Statistical Machine Learning for Reliable Hypothesis Generation in Biomedical Problems

by

Tiffany Tang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Chair
Professor Haiyan Huang
Assistant Professor James B. Brown

Summer 2023

Statistical Machine Learning for Reliable Hypothesis Generation in Biomedical Problems

Abstract

Statistical Machine Learning for Reliable Hypothesis Generation in Biomedical Problems

by

Tiffany Tang

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Chair

Given the ever-growing volume and variety of biomedical data, principled analyses of these rich datasets offer an exciting opportunity to accelerate the scientific discovery process. Here, we advance our goal of extracting reliable scientific hypotheses from such data through (I) the in-context development of interpretable statistical machine learning methods, (II) the demonstration of responsible data science in practice, and (III) the dissemination of open-source software and data for reliable data science.

Throughout this dissertation, we build heavily upon the Predictability, Computability, and Stability (PCS) framework and documentation for veridical (trustworthy) data science (Yu and Kumbier, 2020) to improve the reliability of our scientific conclusions. This framework advocates for the use of predictability as a reality check, computability as an important consideration in algorithmic design and data collection, and stability as a minimum requirement for reproducibility and interpretability in knowledge-seeking and decision-making. Moreover, it calls on the need for transparent documentation of decisions made throughout the data science pipeline.

In Part I, we highlight two statistical machine learning methods, developed within the context of grounded biomedical problems and guided by the PCS framework. First, in Chapter 2, we investigate genetic and epistatic drivers of cardiac hypertrophy in hope of obtaining a more complete understanding of the disease architecture. To this end, we develop a data-driven recommendation system, named the low-signal signed iterative random forest (lo-siRF), to identify candidate genes and gene-gene interactions that are both predictive and stable across various model and data perturbations. We then phenotypically validate these genes and gene-gene interactions via gene-silencing experiments and investigate potential mechanistic explanations for the demonstrated epistases. This leads to a hypothesis in which the identified genes interact through mediating the variable binding of transcription factors that are essential for cardiac contractile function and metabolism. Second, the practical utility of random forests and interpretability tools, not only in the search for epistasis

but in a wide range of scientific problems, motivates the need for reliable tree-based feature importance measures. In Chapter 3, we demonstrate that the mean decrease in impurity (MDI), arguably the most popular random forest feature importance measure, suffers from well-known biases including against highly-correlated and low-entropy features. To overcome these drawbacks, we develop a novel feature importance framework, MDI+, which leverages a connection between MDI and the R^2 value from linear regression. We show that MDI+ improves the reliability and stability of feature importance rankings across an extensive range of data-inspired simulations and two real-data case studies on drug response prediction and breast cancer subtype prediction.

In Part II, we further expand on the theme of reliable data science and demonstrate it in practice through two collaborative projects in cancer -omics. In Chapters 4 and 5, we incorporate principles from the PCS framework while working in close collaboration with scientists and clinicians to identify stable and predictive biomarkers in drug response prediction and the early detection of pancreatic cancer, respectively.

Finally, in Part III, we introduce open-source software and data to promote and facilitate the broader adoption of reliable, transparent data science for statisticians and substantive researchers. In particular, we highlight three tools that support our goals: (1) `simChef`, an R package to simplify the creation of tidy, high-quality simulation studies (Chapter 6); (2) `vdocs`, an interactive virtual lab notebook in R to seamlessly implement, document, and justify human judgment calls throughout the data science pipeline in accordance with the PCS framework (Chapter 7); and (3) a COVID-19 data repository that aided community-wide data science efforts during the height of the pandemic (Chapter 8).

To my family

Contents

Contents **ii**

List of Figures **v**

List of Tables **viii**

1 Overview **1**
1.1 Part I: In-context development of interpretable machine learning methods guided by PCS . . . 2
1.2 Part II: Responsible data science in real-world biomedical problems applying PCS . . . 4
1.3 Part III: Open-source software and data . . . 6

I In-context development of interpretable machine learning methods guided by PCS **8**

2 Low-signal iterative random forests (lo-siRF) for epistasis discovery **9**
2.1 Introduction . . . 9
2.2 Results . . . 12
2.3 Discussion . . . 26
2.4 Methods . . . 27

3 MDI+: A flexible random forest-based feature importance framework **39**
3.1 Introduction . . . 39
3.2 Related Work . . . 42
3.3 Connecting MDI to R^2 Values from Linear Regression . . . 43
3.4 Introducing MDI+ . . . 47
3.5 Data-Inspired Feature Ranking Simulations . . . 51
3.6 MDI+ Overcomes Biases of MDI . . . 55
3.7 Case Studies . . . 58
3.8 Discussion . . . 62

II Responsible data science in real-world biomedical problems applying PCS **64**

4 A stability-driven protocol for drug response interpretable prediction (staDRIP) **65**
4.1 Introduction . . . 65
4.2 Results . . . 66
4.3 Discussion . . . 68
4.4 Methods . . . 69

5 Contribution of the microbiome to a metabolomic signature predictive of risk for pancreatic cancer **76**
5.1 Introduction . . . 76
5.2 Results . . . 77
5.3 Discussion . . . 82
5.4 Methods . . . 87

III Open-source software and data **94**

6 simChef: An R package for high-quality data science simulations using PCS **95**
6.1 Core abstractions of data science simulations . . . 96
6.2 A powerful grammar of data science simulations . . . 97
6.3 Additional Features . . . 99

7 vdocs: An R package for rigorous and transparent PCS documentation **101**
7.1 Core Features . . . 101
7.2 Future Roadmap . . . 103

8 Curating a COVID-19 data repository **105**
8.1 Overview of the COVID-19 data repository . . . 105
8.2 Discussion . . . 110

Bibliography **112**

A Low-signal iterative random forests (lo-siRF) for epistasis discovery **139**
A.1 Supplementary Figures . . . 139
A.2 Supplementary Tables . . . 145
A.3 Supplementary Notes . . . 148

B MDI+: A flexible random forest-based feature importance framework **149**
B.1 Proofs . . . 149

B.2 Feature Ranking Performance Simulations 150
B.3 Additional Data-Inspired Feature Ranking Simulations 164
B.4 Justifying MDI+ Choices 172
B.5 MDI Biases Simulations 176
B.6 PCS Model Recommendations 178
B.7 RF+ Improves Prediction Performance 181
B.8 Case Studies 183

C A stability-driven protocol for drug response interpretable prediction (staDRIP) 196
C.1 Supplementary Figures 196
C.2 Supplementary Tables 197

D Contribution of the microbiome to a metabolomic signature predictive of risk for pancreatic cancer 201
D.1 Supplementary Figures 201
D.2 Supplementary Tables 205

List of Figures

2.1 Schematic of the study workflow. 11
2.2 Low-signal signed iterative random forest (lo-siRF). 13
2.3 lo-siRF finds epistatic genetic drivers of left ventricular hypertrophy. 16
2.4 Gene disruption of identified epistatic interactions in hiPSC-CMs and high-throughput single-cell morphology assessment. 18
2.5 *CCDC141* non-additively interacts with *TTN* and *IGF1R* to modify cardiomyocyte morphology. 20
2.6 *CCDC141* SNPs co-occur with *TTN* and *IGF1R* SNPs in TEAD and GATA transcription factor binding motifs. 23
2.7 Enrichment analysis and weighted gene co-expression network analysis from human heart tissues confirmed interactions between *CCDC141*, *IGF1R*, and *TTN*. 25

3.1 Overview of MDI+. 49
3.2 Regression simulations for MDI+. 53
3.3 Classification simulations for MDI+. 54
3.4 Robust regression simulations for MDI+. 55
3.5 Correlation bias simulations for MDI+. 57
3.6 Entropy bias simulations for MDI+. 57
3.7 Stability results for MDI+ case studies on drug response prediction and breast cancer subtype prediction. 61

4.1 Overview of Cancer Cell Line Encyclopedia (CCLE) data. 70

5.1 Relationship between TMAO and indoleacrylic acid and microbial species. . . . 78
5.2 Predictive performance of the 3-marker microbial panel in the independent newly-diagnosed PDAC cohort. 81
5.3 Absolute 5-year risk estimates for individuals with CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel scores. 83
5.4 Workflow of analyses. 92

6.1 Overview of the four core components in a `simChef Experiment`. 96
6.2 Overview of running a `simChef Experiment`. 97
6.3 Example usage of `simChef`. 98

6.4 Detailed schematic of the `run_experiment` workflow using `simChef`. 100

7.1 Snapshot of the `vdocs` PCS lab notebook template. 102
7.2 Steps to create a new PCS Lab Notebook in RStudio. 102

A.1 Distribution of LVM and LVMi measurements for 30,000 UK Biobank participants. 139
A.2 LVMi GWAS for dimension reduction of candidate SNPs. 140
A.3 Differences in local stability scores between high and low LVMi highlight the importance of the lo-siRF-recommended gene-gene interactions. 141
A.4 Top SNPs from lo-siRF-recommended genes and gene-gene interactions. 142
A.5 Spiral-shaped inertial microfluidic channel for cell focusing and imaging. 143
A.6 Effects of lo-siRF recommended genes and gene-gene interactions on cell size and shape parameters. 144

B.1 Regression simulations for MDI+ under linear response. 152
B.2 Regression simulations for MDI+ under LSS response. 153
B.3 Regression simulations for MDI+ under polynomial interaction response. 154
B.4 Regression simulations for MDI+ under linear+LSS response. 155
B.5 Classification simulations for MDI+ under logistic response. 156
B.6 Classification simulations for MDI+ under logistic LSS response. 157
B.7 Classification simulations for MDI+ under logistic polynomial interaction response. 158
B.8 Classification simulations for MDI+ under logistic linear + LSS response. 159
B.9 Robust regression simulations for MDI+ under linear response with outliers. . . . 160
B.10 Robust regression simulations for MDI+ under LSS response with outliers. . . . 161
B.11 Robust regression simulations for MDI+ under polynomial interaction response with outliers. 162
B.12 Robust regression simulations for MDI+ under linear+LSS response with outliers. 163
B.13 Misspecified linear regression simulations for MDI+. 165
B.14 Misspecified LSS regression simulations for MDI+. 166
B.15 Misspecified polynomial interaction regression simulations for MDI+. 167
B.16 Misspecified linear+LSS regression simulations for MDI+. 168
B.17 Varying sparsity regression simulations for MDI+. 171
B.18 Varying number of features regression simulations for MDI+. 172
B.19 Modeling choices in MDI+ framework with $min\_samples\_per\_leaf$=5. 174
B.20 Modeling choices in MDI+ framework with $min\_samples\_per\_leaf$=1. 175
B.21 Frequency of random forest splits in correlation bias simulations. 176
B.22 Effects of leave-one-out sample splitting in the correlation bias simulations. . . . 177
B.23 Frequency of random forest splits in entropy bias simulations. 178
B.24 Effects of leave-one-out sample splitting in the entropy bias simulations. 178
B.25 PCS model selection simulations for MDI+. 180
B.26 Prediction performance of RF+ for regression and classification. 182
B.27 Full regression prediction results for RF+ (all 24 drugs). 182

B.28 Stability of top genes across train-test splits in drug response prediction case study (full results). 188
B.29 Stability of top genes across train-test splits in drug response prediction case study (MDI+, MDI, TreeSHAP only). 189
B.30 Stability of top genes across RF random seeds in drug response prediction case study (full results). 190
B.31 Stability of top genes across RF random seeds in drug response prediction case study (MDI+, MDI, TreeSHAP only). 191
B.32 Predictive power of top-ranked features from various feature importance methods in drug response prediction case study. 192
B.33 Predictive power of top-ranked features from various feature importance methods in breast cancer subtype prediction case study. 195

C.1 PCA and hierarchical clustering on CCLE RNASeq dataset. 196
C.2 Overview of CCLE data distribution. 199

D.1 Distribution plots for detected microbial-related metabolites across analytical batches in the PLCO specimen set. 202
D.2 Odds ratios and adjusted odds ratios for individual microbial-related metabolites for risk of pancreatic cancer in the Training Set. 203
D.3 Spearman correlation heatmap for microbiome-related metabolites in the Training Set. 204

List of Tables

4.1 List of most stable proteins associated with each drug using staDRIP. 68

5.1 Performance of microbial-related metabolites panels in different learning models in the PLCO Validation Set. 79
5.2 Performance of 3-marker microbial panel and a combined 3-marker microbial panel + 5-marker non-microbial panel for 5-year risk prediction of pancreatic cancer in the set-aside Test Set and the entire PLCO specimen set. 80
5.3 Performance estimates of the CA19-9 and a combined CA19-9+3-marker microbial panel + 5-marker non-microbial panel for 5-year risk prediction of pancreatic cancer in the set-aside Test Set and the entire PLCO specimen set. 86
5.4 Patient and tumor characteristics for PLCO cohort. 88

8.1 Select hospital-level features present in COVID-19 data repository 106
8.2 List of hospital-level datasets available in COVID-19 data repository 106
8.3 Select county-level features present in COVID-19 data repository 107
8.4 List of county-level datasets available in COVID-19 data repository 109

A.1 Characteristics of analyzed participants in the UK Biobank. 145
A.2 Binarization thresholds defining low and high LVMi groups used in siRF fit. . . 145
A.3 Prediction accuracies of methods across different binarization thresholds. 146
A.4 Summary of siRF evaluation metrics for top gene-gene interactions. 147
A.5 Top signed feature importance scores from lo-siRF across binarization thresholds. 148

B.1 Important genes in drug response prediction case study from various feature importance methods. 184
B.2 Summary of predictive power of top 10 features from various feature importance methods in drug response prediction case study. 193
B.3 Test prediction performance from RF+ and RF in breast cancer subtype prediction case study. 194
B.4 Important genes in breast cancer subtype prediction case study from various feature importance methods. 194

C.1 Validation prediction performance across various drug response prediction models. 197
C.2 Summary of best prediction model for each training data source. 197

C.3 Test error for each drug using the RNASeq-based kernel ridge regression model . 198
C.4 Most stable protein and RNASeq signatures for each drug, identified by staDRIP. 200
C.5 Most stable protein for each drug, identified by the elastic net. 200

D.1 Patient and tumor characteristics for the newly-diagnosed PDAC cohort. 205
D.2 Selected microbial-associated metabolites and corresponding model coefficients in LASSO regression. 206
D.3 Stability check of the LASSO regression using perturbed training data and evaluated on the Validation Set for the 3-marker microbial panel. 207
D.4 Selected non-microbial metabolites. 207
D.5 Performance of the 3-marker microbial panel amongst diabetic and non-diabetic individuals in the PLCO set-aside Test Set. 208
D.6 Performance of different learning models based on non-microbial metabolites in the PLCO Validation Set. 209
D.7 Stability check of the 5-marker non-microbial panel using perturbed training data and evaluated on the PLCO Validation Set. 209
D.8 Performance of the 5-marker non-microbial panel in the PLCO set-aside Test Set and the entire specimen set. 210
D.9 Performance of the 5-marker non-microbial panel and alone and in combination with the 3-marker microbial panel stratified by diabetic status. 211
D.10 Performance of the combined metabolite panel plus CA19-9 stratified by diabetic status. 212

Acknowledgments

This PhD would not have been possible without the support of amazing people around me. At every step along my educational and personal journey, they have taught and challenged me to grow both intellectually and personally. I am incredibly fortunate for their kindness, patience, and generosity, and they continue to inspire me every day.

In particular, I would first like to thank my advisor, Bin Yu, who has profoundly shaped my view of scientific research and taught me what it means to truly fight for your students. Even before starting my PhD, her support and mentorship were undeniable, and what I thought were simple phone calls were full of inspirational words of wisdom. I am lucky to have been able to work and learn from such a brilliant role model over the past five years. Through them and the many exciting opportunities and collaborations that she helped provide, she has always pushed me to achieve an exceptionally high standard of rigor while teaching me to think critically to reach that bar. Her leadership and passion for advancing science have also had a tremendous impact on my future goals and belief as to what is possible, and for that, I am especially grateful.

I am also grateful for the opportunity to work with and learn from many exceptional scientific collaborators. I would like to especially thank Qianru Wang, Chad Weldy, Weston Hughes, and Euan Ashley for their many tireless hours and contributions in the Biohub project. This work would not have been possible without Chad and Euan's leading clinical expertise, the deep learning-enabled phenotyping from Weston, and Qianru's novel microfluidics experimentation technology. Qianru, in particular, has been an amazing and gracious co-first-author. I have truly enjoyed working with her so closely. I have also had the pleasure of interacting with many other scientific researchers, to whom I am thankful for their advice and many insightful discussions including: James Priest, Rima Arnaout, Ben Brown, Victoria Parikh, and Atul Butte in the Biohub collaboration; Ehsan Irajizad, Johannes Fahrmann, and Sam Hanash in the MD Anderson collaboration; and Xuewei Wang and Jean-Pierre Kocher in the Mayo Clinic collaboration.

I have also been fortunate to have several other mentors along the way. Thank you to my managers at Genentech (Ning Leng and Jane Fridlyand) and Illumina (Yong Li and Jennifer Zou), who were incredibly patient and generous of their time. This experience outside of academia not only helped me to improve my communication and technical skills, but also helped me to ground my future research directions and interests in impactful real-world problems. Thank you to Ziad Obermeyer for serving on my qual committee and to Haiyan Huang and Ben Brown for serving on my qual and thesis committees. Before coming to Berkeley, I also had the great privilege to conduct research under Genevera Allen at Rice. She started me out on this journey to pursue statistics, and without her mentorship, I certainly would not be interested in statistics nor have had many of the opportunities today.

In addition to these mentors, I am thankful to have collaborated with many students in the Yu group. I am lucky to be surrounded by such a diverse, intelligent, and kind group of people, who not only helped me to grow intellectually but also were incredibly supportive throughout the ups and downs of the PhD. In particular, thank you to Ana Kenney, who

I had the pleasure of collaborating with in several projects and who has been a wonderful mentor and friend; Abhi Agarwal and Yan Shuo Tan for the many fun tree collaborations and long zoom conversations; Omer Ronen for joining me on various genomics project and sharing his passion for new scientific endeavors; James Duncan and Corrine Elliott for challenging me to improve both my programming and communication skills; Xiao Li, Karl Kumbier, and Merle Behr for their mentorship during the early years of my PhD; and other members of the Yu group including Aliyah Hsu, Austin Zane, Briton Park, Chandan Singh, Chengzhong Ye, Dennis Shen, Hue Wang, Nikhil Ghosh, Robbie Netzorg, Theo Saarinen, Wooseok Ha, and Yaxuan Huang, for the countless lunches and their continuous support.

Outside of the Yu group, I would like to give a big thank you to the statistics graduate student community at Berkeley and for all of the great memories at wind-downs, foosball tournaments, and other events. Thank you especially to the best officemates, Melody Huang, Austin Zane, Miyabi Ishihara, and Sara Stoudt; office neighbors, Eric Xia and Addison Hu; and many others including Dan Soriano, Taejoo Ahn, Michelle Yu, Ella Heismayer, and Arisa Sadeghpour, who made the day-to-day PhD life an absolute joy.

I would also like to thank the UC Berkeley Statistics Department staff, especially La Shana Polaris, Keyla Gomez, Tanisha Robinson, Chris Paciorek, and Ryan Lovett, who have truly gone above and beyond to make my life as a student as easy as possible.

Finally, I owe the biggest thank you to my family, who have supported me since the very beginning. They have sacrificed everything to give me the opportunities that led me here today. I would not be who I am nor where I am if it weren't for their unwavering love, support, and encouragement.

Chapter 1

Overview

Rapid technological advances in recent years have led to an unprecedented growth of data in the biomedical sciences. These advances in combination with large-scale coordinated efforts have led to the creation of vast biobanks, such as the UK Biobank (Sudlow et al., 2015) and Japan Biobank (Nagai et al., 2017), as well as comprehensive data consortiums, such as ENCODE (Consortium et al., 2012) and the Cancer Cell Line Encyclopedia (Barretina et al., 2012), among many others. This influx of data not only provides statisticians and data scientists with a vast playground to explore, but also a unique opportunity to collaborate with domain experts on complex scientific problems that could have significant impact.

For instance, in precision medicine, there has been widespread interest in developing increasingly powerful prediction models for tasks such as predicting the risk of a particular disease (Fröhlich et al., 2018; Johnson et al., 2021; Faizal et al., 2021). These models can help clinicians to identify high-risk individuals who should be prioritized for screening, more frequent follow-ups, and other preventative care measures. However, in this example and many other biomedical applications, it is critical to not only build accurate prediction models, but to be able to extract reliable interpretable insights from these models. In the clinic, a physician must be able to intelligibly explain their decision-making to the patient (Vellido, 2020). Likewise, for laboratory scientists, pinpointing the important features in the prediction model is key to generating scientific hypotheses that may inform future experiments (Chen and Ishwaran, 2012; Basu et al., 2018; Behr et al., 2020).

Given the potentially high impact and high-stakes decision-making in biomedical applications, it is thus worrisome that a scientific reproducibility crisis has loomed over the field and is currently ongoing (Chalmers and Glasziou, 2009; Begley and Ellis, 2012; Stupple et al., 2019). Notably, in a 2016 survey of over 1500 scientists, more than 70% of researchers reported that they have tried and failed to reproduce another scientist's experiments, and more than 50% have failed to reproduce even their own experiments (Baker, 2016). These findings combined with the potentially high stakes in the biomedical sciences underscore the need for reliable and veridical (trustworthy) data science (Yu and Kumbier, 2020) in practice.

To this end, the Predictability, Computability, and Stability (PCS) framework and documentation (Yu and Kumbier, 2020) was recently proposed to establish three core principles

to facilitate veridical data science in every step of the data science life cycle – namely, (1) *predictability* as evidence for whether or not the model is an accurate representation of reality; (2) *computability* as an important practical consideration in algorithmic design as well as data collection; and (3) *stability* of conclusions across reasonable data perturbations and human judgment calls as a minimum requirement for interpretability, reproducibility necessary for decision making, and new scientific knowledge. In addition, Yu and Kumbier (2020) call on the need for transparent documentation of the many human judgment calls that are inevitably made in the data science life cycle as they may substantially impact conclusions. These principles were originally motivated by extensive interdisciplinary research (Wu et al., 2016; Basu et al., 2018). It has since demonstrated a strong track record, driving scientific discoveries such as novel gene-gene interactions that drive the red-hair phenotype (Behr et al., 2020), clinically-interpretable subgroups in a randomized drug trial (Dwivedi et al., 2020), and a robust clinical-decision instrument for children after blunt torso trauma (Kornblith et al., 2022).

Rooted in the PCS framework for veridical data science, this dissertation works towards extracting reliable interpretations and generating reliable hypotheses from biomedical data. Here, we emphasize the task of reliable hypothesis generation, in particular, since available data is often noisy and imperfectly-positioned to solve the scientific problem under study outright. We thus view the scientific discovery process as an incremental process, where we first generate reliable scientific hypotheses from data and then subject these hypotheses to further validation (e.g., follow-up laboratory experiments). More often than not, several iterations of this process and numerous forms of validation are required before confidently claiming a scientific discovery.

In this dissertation, we advance our goal of conducting reliable data science and extracting reliable scientific hypotheses from biomedical data via the following three parts:

- **Part I**: We develop new interpretable machine learning methods while grounded in real-world biomedical problems.
- **Part II**: We demonstrate the practice of responsible data science under the PCS framework in two cancer -omics collaborations.
- **Part III**: We disseminate open-source software and data to facilitate reliable data science across the broader community of statisticians and substantive researchers.

We summarize our contributions for each part next.

1.1 Part I: In-context development of interpretable machine learning methods guided by PCS

To advance the boundaries of our scientific understanding, it is becoming increasingly important to develop statistical machine learning methods within the context of grounded scientific

problems. This *in-context development* ensures that the proposed methods are well-equipped to tackle the real-world complications that inevitably arise in these problems, such as low signal-to-noise ratios, complex correlation structures, and high-dimensional features.

Low-signal iterative random forests (lo-siRF) for epistasis discovery

In Chapter 2, we discuss one instance of this in-context methodological development for identifying epistatic gene-gene interactions that drive cardiac hypertrophy. Cardiac hypertrophy is a common heart disease characterized by an enlargement and thickening of the heart wall and carries significant risk for heart failure and sudden cardiac death. Currently, only a small fraction of the disease risk can be explained the known regulating genes. As such, knowledge of new epistatic drivers would not only provide clinicians with more refined understanding of the genetic architecture of cardiac hypertrophy, but also pave the way for more accurate diagnoses, preventative care, and new treatments.

Motivated by this, we develop an end-to-end pipeline for identifying genetic and epistatic drivers of cardiac hypertrophy in this highly-interdisciplinary work, which is joint with Qianru Wang, PIs Bin Yu and Euan Ashley, and many others (Nathan Youlton, Chad Weldy, Ana Kenney, Omer Ronen, Weston Hughes, Elizabeth T. Chin, Shirley C. Sutton, Abhineet Agarwal, Xiao Li, Atul J. Butte, Rima A. Arnaout, James B. Brown, James Priest, and Victoria N. Parikh). This pipeline consists of three major phases. First, using large-scale genetic and cardiac MRI data from the UK Biobank, we identify candidate genes and gene-gene interactions via a novel data-driven recommendation system, the low-signal signed iterative random forest (lo-siRF). At a high level, lo-siRF builds upon the computationally-tractable interaction search engine of signed iterative random forests (Basu et al., 2018; Kumbier et al., 2018), but has been tailored to address major practical challenges inherent with the UK Biobank data and cardiac phenotype — namely, the low signal-to-noise ratio, high-dimensionality, and high correlation between features. To overcome these challenges, lo-siRF leverages (1) deep learning and binarization to extract a refined, denoised phenotype from cardiac MRIs, (2) a genome-wide association study to perform dimension reduction in a domain-inspired manner, and (3) a new feature importance score to exploit the known correlation structure between single nucleotide polymorphisms within a gene locus. Furthermore, building on the PCS framework, we incorporate a prediction screening step to ensure that the learned model fits the data well. We also conduct extensive stability analyses to ensure the robustness of our gene-gene interactions to arbitrary human judgment calls (e.g., the choice of binarization).

Combining the lo-siRF recommendations with prior domain expert knowledge, we then chose a subset of the genes and gene-gene interactions to phenoytpically validate in the second phase via microfluidics-enabled gene-silencing experiments in high-throughput. These gene-silencing experiments demonstrate that cardiac hypertrophy is influenced by the lo-siRF-recommended genes and gene-gene interactions through an epistatic (i.e., non-additive interaction) model, expanding the scope of the genetic regulation of cardiac structure. In the final and third phase, we conclude with an investigation into possible mechanistic expla-

nations for the demonstrated epistases. This leads to a hypothesis in which the identified genes (*CCDC141*, *TTN*, and *IGF1R*) interact through mediating the variable binding of transcription factors that are essential for cardiac contractile function and metabolism.

MDI+: A flexible random forest-based feature importance framework

Inspired by both the need for trustworthy experimental recommendations and the effectiveness of random forests (RF) in the aforementioned work, we explore the reliability of popular RF feature importances in Chapter 3. In particular, mean decrease in impurity (MDI) (Breiman et al., 1984) is arguably the most popular choice for measuring feature importance for RFs, serving as the default RF feature importance method in `scikit-learn`. However, MDI is known to yield inaccurate feature importance rankings when there exists highly correlated features or features of varying entropy levels (Strobl et al., 2007, 2008; Nicodemus and Malley, 2009; Nicodemus, 2011). These complications with correlation and entropy are ubiquitous in practice (Nicodemus and Malley, 2009; Boulesteix et al., 2012).

To help explain these drawbacks of MDI, we draw a new connection between MDI and the R^2 value from linear regression. Moreover, we build upon this R^2 reinterpretation of MDI to develop MDI+, a new feature importance framework which generalizes MDI and overcomes its drawbacks. In short, MDI+ allows the analyst to (1) replace the linear regression model and/or R^2 metric with regularized generalized linear models (GLMs) and metrics better suited for the given data structure and (2) incorporate additional features or knowledge to mitigate known biases of decision trees such as their inefficiency in fitting additive or smooth models. Given the many choices that can be made within the MDI+ framework, we also provide guidance on how practitioners can choose the appropriate GLM and metric using the PCS framework. We demonstrate the effectiveness of MDI+ across an extensive range of data-inspired simulations as well as two real-world case studies on drug response prediction and breast cancer subtype prediction. In both case studies, MDI+ identifies well-established predictive genes and yields the most similar feature importance rankings across RFs trained on different train-test splits, compared to existing feature importance measures. Given that it is highly undesirable for findings to change due to an arbitrary choice such as the train-test split (Yu and Kumbier, 2020), the improved stability of MDI+ is a significant practical advantage and a promising step towards extracting more reliable hypotheses from biomedical data. This work is joint with Abhineet Agarwal, Ana Kenney, Yan Shuo Tan, and Bin Yu.

1.2 Part II: Responsible data science in real-world biomedical problems applying PCS

As advocated by the PCS framework (Yu and Kumbier, 2020; Yu, 2013), a crucial prerequisite for new scientific knowledge is *stability*, not only across reasonable model and data

perturbations (e.g., using different random seeds) but also across human judgment calls (e.g., choices in data collection, data preprocessing/cleaning, modeling, evaluation metrics). If not carefully chosen, arbitrary human judgment calls made throughout the data science pipeline can substantively impact findings and result in spurious downstream conclusions. Aligning ourselves with the PCS framework, we thus emphasize the need to (1) collaborate closely with interdisciplinary researchers to make these human judgment calls in an informed manner, (2) justify these human judgement calls, when possible, via transparent documentation, and (3) conduct rigorous stability analyses to help ascertain the impact of these decisions on conclusions. To exemplify these principles in practice, we highlight two biomedical collaborations in cancer -omics.

A stability-driven protocol for drug response interpretable prediction

In Chapter 4, we present a collaboration with Xiao Li, Xuewei Wang, Jean-Pierre Kocher, and Bin Yu, focusing on drug response prediction. Our goal is to predict the response (or efficacy) of a cancer drug given an individual's molecular -omics profile and to identify the important biomarkers that are most predictive of the drug's response. However, the high heterogeneity and noise in cancer -omics and pharamacological data pose severe practical challenges. One main challenge is that these drug response prediction models and interpretations thereof are often volatile and depend heavily on arbitrary modeling decisions and other human judgment calls. We hence develop a transparent stability-driven pipeline for drug response interpretable predictions, or staDRIP, which builds upon the PCS framework in order to mitigate the impact of human judgment calls on downstream conclusions. This approach places greater weight on the -omics features that were stably important across multiple machine learning models and across reasonable perturbations of the data. In doing so, we identified proteins that have been associated with the drug response in the previous literature at an approximately 20% higher rate than previous methods.

Contribution of the microbiome to a metabolomic signature predictive of risk for pancreatic cancer

Following a similar spirit in Chapter 5, we detail a collaboration with Ehsan Irajizad, Ana Kenney, Bin Yu, Sam Hanash, Johannes Fahrmann, and others, focusing on the early detection of pancreatic cancer. In this work, we investigate whether increases in circulating microbial-related metabolites are associated with pancreatic cancer risk. Using metabolomics profiling of participants in the Prostate, Lung, Colorectal and Ovarian (PLCO) cohort, we develop a three-marker microbial-related metabolite panel, which is predictive of pancreatic cancer up to five years prior to diagnosis (test AUC: 0.64). To enhance the reproducibility and reliability of our finding, we stress-test this 3-marker microbial panel and conduct numerous stability analyses in accordance with the PCS framework. We finally validate

this 3-marker microbial panel using samples from three different PLCO centers and an independent cohort of newly-diagnosed pancreatic cancer cases and controls, demonstrating it achieves similar prediction performance.

1.3 Part III: Open-source software and data

In the last part of this dissertation, we highlight several efforts on the open-source software and data front to facilitate reliable, transparent data science in practice.

simChef: An R package for high-quality data science simulations using PCS

First, we develop an R package, `simChef`, to empower data scientists with an intuitive, extensible, and reusable framework for data science simulations. By removing many of the administrative burdens of simulation design, `simChef` allows data scientists to focus their attention on scientific best practices. In particular, to facilitate high-quality simulations, guided by the PCS framework, `simChef` makes it very easy to vary the data-generating processes, methods, and perturbations thereof in the simulation study. In Chapter 6, we highlight `simChef`'s main features, including (1) its intuitive Tidyverse-inspired grammar for running simulations, (2) its automated documentation and visualization of results, and (3) its flexible utilities for efficient distributed computation, caching, and checkpointing. This work is joint with James P. Duncan, Corrine F. Elliott, Philippe Boileau, and Bin Yu. GitHub code is available at `https://github.com/Yu-Group/simChef`.

vdocs: An R package for rigorous and transparent PCS documentation

In Chapter 7, we introduce `vdocs`, an R package to facilitate transparent PCS documentation for applied scientific research. `vdocs` provides practitioners with an interactive platform (currently, an interactive R Markdown notebook) to seamlessly implement, document, and justify human judgment calls throughout an analysis. This virtual lab notebook is akin to a scientist's physical lab notebook and walks users through a series of questions and stability checks to enhance the reliability of the analysis. This work is joint with Ana Kenney, Ehsan Irajizad, and Bin Yu. GitHub code is available at `https://github.com/Yu-Group/vdocs`.

Curating a COVID-19 data repository

Lastly, we curated and maintained a large open-source COVID-19 data repository to aid community-wide data science efforts during the COVID-19 pandemic. This data repository, detailed in Chapter 8, laid the foundation for our highly collaborative COVID-19 severity prediction work in Altieri et al. (2020) as well as other COVID-19 research efforts from the

broader community. At its peak, we attracted around 12,000 visits, 1,100 unique visitors, and 108 clones on GitHub over the span of two weeks. Though no longer under-going active development, the data repository is open-source on GitHub at `https://github.com/Yu-Group/covid19-severity-prediction`. This work is joint with Yu Wang, Chandan Singh, Bin Yu, and many others (Nick Altieri, Rebecca L. Barter, James P. Duncan, Raaz Dwivedi, Karl Kumbier, Xiao Li, Robert Netzorg, Briton Park, Yan Shuo Tan, and Chao Zhang).

Part I

In-context development of interpretable machine learning methods guided by PCS

Chapter 2

Low-signal iterative random forests (lo-siRF) for epistasis discovery

2.1 Introduction

Cardiac function and cardiac disease risk are intricately linked to the structure of the heart (Weldy and Ashley, 2021). In particular, heart failure is influenced by structural features including ventricular size and wall thickness (Sharir et al., 1994; Bastos et al., 2019; Udelson et al., 1988; Burkhoff et al., 2005). Left ventricular hypertrophy (LVH) — increased thickness of the left ventricular (LV) wall — is a complex phenotypic trait influenced by multiple factors, particularly afterload (Lazzeroni et al., 2016). Progressive LVH further carries significant independent risk for incident heart failure, atrial arrhythmia, and sudden death (Haider et al., 1998; Chrispin et al., 2014; Kawel-Boehm et al., 2019; Bluemke et al., 2008), highlighting the need to understand the genetic determinants of cardiac structure.

Importantly, recent discoveries leveraging cardiac magnetic resonance imaging (MRI) in the UK Biobank have revealed that cardiac structure is in part determined by complex genetics (Pirruccello et al., 2020; Meyer et al., 2020; Harper et al., 2021). Common genetic variants, many located near genetic loci associated with dilated cardiomyopathy and heart failure, have been found to influence LV size and systolic function (Pirruccello et al., 2020). Further, specific genetic variants that influence the amount of LV trabeculation have been shown to affect LV systolic function and overall risk of cardiomyopathy (Meyer et al., 2020). These pivotal findings underscore the importance of complex genetics in cardiac structure and disease risk and have resulted in the discovery of novel gene functions. However, these identified genetic variants remain inadequate to explain the total heritable disease risk (O'Sullivan et al., 2022). Indeed, common genetic variants rarely act independently and additively as modeled by many genome-wide association studies (GWAS) (Guindo-Martínez et al., 2021). There is growing evidence to support a disease risk model in which multiple genes interact non-additively with each other through epistasis (Zeng et al., 2022; Li et al., 2020). LVH is not only a complex phenotypic trait influenced by common gene variation,

but also a distinguishing clinical feature of hypertrophic cardiomyopathy (HCM), which is driven by rare pathogenic monogenic variants in sarcomere genes (Marian and Braunwald, 2017). Recent work has shown that common genetic variation influences HCM susceptibility and expressivity, demonstrating that common genetic variation affecting LVH further impacts Mendelian HCM disease penetrance (Harper et al., 2021). This raises the fascinating possibility that common gene-gene epistatic interactions drive cardiac phenotype, holding enormous potential for uncovering disease mechanisms and developing potential therapeutic strategies.

However, several computational and experimental challenges need to be resolved to ensure robust identification of epistasis. First, detecting epistasis can be computationally prohibitive due to the combinatorial nature of possible interactions. To reduce the computational burden, iterative random forests (iRF) were developed to discover reliable higher-order (not only pairwise) nonlinear interactions in a computationally-tractable manner (Basu et al., 2018). Second, many studies on epistasis were not replicated, raising concerns over the lack of scientific reproducibility (Reimherr and Nicolae, 2011; Murk et al., 2015). To achieve more trustworthy results, a new framework (Yu and Kumbier, 2020) for veridical data science advocates for three core principles - predictability, computability, and stability (PCS) - and calls on the need for transparent documentation of decisions made in data analysis pipelines. Another challenge is the generally small effect size of common genetic variants (O'Sullivan et al., 2022; Koch and Sunyaev, 2021) and the large phenotypic diversity of cardiac hypertrophy, which impedes both the data-driven discovery and functional validation of epistatic interactions. In human biobanks, recent advances in deep-learning-enabled phenotyping (Bai et al., 2018) using cardiac MRI images have led to more refined phenotypes at larger scales than were previously possible. At the cellular level, high-throughput microfluidic technologies (Wang et al., 2019; Di Carlo, 2009; Guan et al., 2013) have been integrated with AI-based image analysis of single-cell morphology (Wu et al., 2020) and human induced pluripotent stem cell (iPSC)-derived cardiomyocytes (hiPSC-CMs) (Dainis et al., 2020), opening up new possibilities for label-free and rapid detection of phenotypic consequences of genetic perturbation.

In this study, we leveraged large-scale genetic and cardiac MRI data from the UK Biobank (Bycroft et al., 2018) and identified a novel role of epistasis in regulating cardiac hypertrophy. More specifically, we innovated an iRF-based pipeline (Figure 2.1), rooted in the new PCS framework for veridical data science. This pipeline integrated a deep-learning-based LV structural analysis (Bai et al., 2018) to identify candidate epistatic drivers from low-signal (noisy) genomics data. This allowed us to identify previously unreported epistatic pairwise interactions between gene loci that mediate LVH. Using image-based microfluidics in combination with a high-throughput single-cell morphology analysis and hiPSC-CMs harboring RNA silencing of selected putative epistatic interactions, we showed that specific pairwise interactions between *CCDC141* and two other genes (*TTN* and *IGF1R*) can rescue the pathologic cardiomyocyte hypertrophy caused by a Mendelian HCM risk variant. We further investigated these genes in transcriptional regulation networks and weighted gene co-expression networks built from more than 300 human heart tissues (Cordero et al., 2019).

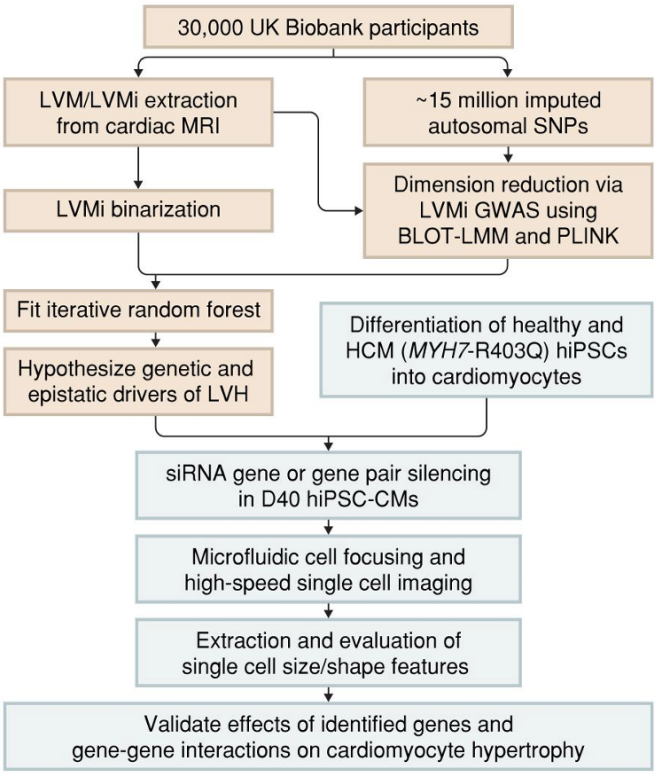


Figure 2.1: Data of genotype and cardiac MRI-derived left ventricular mass indexed to body surface area (LVMi) from 30,000 UK Biobank participants were inputted into an iterative random forest-based search engine to predict genes and epistatic gene-gene interactions that drive left ventricular hypertrophy (LVH). Identified genes and pairwise interactions were silenced in Day 40 cardiomyocytes differentiated from human induced pluripotent stem cells (hiPSCs) derived from healthy donors and patients with hypertrophic cardiomyopathy (HCM). Size and shape features of single cardiomyocytes were imaged and evaluated using an advanced microfluidic technique to validate the impact of hypothesized epistatic interactions between identified genes on cardiomyocyte hypertrophy.

This analysis led to a hypothesis in which the identified genes interact through mediating the variable binding of transcription factors (TFs) that are essential for cardiac contractile function and metabolism. These results demonstrate a complex paradigm for the integration of gene-phenotype associations in cardiac health and disease: common genetic variants affect cardiac hypertrophy through a mechanism of epistasis, and these gene-gene interactions impact rare variant effects on cardiomyocyte phenotype.

2.2 Results

Veridical machine learning enables search for epistatic drivers of left ventricular hypertrophy in low-signal data

We analyzed single-nucleotide polymorphism (SNP) and cardiac MRI data from 30,000 unrelated White British individuals in the UK Biobank (Figure 2.2a, Table A.1). To reliably infer epistatic drivers of LVH from this data, we developed a data-driven pipeline, the low-signal signed iterative random forest (lo-siRF), which builds upon an enhanced version of iRF (Basu et al., 2018; Kumbier et al., 2018) and is heavily guided by the PCS framework for veridical (trustworthy) data science (or machine learning) (Yu and Kumbier, 2020). As a veridical machine learning strategy, lo-siRF aims to extract trustworthy gene-gene interactions by: ensuring that the learned model fits the data well; aligning statistical modeling decisions with the scientific phenomena under study; and conducting stability analyses to improve the robustness of our gene-gene interactions.

More specifically, to first quantify LVH, lo-siRF begins with a careful choice of phenotypic data for LVH by leveraging recent advances in deep learning (Bai et al., 2018) to extract the LV mass indexed to body surface area (LVMi) from cardiac MRIs (Figure 2.2a and Figure A.1, details in *Methods*). lo-siRF then takes several steps to obtain trustworthy gene-gene interactions given the low phenotypic signal, which arises from the small effect size of common genetic variants and phenotypic diversity of cardiac hypertrophy. Namely, lo-siRF incorporates the following steps (details in *Methods*), rooted in the PCS framework:

1. *A biologically-inspired dimension reduction step*: we use GWAS to mitigate challenges when searching across millions of possible SNPs (Figure 2.2b and Figure A.2);

2. *A binarization step*: we binarize the LVMi phenotype measurements into a high and low LVMi category to denoise and facilitate model checking (Figure 2.2c);

3. *A prediction step*: we fit an enhanced version of iRF, namely, signed iRF (siRF) (Basu et al., 2018; Kumbier et al., 2018), a stable and powerful nonlinear interaction search engine that yields a competitive classification accuracy of 55% relative to other common machine learning prediction algorithms (Figure 2.2d); and

4. *An interpretation step*: we develop a novel stability-driven feature importance score (Figure A.3), which leverages the siRF fit and a permutation test, to finally rank the gene-gene interactions (Figure 2.2e). This feature importance score provides the necessary new interpretable machine learning ingredient to complete the lo-siRF discovery pipeline.

A more detailed documentation and justification of our modeling decisions can be found in the supplementary PCS documentation.

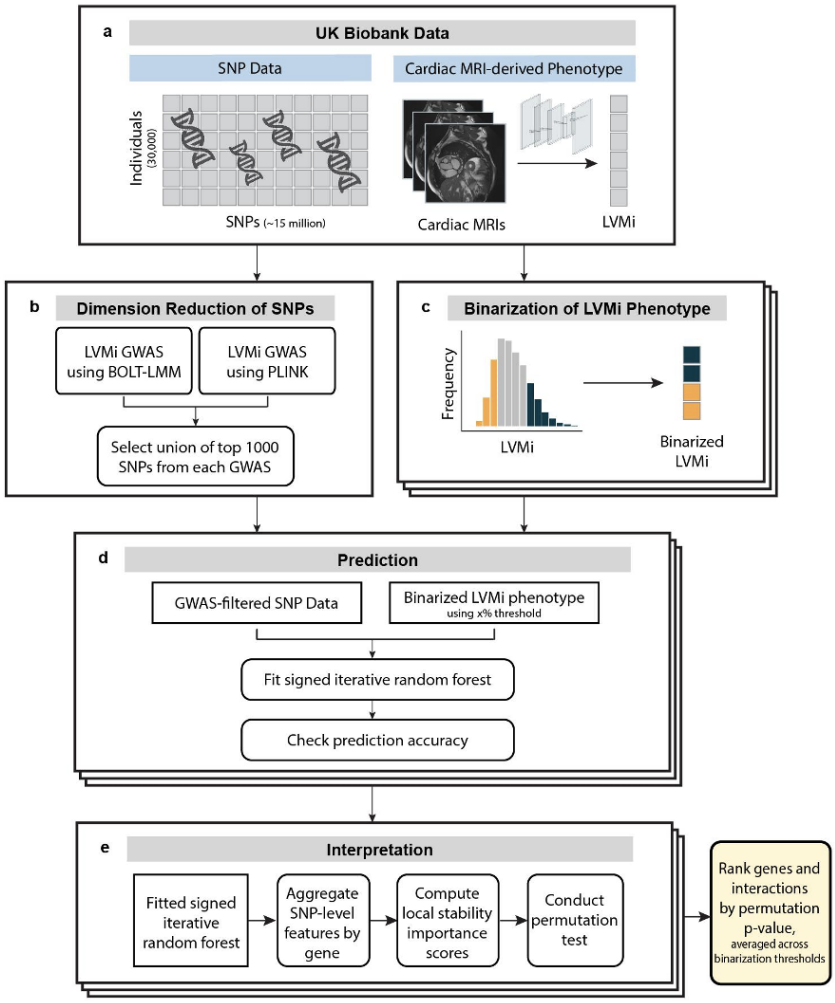


Figure 2.2: (a) Cardiac MRIs and SNP data were pulled from the UK Biobank, and LVMi was derived from the cardiac MRIs via deep learning31. (b) Dimension reduction was performed via GWAS to concentrate the analysis on a smaller set of SNPs. (c) LVMi was binarized into high and low LVMi categories according to three different binarization thresholds (represented by the stacked boxes). (d) For each of the three binarization thresholds, signed iRF was fitted using the GWAS-filtered SNPs to predict the binarized LVMi phenotype, and the validation prediction accuracy was assessed. (e) Genes and gene-gene interactions were ranked according to their importance across the three iRF fits, as measured by a novel stability-driven importance score.

Low-signal signed iterative random forests identify novel epistasis between four genomic loci

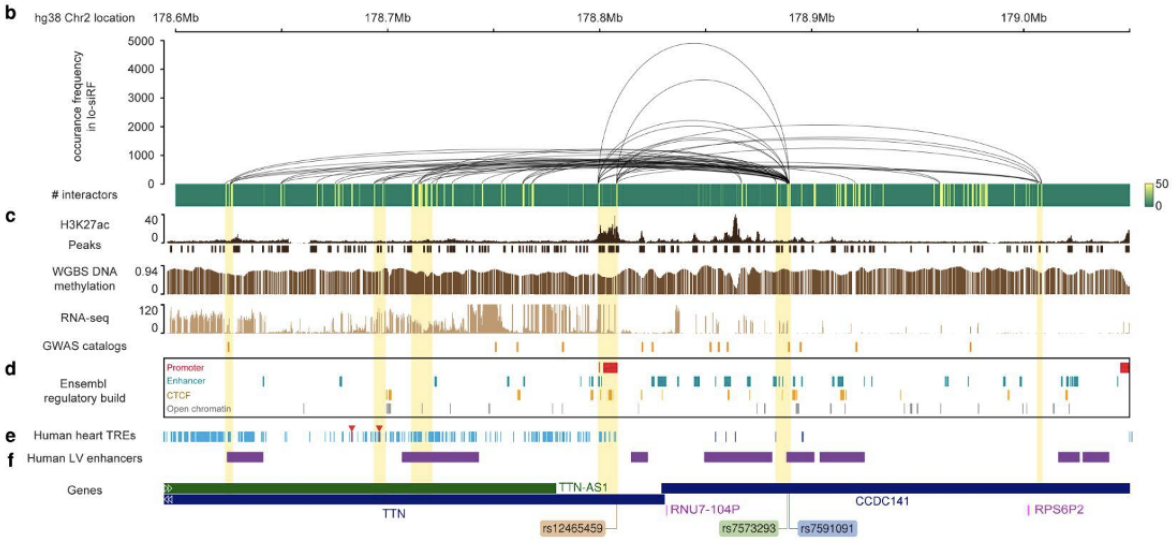
lo-siRF identified pairwise epistasis between SNPs at four genomic loci that exhibited stable and reliable associations with LVMi (highlighted in blue in Figure 2.3a). Specifically, an interaction between the *CCDC141* and *IGF1R* loci (denoted as *CCDC141-IGF1R* in Figure 2.3a) showed the strongest association to LVMi, followed by an interaction between the *CCDC141* and *TTN* loci (*CCDC141-TTN*), and then an interaction between the *CCDC141* locus and the *LOC157273*;*TNKS* intergenic region (*CCDC141-LOC157273*;*TNKS*). lo-siRF provides additional information regarding the sign (or direction) of these interactions; however, given our primary goal of recommending candidates for experimental validation, we omit the signs here and discuss it further in the online PCS documentation.

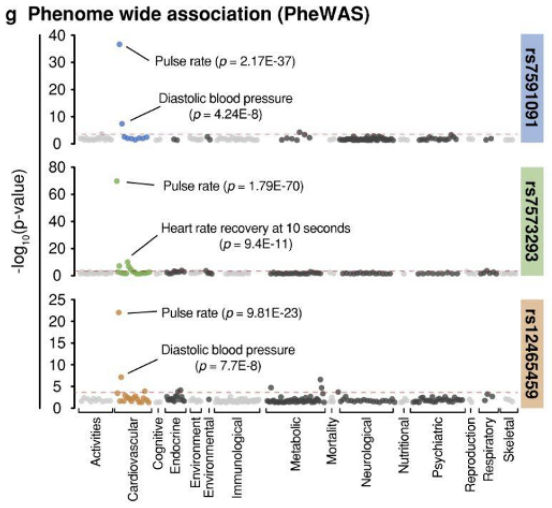
Interestingly, all of these interactions involved the *CCDC141* locus. Furthermore, the *CCDC141* and *TTN* loci exhibit genomic proximity (Figure 2.3b-f). Their interaction, however, does not appear to stem from this proximity. Though previous GWAS studies (Verweij et al., 2018; Thorolfsdottir et al., 2021) have identified missense and intronic variants in *CCDC141* in high linkage disequilibrium (LD) with loci in *TTN*, *CCDC141* and *TTN* have also been marginally associated with LVM/LVMi separately (Khurshid et al., 2023; Aung et al., 2019). Moreover, our lo-siRF results show that all the identified SNPs in *CCDC141* were in low LD with their interacting SNPs found in *TTN* ($R^2 < 0.6$), suggesting independent roles of each genetic locus in their epistasis. Data from GTEx (GTEx Consortium, 2013) also reveal that the 5 SNPs in the *CCDC141* locus with the highest occurrence frequency in lo-siRF (Figure 2.3b), are expression quantitative trait loci (eQTLs) for *CCDC141*, indicating their effects on regulating *CCDC141* gene expression levels. Of the 51 SNPs identified by lo-siRF at the *TTN* locus, 20 are splicing quantitative trait loci (sQTLs) for *TTN*, suggesting their impact on the alternative splicing of *TTN* pre-mRNA. Some of these identified SNPs (e.g., rs7591091 and rs12465459) are eQTLs for both *CCDC141* and *TTN* (Figure 2.3h&j), indicating that these SNPs may regulate the expression of both genes. In contrast, some SNPs (e.g., rs7573293), are eQTLs for *CCDC141* alone and do not appear to impact *TTN* expression (Figure 2.3i). These findings possibly suggest both independent roles of *CCDC141* and *TTN* and their potential interaction in the regulation of cardiac hypertrophy, which we corroborate via mechanistic experiments discussed in the following section.

Beyond this interaction with the *TTN* locus, lo-siRF suggests that the *CCDC141* locus interacts with the *IGF1R* locus and the *LOC157273*;*TNKS* intergenic region. Data from GTEx show that all six SNPs identified at the *IGF1R* locus are eQTLs for *IGF1R* in left ventricles. In addition, lo-siRF identified five loci that exhibited strong marginal associations with LVMi (Figure 2.3a). Three of these identified loci (near *IGF1R*, *TTN*, and *CCDC141*) were involved in the interactions discussed. The remaining two loci are *LSP1* and the intergenic region between *MIR588* and *RSPO3*. *LSP1* encodes the lymphocyte specific protein 1 and has been associated with hypertension and systolic blood pressure (Kanai et al., 2018; Ehret et al., 2016). *RSPO3* encodes R-spondin3, which has been reported to modulate Wnt-signaling and promotes coronary artery formation (Da Silva et al., 2017).

a lo-siRF gene and gene-gene interaction recommendations

Gene / Interaction	Top SNP	Chr	hg38 Position	Risk/Alt Allele	Function	GWAS Beta	GWAS SE	Mean lo-siRF p-value
CCDC141–IGF1R	rs7591091 rs55686521	2 15	178889467 98727212	T/C T/C	intronic intronic	--	--	$< 10^{-3}$
IGF1R	rs62024491	15	98733068	G/A	intronic	-0.0426	0.00962	$< 10^{-3}$
MIR588;RSPO3	rs9401921	6	126925592	A/G	intergenic	0.0474	0.00926	0.002
TTN	rs66733621	2	178799323	A/G	intronic	0.0437	0.00983	0.009
CCDC141–TTN	rs7591091 rs66733621	2 2	178889467 178799323	T/C A/G	intronic intronic	--	--	0.011
LSP1	rs569550	11	1865838	T/G	intronic	0.0423	0.00915	0.017
CCDC141	rs7591091	2	178889467	T/C	intronic	-0.0646	0.00975	0.018
CCDC141–LOC157273;TNKS	rs7591091 rs6999852	2 8	178889467 9478458	T/C G/A	intronic intergenic	--	--	0.056





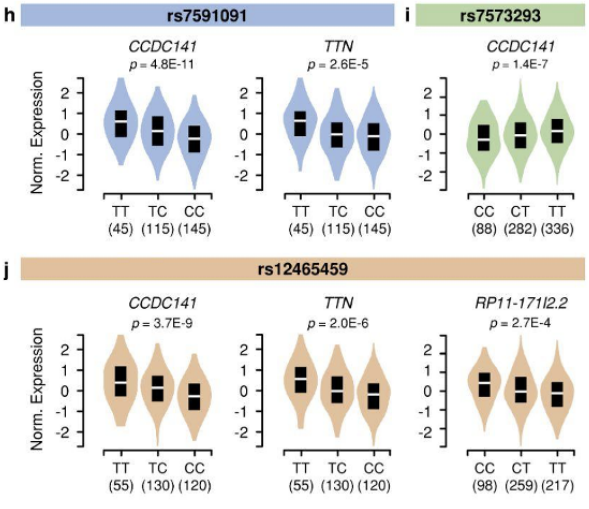


Figure 2.3 *(previous page)*: (a) Top genes and gene-gene interactions (highlighted in blue), along with the SNP(s) most frequently occurring in lo-siRF. Semicolon indicates intergenic regions. Lo-siRF captures associations that GWAS failed to identify (e.g., *IGF1R* and its interaction with *CCDC141*). Chr, chromosome; SE, standard error. Other lo-siRF-prioritized SNPs and SNP-SNP interactions are shown in Figure A.3. (b) Top 50 lo-siRF-prioritized interactions identified between *CCDC141* and *TTN* SNPs. The height of each ellipse indicates the occurrence frequency of the corresponding SNP-SNP interaction in lo-siRF averaged across the three LVMi binarization thresholds (Figure 2.2b). Also shown is a heatmap of the number of interactors for each SNP. These interacting SNPs are overlapped in the genomic context with (c) a comprehensive GTEx dataset from human left ventricle tissues, including H3K27ac Chip-seq and peak tracks, WGBS DNA methylation, RNA-seq coverage, and GWAS catalogs. These interacting SNPs are also overlapped with human heart regulatory regions from (d) Ensembl regulatory build (Martin et al., 2023) and recently published data of (e) transcribed regulatory elements (TREs) and (f) active enhancers (Spurrell et al., 2022). (g) Manhattan plots of PheWAS results ($p < 0.05$) stratified by phenotypic domains for two *CCDC141* SNPs (top, rs7591091, middle, rs7573293) and a *TTN* SNP (bottom, rs12465459) that frequently appear in lo-siRF. Associations with cardiovascular phenotypes are highlighted in blue, green, and orange for these three SNPs, respectively. Red dashed lines represent Bonferroni corrected p values of 2.98E-4, 3.55E-4, and 2.45E-4, respectively. (h) Violin plots of *CCDC141* (left) and *TTN* (right) expression for rs7591091 in GTEx (release v8) indicate that this SNP is an eQTL for both *CCDC141* and *TTN*. In contrast, the other *CCDC141* SNP, rs7573293 (i), is an eQTL for *CCDC141* only. (j) Violin plots of *CCDC141* (left), *TTN* (middle), and *RP11-171I2.2* (right, a novel transcript antisense to titin) expression for eQTL rs12465459 in GTEx. Box plots in (h)-(j) show the allelic effects of the given SNP on normalized expression levels of a given gene, and white lines represent the median values.

Disrupting putative interactions in hiPSC-CMs reveals the epistatic regulation of cardiomyocyte hypertrophy

To phenotypically characterize the identified epistatic interactions, we interrogated how disrupting these putative interaction candidates can affect single cell hypertrophy. We elected to use a genetic model of hypertrophy: hiPSC-CMs derived from patients with and without the foundational Mendelian variant associated with hypertrophic cardiomyopathy (*MYH7* p.R403Q) (Dainis et al., 2020). *MYH7* encodes the cardiac myosin heavy chain 7, a key component of the cardiac sarcomere, which is among the most common genes known to cause this disease (Dainis et al., 2020). Typically, a family history of arrhythmic sudden death and heart failure is seen across multiple generations. This was the case in this family where echocardiography demonstrated severe LVH and a small LV cavity (Dainis et al., 2020).

At the cellular level, hiPSC-CMs carrying the *MYH7*-R403Q mutation showed an elevated mean cell size and prolonged size distribution on the larger side relative to the unaffected control (Figure 2.5a), which concords with reported enlargement of cell size (Eschenhagen and Carrier, 2019) in HCM lines.

To determine if *CCDC141* can act both independently and in epistatic interactions with other genes to attenuate the pathologic cellular hypertrophy caused by *MYH7*-R403Q, we silenced genes *CCDC141*, *IGF1R*, *TTN*, and gene pairs *CCDC141-IGF1R* and *CCDC141-TTN* using siRNAs (220 $\sim$ 316 kbps in length and incorporating locked nucleic acids for specificity, details in *Methods*) in both healthy and *MYH7*-R403Q variant hiPSC-CMs with high efficiency (Figure 2.5b). The resulting cell suspensions were focused into fluid streams using a spiral inertial microchannel (Figure 2.4b and Figure A.5), and photographed to derive cell size and morphology information using a customized image analysis pipeline (Figure 2.4c) in high-throughput. Because the large variety in cardiomyocyte diameters (Figure 2.5a) can lead to differential cell focusing positions (Hood et al., 2016) and thus poor imaging resolution (Stavrakis et al., 2019) of flowing cells, we used inertial microfluidics to address this problem and adopted the Dean flow focusing principle (Guan et al., 2013) to bifurcate the randomly dispersed cardiomyocytes into two streams of large and small cells (Figure 2.4b). These sorted cells were further focused in separate channels to allow high-resolution imaging and rapid extraction of single cell morphology features (details in Figure A.5 and *Methods*).

We analyzed the relative differences between size of cells with silenced genes or pairwise gene interactions compared to their scrambled controls (Figure 2.5c). Bootstrapped hypothesis tests were performed to compare medians and quantiles, for which the p-values are capped below by $p < 10^{-4}$. Additional analyses were performed to ensure that our conclusions are robust against human judgment calls (see *Methods* and Appendix A).

Our results show an epistatic interaction between *CCDC141* and *IGF1R* in reducing hiPSC-CM cell size. More specifically, silencing *IGF1R* alone reduces the median cell size by $5.3\% \pm 0.4\%$ ($p < 10^{-4}$) in *MYH7*-R403Q variant hiPSC-CMs and $6.6\% \pm 0.5\%$ ($p < 10^{-4}$) in healthy cells. Silencing *CCDC141* alone also decreases median cell size relative to the scrambled control by $3.2\% \pm 0.5\%$ ($p < 10^{-4}$) in *MYH7*-R403Q variant cells, but indicates no significant impact on healthy cells. Digenic silencing of *CCDC141-IGF1R* shows a synergic effect on attenuating the pathologic cell hypertrophy in *MYH7*-R403Q variant cells, resulting in an $8.5\% \pm 0.3\%$ ($p < 10^{-4}$) decrease in the median cell size. This is consistent with the case of healthy cells, for which although silencing *CCDC141* alone appears not to affect cell size, the median cell size decreases by $9.3\% \pm 0.5\%$ ($p < 10^{-4}$) when silencing the additional gene *IGF1R*. Moreover, this effect from silencing *CCDC141-IGF1R* appears to be non-additive for both healthy and *MYH7*-R403Q variant cells ($p < 10^{-4}$ for non-additivity), suggestive of an epistatic mechanism. These findings agree with the strongest association of the *CCDC141-IGF1R* interaction identified by lo-siRF (Figure 2.3a).

Similarly, for both healthy and disease samples, digenic silencing of *CCDC141-TTN* leads to the most pronounced reduction in median cell size (by $5.8\% \pm 0.6\%$ for healthy cells and $3.3\% \pm 0.4\%$ for *MYH7*-R403Q variant cells, $p < 10^{-4}$) relative to monogenic silencing. This effect on reducing cell size appears to be non-additive for both healthy

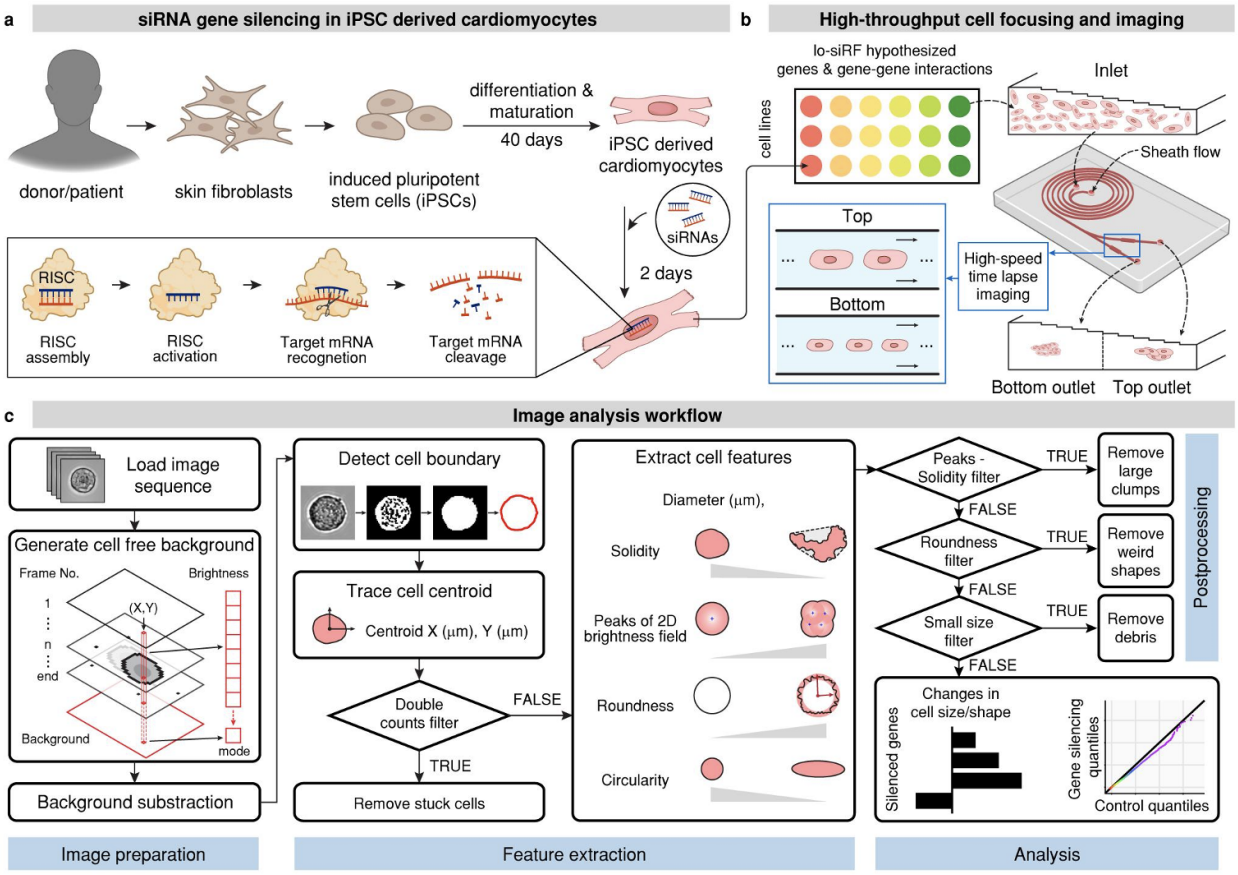


Figure 2.4: (a) Human induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) with and without hypertrophic cardiomyopathy (carrying an *MYH7*-R403Q mutation) were transfected with scramble siRNA or siRNAs specifically targeting single (*CCDC141*, *IGF1R*, and *TTN*) or combined (*CCDC141-IGF1R* and *CCDC141-TTN*) regions recommended by lo-siRF (Figure 2.3b). (b) Gene-silenced hiPSC-CMs were detached from culture dishes and loaded into a spiral microfluidic device using a syringe pump. Interplay of inertial forces focused randomly dispersed large and small cells into the top or bottom microchannel outlets (cell focusing mechanism illustrated in Figure A.4), respectively, which enables high-resolution imaging of flowing cells. (c) Workflow of the image analysis process. Time-lapse image sequences of single cells passing through the top and bottom microchannel outlets were fed into a customized MATLAB-based program that extracts cell size/shape features via a sequential process of bright field background correction, cell boundary detection, cell tracking and stuck cell removal, cell feature extraction, data postprocessing, and data analysis. Extracted cell size/shape features for each gene-silencing condition were compared with their scramble controls to validate the potential role of epistasis in the genetic regulation of cardiomyocyte hypertrophy.

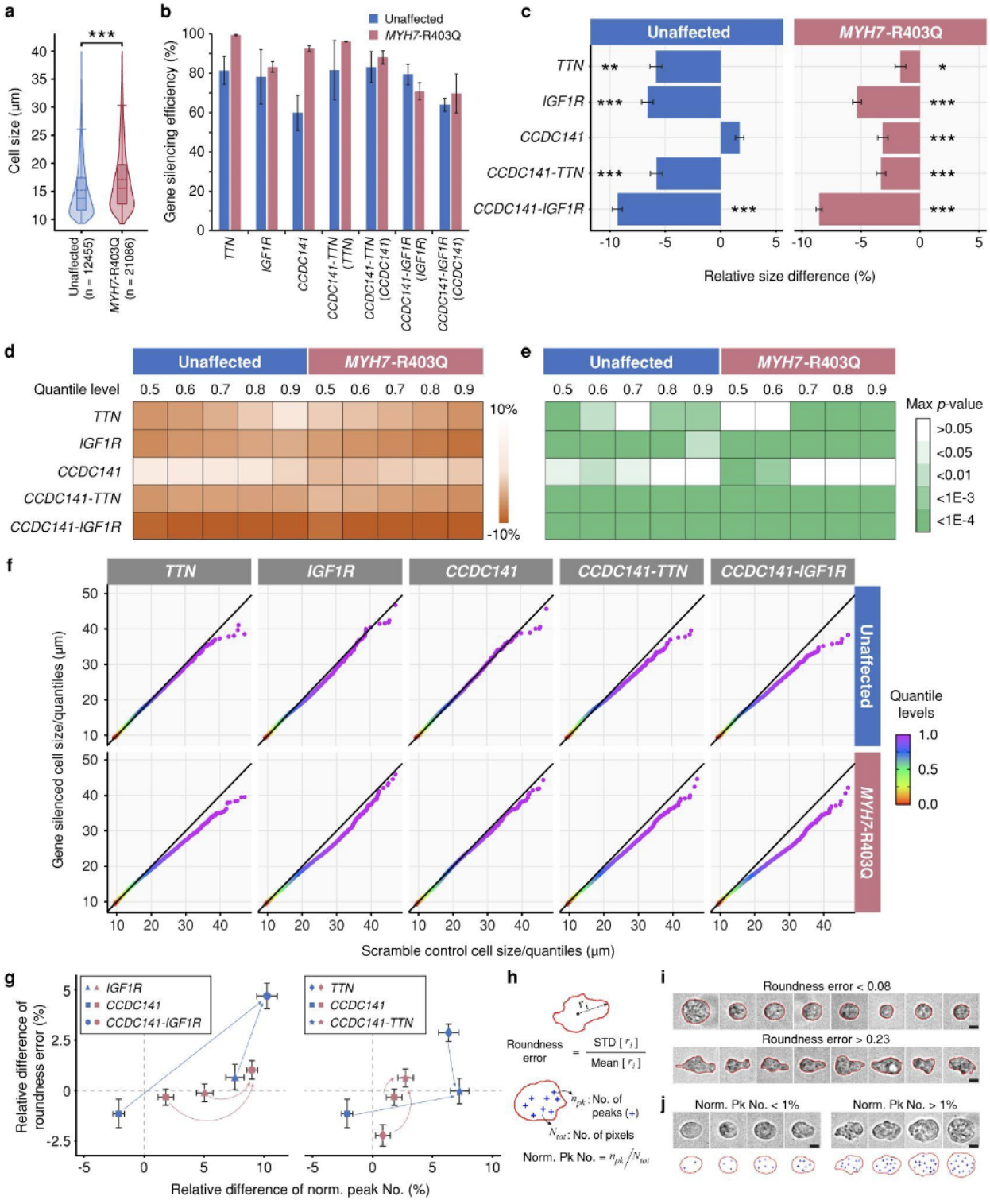

a
Cell size (μm)
***
Unaffected (n = 12455)
MYH7-R403Q (n = 21086)
b
Gene silencing efficiency (%)
Unaffected
MYH7-R403Q
TTN
IGF1R
CCDC141
CCDC141-TTN (TTN)
CCDC141-TTN (CCDC141)
CCDC141-IGF1R (IGF1R)
CCDC141-IGF1R (CCDC141)
c
Unaffected
MYH7-R403Q
TTN
IGF1R
CCDC141
CCDC141-TTN
CCDC141-IGF1R
Relative size difference (%)
d
Unaffected
MYH7-R403Q
Quantile level
0.5 0.6 0.7 0.8 0.9 0.5 0.6 0.7 0.8 0.9
TTN
IGF1R
CCDC141
CCDC141-TTN
CCDC141-IGF1R
10%
-10%
e
Unaffected
MYH7-R403Q
0.5 0.6 0.7 0.8 0.9 0.5 0.6 0.7 0.8 0.9
Max p-value
>0.05
<0.05
<0.01
<1E-3
<1E-4
f
TTN
IGF1R
CCDC141
CCDC141-TTN
CCDC141-IGF1R
Unaffected
MYH7-R403Q
Gene silenced cell size/quantiles (μm)
Scramble control cell size/quantiles (μm)
Quantile levels
1.0
0.5
0.0
g
IGF1R
CCDC141
CCDC141-IGF1R
TTN
CCDC141
CCDC141-TTN
Relative difference of roundness error (%)
Relative difference of norm. peak No. (%)
h
Roundness error = STD [r_i] / Mean [r_i]
n_{pk} : No. of peaks (+)
N_{tot} : No. of pixels
Norm. Pk No. = n_{pk}/N_{tot}
i
Roundness error < 0.08
Roundness error > 0.23
j
Norm. Pk No. < 1%
Norm. Pk No. > 1%

Figure 2.5 *(previous page)*: (a) Violin plots of cell diameters of unaffected (blue) and *MYH7*-R403Q variant (red) hiPSC-CMs indicate a heavily positive-skewed size distribution of cardiomyocytes. Solid and dashed lines in box plots represent median and mean values, respectively. Asterisks indicate significant difference ($^{***}p <$ 1E-36, Wilcoxon signed rank test). (b) Gene-silencing efficiency in unaffected (blue, $n = 5$ to 9) and *MYH7*-R403Q variant (red, $n = 3$) hiPSC-CMs based on RT-qPCR analysis (details in *Methods*). Error bars indicate standard deviations. (c) Fractional change in median cell diameter of gene-silenced hiPSC-CMs relative to scramble control values (relative size difference) indicates that *CCDC141* interacts with *IGF1R* to correct cardiomyocyte hypertrophy. Relative size differences were averaged across data from two to four independent batches of cells. Error bars indicate standard deviations computed on 1000 bootstrap samples of these batches with the following sample size: $n = 13147$ (*TTN*), 19460 (*IGF1R*), 45304 (*CCDC141*), 19979 (*CCDC141-TTN*), and 26135 (*CCDC141-IGF1R*) for unaffected cells and $n = 22134$ (*TTN*), 33801 (*IGF1R*), 21158 (*CCDC141*), 39515 (*CCDC141-TTN*), and 52049 (*CCDC141-IGF1R*) for *MYH7*-R403Q variant cells. Asterisks indicate significant difference between gene-silencing and scramble control conditions based on the maximum p-values of Wilcoxon signed rank test across all batches of cells ($^{*}p < 0.05$, $^{**}p < 0.001$, and $^{***}p < 10^{-4}$). (d) A heatmap of relative differences of cell size quantiles at various levels between gene-silencing and scramble control conditions. Dark color indicates strong reduction of size quantiles in gene-silenced cells thereby a narrower size distribution relative to the scramble control. The corresponding statistical differences (e) were evaluated by the maximum p-values across all batches of cells using a bootstrap quantile test (1000 sample size). (f) Representative QQ-plots of cell size quantiles comparing between gene-silenced cells and scramble controls for both unaffected (top) and *MYH7*-R403Q variant (bottom) hiPSC-CMs indicate a clear size-bias in the effect of silencing *CCDC141-IGF1R* on correcting cardiomyocyte hypertrophy. (g) *CCDC141* non-additively interacts with *IGF1R* (left) and *TTN* (right) to modify boundary and texture features of unaffected (blue) and *MYH7*-R403Q variant (red) hiPSC-CMs. Cell boundary waveness and texture irregularity were measured by the roundness error (h, top) and normalized peak number (h, bottom), respectively. (i) Representative single-cell images overlapped with detected cell boundaries (red lines) show that a higher roundness error indicates increased irregularity of the cell boundary. (j) Representative single-cell images with detected peaks (blue plus signs) of the brightfield intensity distribution enclosed within the cell boundaries (red lines) indicate a varying level of cell textural irregularity.

and disease cells ($p < 10^{-4}$ for non-additivity). In addition, *CCDC141* and *TTN* show distinctive independent roles in repressing cardiomyocyte hypertrophy. In healthy cells, monogenic silencing of *TTN* leads to a larger reduction in the median cell size compared to the case of silencing *CCDC141*. In contrast, *MYH7*-R403Q variant cells display a larger size reduction in response to monogenic silencing of *CCDC141*. Given the diverse expressivity of

cellular hypertrophy among *MYH7*-R403Q variant cells (indicated by the heavily positively-skewed size distribution in Figure 2.5a), we further analyzed size changes of large and small cells focused into different microchannels separately. Our results show that silencing *TTN* ($p = 0.001$) or *CCDC141-TTN* ($p = 0.005$) predominantly targets cells that exhibit cellular hypertrophy (gray bars in Figure A.6a, right) but has a trivial impact on small cells.

To explore this diverse expressivity further, we evaluated how gene silencing reshapes the cell size distribution (Figure 2.5d and p-values in e). QQ-plots showcasing the quantiles of gene-silenced cells against their scrambled controls suggest that digenic gene silencing leads to a stronger effect on larger cells (indicated by the larger deviation of points from the solid black line in Figure 2.5f). For example, the upper quantiles of *MYH7*-R403Q variant cells silencing *CCDC141-IGF1R* are much lower (the 0.9 quantile decreased by 14.4%, $p < 10^{-4}$) than the corresponding scrambled quantiles (Figure 2.5d&f), indicating that this variant favors rescuing hypertrophic cells over small cells. However, this size-dependent effect is mitigated in cells with monogenic silencing of *CCDC141* or *IGF1R*. This discrepancy in how these two genes affect cell size reinforces the hypothesized non-additive interaction between *CCDC141* and *IGF1R* (e.g., $p < 10^{-4}$ for nonadditivity, 0.9 quantile, Figure 2.5f). Importantly, stable non-additive interaction effects for hypertrophic cells (quantile levels higher than 0.6) are also observed for the *CCDC141-TTN* pair in both cell lines. Overall, these experimental results suggest both an epistatic interaction between *CCDC141* and the other two genes (*IGF1R* and *TTN*) as well as independent effects on modifying cardiomyocyte hypertrophy.

Additionally, recent studies have shown that other cellular morphology features, such as cell boundary and textural irregularities, are informative readouts of cytoskeletal structure, which is highly associated with disease state (Wu et al., 2020; Alizadeh et al., 2019). We analyzed relative changes in cell shape and texture (Figure 2.5g and Figure A.6b) by measuring the counts of peak intensities normalized to the total number of pixels enclosed by the cell boundary (Figure 2.5h). Cells with a high normalized peak number display a ruffled texture, which is manifested by an unevenly distributed 2D intensities (Figure 2.5j). Our analysis shows that silencing both *CCDC141* and *IGF1R* (circles in Figure 2.5g, left) yields a larger increase in intensity peak number than the case of silencing *IGF1R* alone (triangles in Figure 2.5g, left) for both cell lines. Specifically in both healthy and disease cells, the *CCDC141-IGF1R* interaction ($p < 10^{-4}$ for non-additivity) exhibits a synergistic epistasis between gene interactors on increasing intensity peak number. We also analyzed cell roundness error, a measure of how far radii measured on the cell outline deviate from the scenario of a circular shape (Figure 2.5h). This parameter increases with an increasing cell boundary waviness or elongation (Figure 2.5i). We show that the silencing of *CCDC141* and *IGF1R* synergistically interact to increase roundness error of HCM cardiomyocytes ($p < 10^{-4}$ for non-additivity, Figure 2.5g, left). In addition, *CCDC141* and *TTN* display antagonistic epistasis and synergistic epistasis in their impact on roundness error for healthy and *MYH7*-R403Q variant cells ($p < 10^{-4}$ for non-additivity, Figure 2.5g, right). Plotting cell roundness error against normalized peak number highlights the non-additive interaction in both gene pairs (*CCDC141-IGF1R* and *CCDC141-TTN*) on modulating cell textural and boundary irregularity (Figure 2.5g).

CCDC141 may interact with *IGF1R* and *TTN* through transcription factor-DNA binding

We explored potential mechanisms for identified epistatic gene interactions. Although the non-coding nature of the discovery SNVs implied it was unlikely protein-protein interaction was responsible for our findings, we first queried the BioGRID (version 4.4.213) (Oughtred et al., 2021) and IntAct (Hermjakob et al., 2004) databases and found no reported protein-protein interactions between *CCDC141*, *IGF1R*, and *TTN*.

We next investigated whether these three genes share transcription factor (TF) DNA binding. We scanned all SNVs that passed the lo-siRF GWAS filter using the Ensembl 2023 database (Martin et al., 2023). Within each of the three genes, we found SNVs that co-occur in regulatory regions harboring a shared set of transcription factor binding motifs (TFBMs) (Figure 2.6a-c). One TFBM binds to a complex formed between two TFs, TEAD4 and GATA3 (denoted as TEAD::GATA motif in Figure 2.6). TEAD4 plays a key role in heart morphogenesis by recruiting a variety of transcriptional co-activators (e.g., YAP, TAZ and VGLLs) that regulate cell proliferation (Currey et al., 2021). A recent study has shown that TFBMs for TEAD4 and other TEAD TFs are significantly enriched in enhancers upregulated in both fetal human hearts and heart tissues from patients with dilated cardiomyopathy (Spurrell et al., 2022), suggesting an important regulatory role of TEAD TFs in activating heart development and disease-specific pathways. In particular, we found that both a *CCDC141* SNP (rs62177296) and an *IGF1R* SNP (rs55686521) reside directly in this TEAD::GATA TFBM. These genetic variants alter the consensus DNA-binding sequences recognized by GATA (5'-GATAAG-3') and TEAD (5'-GGAATG-3') TFs, respectively (Figure 2.6d), suggesting their direct impact on DNA binding affinity (Yang et al., 2022).

In addition to these motif-disrupting variants, several SNPs are located close to (less than 420 bps from) two other TFBMs that are often found to be adjacent to the TEAD::GATA motif. One motif (denoted as GATA motif) is bound by GATA3/4/5 TFs that are critically involved in heart development and cardiac hypertrophy (Hong and Zhang, 2022). GATA4 regulates a spectrum of key cardiac genes, including *NPPA*, *NPPB*, *MYH7*, *MYH6*, *TNNC1*, *TNNI1*, and *MLC1/365*. The other motif (denoted as AP-1 motif) is bound by several important TFs that affect the regulation of cardiac hypertrophy and myocardial infarction, such as members of the AP-1 (e.g., FOS, JDP2, and JUN) and bZIP TF family (XBP1 and CREB3). Indeed, previous studies have shown that many TFs can interdependently bind to DNA, with their individual motifs either packed together (e.g., the TEAD::GATA motif) or proximally separated from each other (Deplancke et al., 2016) (e.g., the cluster of motifs in Figure 2.6a-c). For instance, the collaborative binding of AP-1 (FOS/JUN) and TEAD TFs to separated motifs has been reported in human tumor cells and mouse embryonic fibroblasts (Yang et al., 2022). In addition to the proximal collaborative TF-DNA binding, long-range binding of TF pairs can connect two distant genes in the process of DNA or chromatin looping (Deplancke et al., 2016). The aforementioned TFBM analysis thus suggests one hypothesis in which the demonstrated epistatic effects can be explained by the co-regulation of gene

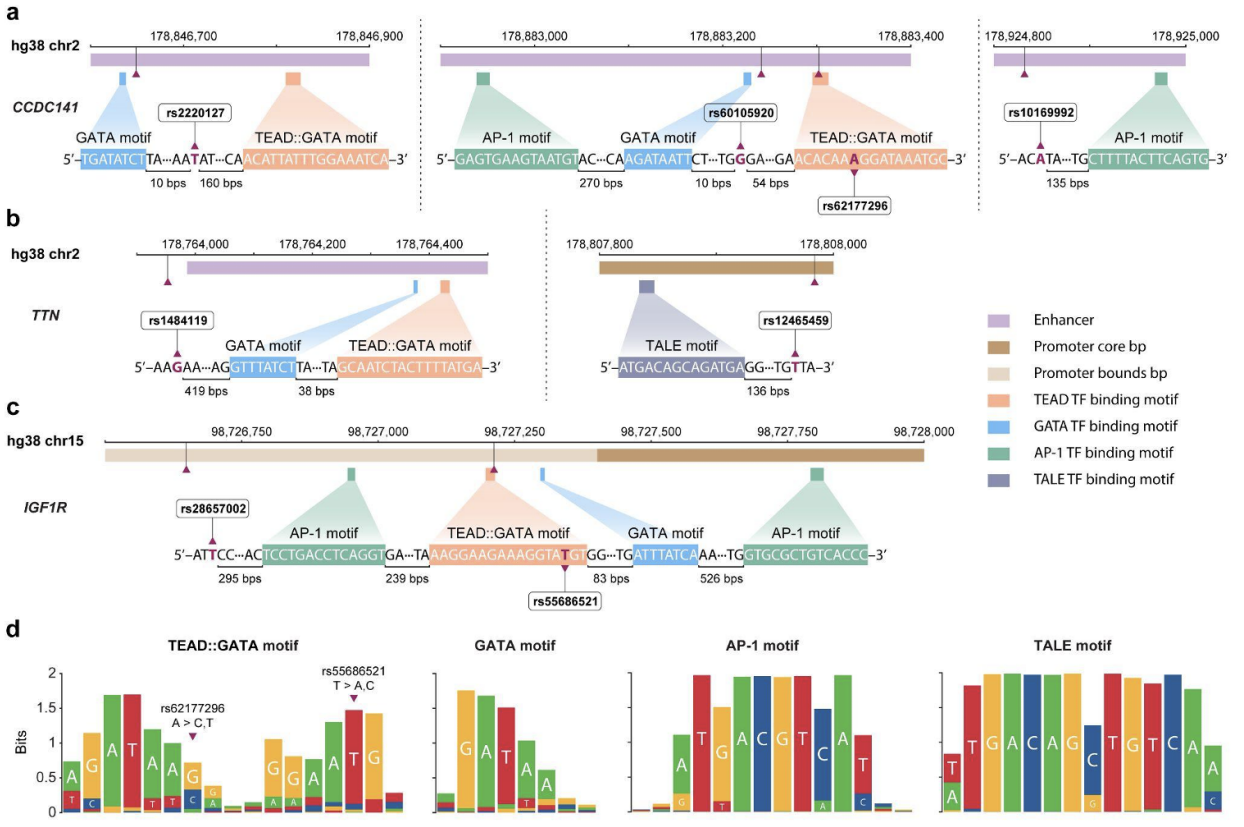


Figure 2.6: (a)-(c), Interacting SNPs identified in *CCDC141* (a), *TTN* (b), and *IGF1R* (c) co-occur in regulatory regions involving TEAD, GATA, and AP-1 transcription factor binding motifs (TFBMs), with their DNA sequences and locations shaded in different colors. The TEAD::GATA TFBM (light orange) is co-bound by TEAD4 and GATA3. The GATA TFBM (light blue) is bound by GATA3/4/5. The AP-1 TFBM (mint green) is bound by XBP1, CREB3, BATF3, JDP2, FOS, ATF7, and JUN. The TALE TFBM (purple gray) is bound by MEIS2, MEIS3, TGIF1, TGIF2, TGIF2LX, PKNOX1, and PKNOX2. d, Motif matrices for TEAD::GATA, GATA, AP-1, and TALE TFBMs. TFBMs were experimentally verified based on the Ensembl 2023 database (Martin et al., 2023).

expression through shared TFs essential for cardiac development (e.g., TEAD, GATA, and AP-1).

To further investigate this hypothesis, we verified the statistical significance of TF overlap between *CCDC141* and *TTN* using known TF enrichment data from Enrichr (Kuleshov et al., 2016). We performed all possible permutations of gene pairs from the GWAS-filtered SNPs and evaluated the degree of overlap in their associated TFs. We found that the TF overlap between *CCDC141* and *TTN* significantly exceeds the TF overlap between random gene pairs (empirical $p = 0.013$, Figure 2.7a-b). On the other hand, *CCDC141* and *IGF1R*

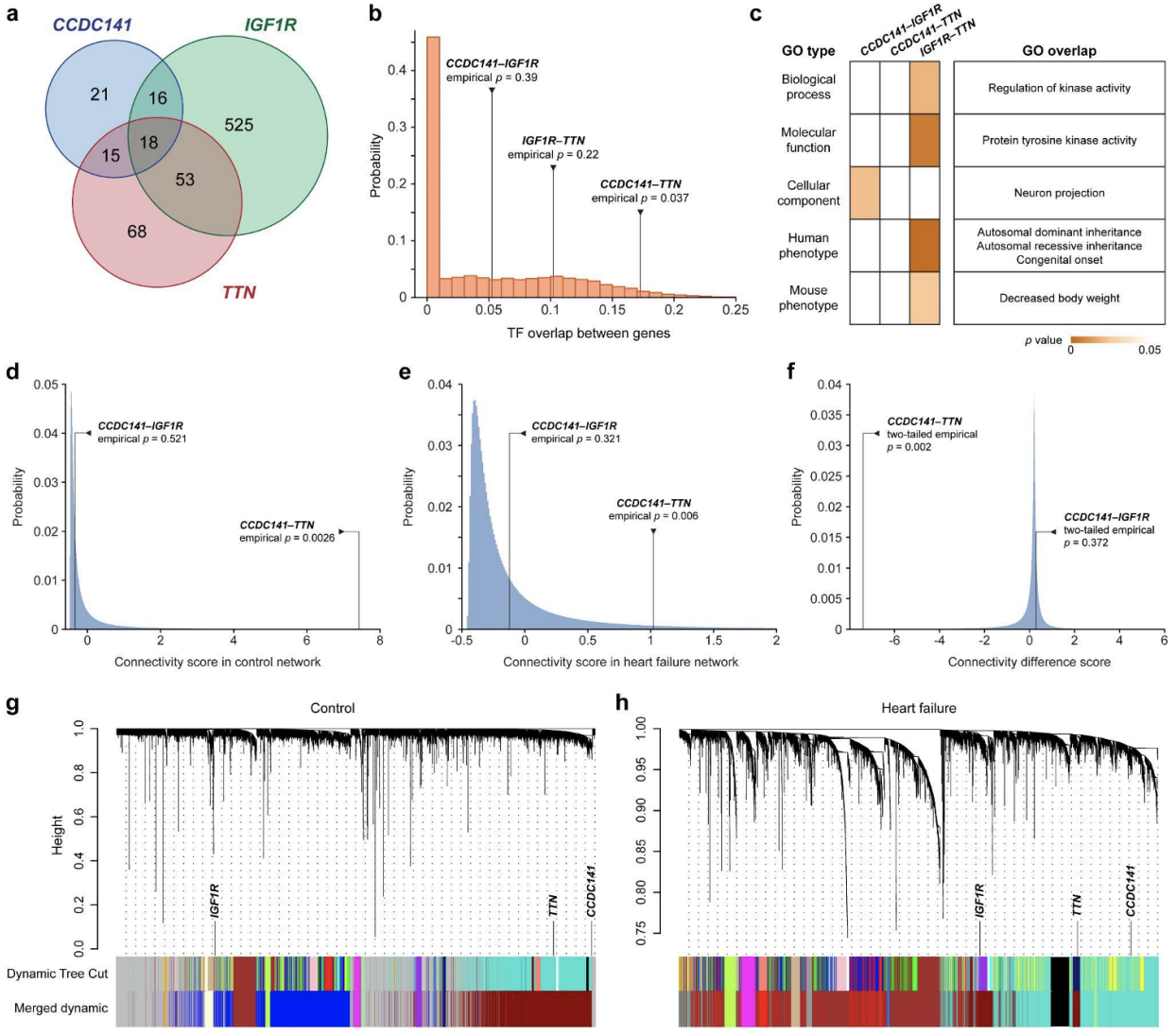

a
CCDC141
IGF1R
TTN
21
16
525
18
15
53
68
b
CCDC141–IGF1R
empirical p = 0.39
IGF1R–TTN
empirical p = 0.22
CCDC141–TTN
empirical p = 0.037
Probability
TF overlap between genes
c
GO type
CCDC141–IGF1R
CCDC141–TTN
IGF1R–TTN
GO overlap
Biological process
Molecular function
Cellular component
Human phenotype
Mouse phenotype
Regulation of kinase activity
Protein tyrosine kinase activity
Neuron projection
Autosomal dominant inheritance
Autosomal recessive inheritance
Congenital onset
Decreased body weight
p value
d
CCDC141–IGF1R
empirical p = 0.521
CCDC141–TTN
empirical p = 0.0026
Probability
Connectivity score in control network
e
CCDC141–IGF1R
empirical p = 0.321
CCDC141–TTN
empirical p = 0.006
Probability
Connectivity score in heart failure network
f
CCDC141–TTN
two-tailed empirical
p = 0.002
CCDC141–IGF1R
two-tailed empirical
p = 0.372
Probability
Connectivity difference score
g
Control
Height
IGF1R
TTN
CCDC141
Dynamic Tree Cut
Merged dynamic
h
Heart failure
IGF1R
TTN
CCDC141

Figure 2.7 *(previous page)*: (a) Number of transcription factors (TFs) that interact with *CCDC141*, *IGF1R*, *TTN*, or any of their combinations reported by Enrichr (Kuleshov et al., 2016). (b) Permutations of all possible gene pairs show that the TF overlap between *CCDC141* and *TTN* significantly exceeds the TF overlap between random gene pairs that pass the GWAS filter (Figure 2.2b). (c) Enrichr reported gene ontology (GO) overlap evaluated using all permutations of gene pairs passing the GWAS filter (Figure 2.2b). The heatmap of empirical p-values indicates significant GO overlaps for *CCDC141–IGF1R* and *IGF1R–TTN* interactions relative to random gene pairs. (d)-(f), Distributions of connectivity score between all possible combinations of genes for the control (d) and heart failure (e) networks and the connectivity difference between these two networks (f) established from Weighted Gene Co-expression Network Analysis (WGCNA) using failing and non-failing human heart tissues (Cordero et al., 2019) *CCDC141* and *TTN* show a significant connectivity in the control and heart failure networks and a significant difference in connectivity between these two networks. (g)-(f), Dendrograms from WGCNA control (g) and heart failure (h) networks.

did not show a significant TF overlap in our analysis (Figure 2.7b). This can be explained by the lack of associated TFs identified for *CCDC141* relative to that for *IGF1R* (Figure 2.7a). However, these two genes show a significant overlap in the gene ontology (GO) of neuron projection (empirical $p = 0.043$, Figure 2.7c), which indicates that these two genes are involved in the same downstream biological processes or pathways.

CCDC141 and *TTN* interact in gene co-expression networks from healthy and failing human hearts

We next explored how the demonstrated epistases contribute to the progression of a failing human heart from its normal state. We compared the difference in connectivity of all the possible gene pairs between two weighted gene co-expression networks established using 177 failing hearts and 136 non-failing hearts (Cordero et al., 2019), respectively. We found a strong connection between *CCDC141* and *TTN* in the healthy control network ($p = 0.0026$, Figure 2.7d). Even though the connectivity is statistically significant ($p = 0.006$, Figure 2.7e) in the heart failure (HF) network, it is substantially weakened ($p = 0.002$, two-tailed empirical for the difference in connectivity, Figure 2.7f). This weakening suggests a potential role of attenuating the *CCDC141-TTN* interaction in the process of heart failure. Comparing gene module memberships of *CCDC141*, *TTN*, and *IGF1R* between the control and HF networks shows that *CCDC141* and *TTN* are co-associated with the module of electron transport chain/metabolism in the control network while *IGF1R* is associated with the module of unfolded protein response (Figure 2.7g). In the HF network, *TTN* and *IGF1R* are co-associated with the module of muscle contraction/cardiac remodeling, whereas

CCDC141 remains associated with the metabolism module (Figure 2.7h). This indicates that the epistatic effect of *CCDC141* and *TTN* on the progression of heart failure may be related to the dependence of contractile function on metabolic efficiency. Although this analysis did not prioritize a direct pairwise interaction between *CCDC141* and *IGF1R*, they may interact through a third gene, such as *TTN*, to contribute to the progression of heart failure.

2.3 Discussion

We have demonstrated evidence for epistasis in the regulation of cardiac hypertrophy through data-driven veridical machine learning recommendations and gene-silencing experiments. Although polygenic and epistatic contributions to disease have been supported by computational models (Zeng et al., 2022; Li et al., 2020), experimental confirmation of these interactions is rare, with notable exceptions including work revealing the oligogenic inheritance of human heart disease (Gifford et al., 2019). Our strategy integrated a deep-learning LV structural analysis and an iRF-based model (lo-siRF), rooted in the PCS framework for veridical data science (or machine learning). This approach allowed us to identify epistatic interactions between *CCDC141* and other two genes (*IGF1R* and *TTN*) that drive cardiac hypertrophy, highlighting the applicability and efficacy of lo-siRF for reliable detection of epistatic drivers of human complex traits. In addition, we phenotypically validated the effect of disrupting putative epistatic gene-gene interactions on cellular morphology in healthy and HCM hiPSC-CMs. We show that high-throughput microfluidics employing imaged-based single cell morphology analysis can rapidly and precisely capture small changes in cellular morphology features, providing a useful approach to assess the phenotypic consequences of common genetic variants with small effect size at scale. This platform can be further integrated with high-dimensional single-cell morphology profiling (Wu et al., 2020) to reveal epistatic effects on hundreds of single-cell morphological features simultaneously that are invisible from single or a few readouts. In addition, a variety of high-throughput cell sorting applications are unlocked by integrating image-based flow cytometry (Schraivogel et al., 2022).

Our mechanistic exploration using bioinformatics databases (e.g., BioGRID (Oughtred et al., 2021), IntAct (Hermjakob et al., 2004), Ensembl (Martin et al., 2023), and Enrichr (Kuleshov et al., 2016)) leads to a hypothesis that *CCDC141*, *IGF1R*, and *TTN* interact through mediating transcriptional regulations rather than physical protein-protein interactions. Since the identified SNPs in our lo-siRF pipeline in each of the three genes co-occur in regulatory regions consisting of similar TF binding motifs, it is possible that variants in these genes interdependently affect the binding affinity of the respective TFs. In other words, regulatory variants in one gene may enhance or repress the expression of the other gene in the demonstrated epistases, and thereby exert non-additive effects on the downstream phenotypic consequences. This hypothesized transcriptional interdependency between genes can partly explain our findings that *CCDC141* synergistically interacts with *TTN* and *IGF1R* to reduce cardiomyocyte hypertrophy (Figure 2.5c). Another possibility is that these three

genes are co-regulated by the same sets of TFs and involved in similar biological processes or pathways. However, due to the complexity of TF-DNA binding and limited data of recognized motifs, further comprehensive analysis is needed to capture different types of genomic data, such as DNA methylation, chromatin states, and DNA structural conformation and its steric characteristics. In addition, *CCDC141* appeared in all the epistatic interactions prioritized by our model. Although *CCDC141* is known to be highly expressed in heart muscle and has been associated with heart rate (Den Hoed et al., 2013; Verweij et al., 2018) and QRS duration on electrocardiogram (Norland et al., 2019), few functional studies exist to elucidate its specific role in cardiac structures (Verweij et al., 2018). *CCDC141* has been hypothesized to act as a regulator of the myosin II pathway, by which the encoded protein interacts with the disrupted-in-schizophrenia 1 (DISC1)-centrosome complex (Fukuda et al., 2010) to control the cortical neuron migration. Therefore, it is possible that *CCDC141* plays a similar role in regulating the myosin II pathway in cardiomyocytes. This hypothesis agrees with our results of weighted gene co-expression network analysis (WGCNA) based upon more than 300 human heart tissues (Cordero et al., 2019). Our analysis suggested that *CCDC141*, *TTN*, and *IGF1R* are associated with gene modules of muscle contraction or metabolism, which are key components of the myosin II pathway.

In summary, this study expands the current scope of the genomic architecture of complex traits, highlighting the importance of epistasis on cardiac structure. The integration of lo-siRF, MRI-based LV structure analysis, and high-throughput microfluidic single-cell phenotyping provides a powerful approach for reliable epistasis discovery. This study also unlocks a number of future research directions, such as the identification of allele-specific binding sites in the demonstrated epistasis and the exploration of higher-order epistatic interactions among more than two genes. Our study also provides important insights for predicting polygenic disease risk as well as suggesting potential therapeutic targets in non-coding regions for diseases with complex traits.

2.4 Methods

Study participants

The UK Biobank is a biomedical database with detailed phenotypic and genetic data from over half a million UK individuals between ages 40 and 69 years at recruitment (Bycroft et al., 2018). In the LVH study, we restricted our analysis to 29,661 unrelated White British individuals (Table A.1) from the UK Biobank cohort with both genetic and cardiac MRI data. More specifically, we considered only those individuals from the UK Biobank cohort who self-reported as White British and have similar genotypic backgrounds based on principal components analysis as done in prior work (Bycroft et al., 2018). We also identified related individuals (i.e., third-degree relatives or closer) via genotyping and omitted all but one individual from each related group in the analysis. Details regarding this cohort refinement have been described and implemented previously (Bycroft et al., 2018; Morgan et al.,

2018). This resulted in a cohort of 337,535 unrelated White British individuals from the UK Biobank, of which 29,661 have both genetic and cardiac MRI data. We randomly split this data into training, validation, and test sets of size 15,000, 5,000, and 9,661 individuals, respectively.

Genotyping and quality control

For the study cohort of 30,000 individuals described above, we leveraged genotype data from approximately 15 million imputed autosomal SNPs. These have been imputed from 800,000 directly assayed SNPs (obtained by the UK Biobank from one of two similar Affymetrix arrays) using the Haplotype Reference Consortium and UK10K reference panels (Bycroft et al., 2018). Imputed variants were subject to several quality-control filters, including outlier-based filtration on effects due to batch, plate, sex, array, and discordance across control replicates. Further, we excluded variants due to extreme heterozygosity, missingness, minor allele frequency ($< 10^{-4}$), Hardy-Weinburg equilibrium ($< 10^{-10}$), and poor imputation quality (< 0.9). Further details can be found in previous studies (Bycroft et al., 2018; Morgan et al., 2018).

Quantification of left ventricular hypertrophy

We retrieved 44,503 cardiac MRI studies from the UK Biobank taken during imaging visits and first follow-up imaging visits. Using a previously described segmentation algorithm (Bai et al., 2018), we determined areas of the LV cavity and myocardium from each short axis frame, from which we computed the volume of the LV myocardium at the end diastole. After quality control, 44,220 segmentations remained. This volume was then converted to a mass (LVM) using a standard density estimate of 1.05 g/mL (Grothues et al., 2002). The LVMi was computed by dividing the LVM by an estimate of body surface area based on height and body weight calculated using the Du Bois formula (Du Bois and Du Bois, 1916). From the 44,220 segmentations, we refined the analysis to LVMi measurements for 29,661 White British unrelated individuals using the measurements from their most recent imaging visit if multiple imaging visits were recorded.

Low-signal iterative random forests (lo-siRF)

Step 1. Dimension reduction of SNPs via GWAS

As the first step in the lo-siRF pipeline, we performed a GWAS on the training data for the rank-based inverse normal-transformed LVMi using two algorithms, PLINK (Chang et al., 2015) and BOLT-LMM (Loh et al., 2015), and filtered the number of features by p-value ranking in a domain-inspired dimension reduction step (Figure 2.2b). This step is akin to typical preprocessing phases in fine-mapping (Schaid et al., 2018). Since BOLT-LMM and PLINK rely on different statistical models, we employ both software packages to mitigate

the dependence of downstream conclusions on this arbitrary choice. Specifically, for the first GWAS run, we fitted a linear regression model, implemented via `glm` in PLINK (Chang et al., 2015). For the second GWAS run, we used BOLT-LMM (Loh et al., 2015), a fast Bayesian-based linear mixed model method, to assess the association between each SNP and the LVMi phenotype. Each GWAS was adjusted for the first five principal components of ancestry, gender, age, weight, and height. After ranking the SNPs by significance (i.e., the GWAS p-value) for each GWAS run separately, we took the union of the top 1000 SNPs (without clumping) from each of the two GWAS runs and proceeded with this resulting set of 1405 SNPs for the remainder of the lo-siRF pipeline. We chose to use the top 1000 SNPs per GWAS method (without clumping) as it (1) struck a balance between the amount of information loss and the computational cost of downstream modeling and (2) yielded the highest validation prediction accuracy compared to choosing other possible thresholds (500 and 2000 SNPs per GWAS) with and without clumping. We note that these 1405 SNPs strictly contain the SNPs that passed the genome-wide GWAS significance threshold ($p = 10^{-8}$).

Step 2. Binarization of the LVMi phenotype

Next, we binarized the raw (continuous) LVMi phenotype into a low and a high LVMi group before fitting siRF (Figure 2.2c). That is, for a given threshold x, we took the individuals in the top and bottom $x\%$ of LVMi values into two classes with the high and low LVMi values, respectively, while omitting the individuals in the middle quantile range. Due to the sex-specific biological variation of LVMi (see PCS documentation), we performed the binarization for males and females separately. For males, low and high LVMi was considered under 43.8-46.0 g/m^2 and above 55.4-58.5 g/m^2, respectively, depending on the choice of binarization threshold. For females, low and high LVMi was defined as under 35.1-36.8 g/m^2 and above 43.8-46.1 g/m^2, respectively (Table A.2). This binarization simplified the original low-signal regression problem into a relatively easier classification task: to predict whether an individual has a very high or very low LVMi, in essence, denoising the LVMi signal in the data. Moreover, as we will see in the next section, this binarization helped us more readily interpret and assess the performance of the prediction method with respect to the prediction screening step of the PCS framework (Yu and Kumbier, 2020). Since the specific threshold choice is arbitrary, we ran the remainder of the lo-siRF pipeline using three different binarization thresholds (15%, 20%, 25%) that balance the amount of denoising and data lost. In the end, we consolidated the results that were stable across all three binarization thresholds, described in *4.3. Gene and gene-gene interaction rankings.*

Step 3. Prediction

3.1. Fitting siRF on the binarized LVMi phenotype. For each binarization threshold, we trained the signed version of iRF (siRF) (Kumbier et al., 2018) using the SNPs that passed the GWAS filter to predict the binarized LVMi phenotype and generate candidate

interactions for further investigation (Figure 2.2c). Briefly, siRF iteratively re-weights features in a random forest to stabilize decision paths with an added outer loop of bootstrap and extracts nonlinear higher-order interactions based on commonly co-occurring features on a decision path, provided that the siRF model gives good prediction performance. In addition to giving a computationally-tractable interaction search engine, part of the attraction of siRF for detecting gene-gene epistatic interactions is the resemblance between the thresholding behavior of its decision trees and the thresholding (or switch-like) behavior frequently observed in biomolecular interactions (Nelson et al., 2008). In siRF, we not only keep track of which features are on the decision path, but also the direction of the feature split, i.e., whether low values (denoted $feature-$) or high values (denoted $feature+$) of the feature, appear on the decision path. Thus, a signed interaction consists of two or more signed features appearing on the same decision path. Note when applying siRF to SNP data, $SNP+$ typically represents a heterozygous or homozygous mutation and $SNP-$ typically represents no mutation at the locus. The following hyperparameters were used to train siRF using the iRF2.0 R package: number of iterations = 3, number of trees = 500, number of bootstrap replicates = 50, depth of random intersection tree (RIT) = 3, number of RIT = 500, number of children in RIT = 5, and minimum node size in RIT = 1. To fit siRF, we used 10,000 training samples (randomly sampled out of the 15,000 total training samples) and reserved the remaining 5,000 training samples for selecting genes and gene-gene interactions for the permutation test (see *4.2. Permutation test for difference in local stability importance scores*).

3.2. Prediction check. Per the PCS framework for veridical data science (or machine learning) (Yu and Kumbier, 2020), we next checked the validation prediction accuracy of siRF, which serves as one indication of whether the learnt model is capturing some phenotypic signal, relevant to reality, rather than simply noise. We observed that the predictive power of siRF for the binarized LVMi, though weak ($\sim 55\%$ balanced classification accuracy), was greater or on par with other commonly used prediction methods (i.e., LASSO, ridge, RF, and support vector machines) across a variety of metrics including classification accuracy, area under the receiver operator curve (AUROC), and area under the precision-recall curve (AUPRC) (Table A.3). Given that siRF performed better than random guessing (i.e., $> 50\%$ balanced classification accuracy, which is not guaranteed given the high phenotypic diversity and complexity of the LVMi trait) and yielded the best prediction power compared to these alternative, common prediction methods (except for the 15% binarization threshold where it performed on par with respect to classification accuracy), we deemed that the siRF fit for LVMi passed the prediction screening step of the PCS framework and hence proceeded to interpret this siRF model to extract candidate interactions. We further note that this prediction check played a key role in our choice of phenotypic data. Prior to studying LVMi, we attempted to run a similar analysis to predict HCM diagnosis, defined as any ICD10 billing code diagnosis of I42.1 or I42.2 in the UK Biobank data. However, neither siRF nor the other aforementioned prediction methods passed the 50% balanced classification accuracy requirement for predicting HCM diagnosis. We thus chose not to proceed with the HCM

analysis given the poor prediction accuracy and uncertain connection between the prediction models and the underlying biological processes. This failed prediction check demonstrated a need for a more refined phenotypic measure of cardiac hypertrophy, which ultimately led to the automated extraction of cardiac MRI-derived LVMi, our final choice of phenotype in this work. Further discussion of the HCM analysis can be found in the PCS Documentation.

Step 4. Interpretation

4.1. Local stability importance score. To interpret the siRF fit for LVMi, we developed a new local (i.e., on a per-individual basis) stability importance score that aggregates weak, unstable SNP-level importances into stronger, more stable gene-level importances (Figure A.3). This gene-level aggregation, which takes advantage of local feature importance scores to account for genes of varying lengths, is necessary since the individual SNP-level importances are incredibly unstable given the weak phenotypic signal and high correlation between SNPs (see PCS documentation for additional discussion). Specifically, given the fitted siRF, we used Annovar (Wang et al., 2010) to map each SNP in the fitted forest to its corresponding gene region according to the hg19 refGene annotations (i.e., given by the `Gene.refGene` column in the Annovar output). Note that using these annotations, each SNP gets mapped to exactly one gene region and that this region may correspond to an intergenic region between two genes. Then, given this forest T, for each individual i and signed gene or gene-gene interaction G, we computed a quantity named the local stability importance score, $LSI_T(G, i)$, which measures how important the signed gene or gene-gene interaction G is for making the prediction for individual i. Put concretely, the local stability importance score, $LSI_T(G, i)$, is defined as $D_T(G, i)$ / (# trees in the forest T), where $D_T(G, i)$ is the number of decision paths in the forest T such that individual i appears in its terminal node and at least one signed SNP from each signed gene $g \in G$ was used in a decision split along the path. In other words, $LSI_T(G, i)$ is the proportion of trees in the forest for which at least one SNP from each signed gene in the signed gene or gene-gene interaction was used in making the prediction for individual i. Consequently, a higher score indicates greater importance of the signed gene or gene-gene interaction for individual i's prediction (Figure A.3).

4.2. Permutation test for difference in local stability importance scores. Once we obtained the local stability importance scores for each individual, we performed a two-sample permutation test to assess whether the local stability importance scores (i.e. the importance) for a given signed gene or gene-gene interaction, G, are different between individuals with high and low LVMi (conditioned on the rest of the fitted forest). More formally, the proposed permutation test tests the null hypothesis $L = H$ versus the alternative hypothesis $L \neq H$, where L and H are the distributions of local stability importance scores for individuals with low and high LVMi, respectively. If the local stability importance scores are indeed different between high and low LVMi individuals (which corresponds to a small p-value from the proposed permutation test), then this suggests that G is able to differentiate between individuals with high versus low LVMi and hence may be an important gene or gene-gene

interaction for LVMi. We performed this permutation test using the 5,000 validation data (which were not used in training siRF), the difference in means as the test statistic, and 10,000 permutations. To reduce the large multiple testing and computational burden, we tested only the top 25 genes, ranked by their average local stability importance scores across 5,000 samples. These 5,000 samples were previously set-aside from within the 15,000 training samples and were not used in fitting the siRF (see *3.1. Fitting siRF on the binarized LVMi phenotype*). Furthermore, we tested all signed gene-gene interactions picked up by siRF that were "stable" across 50 bootstrap replicates. Here, a "stable" interaction is defined as one that passed the following siRF stability metric thresholds: stability score > 0.5, stability score for mean increase in precision > 0, and stability score for independence of feature selection > 0 (Table A.4) (Basu et al., 2018; Kumbier et al., 2018). Details on the siRF interaction and stability metrics can be found in previous work (Kumbier et al., 2018).

4.3. Gene and gene-gene interaction rankings. Finally, to obtain the top lo-siRF recommendations for follow-up experiments, we considered only those signed genes and gene-gene interactions that passed the aforementioned siRF stability filters (see *4.2. Permutation test for difference in local stability importance scores*) and underwent the permutation test in all three binarization runs. For these signed genes and gene-gene interactions that were stably identified by siRF and yielded a p-value < 0.1 in all three binarization runs, we finally ranked them by their mean permutation p-value, averaged across the three binarization thresholds (Table A.5). Because of our emphasis on recommending candidates for experimental validation, if both the $+$ and $-$ version of the signed gene appear, the final gene recommendations (Figure 2.3a) are ranked according to the smaller one of the two p-values. We note that though the signed information is not pertinent to our goal of recommending candidates for experiments, the signed information from siRF provides more granular information that can improve our interpretation of the fit, and we discuss this further in the PCS documentation.

lo-siRF PCS documentation and additional stability analyses

In an effort to facilitate transparency of human judgment calls that are inevitably made here and throughout our veridical machine learning analysis, we provided extensive documentation of code and more importantly, justification of these human judgment calls in the supplementary PCS documentation. We also performed additional stability analyses in accordance with the PCS framework (Yu and Kumbier, 2020) to ensure that our findings are robust to these human judgment calls (e.g., the choice of GWAS method, binarization threshold) and to bolster the reproducibility of our findings (see the supplementary PCS documentation for details).

hiPSC-cardiomyocyte differentiation

The studied patient-specific hiPSCs were derived from a 45 yr old female proband with a heterozygous *MYH7*-R403Q mutation. Derivation and maintenance of hiPSC lines was performed following Dainis et al. (2020). Briefly, hiPSCs were maintained in MTeSR (StemCell Technologies) and split at a low density (1:12) onto fresh 1:200 matrigel-coated 12 well plates. Following the split, cells were left in MTeSR media supplemented with 1 µM Thiazovivin. hiPSCs were maintained in MTeSR until cells reached 90% confluency, which began Day 0 of the iPSC-CM differentiation protocol. Cardiomyocytes were differentiated from hiPSCs using small molecule inhibitors. For Days 0-5, cells were given RPMI 1640 medium + L-glutamine and B27 - insulin. On Days 0 and 1, the media was supplemented with 6 µM of the GSK3β inhibitor, CHIR99021. On Days 2 and 3, the media was supplemented with 5 µM of the Wnt inhibitor, IWR-1. Media was switched to RPMI 1640 medium + L-glutamine and B27 + insulin on Days 6-8. On Days 9-12, cells were maintained in RPMI 1640 medium + L-glutamine - glucose, B27 + insulin, and sodium lactate. On Day 13, cells were detached using Accutase for 7-10 minutes at 37°C and resuspended in neutralizing RPMI 1640 medium + L-glutamine and B27 + insulin. This mixture was centrifuged for 5 minutes at 1000 rpm (103 rcf). The cell pellet was resuspended in 1 µM thiazovivin supplemented RPMI 1640 medium + L-glutamine and B27 + insulin. For the rest of the protocol (Days 14-40), cells were exposed to RPMI 1640 medium + L-glutamine - glucose, B27 + insulin, and sodium lactate. Media changes occurred every other day on Days 14-19 and every three days for Days 20-40. On Day 40, cardiomyocytes reached maturity.

hiPSC-cardiomyocyte siRNA gene silencing

Mature hiPSC-derived cardiomyocytes were transfected with Silencer Select siRNAs (Thermofisher) using TransIT-TKO Transfection reagent (Mirus Bio). Cells were incubated for 48 hours with 75 nM siRNA treatments. Four wells of cells were transfected with each of the six siRNAs: scramble, *CCDC141* (ID s49797), *IGF1R* (ID s223918), *TTN* (ID s14484), *CCDC141* and *IGF1R*, and *CCDC141* and *TTN*. After 2 days, hiPSC-CMs were collected for RNA extraction.

RT-qPCR analysis for siRNA gene silencing efficiency

Following cell morphology measurement, all cells for each condition were centrifuged for 5 minutes at 1000 rpm (103 rcf). Cell pellets were frozen at -80°C prior to RNA extraction. RNA was extracted using Trizol reagent for RT-qPCR to confirm gene knockdown occurred. Reverse Transcription of RNA was done using High-Capacity cDNA Reverse Transcription Kit (Thermofisher). qPCR of the single stranded cDNA was performed using TaqMan Fast Advanced MM (Thermofisher) with the following annealing temperatures: 95°C 20" and 40 cycles of 95°C 1" and 60°C 20". qPCR of the silenced genes was performed using TaqMan® Gene Expression Assays, including *CCDC141* (Hs00892642_m1), *IGF1R*

(Hs00609566_m1), and *TTN* (Hs00399225_m1). For gene silencing efficiency analysis, the gene *RPLP0* (Hs00420895_gH) was used as a reference gene. Data were analyzed using the delta-delta Ct method.

Cell sample preparation for cell morphology measurement

Following siRNA treatments, cells were detached for microfluidic single cell imaging using a mixture of 5 parts Accutase and 1 part TrypLE for 6 minutes at 37°C. Cells were then added to the neutralizing RPMI 1640 medium + L-glutamine and B27 + insulin. These mixtures were centrifuged for 5 minutes at 1000 rpm (103 rcf). For each gene silencing condition, the four wells of cells were resuspended in 4 mL of the MEM medium, which is composed of MEM (HBSS balanced) medium, 10% FBS, and 1% Pen Strep (Gibco). Cells were filtered with 100 µm strainers (Corning) before adding into the microfluidic devices.

Microfluidic inertial focusing device

We developed a new spiral inertial microfluidics system on the basis of the study by Guan et al. (2013) to focus randomly suspended cells into separate single streams based on cell size for high-resolution and high-throughput single cell imaging. The microfluidic device (Figure A.5) contains 5 loops of spiral microchannel with a radius increasing from 3.3 mm to 7.05 mm. The microchannel has a cross-section with a slanted ceiling, resulting in 80 µm and 150 µm depths at the inner and outer side of the channel, respectively. The channel width is fixed to 600 µm. The 495 µm wide slanted region of the channel ceiling is composed of ten 7 µm deep stairs. This particular geometry induces strong Dean vortices in the outer half of the channel cross-section, leading to high sensitivity of size separation and cell focusing. The device has two inlets at the spiral center to introduce cell suspensions and sheath flow of fresh medium. At the outlet region, the channel is expanded in width and split into two outlet channels with a width of 845 µm for the top outlet and 690 µm for the bottom outlet. Depths of the two outlet channels are designed to create equal hydraulic resistance. The top and bottom outlet channels are connected to 80 µm and 50 µm deep straight observation channels for high-throughput cell imaging.

Microdevice fabrication

The spiral microchannel was fabricated by CNC micromachining a piece of laser-cut poly (methyl methacrylate) (PMMA) sheet, which was bonded with a PMMA chip machined only with the inlet channels and another blank PMMA chip using a solvent-assisted thermal binding process to form the enclosed channel (Wang et al., 2019). Before bonding, PMMA chips were cleaned with acetone, methanol, isopropanol and deionized water in sequence. Droplets of a solvent mixture (47.5% DMSO, 47.5% water, 5% methanol) were evenly spread over the cleaned chips. The PMMA chips were assembled appropriately and clamped using a customized aluminum fixture, and then heated in a ThermoScientific Lindberg Blue M

oven at 96°C for 2 hrs. After bonding, fluid reservoirs (McMaster) were then attached to the chips using a two-part epoxy (McMaster). Microchannels were flushed with 70% ethanol followed by DI water for sterilization.

High-throughput single cell imaging

Before each experiment, microchannels were flushed with 3 mL of the MEM medium. Prepared cell samples and fresh MEM medium were loaded into 3 mL syringes, which were connected to the corresponding microchannel inlets using Tygon PVC tubing (McMaster). Both cells and the fresh MEM medium were infused into the microchannel using a Pico Plus Elite syringe pump (Harvard Apparatus) at 1.2 mL/min. Microscope image sequences of cells focused to the top and bottom observation channels were captured using a VEO 710S high-speed camera (Phantom) with a sampling rate of 700 fps and a 5 µsec light exposure.

Image analysis for cell feature extraction

For each gene silencing condition of each biological repeat, 21,000 images were processed to extract cell morphology features. To analyze cell size and shape changes induced by gene silencing, we developed a MATLAB-based image analysis pipeline, which includes three major steps: image preparation, feature extraction, and image postprocessing (Figure 2.4c). In step one, image sequences were fed into our program and subtracted from the corresponding background image to correct any inhomogeneous illumination. The program automatically generates background images, in which each pixel value is computed as the mode pixel intensity value among the same pixel of the entire corresponding image sequence. After illumination correction, step two detects cell edges by looking for the local maxima of the bright field intensity gradient, following which the program closes edges gaps, removes cells connected to the image borders, cleans small features (noise), and then fills holes to generate binary images and centroid positions for each single cell. Cell locations were then traced and stuck cells were removed by the double-counts filter if in presence. The double-counts filter excludes measurements collected around the same location with similar cell sizes using a Gaussian kernel density method (bandwidth = 0.09) when the estimated density for a certain location and size exceeds a particular threshold. The maximal density value for experimental runs where no repeated measurements were observed was used as the threshold. This procedure was validated using visual inspection of the removed cells. Binary images passing the double-counts filter were used to create coordinates (X, Y) of cell outlines, which leads to a range of cell size and shape parameters, including cell diameter and area, solidity, roundness error, circularity, and intensity spatial relationship enclosed within the cell. Cell area was computed as the 2D integration of the cell outline, and the cell diameter was computed as $2\sqrt{\text{Area}/\pi}$. Solidity is the ratio of cell area to the area of the smallest convex polygon that can contain the cell region. Roundness error was computed as the ratio between the standard deviation and mean of radii on the cell outline measured from the centroid. Circularity was calculated as $4 \cdot \text{Area} \cdot \pi/\text{Perimeter}^2$. The 2D intensity distributions within cell outlines

were used to derive peak locations and count peak numbers, which is a measure of intensity spatial relationship and a gating parameter to remove clumped cells. In the postprocessing step, data were cleaned using three filters with the following gating threshold. To remove large clumps, the peak-solidity filter removes data fall out of the polygonal region defined by (0.9, 0), (0.9, 3.2), (0.934, 8.26), (1, 28), (1, 0) in the (solidity, peak No.) space. Then, the roundness filter removes cells with weird shapes by excluding data with a roundness error higher than 0.3 or a circularity lower than 0.6. Finally, the small size filter removes cell debris whose major diameter is lower than 15 µm (12 µm) or minor diameter is lower than 12 µm (10 µm) for images photographed at the top (bottom) outlet microchannels.

Post-iRF analysis on co-associated transcription factors and gene ontologies between gene pairs

To explore whether the iRF-identified SNPs have any regulatory effects, we searched the Ensembl 2023 Regulatory Build database (Martin et al., 2023), and aligned the iRF-identified loci with known regulatory regions and motif features (Figure 2.6), downloaded from `https://ftp.ensembl.org/pub/release-109/regulation/homo_sapiens/`. This database provides increased strictness of motif feature annotation and only displays motif features that overlap a ChIP-seq peak in at least one cell type. This enhances the confidence that the used motifs have biological significance. We next evaluated the degree of overlap in co-associated transcription factors (TFs) and gene ontology (GO) terms between the demonstrated genes in epistasis. To this end, we used Enrichr (Kuleshov et al., 2016) to search for the enriched TFs and GOs for all the genes prioritized by GWAS (see *Step 1. Dimension reduction of SNPs via GWAS* in the *Low-signal iterative random forests (lo-siRF)* methods section). For each possible combination of gene pair (gene A and gene B), we measured the degree of overlap in shared TFs or GOs by $R_{\text{overlap}} = N_{(A \cap B)}/N_{(A \cup B)}$. Here, $N_{(A \cap B)}$ is the number of TFs or GOs shared between gene A and gene B, and $N_{(A \cup B)}$ is the number of TFs or GOs associated with either gene A or gene B. For cases where $N_{(A \cup B)} = 0$, we define $R_{\text{overlap}} = 0$ to indicate that no common TFs or GOs were found between the respective gene-gene combinations. All possible permutations of gene pairs were performed to generate the distributions of R_{overlap} for TFs (Figure 2.7b) and GOs in different categories (Figure 2.7c). Empirical p-values represent the number of random gene pairs with an R_{overlap} higher than the value for the respective epistatic gene-gene interaction.

Disease-state-specific gene co-expression network analysis

In order to evaluate the connectivity between genes and their potential roles in the transition from healthy to failing myocardium, we compared gene-gene connectivity and changes in the topological structure between the gene co-expression networks for healthy and failing human heart tissues. To construct the gene co-expression networks, cardiac tissue samples from 177 failing hearts and 136 donor, non-failing (control) hearts were collected from operating rooms and remote locations for RNA expression measurements. We performed weighted

gene co-expression network analysis (WGCNA) on the covariate-corrected RNA microarray data for the control and heart failure networks separately (Figure 2.7g-h). Detailed steps for generating these co-expression networks, which included calculating the correlation matrix, TOM transformation, and Dynamic Tree Cut module finding, are described in our previous study (Cordero et al., 2019), and data for these networks is available at `https://doi.org/10.5281/zenodo.2600420`. To evaluate the degree of connectivity between interacting genes in each of the networks, we compared the edge weights between the interacting genes demonstrated in this study relative to the distribution of all possible pairwise combinations of genes (Figure 2.7d-e). We also evaluated the difference of edge weights (Z-score normalized) between the control and heart failure networks to understand how these gene-gene interactions change between non-failing and failing hearts (Figure 2.7f). The two-tailed empirical p-value represents the proportion of the absolute difference in edge weights of all gene pairs that exceed the absolute difference score for the gene pairs of interest. We then compared the structure of modules derived from dendrograms on the WGCNA control and heart failure networks (Figure 2.7g-h). Modules were labeled according to Reactome Enrichment analysis of genes within each module. The full gene module descriptions and Benjamini-Hochberg-adjusted enrichment p-values can be found in the Supplementary Data 5 and 6 in the study by Cordero et al. (2019).

Statistical analysis of experimental results

All statistical analysis involved in the lo-siRF pipeline has been elucidated in the corresponding lo-siRF Methods sections. For experimental results, we compared differences in cell size/shape parameters between cells with silenced genes or gene pairs and their scramble controls using various summary statistics. To study the centers of the distributions, we compared the medians using a traditional Wilcoxon signed rank test (Figure 2.5c) and a bootstrap quantile test at the 0.5 quantile level. We also performed a bootstrap-t for trimmed means of cell size distributions and evaluated the stability of results by varying the amount of trimming (ranging from 0-0.3) to ensure that outliers (e.g., from large clumps of cells) do not drive the conclusions (data not shown). In addition to comparisons of central behavior, we performed the bootstrap quantile test across increasing quantile levels (ranging from 0.5-0.9) to study size changes (Figure 2.5d-e) that favor hypertrophic cells, which is more relevant to the pathologic phenotype of cardiac hypertrophy. All tests are performed on each experimental batch, and the maximum p-value across batches are reported in the main text. To formally quantify whether interactions are non-additive, for each interaction (i.e.,*CCDC141-TTN* and *CCDC141-IGF1R*), we tested whether the interaction effect is significant when regressing cell size on the interaction and main effects for each relevant gene. In accordance with the PCS framework, we performed a stability analysis by comparing results across different regression parameters — namely, quantile regression with increasing quantile levels (ranging from 0.5-0.8). In each regression, percentile bootstrap t-tests were used to compute p-values. Traditional t-test p-values support these results and are provided in Appendix A. Since these regressions require comparisons between gene silencing

experiments (e.g., silencing *CCDC141* and *TTN* vs only silencing *CCDC141*), and the gene silencing efficiency varied across experiments, batches were merged based on those with the highest efficiency for each regression.

Data and code availability

All genotype and cardiac MRI data used as input to the lo-siRF pipeline are available from the UK Biobank (`https://www.ukbiobank.ac.uk/`). This work was conducted under the UK Biobank application 22282. The lo-siRF pipeline was performed using R version 3.6.1 and iRF2.0 (`https://github.com/karlkumbier/iRF2.0`) with code available via GitHub (`https://github.com/Yu-Group/epistasis-cardiac-hypertrophy`). The LVMi derivation from cardiac MRI images and corresponding code have been published elsewere (`https://github.com/baiwenjia/ukbb_cardiac`) (Bai et al., 2018). ANNOVAR (`https://annovar.openbioinformatics.org/en/latest/`) was used to map each SNP to its associated gene region. Single cell image analysis was performed using MATLAB version R2019a. This code was used to extract single cell size/shape information from image sequences of flowing cells (Figure 2.4c). Statistical analysis of experimental results was performed using R version 3.6.1 with code available via GitHub (`https://github.com/Yu-Group/epistasis-cardiac-hypertrophy`).

Contributions

T.T., Q.W., C.W., W.H., A.B., R.A., J.P., J.B., V.P., B.Y., and E.A.A. conceived and designed research. W.H. performed LVMi extraction from cardiac MRI images. T.T., A.A., and X.L. performed exploratory data investigations leading to development of lo-siRF. T.T., A.A., and B.Y. developed the lo-siRF pipeline; T.T. performed the lo-siRF analysis. Q.W. designed and created microfluidic devices; N.Y., S.S. and V.P. created gene-silenced hiPSC-CM lines; Q.W. and N.Y. performed experiments. Q.W., O.R., and A.K. performed single cell image analysis. Q.W. performed TF enrichment analysis. E.T.C. evaluated differences of gene connectivity between cardiac co-expression networks of failing and non-failing hearts. T.T., Q.W., A.K., O.R., C.W., V.P., W.H., B.Y., and E.A.A. interpreted results of experiments; T.T., Q.W., A.K., and C.W. prepared figures; T.T., Q.W., and C.W. drafted manuscript; All authors contributed in editing and revising manuscript.

Chapter 3

MDI+: A flexible random forest-based feature importance framework

3.1 Introduction

Random forests (RFs) (Breiman, 2001) are amongst the most popular supervised learning algorithms. They achieve state-of-the-art prediction performance over a wide class of learning problems (Caruana et al., 2008; Fernández-Delgado et al., 2014; Olson et al., 2018), often outperforming deep learning methods on small or moderately-sized tabular datasets (Shwartz-Ziv and Armon, 2022). These types of datasets frequently arise in high-stakes applications such as biology, medicine, and the social sciences due to high costs of data collection.

Given their strong predictive performance in these settings, practitioners are also keen on using RFs to extract new scientific insights (Basu et al., 2018; De Rosa et al., 2022). To illustrate the utility of RFs in real-world scientific problems, we focus on the two following case studies: (i) predicting the efficacy of cancer drugs given a patient's genetic information in The Cancer Cell Encyclopedia (CCLE) (Atlas, 2012), and (ii) classifying breast cancer subtypes using gene expression data catalogued in the Cancer Genome Atlas (TCGA) (Parker et al., 2009). In both problems, RFs have not only shown strong predictive performance (Costello et al., 2014; Wan and Pal, 2014; List et al., 2014), but have also been used to understand the genetic risk factors that drive drug response or are predictive of a particular breast cancer subtype.

Determining these genetic risk factors via RFs is done via feature importance measures, which summarize each feature's contribution to a model's predictions. Mean decrease in impurity (MDI) (Breiman et al., 1984) is arguably the most popular feature importance method for RFs, serving as the default measure in `scikit-learn` (Pedregosa et al., 2011). It measures the importance for a given feature X_k by tallying up the decrease in variance associated to each split on X_k in the ensemble. Scientists have used MDI in both case studies to pinpoint genes predictive of a drug's response on cancer cell lines (Wan and

Pal, 2014; Lind and Anderson, 2019) and to identify chemical pathways that differentiate between metaplastic and luminal breast cancer subtypes (List et al., 2014). Using feature importance measures to identify these biological factors provides a pathway to use powerful machine learning (ML) techniques such as RFs for improved diagnosis, early prevention, and development of novel therapeutics.

Despite these successes, MDI suffers from notable "biases" when used to identify important features. It favors features with higher entropy (e.g., continuous variables) or lower correlation with other features, and does so in a way that is independent of their relationship with the response function (Strobl et al., 2007; Nicodemus, 2011; Strobl et al., 2008; Nicodemus and Malley, 2009; Nembrini, 2019). This is unfortunate, because causal genes with low entropy or whose expression is strongly correlated with other nearby genes are commonly observed in datasets such as CCLE and TCGA. Consequently, MDI may produce inaccurate and misleading feature importance rankings, often leading to large downstream costs when used in high-stakes decisions such as deciding which genes to prioritize for future investigation.

Separately, Tan et al. (2021) established that trees are statistically inefficient at fitting regression functions with additive components. Such structure is ubiquitous in many scientific problems (Friedman and Popescu, 2008) including genomics, as seen by the prevalent use of linear models not only in both case studies (Jang et al., 2014; Parker et al., 2009), but also in other genomics applications such as understanding complex autoimmune conditions (Visscher et al., 2017). Since MDI is based on the fitted model, its fidelity to the true feature importances degrades when there is poor model fit (Murdoch et al., 2019; Yu and Kumbier, 2020).

To address these challenges, we propose a new framework for feature importance measures known as MDI+. The starting point of our framework is a recently discovered connection between decision trees and linear models (Klusowski, 2021; Agarwal et al., 2022). That is, a decision tree is the best fit linear model on a collection of engineered features associated with splits (i.e., local decision stumps) from the tree. Building on this connection, we *establish a new interpretation of MDI as an R^2 value in this linear regression*, and use this interpretation to explain the drawbacks listed above as well as to reveal several new ones. MDI+ then leverages this interpretation to overcome these shortcomings and generalize MDI by:

1. Using more flexible models (e.g., regularized generalized linear models (GLMs)) and similarity metrics in place of linear regression and R^2.
2. Appending smooth, linear features (e.g., the raw feature) and possibly other features (e.g., engineered from prior knowledge) to the nonlinear local decision stump representation.
3. Introducing computationally-efficient sample splitting for evaluating similarty metrics such as R^2.

As will be explained in subsequent sections, regularization and sample splitting helps to overcome the known biases of MDI, while appending linear features helps to overcome the model mismatch between trees and additive regression functions. The increased flexibility provided by the MDI+ framework offers practitioners the ability to tailor the feature importance computation to the problem structure and include any prior information via the choice of GLM, similarity metric, and expanded feature representation. To further assist practitioners with these choices, we provide a data- and stability-driven procedure to guide model selection and aggregation (i.e., combining importance rankings from multiple models). Additionally, we recommend that a practitioner investigate feature rankings produced by different MDI+ models in order to gain possible additional scientific insight. This procedure and holistic approach of investigating the output of different models is grounded in the Predictability, Computability, and Stability (PCS) framework for veridical data science (Yu and Kumbier, 2020). The PCS framework builds upon the three core principles in its name to bridge, unify, and expand on the best ideas from machine learning and statistics for the entire data science life cycle. Moreover, as a by-product of our MDI+ framework, the flexible use of GLMs and expanded feature representations yields a new class of powerful prediction models, RF+, which often improves upon the prediction performance of RFs and provides a bridge between classical statistical models (i.e., GLMs) with modern ML.

To demonstrate the effectiveness of MDI+, we conduct an extensive data-inspired simulation study, showing that MDI+ significantly outperforms other popular feature importance measures (e.g., TreeSHAP) in its ability to identify signal features. These data-inspired simulations were carefully designed to reflect a diverse range of problem structures (e.g., regression/classification, linear/non-linear) and challenges (e.g., low signal-to-noise, outliers, omitted variables, small sample sizes) commonly encountered in the real world. We also establish that MDI+ overcomes the biases of MDI through simulations with highly correlated features and features with varying levels of entropy. Finally, we return to the two motivating case studies and show that MDI+ feature rankings concur with established domain knowledge with significantly greater stability than other feature importance measures. As advocated by the PCS framework (Yu and Kumbier, 2020), methods that are more trustworthy when they are stable to reasonable perturbations. We also provide further evidence for the features selected by MDI+ by examining the predictive power of the top-ranked features. Furthermore, RF+ increases prediction accuracy over RF in these case studies by approximately 5%, further enhancing the credibility of the feature ranking from MDI+. The strong predictive performance of RF+ and stability of MDI+ in both case studies thus demonstrate its utility as a powerful and practical tool to extract reliable scientific insights in real-world problems.

Organization In Section 3.2 we review other popular feature importance measures. In Section 3.3, we establish our connection between MDI and R^2 values, and discuss how this interpretation explains new drawbacks, and reveals new ones. In Section 3.4, we introduce MDI+ and RF+. In Section 3.5, we run data-inspired simulations that show the efficacy of MDI+ in identifying relevant features. In Section 3.6, we show that MDI+ is able to to overcome the biases of MDI against highly correlated and low entropy features. In Section

3.7, we apply MDI+ to both of our motivating case studies, and investigate the stability and accuracy of its feature rankings in comparison with other feature importance measures. Finally, we conclude with a discussion in Section 3.8.

3.2 Related Work

In this section, we briefly review existing approaches to combat the known biases of MDI as well as other RF feature importance measures that are commonly used in practice.

MDI. Previous work (Strobl et al., 2007; Nicodemus, 2011) establish that MDI is biased towards high entropy features (e.g., continuous features). Many strategies have been proposed to mitigate this bias. Sandri and Zuccolotto (2008) proposed creating artificial uninformative features to evaluate bias, and Nembrini et al. (2018) created a fast implementation of this computationally intensive procedure. Zhou and Hooker (2021) and Li et al. (2019) proposed UFI and MDI-oob respectively, which both use out-of-bag samples to evaluate MDI in slightly different ways. Additionally, a penalized framework which combines in-bag and out-of-bag samples has been shown to debias MDI (Loecher, 2022a,b). Researchers have also tried to understand MDI theoretically by analyzing the population MDI values under some distributional assumptions (Louppe et al., 2013; Scornet, 2020) as well as obtaining consistency guarantees (Scornet, 2020). In particular, Scornet (2020) shows that the normalized sum of the MDI values of a single tree is the global R^2 (total variance explained) of the model, but does not connect individual MDI values to linear regression R^2 values. Finally, Klusowski and Tian (2021) show a connection between MDI and linear correlation for decision stump models and prove nonparametric variable selection consistency under some assumptions on the data-generating process.

MDA. In addition to MDI, Breiman (2001) proposed a permutation-based feature importance measure called Mean Decrease in Accuracy (MDA). To measure the importance of a feature X_k, MDA permutes the values of X_k marginally for out-of-bag samples, and calculates the excess prediction loss incurred on these samples. Recent work has studied MDA and variants thereof both empirically (Strobl et al., 2008; Genuer et al., 2008; Grömping, 2009; Genuer et al., 2010; Hooker and Mentch, 2019) and theoretically (Zhu et al., 2015; Gregorutti et al., 2017; Ramosaj and Pauly, 2019; Bénard et al., 2021).

TreeSHAP. Unlike MDI and MDA, SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) is a model-agnostic feature attribution procedure that provides local (i.e, sample-specific) feature importance scores. SHAP uses Shapley values from economic game theory to compute the contribution of each feature to a prediction for any given sample. These values can be summarized into a global feature importance measure by taking a mean over the samples. TreeSHAP (Lundberg et al., 2020) is a computationally-efficient implementation of SHAP values for tree-based methods.

Other feature importance methods and comparisons. To the best of our knowledge, MDI, MDA, and TreeSHAP are the most popular feature importance measures for RFs,

although both global and local scores can be defined in other ways (Ishwaran, 2007; Ishwaran et al., 2010; Kazemitabar et al., 2017; Epifanio, 2017; Saabas, 2022; Sutera et al., 2021; Klusowski and Tian, 2021). Also related are model-agnostic conditional dependence scores (Azadkia and Chatterjee, 2021; Zhang and Janson, 2020), model-agnostic confidence intervals for feature importances (Gan et al., 2022), and importance measures for interactions, such as via iterative random forests that have been widely adopted in the genomics community (Basu et al., 2018; Kumbier et al., 2018; Behr et al., 2020).

Dealing with correlated features. Empirical studies show that neither MDI nor MDA work well when features are highly correlated (Strobl et al., 2008; Nicodemus and Malley, 2009; Nicodemus, 2011; Nembrini, 2019). In particular, permutation-based measures such as MDA have been criticized because permutations break dependencies between features in the dataset (Hooker and Mentch, 2019). Hence, they result in evaluations of the model out-of-distribution, i.e., on regions of the covariate space with little or no data. These scores may have little connection to the underlying data-generating process because RFs are known to extrapolate to such regions in unreliable ways. This problem of out-of-distribution evaluation also affects TreeSHAP. To overcome this, variants of permutation scores have been proposed (Strobl et al., 2008; Hooker and Mentch, 2019) while other works have investigated altering the RF algorithm altogether (Hothorn et al., 2006).

3.3 Connecting MDI to R^2 Values from Linear Regression

Assume we are given a dataset $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with covariates $\mathbf{x}_i \in \mathbb{R}^p$ and responses $y_i \in \mathbb{R}$. An RF is an ensemble of classification or regression trees (CART) (Breiman et al., 1984; Breiman, 2001) that are fitted independently of one another on bootstrapped versions $\mathcal{D}_n^*$ of $\mathcal{D}_n$. Each CART model is fit by performing recursive axis-aligned splits according to a minimum impurity decrease criterion.

In more detail, a *potential split* s of a node $\mathtt{t}$ partitions it into two children nodes $\mathtt{t}_L = \{\mathbf{x}_0 \in \mathtt{t}\colon x_{0,k} \leq \tau\}$ and $\mathtt{t}_R = \{\mathbf{x}_0 \in \mathtt{t}\colon x_{0,k} > \tau\}$ for some feature index k and threshold τ. The *impurity decrease* of s is defined as

$$\hat{\Delta}(s, \mathcal{D}_n^*) \coloneqq N(\mathtt{t})^{-1} \left(\sum_{\mathbf{x}_i \in \mathtt{t}} (y_i - \bar{y}_{\mathtt{t}})^2 - \sum_{\mathbf{x}_i \in \mathtt{t}_L} (y_i - \bar{y}_{\mathtt{t}_L})^2 - \sum_{\mathbf{x}_i \in \mathtt{t}_R} (y_i - \bar{y}_{\mathtt{t}_R})^2 \right), \tag{3.1}$$

where all summations are over samples in $\mathcal{D}_n^*$, $N(\mathtt{t})$ represents the number of bootstrap samples in node $\mathtt{t}$, and $\bar{y}_{\mathtt{t}}, \bar{y}_{\mathtt{t}_L}, \bar{y}_{\mathtt{t}_R}$ are the mean responses in each node. Starting with the entire covariate space as the root node and until it reaches a stopping condition, CART recursively splits each node by picking and actualizing the potential split with the largest impurity decrease. When CART is used as part of RF, a random subset of features is chosen at each node and the best split is chosen from within this subset. This is to introduce

further diversity amongst the trees in the ensemble. Finally, at prediction time, a CART model identifies the unique leaf node containing the query point and predicts the mean response over that node.

To define the MDI values for a CART model, first let $\mathcal{S} = \{s_1, \ldots, s_m\}$ denote all the splits it contains and call this its *tree structure*. For each $k = 1, \ldots, p$, the MDI of feature X_k is defined as

$$\mathrm{MDI}_k(\mathcal{S}, \mathcal{D}_n^*) \coloneqq \sum_{s \in \mathcal{S}^{(k)}} n^{-1} N\left(\mathfrak{t}(s)\right) \hat{\Delta}(s, \mathcal{D}_n^*), \tag{3.2}$$

where $\mathcal{S}^{(k)}$ is the subset of splits in $\mathcal{S}$ that split on the feature X_k and $\mathfrak{t}(s)$ is the node being split by a split s. For RFs, the MDI of feature X_k is the mean of these values across all CART trees in the ensemble.

Now to build the connection between MDI and R^2 values given a tree structure $\mathcal{S}$ and data $\mathcal{D}_n$, we associate to each split $s \in \mathcal{S}$ the local decision stump function

$$\psi(\mathbf{x}; s, \mathcal{D}_n) = \frac{N\left(\mathfrak{t}_R\right) \mathbf{1}\{\mathbf{x} \in \mathfrak{t}_L\} - N\left(\mathfrak{t}_L\right) \mathbf{1}\{\mathbf{x} \in \mathfrak{t}_R\}}{\sqrt{N\left(\mathfrak{t}_L\right) N\left(\mathfrak{t}_R\right)}}. \tag{3.3}$$

Intuitively, ψ is a tri-valued function that indicates whether the sample $\mathbf{x}$ lies to the left of the threshold, lies to the right of the threshold, or is not contained in node $\mathfrak{t}$ at all. If $\mathcal{S}$ has m splits, concatenating these m functions yields the learned feature map $\Psi(\mathbf{x}; \mathcal{S}, \mathcal{D}_n) \coloneqq (\psi(\mathbf{x}; s_1, \mathcal{D}_n), \ldots, \psi(\mathbf{x}; s_m, \mathcal{D}_n))$ and its corresponding transformed dataset $\Psi\left(\mathbf{X}; \mathcal{S}, \mathcal{D}_n\right) \in \mathbb{R}^{n \times m}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the original data matrix. Klusowski (2021) showed that the CART model $\hat{f}(-; \mathcal{S}, \mathcal{D}_n)$ is equivalent to the best fit linear model of the responses $\mathbf{y}$ on the transformed dataset $\Psi\left(\mathbf{X}; \mathcal{S}, \mathcal{D}_n\right)$. We further note that there is a natural partition of decision stumps according to the (original) feature that they split on. That is, we have $\mathcal{S} = \mathcal{S}^{(1)} \sqcup \cdots \sqcup \mathcal{S}^{(p)}$, and can write $\Psi(\mathbf{x}; \mathcal{S}, \mathcal{D}_n) = \left(\Psi(\mathbf{x}; \mathcal{S}^{(1)}, \mathcal{D}_n), \ldots, \Psi(\mathbf{x}; \mathcal{S}^{(p)}, \mathcal{D}_n)\right)$. One can check that the corresponding blocks of $\Psi(\mathbf{X}; \mathcal{S}, \mathcal{D}_n)$ are orthogonal, enabling us to extend Klusowski (2021)'s result and derive the following new connection between MDI and R^2 values.

Theorem 1. *Assume we have a tree structure $\mathcal{S}$ and a dataset $\mathcal{D}_n = (\mathbf{X}, \mathbf{y})$, with $\mathbf{X}$ denoting the matrix of covariates and $\mathbf{y}$ denoting the response vector. For any feature X_k, we have the following identity:*

$$\frac{\mathrm{MDI}_k(\mathcal{S}, \mathcal{D}_n)}{n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n \left(y_i - \hat{y}_i^{(k)}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} =: R^2(\mathbf{y}, \hat{\mathbf{y}}^{(k)}), \tag{3.4}$$

where $\hat{\mathbf{y}}^{(k)} = \left(\hat{y}_1^{(k)}, \hat{y}_2^{(k)}, \ldots, \hat{y}_n^{(k)}\right)$ is the vector of fitted response values when regressing $\mathbf{y} \sim \Psi(\mathbf{X}; \mathcal{S}^{(k)}, \mathcal{D}_n)$.[1]

[1] If $\mathcal{S}^{(k)} = \emptyset$, the fitted model is the constant intercept model, and both the MDI and the R^2 value of this regression are equal to 0.

Here, the fitted values $\hat{\mathbf{y}}^{(k)}$, or *partial model predictions*, are the resulting model predictions using only those decision stumps that split on X_k. Theorem 1 thus formalizes the intuition that a more important feature X_k leads to more accurate predictions based solely on X_k and hence a larger MDI.

We also note that the partial model predictions $\hat{\mathbf{y}}^{(k)}$ are precisely the Saabas (2022) local feature importance scores for X_k. As such, Theorem 1 is implied by but also helps to clarify Proposition 1 in Li et al. (2019), which equates the MDI of X_k to the sample covariance between the Saabas scores and the responses but does not derive this in a linear regression setting. We prove Theorem 1 directly in Appendix B.1.

Reinterpreting MDI via OLS

Theorem 1 enables us to reinterpret the computation of MDI for a single tree in an RF via the following procedure (Figure 3.1, left). Given a fitted tree, which was trained on a bootstrapped dataset $\mathcal{D}_n^* = (\mathbf{X}^*, \mathbf{y}^*)$ and has the tree structure $\mathcal{S} = \mathcal{S}(\mathcal{D}_n^*)$, MDI can be computed by:

Step 1: Obtain transformed dataset on in-bag samples. Construct the feature map $\Psi(\mathbf{x}; \mathcal{S}, \mathcal{D}_n^*)$ and use it to obtain the transformed in-bag dataset:

$$\Psi(\mathbf{X}^*; \mathcal{S}, \mathcal{D}_n^*) = [\Psi(\mathbf{X}^*; \mathcal{S}^{(1)}, \mathcal{D}_n^*), \ldots, \Psi(\mathbf{X}^*; \mathcal{S}^{(p)}, \mathcal{D}_n^*)].$$

For ease of notation, we will henceforth suppress the dependence on $\mathcal{S}$ and $\mathcal{D}_n^*$ when describing the MDI procedure, and denote $\Psi^k(-) = \Psi(-; \mathcal{S}^{(k)}, \mathcal{D}_n^*)$.

Step 2: Fit linear model. Fit an ordinary least squares (OLS) model for $\mathbf{y}$ on $\Psi(\mathbf{X}^*)$, obtaining the estimated regression coefficient $\hat{\boldsymbol{\beta}}$.

Step 3: Make partial model predictions. For each $k = 1, \ldots, p$, obtain the partial model predictions $\hat{\mathbf{y}}^{(k)}$. Due to the orthogonality and centering of $\Psi(\mathbf{X}^*)$, $\hat{\mathbf{y}}^{(k)}$ as defined in Theorem 1 is equivalent to imputing the mean value for all stump features in $\Psi(\mathbf{X}^*)$ that do not split on X_k and then multiplying this modified matrix by $\hat{\boldsymbol{\beta}}$. Formally,

$$\hat{\mathbf{y}}^{(k)} = [\bar{\Psi}^1(\mathbf{X}^*), \ldots, \bar{\Psi}^{k-1}(\mathbf{X}^*), \Psi^k(\mathbf{X}^*), \bar{\Psi}^{k+1}(\mathbf{X}^*), \ldots, \bar{\Psi}^p(\mathbf{X}^*)]\hat{\boldsymbol{\beta}},$$

where $\bar{\Psi}^j(\mathbf{X}^*) = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\Psi^j(\mathbf{X}^*)$ and $\mathbf{1}_n$ is an $n \times 1$ vector of all ones.

Step 4: Evaluate predictions via R^2. For each $k = 1, \ldots, p$, the MDI for feature k in the given tree is precisely the unnormalized R^2 value between the observed responses $\mathbf{y}$ and the partial model predictions $\hat{\mathbf{y}}^{(k)}$, as shown by Theorem 1.

Drawbacks of MDI

Viewing MDI via this new linear regression lens allows us to explain known drawbacks as well as reveal new challenges. We consolidate these into two main issues, differential optimism bias and model mismatch, and then discuss possible ways to overcome them.

Differential Optimism Bias.

For a fixed design with homoskedastic noise, it is known that the average difference between training and held-out test mean-squared error (i.e. the *optimism bias*) for linear models is twice the number of degrees of freedom scaled by the noise variance (Hastie et al., 2009). Combining Theorem 1 with a similar set of calculations leads to the following optimism bias of MDI.

Proposition 1. *Suppose $\mathcal{D}_n$ is generated according to a fixed design with fixed covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and responses $y_i = f(\mathbf{x}_i) + \epsilon_i$ with $\epsilon_1, \ldots, \epsilon_n$ independent and satisfying $Var(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$. Let $\mathcal{D}_n^0 = \{(\mathbf{x}_i, f(\mathbf{x}_i))\}$ denote the noiseless dataset. Then for any fixed tree structure $\mathcal{S}$, we have*

$$\mathbb{E}\{\mathrm{MDI}_k(\mathcal{S}, \mathcal{D}_n)\} = \mathrm{MDI}_k(\mathcal{S}, \mathcal{D}_n^0) + \frac{\sigma^2 \left|\mathcal{S}^{(k)}\right|}{n}.$$

The noiseless MDI value, $\mathrm{MDI}_k(\mathcal{S}, \mathcal{D}_n^0)$, is the proportion of variance in $f(\mathbf{x})$ explained by $\Psi(\mathbf{x}; \mathcal{S}^{(k)}, \mathcal{D}_n)$. Proposition 1 reveals that $\mathrm{MDI}(k; \mathcal{S}, \mathcal{D}_n)$ is optimistic for $\mathrm{MDI}(k; \mathcal{S}, \mathcal{D}_n^0)$ and, further that the optimism bias is larger for features X_k that have more splits, leading to a *differential optimism bias*. This differential optimism is problematic because the split frequency of a feature X_k reflects not only the feature's inherent influence on the response, but also some biases due to the tree-growing process of CART. In particular, CART's splitting rule results in highly-correlated and low-entropy features receiving fewer splits (see Section 3.6). Signal features with these characteristics hence often have undesirably lower MDI than non-signal features without these characteristics, as a result of the differential optimism bias.

Though Proposition 1 requires a fixed tree structure, when $\mathcal{S}$ is random, the result above can be applied conditionally on $\mathcal{S}$, as long as it is independent of $\mathcal{D}_n$. This independence can be achieved if out-of-bag samples are used to compute MDI, as is sometimes done in practice. However, Proposition 1 shows that this does not resolve the differential optimism bias. More commonly, in-bag samples are used to both generate $\mathcal{S}$ and compute MDI. In this case, Proposition 1 does not apply directly, but because the splits $\mathcal{S}$ tend to be correlated with the noise, the amount of overfitting (i.e., differential optimism bias) should be worse.

Mitigating Differential Optimism Bias. To remedy the differential optimism bias, we can (1) replace OLS (i.e., step 2 of MDI procedure) with a regularized regression algorithm such as ridge (Hoerl and Kennard, 1970) or LASSO (Tibshirani, 1996) to reduce the effective degrees of freedom (i.e., number of splits) and stabilize the feature importance calculation. (2) Traditionally, MDI is evaluated on the same data (i.e., in-bag) used to train the tree, thereby amplifying biases in the learned model. To overcome this issue, Li et al. (2019) proposed a sample-splitting technique, known as MDI-oob, which computes the covariance between the out-of-bag (OOB) partial model predictions $\hat{\mathbf{y}}^{(k)}$ and OOB responses $\mathbf{y}$. This is equivalent to computing the R^2 value for the fitted linear model (trained on in-bag samples) when evaluated on OOB samples. That is, sample-splitting approaches mitigate differential optimism bias by using one portion of the data (e.g., in-bag data) to learn the linear model,

and the other (out-of-bag data) to evaluate its generalization error. However, this comes at the cost of using fewer samples to *both* fit OLS and estimate its out-of-sample error. To more efficiently use samples, one can instead use a leave-one-out (LOO) scheme that leverages the entire dataset (in- and out-of-bag samples) to both learn the linear model and estimate its out-of-sample error.

Model Mismatch

While RF is a nonparametric method that can in principle approximate any functional structure, it comes with a set of inductive biases that allow it to adapt better to some types of structure rather than others. First, trees are statistically inefficient at estimating smoothly-varying functions (Tsybakov, 2004) and additive generative models (Tan et al., 2021), which are ubiquitous in real-world datasets (Hastie and Tibshirani, 1986). This inefficiency results from the local decision stump features being poorly adapted to fit smooth or additive functions. Specifically, their piecewise constant nature makes them inefficient at representing smooth relationships, while their locality makes them inefficient at representing additive structure. Second, our reinterpretation shows MDI implicitly relies on OLS and R^2 to measure feature importance. However, OLS and R^2 may not be well-suited for all problem structures (e.g., in situations with a categorical response or gross outliers).

Mitigating Model Mismatch. To alleviate the inductive biases of RFs against additive or smooth structures, we can perform more flexible feature engineering by augmenting the k^{th} block of local decision stump features $\Psi^{(k)}(\mathbf{X})$ with other engineered features based on X_k. In particular, simply augmenting X_k itself can help bridge the gap between RFs and linear models. Secondly, we can replace OLS and the R^2 metric with prediction models and metrics that may be better suited for the given problem and data structure. For example, logistic regression with negative log-loss or Huber regression with negative Huber loss might be more appropriate for classification problems or problems with outliers, respectively.

3.4 Introducing MDI+

In this section, we build upon the ideas introduced in the previous section to develop our feature importance framework, MDI+.

The MDI+ Framework

MDI+ improves upon MDI by directly addressing its drawbacks described in the previous section and by offering practitioners a flexible and adaptable framework for computing feature importances. In particular, within this framework, practitioners are allowed several choices (i.e., the choice of feature augmentation, GLM, and similarity metric) that can be tailored to the data or problem structure. We discuss how these choices can be made in a

principled manner in more detail in Section 3.4 and proceed now to introduce the overarching MDI+ framework for a given choice of feature augmentation, GLM, and similarity metric.

As a reminder of relevant notation, the original dataset is denoted by $\mathcal{D}_n = (\mathbf{X},\ \mathbf{y})$, and each tree is fitted on a bootstrapped dataset $\mathcal{D}_n^* = (\mathbf{X}^*,\ \mathbf{y}^*)$. Also, recall we denote the tree structure as $\mathcal{S} = \mathcal{S}(\mathcal{D}_n^*)$, and the feature map as $\Psi(\mathbf{x};\ \mathcal{S},\ \mathcal{D}_n^*)$. We next introduce MDI+ for a single tree in an RF (Figure 3.1, right) using an analogous scaffolding as MDI (see Section 3.3) but with the main differences highlighted in bold.

*Step 1: Obtain **augmented** transformed dataset on **in- and out-of-bag** samples.* For features $k = 1, \ldots, p$, construct an augmented feature map $\tilde{\Psi}(\mathbf{x};\ \mathcal{S}^{(k)}, \mathcal{D}_n^*)$ by appending the raw feature X_k to the feature map $\Psi(\mathbf{x}; \mathcal{S}^{(k)}, \mathcal{D}_n^*)$.[2] That is, let $\tilde{\Psi}(\mathbf{x};\ \mathcal{S}^{(k)}, \mathcal{D}_n^*) = [\Psi(\mathbf{x};\ \mathcal{S}^{(k)}, \mathcal{D}_n^*),\ x_k])$ if $\mathcal{S}^{(k)} \neq \emptyset$. Let $\tilde{\Psi}(\mathbf{x};\ \mathcal{S},\ \mathcal{D}_n^*) = \Big(\tilde{\Psi}(\mathbf{x};\ \mathcal{S}^{(1)},\ \mathcal{D}_n^*), \ldots, \tilde{\Psi}(\mathbf{x};\ \mathcal{S}^{(p)},\ \mathcal{D}_n^*)\Big)$, and apply $\tilde{\Psi}(\mathbf{x};\ \mathcal{S},\ \mathcal{D}_n^*)$ to the entire dataset (in- and out-of-bag samples) to get $\tilde{\Psi}(\mathbf{X};\ \mathcal{S},\ \mathcal{D}_n^*)$. Henceforth, we denote $\tilde{\Psi}(\mathbf{X}) = \tilde{\Psi}(\mathbf{X};\ \mathcal{S},\ \mathcal{D}_n^*)$ and $\tilde{\Psi}^{(k)}(\mathbf{X}) = \tilde{\Psi}(\mathbf{X};\ \mathcal{S}^{(k)};\ \mathcal{D}_n^*)$.

*Step 2: Fit **regularized generalized** linear model (GLM).* Fit a regularized GLM $\mathcal{M}$ with link function g and penalty parameter λ for $\mathbf{y}$ using $\tilde{\Psi}(\mathbf{X})$ to obtain the estimated regression coefficients $\hat{\boldsymbol{\beta}}_\lambda$ and intercept $\hat{\alpha}_\lambda$. We tune λ via the approximate leave-one-out (LOO) method described in Rad and Maleki (2020), which does not require re-fitting the GLM n times. Note that we use the full dataset $(\mathbf{X},\ \mathbf{y})$ comprising both in- and out-of-bag samples to perform this as well as subsequent steps.

*Step 3: Make partial model prediction via **LOO**.* For a feature X_k, impute the mean value for all features in $\tilde{\Psi}(\mathbf{X})$ not derived from X_k. Then, for each feature $k = 1, \ldots, p$, and each sample $i = 1, \ldots, n$, obtain the LOO partial model predictions:

$$\hat{y}_i^k = g^{-1}\left(\left[\bar{\tilde{\Psi}}^1(\mathbf{X}), \ldots, \bar{\tilde{\Psi}}^{k-1}(\mathbf{X}), \tilde{\Psi}^k(\mathbf{X}), \bar{\tilde{\Psi}}^{k+1}(\mathbf{X}), \ldots, \bar{\tilde{\Psi}}^p(\mathbf{X})\right] \hat{\boldsymbol{\beta}}_{-i,\lambda} + \alpha_\lambda\right), \tag{3.5}$$

where $\bar{\tilde{\Psi}}^j(\mathbf{X}) = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\tilde{\Psi}^j(\mathbf{X})$, $\mathbf{1}_n$ is an $n \times 1$ vector of all ones, and $\hat{\boldsymbol{\beta}}_{-i,\lambda}$ is the LOO coefficient vector learnt without using sample $\mathbf{x}_i$. Again, we use the approximate LOO method of Rad and Maleki (2020) to obtain $\hat{\boldsymbol{\beta}}_{-i,\lambda}$ without refitting the GLM. Denote the LOO partial model predictions for feature k by $\hat{\mathbf{y}}^{(k)} = \big(\hat{y}_1^k, \ldots,\ \hat{y}_n^k\big)$.

*Step 4: Evaluate predictions via **similarity metric**.* Pick a similarity metric m and use it to evaluate the similarity[3] between the true responses and the partial model predictions. In other words, given the tree structure $\mathcal{S}$, dataset $\mathcal{D}_n^*$, together with choices of $\tilde{\Psi}$, $\mathcal{M}$ and m for Steps 1, 2 and 4, we define, for each $k = 1, \ldots, p$, the MDI+ for feature k as

$$\mathrm{MDI}_k^+(\mathcal{S}, \mathcal{D}_n^*, \tilde{\Psi}, \mathcal{M}, m) \coloneqq m\left(\mathbf{y}, \hat{\mathbf{y}}^{(k)}\right). \tag{3.6}$$

For a RF, the MDI+ of a feature k can be computed by taking the average across the trees.[4]

[2] More generally, we can augment this feature map with any number of engineered features derived from X_k.

[3] The metric should attain larger values on closer arguments.

[4] If a feature is never split on amongst all trees in the RF, we set its MDI+ to be $-\infty$.

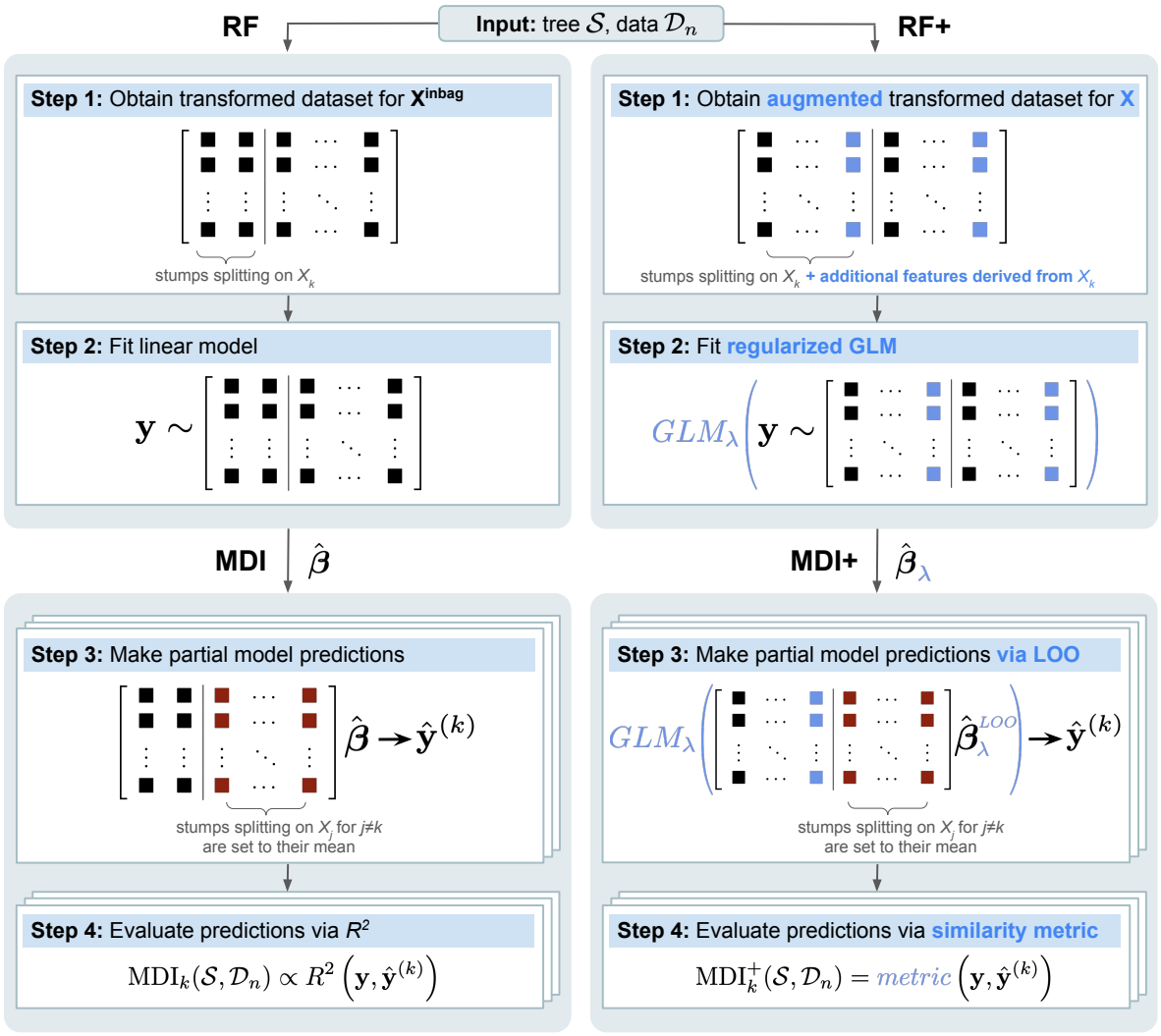


Figure 3.1: Overview of MDI+ for a single tree. For each tree $\mathcal{S}$ in the random forest, **Step 1:** Obtain the transformed dataset on the in- and out-of-bag samples using stumps from the tree and append the raw and/or any additional (possibly engineered) features. **Step 2:** Fit a regularized GLM. **Step 3:** Using the fitted GLM, make partial model predictions $\hat{\mathbf{y}}^{(k)}$ for each feature $k = 1, \ldots, p$ (stacked boxes) using a leave-one-out (LOO) data splitting scheme. **Step 4:** For each $k = 1, \ldots, p$, evaluate partial model predictions via any user-defined similarity metric to obtain the MDI+ for feature k in tree $\mathcal{S}$.

RF+. The GLM $\mathcal{M}$ fitted on the augmented transformed dataset $\tilde{\Psi}(\mathbf{X})$ serves as a new class of prediction models, which we refer to as RF+. RF+ can be viewed as a generalization of RFs that furthermore provides a natural link between classical statistical models (i.e., GLMs) and modern ML (i.e., RFs). In Section 3.7 and Appendix B.7, we show that RF+ improves upon prediction performance of RFs by approximately 5% across a variety of real-world datasets in the regression and classification settings. Though the main focus of this work is feature importance, the strong prediction performance of RF+ suggests that it better captures the underlying data-generating process, giving additional credence to MDI+ (Murdoch et al., 2019; Yu and Kumbier, 2020).

PCS-Informed Model Recommendation

One rarely has sufficient prior knowledge of the problem structure in order to make definitive choices for $\tilde{\Psi}$, $\mathcal{M}$ and m. Further, even if there is prior information, a practitioner might be left with multiple choices (e.g., using lasso or ridge for $\mathcal{M}$). To address this challenge, we briefly discuss two approaches inspired by the PCS framework (Yu and Kumbier, 2020) to make these modeling choices in a data-driven manner, and defer details to Appendix B.6.

Stability-based selection for $\tilde{\Psi}$, $\mathcal{M}$, m. Let $h = \left(\tilde{\Psi}(\mathbf{X}), \mathcal{M}, m\right)$ denote an MDI+ model defined by the choices of augmented feature representation, GLM, and similarity metric, respectively. Accordingly, let $\mathcal{H} = \{h_1, \ldots, h_N\}$ denote the set of possible MDI+ models under consideration.

Step 1: Prediction Screening. For each $h \in \mathcal{H}$, evaluate the prediction performance of $\mathcal{M}$ on a held-out test set, and filter out all h whose prediction performance is worse than that of RF. This prediction check ensures that the interpreted model is a reasonably faithful approximation of the underlying data generating process (Murdoch et al., 2019; Yu and Kumbier, 2020).

Step 2: Stability Selection. Generate B bootstrapped samples of the fitted trees in the ensemble. For each h that passed the prediction screening, evaluate MDI+ for each of the B bootstrapped samples, and choose the h which yields the most similar (or stable) feature rankings across the B bootstrapped samples. We measure the similarity between different bootstrap samples via Rank-based Overlap (Webber et al., 2010). While we measure stability via bootstrap sampling, one can also measure stability over different algorithmic perturbations (e.g., random seeds used to train the RF).

Model aggregation via PCS principles. Instead of choosing a single best model for MDI+, it may make sense to compare and ensemble multiple models that have similar predictive performance with respect to the criteria identified above, as it can be argued that they all provide competing descriptions of reality that have approximately equal statistical evidence (Murdoch et al., 2019; Rudin et al., 2021). Further, investigating and comparing the feature rankings of different predictive-screened MDI+ models might be of independent scientific interest.

In Appendix B.6, we perform a preliminary simulation study to establish the efficacy of both approaches. In particular, we show that the GLM and metric selected by the stability-based procedure leads to the best feature ranking performance across different candidate MDI+ models. While the model aggregation approach does not lead to the best feature ranking performance, it performs competitively, and we present it here since both approaches might be useful in practice. However, we leave a more thorough investigation of these model recommendation techniques to future work.

3.5 Data-Inspired Feature Ranking Simulations

In this section, we describe a number of simulations that illustrate the effectiveness of MDI+ to identify signal features in three common statistical settings: regression, classification, and robust regression (i.e., presence of outliers). In all settings, we highlight the flexibility of MDI+ choosing the appropriate GLM and evaluation metric for the data at hand. For all of our simulations in this section, we use covariate matrices coming from real-world datasets to capture naturally occurring structure. Using these covariate matrices, we simulate responses with analytical functions that depend on a sparse set of features. These response functions were chosen to reflect both canonical forms as well as domain knowledge. Then, we measure the ability of MDI+ and other feature importance methods to recover these signal features, with the specific metric for recovery discussed below. Additional simulations under varying sparsity levels, number of features, and misspecified model settings with omitted variables are provided in Appendix B.3, further supporting the strong empirical performance of MDI+.

Simulation Setup

Real-world datasets used. For our covariate matrices, we use the following datasets: (i) Juvenile dataset ($n = 3640$, $p = 277$) (Osofsky, 1997); (ii) a subset of the Cancer Cell Line Encyclopedia (CCLE) RNASeq gene expression dataset ($n = 472$, $p = 1000$) (Barretina et al., 2012); (iii) Enhancer dataset ($n = 7809$, $p = 41$) (Basu et al., 2018); and (iv) Splicing dataset ($n = 23823$, $p = 264$) (Basu et al., 2018).

Simulated responses. Using each dataset above as the covariate matrix $\mathbf{X}$, we consider the following response functions.

1. Linear model: $\mathbb{E}[Y \mid X] = \sum_{j=1}^{5} X_j$;
2. Locally-spiky-sparse (LSS) model (Behr et al., 2021): $\mathbb{E}[Y \mid X] = \sum_{m=1}^{3} \mathbf{1}(X_{2m-1} > 0)\mathbf{1}(X_{2m} > 0)$;
3. Polynomial interaction model: $\mathbb{E}[Y \mid X] = \sum_{m=1}^{3} X_{2m-1} + \sum_{m=1}^{3} X_{2m-1}X_{2m}$.
4. Linear + LSS model: $\mathbb{E}[Y \mid X] = \sum_{m=1}^{3} X_{2m-1} + \sum_{m=1}^{3} \mathbf{1}(X_{2m-1} > 0)\mathbf{1}(X_{2m} > 0)$;

These regression functions were chosen to reflect several archetypes of real DGPs. (1) and (3) reflect well-studied models in the statistics literature. (2) exhibits the discontinuous interactions observed in biological processes (Nelson et al., 2008; Behr et al., 2021). (4) reflects a combination of linear and discontinuous interactions also thought to be prevalent in genomics. For our classification simulations, we pass the mean response function (i.e., $\mathbb{E}[Y \mid X]$) through a logistic link function to convert the responses to be binary.

Feature importance methods under consideration. We compare MDI+ to a number of popular feature importance methods for RFs: MDI (Breiman et al., 1984), MDI-oob (Li et al., 2019), MDA (Breiman, 2001), and TreeSHAP (Lundberg and Lee, 2017).

RF settings. For the regression (Section 3.5) and robust regression (Section 3.5) settings, we train a RF regressor using `scikit-learn` (Pedregosa et al., 2011) with *n_estimators*=100 (i.e., number of trees), *max_features*=$p/3$ (i.e., proportion of features subsampled at each node), and *min_samples_leaf*=5 alongside other default parameters. In the classification setting (Section 3.5), we use `scikit-learn`'s RF classifier with *n_estimators*=100, *max_features*=$\sqrt{p}$, and *min_samples_leaf*=1 alongside other default parameters.

Feature Ranking Performance Metric. As in previous work (Li et al., 2019; Yu and Kumbier, 2020), we evaluate the performance of each feature importance method by how well it can be used to classify features used in the regression function (signal) vs. those that are not (non-signal). Each set of feature importance scores induces a ranking of the features, which can then be evaluated for this classification problem using AUROC. A high AUROC scores indicates that the signal features are ranked higher (i.e., more important) than the non-signal features. The performance of each feature importance method is averaged across 50 simulation replicates. In each simulation replicate, we choose the signal features randomly from $\mathbf{X}$.

Regression Simulations

We simulate responses as discussed above, but introduce additive Gaussian noise. That is, we simulate responses as $Y = \mathbb{E}[Y \mid X] + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ We examine performance across various signal-to-noise ratios as measured by the proportion of variance explained,[5] defined as $PVE = \text{Var}\{\mathbb{E}\{Y|X\}\}/\text{Var}\{Y\}$. More specifically, we vary across PVE in $\{0.1, 0.2, 0.4, 0.8\}$ (or equivalently, signal-to-noise ratio in $\{0.11, 0.25, 0.66, 4\}$) and across a range of sample sizes n. For CCLE, we vary across n in $\{100, 250, 472\}$, and for the other three datasets, we vary across n in $\{100, 250, 500, 1000, 1500\}$. We use l_2-regularized regression (i.e., ridge regression) as the GLM, and R^2 as our metric of choice.

We display our results for the Splicing dataset for the LSS, and polynomial interaction model (See Section 3.5 for details) in Figure 3.2. Results for other datasets and regression

[5] PVE is a monotone transformation of the signal-to-noise ratio (SNR) such that its range is bounded between 0 and 1 and thus more interpretable. It is also a standard measurement of noise used in many fields (e.g., it is often called *heritability* in genomics). PVE in genomics is estimated to range from between 0.05 to 0.4 (Wang et al., 2020).

functions are deferred to Appendix B.2 for brevity, but are similar to those displayed in Figure 3.2. Across all simulation scenarious, MDI+ produces more accurate feature rankings, often enjoying more than a 10% improvement in AUROC over its closest competitor (typically MDI-oob or MDA). In particular, MDI+ produces the most significant improvement for low PVEs, which is especially important since real-world data in fields such as biology, medicine, and social sciences have low SNRs. The strong performance of MDI+ relies on a number of choices made to mitigate the aforementioned biases such as: using l_2 regularizaton, including the raw feature, and evaluating predictions via LOO. In Appendix B.4, we provide simulations that show how each of these choices individually leads to an increase in the ranking accuracy of MDI+.

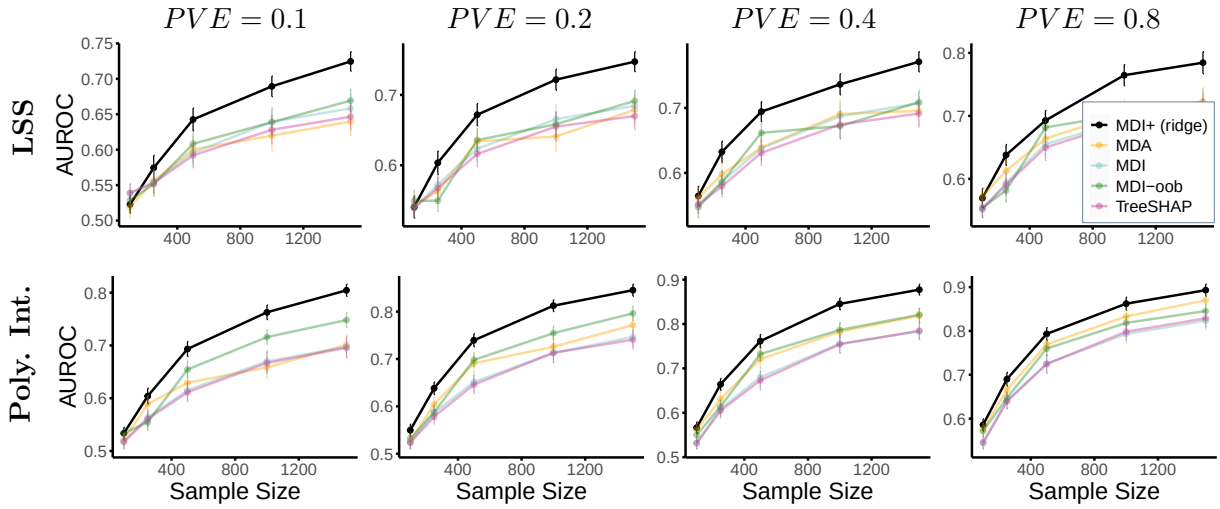


Figure 3.2: MDI+ outperforms all other feature importance methods for the data-inspired regression simulations described in Section 3.5 using the Splicing dataset. This pattern is evident across various regression functions (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

Classification Simulations

We simulate responses according to the response functions defined in Section 3.5, and introduce noise by randomly flipping a percentage of the binary response labels to the opposite class. We measure AUROC as we vary the percentage of corrupted labels in $\{0\%, 5\%, 15\%, 25\%\}$ and the number of samples as before. To tailor MDI+ to the classification setting, we use l_2-regularized logistic regression and negative log-loss as our choice of GLM and metric, respectively. We compare these choices to those used in the regression

setting (i.e., ridge regression and R^2). Henceforth, we shall refer to these particular settings as MDI+ (logistic) and MDI+ (ridge), respectively.

We display results for the CCLE dataset for the linear, and linear + LSS model (See Section 3.5 for details) in Figure 3.3. We defer results for other datasets and regression functions to Appendix B.2, but are similar to those displayed in Figure 3.2. Both MDI+ (logistic) and MDI+ (ridge) outperform competitors by over 10% across most simulation scenarios Further, we see that MDI+ (logistic) improves upon MDI+ (ridge), indicating the benefit of tailoring GLMs and metric to the statistical problem at hand. Exploring other GLMs and error metrics may further improve performance.

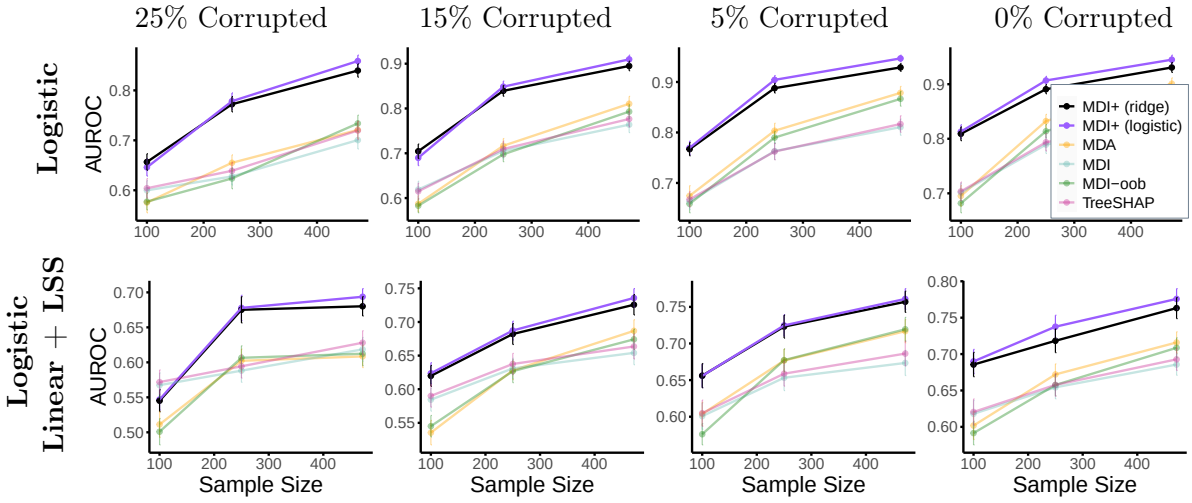


Figure 3.3: Both MDI+ (ridge) and MDI+ (logistic) outperform all other feature importance methods for the data-inspired classification simulations described in Section 3.5 using the CCLE RNASeq dataset. Furthermore, MDI+ (logistic) slightly outperforms MDI+ (ridge) in the majority of settings, indicating the benefit of tailoring the choices of MDI+ to the data at hand. This pattern is evident across various regression functions (specified by row), proportions of corrupted labels (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

Robust Regression Simulations

We illustrate another use-case of our framework by showing how MDI+ can be tailored to account for presence of outliers. We simulate responses as done in the regression setting, and introduce outliers as follows. We first select samples in the top and bottom select samples in the top and bottom $q/2\%$ quantiles for a randomly chosen *non-signal feature*. For the selected samples, we corrupt their responses by drawing their responses from $N(\mu_{corrupt}, 1)$

and $N(-\mu_{corrupt}, 1)$ for the bottom and top quantile samples respectively. In our simulations, we vary across q in $\{0, 1, 2.5, 5\}$ and $\mu_{corrupt}$ in $\{10, 25\}$ for a variety of sample sizes n. We tailor MDI+ to this setting by using a robust version of ridge regression (Owen, 2007) as our choice of regularized GLM, and negative Huber loss as our similarity metric. We compare these choices to those used in the regression setting (ridge regression and R^2). We shall refer to these particular settings as MDI+ (Huber) and MDI+ (ridge).

Simulation results for the Enhancer dataset are shown in Figure 3.4, with responses simulated via the LSS function (see Section 3.5 for details). Results for other datasets and regression functions are shown in Appendix B.2. We observe that MDI+ (Huber)'s performance is more robust than competing methods including MDI+ (ridge) and often remains accurate as both $\mu_{corrupt}$ and the percentage of outliers increase.

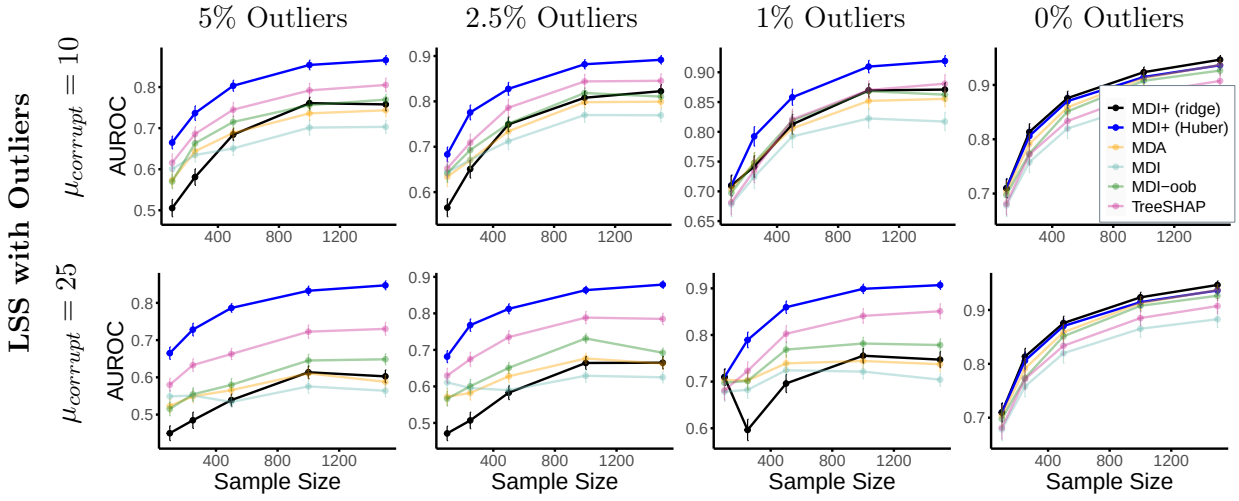


Figure 3.4: Under the LSS with outliers regression setting using the Enhancer dataset, MDI+ (Huber)'s performance remains suffers far less than other methods including MDI+ (Ridge) as the level of corruption $\mu_{corrupt}$ (specified by row) and the proportion of outliers (specified by column) grow. This pattern also holds across sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

3.6 MDI+ Overcomes Biases of MDI

As established in the literature and further discussed in Section 3.3, MDI is biased against highly-correlated and low-entropy features, which are commonly found in practice. In this section, we run simulations showing that MDI+ overcomes these biases.

Correlated Feature Bias

Experimental details. We draw $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{x}_i \in \mathbb{R}^{100}$ and $\mathbf{\Sigma} \in \mathbb{R}^{100\times100}$ has the following block-covariance structure: Features $X_1, \ldots, X_{50}$ have pairwise correlation ρ, and features $X_{51}, \ldots, X_{100}$ are independent of all other features. We then simulate responses from the linear+LSS model: $Y = \sum_{m=1}^{3} X_{2m-1} + \sum_{m=1}^{3} \mathbf{1}(X_{2m-1} > 0)\mathbf{1}(X_{2m} > 0)$ with PVE in $\{0.1, 0.4\}$. Denote the group of signal features (i.e., $X_1, \ldots, X_6$) as Sig, the non-signal features that have non-zero correlation with the signal features (i.e., $X_7, \ldots, X_{50}$) as C-NSig, and the non-signal uncorrelated features (i.e., $X_{51}, \ldots, X_{100}$) as NSig. We generate $n = 250$ samples, and vary the correlation across ρ in $\{0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$. For all feature importance methods used in our simulation study in Section 3.5, we compute the average ranks for features within each group (Sig, C-NSig, and NSig).

Results. The results for the average ranks of features per group are shown in Figure 3.5. Across all levels of correlation and PVEs, MDI+, MDI-oob, and MDA rank the true signal features (Sig, dark green) as the most important feature group and the uncorrelated non-signal group (NSig, red) as the least important feature group by a sizeable margin. MDI's behavior can be explained by Proposition 1, and Figure B.23, which displays the average percentage of RF splits per feature in each group. As ρ increases, Figure B.23 shows the percentage of splits on NSig features increases, while that of Sig and C-NSig decreases. Intuitively, this occurs because Sig and C-NSig features, being correlated with each other, result in similar decision stump functions and must compete over splits. Since MDI grows with the number of splits as established in Proposition 1, this leads to an overestimate of MDI for the NSig features. Moreover, MDI and TreeSHAP are measured on in-bag (training) samples, thereby amplifying biases that are learned during the tree construction. In contrast, MDI+, MDI-oob, and MDA use sample-splitting, which helps to mitigate this overfitting issue. A direct comparison between MDI+ with LOO sample splitting versus no sample splitting in Appendix B.5 (Figure B.22) helps to confirm this intuition.

Entropy Bias

Experimental details. Following the simulation setup proposed in Li et al. (2019), we sample 5 features: X_1 from a Bernoulli distribution with $p = 0.5$, X_2 from a standard Gaussian distribution, and X_3, X_4, and X_5 from a uniform discrete distribution with 4, 10, and 20 categories, respectively. To investigate entropy bias, we simulate the response as a function of only the lowest entropy feature, X_1, under (1) the regression setting via $Y = X_1 + N(0, \sigma^2)$ where σ^2 is chosen to achieve $PVE = 0.1$ and (2) the binary classification setting via $\mathbb{P}(Y = 1 \mid X) = \frac{1+X_1}{3}$. We vary the number of samples $n \in \{50, 100, 250, 500, 1000\}$ and measure the rank and proportion of splits in RF for each feature across 50 replicates.

Results. Figures 3.6 and B.23 display the results for the feature rankings and proportion of RF splits per feature respectively. Only MDI is able to identify X_1 as the signal features. This occurs because X_1 has the lowest entropy relative to other features, and is hence split

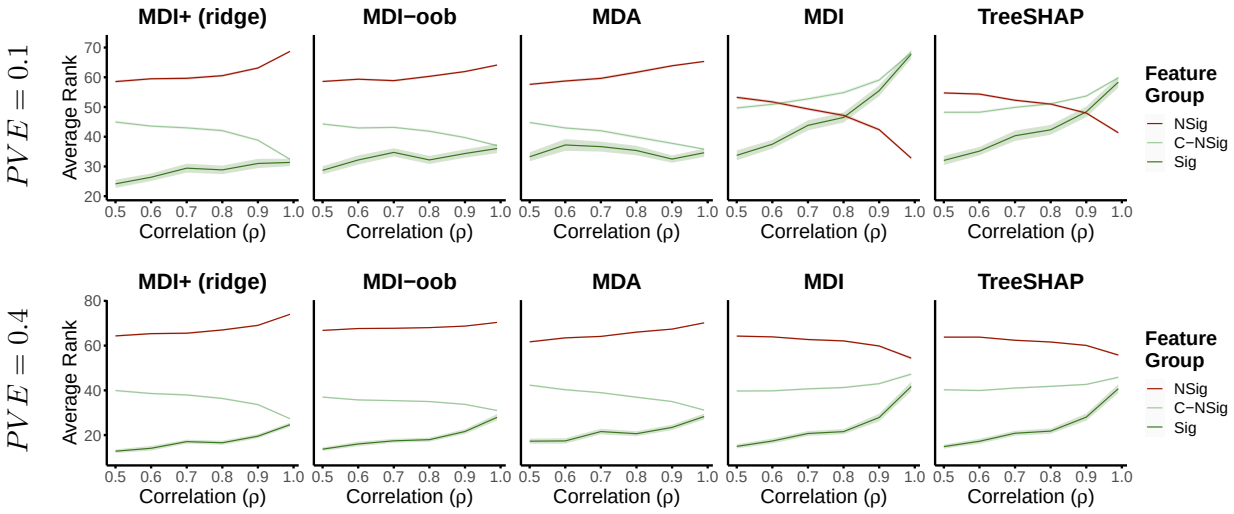


Figure 3.5: MDI+ improves upon the bias that MDI has against selecting features that are highly correlated. We show the average ranks ($\pm 1SE$) within feature groups (Sig, C-NSig, NSig) for various correlation (ρ) levels over 50 replicates. When the signal is moderate-to-high ($PVE = 0.4$, bottom), all methods rank the true signal features (Sig, dark green) as more important than the non-signal features (NSig, red) for all ρ; however, the gap is small for MDI and TreeSHAP. When the signal is low ($PVE = 0.1$, top), MDI+ , MDI-oob, and MDA are still able to identify the true signal features (Sig, dark green) as most important. In contrast, MDI and TreeSHAP rank the non-signal features (NSig, red) as more important than the true signal (Sig, dark green) and the correlated, non-signal (C-NSig, light green) feature groups when the ρ is large.

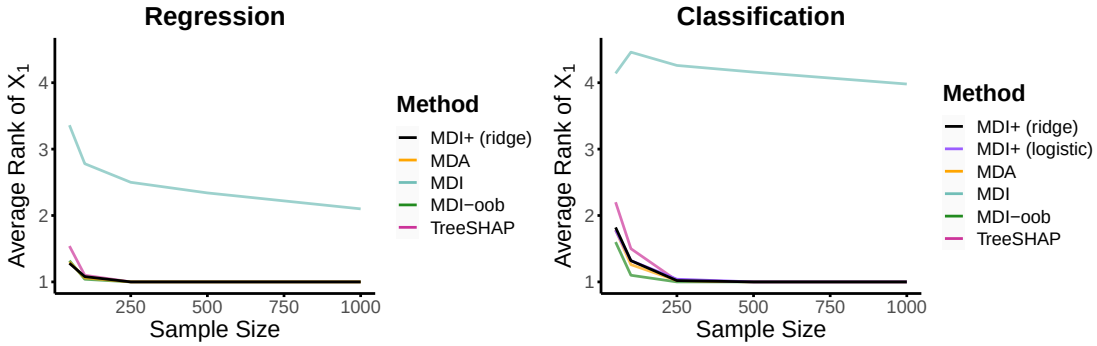


Figure 3.6: MDI+ improves upon the bias that MDI has against selecting features with lower entropy in both the regression (left) and classification (right) simulation settings described in Section 3.6. The feature ranking of the solo signal feature, X_1, averaged across 50 replicates, is shown on the y-axis. Here, a lower value indicates greater importance.

upon the least (Figure B.23). As a result of Proposition 1, this leads to an overestimate of MDI for high-entropy features $X_2, \ldots, X_5$ that are split upon more often. To combat this, MDI+ employs regularization to control the effective degrees of freedom and sample-splitting to mitigate biases learned during the tree construction. A direct comparison of MDI+ with and without ridge regularization and LOO sample-splitting in Figure B.24 helps confirm this.

3.7 Case Studies

In this section, we revisit our two driving case studies and use MDI+ as well as other feature importance measures to identify important features in predicting drug responses and breast cancer subtypes. In both case studies, we show that RF+ increases the prediction accuracy over RF and that the feature rankings from MDI+ agree with established domain knowledge with significantly greater stability (i.e., across different train-test splits and random seeds used to train the RF) than other feature importance measures. Given that both predictability and stability are important prerequisites for interpretability (Murdoch et al., 2019; Yu and Kumbier, 2020), these findings showcase the effectiveness of MDI+ for extracting reliable interpretations in real-world scientific problems.

Case Study I: Drug Response Prediction

Accurately predicting a cancer drug's efficacy for a patient before prescribing it can tremendously improve both the patient's health and financial well-being given the exorbitant costs of many cancer drugs. Moreover, identifying the important biological predictors, such as genes, that are influential of the drug response can provide valuable insights into potential targets and novel candidates for future preclinical research. Towards this end, we leverage data from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) to build accurate drug response models and identify genes whose expression levels are highly predictive of a drug's response.

Data and Methods. More specifically, we have gene expression data $\mathbf{X} \in \mathbb{R}^{472 \times 5000}$, measured via RNASeq, from $n = 472$ cell lines and $p = 5000$ genes after filtering. For each cell line, the response of 24 different drugs was measured, yielding a multivariate response matrix $\mathbf{Y} \in \mathbb{R}^{472 \times 24}$. We split the data into 80% training and 20% test. Details regarding the data and preprocessing can be found in Appendix B.8. Using the training data, we fit 24 separate RFs, one to predict the response for each drug using the gene expression data $\mathbf{X}$, and investigate the gene importance rankings for each drug. We repeat this for 32 different train-test splits. Here, the RF settings and feature importance methods used are the same as those in Section 3.5 for regression.

Prediction accuracy. For 23 out of the 24 drugs, the test R^2, averaged across the 32 train-test splits, was higher for RF+ than RF. Furthermore, RF+ improved the test R^2 by an

average of 4% across the 18 drugs with RF test $R^2 > 0.1$. We use this threshold of $R^2 > 0.1$ to focus on models that have non-trivial predictive power; however, the improvement from RF+ is even higher without this filter. Details and the full prediction results can be found in Appendix B.7.

Accuracy of gene importance rankings. In addition to the increase in prediction accuracy from RF+, MDI+ was able to identify several well-known genes that have been previously identified as drug targets in the previous literature. In particular, every known gene expression predictor of drug response identified in the original work on CCLE (Barretina et al., 2012), except for one, was ranked in the top 5 by MDI+ for their respective drugs. The one gene outside of the top 5 was *MDM2*, a known predictor for the drug Nutlin-3 (Barretina et al., 2012). Still, MDI+ ranked *MDM2* 17^{th}, which is higher than the rankings given by its competitors (MDI-oob: 22^{nd}, TreeSHAP: 30^{th}, MDI: 35^{th}, MDA: 53^{rd}). A full list of the top 5 genes for each drug using various feature importance methods is provided in Appendix B.8.

To provide another assessment of the gene importances beyond evidence from the scientific literature, we also evaluated the predictive power of the top 10 genes from each feature importance method. We found that the top 10 genes from MDI+ were often more predictive than that from other feature importance methods. Per the predictability principle of the PCS framework, strong prediction performance suggests that the model (and in this case, the top-ranked features) may better capture the underlying data-generating process. For brevity, we defer details and additional discussion regarding this prediction analysis to Appendix B.8.

Stability of gene importance rankings. Lastly, for each feature importance method, we investigated the stability (or similarity) of gene importance rankings across the 32 train-test splits. Methods that exhibit greater stability are highly advantageous in practice since it is undesirable for interpretations to change due to arbitrary choices like train-test splits and random seeds. In Figure 3.7a, we display one measure of stability, namely, the number of distinct genes ranked in the top 10 across the 32 train-test splits. For 22 out of the 24 drugs (the exceptions being PD-0325901 and Panobinostat), MDI+ had the fewest number of distinct genes that were ranked in the top 10 across the 32 train-test splits. Moreover, for all 24 drugs, we found that MDI+ had the highest number of genes that were always ranked in the top 10 across all train-test splits. Interestingly, sample-splitting techniques such as MDA and MDI-oob were significantly less stable than other methods, highlighting the downside of using fewer samples to measure feature importances. On the other hand, MDI+, which leverages LOO and regularization, overcomes this drawback. To further establish the stability of MDI+, we also examined the distribution of feature rankings of each method for the five most important genes across the 32 splits. As visualized in Figure 3.7b (results for all drugs in Figures B.28 and B.29), feature rankings of MDI+ tend to have smaller variance as compared to other methods. This remains true for PD-0325901 and Panobinostat, for which MDI+ did not perform optimally according to the aforementioned top 10 stability metric. A similar improvement in stability is also seen when fixing the training data and only altering the random seed used for fitting the RF (see Appendix B.8). This improved stability of

MDI+, in addition to the increase in prediction performance from RF+, demonstrate the practical advantages of MDI+ in this real-data case study on drug response prediction.

Case Study II: Breast Cancer Subtype Prediction

Breast cancer is known to be a heterogeneous disease (Atlas, 2012) and is thus often classified into various molecular subtypes. PAM50 is one prominent breast cancer subtyping method, which is frequently used to inform treatment plans (Parker et al., 2009). PAM50 divides breast cancers into five intrinsic subtypes: Luminal A, Luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, Basal-like, and Normal-like. Each subtype differs based upon their biological properties and generally corresponds to a different prognosis. Luminal A typically has the best prognosis while HER2-enriched and Basal-like are more aggressive (Caan et al., 2014). In this study, we aim to predict the PAM50-defined breast cancer subtypes using gene expression data from The Cancer Genome Atlas (TCGA) and more importantly, identify the genes that heavily determine the breast cancer subtype.

Data and Methods. Using data from the TCGA, we have RNASeq gene expression data from $n = 1083$ individuals and $p = 5000$ genes after filtering (details in Appendix B.8). We also have the PAM50-defined breast cancer subtype for each of these individuals, giving rise to a multi-class classification problem with five classes. As in Section 3.7, we split the data into 80% training and 20% test and evaluate results across 32 train-test splits. The RF settings and feature importance methods used are the same as those in Section 3.5 for classification. Note however that MDI-oob has not been implemented for multi-class classification and is thus omitted from this analysis.

Prediction accuracy. Comparing RF and RF+ across the 32 train-test splits, we found that RF+ (logistic) yielded the best prediction performance over a variety of metrics (see Table B.3). In particular, RF+ (logistic) yielded an average classification accuracy of 88.4%, compared to 86.1% from RF. RF+ (ridge) also improved the prediction performance over RF (accuracy: 87.3%), but was slightly worse than RF+ (logistic). The improvement of RF+ (logistic) over RF+ (ridge) provides further evidence of the benefit of tailoring the GLM to the problem structure.

Accuracy of gene importance rankings. With respect to the gene importance rankings, both MDI+ (ridge) and MDI+ (logistic) produced the same set of top 10 genes, albeit in a different ranking order. For each of these 10 genes (Table B.4), there is significant literature, supporting the gene's involvement in the development and progression of breast cancer. For example, *ESR1*, the top-ranked gene across all feature importance methods, has been widely studied over the last decade due to its prominent role in the pathogenesis of breast cancer (Toy et al., 2013; Chandarlapaty et al., 2016) as well as in novel promising therapies (Brett et al., 2021). *ESR1* encodes the estrogen receptor-α protein and is known to cause increased resistance to standard-of-care endocrine therapy (Brett et al., 2021). Several other genes in the top 10, namely, *FOXA1*, *FOXM1*, and *MLPH*, are also associated with estrogen

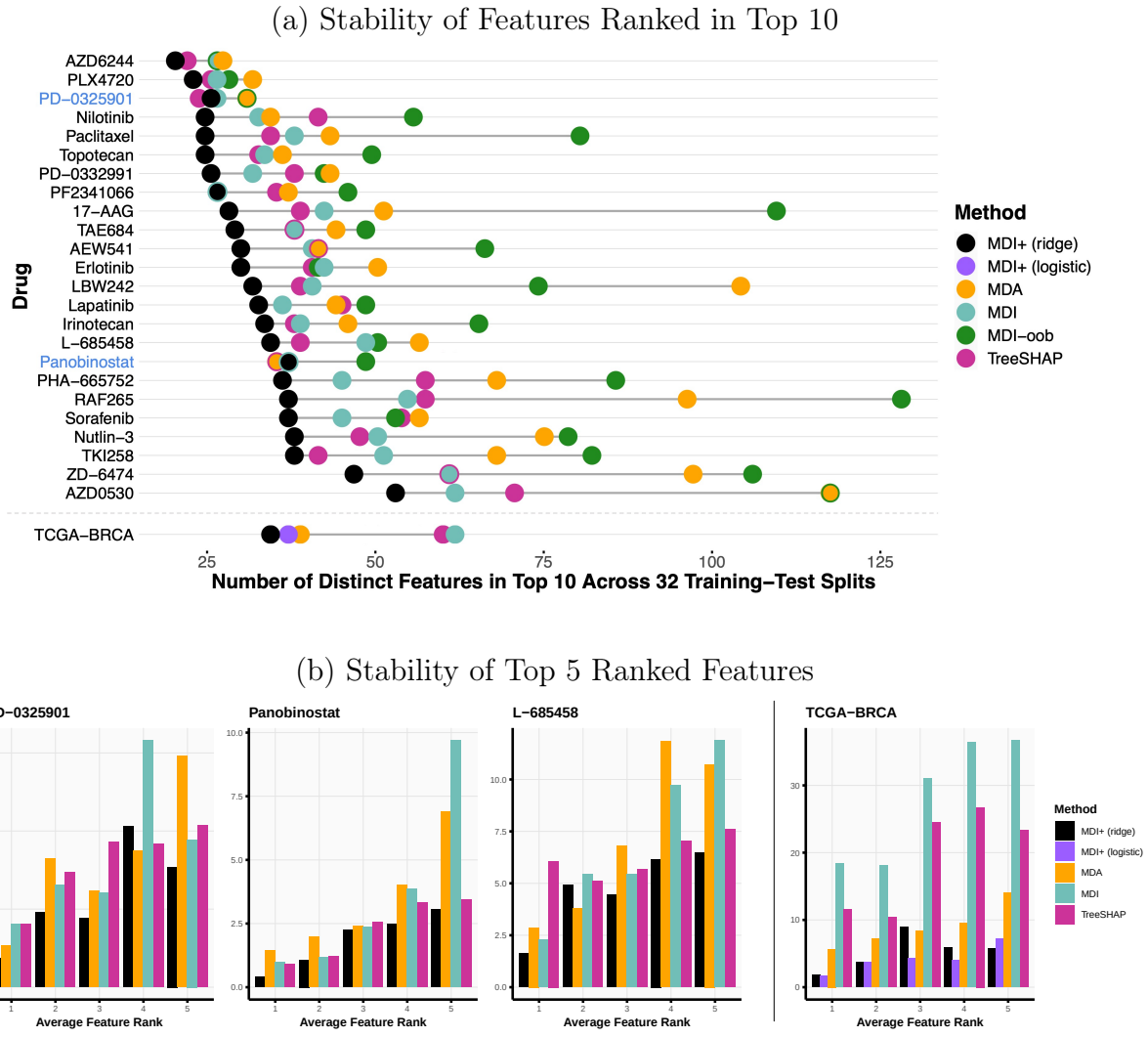


Figure 3.7: In the CCLE drug response and TCGA breast cancer (TCGA-BRCA) subtype case studies, MDI+ provided the most stable feature importance rankings across 32 train-test splits. Specifically, (a) for 22 out of the 24 drugs in the CCLE case study (the exceptions being PD-0325901 and Panobinostat, highlighted in blue) and in the TCGA-BRCA case study, MDI+ had the fewest number of distinct features that were ranked in the top 10 across the 32 train-test splits. Note that the outlined points denote ties and that MDI+ (logistic) is only used in the classification setting (i.e., for TCGA-BRCA). Moreover, (b) for the top five features as ranked by their average ranking across the 32 train-test splits, the standard deviation of the feature importance rankings across the 32 train-test splits is smaller for MDI+ compared to other competitor methods. We provide the results for three CCLE drugs and the TCGA-BRCA subtypes here. MDI-oob is also omitted due to its high instability.

receptor-positive breast cancer and increased resistance to endocrine therapy (Fu et al., 2019; Bergamaschi et al., 2014; Thakkar et al., 2010). Beyond this, *GATA3* (ranked 2$^{\text{nd}}$ in MDI+ (logistic) and 4$^{\text{th}}$ in MDI+ (ridge)) was shown to control luminal breast cancer predominantly through differential regulation of other genes including *THSD4* (Cohen et al., 2014) while *FOXC1* can counteract *GATA3* binding and impact endocrine resistance (Yu-Rice et al., 2016). Finally, *TPX2* and *TBC1D9* have been found to be over and under-expressed in triple-negative breast cancer patients respectively (Jiang et al., 2019; Kothari et al., 2020), and at the protein level, *AGR3* is over-expressed in human breast tumors (Garczyk et al., 2015).

In addition to this supporting literature evidence, we further show in Figure B.33 that the top 25 genes from MDI+ (logistic) and MDI+ (ridge) are more predictive of the PAM50 breast cancer subtypes than the top 25 genes from other feature importance measures. This top 25 from MDI+ includes genes such as *RRM2* and *SFRP1* that are not in the top 25 from any of the competing feature importance methods. However, these genes are known to play roles in breast cancer. For instance, up-regulation of *RRM2* was shown to enhance breast cancer cell proliferation (Liang et al., 2019), while loss of *SFRP1* expression is associated with breast tumor stage and poor prognosis when present in early stage breast tumors (Klopocki et al., 2004).

Stability of gene importance rankings. As in the previous CCLE case study, we evaluated the stability of the gene importance rankings across 32 train-test splits and found that MDI+ gave the most stable rankings. In particular, in Figure 3.7a, MDI+ (ridge), followed by MDI+ (logistic), yielded the smallest number of distinct features in the top 10 across the 32 train-test splits. Moreover, for the top five features, the variability of the feature importance rankings across the 32 train-test splits is much smaller for MDI+ (ridge) and MDI+ (logistic) compared to its competitors (Figure 3.7b). Given the need for reproducibility in high-stakes applications where feature importances are commonly used, this increased stability of MDI+, demonstrated in realistic settings such as the TCGA and CCLE case studies, is a vital practical advantage over other feature importance measures.

3.8 Discussion

Feature importance measures are used not only to describe how complicated ML models make predictions, but also to identify features that are most relevant to the underlying data-generating process. MDI+ provides a novel feature importance framework that generalizes the popular MDI importance measure. Specifically, MDI+ builds upon a recent interpretation of decision trees as a representation learning algorithm and connects MDI to the R^2 value from a linear regression on this learned representation. Leveraging this interpretation, MDI+ allows one to (1) replace the regression model and/or R^2 metric with other models and/or metrics that are tailored to a given data structure and (2) incorporate additional features or knowledge to mitigate the known biases of decision tree-based techniques. Fur-

thermore, MDI+ adds regularization and efficient sample-splitting to overcome known biases of MDI against correlated features and features with low entropy.

We demonstrate the utility of MDI+ via a wide array of data-inspired simulations designed to reflect a number of real-world scenarios, and via two real-world case studies on drug response prediction and breast cancer subtype prediction. Our simulation studies show that tailoring MDI+ to different problem structures results in more accurate feature rankings, compared to other popular feature importance measures. Moreover, applying MDI+ to both case studies shows the robustness of our method to different data perturbations (e.g., different train-test splits and random seeds) as well as its ability to identify predictive features. However, given the potentially high-stakes and downstream impacts of using feature importance measures in real-world scientific problems, our data-inspired simulations and real-world case studies only represent a first step in establishing the trustworthiness of MDI+, or any feature importance measure. Towards this end, developing a formal protocol and benchmarks to evaluate the effectiveness of feature importance measures will have a significant impact on the applicability of these methods in practice. One possible pathway to constructing this protocol is via the PCS framework. In particular, the PCS framework advocates for holistic evaluation of feature importances including its predictability, and stability, as we have begun to explore in our case studies. We leave it as important future work to formalize and expand these ideas.

Beyond this, MDI+ opens the door to many other natural directions for future work. First, MDI+ need not be limited to RFs and can be defined for any tree ensemble, and in particular, XGBoost (Chen and Guestrin, 2016). Second, as we illustrated in Section 3.5, exploiting the flexibility of MDI+ can substantially improve performance. Though we demonstrated this in classification and robust regression settings, there are many other contexts to consider such as multitarget prediction, survival analysis, longitudinal data analysis, etc. In addition, while we have provided practitioners with a PCS-based approach to assist practitioners with modeling choices in the MDI+ framework, further investigation is needed (e.g., different approaches to combining scores from different MDI+ models). Thirdly, the connection to linear models, and now GLMs, can also be further utilized to adapt traditional hypothesis testing tools so as to convert our importance measures into significance scores. Finally, the flexibility of RF+ as a general prediction method may be of significant independent interest. Beyond fitting GLMs on the augmented transformed dataset, RF+ allows practitioners to fit any machine learning algorithm on the augmented transformed dataset. This provides a novel avenue to incorporate the advantages of tree-based methodology and other modern machine learning tools for prediction.

Part II

Responsible data science in real-world biomedical problems applying PCS

Chapter 4

A stability-driven protocol for drug response interpretable prediction (staDRIP)

4.1 Introduction

A critical goal in precision medicine oncology revolves around predicting a patient's response to therapeutic drugs given the patient's unique molecular profile (Rubin, 2015; Kohane, 2015). Accurate personalized drug response predictions can immediately shed light on therapies that are likely to be ineffective or toxic and aid clinicians in deciding the most promising treatment for their patients (Azuaje, 2016). Moreover, interpreting these drug response prediction models can help to improve recommendations of compounds and target genes to prioritize in future preclinical research (Caponigro and Sellers, 2011).

While several community-wide, public efforts (Barretina et al., 2012; Costello et al., 2014) and many other works have made progress towards improving the predictive accuracy of drug response predictions, identifying the important disease signatures (i.e., proteins, genes, and other biomarkers) that drive the drug response prediction models has received less attention. To date, previous works have typically focused on feature selection within one specific model such as elastic nets (Jang et al., 2014; Barretina et al., 2012) and random forest (Riddick et al., 2011). However, because molecular profiling data is often heterogeneous, noisy, and high-dimensional, these results are highly sensitive to modeling decisions made by humans including the type of model, the amount of training data, and the choice of algorithm.

In this work, we focus on this goal of detecting stable, interpretable, and predictive -omic signatures that drive a cell line's drug response. To overcome the aforementioned challenges, we develop a transparent stability-driven pipeline for drug response interpretable prediction called staDRIP that is rooted in the PCS framework for veridical data science (Yu and Kumbier, 2020). At its core, the PCS framework builds its foundation on three principles: *predictability* as a reality check, *computability* as an important consideration in

algorithmic design and data collection, and *stability* as an overarching principle and minimal requirement for scientific knowledge extraction. These principles were motivated by extensive interdisciplinary research such as Wu et al. (2016), which analyzed the gap-gene network of *Drosophila*, and Basu et al. (2018), which discovered stable transcription factor interactions in *Drosophila* embryos. Since its conception, the PCS framework has further demonstrated a strong track record of driving many scientific discoveries including novel gene-gene interactions for the red-hair phenotype (Behr et al., 2020) and clinically-interpretable subgroups in a randomized drug trial (Dwivedi et al., 2020).

Here, using integrative -omics and drug response data from the Cancer Cell Line Encyclopedia (CCLE)[1] (Barretina et al., 2012), we employ the PCS framework to develop staDRIP and provide extensive documentation[2] of our modeling choices to arrive at stable biological discoveries of proteins and genes that are predictive of cancer drug responses. Unlike previous works whose results depend heavily on human decisions, staDRIP finds predictive -omic features that are stable across various models and data perturbations, thus mitigating the impact of human judgment calls. We further show that 18 of the top 24 -omic features identified by staDRIP have been previously implicated in the scientific literature, and in doing so, hint at novel candidates for future preclinical research.

4.2 Results

Prediction Accuracy

Building on the PCS framework, staDRIP first uses predictive accuracy as a reality check to filter out models that are poor fits for the observed data before turning to our primary goal of identifying important biomarkers for drug response prediction. Specifically, we define drug response as the area under the fitted dose-response curve of growth inhibition (see Section 4.4 for details). For each of the available 24 anticancer drugs, we divide the data into a 50-25-25% training-validation-test split and use the training data to fit (1) an elastic net tuned with cross-validation (CV), which has been widely used and advocated by previous studies (Barretina et al., 2012; Jang et al., 2014), (2) Lasso tuned with CV, (3) Lasso tuned with ESCV, an alternative CV metric that incorporates stability to yield more stable estimation properties with minimal loss of accuracy (Lim and Yu, 2016), (4) Gaussian kernel ridge regression, and (5) random forest to predict the drug response given the miRNA, RNASeq, methylation, and protein expression profiles separately. We also fit several data integration methods including concatenated versions of the aforementioned methods, the

[1]The CCLE is one of the most comprehensive public databases for developing detailed genetic and pharmacologic characterizations of human cancer cell lines. After preprocessing (see Section 4.4), we arrive at a panel of 370 human cancer cell lines that have both high-throughput molecular profiling of RNASeq gene expression (5000 genes), microRNA expression (734 miRNAs), DNA methylation (4000 transcription start sites), and protein expression (214 proteins) as well as pharmacological data for 24 anticancer drugs.

[2]PCS documentation can be found at `https://github.com/Yu-Group/staDRIP`.

recently proposed X-shaped Variational Autoencoder (X-VAE) (Simidjievski et al., 2019), and the winner of the DREAM 7 challenge,[3] the Bayesian multi-task multiple kernel learning method (BMTMKL) (Costello et al., 2014).

For each of these fits, we report in Table C.1 the average validation accuracy across all 24 drugs as measured by the R^2 value and the WPC-index, a weighted probabilistic concordance index, which has been used in previous studies and measures how well the predicted rankings agree with the true responses (Costello et al., 2014). From Table C.1, we see that kernel ridge regression trained only on the RNASeq data yields the best predictive performance. However, considering that our primary goal is not purely prediction, the differences between model prediction accuracies shown in Table C.1 are relatively small from a practical viewpoint. In our inferential procedure discussed next, we will see that leveraging the stability across these methods with similar predictive accuracies is key to our staDRIP pipeline for identifying genes and proteins that are stable predictive features underlying the drug response models.

Nonetheless, for completeness, we report the test accuracy from the best model, the RNASeq-based kernel ridge regression, to have an R^2 ($\pm$1SD) of 0.204 ($\pm$ 0.038) and WPC-index of 0.620 ($\pm$ 0.0075) across the 24 drugs. For brevity, we leave more detailed individual drug-level prediction results to Appendix C.

Identifying predictive -omic features with PCS inference

Beyond predictability, the PCS framework emphasizes stability throughout the data science life cycle so as to reduce reliance on particular human judgment calls. Accordingly, we leverage and quantify the stability of important features under numerous data and model perturbations in staDRIP as follows: for each of the 24 drugs separately,

1. **Use predictability as reality check:** select a set $\mathcal{M}$ of models with high predictive accuracy across a variety of metrics on the validation data.
2. **Compute stability of predictive features across data perturbations:** for each model $M \in \mathcal{M}$, refit the model M to B bootstrap replicates of the data, and compute the stability score of each feature as the proportion of B bootstrap samples where the feature is selected (details in Section 4.4). Let F_M denote the subset of features with high stability scores (e.g., top 10).
3. **Select predictive features that are stable across model perturbations:** take the intersection $\cap_{M \in \mathcal{M}} F_M$ as the stable predictive -omic features across data and model perturbations.

In our work, we are primarily interested in identifying proteins and genes that are predictive of drug responses as many drugs are directly related to known proteins and genes. Hence, considering the five models trained on the RNAseq and protein data separately, we take $\mathcal{M} = \{\text{RF}, \text{Lasso (ESCV)}, \text{Elastic Net}\}$. Note that while kernel ridge has the highest

[3]The DREAM 7 challenge was a public competition where teams were tasked to integrate multiple –omics measurements and predict drug sensitivity in cancer cell lines.

accuracy, it is omitted from $\mathcal{M}$ since there is no straightforward, computationally efficient method to select features from kernel ridge to the best of our knowledge. We also omit the Lasso from $\mathcal{M}$ as it generally has the worst predictive accuracy. For each remaining model in $\mathcal{M}$, we then take F_M to be the 10 features with the highest stability scores and list those genes and proteins in the top 10 most stable features across all three models in Table C.4 in Appendix C.

In Table 4.1, we provide our main evidence for the utility of staDRIP, listing the single most stable protein for each drug along with independent publications that support these findings. Specifically, of the 24 proteins identified as most stable by staDRIP, 18 have been associated with the drug sensitivity or identified as a known or possible drug target in prior preclinical studies. See Section 4.4 for details of this literature evidence.

Now in contrast to staDRIP, which finds stable predictive features across models with similar predictive accuracies, previous state-of-the-art methods (Barretina et al., 2012; Jang et al., 2014) use only an elastic net to identify predictive -omics features of drug responses. To compare staDRIP to this elastic net approach, we extract the proteins with the highest stability score for each drug when taking $\mathcal{M} = \{\text{Elastic Net}\}$. Repeating the same literature search procedure as we did for the proteins identified by staDRIP, we found only 14 of the 24 proteins identified by the elastic net are known from previous clinical studies (see Table C.5 in Appendix C). Detailed comparisons of the results of our method, staDRIP, and that of the elastic net can be found in Section 4.4.

Table 4.1: Most stable protein associated with each drug, as identified by staDRIP, along with literature that supports the association between the protein and drug sensitivity.

Drug	**Protein**	**Supporting Literature**	**Drug**	**Protein**	**Supporting Literature**
17-AAG	Bax	He et al. (2013)	PD-0332991	Bcl-2	Chen and Pan (2017)
AEW541	Akt	Attias-Geva et al. (2011)	PF2341066	c-Met	Camidge et al. (2014)
AZD0530	p38	Yang et al. (2010)	PHA-665752	MEK1	–
AZD6244	PI3K-p85	Balmanno et al. (2009)	PLX4720	MEK1	Emery et al. (2009)
Erlotinib	EGFR	McDermott et al. (2007)	Paclitaxel	Src	Le and Bast (2011)
Irinotecan	MDMX_MDM4	Ling et al. (2014)	Panobinostat	VEGFR2	Strickler et al. (2012)
L-685458	YAP	–	RAF265	PI3K-p85	Mordant et al. (2010)
LBW242	ASNS	–	Sorafenib	Bcl-2	Tutusaus et al. (2018)
Lapatinib	HER2	Esteva et al. (2010)	TAE684	PTEN	–
Nilotinib	STAT5	Warsch et al. (2011)	TKI258	CD49b	–
Nutlin.3	Bcl-2	Drakos et al. (2011)	Topotecan	–	–
PD-0325901	MEK1	Henderson et al. (2010)	ZD-6474	c-Kit	Yang et al. (2006)

4.3 Discussion

Rooted by the PCS framework, we emphasize the importance of predictability, (computability), and stability as minimum requirements for extracting scientific knowledge throughout

the staDRIP pipeline. We show that, guided by good prediction performance, incorporating a number of stability checks and extracting the stable parts of top-performing models can help to avoid the poor generalization exhibited by existing methods and can successfully identify candidate therapeutic targets for future preclinical research. We also acknowledge that while many stability considerations are built into staDRIP, there are inevitably human judgment calls that still impact our analysis. For example, we make a number of judgement calls in the data preprocessing stage, which we detail in Section 4.4. Additionally, many other reasonable models such as ridge regression and gradient boosting could be considered in the staDRIP pipeline. We thus provide transparent and extensive documentation at `https://github.com/Yu-Group/staDRIP` to justify these decisions using domain knowledge when possible.

4.4 Methods

Data

To begin building the personalized drug response models, we leverage data from a panel of 397 human cancer cell lines that have both high-throughput molecular profiling and pharmacological data for 24 anticancer drugs from the Cancer Cell Line Encyclopedia (CCLE) project (Barretina et al., 2012). Specifically, -omics data from the CCLE was downloaded from DepMap Public 18Q3 (`https://depmap.org/portal/download/`). These cell lines encompass 23 different tumor sites and have been profiled for gene expression, microRNA expression, DNA methylation, and protein expression. Note that though the CCLE contains data from 947 cell lines, only 397 of these cell lines had data from all four molecular profiles of interest and pharmacological profiling.

In addition to these molecular profiles, we obtained pharmacological profiling of 24 chemotherapy and target therapy drugs from the CCLE (Barretina et al., 2012). For each cell line-drug combination, the CCLE incorporated a systematic framework to measure molecular correlates of pharmacological sensitivity in vitro across eight dosages. We refer to Barretina et al. (2012) for details on this procedure, but given the fitted dose-response curves of growth inhibition from these experiments, we took the activity area, or AUC, to be the primary response of interest in this work. The AUC is defined as the area between the response curve and 0 (i.e., the no response reference level) and is a well-accepted measure of drug sensitivity (Jang et al., 2014; Barretina et al., 2012). In this case, the AUC is measured on an 8-point scale with 0 corresponding to an inactive compound and 8 corresponding to a compound with 100% inhibition at all 8 dosages.

In Figure 4.1, we provide a graphical summary of the raw molecular and pharmacological profiling data sets.

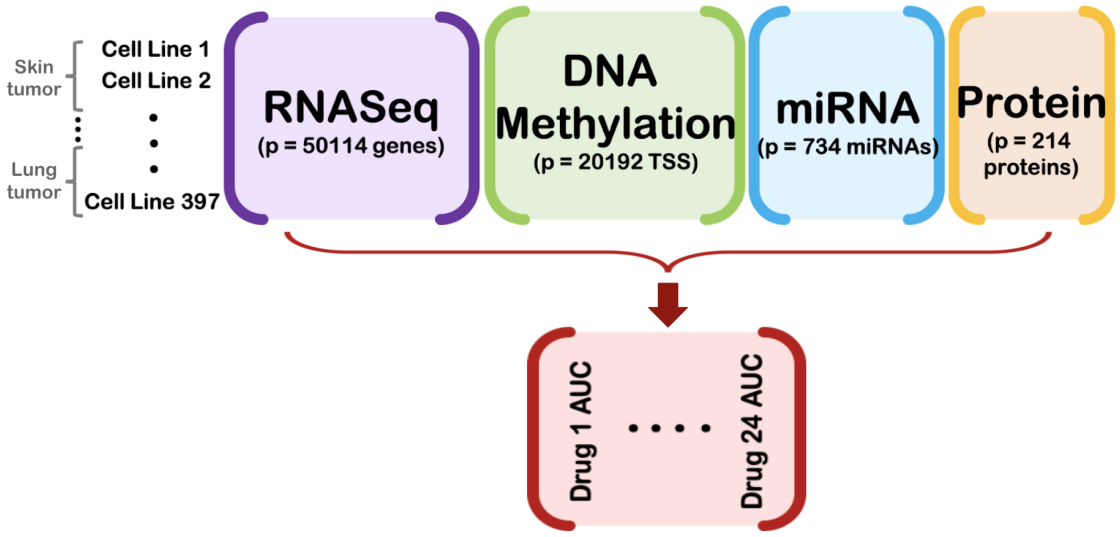


Figure 4.1: A graphical overview of the raw CCLE molecular profiling data sets, which are used to predict the drug responses of 24 therapeutic drugs, as measured via the drug response AUC.

Data preprocessing

Given the raw data described above, there are a couple areas of initial concern that warrant preprocessing. First, the cancer cell lines encompass 23 different tumor sites, and cell lines from the same tumor site tend to have more similar expression profiles than cell lines from different sites. To illustrate, we observe clusters of cell lines by tumor site when performing both hierarchical clustering and PCA on the RNASeq profile in Figure C.1. Due to these inherent differences between tumors, we chose to omit the cell lines from tumor sites with < 8 cell lines. This reduces our sample size to 370 cell lines from 16 tumor sites. Here, we chose the threshold 8 to ensure we have at least 2 cell lines from each tumor site in each of the training, validation, and test splits (using a 50-25-25% partitioning scheme).

In addition to reducing the number of samples in our analysis, we reduced the number of features to more manageable sizes before continuing with our analyses. Originally, the molecular profiling data consisted of 734 miRNAs, 50114 genes, 20192 TSS, and 214 proteins. With only 370 cell lines, we aggressively preprocessed the number of genes and TSS by taking the top 10% of genes (or 5000 genes) and top 20% of TSS (or 4000 TSS) with the highest variance. We also transformed the miRNA and RNASeq expression values using the log-transformation log(x+1) in order to mitigate potential problems with highly skewed positive count values.

We recognize however that there were many other reasonable ways to preprocess this data. For instance, we could have taken the top 20% of genes and top 40% of TSS with the highest variance. Another common alternative would have been to filter features using

marginal correlations with the response or using a multivariate prediction model (e.g., the Lasso). To assess robustness to these choices, we reran our prediction analysis using these alternative preprocessing procedures and saw that the prediction accuracies are higher using the variance-filtering preprocessing pipelines, as compared to the correlation-filtering and Lasso-filtering pipelines (see PCS documentation at `https://github.com/Yu-Group/staDRIP`). Further, the smaller variance-filtered model gives similar prediction accuracies as the larger variance-filtered model. Thus, for simplicity moving forward, we use and focus primarily on the initially proposed variance-filtering procedure as it is less computationally expensive than the model with twice as many features and maintains similarly high accuracy.

To summarize, after this preprocessing, we have 370 cell lines with data across the four molecular profiles of interest with 734 miRNAs (log-transformed), 5000 genes (log-transformed), 4000 TSS, and 214 proteins and pharmacological data, measured via the AUC drug response scores, for 24 anticancer compounds. We provide a visual summary of the preprocessed data and plot the overall distribution of features in the four molecular profiles as well as the distribution of the 24 drug responses in Figure C.2.

Prediction Models

Like in data preprocessing, human judgment calls play a significant role in the modeling stage, including the decision of which methods to fit. Ideally, the chosen methods should have some justified connection to the biological problem at hand, but in our case, it is unclear which models or assumptions best fit the biological drug response mechanism a priori. Nevertheless, we have reasons to believe that the Lasso, elastic net, RF, and kernel ridge regression are particularly appealing fits for this problem.

First, the Lasso assumes a sparse linear model, meaning that the effect of each feature is additive and only a sparse number of the features contribute to the drug sensitivity. The simplicity and interpretability of the Lasso makes it a popular tool for bioinformatics prediction tasks, so we choose to use the Lasso as a baseline model for our analysis. The elastic net is perhaps even more popular than the Lasso in drug response prediction studies (Jang et al., 2014; Barretina et al., 2012). Similar to the Lasso, the elastic net assumes linearity and some sparsity but is also able to better handle correlated features. Beyond linearity, kernel ridge regression with a Gaussian kernel allows for more flexible, but less interpretable, functional relationships that are not necessarily linear. Kernel methods have been applied in previous case studies with great success (Costello et al., 2014) and are hence promising candidates for our study as well. Lastly, random forest can be viewed as a collection of localized, non-linear thresholded decisions rules (like on-off switches), which are well-suited for many biological processes that match the combinatorial thresholding (or switch-like) behavior of decision trees (Nelson et al., 2008). Random forests are also invariant to the scale of the data. This is especially advantageous for integrating different data sets with varying scales and domain types (e.g., count-valued RNASeq expression, proportion-valued methylation data, continuous-valued protein expression).

In addition to fitting the aforementioned methods on each of the molecular profiles separately, we also tried fitting various data integration methods since incorporating multiple sources of -omics data can sometimes result in more accurate predictions than models built using only a single -omics sources (Costello et al., 2014; Güvenç Paltun et al., 2019; Simidjievski et al., 2019). The most natural integration idea is to concatenate the -omics data sets together and to fit a single model (e.g., the elastic net) on the concatenated data. When fitting models like the Lasso, elastic net, and kernel ridge regression which are not scale-invariant, the molecular profiles are scaled to have columns with mean 0 and variance 1 to allow for fair comparisons between molecular types. We refer to this method as the concatenated data approach and use this as a baseline for evaluating data integration methods. More sophisticated methodology has also been proposed to integrate -omics data, including recent work using the X-VAE, a variational autoencoder for cancer data integration (Simidjievski et al., 2019), and the BMTMKL, a Bayesian multitask multiple kernel learning method which won the NCI-DREAM 7 challenge (Costello et al., 2014).

Note that though an alternative approach would have been to develop new methodology, we instead leverage these existing machine learning methods that have been rigorously vetted and have been shown to work well in many related problems. In fact, by examining the stable properties across these existing methods, we obtain high-quality scientific findings, as made evident by the abundance of supporting literature (see Table 4.1).

Model hyperparameters

To select hyperparameters in each of these methods, we use 5-fold cross validation, where the folds are stratified by tumor type. We also investigate using the estimation stability cross validation (ESCV) metric for selecting the Lasso's hyperparameter. This ESCV metric combines a stability measure within the cross-validation framework to yield more stable estimation properties with minimal loss of accuracy when using the Lasso (Lim and Yu, 2016).

For the X-VAE model, we adapt an X-shaped network architecture to train a variation autoencoder that learns joint representation of the RNAseq and protein data. In particular, we take the 2,000 RNAseq features with highest variance, since the number of cell lines is too small compared with the original number of RNAseq features. In our experiment, both the encoder and the encoder have one hidden layer. There are 128 neurons corresponding to the RNAseq protein in the hidden layer of the encoder and the decoder, and 32 neurons corresponding to the protein features. The latent representation has a dimension of 32. The dimension of the hidden layers and the latent representation are based on the recommendation of (Simidjievski et al., 2019), and are not tuned. We used ELU activation and employed batch normalization and a dropout component with rate 0.2, as recommended by (Simidjievski et al., 2019). The models were trained for 500 epochs using an Adam optimizer with a learning rate of 0.001.

Evaluation metrics

We primarily consider two evaluation metrics for prediction accuracy as each captures a different aspect of prediction: (1) R^2 value and (2) probabilistic concordance-index (PC-index). R^2 is defined as $1 - \frac{\text{MSE}(Y,\hat{Y})}{\text{Var}(Y)}$, where $\text{Var}(Y)$ denotes the variance of the observed responses, and $\text{MSE}(Y, \hat{Y})$ denotes the mean sum of squared errors between the predicted responses $\hat{Y}$ and observed responses Y. R^2 is a rescaling of the MSE that accounts for the amount of variation in the observed response and thus allows us to easily compare accuracies between drug response models with different amounts of variation in the observed response, but as with the MSE, R^2 can be heavily influenced by outliers. PC-index is a measure of how well the predicted rankings agree with the true responses. This metric takes into account the variance of the drug responses but it also assumes that the drug responses follow a Gaussian distribution, which may not be true in some cases. We consider this metric because it is the primary method of evaluation in the NCI-DREAM 7 competition (Costello et al., 2014). Given the large scale and breadth of this challenge, we compare our results to this work. For further details on the PC-index, we refer to Costello et al. (2014).

In each of the evaluation metrics above, we receive a separate score for each of the 24 drug response models. It may also be beneficial to aggregate the 24 scores into a single number for concrete evaluation. In particular, Costello et al. (2014) used a weighted average of the PC-indices to compare various models and referred to this evaluation metric as the weighted PC-index (WPC-index). To compare our results with the benchmark in Costello et al. (2014), we also consider the WPC-index in evaluating our models.

PCS Inference

Detailed description on the computation of stability scores

We next describe in detail how to compute the stability scores in the PCS-driven disease signature identification pipeline. Note that the following procedure is repeated for each of the 24 drugs.

We randomly draw $B = 100$ bootstrap samples $\mathcal{D}^{(b)}, b = 1, \ldots, 100$ from the training data. Then, for each bootstrap sample, we fit (1) an elastic net with tuning parameter selected by CV (2) a Lasso with tuning parameter selected by ESCV and (3) a random forest. Each model is fitted using the protein and RNAseq data separately since the integration approaches did not improve the prediction accuracy over simply using the RNASeq data only (see Table C.1). Next, for each feature X_j from either the protein or RNAseq data set, let $\omega_j^{(b)}$ be defined in the following way: for the Lasso and elastic net, $\omega_j^{(b)} = 1$ if the coefficient of X_j is non-zero, and $\omega_j^{(b)} = 0$ otherwise; for the random forest, $\omega_j^{(b)}$ is the MDI feature importance of X_j. We then define the stability score $\text{sta}(X_j)$ of each feature X_j as $\text{sta}(X_j) = \frac{1}{B}\sum_{b=1}^{B} \omega_j^{(b)}$ and rank the proteins and genes separately by the stability scores of the features.

Discussion on the disease signatures identified by the PCS pipeline

In Table C.4, we list the proteins and genes which we found to be stable and among the top 10 features for all three methods. Among these stable features, we list them in decreasing order by the sum of stability score rankings. Though we identify fewer stable genes, this is most likely due to two reasons. First, there are 5000 genes in the model, compared to only 214 proteins, so thresholding at the top 10 genes is extremely conservative. Secondly, the average correlation between genes is higher than that between proteins, adding to the instability.

With regards to the identified protein signatures, we can roughly classify them into three categories. The first category contains those that are known targets of the corresponding target therapy drugs. For example, Erlotinib is a medication used to treat non-small cell lung cancer (NSCLC) and pancreatic cancer. It is an EGFR inhibitor and is specifically used for NSCLC patients with tumors positive for EGFR exon 19 deletions (del19) or exon 21 (L858R) substitution mutations. Correspondingly, EGFR is ranked in the top ten stable proteins in all three models. Other such examples include the drug Lapatinib and its target HER2, PD-0325901 and its target MEK, PHA-665752 and its target c-Met, and ZD-6474 and its target c-Kit.

The second category contains those that are not known to be direct targets of the drug but have been shown in preclinical studies to be potential therapeutic targets or are associated with drug resistance. For example, Ling et al. (2014) identified a potential application of the drug Irinotecan as an MdmX inhibitor for targeted therapies, and in our pipeline, MdmX had the highest stability score for all three models. As another instance, we identified MEK1 as a top protein signature, ranked by stability score, for the drug PLX4720 while Emery et al. (2009) showed that MEK1 mutations confer resistance to PLX4720.

The third category are proteins that do not belong to the two categories above. Still, these biomarkers are predictive of the drug response under various model and data perturbations. Given the evidence in the scientific literature that supports many of our identified features, the proteins in this category may be potential candidates for future preclinical investigation.

Among the list of overlapping stable features in Table C.4, we list in Table 4.1 the one with the highest stability score ranking along with recent biomedical publications, supporting the association between the protein and the drug. The procedure of this literature search is as follows: we first searched for papers where the protein and the drug co-occurs. Then for each paper, we read the introduction section to understand their main conclusions. Each of the 18 papers listed in Table 4.1 includes sentences such as "our findings suggest that the over-expression of this protein will increase drug sensitivity/resistance" or "this protein is a potential (or known) therapeutic target for the drug". Out of the 24 predictive protein signatures that we identify as most stable, 18 of them have existing preclinical studies that confirm the effectiveness of our stability analysis.

In Table C.5, we list the protein with the highest stability score when fitting an elastic net to 100 bootstrap samples of the training data for each of the 24 drugs. This approach of finding predictive -omics features was previously used in (Barretina et al., 2012; Jang et al.,

2014). Compared with staDRIP, which searches for stable features across different models, this approach only uses a single model (i.e, an elastic net) for feature selection. We repeat the same literature search procedure as we did for our findings and found that among the 24 most stable protein features identified by elastic net, only 14 are known from previous clinical studies. For 10 drugs, the most stable protein from elastic net and that from staDRIP is the same, and among these 10 proteins, 9 are implicated in the existing literature. For the other 14 drugs, 5 proteins identified by elastic net are implicated in the existing literature, while 9 protein features identified by staDRIP are implicated in the existing literature.

Chapter 5

Contribution of the microbiome to a metabolomic signature predictive of risk for pancreatic cancer

5.1 Introduction

Pancreatic cancer is highly lethal and is projected to become the second-leading cause of death in the United States by 2040 (Rahib et al., 2021). Surgical resection of localized disease represents the greatest chance for curative therapy. Unfortunately, only a minority ($15-20\%$) of patients present with surgically resectable disease (Kleeff et al., 2016; Ryan et al., 2014).

The low incidence of pancreatic cancer in the average-risk population (approximately $8-12$ per $100,000$) (Bray et al., 2018; Rawla et al., 2019) makes it challenging to implement effective screening programs for pancreatic cancer. The United States Preventative Services Task Force (USPSTF) currently recommends against screening for pancreatic cancer in the general population using any method (Force, 2019). Yet, the USPSTF recognizes that screening in persons who are at an increased risk may be warranted (Force, 2019). There remains an opportunity to develop blood-based signatures that can identify individuals at increased risk who would benefit from screening and potentially from preventive interventions.

The microbiota is a complex ecosystem integral to human health. Microbial diversity is site-specific and varies depending on the organ location (Consortium, 2012). Increasing evidence suggests that alterations in the microbiome are associated with risk for certain cancers, including pancreatic cancer (Huybrechts et al., 2020). Studies suggest that loss of microbial diversity and community stability coupled with increases in pathogenic microbes increase cancer susceptibility (Bhatt et al., 2017). In the context of pancreatic cancer, the composition of the microbiome has been linked to alterations in the local microenvironment and to promotion of oncogenesis through immune suppression (Barbour et al., 1997; Pushalkar et al., 2018; Riquelme et al., 2019) with implications for response to therapy and survival

(Zhang et al., 2020).

Microbiome colonization has been associated with metabolic changes that can perpetuate inflammation and increase an individual's risk of developing cancer (Brennan and Garrett, 2016; Consortium, 2012; Risch, 2012; Wei et al., 2019). Microbiome-related metabolites include short-chain fatty acids, butyrate and acetate, secondary bile acids, indole-derivatives, cadaverine, trimethylamine N-oxide (TMAO), and lipopolysaccharides (Kiss et al., 2020). A study of serum methionine-related metabolites identified elevated serum levels of TMAO, a gut microbiota-derived metabolite (Yang et al., 2019), as associated with pancreatic cancer (Huang et al., 2020; Mayers et al., 2014). Other metabolites consisting of indoleacrylic acid and indole-3-acetate have been shown to differentiate newly-diagnosed pancreatic cancer cases from controls (Xie et al., 2015).

We designed our study to quantify the extent to which microbiome-related and other metabolites in circulation are elevated among subjects that were subsequently diagnosed with pancreatic cancer using sera collected from participants in the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Using a training and testing approach, we established a microbial-related metabolite panel for 5-year risk assessment of pancreatic cancer. The performance of the microbiome metabolite panel for risk prediction of pancreatic cancer was further evaluated in an independent cohort of newly-diagnosed pancreatic cancer patients compared to non-cancer controls. The complementary value of other non-microbial-related metabolites as well as CA19-9 was also determined.

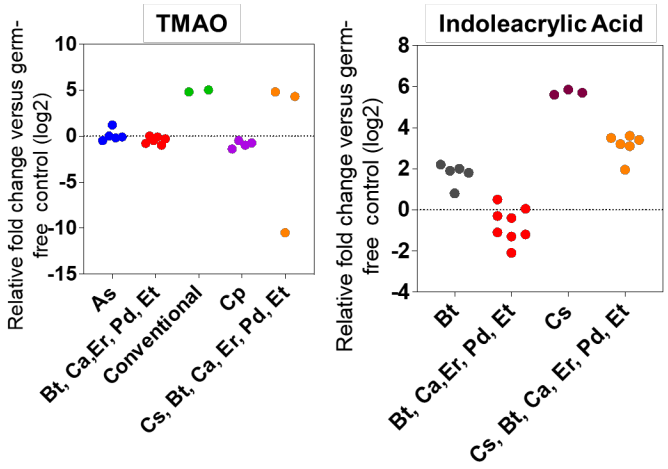
5.2 Results

Quantification of microbial-related metabolites

Using untargeted metabolomics, we screened for microbial-derived metabolites in sera from 172 cases diagnosed within 5 years of blood draw and 863 non-case participants from the PLCO screening trial (Table 5.4). A total of 14 microbial-related metabolites were detected and quantified across all specimens, including 9 indole-derivatives (Jaglin et al., 2018; Lee and Lee, 2010), two secondary bile acids (Hylemon et al., 2018; Ridlon et al., 2014), 5-hydroxy-tryptophan (Stoll et al., 2016), acetylcadaverine (Pugin et al., 2017), and TMAO (Brunt et al., 2021; Xu et al., 2015). Of the 14 metabolites, indoleacrylic acid, TMAO, and indole-derivative_2 had adjusted odds ratios (ORs) per unit standard deviation (SD) increase ≥ 1.2 for risk of pancreatic cancer (Supplementary Figure D.2; Supplementary Figure D.3). Elevated levels of TMAO and indoleacrylic acid have been associated with phyla of Bacillota, Bacteroidota, Actinomycetota, and Pseudomonadota (species of Clostridium sporogenes (Cs), Eubacterium rectale (Er), Bacteroides thetaiotaomicron (Bt), Parabacteroides distasonis (Pd), Collinsella aerofaciens (Ca), and Edwardsiella tarda (Et)) (Han et al., 2021), all of which have relevance to pancreatic cancer (Figure 5.1A-B) (Half et al., 2019; Kamiyama et al., 2019; Matsukawa et al., 2021; Zhou et al., 2021).

A

Annotation	Complete name
As	Anaerostipes sp
Cp	Citrobacter portucalensis
Cs	Clostridium sporogenes
Er	Eubacterium rectale
Bt	Bacteroides thetaiotaomicron
Pd	parabacteroides distasonis
Ca	collinsella aerofaciens
Et	Edwardsiella tarda



B

Species	Phylum	Class	Order	Associated with PDAC	Reference
Clostridium sporogenes (Cs)	Bacillota	Clostridia	Clostridiales	Increased in fecal samples of PDAC patients compared to controls	Zhou et al., 2021
Eubacterium rectale (Er)	Bacillota	Clostridia	Clostridiales	Decreased in fecal samples of PDAC patients compared to controls	Zhou et al., 2021
Bacteroides thetaiotaomicron (Bt)	Bacteroidota	Bacteroidia	Bacteroidales	Bacteriodales species elevated in fecal samples of PDAC patients compared to controls	Half et al., 2019 Matsukawa et al., 2021)
Parabacteroides distasonis (Pd)	Bacteroidota	Bacteroidia	Bacteroidales	Bacteriodales species elevated in fecal samples of PDAC patients compared to controls	Half et al., 2019 Matsukawa et al., 2021
Collinsella aerofaciens (Ca)	Actinomycetota	Coriobacteriia	Coriobacteriales	Relative abundances in fecal samples are associated with poor prognosis among PDAC patients	Matsukawa et al., 2021
Edwardsiella tarda (Et)	Pseudomonadota	Gammaproteobacteria	Enterobacterales	Associated with hepatobiliary disease including pancreatic cancer	Kamiyama et al., 2019

Figure 5.1: Relationship between TMAO and indoleacrylic acid and microbial species. (A) Association between TMAO and indoleacrlyic acid with different microbial species. Data was derived from the Metabolomics Data Explorer database (see Methods) (Han et al., 2021). (B) Association between referenced microbial species and pancreatic cancer.

Model	Hyperparameters	AUC (95% CI)	Adj OR‡
Logistic regression	–	0.57 (0.46-0.67)	1.30 (0.85-2.02)
Logistic regression with ridge (L_2) regularization	Penalty weight = 0.22	0.58 (0.48-0.68)	1.32 (0.87-2.05)
Logistic regression with LASSO (L_1) regularization	Penalty weight = 0.023, number of selected features = 3	0.64 (0.54-0.73)	1.42 (0.94-2.13)
Iterative Random Forest	Number of iterations = 4	0.52 (0.41-0.62)	1.28 (0.80-1.77)
Deep neural network model	Number of cross-validation folds = 4, hidden layers = 2 with 64 nodes in each layer	0.55 (0.45-0.65)	1.17 (0.75-1.80)
GBM	Number of trees = 36, max depth= 6	0.53 (0.41-0.65)	1.12 (0.76-1.58)
Auto ML	Selected model = randomized trees	0.57 (0.45-0.68)	1.04 (0.64-1.63)

Table 5.1: Performance of microbial-related metabolites panels in different learning models in the PLCO Validation Set. ‡Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs).

Model building and testing of microbial-related metabolite panel

To establish a combination rule, all 14 microbial-related metabolites were considered. Seven different models were trained and optimized in the Development Set (Figure 5.4). LASSO regression with three selected features achieved the highest prediction performance amongst all models in the Validation Set, yielding an AUC of 0.64 (95% CI: $0.54 - 0.73$) and an adjusted OR of 1.42 (95% CI $0.94 - 2.13$) per unit SD increase for 5-year probability of pancreatic cancer (Table 5.1; Supplementary Table D.2). To verify the reproducibility of our finding, we adhered to the predictability, computability and stability (PCS) framework (Yu and Kumbier, 2020) and stress-tested the 3-marker microbial panel to assure its reliability. Stable performance in terms of AUC and adjusted odds ratio across various data perturbations and stability checks demonstrated robustness of the 3-marker microbial panel (Supplementary Table D.3).

Performance of the 3-marker microbial panel in the Test Set

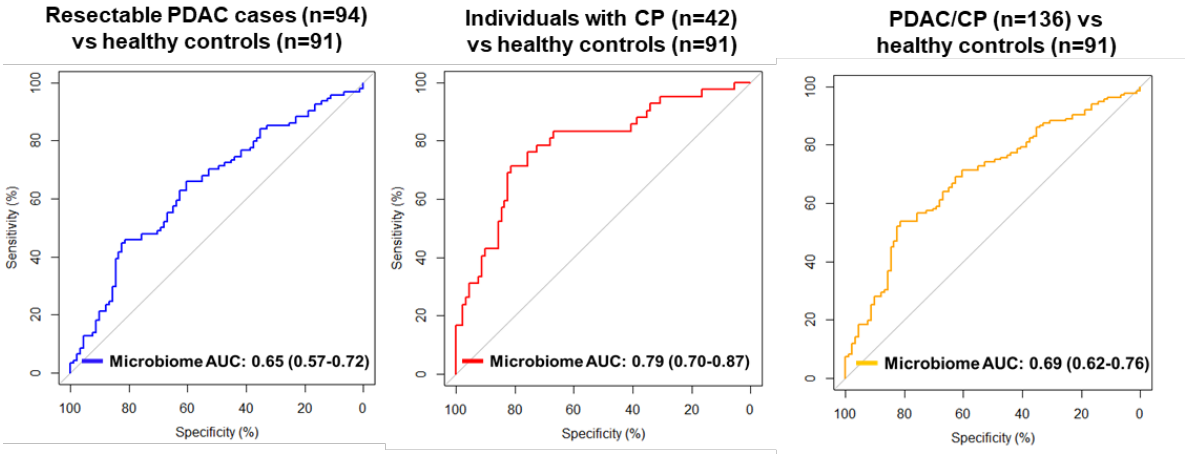
In the Test Set, the 3-marker microbial panel yielded an AUC of 0.64 (95% CI: 0.53-0.76) and an adjusted OR of 1.72 (95% CI: 1.25-2.37) per unit SD increase for 5-year probability of pancreatic cancer (Table 5.2). When considering cases diagnosed within 2 years of blood draw, the 3-marker microbial panel yielded an AUC of 0.61 (95% CI: 0.48-0.74) and an adjusted OR of 1.43 (95% CI: 0.98 – 2.03) per unit SD increase for risk prediction of pancreatic cancer. (Table 5.2). Prediction performance of the 3-marker microbial panel for risk assessment of pancreatic cancer was similar among diabetic and non-diabetic individuals (Supplementary Table D.5).

Set-aside Test Set								
			3-marker microbial panel			**3-marker microbial panel + 5-marker non-microbial panel**		
Time to Dx	Cases N	Non-cases N	AUC (95% CI)	Adj. OR† (95% CI)	P-value	AUC (95% CI)	Adj. OR† (95% CI)	P-value
[0-5)	37	225	0.64 (0.53 - 0.76)	1.72 (1.25 - 2.37)	<0.001	0.79 (0.71-0.88)	3.13 (2.08-4.98)	<0.001
[0-2)	24	225	0.61 (0.48 - 0.74)	1.43 (0.98 - 2.03)	0.04	0.82 (0.72-0.93)	3.8 (2.33-6.74)	<0.001
[2-5)	13	225	0.70 (0.50 - 0.90)	2.11 (1.33 - 3.43)	<0.001	0.74 (0.60-0.86)	1.90 (1.08-3.37)	0.02
Entire Set								
			3-marker microbial panel			**3-marker microbial panel + 5-marker non-microbial panel**		
Time to Dx	Cases N	Non-cases N	AUC (95% CI)	Adj. OR† (95% CI)	P-value	AUC (95% CI)	Adj. OR† (95% CI)	P-value
[0-5)	172	861	0.62 (0.57 - 0.67)	1.50 (1.28 - 1.76)	<0.001	0.76 (0.72-0.80)	2.75 (2.25-3.38)	<0.001
[0-2)	92	861	0.60 (0.54 - 0.67)	1.43 (1.18 - 1.74)	<0.001	0.81 (0.76-0.86)	3.66 (2.81-4.84)	<0.001
[2-5)	80	861	0.64 (0.57 - 0.70)	1.53 (1.28 - 1.87)	<0.001	0.69 (0.63-0.75)	1.92 (1.51-2.44)	0.02

Table 5.2: Performance estimates of the 3-marker microbial panel and a combined 3-marker microbial panel + 5-marker non-microbial panel for 5-year risk prediction of pancreatic cancer in the set-aside Test Set and the entire PLCO specimen set. †Age, sex, bmi and smoking status were included as co-variables in adjusted odd ratios.

Prediction performance of the 3-marker microbial panel in an independent newly-diagnosed PDAC cohort

We further assessed the prediction performance of the 3-marker microbial panel in an independent set of samples from 99 newly diagnosed, resectable PDAC cases, 50 patients with CP, and 100 healthy controls. Compared to healthy controls, the 3-marker microbial panel had an OR of 1.55 (95% CI: 1.13-2.23) per unit SD increase for probability of pancreatic cancer and an OR of 2.07 (95% CI: 1.45-3.18) for pancreatic disease (cancer or chronic pancreatitis) (Figure 5.2).



Odds Ratio (95% CI)		
Resectable PDAC cases vs healthy controls	Individuals with CP vs healthy controls	PDAC/CP vs healthy controls
1.55 (1.13-2.23)	2.83 (1.83-4.82)	2.07 (1.45-3.18)

Figure 5.2: Predictive performance of the 3-marker microbial panel in the independent newly-diagnosed PDAC cohort. Abbreviation: CP- chronic pancreatitis. A subset samples were excluded due to insufficient sample volume or not having passed quality control criteria.

Contributions of non-microbial metabolites for improved risk prediction of pancreatic cancer

We assessed the contribution of non-microbial metabolites for pancreatic cancer risk assessment. A total of 1,009 non-microbial metabolites were quantified in the PLCO specimen set. Five non-microbial metabolites (cholesterol glucuronide, 2-hydroxyglutarate, galactosamine, glucose, and erythritol) exhibited statistically significant adjusted OR in the Development Set (Supplementary Table D.4). We subsequently applied the PCS framework to develop and stress-test a model based on the five non-microbial metabolites. A logistic regression model was selected based on exhibiting the highest predictive performance in the Validation Set, with a resultant AUC of 0.72 (95% CI: 0.65-0.97) and an adjusted OR of 2.10 (95% CI: 1.04-2.80) for 5-year risk prediction of pancreatic cancer (Supplementary Table D.6-D.7). In the set-aside Test Set, the 5-marker non-microbial panel yielded an AUC of 0.74 (95% CI: 0.65-0.83) and an adj OR of 2.72 (95% CI: 1.83-4.24) for 5-year risk prediction of pancreatic cancer.

To assess the contributions of the 3-marker microbial panel and the 5-marker non-microbial panel, we fitted a logistic regression with the 3-marker microbial panel scores and the 5-marker non-microbial panel scores as two separate predictors. The combined metabolite panel yielded an AUC of 0.79 (95% CI: 0.71-0.88) and an adj OR of 3.13 (95% CI:

2.08-4.98) per unit SD increase for 5-year probability of pancreatic cancer in the set-aside Test Set (Table 5.2, Supplementary Table D.8-D.9). When considering cases diagnosed within 0-2 years and 2-5 years of blood draw, the combined metabolite panel had respective AUCs of 0.82 (95% CI: 0.72-0.93) and 0.74 (95% CI: 0.60-0.86) (Table 5.2, Supplementary Table D.8-D.9).

Contribution of the combined metabolite panel with CA19-9 for pancreatic cancer risk assessment

We previously demonstrated that levels of CA19-9 were increased in PDAC cases in the PLCO cohort with an exponential rise starting two years prior to diagnosis (Fahrmann et al., 2021). We therefore assessed whether the combined metabolite panel (3-marker microbial panel plus the 5-marker non-microbial panel) would be complementary with CA19-9 for risk prediction of pancreatic cancer. In the set-aside Test Set, the combined metabolite panel + CA19-9 had an AUC of 0.84 (95% CI: 0.76-0.91) and an adj OR of 9.67 (95% CI: 4.56-23.30) per unit SD increase for 5-year probability of pancreatic cancer (Table 5.3). For cases diagnosed within 2 years after blood draw, the combined metabolite panel + CA19-9 yielded an AUC of 0.86 (95% CI: 0.77-0.95), which was markedly improved compared to CA19-9 alone (AUC: 0.70 (0.57-0.82), comparison of AUCs p-value: 0.006) (Table 5.3, Supplementary Table D.10).

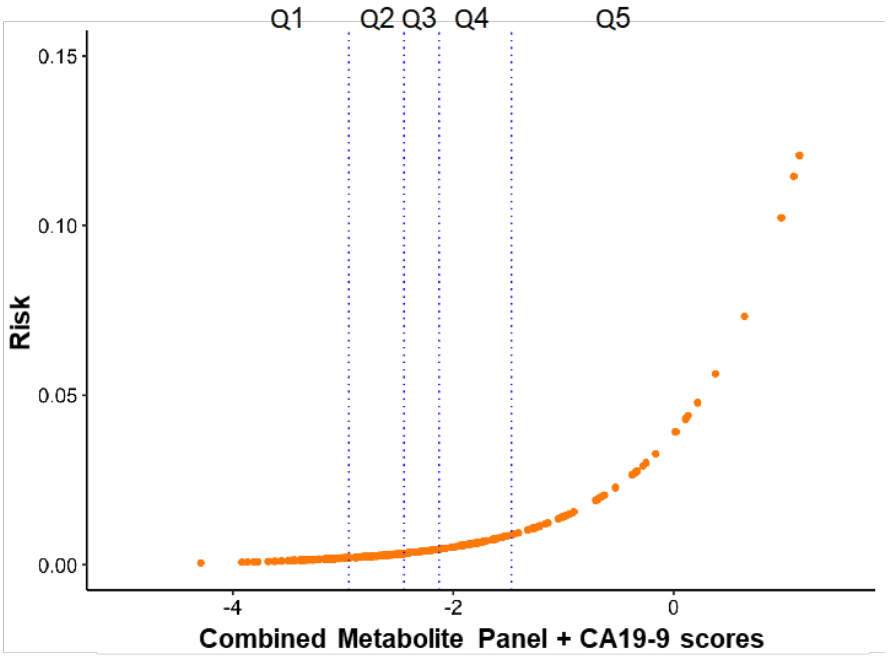
Performance of the combined metabolite panel plus CA19-9 for 5-year risk assessment of pancreatic cancer in the entire PLCO specimen set

In the entire PLCO specimen set, the combined metabolite panels + CA19-9 had an AUC of 0.80 (95% CI: 0.75-0.83) and an adjusted OR of 8.44 (95% CI: 5.80-12.20) for 5-year probability of pancreatic cancer and an AUC of 0.87 (95% CI: 0.83-0.91) with an adjusted OR of 20.02 (95% CI: 11.51-36.97) per unit SD increase for 2-year probability of pancreatic cancer (Table 5.3; Supplementary Table D.10).

5-year absolute risk estimates adjusted for prevalence of disease based on the entire intervention arm of the PLCO population (Anderson et al., 2012; Prorok et al., 2000) for individuals with combined metabolite panel + CA19-9 model scores in the 80th, 90th, and 95th percentile was 1.07%, 2.05%, and 4.52%, respectively (Figure 5.3).

5.3 Discussion

Meaningful reductions in pancreatic cancer-related mortality may be realized through effective screening programs for earlier detection of disease. The low incidence of pancreatic cancer necessitates that a screening test for the general population yield adequate sensitivity



	5-year absolute risk (%)		
Percentiles	**CA19-9**	**Combined Metabolite Panel**	**Combined Metabolite Panel + CA19-9**
20.0%	0.450	0.262	0.227
40.0%	0.517	0.425	0.350
60.0%	0.609	0.652	0.528
80.0%	0.870	1.245	1.066
90.0%	1.389	1.890	2.049
95.0%	2.159	2.880	4.521
97.5%	10.060	4.740	13.330

Figure 5.3: Absolute 5-year risk estimates for individuals with CA19-9 + 3-marker microbial panel+ 5-marker non-microbial panel scores. Vertical lines represent 20, 40, 60, and 80 percentiles values. Table beneath provides absolute 5-year risk estimates for individuals with CA19-9, combined metabolite panel (3-marker microbial metabolite panel + 5-marker non-microbial metabolite panel) and the combined metabolite panel + CA19-9 scores.

at exceptionally high specificity. No such tests yet exist that meet performance criteria necessary for implementation for pancreatic cancer screening in the general population. However, the USPSTF has recognized that high-risk individuals, such as those with inherited risk or individuals with a history of chronic pancreatitis, may benefit from screening (Force, 2019). Enriching for individuals who are at high-risk of pancreatic cancer increases the positive predictive value of pertinent tests while reducing the number of false-positive tests. Here, we performed a metabolite screen for reported microbial-related metabolites in the blood and evaluated their association with pancreatic cancer risk by following the PCS framework for reliable and reproducible data analysis. The predictive performance of this microbial-related metabolites panel was assessed and corroborated on an independent newly-diagnosed cohort of pancreatic cancer. A broader metabolite screen resulted in a blood-based metabolite panel consisting of microbial and non-microbial metabolites that yielded further improvements for 5-year risk assessment of pancreatic cancer. We further showed that the combined (microbial + non-microbial) metabolite panel is additive to CA19-9 for improved pancreatic cancer risk prediction. Such a test may enrich for individuals at increased risk of pancreatic cancer that would benefit from inclusion in screening programs.

The microbial-related metabolite panel includes indoleacyrlic acid, an indole-derivative, and TMAO. TMAO and indoleacrylic acid producing bacteria include those in the phyla of Bacillota, Bacteroidota, Actinomycetota, and Pseudomonadota. Bacillota species such as Clostridium sporogenes, Eubacterium rectale and Bacteriodota species including Bacteroides thetaiotaomicron and Parabacteroides distasonis have been shown to be increased in fecal samples of PDAC patients compared to controls (Half et al., 2019; Matsukawa et al., 2021; Zhou et al., 2021). Relative abundances of fecal Collinsella aeofaciens, a species of Actinomycetota, is associated with poor prognosis in PDAC (Matsukawa et al., 2021).

Indole and associated derivatives are derived through the catabolism of tryptophan via the microbiome that may serve as ligands for the aryl hydrocarbon receptor (AHR) to modulate the immune and inflammatory response (Hezaveh et al., 2022; Hubbard et al., 2015; Wlodarska et al., 2017). Notably, indole and indole-derivatives are thought to be largely derived from commensal microbes with reported anti-inflammatory properties (Roager and Licht, 2018).

TMAO is a gut microbiota-derived metabolite of dietary choline, betaine, and L-carnitine that has been reported to be associated with increased risk of several cancer types including pancreatic cancer (Huang et al., 2017, 2020; Jaglin et al., 2018; Jiao et al., 2019). Prior studies have shown that TMAO is elevated in pancreatic cystic fluid of individuals presenting with high-risk IPMN or pancreatic cancer compared to those harboring non-cancerous cysts (Morgell et al., 2021). Moreover, levels of TMAO in cystic fluid were positively correlated with bacterial clusters corresponding to Enterobacteriacea, Granulicatella, Klebsiella, Stenotrophomonas, Streptococcus, Haemophilus, and Fusobacterium (Morgell et al., 2021), which have previously been reported to be associated with pancreatic cancer (Rogers et al., 2017; Wei et al., 2019). Mechanistically, studies have shown that TMAO induces activation of inflammatory pathways, including the NF-Kappa B pathway and the thioredoxin-interactive protein (TXNIP)-NLRP3 inflammasome, resulting in increased oxidative stress, DNA dam-

age, and release of inflammatory cytokines that may potentiate cancer development (Arlt et al., 2012; Missiroli et al., 2021; Seldin et al., 2016; Sun et al., 2016). We observed TMAO to also be particularly elevated in patients presenting with chronic pancreatitis, further suggesting a relationship between TMAO, inflammation of pancreas tissues and pancreatic cancer risk (Farrow and Evers, 2002; Whitcomb, 2004).

Non-microbial metabolites in the metabolite panel included 2-hydroxyglutarate, cholesterol glucuronide, galactosamine, glucose, and erythritol. Production of the oncometabolite 2-hydroxyglutarate that is largely associated with mutations in isocitrate dehydrogenase 1 (IDH1) and IDH2, neomorphic enzymes that convert α-ketoglutarate to 2-hydroxyglutarate (Dang et al., 2009). 2-hydroxyglutarate can also be produced through alternative metabolic pathways with pro-tumoral effects. For instance, recent data also suggests that, under hypoxic conditions, lactate dehydrogenase produces 2-hydroxyglutarate to maintain stemness and facilitate immune evasion in pancreatic cancer (Gupta et al., 2021). Cholesterol glucuronide is a natural metabolite of cholesterol generated in the liver by UDP glucuonyltransferase. Prior studies have shown that elevated levels of cholesterol glucuronide is prognostic for poor survival in patients with pancreatic cancer (Chen et al., 2021).

The onset of diabetes is often a manifestation that precedes diagnosis of pancreatic cancer and new-onset glucose intolerance is a frequent and characteristic feature of pancreatic cancer (Pannala et al., 2009; Sharma et al., 2018). To this end, in a prior population-based case-control study of 736 pancreatic cancer cases and 1,875 age- and sex-matched controls, 40.2% of pancreatic cancer cases had diabetes (Pannala et al., 2009). In another study, 50% of patients with stage I and II pancreatic cancer had diabetes (Pannala et al., 2008, 2009; Sharma et al., 2018). Thus, elevated levels of glucose and galactosamine, a hexosamine derived from galactose (Theocharis et al., 2000), likely reflect an onset of diabetes that temporally occurs with the development of pancreatic cancer.

There are some considerations to our study. 16S sequencing data to assess stool or tissue-level microbial diversity and composition was not available for analyzed samples, thus preventing direct correlative studies between specific microbial species and the established microbial-related metabolite panel. Time-dependent performance estimates were derived based on availability of serum samples at various time points preceding cancer diagnosis from individual patients. Availability of serial samples would allow for development of longitudinal algorithms for assessment of pancreatic cancer risk. Whether the metabolite panel to inform on risk of other cancer types warrants consideration. Specificity of the metabolite panel for risk of pancreatic cancer can be improved through testing of recognized high-risk populations, including those with inherited risk (Canto et al., 2018; Petersen, 2016), mucinous cysts of the pancreas (Ohno et al., 2018), or individuals older than 50 with new onset diabetes (Pannala et al., 2009; Sharma et al., 2018).

In conclusion, the metabolite panel has potential to identify individuals at high risk of pancreatic cancer who may benefit from screening for earlier detection. Integration of the panel with other risk models of pancreatic cancer may yield further improvements for risk assessment.

Set-aside Test Set											
		CA19-9			CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel			Difference			
Time to Dx	**Sample Size**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**	**Diff. Of AUCs (95% CI)**	**P-value of Difference**	**Diff. of Adj. OR (95% CI)**	**P-value of Difference**
[0-5)	N0 = 225 N1 = 37	0.66 (0.55 - 0.77)	2.2 (1.53 - 3.30)	<0.001	0.84 (0.76 - 0.91)	9.67 (4.56 - 23.30)	<0.001	0.18 (0.08 - 0.25)	<0.001	7.47 (2.10 - 15.97)	0.003
[0-2)	N0 = 225 N1 = 24	0.70 (0.57 - 0.82)	2.55 (1.66 - 4.19)	<0.001	0.86 (0.77 - 0.95)	14.99 (5.76 - 47.66)	<0.001	0.16 (0.05-0.29)	0.006	12.44 (2.30 - 47.40)	0.01
[2-5)	N0 = 225 N1 = 13	0.60 (0.40 - 0.81)	1.64 (0.94 - 2.89)	0.01	0.79 (0.67 - 0.90)	5.1 (1.93 – 15.88)	0.002	0.19 (0.02-0.37)	0.02	3.46 (-0.06– 13.20)	0.06
Entire Set											
		CA19-9			CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel			Difference			
Time to Dx	**Sample Size**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**	**Diff. Of AUCs (95% CI)**	**P-value of Difference**	**Diff. of Adj. OR (95% CI)**	**P-value of Difference**
[0-5)	N0 = 861 N1 = 172	0.68 (0.63 - 0.73)	2.27 (1.89 - 2.76)	<0.001	0.8 (0.75 - 0.83)	8.44 (5.80 – 12.20)	<0.001	0.12 (0.07 - 0.16)	<0.001	6.17 (1.80 – 8.77)	0.004
[0-2)	N0 = 861 N1 = 92	0.75 (0.69 - 0.81)	3.21 (2.50 - 4.20)	<0.001	0.87 (0.83 – 0.91)	20.02 (11.51- 36.97)	<0.001	0.12 (0.07-0.16)	<0.001	16.81 (2.10- 27.31)	<0.001
[2-5)	N0 = 861 N1 = 80	0.60 (0.53 - 0.67)	1.48 (1.18 - 1.87)	<0.001	0.71 (0.65 – 0.77)	3.52 (2.36- 5.32)	<0.001	0.11 (0.57-0.70)	0.001	2.04 (1.20 - 5.86)	0.04

Table 5.3: Performance estimates of the CA19-9 and a combined CA19-9 + 3-marker microbial panel + 5-marker non-microbial panel for 5-year risk prediction of pancreatic cancer in the set-aside Test Set and the entire PLCO specimen set. †Age, sex, bmi and smoking status were included as co-variables in adjusted odd ratios. Log transformation of the values were considered for adjusted odds ratio calculation. N0: number of non-cases. N1: number of cases.

5.4 Methods

PLCO Cohort

The PLCO Cancer Screening Trial is a randomized multicenter trial in the United States that aimed to evaluate the impact of early detection procedures for prostate, lung, colorectal and ovarian cancer on disease-specific mortality. Detailed information regarding the PLCO cohort is provided elsewhere (Fahrmann et al., 2021; Prorok et al., 2000). The study included 173 pancreatic cancer cases that were diagnosed within 5 years of blood draw and 863 matched non-cases from 10 participating PLCO study centers (Table 5.4). Pancreatic cancer cases were identified by self-report in annual mail-in surveys, state cancer registries, death certificates, physician referrals and reports from next of kin for deceased individuals. All medical and pathologic records related to pancreatic cancer diagnosis and supporting documentation were obtained and confirmed by PLCO staff. Pancreatic cancers were classified as localized, regional, distant, or unstaged using the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) historic staging system. Non-cases, alive at the time when the index case was diagnosed, were matched to cases at a ratio of 5:1 (non-case:case) based on the distribution of age, race, sex, and calendar date of blood draw in 2-month blocks within the case cohort.

Newly diagnosed pancreatic cancer cohort

An independent test set consisted of plasma samples from 99 patients with resected pancreatic ductal adenocarcinoma (PDAC), 50 patients with chronic pancreatitis, and 100 healthy controls as previously described (Supplemental Table D.1) (Fahrmann et al., 2021). Patients with pancreatic cancer provided informed written consent to blood collection pretreatment and to clinical data abstraction. Patients with PDAC were recruited from cancer clinics at Dana-Farber Cancer Institute/Brigham and Women's Hospital (DFCI/BWH), Beth Israel Deaconess Medical Center (BIDMC), and Columbia University Irving Medical Center (CUIMC). Healthy controls were recruited from DFCI/BWH and CUIMC and consisted of subjects undergoing screening colonoscopy or accompanying a non–blood-related patient to an appointment at a gastrointestinal cancer clinic. Healthy controls had no history of cancer in the 5 years before sample collection. Patients with pancreatic cancer and healthy controls were matched on gender and age at the time of blood collection. Patients with chronic pancreatitis (CP) were recruited from gastroenterology clinics at DFCI/BWH, BIDMC, and CUIMC. Patients were included if clinic notes from a gastroenterologist indicated a diagnosis of CP. Patients with pancreatic cancer or CP were not gender or age matched. Clinical data abstraction was performed identically across the sites with data uploaded to a password-protected REDCap database. All plasma samples were collected and processed according to a uniform, standardized protocol across the sites and patient groups.

	Case/Control Status			
	Non-Case		***Case***	
	N	***%***	***N***	***%***
Total	863	100	173	100
Gender				
Female	357	41.4	72	41.6
Male	506	58.6	101	58.4
Age At Randomization				
<= 59	183	21.2	37	21.4
60-64	206	23.9	41	23.7
65-69	321	37.2	64	37
>= 70	153	17.7	31	17.9
Race				
White	783	90.7	157	90.8
Black	30	3.5	6	3.5
Other	50	5.8	10	5.9
Cigarette Smoking Status				
Never Smoked Cigarettes	420	48.7	63	36.4
Current Cigarette Smoker	74	8.6	36	20.8
Former Cigarette Smoker	369	42.8	74	42.8
BMI at Baseline (In kg/m2)				
Not Answered	7	0.8	0	0
0-18.5	8	0.9	3	1.7
18.5-25	300	34.8	56	32.4
25-30	365	42.3	71	41
30+	183	21.2	43	24.9
Diabetic Status				
Unknown	1	0.1	0	0
Yes	55	6.4	22	12.7
No	807	93.5	151	87.3
SEER Staging (cases only)				
Unknown	-	-	15	8.7
Localized	-	-	35	20.2
Regional	-	-	33	19.1
Distant	-	-	90	52

Table 5.4: Patient and tumor characteristics for PLCO cohort.

Metabolomic analysis

Sample Extraction

Serum and plasma metabolites were extracted from pre-aliquoted biospecimen (15 µL) with 45 µL of LCMS grade methanol (ThermoFisher) in a 96-well microplate (Eppendorf). Plates were heat sealed, vortexed for 5 min at 750 rpm, and centrifuged at 2000 × g for 10 minutes at room temperature. The supernatant (30 µL) was carefully transferred to a 96-well plate, leaving behind the precipitated protein. The supernatant was further diluted with 60 µL of 100 mM ammonium formate, pH3 (Fisher Scientific). For Hydrophilic Interaction Liquid Chromatography (HILIC) positive ion analysis, 15 µL of the supernatant and ammonium formate mix were diluted with 195 µL of 1:3:8:144 water (GenPure ultrapure water system, Thermofisher): LCMS grade methanol (ThermoFisher): 100 mM ammonium formate, pH3 (Fisher Scientific): LCMS grade acetonitrile (ThermoFisher). For C18 analysis, 15 µL of the supernatant and ammonium formate mix were diluted with 90 µL water (GenPure ultrapure water system, Thermofisher) for positive ion mode. Each sample solution was transferred to 384-well microplate (Eppendorf) for LCMS analysis.

Untargeted Metabolomic Analyses

Untargeted metabolomics analysis was conducted on Waters Acquity™ UPLC system with 2D column regeneration configuration (I-class and H-class) coupled to a Xevo G2-XS quadrupole time-of-flight (qTOF) mass spectrometer as previously described (Fahrmann et al., 2019, 2021, 2020; Johannes et al., 2021). Chromatographic separation was performed using HILIC (Acquity™ UPLC BEH amide, 100 Å, 1.7 µm, 2.1 × 100 mm, Waters Corporation, Milford, U.S.A) and C18 (Acquity™ UPLC HSS T3, 100 Å, 1.8 µm, 2.1×100 mm, Water Corporation, Milford, U.S.A) columns at 45°C.

Quaternary solvent system mobile phases were (A) 0.1% formic acid in water, (B) 0.1% formic acid in acetonitrile and (D) 100 mM ammonium formate, pH 3. Samples were separated on the HILIC using the following gradient profile at 0.4 mL/min flow rate: (95% B, 5% D) linear change to (70% A, 25% B and 5% D) over 5 min; 100% A for 1 min; and 100% A for 1 min. For C18 separation, the chromatography gradient was as follows at 0.4 mL/min flow rate: 100% A with a linear change to (5% A, 95% B) over 5 min; (95% B, 5% D) for 1 min; and 1 min at (95% B, 5% D).

A binary pump was used for column regeneration and equilibration. The solvent system mobile phases were (A1) 100 mM ammonium formate, pH 3, (A2) 0.1% formic in 2-propanol and (B1) 0.1% formic acid in acetonitrile. The HILIC column was stripped using 90% A2 for 5 min at 0.25 mL/min flow rate, followed by a 2 min equilibration using 100% B1 at 0.3 mL/min flow rate. Reverse phase C18 column regeneration was performed using 95% A1, 5% B1 for 2 min followed by column equilibration using 5% A1, 95% B1 for 5 min at 0.4 mL/min flow rate.

Mass Spectrometry Data Acquisition

Mass spectrometry data was acquired using 'sensitivity' mode in positive electrospray ionization mode within 50-800 Da range. For the electrospray acquisition, the capillary voltage was set at 1.5 kV (positive), sample cone voltage 30 V, source temperature at 120°C, cone gas flow 50 L/h and desolvation gas flow rate of 800 L/h with scan time of 0.5 sec in continuum mode. Leucine Enkephalin; 556.2771 Da (positive) was used for lockspray correction and scans were performed at 0.5 sec. The injection volume for each sample was 6 µL. The acquisition was carried out with instrument auto gain control to optimize instrument sensitivity over the samples acquisition time.

Data Processing

LC-MS and LC-MSe data were processed using Progenesis QI (Nonlinear, Waters). Peak picking and retention time alignment of LC-MS and MSe data were performed using Progenesis QI software (Nonlinear, Waters). Data processing and peak annotations were performed using an in-house automated pipeline as previously described (Fahrmann et al., 2019, 2021, 2020; Vykoukal et al., 2020). Annotations were determined by matching accurate mass and retention times using customized libraries created from authentic standards and by matching experimental tandem mass spectrometry data against the NIST MSMS, LipidBlast or HMDB v3 theoretical fragmentations. To correct for injection order drift, each feature was normalized using data from repeat injections of quality control samples collected every 10 injections throughout the run sequence. Measurement data were smoothed by Locally Weighted Scatterplot Smoothing (LOESS) signal correction (QC-RLSC) as previously described. Values are reported as ratios relative to the median of historical quality control reference samples run with every analytical batch for the given analyte (Fahrmann et al., 2019, 2021, 2020; Vykoukal et al., 2020).

Microbial-associated Metabolite Database

To evaluate the association between the microbial-associated metabolites identified in the PLCO specimen sets with distinct microbial species, we used the Metabolomics Data Explorer database[1] developed by Han et al. (2021). The database reports the metabolic profiles of 178 gut microorganism strains; microbiota-dependent metabolites were established in diverse biological fluids from gnotobiotic and conventionally colonized mice and traced back to the corresponding metabolomic profiles of cultured bacteria (Han et al., 2021).

Statistical analysis

Predictive performance estimates for individual microbial-related metabolites identified and quantified through metabolomic profiling of sera were assessed using receiver operating char-

[1] `https://sonnenburglab.github.io/Metabolomics_Data_Explorer/#/invivo`

acteristic curve (ROC). Time-dependent ROC analyses were performed using pROC (version 1.15.3) in the R software environment (version 3.6.1, The R Foundation, https://www.r-project.org). The 95% confidence intervals (CI) for AUCs were estimated using the Delong method (DeLong et al., 1988). Corresponding 95% confidence intervals for odds ratios, adjusted odds ratios, specificity, sensitivity and the difference measurements were calculated using 1,000 bootstrap samples. Age, sex, BMI, and smoking status were included as covariates in the adjusted odds ratio.

Throughout the statistical analysis, we adhered to the PCS (Predictability, Computability and Stability) framework for veridical (trustworthy) data science (Yu and Kumbier, 2020), which has proven valuable in many previous scientific discoveries including novel gene-gene interaction for the red-hair phenotype (Behr et al., 2020), clinically-relevant subgroups in a randomized drug trial (Dwivedi et al., 2020), and interpretable drug response prediction (Li et al., 2020). For the modeling stage as in this paper, the PCS framework uses predictability as a reality check, and for reproducibility, it advocates for a stability analysis across different reasonable perturbations of the data and models that pass the prediction check. Under this framework, the entire PLCO specimen set was divided into (1) a Development Set that was used for training and tuning the models (Training Set) and model selection (Validation Set) and (2) a set-aside Test Set for obtaining an unbiased evaluation of the selected final model (Figure 5.4). The Development Set consisted of case and non-case sera from seven of the ten PLCO study centers; the set-aside Test Set consisted of case and non-case sera from the remaining three PLCO study centers.

Seven different learning algorithms were evaluated including a deep learning model (fully-connected feed-forward network), gradient boosting machine, auto-machine learning, iterative random forest, logistic regression with LASSO (L_1) regularization, logistic regression with ridge (L_2) regularization, and logistic regression models. Deep neural network, extreme gradient boosting, and auto machine learning algorithms were performed using the h2o package in R (Candel et al., 2016). Iterative random forest was run using the iRF package in R (Basu et al., 2018). To further evaluate model stability in accordance with PCS framework, data perturbations (e.g. via random selection and replacement) were introduced to the Development Set and the performance re-assessed. Based on AUC, a LASSO regression model was selected for subsequent testing in the set-aside Test Set as well as the independent newly-diagnosed PDAC cohort.

To select the non-microbiome metabolites, the adjusted odds ratio and corresponding p-value for each feature were calculated and corrected using Benjamini-Hochberg in the training set in which 12 metabolites showed an adjusted odds ratio greater than 1 with adjusted p-value less than 0.05. Five out of 12 features yielded significant p-values and adjusted odds ratio greater than 1 in the Validation Set. The prediction performance of the combined five non-microbiome features trained in the training set using logistic regression was evaluated against the microbiome metabolite panel and CA19-9 in the testing set.

For the combination of 3-marker microbial-related metabolite panel, non-microbiome metabolite panel and CA19-9, we fit a logistic regression with three separate predictors, one corresponding to each of the aforementioned features. This model was developed in the

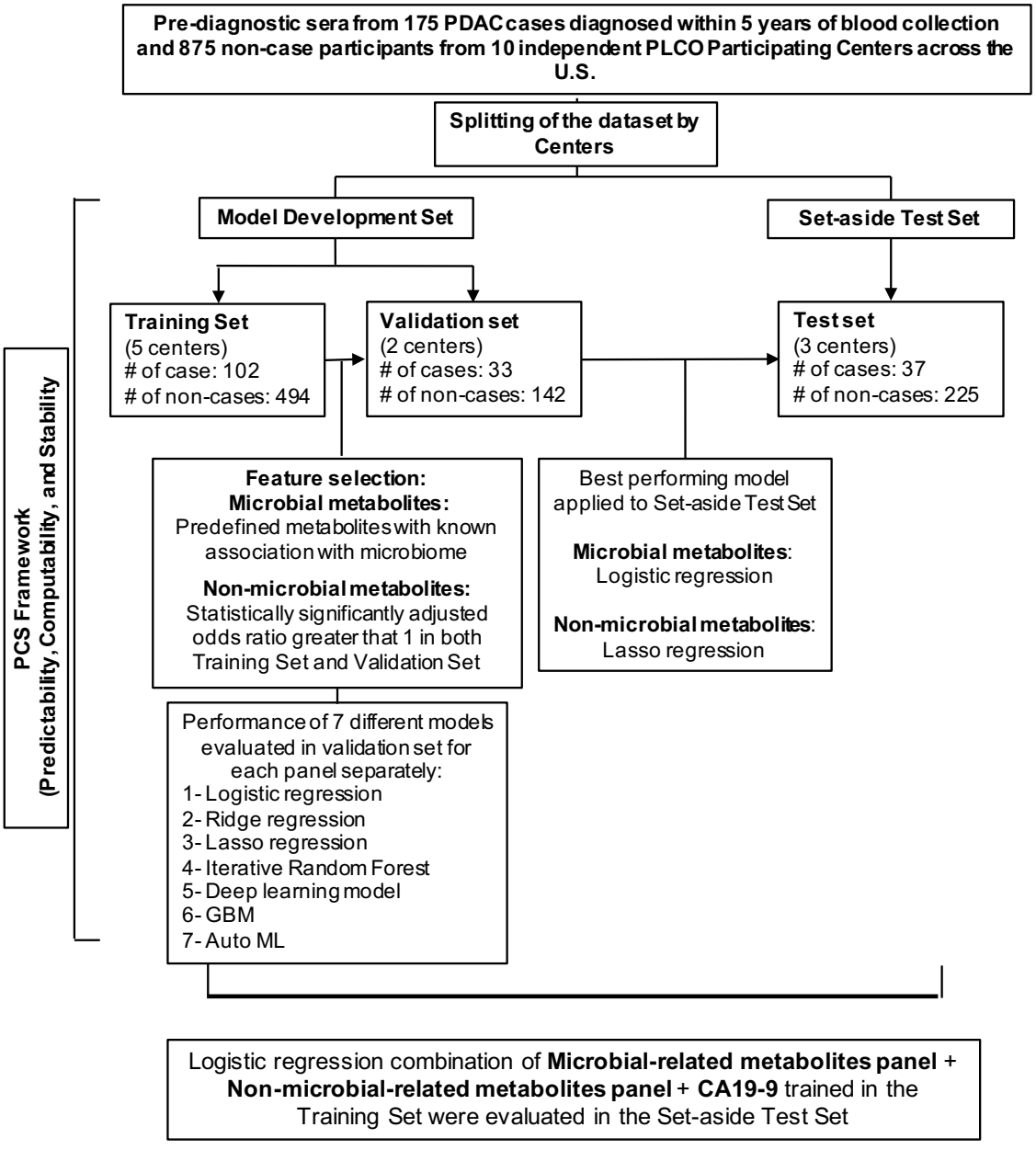


Figure 5.4: Workflow of analyses.

Development Set and validated in the set-aside Test Set.

Samples assayed via metabolomics herein reflect a nested case-control cohort that enriches for cases and, therefore, do not reflect the true risk of pancreatic cancer in the general population. In order to determine the 0.5%, 1%, 1.5% and 2% 5-year risk of pancreatic cancer, we thus adjust the estimates to reflect the entire PLCO study population using the approach of Prentice and Pyke (1979). In this approach, a prospective logistic model is estimated from the case-control study that includes an offset term to the logistic model. The offset term is the logit of the prevalence in the population minus the logit of the prevalence in the analyzed dataset. Briefly, absolute risk values for each biomarker were estimated by calculating coefficients of a logistic regression in the training set and the intercept adjusted using the following equation:

$$Risk = \frac{\exp\left(\beta_0' + \beta_1 \times (model)\right)}{1 + \exp(\beta_0' + \beta_1 \times (model))},$$

where

$$\beta_0' = \beta_0 - \log\left(\frac{P_{\text{data}}}{1 - P_{\text{data}}}\right) + \log\left(\frac{P_{\text{Population}}}{1 - P_{\text{Population}}}\right).$$

In this equation, β_0 is the intercept derived from logistic regression in the nested case-control within a cohort, P_{data} is the prevalence of the disease in our case-enriched dataset, $P_{\text{Population}}$ is the prevalence of the disease in the general population, *model* represents the predicted score derived from the selected model and β_1 is the corresponding coefficient for the model score.

Part III

Open-source software and data

Chapter 6

simChef: An R package for high-quality data science simulations using PCS

Data science simulation studies occupy an important role in data science research as a means to gain insight into new and existing statistical methods. Whether as a means to establish comprehensive benchmarks of existing procedures for a common task, to demonstrate the strengths and weaknesses of novel methodology applied to synthetic and real-world data, or to probe the validity of a theoretical analysis, simulations serve as statistical sandboxes that open a path toward otherwise inaccessible discoveries. Yet creating high-quality simulation studies typically involves a number of repetitive and error-prone coding tasks, such as implementing data-generating processes (DGPs) and statistical methods, sampling from these DGPs, parallelizing computation of simulation replicates, summarizing metrics, and visualizing, documenting, and saving results. While this administrative overhead is necessary to reach the end goals of a given data science simulation, it is not sufficient, as the data scientist must navigate a number of important judgment calls such as the choice of data settings, baseline statistical methods, associated parameters, and evaluation metrics for scientific relevancy. The scientific context varies drastically from one study to the next while the simulation scaffolding remains largely similar; yet simulation code repositories often lack the flexibility to easily allow for reuse in novel settings or even simple extension when new questions arise in the original context.

`simChef` addresses the need for an intuitive, extensible, and reusable framework for data science simulations. Drawing substantially from the Predictability, Computability, and Stability (PCS) framework (Yu and Kumbier, 2020), `simChef` empowers data scientists to focus their attention toward the scientific best practices encompassed by PCS by removing many of the administrative burdens of simulation design with an intuitive tidy grammar[1] of data science simulations and automated interactive R Markdown documentation.

[1] `https://design.tidyverse.org/`

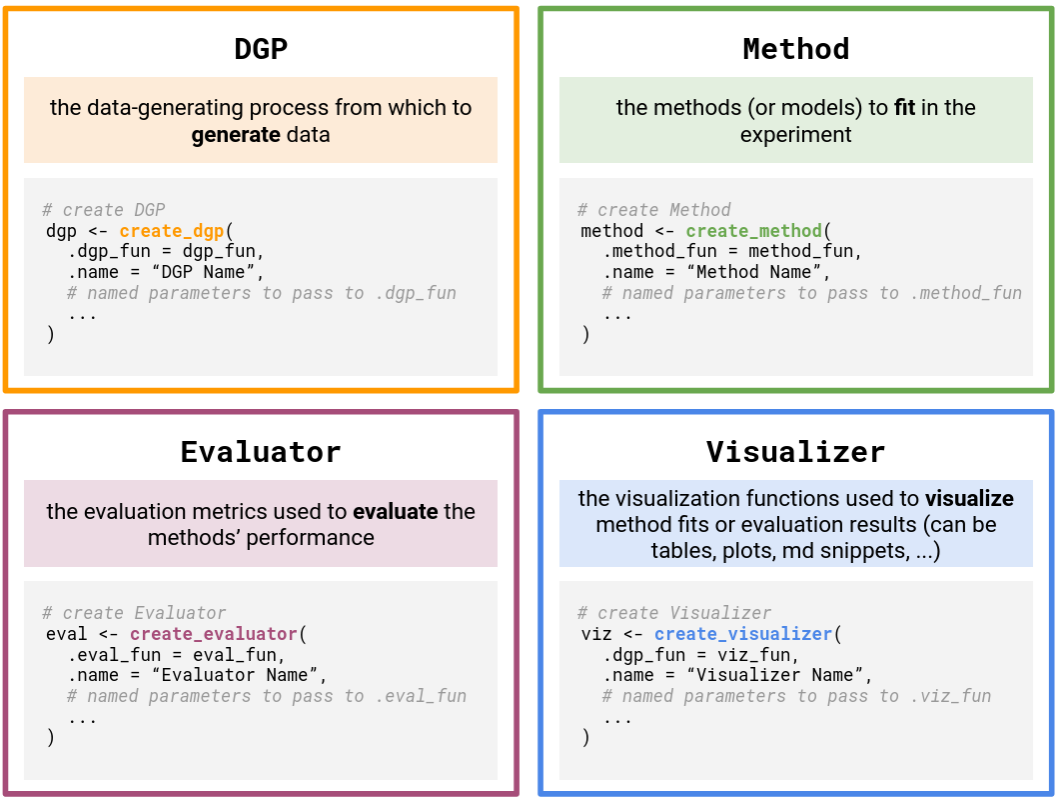


Figure 6.1: `simChef` provides four classes which implement distinct simulation objects in an intuitive and modular manner: `DGP`, `Method`, `Evaluator`, and `Visualizer`.

6.1 Core abstractions of data science simulations

At its core, `simChef` breaks down a simulation experiment into four modular components (Figure 6.1), each implemented as an `R6` class (Chang, 2022):

1. `DGP`: the data-generating processes from which to *generate* data
2. `Method`: the methods (or models) to *fit* in the experiment
3. `Evaluator`: the evaluation metrics used to *evaluate* the methods performance
4. `Visualizer`: the visualization functions used to *visualize* outputs from the method fits or evaluation results (can be tables, plots, or even R Markdown snippets to display)

Using these classes, users can create or reuse custom functions (i.e., `dgp_fun`, `method_fun`, `eval_fun`, and `viz_fun` in Figure 6.1) aligned with their scientific goals. The custom functions are then optionally parameterized and encapsulated in one of the corresponding classes via a `create_*` method together with optional constant parameters.

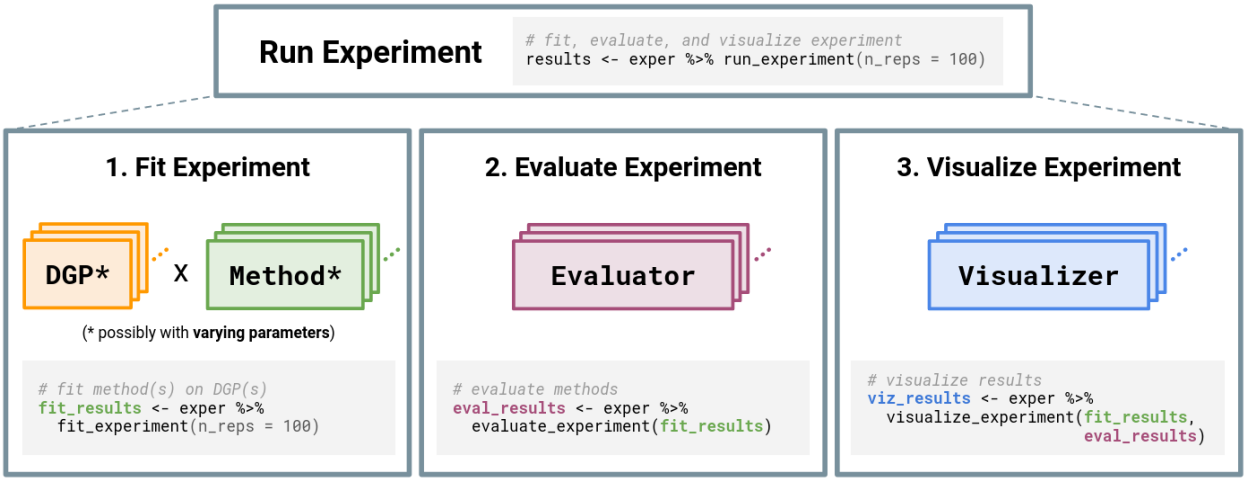


Figure 6.2: The `Experiment` class handles relationships between the four classes portrayed in Figure 6.1. Experiments may have multiple `DGP`s and `Method`s, which are combined across the Cartesian product of their varying parameters (represented by). Once computed, each `Evaluator` and `Visualizer` takes in the fitted simulation replicates, while `Visualizer` additionally receives evaluation summaries.

A fifth `R6` class, `Experiment`, unites the four components above and serves as a concrete implementation of the user's intent to answer a specific scientific question. Specifically, the `Experiment` stores references to the `DGP`(s), `Method`(s), `Evaluator`(s), and `Visualizer`(s) along with the `DGP` and `Method` parameters that should be varied and combined during the simulation run.

6.2 A powerful grammar of data science simulations

Inspired by the tidyverse (Wickham et al., 2019), `simChef` develops an intuitive grammar for running simulation studies using the aforementioned `R6` classes. We provide an illustrative example in Figure 6.3.

In Figure 6.3, `DGP`(s), `Method`(s), `Evaluator`(s), and `Visualizer`(s) are first created via `create_*()`. These simulation objects can then be combined into an `Experiment` using either `create_experiment()` and/or `add_*()`.

In an `Experiment`, `DGP`(s) and `Method`(s) can also be varied across one or multiple parameters via `add_vary_across()`. For instance, in the example `Experiment`, there are two `DGP` instances, both of which are varied across three values of `n` and one of which is additionally varied across two values of `sparse`. This effectively results in nine distinct configurations for data generation. For the single `Method` in the experiment, we use three values of `scalar_valued_param`, two of `vector_valued_param`, and another two of

```r
library(simChef)

dgp1 <- create_dgp(dgp_fun1, "my_dgp1", sd = 0.5)
dgp2 <- create_dgp(dgp_fun2, "my_dgp2")
method <- create_method(method_fun, "my_method")
eval <- create_evaluator(eval_fun)
viz <- create_vizualizer(viz_fun)

exper <- create_experiment(dgp_list = list(dgp1, dgp2)) %>%
  add_method(method) %>%
  add_vary_across(
    list(dgp1, dgp2),
    n = c(1e2, 1e3, 1e4)
  ) %>%
  add_vary_across(
    dgp2,
    sparse = c(FALSE, TRUE)
  ) %>%
  add_vary_across(
    method,
    scalar_valued_param = c(0.1, 1.0, 10.0),
    vector_valued_param = list(c(1, 2, 3), c(4, 5, 6)),
    list_valued_param = list(list(a1=1, a2=2, a3=3),
                             list(b1=3, b2=2, b3=1))
  ) %>%
  add_evaluator(eval) %>%
  add_viz(viz)

future::plan(multicore, workers = 64)

results <- exper %>%
  run_experiment(n_reps = 100, save = TRUE)

new_method <- create_method(new_method_fun, 'my_new_method')

exper <- exper %>%
  add_method(new_method)

results <- exper %>%
  run_experiment(n_reps = 100, use_cached = TRUE)

init_docs(exper)
render_docs(exper)
```

Figure 6.3: Example usage of `simChef`.

`list_valued_param`, giving 12 distinct configurations. Hence, there are a total of 108 DGP-method-parameter combinations in the `Experiment`.

Thus far, we have simply instantiated an `Experiment` object (akin to creating a recipe for an experiment). To compute and run the simulation experiment, we call `run_experiment` with the desired number of replicates. As summarized in Figure 6.2, running the experiment will (1) *fit* each `Method` on each `DGP` (and for each of the varying parameter configurations), (2) *evaluate* the experiment according to the given `Evaluator`(s), and (3) *visualize* the experiment according to the given `Visualizer`(s). Furthermore, the number of replicates per combination of `DGP`, `Method`, and parameters specified via `add_vary_across` is determined by the `n_reps` argument to `run_experiment`. Because replication happens at the per-combination level, the effective total number of replicates in the `Experiment` depends on the number of DGPs, methods, and varied parameters. In the given example, there are 108 DGP-method-parameter combinations, each of which is replicated 100 times. To reduce the computational burden, the `Experiment` class flexibly handles the computation of simulation replicates in parallel using the `future` package (Bengtsson, 2021). Figure 6.4 provides a detailed schematic of the `run_experiment` workflow, along with the expected inputs to and outputs from user-defined functions.

6.3 Additional Features

In addition to the ease of parallelization, `simChef` enables caching of results to further alleviate the computational burden. Here, users can choose to save the experiment's results to disk by passing `save = TRUE` to `run_experiment`. Once saved, the user can add new `DGP` and `Method` objects to the experiment and compute additional replicates without re-computing existing results via the `use_cached` option. Considering the example above, when we add `new_method` and call `run_experiment` with `use_cached = TRUE`, `simChef` finds that the cached results are missing combinations of `new_method`, existing DGPs, and their associated parameters, giving nine new configurations. Replicates for the new combinations are then appended to the cached results.

`simChef` also provides users with a convenient API to automatically generate an R Markdown document. This documentation gathers the scientific details, summary tables, and visualizations side-by-side with the user's custom source code and parameters for data-generating processes, statistical methods, evaluation metrics, and plots. A call to `init_docs` generates empty markdown files for the user to populate with their overarching simulation objectives and with descriptions of each of the `DGP`, `Method`, `Evaluator`, and `Visualizer` objects included in the `Experiment`. Finally, a call to `render_docs` prepares the R Markdown document, either for iterative design and analysis of the simulation or to provide a high-quality overview that can be easily shared. We provide an example of the simulation documentation at `https://philboileau.github.io/simChef-case-study/results/empirical-fdr-comparison/empirical-fdr-comparison.html`. Corresponding R source code is available on GitHub at `https://github.com/PhilBoileau/simChef-case-study`.

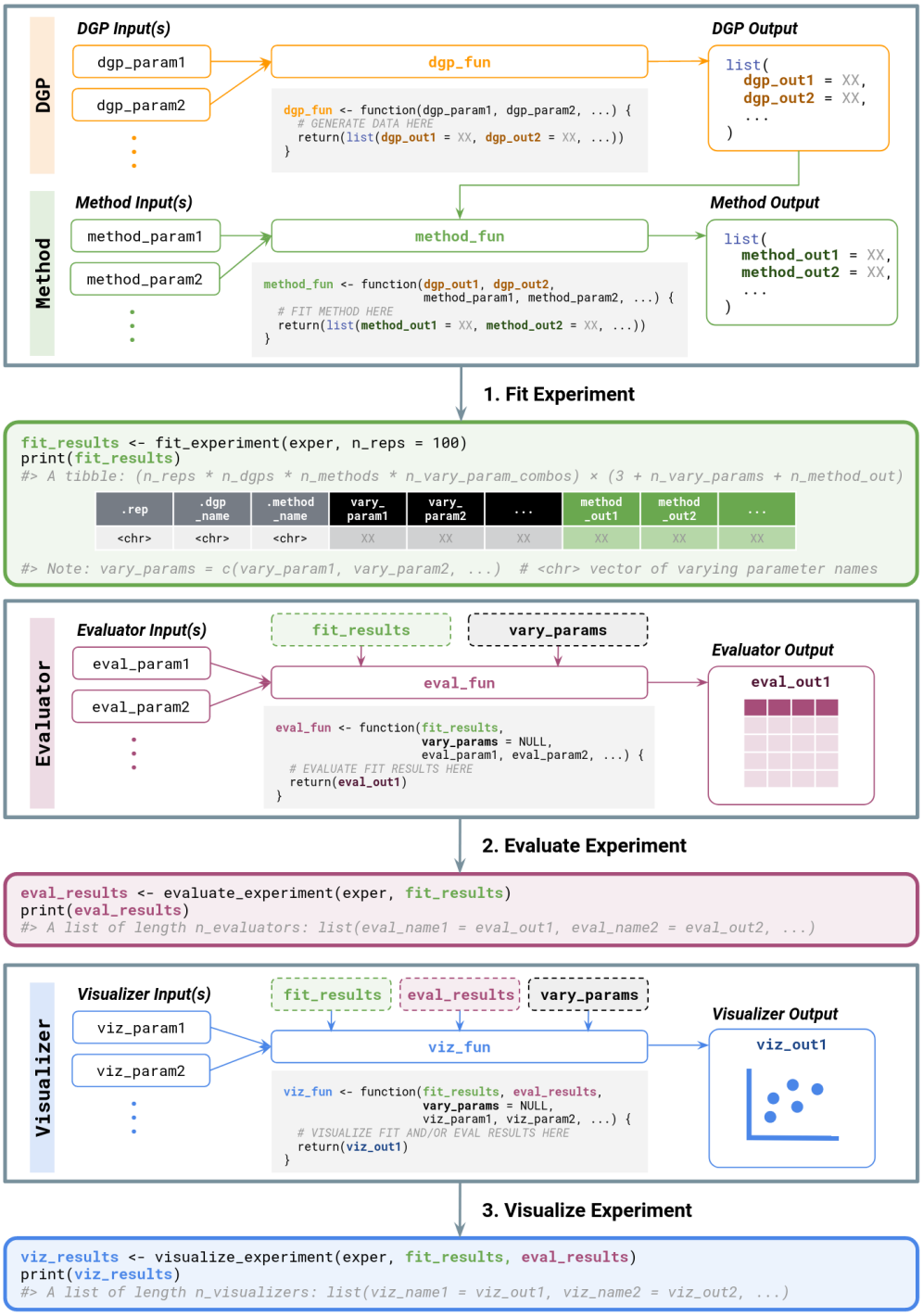


Figure 6.4: Detailed schematic of the `run_experiment` workflow. Expected inputs to and outputs from user-defined functions are also provided.

Chapter 7

vdocs: An R package for rigorous and transparent PCS documentation

In Part II, we detailed two collaborative projects, which were heavily driven by the need for stable, reliable conclusions and the PCS framework for veridical data science. These projects also illustrated the necessity for transparent documentation of the many human judgment calls and choices that are inevitably made throughout the data science pipeline. For example, what models were tried and why? How was the data cleaned and why? Which evaluation metrics are used, and why? Being transparent about these decisions is critical for scientific reproducibility (Yu and Kumbier, 2020). However, creating such documentation often requires a considerable amount of work and effort and is thus not common practice.

To lower the activation barrier for creating rigorous and transparent documentation, we developed an R package named `vdocs`. `vdocs` provides a beautiful, user-friendly R Markdown template to facilitate PCS-style documentation and encourage more veridical data science in practice. A snapshot of this template is shown in Figure 7.1. Like scientists with their lab notebooks, data scientists should provide a narrative of their discoveries and justification for human judgment calls made along the way in a lab notebook designed for veridical data science. `vdocs` makes this documentation via a *PCS lab notebook* quick and easy with the click of a button and minimal code.

`vdocs` is still under active development. We detail its currently available features in Section 7.1 and provide a brief roadmap of features planned for the future in Section 7.2.

7.1 Core Features

The main feature of `vdocs` is the *PCS lab notebook* template. In RStudio, the easiest way to create a new PCS lab notebook is to open a new R Markdown file from template: go to `File > New File > R Markdown... > From Template > PCS Lab Notebook > OK` (see Figure 7.2). A directory, whose name was specified in the "Name" dialog box, is then created and contains the .Rmd PCS lab notebook template.

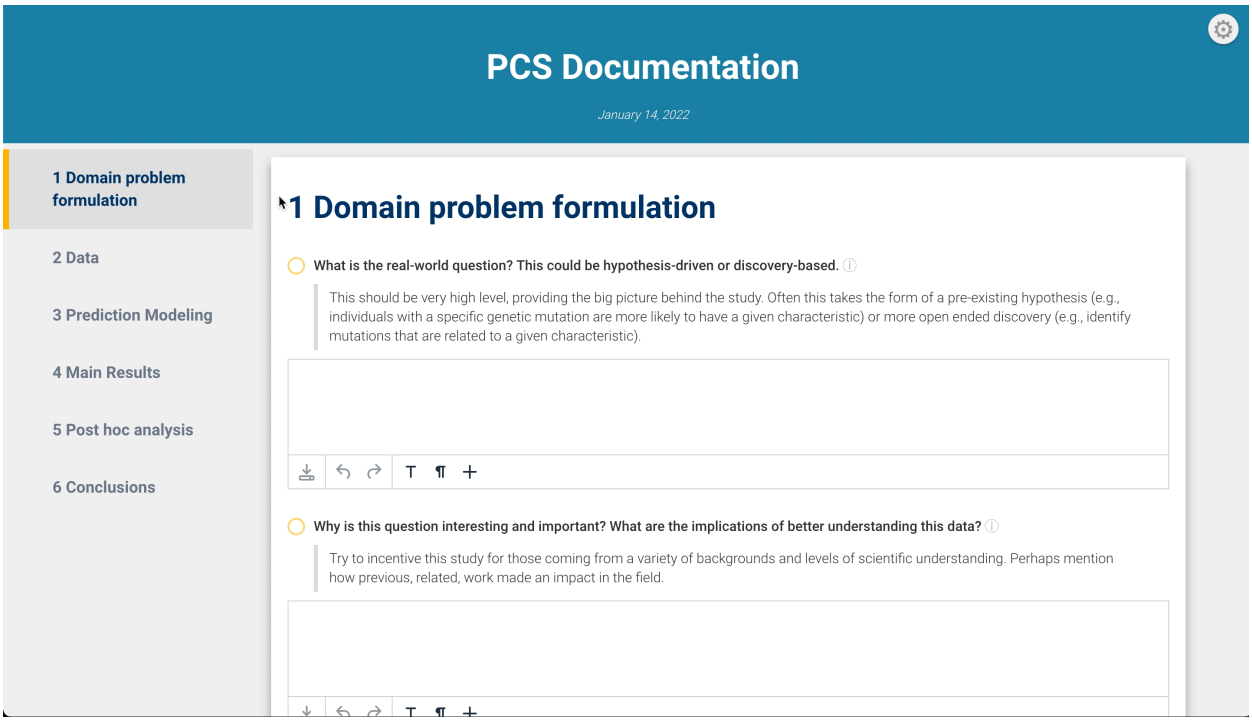


Figure 7.1: Snapshot of the `vdocs` PCS lab notebook template.

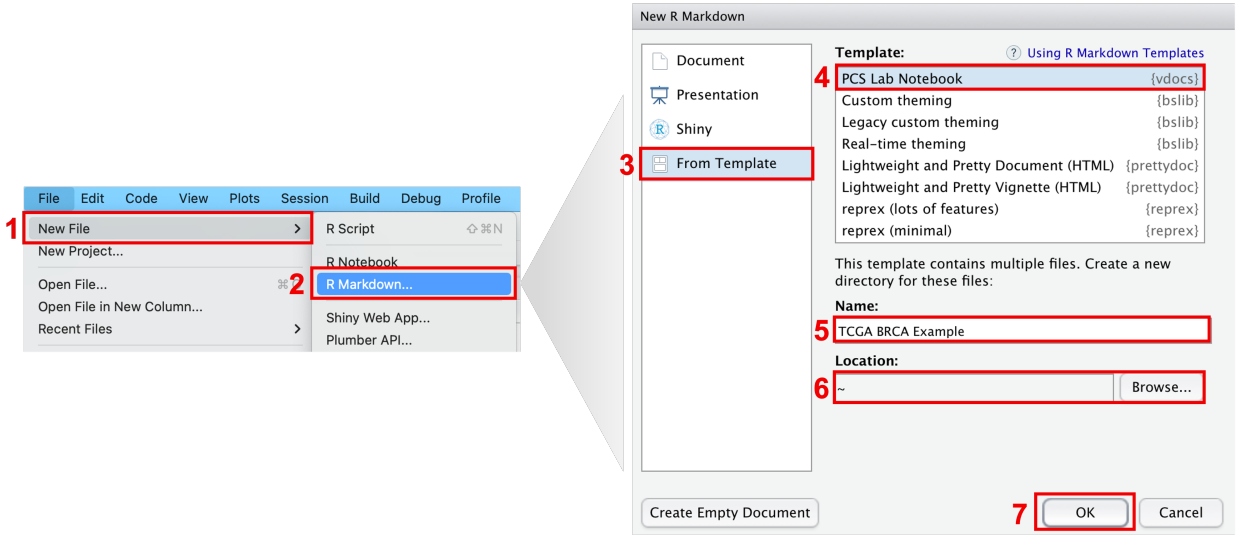


Figure 7.2: Steps to create a new PCS Lab Notebook in RStudio: go to `File > New File > R Markdown... > From Template > PCS Lab Notebook > OK`.

This PCS lab notebook template features:

1. A basic checklist of questions that should be considered and documented throughout the analysis pipeline
2. Interactive textboxes to easily record responses to these questions and document other human judgment calls
3. Additional tips when performing the analysis in order to comply with principles from the PCS framework
4. Starter code to run a PCS-style analysis with basic prediction and stability checks

We note that the provided PCS lab notebook template has been auto-populated with R code to perform a PCS-style analysis for a basic supervised learning task However, any or all of the R code can be changed. The PCS lab notebook still adds much-needed value by providing a checklist of questions that initiates further discussions and thinking of the PCS principles throughout the data analysis. While responding to these questions takes time, we highly encourage every data scientist or practitioner to put in this extra effort as a step towards our greater goal of ensuring scientific reproducibility.

Details on how to use or modify the PCS lab notebook can be found in our vignette found at `https://yu-group.github.io/vdocs/articles/vdocs.html`. We also provide an example usage at `https://yu-group.github.io/vdocs/TCGA-BRCA-Example.html`. In this example, we walk through an analysis using real-world breast cancer data from The Cancer Genome Atlas (Atlas, 2012). The goal in this example analysis is to (1) predict the breast cancer subtype (known as the PAM50 subtype (Parker et al., 2009)) using gene expression data and (2) to identify the important genes that lead to these predictions.

7.2 Future Roadmap

Currently, the main limitation of `vdocs` is its accessibility. As is, the PCS lab notebook is an R Markdown template and hence most accessible to users who are familiar with R. To broaden the reach of `vdocs`, we plan to build a Shiny counterpart, which shares many of the features of the R Markdown PCS lab notebook template. This Shiny application would provide a more interactive, click-and-point interface, tailored for practitioners with little to no coding experience. Other planned improvements to `vdocs` include the following:

- Expand the code scaffolding to accommodate tasks beyond prediction and basic interpretations of these prediction models
- Provide additional case studies and examples of PCS-style analyses using the `vdocs` template

- Migrate from R Markdown to Quarto to leverage the more flexible capabilities of Quarto for generating production quality output

Chapter 8

Curating a COVID-19 data repository

At the height of the COVID-19 pandemic (around March 2020), we began curating a large open-source data repository on GitHub with COVID-19-related data. The goal of this COVID-19 data repository was to aid community-wide data science efforts by providing easy access to a large corpus of cleaned COVID-19-related data. In particular, our COVID-19 data repository laid the foundation for highly-collaborative work on COVID-19 severity prediction (Altieri et al., 2020). This work was used to allocate over 65,000 face shields to twenty-five recipients in fifteen states through the non-profit organization, Response4Life. The data repository is open-source on GitHub at `https://github.com/Yu-Group/covid19-severity-prediction`.

8.1 Overview of the COVID-19 data repository

As of June 20, 2020, the COVID-19 data repository consists of both *hospital-level* and *county-level* data from over twenty public data sources, highlighted next.

Hospital-level data

At the hospital-level, our data repository covers over 7000 US hospitals and over 30 features including the hospital's CMS certification number (a unique ID of each hospital used by Centers for Medicare and Medicaid Services), the hospital's location, the number of ICU beds, the hospital type (e.g., short-term acute care and critical access), and numerous other hospital statistics. We provide a feature-level snapshot of the different types of hospital-level data available in our repository in Table 8.1. Alternatively, in Table 8.2, we provide an overview of the hospital-level data sources in our repository, organized by the dataset.

County-level data

There are more than 3,100 counties in the US. At the county-level, our repository includes:

Description of Hospital-level Features	Data Source(s)
CMS certification number	Centers for Medicares & Medicaid Services (2018) (Case Mix Index File)
Case Mix Index	Centers for Medicares & Medicaid Services (2018) (Case Mix Index File); Centers for Medicares & Medicaid Services (2020) (Teaching Hospitals)
Hospital location (latitude and longitude)	Homeland Infrastructure Foundation-Level Data (2020); Definitive Healthcare (2020)
# of ICU/staffed/licensed beds and beds utilization rate	Definitive Healthcare (2020)
Hospital type	Homeland Infrastructure Foundation-Level Data (2020); Definitive Healthcare (2020)
Trauma Center Level	Homeland Infrastructure Foundation-Level Data (2020)
Hospital website and telephone number	Homeland Infrastructure Foundation-Level Data (2020)

Table 8.1: A list of select relevant features from across all hospital-level datasets contained in our COVID-19 repository. See Table 8.2 for an overview of each hospital-level dataset.

Hospital-level Dataset	Description
Homeland Infrastructure Foundation-Level Data (2020)	Includes number of ICU beds, and location for US hospitals
Definitive Healthcare (2020)	Provides data on number of licensed beds, staffed beds, ICU beds, and the bed utilization rate for hospitals in the US
Centers for Medicares & Medicaid Services (2018) (Case Mix Index File)	Reports the Case Mix Index (CMI) for each hospital
Centers for Medicares & Medicaid Services (2020) (Teaching Hospitals)	Lists teaching hospitals along with address (2020)

Table 8.2: A list of hospital-level datasets contained within in our COVID-19 repository. Currently, all hospital-level sources are static. See Table 8.1 for an overview of select features from these hospital-level datasets.

(i) daily recorded COVID-19-related case count and (recorded) death count by The New York Times (2020) (hereafter NYT (2020)) and USAFacts (2020);

(ii) demographic features such as population distribution by age and population density;

(iii) socioeconomic factors including poverty levels, unemployment, education, and social vulnerability measures;

(iv) health resource availability such as the number of hospitals, ICU beds, and medical staff;

(v) health risk indicators including heart disease, chronic respiratory disease, smoking, obesity, and diabetes prevalence;

(vi) mobility measures such as the percent change in mobility from a pre-COVID-19 baseline; and

(vii) other relevant information such as county-level presidential election results from 2000 to 2016, county-level commute data that includes the number of workers in the commuting flow, and airline ticket survey data that includes origin, destination, and other itinerary details.

In total, there are over 8000 features in the county-level dataset. We provide a feature-level snapshot of the different types of county-level data available in our repository in Table 8.3. Alternatively, in Table 8.4, we provide an overview of the county-level data sources in our repository, organized by the dataset.

Table 8.3: A list of select relevant features from across all county-level datasets contained in our COVID-19 repository grouped by feature topic. See Table 8.4 for an overview of each of the individual county-level datasets.

Description of County-level Features	Data Source(s)
COVID-19 Cases/Deaths	
Daily # of COVID-19-related recorded cases by US county	USAFacts (2020); The New York Times (2020)
Daily # of COVID-19-related deaths by US county	USAFacts (2020); The New York Times (2020)
Demographics	
Population estimate by county (2018)	Health Resources and Services Administration (2019) (Area Health Resources Files)
Census population by county (2010)	Health Resources and Services Administration (2019) (Area Health Resources Files)
Age 65+ population estimate by county (2017)	Health Resources and Services Administration (2019) (Area Health Resources Files)

Table 8.3: *(continued)*

Description of County-level Features	Data Source(s)
Median age by county (2010)	Health Resources and Services Administration (2019) (Area Health Resources Files)
Population density per square mile by county (2010)	Health Resources and Services Administration (2019) (Area Health Resources Files)
Socioeconomic Factors	
% uninsured by county (2017)	County Health Rankings & Roadmaps (2020)
High school graduation rate by county (2016-17)	County Health Rankings & Roadmaps (2020)
Unemployment rate by county (2018)	County Health Rankings & Roadmaps (2020)
% with severe housing problems in each county (2012-16)	County Health Rankings & Roadmaps (2020)
Poverty rate by county (2018)	United States Department of Agriculture, Economic Research Service (2018)
Median household income by county (2018)	United States Department of Agriculture, Economic Research Service (2018)
Social vulnerability index for each county	Centers for Disease Control and Prevention et al. (2018) (Social Vulnerability Index)
Health Resources Availability	
# of hospitals in each county	Kaiser Health News (2020)
# of ICU beds in each county	Kaiser Health News (2020)
# of full-time hospital employees in each county (2017)	Health Resources and Services Administration (2019) (Area Health Resources Files)
# of MDs in each county (2017)	Health Resources and Services Administration (2019) (Area Health Resources Files)
Health Risk Factors	
Heart disease mortality rate by county (2014-16)	Centers for Disease Control and Prevention (2018) (Interactive Atlas of Heart Disease and Stroke)
Stroke mortality rate by county (2014-16)	Centers for Disease Control and Prevention (2018) (Interactive Atlas of Heart Disease and Stroke)
Diabetes prevalence by county (2016)	Centers for Disease Control and Prevention et al. (2016) (Diagnosed Diabetes Atlas)
Chronic respiratory disease mortality rate by county (2014)	Institute for Health Metrics and Evaluation (2017)
% of smokers by county (2017)	County Health Rankings & Roadmaps (2020)
% of adults with obesity by county (2016)	County Health Rankings & Roadmaps (2020)
Crude mortality rate by county (2012-16)	United States Department of Health and Human Services et al. (2017)
Mobility	
Start date of stay at home order by county	Killeen et al. (2020)
% change in mobility at parks, workplaces, transits, groceries/pharmacies, residential, and retail/recreational areas	Google LLC (2020)

Table 8.4: A list of county-level datasets contained within in our COVID-19 repository grouped by data category. Several datasets are relevant to multiple categories and are thus listed multiple times. See Table 8.3 for an overview of select features from these county-level datasets.

County-level Dataset	Description
COVID-19 Cases/Deaths Data	
USAFacts (2020)	Daily cumulative number of reported COVID-19-related death and case counts by US county, dating back to Jan. 23, 2020
The New York Times (2020)	Similar to the USAFacts dataset, but includes aggregated death counts in New York City without county breakdowns
Demographics and Socioeconomic Factors	
Health Resources and Services Administration (2019) (Area Health Resources Files)	Includes data on health facilities, professions, resource scarcity, economic activity, and socioeconomic factors (2018-2019)
County Health Rankings & Roadmaps (2020)	Estimates of various health behaviors and socioeconomic factors (e.g., unemployment, education)
Centers for Disease Control and Prevention et al. (2018) (Social Vulnerability Index)	Reports the CDC's measure of social vulnerability from 2018
United States Department of Agriculture, Economic Research Service (2018)	Poverty estimates and median household income for each county
Health Resources Availability	
Health Resources and Services Administration (2019) (Area Health Resources Files)	Includes data on health facilities, professions, resource scarcity, economic activity, and socioeconomic factors (2018-2019)
Health Resources and Services Administration (2020) (Health Professional Shortage Areas)	Provides data on areas having shortages of primary care, as designated by the Health Resources & Services Administration
Kaiser Health News (2020)	# of hospitals, hospital employees, and ICU beds in each county
Health Risk Factors	
County Health Rankings & Roadmaps (2020)	Estimates of various socioeconomic factors and health behaviors (e.g., % of adult smokers, % of adults with obesity)
Centers for Disease Control and Prevention (2018) (Interactive Atlas of Heart Disease and Stroke)	Estimated heart disease and stroke death rate per 100,000 (all ages, all races/ethnicities, both genders, 2014-2016)
Centers for Disease Control and Prevention et al. (2016) (Diagnosed Diabetes Atlas)	Estimated percentage of people who have been diagnosed with diabetes per county (2016)
Institute for Health Metrics and Evaluation (2017)	Estimated mortality rates of chronic respiratory diseases (1980-2014)

Table 8.4: *(continued)*

County-level Dataset	Description
Centers for Medicare & Medicaid Services (2017) (Chronic Conditions)	Prevalence of 21 chronic conditions based upon CMS administrative enrollment and claims data for Medicare beneficiaries
United States Department of Health and Human Services et al. (2017)	Overall mortality rates (2012-2016) for each county from the National Center for Health Statistics
Mobility	
Killeen et al. (2020) (JHU Date of Interventions)	Dates that counties (or states governing them) took measures to mitigate the spread by restricting gatherings
Google LLC (2020) (Google Community Mobility Reports)	Reports relative movement trends over time by geography and across different categories of places (e.g., retail/recreation, groceries/pharmacies)
Apple Inc (2020) (Apple Mobility Trends)	Uses Apple maps data to report relative (to Jan. 13, 2020) volume of directions requests per country/region, sub-region or city
Miscellaneous	
United States Census Bureau (2018) (County Adjacency File)	Lists each US county and its neighboring counties; from the US Census
Bureau of Transportation Statistics (2020) (Airline Origin and Destination Survey)	Survey data with origin, destination, and itinerary details from a 10% sample of airline tickets in 2019
MIT Election Data and Science Lab (2018) (County Presidential Data)	County-level returns for presidential elections from 2000 to 2016 according to official state election data records

8.2 Discussion

The full corpus of data, along with further details and extensive documentation, are available on GitHub. In particular, we have created a comprehensive data dictionary with the available data features, their descriptions, and source dataset for ease of navigation.[1] We have also provided a quick-start guide for accessing the unabridged county-level and hospital-level datasets with a single Python code line.

While the COVID-19 data repository was originally created to facilitate our work with Response4Life to aid medical supply allocation efforts, the data repository has continually grown over time. Consequently, our COVID-19 data repository supports investigations into a wide range of COVID-19-related questions. For instance, using the breadth of travel informa-

[1] `https://github.com/Yu-Group/covid19-severity-prediction/blob/master/data/list_of_columns.md`

tion in our repository, including (aggregated) air travel and work commute data, researchers can investigate the impact of both local and between-city travel patterns on the spread of COVID-19. Our repository also includes data on the prevalence of various COVID-19 health risk factors, including diabetes, heart disease, and chronic respiratory disease, which can be used to stratify counties. Furthermore, our data enables an investigation into the connection between socioeconomic and demographic information with health resource data (e.g., number of ICU beds, medical staff) to gain a deeper understanding of the severity of the pandemic at the county-level. Stratification using these covariates is particularly crucial for assessing the COVID-19 status of rural communities, which are not directly comparable, both in terms of people and resources, to the larger cities and counties that have received more attention.

Beyond research, our COVID-19 data repository was widely used by researchers at other universities and classrooms for final projects (e.g., a graduate-level machine learning course at the University of Illinois Urbana-Champaign, DATA100 at the University of California, Berkeley with over 1000 students). At its peak, we had around 12,000 visits, 1,100 unique visitors, and 108 clones on GitHub over the span of two weeks.

Bibliography

Agarwal, A., Y. S. Tan, O. Ronen, C. Singh, and B. Yu (2022). Hierarchical shrinkage: Improving the accuracy and interpretability of tree-based models. In *International Conference on Machine Learning*, pp. 111–135. PMLR.

Alizadeh, E., W. Xu, J. Castle, J. Foss, and A. Prasad (2019, June). TISMorph: A tool to quantify texture, irregularity and spreading of single cells. *PLoS One 14*(6), e0217346.

Altieri, N., R. L. Barter, J. Duncan, R. Dwivedi, K. Kumbier, X. Li, R. Netzorg, B. Park, C. Singh, Y. S. Tan, T. Tang, Y. Wang, C. Zhang, and B. Yu (2020, 11). Curating a covid-19 data repository and forecasting county-level death counts in the united states. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/p6isyf0g.

Anderson, K., S. Mongin, R. Sinha, R. Stolzenberg-Solomon, M. Gross, R. Ziegler, J. Mabie, A. Risch, S. Kazin, and T. Church (2012). Pancreatic cancer risk: Associations with meat-derived carcinogen intake in the prostate, lung, colorectal, and ovarian cancer screening trial (plco) cohort. *Molecular carcinogenesis 51*, 128–137.

Apple Inc (2020). Apple Mobility Trends Reports. Accessed on 05-15-2020 at `https://www.apple.com/covid19/mobility`.

Arlt, A., H. Schäfer, and H. Kalthoff (2012). The 'n-factors' in pancreatic cancer: functional relevance of nf-κb, nfat and nrf2 in pancreatic cancer. *Oncogenesis 1*(11), e35–e35.

Atlas, T. C. G. (2012). Comprehensive molecular portraits of human breast tumours. *Nature 490*(7418), 61–70.

Attias-Geva, Z., I. Bentov, A. Fishman, H. Werner, and I. Bruchim (2011). Insulin-like growth factor-i receptor inhibition by specific tyrosine kinase inhibitor nvp-aew541 in endometrioid and serous papillary endometrial cancer cell lines. *Gynecologic oncology 121*(2), 383–389.

Aung, N., J. D. Vargas, C. Yang, C. P. Cabrera, H. R. Warren, K. Fung, E. Tzanis, M. R. Barnes, J. I. Rotter, K. D. Taylor, A. W. Manichaikul, J. A. C. Lima, D. A. Bluemke, S. K. Piechnik, S. Neubauer, P. B. Munroe, and S. E. Petersen (2019, October). Genome-Wide analysis of left ventricular Image-Derived phenotypes identifies fourteen loci associated

with cardiac morphogenesis and heart failure development. *Circulation 140*(16), 1318–1330.

Azadkia, M. and S. Chatterjee (2021). A simple measure of conditional dependence. *The Annals of Statistics 49*(6), 3070–3102.

Azuaje, F. (2016). Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics 18*(5), 820–829.

Bai, W., M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, V. Carapella, Y. J. Kim, H. Suzuki, B. Kainz, P. M. Matthews, S. E. Petersen, S. K. Piechnik, S. Neubauer, B. Glocker, and D. Rueckert (2018, September). Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J. Cardiovasc. Magn. Reson. 20*(1), 65.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature 533*(7604).

Balmanno, K., S. D. Chell, A. S. Gillings, S. Hayat, and S. J. Cook (2009). Intrinsic resistance to the mek1/2 inhibitor azd6244 (arry-142886) is associated with weak erk1/2 signalling and/or strong pi3k signalling in colorectal cancer cell lines. *International journal of cancer 125*(10), 2332–2341.

Barbour, S., K. Nakashima, J.-B. Zhang, S. Tangada, C.-L. Hahn, H. Schenkein, and J. Tew (1997). Tobacco and smoking: environmental factors that modify the host response (immune system) and have an impact on periodontal health. *Critical Reviews in Oral Biology & Medicine 8*, 437–460.

Barretina, J., G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature 483*(7391), 603–607.

Bastos, M. B., D. Burkhoff, J. Maly, J. Daemen, C. A. d. Uil, K. Ameloot, M. Lenzen, F. Mahfoud, F. Zijlstra, J. J. Schreuder, and N. M. Van Mieghem (2019). Invasive left ventricle pressure–volume analysis: overview and practical clinical implications. *Eur. Heart J. 41*(12), 1286–1297.

Basu, S., K. Kumbier, J. B. Brown, and B. Yu (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences 115*(8), 1943–1948.

Begley, C. G. and L. M. Ellis (2012). Raise standards for preclinical cancer research. *Nature 483*(7391), 531–533.

Behr, M., K. Kumbier, A. Cordova-Palomera, M. Aguirre, E. Ashley, A. J. Butte, R. Arnaout, B. Brown, J. Priest, and B. Yu (2020). Learning epistatic polygenic phenotypes with boolean interactions. *bioRxiv*, 2020–11.

Behr, M., Y. Wang, X. Li, and B. Yu (2021). Provable boolean interaction recovery from tree ensemble obtained via random forests. *arXiv preprint arXiv:2102.11800*.

Bénard, C., S. Da Veiga, and E. Scornet (2021). Mda for random forests: inconsistency, and a practical solution via the sobol-mda. *arXiv preprint arXiv:2102.13347*.

Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal 13*(2), 208.

Bergamaschi, A., Z. Madak-Erdogan, Y. J. Kim, Y.-L. Choi, H. Lu, and B. S. Katzenellenbogen (2014). The forkhead transcription factor foxm1 promotes endocrine resistance and invasiveness in estrogen receptor-positive breast cancer by expansion of stem-like cancer cells. *Breast Cancer Research 16*, 1–18.

Bhatt, A., M. Redinbo, and S. Bultman (2017). The role of the microbiome in cancer development and therapy. *CA Cancer J Clin 67*, 326–344.

Bluemke, D. A., R. A. Kronmal, J. A. C. Lima, K. Liu, J. Olson, G. L. Burke, and A. R. Folsom (2008). The Relationship of Left Ventricular Mass and Geometry to Incident Cardiovascular Events The MESA (Multi-Ethnic Study of Atherosclerosis) Study. *J. Am. Coll. Cardiol. 52*(25), 2148–2155.

Boulesteix, A.-L., A. Bender, J. Lorenzo Bermejo, and C. Strobl (2012). Random forest gini importance favours snps with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics 13*(3), 292–304.

Bray, F., J. Ferlay, I. Soerjomataram, R. Siegel, L. Torre, and A. Jemal (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin 68*, 394–424.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and Regression Trees*. CRC press.

Brennan, C. and W. Garrett (2016). Gut microbiota, inflammation, and colorectal cancer. *Annu Rev Microbiol 70*, 395–411.

Brett, J. O., L. M. Spring, A. Bardia, and S. A. Wander (2021). Esr1 mutation as an emerging clinical biomarker in metastatic hormone receptor-positive breast cancer. *Breast Cancer Research 23*, 1–15.

Brunt, V., T. LaRocca, A. Bazzoni, Z. Sapinsley, J. Miyamoto-Ditmon, R. Gioscia-Ryan, A. Neilson, C. Link, and D. Seals (2021). The gut microbiome–derived metabolite trimethylamine n-oxide modulates neuroinflammation and cognitive function with aging. *GeroScience 43*, 377–394.

Bureau of Transportation Statistics (2020). Airline origin and destination survey (db1b). Accessed on 04-20-2020 at `https://transtats.bts.gov/Databases.asp?Mode_ID=1&Mode_Desc=Aviation&Subject_ID2=0`.

Burkhoff, D., I. Mirsky, and H. Suga (2005). Assessment of systolic and diastolic ventricular properties via pressure-volume analysis: a guide for clinical, translational, and basic researchers. *American Journal of Physiology-Heart and Circulatory Physiology 289*(2), H501–H512.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature 562*(7726), 203–209.

Caan, B. J., C. Sweeney, L. A. Habel, M. L. Kwan, C. H. Kroenke, E. K. Weltzien, C. P. Quesenberry Jr, A. Castillo, R. E. Factor, L. H. Kushi, et al. (2014). Intrinsic subtypes from the pam50 gene expression assay in a population-based breast cancer survivor cohort: prognostication of short-and long-term outcomes. *Cancer epidemiology, biomarkers & prevention 23*(5), 725–734.

Camidge, D. R., S.-H. I. Ou, G. Shapiro, G. A. Otterson, L. C. Villaruz, M. A. Villalona-Calero, A. J. Iafrate, M. Varella-Garcia, S. Dacic, S. Cardarella, et al. (2014). Efficacy and safety of crizotinib in patients with advanced c-met-amplified non-small cell lung cancer (nsclc).

Candel, A., V. Parmar, E. LeDell, and A. Arora (2016). *Deep learning with H2O*. H2O ai Inc.

Canto, M., J. Almario, R. Schulick, C. Yeo, A. Klein, A. Blackford, E. Shin, A. Sanyal, G. Yenokyan, and A. Lennon (2018). Risk of neoplastic progression in individuals at high risk for pancreatic cancer undergoing long-term surveillance. *Gastroenterology 155*, 740–751 742.

Caponigro, G. and W. R. Sellers (2011). Advances in the preclinical testing of cancer therapeutic hypotheses. *Nature reviews Drug discovery 10*(3), 179.

Caruana, R., N. Karampatziakis, and A. Yessenalina (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine learning*, pp. 96–103.

Centers for Disease Control and Prevention (2018). HIV in the United States and Dependent Areas. Accessed on 10-24-2020 at `http://nccd.cdc.gov/DHDSPAtlas`.

Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry, and Geospatial Research, Analysis, and Services Program (2018). Social Vulnerability Index Database. Accessed on 04-03-2020 at `https://svi.cdc.gov/data-and-tools-download.html`.

Centers for Disease Control and Prevention, Division of Diabetes Translation, and US Diabetes Surveillance System (2016). Diagnosed diabetes atlas. Accessed on 04-02-2020 at `https://www.cdc.gov/diabetes/data`.

Centers for Medicare & Medicaid Services (2017). Chronic Conditions Prevalence State/County Level: All Beneficiaries by Age, 2007-2017. Accessed on 04-02-2020 at `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CC_Main`.

Centers for Medicares & Medicaid Services (2018). Case mix index file. Accessed on 04-01-2020 at `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2020-IPPS-Final-Rule-Home-Page-Items/FY2020-IPPS-Final-Rule-Data-Files`.

Centers for Medicares & Medicaid Services (2020). 2020 reporting cycle: Teaching hospital list. Accessed on 04-01-2020 at `https://www.cms.gov/OpenPayments/Downloads/2020-Reporting-Cycle-Teaching-Hospital-List-PDF-.pdf`.

Chalmers, I. and P. Glasziou (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet 374*(9683), 86–89.

Chandarlapaty, S., D. Chen, W. He, P. Sung, A. Samoila, D. You, T. Bhatt, P. Patel, M. Voi, M. Gnant, et al. (2016). Prevalence of esr1 mutations in cell-free dna and outcomes in metastatic breast cancer: a secondary analysis of the bolero-2 clinical trial. *JAMA oncology 2*(10), 1310–1315.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee (2015, February). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience 4*, 7.

Chang, W. (2022). *R6: Encapsulated Classes with Reference Semantics*. https://r6.r-lib.org, https://github.com/r-lib/R6/.

Chen, K., C. Zhou, Y. He, J. Liu, and X. Yang (2021). Metabolomics profiling of eus-fna sample predicts advanced pancreatic adenocarcinoma prognosis. In, (Research Square).

Chen, L. and J. Pan (2017). Dual cyclin-dependent kinase 4/6 inhibition by pd-0332991 induces apoptosis and senescence in oesophageal squamous cell carcinoma cells. *British journal of pharmacology 174*(15), 2427–2443.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chen, X. and H. Ishwaran (2012). Random forests for genomic data analysis. *Genomics 99*(6), 323–329.

Chrispin, J., A. Jain, E. Z. Soliman, E. Guallar, A. Alonso, S. R. Heckbert, D. A. Bluemke, J. A. C. Lima, and S. Nazarian (2014). Association of electrocardiographic and imaging surrogates of left ventricular hypertrophy with incident atrial fibrillation.

Cohen, H., R. Ben-Hamo, M. Gidoni, I. Yitzhaki, R. Kozol, A. Zilberberg, and S. Efroni (2014). Shift in gata3 functions, and gata3 mutations, control progression and clinical presentation in breast cancer. *Breast Cancer Research 16*, 1–16.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature 489*(7414), 57.

Consortium, H. (2012). Structure, function and diversity of the healthy human microbiome. *nature 486*, 207.

Cordero, P., V. N. Parikh, E. T. Chin, A. Erbilgin, M. J. Gloudemans, C. Shang, Y. Huang, A. C. Chang, K. S. Smith, F. Dewey, K. Zaleta, M. Morley, J. Brandimarto, N. Glazer, D. Waggott, A. Pavlovic, M. Zhao, C. S. Moravec, W. H. W. Tang, J. Skreen, C. Malloy, S. Hannenhalli, H. Li, S. Ritter, M. Li, D. Bernstein, A. Connolly, H. Hakonarson, A. J. Lusis, K. B. Margulies, A. A. Depaoli-Roach, S. B. Montgomery, M. T. Wheeler, T. Cappola, and E. A. Ashley (2019, June). Pathologic gene network rewiring implicates PPP1R3A as a central regulator in pressure overload heart failure. *Nat. Commun. 10*(1), 2760.

Costello, J. C., L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-Ud-Din, P. Hintsanen, S. A. Khan, et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology 32*(12), 1202–1212.

County Health Rankings & Roadmaps (2020). County Health Rankings & Roadmaps 2020 Measures. Accessed on 04-02-2020 at `https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/2020-measures`.

Currey, L., S. Thor, and M. Piper (2021, June). TEAD family transcription factors in development and disease. *Development 148*(12).

Da Silva, F., A. S. Rocha, F. J. Motamedi, F. Massa, C. Basboga, H. Morrison, K. D. Wagner, and A. Schedl (2017, August). Coronary artery formation is driven by localized expression of r-spondin3. *Cell Rep. 20*(8), 1745–1754.

Dainis, A., K. Zaleta-Rivera, A. Ribeiro, A. C. H. Chang, C. Shang, F. Lan, P. W. Burridge, W. R. Liu, J. C. Wu, A. C. Y. Chang, B. L. Pruitt, M. Wheeler, and E. Ashley (2020, July). Silencing of MYH7 ameliorates disease phenotypes in human iPSC-cardiomyocytes. *Physiol. Genomics 52*(7), 293–303.

Dang, L., D. White, S. Gross, B. Bennett, M. Bittinger, E. Driggers, V. Fantin, H. Jang, S. Jin, and M. Keenan (2009). Cancer-associated idh1 mutations produce 2-hydroxyglutarate. *Nature 462*, 739–744.

De Rosa, A., A. Fontana, T. Nuzzo, M. Garofalo, A. Di Maio, D. Punzo, M. Copetti, A. Bertolino, F. Errico, A. Rampino, et al. (2022). Machine learning algorithm unveils glutamatergic alterations in the post-mortem schizophrenia brain. *Schizophrenia 8*(1), 8.

Definitive Healthcare (2020). Definitive Healthcare: USA Hospital Beds. Accessed on 04-01-2020 at `https://coronavirus-resources.esri.com/datasets/definitivehc::definitive-healthcare-usa-hospital-beds`.

DeLong, E., D. DeLong, and D. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics 44*, 837–845.

Den Hoed, M., M. Eijgelsheim, T. Esko, B. J. Brundel, D. S. Peal, D. M. Evans, I. M. Nolte, A. V. Segrè, H. Holm, R. E. Handsaker, et al. (2013). Identification of heart rate–associated loci and their effects on cardiac conduction and rhythm disorders. *Nature genetics 45*(6), 621–631.

Deplancke, B., D. Alpern, and V. Gardeux (2016, July). The genetics of transcription factor DNA binding variation. *Cell 166*(3), 538–554.

Di Carlo, D. (2009, November). Inertial microfluidics. *Lab Chip 9*(21), 3038–3046.

Drakos, E., R. R. Singh, G. Rassidakis, E. Schlette, J. Li, F.-X. Claret, R. Ford, F. Vega, and L. J. Medeiros (2011). Activation of the p53 pathway by the mdm2 inhibitor nutlin-3a overcomes bcl2 overexpression in a preclinical model of diffuse large b-cell lymphoma associated with t (14; 18)(q32; q21). *Leukemia 25*(5), 856–867.

Du Bois, D. and E. F. Du Bois (1916). Clinical calorimetry: tenth paper a formula to estimate the approximate surface area if height and weight be known. *Archives of internal medicine 17*(6_2), 863–871.

Dwivedi, R., Y. Tan, B. Park, M. Wei, K. Horgan, D. Madigan, and B. Yu (2020). Stable discovery of interpretable subgroups via calibration in causal studies. *International Statistical Review 88*, 135– 178.

Ehret, G. B., T. Ferreira, D. I. Chasman, A. U. Jackson, E. M. Schmidt, T. Johnson, G. Thorleifsson, J. Luan, L. A. Donnelly, S. Kanoni, et al. (2016). The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nature genetics 48*(10), 1171–1184.

Emery, C. M., K. G. Vijayendran, M. C. Zipser, A. M. Sawyer, L. Niu, J. J. Kim, C. Hatton, R. Chopra, P. A. Oberholzer, M. B. Karpova, et al. (2009). Mek1 mutations confer resistance to mek and b-raf inhibition. *Proceedings of the National Academy of Sciences 106*(48), 20411–20416.

Epifanio, I. (2017). Intervention in prediction measure: a new approach to assessing variable importance for random forests. *BMC bioinformatics 18*(1), 1–16.

Eschenhagen, T. and L. Carrier (2019, May). Cardiomyopathy phenotypes in human-induced pluripotent stem cell-derived cardiomyocytes-a systematic review. *Pflugers Arch. 471*(5), 755–768.

Esteva, F. J., D. Yu, M.-C. Hung, and G. N. Hortobagyi (2010). Molecular predictors of response to trastuzumab and lapatinib in breast cancer. *Nature reviews Clinical oncology 7*(2), 98.

Fahrmann, J., L. Bantis, M. Capello, G. Scelo, J. Dennison, N. Patel, E. Murage, J. Vykoukal, D. Kundnani, and L. Foretova (2019). A plasma-derived protein-metabolite multiplexed panel for early-stage pancreatic cancer. *J Natl Cancer Inst 111*, 372–379.

Fahrmann, J., E. Irajizad, M. Kobayashi, J. Vykoukal, J. Dennison, E. Murage, R. Wu, J. Long, K. Do, and J. Celestino (2021). A myc-driven plasma polyamine signature for early detection of ovarian cancer. *Cancers (Basel 13*.

Fahrmann, J., C. Schmidt, X. Mao, E. Irajizad, M. Loftus, J. Zhang, N. Patel, J. Vykoukal, J. Dennison, and J. Long (2021). Lead-time trajectory of ca19-9 as an anchor marker for pancreatic cancer early detection. *Gastroenterology 160*, 1373–1383. e1376.

Fahrmann, J., J. Vykoukal, A. Fleury, S. Tripathi, J. Dennison, E. Murage, P. Wang, C. Yu, M. Capello, and C. Creighton (2020). Association between plasma diacetylspermine and tumor spermine synthase with outcome in triple-negative breast cancer. *J Natl Cancer Inst 112*, 607–616.

Faizal, A. S. M., T. M. Thevarajah, S. M. Khor, and S.-W. Chang (2021). A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. *Computer methods and programs in biomedicine 207*, 106190.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Farrow, B. and B. Evers (2002). Inflammation and the development of pancreatic cancer. *Surgical oncology 10*, 153–169.

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research 15*(1), 3133–3181.

Force, U. (2019). Screening for pancreatic cancer: Us preventive services task force reaffirmation recommendation statement. *JAMA 322*, 438–444.

Friedman, J. H. and B. E. Popescu (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics 2*(3), 916–954.

Fröhlich, H., R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, et al. (2018). From hype to reality: data science enabling personalized medicine. *BMC medicine 16*(1), 1–15.

Fu, X., R. Pereira, C. De Angelis, J. Veeraraghavan, S. Nanda, L. Qin, M. L. Cataldo, V. Sethunath, S. Mehravaran, C. Gutierrez, et al. (2019). Foxa1 upregulation promotes enhancer and transcriptional reprogramming in endocrine-resistant breast cancer. *Proceedings of the National Academy of Sciences 116*(52), 26823–26834.

Fukuda, T., S. Sugita, R. Inatome, and S. Yanagi (2010, December). CAMDI, a novel disrupted in schizophrenia 1 (DISC1)-binding protein, is required for radial migration*. *J. Biol. Chem. 285*(52), 40554–40561.

Gan, L., L. Zheng, and G. I. Allen (2022). Inference for interpretable machine learning: Fast, model-agnostic confidence intervals for feature importance. *arXiv preprint arXiv:2206.02088*.

Garczyk, S., S. von Stillfried, W. Antonopoulos, A. Hartmann, M. G. Schrauder, P. A. Fasching, T. Anzeneder, A. Tannapfel, Y. Ergönenc, R. Knüchel, et al. (2015). Agr3 in breast cancer: prognostic impact and suitable serum-based biomarker for early cancer detection. *PloS one 10*(4), e0122106.

Genuer, R., J.-M. Poggi, and C. Tuleau (2008). Random forests: some methodological insights. *arXiv preprint arXiv:0811.3619*.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot (2010). Variable selection using random forests. *Pattern recognition letters 31*(14), 2225–2236.

Gifford, C. A., S. S. Ranade, R. Samarakoon, H. T. Salunga, T. Y. de Soysa, Y. Huang, P. Zhou, A. Elfenbein, S. K. Wyman, Y. K. Bui, K. R. Cordes Metzler, P. Ursell, K. N. Ivey, and D. Srivastava (2019, May). Oligogenic inheritance of a human heart disease involving a genetic modifier. *Science 364*(6443), 865–870.

Google LLC (2020). Google COVID-19 Community Mobility Reports. Accessed on 05-15-2020 at `https://www.google.com/covid19/mobility/`.

Gregorutti, B., B. Michel, and P. Saint-Pierre (2017). Correlation and variable importance in random forests. *Statistics and Computing 27*(3), 659–678.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician 63*(4), 308–319.

Grothues, F., G. C. Smith, J. C. Moon, N. G. Bellenger, P. Collins, H. U. Klein, and D. J. Pennell (2002). Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy. *The American journal of cardiology 90*(1), 29–34.

GTEx Consortium (2013, June). The Genotype-Tissue expression (GTEx) project. *Nat. Genet. 45*(6), 580–585.

Guan, G., L. Wu, A. A. Bhagat, Z. Li, P. C. Y. Chen, S. Chao, C. J. Ong, and J. Han (2013). Spiral microchannel with rectangular and trapezoidal cross-sections for size based particle separation. *Sci. Rep. 3*, 1475.

Guindo-Martínez, M., R. Amela, S. Bonàs-Guarch, M. Puiggròs, C. Salvoro, I. Miguel-Escalada, C. E. Carey, J. B. Cole, S. Rüeger, E. Atkinson, A. Leong, F. Sanchez, C. Ramon-Cortes, J. Ejarque, D. S. Palmer, M. Kurki, FinnGen Consortium, K. Aragam, J. C. Florez, R. M. Badia, J. M. Mercader, and D. Torrents (2021, April). The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun. 12*(1), 2436.

Gupta, V., N. Sharma, B. Durden, V. Garrido, K. Kesh, D. Edwards, D. Wang, C. Myer, B. Mateo-Victoriano, and S. Kollala (2021). Hypoxia-driven oncometabolite l-2hg maintains stemness-differentiation balance and facilitates immune evasion in pancreatic cancer. *Cancer Res 81*, 4001–4013.

Güvenç Paltun, B., H. Mamitsuka, and S. Kaski (2019). Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. *Briefings in Bioinformatics*.

Haider, A. W., M. G. Larson, E. J. Benjamin, and D. Levy (1998, November). Increased left ventricular mass and hypertrophy are associated with increased risk for sudden death. *J. Am. Coll. Cardiol. 32*(5), 1454–1459.

Half, E., N. Keren, L. Reshef, T. Dorfman, I. Lachter, Y. Kluger, N. Reshef, H. Knobler, Y. Maor, and A. Stein (2019). Fecal microbiome signatures of pancreatic cancer patients. *Scientific reports 9*, 1–12.

Han, S., W. Treuren, C. Fischer, B. Merrill, B. DeFelice, J. Sanchez, S. Higginbottom, L. Guthrie, L. Fall, and D. Dodd (2021). A metabolomics pipeline for the mechanistic interrogation of the gut microbiome. *Nature 595*, 415–420.

Harper, A. R., A. Goel, C. Grace, K. L. Thomson, S. E. Petersen, X. Xu, A. Waring, E. Ormondroyd, C. M. Kramer, C. Y. Ho, S. Neubauer, HCMR Investigators, R. Tadros, J. S. Ware, C. R. Bezzina, M. Farrall, and H. Watkins (2021, February). Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. *Nat. Genet. 53*(2), 135–142.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science*, 297–310.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Volume 2. Springer.

He, K., X. Zheng, L. Zhang, and J. Yu (2013). Hsp90 inhibitors promote p53-dependent apoptosis through puma and bax. *Molecular cancer therapeutics 12*(11), 2559–2568.

Health Resources and Services Administration (2019). Area Health Resources Files. Accessed on 04-02-2020 at `https://data.hrsa.gov/data/download`.

Health Resources and Services Administration (2020). Health Professional Shortage Areas - Primary Care. Accessed on 04-04-2020 at `https://data.hrsa.gov/data/download`.

Henderson, Y. C., Y. Chen, M. J. Frederick, S. Y. Lai, and G. L. Clayman (2010). Mek inhibitor pd0325901 significantly reduces the growth of papillary thyroid carcinoma cells in vitro and in vivo. *Molecular cancer therapeutics 9*(7), 1968–1976.

Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler (2004, January). IntAct: an open source molecular interaction database. *Nucleic Acids Res. 32*(Database issue), D452–5.

Hezaveh, K., R. Shinde, A. Klötgen, M. Halaby, S. Lamorte, M. Ciudad, R. Quevedo, L. Neufeld, Z. Liu, and R. Jin (2022). Tryptophan-derived microbial metabolites activate the aryl hydrocarbon receptor in tumor-associated macrophages to suppress anti-tumor immunity. *Immunity 55*, 324–340 328.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Homeland Infrastructure Foundation-Level Data (2020). Hospitals. Accessed on 06-23-2020 at `https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0`.

Hong, J.-H. and H.-G. Zhang (2022). Transcription factors involved in the development and prognosis of cardiac remodeling. *Front. Pharmacol. 13*.

Hood, K., S. Kahkeshani, D. Di Carlo, and M. Roper (2016, August). Direct measurement of particle inertial migration in rectangular microchannels. *Lab Chip 16*(15), 2840–2850.

Hooker, G. and L. Mentch (2019). Please stop permuting features: An explanation and alternatives. *arXiv e-prints*, arXiv–1905.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics 15*(3), 651–674.

Huang, J., L. Butler, Ø. Midttun, R. Wang, A. Jin, Y.-T. Gao, P. Ueland, W.-P. Koh, and J.-M. Yuan (2017). Serum choline, methionine, betaine, dimethylglycine, and trimethylamine-n-oxide in relation to pancreatic cancer risk in two nested case-control studies in asian populations. *Cancer Research 77*(13_Supplement), 2273–2273.

Huang, J. Y., H. N. Luu, L. M. Butler, Ø. Midttun, A. Ulvik, R. Wang, A. Jin, Y.-T. Gao, Y. Tan, P. M. Ueland, et al. (2020). A prospective evaluation of serum methionine-related metabolites in relation to pancreatic cancer risk in two prospective cohort studies. *International Journal of Cancer 147*(7), 1917–1927.

Hubbard, T., I. Murray, W. Bisson, T. Lahoti, K. Gowda, S. Amin, A. Patterson, and G. Perdew (2015). Adaptation of the human aryl hydrocarbon receptor to sense microbiota-derived indoles. *Scientific Reports 5*, 12689.

Huybrechts, I., S. Zouiouich, A. Loobuyck, Z. Vandenbulcke, E. Vogtmann, S. Pisanu, I. Iguacel, A. Scalbert, I. Indave, and V. Smelov (2020). The human microbiome in relation to cancer risk: A systematic review of epidemiologic studies. *Cancer Epidemiol Biomarkers Prev 29*, 1856–1868.

Hylemon, P., S. Harris, and J. Ridlon (2018). Metabolism of hydrogen gases and bile acids in the gut microbiome. *FEBS letters 592*, 2070–2082.

Institute for Health Metrics and Evaluation (2017). United States Chronic Respiratory Disease Mortality Rates by County 1980-2014. Accessed on 04-02-2020 at `http://ghdx.healthdata.org/record/ihme-data/united-states-chronic-respiratory-disease-mortality-rates-county-1980-2014`.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics 1*, 519–537.

Ishwaran, H., U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association 105*(489), 205–217.

Jaglin, M., M. Rhimi, C. Philippe, N. Pons, A. Bruneau, B. Goustard, V. Daugé, E. Maguin, L. Naudon, and S. Rabot (2018). Indole, a signaling molecule produced by the gut microbiota, negatively impacts emotional behaviors in rats. *Frontiers in neuroscience 12*, 216.

Jang, I. S., E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Biocomputing 2014*, pp. 63–74. World Scientific.

Jiang, Y., Y. Liu, X. Tan, S. Yu, and J. Luo (2019). Tpx2 as a novel prognostic indicator and promising therapeutic target in triple-negative breast cancer. *Clinical breast cancer 19*(6), 450–455.

Jiao, L., S. Maity, C. Coarfa, K. Rajapakshe, L. Chen, F. Jin, V. Putluri, L. Tinker, Q. Mo, and F. Chen (2019). A prospective targeted serum metabolomics study of pancreatic cancer in postmenopausal women. *Cancer prevention research 12*, 237–246.

Johannes, F., P. Carla, and W. Amanda (2021). A blood-based polyamine signature associated with men1 duodenopancreatic neuroendocrine tumor progression.

Johnson, K. B., W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon (2021). Precision medicine, ai, and the future of personalized health care. *Clinical and translational science 14*(1), 86–93.

Kaiser Health News (2020). ICU Beds by County. Accessed on 04-02-2020 at `https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/`.

Kamiyama, S., A. Kuriyama, and T. Hashimoto (2019). Edwardsiella tarda bacteremia. *Emerging Infectious Diseases 25*, 1817.

Kanai, M., M. Akiyama, A. Takahashi, N. Matoba, Y. Momozawa, M. Ikeda, N. Iwata, S. Ikegawa, M. Hirata, K. Matsuda, M. Kubo, Y. Okada, and Y. Kamatani (2018, March). Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nat. Genet. 50*(3), 390–400.

Kawel-Boehm, N., R. Kronmal, J. Eng, A. Folsom, G. Burke, J. J. Carr, S. Shea, J. A. C. Lima, and D. A. Bluemke (2019, October). Left ventricular mass at MRI and long-term risk of cardiovascular events: The Multi-Ethnic study of atherosclerosis (MESA). *Radiology 293*(1), 107–114.

Kazemitabar, J., A. Amini, A. Bloniarz, and A. S. Talwalkar (2017). Variable importance using decision trees. *Advances in neural information processing systems 30*.

Khurshid, S., J. Lazarte, J. P. Pirruccello, L.-C. Weng, S. H. Choi, A. W. Hall, X. Wang, S. F. Friedman, V. Nauffal, K. J. Biddinger, K. G. Aragam, P. Batra, J. E. Ho, A. A. Philippakis, P. T. Ellinor, and S. A. Lubitz (2023, March). Clinical and genetic associations of deep learning-derived cardiac magnetic resonance-based left ventricular mass. *Nat. Commun. 14*(1), 1558.

Killeen, B. D., J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, and M. Unberath (2020). A county-level dataset for informing the United States' response to COVID-19. *arXiv preprint arXiv:2004.00756*.

Kiss, B., E. Mikó, É. Sebő, J. Toth, G. Ujlaki, J. Szabó, K. Uray, P. Bai, and P. Árkosy (2020). Oncobiosis and microbial metabolite signaling in pancreatic adenocarcinoma. *Cancers 12*(5), 1068.

Kleeff, J., M. Korc, M. Apte, C. Vecchia, C. Johnson, A. Biankin, R. Neale, M. Tempero, D. Tuveson, R. Hruban, and J. Neoptolemos (2016). Pancreatic cancer. *Nat Rev Dis Primers 2*, 16022.

Klopocki, E., G. Kristiansen, P. J. Wild, I. Klaman, E. Castanos-Velez, G. Singer, R. Stöhr, R. Simon, G. Sauter, H. Leibiger, et al. (2004). Loss of sfrp1 is associated with breast cancer progression and poor prognosis in early stage tumors. *International journal of oncology 25*(3), 641–649.

Klusowski, J. and P. Tian (2021). Nonparametric variable screening with optimal decision stumps. In *International Conference on Artificial Intelligence and Statistics*, pp. 748–756. PMLR.

Klusowski, J. M. (2021). Universal consistency of decision trees in high dimensions. *arXiv preprint arXiv:2104.13881*.

Koch, E. M. and S. R. Sunyaev (2021, November). Maintenance of complex trait variation: Classic theory and modern data. *Front. Genet. 12*, 763363.

Kohane, I. S. (2015). Ten things we have to do to achieve precision medicine. *Science 349*(6243), 37–38.

Kornblith, A. E., C. Singh, G. Devlin, N. Addo, C. J. Streck, J. F. Holmes, N. Kuppermann, J. Grupp-Phelan, J. Fineman, A. J. Butte, et al. (2022). Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. *PLOS Digital Health 1*(8), e0000076.

Kothari, C., M. A. Osseni, L. Agbo, G. Ouellette, M. Déraspe, F. Laviolette, J. Corbeil, J.-P. Lambert, C. Diorio, and F. Durocher (2020). Machine learning analysis identifies genes differentiating triple negative breast cancers. *Scientific reports 10*(1), 10464.

Kuleshov, M. V., M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan (2016, July). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res. 44*(W1), W90–7.

Kumbier, K., S. Basu, J. B. Brown, S. Celniker, and B. Yu (2018). Refining interaction search through signed iterative random forests. *arXiv preprint arXiv:1810.07287*.

Lazzeroni, D., O. Rimoldi, and P. G. Camici (2016, February). From left ventricular hypertrophy to dysfunction and failure. *Circ. J. 80*(3), 555–564.

Le, X.-F. and R. C. Bast, Jr (2011). Src family kinases and paclitaxel sensitivity. *Cancer biology & therapy 12*(4), 260–269.

Lee, J.-H. and J. Lee (2010). Indole as an intercellular signal in microbial communities. *FEMS microbiology reviews 34*, 426–444.

Li, X., T. M. Tang, X. Wang, J.-P. A. Kocher, and B. Yu (2020). A stability-driven protocol for drug response interpretable prediction (stadrip). *arXiv preprint arXiv:2011.06593*.

Li, X., Y. Wang, S. Basu, K. Kumbier, and B. Yu (2019). A debiased mdi feature importance measure for random forests. *Advances in Neural Information Processing Systems 32*.

Li, Y., H. Cho, F. Wang, O. Canela-Xandri, C. Luo, K. Rawlik, S. Archacki, C. Xu, A. Tenesa, Q. Chen, et al. (2020). Statistical and functional studies identify epistasis of cardiovascular risk genomic variants from genome-wide association studies. *Journal of the American Heart Association 9*(7), e014146.

Liang, W.-H., N. Li, Z.-Q. Yuan, X.-L. Qian, and Z.-H. Wang (2019). Dscam-as1 promotes tumor growth of breast cancer by reducing mir-204-5p and up-regulating rrm2. *Molecular carcinogenesis 58*(4), 461–473.

Lim, C. and B. Yu (2016). Estimation stability with cross-validation (escv). *Journal of Computational and Graphical Statistics 25*(2), 464–492.

Lind, A. P. and P. C. Anderson (2019). Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PloS one 14*(7), e0219774.

Ling, X., C. Xu, C. Fan, K. Zhong, F. Li, and X. Wang (2014). Fl118 induces p53-dependent senescence in colorectal cancer cells by promoting degradation of mdmx. *Cancer research 74*(24), 7487–7497.

List, M., A.-C. Hauschild, Q. Tan, T. A. Kruse, J. Baumbach, and R. Batra (2014). Classification of breast cancer subtypes by combining gene expression and dna methylation data. *Journal of integrative bioinformatics 11*(2), 1–14.

Liu, Y., Z. Wang, S. Q. Kwong, E. L. H. Lui, S. L. Friedman, F. R. Li, R. W. C. Lam, G. C. Zhang, H. Zhang, and T. Ye (2011). Inhibition of pdgf, tgf-β, and abl signaling and reduction of liver fibrosis by the small molecule bcr-abl tyrosine kinase antagonist nilotinib. *Journal of hepatology 55*(3), 612–625.

Loecher, M. (2022a). Debiasing mdi feature importance and shap values in tree ensembles. In *Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23–26, 2022, Proceedings*, pp. 114–129. Springer.

Loecher, M. (2022b). Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods 51*(5), 1413–1425.

Loh, P.-R., G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price (2015, March). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet. 47*(3), 284–290.

Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts (2013). Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems 26*.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence 2*(1), 56–67.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*.

Marian, A. J. and E. Braunwald (2017, September). Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circ. Res. 121*(7), 749–770.

Martin, F. J., M. R. Amode, A. Aneja, O. Austine-Orimoloye, A. G. Azov, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai, et al. (2023). Ensembl 2023. *Nucleic Acids Research 51*(D1), D933–D941.

Matsukawa, H., N. Iida, K. Kitamura, T. Terashima, J. Seishima, I. Makino, T. Kannon, K. Hosomichi, T. Yamashita, and Y. Sakai (2021). Dysbiotic gut microbiota in pancreatic cancer patients form correlation networks with the oral microbiota and prognostic factors. *American Journal of Cancer Research 11*, 3163.

Mayers, J., C. Wu, C. Clish, P. Kraft, M. Torrence, B. Fiske, C. Yuan, Y. Bao, M. Townsend, and S. Tworoger (2014). Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nature Medicine 20*, 1193–1198.

McDermott, U., S. V. Sharma, L. Dowell, P. Greninger, C. Montagut, J. Lamb, H. Archibald, R. Raudales, A. Tam, D. Lee, et al. (2007). Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proceedings of the National Academy of Sciences 104*(50), 19936–19941.

Meyer, H. V., T. J. W. Dawes, M. Serrani, W. Bai, P. Tokarczuk, J. Cai, A. d. Marvao, A. Henry, R. T. Lumbers, J. Gierten, T. Thumberger, J. Wittbrodt, J. S. Ware, D. Rueckert, P. M. Matthews, S. K. Prasad, M. L. Costantino, S. A. Cook, E. Birney, and D. P. O'Regan (2020). Genetic and functional insights into the fractal structure of the heart. *Nature 584*(7822), 589–594.

Missiroli, S., M. Perrone, C. Boncompagni, C. Borghi, A. Campagnaro, F. Marchetti, G. Anania, P. Greco, F. Fiorica, P. Pinton, and C. Giorgi (2021). Targeting the NLRP3 Inflammasome as a New Therapeutic Option for Overcoming Cancer. Cancers (Basel.

MIT Election Data and Science Lab (2018). County Presidential Election Returns 2000-2016.

Mordant, P., Y. Loriot, C. Leteur, J. Calderaro, J. Bourhis, M. Wislez, J.-C. Soria, and E. Deutsch (2010). Dependence on phosphoinositide 3-kinase and ras-raf pathways drive the activity of raf265, a novel raf/vegfr2 inhibitor, and rad001 (everolimus) in combination. *Molecular cancer therapeutics 9*(2), 358–368.

Morgan, M. D., E. Pairo-Castineira, K. Rawlik, O. Canela-Xandri, J. Rees, D. Sims, A. Tenesa, and I. J. Jackson (2018, December). Genome-wide study of hair colour in UK biobank explains most of the SNP heritability. *Nat. Commun. 9*(1), 5271.

Morgell, A., J. Reisz, Z. Ateeb, H. Davanian, S. Reinsbach, A. Halimi, R. Gaiser, R. Valente, U. Arnelo, and M. Del Chiaro (2021). Metabolic characterization of plasma and cyst fluid from cystic precursors to pancreatic cancer patients reveal metabolic signatures of bacterial infection. *J Proteome Res 20*, 2725–2738.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences 116*(44), 22071–22080.

Murk, W., M. B. Bracken, and A. T. DeWan (2015, August). Confronting the missing epistasis problem: on the reproducibility of gene-gene interactions. *Hum. Genet. 134*(8), 837–849.

Nagai, A., M. Hirata, Y. Kamatani, K. Muto, K. Matsuda, Y. Kiyohara, T. Ninomiya, A. Tamakoshi, Z. Yamagata, T. Mushiroda, et al. (2017). Overview of the biobank japan project: study design and profile. *Journal of epidemiology 27*(Supplement_III), S2–S8.

Naito, A. T., S. Okada, T. Minamino, K. Iwanaga, M.-L. Liu, T. Sumida, S. Nomura, N. Sahara, T. Mizoroki, A. Takashima, et al. (2010). Promotion of chip-mediated p53 degradation protects the heart from ischemic injury. *Circulation research 106*(11), 1692.

Nelson, D. L., A. L. Lehninger, and M. M. Cox (2008). *Lehninger principles of biochemistry*. Macmillan.

Nembrini, S. (2019). Bias in the intervention in prediction measure in random forests: illustrations and recommendations. *Bioinformatics*.

Nembrini, S., I. R. König, and M. N. Wright (2018). The revival of the gini importance? *Bioinformatics 34*(21), 3711–3718.

Nicodemus, K. K. (2011). On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics 12*(4), 369–373.

Nicodemus, K. K. and J. D. Malley (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics 25*(15), 1884–1890.

Nishioka, C., T. Ikezoe, A. Takeshita, J. Yang, T. Tasaka, Y. Yang, Y. Kuwayama, N. Komatsu, K. Togitani, H. Koeffler, et al. (2007). Zd6474 induces growth arrest and apoptosis of human leukemia cells, which is enhanced by concomitant use of a novel mek inhibitor, azd6244. *Leukemia 21*(6), 1308–1310.

Norland, K., G. Sveinbjornsson, R. B. Thorolfsdottir, O. B. Davidsson, V. Tragante, S. Rajamani, A. Helgadottir, S. Gretarsdottir, J. van Setten, F. W. Asselbergs, J. T. Sverrisson, S. S. Stephensen, G. Oskarsson, E. L. Sigurdsson, K. Andersen, R. Danielsen, G. Thorgeirsson, U. Thorsteinsdottir, D. O. Arnar, P. Sulem, H. Holm, D. F. Gudbjartsson, and K. Stefansson (2019, October). Sequence variants with large effects on cardiac electrophysiology and disease. *Nat. Commun. 10*(1), 4803.

Ohno, E., Y. Hirooka, H. Kawashima, T. Ishikawa, A. Kanamori, H. Ishikawa, Y. Sasaki, K. Nonogaki, K. Hara, and S. Hashimoto (2018). Natural history of pancreatic cystic lesions: A multicenter prospective observational study for evaluating the risk of pancreatic cancer. *J Gastroenterol Hepatol 33*, 320–328.

Olson, R. S., W. L. Cava, Z. Mustahsan, A. Varik, and J. H. Moore (2018). Data-driven advice for applying machine learning to bioinformatics problems. In *Biocomputing 2018: Proceedings of the Pacific Symposium*, pp. 192–203. World Scientific.

Osofsky, J. D. (1997). The effects of exposure to violence on young children (1995). *Carnegie Corporation of New York Task Force on the Needs of Young Children; An earlier version of this article was presented as a position paper for the aforementioned corporation.*.

O'Sullivan, J. W., S. Raghavan, C. Marquez-Luna, J. A. Luzum, S. M. Damrauer, E. A. Ashley, C. J. O'Donnell, C. J. Willer, P. Natarajan, and American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology

and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease (2022, August). Polygenic risk scores for cardiovascular disease: A scientific statement from the american heart association. *Circulation 146*(8), e93–e118.

Oughtred, R., J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, and M. Tyers (2021, January). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. 30*(1), 187–200.

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics 443*(7), 59–72.

Pannala, R., A. Basu, G. Petersen, and S. Chari (2009). New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer. *Lancet Oncol 10*, 88–95.

Pannala, R., J. Leirness, W. Bamlet, A. Basu, G. Petersen, and S. Chari (2008). Prevalence and clinical profile of pancreatic cancer-associated diabetes mellitus. *Gastroenterology 134*, 981–987.

Parker, J. S., M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology 27*(8), 1160.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research 12*, 2825–2830.

Petersen, G. (2016). Familial pancreatic cancer. *Semin Oncol 43*, 548–553.

Pirruccello, J. P., A. Bick, M. Wang, M. Chaffin, S. Friedman, J. Yao, X. Guo, B. A. Venkatesh, K. D. Taylor, W. S. Post, S. Rich, J. A. C. Lima, J. I. Rotter, A. Philippakis, S. A. Lubitz, P. T. Ellinor, A. V. Khera, S. Kathiresan, and K. G. Aragam (2020). Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun. 11*(1), 2254.

Prentice, R. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*, 403–411.

Prorok, P., G. Andriole, R. Bresalier, S. Buys, D. Chia, E. Crawford, R. Fogel, E. Gelmann, F. Gilbert, and M. Hasson (2000). Design of the prostate, lung, colorectal and ovarian (plco) cancer screening trial. *Controlled clinical trials 21*, 273 –309.

Pugin, B., W. Barcik, P. Westermann, A. Heider, M. Wawrzyniak, P. Hellings, C. Akdis, and L. O'Mahony (2017). A wide diversity of bacteria from the human gut produces and degrades biogenic amines. *Microb Ecol Health Dis 28*, 1353881.

Pushalkar, S., M. Hundeyin, D. Daley, C. Zambirinis, E. Kurz, A. Mishra, N. Mohan, B. Aykut, M. Usyk, and L. Torres (2018). The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer discovery 8*, 403–416.

Rad, K. R. and A. Maleki (2020). A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(4), 965–996.

Rahib, L., M. Wehner, L. Matrisian, and K. Nead (2021). Estimated projection of us cancer incidence and death to 2040. *JAMA Network Open 4*, 214708– 214708.

Ramosaj, B. and M. Pauly (2019). Asymptotic unbiasedness of the permutation importance measure in random forest models. *arXiv preprint arXiv:1912.03306*.

Rawla, P., T. Sunkara, and V. Gaduputi (2019). Epidemiology of pancreatic cancer: Global trends, etiology and risk factors. *World journal of oncology 10*, 10–27.

Reimherr, M. and D. L. Nicolae (2011, January). You've gotta be lucky: Coverage and the elusive gene-gene interaction. *Ann. Hum. Genet. 75*(1), 105–111.

Riddick, G., H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine (2011). Predicting in vitro drug sensitivity using random forests. *Bioinformatics 27*(2), 220–224.

Ridlon, J., D. Kang, P. Hylemon, and J. Bajaj (2014). Bile acids and the gut microbiome. *Current opinion in gastroenterology 30*, 332.

Riquelme, E., Y. Zhang, L. Zhang, M. Montiel, M. Zoltan, W. Dong, P. Quesada, I. Sahin, V. Chandra, and A. San Lucas (2019). Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell 178*, 795–806 712.

Risch, H. (2012). Pancreatic cancer: Helicobacter pylori colonization, n-nitrosamine exposures, and abo blood group. *Molecular carcinogenesis 51*, 109–118.

Roager, H. and T. Licht (2018). Microbial tryptophan catabolites in health and disease. *Nature Communications 9*, 3294.

Rogers, M., V. Aveson, B. Firek, A. Yeh, B. Brooks, R. Brower-Sinning, J. Steve, J. Banfield, A. Zureikat, and M. Hogg (2017). Disturbances of the perioperative microbiome across multiple body sites in patients undergoing pancreaticoduodenectomy. *Pancreas 46*, 260–267.

Rubin, M. A. (2015). Health: Make precision medicine work for cancer care. *Nature News 520*(7547), 290.

Rudin, C., C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*.

Ryan, D., T. Hong, and N. Bardeesy (2014). Pancreatic adenocarcinoma. *The New England journal of medicine 371*, 1039–1049.

Saabas, A. (2022). reeinterpreter python package. `https://github.com/andosa/treeinterpreter`.

Sandri, M. and P. Zuccolotto (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics 17*(3), 611–628.

Schaid, D. J., W. Chen, and N. B. Larson (2018, August). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet. 19*(8), 491–504.

Schraivogel, D., T. M. Kuhn, B. Rauscher, M. Rodríguez-Martínez, M. Paulsen, K. Owsley, A. Middlebrook, C. Tischer, B. Ramasz, D. Ordoñez-Rueda, M. Dees, S. Cuylen-Haering, E. Diebold, and L. M. Steinmetz (2022, January). High-speed fluorescence image-enabled cell sorting. *Science 375*(6578), 315–320.

Scornet, E. (2020). Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*.

Seldin, M. M., Y. Meng, H. Qi, W. Zhu, Z. Wang, S. L. Hazen, A. J. Lusis, and D. M. Shih (2016). Trimethylamine n-oxide promotes vascular inflammation through signaling of mitogen-activated protein kinase and nuclear factor-κb. *Journal of the American Heart Association 5*(2), e002767.

Shao, J., Z. Xu, X. Peng, M. Chen, Y. Zhu, L. Xu, H. Zhu, B. Yang, P. Luo, and Q. He (2016). Gefitinib synergizes with irinotecan to suppress hepatocellular carcinoma via antagonizing rad51-mediated dna-repair. *PLoS One 11*(1), e0146968.

Sharir, T., M. D. Feldman, H. Haber, A. M. Feldman, A. Marmor, L. C. Becker, and D. A. Kass (1994). Ventricular systolic assessment in patients with dilated cardiomyopathy by preload-adjusted maximal power. Validation and noninvasive application. *Circulation 89*(5), 2045–2053.

Sharma, A., H. Kandlakunta, S. Nagpal, Z. Feng, W. Hoos, G. Petersen, and S. Chari (2018). Model to determine risk of pancreatic cancer in patients with new-onset diabetes. *Gastroenterology 155*, 730–739 733.

Shwartz-Ziv, R. and A. Armon (2022). Tabular data: Deep learning is not all you need. *Information Fusion 81*, 84–90.

Simidjievski, N., C. Bodnar, I. Tariq, P. Scherer, H. Andres-Terre, Z. Shams, M. Jamnik, et al. (2019). Variational autoencoders for cancer data integration: Design principles and computational practice. *BioRxiv*, 719542.

Singh, C., K. Nasseri, Y. S. Tan, T. Tang, and B. Yu (2021). imodels: a python package for fitting interpretable models. *Journal of Open Source Software 6*(61), 3192.

Spurrell, C. H., I. Barozzi, M. Kosicki, B. J. Mannion, M. J. Blow, Y. Fukuda-Yuzawa, N. Slaven, S. Y. Afzal, J. A. Akiyama, V. Afzal, S. Tran, I. Plajzer-Frick, C. S. Novak, M. Kato, E. A. Lee, T. H. Garvin, Q. T. Pham, A. N. Kronshage, S. Lisgo, J. Bristow, T. P. Cappola, M. P. Morley, K. B. Margulies, L. A. Pennacchio, D. E. Dickel, and A. Visel (2022, September). Genome-wide fetalization of enhancer architecture in heart disease. *Cell Rep. 40*(12), 111400.

Stavrakis, S., G. Holzner, J. Choo, and A. deMello (2019, February). High-throughput microfluidic imaging flow cytometry. *Curr. Opin. Biotechnol. 55*, 36–43.

Stoll, M., R. Kumar, E. Lefkowitz, R. Cron, C. Morrow, and S. Barnes (2016). Fecal metabolomics in pediatric spondyloarthritis implicate decreased metabolic diversity and altered tryptophan metabolism as pathogenic factors. *Genes & Immunity 17*, 400–405.

Strickler, J. H., A. N. Starodub, J. Jia, K. L. Meadows, A. B. Nixon, A. Dellinger, M. A. Morse, H. E. Uronis, P. K. Marcom, S. Y. Zafar, et al. (2012). Phase i study of bevacizumab, everolimus, and panobinostat (lbh-589) in advanced solid tumors. *Cancer chemotherapy and pharmacology 70*(2), 251–258.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC bioinformatics 9*(1), 1–11.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics 8*(1), 1–21.

Stupple, A., D. Singerman, and L. A. Celi (2019). The reproducibility crisis in the age of digital medicine. *NPJ digital medicine 2*(1), 2.

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. (2015). Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine 12*(3), e1001779.

Sun, X., X. Jiao, Y. Ma, Y. Liu, L. Zhang, Y. He, and Y. Chen (2016). Trimethylamine n-oxide induces inflammation and endothelial dysfunction in human umbilical vein endothelial cells via activating ros-txnip-nlrp3 inflammasome. *Biochem Biophys Res Commun 481*, 63–70.

Sutera, A., G. Louppe, V. A. Huynh-Thu, L. Wehenkel, and P. Geurts (2021). From global to local mdi variable importances for random forests and when they are shapley values. *Advances in Neural Information Processing Systems 34*.

Tan, Y. S., A. Agarwal, and B. Yu (2021). A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds.

Thakkar, A. D., H. Raj, D. Chakrabarti, Ravishankar, N. Saravanan, B. Muthuvelan, A. Balakrishnan, and M. Padigaru (2010). Identification of gene expression signature in estrogen receptor positive breast carcinoma. *Biomarkers in cancer 2*, BIC–S3793.

The New York Times (2020). COVID-19 Data in the United States. `https://github.com/nytimes/covid-19-data`. Accessed on 04-01-2020 at `https://github.com/nytimes/covid-19-data`.

Theocharis, A., M. Tsara, N. Papageorgacopoulou, D. Karavias, and D. Theocharis (2000). Pancreatic carcinoma is characterized by elevated content of hyaluronan and chondroitin sulfate with altered disaccharide composition. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1502*, 201–206.

Thorolfsdottir, R. B., G. Sveinbjornsson, H. M. Aegisdottir, S. Benonisdottir, L. Stefansdottir, E. V. Ivarsdottir, G. H. Halldorsson, J. K. Sigurdsson, C. Torp-Pedersen, P. E. Weeke, et al. (2021). Genetic insight into sick sinus syndrome. *European heart journal 42*(20), 1959–1971.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Toy, W., Y. Shen, H. Won, B. Green, R. A. Sakr, M. Will, Z. Li, K. Gala, S. Fanning, T. A. King, et al. (2013). Esr1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nature genetics 45*(12), 1439–1445.

Tsybakov, A. B. (2004). Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the 9*(10).

Tutusaus, A., M. Stefanovic, L. Boix, B. Cucarull, A. Zamora, L. Blasco, P. G. de Frutos, M. Reig, J. C. Fernandez-Checa, M. Marí, et al. (2018). Antiapoptotic bcl-2 proteins determine sorafenib/regorafenib resistance and bh3-mimetic efficacy in hepatocellular carcinoma. *Oncotarget 9*(24), 16701.

Udelson, J. E., R. O. C. 3rd, S. L. Bacharach, T. F. Rumble, and R. O. Bonow (1988). Beta-adrenergic stimulation with isoproterenol enhances left ventricular diastolic performance in hypertrophic cardiomyopathy despite potentiation of myocardial ischemia. Comparison to rapid atrial pacing. *Circulation 79*(2), 371–382.

United States Census Bureau (2018). County Adjacency File. Accessed on 05-15-2020 at `https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html`.

United States Department of Agriculture, Economic Research Service (2018). Poverty estimates for the U.S., states, and counties. Accessed on 04-24-2020 at `https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/`.

United States Department of Health and Human Services, Centers for Disease Control and Prevention , and National Center for Health Statistics (2017). Compressed Mortality File (CMF) on CDC WONDER Online Database, 2012-2016. Accessed on 04-02-2020 at `https://wonder.cdc.gov/cmf-icd10.html`.

USAFacts (2020). COVID-19 Deaths Data. Accessed on 03-31-2020 at `https://www.reuters.com/article/us-health-coronavirus-who/covid-19-spread-map`.

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications 32*(24), 18069–18083.

Verweij, N., Y. J. van de Vegte, and P. van der Harst (2018, March). Genetic study links components of the autonomous nervous system to heart-rate profile during exercise. *Nat. Commun. 9*(1), 898.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics 101*(1), 5–22.

Vykoukal, J., J. Fahrmann, J. Gregg, Z. Tang, S. Basourakos, E. Irajizad, S. Park, G. Yang, C. Creighton, and A. Fleury (2020). Caveolin-1-mediated sphingolipid oncometabolism underlies a metabolic vulnerability of prostate cancer. *Nat Commun 11*, 4279.

Wan, Q. and R. Pal (2014). An ensemble based top performing approach for nci-dream drug sensitivity prediction challenge. *PloS one 9*(6), e101183.

Wang, G., A. Sarkar, P. Carbonetto, and M. Stephens (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(5), 1273–1300.

Wang, K., M. Li, and H. Hakonarson (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research 38*(16), e164–e164.

Wang, Q., A.-A. D. Jones, 3rd, J. A. Gralnick, L. Lin, and C. R. Buie (2019, January). Microfluidic dielectrophoresis illuminates the relationship between microbial cell envelope polarizability and electrochemical activity. *Sci Adv 5*(1), eaat5664.

Warsch, W., K. Kollmann, E. Eckelhart, S. Fajmann, S. Cerny-Reiterer, A. Hölbl, K. V. Gleixner, M. Dworzak, M. Mayerhofer, G. Hoermann, et al. (2011). High stat5 levels mediate imatinib resistance and indicate disease progression in chronic myeloid leukemia. *Blood, The Journal of the American Society of Hematology 117*(12), 3409–3420.

Webber, W., A. Moffat, and J. Zobel (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS) 28*(4), 1–38.

Wei, M.-Y., S. Shi, C. Liang, Q.-C. Meng, J. Hua, Y.-Y. Zhang, J. Liu, B. Zhang, J. Xu, and X.-J. Yu (2019). The microbiota and microbiome in pancreatic cancer: more influential than expected. *Molecular cancer 18*, 1–15.

Weldy, C. S. and E. A. Ashley (2021). Towards precision medicine in heart failure. *Nat. Rev. Cardiol.*, 1–18.

Whitcomb, D. (2004). Inflammation and cancer v. chronic pancreatitis and pancreatic cancer. *American Journal of Physiology-Gastrointestinal and Liver Physiology 287*, 315– 319.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani (2019, November). Welcome to the Tidyverse. *Journal of Open Source Software 4*(43), 1686.

Wlodarska, M., C. Luo, R. Kolde, E. d'Hennezel, J. Annand, C. Heim, P. Krastel, E. Schmitt, A. Omar, and E. Creasey (2017). Indoleacrylic acid produced by commensal peptostreptococcus species suppresses inflammation. *Cell host & microbe 22*, 25–37. e26.

Wu, P.-H., D. M. Gilkes, J. M. Phillip, A. Narkar, T. W.-T. Cheng, J. Marchand, M.-H. Lee, R. Li, and D. Wirtz (2020, January). Single-cell morphology encodes metastatic potential. *Sci Adv 6*(4), eaaw6938.

Wu, S., A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences 113*(16), 4290–4295.

Wu, Y.-J., Y.-J. Jan, B.-S. Ko, S.-M. Liang, and J.-Y. Liou (2015). Involvement of 14-3-3 proteins in regulating tumor progression of hepatocellular carcinoma. *Cancers 7*(2), 1022–1036.

Xie, G., L. Lu, Y. Qiu, Q. Ni, W. Zhang, Y. Gao, H. Risch, H. Yu, and W. Jia (2015). Plasma metabolite biomarkers for the detection of pancreatic cancer. *J Proteome Res 14*, 1195–1202.

Xu, R., Q. Wang, and L. Li (2015). A genome-wide systems analysis reveals strong link between colorectal cancer and trimethylamine n-oxide (tmao), a gut microbial metabolite of dietary meat and fat. *BMC genomics 16*, 1–9.

Yang, J. C., L. Bai, S. Yap, A. C. Gao, H.-J. Kung, and C. P. Evans (2010). Effect of the specific src family kinase inhibitor saracatinib on osteolytic lesions using the pc-3 bone model. *Molecular cancer therapeutics 9*(6), 1629–1637.

Yang, M. G., E. Ling, C. J. Cowley, M. E. Greenberg, and T. Vierbuchen (2022, August). Characterization of sequence determinants of enhancer function using natural genetic variation. *Elife 11*, e76500.

Yang, S., X. Li, F. Yang, R. Zhao, X. Pan, J. Liang, L. Tian, X. Li, L. Liu, Y. Xing, and M. Wu (2019). Gut microbiota-dependent marker tmao in promoting cardiovascular disease: Inflammation mechanism, clinical prognostic, and potential as a therapeutic target. *Front Pharmacol 10*, 1360.

Yang, Y., T. Ikezoe, C. Nishioka, T. Taguchi, W.-g. Zhu, H. P. Koeffler, and H. Taguchi (2006). Zd6474 induces growth arrest and apoptosis of gist-t1 cells, which is enhanced by concomitant use of sunitinib. *Cancer science 97*(12), 1404–1409.

Yu, B. (2013). Stability. *Bernoulli 19*(4), 1484–1500.

Yu, B. and K. Kumbier (2020). Veridical data science. *Proceedings of the National Academy of Sciences 117*(8), 3920–3929.

Yu-Rice, Y., Y. Jin, B. Han, Y. Qu, J. Johnson, T. Watanabe, L. Cheng, N. Deng, H. Tanaka, B. Gao, et al. (2016). Foxc1 is involved in erα silencing by counteracting gata3 binding and is implicated in endocrine resistance. *Oncogene 35*(41), 5400–5411.

Zeng, L., S. Moser, N. Mirza-Schreiber, C. Lamina, S. Coassin, C. P. Nelson, T. Annilo, O. Franzén, M. E. Kleber, S. Mack, T. F. M. Andlauer, B. Jiang, B. Stiller, L. Li, C. Willenborg, M. Munz, T. Kessler, A. Kastrati, K.-L. Laugwitz, J. Erdmann, S. Moebus, M. M. Nöthen, A. Peters, K. Strauch, M. Müller-Nurasyid, C. Gieger, T. Meitinger, E. Steinhagen-Thiessen, W. März, A. Metspalu, J. L. M. Björkegren, N. J. Samani, F. Kronenberg, B. Müller-Myhsok, and H. Schunkert (2022, March). Cis-epistasis at the LPA locus and risk of cardiovascular diseases. *Cardiovasc. Res. 118*(4), 1088–1102.

Zhang, L. and L. Janson (2020). Floodgate: inference for model-free variable importance. *arXiv preprint arXiv:2007.01283*.

Zhang, X., Q. Liu, Q. Liao, and Y. Zhao (2020). Pancreatic cancer, gut microbiota, and therapeutic efficacy. *J Cancer 11*, 2749–2758.

Zhou, W., D. Zhang, Z. Li, H. Jiang, J. Li, R. Ren, X. Gao, J. Li, X. Wang, and W. Wang (2021). The fecal microbiota of patients with pancreatic ductal adenocarcinoma and autoimmune pancreatitis characterized by metagenomic sequencing. *Journal of Translational Medicine 19*, 1–12.

Zhou, Z. and G. Hooker (2021). Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data (TKDD) 15*(2), 1–21.

Zhu, R., D. Zeng, and M. R. Kosorok (2015). Reinforcement learning trees. *Journal of the American Statistical Association 110*(512), 1770–1784.

Appendix A

Low-signal iterative random forests (lo-siRF) for epistasis discovery

A.1 Supplementary Figures

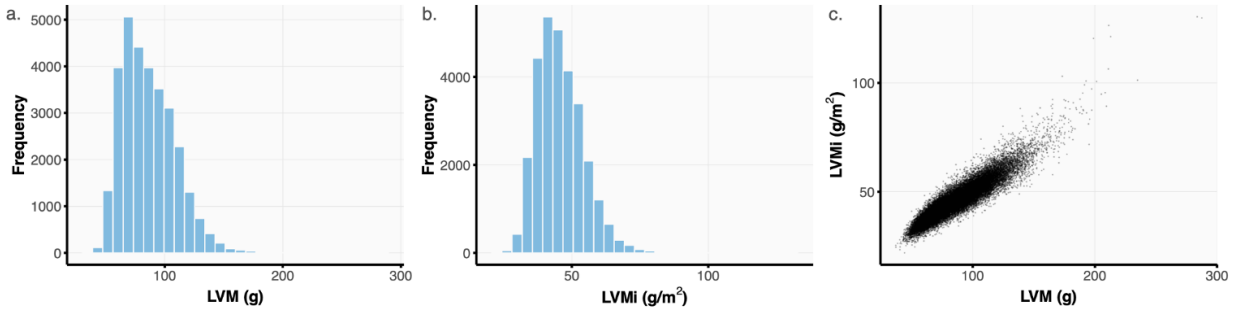


Figure A.1: Left ventricular mass (LVM, a) and LVM indexed to body surface area (LVMi, b) measurements were extracted from cardiac MRIs for 30,000 White British unrelated individuals via deep learning (Bai et al., 2018). (c) A high Pearson correlation of 0.92 was observed between these LVM and LVMi measurements.

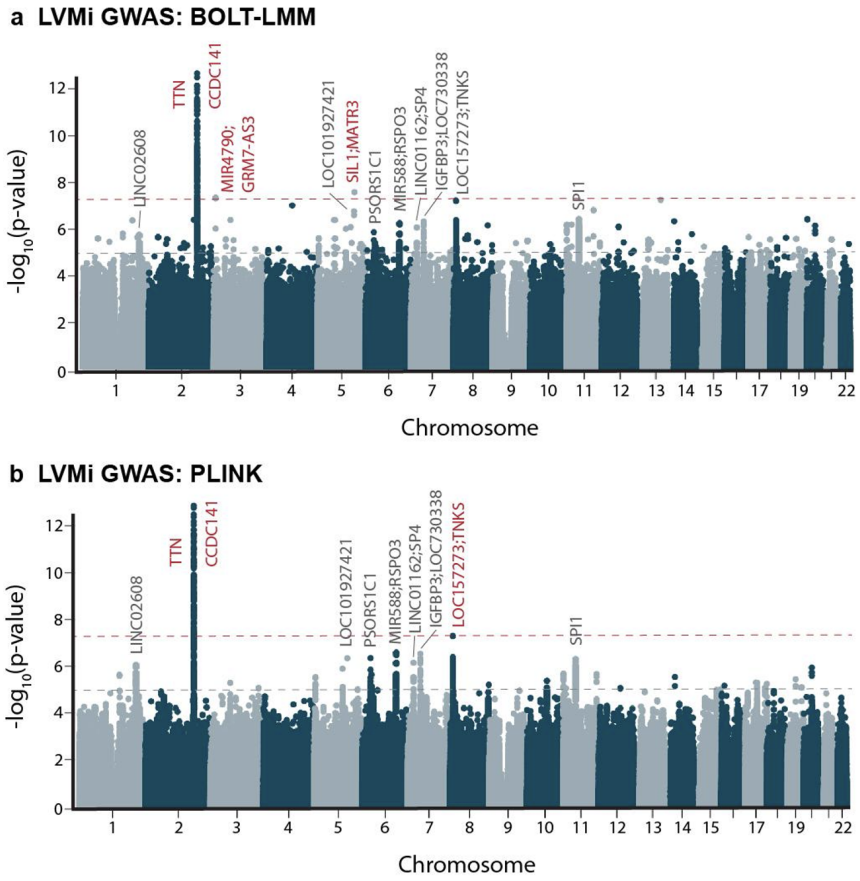


Figure A.2: GWAS using BOLT-LMM (a) and PLINK (b) identified associations with LVMi, of which *TTN* and *CCDC141* show the highest significance and stability. Loci meeting the genome-wide significance ($p < 5E\text{-}8$, red dashed line) are annotated in red. Loci, ranked in the top 10 by average p-value from BOLT-LMM and PLINK, are annotated in dark gray. Suggestive significance ($p < 1E\text{-}5$) is given by the gray dashed line.

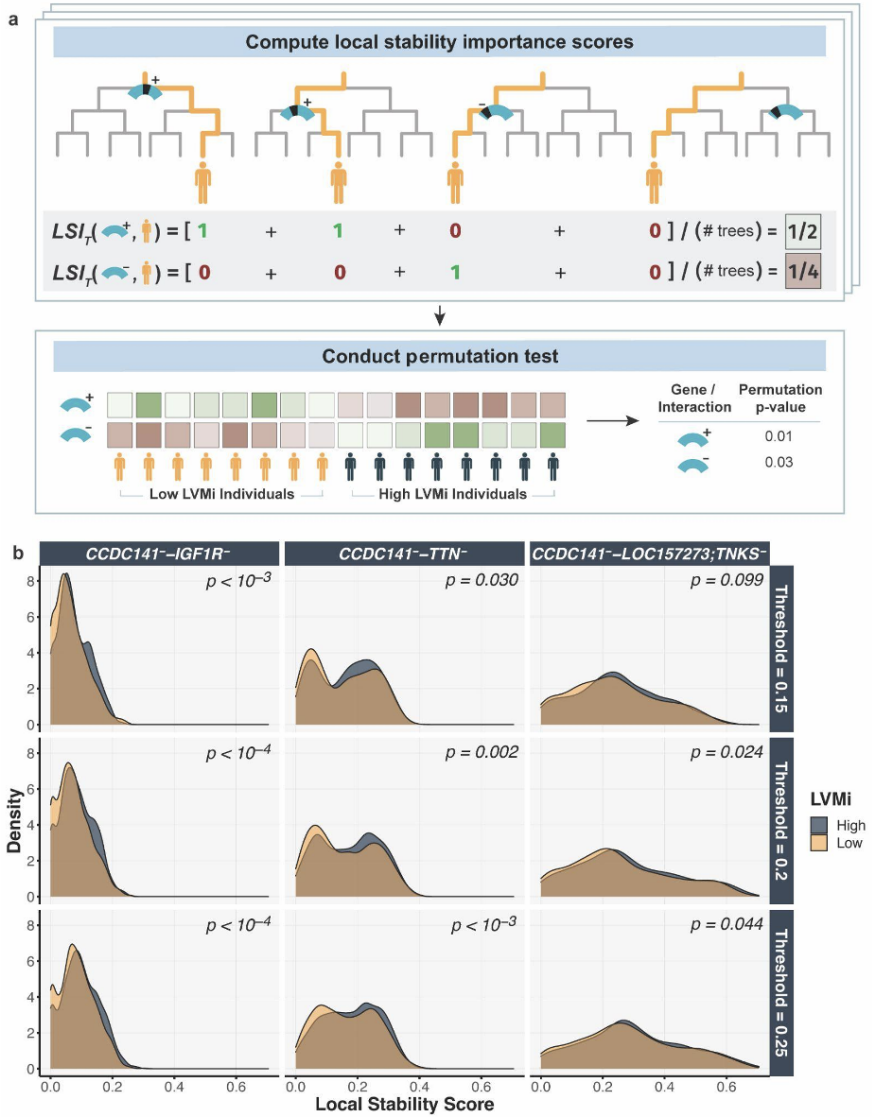


Figure A.3: (a) Schematic of local stability importance score computation. Given a gene (light blue transcript), the local stability importance score for an individual is defined as the proportion of trees for which at least one SNP (shaded black region) in the gene is used in the individual's decision path. This computation (top) was performed for each individual (denoted by the stacked boxes). Then, a permutation test was conducted to assess the difference in these local stability importance scores between the low and high LVMi individuals (bottom). (b) Differences in the distribution of local stability importance scores suggest that the identified gene-gene interactions are important for differentiating individuals with high (dark blue) and low (orange) LVMi in the iRF fit. This result, evaluated on the validation data, is stable across the three binarization thresholds and is quantified by a permutation p-value given in the top right corner of each subplot.

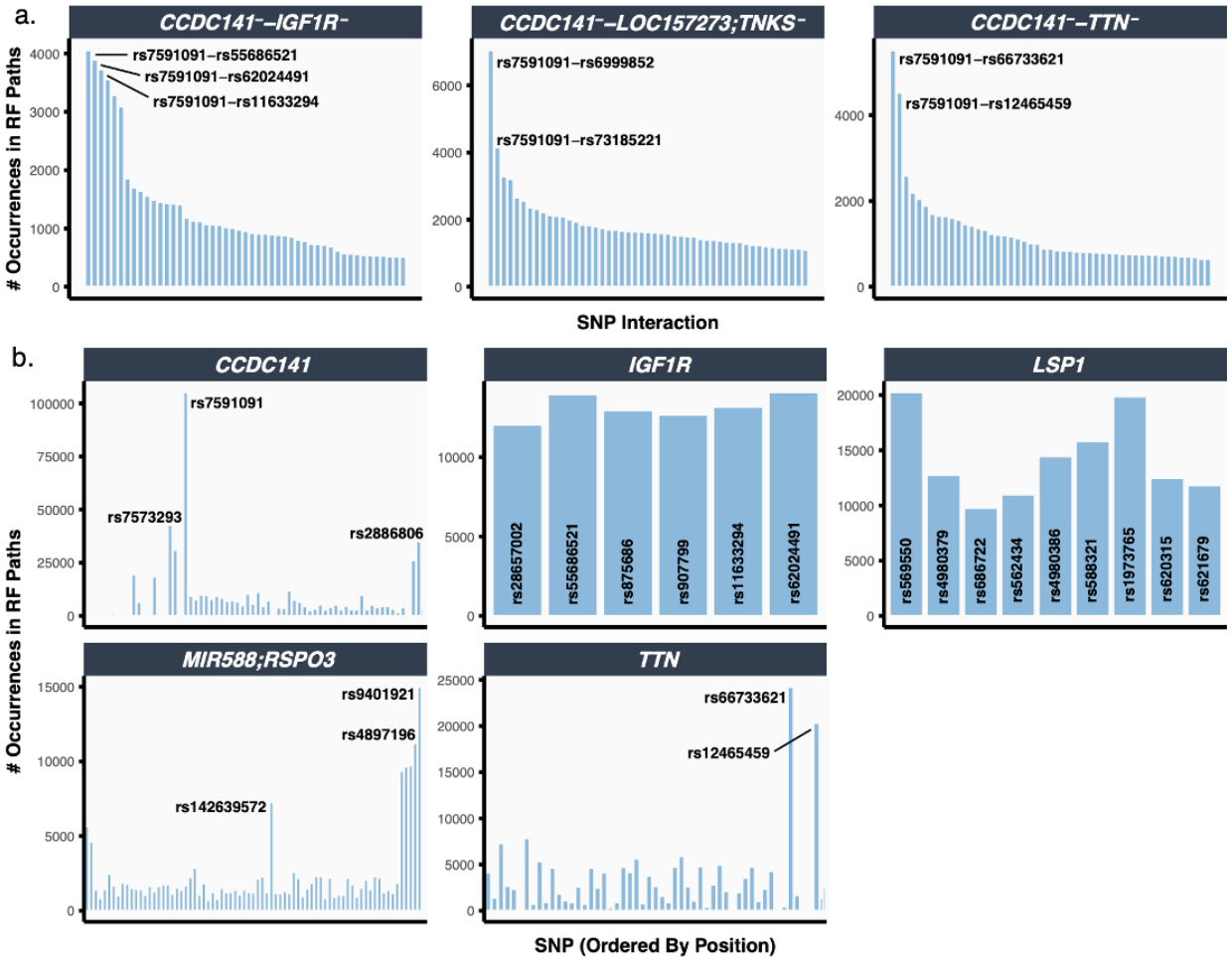


Figure A.4: The most important SNPs and SNP-SNP interactions, as measured by their frequency of occurrence in the siRF fit (binarization threshold = 0.2), are annotated for the top lo-siRF-recommended gene-gene interactions in (a) and top genes in (b). In each of the interactions, the SNP rs7591091 in *CCDC141* appears most frequently, suggesting a key role in cardiac hypertrophy.

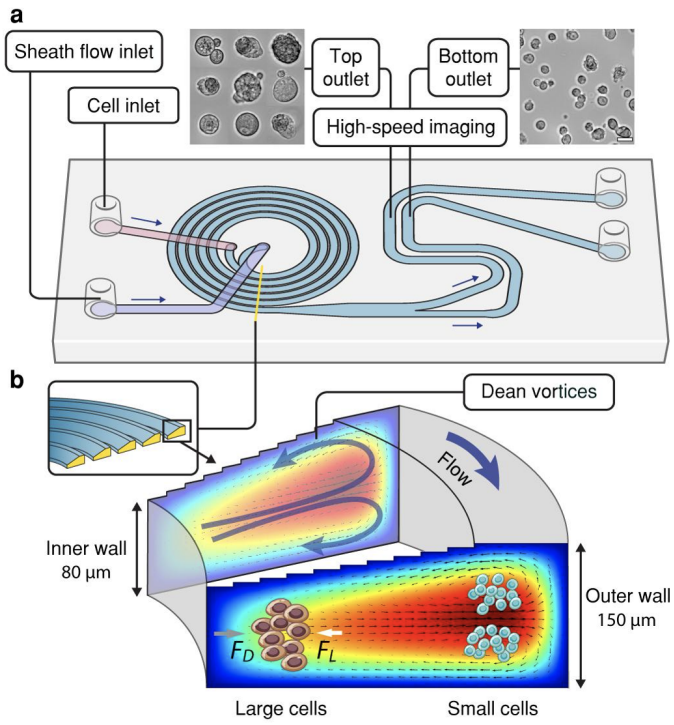


Figure A.5: (a) Schematic of an inertial microfluidic cell focusing device. Cell suspensions and fresh medium were introduced into the microfluidic device through the cell and sheath flow inlets, respectively, using a syringe pump and flowed down the 5-loop spiral microchannel with the same flow rate (1.2 mL/min). Inserted microscopy images show that randomly dispersed cells were separated by size and bifurcated into the top (large cells) or bottom (small cells) outlets. Scale bar, 10 µm. Outlet channels are connected to straight observation channels where flowing cells were further focused in the channel height direction and imaged using a high-speed camera for cell feature extraction (Figure 2.4c). (b) Schematic of the cell focusing principle. The spiral microchannel has a cross-section with a slanted ceiling, resulting in different depths at the inner and outer side of the microchannel. This geometry induces strong Dean vortices (counter rotating vortices in the plane perpendicular to the main flow direction) in the outer half of the microchannel cross-section. The interplay between drag forces (FD) induced by Dean vortices and lift forces (FL) due to shear gradient and the channel wall drives cell transverse migration towards equilibrium positions where the net force is zero. As a result, large cells in a heterogeneous population progressively migrate closer to the inner channel wall, while smaller cells move towards the outer channel wall. Details about microchannel dimensions can be found in *Methods*.

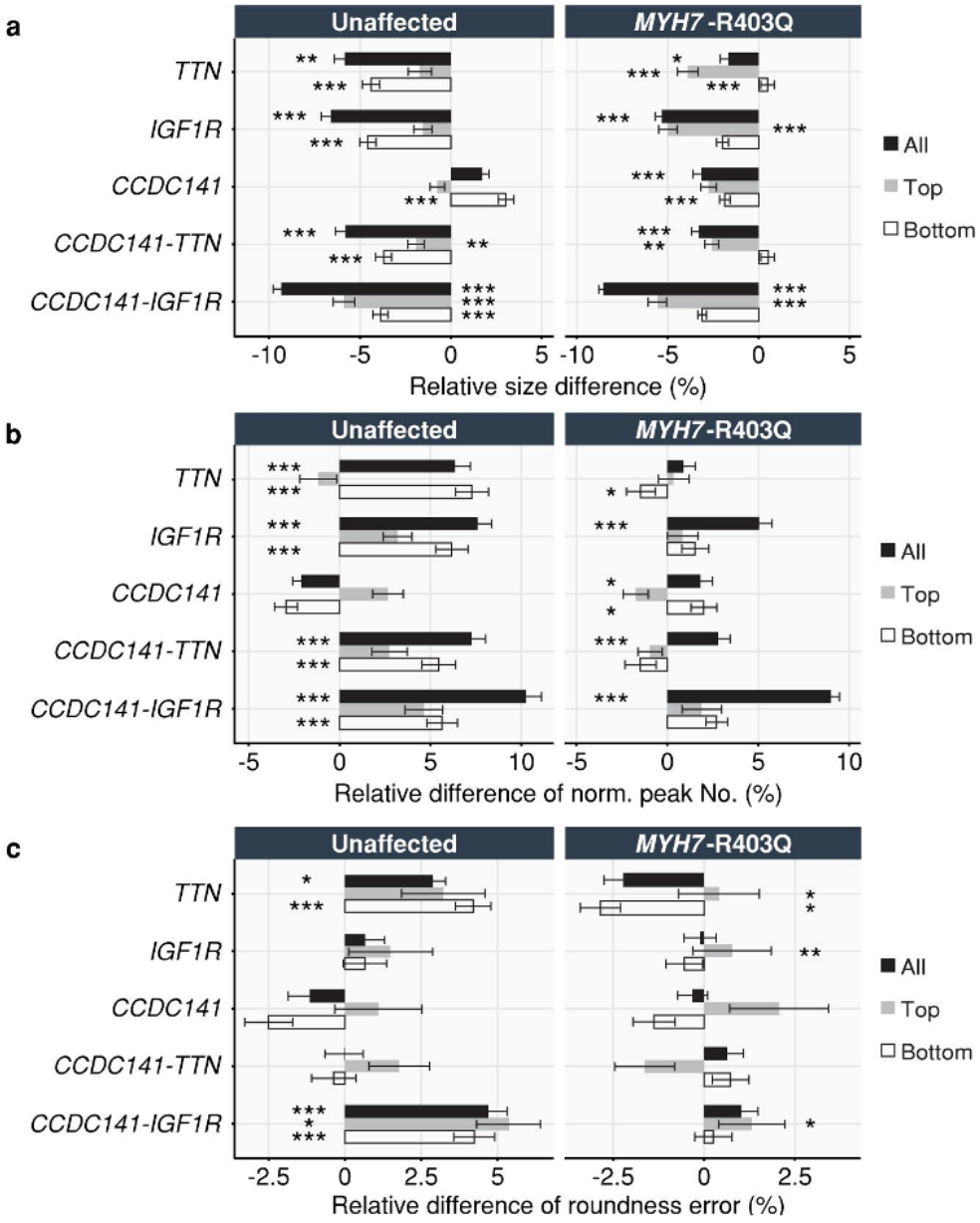


Figure A.6: Relative differences in median cell size (a), normalized peak number (b), and roundness error (c) of gene-silenced hiPSC-CMs compared to scramble controls computed for cells in the top (gray), bottom (white), or both (black) outlets. Error bars indicate standard deviation calculated from bootstrapping samples of 2 to 4 batches of cells. Asterisks indicate significant differences compared to the scramble control based on the maximum p-values of Wilcoxon signed rank test across all batches of cells ($^{*}p < 0.05$, $^{**}p < 0.001$, and $^{***}p <$ 1E-4).

A.2 Supplementary Tables

Table A.1: Summary statistics of the 30,000 White British unrelated individuals analyzed in this study. Means and standard deviations (in parentheses) are reported for continuous measurements (iLVM, LVM, age, height, and weight) alongside the percentage of individuals with various cardiac hypertrophy-related diseases or on blood pressure medication.

	Female	Male
N	15394	14606
LVMi (g/m2)	40.7 (5.7)	51.3 (7.8)
LVM (g)	70.7 (12.3)	102.6 (18.7)
Age (y)	63.5 (7.5)	64.9 (7.7)
Height (cm)	162.9 (6.2)	176.2 (6.6)
Weight (kg)	69.0 (13.0)	83.7 (13.3)
Hypertensive Diseases	20.2%	29.8%
Aortic Stenosis	0.1%	0.2%
Heart Failure	0.1%	0.6%
Type II Diabetes	1.6%	3.2%
Blood Pressure Medication	12.1%	19.9%

Table A.2: For each of the three binarization thresholds used in lo-siRF (corresponding to the bottom/top 15^{th}, 20^{th}, and 25^{th} quantiles), we provide the gender-specific LVMi cutoffs for the low and high LVMi groups. All thresholds were measured in g/m^2.

	Male		Female	
Binarization Threshold	Low LVMi Threshold	High LVMi Threshold	Low LVMi Threshold	High LVMi Threshold
0.15	43.8	58.5	35.1	46.1
0.20	45.1	56.8	36.0	44.9
0.25	46.0	55.4	36.8	43.8

Table A.3: Maximum prediction accuracies highlighted in bold. iRF performs better or on par with other commonly used machine learning methods when predicting the binarized LVMi phenotype. This result holds across all three binarization thresholds and three different classification metrics, i.e., accuracy, area under the receiver operator curve (AUROC), and area under the precision-recall curve (AUPRC). In accordance with the prediction check component of the PCS framework, iRF is an appropriate fit for the given data.

	Binarization Threshold = 0.15			**Binarization Threshold = 0.20**			**Binarization Threshold = 0.25**		
Method	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC	Accuracy	AUROC	AUPRC
siRF	0.554	**0.585**	**0.579**	**0.557**	**0.583**	**0.562**	**0.563**	**0.582**	**0.556**
Lasso	0.547	0.559	0.541	0.545	0.556	0.529	0.534	0.550	0.526
RF	0.546	0.572	0.571	0.553	0.569	0.549	0.554	0.569	0.555
Ridge	**0.559**	0.567	0.550	0.539	0.563	0.539	0.541	0.555	0.536
SVM	0.553	0.566	0.552	0.551	0.565	0.541	0.544	0.558	0.541

Table A.4: Though prediction accuracy is weak (indicated by precision scores close to 0.5), the lo-siRF-recommended interactions are stable across binarization thresholds and across bootstrap replicates (indicated by all types of stability scores being close or equal to 1). Here, prevalence measures the proportion of high LVMi individuals for which the interaction appears. Precision measures the probability of having high LVMi given that the interaction is active. The class difference in prevalence is the prevalence of the interaction in high LVMi individuals minus the prevalence in low LVMi individuals. Independence of feature selection evaluates whether the interaction is collectively or individually associated with the responses. The stability of each of these metrics evaluates how stable the respective scores are across 50 bootstrap replicates. The overall stability score (last column) is the proportion of times that the interaction is identified by iRF across 50 bootstrapped runs. Higher scores for each listed metric indicate greater importance.

Threshold	Prevalence	Precision	Class Difference in Prevalence	Stability of Class Difference in Prevalence	Independence of Feature Selection	Stability of Independence of Feature Selection	Increase in Precision	Stability of Increase in Precision	Stability
CCDC141⁻–IGF1R⁻									
Binarization Quantile = 0.15	0.064	0.56	0.012	1.0	0.00094	0.78	0.017	1.0	0.68
Binarization Quantile = 0.2	0.062	0.54	0.0095	1.0	0.0044	1.0	0.014	1.0	0.64
Binarization Quantile = 0.25	0.077	0.54	0.012	1.0	0.011	1.0	0.019	1.0	0.82
CCDC141⁻–LOC157273;TNKS⁻									
Binarization Quantile = 0.15	0.20	0.56	0.040	1.0	0.012	1.0	0.020	1.0	1.0
Binarization Quantile = 0.2	0.19	0.56	0.037	1.0	0.020	1.0	0.028	1.0	1.0
Binarization Quantile = 0.25	0.23	0.55	0.043	1.0	0.0068	1.0	0.022	1.0	1.0
CCDC141⁻–TTN⁻									
Binarization Quantile = 0.15	0.12	0.56	0.023	1.0	0.0066	1.0	0.016	1.0	1.0
Binarization Quantile = 0.2	0.14	0.55	0.024	1.0	0.0042	0.96	0.017	1.0	1.0
Binarization Quantile = 0.25	0.085	0.52	0.0076	1.0	-0.0036	0.020	0.0024	0.88	0.92

Table A.5: A list of the top signed genes and gene-gene interactions, recommended by losiRF, that were stably important across all three binarization thresholds. These genes and gene-gene interactions are ranked by the mean p-value, averaged across the three binarization thresholds.

	Binarization Threshold			
Gene / Interaction	**0.15**	**0.20**	**0.25**	**Mean p-value**
CCDC141⁻–IGF1R⁻	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$	**$< 10^{-3}$**
IGF1R⁻	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	**$< 10^{-3}$**
MIR588;RSPO3⁺	0.002	0.004	$< 10^{-4}$	**0.002**
TTN⁻	0.022	0.006	$< 10^{-3}$	**0.009**
CCDC141⁻–TTN⁻	0.030	0.002	$< 10^{-3}$	**0.011**
TTN⁺	0.030	0.005	$< 10^{-3}$	**0.012**
MIR588;RSPO3⁻	0.016	0.014	0.009	**0.013**
LSP1⁻	0.029	0.019	0.002	**0.017**
CCDC141⁻	0.033	0.007	0.015	**0.018**
CCDC141⁻–LOC157273;TNKS⁻	0.099	0.024	0.044	**0.056**

A.3 Supplementary Notes

Definition of medical conditions in Table A.1. In Table A.1, we define hypertensive diseases as self-reported hypertension, high blood pressure as diagnosed by a doctor, or any ICD10 billing code diagnosis in I10-I16; aortic stenosis as self-reported aortic stenosis or an ICD10 billing code diagnosis of I35; heart failure as self-reported heart failure or an ICD10 billing code diagnosis of I50; and type II diabetes as self-reported type II diabetes or an ICD10 billing code diagnosis of E11.

Appendix B

MDI+: A flexible random forest-based feature importance framework

B.1 Proofs

Proof of Theorem 1

Using the law of total variance, we may rewrite the formula for impurity decrease for a node $\mathtt{t}$ via

$$\begin{aligned}\hat{\Delta}(\mathtt{t}) &= N\left(\mathtt{t}\right)^{-1}\left(\sum_{\mathbf{x}_i\in\mathtt{t}}\left(y_i-\bar{y}_{\mathtt{t}}\right)^2-\sum_{\mathbf{x}_i\in\mathtt{t}_L}\left(y_i-\bar{y}_{\mathtt{t}_L}\right)^2-\sum_{\mathbf{x}_i\in\mathtt{t}_R}\left(y_i-\bar{y}_{\mathtt{t}_R}\right)^2\right)\\ &= \frac{N\left(\mathtt{t}_L\right)N\left(\mathtt{t}_R\right)}{N\left(\mathtt{t}\right)^2}\left(\bar{y}_{\mathtt{t}_L}-\bar{y}_{\mathtt{t}_R}\right)^2.\end{aligned}$$

Note that we use a binary indicator of a sample lying in the right child node as the conditioning variable. Next, using equation (3.3) and using $\psi(\mathtt{t})$ to denote the resulting feature vector on the training set, we compute

$$\begin{aligned}\psi(\mathtt{t})^T\mathbf{y} &= \frac{N(\mathtt{t}_R)\sum_{\mathbf{x}_i\in\mathtt{t}_L}y_i-N(\mathtt{t}_L)\sum_{\mathbf{x}_i\in\mathtt{t}_R}y_i}{\sqrt{N(\mathtt{t}_L)N(\mathtt{t}_R)}}\\ &= \sqrt{N(\mathtt{t}_L)N(\mathtt{t}_R)}(\bar{y}_{\mathtt{t}_L}-\bar{y}_{\mathtt{t}_R}).\end{aligned}$$

By combining the above two equations, we obtain the formula

$$\hat{\Delta}(\mathtt{t}) = \frac{\left(\psi(\mathtt{t})^T\mathbf{y}\right)^2}{N(\mathtt{t})^2}. \tag{B.1}$$

Now, let $\mathtt{t}_1, \ldots, \mathtt{t}_m$ denote the nodes splitting on feature X_k, which means that we have $\Psi^k = [\psi(\mathtt{t}_1), \cdots, \psi(\mathtt{t}_m)]$. We may then compute

$$\begin{aligned}\left\|\Psi^k(\mathbf{X})\big(\Psi^k(\mathbf{X})^T\Psi^k(\mathbf{X})\big)^{-1}\Psi^k(\mathbf{X})^T\mathbf{y}\right\|_2^2 &= \mathbf{y}^T\Psi^k(\mathbf{X})\big(\Psi^k(\mathbf{X})^T\Psi^k(\mathbf{X})\big)^{-1}\Psi^k(\mathbf{X})^T\mathbf{y} \\ &= \sum_{i=1}^{m}\frac{\big(\psi(\mathtt{t}_i)^T\mathbf{y}\big)^2}{N(\mathtt{t}_i)} \\ &= \sum_{i=1}^{m} N(\mathtt{t}_i)\hat{\Delta}(\mathtt{t}_i),\end{aligned}$$

where the second equality comes from recognizing that $\big(\Psi^k(\mathbf{X})^T\Psi^k(\mathbf{X})\big)^{-1}$ is a diagonal matrix with i-th diagonal entry equal to $N(\mathtt{t}_i)^{-1}$, and the third equality comes from plugging in (B.1). Recognizing that the right-hand side is simply $n\text{MDI}(\hat{f}, k)$ completes the proof of the left equality in (3.4). The right equality follows from the definition of training R^2 as measuring the fraction of variance in the responses explained by the fitted values.

Proof of Proposition 1

Denote $\mathbf{P} \coloneqq \Psi^k(\mathbf{X})\big(\Psi^k(\mathbf{X})^T\Psi^k(\mathbf{X})\big)^{-1}\Psi^k(\mathbf{X})^T$. We have shown in the proof of Theorem 1 that

$$\text{MDI}_k(\mathcal{S}, \mathcal{D}_n) = \frac{1}{n}\|\mathbf{P}\mathbf{y}\|_2^2. \tag{B.2}$$

Now write $\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}$ where $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Under a fixed design, $\mathbf{f}$ is deterministic while $\boldsymbol{\epsilon}$ is random. Taking expectations therefore gives

$$\begin{aligned}\mathbb{E}\{\text{MDI}_k(\mathcal{S}, \mathcal{D}_n)\} &= \mathbb{E}\left\{\frac{1}{n}\|\mathbf{P}(\mathbf{f}+\boldsymbol{\epsilon})\|_2^2\right\} \\ &= \frac{1}{n}\|\mathbf{P}\mathbf{f}\|_2^2 + \frac{1}{n}\mathbb{E}\big\{\|\mathbf{P}\boldsymbol{\epsilon}\|_2^2\big\} \\ &= \frac{1}{n}\|\mathbf{P}\mathbf{f}\|_2^2 + \frac{\sigma^2\text{Trace}(\mathbf{P})}{n}.\end{aligned} \tag{B.3}$$

By orthogonality of local decision stumps, $\Psi^k(\mathbf{X})$ has full rank, and $\text{Trace}(\mathbf{P})$ is just the number of columns, which is equal to $\big|\mathcal{S}^{(k)}\big|$. Finally, notice that

$$\text{MDI}_k(\mathcal{S}, \mathcal{D}_n^0) = \frac{1}{n}\|\mathbf{P}\mathbf{f}\|_2^2. \tag{B.4}$$

B.2 Feature Ranking Performance Simulations

In this section, we provide additional simulation experiments to supplement those already provided in Section 3.5. Unless specified otherwise, we follow the simulation protocol de-

scribed in Section 3.5, and all plots show the mean evaluation metric, averaged across 50 experimental replicates, with error bars denoting ± 1SE.

Data and Code Availability

The Juvenile dataset can be downloaded using the `imodels` python package (Singh et al., 2021). The Enhancer and Splicing datasets were taken from (Basu et al., 2018). The CCLE dataset can be downloaded from DepMap Public 18Q3 (`https://depmap.org/portal/download/`).

Data Preprocessing. We performed basic data cleaning on these datasets before using them as covariate matrices in the simulation study. Specifically, for all four datasets, we removed constant and duplicated columns. We also applied a $\log(x + 1)$ transformation to all values in the Enhancer and CCLE datasets as these are count-valued and highly right-skewed. For the Enhancer dataset, the raw data contained repeated measurements (i.e., multiple feature columns) for the same transcription factor; we thus removed all but one measurement (i.e., column) for each transcription factor. Note finally that the CCLE dataset contains 50114 features, but we chose a random subset of 1000 features to use in the covariate matrix in each simulation replicate.

Code to run all simulations can be found on GitHub at `https://github.com/Yu-Group/imodels-experiments`.

Regression Simulations

Following the simulation set-up discussed in Sections 3.5 and 3.5 , we provide the results for all datasets and regression functions in the regression setting. We show the results for all datasets for the linear function in Figure B.1, lss function in Figure B.2, polynomial interaction in Figure B.3, and the linear + LSS function in Figure B.4.

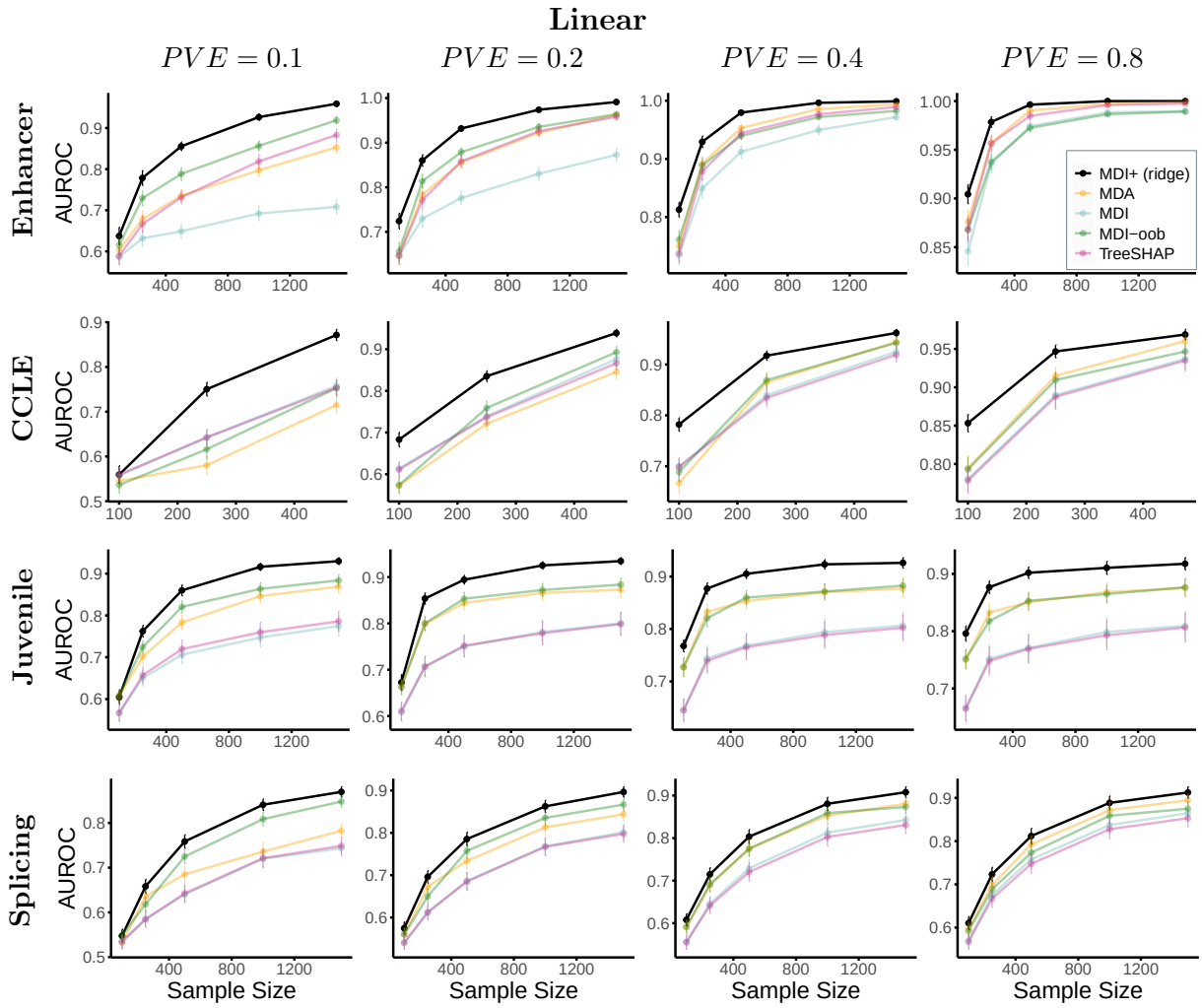


Figure B.1: MDI+ (ridge) outperforms other feature importance methods for the linear regression function described in Section 3.5. This pattern is evident across various datasets (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

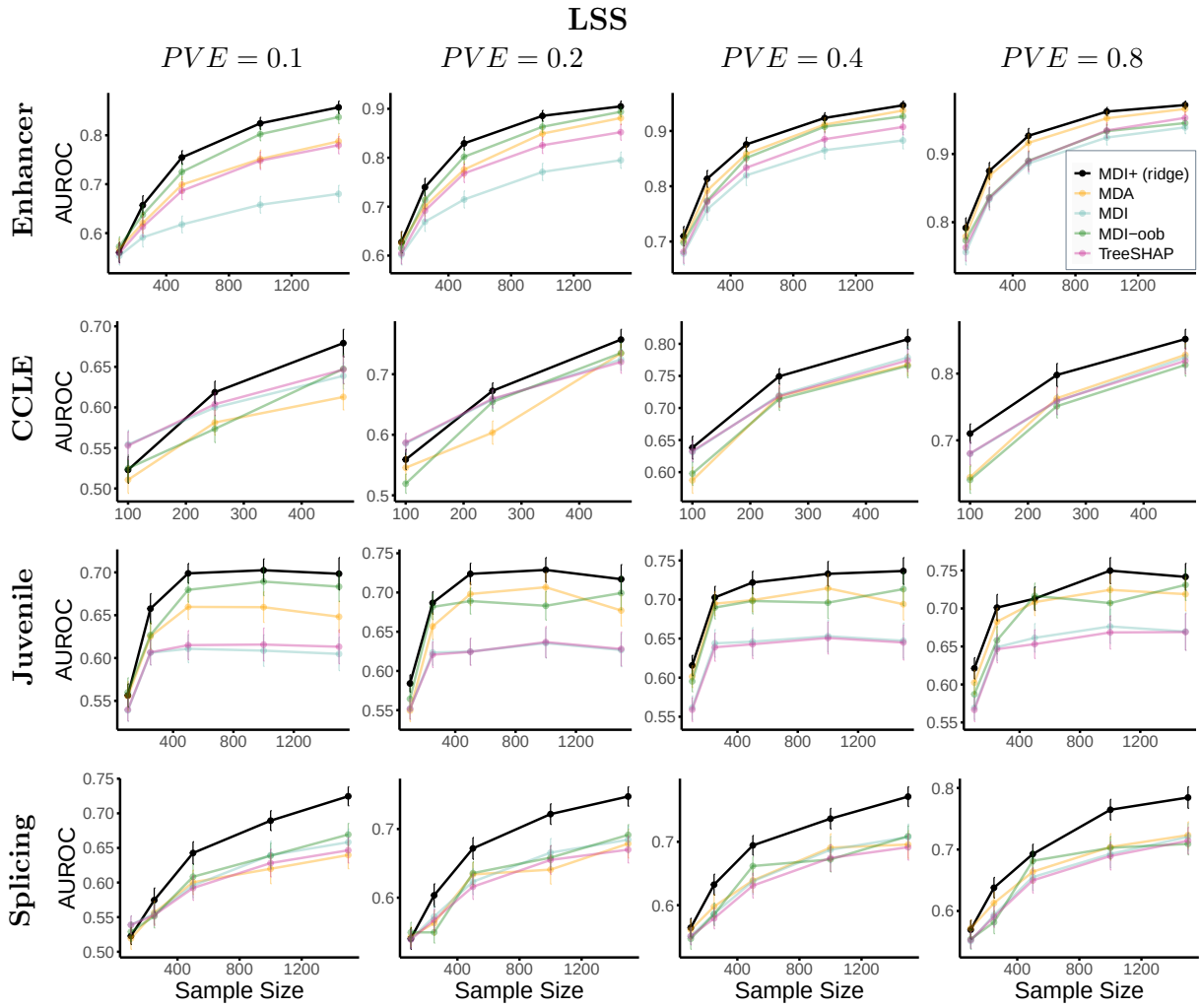


Figure B.2: MDI+ (ridge) outperforms other feature importance methods for the LSS function described in Section 3.5. This pattern is evident across various datasets (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

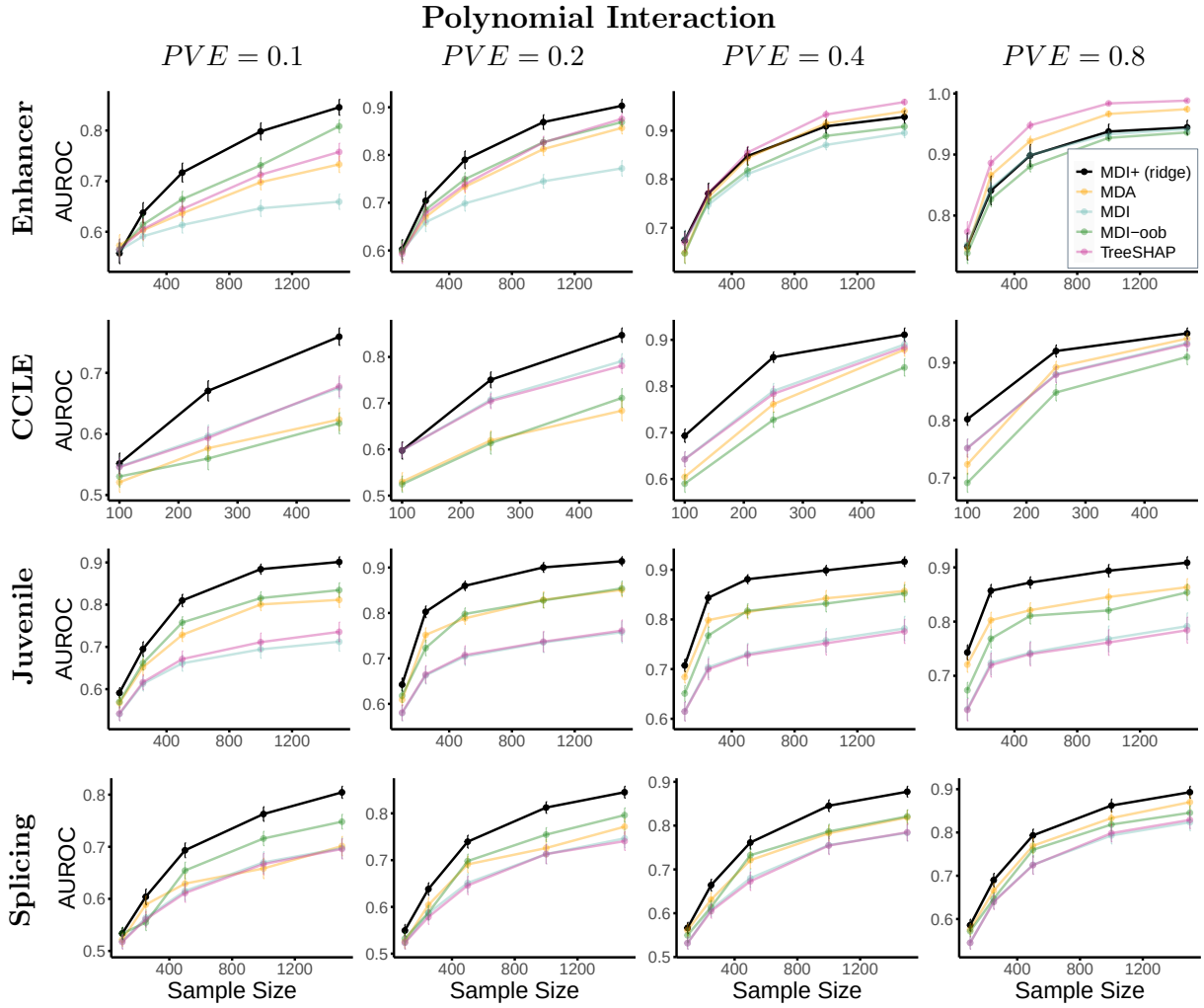


Figure B.3: MDI+ (ridge) outperforms other feature importance methods for the polynomial interaction regression function described in Section 3.5. This pattern is evident across various datasets (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

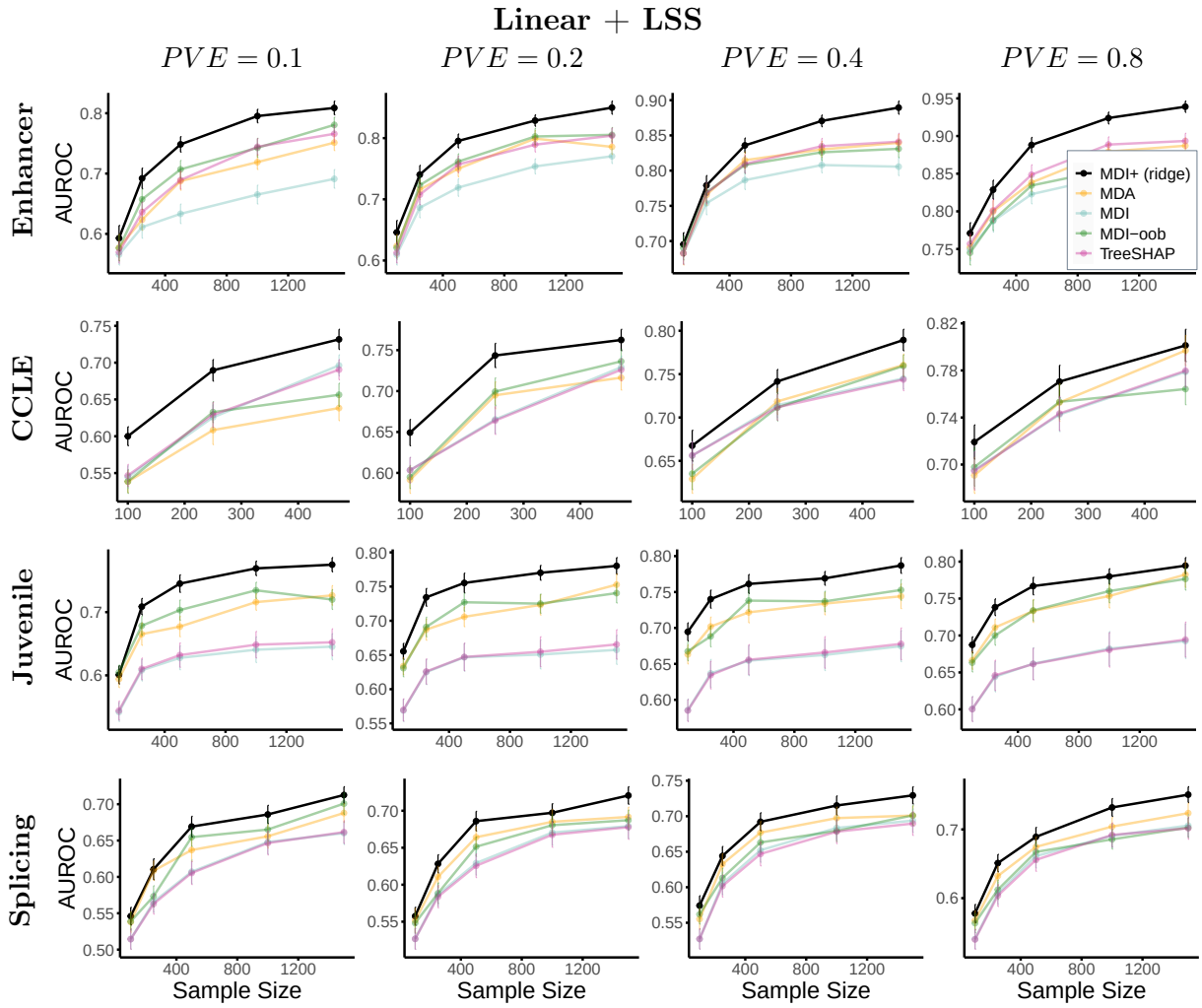


Figure B.4: MDI+ (ridge) outperforms other feature importance methods for the linear + LSS regression function described in Section 3.5. This pattern is evident across various datasets (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

Classification Simulations

Following the simulation set-up discussed in Sections 3.5 and 3.5 , we provide the results for all datasets and regression functions in the classification setting. We show the results for all datasets for the linear function in Figure B.1, LSS function in Figure B.2, polynomial interaction in Figure B.3, and the linear + LSS function in Figure B.4. All regression functions in the classification setting were passed through the logistic link function to ensures

that the responses are binary.

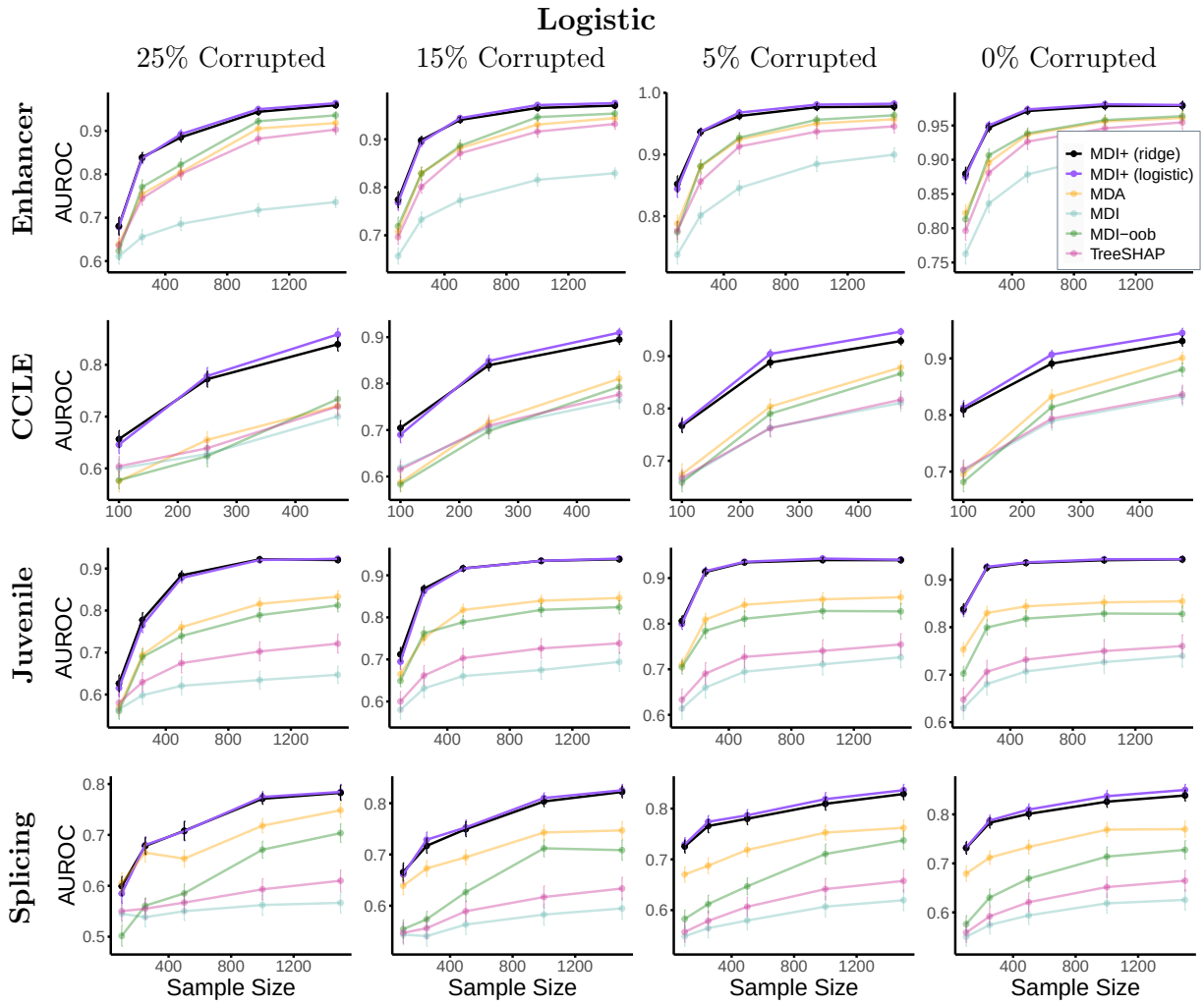


Figure B.5: Both MDI+ (ridge) and MDI+ (logistic) outperform all other feature importance methods for the logistic linear regression function described in Section 3.5. This pattern is evident across various regression functions (specified by panel), datasets with different covariate structures (specified by row), proportions of corrupted labels (specified by column), and sample sizes (on the x-axis). Furthermore, MDI+ (logistic) outperforms MDI+ (ridge) for some datasets, indicating the benefit of tailoring the choices of MDI+ to the data at hand. In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

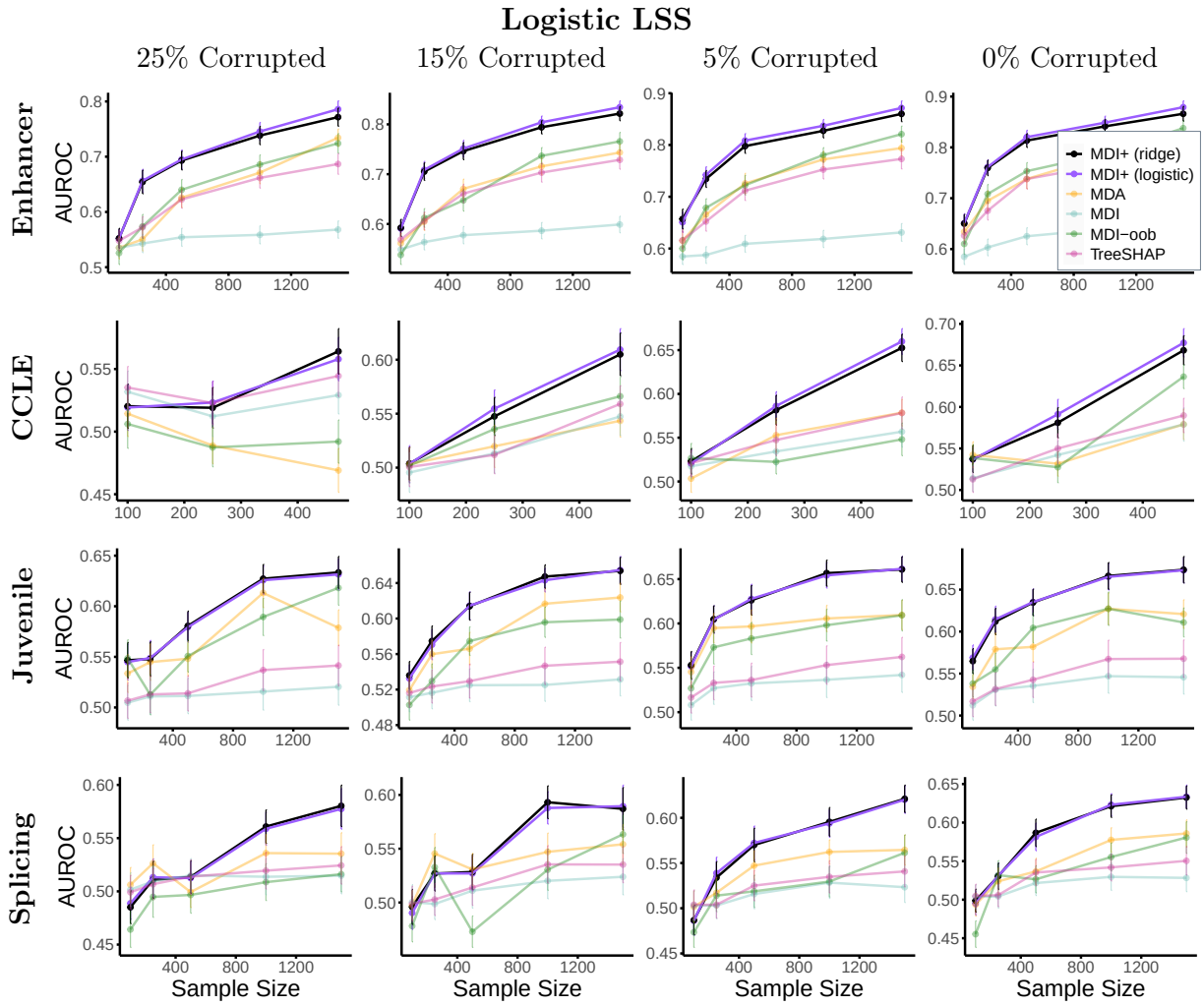


Figure B.6: Both MDI+ (ridge) and MDI+ (logistic) outperform all other feature importance methods for the logistic LSS regression function described in Section 3.5. This pattern is evident across various regression functions (specified by panel), datasets with different covariate structures (specified by row), proportions of corrupted labels (specified by column), and sample sizes (on the x-axis). Furthermore, MDI+ (logistic) outperforms MDI+ (ridge) for some datasets, indicating the benefit of tailoring the choices of MDI+ to the data at hand. In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

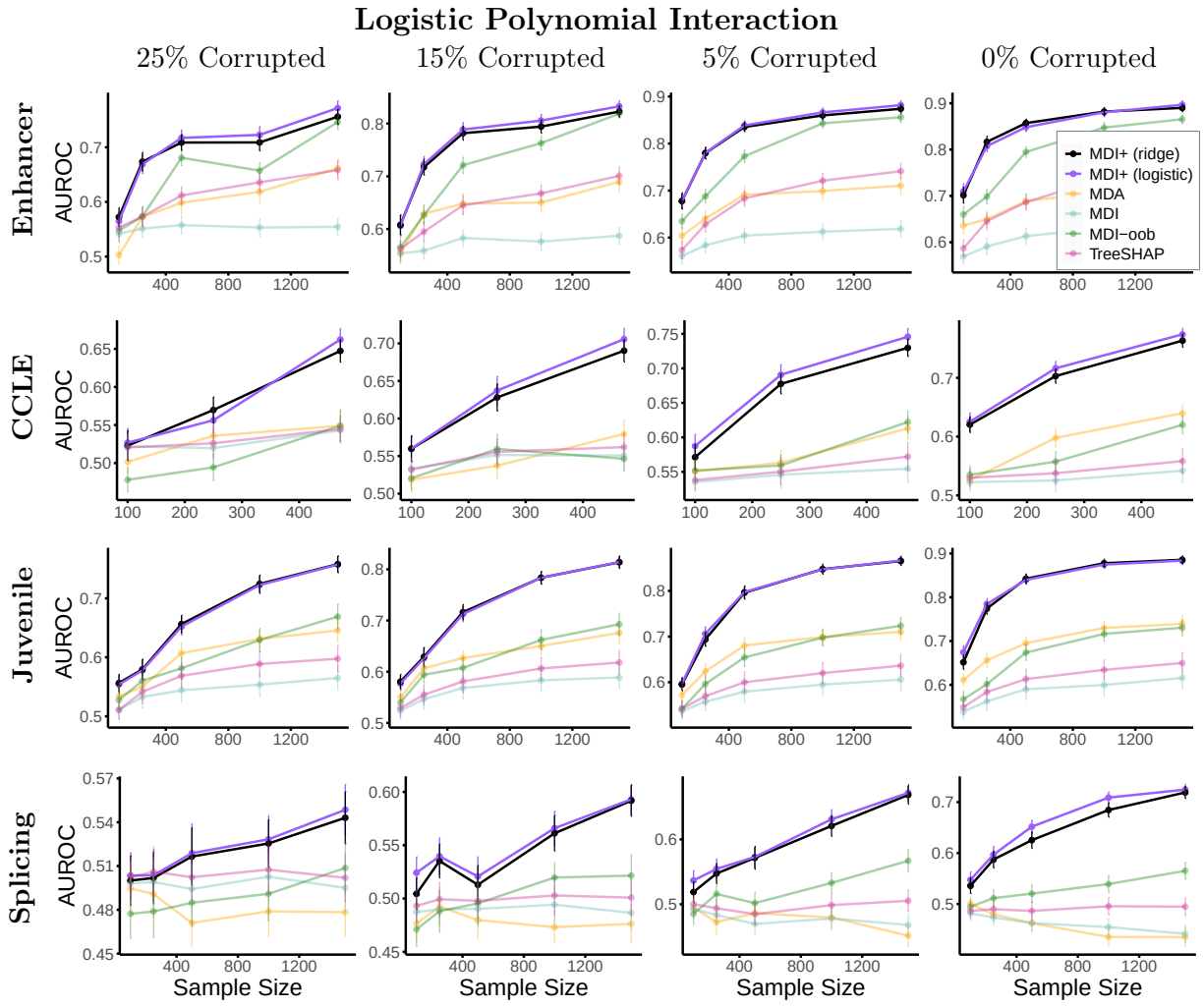


Figure B.7: Both MDI+ (ridge) and MDI+ (logistic) outperform all other feature importance methods for the logistic polynomial interaction regression function described in Section 3.5. This pattern is evident across various regression functions (specified by panel), datasets with different covariate structures (specified by row), proportions of corrupted labels (specified by column), and sample sizes (on the x-axis). Furthermore, MDI+ (logistic) outperforms MDI+ (ridge) for some datasets, indicating the benefit of tailoring the choices of MDI+ to the data at hand. In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

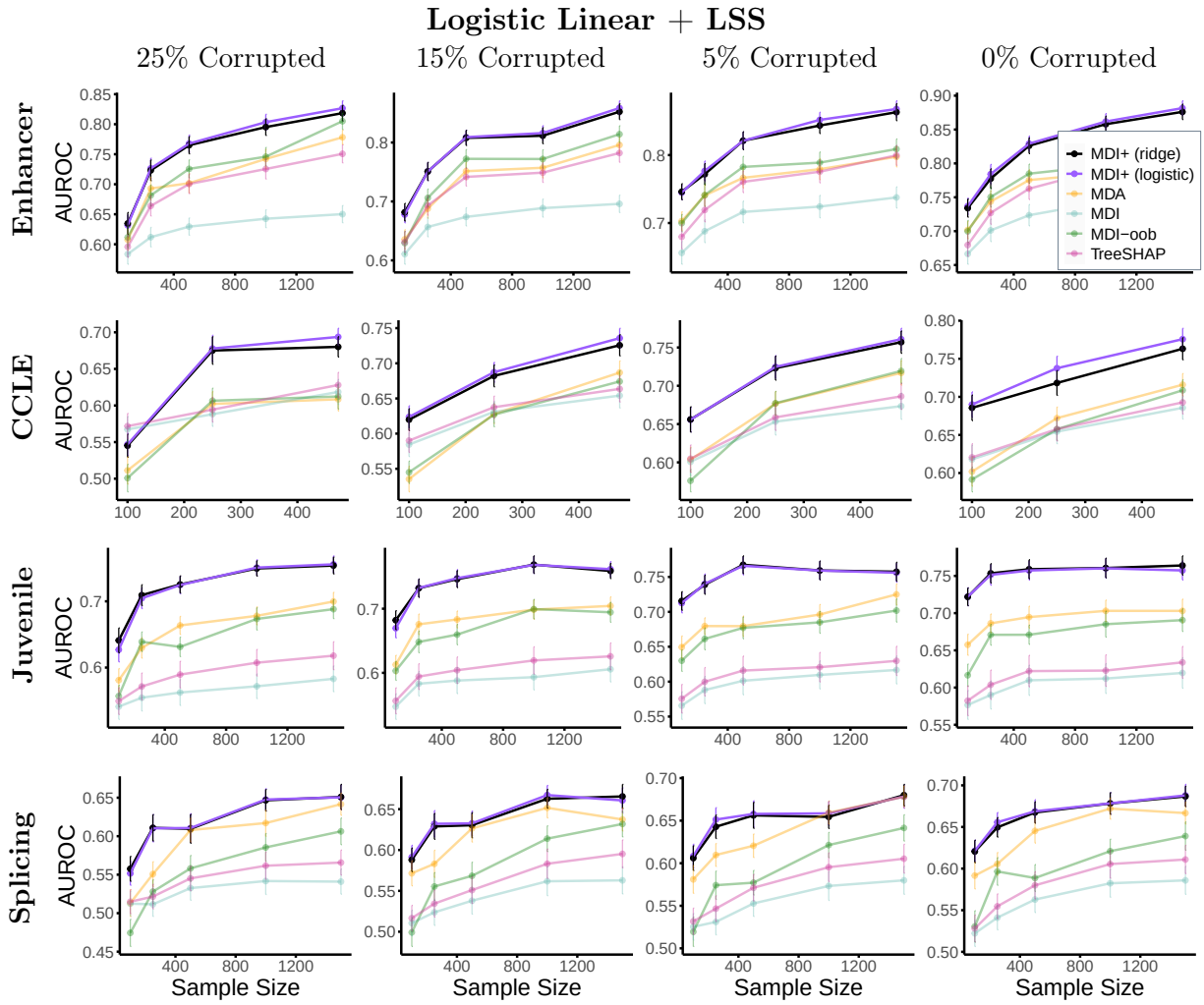


Figure B.8: Both MDI+ (ridge) and MDI+ (logistic) outperform all other feature importance methods for the logistic polynomial interaction regression function described in Section 3.5. This pattern is evident across various regression functions (specified by panel), datasets with different covariate structures (specified by row), proportions of corrupted labels (specified by column), and sample sizes (on the x-axis). Furthermore, MDI+ (logistic) outperforms MDI+ (ridge) for some datasets, indicating the benefit of tailoring the choices of MDI+ to the data at hand. In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

Robust Regression Simulations

Following the simulation set-up discussed in Sections 3.5 and 3.5 , we provide the results for the Enhancer and CCLE datasets and regression functions in the robust regression setting.

We show the results for these two datasets for the linear function in Figure B.9, lss function in Figure B.10, polynomial interaction in Figure B.11, and the linear + LSS function in Figure B.12.

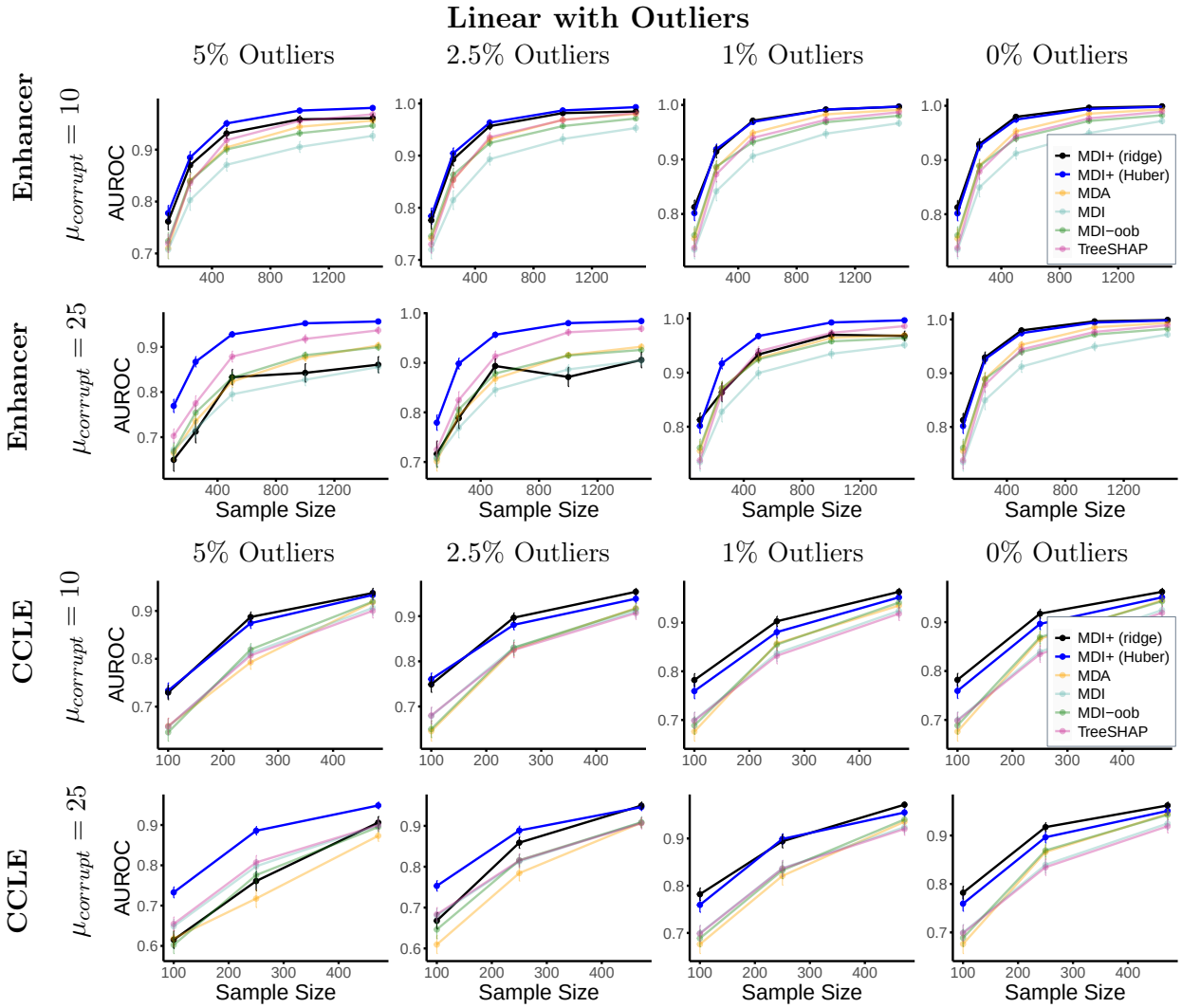


Figure B.9: Under the Linear model with outliers regression setting (See Sections 3.5 and 3.5 for simulation details) using the Enhancer and CCLE dataset, MDI+ (Huber)'s performance remains suffers far less than other methods including MDI+ (Ridge) as the level of corruption $\mu_{corrupt}$ (specified by row) and the proportion of outliers (specified by column) grow. This pattern also holds across sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

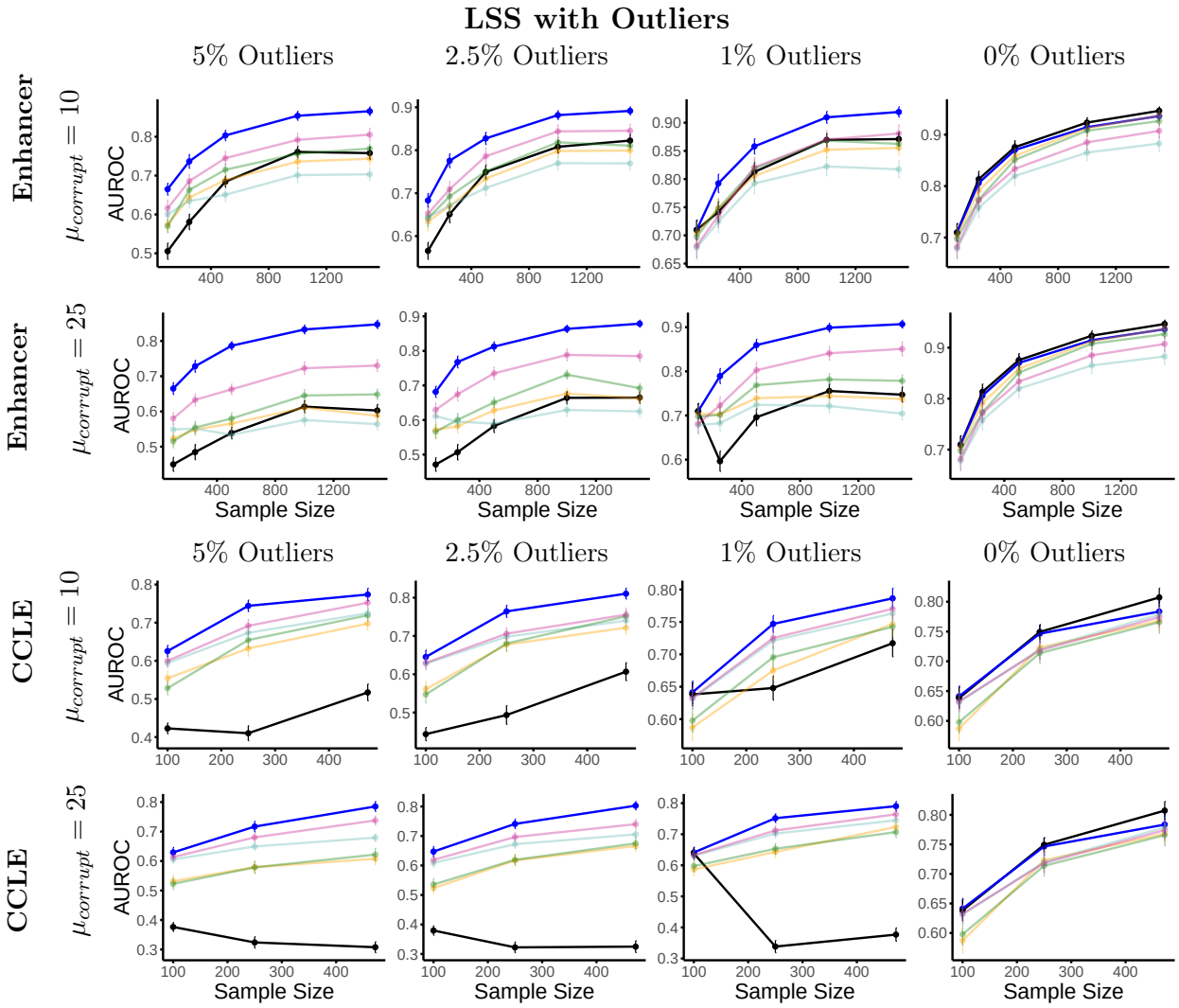


Figure B.10: Under the LSS model with outliers regression setting (See Sections 3.5 and 3.5 for simulation details) using the Enhancer and CCLE dataset, MDI+ (Huber)'s performance remains suffers far less than other methods including MDI+ (Ridge) as the level of corruption $\mu_{corrupt}$ (specified by row) and the proportion of outliers (specified by column) grow. This pattern also holds across sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

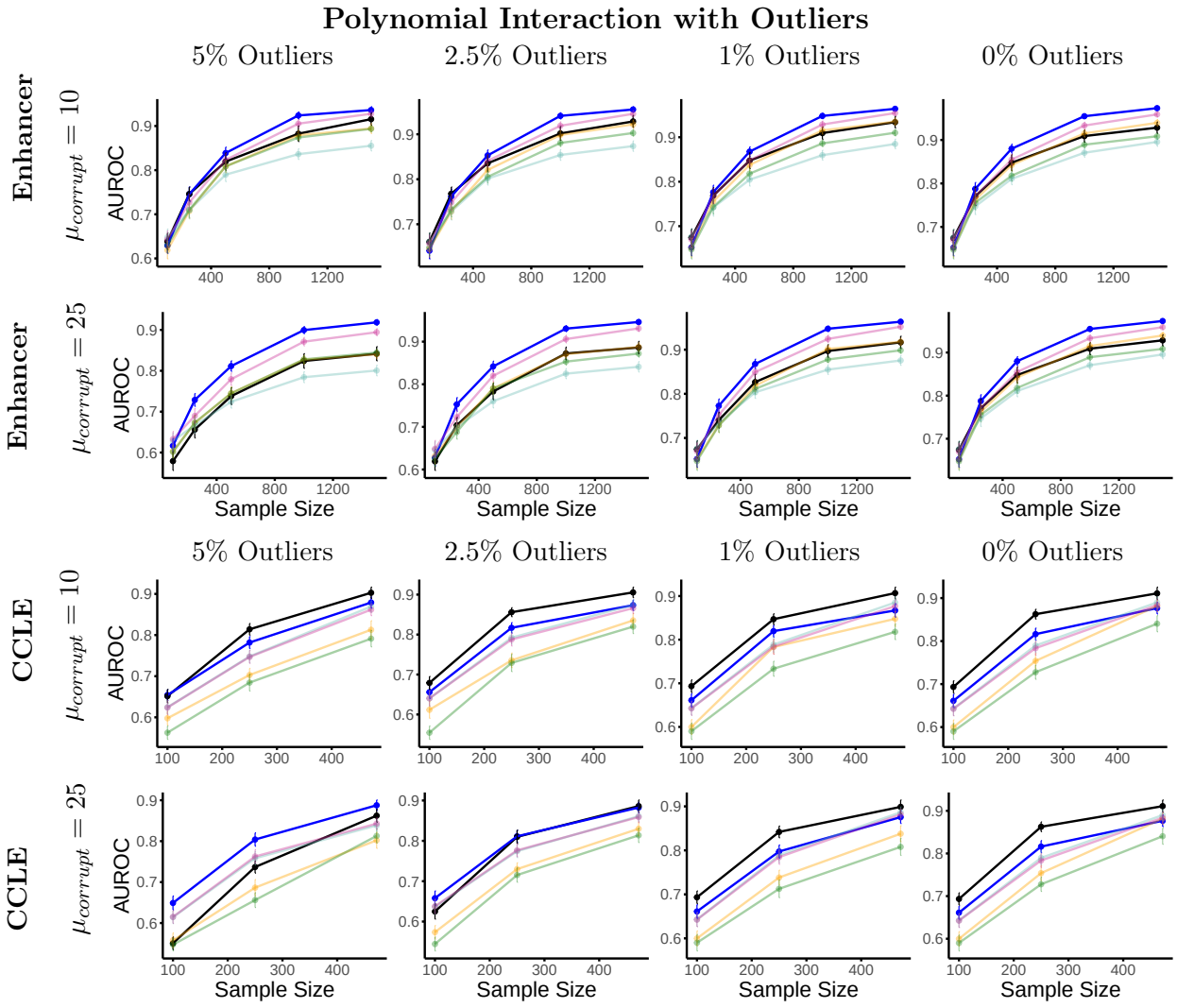


Figure B.11: Under the polynomial interaction model with outliers regression setting (See Sections 3.5 and 3.5 for simulation details) using the Enhancer and CCLE dataset, MDI+ (Huber)'s performance remains suffers far less than other methods including MDI+ (Ridge) as the level of corruption $\mu_{corrupt}$ (specified by row) and the proportion of outliers (specified by column) grow. This pattern also holds across sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

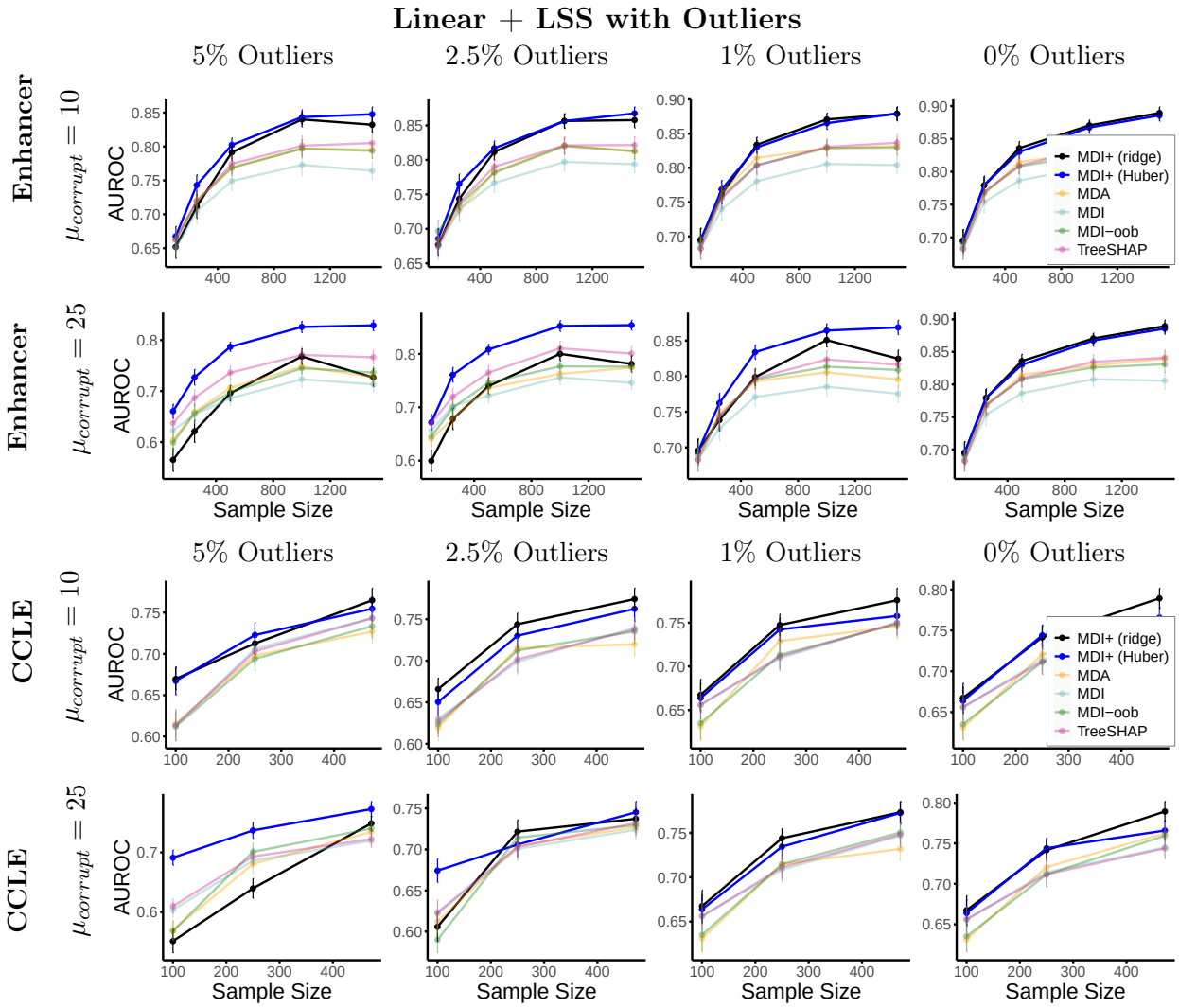


Figure B.12: Under the linear + LSS model with outliers regression setting (See Sections 3.5 and 3.5 for simulation details) using the Enhancer and CCLE dataset, MDI+ (Huber)'s performance remains suffers far less than other methods including MDI+ (Ridge) as the level of corruption $\mu_{corrupt}$ (specified by row) and the proportion of outliers (specified by column) grow. This pattern also holds across sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE.

B.3 Additional Data-Inspired Feature Ranking Simulations

In this section, we describe other data-inspired simulations we perform to establish the efficacy of MDI+ in a number of settings. Specifically, we perform simulations under a misspecificied setting (e.g., presence of unobserved features), as we vary the sparsity of the generating function, and as we vary the number of features.

Misspecified Regression Simulations

In practice, we are often unable to observe all covariates that are relevant for the response (Yu and Kumbier, 2020). To investigate the performance of MDI+ under this type of misspecified model scenario, we consider the following simulation setup.

Experimental details. We simulate data according to the four regression functions described in Section 3.5 (i.e., linear, LSS, polynomial interaction, and linear+LSS) but omit the first two signal features (i.e., X_1, X_2) from the covariate matrix $\mathbf{X}$ before fitting the RF and feature importance method under study. For example, in the linear regression simulation, we simulate the response y via $Y = \sum_{j=1}^{5} X_j + \varepsilon$, where $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$; however, we fit the RF and compute the feature importance measure using only y and $X_3, \ldots, X_p$. The rest of the experimental details are identical to those described previously.

Results. For each of the four regression functions, the AUROC results under the misspecified model regime are summarized in Figures B.13-B.16. In terms of the AUROC, MDI+ improves the ranking performance compared to the existing methods under this misspecified model scenario across a variety of regression functions, datasets with different covariate structures, proportions of variance explained, and sample sizes. Note here that the AUROC is computed with respect to only the observed covariates and ignore the omitted variables X_1 and X_2.

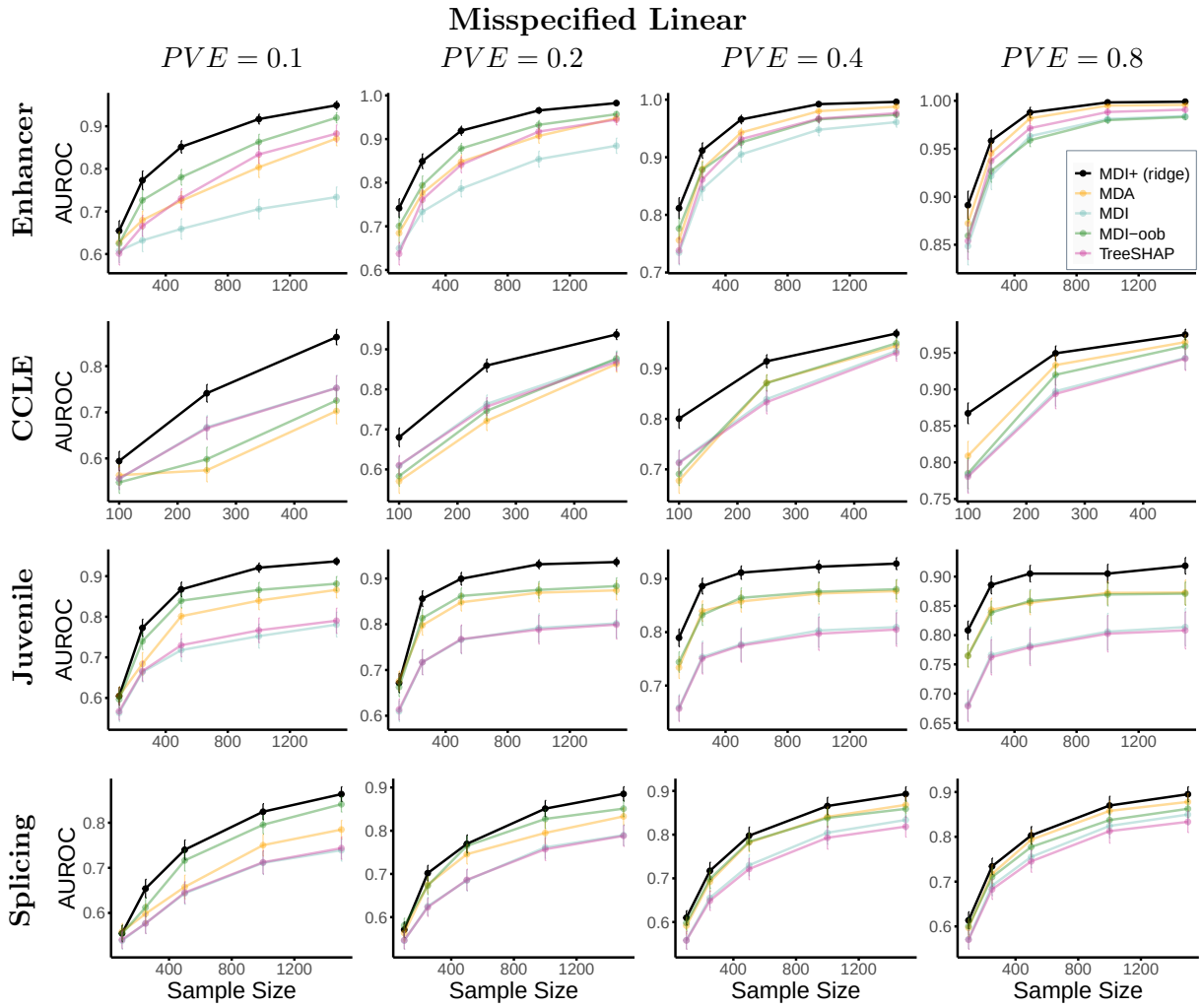


Figure B.13: MDI+ (ridge) outperforms other feature importance methods across almost all misspecified linear regression simulations described in Appendix B.3. This pattern is evident across various datasets with different covariate structures (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

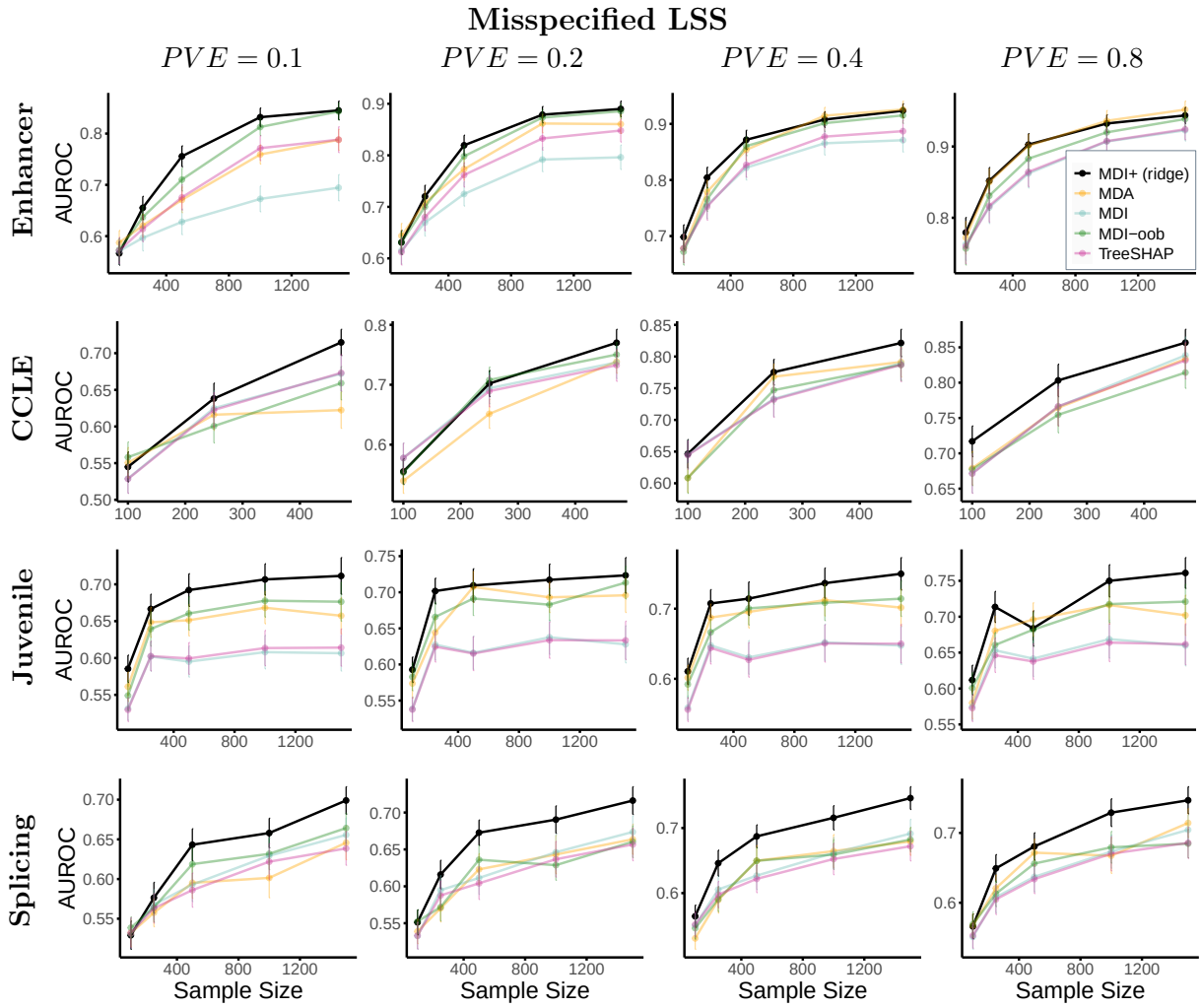


Figure B.14: MDI+ (ridge) outperforms other feature importance methods across almost all misspecified LSS regression simulations described in Appendix B.3. This pattern is evident across various datasets with different covariate structures (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

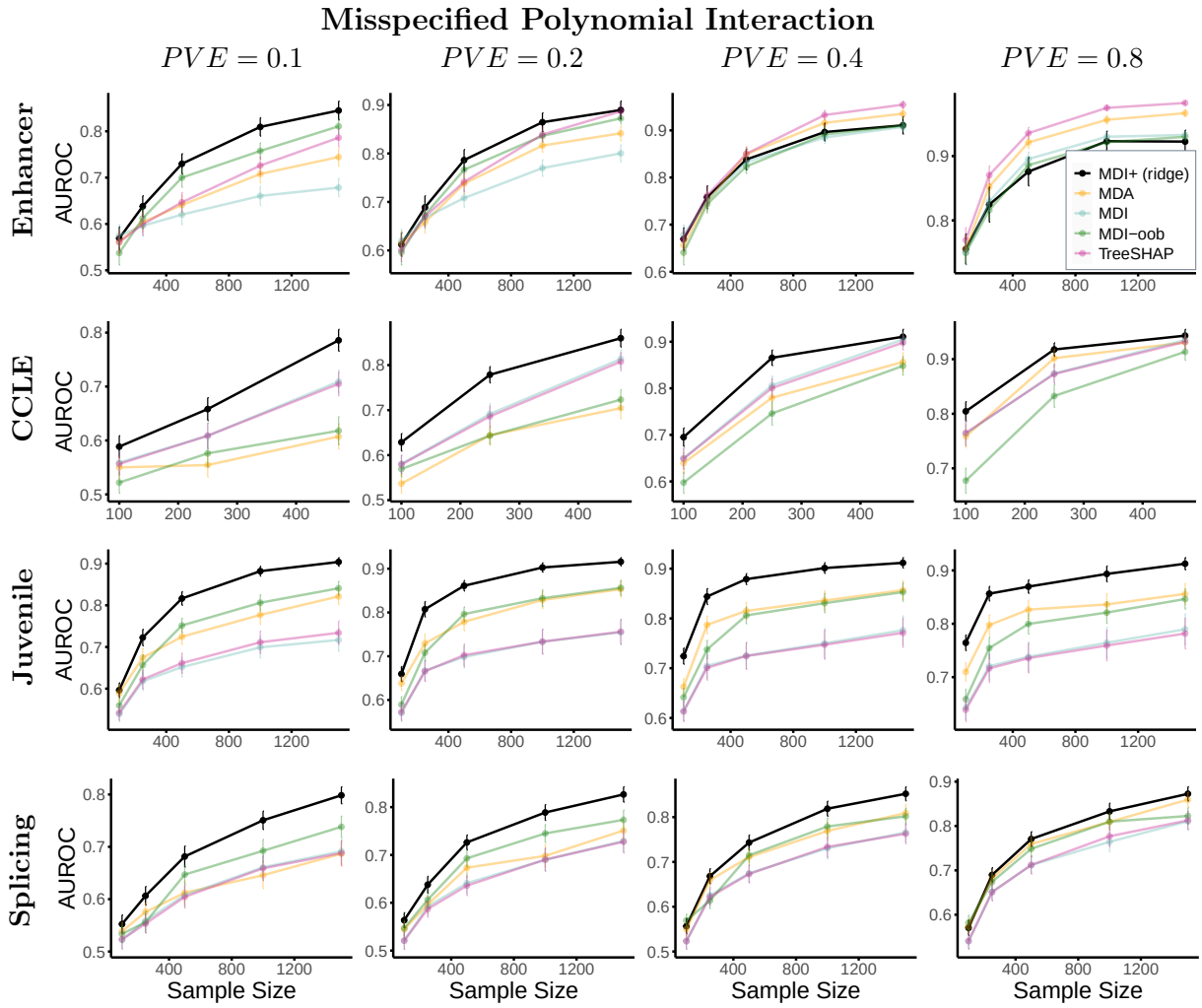


Figure B.15: MDI+ (ridge) outperforms other feature importance methods across almost all misspecified polynomial interaction regression simulations described in Appendix B.3. This pattern is evident across various datasets with different covariate structures (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

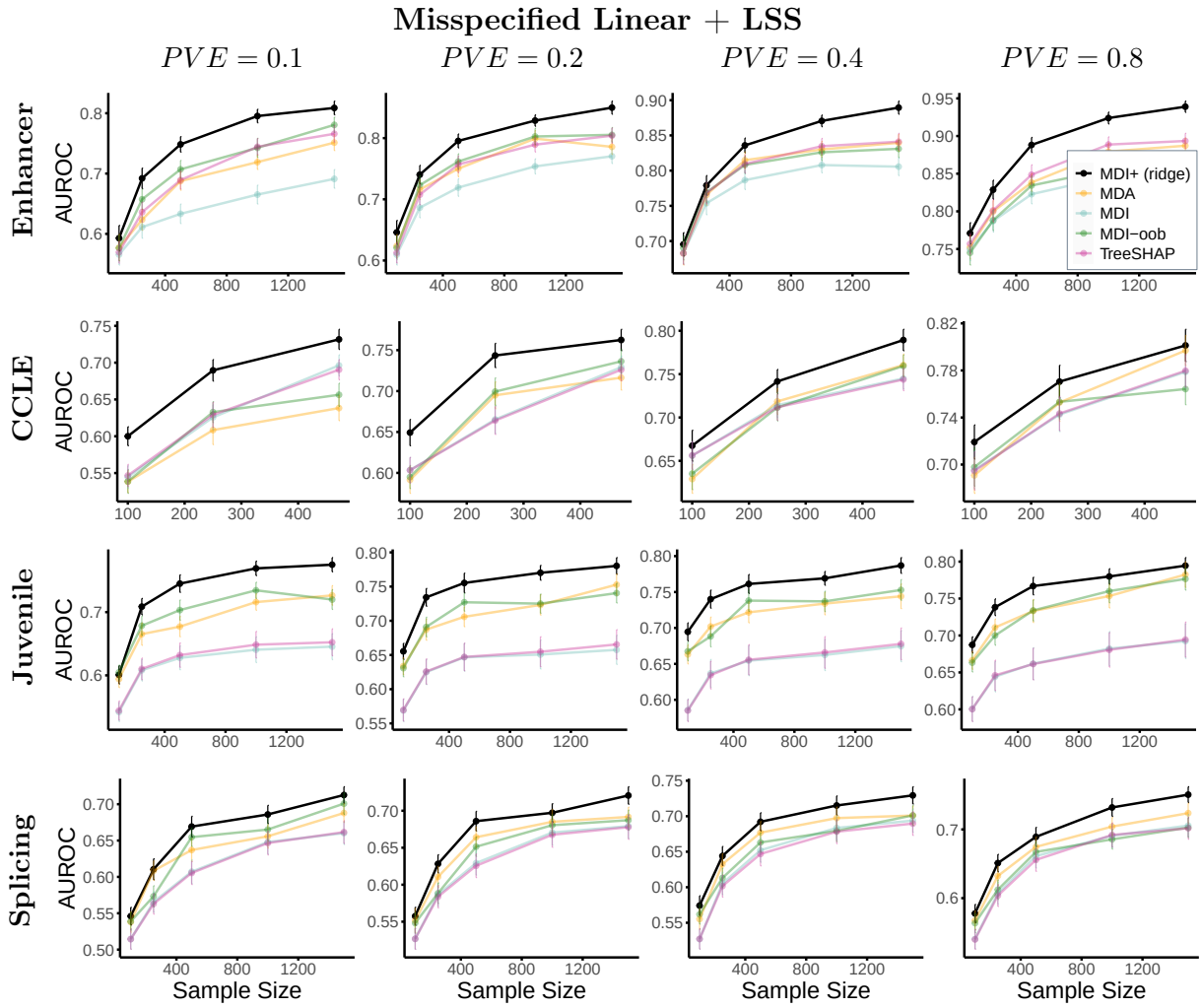


Figure B.16: MDI+ outperforms other feature importance methods for the linear + LSS regression function described in Section 3.5. This pattern is evident across various datasets (specified by row), proportions of variance explained (specified by column), and sample sizes (on the x-axis). In all subplots, the AUROC has been averaged across 50 experimental replicates, and error bars represent $\pm$ 1SE

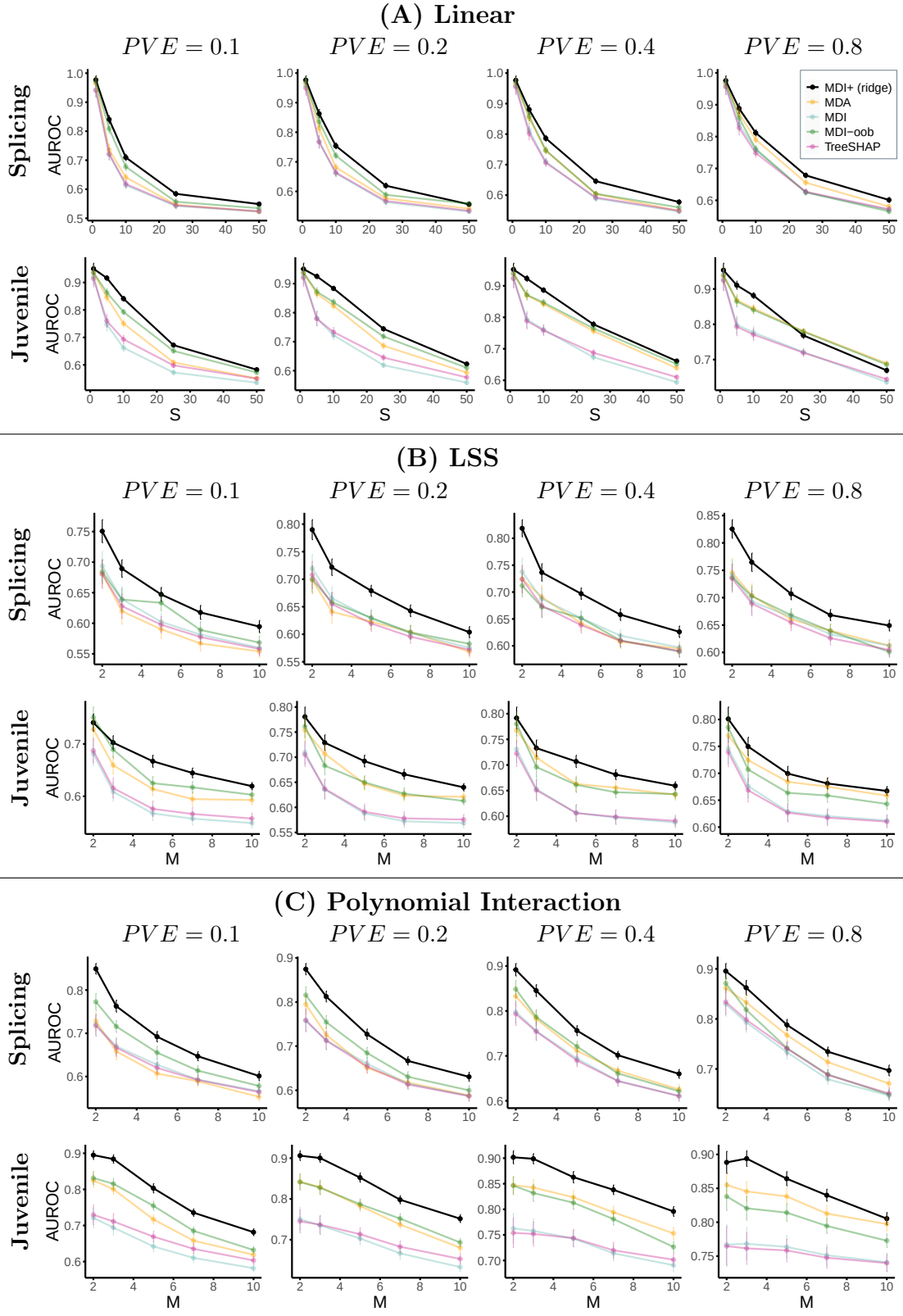
Varying Sparsity Simulations

In this section, we compare the performance of MDI+ against its competitors as we vary the sparsity of the regression function.

Experimental details. Using the Juvenile and Splicing datasets as the covariate matrices $\mathbf{X}$, we simulate the responses $\mathbf{y}$ according to the four regression functions described in Section 3.5 (i.e., linear, LSS, polynomial interaction, and linear+LSS), and we vary the sparsity

level of the regression function. Specifically, for the linear regression function, we vary S, which denotes the number of signal features used in the regression function. For the LSS, polynomial interaction, and linear+LSS models, we vary M, which represents the number of interaction terms. For the simulations described in this section, we take $n = 1000$ samples and evaluate the performance across varying proportions of variance explained ($PVE = 0.1, 0.2, 0.4, 0.8$).

Results. Our results are summarized in Figure B.17, which shows that MDI+ significantly outperforms competitors across various sparsity levels, as measured by AUROC.


(A) Linear
$PVE = 0.1$
$PVE = 0.2$
$PVE = 0.4$
$PVE = 0.8$
Splicing
Juvenile
AUROC
S
MDI+ (ridge)
MDA
MDI
MDI−oob
TreeSHAP
(B) LSS
M
(C) Polynomial Interaction

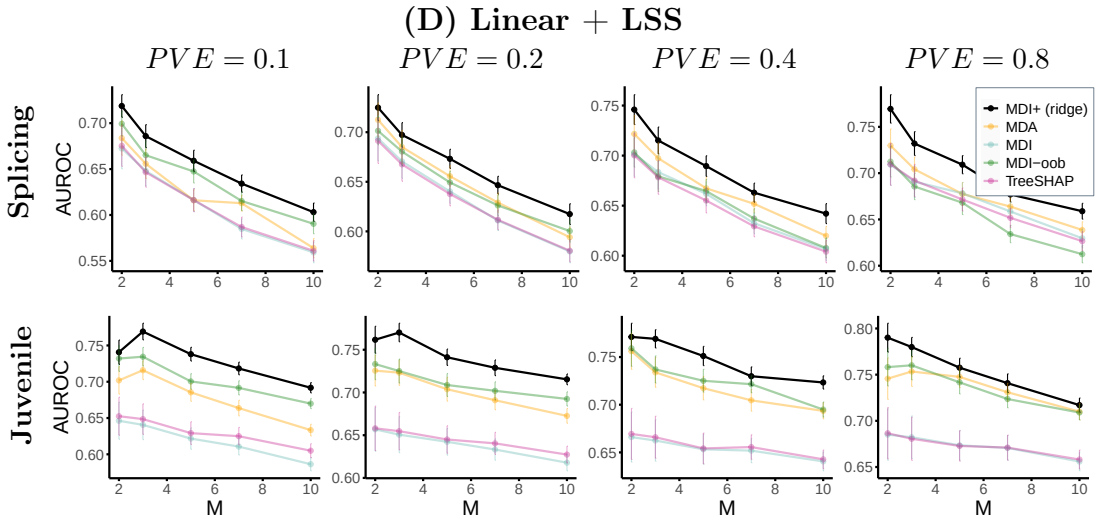


Figure B.17: In the regression simulations described in Appendix B.3, MDI+ outperforms other feature importance methods across a variety of sparsity levels (specified on the x-axis). This pattern is evident across various regression functions (specified by panel), datasets with different covariate structures (specified by row), and proportions of variance explained (specified by column).

Varying Number of Features Simulations

In this section, we compare the performance of MDI+ against its competitors under a variety of real data-inspired simulations as we vary the number of features in the covariate matrix $\mathbf{X}$.

Experimental details: Using the CCLE gene expression dataset as the covariate matrix, we simulate the responses $\mathbf{y}$ according to the four regression functions described in Section 3.5 (i.e., linear, LSS, polynomial interaction, and linear+LSS), and we vary the number of features p in the covariate matrix $\mathbf{X}$. Specifically, for each choice of $p = 10, 25, 50, 100, 250, 500, 1000, 2000$, we subsample the desired number of columns from the full CCLE gene expression dataset, which originally consists of $51,658$ genes (or features). For the simulations described in this section, we take the max number of samples in the CCLE dataset ($n = 472$) and evaluate the performance across varying proportions of variance explained ($PVE = 0.1, 0.2, 0.4, 0.8$).

Results: Our results are summarized in Figure B.18, which shows that MDI+ significantly outperforms competitors in terms of AUROC, regardless of the number of features in the covariate matrix $\mathbf{X}$.

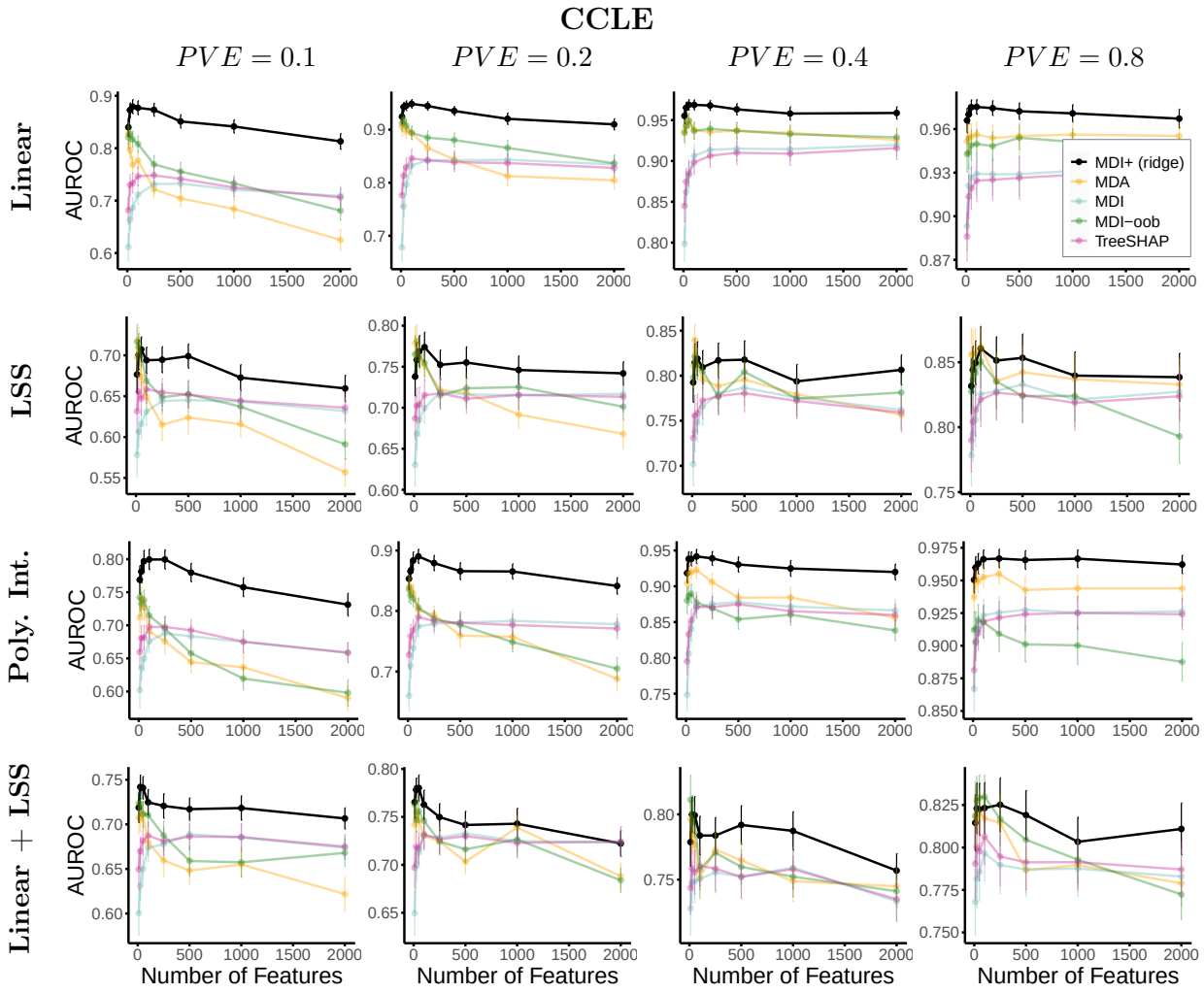


Figure B.18: In the regression simulations described in Appendix B.3, MDI+ (ridge) outperforms other feature importance methods regardless of the number of features in the covariate matrix $\mathbf{X}$ (specified on the x-axis). This pattern is evident across various regression functions (specified by row) and proportions of variance explained (specified by column).

B.4 Justifying MDI+ Choices

In this section, we provide justification for a number of MDI+ choices: regularization, including the raw feature, and evaluating predictions via LOO. That is, we provide simulations that show how inclusion of each of these choices in MDI+ leads to an increase in feature ranking AUROC. Our experimental details and results are as follows.

Experimental details. We use the Enhancer and CCLE gene expression datasets as our

covariate matrix $\mathbf{X}$, and simulate responses $\mathbf{y}$ according to the linear and polynomial interaction function described in Section 3.5. As in Section 3.5, we vary the number of samples n across $\{100, 250, 500, 1000, 1500\}$ for the Enhancer dataset and $\{100, 250, 472\}$ for the CCLE gene expression dataset. We also vary PVE in $\{0.1, 0.2, 0.4, 0.8\}$. In order to illustrate the impact of the MDI+ modeling choices, we consider the following sequence of models:

1. **MDI**: equivalent to MDI+ using OLS as the GLM without the raw feature and evaluating R^2 on the in-bag samples.
2. **MDI-oob**: equivalent to MDI+ using OLS as the GLM without the raw feature and evaluating R^2 on the out-of-bag samples.
3. **MDI+ (raw only)**: MDI+ using OLS as the GLM with the raw feature and in-bag evaluation
4. **MDI+ (loo only)**: MDI+ using OLS as the GLM without the raw feature but using LOO evaluation on the entire dataset
5. **MDI+ (raw+loo only)**: MDI+ using OLS as the GLM with the raw feature and LOO evaluation on the entire dataset
6. **MDI+ (ridge+raw+loo)**: MDI+ using ridge as the GLM with the raw feature and LOO evaluation (i.e., the default MDI+ settings described in Section 3.5).

We perform the simulation with a RF regressor using $min\_samples\_leaf = \{1, 5\}$ alongside other default parameters and average the performance of each feature importance method across 50 simulation replicates. The results for $min\_samples\_leaf = \{1, 5\}$ are displayed in Figures B.20 and B.19 respectively.

Results. From Figures B.19 and B.20, MDI+ with the default regression settings, labeled in black as MDI+ (ridge+raw+loo), most consistently outperforms MDI, MDI-oob, and other MDI+ configurations in terms of the AUROC across the various regression functions, datasets, proportions of variance explained, and choices of *min_samples_per_leaf*. Moreover, when using a RF regressor with *min_samples_per_leaf*=5 in Figure B.19, we see the added benefits of including the raw feature and/or using LOO evaluation within the MDI+ framework, compared to MDI which does not include the raw feature and uses in-bag evaluation. This point becomes more obscure when using a RF regressor trained to purity with *min_samples_per_leaf*=1. Here, MDI outperforms MDI-oob, MDI+ (loo only), and MDI+ (raw+loo only) on the CCLE gene expression dataset, that is, in a small n, large p setting. We hypothesize that this is due to overfitting and thus high instability given that the trees are being grown to purity. By incorporating shrinkage using ridge regression (Agarwal et al., 2022) as opposed to OLS as the GLM within the MDI+ framework, we are able to mitigate this instability and regain the added benefits of including the raw feature and LOO evaluation, as illustrated by the strong performance of MDI+ (ridge+raw+loo).

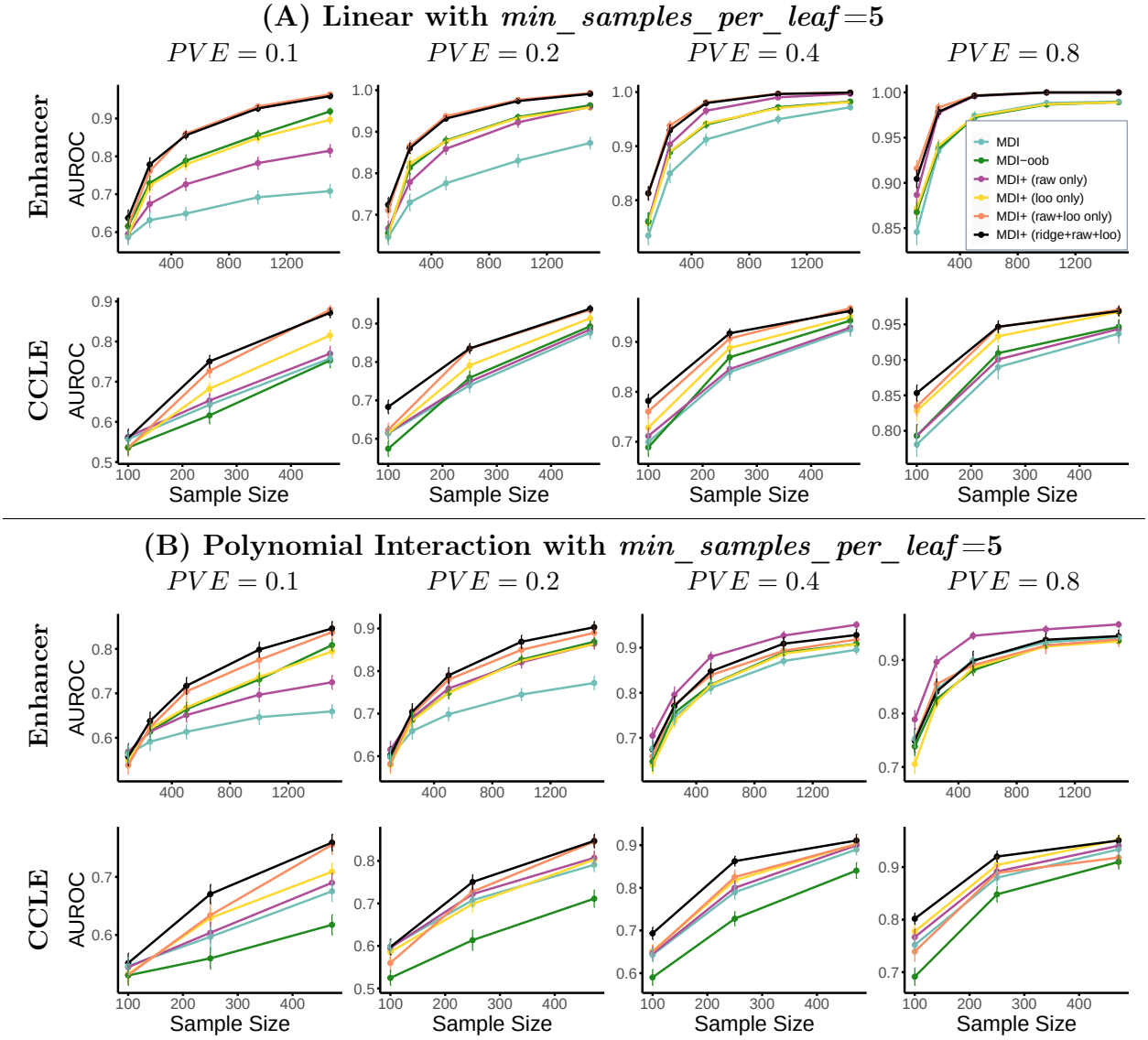


Figure B.19: We illustrate the impact of various MDI+ choices, namely, regularization (via ridge regression), including the raw feature, and evaluating predictions via LOO, applied to a RF regressor with *min_samples_per_leaf*=5. MDI+ with ridge (i.e., shrinkage), including the raw feature, and LOO evaluation (black) consistently outperforms or is on par with other MDI+ modeling choices for random forests across various regression functions (specified by panel), datasets with different covariate structures (specified by row), and proportions of variance explained (specified by column).

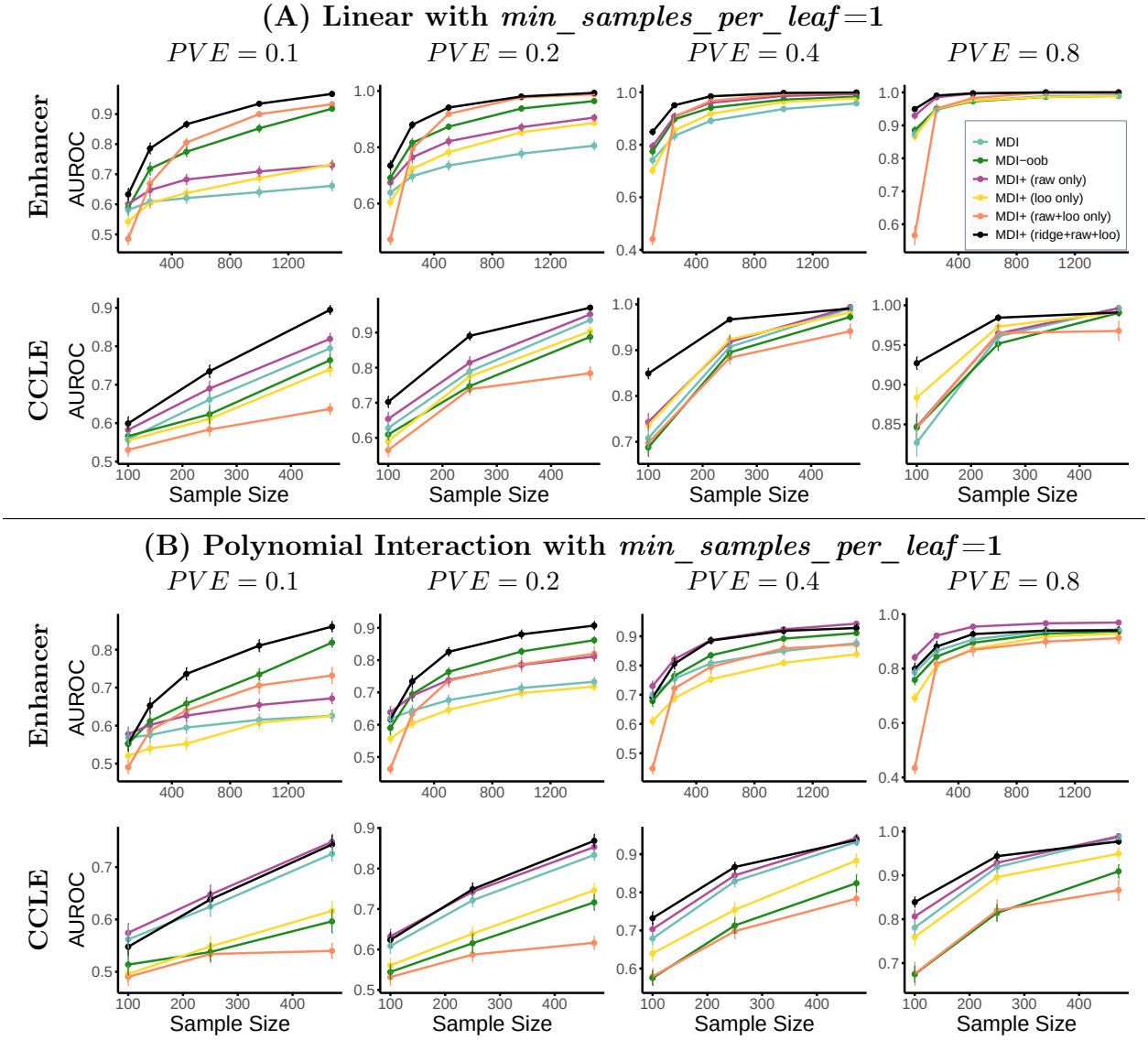


Figure B.20: We illustrate the impact of various MDI+ choices, namely, regularization (via ridge regression), including the raw feature, and evaluating predictions via LOO, applied to a RF regressor with *min_samples_per_leaf*=1. MDI+ with ridge (i.e., shrinkage), including the raw feature, and LOO evaluation (black) consistently outperforms or is on par with other MDI+ modeling choices for random forests across various regression functions (specified by panel), datasets with different covariate structures (specified by row), and proportions of variance explained (specified by column).

B.5 MDI Biases Simulations

In this section, we provide additional simulations regarding the biases of MDI against highly-correlated and low-entropy features, as well as the ability of MDI+ to overcome these biases. Throughout this section, we follow the simulation set-up discussed in Section 3.6.

Correlation Bias

In Figure B.21, we display average percentage of RF splits per feature in each group (i.e., Sig, CSig, and NSig). Further, in Figure B.22 we examine the performance of MDI+ with and without the LOO data splitting scheme. As seen in Figure B.22, for both $PVE = 0.1, 0.4$, LOO sample-splitting overcomes the correlation bias that using in-bag samples induces.

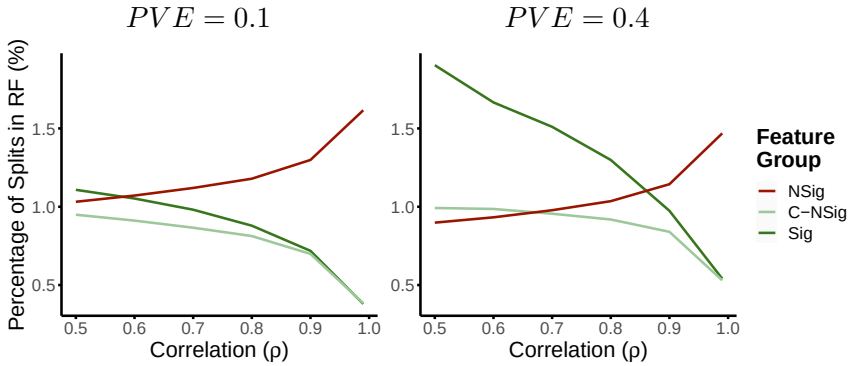


Figure B.21: As the correlation increases, the percentage of splits in the RF that are made using features from the correlated group (Sig or C-NSig) decreases. This pattern is true for both $PVE = 0.1$ (left) and $PVE = 0.4$ (right) under the correlation simulation setup described in Section 3.6.

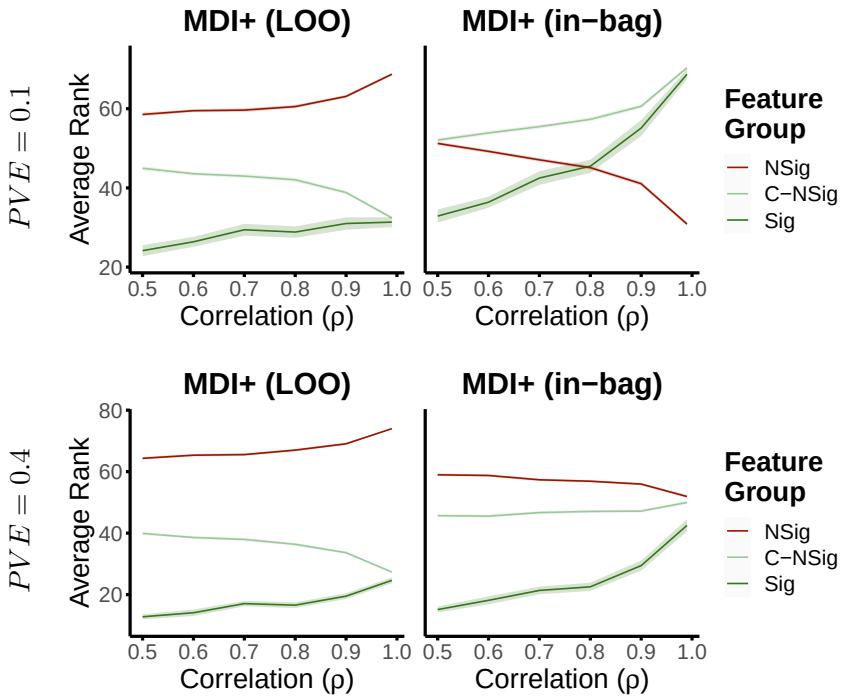


Figure B.22: While MDI+ using the LOO partial model predictions is able to mitigate the correlated feature bias, MDI+ without the LOO scheme suffers from the correlated feature bias like MDI. This pattern holds for both the $PVE = 0.1$ (top) and the $PVE = 0.4$ (bottom) simulation settings described in Section 3.6.

Entropy Bias

In Figure B.23, we display average percentage of RF splits per feature in both the regression and classification setting. Further, in Figure B.24 we examine the performance of MDI+ with and without ridge regularization and the LOO data splitting scheme in both the regression and classification setting. As seen in Figure B.24, LOO sample-splitting overcomes the entropy bias in both settings. In the regression setting, l_2 regularization also helps to mitigate the entropy bias of MDI.

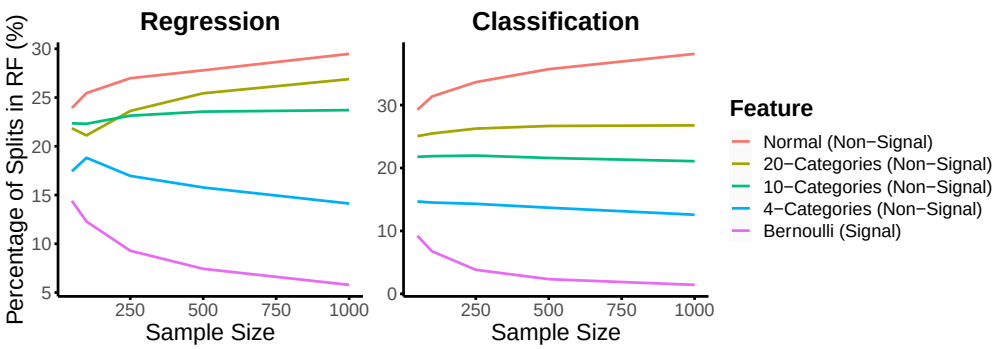


Figure B.23: Noisy features with higher entropy are inherently split on more frequently in the RF. This pattern is true across various sample sizes under both the regression and classification simulation settings described in Section 3.6.

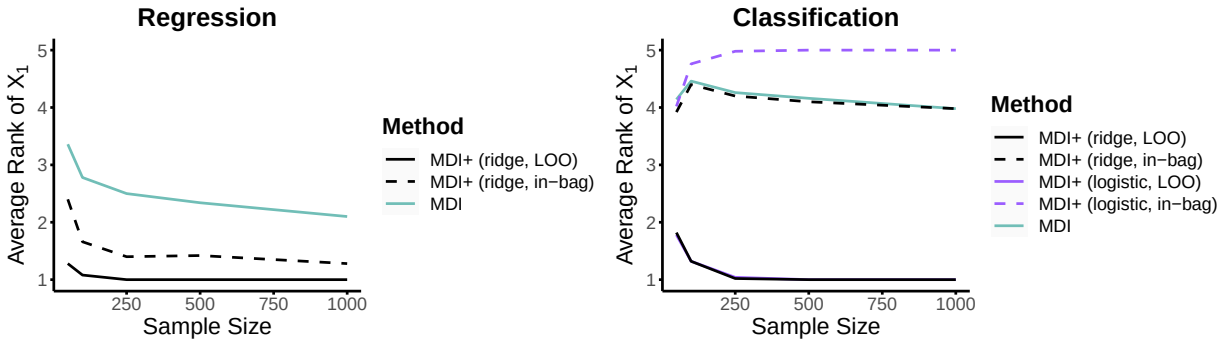


Figure B.24: While MDI+ using the LOO partial model predictions is able to mitigate the entropy bias, MDI+ without the LOO scheme suffers from the entropy bias like MDI. This pattern holds for both the regression (left) and the classification (bottom) simulation settings described in Section 3.6.

B.6 PCS Model Recommendations

In this section, we provide further details on the stability-driven selection and model aggregation procedure discussed in Section 3.4. We also establish the efficacy of both approaches through a data-inspired simulation study.

For both approaches, the first step is to evaluate the prediction performance of each $h \in \mathcal{H}$ on an independent test set (obtained from a train-test split). Filter out all h whose predictive performance is worse than that of RF. Given this set of screened models $\{h_1, \ldots, h_M\}$, we propose the following two approaches to compute feature rankings.

Stability-based selection. Let $\mathcal{T} = \{\mathcal{S}_1, \ldots, \mathcal{S}_T\}$ denote the fitted trees of the RF trained on the data $\mathcal{D}_n$. Generate bootstrap samples of the fitted trees $\mathcal{T}_b$, $b = 1, \ldots, B$. For each bootstrap sample, $\mathcal{T}_b$, compute feature rankings $R_b(h_i)$ for the MDI+ model h_i. Denote the set of feature rankings generated by h_i as $\mathcal{R}_i = \{R_1(h_i), \ldots, R_B(h_i)\}$. To evaluate the stability of h_i, we measure the similarity between rankings from all pairs of the B bootstraps, and take the average across these $\binom{B}{2}$ similarity scores. To measure this similarity between two ranked lists, we use Rank-based Overlap (RBO) (Webber et al., 2010). RBO measures how frequently two ranked lists agree on ordering of items, weighing agreement between higher ranks more heavily than lower ranks. This makes it appropriate for feature importance rankings where we are often concerned with the most important features (i.e., the highest ranking features). Finally, we choose h^* to be the $h \in \{h_1, \ldots, h_M\}$ with the highest RBO score averaged across all $\binom{B}{2}$ pairs of bootstraps.

Model aggregation. While the approach above results in feature importance rankings that correspond to a single MDI+ model, it is sometimes desirable to obtain rankings that do not depend on a single choice of MDI+ model. Hence, we propose an ensemble approach where we rank features based upon the median ranking of each feature X_k across all prediction-screened MDI+ models $h \in \{h_1, \ldots, h_M\}$. We refer to this as MDI+ (ensemble). Other ways of aggregating feature importances can also be performed; we leave this to future work.

Simulation Study. We evaluate the effectiveness of these two PCS model selection techniques for feature importance ranking via the following simulation study. We follow the simulation set-up described in Section 3.5. We consider ridge and LASSO as the two (regularized) GLMs, and R^2 and mean absolute error (MAE) as the two metrics, producing a total of four candidate MDI+ models. We present here the results using the CCLE dataset with responses generated by a linear and polynomial interaction model as described in Section 3.5. We note that all four MDI+ models passed the predictive check. In Figure B.25, we thus plot the RBO for all four MDI+ models, where a higher RBO indicates greater ranking stability. Additionally, we display the feature ranking performance (as measured by AUROC) for these four MDI+ models, the ensemble approach (MDI+ (ensemble)), as well as MDI and MDI-oob. As seen in Figure B.25, MDI+ (ridge, R^2) has the highest stability score, which often translates to the best feature ranking performance across various sample sizes and $PVEs$. MDI+ (ensemble) also performs reasonably well, typically yielding an AUROC for feature ranking accuracy in between the different MDI+ versions and better than the baselines of MDI and MDI-oob. Despite approach 1 providing the best feature ranking accuracy in these simulations, these simulations may not capture all subtleties of applying these methods in practice. Thus, MDI+ (ensemble) may prove useful in practice.

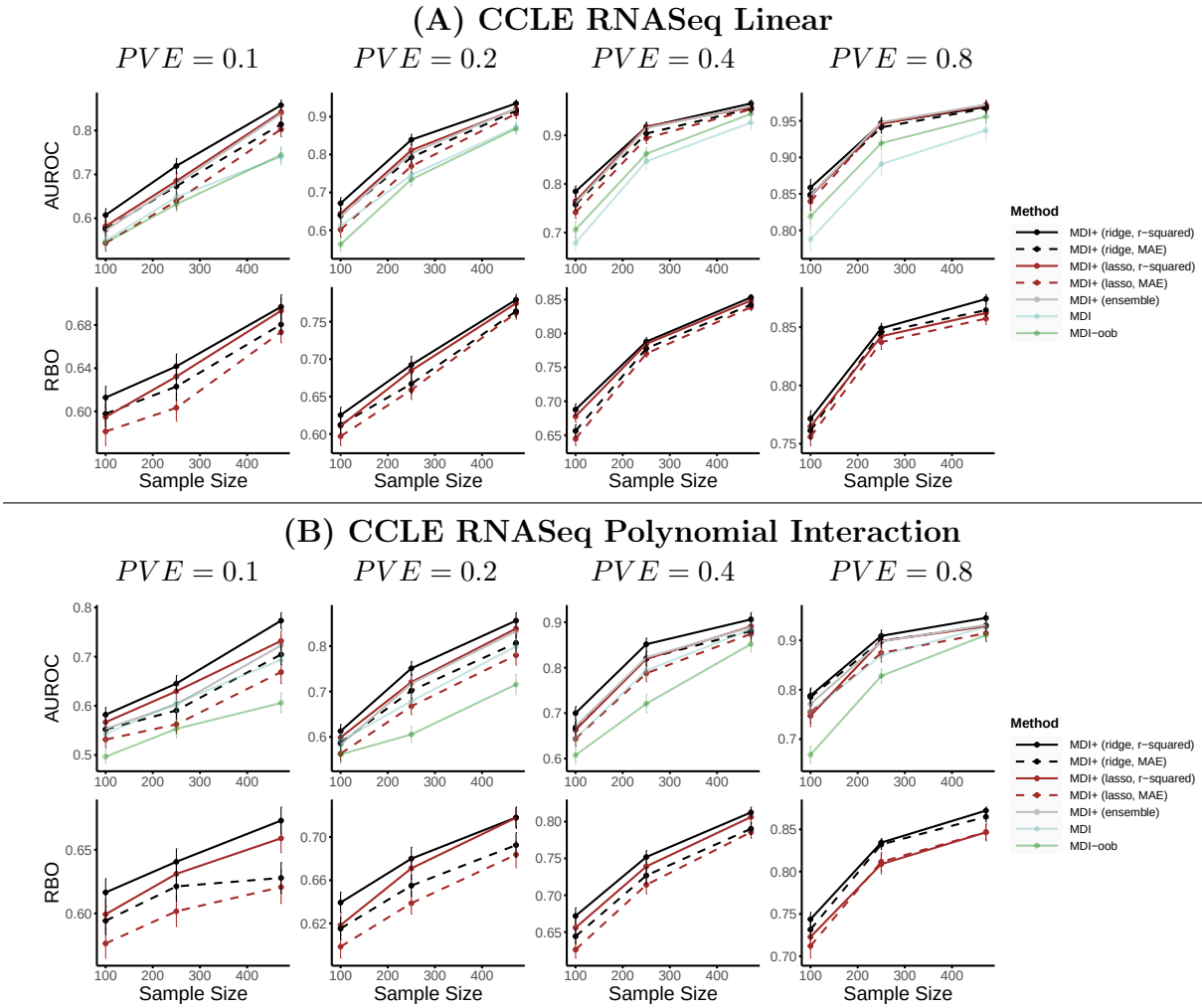


Figure B.25: In each panel, the top row shows the accuracy of the feature importance rankings using various methods, and the bottom row shows the stability scores, as measured by RBO, for each choice of GLM and metric used in MDI+. The GLM and metric yielding the most stable feature rankings generally gives the most accurate feature importance rankings. MDI+ (ensemble) typically falls in between the different MDI+ versions and outperforms existing methods (MDI and MDI-oob) in terms of feature importance ranking accuracy. These patterns hold across various choices of sample sizes (specified on the x-axis) and regression functions (specified by panel).

B.7 RF+ Improves Prediction Performance

In this section, we show that RF+ increases prediction performance for various real-world datasets. Alongside stability, predictability is a crucial pre-requisite for interpretation (Murdoch et al., 2019). The improved predictive performance of RF+ suggests it better fits the underlying DGP, giving additional credence to MDI+.

Regression. We use the CCLE RNASeq gene expression data described in Section 3.7 to predict the response for 24 drugs, each independently in separate regression problems. We split the data into 80% training and 20% testing. The prediction models under study are the default `scikit-learn` RF regressor and RF+ using ridge regression as the GLM and the R^2 metric. We evaluate the relative difference in prediction performance as measured by the R^2 between RF+ and RF for each drug averaged across 32 different random train-test splits. We present the results for drugs, where the vanilla RF yielded an average test-set $R^2 > 0.1$ in Figure B.26(A). We use this prediction screening threshold to focus on models that have non-trivial predictive power. Here, RF+ increases R^2 performance for 17 out of 18 drugs. Averaged across all drugs, RF+ increases performance by an average of 4.4% relative to RF. The full results including all drugs can be found in Figure B.27.

Classification. We use the Enhancer, Splicing and Juvenile datasets described in Section 3.5. Additionally, we use gene expression data from The Cancer Genome Atlas (TCGA) to predict the PAM50 breast cancer subtypes (a multiclass classification problem with five different subtype labels) (Atlas, 2012). We split the data into 80% training and 20% testing. We use the default `scikit-learn` RF classifier and RF+ using l_2-regularized logistic regression as the GLM with the log-loss metric. We evaluate the relative difference in prediction performance as measured by the $F1$ score and AUPRC between RF+ and RF, averaged across 32 different random train-test splits. The results are displayed in Figure B.26(B). RF+ increases the F1 score for three of the four datasets and has approximately the same F1 score for the splicing dataset. On average, RF+ increases F1 score by an average of 3.6% relative to RF. RF+ also increases the AUPRC score for all four datasets, and on average by 2.3% relative to RF.

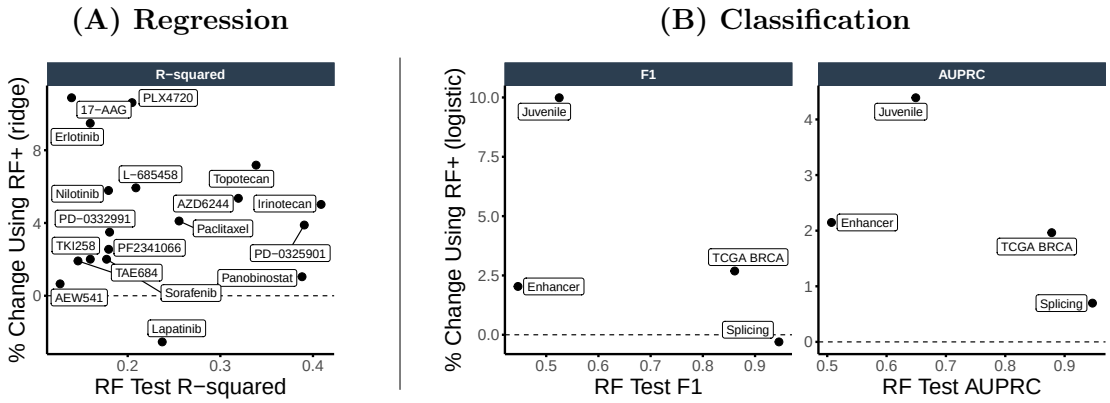


Figure B.26: Relative Performance of RF+ (ridge) as compared to RF in both the (A) regression and (B) classification settings. In the regression setting, RF+ (ridge) increases performance by an average of 4% averaged across all 18 drugs that have test set $R^2 > 0.1$. In the classification setting, RF+ (logistic) increases performance according to F1 score for three of the four datasets, and on average by 3.75%. RF+ (logistic) increases AUPRC for all datasets, and on average by 2.5%.

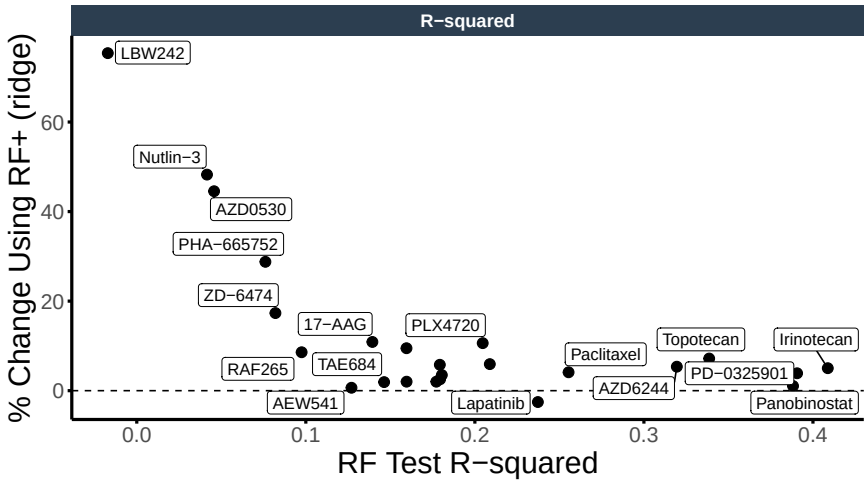


Figure B.27: Relative performance of RF+ (ridge) as compared to RF for all 24 CCLE drugs in the regression setting. Here, we measure prediction performance using the test set R^2. Averaged across 32 train-test splits, RF+ (ridge) yields higher test prediction performance than RF for 23 out of 24 drugs.

B.8 Case Studies

In this section, we provide additional background and results regarding the two case studies on drug response prediction and breast cancer subtype prediction.

Drug Response Prediction

Data preprocessing. Originally, the CCLE RNASeq gene expression data set consisted of 50114 genes. To reduce the number of features to a more manageable size for our analysis, we took the top 5000 genes with the highest variance. Filtering the number of features in this way is both common and often beneficial especially in this ultra-high-dimensional regime (Fan and Lv, 2008). The unprocessed CCLE data can be downloaded from DepMap Public 18Q3 (`https://depmap.org/portal/download/`).

Details of drug response outcome variable. In this case study, the outcome of interest is the efficacy, or response, of a drug. To obtain a measure of the drug response for each cell line, the CCLE project measured the pharmacological sensitivity of each drug in vitro across eight different dosages and quantified this using the *activity area*, or area under the dose-response curve. This activity area is measured on an 8-point scale with 0 corresponding to an inactive compound (i.e., a drug that did not inhibit the growth of the cancer cells across all 8 dosages) and 8 corresponding to a compound with 100% inhibition of cancer cell growth at all 8 dosages. Further details on the CCLE data can be found in Barretina et al. (2012).

Results. To supplement the discussion in Section 3.7, we provide additional tables and figures below.

In Table B.1, we list the top 5 gene expression predictors for each drug according to various feature importance methods. Since we evaluated the feature importance methods across 32 train-test splits (details in Section 3.7), the genes are ranked according to their average feature importance ranking across the 32 splits.

We also provide the stability distribution plots for all 24 drugs in Figure B.28. Specifically, for the top 5 genes using each feature importance method under study (i.e., those listed in Table B.1), we plot the standard deviation of the gene's feature importance rankings across the 32 train-test splits. A small standard deviation indicates a more stable feature importance measure. For many drugs, these stability distribution plots are skewed by the high instability of MDI-oob and MDA. We thus show the same plot in Figure B.29, excluding MDI-oob and MDA. Across all drugs, we see that MDI+ generally yields the most stable rankings for the top 5 features across the 32 train-test splits.

To further evaluate the stability of MDI+ across only the randomness in the RF training, we performed an analogous stability analysis in Figures B.30 and B.31, where we varied the random seed used to train the RF but kept the training data fixed. Specifically, we trained 32 RFs on the full CCLE data, each RF using a different random seed. We then evaluate the stability of the feature rankings across these 32 RF fits. We see in Figures B.30 and B.31 that MDI+ remains the most stable across the different RF fits. This highlights another

practical advantage of MDI+, as it is highly undesirable for the feature importance rankings to change due to an arbitrary choice such as the random seed used in training the RF.

Lastly, in Figure B.32, we evaluated the predictive power of the top k genes from each feature importance method across the 32 train-test (80-20%) splits. Specifically, for each k and each feature importance method, we took the top k ranked genes from the given feature importance method and trained an RF using the training data, restricted to only these k features. We then evaluated the prediction accuracy of the fitted RF on the test set and average these results across the 32 train-test splits. While evidence from the existing scientific literature remains the main source of validation, supporting the top-ranked genes from MDI+, the prediction accuracy of the top-ranked genes (as shown in Figure B.32) can provide another check. In Table B.2, we summarize the prediction results when $k = 10$ by counting the number of drugs, for which each feature importance method gave the best test R^2, second-best test R^2, etc. We see that for 12 out of the 24 drugs, the top 10 genes from MDI+ (ridge) generally had the highest prediction power compared to other methods. In accordance with the predictability principle of the PCS framework, strong prediction performance suggests that the model (and in this case, the top-ranked features) may better capture the underlying data-generating process.

It is important to note, however, that correlated variables can pose subtle issues when interpreting these top k prediction results. For concreteness, consider the scenario where the top k ranked features from Method A tend to be independent while the top-ranked features from Method B tend to be highly-correlated. It is likely that the predictive power from the top k features from Method A is higher than that from Method B simply because the k independent features from Method A inherently contain more information than the k highly-correlated features from Method B. That is, the prediction accuracies from these two sets of k features are not directly comparable and requires additional investigation into the correlation structure of $\mathbf{X}$. From preliminary explorations, this complication is typically most apparent for small values of k. Intuitively, when k is small, correlation structures between two sets of features can be very different. As k grows, the two correlation structures tend to become most similar to each other. This motivated our choice of $k = 10$ in Table B.2. However, further investigation is warranted, and we leave a deeper investigation of this phenomenon to ongoing and future work.

Table B.1: Top 5 most important genes for each drug's response according to various feature importance methods. Genes are ranked by their average feature importance ranking across 32 train-test splits (shown in parentheses).

	MDI+	MDI	TreeSHAP	MDI-oob	MDA
17-AAG					
1	PCSK1N (1.47)	PCSK1N (2.19)	PCSK1N (2.19)	PCSK1N (2.5)	PCSK1N (9.16)
2	MMP24 (3.41)	MMP24 (4.97)	MMP24 (4.06)	NQO1 (9.03)	MMP24 (194.5)
3	RP11-109D20.2 (4.59)	ZSCAN18 (6.94)	ZSCAN18 (8.56)	ZNF667-AS1 (26.59)	ZNF667-AS1 (241.38)
4	ZSCAN18 (8.09)	RP11-109D20.2 (7.41)	RP11-109D20.2 (10.09)	ZSCAN18 (44.03)	RP11-109D20.2 (525.22)
5	NQO1 (8.84)	NQO1 (9.53)	NQO1 (11.81)	TST (49.38)	SH3BP1 (587.94)

Table B.1: *(continued)*

	MDI+	MDI	TreeSHAP	MDI-oob	MDA
AEW541					
1	TXNDC5 (1.41)	TCEAL4 (1.62)	TXNDC5 (1.75)	TXNDC5 (1.5)	TCEAL4 (5.59)
2	ATP8B2 (4.34)	TXNDC5 (3.53)	ATP8B2 (3.84)	ATP8B2 (4.69)	IQGAP2 (238.8)
3	VAV2 (6.03)	ATP8B2 (8.38)	VAV2 (5.41)	VAV2 (5.84)	RP11-343H19.2 (303.25)
4	TNFRSF17 (8.53)	VAV2 (10.47)	TCEAL4 (9.44)	TCEAL4 (6.5)	TXNDC5 (312.62)
5	TCEAL4 (9.03)	PLEKHF1 (19.25)	TNFRSF17 (9.69)	TNFRSF17 (13.56)	ATP8B2 (318.59)
AZD0530					
1	PRSS57 (5.16)	SYTL1 (17.62)	PRSS57 (7.69)	PRSS57 (12.09)	VTN (105.69)
2	SYTL1 (12.31)	PRSS57 (32.5)	SYTL1 (42.94)	DDAH2 (440.16)	SYTL1 (216.62)
3	STXBP2 (15.38)	SFTA1P (36.5)	NFE2 (43.06)	SLC16A9 (484.34)	STXBP2 (245.31)
4	NFE2 (23.12)	STXBP2 (62.59)	STXBP2 (61.62)	STXBP2 (486.09)	ZBED2 (466.48)
5	THEM4 (34.41)	CLDN16 (67.28)	SLC16A9 (61.81)	RAPGEF3 (514.2)	DDAH2 (472.06)
AZD6244					
1	LYZ (1.66)	TOR4A (2.31)	LYZ (1.72)	LYZ (2.53)	LYZ (2.09)
2	SPRY2 (2.34)	SPRY2 (3.31)	RP11-1143G9.4 (3.59)	SPRY2 (2.59)	RP11-1143G9.4 (3.59)
3	RP11-1143G9.4 (2.84)	LYZ (3.69)	SPRY2 (3.91)	TOR4A (3.25)	TOR4A (3.66)
4	ETV4 (5.22)	ETV4 (5.19)	TOR4A (6.41)	RP11-1143G9.4 (4.75)	SPRY2 (4.5)
5	TOR4A (6.41)	RP11-1143G9.4 (6.84)	RNF125 (6.66)	ETV4 (6.91)	ETV4 (6.34)
Erlotinib					
1	CDH3 (1.47)	CDH3 (1.84)	CDH3 (2.03)	CDH3 (1.97)	CDH3 (1.88)
2	RP11-615I2.2 (2.28)	RP11-615I2.2 (3.28)	RP11-615I2.2 (2.88)	RP11-615I2.2 (3.53)	RP11-615I2.2 (3.16)
3	EGFR (4.34)	SPRR1A (3.97)	EGFR (3.97)	SPRR1A (6.25)	SPRR1A (3.84)
4	SPRR1A (4.44)	SYTL1 (7.84)	SPRR1A (4.19)	GJB3 (8.78)	EGFR (8.31)
5	GJB3 (7.44)	EGFR (8.69)	KRT16 (11.41)	EGFR (9.31)	SYTL1 (8.72)
Irinotecan					
1	SLFN11 (1)	SLFN11 (1)	SLFN11 (1)	SLFN11 (1)	SLFN11 (1)
2	S100A16 (4.12)	S100A16 (3.75)	S100A16 (3.84)	S100A16 (3.25)	WWTR1 (6.03)
3	IFITM10 (4.19)	IFITM10 (4.09)	WWTR1 (4.28)	WWTR1 (4.12)	TRIM16L (150.38)
4	WWTR1 (4.94)	WWTR1 (4.78)	IFITM10 (8.03)	RP11-359P5.1 (8.22)	IFITM10 (163.44)
5	PPIC (7.81)	PPIC (10.22)	RP11-359P5.1 (8.47)	IFITM10 (8.41)	S100A16 (182.19)
L-685458					
1	PXK (2.03)	DEF6 (4.62)	PXK (2.03)	PXK (3)	PXK (2.94)
2	DEF6 (4.28)	PXK (4.94)	DEF6 (4.38)	IKZF1 (3.44)	CXorf21 (4.62)
3	CXorf21 (4.84)	CXorf21 (5.44)	CXorf21 (5.62)	CXorf21 (5.75)	DEF6 (4.66)
4	IKZF1 (6.03)	IKZF1 (5.75)	IKZF1 (7.91)	DEF6 (10.31)	IKZF1 (5.47)
5	RP11-359P5.1 (9.09)	RP11-359P5.1 (9.94)	RP11-359P5.1 (12.81)	CTNNA1 (13.88)	RP11-359P5.1 (10.06)
LBW242					
1	SERPINB6 (1.12)	SERPINB6 (1)	SERPINB6 (1.66)	SERPINB6 (1.31)	SERPINB6 (1.56)
2	RGS14 (5.12)	RGS14 (6.66)	RGS14 (3.66)	GPT2 (45.62)	HERC5 (54.31)
3	HERC5 (5.41)	MAGEC1 (7.5)	MAGEC1 (5.53)	GBP1 (179.28)	ITGA1 (73.34)
4	MAGEC1 (7.62)	ITGA1 (10.53)	GBP1 (5.78)	ZNF32 (222.03)	PTGS1 (296.62)
5	GBP1 (8.22)	HERC5 (12.41)	CCL2 (13.5)	IGSF3 (257.88)	GPT2 (316.12)
Lapatinib					
1	ERBB2 (1.06)	ERBB2 (1.5)	ERBB2 (1.41)	ERBB2 (1.69)	ERBB2 (1.47)
2	PGAP3 (3.09)	NA (6.81)	PGAP3 (3.44)	PGAP3 (8.03)	NA (4.31)
3	NA (5.03)	PGAP3 (12.41)	IKBIP (6.19)	C2orf54 (13.09)	PGAP3 (14.03)
4	C2orf54 (6.91)	DPYSL2 (16.16)	NA (6.22)	DPYSL2 (15.41)	PKP3 (20.31)
5	IKBIP (8.28)	PKP3 (16.47)	C2orf54 (7.41)	EMP3 (20.47)	EMP3 (22.38)
Nilotinib					
1	SPN (1.25)	SPN (1.81)	SPN (1.62)	SPN (1.47)	SPN (3.59)
2	GPC1 (3.5)	GPC1 (4.38)	GPC1 (3.5)	GPC1 (3.44)	SELPLG (9.03)
3	TRDC (6.62)	SELPLG (7.5)	TRDC (10.16)	SELPLG (9.97)	KLF13 (26.22)
4	SELPLG (6.78)	KLF13 (16.97)	LMO2 (10.44)	TRDC (10.22)	BCL2 (51.09)
5	LMO2 (9.44)	TRDC (20.22)	CISH (11.03)	LMO2 (10.75)	GPC1 (166.19)
Nutlin-3					
1	RP11-148O21.4 (1.41)	MET (2.53)	RP11-148O21.4 (1.59)	RP11-148O21.4 (1.72)	RAPGEF5 (37)
2	MET (2.53)	RP11-148O21.4 (4.75)	MET (4.06)	LRRC16A (9.56)	G6PD (63.53)
3	BLK (5.12)	LAYN (6.41)	BLK (5.25)	BLK (10.97)	MET (147.78)
4	LRRC16A (5.16)	RPS27L (12.94)	LRRC16A (5.97)	MET (26.09)	BLK (160.06)

Table B.1: *(continued)*

	MDI+	MDI	TreeSHAP	MDI-oob	MDA
5	LAT2 (7.34)	ADD3 (21.03)	LAT2 (7.78)	LAYN (138.84)	RP11-148O21.4 (164.94)
PD-0325901					
1	SPRY2 (1.16)	SPRY2 (1.53)	SPRY2 (1.75)	SPRY2 (1.19)	SPRY2 (1.75)
2	LYZ (2.72)	ETV4 (2.88)	LYZ (2.09)	LYZ (2.88)	LYZ (2.62)
3	ETV4 (2.72)	LYZ (3.59)	ETV4 (3.66)	ETV4 (3.38)	ETV4 (3.47)
4	RP11-1143G9.4 (4.34)	TOR4A (4.56)	RP11-1143G9.4 (4.53)	TOR4A (4.72)	TOR4A (4.88)
5	PLEKHG4B (5.62)	PLEKHG4B (4.66)	PLEKHG4B (5.31)	RP11-1143G9.4 (5.38)	PLEKHG4B (5.5)
PD-0332991					
1	SH2D3C (4.31)	SH2D3C (6.81)	SH2D3C (4.56)	SH2D3C (8)	KRT15 (10.56)
2	FMNL1 (6.56)	FMNL1 (8.38)	HSD3B7 (6.75)	AL162151.3 (9.62)	HSD3B7 (14.5)
3	HSD3B7 (6.59)	AL162151.3 (11.59)	FMNL1 (7.09)	HSD3B7 (11.03)	SEPT6 (16.44)
4	KRT15 (7.19)	TWF1 (12.03)	KRT15 (7.78)	KRT15 (11.97)	PPIC (17.03)
5	AL162151.3 (8.84)	KRT15 (12.56)	TWF1 (8.97)	FMNL1 (16.34)	AL162151.3 (18.34)
PF2341066					
1	ENAH (1.03)	ENAH (1)	ENAH (1.31)	ENAH (1.12)	ENAH (1.06)
2	SELPLG (2.09)	SELPLG (2.81)	SELPLG (2.06)	SELPLG (2.28)	SELPLG (2.47)
3	HGF (3.62)	HGF (3.94)	MET (5.44)	HGF (7.03)	CTD-2020K17.3 (10.31)
4	CTD-2020K17.3 (9.72)	CTD-2020K17.3 (9.41)	HGF (6)	MET (10.69)	HGF (14.06)
5	MET (10.53)	MET (11.5)	MLKL (12)	CTD-2020K17.3 (11.88)	DOK2 (14.41)
PHA-665752					
1	ARHGAP4 (1)	ARHGAP4 (1.06)	ARHGAP4 (1.06)	ARHGAP4 (1.09)	ARHGAP4 (1)
2	CTD-2020K17.3 (2.88)	CTD-2020K17.3 (2.66)	CTD-2020K17.3 (4.25)	CTD-2020K17.3 (2.56)	FMNL1 (4.22)
3	FMNL1 (4.88)	FMNL1 (7.44)	PFN2 (10.84)	PFN2 (24.31)	CTD-2020K17.3 (5.44)
4	PFN2 (8.06)	PGPEP1 (13.28)	FMNL1 (11.78)	FMNL1 (33.12)	INHBB (216.5)
5	PGPEP1 (10.12)	INHBB (18.88)	FDFT1 (18.66)	MICB (59.69)	PGPEP1 (335.97)
PLX4720					
1	RXRG (1)	RXRG (1.16)	RXRG (1)	RXRG (1)	RXRG (1.03)
2	MMP8 (5.19)	MMP8 (3.97)	MMP8 (5.16)	MMP8 (5.56)	MMP8 (4.88)
3	RP11-164J13.1 (6.28)	RP11-599J14.2 (7.22)	MYO5A (7.09)	MYO5A (6.81)	MYO5A (8.75)
4	RP11-599J14.2 (7)	AP1S2 (8.47)	LYST (7.97)	LYST (11.31)	AP1S2 (10.53)
5	RP4-718J7.4 (7.84)	LYST (9.34)	RP11-599J14.2 (8.41)	RP4-718J7.4 (13.22)	RP11-599J14.2 (12.69)
Paclitaxel					
1	MMP24 (1.09)	MMP24 (1.28)	MMP24 (1.22)	MMP24 (2.38)	SH3BP1 (10.38)
2	AGAP2 (3.16)	SH3BP1 (2.75)	AGAP2 (3.44)	SH3BP1 (2.88)	PRODH (60.41)
3	SH3BP1 (3.5)	AGAP2 (4.06)	SH3BP1 (3.78)	SLC38A5 (3.84)	AGAP2 (157.41)
4	SLC38A5 (4.34)	SLC38A5 (4.22)	PTTG1IP (3.91)	AGAP2 (5.88)	SLC38A5 (179.34)
5	PTTG1IP (4.72)	PTTG1IP (4.34)	SLC38A5 (4.31)	PTTG1IP (6.97)	MMP24 (308.66)
Panobinostat					
1	AGAP2 (1.12)	AGAP2 (1.56)	AGAP2 (1.81)	AGAP2 (2.16)	CYR61 (2.56)
2	CYR61 (2.44)	CYR61 (2.09)	CYR61 (2.16)	CYR61 (2.78)	AGAP2 (3.25)
3	RPL39P5 (4.19)	RPL39P5 (4.41)	RPL39P5 (3.75)	RPL39P5 (3.88)	RPL39P5 (152.44)
4	WWTR1 (5.16)	WWTR1 (5.78)	WWTR1 (5.94)	WWTR1 (6.53)	S100A2 (316.66)
5	MYOF (6.56)	MYOF (6.16)	IKZF1 (12.41)	IKZF1 (9.72)	MYOF (366.28)
RAF265					
1	CMTM3 (1.34)	CMTM3 (1.5)	CMTM3 (1.69)	SH2B3 (34)	CMTM3 (7.06)
2	SYT17 (5.69)	SYT17 (5.66)	SYT17 (7.69)	CMTM3 (155.25)	SH2B3 (470.66)
3	SH2B3 (6.03)	SH2B3 (8.91)	SH2B3 (17.5)	SLC29A3 (159)	SYT17 (652.84)
4	EMILIN2 (11.94)	SLC29A3 (11.84)	STAT5A (19.66)	PRKCQ (235)	RGS16 (713.47)
5	STAT5A (12.47)	NA (19.91)	SLC29A3 (22.22)	LCP2 (259.7)	AC007620.3 (1087.11)
Sorafenib					
1	PXK (4.47)	TP63 (6.72)	PXK (4.12)	FAM212A (6.41)	PXK (9.34)
2	P2RX1 (4.62)	P2RX1 (7.78)	FAM212A (5.75)	P2RX1 (8)	ARHGAP9 (34.25)
3	FAM212A (4.69)	PXK (8.88)	STAC3 (7.16)	SEC31B (41.69)	P2RX1 (156.91)
4	STAC3 (5.16)	FAM212A (16.97)	P2RX1 (7.25)	ARHGAP9 (43.19)	FAM212A (187.31)
5	ARHGAP9 (7.91)	STAC3 (20.16)	TP63 (40.97)	CXCL8 (57.72)	SEC31B (222.41)
TAE684					
1	SELPLG (1.09)	SELPLG (1.06)	SELPLG (1.12)	SELPLG (1.03)	SELPLG (1.34)
2	IL6R (3.19)	ARID3A (8.12)	IL6R (3.34)	IL6R (6.41)	ARID3A (18.31)
3	NFIL3 (6.34)	GALNT18 (8.62)	NFIL3 (6.25)	NFIL3 (8.06)	FMNL1 (25.25)

Table B.1: *(continued)*

	MDI+	MDI	TreeSHAP	MDI-oob	MDA
4	ARID3A (7)	IL6R (10.16)	RRAS2 (10.19)	ARID3A (16.38)	RP11-334A14.2 (144.34)
5	RRAS2 (8.66)	PPP2R3A (15.69)	FMNL1 (10.19)	RRAS2 (17.97)	PPP2R3A (165.84)
TKI258					
1	TWF1 (2.44)	TWF1 (3.31)	TWF1 (2.41)	TWF1 (2.28)	LAPTM5 (20.84)
2	SLC43A1 (2.88)	GPR162 (3.75)	PRTN3 (3.31)	LAPTM5 (5.34)	SLC43A1 (156.12)
3	PRTN3 (5.25)	SLC43A1 (6.84)	SLC43A1 (6.06)	PRTN3 (6.12)	TWF1 (163.78)
4	LAPTM5 (5.78)	TTC28 (8.94)	LAT2 (6.62)	SLC43A1 (14.91)	GPR162 (169.66)
5	LAT2 (5.94)	LAPTM5 (11.53)	LAPTM5 (8.19)	LAT2 (17.41)	LYL1 (387.97)
Topotecan					
1	SLFN11 (1)	SLFN11 (1)	SLFN11 (1)	SLFN11 (1)	SLFN11 (1)
2	HSPB8 (2.16)	HSPB8 (2.28)	HSPB8 (2.84)	HSPB8 (2.06)	HSPB8 (2.62)
3	PPIC (5.69)	OSGIN1 (5.28)	OSGIN1 (5.31)	PPIC (7.81)	OSGIN1 (7.19)
4	OSGIN1 (5.88)	AGAP2 (8.81)	PPIC (6.81)	RP11-359P5.1 (10.75)	AGAP2 (15.25)
5	AGAP2 (6.53)	PPIC (8.91)	RP11-359P5.1 (8.25)	CORO1A (12.06)	HMGB2 (21.53)
ZD-6474					
1	MAP3K12 (1)	MAP3K12 (1)	MAP3K12 (1)	MAP3K12 (1.09)	MAP3K12 (1)
2	PIM1 (5.47)	CTSH (21.88)	PIM1 (10.28)	PIM1 (31.44)	SCD5 (339.58)
3	PRKCQ (9.03)	TIMP1 (26.31)	PRKCQ (15.28)	DYNLT3 (192.12)	ITGA10 (527.41)
4	CTSH (12.91)	PRKCQ (26.47)	CTSH (18.16)	TIMP1 (211.25)	TIMP1 (569.81)
5	ITGA10 (20.81)	PIM1 (33.81)	ANXA5 (78.78)	EPHA1 (320.22)	CTSH (631.5)

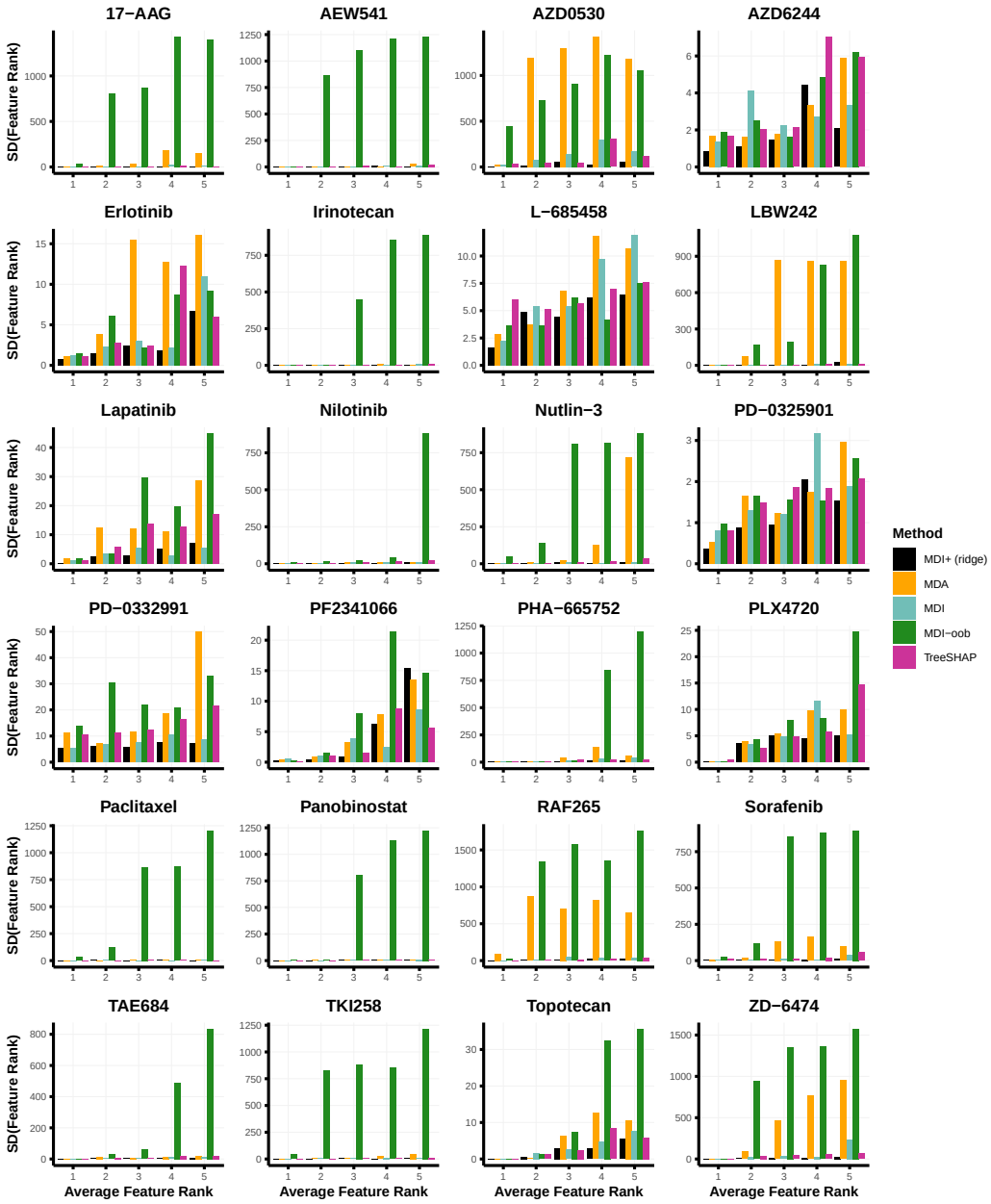


Figure B.28: Stability of top 5 genes for each drug response prediction model across 32 train-test splits. The x-axis corresponds to the top 5 features for each method, ranked according to their average feature ranking across 32 train-test splits. On the y-axis, we provide one measure of stability – namely, the standard deviation of the feature rankings across the 32 train-test splits. MDI+ generally provides the most stable feature importance rankings for these top 5 genes. Results from all feature importance methods under study are shown here.

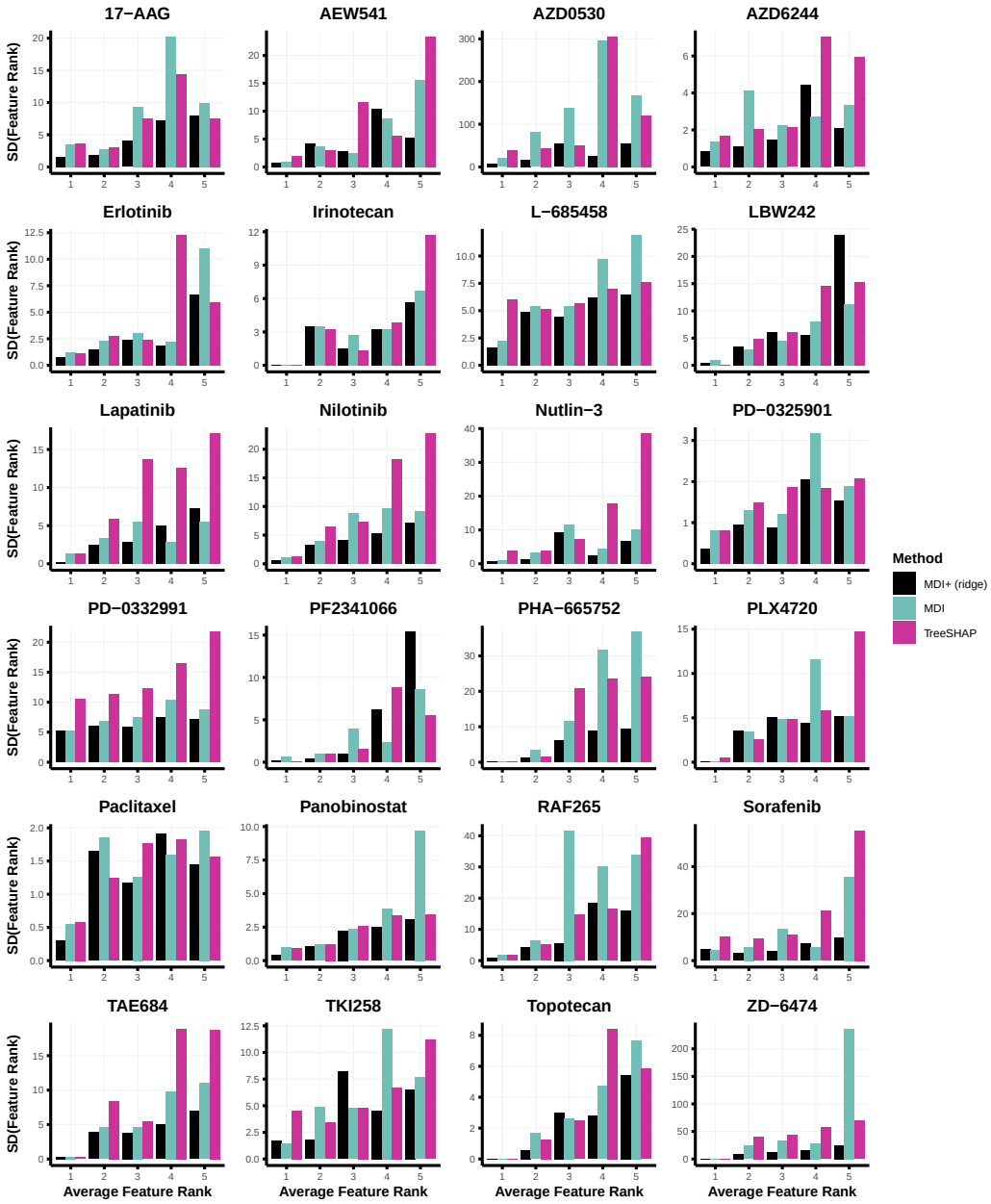


Figure B.29: Stability of top 5 genes for each drug response prediction model across 32 train-test splits. The x-axis corresponds to the top 5 features for each method, ranked according to their average feature ranking across 32 train-test splits. On the y-axis, we provide one measure of stability – namely, the standard deviation of the feature rankings across the 32 train-test splits. MDI+ generally provides the most stable feature importance rankings for these top 5 genes. Results from the feature importance methods under study, excluding MDI-oob and MDA, are shown here.

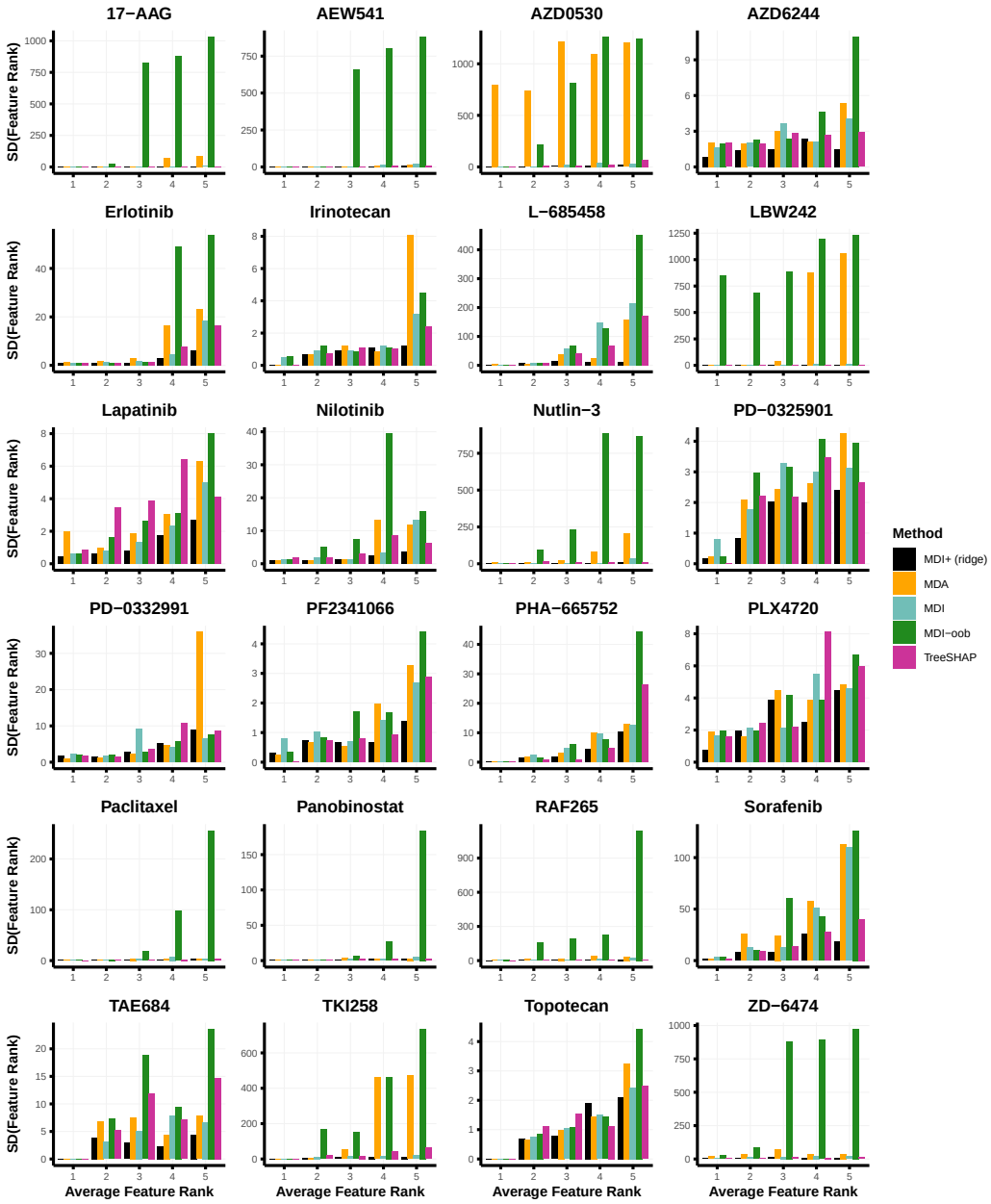


Figure B.30: Stability of top 5 genes for each drug response prediction model across 32 RF fits, trained using different random seeds. The x-axis corresponds to the top 5 features for each method, ranked according to their average feature ranking across the 32 RF fits. On the y-axis, we provide one measure of stability – namely, the standard deviation of the feature rankings across the 32 RF fits. MDI+ generally provides the most stable feature importance rankings for these top 5 genes. Results from all feature importance methods under study are shown here.

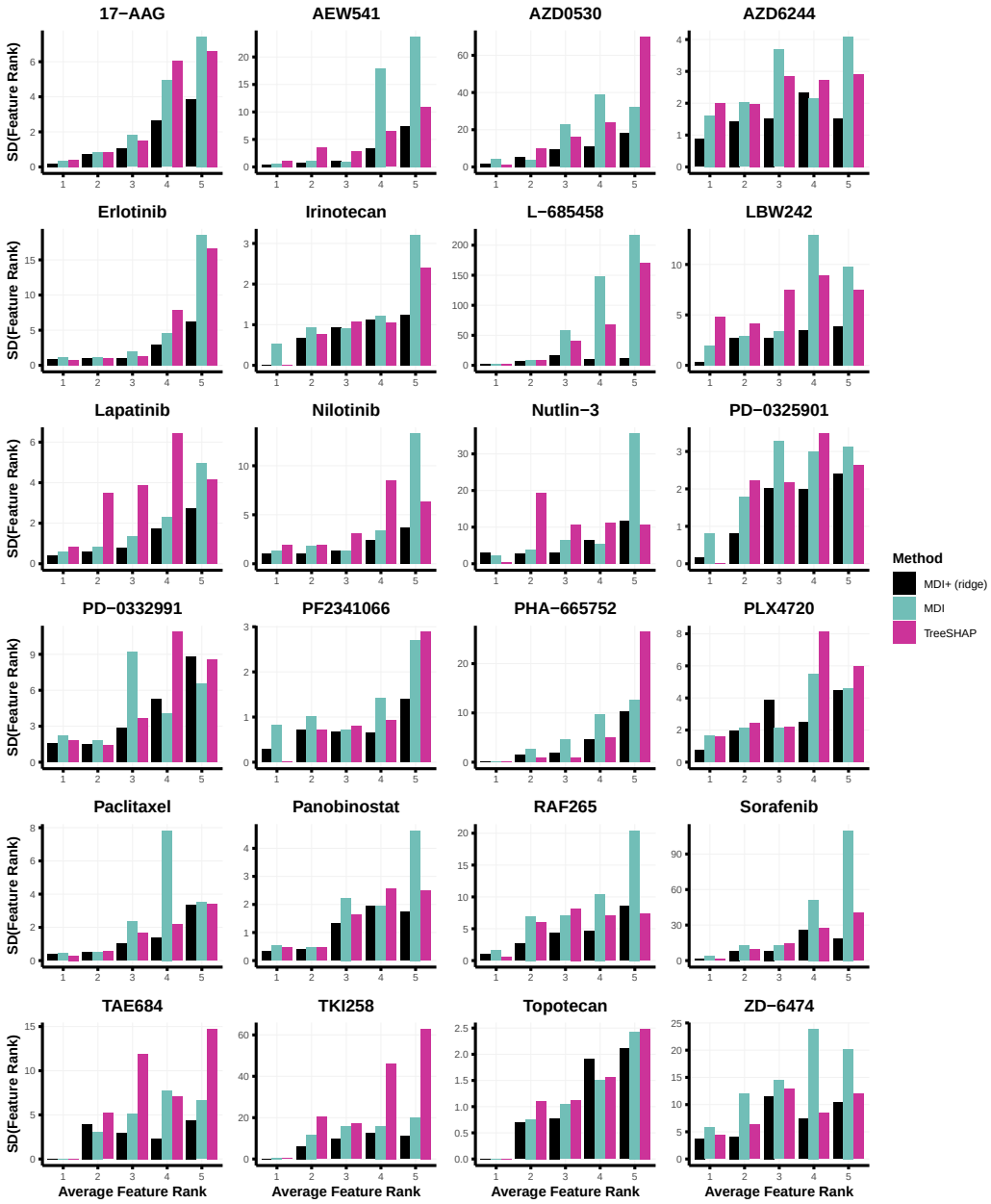


Figure B.31: Stability of top 5 genes for each drug response prediction model across 32 RF fits, trained using different random seeds. The x-axis corresponds to the top 5 features for each method, ranked according to their average feature ranking across the 32 RF fits. On the y-axis, we provide one measure of stability – namely, the standard deviation of the feature rankings across the 32 RF fits. MDI+ generally provides the most stable feature importance rankings for these top 5 genes. Results from the feature importance methods under study, excluding MDI-oob and MDA, are shown here.

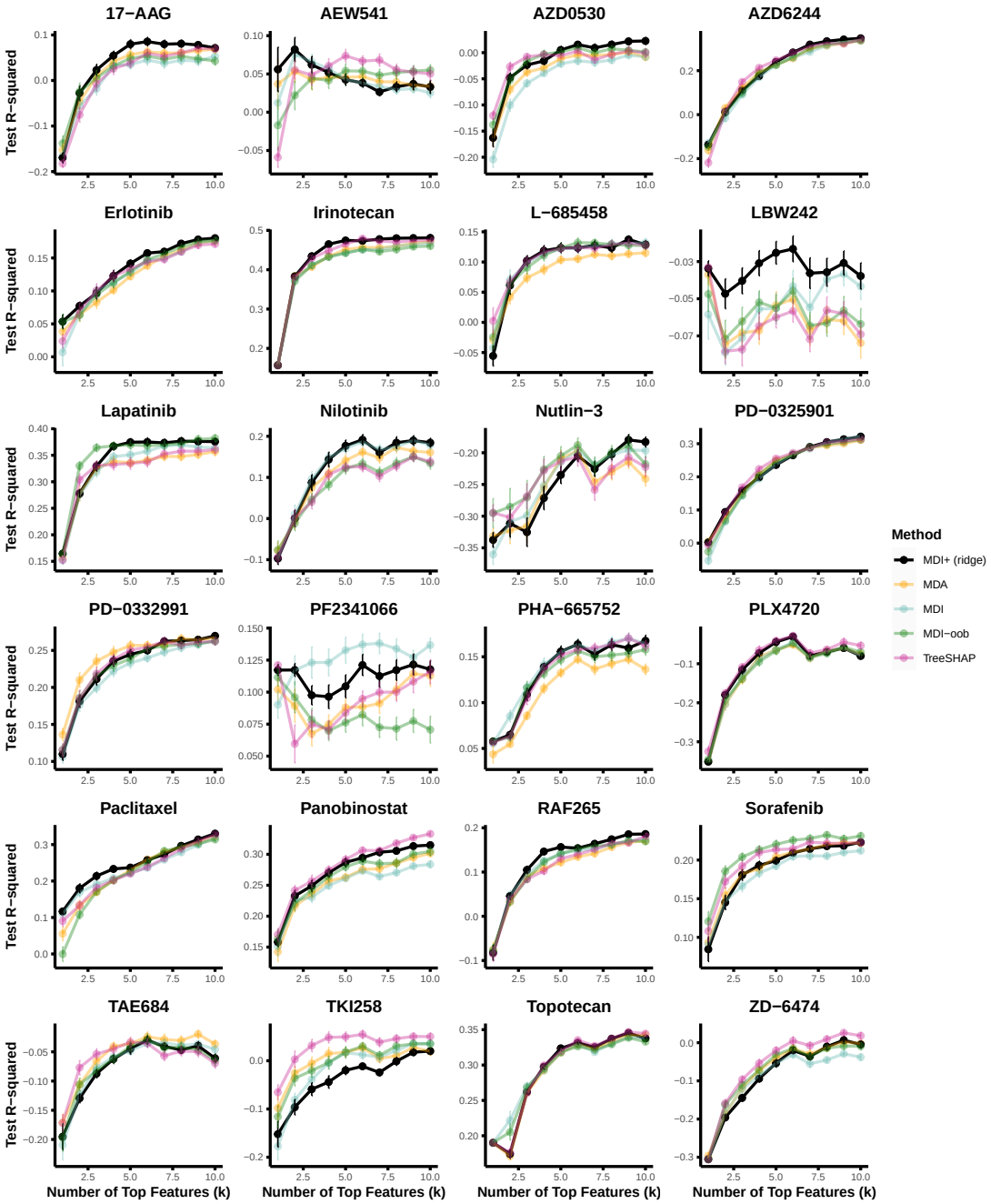


Figure B.32: RF prediction performance, measured via test R^2, using the top k features from each feature importance methods across the 24 drugs in the CCLE case study. Results are averaged across 32 train-test data splits.

Table B.2: Summary of prediction power using the top 10 features from each feature importance method in the CCLE drug response case study. For each drug, we evaluated the average test R^2 (averaged across 32 train-test splits) from an RF trained using only the top 10 features from each feature importance method. We then rank the feature importance methods by this average test R^2 (1 = best test R^2, 5 = worst test R^2) and display the number of drugs, for which that rank was achieved. In particular, taking the top 10 genes from MDI+ (ridge) gave the best prediction performance for 12 out of the 24 drugs.

Rank	**MDI+ (ridge)**	**TreeSHAP**	**MDI**	**MDI-oob**	**MDA**
1	12	5	3	3	1
2	6	6	6	3	3
3	2	6	6	4	6
4	3	5	5	5	6
5	1	2	4	9	8

Breast Cancer Subtype Prediction

Data preprocessing. We pulled data from the TCGA breast cancer project using the `TCGAbiolinks` R package. This gene expression dataset originally consisted of 19,947 genes. As before, we reduced the number of features under consideration by taking only the top 5000 features with the highest variance.

Results. In Table B.3, we summarize the test prediction performance for RF, RF+ (ridge), and RF+ (logistic), applied to the TCGA case study. We show the average test classification accuracy, AUROC, and area under the precision recall curve (AUPRC), averaged across 32 train-test splits. In Table B.4, we list the top 25 gene expression predictors according to each feature importance method. These genes are ranked according to their average feature importance ranking across the 32 train-test splits. We also show the RF predictive power of the top k genes from each feature importance method in Figure B.32. Details and discussion on this procedure were provided in the previous section. As discussed, correlated features can make it challenging to directly compare the prediction accuracy using the top k genes from various feature importance methods. However, it is worth noting that the top 15 features from MDI+ (ridge) and MDI+ (logistic) yield a higher test prediction performance (in terms of both AUROC and classification accuracy) as the top 25 features from the other competing feature importance methods. This improvement in predictive power further supports the practical utility of MDI+ for feature ranking (Yu and Kumbier, 2020).

Table B.3: Test prediction performance for various methods, averaged across 32 train-test splits, on the TCGA case study. Standard errors are shown in parentheses.

Model	Classification Accuracy	AUROC	AUPRC
RF+ (logistic)	**0.884 (0.003)**	**0.981 (0.001)**	**0.895 (0.004)**
RF+ (ridge)	0.873 (0.003)	0.981 (0.001)	0.892 (0.004)
RF	0.861 (0.003)	0.978 (0.001)	0.878 (0.005)

Table B.4: Top 25 most important genes for predicting breast cancer subtype according to various feature importance methods. Genes are ranked by their average feature importance ranking across 32 train-test splits (shown in parentheses).

Rank	MDI+ (ridge)	MDI+ (logistic)	MDA	TreeSHAP	MDI
1	ESR1 (1.91)	ESR1 (1.91)	ESR1 (4.5)	ESR1 (7.62)	ESR1 (13.91)
2	FOXA1 (4.25)	GATA3 (4.5)	GATA3 (6.38)	TPX2 (10.41)	TPX2 (15.34)
3	FOXC1 (6.12)	FOXA1 (5.09)	FOXA1 (8.11)	GATA3 (19.62)	FOXM1 (22.84)
4	GATA3 (6.97)	TPX2 (6.81)	TPX2 (10.12)	FOXM1 (20.06)	MLPH (24.97)
5	AGR3 (7.94)	AGR3 (10.22)	MLPH (10.16)	FOXA1 (20.72)	FOXA1 (25.66)
6	MLPH (8.16)	FOXC1 (12.94)	AGR3 (12.94)	CDK1 (22.38)	GATA3 (30.44)
7	TPX2 (11.03)	MLPH (15.69)	TBC1D9 (14.22)	MLPH (22.53)	CDK1 (31.41)
8	TBC1D9 (14.44)	FOXM1 (18.12)	FOXC1 (15.09)	AGR3 (25.88)	THSD4 (34.69)
9	FOXM1 (18.66)	TBC1D9 (21.03)	FOXM1 (19.88)	PLK1 (28.47)	FOXC1 (35)
10	THSD4 (21.78)	THSD4 (23.66)	THSD4 (21.28)	TBC1D9 (29.84)	TBC1D9 (35.44)
11	SPDEF (25.81)	CDK1 (24.44)	CDK1 (21.77)	FOXC1 (30.44)	AGR3 (36.44)
12	CA12 (29.44)	MYBL2 (25.34)	XBP1 (24.88)	MYBL2 (33.06)	PLK1 (36.81)
13	CDK1 (36.09)	RACGAP1 (26.81)	KIF2C (27.95)	THSD4 (33.53)	MYBL2 (45.22)
14	GABRP (36.72)	ASPM (27.56)	PLK1 (28.58)	KIF2C (35.56)	KIF2C (46.22)
15	PLK1 (37.97)	PLK1 (28.84)	MYBL2 (30.97)	ASPM (40.59)	ASPM (49.72)
16	FAM171A1 (38.59)	UBE2C (30.41)	GABRP (31.94)	GMPS (41.41)	GMPS (52.03)
17	ASPM (39.5)	SPAG5 (31.81)	ASPM (35.97)	XBP1 (50.44)	SPDEF (58.88)
18	SFRP1 (40.25)	GMPS (35.34)	FAM171A1 (38.55)	CENPF (56.69)	FAM171A1 (68.56)
19	XBP1 (40.44)	KIF2C (36.03)	CA12 (40.16)	MKI67 (59.69)	CA12 (69.59)
20	MYBL2 (43.12)	CA12 (37.31)	CDC20 (48.45)	CA12 (59.91)	UBE2C (69.78)
21	TFF3 (43.12)	RRM2 (46.09)	SPDEF (53.05)	RACGAP1 (60.53)	TFF3 (70)
22	KIF2C (44.62)	CENPF (46.31)	UBE2C (54.27)	SPAG5 (61.94)	CENPF (70.31)
23	PRR15 (44.88)	GABRP (47.59)	ANXA9 (55.41)	FAM171A1 (63)	RACGAP1 (70.94)
24	AGR2 (45)	SFRP1 (49.34)	C1orf64 (57.22)	KIF11 (63.56)	SPAG5 (71.81)
25	MIA (45.31)	XBP1 (49.78)	RACGAP1 (57.97)	ANLN (64.28)	KIF11 (73.12)

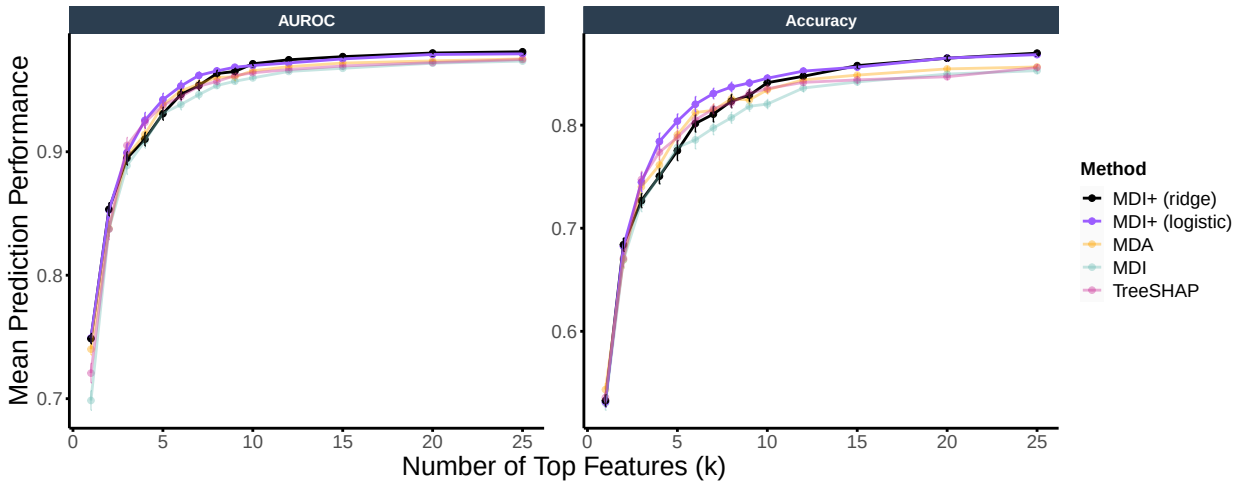


Figure B.33: RF prediction performance, measured via test AUROC and classification accuracy, using the top k features from each feature importance methods in the TCGA case study. Results are averaged across 32 train-test data splits.

Appendix C

A stability-driven protocol for drug response interpretable prediction (staDRIP)

C.1 Supplementary Figures

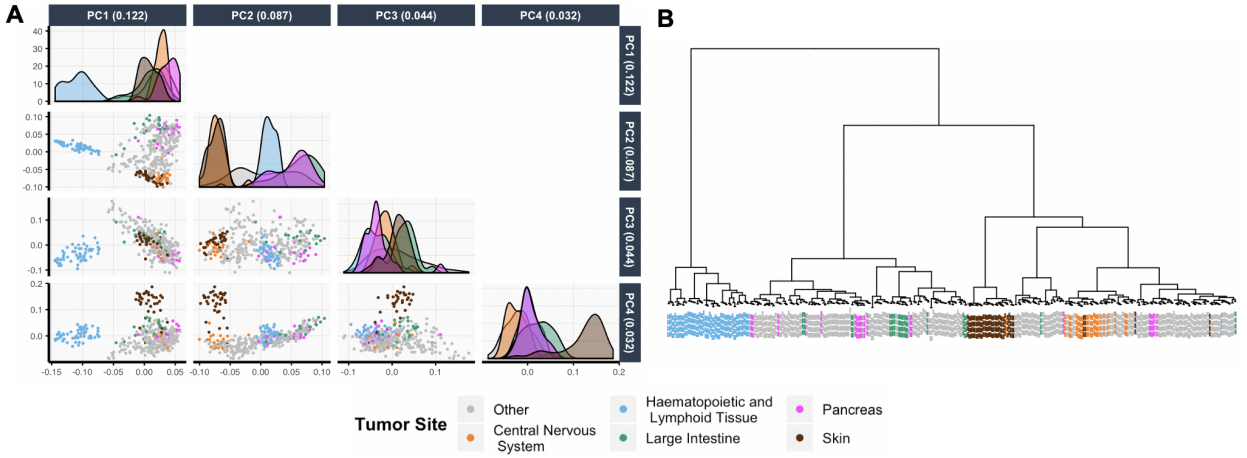


Figure C.1: We apply (A) PCA and (B) hierarchical clustering (with Ward's linkage) to the log-transformed RNASeq data set and color the samples by their tumor site. For simplicity, we use color to distinguish between five prominent tumor sites and show the remaining tumor sites in grey. We also show the proportion of variance explained by each principal component in the subplot titles of (A). In both the PC plots and the hierarchical clustering dendrogram, we can see clusters of tumor sites, illustrating the inherent differences between tumor sites.

C.2 Supplementary Tables

In Table C.1, for various methods, we summarize the validation accuracy across all 24 drugs as measured by the average R^2 value and WPC-index. In Tables C.2 and C.3, we provide additional insights into the drug response prediction accuracies at the individual drug level. In Table C.2, we see that the best model depends on the particular drug, but the kernel ridge regression model works best on average. In Table C.3, we show the test errors from the kernel ridge regression fit for each drug separately.

Table C.1: Validation WPC-index and average R^2 across all 24 drug response models for various methods trained on each molecular profile separately and together. Higher values of R^2 and WPC-index indicate better fits.

	Validation Set WPC-Index					**Validation Set R^2**				
	Methyl.	miRNA	Protein	RNASeq	Integrated	Methyl.	miRNA	Protein	RNASeq	Integrated
Kernel Ridge	0.600	0.603	0.617	**0.631**	0.624	0.111	0.104	0.168	**0.231**	0.200
Elastic Net	0.602	0.606	0.608	0.626	0.625	0.102	0.124	0.126	0.183	0.162
Lasso	0.597	0.605	0.609	0.620	0.620	0.117	0.105	0.121	0.172	0.176
Lasso (ESCV)	0.600	0.601	0.609	0.623	0.618	0.114	0.113	0.129	0.195	0.141
RF	0.599	0.594	0.606	0.626	0.622	0.124	0.088	0.123	0.214	0.196
X-VAE	–	–	–	–	0.617	–	–	–	–	0.188
BMTMKL	–	–	–	–	0.613	–	–	–	–	0.179

Table C.2: For each molecular profile (or the integrated profile) used for training, we count the number of drugs (out of 24) for which each method performed the best and gave the highest validation R^2 compared to its six other competitors.

	Methyl.	miRNA	Protein	RNASeq	Integrated
Kernel Ridge	7	5	16	12	6
RF	9	5	2	6	5
Elastic Net	1	8	1	1	2
Lasso (ESCV)	4	4	3	4	0
Lasso	3	2	2	1	5
X-VAE	–	–	–	–	2
BMTMKL	–	–	–	–	2

Table C.3: Test error for each drug using the RNASeq-based kernel ridge regression model

Drug	R^2	**PC-Index**
17-AAG	0.000	0.574
AEW541	0.034	0.558
AZD0530	0.037	0.560
AZD6244	0.425	0.675
Erlotinib	0.254	0.615
Irinotecan	0.307	0.644
L-685458	0.210	0.624
LBW242	-0.001	0.511
Lapatinib	0.208	0.607
Nilotinib	0.258	0.590
Nutlin-3	0.022	0.549
PD-0325901	0.543	0.701
PD-0332991	0.218	0.596
PF2341066	0.091	0.564
PHA-665752	0.115	0.559
PLX4720	0.305	0.585
Paclitaxel	0.369	0.670
Panobinostat	0.446	0.679
RAF265	0.215	0.625
Sorafenib	0.242	0.567
TAE684	0.024	0.576
TKI258	0.183	0.585
Topotecan	0.240	0.630
ZD-6474	0.155	0.591

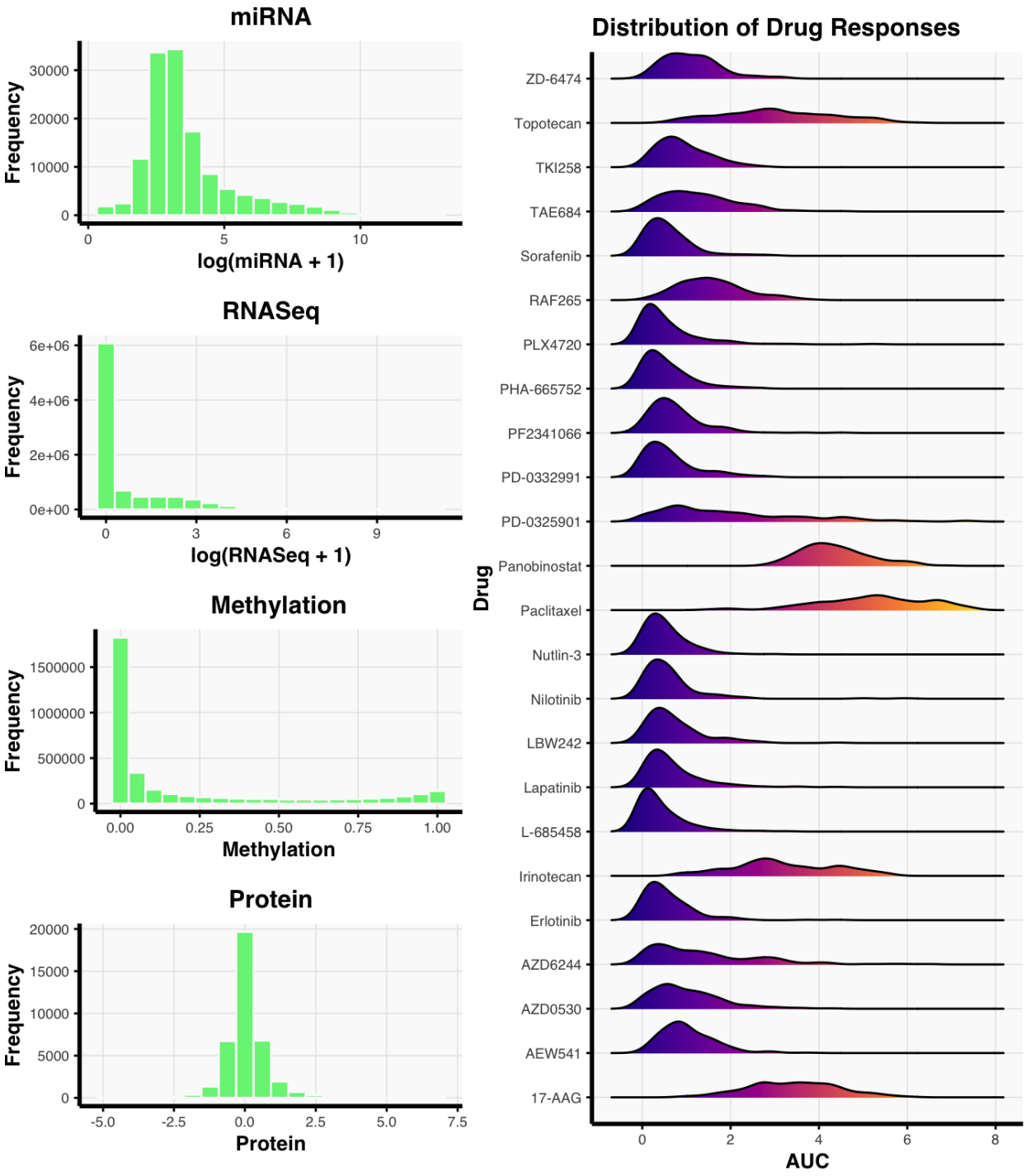


Figure C.2: Left: Distribution of features in each of the four molecular profiles. Right: Distribution of the drug responses for each of the 24 drugs.

Table C.4: Stable protein and RNAseq signatures. A feature is included if it is among the top 10 most stable features under 3 different machine learning models (i.e., elastic net, Lasso (ESCV), and random forests). The stability of the features are computed from the PCS inference framework in staDRIP. Blank cells indicate that no features appeared among the top 10 most stable features for all three models.

Drug name	**Protein signature**	**RNAseq Signature**
17-AAG	Bax, p53, Caspase-7, eIF4E	CTD, AP2S1, BZW2
AEW541	Akt, Smad1, p27, PTEN, RAD51	B4GALT3, SEMA3B
AZD0530	p38, c-Kit	HPGD
AZD6244	PI3K-p85, TFRC, Bax	SPRY2, RP11, LYZ, DUSP6, PRSS57
Erlotinib	EGFR, Beclin, P-Cadherin	PIP4K2C, SEC61G
Irinotecan	MDMX_MDM4, Src	
L-685458	YAP, VEGFR2, Src	
LBW242	ASNS	MRPL24
Lapatinib	HER2, HER3, EGFR, Rab25, Heregulin	STARD3
Nilotinib	STAT5, c-Kit, SHP-2, Src, p27	
Nutlin.3	Bcl-2, Bax	
PD-0325901	MEK1, Bax, TFRC, PI3k	SPRY2, DUSP6, ETV4
PD-0332991	Bcl-2, MDMX_MDM4, Src	
PF2341066	c-Met	CAPZA2
PHA-665752	MEK1, c-Met	FMNL1
PLX4720	MEK1, Bax, PREX1, Beclin	FABP7
Paclitaxel	Src, beta-Catenin	ORMDL2
Panobinostat	VEGFR2, Src	
RAF265	PI3K-p85, FOXO3a, eEF2K	RETN
Sorafenib	Bcl-2, Src	
TAE684	PTEN, Akt, p70S6K, Bcl-2	H1FX
TKI258	CD49b, C-Raf	
Topotecan	c-Met	OSGIN1
ZD.6474	c-Kit, STAT5-alpha	

Table C.5: Most stable protein associated with each drug, as identified by the elastic net, along with preclinical evidence that supports the association between the listed protein and drug sensitivity.

Drug	**Protein**	**Supporting Literature**	**Drug**	**Protein**	**Supporting Literature**
17-AAG	p53	Naito et al. (2010)	PD-0332991	Bcl-xL	Chen and Pan (2017)
AEW541	Akt	Attias-Geva et al. (2011)	PF2341066	PEA15	–
AZD0530	p38	Yang et al. (2010)	PHA-665752	MEK1	–
AZD6244	CD20	–	PLX4720	MEK1	Emery et al. (2009)
Erlotinib	P-Cadherin	–	Paclitaxel	Src	Le and Bast (2011)
Irinotecan	RAD51	Shao et al. (2016)	Panobinostat	Src	–
L-685458	VEGFR2	–	RAF265	PI3K-p85	Mordant et al. (2010)
LBW242	Caspase-7	–	Sorafenib	14-3-3 epsilon	Wu et al. (2015)
Lapatinib	HER2	Esteva et al. (2010)	TAE684	Akt	–
Nilotinib	p27	Liu et al. (2011)	TKI258	14-3-3 epsilon	–
Nutlin.3	Bcl-2	Drakos et al. (2011)	Topotecan	14-3-3 epsilon	–
PD-0325901	MEK1	Henderson et al. (2010)	ZD-6474	c-Kit	Nishioka et al. (2007)

Appendix D

Contribution of the microbiome to a metabolomic signature predictive of risk for pancreatic cancer

D.1 Supplementary Figures

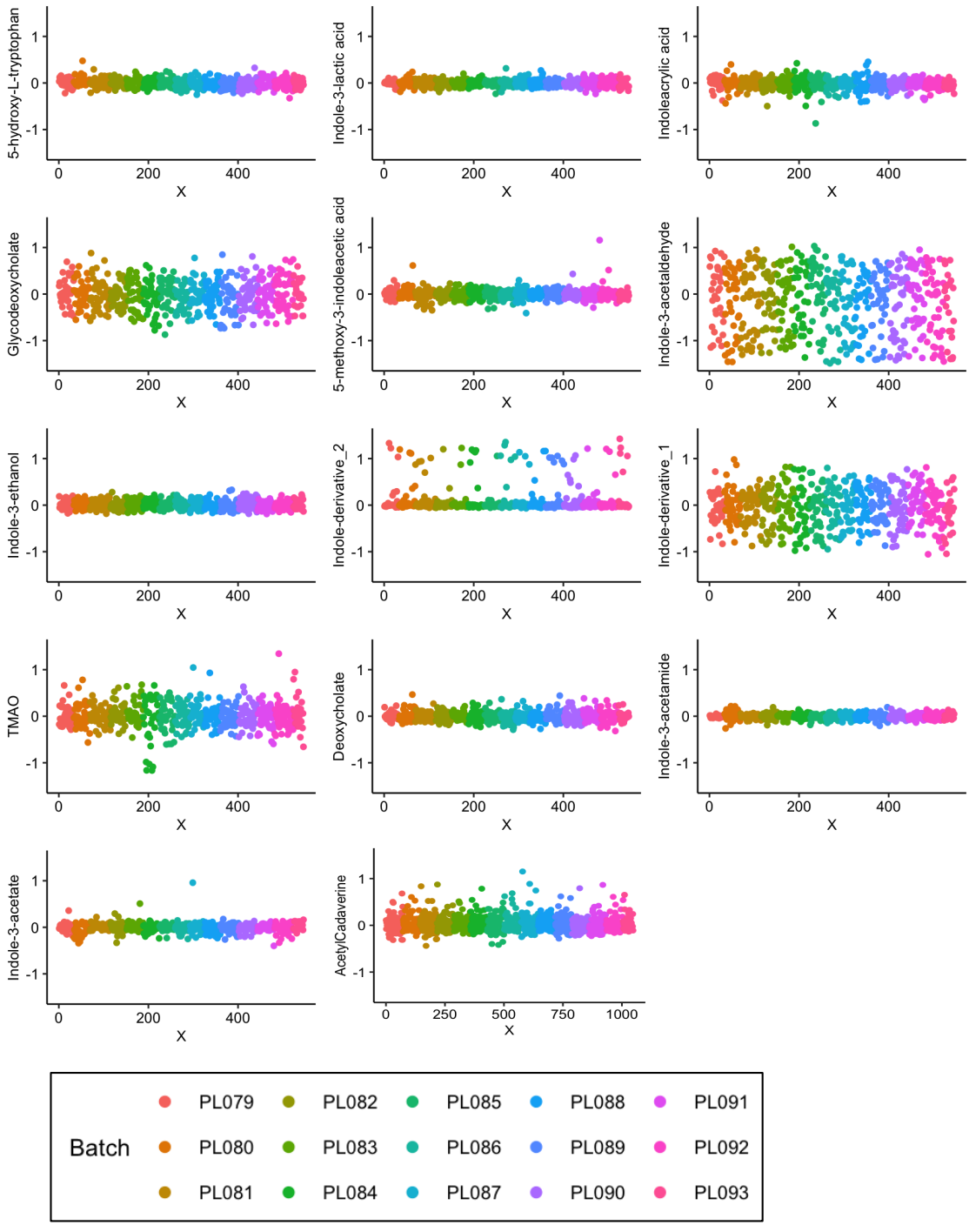


Figure D.1: Distribution plots for detected microbial-related metabolites across analytical batches in the PLCO specimen set. X-axis represents individual specimens.

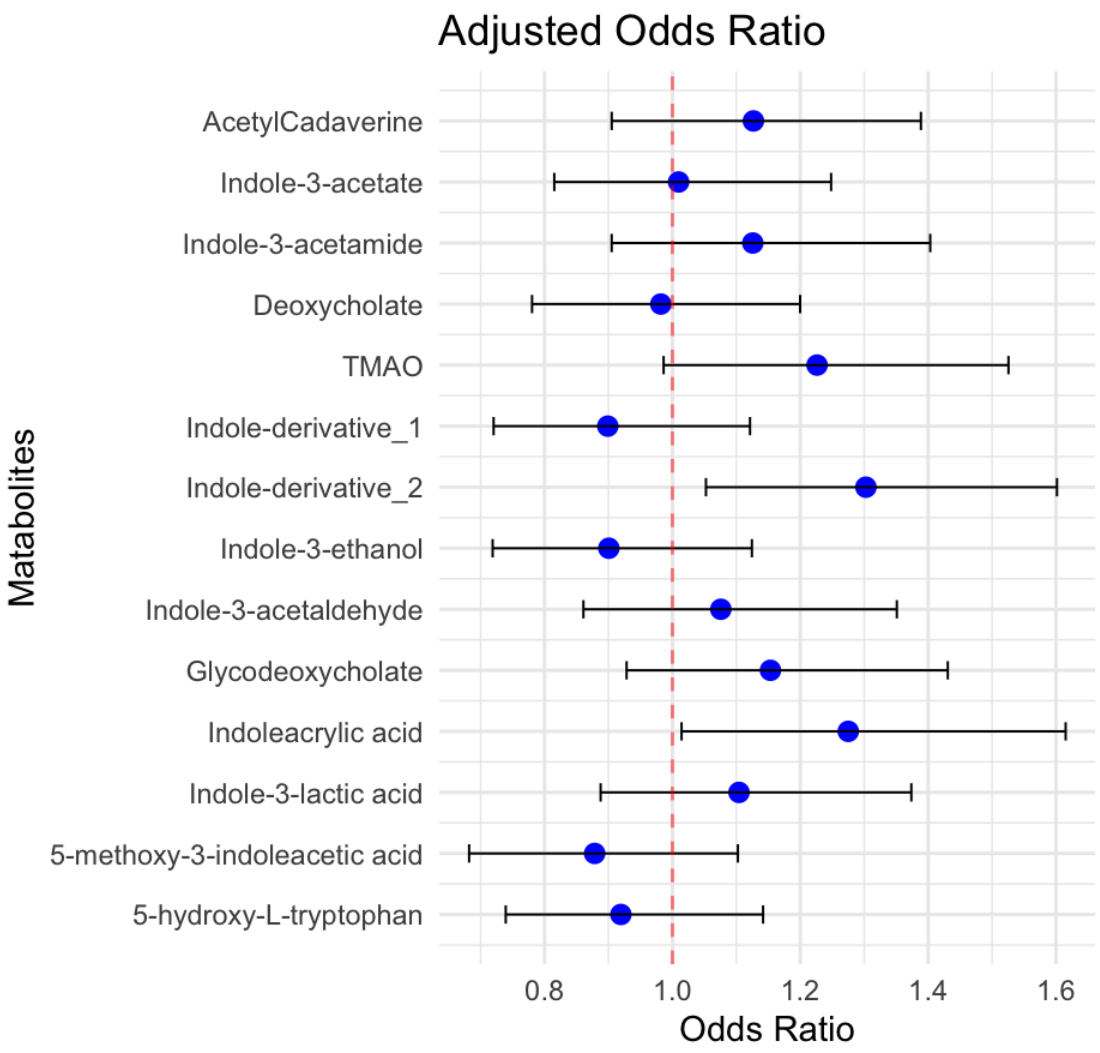


Figure D.2: Odds ratios and adjusted odds ratios for individual microbial-related metabolites for risk of pancreatic cancer in the Training Set. Sex, age, smoking status, and BMI were included as covariates in adjusted odds ratios.

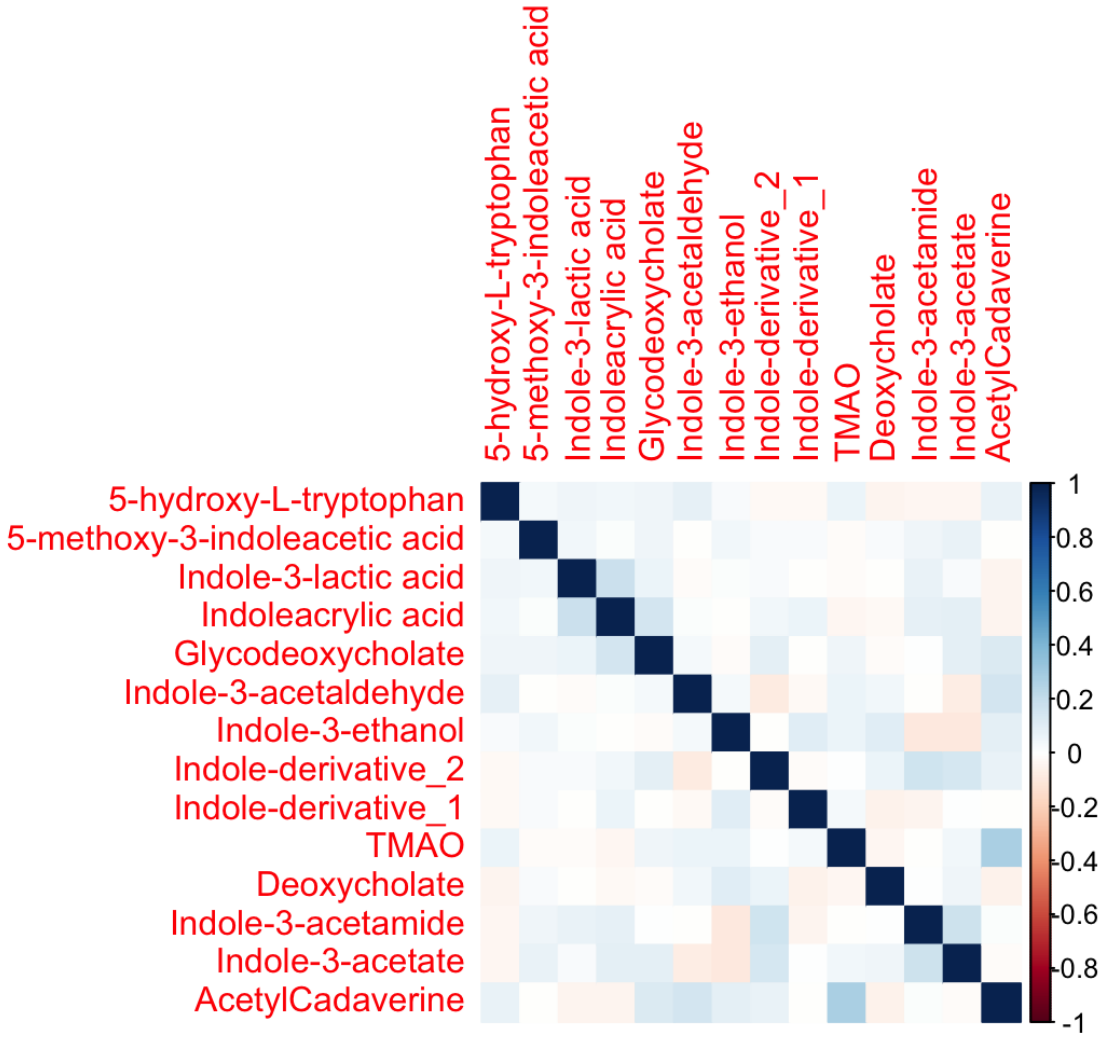


Figure D.3: Spearman correlation heatmap for microbiome-related metabolites in the Training Set.

D.2 Supplementary Tables

Table D.1: Patient and tumor characteristics for the newly-diagnosed PDAC cohort. DF/BWCC: Dana-Farber/Brigham and Women's Cancer Center; BIDMC: Beth Israel Deaconess Medical Center; CUMC: Columbia University Medical Center. AJCC: American Joint Committee on Cancer, PDAC: Pancreatic ductal adenocarcinoma, BMI: Body mass index. [a]Patients who underwent up-front surgical resection. [b]Patients who received neoadjuvant treatment and then underwent surgical resection. [c]The median (IQR) follow-up time was 15.0 (7.2-23.2) months for patients without cancer recurrence.

Variable	**PDAC Case (N=99)**		**Chronic Pancreatitis (N=50)**		**Healthy Control (N=100)**	
	No.	**%**	**No.**	**%**	**No.**	**%**
Institution						
DF/BWCC	69	70%	30	60%	94	94%
BIDMC	15	15%	15	30%	0	0%
CUMC	15	15%	5	10%	6	6%
Age (year), median (IQR)	69.8 (62.5-74.8)		65.4 (54.7-72.2)		63.7 (55.7-70.6)	
Gender						
Male	51	52%	33	66%	51	51%
Female	48	48%	17	34%	49	49%
Race						
White	94	95%	42	84%	84	86%
Black/African-American	0	0%	5	10%	5	5%
Asian	1	1%	0	0%	2	2%
Other	4	4%	3	6%	7	7%
Blood collection year						
2015-2016	19	19%	2	4%	0	0%
2017-2019	80	81%	48	96%	100	100%
Smoking Status						
Current Smoker	6	6%	11	22%	4	4%
Past smoker	50	51%	17	34%	42	42%
Never smoker	43	43%	22	44%	54	54%
BMI (kg/m^2), Meidan (IQR)	27.4 (24.0-30.0)		25.0 (22.8-27.6)		27.5 (24.3-32.0)	
Diabetes						
No	64	65%	23	46%	93	93%
Yes	35	35%	27	54%	7	7%
Etiology of chronic pancreatitis						
Alcohol	-	-	16	32%	-	-
Autoimmune	-	-	2	4%	-	-
Congenital anatomical variant	-	-	3	6%	-	-
Duct stricture or stones	-	-	7	14%	-	-
Idiopathic	-	-	21	42%	-	-
Other	-	-	1	2%	-	-

Table D.1: *(continued)*

Variable	PDAC Case (N=99)		Chronic Pancreatitis (N=50)		Healthy Control (N=100)	
	No.	%	No.	%	No.	%
AJCC 8th edition staging pTNM[a]						
T0-2N0M0	15	24%	-	-	-	-
T3-4N0M0	2	3%	-	-	-	-
T1-4N1M0	28	45%	-	-	-	-
T1-4N2M0	17	28%	-	-	-	-
AJCC 8th edition staging ypTNM[b]						
T0-2N0M0	24	64%	-	-	-	-
T3-4N0M0	1	3%	-	-	-	-
T1-4N1M0	7	19%	-	-	-	-
T1-4N2M0	5	14%	-	-	-	-
PDAC recurrence						
No[c]	56	57%	-	-	-	-
Yes	43	43%	-	-	-	-

Table D.2: Selected microbial-associated metabolites and corresponding model coefficients in LASSO regression.

	Development Set					
	Training Set		Validation Set		Set-Aside Test Set	
	Non-cases	Cases	Non-cases	Cases	Non-cases	Cases
Total	494	102	142	33	225	37
Gender, N (%)						
Female	204 (41)	41 (40)	61 (43)	17 (52)	91 (40)	13 (35)
Male	290 (59)	61 (60)	81 (57)	16 (48)	134 (60)	24 (65)
Age At Randomization, N (%)						
<= 59	116 (23)	21 (21)	22 (15)	11 (33)	45 (20)	5 (14)
60-64	108 (22)	24 (24)	34 (24)	3 (9)	63 (28)	14 (38)
65-69	192 (39)	41 (40)	50 (35)	9 (27)	79 (35)	13 (35)
>= 70	78 (16)	16 (16)	36 (25)	10 (30)	38 (17)	5 (14)
Race, N (%)						
White	463 (94)	99 (97)	107 (75)	24 (73)	211 (94)	33 (89)
Black	22 (4)	3 (3)	2 (1)	1 (3)	6 (3)	2 (5)
Other	9 (2)	0 (0)	33 (23)	8 (24)	8 (4)	2 (5)

Table D.3: Stability check of the LASSO regression using perturbed training data and evaluated on the Validation Set for the 3-marker microbial panel. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs).

	Perturbations	**AUC (95% CI)**	**Adj OR†**
Lasso regression with 3 selected features	2 randomly selected centers	0.63 (0.44-0.82)	1.37 (0.89-2.09)
	2 randomly selected centers	0.73 (0.60-0.86)	2.33 (1.52-3.77)
	2 randomly selected centers	0.54 (0.41-0.68)	1.25 (0.90-1.73)
	2 randomly selected centers	0.55 (0.45-0.63)	1.27 (0.92-1.72)
	3 randomly selected centers	0.64 (0.54-0.73)	1.65 (1.23-2.24)
	300 random samples	0.60 (0.51-0.68)	1.40 (1.02-1.90)

Table D.4: Selected non-microbial metabolites. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. ‡Benjamini and Hochberg-adjusted p-values. £Raw p-values.

	Training - 5 centers		**Validation- 2 centers**	
Name	**Adj. Odds Ratio†**	**P-value‡ (FDR)**	**Adj. Odds Ratio†**	**P-value£**
Cholesterol glucuronide	1.735	<0.001	1.72	0.006
Galactosamine	1.749	<0.001	1.514	0.035
2-Hydroxyglutarate	1.857	<0.001	1.738	0.006
Erythritol	1.688	<0.001	1.532	0.03
Glucose	1.744	<0.001	1.662	0.018

Table D.5: Performance of the 3-marker microbial panel amongst diabetic and non-diabetic individuals in the PLCO set-aside Test Set. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. N0: Number of non-cases, N1: Number of cases.

	Diabetics				**Non-Diabetic**			
3MMP	**Sample Size**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**
PLCO Testing Set	N0 = 14 N1 = 4	0.62 (0.22-1.00)	0.8 (0.09-3.61)	0.77	N0 = 210 N1 = 33	0.64 (0.53-0.77)	1.84 (1.32-2.61)	<0.001
All PLCO samples	N0 = 55 N1 = 22	0.6 (0.46-0.74)	1.56 (0.88-2.95)	0.13	N0 = 805 N1 = 150	0.62 (0.57-0.67)	1.5 (1.27-1.77)	<0.001

Table D.6: Performance of different learning models based on non-microbial metabolites in the PLCO Validation Set. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase.

Model	**Hyperparameters**	**AUC (95% CI)**	**Adj OR† (95% CI)**
Logistic regression	-	0.72 (0.63-0.81)	2.10 (1.04-2.90)
Logistic regression with ridge (L_2) regularization	Penalty weight = 0.18	0.69 (0.58-0.78)	1.74 (1.20-2.25)
Logistic regression with LASSO (L_1) regularization	Penalty weight = 0.01, number of selected features = 4	0.71 (0.54-0.73)	2.08 (0.94-2.83)
Iterative Random Forest	Number of iterations = 3	0.60 (0.49-0.72)	1.44 (0.90-1.90)
Deep neural network model	Number of cross-validation folds = 6, hidden layers = 3 with 32 nodes in each layer	0.59 (0.48-0.68)	1.43 (0.95-2.10)
GBM	Number of trees = 42, max depth= 5	0.58 (0.46-0.67)	1.30 (0.93-1.87)
Auto ML	Selected model = randomized trees	0.66 (0.52-0.72)	1.85 (1.50-2.02)

Table D.7: Stability check of the 5-marker non-microbial panel using perturbed training data and evaluated on the PLCO Validation Set. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase.

	Perturbations	**AUC (95% CI)**	**Adj OR†**
Logistic regression with 5 selected features	2 randomly selected centers	0.71 (0.52-0.87)	2.10 (1.10-2.94)
	2 randomly selected centers	0.74 (0.61-0.91)	2.33 (1.42-4.10)
	2 randomly selected centers	0.69 (0.59-0.80)	2.11 (0.90-2.73)
	2 randomly selected centers	0.67 (0.45-0.85)	1.90 (1.12-2.72)
	3 randomly selected centers	0.60 (0.52-0.68)	1.65 (1.23-2.24)
	300 random samples	0.64 (0.55-0.71)	1.73 (1.52-2.20)

Table D.8: Performance of the 5-marker non-microbial panel in the PLCO set-aside Test Set and the entire specimen set. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. N0: number of non-cases. N1: number of cases. [a]Non-microbial-related metabolite signature includes cholesterol glucuronide, hydroxyglutarate, galactosamine, glucose, and erythritol.

Set-aside Test Set				
		5-marker non-microbial panel[a]		
Time to Dx	Sample Size	AUC (95% CI)	Adj. OR† (95% CI)	*P-value*
[0-5)	N0 = 225 N1 = 37	0.74 (0.65 - 0.83)	2.72 (1.83 - 4.24)	<0.001
[0-2)	N0 = 225 N1 = 24	0.82 (0.72 - 0.92)	4.03 (2.41 - 7.32)	<0.001
[2-5)	N0 = 225 N1 = 13	0.59 (0.44 - 0.72)	1.32 (0.71 - 2.41)	0.36
Entire Set (Development + Set-aside Test Set)				
		5-marker non-microbial panel[a]		
Time to Dx	Sample Size	AUC (95% CI)	Adj. OR† (95% CI)	*P-value*
[0-5)	N0 = 861 N1 = 172	0.74 (0.67 - 0.77)	2.59 (2.13 - 3.18)	<0.001
[0-2)	N0 = 861 N1 = 92	0.8 (0.75 - 0.85)	3.69 (2.83 - 4.91)	<0.001
[2-5)	N0 = 861 N1 = 80	0.65 (0.59 - 0.72)	1.74 (1.37 – 2.21)	<0.001

Table D.9: **Performance of the 5-marker non-microbial panel and alone and in combination with the 3-marker microbial panel stratified by diabetic status.** †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. N0: number of non-cases N1: number of cases.

	Diabetics				Non-Diabetics			
5-marker non-microbial panel	Sample Size	AUC (95% CI)	Adj. OR† (95% CI)	*P-value*	Sample Size	AUC (95% CI)	Adj. OR† (95% CI)	*P-value*
PLCO Testing Set	N0 = 14 N1 = 4	0.65 (0.27-1.00)	1.93 (0.45-17.61)	0.43	N0 = 210 N1 = 33	0.75 (0.65-0.84)	2.74 (1.83-4.32)	<0.001
All PLCO samples	N0 = 55 N1 = 22	0.67 (0.52-0.82)	2.67 (1.44-5.72)	0.004	N0 = 805 N1 = 150	0.74 (0.70-0.78)	2.95 (2.12-3.20)	<0.001

	Diabetics				Non-Diabetics			
5-marker non-microbial panel + 3-marker microbial panel	Sample Size	AUC (95% CI)	Adj. OR† (95% CI)	*P-value*	Sample Size	AUC (95% CI)	Adj. OR† (95% CI)	*P-value*
PLCO Testing Set	N0 = 14 N1 = 4	0.65 (0.29-1.00)	1.7 (0.38-13.34)	0.52	N0 = 210 N1 = 33	0.81 (0.72-0.89)	3.39 (2.19-5.61)	<0.001
All PLCO samples	N0 = 55 N1 = 22	0.67 (0.53-0.81)	2.71 (1.44-5.84)	0.004	N0 = 805 N1 = 150	0.76 (0.72-0.80)	2.79 (2.27-3.46)	<0.001

Table D.10: Performance of the combined metabolite panel plus CA19-9 stratified by diabetic status. †Age, sex, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. N0: number of non-cases. N1: number of cases.

	Diabetics				**Non-Diabetic**			
5-marker non-microbial panel + 3-marker microbial panel + CA19.9	**Sample Size**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR† (95% CI)**	**P-value**
PLCO Testing Set	N0 = 14 N1 = 4	0.78 (0.50-1.00)	6.82 (1.14-210.61)	0.1	N0 = 210 N1 = 33	0.84 (0.76-0.92)	10.21 (4.55-26.61)	<0.001
All PLCO samples	N0 = 55 N1 = 22	0.71 (0.60-0.84)	3.75 (1.81-9.72)	0.001	N0 = 805 N1 = 150	0.8 (0.76-0.84)	9.54 (6.36-14.75)	<0.001