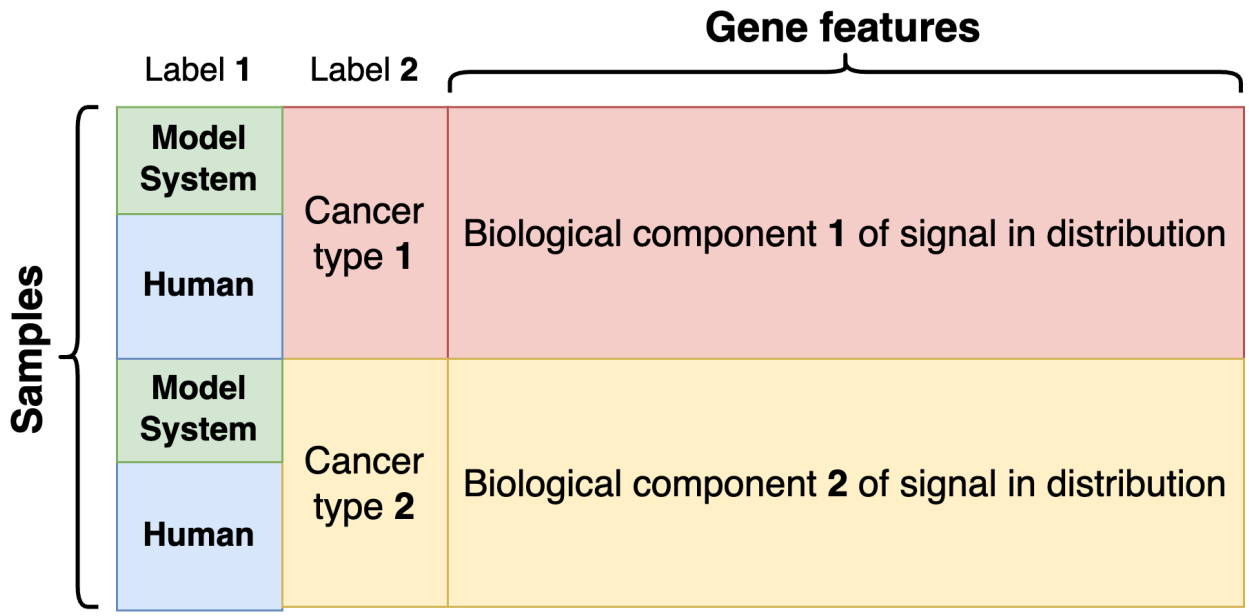


Fig. 5-2 Dual-label data structure prepared for machine learning. Actual cancer model system data structures may contain more than two classes within each label. Limited sample sizes and class imbalances, denoted by the difference in height of the green and the blue boxes, may be severe thus presenting challenges for model learning

Modeling challenges

The three primary challenges in modeling with these data, as summarized in Fig. 5-3, were limited sample sizes, class imbalances, and incomplete coverage of data types and cancer types. Limited statistical power due to small sample sizes presented training challenges for both the traditional ML models used in the evaluation and the neural network models used in the correction. Additionally, algorithms are often designed to optimize overall accuracy, which can be dominated by the majority class[141]. In extreme cases, the minority class may be ignored altogether. This leads to poor predictive performance on the minority class such as a rare cancer type or model system with relatively fewer samples. Further, across model systems, the intersection of data types and cancer types is incomplete. These data characteristics limited the available set of testable experimental configurations within the constraints of presently available ML tools.

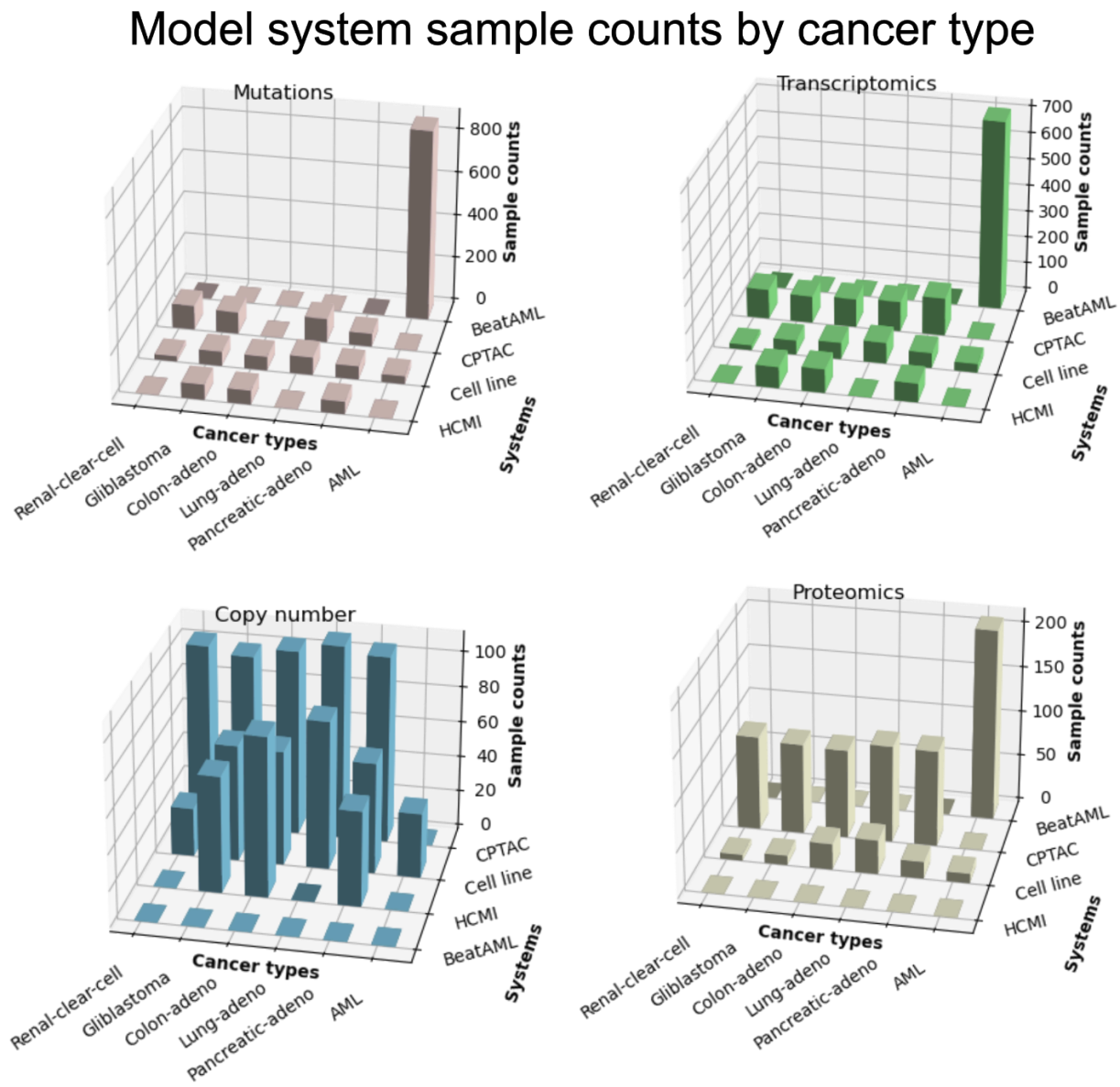


Fig. 5-3 Model system sample counts by cancer type faceted by data type. Mutation, transcriptomic, copy number, and proteomic data types by 5 solid and 1 blood cancer types. Model system combination varied by several orders of magnitude resulting in class imbalances and limited power. Data shown for the five most-common cancer types intersecting the four model system projects. HCMI data were first pre-filtered to include only 3D organoid samples

Model approaches and evaluations

Within these constraints of small sample-sizes, class imbalances, and limitations in data type-coverage, modeling configurations were explored to develop both the correction and

evaluation components of the pipeline. A survey of single-cell batch correction algorithms, such as the popular ComBat-seq[142], was conducted to identify a method amenable to our data and establish a baseline of comparison for the GDAN-TMP quantile rescaling algorithm and the VAE. MBatch[143] was identified as a tractable solution and deployed in a comparison pipeline with four custom VAEs developed and deployed in collaboration with Raphael Kirchgaessner, a PhD candidate also with the Ellrott Lab. The VAE development drew from a variety of VAE-tuning methods with potential to achieve our goal of preserving biological signal while removing both the biological artifacts arising from different model system sources. This process is termed disentanglement of latent representations and a multitude of approaches have been reported; one example implementation is adding a penalty term to the composite loss function of a VAE[144–146]. Representative results of batch correction method and evaluation method comparisons for a multicategorical cancer type comparison are shown in Fig. 5-4. The five cancer types comprising this primary tissue prediction are the solid tumors shown in Fig. 5-3 i.e. all but AML.

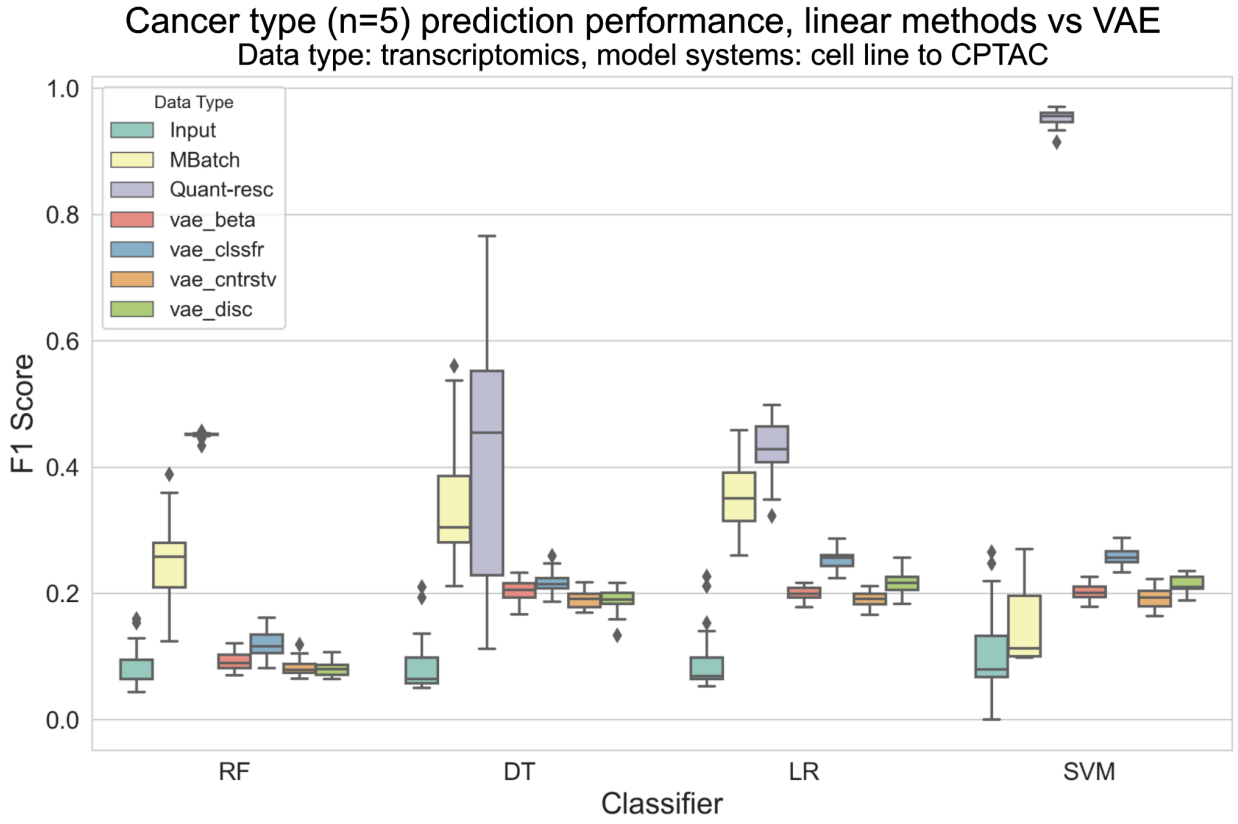


Fig. 5-4 Comparison of batch correction methods with multicategorical cancer-type classification. Performance improvement from uncorrected data was measured with four Scikit-learn evaluation models:

random forest, decision tree, logistic regression, and support vector machine. The uncorrected data is shown in green with the linear correction method MBatch in yellow, the GDAN-TMP quantile rescaling in purple, and our four VAE correction methods developed at OHSU: VAE beta classifier, VAE classifier, VAE contrastive loss, and VAE discriminative

In comparing the performance of our GDAN-TMP quantile rescaling, MBatch, and the four VAEs, the rescaling method consistently returns the largest increase in cancer type prediction performance. For three of the four evaluation methods, the MBatch correction method raised the mean prediction above the upper quartile of the uncorrected data. Our four VAE methods showed similar performance results for each of the four evaluation classifiers with the lower quartile of all four VAE-improved predictions above the upper quartile of the uncorrected data for the decision tree and logistic regression evaluation classifiers. The classifier VAE (vae_clssfr) was the most performant VAE architecture although the overall difference among the VAEs was minimal.The supervised categorical prediction evaluation was also applied using the model system sources as the y-labels (results not shown). Here, a reduction in prediction performance on model system labels indicates successful batch correction.

Next, a UMAP dimensionality reduction was performed on both the cancer type and model system source as an additional evaluation of correction performance. The expected behavior was analogous to the classification scheme where post-correction clustering effect would diminish for model system source and the clustering effect would improve for biological signal as indicated by cancer type. A UMAP cluster comparison on system source labels, shown in Fig. 5-5, shows a similar reduction in clustering performance for both MBatch and the VAE, over two cancer types, which indicates in both cases the intended removal of batch effect.

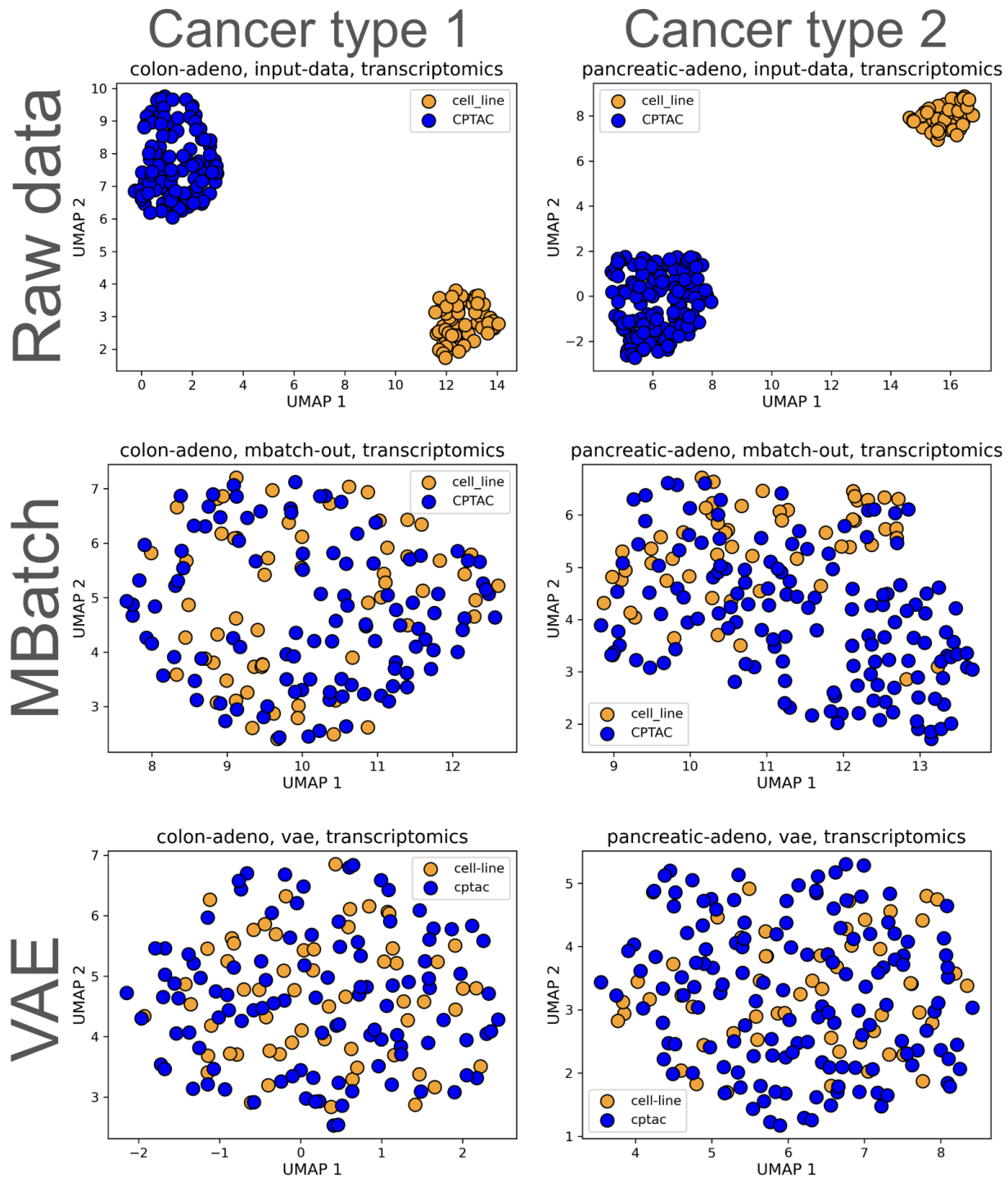


Fig. 5-5 Dimensionality reduction pre- and post-artifact correction. MBatch vs VAE — both the linear MBatch and non-linear VAE correction methods transformed the data distributions so that samples that were originally clustered by the model system data source did not cluster after the correction. Variation in this effect was observed over combinations of methods and cancer types

Discussion

This experiment addressed the challenge that observations in cancer model systems, such as drug screening in cell-lines, have poor transferability to humans[147]. In mouse models, for example, this is the result of significant differences in gene expression that relate to differences in metabolism, life span, and tumor-type susceptibility[148]. Cell-lines lack an immune system and microbiome altogether which also degrades the capability of recapitulating the effect of drug perturbations in humans. Similar challenges exist for organoids. These sources of biological noise are distinct from the phenotype of interest and are what batch correction, widely developed for single-cell data integration, attempts to overcome[137]. In addition to these biological factors, the aggregation of assays with differences in time-of-capture, operations personnel, reagents, and instruments results in technical artifacts between groups of samples in the data. Together, these biological and non-biological factors combine to result in what are termed nested, confounded, or hierarchical data effects.

Our approach sought to overcome these data platform incongruences by leveraging the non-linear modeling capabilities of generative neural networks i.e. a VAE to transform the distributions of cancer model system omics to be of utility in prediction problems such as drug response with human data. As an initial step toward this larger goal, we structured our modeling around cancer tissue-of-origin prediction. Our pipelines consisted of an evaluation framework and the correction methods under comparison. The evaluation framework utilized non-neural network ML classifiers to compare the effect of cancer type predictability on both the cancer labels and the model system source labels. F1 scoring was used to quantify the prediction accuracies to account for class imbalances; ideally, a model system dataset for developing correction methods would have balanced classes to not confound the source effects. A UMAP clustering provided a corresponding qualitative assessment of the correction effectiveness. For both the label prediction and clustering, diminished performance on the model system source labels served as an indicator of the intended effect of aligning omic data distributions.

Over the various permutations of data type, prediction goal, correction methods, and evaluation methods we observed variation in the correction results. However, the general trend was that the GDAN-TMP quantile rescaling algorithm showed the greatest improvement in prediction of phenotypic class followed by the MBatch linear method and our four VAEs. We worked toward comparisons of multiple correction methods using multiple evaluations. Future work should continue in this direction by integrating explicit comparisons across methods such as ComBat, MNN, and contrastive learning and incorporating direct measures of evaluation such as kBET and silhouette.

Conclusions

The utility of cancer model systems is constrained by nested batch effects that include both biological and technical artifacts. Clustering and supervised categorical prediction can be implemented with multiple methods to build robust evaluation schemes for correcting data source effects in cancer model systems. The interactive effects between model system data attributes, different evaluation frameworks, and different correction methods can lead to significantly different conclusions in terms of both the degree to which the biological signal of interest has been preserved and the degree to which the batch effects have been removed. This highlights that an arbitrary choice of evaluation classifier i.e. SVM can lead to different conclusions than if using multiple evaluation methods. An ideal model for transforming cancer model system data distributions would diminish data source effects across data types and prediction tasks while improving detectability of the phenotypic signal of interest.

Methods

Data provenance and structuring

Data for copy number variation, mutations, gene expression, and proteomics were obtained for four model system platforms: CPTAC, HCMI, cell-line, and BeatAML using the CoderData data portal at https://pnnl-compbio.github.io/coderdata/. Age ranges, male-to-female ratios, and the disease stages are shown for the five cancer types reported herein in Table 5-1.

Table 5-1 CPTAC demographics

cancer_type	N	Age Range (years)	Sex (Female / Male)	Disease Stage (Early / Late)
Colon-adeno	106	35-88	61 / 45	52 / 53
Glioblastoma	100	24-88	44 / 56	NaN
Lung-adeno	111	35-81	38 / 73	89 / 22
Pancreatic-adeno	154	31-85	68 / 86	87 / 54
Renal-clear-cell	109	30-89	29 / 80	64 / 45

The model system labels used in this experiment do not match the current data offerings due to ongoing updates to CoderData. The HCMI data included model systems other than organoid such as cell-line. These non-organoid samples were removed rather than added to the other cell-line data as to avoid introducing additional source noise; this came at the expense of increased statistical power for the cell-line data type.

An interactive data extraction and re-structuring notebook was built with python in Jupyter Lab using the cd.load() function in the coderdata package:

```
>> import coderdata as cd
```

The evaluation framework, which included the ML classifiers and plotting functions, described in detail below, was built into the same notebook to enable rapid exploratory comparisons across stages of data correction. The R-based MBatch read separate input files from disk in a transposed format compared with the typical pandas dataframe input format for the Python-based VAEs.

Modeling

During development, multiple cancer label combinations were tested within binary and multi-class formulations of the problem. The prediction results shown in Fig. 5-4 used transcriptomics data for predicting five cancer types: renal-clear-cell, glioblastoma, colon-adeno, lung-aeno, and pancreatic-adeno. The evaluation framework consisted of four scikit-learn models: random forest, decision tree, logistic regression, and support vector machine. Hyperparameter settings were set to default.

A development VAE was built by modifying the Tybalt[120] model by adding a penalty and reward term to the composite loss function. The two distribution distance metrics tested in this phase were cosine similarity and euclidean distance. The four models built by project collaborator Raphael Kirchgaessner were deployed on OHSU's Exacloud using a push-only to production strategy via a shared GitHub Repo. The four variations of distentanglement VAEs tested were: beta, contrastive, classifier, discriminative.

Matplotlib and seaborn were used to generate the figures.

Hardware and software

Exacloud: https://www.ohsu.edu/advanced-computing-center/acc-cluster-computing

* Note: Exacloud has since rebranded to Advanced Computing Center (ACC) Research Cluster (ARC)

UMAP: https://umap-learn.readthedocs.io/en/latest/

Scikit-learn: https://scikit-learn.org/stable/

MBatch: https://bioinformatics.mdanderson.org/public-software/mbatch/

GDAN-TMP quantile rescaling algorithm: https://github.com/NCICCGPO/gdan-tmp-models/tree/main/tools/quantile_rescale.py

Chapter 6 — Conclusions, overall scientific contributions, and individual contributions

Brian Karlberg

Publishing/permissions: NA

The complexity of cancer - a chemical system of interwoven networks for communication, control, biosynthesis, and energy production[149–151] - requires modeling approaches that transcend established disciplinary boundaries[152,153]. Machine learning (ML) is isomorphic to such a system of complexity in providing a means for integrative modeling; capabilities beyond traditional analytical methods. This dissertation presented ML-based analyses of cancer as molecularly defined by sequencing-based omics. These molecular profiles of genomic and transcriptomic measurements were studied over the majority of solid tumors in both humans and cancer model systems. Our findings show the mRNA data is of particular utility in delineating cancer subtypes and developing ML pipelines with potential for cost-effective clinical implementation. Another overarching theme is the identification of feature sets, classifiers, or generative models that are specific to data types and can be tailored to goals ranging from cancer detection, defining subtypes, synthetic sample generation, and batch correction.

Prediction of TCGA molecular subtypes

The GDAN-TMP project utilized an interdisciplinary approach[154] — clinicians, biologists, computer scientists, and engineers building ML models for translating high-dimensional TCGA molecular data into clinically applicable knowledge and software tools. Our work in applying five distinct ML approaches with five molecular data types across the majority of solid tumors revealed genomic and transcriptomic gene signatures specific to cancer types. From this development work, several key findings emerged: cross validation and performance scoring to account for imbalanced classes can facilitate benchmarking of orthogonal ML methods, compact feature sets of potential clinical utility, that gene-centric molecular profiles can enhance interpretability, gene expression signatures are comparatively performant in capturing complex underlying biology, external validation frameworks can be built with rigor, the statistical effects of limited sample sizes can be characterized, and that pre-trained ML pipelines for molecular subtype prediction can be shipped to facilitate adoption by the broader community.

Mutational landscape vs transcriptional state

Memoization-based sorting can be applied to gene-level binary mutation values for cancer cohorts at both the primary tumor and subtype levels. This can be done either with or without a priori feature selection and reveals patterns of mutually exclusivity in mutation data as a basis for analysis of alterations to DNA coding regions. Feature selection was conducted within a repeated sub-sampling framework to produce selection frequency-based ranked sets of expression features. Subtype-specific interaction networks can be built to introspect cancer molecular subtype biology. Class-specific feature importance quantification was developed to analyze relative feature importance with subtype resolution. Combining memo-sort-identified mutation gene sets with expression gene sets via Pathway Commons interaction mapping yields a route constructing gene interaction networks with either primary tumor-type or subtype specificity. Gene expression features predict molecular subtype with equivalent F1 score and prediction confidence across the TCGA.

Synthetic sample generation

Here, we demonstrated how generative deep learning can be leveraged to synthesize transcriptomic data. We utilized transfer learning, fine tuning, and then sub-sampling with a variational autoencoder to generate latent-space representations of mRNA profiles specific to 106 TCGA cancer subtypes. This approach used the prior information of thousands of samples for learning distributions of rare subtypes with limited statistical power. Our novel data augmentation technique was demonstrated to be robust under both quantitative and qualitative evaluations that included supervised categorical prediction, an RMSE analysis, UMAP and Scipy clustering, maximum mean discrepancy distributional analysis, feature importance calculations, and application to single-cell data. In sum, this experiment showed how deep learning can integrate with and boost the performance of traditional, interpretable ML methods.

Cancer model systems batch correction

Cancer model systems such as cell-lines and organoids provide a means for testing the efficacy of putative chemotherapy drugs. However, molecular profiling data derived from such model systems suffer from inherent limitations relating to both technical and biological artifacts. These artifacts result in what is termed “nested batch effects” and inhibit the transferability of observations to confident conclusions of drug-response in humans. A preprocessing pipeline was developed to extract model system data from various sources and prepare dual labeled — system and cancer type — machine learning ready. We explored the utility of adapting batch-correction methods from single cell methods and worked to build batch-correcting terms into variational autoencoders. Our evaluation framework included unsupervised clustering and supervised categorical prediction on both label sets to compare the performance of various correction methods. The quantile rescaling algorithm from the GDAN-TMP project was observed to be particularly performant under one set of evaluations, however this result underscores the theme of importance in comparing multiple evaluation methods in drawing conclusions from the application of ML to molecular profile data of cancer.

Overall scientific contributions

This dissertation presents a body of work demonstrating the application of machine learning to cancer molecular profiles to advance precision oncology. The primary contributions are organized across four key areas: the prediction of molecular subtypes, the interpretation of the underlying biology, the improvement of modeling performance through data synthesis, and the extension of these methods to bridge cancer model systems with human data.

First, this work established a comprehensive and robust framework for the clinical translation of cancer subtyping. We developed a public library of machine learning classifiers capable of predicting molecular subtypes for new tumor samples across 26 different cancer types from The Cancer Genome Atlas (TCGA). A key finding was that gene expression (mRNA) signatures were frequently the most informative data type for classification, often outperforming the genomic and epigenomic features originally used to define the subtypes. The generalizability of

these models was confirmed through rigorous external validation, and we characterized the relationship between sample size and predictive accuracy, providing guidance for future study design.

Second, we investigated the complex relationship between the mutational landscape and the resulting transcriptomic state in cancer. A novel feature selection framework was developed, combining a memo-sort algorithm to identify patterns of mutual exclusivity among mutated oncogenes with a sub-sampling approach to pinpoint corresponding mRNA expression signatures. This analysis revealed that while specific mutations are critical, mRNA signatures effectively integrate the complex downstream effects of numerous genomic and proteomic alterations. This provides a strong rationale for their utility in molecular classification, as they consistently demonstrate high predictive power even in cases where common oncogene panels fail to detect a mutation.

Third, to address the common challenge of limited data for rare cancer subtypes, we developed a novel deep learning method for synthetic data generation. The tool, SyntheVAEiser, utilizes a variational autoencoder (VAE) trained via a transfer learning strategy to generate high-fidelity, synthetic gene expression samples. We demonstrated that augmenting training datasets with these synthetic samples significantly improves the accuracy of machine learning classifiers, thereby overcoming limitations posed by small sample sizes and imbalanced classes. This contribution provides a powerful method for improving statistical power in genomic studies.

Finally, this research tackled the critical challenge of translating findings from cancer model systems (e.g., cell lines, organoids) to human patients. We developed and evaluated a multi-way batch correction strategy to remove the technical and biological artifacts that confound the integration of these disparate datasets. By employing and comparing various methods, including novel VAE-based approaches, we demonstrated a pipeline to align data distributions while preserving the essential biological signals. This work represents an important step toward enabling the prediction of patient drug response based on data from preclinical model systems.

Collectively, these contributions demonstrate a multi-faceted machine learning approach to address fundamental challenges in cancer genomics. This dissertation delivers not only predictive models and software tools for the research community but also provides deeper

insights into cancer biology and a clear path toward more precise, molecularly-guided clinical decision-making.

Contributions, individual

Chapter 1 was written by B.K. with general guidance for content and formatting from the DAC committee in ideation and revisions. The three topics comprising Chapter 2 were extracted from the GDAN-TMP publication per B.K. primary contributions. K.E. and J.L were involved in the grid search and feature selection results generation with the TMP consortia providing experimental and figure ideation and revision guidance for the validation and sample collection components. Unless otherwise noted, figures were generated by B.K. and reproduced without modification from the Cancer Cell publication. Chapter 3 was written by B.K upon results generated by B.K. in iterative ideation and revision with K.E. All figures were created by B.K. Chapter 4 was written by B.K. via direct reproduction, without modification, of the Genome Biology publication.All figures were created by B.K. Chapter 5 was written by B.K. based on the AACR poster. All figures were created by B.K. Chapter 5 was written by B.K with revised content based on committee comments.

Bibliography

1. Mattiuzzi, C. & Lippi, G. Current Cancer Epidemiology. *J. Epidemiol. Glob. Health* **9**, 217–222 (2019).
2. Ferlay, J. *et al.* Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int. J. Cancer* **144**, 1941–1953 (2019).
3. Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **74**, 229–263 (2024).
4. Fidler, M. M., Soerjomataram, I. & Bray, F. A global view on cancer incidence and national levels of the human development index. *Int. J. Cancer* **139**, 2436–2446 (2016).
5. Cronin, K. A. *et al.* Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. *Cancer* **124**, 2785–2800 (2018).
6. Harris, T. J. R. & McCormick, F. The molecular pathology of cancer. *Nat. Rev. Clin. Oncol.* **7**, 251–265 (2010).
7. Golub, T. R. *et al.* Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286**, 531–537 (1999).
8. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.* **99**, 6567–6572 (2002).
9. Bhinder, B., Gilvary, C., Madhukar, N. S. & Elemento, O. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov.* **11**, 900–915 (2021).
10. Ellrott, K. *et al.* Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets. *Cancer Cell* **43**, 195-212.e11 (2025).
11. Jiao, W. *et al.* A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728 (2020).
12. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine

learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015).

13. Karlberg, B. *et al.* SyntheVAEiser: augmenting traditional machine learning methods with VAE-based gene expression sample generation for improved cancer subtype predictions. *Genome Biol.* **25**, 309 (2024).
14. Amin, M. B. *et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more ‘personalized’ approach to cancer staging. *CA. Cancer J. Clin.* **67**, 93–99 (2017).
15. Collins, F. Cancer: A Disease of the Genome. *Cancer Res.* **67**, PL01-01 (2007).
16. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* **173**, 283–285 (2018).
17. Blum, A., Wang, P. & Zenklusen, J. C. SnapShot: TCGA-Analyzed Tumors. *Cell* **173**, 530 (2018).
18. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
19. Edwards, N. J. *et al.* The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **14**, 2707–2713 (2015).
20. Tonsing-Carter, E. *et al.* Abstract 4681: Human Cancer Models Initiative (HCMI): A community resource of next-generation cancer models and associated data. *Cancer Res.* **83**, 4681 (2023).
21. Jiang, Y., García-Durán, A., Losada, I. B., Girard, P. & Terranova, N. Generative models for synthetic data generation: application to pharmacokinetic/pharmacodynamic data. *J. Pharmacokinet. Pharmacodyn.* **51**, 877–885 (2024).
22. Penson, A. *et al.* Development of Genome-Derived Tumor Type Prediction to Inform Clinical Cancer Care. *JAMA Oncol.* **6**, 84–91 (2020).
23. Heller, M. J. DNA Microarray Technology: Devices, Systems, and Applications. *Annu. Rev.*

Biomed. Eng. **4**, 129–153 (2002).

24. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
25. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).
26. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **95**, 14–18 (2003).
27. Liu, H. & Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*. (Springer Science & Business Media, 2012).
28. Remeseiro, B. & Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **112**, 103375 (2019).
29. Zhang, E. *et al.* LLM-Lasso: A Robust Framework for Domain-Informed Feature Selection and Regularization. Preprint at https://doi.org/10.48550/arXiv.2502.10648 (2025).
30. Shuaibi, A., Chitra, U. & Raphael, B. J. A latent variable model for evaluating mutual exclusivity and co-occurrence between driver mutations in cancer. *bioRxiv* 2024.04.24.590995 (2024) doi:10.1101/2024.04.24.590995.
31. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. (Wiley-Interscience, Hoboken, NJ, 2001).
32. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*. (Springer, New York, NY, 2009). doi:10.1007/978-0-387-84858-7.
33. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
34. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
35. Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2**, 160 (2021).

36. Shrestha, A. & Mahmood, A. Review of Deep Learning Algorithms and Architectures. *IEEE Access* **7**, 53040–53065 (2019).

37. Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. (MIT Press, 2012).

38. Japkowicz, N. Assessment Metrics for Imbalanced Learning. in *Imbalanced Learning* 187–206 (John Wiley & Sons, Ltd, 2013). doi:10.1002/9781118646106.ch8.

39. Diallo, R., Edalo, C. & Awe, O. O. Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. in *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA* (eds. Awe, O. O. & A. Vance, E.) 283–312 (Springer Nature Switzerland, Cham, 2025). doi:10.1007/978-3-031-72215-8_12.

40. Mortara, I. & Executive Director, International Union Against Cancer (UICC), Geneva. The International Union Against Cancer. *Oncol. Hematol. Rev. US* 22 (2007) doi:10.17925/OHR.2007.00.2.22.

41. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).

42. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLOS Comput. Biol.* **15**, e1006953 (2019).

43. Lu, M. Y. *et al.* AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).

44. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).

45. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).

46. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).

47. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*

531, 47–52 (2016).

48. Rouzier, R. *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **11**, 5678–5685 (2005).
49. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563 (2016).
50. Bagaev, A. *et al.* Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell* **39**, 845-865.e7 (2021).
51. Grewal, J. K. *et al.* Application of a Neural Network Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. *JAMA Netw. Open* **2**, e192597 (2019).
52. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. Preprint at https://doi.org/10.48550/arXiv.1201.0490 (2018).
53. Uzunangelov, V., Wong, C. K. & Stuart, J. M. Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. *PLOS Comput. Biol.* **17**, e1008878 (2021).
54. Bressler, R. *et al.* CloudForest: A Scalable and Efficient Random Forest Implementation for Biological Data. *PLOS ONE* **10**, e0144820 (2015).
55. Tsamardinos, I. *et al.* Just Add Data: automated predictive modeling for knowledge discovery and feature selection. *Npj Precis. Oncol.* **6**, 1–17 (2022).
56. Grandini, M., Bagli, E. & Visani, G. Metrics for Multi-Class Classification: an Overview. Preprint at https://doi.org/10.48550/arXiv.2008.05756 (2020).
57. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
58. Garcia-Recio, S. *et al.* Multiomics in primary and metastatic breast tumors from the AURORA US network finds microenvironment and epigenetic drivers of metastasis. *Nat. Cancer* **4**, 128–147 (2023).

59. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

60. Esteve-Codina, A. *et al.* A Comparison of RNA-Seq Results from Paired Formalin-Fixed Paraffin-Embedded and Fresh-Frozen Glioblastoma Tissue Samples. *PLOS ONE* **12**, e0170632 (2017).

61. Anderson, J. R. The Architecture of Cognition. in (Psychology Press, 2013). doi:10.4324/9781315799438.

62. Figueroa, R. L., Zeng-Treitler, Q., Kandula, S. & Ngo, L. H. Predicting sample size required for classification performance. *BMC Med. Inform. Decis. Mak.* **12**, 8 (2012).

63. Mukherjee, S. *et al.* Estimating Dataset Size Requirements for Classifying DNA Microarray Data. *J. Comput. Biol.* **10**, 119–142 (2003).

64. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: a primer. *Nat. Methods* **14**, 1119–1120 (2017).

65. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.* **48**, D489–D497 (2020).

66. Grayson, S., Suver, C., Wilbanks, J. & Doerr, M. Open Data Sharing in the 21st Century: Sage Bionetworks' Qualified Research Program and Its Application in mHealth Data Release. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.3502410 (2019).

67. Tan, H. Somatic mutation in noncoding regions: The sound of silence. *EBioMedicine* **61**, 103084 (2020).

68. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).

69. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

70. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci.* **112**, 5473–5478 (2015).

71. Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L. & Siu, L. L. Molecular profiling for

precision cancer therapies. *Genome Med.* **12**, 8 (2020).

72. Hou, Y.-C. C., Neidich, J. A., Duncavage, E. J., Spencer, D. H. & Schroeder, M. C. Clinical whole-genome sequencing in cancer diagnosis. *Hum. Mutat.* **43**, 1519–1530 (2022).
73. Huang, S. C. *et al.* Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling. *PLOS Comput. Biol.* **9**, e1002887 (2013).
74. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front. Bioinforma.* **2**, 927312 (2022).
75. COSMIC. https://www.sanger.ac.uk/tool/cosmic/.
76. Alberts, B. *et al.* Finding the Cancer-Critical Genes. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).
77. Housden, B. E. *et al.* Loss-of-function genetic tools for animal models: cross-species and cross-platform differences. *Nat. Rev. Genet.* **18**, 24–40 (2017).
78. Zeng, Z. & Bromberg, Y. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Front. Genet.* **10**, (2019).
79. Hess, J. M. *et al.* Passenger Hotspot Mutations in Cancer. *Cancer Cell* **36**, 288-301.e14 (2019).
80. Chang, M. T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
81. Küçükosmanoglu, A. *et al.* Oncogenic composite mutations can be predicted by co‐mutations and their chromosomal location. *Mol. Oncol.* **18**, 2407–2422 (2024).
82. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
83. https://manual.cytoscape.org/en/stable/Supported_Network_File_Formats.html#sif-format - Google Search.

https://www.google.com/search?q=https://manual.cytoscape.org/en/stable/Supported_Network_File_Formats.html%23sif-format.

84. Robson, M. Multigene Panel Testing: Planning the Next Generation of Research Studies in Clinical Cancer Genetics. *J. Clin. Oncol.* **32**, 1987–1989 (2014).

85. Kurian, A. W. *et al.* Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **32**, 2001–2009 (2014).

86. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

87. Michie, D. “Memo” Functions and Machine Learning. *Nature* **218**, 19–22 (1968).

88. Park, J. & Paramasivam, N. PyOncoPrint: a python package for plotting OncoPrints. *Genomics Inform.* **21**, e14 (2023).

89. Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M. & Tsamardinos, I. Feature selection with the r package mxm: Discovering statistically equivalent feature subsets. *J. Stat. Softw.* **80**, 1–25 (2017).

90. Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

91. Dealmakers, B. Generative AI platforms drive drug discovery dealmaking. *Biopharma Deal.* (2024) doi:10.1038/d43747-024-00084-w.

92. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

93. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **46**, 389–422 (2002).

94. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at http://arxiv.org/abs/1705.07874 (2017).

95. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).

96. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23**, 40–55 (2022).

97. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. Poznan Pol.* **19**, A68-77 (2015).

98. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506–519 (2015).

99. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).

100. Roh, W. *et al.* High-Resolution Profiling of Lung Adenocarcinoma Identifies Expression Subtypes with Specific Biomarkers and Clinically Relevant Vulnerabilities. *Cancer Res.* **82**, 3917–3931 (2022).

101. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

102. Fishbein, L. *et al.* Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. *Cancer Cell* **31**, 181–193 (2017).

103. Picornell, A. C. *et al.* Breast cancer PAM50 signature: correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. *BMC Genomics* **20**, 452 (2019).

104. Jensen, M.-B. *et al.* The Prosigna 50-gene profile and responsiveness to adjuvant anthracycline-based chemotherapy in high-risk breast cancer patients. *NPJ Breast Cancer* **6**, 7 (2020).

105. de Melo, C. M. *et al.* Next-generation deep learning based on simulators and synthetic data. *Trends Cogn. Sci.* **26**, 174–187 (2022).

106. Hosna, A. *et al.* Transfer learning: a friendly introduction. *J. Big Data* **9**, 102 (2022).

107. Ruthotto, L. & Haber, E. An introduction to deep generative modeling. *GAMM-Mitteilungen* **44**, e202100008 (2021).

108. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. Preprint at https://doi.org/10.48550/arXiv.1401.4082 (2014).

109. Kingma, D. P., Salimans, T. & Welling, M. Variational Dropout and the Local Reparameterization Trick. Preprint at https://doi.org/10.48550/arXiv.1506.02557 (2015).

110. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2022).

111. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. & Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **12**, e1608 (2022).

112. Kokol, P., Kokol, M. & Zagoranski, S. Machine learning on small size samples: A synthetic knowledge synthesis. *Sci. Prog.* **105**, 368504211029777 (2022).

113. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

114. Blagus, R. & Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* **14**, 106 (2013).

115. Osuala, R. *et al.* Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging. *Med. Image Anal.* **84**, 102704 (2023).

116. Shorten, C. & Khoshgoftaar, T. M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **6**, 60 (2019).

117. Baur, C., Albarqouni, S. & Navab, N. MelanoGANs: High Resolution Skin Lesion Synthesis with GANs. Preprint at https://doi.org/10.48550/arXiv.1804.04338 (2018).

118. Ahmed, K. T., Sun, J., Cheng, S., Yong, J. & Zhang, W. Multi-omics data integration by generative adversarial network. *Bioinforma. Oxf. Engl.* **38**, 179–186 (2021).

119. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).

120. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **23**, 80–91 (2018).

121. Caruana, R. & Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. in *Proceedings of the 23rd international conference on Machine learning - ICML '06* 161–168 (ACM Press, Pittsburgh, Pennsylvania, 2006). doi:10.1145/1143844.1143865.

122. Kim, A. A., Rachid Zaim, S. & Subbian, V. Assessing reproducibility and veracity across machine learning techniques in biomedicine: A case study using TCGA data. *Int. J. Med. Inf.* **141**, 104148 (2020).

123. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).

124. Shen, H. *et al.* Integrated Molecular Characterization of Testicular Germ Cell Tumors. *Cell Rep.* **23**, 3392–3406 (2018).

125. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A Kernel Two-Sample Test. *J. Mach. Learn. Res.* **13**, 723–773 (2012).

126. Grandvalet, Y., Canu, S. & Boucheron, S. Noise injection: theoretical prospects. *Neural Comput* **9**, 1093–1108 (1997).

127. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. Preprint at https://doi.org/10.48550/arXiv.1109.2378 (2011).

128. github.com/ohsu-comp-bio/syntheVAEiser. Oregon Health and Science University Computational Biology (2024).

129. Medicine, O. H. & S. U. S. of. SyntheVAEiser codebase on Zenodo: zenodo.org/records/13948571. (2024).

130. Welm, B. E., Vaklavas, C. & Welm, A. L. Toward improved models of human cancer. *APL Bioeng.* **5**, 010901 (2021).

131. Sajjad, H. *et al.* Cancer models in preclinical research: A chronicle review of advancement in effective cancer research. *Anim. Models Exp. Med.* **4**, 87–103 (2021).
132. Martinez-Pacheco, S. & O'Driscoll, L. Pre-Clinical In Vitro Models Used in Cancer Research: Results of a Worldwide Survey. *Cancers* **13**, 6033 (2021).
133. Tosca, E. M., Ronchi, D., Facciolo, D. & Magni, P. Replacement, Reduction, and Refinement of Animal Experiments in Anticancer Drug Development: The Contribution of 3D In Vitro Cancer Models in the Drug Efficacy Assessment. *Biomedicines* **11**, 1058 (2023).
134. Jung, J. Human Tumor Xenograft Models for Preclinical Assessment of Anticancer Drug Development. *Toxicol. Res.* **30**, 1–5 (2014).
135. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
136. Thomas, R. M., Van Dyke, T., Merlino, G. & Day, C.-P. Concepts in Cancer Modeling: A Brief History. *Cancer Res.* **76**, 5921–5925 (2016).
137. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
138. github.com/PNNL-CompBio/coderdata. Computational Biology @Pacific Northwest National Laboratory (2025).
139. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).
140. Hirbe, A. C. *et al.* Contemporary Approach to Neurofibromatosis Type 1–Associated Malignant Peripheral Nerve Sheath Tumors. *Am. Soc. Clin. Oncol. Educ. Book* **44**, e432242 (2024).
141. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).
142. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for

RNA-seq count data. *NAR Genomics Bioinforma.* **2**, lqaa078 (2020).

143. Zhang, N. *et al.* PCA-Plus: Enhanced principal component analysis with illustrative applications to batch effects and their quantitation. *bioRxiv* 2024.01.02.573793 (2024) doi:10.1101/2024.01.02.573793.
144. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
145. Higgins, I. *et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. in (2017).
146. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
147. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
148. Fischer, M. Mice Are Not Humans: The Case of p53. *Trends Cancer* **7**, 12–14 (2021).
149. Lecca, P. Control Theory and Cancer Chemotherapy: How They Interact. *Front. Bioeng. Biotechnol.* **8**, 621269 (2021).
150. Gyamfi, J., Kim, J. & Choi, J. Cancer as a Metabolic Disorder. *Int. J. Mol. Sci.* **23**, 1155 (2022).
151. Ghaffari, P., Mardinoglu, A. & Nielsen, J. Cancer Metabolism: A Modeling Perspective. *Front. Physiol.* **6**, 382 (2015).
152. Buehler, M. J. Accelerating Scientific Discovery with Generative Knowledge Extraction, Graph-Based Representation, and Multimodal Intelligent Graph Reasoning. Preprint at https://doi.org/10.48550/arXiv.2403.11996 (2024).
153. Alber, M. *et al.* Integrating Machine Learning and Multiscale Modeling: Perspectives, Challenges, and Opportunities in the Biological, Biomedical, and Behavioral Sciences. *Npj Digit. Med.* **2**, 115 (2019).

154. Mambetsariev, I. *et al.* Clinical Network Systems Biology: Traversing the Cancer Multiverse. *J. Clin. Med.* **12**, 4535 (2023).