METABOLIC PHENOTYPING MEETS MACHINE LEARNING: DETECTING RENAL CELL CARCINOMA IN URINE

by

OLATOMIWA O. BIFARIN

(Under the Direction of ARTHUR EDISON)

ABSTRACT

Renal cell carcinoma (RCC) is one of the deadliest urogenital cancers today. Detecting and staging renal cell carcinoma involves expensive imaging tests and biopsy, which is invasive and can be riddled with sampling errors. Alternative, non-invasive, cost-effective diagnostic methods will significantly reduce the burden of RCC in the world. Given that metabolic rewiring is required for the onset and progression of RCC, and the proximity of urine with the kidney, I set out to discover a urinary metabolic biomarker for RCC using advances in machine learning. Metabolomics is the study of small molecules in biological samples – and as the apogee of the omics trilogy, it is the closest to an organism phenotype. In this dissertation, liquid chromatography-mass spectrometry (LC-MS) and nuclear magnetic spectroscopy (NMR) were used for untargeted metabolite profiling for broad analyte coverage. I employed the use of machine learning to mine the metabolomics data generated. Machine learning (ML) is a type of artificial intelligence that entails computational techniques for learning patterns in a complex dataset. I conducted three categories of ML tasks in the dissertation – binary classification, regression, and ML model interpretations. All data modalities are tabular. Using untargeted metabolomics and ML, I

utilized human urine samples to discriminate between healthy controls and RCC to identify biomarkers that can be used for RCC detection. In addition, I predicted RCC primary tumor sizes using selected urinary metabolites, as well as the discrimination of early-stage RCC from advanced-stage RCC. Furthermore, I introduced a state-of-the-art interpretable machine learning (IML) technique called Shapley Additive Explanations (SHAP). SHAP was used to interpret ML models' predictions for publicly available clinical metabolomics dataset – and also to interpret a ML model prediction for RCC detection. These studies led to the accurate detection and staging of RCC in the study cohort and the identification of some novel metabolic markers. In addition, the ML methods presented in the thesis can be used to advance biomarker discoveries in other omics fields.

INDEX WORDS: renal cell carcinoma, metabolomics, machine learning, liquid chromatography mass spectrometry, nuclear magnetic resonance spectroscopy, interpretability, IML

METABOLIC PHENOTYPING MEETS MACHINE LEARNING: DETECTING RENAL CELL CARCINOMA IN URINE

by

OLATOMIWA O. BIFARIN

BS, Obafemi Awolowo University, Nigeria, 2012

MS, Catholic University of America, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

METABOLIC PHENOTYPING MEETS MACHINE LEARNING: DETECTING RENAL CELL CARCINOMA IN URINE

by

OLATOMIWA O. BIFARIN

Major Professor:	Arthur S. Edison
Committee:	Natarajan Kannan
	Edward T. Kipreos
	Kelly W. Moremen

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

To my grandmother, Mary Babatola Bifarin

ACKNOWLEDGEMENTS

I decided to join Arthur Edison's lab in December 2015 because of a strong belief in the following maxim: "who you are working for (and with) is more important than what you are doing." Now, to use a machine learning lingua – which you will come across should you read this thesis – I would say the *decision function* was quite impressive. Looking around, I am not sure I would have had a better mentor. He is brilliant, very kind, and critical, all at the same time. To the entire Edison lab crew over the past six years, I would like to thank everyone for being a kind soul, notably Laura Morris, who solves your problem even before you say it – such a blessing. To my committee members, Dr. Natarajan Kannan, Dr. Edward T. Kipreos, Dr. Kelly W. Moremen, for the much-needed guidance, I say thank you. I would like to thank my collaborators, Petros Lab at Emory University and Fernandez lab at Georgia Tech, for the advice and help in the course of the RCC study. To all my teachers over the years, Mr. Jimoh for that English writing lessons in high school, Prof. Anthonia O. Oluduro, and Prof. Bridget Omafuvbe for the motherly touch in College, Dr. John Choy, Dr. Ayça Akal-Strader, and Dr. Frank Portugal for the help in transiting to academia in the United States. I would like to thank my mum, dad, and my siblings for all the support during the program. I would also like to thank my sweet aunt, Dr. Ebun Ajibola – for everything.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

GENERAL INTRODUCTION

1.1 Renal Cell Carcinoma

Renal cell carcinoma (RCC) is a group of cancers that originate from renal tubular epithelial cells, and it makes up to 90% of kidney cancer diagnoses[1-2]. RCC is one of the top ten most common cancers in the world[1], with major histological subtypes including clear cell RCC (ccRCC), papillary RCC (pRCC), and chromophobe RCC (chRCC). ccRCC is the most prevalent RCC subtype, with about 75% in metastatic presentations [3]. The organ primarily affected, the kidney, are two bean-shaped organs about 4-5 inches long. The kidney's function is to maintain the homeostatic balance of solutes in the body, and it filters out waste products from the blood that – in turn – get converted into urine. Blood with waste comes into the kidney through the renal artery, filtered blood leaves *via* the renal vein, while urine leaves the kidney through the ureter into the bladder (**Figure 1-1**). A nephron is the functional unit of the kidney composing of the glomerulus, which filters the blood, and the renal tubules, which engage in the reabsorption and secretion of substances. The secreted substances include wastes and water, which is urine.

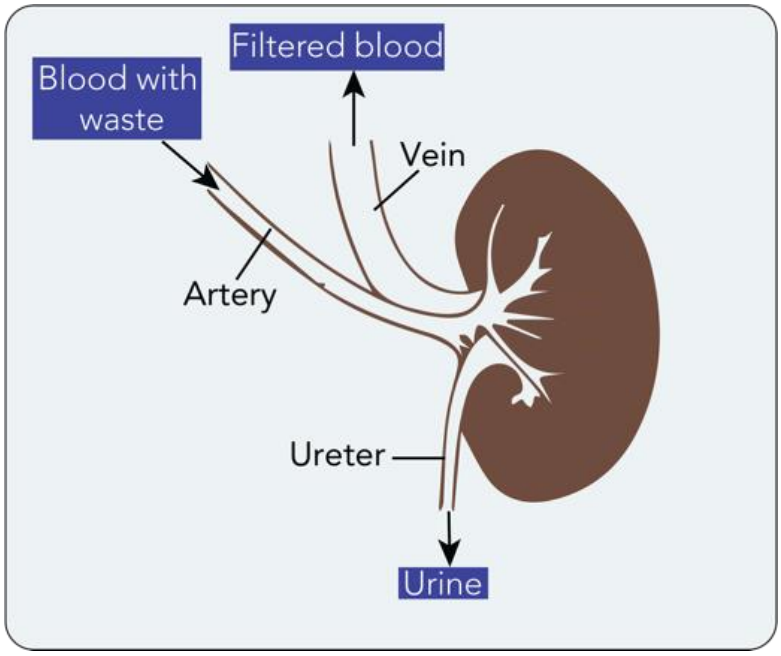


Figure 1-1. The Kidney: Showing the Movement of Biofluids In and Out of the Organ.

1.1.1 Epidemiology and Risk Factors

About 2% of all cancer diagnoses originate from the kidney[2]. Every year, ~300,000 kidney cancer cases are reported, with about 145,000 deaths. In the United States, 13,780 people are estimated to die from kidney cancer in 2021, with a projected 76,080 diagnoses[4]. According to the NCI Surveillance, Epidemiology, and End Results (SEER) database, kidney cancer occurrence has increased in the developed world over the past several decades, with incidence rates doubling in the United States since 1975[5]. In addition, a female is diagnosed for every two males diagnosis, and the lifetime risk of men and women developing kidney cancer is about 2% and 1%, respectively[5]. According to the American Cancer Society, the average age at the time of diagnosis is 64 years[6]. The chances of developing kidney cancer increase as we age, and diagnosis at a younger age (less than 45 years old) might indicate a hereditary kidney cancer syndrome[7]. Localized kidney cancer

has a 5-year relative survival rate of 93%, a 70% rate for regional RCC, and a 13% rate for distant RCC. Localized means no sign of tumor spread beyond the kidney, regional means that cancer has spread beyond the kidney to nearby structures, while distant implies that cancer has spread to distant organs of the body[8]. In aggregate, there is a 75% 5-year survival rate of RCC patients in the United States[8].

As indicated above, age and sex are two of the unmodifiable risks factor for RCC, with older people and men more likely to have the disease. In the United States, SEER age-adjusted incidence rates between 2000-2018 also indicate that RCC rates are lowest amongst Asian/Pacific Islanders and highest amongst American Indian/Alaska natives[5]. Apart from the unmodifiable risk factors, there are also modifiable risk factors for RCC, including smoking, obesity, diet, alcohol intake, and hypertension. As is the case with lung and bladder cancer, smoking is one of the most significant modifiable risk factors for RCC. Tobacco smoke consists of several compounds, many of which are carcinogenic, such as Polycyclic aromatic hydrocarbons (PAHs) and Tobacco-specific nitrosamines (TSNAs)[9]. As these compounds are filtered through the kidney, the compounds get metabolized and lead to inflammation and DNA damage, paving the way for carcinogenesis[2]. In a retrospective study including 845 eligible patients with advanced RCC, former smokers had a 1.6-fold increased odds of advanced RCC, while current smokers have a 1.5-fold increased odd[10]. In a meta-analysis of 24 studies of the effect of cigarette smoking in renal cell carcinoma occurrence, the relative risk (RR) of RCC for smokers to non-smokers was 1.38, and increased pack sizes smoked per day increases the RR of RCC[11]. In addition, smoke cessation for greater than ten years reduces the RR compared to the cohort with smoke cessation between 1-10 years[11].

Obesity is a risk factor for many cancer types, such as colorectal cancer and breast cancer, including RCC. For example, Insulin and insulin-like growth factors (IGF) levels have been indicated to influence cancer risk and cancer prognosis [12]. In addition, positive associations have been reported between adipocytes cytokines, leptin and adiponectin, and the risk of RCC[13]. A case-control study from Italy that studied the effect of lifetime physical activity and risk of renal cell cancer concludes that 9% of RCC cases in Italy could be avoided by increasing physical activities[14]. For the effects of diet, a case-cohort study has shown that cruciferous vegetable consumption decreases RCC risk[15]. In contrast, studies on the effect of meat consumption on RCC have been contradictory[16-17]. A meta-analysis of 23 studies indicates a positive association between processed and red meat consumption and RCC risk[16], while other studies suggest no such risk[17]. Also, moderate alcohol consumption has been shown to reduce RCC risk[18]. Genetic factors also influence RCC risk, with alterations in at least 11 genes been identified. These genes include *VHL, TCS2, TSC1, SDHD, SDHC, SDHB, PTEN, MET, FH, FLCN, BAP1*[19]. Familial history of RCC increases the risk of developing the disease by two-fold[20].

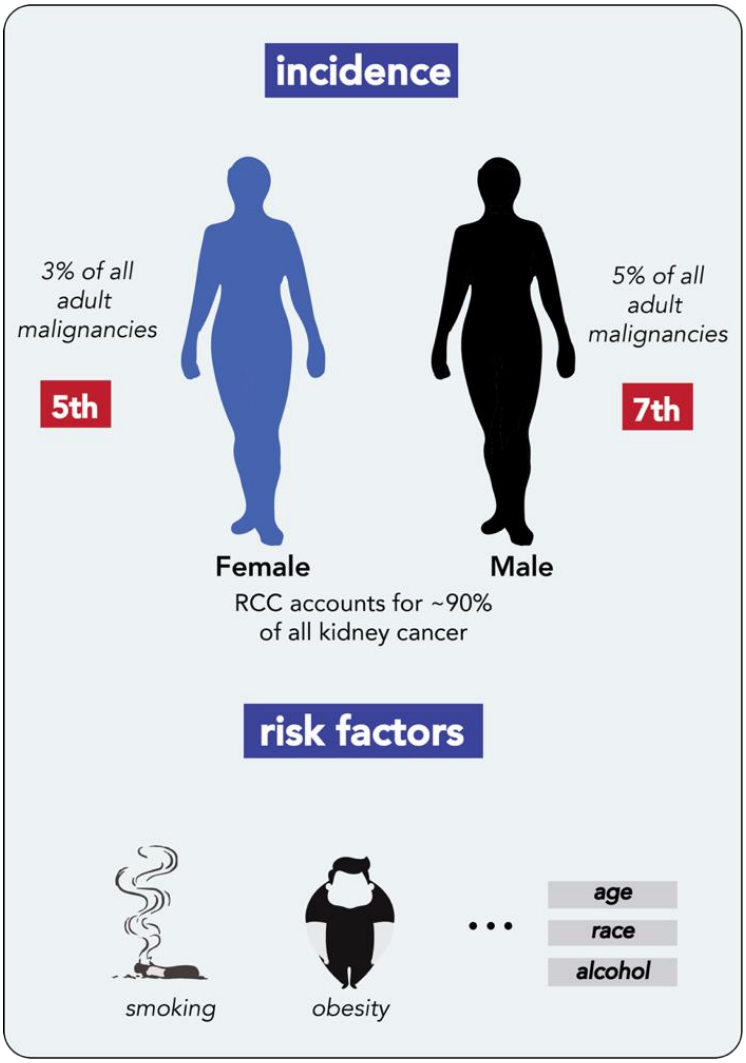


Figure 1-2. Summary of Incidence and Risk Factors of RCC.

Kidney cancer is responsible for 3% of all adult malignancies in females and 5% of all adult malignancies in males. In females, kidney cancer is the 5th most prevalent cancer type and the 7th most prevalent cancer type in males[4].

1.1.2 Pathophysiology and Mechanisms

von Hippel-Lindau (*VHL*) tumor suppressor gene is the most altered gene in clear cell renal cell carcinoma (ccRCC)[21-22]. Mutations include indels, point mutations, 3p25 loss, and epigenetic modification such as promoter methylation[23-24]. The *VHL* gene codes

for the *VHL* protein, a substrate for an E3 ligase complex, which binds hypoxia-inducible factors (HIFs) for proteasome-mediated degradation[1]. As such, loss of the *VHL* gene leads to HIF accumulation. Under normoxemia and in the presence of VHL protein (pVHL), proline residues on HIF 1-alpha (HIF1A) and HIF 2-alpha (HIF2A) are hydroxylated, and they bind pVHL where they get targeted for polyubiquitination and in turn proteasome degradation. On the other hand, in the absence of pVHL, HIFA dimerizes with HIF beta (HIFB) to form a complex and migrate into the nucleus, where it acts as a transcription factor to upregulate proangiogenic genes. These genes include vascular endothelial growth factor (*VEGF*)[23], platelet-derived growth factor-beta (*PDGFB*), and transforming growth factor-alpha (*TGFA*)[25], in addition to upregulation of erythropoietin and extracellular matrix – laying the stage for tumorigenesis. However, studies in mice and humans indicate that *VHL* loss is not a sole requirement to induce RCC[26-27].

In that vein, genomic sequencing studies have implicated more gene alterations in ccRCC apart from *VHL*, and they include tumor suppressor genes protein polybromo-1 (*PBRM1*), BRCA1-associated protein 1 (*BAP1*), and SET domain-containing protein 2 (*SETD2*). They encode chromatin and histone regulating proteins, and they are located at 3p21[21, 28-30]. While there are no predictive genetic biomarkers for ccRCC, some clinical correlations in genetic mutations with ccRCC have been reported. For example, *VHL* mutation is considered the founding process for ccRCC and does not have any strong relationships with clinical outcomes. On the other hand, mutations in *PBRM1*, *BAP1*, and *SETD2* have been linked to disease progression and have associations with aggressive clinical phenotypes[31-33].

Importantly, renal cell carcinoma is driven by metabolic alterations, like other cancers[34]. For example, in ccRCC, the consequence of the pseudohypoxia induced by the inactivation of VHL implies significant metabolic reprogramming to accommodate cellular proliferation[35]. Increased glucose metabolism and reductive carboxylation had been reported in ccRCC[36]. Increased glucose metabolism is a hallmark of tumors, and reductive carboxylation allows for tumor growth by using citrate for fatty acid synthesis[37]. In addition, pentose phosphate metabolism is upregulated in ccRCC to support nucleotide biosynthesis *via* the synthesis of ribose sugars[21, 36].

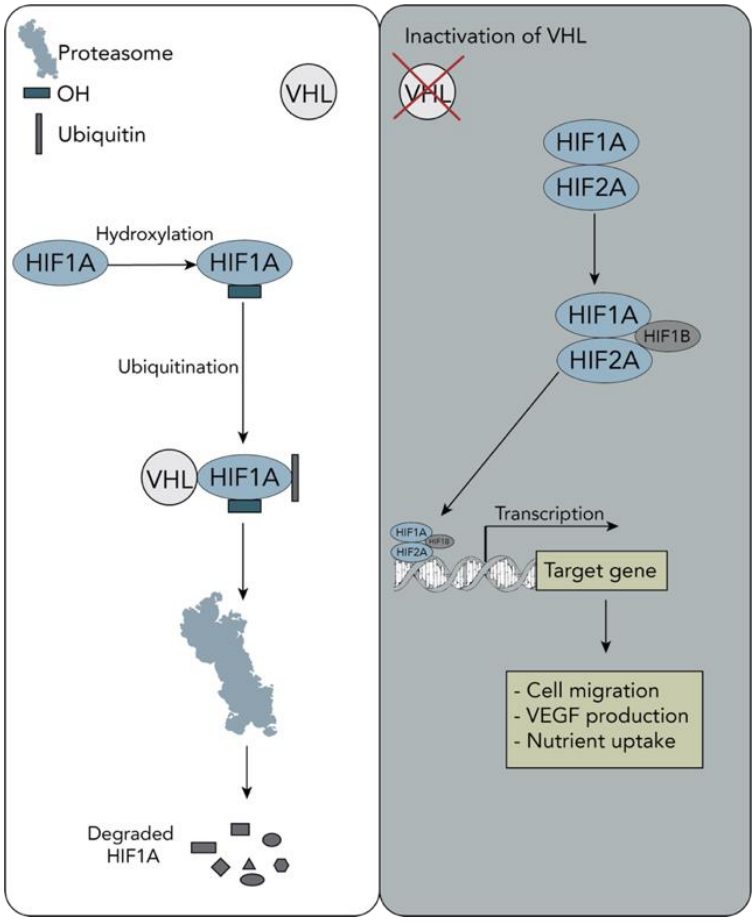


Figure 1-3. von Hippel-Lindau (*VHL*) Tumor Suppressor Gene and Hypoxia-inducible Factors in Clear Cell Renal Cell Carcinoma

1.1.3 Urine Metabolomics of RCC

Several studies have been undertaken to identify urine biomarkers in RCC. In **Table 1-1**, some of the notable urine metabolomics studies that compare RCC to controls are highlighted, indicating the study sample size, the race or location of sample collection, the analytical platform used for measurements, and the biomarker identified.

Table 1-1. Some Notable Urine-based Metabolomics Studies Comparing RCC to Controls

Study	Sample Size	Race/Location of Sample Collection	Platform	Biomarker (Increase in RCC)	Biomarker (Decrease in RCC)
Kind et al. 2007[38]	6C, 6N	Tennessee, US	HILIC-LC-MS, GC-TOF-MS, RP-UHPLC-MS	No compound identification	No compound identification
Kim et al. 2009[39]	11C*, 15N	Texas, US California, US	HILIC LC-MS	No compound identification	
Kim et al. 2011[40]	29C, 33N	California, US	UHPLC-MS, GC-MS	Quinolinate, alpha-ketoglutarate	Gentisate
Monteiro, M. et al. 2016[41]	42C, 49N	Portugal	NMR	2-ketoglutarate, N-methyl-2-pyridone-5-carboxamide, bile acids, galactose, pyruvate, succinate, and valine	4-hydroxyhippurate, 4-hydroxyphenylacetate, acetone, GAA, glycine, hippurate, malonate, phenylacetylglutamine, tartrate, trigonelline

Ragone et al. 2016[42]	40C*, 29N	Italy	NMR	Creatine, alanine, lactate, and pyruvate	Hippurate, citrate, and betaine
Monteiro, M. et al. 2017[43]	30C, 37N	Portugal	GC-MS	2-oxopropanal	2,5,8-trimethyl-1,2,3,4-tetrahydronaphthalene-1-ol
Niziol et al. 2018[44]	7C*, 15N	Poland	LC-HRMS	Hydroxybutyrylcarnitine, decanoylcarnitine, propanoylcarnitine, carnitine, dodecanoylcarnitine, and norepinephrine sulfate.	riboflavin, acetylaspartylglutamate
Liu et al. 2019[45]	100C, 129N	China (all Chinese subjects)	LC-MS	N-Jasmonoyltyrosine, Androstenedione, Dopamine 4-sulfate, 3-Methylazelaic acid, 7alpha-hydroxy-3-oxochol-4-en-24-oic acid, Lithocholyltaurine, 11-Dodecenoic acid	Tetrahydroaldosterone-3-glucuronide, Cortolone-3-glucuronide
Wang et al. 2019[46]	117 C**, 98N	China	UPLC-MS	α-CEHC, flunisolide, glycerol tripropanoate	β-cortolone, deoxyinosine, 11b,17a,21-trihydroxypregnenolone
Zhang et al. 2020[47]	39C, 68N	China	LC-MS		Aminoadipic acid, 2-(formamido)-N1-(5-phosphod-ribosyl) acetamidine and alpha-N-phenylacetyl-l-glutamine

C* = clear cell renal cell carcinoma. C = renal cell carcinoma, N=healthy controls, C** = 53 bladder cancer patients and 64 RCC. HILIC-LC-MS: Hydrophilic Interaction

Chromatography Liquid Chromatography Mass Spectrometry, GC-TOF-MS: Gas Chromatography Time-Of-Flight Mass Spectrometry, RP-UHPLC-MS: Reverse Phase Ultra High-Performance Liquid Chromatography Mass Spectrometry, LC-HRMS: Liquid Chromatography-High Resolution Mass Spectrometry, VOC: Volatile organic compounds.

Studies comparing benign renal tumors and RCC were not included in this review. In Monteiro, M. *et al.* 2016, hypoxanthine and isoleucine are potentially confounded[41]. In Monteiro, M. *et al.* 2017, the biomarkers were selected from the 21 initial VOCs, using two independent small sample sets[43]. In Niziol *et al.* 2018, the biomarkers in urine were also present in the tissue[44]. In Wang *et al.* 2019[46], bladder cancer patients are confounders; the biomarkers identified were validated in an external cohort consisting of 30 RCC and 44 controls; and in addition, only early-stage RCC patients (T1 and T2 stages) were considered.

1.2 Metabolomics for Biomarker Discovery

Metabolomics is a field of study that investigates metabolites profile within biological samples[48]. It measures the small-molecule compounds, both endogenous and exogenous molecules, which are substrates, intermediates, and products of cellular processes, collectively called the metabolome[49]. As such, metabolomics studies the biological phenotype of an organism (i.e., the metabolome), with imprints of genetics and the environment. Metabolomics experiments often proceed with two main experimental methodologies: untargeted metabolomics and targeted metabolomics[50]. In the former, experiments are carried out to measure thousands of metabolic features without explicit hypotheses; in contrast, in the latter, only 'targeted' metabolites or groups of metabolites

are measured to test a specific hypothesis. A biomarker is a biological feature used to track the presence, absence, or progression of a diseased or physiological condition of an organism. Therefore, biomarker discoveries in metabolomics are mostly achieved *via* untargeted (or global) metabolomics. This is accomplished *via* comparing the metabolic profile of healthy (or control) samples with diseased samples to identify discriminative metabolites (i.e., the biomarkers).

Metabolomics experiments are often carried out by mass spectrometry (MS) and/or nuclear magnetic resonance (NMR), as they are used to characterize the molecular composition of a sample[51]. NMR leverages atomic nuclei's magnetic property to characterize the molecular composition of samples. In MS, samples are ionized, and the ions are separated, and the mass-to-charge ratio is detected and measured to elucidate molecular compounds. MS has higher sensitivity compared to NMR[52]. As such, MS affords a much more extensive metabolite coverage than NMR. Liquid chromatography (LC) and gas chromatography (GC) are used for complex mixture separation before detection with MS. In that vein, in MS experiments, more complicated sample preparation is required, while minimal sample preparation is needed for NMR. This makes NMR desirable for large sample size studies, as the time of sample preparation is lower compared to MS. Another advantageous quality of NMR is that, because samples are not ionized for detection, recovery of samples after data acquisition is possible. In addition, NMR has high reproducibility. It is important to note that data collection is just the first step in metabolomics, as it involves chemistry (metabolite identification), statistics, and machine learning [53-54]. In the following sections, NMR will be discussed further, especially in the

context of urine metabolomics. In addition, applications of machine learning techniques for data mining are discussed.

1.2.1 Nuclear Magnetic Resonance Spectroscopy (NMR)

Nuclei such as ^{1}H, ^{13}C, ^{15}N, and ^{31}P possess nuclear spins of ½, and NMR takes advantage of this property to interrogate the chemical environment of the nuclei, and hence the molecular composition of samples. In the absence of a magnetic field, they are randomly oriented; however, when an external magnetic field is applied, they become parallel to the applied magnetic field, either aligned with or opposed. This creates a spin configuration: the spin-aligned state is the lower energy state, and the spin-opposed state is the higher energy state.

The distribution of the nuclei between different energy states in thermal equilibrium is given by Boltzmann distribution:

$$\frac{N_\beta}{N_\alpha} = e^{-\Delta E/k_B T} \tag{1.1}$$

Where N_β and N_α represents the number of nuclei in the upper energy and lower energy states respectively. ΔE is the energy difference between the upper and lower energy states, k_B is the Boltzmann constant, and T is the absolute temperature in Kelvin. In proton and other nuclides, ΔE is very small compared to the $k_B T$, which is the average energy of the thermal motions, as such the populations at the two energy states are almost equal with only a small excess number of spins aligned (lower state) – which is in the region of parts per million (ppm).

In NMR, in order to extract information about the nuclei, an electromagnetic field is used to flip this spin state (spin-flip). The energy difference between the spin states (ΔE) equals the energy required for the spin-flip, which corresponds to the radio frequency range

of the electromagnetic spectrum (also called resonance frequency). The resonance frequency of a particular nucleus is, in turn, dependent on the magnetic field strength at the nucleus.

In brief, when NMR samples are placed in the magnetic field, the active nuclei will precess at a particular resonance frequency. The resonance frequency is measured in the time domain as free induction decays (FIDs), and FIDs are converted to the frequency domain via Fourier transformation (FT), resulting in the NMR spectrum (**Figure 1-4**). As such, the signal in the NMR spectrum is the resonance, and the frequency of the resonance is the chemical shift. The chemical shift of a resonance is defined in reference to the resonance of a reference compound, and it is independent of spectrometer frequency. The formula for computing chemical shift δ is shown below:

$$\delta = \frac{signal\ frequency - reference\ frequency}{spectrometer\ frequency} \times 10^6$$

Because of the natural abundance of hydrogen (99.97%), ^{1}H NMR is the most common type of NMR spectroscopy. Sample preparation for biofluids such as urine for proton NMR experiments involves the addition of 1) a deuterated solvent such as deuterium oxide (D_2O), which ensures the stability of the magnetic field for quality data output, 2) a reference compound standard, as described above and 3) a buffer (saline or phosphate) that controls pH differences between samples. Henceforth, ^{1}H NMR will be simply referred to as NMR.

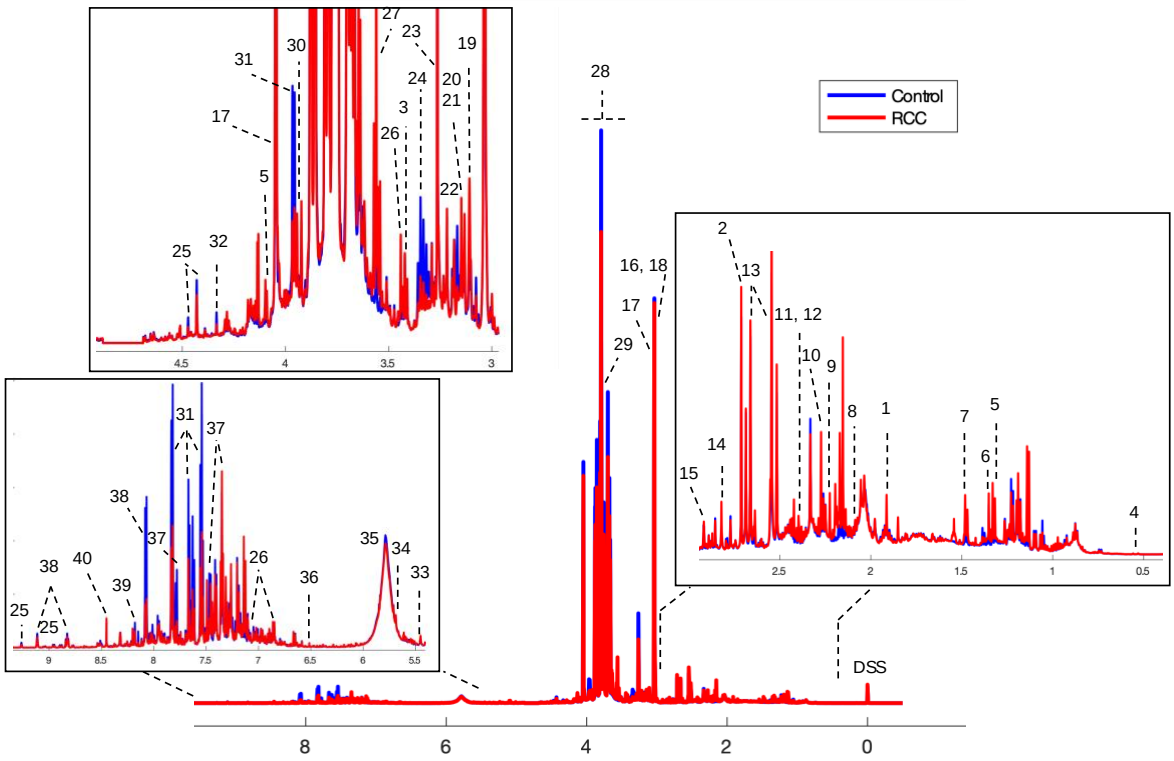


Figure 1-4. A Sample 1D ^{1}H NMR Spectra.

These data are used in chapter 3. Numbers indicate identified metabolites.

1.2.1.1 Experimental Design

In the study carried out for this thesis, urine samples were stored at -80^{0} C – as is the recommendation for other biofluids used in metabolic profiling – to prevent metabolic alterations [55-56]. After sample collection, experiments were carefully designed to detect any confounding variables such as instrumental or sample preparation bias. The methodology used for untargeted NMR metabolomics experimental design will be summarized (also see **Figure 1-5**). The first important concept is sample randomization, as all samples in the study are randomized to correct any bias that might emanate from sample preparation and NMR data collection. In addition, multiple experimental controls are deployed for the

study. First are the NMR buffer blanks, which help detect any carry-over of biofluids from one sample to the next during high-throughput NMR measurements. In addition, it is used to confirm that there are no contaminations during sample preparation.

The NMR buffer blank constitutes high-performance liquid chromatography (HPLC) grade water, saline/phosphate buffer, and deuterated solvent. The buffer blank should consist of 5% of the total samples in the study and should be arranged as such: for each NMR racks, one buffer blank is placed at the beginning, middle, and at the end of each rack (a rack is the equivalent of one run), while the rest are randomized with the study samples. Furthermore, there is a need for external pooled controls that helps to check for the stability of NMR instrumentations throughout the measurement period. Five percent of the total samples in the study representing external pooled controls is also recommended. To that end, one external control sample is placed at the beginning of a run after an NMR buffer blank, and at the end of the run, just before the NMR buffer blank; while other external pooled controls were randomized with the study samples. After NMR data collections and data processing, the external pooled controls are supposed to cluster tightly together in a principal component analysis (PCA) plot, as shown in **Figure 1-6**. Finally, the internal pooled controls are included. It consists of all experimental samples in the study; as such, two-dimensional (2D) NMR measurements are collected on these samples as they aid metabolite identification, which will be discussed in detail in later sections.

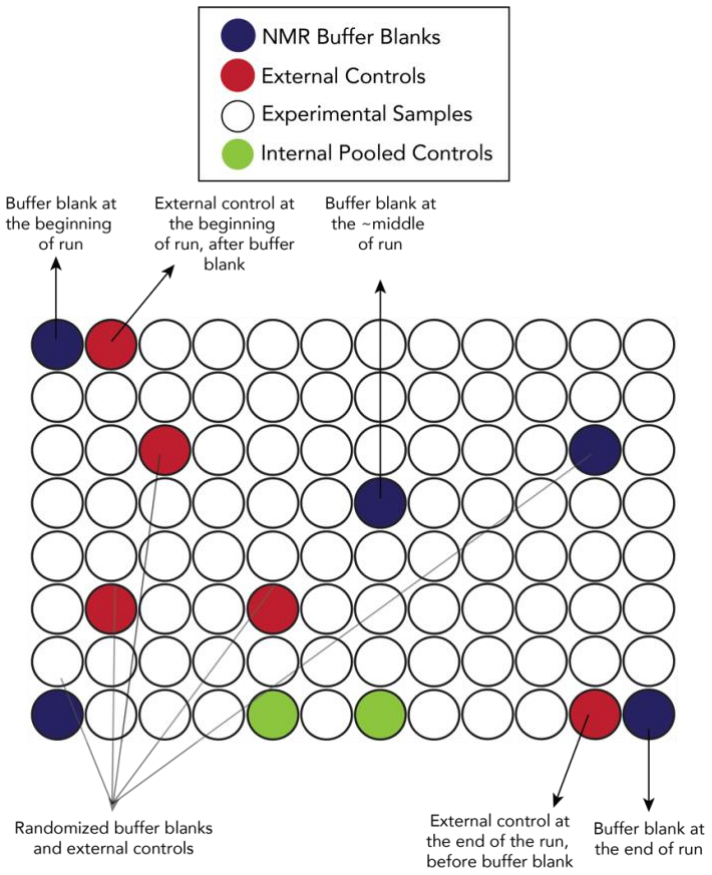


Figure 1-5. An Illustrated Experimental Design for NMR Metabolomics.

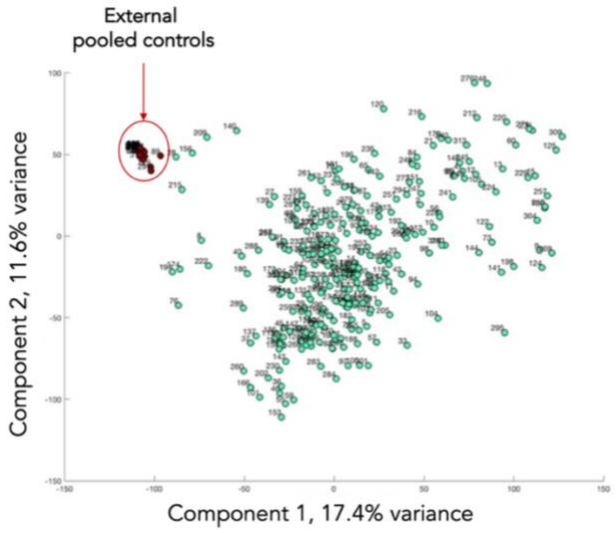


Figure 1-6. A Sample PCA Score Plot Shows the Clustering of the External Pooled Controls.

This indicates the reliability of the NMR instrumentation over the measurement time. In red are the external pooled controls, while in green are the experimental samples.

1.2.1.2 NMR Spectra and Data Processing

After NMR data collection, spectral processing, and data post-processing are conducted before NMR metabolite identification can proceed. The first step is phase correction which involves mixing the real and the imaginary components of the spectrum to get a pure absorptive signal. An unphased spectrum will have peaks pointing upwards and downwards, with broad tails while a phased spectrum will have narrow, symmetrical peaks pointing upwards. The next step is chemical shift referencing, which is important for metabolite ID, spectra alignment, and other downstream analysis[57-58]. Chemical shift referencing entails adjusting the chemical shift axis to a reference compound, for example, adjusting 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) to the 0 ppm mark. After chemical shift referencing comes the baseline correction, which involves making regions with no signals flat with zero intensity. The process is trivial for the NMR spectrum with a few peaks; however, it is not with a spectrum with thousands of peaks, as is the case with urine. Correcting the baseline involves a semi-automated step where baselines are identified manually, and NMR programs complete the baseline correction steps. This correction is essential for spectral alignment and accurate peak integrations. The sub-spectral selection follows, with unwanted spectral regions removed for the NMR spectra. Typically, this includes the ends of the spectra (less than -0.50 ppm, greater than 10.0 ppm) and water regions (between ~4.50 to 4.9 ppm). Afterward, spectral alignment is conducted. Spectral alignment involves adjusting peaks so that peaks that belong to the same

compound align on the spectra. Chemical shift misalignments occur because of non-systematic misalignment sources such as temperature changes, intermolecular interactions, instrumental factors, and systematic sources which are of biological origins, such as sample pH.[59-60] In response to this challenge, computational methods are employed to align the spectra, some of which include: correlation optimized warping (COW)[61], interval correlated optimized shifting (icoshift)[62], and fuzzy warping[60]. Next, aligned spectra are normalized. Unlike biofluids like serum and plasma, urine is not under tight physiological control. Therefore, metabolite concentration variations exist between urine samples[63-64] due to differences in water intake and other physiological reasons. Hence, a sample-to-sample normalization operation is required. Normalization methods often used in NMR metabolomics include total integral intensity[65], histogram matching (HM)[66], and probabilistic quotient normalization (PQN)[67]. PQN involves scaling the NMR spectra using a 'probable dilution factor.' The process consists of selecting a reference spectrum (usually the median spectrum), calculating the quotients of the variables of a test spectrum using the reference, and computing the median of the quotients. The median of the quotient is, in turn, used to scale the test spectrum. Finally, before multivariate statistical analysis, scaling is carried out to make metabolites comparable. Computational methods used include range scaling, Pareto scaling, level scaling, and autoscaling[68-69].

1.2.1.3 Metabolite Identification

Given the high natural abundance of ^{1}H, as indicated above, 1D ^{1}H NMR metabolomics gives rise to hundreds of peaks with a considerable amount of overlap in complex mixtures like urine. This makes metabolite identification for NMR-based urine

metabolomics a very challenging step. As such, metabolite identification is usually directed towards statistical differences between peaks emanating from, for example, diseased samples compared to the metabolite profile of healthy controls, rather identifying all metabolites. Many strategies, techniques, and methods are employed to assist in metabolite identification. These include the use of Statistical Total Correlation Spectroscopy (STOCSY)[70], commercial packages such as Assure NMR (Bruker), open-source software such as COLMARm[71], two-dimensional (2D) NMR experiments, and confirmation with chemical standards. This section will discuss the identification of 'known unknowns' that is unknown metabolites in a sample that had been previously identified, and not 'unknown unknowns,' which are unknown metabolites that had not been previously identified[72].

STOCSY leverages the fact that NMR can have multiple resonances that correspond to the same metabolite[70]. Therefore, resonances from the same metabolite will have the same relative intensities to each other, independent of metabolite concentrations. In brief, STOCSY involves the selection of a driver peak and correlates it with all the peaks in the NMR spectra across the entire study. The highly correlated peak clusters are metabolite candidates present in the complex mixture. In addition, software programs such as AssureNMR (Bruker) and open-access databases such as Human Metabolome Database (HMDB)[73] and BioMagResBank (BMRB)[74] can be used to identify the metabolites from the chemical shift assignments from ^{1}H NMR experiments, selected by a chemist or from STOCSY analyses.

While 1D ^{1}H NMR gives chemical information like the chemical shift, and multiplicity, and the shape of resonances, for greater confidence in metabolite identification, two-dimensional (2D) NMR experiments are conducted. 2D NMR data gives two-dimensional

chemical information that helps to resolve the data and tackle 1D ^{1}H NMR overlap issues. Some of the homonuclear and heteronuclear 2D experiments typically carried out include ^{1}H–^{1}H J-resolved (*J-Res*)[75-76], ^{1}H–^{13}C Heteronuclear Single- Quantum Coherence (HSQC)[77], ^{1}H–^{1}H Total Correlation Spectroscopy (TOCSY)[77-78] and a combination of the last two, ^{1}H-^{13}C heteronuclear single quantum correlation-total correlation spectroscopy (HSQC-TOCSY). *J*-coupling values can be derived from *J-Res* in addition to the multiplicity of signals. HSQC, on the other hand, gives the correlation of directly coupled hydrogen and carbon atoms in the complex mixture. In contrast, TOCSY provides the correlations between all protons in a given spin system. While 2D NMR data can be invaluable, the experiments can take several hours to complete; on the other hand, ^{1}H NMR experiments are completed in minutes. Therefore, 2D experiments are typically conducted on internal pooled controls, as described under the experiment design section; and databases like COLMARm (http://spin.ccic.ohio-state.edu/index.php/colmarm/index) web server[71] allows for the semi-automated query of ^{1}H–^{13}C HSQC, HSQC-TOCSY, and TOCSY. For conclusive evidence for metabolite identification, chemical standards of the metabolite are spiked into the complex mixtures. However, this is not feasible for all metabolites in a complex mixture but only for metabolites of the highest import. In the Edison Laboratory, we therefore assign a confidence score to identified metabolites, based on the strength of evidence available. The scores are defined as follows: (1) putatively characterized compound classes or annotated compounds, (2) matches from 1D NMR to literature and/or 1D BBiorefcode compound (AssureNMR) or other database libraries such as BMRB and HMDB (3) matched to HSQC, (4) matched to HSQC and validated by

HSQC–TOCSY (COLMARm), and (5) validated by spiking the authentic compound into the sample.

1.3 Machine Learning

Given the rise of high throughput omics technologies over the past couple of decades, data mining had become inevitable in fields such as metabolomics, especially in untargeted metabolomics, as is the case in this thesis. Tom Mitchell gave a widely accepted definition of machine learning (ML) given as such: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task T, as measured by P, improves with experience E." [79] In a metabolomics context, the experience E is the metabolomics data, task T could be the classification of samples to either disease group or control group, while performance measure P could be the accuracy of the classification task. In essence, ML entails finding predictive patterns in datasets, be it text, images, or tables. Our focus will be on tabular datasets in this study, as the abundance of metabolites are presented in tables.

1.3.1 Machine Learning Systems

One approach for classifying machine learning systems is whether they are under human supervision or not. To this end, we have methods like supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning. Supervised and unsupervised learning are carried out in this thesis; therefore, introductory comments will be limited to these topics. Supervised learning entails training ML models with labeled data, and it includes classification and regression tasks. Classification tasks involve training and predicting with discrete labels as in binary classification (two groups) and multi-class

classification (> two groups). On the other hand, regression tasks use continuous labels (numeric data) for training and predictions. For example, in the studies presented in this thesis, classification tasks were carried out to detect RCC and for RCC stage stratification. At the same time, regression tasks were conducted for RCC tumor size predictions. In all cases, urinary metabolites are the predictor variables.

Unsupervised learning applications are more prevalent in biology research, and it involves 'training' with unlabeled data, with the analogy that a ML system tries to learn without a teacher. Unsupervised learning techniques can be further classified into clustering and dimensionality reduction techniques. Clustering aims to detect the similarity between groups, and it involves methods like k-Means, and hierarchical cluster analysis (HCA). On the other hand, dimensionality reduction reduces the dimension of the dataset and projects it on a 2D or 3D dimensional space, all the while reducing information loss in the process. Principle component analysis (PCA) belongs to this class of ML systems. See **Figure 1-7** for ML systems and methods used in the thesis. ML explanations refer to the interpretation of the cause of the predictive output of ML models. This topic is discussed in some detail in session 1.3.3.

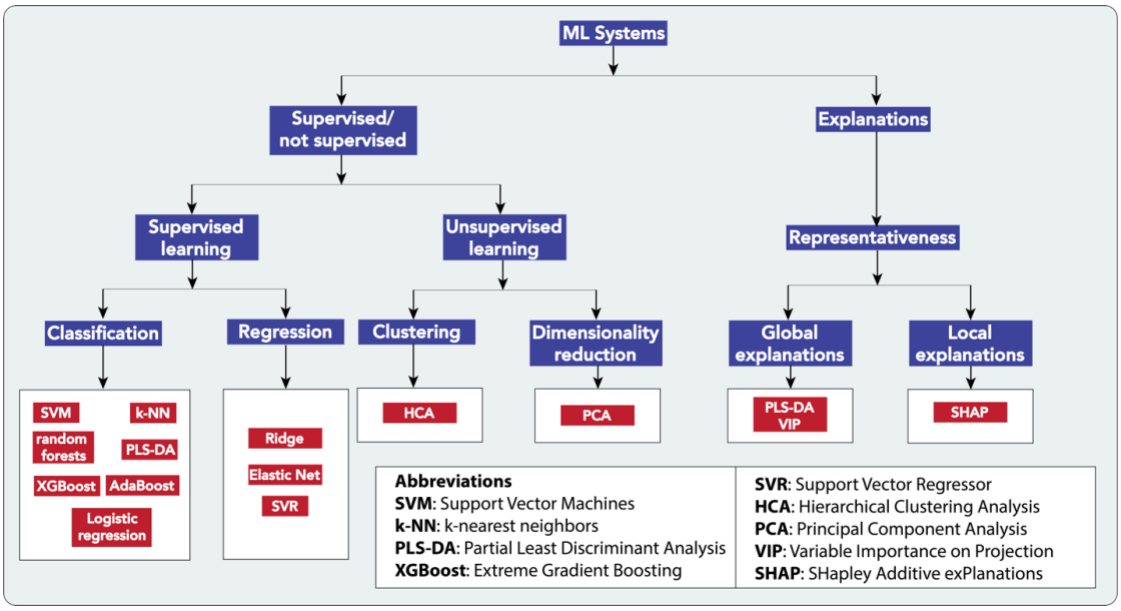


Figure 1-7. Machine Learning Systems and Techniques Used in the Thesis.

Table 1-2. The Basis of Supervised and Unsupervised Learning Techniques Used in the Study.

Algorithms	**Basis**
Support vector machines[80] (**SVM**)	- It fits the largest possible margin between classes while limiting margin violations - Learns linear and non-linear relationship in datasets - Linear SVM separate linearly separable classes with a hyperplane - Kernelized SVM is used for linear inseparable datasets such as polynomial kernel and Gaussian Radial Basis Function (RBF) Kernel.
SVM Regressor[81]	- An extension of SVM for regression tasks

(**SVR**)	- It fits instances within a margin, closer to the hyperplane, while limiting margin violations.
Random Forests (**RF**)	- An ensemble of decision trees[82] - Models are typically trained via the bagging method[83] - Can learn complex relationships in datasets
Adaptive Boosting[84] (**AdaBoost**)	- It belongs to the family of boosting algorithms, where weak learners are transformed into strong learners by sequentially training predictors. - The base learner is typically a decision tree but could be any algorithm such as SVM. - AdaBoost assigns higher weights to misclassified instances in the previous predictor so that the next predictor focuses more on incorrectly classified instances.
Extreme Gradient Boosting[85] (**XGBoost**)	- An advanced gradient boosting algorithm. - It transforms weak learners into strong learners via the summation of decision tree residuals.
Logistic Regression	- A regression algorithm used for classification tasks. - Maps regression output to probabilities *via* a sigmoid function.
k-Nearest Neighbors (***k*-NN**)	- An instance-based learning algorithm. - Induction bias: Closer instances belong to the same class as defined by *k*- nearest neighbors.

Partial Least Square Discriminant Analysis **(PLS-DA)**	- PLS-DA is a supervised algorithm extension of PLS regression, and it is used widely in metabolomics. - It projects high dimensional datasets into a lower dimension space and maximizes the covariance between the independent and response variables[86].
Ridge regression	- A regularized form of linear regression
Elastic net regression	- Linearly combines the regularized form ridge and lasso regression
Hierarchical clustering analysis (**HCA**)	- A cluster analysis algorithm - Clusters instances based on data similarities using varied metric and linkage criteria.
Principal component analysis (**PCA**)	- Reduces the dimensions of a high-dimensional dataset while preserving the maximum variance in the dataset.

1.3.2 Resampling Methods and Performance Metrics

Resampling methods in machine learning involve sampling instances to assess the quality of models. These methods include the validation set method, *k*-fold cross-validation method, and leave one out cross-validation method (LOOCV). The validation set method involves randomly distributing samples into a training set and a test or validation set. A training set is used to train the model, while the test set assesses the model's performance. In the *k*-fold cross-validation method, the samples are divided into training set and test set

k-times, such that each sample is used for both training and testing. *k*-fold cross-validation is usually employed when the sample size is small. Finally, LOOCV is a special case of the *k*-fold cross-validation method where $k = 1$. In this formulation, the test set is restricted to only one sample for each round of testing.

Assessing the quality of ML models depends on the task at hand. For classification tasks, accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC ROC) are popular choices. Accuracy is simply the percentage of the correctly predicted class. Sensitivity is the percentage of correctly predicted positive cases, and specificity is the percentage of correctly predicted negative cases. For biomedical applications, healthy controls are often encoded as negative cases, and diseased samples are encoded as positive cases. The ROC curve plots the true positive rate against the false-positive rate, showing the performance of a model at different classification thresholds. AUC, in turn, measures the area covered by the entire ROC curve; as such, it gives an aggregated predictive power of a model under all classification thresholds possibilities. For regression tasks, residual plots are used to access the quality of models. It plots residuals ($y - \hat{y}$) against the predicted numeric value ($\hat{y}$) (**Figure 1-8**), giving a visual guide for accessing regression models. A regression metric is the coefficient of determination, R^2 described mathematically below:

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

Where i is the index for a sample, y is the ground truth, $\hat{y}$ is the predicted value, and $\bar{y}$ is the mean ground truth. In brief, R^2 is the proportion of the variance in the prediction accounted for by the predictors, X.

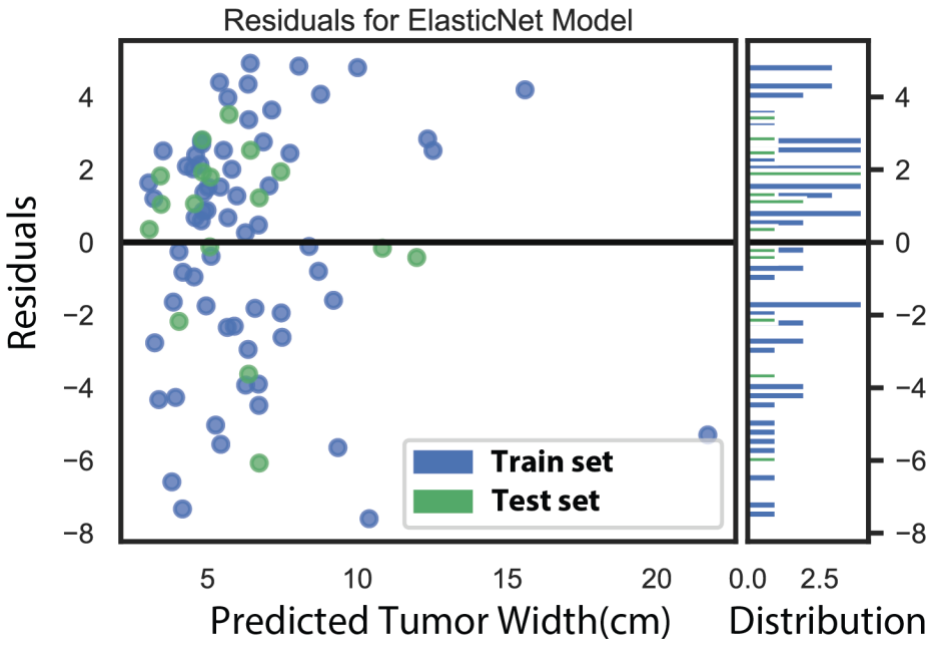


Figure 1-8. A Sample Residual Plot.

This result is presented and discussed in chapter 3 of the thesis with regard tumor size predictions.

1.3.3 Machine Learning Explanations

The goal of machine learning models is to unearth patterns buried within datasets. As ML models had become increasingly successful at such pattern recognitions, the result led to more complex models. A bane of this increased complexity of models is that these models are typically impervious to interpretations. As such, there exist an accuracy interpretability trade-off trend in machine learning: the more accurate a ML model is, the less interpretable it is likely to be. (**Figure 1-9**). These opaque models are called 'black box' models. As a result, in recent years, there has been increased progress in the field of explainable artificial intelligence (XAI)/ interpretable machine learning (IML)[87-88].

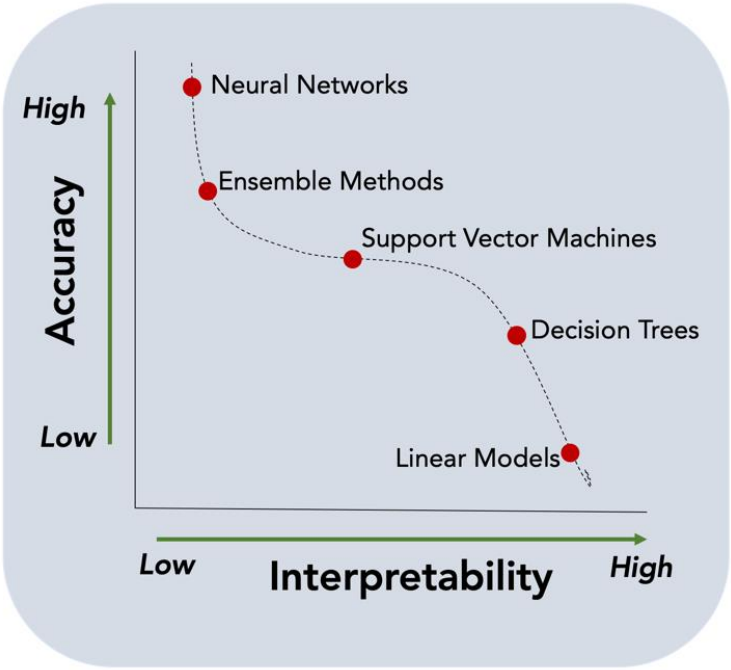


Figure 1-9. The Accuracy Interpretability Trade-off of Machine Learning Models. More complex models tend to be more accurate and less interpretable[89].

IML methods are classified based on several criteria[90]. One criterion is whether the interpretation is accomplished using the same ML model for prediction (intrinsic interpretations) versus if interpretations are carried out after model building (post-hoc interpretations). A post-hoc method will be desirable as it allows for the flexibility of ML choice for model building, as they tend to be model agnostic methods. Another criterion is the results of the explanations. For example, some IML methods provide feature summary statistics such as feature importance scores. This is the case of variable importance of projection (VIP) score as in PLS-DA[91] and Gini index in random forests[92]. Another type of explanation result comes in the form of feature summary visualization, as is the case in partial dependence plots (PDP)[93] and individual conditional expectation (ICE) plots[94]. Another criterion for classification is the scope of explanations. This could range from a

method that can explain entire model behavior (global explanations) to individual predictions (local explanations).

As an addition to the classification above, there are important properties of interpretation methods that require some highlight. This includes 1) the algorithmic complexity of explanation methods, 2) the fidelity of the method, that is the accuracy of the explanations, 3) the representativeness of explanations, that is the scope of explanations, 4) translucency of method: the extent to which an IML methods looks inside the ML model to explain model behavior, 5) portability: the scope of the ML method it can interpret, and 6) human comprehensibility of explanation results.

Unlike some other fields of studies/applications where accuracy is the sole goal, interpreting machine learning models is paramount in biology. In metabolomics, interpretations are primarily achieved *via* PLS-DA Variable Importance on Projection (VIP) scores[91, 95]. PLS-DA is a linear-based model whose linear assumptions permit model interpretation, making it an intrinsic interpretation method. However, scientists can be needlessly restricted to PLS-DA in the event that it's predictive accuracy is inferior to other machine learning models. Therefore, a post-hoc/model agnostic interpretable method for explaining metabolomics ML models might be desirable. Such post-hoc IML methods include Local Interpretable Model-agnostic Explanations (LIME)[96], which can explain individual instances; permutation feature importance, which gives global explanations – explains the entire model behavior[97-98]; Partial Dependence Plot (PDP), which computes and visualizes the impact of one or two features on prediction outcome[93]; SHapley Additive exPlanations (SHAP) gives both local and global explanations[99-100]; and several other methods[87].

1.4 Scope of Thesis

In this thesis, I hypothesized that the metabolic reprogramming that supports RCC[35] would allow for the identification of metabolite biomarkers in human urine. To improve upon previously published studies, 1) a large cohort (105 RCC patients and 179 controls) – compared to many published studies in the field – was selected for the study. 2) While many studies use either Liquid Chromatography−Mass Spectrometry (LC−MS) or Nuclear Magnetic Resonance (NMR) for metabolite profiling, both LC-MS and NMR were used in this thesis to ensure a broad analyte coverage. 3) Rather than relying on machine learning software for chemists that are often very restrictive, custom machine learning algorithms were built using Python programming language to explore the benefits of several induction biases that various ML models afford. 4) Given the mounting evidence of the potential role of the exposome – the totality of environmental exposures of an individual in a lifetime – in cancer[101-102], biomarkers were not restricted to endogenous metabolites, as is usually the case in many biomarker studies.

In Chapter 2, a RCC detection study was conducted by comparing RCC patients with healthy controls. This chapter is a research paper accepted for publication in the *Journal of Proteome Research* titled "*machine learning-enabled renal cell carcinoma status prediction using multiplatform urine-based metabolomics*." In chapter 3, the metabolomics data of the RCC cohort was used to identify metabolites associated with RCC progression. This task was accomplished by 1) predicting the primary tumor size of RCC and 2) discrimination of early RCC (stage I and II) from advanced RCC (stage III and IV). The chapter is a manuscript written for submission to the journal *Cancers* and

tentatively titled "*Urine-Based Metabolomics and Machine Learning Reveals Metabolites Associated with Renal Cell Carcinoma Progression*."

In chapter 4, machine learning interpretations were tackled. PLS-DA VIP score is the contemporary IML method in metabolomics, with the downside that ML choices are often restricted to PLS-DA and explanations are not local. I introduced Shapley values in game theory and Shapley Additive Explanations (SHAP) – the adaptation of Shapley values in IML. Random forest (RF) and extreme gradient boosting (XGBoost) were used to predict binary classes in previously published clinical metabolomics datasets. Tree-based SHAP (Tree SHAP) explained the machine learning predictions, and its utility was validated. In addition, given the large sample size of the RCC detection study, random forest model was built using the selected biomarker panel in chapter 2, and its prediction is explained with Tree SHAP. This chapter is primarily based on a manuscript that had been submitted to the journal *Metabolites*, and it has been titled "*Digging Deep: Applying Tree-based Shapley Additive Explanations to Metabolomics Datasets*."

In chapter 5, I concluded the thesis by summarizing the contribution of my studies to the field and offered insights as to the path required to building on my findings and ultimately making urine testing for RCC available in the clinic.

1.5 References

1. Hsieh, J. J.; Purdue, M. P.; Signoretti, S.; Swanton, C.; Albiges, L.; Schmidinger, M.; Heng, D. Y.; Larkin, J.; Ficarra, V., Renal cell carcinoma. *Nat Rev Dis Primers* **2017,** *3*, 17009.

2. Padala, S. A.; Barsouk, A.; Thandra, K. C.; Saginala, K.; Mohammed, A.; Vakiti, A.; Rawla, P.; Barsouk, A., Epidemiology of Renal Cell Carcinoma. *World J Oncol* **2020,** *11* (3), 79-87.

3. Xu, W. H.; Qu, Y. Y.; Wang, J.; Wang, H. K.; Wan, F. N.; Zhao, J. Y.; Zhang, H. L.; Ye, D. W., Elevated CD36 expression correlates with increased visceral adipose tissue and predicts poor prognosis in ccRCC patients. *J Cancer* **2019,** *10* (19), 4522-4531.

4. Siegel, R. L.; Miller, K. D.; Fuchs, H. E.; Jemal, A., Cancer Statistics, 2021. *CA Cancer J Clin* **2021,** *71* (1), 7-33.

5. SEER*Explorer: An interactive website for SEER cancer statistics [Internet] https://seer.cancer.gov/explorer/. *Surveillance Research Program, National Cancer Institute. [Cited 2021 April 15].*

6. American Cancer Society: Key Statistics About Kidney Cancer [Internet]: https://www.cancer.org/cancer/kidney-cancer/about/key-statistics.html. *Retrieved 2021 May 27.*

7. Haas, N. B.; Nathanson, K. L., Hereditary kidney cancer syndromes. *Adv Chronic Kidney Dis* **2014,** *21* (1), 81-90.

8. American Cancer Society: Survival Rates for Kidney Cancer [Internet]: https://www.cancer.org/cancer/kidney-cancer/detection-diagnosis-staging/survival-rates.html. *Retrieved 2021 May 27.*

9. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis of Smoking-Attributable Disease, A Report of the Surgeon General. **2010**.

10. Tsivian, M.; Moreira, D. M.; Caso, J. R.; Mouraviev, V.; Polascik, T. J., Cigarette smoking is associated with advanced renal cell carcinoma. *J Clin Oncol* **2011,** *29* (15), 2027-31.

11. Hunt, J. D.; van der Hel, O. L.; McMillan, G. P.; Boffetta, P.; Brennan, P., Renal cell carcinoma in relation to cigarette smoking: meta-analysis of 24 studies. *Int J Cancer* **2005,** *114* (1), 101-8.

12. Pollak, M., Insulin and insulin-like growth factor signalling in neoplasia. *Nat Rev Cancer* **2008,** *8* (12), 915-28.

13. Liao, L. M.; Schwartz, K.; Pollak, M.; Graubard, B. I.; Li, Z.; Ruterbusch, J.; Rothman, N.; Davis, F.; Wacholder, S.; Colt, J.; Chow, W. H.; Purdue, M. P., Serum leptin and adiponectin levels and risk of renal cell carcinoma. *Obesity (Silver Spring)* **2013,** *21* (7), 1478-85.

14. Tavani, A.; Zucchetto, A.; Dal Maso, L.; Montella, M.; Ramazzotti, V.; Talamini, R.; Franceschi, S.; La Vecchia, C., Lifetime physical activity and the risk of renal cell cancer. *Int J Cancer* **2007,** *120* (9), 1977-80.

15. Liu, B.; Mao, Q.; Wang, X.; Zhou, F.; Luo, J.; Wang, C.; Lin, Y.; Zheng, X.; Xie, L., Cruciferous vegetables consumption and risk of renal cell carcinoma: a meta-analysis. *Nutr Cancer* **2013,** *65* (5), 668-76.

16. Zhang, S.; Wang, Q.; He, J., Intake of red and processed meat and risk of renal cell carcinoma: a meta-analysis of observational studies. *Oncotarget* **2017,** *8* (44), 77942-77956.

17. Lee, J. E.; Spiegelman, D.; Hunter, D. J.; Albanes, D.; Bernstein, L.; van den Brandt, P. A.; Buring, J. E.; Cho, E.; English, D. R.; Freudenheim, J. L.; Giles, G. G.;

Graham, S.; Horn-Ross, P. L.; Hakansson, N.; Leitzmann, M. F.; Mannisto, S.; McCullough, M. L.; Miller, A. B.; Parker, A. S.; Rohan, T. E.; Schatzkin, A.; Schouten, L. J.; Sweeney, C.; Willett, W. C.; Wolk, A.; Zhang, S. M.; Smith-Warner, S. A., Fat, protein, and meat consumption and renal cell cancer risk: a pooled analysis of 13 prospective studies. *J Natl Cancer Inst* **2008,** *100* (23), 1695-706.

18. Setiawan, V. W.; Stram, D. O.; Nomura, A. M.; Kolonel, L. N.; Henderson, B. E., Risk factors for renal cell cancer: the multiethnic cohort. *Am J Epidemiol* **2007,** *166* (8), 932-40.

19. Nielsen, S. M.; Rhodes, L.; Blanco, I.; Chung, W. K.; Eng, C.; Maher, E. R.; Richard, S.; Giles, R. H., Von Hippel-Lindau Disease: Genetics and Role of Genetic Counseling in a Multiple Neoplasia Syndrome. *J Clin Oncol* **2016,** *34* (18), 2172-81.

20. Karami, S.; Schwartz, K.; Purdue, M. P.; Davis, F. G.; Ruterbusch, J. J.; Munuo, S. S.; Wacholder, S.; Graubard, B. I.; Colt, J. S.; Chow, W. H., Family history of cancer and renal cell cancer risk in Caucasians and African Americans. *Br J Cancer* **2010,** *102* (11), 1676-80.

21. Cancer Genome Atlas Research, N., Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **2013,** *499* (7456), 43-9.

22. Gnarra, J. R.; Tory, K.; Weng, Y.; Schmidt, L.; Wei, M. H.; Li, H.; Latif, F.; Liu, S.; Chen, F.; Duh, F. M.; et al., Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat Genet* **1994,** *7* (1), 85-90.

23. Hakimi, A. A.; Pham, C. G.; Hsieh, J. J., A clear picture of renal cell carcinoma. *Nat Genet* **2013,** *45* (8), 849-50.

24. Linehan, W. M.; Srinivasan, R.; Schmidt, L. S., The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol* **2010,** *7* (5), 277-85.

25. Nabi, S.; Kessler, E. R.; Bernard, B.; Flaig, T. W.; Lam, E. T., Renal cell carcinoma: a review of biology and pathophysiology. *F1000Res* **2018,** *7*, 307.

26. Kaelin, W. G., Von Hippel-Lindau disease. *Annu Rev Pathol* **2007,** *2*, 145-73.

27. Kapitsinou, P. P.; Haase, V. H., The VHL tumor suppressor and HIF: insights from genetic studies in mice. *Cell Death Differ* **2008,** *15* (4), 650-9.

28. Pena-Llopis, S.; Vega-Rubin-de-Celis, S.; Liao, A.; Leng, N.; Pavia-Jimenez, A.; Wang, S.; Yamasaki, T.; Zhrebker, L.; Sivanand, S.; Spence, P.; Kinch, L.; Hambuch, T.; Jain, S.; Lotan, Y.; Margulis, V.; Sagalowsky, A. I.; Summerour, P. B.; Kabbani, W.; Wong, S. W.; Grishin, N.; Laurent, M.; Xie, X. J.; Haudenschild, C. D.; Ross, M. T.; Bentley, D. R.; Kapur, P.; Brugarolas, J., BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet* **2012,** *44* (7), 751-9.

29. Hakimi, A. A.; Chen, Y. B.; Wren, J.; Gonen, M.; Abdel-Wahab, O.; Heguy, A.; Liu, H.; Takeda, S.; Tickoo, S. K.; Reuter, V. E.; Voss, M. H.; Motzer, R. J.; Coleman, J. A.; Cheng, E. H.; Russo, P.; Hsieh, J. J., Clinical and pathologic impact of select chromatin-modulating tumor suppressors in clear cell renal cell carcinoma. *Eur Urol* **2013,** *63* (5), 848-54.

30. Sato, Y.; Yoshizato, T.; Shiraishi, Y.; Maekawa, S.; Okuno, Y.; Kamura, T.; Shimamura, T.; Sato-Otsubo, A.; Nagae, G.; Suzuki, H.; Nagata, Y.; Yoshida, K.; Kon, A.; Suzuki, Y.; Chiba, K.; Tanaka, H.; Niida, A.; Fujimoto, A.; Tsunoda, T.; Morikawa, T.; Maeda, D.; Kume, H.; Sugano, S.; Fukayama, M.; Aburatani, H.; Sanada, M.; Miyano, S.;

Homma, Y.; Ogawa, S., Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* **2013,** *45* (8), 860-7.

31. Hakimi, A. A.; Ostrovnaya, I.; Reva, B.; Schultz, N.; Chen, Y. B.; Gonen, M.; Liu, H.; Takeda, S.; Voss, M. H.; Tickoo, S. K.; Reuter, V. E.; Russo, P.; Cheng, E. H.; Sander, C.; Motzer, R. J.; Hsieh, J. J.; cc, R. C. C. C. G. A. R. N. i., Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network. *Clin Cancer Res* **2013,** *19* (12), 3259-67.

32. Kapur, P.; Pena-Llopis, S.; Christie, A.; Zhrebker, L.; Pavia-Jimenez, A.; Rathmell, W. K.; Xie, X. J.; Brugarolas, J., Effects on survival of BAP1 and PBRM1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation. *Lancet Oncol* **2013,** *14* (2), 159-167.

33. Nam, S. J.; Lee, C.; Park, J. H.; Moon, K. C., Decreased PBRM1 expression predicts unfavorable prognosis in patients with clear cell renal cell carcinoma. *Urol Oncol* **2015,** *33* (8), 340 e9-16.

34. Seyfried, T. N.; Flores, R. E.; Poff, A. M.; D'Agostino, D. P., Cancer as a metabolic disease: implications for novel therapeutics. *Carcinogenesis* **2014,** *35* (3), 515-27.

35. Linehan, W. M.; Schmidt, L. S.; Crooks, D. R.; Wei, D.; Srinivasan, R.; Lang, M.; Ricketts, C. J., The Metabolic Basis of Kidney Cancer. *Cancer Discov* **2019,** *9* (8), 1006-1021.

36. Hakimi, A. A.; Reznik, E.; Lee, C. H.; Creighton, C. J.; Brannon, A. R.; Luna, A.; Aksoy, B. A.; Liu, E. M.; Shen, R.; Lee, W.; Chen, Y.; Stirdivant, S. M.; Russo, P.; Chen,

Y. B.; Tickoo, S. K.; Reuter, V. E.; Cheng, E. H.; Sander, C.; Hsieh, J. J., An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancer Cell* **2016,** *29* (1), 104-116.

37. Mullen, A. R.; Wheaton, W. W.; Jin, E. S.; Chen, P. H.; Sullivan, L. B.; Cheng, T.; Yang, Y.; Linehan, W. M.; Chandel, N. S.; DeBerardinis, R. J., Reductive carboxylation supports growth in tumour cells with defective mitochondria. *Nature* **2011,** *481* (7381), 385-8.

38. Kind, T.; Tolstikov, V.; Fiehn, O.; Weiss, R. H., A comprehensive urinary metabolomic approach for identifying kidney cancerr. *Anal Biochem* **2007,** *363* (2), 185-95.

39. Kim, K.; Aronov, P.; Zakharkin, S. O.; Anderson, D.; Perroud, B.; Thompson, I. M.; Weiss, R. H., Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Mol Cell Proteomics* **2009,** *8* (3), 558-70.

40. Kim, K.; Taylor, S. L.; Ganti, S.; Guo, L.; Osier, M. V.; Weiss, R. H., Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. *OMICS* **2011,** *15* (5), 293-303.

41. Monteiro, M. S.; Barros, A. S.; Pinto, J.; Carvalho, M.; Pires-Luis, A. S.; Henrique, R.; Jeronimo, C.; Bastos, M. L.; Gil, A. M.; Guedes de Pinho, P., Nuclear Magnetic Resonance metabolomics reveals an excretory metabolic signature of renal cell carcinoma. *Sci Rep* **2016,** *6*, 37275.

42. Ragone, R.; Sallustio, F.; Piccinonna, S.; Rutigliano, M.; Vanessa, G.; Palazzo, S.; Lucarelli, G.; Ditonno, P.; Battaglia, M.; Fanizzi, F. P.; Schena, F. P., Renal Cell Carcinoma: A Study through NMR-Based Metabolomics Combined with Transcriptomics. *Diseases* **2016,** *4* (1).

43. Monteiro, M.; Moreira, N.; Pinto, J.; Pires-Luis, A. S.; Henrique, R.; Jeronimo, C.; Bastos, M. L.; Gil, A. M.; Carvalho, M.; Guedes de Pinho, P., GC-MS metabolomics-based approach for the identification of a potential VOC-biomarker panel in the urine of renal cell carcinoma patients. *J Cell Mol Med* **2017,** *21* (9), 2092-2105.

44. Niziol, J.; Bonifay, V.; Ossolinski, K.; Ossolinski, T.; Ossolinska, A.; Sunner, J.; Beech, I.; Arendowski, A.; Ruman, T., Metabolomic study of human tissue and urine in clear cell renal carcinoma by LC-HRMS and PLS-DA. *Anal Bioanal Chem* **2018,** *410* (16), 3859-3869.

45. Liu, X.; Zhang, M.; Liu, X.; Sun, H.; Guo, Z.; Tang, X.; Wang, Z.; Li, J.; Li, H.; Sun, W.; Zhang, Y., Urine Metabolomics for Renal Cell Carcinoma (RCC) Prediction: Tryptophan Metabolism as an Important Pathway in RCC. *Front Oncol* **2019,** *9*, 663.

46. Wang, Z.; Liu, X.; Liu, X.; Sun, H.; Guo, Z.; Zheng, G.; Zhang, Y.; Sun, W., UPLC-MS based urine untargeted metabolomic analyses to differentiate bladder cancer from renal cell carcinoma. *BMC Cancer* **2019,** *19* (1), 1195.

47. Zhang, M.; Liu, X.; Liu, X.; Li, H.; Sun, W.; Zhang, Y., A pilot investigation of a urinary metabolic biomarker discovery in renal cell carcinoma. *Int Urol Nephrol* **2020,** *52* (3), 437-446.

48. Nicholson, J. K.; Lindon, J. C., Systems biology: Metabonomics. *Nature* **2008,** *455* (7216), 1054-6.

49. Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.;

Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L., HMDB: the Human Metabolome Database. *Nucleic Acids Res* **2007,** *35* (Database issue), D521-6.

50. Bingol, K., Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods. *High Throughput* **2018,** *7* (2).

51. Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L., Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **2011,** *40* (1), 387-426.

52. Pan, Z.; Raftery, D., Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Anal Bioanal Chem* **2007,** *387* (2), 525-7.

53. Zamboni, N.; Saghatelian, A.; Patti, G. J., Defining the metabolome: size, flux, and regulation. *Mol Cell* **2015,** *58* (4), 699-706.

54. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S., Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **2016,** *44* (D1), D463-70.

55. Pinto, J.; Domingues, M. R.; Galhano, E.; Pita, C.; Almeida Mdo, C.; Carreira, I. M.; Gil, A. M., Human plasma stability during handling and storage: impact on NMR metabolomics. *Analyst* **2014,** *139* (5), 1168-77.

56. Maher, A. D.; Zirah, S. F.; Holmes, E.; Nicholson, J. K., Experimental and analytical variation in human urine in 1H NMR spectroscopy-based metabolic phenotyping studies. *Anal Chem* **2007,** *79* (14), 5204-11.

57. Emwas, A. H., The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol Biol* **2015,** *1277*, 161-93.

58. James S. Nowick; Omid Khakshoor; Mehrnoosh Hashemzadeh; Brower, J. O., DSA: A New Internal Standard for NMR Studies in Aqueous Solution. *Org. Lett.* **2003,** *5* (19), 3511–3513.

59. Guro F.Giskeødegårda; Tom G.Bloembergb; Geert Postma; Beathe Sitter; May-Britt Tessem; Ingrid S.Gribbestad; Tone F.Bathen; M.C.Buydensb, L., Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Analytica Chimica Acta* **2010,** *683* (1), 1-11.

60. Wu, W.; Daszykowski, M.; Walczak, B.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N.; Crowther, D. J.; Gill, R. W.; Lutz, M. W., Peak alignment of urine NMR spectra using fuzzy warping. *J Chem Inf Model* **2006,** *46* (2), 863-75.

61. Niels-Peter Vest Nielsen; Jens Michael Carstensen; Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **1998,** *805* (1-2), 17-35.

62. Savorani, F.; Tomasi, G.; Engelsen, S. B., icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J Magn Reson* **2010,** *202* (2), 190-202.

63. Emwas, A. H.; Saccenti, E.; Gao, X.; McKay, R. T.; Dos Santos, V.; Roy, R.; Wishart, D. S., Recommended strategies for spectral processing and post-processing of 1D (1)H-NMR data of biofluids with a particular focus on urine. *Metabolomics* **2018,** *14* (3), 31.

64. Emwas, A. H.; Roy, R.; McKay, R. T.; Ryan, D.; Brennan, L.; Tenori, L.; Luchinat, C.; Gao, X.; Zeri, A. C.; Gowda, G. A.; Raftery, D.; Steinbeck, C.; Salek, R. M.; Wishart, D. S., Recommendations and Standardization of Biomarker Quantification Using NMR-Based Metabolomics with Particular Focus on Urinary Analysis. *J Proteome Res* **2016,** *15* (2), 360-73.

65. Bollard, M. E.; Stanley, E. G.; Lindon, J. C.; Nicholson, J. K.; Holmes, E., NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR Biomed* **2005,** *18* (3), 143-62.

66. R. J. O. Torgrip; K. M. Åberg; E. Alm; I. Schuppe-Koistinen; Lindberg, J., A note on normalization of biofluid 1D 1H-NMR data. *Metabolomics* **2008,** *4*, 114-121.

67. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H., Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **2006,** *78* (13), 4281-90.

68. Ebbels, T. M.; Lindon, J. C.; Coen, M., Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles. *Methods Mol Biol* **2011,** *708*, 365-88.

69. Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C., Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem* **2006,** *78* (7), 2262-7.

70. Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J., Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal Chem* **2005,** *77* (5), 1282-9.

71. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. *Anal Chem* **2016,** *88* (24), 12411-12418.

72. Wishart, D. S., Computational strategies for metabolite identification in metabolomics. *Bioanalysis* **2009,** *1* (9), 1579-96.

73. Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.; Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I., HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* **2009,** *37* (Database issue), D603-10.

74. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L., BioMagResBank. *Nucleic Acids Res* **2008,** *36* (Database issue), D402-8.

75. Fonville, J. M.; Maher, A. D.; Coen, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K., Evaluation of full-resolution J-resolved 1H NMR projections of biofluids for metabonomics information retrieval and biomarker identification. *Anal Chem* **2010,** *82* (5), 1811-21.

76. Ludwig, C.; Viant, M. R., Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochem Anal* **2010,** *21* (1), 22-32.

77. Dona, A. C.; Kyriakides, M.; Scott, F.; Shephard, E. A.; Varshavi, D.; Veselkov, K.; Everett, J. R., A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. *Comput Struct Biotechnol J* **2016,** *14*, 135-53.

78. Liu, M.; Nicholson, J. K.; Lindon, J. C., High-resolution diffusion and relaxation edited one- and two-dimensional 1H NMR spectroscopy of biological fluids. *Anal Chem* **1996,** *68* (19), 3370-6.

79. Mitchell, T., *Machine Learning*. McGraw Hill: 1997.

80. Noble, W. S., What is a support vector machine? *Nature Biotechnology* **2006,** *24* (12), 1565–1567.

81. Awad M.; R., K., Support Vector Regression. In: Efficient Learning Machines. . Apress, Berkeley, CA. : 2015.

82. Ho, T. K., Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition* **1995,** *1*, 278-282.

83. Breiman, L., Bagging predictors. *Machine Learning* **1996,** *24*, 123–140.

84. Yoav Freund; Schapire, R. E., A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **1997,** *55* (1), 119-139.

85. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**.

86. Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R., A tutorial review: Metabolomics and partial least squares-discriminant

analysis--a marriage of convenience or a shotgun wedding. *Anal Chim Acta* **2015,** *879*, 10-23.

87. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S., Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel)* **2020,** *23* (1).

88. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D., A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* **2018,** *51* (5), 42.

89. Morocho-Cayamcela, M. E.; Lee, H.; Lim, W., Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions. *IEEE Access* **2019,** *7*, 137184-137206.

90. Molnar, C., *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019.

91. Galindo-Prieto, B.; Eriksson, L.; Trygg, J., Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *Journal of Chemometrics* **2014,** *28* (8), 623-632.

92. Nembrini, S.; Konig, I. R.; Wright, M. N., The revival of the Gini importance? *Bioinformatics* **2018,** *34* (21), 3711-3718.

93. Friedman, J. H., Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **2001,** *29* (5), 1189-1232

94. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E., Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *arXiv* **2014**.

95. Worley, B.; Powers, R., Multivariate Analysis in Metabolomics. *Curr Metabolomics* **2013,** *1* (1), 92-107.

96. Ribeiro, M.; Singh, S.; Guestrin, C., "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 1135–1144.

97. Fisher, A.; Rudin, C.; Dominici, F., All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research* **2019,** *20* (177), 1-81.

98. Breiman, L., Random Forests. *Machine Learning* **2001,** *45*, 5–32.

99. Lundberg, S.; Lee, S.-I., A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**.

100. Lundberg, S. M.; Erion, G. G.; Lee, S.-I., Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2018**.

101. Bessonneau, V.; Rudel, R. A., Mapping the Human Exposome to Uncover the Causes of Breast Cancer. *Int J Environ Res Public Health* **2019,** *17* (1).

102. Wild, C. P.; Scalbert, A.; Herceg, Z., Measuring the exposome: a powerful basis for evaluating environmental exposures and cancer risk. *Environ Mol Mutagen* **2013,** *54* (7), 480-99.

CHAPTER 2

MACHINE LEARNING-ENABLED RENAL CELL CARCINOMA STATUS PREDICTION USING MULTI-PLATFORM URINE-BASED METABOLOMICS[1]

[1] Olatomiwa O. Bifarin, David A. Gaul, Samyukta Sah, Rebecca S. Arnold, Kenneth Ogan, Viraj A. Master, David L. Roberts, Sharon H. Bergquist, John A. Petros, Facundo M. Fernández, and Arthur S. Edison. *J. Proteome Res.* 2021, 20, 7, 3629–3641. Reprinted here with permission of publisher.

Contributing Authors

Rebecca S. Arnold, Kenneth Ogan, Viraj A. Master, David L. Roberts, Sharon H. Bergquist, and John A. Petros are the team of scientists, surgeons, and physicians from the departments of Urology and Medicine at Emory University that were involved in the selection of healthy controls and RCC patient cohort, and collection of urine samples. David A. Gaul and Samyukta Sah conducted LC-MS experiments and compound annotations. Olatomiwa O. Bifarin conducted NMR-based metabolomics, data analysis, machine learning experiments, and determined the biological relevance of discriminant metabolites. Facundo M. Fernández, and Arthur S. Edison are senior authors.

2.1 Abstract

Renal cell carcinoma (RCC) is diagnosed through expensive cross-sectional imaging, frequently followed by renal mass biopsy, which is not only invasive but also prone to sampling errors. Hence, there is a critical need for a non-invasive diagnostic assay. RCC exhibits altered cellular metabolism combined with the close proximity of the tumor(s) to the urine in the kidney, suggesting urine metabolomic profiling is an excellent choice for assay development. Here, we acquired liquid chromatography-mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR) data followed by the use of machine learning (ML) to discover candidate metabolomic panels for RCC. The study cohort consisted of 105 RCC patients and 179 controls separated into two sub-cohorts: the model cohort and the test cohort. Univariate, wrapper, and embedded methods were used to select discriminatory features using the model cohort. Three ML techniques, each with different induction biases, were used for training and hyperparameter tuning. Assessment of RCC status prediction was evaluated using the test cohort with the selected biomarkers and the

optimally-tuned ML algorithms. A seven-metabolite panel predicted RCC in the test cohort with 88% accuracy, 94% sensitivity, and 85% specificity, and an AUC of 0.98. Metabolomics Workbench Study IDs: ST001705 and ST001706.

2.2 Introduction

In the United States, kidney cancer is one of the most lethal urinary cancers. In 2021, an estimated 76,080 patients will be diagnosed, with a death toll of 13,780.[1] Approximately 90% of kidney and renal pelvis cancers are renal cell carcinomas (RCC). RCC lacks specific symptoms in the early stages, and the latest statistics indicate that over 50% of patients are diagnosed incidentally.[2-3] Diagnosis is typically performed *via* expensive imaging tests[4-5] and biopsies, the latter being highly invasive and prone to sampling errors.[2, 6-7] Current treatments and early diagnosis, when tumors are localized, results in a 92.6% 5-year survival, while late diagnosis results in the decrease of 5-year survival to 13.0%.[2] An improved, non-invasive and cost-effective diagnostic test is urgently needed to diagnose RCC earlier in the course of the disease.

As early as in the middle ages, physical properties including taste, smell and color of urine were used to diagnose disease, and these properties are influenced by urine metabolites.[8] Today, analytical chemistry platforms such as nuclear magnetic resonance (NMR) spectroscopy and liquid chromatography mass spectrometry (LC-MS) can determine the chemical composition of urine in an high throughput fashion for biomarker discovery and diagnostics.[9-10] The metabolome closeness to the phenotype of biological systems supports its utility to investigate the biology of cancer, which is considered by many to effectively be a metabolic disease.[11-12] Close proximity of RCC tumor(s) to the

urine suggests metabolomic alterations may be ideally detected in this non-invasively collected biofluid.

The high-throughput nature of metabolomics experiments and the broad analyte coverage by both NMR and LC-MS often results in enormous datasets that frequently require machine learning approaches to investigate biological alterations.[13] Machine learning is a branch of artificial intelligence that uses algorithms to uncover patterns in complex data without explicit programming.[14] These models allow for the prediction of output(s) based on a set of inputs, such as the prediction of RCC status using a panel of metabolite abundances selected from the urine feature dataset.

Several previous studies have investigated urine metabolome changes associated with RCC.[15-30] Kim *et al.* found 4-hydroxybenzoate, quinolinate, and gentisate to be differentially expressed at a false discovery rate of 0.28 between RCC (*n*=29) and controls (*n*=33) using ultra high-performance liquid chromatography–mass spectrometry (UHPLC-MS) and gas chromatography–mass spectrometry (GC-MS).[23] Monteiro *et al*. reported a 32-metabolite resonance signature from NMR urine metabolomics that discriminated RCC patients (*n*=42) from controls (*n*=49) using unsupervised learning.[20] Urinary volatile metabolic profiling using GC-MS led to the discovery of a panel of 21 volatile organic compounds correlated with RCC when 30 RCC patients were compared to 37 controls, with 2,5,8-trimethyl-1,2,3,4 tetrahydronaphthalene-1-ol and 2-oxopropanal subsequently validated as potential RCC biomarkers in a small independent sample set.[21] In 2019, Liu *et al*. used LC-MS to identify androstenedione, 7-alpha-hydroxy-3-oxochol-4-en-24-oic acid and lithocholyltaurine to be the most significantly altered metabolites between RCC (n=100) and controls (n=129).[22] In 2020, Zhang *et al*. identified aminoadipic acid, 2-

(formamido)-N1-(5-phospho-d-ribosyl) acetamidine and alpha-N-phenylacetyl-l-glutamine to be predictive of RCC in a cohort of 68 healthy controls, and 39 RCC patients using LC-MS.[15] Unfortunately, none of these highlighted studies made their data widely available, complicating progress in the field.

To improve our understanding of metabolome alterations associated with RCC and to build on prior research conducted in the field, we here report on a multiplatform (NMR + Hydrophilic Interaction Liquid Chromatography (HILIC) LC-MS) metabolomics study on a patient cohort of larger size than most previously-published studies (healthy control = 179, RCC patients = 105). The use of custom-built machine learning models enabled to investigate algorithms with different inductive biases, and hyperparameter tuning. In addition, the dataset was not filtered to retain only endogenous metabolites, therefore allowing for the inclusion of xenobiotics and exposure metabolites such as 2-mercaptobenzothiazole and dibutylamine in the discriminatory panel. We have shown that seven MS-derived metabolites, which discriminated RCC patients from healthy controls with 88% accuracy in the test cohort, could be identified. In addition, four NMR-derived diagnostic markers discriminated RCC patients from healthy controls with an accuracy of 78%. These results underlie the promise of RCC detection using urine metabolomics, providing additional evidence for metabolic perturbations in RCC.

2.3 Materials and Methods

2.3.1 Chemicals

Optima (ThermoFisher Scientific) LC-MS grade water and acetonitrile were used to prepare all mobile phase components. Ammonium acetate (Sigma, molecular biology grade) and ammonium hydroxide 28-30% solution (Fisher Chemical) were used as

additives for mobile phases. For NMR samples, D_2O and 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) were obtained from Cambridge Isotope Laboratories (Andover, MA, USA).

2.3.2 Urine Collection

Patients at Emory University Hospital with a solid renal mass with potential for RCC, and subsequently confirmed to be RCC following surgery were identified prospectively. Healthy controls were identified during an annual physical exam. All patients provided informed consent (Emory University approvals IRB00058903, IRB00054812, IRB00085068, and IRB00055316). Urine was collected at either a clinic appointment or at time of surgery in a urine collection cup and placed on ice. Urine was mixed by turning the cup upside down 5 times and 15 mL were transferred to a sterile tube followed by centrifugation at 1800 g for 20 min at 4 °C. Ten mL of the supernatant were transferred to a clean, sterile tube, and one tablet of Complete Protease Inhibitor Cocktail (Sigma, St. Louis) was added to the tube. The tube was placed on ice for 10 min with periodic vortex mixing to dissolve the tablet. This urine was then transferred into 5 X 1.5 mL aliquots and stored at -80 °C.

2.3.3 Hydrophilic Interaction Liquid Chromatography High Resolution Mass Spectrometry

Urine samples were thawed on ice, and proteins were precipitated with addition of methanol in a 5:1 volume ratio to 50 µL of urine. Samples were vortex-mixed for 30 s and, after centrifugation at 21,100 x g for 5 min, the supernatant was transferred to a snap-on

cap LC vial and stored at 4 °C until analysis. A sample preparation blank was analyzed jointly with the samples, and a pooled sample was created for use as quality control and to correct for instrument drift. Samples were analyzed in randomized order, and the pooled sample was included in approximately every tenth injection over the course of the batch.

Compounds were separated using an Ultimate3000 (ThermoFisher Scientific), fitted with a Waters Acquity UPLC BEH HILIC column (2.1 x 75 mm, 1.7 µm particle size). The compounds were eluted with the following gradient: 95:5 10 mM ammonium acetate with ~0.014% ammonium hydroxide: acetonitrile (mobile phase A) and acetonitrile with ~0.014% ammonium hydroxide (mobile phase B) using the following gradient program: 0 min 5% A; 3 min 63% A; 7 min 63% A; 7.1 min 5% A; 9.9 min 5%. The flow rate was set at 0.30 mL min^{-1} for 0-7.1 min; increased to 0.5 mL min^{-1} from 7.1-7,2 min; 7.2-9.5 min at 0.5 mL min^{-1}; and decreased to 0.30 mL min^{-1} from 9.5 – 10.0 min. The column temperature was set to 50 °C, and the injection volume was 2 µL. A high-resolution accurate mass Q Exactive HF mass spectrometry system (ThermoFisher Scientific) was used for all measurements. The heated electrospray ionization (HESI) source was operated at a capillary temperature of 275 ^{0}C, a spray voltage of 3.5 kV, and sheath, auxiliary, and sweep gas flow rates of 48, 11, and 2 arbitrary units, respectively. MS data were acquired in the 70-1050 *m/z* range in both positive and negative ionization modes. MS/MS experiments were performed by acquiring mass spectra in a data-dependent acquisition fashion. Survey MS were collected with a resolution setting of 120,000 and the top 10 dd-MS^2 were collected at a resolution of 30,000 and an isolation window of 0.4 *m/z*. Stepped normalized collision energies of 10, 30, and 50 fragmented selected precursor ions in the

HCD cell prior to combining all ions for Orbitrap analysis. Dynamic exclusion was set at 10 s and ions with charges greater than 2 were omitted.

Data acquisition and processing were carried out using Xcalibur V4.0 (ThermoFisher Scientific) and Compound Discoverer V3.0 (ThermoFisher Scientific), respectively. Pooled QC injections were used to adjust for instrument drift using a LOESS algorithm. Background peaks were filtered from the dataset when signals were less than 5x of corresponding features in sample blank injections. A feature was filtered if it was present in less than 50% of the QC sample injections or if a relative standard deviation was observed greater than 30% in the QC injections.

Once a panel of discriminant features was selected, additional experiments were conducted with an Orbitrap ID-X Tribrid mass spectrometer (ThermoFisher Scientific) using data dependent acquisition methods to collect MS^2 data for features that were missed during the original DDA data collection. For these experiments, a Waters Acquity UPLC BEH amide column (2.1 x 150mm, 1.7 µm particle size) was used with the following mobile phases: 80:20 10 mM ammonium formate with 0.1% formic acid: acetonitrile (mobile phase A) and acetonitrile with 0.1% formic acid (mobile phase B). The gradient used was as follows: 0 min 5% A; 0.5 min 5% A; 8 min 60% A; 9.4 min 60% A; 11 min 5% A. The flow rate was set to 0.40 mL min^{-1}, the column temperature was set to 40 °C, and the injection volume was 2 µL. Tandem MS spectra were collected for an inclusion list of precursors if they were above an intensity threshold of 6.0E3, using an isolation window of 0.8 m/z. Survey mass spectra were collected with a resolution of 60,000. Stepped normalized collision energies of 15, 30, and 45 fragmented the precursors in the HCD cell,

followed by Orbitrap analysis at a resolution of 30,000. Precursor ions were also sequentially fragmented with a CID collision energy of 45, and analyzed in the ion trap.

Data processing was performed with Compound Discoverer v3.0 (ThermoFisher Scientific), which included elemental formula prediction based on exact masses and isotope patterns. When elemental formula prediction was not achieved in the automated fashion *via* Compound Discoverer, the feature was manually analyzed using Xcalibur v3.0 to assign elemental formula. Tentative annotations were assigned based on searches against literature and metabolomic databases, such as the Human Metabolome Database (HMDB), Metlin, mzCloud, and MassBank. Elemental formulas and exact masses with a mass error of 10mDa were used in this case. Fragmentation patterns were also analyzed and matched against tandem MS databases such as mzCloud and locally-built mzVault libraries in order to assign annotations.

2.3.4 Nuclear Magnetic Resonance Spectroscopy

Urine samples were thawed in a 4 °C cold room followed by centrifugation at 20,200 relative centrifugal force (rcf) for 20 min at 4 °C to remove any precipitated materials. A sample preparation robot (SamplePro, Bruker Biospin, Rheinstetten Germany) was used to dispense 60 µL of NMR buffer into 5 mm SampleJet NMR tubes (Bruker Biospin, Billerica, MA, USA), followed by the transfer of 540 µL of urine sample and sample mixing. The NMR buffer used was 1.5 M KH_2PO_4/K_2HPO_4 buffer with a pH of 7.0 in D_2O, containing 0.11 mM of 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS). DSS is used as a chemical shift reference (0.0 ppm). Quality assurance and quality control for this study is described in appendix A section A-1. NMR spectra were acquired using an Avance III

HD 600 MHz Bruker NMR spectrometer with a Bruker SampleJet cooled to 5.6 °C. The following NMR experiments were conducted: one-dimensional nuclear Overhauser effect pulse sequence with pre-saturation of water resonance (NOESYPR1D), two-dimensional (2D) ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC) and HSQC–TOCSY (HSQC–total correlation spectroscopy). For 1D ^{1}H NMR metabolomics spectra, phase and baseline correction, and referencing were carried out with Bruker's TopSpin software. Referencing to DSS was confirmed using the Edison laboratory in-house MATLAB scripts (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA). In addition, the ends of NMR spectra (less than -0.50 ppm, greater than 10.0 ppm) and water regions (between 4.89 ppm and 4.68 ppm) were removed from all samples. Urine NMR spectra were aligned using constrained correlation optimized warping (CCOW),[31] and normalized using probabilistic quotient normalization (PQN).[32] NMRPipe was used to pre-process the 2D NMR Data (HSQC, and HSQC-TOCSY).[33] Metabolites were identified using the AssureNMR software (Bruker Biospin, USA) with BBiorefcode metabolite database and COLMARm.[34] Metabolites were assigned a confidence score from 1 to 5, with 5 as the highest confidence score. The scores were defined as follows: (1) putatively characterized compound classes or annotated compounds, (2) matches from 1D NMR to literature and/or 1D BBiorefcode compound (AssureNMR) or other database libraries such as BMRB[35] and HMDB,[36] (3) matched to HSQC, (4) matched to HSQC and validated by HSQC–TOCSY (COLMARm), and (5) validated by spiking the authentic compound into the sample. Fifty metabolomic features in the aligned and normalized 1D ^{1}H NMR spectra were quantified by taking spectral areas for integration, and combined with MS features for downstream analysis (See **Appendix A; Section A-1 and scheme A-1** for NMR peak picking and

integration details). Of the 50 metabolomic features quantified *via* NMR, thirty metabolites were identified with some metabolites having multiple resonances quantified, in addition to 11 unknown resonances.

2.3.5 Sample Cohort Selection

Propensity score matching[37] was used to reduce the sample selection bias effect while balancing potential confounders amongst control and RCC patient groups. The covariates considered included: age, gender, BMI, race, and smoking history. The propensity score was computed *via* a logistic regression model using the default parameters of the Scikit-learn[38] linear model module in Python. A one-to-one propensity score matching with the caliper method, which allowed for a maximum distance of 1e-5 between the propensity scores of matched pairs, resulted in the selection of 31 control subjects and 31 subjects with RCC to form the model cohort.

2.3.6 Feature Selection for RCC Prediction

Features were selected using the 62-model cohort. The normalized abundances of the 50 metabolomic features that were quantified by NMR and the 7,097 normalized MS features were merged into one feature table in Python. The combined feature table was subjected to both filtering and wrapper feature selection methods.[39-40] The features were filtered *via* the following sequential criteria: 1) features with greater than 1-fold difference between the two groups were retained; 2) features with a *q*-value lower than 0.05 were retained (*q*-value is defined as the *p*-value obtained from a student *t*-test followed by Benjamini-Hochberg false discovery rate correction[41]); and 3) one of two highly correlated features were

removed, with a Pearson correlation coefficient cut-off of 0.8. The resulting features were auto-scaled prior further feature selection. A recursive feature elimination method under stratified five-fold cross validation conditions was implemented using random forests (RF-RFECV). The Scikit-learn[38] default hyperparameters were used with the number of estimators set to 100 decision trees. In addition, a PLS regression method was applied on the same reduced feature set using the default PLS regression method in the cross-decomposition module in Scikit-learn. For each method, features were ranked based on importance for discriminating RCC patients from healthy controls. The Gini index was used in RF-RFECV, while variable importance in projection (VIP) scores were used in PLS regression. Finally, a voting-based system for potential biomarkers was used; the overlapping features among the top features from each method were selected as the final potential biomarkers. Variants of this method were used for selecting only upregulated biomarkers and NMR biomarkers in the study.

2.3.7 Machine Learning (ML) Methods for RCC Prediction

Random forest (RF), *k*-nearest neighbors (*k*-NN), linear kernel support vector machine (SVM-Lin) and radial basis kernel support vector machines (SVM-RBF) were used for predictions. Optimized hyperparameters for each ML method used the model cohort and the selected metabolite panel. A linear search for a single hyperparameter or a grid search for two (or more) hyperparameters were used under five-fold cross validation conditions. These tuned ML models were used to predict RCC status in the test cohort.

Random Forest

Random forests are a collection of decision trees built using bootstrapped training samples, where decision trees are constructed using a random subset of metabolomic features as candidates for node splitting. The decision tree is an inverted tree starting with the root node at the top of the tree, followed by internal nodes, and finally leaf nodes. The root node and internal nodes are assigned specific metabolomic features, while the leaf nodes indicate the final prediction.

k-Nearest Neighbor

k-NN classifiers are an instance-based learning algorithm that classifies samples *via* the vote of the majority of a *k* (to be defined) closest neighbors. Distance measures considered for determining nearest neighbors during hyperparameter tuning include Euclidean (*E*) and Manhattan (*M*) distances.

$$E = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{2.1}$$

$$M = \sum_{i=1}^{k} |x_i - y_i| \tag{2.2}$$

Linear Kernel Support Vector Machine

For binary classification, the goal of SVM-Lin is to generate a separating hyperplane that separates the classes in a *j*-dimensional space, where *j* is the number of features. Given *n* numbers of training samples $x_1, \ldots x_n \in R^j$ with class membership of $y_1, \ldots, y_n \in (-1, 1)$ where -1 represents controls and 1 represents RCC, the function of the separating hyperplane, defined here as the *RCC metabolic score*, is given by the following:

$$\text{RCC metabolic score} = \beta_0 + \sum_{j=1}^{j} \beta_j x_{ij} \quad (2.3)$$

Where β_0 and β_j are the bias and the weight parameters respectively, determined during training. The class membership of a new observation was defined by the sign of the RCC metabolic score (negative for control and positive for RCC). The function $\beta_0 + \beta x' = 0$ is the separating hyperplane that maximized the margin between the two classes, while the margin is defined as the following:

$$\beta_0 + \beta x' \geq 1, \qquad c = +1 \quad (2.4)$$

$$\beta_0 + \beta x' \leq -1, \qquad c = -1 \quad (2.5)$$

The only hyperparameter to be tuned in the SVM-Lin is the non-negative regularization parameter cost (*C*) which allows for the flexibility of misclassification by the hyperplane margin. *C* in SVM controls the bias-variance tradeoff associated with statistical learning algorithms.

Radial Basis Function Kernel Support Vector Machine

The RBF kernel is a kernel method that projects data in a higher dimensional space for the purpose of a linear separation, which is equivalent to a non-linear decision boundary in the original feature space. SVM-RBF is defined by the following function:

$$K(x_i x_{i'}) = exp\ (-\gamma \sum_{j=1}^{j} (\ x_{ij} - x_{i'j})^2) \quad (2.6)$$

Where x_i are training data, $x_{i'}$ are test data, and γ, gamma is a positive tuning parameter. γ and *C* are the hyperparameters considered for tuning.

See **Appendix A; Section A-2** for model evaluation metrics.

2.3.8 Unsupervised Learning Methods

Hierarchical clustering analysis was conducted on 435 metabolic features, which were the top differential metabolites between RCC and controls, with greater than one-fold change and *q* value lower than 0.05. Of the 435 features, 433 were from LC-MS and 2 features were from NMR. The cluster map function in Seaborn was used.[42] The linkage method for calculating clusters was weighted, while the distance metric was Euclidean. All features were autoscaled prior to analysis.

2.3.9 Data Availability and Implementation Environment

NMR data analysis was carried out using the Edison's Lab in-house MATLAB scripts (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA, Matlab R2017b, The Mathworks, Inc.). Post metabolic features normalization computations were carried out in the Python 3.7.0 programming language using the following packages: Pandas for data handling,[43] Matplotlib/Seaborn for data visualization,[44] Numpy and Scipy for numerical computations,[45-46] Statsmodel for statistical computations,[47] and Sci-kit learn for machine learning.[38] A Jupyter notebook was used as the integrated development environment (IDE).[48] All Jupyter notebooks used in this study can be found here: https://github.com/artedison/RCC_MLprediction. The datasets collected in this work are available through the NIH Metabolomics Workbench[49] with the project ID of PR001091, and study IDs ST001705 and ST001706. The data set can be accessed via http://dx.doi.org/10.21228/M8P97V.

2.4 Results

2.4.1 Patient Selection

NMR measurements were conducted on 179 controls and 105 renal cell carcinoma (RCC) patient urine samples, while LC-MS measurements were conducted on 178 controls and 102 RCC patient urine samples. The subset of controls (n=174) and RCC (n=82) samples that was analyzed by both methods was selected for further investigation. While all control urine samples were collected in the clinic, RCC patient urine samples were collected both in the clinic and in the operating room. Pre-operative procedures added cofounders to the samples collected in the operating room, and were therefore not ideal for use in feature selection due to the potential for introducing bias in the RCC group. However, these operating room samples still had utility as part of the test cohort and were retained. The strategy for grouping of the samples into either model or test cohorts is presented in **Figure 2-1**. In the model cohort used for training purposes, 31 RCC urine samples collected in the clinic were matched via one-to-one propensity score matching (PSM)[37] to 31 control urine samples. PSM seeks to balance the population characteristics of the case vs. control samples in terms of characteristics such as age, BMI and smoking history, and is essential to obtain unbiased machine learning results and robust biomarker panels. In general, the model cohort consists of samples collected in the clinic. As such, features were selected, and models were trained solely using clinic samples. This addresses any sample collection bias concerns, as model training was not carried out using the test cohort. Moreover, all discriminating features identified in the study were statistically insignificant (independent t-Test, BH-FDR $q \geq 0.05$) when RCC samples collected in the clinic *vs*. those collected in the operating room were compared (**Appendix A; Figure A-1**).

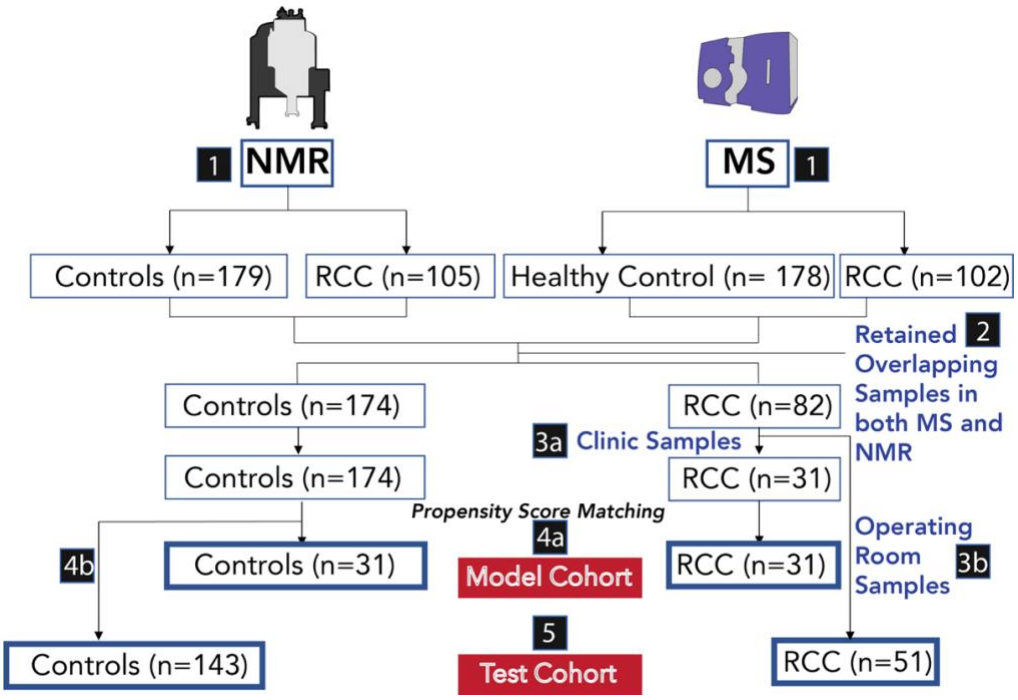


Figure 2-1. Flow Chart for Patient Selection.

Samples for which NMR and MS measurements were collected **(1)**. A total of 284 samples, with 174 control individuals and 82 RCC patients have their urine samples analyzed by both NMR and LC-MS methods **(2).** RCC samples collected in the clinic are selected for the model cohort **(3a)**, while the operating room RCC samples are selected for the test cohort **(3b).** The model cohort was selected *via* propensity score matching from those samples collected in the clinic (31 RCC samples: 31 control samples) **(4)**. The test cohort contained 51 RCC samples collected in the operating room and 143 controls collected in the clinic **(5)**.

Figure 2-2a and **Appendix A; Table A-1** shows the comparative statistics of the pre-PSM and post-PSM model cohorts. Adjusted co-variates included gender, age, BMI, race, and smoking history. Following PSM, the cohorts were gender matched (17 males, 14 females), and had statistically-insignificant differences in age (*p*-value=0.64) and BMI (*p*-

value=0.06). Smoking history and race statistics also improved considerably when compared to the pre-matched cohort. In addition, all RCC stages were represented in the model cohort, early stage RCC (Stage I and II) represented 55% of the cohort, while late stage RCC (Stage III and IV) represented 45% (**Figure 2-2b, Appendix A; Table A-2**). The second sub-cohort in the study, the test cohort, was constructed from the remainder of the samples following removal of the model cohort. It was composed of 143 controls and 51 RCC patients (**Figure 2-2c** and **Appendix A; Table A-3**.) The imbalance of gender, age, BMI, and smoking history in the test cohort made it a good candidate for a challenging test of the utility of the metabolic panel selected by modeling the PSM-adjusted model cohort.

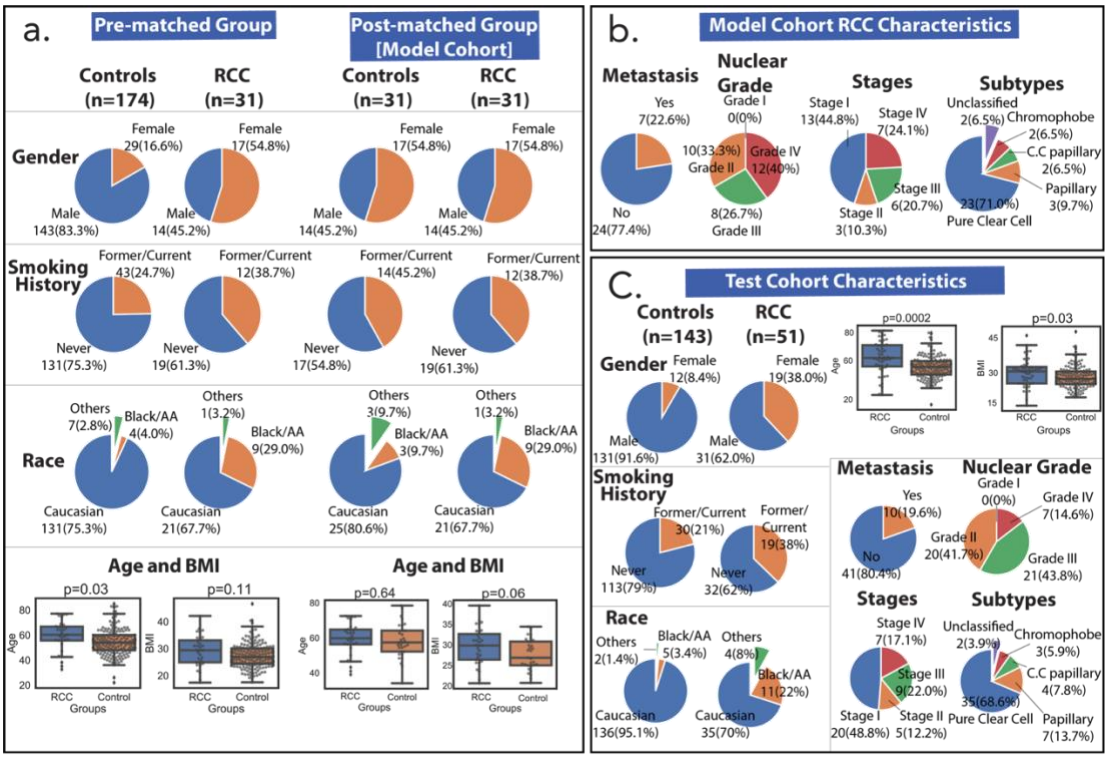


Figure 2-2. Study Cohort Characteristics

(a) Model cohort characteristics (gender, smoking history, race, age, BMI in no particular order) are shown before and after propensity matching. *p*-values were calculated for unequal and equal sample sizes using Welch and Student *t*-tests, respectively. **(b)** Additional model cohort RCC characteristics (metastasis, nuclear grade, stage, and RCC subtype) show the majority of the group was early stage RCC and pure clear cell subtype. There was one nuclear grade datum unreported, and two cancer stages that were not reported due to inconclusive TNM staging information. **(c)** Test cohort characteristics show differences useful in testing the feature panels selected using the model cohort. All *p*-values were calculated using the Welch t-test (unequal sample size). Three samples had unreported nuclear grades and ten samples did not have RCC staging due to inconclusive TNM staging. Abbreviations. AA: African American; BMI: Body Mass Index; RCC: Renal Cell Carcinoma; C.C. Papillary: Clear cell papillary

2.4.2 Metabolomics Analysis and Machine Learning Pipeline

After NMR data collection, ends of spectra and water regions were removed. Several alignment methods were attempted with CCOW[31] giving the most reliable alignment, followed by data normalization. A total of fifty metabolic features were quantitated with NMR and 30 metabolites confirmed with ^{1}H NMR, and/or HSQC and HSQC-TOCSY as described in Materials and Methods (**Appendix A; Table A-4** and **Figure 2-3a**). A total of 7,097 features were detected with LC-MS (4623 from positive mode and 2474 from negative mode), as described under Materials and Methods (**Figure 2-3b** and **c**).

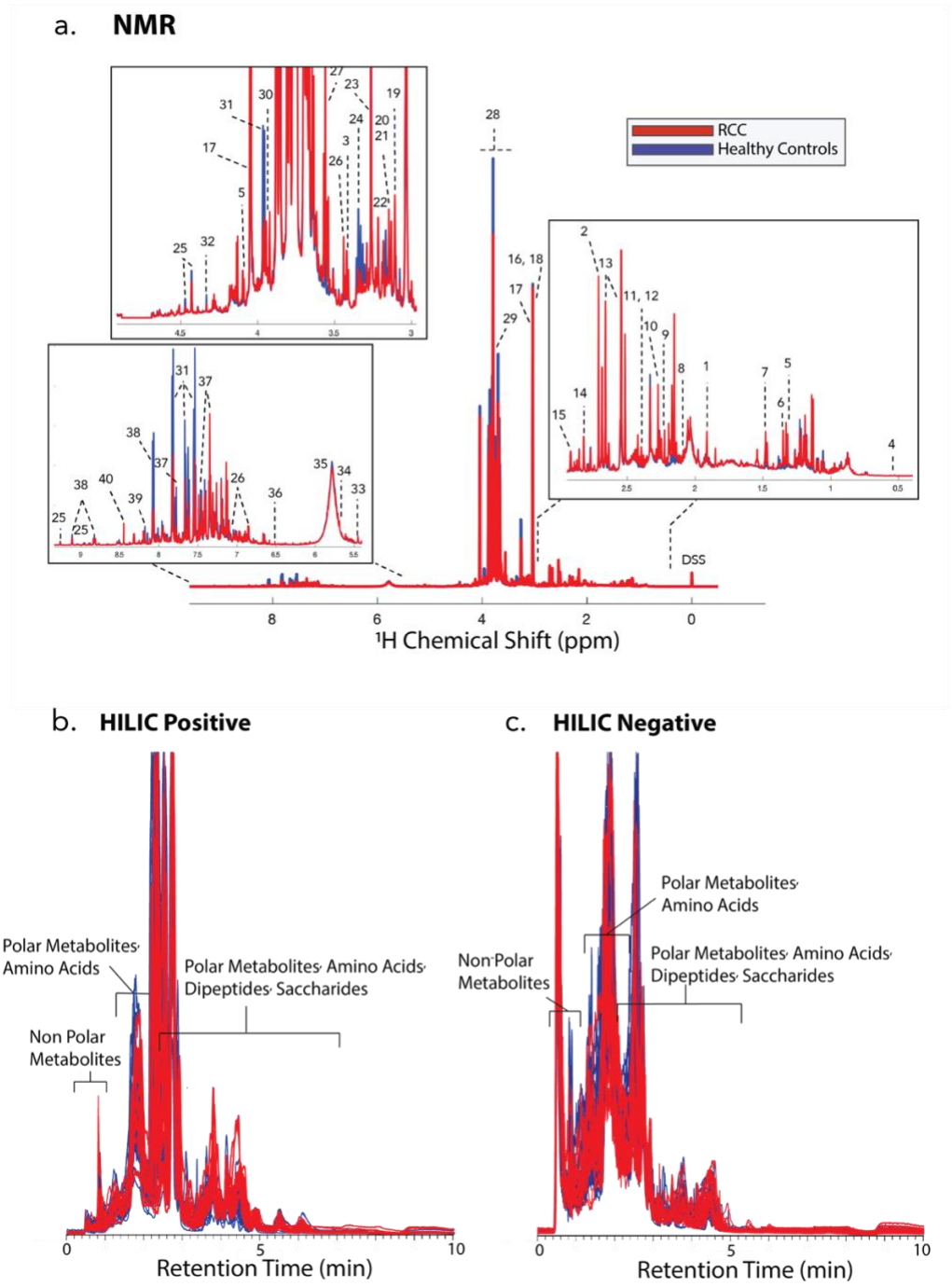


Figure 2-3. Raw Data for Various Metabolomics Platforms.

(a) Average 600MHz ^{1}H 1D NOESY-PR NMR spectra of all urine samples tested in the study. 1) Acetate 2) dimethylamine (DMA) 3) taurine 4) bile acid (tentative assignment) 5) lactate 6) alpha-hydroxyisobutyrate (HIBA) 7) alanine 8) acetyl phosphate 9) acetone

10) acetoacetate 11) succinate 12) pyruvate 13) citrate 14) methylguanidine 15) N,N dimethylglycine (DMG) 16) creatine 17) creatinine 18) creatine phosphate 19) cis-aconitate 20) dimethylsulfone (DMS) 21) ethanolamine 22) choline 23) betaine 24) syllo-inositol 25) trigonellinamide 26) 4-hydroxyphenylacetate (4-HPA) 27) glycine 28) mannitol 29) guadinoacetate 30) glycolate 31) hippurate 32) tatrate 33) allantoin 34) cis-aconitate 35) urea 36) fumarate 37) indoxyl sulfate 38) trigonelline 39) hypoxanthine 40) formate 41) 3-hydroxyisovaleric acid 42) 4-aminohippuric acid 43) 4-hydroxyhippuric acid 44) valine **(b)** HILIC LC-MS positive ion mode data, displaying all samples. **(c)** HILIC LC-MS negative ion mode data, displaying all samples.

All 7,147 metabolomic features from both platforms were merged and data analysis proceeded according to the ML pipeline shown in **Figure 2-4**. The dataset was filtered to include 435 features with greater than one-fold change between RCC and controls, and a Student's *t*-test with Benjamini-Hochberg false discovery rate correction ($q<0.05$) performed. **Figure 2-5** shows the hierarchical clustering of the 435 features selected in this analysis. To minimize the effect of feature multicollinearity, one out of a pair of highly correlated features (Pearson correlation, $r > 0.8$) was retained resulting in 128 features for further analysis. The top 20 features with PLS-DA ranked by VIP scores, and the top 20 features with RF-RFECV ranked by Gini Index were selected from this set of 128 features. Ten features were present on both feature lists selected by PLS-DA and RF-RFECV, leading to the 10-metabolite panel. **(Appendix A; Table A-5 and Figure A-2).** This voting strategy was used to minimize bias from using only one machine learning algorithm for feature selection. Also, as a way of comparison with a more conventional workflow that

relies less on machine learning, features with the top ten highest *q*-values from the univariate analysis were selected, and a classification task was performed for the model cohort with logistic regression using the Metaboanalyst 5.0 biomarker analysis platform. Classification results showed an AUC of 0.86 and an accuracy of 83.3% (**Appendix A; Figure A-3**). These were lower performance scores compared to the ten features selected via the voting-based feature selection methods and employed in the k-NN classifier (0.96 AUC and 95% accuracy) for the model cohort (**Appendix A; Table A-7**). As such, we proceeded with the vote-based ML-derived features.

For predicting RCC status, four machine learning (ML) algorithms were used: random forest (RF), k-nearest neighbor (k-NN), support vector machine with radial basis function (SVM-RBF) and the linear kernel support vector machine (SVM-Lin). Selected hyperparameters were tuned using the 62-model cohort under 5-fold cross validation conditions (**Appendix A; Table A-6**). The tuned ML models were then used to predict RCC status in the test cohort. Overall, k-NN gave the best prediction with an AUC of 0.96, 87% accuracy, 83% specificity, and a sensitivity of 96% (**Appendix A; Table A-7**).

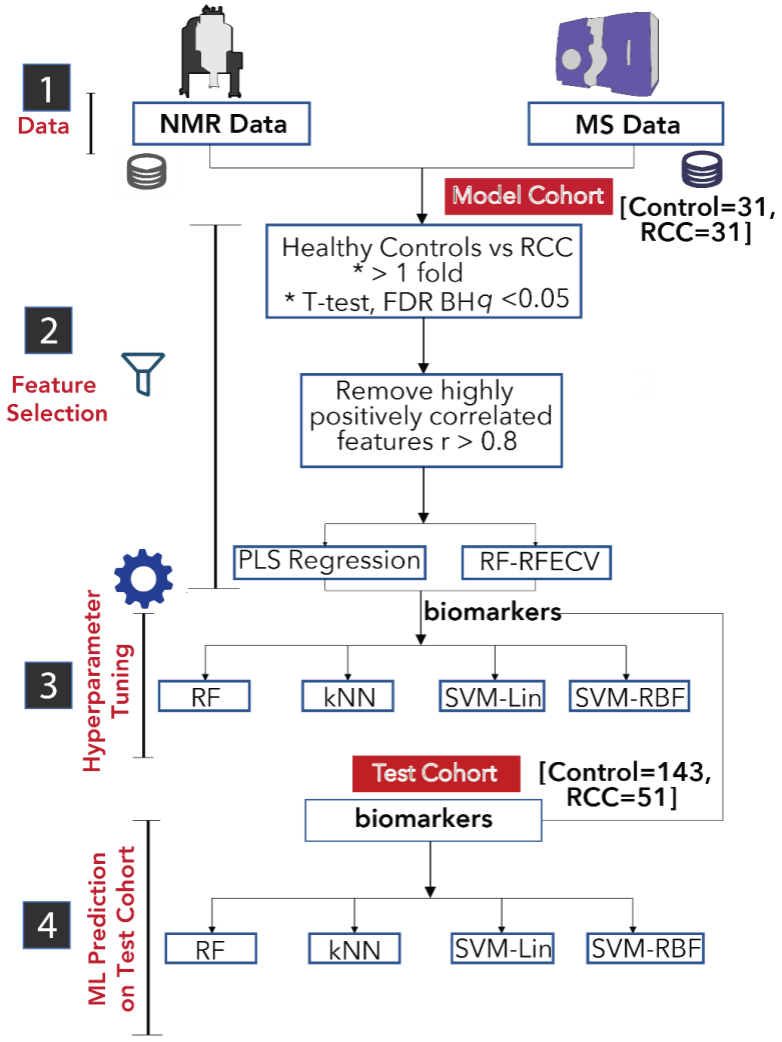


Figure 2-4. Machine Learning Pipeline for RCC Detection Biomarker.

Using the model cohort, a hybrid method of feature selection resulted in a panel of ten metabolites. Hyperparameters for four different machine learning models were tuned using the model cohort and the 10-metabolite panel. The RCC status of the test cohort was predicted with four models. **PLS:** partial least squares; **RF- RFECV:** random forest recursive feature elimination – cross validation; **FDR-BH:** false discovery rate Benjamini Hochberg procedure; ***k*-NN:** *k*-nearest neighbors; and **SVM:** support vector machines (Lin: linear, RBF: radial basis function).

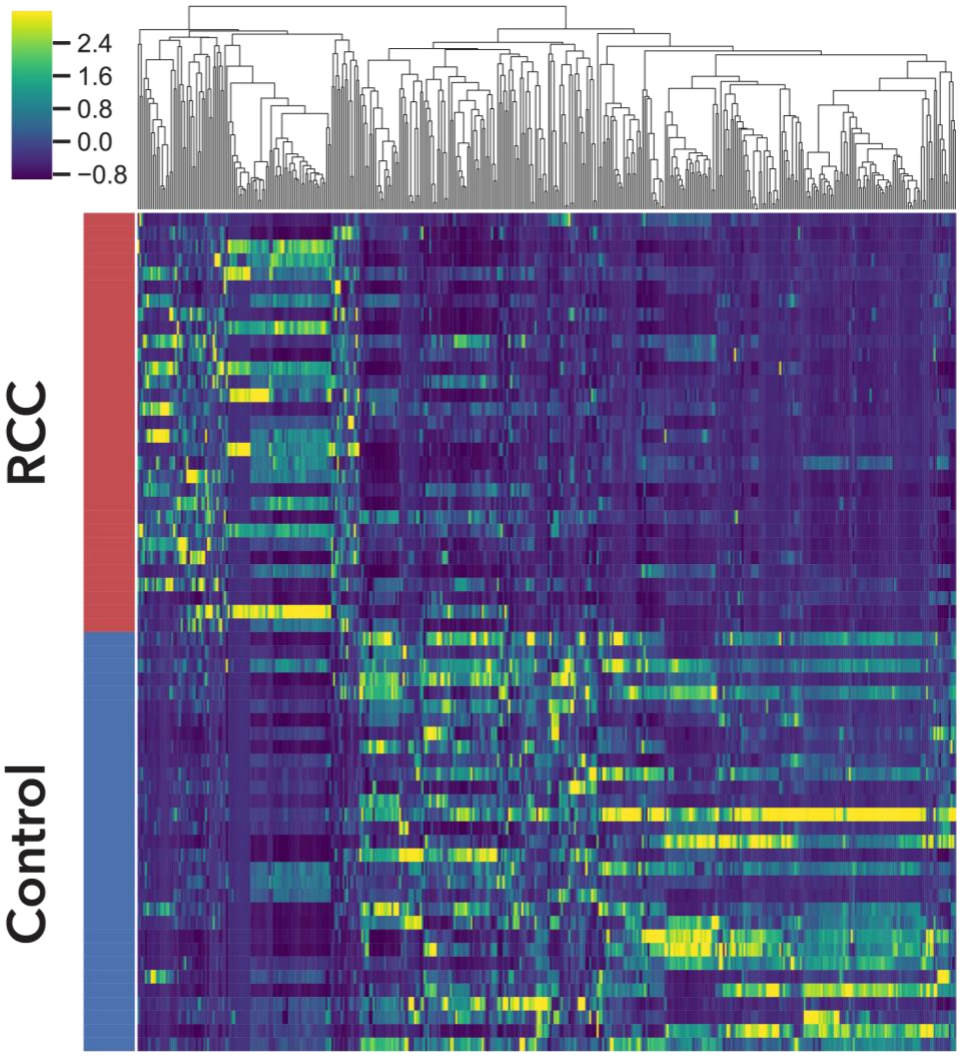


Figure 2-5. Hierarchical Clustering of Top Differential Metabolites.

435 metabolomic features with *q* values < 0.05 and > 1-fold change in the model cohort. z-Scores are represented as shown in the color bar. Yellow represents higher abundances in RCC, while dark blue represents higher abundances in the controls. See **Appendix A**; **Table A-15** for details of metabolomic features.

Eight of the 10 selected markers were in lower relative abundance in RCC samples (**Appendix A; Figure A-2**) *vs.* control samples, so we identified another panel containing features with higher relative abundance in the RCC patients' urine versus control urine, as measuring increased abundance upon appearance of disease is favored in clinical practice. **Figure A-4** describes the machine learning pipeline for upregulated metabolic features in RCC which resulted in a five-metabolite panel (**Appendix A; Table A-8 and Figure A-5**). Again, selected hyperparameters were tuned using the 62-model cohort under 5-fold cross validation conditions (**Appendix A; Table A-9**). The tuned ML models were then used to predict RCC status in the test cohort. It was found that k-NN yielded the best prediction of the test cohort with an AUC of 0.92, an accuracy of 81%, sensitivity of 86%, and specificity of 79% (**Appendix A; Table A-10**), which was a slightly lower performance than for the 10-metabolite panel (**Appendix A; Table A-7**).

High resolution MS and tandem MS experiments were performed for metabolite annotation. Through standard procedures such as analyzing exact masses, isotopic relative ion abundances and fragmentation patterns, five metabolites in the ten-metabolite panel (**Appendix A; Table A-5**) and four of the five in the upregulated metabolite panel (**Appendix A; Table A-8**) were annotated. A third metabolite panel was formed to include only annotated features from the first two panels. **Table 2-1** and **Figure 2-6** shows the results using this last panel, namely a 7-metabolite panel for RCC which included 2-phenylacetamide, Lys-Ile (or Lys-leu), dibutylamine, hippuric acid, mannitol hippurate, 2-mercaptobenzothiazole, and N-acetyl-glucosaminic acid (**Appendix A; Table A-11**). ML hyperparameters were tuned using the 62-model cohort as described above (**Appendix A; Table A-12**). ML models were used to predict RCC status in the test cohort with the most

accurate model being linear SVM with an AUC of 0.98, accuracy of 88%, sensitivity of 94%, and a specificity of 85% (**Table 2-2**).

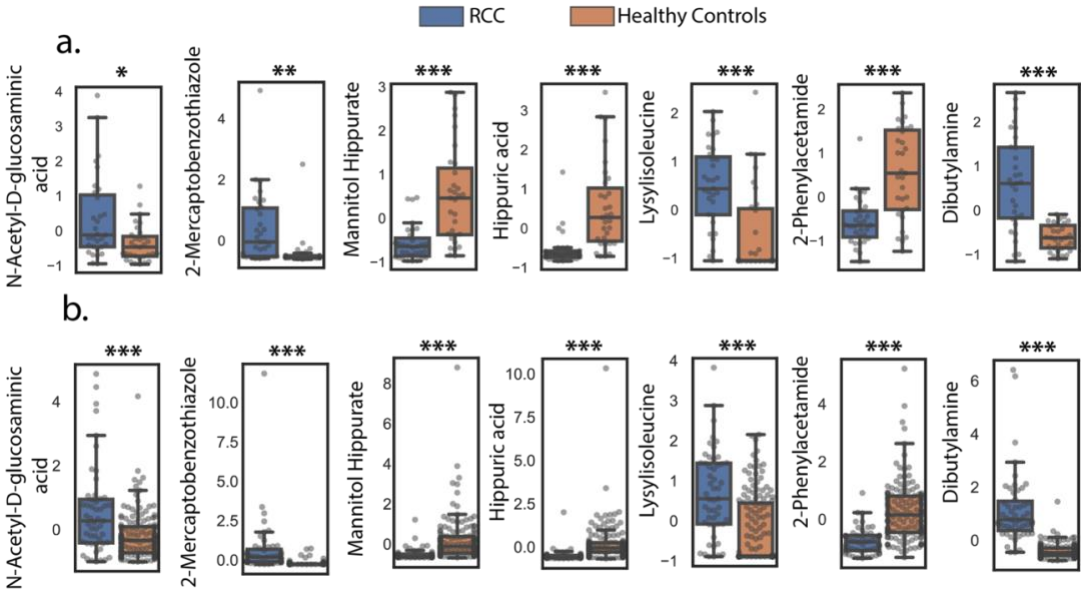


Figure 2-6. Relative Abundances for the 7 Metabolite-panel for RCC Detection.

(a) In the model cohort. After selecting features with greater than one-fold changes between controls and RCC groups, *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test. ($* q \leq 0.05, ** q \leq 0.01, *** q \leq 0.001$) (b) Relative abundances in the test cohort, *p*-values from the Welch *t*-test were reported (unequal sample size). ($* p \leq 0.05, ** p \leq 0.01, *** p \leq 0.001$). Raw data were transformed *via* autoscaling for visualization.

Table 2-1. Compound Annotation and Identification for the 7-Metabolite Panel for RCC Detection.

ID no.	Retention Time (min)	m/z		Adduct Type	Mass error (ppm)	Elemental Formula	Name
		Theoretical	Experimental				
720	5.68	136.0757	136.0755	$[M+H]^+$	-1.47	C_8H_9NO	2-Phenylacetamide

1481	8.83	260.1969	260.1969	$[M+H]^+$	0.00	$C_{12}H_{25}N_3O_3$	Lys-Ile or Lys-leu
2102	4.39	130.1590	130.1591	$[M+H]^+$	0.77	$C_8H_{19}N$	Dibutylamine (alkyl chain branching not determined, isomers possible)
3804	2.59	202.0475	202.0478	$[M+Na]^+$	1.48	$C_9H_9NO_3Na$	Hippuric acid
6262	2.67	376.1249, 358.1143	376.1246, 358.1147	$[M+H_2O-H]^-$ $[M-H]$	-0.68	$C_{15}H_{21}NO_9$	Hippurate-mannitol derivative
6578	1.09	165.9790	165.9784	$[M-H]^-$	-3.61	$C_7H_5NS_2$	2-Mercaptobenzothiazole
6594	6.89	236.0776	236.0777	$[M+H]^+$	0.42	$C_8H_{15}NO_7$	N-acetyl-glucosaminic acid

Table 2-2. Machine Learning Performance for the 7-Metabolite Biomarker Panel for RCC Detection.

Algorithm	**RF**	**K-NN**	**SVM-RBF**	**Linear SVM**
AUC	0.96 +/- 0.04 (0.99)	0.96 +/- 0.05 (0.94)	0.97 +/- 0.04 (0.94)	0.97 +/- 0.04 (**0.98**)
Accuracy	0.9 +/- 0.06 (87%)	0.92 +/- 0.07 (80%)	0.88 +/- 0.09 (78%)	0.87 +/- 0.1 (**88%**)
Sensitivity	0.87 +/- 0.12 (100%)	0.83 +/- 0.15 (92%)	0.8 +/- 0.19 (90%)	0.77 +/- 0.23 (**94%**)
Specificity	0.93 +/- 0.08 (83%)	1.0 +/- 0.0 (76%)	0.97 +/- 0.07 (73%)	0.97 +/- 0.07 **(85%)**

When combined with the 7097 LC-MS features, the 50 NMR features were not selected by machine learning procedures in any of the final panels. This is likely caused by the over-representation of MS features in the final feature list. To further investigate the utility of NMR features, the dataset was filtered with Student's *t*-test with Benjamini-Hochberg false discovery rate correction (q<0.05). Following that, metabolomic features representing the same metabolites were removed *via* a Pearson's correlation cut-off of 0.80 to retain only one feature representing a metabolite (**Appendix A**; **Figure A-6**). This gave rise to a four-metabolite panel consisting of hippurate, trigonellinamide, lactate, and mannitol **(Appendix A; Figure A-7**). As with other panels, selected hyperparameters were tuned using the 62-model cohort under 5-fold cross validation conditions (**Appendix A; Table A-13**). The tuned ML models were then used to predict RCC status in the test cohort. SVM-RBF yielded the best prediction in the test cohort with an AUC of 0.89, an accuracy of 78%, 86% sensitivity, and a specificity of 76% (**Appendix A; Table A-14**).

2.5 Discussion

Machine learning enabled the accurate selection of metabolite markers that accurately distinguished urine samples from RCC patients to those from controls following propensity score matching of the cohorts. Because different machine learning techniques are driven by different induction biases, we used a variety of feature selection strategies to better down-select biomarkers. As initial feature filters, univariate statistical methods such as *t*-tests, fold changes, and Pearson's correlations were used for down-sizing the metabolic feature set. The last few steps of the machine learning pipeline were based on two ML methods with differing inductive biases. PLS-DA assumes linear statistical relationships[50] while random forests can model more complex relationships in the dataset.[51] This step was

followed by voting for the top ranking overlapping metabolic features from the different methods tested. For the classification tasks, hyperparameter tuning of machine learning algorithms was carried out, culminating in excellent predictions of the test cohorts. These data analysis pipelines resulted in a ten-metabolite panel, a five-metabolite panel including only metabolites upregulated in RCC, and a four-metabolite marker containing only metabolites detected by NMR. The seven-identified metabolites biomarker proposals in the study gave an accuracy of 88% and an AUC of 0.98. This is likely a conservative assessment of the robustness of the biomarker given the small size of the training dataset *vs.* a relatively large test cohort, given the constraint of patient selection. In general, many of the markers identified in these panels were novel, but a handful had already been reported in the literature, validating the approach used in this study.

Examination of the biological role of the metabolites in the various panels constructed led to new insights into potential origins and mechanisms of disease progression in RCC. The metabolite 2-phenylacetamide decreased in RCC urine samples, indicating a downregulation of phenylalanine metabolism. Indeed, downregulation of phenylalanine metabolism has been reported in RCC cancer cells,[52] while RCC urine metabolomics studies have also reported the downregulation of metabolites in the phenylalanine pathway such as 4-hydroxyphenylacetate and phenylacetyl-L-glutamine.[15, 20]

The dipeptide lysyl-isoleucine/lysyl-leucine (Lys-Ile/Lys-leu) was observed to be increased in RCC urine samples. Upregulation of other types of dipeptides has been linked to RCC.[22, 53] For example, in a paired normal/clear cell renal cell carcinoma tissue metabolomics study by Hakimi and co-workers, numerous dipeptides were detected as

being upregulated in RCC.[53] In addition, dipeptides such as aspartyl-phenylalanine and glutamyl-threonine have been reported to be upregulated in a urine RCC metabolomics study.[22] Increased dipeptide abundances are typically associated with the increased protein degradation/reutilization processes in tumors.[53]

Reduced levels of hippuric acid and feature $C_{15}H_{21}NO_9$, likely a hippurate and mannitol derivative, in RCC patient urine was in line with the disrupted renal function that arises as a result of a disease, which may lead to the disruption of hippurate elimination or production.[54] Hippurate is formed *via* the conjugation of glycine and benzoic acid, which takes place in the kidney, and this metabolite has been reported to have a strong association with diet and the gut microbiota.[54] Reduced levels of hippurate in RCC patient urine were also reported in studies with smaller cohorts.[20, 28] In addition, reduced level of hippurate have been reported in several RCC-predisposing conditions such as obesity[55-56] and high blood pressure.[57]

N-acetyl-D-glucosaminic acid, an acylaminosugar, was elevated in RCC in our study. Increased glucose uptake might be driving the elevation of the acylaminosugar *via* the hexosamine biosynthetic pathway (HBP). This increased HBP flux has been implicated in many cancer types[58-62] as this pathway plays a central role in DNA repair, cellular signaling, and metastasis.[63]

In addition to endogenous metabolites, two exogenous metabolites were also selected as markers, 2-mercaptobenzothiazole (2-MBT) and dibutylamine. 2-MBT was found at higher levels in RCC patients' urine. 2-MBT is used in acceleration of vulcanization, as such it can be found in car tires. Other commodities that might contain 2-MBT include cables, rubber gloves, shoes, rubber bands and toys.[64-65] Humans are exposed

to 2-MBT *via* inhalation, dermal or oral intake, and the compound has been detected in human urine.[64-65] It has been identified as a marker for traffic intensity because of the tire tread wear linked to car usage, and calls were made for the revision of the risk assessment to 2-MBT.[66] The International Agency for Research has classified it as 'probably carcinogenic to humans',[66] while it has also been linked to an increase in the bladder cancer risk.[65]

Higher levels of dibutylamine (DBA) were also present in RCC patient urine. Dibutylamine is a precursor to N-nitrosodibutylamine (NDBA), a nitrosamine.[67] Nitrosamines are environmental carcinogens that can produce tumors in many organs in the body,[68] NBDA being one of the most potent bladder cancer carcinogens.[69] Increased human urinary excretion of nitrosamines, including NBDA, has also been associated with esophageal cancer.[70-71] Sources of amines and nitrosamines include drinking water[67] and meat products.[72-73]

Hippuric acid, lactate, trigonellinamide, and mannitol were selected as markers in the NMR-only panel. Hippuric acid was also selected in the 10-metabolite panel, making the selection in the NMR-only panel unsurprising. The reduction in abundance of trigonellinamide (1-methylnicotinamide) in RCC patient urine could be indicative of a dysregulated nicotinate and nicotinamide metabolism,[74] particularly considering that our study also identified reduced level of trigonelline, a metabolite that showed a similar trend in a separate NMR study.[20] Increased levels of lactate might reflect the activation of oncogenic aerobic glycolysis, the Warburg effect, which is a hallmark of cancerous cells.[75] In addition, upregulation of lactate dehydrogenase A levels has been reported in RCC cells and tissues.[76] Decreased mannitol excretion in RCC patients might be caused by

dysregulations in energy metabolism, with this trend being reported in a separate urine metabolomics study.[29]

2.6 Conclusions

We have shown the potential utility of a urine assay in the clinical setting for RCC detection. This study, like others of its kind, has the limitation of numerous potential confounders that could impact biomarker discovery results. While randomized control trials (RCTs) are gold standards for epidemiology research, observational studies remain inescapable for studies like this, as randomizing the intervention (RCC) is impossible. As such, to argue for the reduction in selection bias, we adjusted for five potential confounders in the study: age, BMI, gender, smoking history, and race. Of these, four adjustments were largely successful. Going forward, a much larger cohort, representing the diversity of race and geographical locations would be required for the validation of our biomarker proposals.

2.7 References

1. Siegel, R. L.; Miller, K. D.; Fuchs, H. E.; Jemal, A., Cancer Statistics, 2021. *CA Cancer J Clin* **2021,** *71* (1), 7-33.

2. Escudier, B.; Porta, C.; Schmidinger, M.; Rioux-Leclercq, N.; Bex, A.; Khoo, V.; Grunwald, V.; Gillessen, S.; Horwich, A.; clinicalguidelines@esmo.org, E. G. C. E. a., Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* **2019,** *30* (5), 706-720.

3. Gray, R. E.; Harris, G. T., Renal Cell Carcinoma: Diagnosis and Management. *Am Fam Physician* **2019,** *99* (3), 179-184.

4. Diaz de Leon, A.; Pedrosa, I., Imaging and Screening of Kidney Cancer. *Radiol Clin North Am* **2017,** *55* (6), 1235-1250.

5. Kang, S. K.; Chandarana, H., Contemporary imaging of the renal mass. *Urol Clin North Am* **2012,** *39* (2), 161-70, vi.

6. Patel, H. D.; Johnson, M. H.; Pierorazio, P. M.; Sozio, S. M.; Sharma, R.; Iyoha, E.; Bass, E. B.; Allaf, M. E., Diagnostic Accuracy and Risks of Biopsy in the Diagnosis of a Renal Mass Suspicious for Localized Renal Cell Carcinoma: Systematic Review of the Literature. *J Urol* **2016,** *195* (5), 1340-1347.

7. Haifler, M.; Kutikov, A., Update on Renal Mass Biopsy. *Curr Urol Rep* **2017,** *18* (4), 28.

8. Nicholson, J. K.; Lindon, J. C., Systems biology: Metabonomics. *Nature* **2008,** *455* (7216), 1054-6.

9. Zhang, A.; Sun, H.; Wu, X.; Wang, X., Urine metabolomics. *Clin Chim Acta* **2012,** *414*, 65-9.

10. Fiehn, O., Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp Funct Genomics* **2001,** *2* (3), 155-68.

11. Seyfried, T. N.; Shelton, L. M., Cancer as a metabolic disease. *Nutr Metab* **2010,** *7* (1), 7.

12. Linehan, W. M.; Srinivasan, R.; Schmidt, L. S., The genetic basis of kidney cancer: a metabolic disease. *Nat Rev Urol* **2010,** *7* (5), 277-85.

13. Xu, C.; Jackson, S. A., Machine learning and complex biological data. *Genome Biol* **2019,** *20* (1), 76.

14. Mitchell, T. M., *Machine Learning*. McGraw-Hill: New York, 1997; p xvii, 414 p.

15. Zhang, M.; Liu, X.; Liu, X.; Li, H.; Sun, W.; Zhang, Y., A pilot investigation of a urinary metabolic biomarker discovery in renal cell carcinoma. *Int Urol Nephrol* **2020,** *52* (3), 437-446.

16. Wang, Z.; Liu, X.; Liu, X.; Sun, H.; Guo, Z.; Zheng, G.; Zhang, Y.; Sun, W., UPLC-MS based urine untargeted metabolomic analyses to differentiate bladder cancer from renal cell carcinoma. *BMC Cancer* **2019,** *19* (1), 1195.

17. Rodrigues, D.; Monteiro, M.; Jeronimo, C.; Henrique, R.; Belo, L.; Bastos, M. L.; Guedes de Pinho, P.; Carvalho, M., Renal cell carcinoma: a critical analysis of metabolomic biomarkers emerging from current model systems. *Transl Res* **2017,** *180*, 1-11.

18. Oto, J.; Fernandez-Pardo, A.; Roca, M.; Plana, E.; Solmoirago, M. J.; Sanchez-Gonzalez, J. V.; Vera-Donoso, C. D.; Martinez-Sarmiento, M.; Espana, F.; Navarro, S.; Medina, P., Urine metabolomic analysis in clear cell and papillary renal cell carcinoma: A pilot study. *J Proteomics* **2020,** *218*, 103723.

19. Niziol, J.; Bonifay, V.; Ossolinski, K.; Ossolinski, T.; Ossolinska, A.; Sunner, J.; Beech, I.; Arendowski, A.; Ruman, T., Metabolomic study of human tissue and urine in clear cell renal carcinoma by LC-HRMS and PLS-DA. *Anal Bioanal Chem* **2018,** *410* (16), 3859-3869.

20. Monteiro, M. S.; Barros, A. S.; Pinto, J.; Carvalho, M.; Pires-Luis, A. S.; Henrique, R.; Jeronimo, C.; Bastos, M. L.; Gil, A. M.; Guedes de Pinho, P., Nuclear Magnetic Resonance metabolomics reveals an excretory metabolic signature of renal cell carcinoma. *Sci Rep* **2016,** *6*, 37275.

21. Monteiro, M.; Moreira, N.; Pinto, J.; Pires-Luis, A. S.; Henrique, R.; Jeronimo, C.; Bastos, M. L.; Gil, A. M.; Carvalho, M.; Guedes de Pinho, P., GC-MS metabolomics-based approach for the identification of a potential VOC-biomarker panel in the urine of renal cell carcinoma patients. *J Cell Mol Med* **2017,** *21* (9), 2092-2105.

22. Liu, X.; Zhang, M.; Liu, X.; Sun, H.; Guo, Z.; Tang, X.; Wang, Z.; Li, J.; Li, H.; Sun, W.; Zhang, Y., Urine Metabolomics for Renal Cell Carcinoma (RCC) Prediction: Tryptophan Metabolism as an Important Pathway in RCC. *Front Oncol* **2019,** *9*, 663.

23. Kim, K.; Taylor, S. L.; Ganti, S.; Guo, L.; Osier, M. V.; Weiss, R. H., Urine metabolomic analysis identifies potential biomarkers and pathogenic pathways in kidney cancer. *OMICS* **2011,** *15* (5), 293-303.

24. Kim, K.; Aronov, P.; Zakharkin, S. O.; Anderson, D.; Perroud, B.; Thompson, I. M.; Weiss, R. H., Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Mol Cell Proteomics* **2009,** *8* (3), 558-70.

25. Ganti, S.; Weiss, R. H., Urine metabolomics for kidney cancer detection and biomarker discovery. *Urol Oncol* **2011,** *29* (5), 551-7.

26. Ganti, S.; Taylor, S. L.; Abu Aboud, O.; Yang, J.; Evans, C.; Osier, M. V.; Alexander, D. C.; Kim, K.; Weiss, R. H., Kidney tumor biomarkers revealed by simultaneous multiple matrix metabolomics analysis. *Cancer Res* **2012,** *72* (14), 3471-9.

27. Kind, T.; Tolstikov, V.; Fiehn, O.; Weiss, R. H., A comprehensive urinary metabolomic approach for identifying kidney cancerr. *Anal Biochem* **2007,** *363* (2), 185-95.

28. Rosa Ragone, F. S., Sara Piccinonna, Monica Rutigliano, Galleggiante Vanessa, Silvano Palazzo, Giuseppe Lucarelli, Pasquale Ditonno, Michele Battaglia, Francesco

Paolo Fanizzi, Francesco Paolo Schena, Renal Cell Carcinoma: A Study Through NMR-Based Metabolomics Combined With Transcriptomics. *Diseases* **2016,** *4* (1), 7.

29. Oluyemi S. Falegan, M. W. B., Rustem A. Shaykhutdinov, Phillip M. Pieroraio, Farshad Farshidfar, Hans J. Vogel, Mohamad E. Allaf, Matthew E. Hyndman, Urine and Serum Metabolomics Analyses May Distinguish between Stages of Renal Cell Carcinoma. *Metabolites* **2017,** *7* (1), 6.

30. Oluyemi S. Falegan, S. A. A. E., Andries Zijlstra, M. Eric Hyndman, Hans J. Vogel, Urinary Metabolomics Validates Metabolic Differentiation Between Renal Cell Carcinoma Stages and Reveals a Unique Metabolic Profile for Oncocytomas. *Metabolites* **2019,** *9* (8), 155.

31. Niels-Peter Vest Nielsen, J. M. C., Jørn Smedsgaarda, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr* **1998,** *805* (1-2), 17-35.

32. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H., Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* **2006,** *78* (13), 4281-90.

33. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **1995,** *6* (3), 277-93.

34. Bingol, K.; Li, D. W.; Zhang, B.; Bruschweiler, R., Comprehensive Metabolite Identification Strategy Using Multiple Two-Dimensional NMR Spectra of a Complex Mixture Implemented in the COLMARm Web Server. *Anal Chem* **2016,** *88* (24), 12411-12418.

35. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L., BioMagResBank. *Nucleic Acids Res* **2008,** *36* (Database issue), D402-8.

36. Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L., HMDB: the Human Metabolome Database. *Nucleic Acids Res* **2007,** *35* (Database issue), D521-6.

37. Paul Rosenbaum, D. R., The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983,** *70* (1), 41–55.

38. Fabian Pedregosa, G. V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, Scikit-learn: Machine Learning in Python. *JMLR* **2011,** *12* (85), 2825−2830.

39. Wang, L.; Wang, Y.; Chang, Q., Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods* **2016,** *111*, 21-31.

40. Girish Chandrashekar, F. S., A survey on feature selection methods. *Computers and Electrical Engineering* **2014,** *40* (1), 16-28.

41. Yoav Benjamini, Y. H., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* **1995,** *57* (1), 12.

42. Waskom, M. L., Seaborn: statistical data visualization. *Journal of Open Source Software* **2021,** *6* (60), 3021.

43. McKinney, W., Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* **2010,** *445*, 56-61.

44. Hunter, J. D., Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007,** *9* (3), 90-95.

45. Stéfan van der Walt, C. C., Gaël Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **2011,** *13* (2), 22-30.

46. Oliphant, T. E., Python for Scientific Computing. *Computing in Science & Engineering* **2007,** *9* (3), 10-20.

47. Seabold Skipper, J. P., Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference* **2010**.

48. Fernando Pérez, B. E. G., IPython: A System for Interactive Scientific Computing. *Computing in Science and Engineering* **2007,** *9* (3), 21-29.

49. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S., Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **2016,** *44* (D1), D463-70.

50. Worley, B.; Powers, R., Multivariate Analysis in Metabolomics. *Curr Metabolomics* **2013,** *1* (1), 92-107.

51. Boulesteix, A. L.; Bender, A.; Lorenzo Bermejo, J.; Strobl, C., Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Brief Bioinform* **2012,** *13* (3), 292-304.

52. Pandey, N.; Lanke, V.; Vinod, P. K., Network-based metabolic characterization of renal cell carcinoma. *Sci Rep* **2020,** *10* (1), 5955.

53. Hakimi, A. A.; Reznik, E.; Lee, C. H.; Creighton, C. J.; Brannon, A. R.; Luna, A.; Aksoy, B. A.; Liu, E. M.; Shen, R.; Lee, W.; Chen, Y.; Stirdivant, S. M.; Russo, P.; Chen, Y. B.; Tickoo, S. K.; Reuter, V. E.; Cheng, E. H.; Sander, C.; Hsieh, J. J., An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancer Cell* **2016,** *29* (1), 104-116.

54. Lees, H. J.; Swann, J. R.; Wilson, I. D.; Nicholson, J. K.; Holmes, E., Hippurate: the natural history of a mammalian-microbial cometabolite. *J Proteome Res* **2013,** *12* (4), 1527-46.

55. Calvani, R.; Miccheli, A.; Capuani, G.; Tomassini Miccheli, A.; Puccetti, C.; Delfini, M.; Iaconelli, A.; Nanni, G.; Mingrone, G., Gut microbiome-derived metabolites characterize a peculiar obese urinary metabotype. *Int J Obes (Lond)* **2010,** *34* (6), 1095-8.

56. Waldram, A.; Holmes, E.; Wang, Y.; Rantalainen, M.; Wilson, I. D.; Tuohy, K. M.; McCartney, A. L.; Gibson, G. R.; Nicholson, J. K., Top-down systems biology modeling of host metabotype-microbiome associations in obese rodents. *J Proteome Res* **2009,** *8* (5), 2361-75.

57. Holmes, E.; Loo, R. L.; Stamler, J.; Bictash, M.; Yap, I. K.; Chan, Q.; Ebbels, T.; De Iorio, M.; Brown, I. J.; Veselkov, K. A.; Daviglus, M. L.; Kesteloot, H.; Ueshima, H.;

Zhao, L.; Nicholson, J. K.; Elliott, P., Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* **2008,** *453* (7193), 396-400.

58. Zhu, Q.; Zhou, L.; Yang, Z.; Lai, M.; Xie, H.; Wu, L.; Xing, C.; Zhang, F.; Zheng, S., O-GlcNAcylation plays a role in tumor recurrence of hepatocellular carcinoma following liver transplantation. *Med Oncol* **2012,** *29* (2), 985-93.

59. Xu, D.; Wang, W.; Bian, T.; Yang, W.; Shao, M.; Yang, H., Increased expression of O-GlcNAc transferase (OGT) is a biomarker for poor prognosis and allows tumorigenesis and invasion in colon cancer. *Int J Clin Exp Pathol* **2019,** *12* (4), 1305-1314.

60. Gu, Y.; Mi, W.; Ge, Y.; Liu, H.; Fan, Q.; Han, C.; Yang, J.; Han, F.; Lu, X.; Yu, W., GlcNAcylation plays an essential role in breast cancer metastasis. *Cancer Res* **2010,** *70* (15), 6344-51.

61. Lynch, T. P.; Ferrer, C. M.; Jackson, S. R.; Shahriari, K. S.; Vosseller, K.; Reginato, M. J., Critical role of O-Linked beta-N-acetylglucosamine transferase in prostate cancer invasion, angiogenesis, and metastasis. *J Biol Chem* **2012,** *287* (14), 11070-81.

62. de Queiroz, R. M.; Oliveira, I. A.; Piva, B.; Bouchuid Catao, F.; da Costa Rodrigues, B.; da Costa Pascoal, A.; Diaz, B. L.; Todeschini, A. R.; Caarls, M. B.; Dias, W. B., Hexosamine Biosynthetic Pathway and Glycosylation Regulate Cell Migration in Melanoma Cells. *Front Oncol* **2019,** *9*, 116.

63. Ma, Z.; Vosseller, K., Cancer metabolism and elevated O-GlcNAc in oncogenic signaling. *J Biol Chem* **2014,** *289* (50), 34457-65.

64. Gries, W.; Kupper, K.; Leng, G., Rapid and sensitive LC-MS-MS determination of 2-mercaptobenzothiazole, a rubber additive, in human urine. *Anal Bioanal Chem* **2015,** *407* (12), 3417-23.

65. Murawski, A.; Schmied-Tobies, M. I. H.; Schwedler, G.; Rucic, E.; Gries, W.; Schmidtkunz, C.; Kupper, K.; Leng, G.; Conrad, A.; Kolossa-Gehring, M., 2-Mercaptobenzothiazole in urine of children and adolescents in Germany - Human biomonitoring results of the German Environmental Survey 2014-2017 (GerES V). *Int J Hyg Environ Health* **2020,** *228*, 113540.

66. Avagyan, R.; Sadiktsis, I.; Bergvall, C.; Westerholm, R., Tire tread wear particles in ambient air--a previously unknown source of human exposure to the biocide 2-mercaptobenzothiazole. *Environ Sci Pollut Res Int* **2014,** *21* (19), 11580-6.

67. Wang, W.; Ren, S.; Zhang, H.; Yu, J.; An, W.; Hu, J.; Yang, M., Occurrence of nine nitrosamines and secondary amines in source water and drinking water: Potential of secondary amines as nitrosamine precursors. *Water Res* **2011,** *45* (16), 4930-8.

68. Hecht, S. S., Approaches to cancer prevention based on an understanding of N-nitrosamine carcinogenesis. *Proc Soc Exp Biol Med* **1997,** *216* (2), 181-91.

69. Diana, M.; Felipe-Sotelo, M.; Bond, T., Disinfection byproducts potentially responsible for the association between chlorinated drinking water and bladder cancer: A review. *Water Res* **2019,** *162*, 492-504.

70. Zhao, C.; Lu, Q.; Gu, Y.; Pan, E.; Sun, Z.; Zhang, H.; Zhou, J.; Du, Y.; Zhang, Y.; Feng, Y.; Liu, R.; Pu, Y.; Yin, L., Distribution of N-nitrosamines in drinking water and human urinary excretions in high incidence area of esophageal cancer in Huai'an, China. *Chemosphere* **2019,** *235*, 288-296.

71. Zhao, C.; Zhou, J.; Gu, Y.; Pan, E.; Sun, Z.; Zhang, H.; Lu, Q.; Zhang, Y.; Yu, X.; Liu, R.; Pu, Y.; Yin, L., Urinary exposure of N-nitrosamines and associated risk of esophageal cancer in a high incidence area in China. *Sci Total Environ* **2020,** *738*, 139713.

72. Bouma, K.; Schothorst, R. C., Identification of extractable substances from rubber nettings used to package meat products. *Food Addit Contam* **2003,** *20* (3), 300-7.

73. Fiddler, W.; Doerr, R. C., Gas chromatographic/chemiluminescence detection (thermal energy analyzer-nitrogen mode) method for the determination of dibutylamine in hams. *J AOAC Int* **1993,** *76* (3), 578-81.

74. Slominska, E. M.; Smolenski, R. T.; Szolkiewicz, M.; Leaver, N.; Rutkowski, B.; Simmonds, H. A.; Swierczynski, J., Accumulation of plasma N-methyl-2-pyridone-5-carboxamide in patients with chronic renal failure. *Mol Cell Biochem* **2002,** *231* (1-2), 83-8.

75. Warburg, O., On the origin of cancer cells. *Science* **1956,** *123* (3191), 309-14.

76. Perroud, B.; Ishimaru, T.; Borowsky, A. D.; Weiss, R. H., Grade-dependent proteomics characterization of kidney cancer. *Mol Cell Proteomics* **2009,** *8* (5), 971-85.

CHAPTER 3

URINE-BASED METABOLOMICS AND MACHINE LEARNING REVEALS METABOLITES ASSOCIATED WITH RENAL CELL CARCINOMA PROGRESSION[2]

[2] Olatomiwa O. Bifarin, David A. Gaul, Samyukta Sah, Rebecca S. Arnold, Kenneth Ogan, Viraj A. Master, John A. Petros, Facundo M. Fernández, and Arthur S. Edison. To be submitted to *Cancers*.

Contributing Authors

The data used for modelling in this chapter was generated in the previous chapter. David A. Gaul and Samyukta Sah conducted LC-MS compound annotations for new discriminant metabolites. Olatomiwa O. Bifarin conducted data analysis, machine learning experiments, and determined the biological relevance of discriminant metabolites. Facundo M. Fernández, and Arthur S. Edison are senior authors.

3.1 Simple Summary

Every year, hundreds of thousands of cases of renal carcinoma (RCC) are reported worldwide. Accurate staging of the disease is important for treatment and prognosis purposes; however, contemporary methods such as computerized tomography (CT) and biopsies are expensive and prone to sampling errors, respectively. As such, a non-invasive diagnostic assay for staging would be beneficial. This study aims to investigate urine metabolites as potential biomarkers to stage RCC. In the study, we identified a panel of such urine metabolites with machine learning techniques.

3.2 Abstract

Urine metabolomics profiling is an excellent non-invasive tool for staging RCC, in addition to providing metabolic insights into the disease progression. In this study, we utilized liquid chromatography-mass spectrometry (LC-MS), nuclear magnetic resonance (NMR), and machine learning (ML) for the discovery of urine metabolites associated with RCC progression. Two machine learning problems were posed in the study: RCC tumor size prediction with regression analysis and binary classification into early RCC (stage I and II) and advanced RCC stages (stage III and IV). 82 RCC patients with tumor size and

metabolomic measurements were used for the regression task, and 70 RCC patients with complete tumor-nodes-metastasis (TNM) staging information were used for the classification tasks under ten-fold cross-validation conditions. A voting ensemble regression model consisting of elastic net, ridge, and support vector regressor predicted tumor size of RCC with a R^2 value of 0.58. A voting classifier model consisting of random forest, support vector machines, logistic regression, and adaptive boosting gave an AUC of 0.96 and an accuracy of 87%. Some identified metabolites associated with renal cell carcinoma progression include 4-guanidinobutanoic acid, 7-aminomethyl-7-carbaguanine, 3-hydroxyanthranilic acid, lysylglycine, glycine, citrate, and pyruvate. Overall, we identified urine metabolites associated with renal cell carcinoma progression, espousing the promise of a urine-based metabolomic assay for staging the disease.

3.3 Introduction

Kidney cancer is one of the deadliest urinary cancers, with an advanced stage (stage III and IV) 5-year survival rate of 12%[1]. In the United States, 76,080 patients are projected to be diagnosed with the disease in 2021, with an estimated death toll of 13,780[2]. Renal cell carcinomas constitute approximately 90% of kidney and renal pelvis cancers. Because RCC prognosis and treatment depend on accurate staging, innovations in clinical staging are warranted. Accurate staging is currently carried out via computerized tomography (CT) scans and biopsy, which are expensive and highly invasive, respectively[3]. Non-invasive staging assays using urine samples, therefore, have the potential of being highly beneficial. In chapter 2, machine learning and multiplatform metabolomics was applied to detect RCC in urine samples. This current study investigates the discrimination of early and advanced

RCC stages using the RCC cohort datasets from our previous RCC detection metabolomics study.

Metabolic reprogramming in cancer contributes to cancer progression[4-6]. As such, changes in metabolite profiles in biofluids such as urine could enable RCC stage stratification and monitoring. Given the proximity of the kidney and urine, the case for a urine-based surveillance method for RCC is further strengthened. Mass spectrometry (MS) and Nuclear Magnetic Resonance (NMR) spectroscopy are two popular platforms for metabolomics profiling. In this study, both platforms were combined for maximum coverage. Omics research has been one of the hallmarks of biology research in the 21st century, marked by the rapid growth in the interrogation of large datasets by modern statistical techniques such as machine learning. The metabolomics literature has reflected this technological revolution[7-9]. Machine learning (ML) is a subfield of artificial intelligence that involves computer learning of patterns buried in data without being explicitly programmed to do so[10]. This characteristic makes machine learning a powerful tool for biomarker discoveries[8, 11].

While many studies have focused on RCC detection urine biomarkers, only a handful of studies have investigated biomarkers for RCC staging. In 2020, Liu and co-workers presented a metabolic panel for discriminating early RCC and late RCC, using liquid chromatography (LC)-MS. In their study, early RCC consisted of primary tumor stages 1 and 2 (pT1 and pT2), while advanced RCC consisted of primary tumor stages 3 and 4 (pT3 and pT4). The discriminant metabolic panel consisted of thymidine, cholic acid glucuronide, alanyl-proline, isoleucyl-hydroxyproline, and myristic acid (a fatty acid)[12]. In addition, Falegan *et al.* showed the potential for discrimination between pT1 and pT3 tumor

stages using gas chromatography (GC)-MS of serum samples coupled to Partial Least Squares-Discriminant Analysis (PLS-DA) modeling but did not identify the chemical structure of the metabolites responsible for such separation[13]. Furthermore, Manzi and co-workers reported a 26-lipid panel that discriminates early clear cell RCC from late-stage clear cell RCC in human serum samples[14].

Given the success of detecting RCC in urine in chapter 2 (AUC of 0.98 and accuracy of 88%), and because of the limited study on RCC metabolomics stage stratification, we used the data from that study to apply machine learning methods for RCC stage stratification. The original study focused on discriminating RCC from healthy controls. In this study, comprehensive tumor, nodes, and metastases (TNM) staging were carried out considering 1) the size of the primary tumor, 2) presence or absence of metastasis in the regional lymph nodes, and 3) the presence or absence of distant metastasis. NMR and LC-MS were used for non-targeted metabolic profiling, and ML was used for selecting the best set of metabolites and discrimination of early and advanced RCC. In addition, the size of tumors was predicted using ML through preselected urinary metabolites. In short, we provide evidence that a patient's urine metabolic profile can be used for monitoring RCC progression.

3.4 Materials and Methods

RCC patients were identified at Emory University Hospital. Urine samples were collected and stored at -80 °C, and hydrophilic interaction liquid chromatography high-resolution mass spectrometry and nuclear magnetic spectroscopy experiments were conducted for metabolic profiling in our published study on RCC detection. Tandem MS was performed for the identification of discriminant features, while NMR experiments carried out

included: one-dimensional nuclear Overhauser effect pulse sequence with pre-saturation of water resonance (NOESYPR1D), two-dimensional (2D) ^{1}H-^{13}C heteronuclear single quantum correlation (HSQC), and HSQC–TOCSY (HSQC–total correlation spectroscopy). NMR metabolomic features/metabolites are reported with resonances signatures and confidence scores in **Appendix B; Table B-1**. For LC-MS, seven thousand ninety-seven spectral features resulted from the analysis, with 4623 from positive mode and 2474 from negative mode. See chapter 2 for more complete experimental details. Subsequent machine learning analyses were carried out on the set of combined LC-MS and NMR metabolomic features.

3.4.1 Tumor Size Predictions

Maximum tumor width was used as a proxy for tumor size. Out of the 82 patients in the study with NMR and MS metabolomics measurements, only two had missing tumor size. These missing data were replaced with mean imputations. Pearson's correlations between metabolites and tumor size were used for metabolomic feature selection, with a cut-off value of 0.55. Elastic net, support vector, ridge, and voting ensemble regression models were used in tumor size predictions. The default parameters in the Scikit-learn library[15] in Python were used for modeling, and 80% and 20% of the data were used for training and testing purposes, respectively.

Ridge regression is a regularized linear model with the goal of minimizing the following objective function during training:

$$R(\boldsymbol{\beta},\, \lambda) = \|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2{}^2 \qquad (3.1)$$

Where $\boldsymbol{\beta} = (\beta_1 \ldots, \beta_p)'$ is a vector of slope regression coefficients, $\|\cdot\|_2$ is the L_2 norm, and λ is a tuning parameter that denotes the regularization strength. λ was set at 1.0.

Elastic net regression combines the *l*2 regularization of linear model (ridge regression), and *l*1 regularization of linear model (lasso regression). The objective function is as follow:

$$E(\boldsymbol{\beta}, \lambda, \alpha) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2^2 + \alpha\lambda\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda(1-\alpha)\|\boldsymbol{\beta}\|_2{}^2 \qquad (3.2)$$

Where $\boldsymbol{\beta} = (\beta_1 \ldots, \beta_p)'$ is a vector of slope regression coefficients, $\|\cdot\|_2$ is the L_2 norm, $\|\cdot\|_1$ is the L_1 norm, λ is a tuning parameter as described above, and α is the mixing parameter between ridge and lasso regression. λ and α were set to 1.0 and 0.5, respectively.

Support Vector Regressor (SVR) is a nonparametric technique as it relies on kernel functions. The objective function of SVR, as opposed to ordinary least square methods, involves minimizing the *l*2 norm of the coefficient vector ($\frac{1}{2}\|\boldsymbol{\beta}\|^2$), and not the squared error term. The objective function is as follows:

$$\frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}|\xi_i| \qquad (3.3a)$$

With the following constraint:

$$|y_i - \beta_i x_i| \leq \varepsilon + |\xi_i| \qquad (3.3b)$$

Where $\boldsymbol{\beta}$ is the coefficient vector, ε is the maximum error – which defines the margin of error acceptable to the model. Additional errors beyond ε are the slack parameters ξ. C is a regularization parameter that accommodates or penalizes ξ. In short, the objective function will be minimized with the constraint that the absolute difference between tumor sizes and predicted tumor sizes must be less or equal to the maximum error and absolute

slack parameters for samples during training. C and ε were set to 1.0 and 0.1, respectively. The radial basis function was the kernel used.

The voting ensemble regressor is an ensemble of the three regression models above. The base regressors were fit to the dataset, and the average of the output of the individual predictions for each base regressor was computed. All models were evaluated using the coefficient of determination (R^2), which describes the proportion of variance for the tumor size explained by the urine metabolites predictors. The formula is given below:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3.4}$$

Where y_i is the RCC tumor size of patient i, $\hat{y}_i$ is the predicted RCC tumor size of patient i, and $\bar{y}$ is the mean RCC tumor size of all patients.

3.4.2. Feature Selection for the RCC Stage Stratification

The normalized abundances of 50 metabolomic features quantified with NMR and > 7000 features from LC-MS were combined into one feature table in Python. Features for RCC stratification were retained through the following sequential steps: 1) 1-fold change between the two groups. 2) Student *T*-test with a *p*-value < 0.05, 3) One of two positively highly correlated features were retained (Pearson correlation > 0.8). Before further feature selection, all features were auto-scaled. Partial least square discriminant analysis (PLS-DA) was carried out, and the variable importance in projection (VIP) scores were used to select top-ranked features. Similarly, random forest recursive feature elimination (RF-RFE) was conducted, and its Gini Index was used to select top-ranked features. Overlapping features from the top 35 ranked features were selected as a metabolite panel for this study.

3.4.3 Machine Learning-enabled RCC Stage Stratification

RCC stage stratification was done by predicting early RCC (stage I and II) and advanced RCC (stage III and IV) with random forest, support vector machine, logistic regression, adaptive boosting, and a voting ensemble classifier. The default parameters in the Scikit-learn library[15] in Python were used for modeling. For training and testing purposes, a 10-fold cross-validation method was applied.

Random forest classification is a collection of decision tree estimators that are constructed with bootstrapped training samples. A decision tree is an inverted tree with a root node, an internal node, and a leaf node. The root and internal nodes are assigned metabolomic features that drive the decisions, while the leaf nodes give the final prediction of either early or advanced RCC. One hundred trees were used in the forest, and the quality of the split was measured by Gini impurity[16].

In support vector machines, the algorithm's goal is to discover a separating hyperplane, in this case, for a binary classification problem. The decision function takes the following form:

$$\text{RCC score} = \beta_0 + \sum_{j=1}^{j} \beta_j x_{ij} \tag{3.5}$$

β_0 and β_j are the bias and the weight parameters, respectively, of the model. The index *i* indicates the sample, and *j* represents the metabolomic features. The RCC score determines the class membership. In this formulation, a negative score indicates early RCC, while a positive score indicates advanced RCC, as the separating hyperplane takes the form $\beta_0 + \beta x' = 0$. A radial basis function (RBF) kernel was used, and the regulation parameter *C* was set at 1.0.

Logistic regression is an extension of linear regression where predictions are mapped to a class membership via the sigmoid function. The objective function is:

$$(\hat{y}, y) = -[y \log \hat{y} + (1 - y) log\ (1 - \hat{y})] \quad (3.6)$$

Where y indicates actual tumor size, and $\hat{y}$ the predicted tumor size.

Adaptive boosting (AdaBoost) is an ensemble of decision tree classifiers. AdaBoost involves the sequential boosting of its base classifier by ascribing larger weights to misclassified samples to induce the corrections of misclassifications in subsequent decision trees classifier. A linear combination of all base classifiers results in the final decision function. The learning rate was set to 1.0, while the number of decision trees was set to 50. The voting classifier is an ensemble of the four classifiers above. Soft voting was used, where the average probability outputs of the base learners are the voting classifier's final output.

Binary classifiers were evaluated using the area under the curve (AUC), accuracy, sensitivity, and specificity. AUC is the area under the curve of a receiver operating characteristics (ROC) curve. The ROC curve plots the true positive rate against the false-positive rate, displaying the model's performance at all classification thresholds. As such, this makes AUC the most desirable metric for binary classification with an unbalanced dataset. AUC was used to select the best models in the study.

Accuracy is calculated as the percentage of all correctly-predicted RCC stage samples.

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (3.7)$$

Sensitivity is the percentage of correctly predicted advanced RCC patients out of the total advanced RCC samples.

$$\text{Sensitivity} = TP/(TP + FN) \quad (3.8)$$

Specificity is the percentage of correctly predicted early-stage RCC patients out of the total early RCC samples.

$$\text{Specificity} = TN/(TN + FP) \quad (3.9)$$

Early RCC is denoted as a negative sample and advanced RCC as a positive sample in this setting. FP is false positive, FN is false negative, TP is true positive, and TN is true negative.

3.4.4 Implementation Environment and Computational Libraries

Edison Lab's in-house MATLAB Metabolomics Toolbox (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA, Matlab R2017b, The Mathworks, Inc.) was used for NMR data analysis. Combination of LC-MS and NMR metabolomic features and subsequent computational analysis was carried out using the Python 3.7.0 programming language. Pandas 1.0.5 was used for data handling and manipulations[17]. Matplotlib 3.3.0 and Seaborn 0.10.1 was used in data visualization[18]. NumPy 1.19.1 and SciPy 1.5.1 were used for numerical computing[19-20]. Statsmodel 0.11.1 was the statistical package used[21]. Sci-kit learn 0.23.2, and Yellowbrick 1.3.post1 was used for machine learning[15], and the integrated development environment used was Jupyter notebook[22]. All Jupyter notebooks used in the study can be found on GitHub (https://github.com/artedison/RCC-staging).

3.5 Results

3.5.1 Patient Cohort Characteristics

The TNM staging protocol used for tumor stratification is shown in **Figure 3-1**[3]. T indicates the size and extent of the primary tumor, N indicates the presence or absence of

metastasis in the regional lymph nodes, and M indicates the presence or absence of distant metastasis. Stage I and II are classified as early RCC, with the tumor confined to the kidney, while Stage III and IV are classified as advanced RCC, with the tumor spreading from the kidney. Out of the 82 urine samples with metabolomics measurements, twelve samples with inconclusive TNM staging were removed from the RCC stratification modeling. However, all samples were used for the tumor size predictions (**Appendix B; Table B-2**).

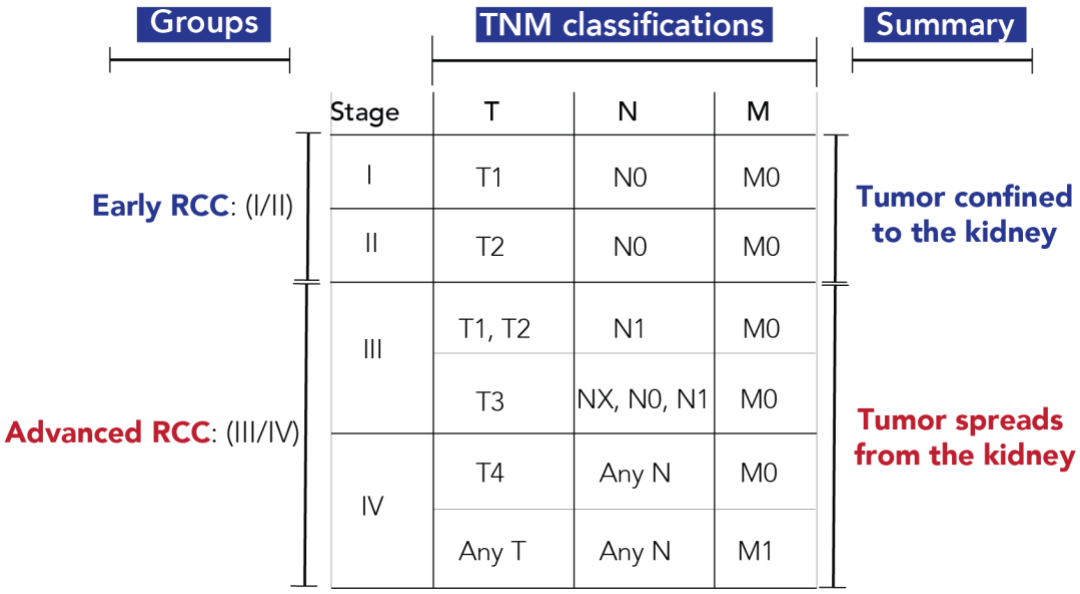


Figure 3-1. Classification of Early-stage and Advanced-stage RCC.

Abbreviations: T, primary tumor; T1, the tumor is 7 cm or less in its greatest dimension and limited to the kidney; T2, the tumor is greater than 7 cm in its greatest dimension but limited to the kidney; T3, the tumor extends into major veins or perinephric tissues but not into the ipsilateral adrenal gland and not beyond Gerota fascia; T4, tumor invades beyond Gerota fascia (including contiguous extension into the ipsilateral adrenal gland); N, regional lymph nodes; NX, regional lymph nodes cannot be assessed; N0, no regional

lymph node metastasis; N1, metastasis in regional lymph node(s); M, distant metastasis; M0, no distant metastasis; M1, distant metastasis[3].

The relevant clinical and demographics information for the studied cohort are shown in **Table 3-1**. We used 41 and 29 urine samples for early and advanced stage RCC, respectively, with no statistical significant difference between the groups concerning BMI (p=0.63, Student's T-Test) and age (p=0.14, Student's T-Test) (**Appendix B; Figure B-1 d & e**). The predominant race in both early RCC (n=26, 63.4%) and advanced RCC (n=21, 72.4%) was Caucasian, with a greater proportion of subjects who never smoked in both early RCC (n=24, 58.5%) and advanced RCC (n=19, 65.5%). The proportion of female patients in early RCC was 53.7% (n=21), and 31.1% (n=9) in advanced RCC. To test whether these covariates were potential confounders in the study, principal component analysis (PCA) was applied using the final 24-metabolite proposed in the study as features. PCA score plots showed no clustering based on any of the previously named variables. (**Appendix B; Figure B-1 a-c**).

Table 3-1. Patient Cohort Characteristics for RCC Staging

Characteristic	Early RCC	Advanced RCC
No of Urine Samples	41	29
Mean Age ± SD	60.1 ± 13.3	61.6 ± 13.2 [a]
Mean BMI± SD	29.9 ± 5.2	27.9 ± 6.2 [b]
Race		
Caucasian	26 (63.4%)	21 (72.4%)
Black/African American	14 (34.1%)	5 (17.2%)
American-Indian/Alaskan-Native	1 (2.4%)	1 (3.4%)
Mixed	-	1 (3.4%)
Unknown/Missing	-	1 (3.4%)
Smoker		

Never	24 (58.5%)	19 (65.5%)
Former/Current	17 (41.5%)	10 (34.5%)
Gender		
Male	19 (46.3%)	20 (68.9%)
Female	22 (53.7%)	9 (31.1%)
Histological Subtypes		
Pure Clear Cell	23 (56.1%)	26 (89.6%)
Papillary	9 (21.9%)	1 (3.4%)
Clear Cell Papillary	4 (9.8%)	-
Chromophobe	4 (9.8%)	-
Unclassified	1 (2.4%)	2 (6.9%)
Nuclear Grade		
1	-	-
2	21 (51.2%)	3 (10.3%)
3	17 (41.5%)	10 (34.5%)
4	3 (7.3%)	16 (55.2%)
RCC Stage		
I	33 (80.5%)	-
II	8 (19.5%)	-
III	-	15 (51.7%)
IV	-	14 (48.3%)

p-values were calculated using the Student *T*-test. [a]Age *p*-value = 0.63, [b] BMI *p*-value = 0.14. Twelve samples with missing TNM staging information were excluded.

3.5.2 Predicting RCC Tumor Size with Urine Metabolites

An important characteristic of cancer is the primary tumor size, which is the original or first tumor. As such, we investigated the associations between RCC tumor sizes and urine metabolites. To do this, Pearson correlations were computed between the two variables. Eighty-two samples with associated tumor size were used for the analysis, with **Appendix B; Table B-2** describing the clinical and demographic characteristics of this cohort. **Figure 3-2a** shows correlation plots of the four metabolites with the highest associations with tumor size in the study ($r > 0.55$) after removing inconsistent features. Three of the four metabolites were identified and they included, cytosine dimer (r=0.58, p=9.8×10^{-8}), dihydrouridine (r=0.56, p=3.8×10^{-8}), and asparaginyl-hydroxyproline (r=0.57,

p=1.8 × 10^{-8}) (**Table 3-2**). Given this positive association of urine metabolites with RCC tumor size, we predicted tumor size using these metabolites with elastic net regressor, support vector regressor, ridge regressor, and a voting ensemble regressor combining all three previous regression models. The models were trained on 80% of the data, while the test set consisted of 20% of the data. The best prediction was modeled using the vote regressor with a test set R^2 value of 0.58 (**Figure 3-2b & c**). The results for other models such as the elastic net regressor (train R^2=0.46, test set R^2= 0.56), support vector regressor (train R^2=0.23, test set R^2= 0.51), and ridge regressor model (train R^2=0.46, test set R^2= 0.56), are shown in **Appendix B; Figure B-2**.

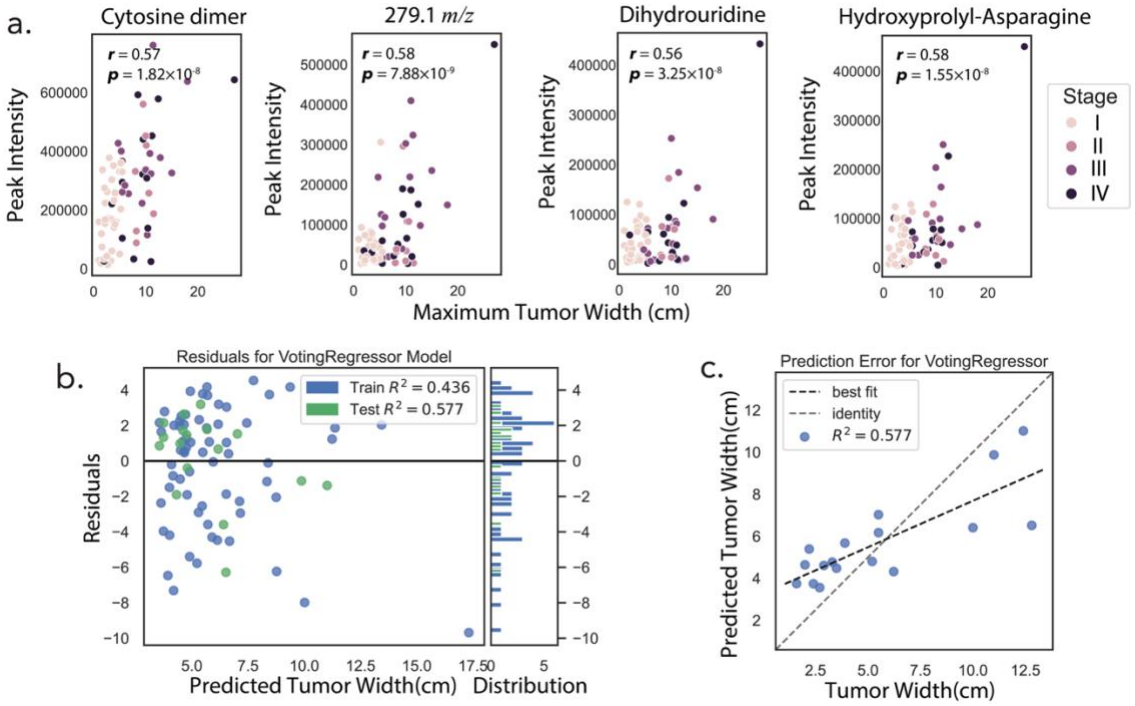


Figure 3-2. Correlation Between Tumor Size and Urine Metabolites, and Tumor Size Predictions.

a) Metabolites with the highest correlation with maximum tumor width. Pearson correlation coefficient and *p*-values for testing non-correlation are provided. The threshold for the correlation coefficient was $r > 0.55$. **b)** Residual plots **c)** prediction error plot.

Table 3-2. Compound Annotation and Identification for the Metabolites with the Highest Correlation (r >0.55) with Tumor Size of RCC Patients.

ID no.	Retention Time (min)	*m/z*		Adduct Type	Mass error (ppm)	Elemental Formula	Metabolite Identity	Confidence Level
		Theoretical	Experimental					
2745	1.87	223.0938	223.0936	$[M+H]^+$	-0.64	$C_8 H_{10} N_6 O_2$	cytosine dimer	2
3163	3.53	279.1187	279.1194	$[M+H]^+$	2.54	$C_{10} H_{18} N_2 O_7$	--	4
5362	3.46	245.0774	245.0775	$[M-H]^-$	0.61	$C_9 H_{14} N_2 O_6$	dihydrouridine	2
6681	2.80	244.0933	244.0934	$[M-H]^-$	0.31	$C_9 H_{15} N_3 O_5$	hydroxyprolyl-asparagine/ asparaginylhydroxyproline	2

3.5.3 Machine Learning Discriminates Early Stage RCC and Advanced Stage RCC

NMR and LC-MS metabolic features were combined into a single feature table, and 16 discriminating metabolites were selected using multiple methods. The chart in **Appendix B; Figure B-3** describes this machine learning pipeline. First, two filter-based approaches were used for feature selection: 1) metabolic features with greater than 1-fold difference between the early and advanced RCC, followed by 2) a Student *t*-Test between the two groups (cut off value, $p < 0.05$). This resulted in 171 LC-MS metabolic features (**Appendix B; Table B-3**). Two, a Pearson correlation-based method was used to remove potentially redundant features that might degrade machine learning predictions: for two highly correlated features, one feature was dropped (cut-off value, $r > 0.8$). This leaves 99

metabolic features remaining in the feature table. Finally, embedded feature selection techniques were layered inside a voting-based system for the final biomarker selection. A partial least squares regression technique was used to rank feature importance via its Variable Importance in Projection (VIP) scores, and the top metabolic 35 features were selected. In addition, a random forest recursive feature elimination technique was used to rank feature importance via its Gini index, and the top 35 features were also selected. As a voting system, features that appear on both lists were selected after the removal of inconsistent features. This led to selecting a 16 urine metabolites panel for RCC stage stratification (**Table 3-3**). The identified metabolites include 4-guanidinobutanoic acid, 7-aminomethyl-7-carbaguanine, N-alpha-N-alpha-dimethyl-L-histidine, diethyl-2-methyl-3-oxosuccinate, 3-hydroxyanthranilic acid, apo-[3-methylcrotonoyl-CoA:carbon-dioxide ligase (ADP-forming)], lys-gly/gly-lys, and succinic anhydride. All the markers were detected by LC-MS. Random forest (RF), adaptive boosting (AdaBoost), support vector machine with radial basis function kernel (SVM-RBF), logistic regression, and a voting ensemble combining all four methods were used for stratification under ten-fold cross-validation conditions. **Appendix B; Figure B-4** shows the ML predictions for RCC staging using this panel. The voting ensemble models gave the best predictions, with an AUC of 0.95, accuracy of 86%, sensitivity of 80%, and specificity of 91%.

Table 3-3. Compound Annotation and Identification of the 16-Metabolite Panel for RCC Staging.

ID	Retention Time (min)	Theoretical *m/z*	Experimental *m/z*	Adduct Type	Mass Error (ppm)	Elemental formula	Metabolite Identity	Confidence Score
1372	3.94	146.0924	146.0924	$[M+H]^+$	0.03	$C_5 H_{11} N_3 O_2$	4-guanidinobutanoic acid	2

1904	4.00	180.0879	180.0880	$[M+H]^+$	0.08	$C_7 H_9 N_5 O$	7-aminomethyl-7-carbaguanine	2
2122	1.20	184.1081	184.1080	$[M+H]^+$	-0.36	$C_8 H_{13} N_3 O_2$	N- _N- dimethyl-histidine	2
2317	0.89, 0.89	203.0913, 422.2020	203.0912, 422.2023	$[M+H]^+$, $[2M+NH4]^{+1}$	-0.44 0.71	$C_9 H_{14} O_5$	diethyl-2-methyl-3-oxosuccinate	3
2465	0.89, 0.89	154.0498 136.0393	154.0497, 136.0392	$[M+H]^+$, $[M+H-H2O]^+$	-0.62 -0.73	$C_7 H_7 N O_3$	3-hydroxyanthranilic acid	2
3163	3.53	279.1187	279.1194	$[M+H]^+$	2.54	$C_{10} H_{18} N_2 O_7$	--	4
3766	3.63	174.1237	174.1238	$[M+H]^+$	0.37	$C_7 H_{15} N_3 O_2$	apo-[3-methylcrotonoyl-CoA:carbon-dioxide ligase (ADP-forming)]	2
4116	3.79	119.0577	119.0580	$[M+H]^+$	4.51	$C_4 H_8 N O_3$	--	4
5045	3.49	218.0129	218.0123	$[M-H]^-$	-3.50	$C_7 H_9 N O_5 S$	--	4
5420	3.38	205.0526	205.0535	$[M-H]^-$	4.32	$C_4 H_{12} N_6 P_2$	--	4
5437	0.76	123.0114	123.0108	$[M-H]^-$	-4.47	$C_9 H_2 N$	--	4
5713	1.23	305.0990	305.0989	$[M-H]^-$	-0.58	$C_{11} H_{18} N_2 O_8$	--	4
5737	3.99	202.1197	202.1190	$[M-H]^-$	-3.58	$C_8 H_{17} N_3 O_3$	lys-gly/ gly-lys	2
5985	0.94	99.0087	99.0088	$[M-H]^-$	0.21	$C_4 H_4 O_3$	succinic anhydride	2
6687	0.86	369.0517	369.0502	$[M-H]^-$	-4.30	$C_6 H_{14} N_{10} O_5 S_2$	--	4

m/z = mass-to-charge ratio, min = minutes, ppm = part per million. Metabolite identification level was assigned based on the following criteria: 1) exact mass, isotopic pattern, retention time, and MS/MS spectrum of standard matched to the feature. 2) exact mass, isotopic pattern, and MS/MS spectrum matched with literature spectra or fragmentation ions observed are consistent with the proposed structure. 3) tentative ID assignment based on elemental formula match with literature. 4) unknowns.

To further improve prediction scores, we included metabolites selected for tumor size predictions that were missing in the 16-metabolite panel. These include cytosine dimer, dihydrouridine, and hydroxyprolyl-asparagine (**Table 3-2**). In addition, because NMR metabolomic features were not selected in the metabolite-panel, perhaps due to the over-abundance of MS features, we included NMR metabolomic features with a *p*-value

less than 0.05 (Student's *t*-Test) in the panel. These metabolites included citrate, glycine, choline, acetone, and pyruvate (**Table 3-4**).

Table 3-4. Compound Annotation and Identification for the NMR Metabolites with a *p*-Value of Less Than 0.05, for RCC Staging.

Metabolite/ Features	^{1}H (ppm)	^{13}C (ppm)	Peak patterns	Confidence Score	Fold Change	*p*-value
acetone	2.23	32.40	(s)	3	0.49	0.029
pyruvate	2.41	-	(s)	2	0.31	0.028
citrate	2.53	48.52	(d)	3	-0.54	0.003
choline	3.19	56.69	(s)	3	0.22	0.026
glycine	3.56	44.18	(s)	3	-0.66	0.032

s=singlet, d=doublet. Fold change (FC) was calculated as the base 2 logarithm of the mean integral ratios between advanced RCC and early RCC samples. Positive FC values indicate increased abundance in advanced RCC, while negative values indicate higher abundance in early RCC. *p*-values were Student's *t*-Test. Confidence score: (1) putatively characterized compound classes, or annotated compounds, (2) matches from 1D NMR to literature and/or 1D BBiorefcode compound (AssureNMR) or other database libraries such as Biological Magnetic Resonance Bank (BMRB) and Human Metabolome Database (HMDB) (3) matched to Heteronuclear Single Quantum Coherence (HSQC).

These operations gave rise to the final 24-metabolite panel. Normalized relative abundances for this panel are shown in **Figure 3-3**. As above, random forest, AdaBoost,

SVM-RBF, logistic regression, and a voting ensemble combining all four methods were used for stratification under ten-fold cross-validation conditions. **Figure 3-4** shows the ML predictions of RCC staging using the 24-metabolite panel. The voting ensemble classifier gave the best predictions with an AUC of 0.96, a slightly higher classification score than the 16-metabolite panel. Other prediction scores include 87% accuracy, 80% sensitivity, and specificity of 93%.

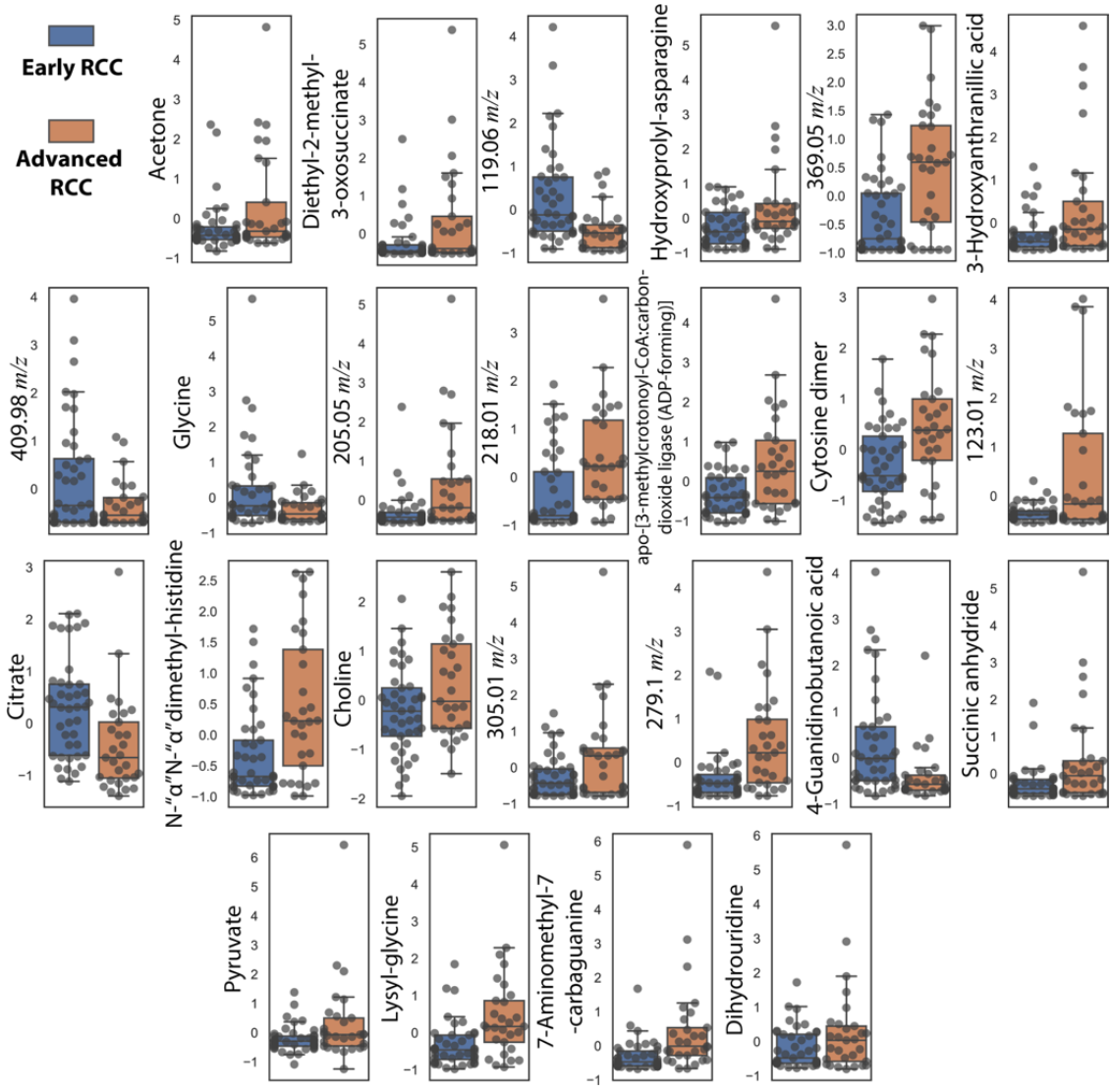


Figure 3-3. Box Plots Showing the Auto-scaled Normalized Relative Abundance of 24 Metabolite-panel that Distinguish Early-stage RCC from Advanced-stage RCC.

The mean, upper quartile, lower quartile, minimum, and maximum values are shown. All metabolites have *p*-values < 0.05 (Student *t*-test).

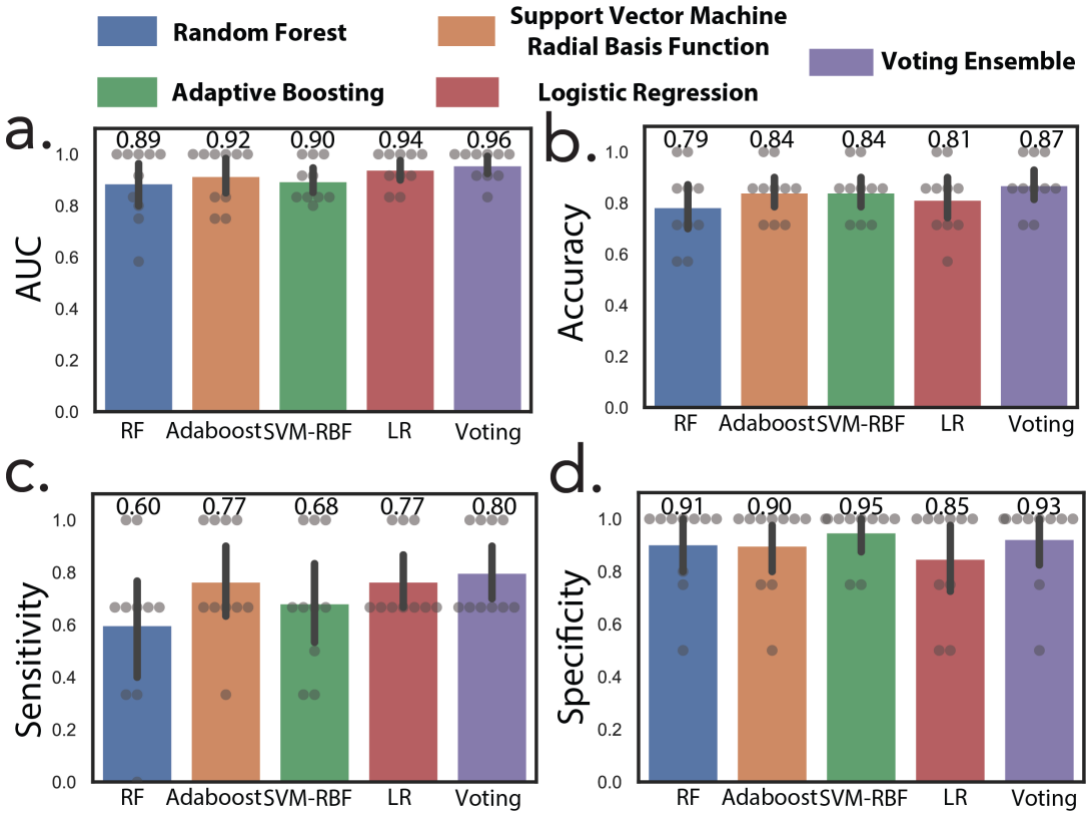


Figure 3-4. Machine Learning Discriminates Between Early-stage RCC and Late-stage RCC.

Machine learning predictions by random forest, AdaBoost, support vector machine radial basis function (SVM-RBF), logistic regression (LR), and voting ensembles using the 24-metabolite panel. (a) The area under the ROC curve (b) Accuracy (c) Sensitivity (d) Specificity.

3.6 Discussion

Monitoring tumor progression *via* systemic metabolite profiles in biofluids in the clinic presents a great opportunity. One characteristic of tumor progression is metabolic

rewiring[23]. Comparison of normal and tumor tissues has revealed dysregulation in the nucleotide biosynthesis, oxidative phosphorylation, glycolysis, and pentose phosphate pathway, amongst others[23]. Results in this study reinforce those findings, promising the capability to monitor RCC progression *via* urine-based metabolomics. We used LC-MS and NMR for metabolic profiling and machine learning for mining the dataset to identify the most discriminating metabolic features between early-stage and advanced-stage RCC. Twenty-four metabolites detected by both NMR and LC-MS were used for RCC staging classifications.

Evidence of upregulated nucleotide metabolism was observed in the metabolite-panel, with an increase in the abundances of cytosine dimer, 7-aminomethyl-7-carbaguanine, and dihydrouridine (DHU) in advanced stage RCC urine samples. Increased nucleotide metabolism is a hallmark of tumorigenesis, as it directly supports uncontrolled cell growth[24] *via* the pentose phosphate pathway[25]. This explains the predictive power of urinary cytosine dimer and DHU for RCC tumor size. Cytosine is present in both DNA and RNA, while DHU is found in tRNA as a nucleoside. Together, these metabolites could be indicative of nucleotide degradation in RCC[26]. An additional pyrimidine metabolite in the panel, 7-aminomethyl-7-carbaguanine, is one of the precursors for the synthesis of queuosine, a modified analogue of guanosine found in the first anticodon loop of tRNAs for histidine, aspartic acid, asparagine, and tyrosine[27]. This modification of tRNAs has been reported to promote cell proliferation in cancer in mouse models[28]. In addition, queuine tRNA-ribosyltransferase (QTRT1), the enzyme that catalyzes the hypermodification of queuosine using 7-aminomethyl-7-carbaguanine, is highly expressed in lung adenocarcinoma (LUAD)[29]. It has been identified as a risk factor for the progression of

LUAD[29]. In addition, this trend has been reported in other human malignant tumors[30-31], and breast cancer cell lines[32].

The relative abundance of 3-hydroxyanthranilic acid was increased in advanced RCC patients' urine samples. 3-Hydroxyanthranilic acid is an intermediate of tryptophan metabolism, a metabolic pathway that had been implicated in a recent urine metabolomics RCC study[12]. In that study, N-formylkynurenine, a metabolite upstream of hydroxyanthranilic acid, was selected as a putative marker that discriminated malignant RCC tumors from the healthy cohort and benign RCC tumors. In fact, studies from as early as 1975 have reported higher levels of 3-hydroxyanthranilic acid in untreated bladder and kidney carcinoma patients[33]. Indeed, 3-hydroxyanthranilic acid has been shown to promote tumor immune evasion[34-36]. Two dipeptides, hydroxyprolyl-asparagine, and lysyl-glycine had elevated levels in advanced RCC urine samples in our study. Numerous dipeptides had been reported to increase at advanced RCC stages (III and IV) in a paired clear cell renal cell carcinoma (ccRCC)/normal tissue study[37]. The presence of these increased dipeptide levels might be indicative of various protein degradation and reutilization processes[38-39]. In addition, in a urine metabolomics study, the dipeptides alanyl-proline and isoleucyl-hydroxyproline were seen as being elevated in RCC pT3 and pT4 stages[12]. Lower levels of guanidinobutanoic acid, a gamma-amino acid and uremia toxin, was found in the advanced stages of RCC. This might be due to the progressive retention of the metabolite that is otherwise excreted normally in healthy kidneys[40]. Apo-[3-methylcrotonoyl-CoA:carbon-dioxide ligase (ADP-forming)] is involved in the biotin metabolism pathway, indicating a possible alteration in biotin metabolism in advanced RCC. Likewise, diethyl-2-methyl-3-oxosuccinate and N,N-dimethyl-histidine might indicate alterations in succinate and

histidine metabolism, respectively; while succinic anhydride is likely an exogenous metabolite that is used in food additives[41]. Succinic anhydride is a Class 3 carcinogen, according to the WHO International Agency for Research on Cancer (IARC).

NMR-derived metabolites in the panel included citrate, glycine, choline, acetone, and pyruvate. Reduced levels of citrate and increased levels of pyruvate suggest a dysregulated aerobic glycolytic pathway in RCC[42]. This dysregulation is required to keep up with the cell proliferation that characterizes tumors, and the differences in the abundance of metabolites of this pathway between early and advanced RCC, are expected as the tumor progresses. Abundance of citrate has been reported to decrease in urine metabolomics studies comparing healthy controls or benign RCC tumors with malignant RCC tumors[43-44]. Citrate has been linked to drive increased fatty acid synthesis in tumors[45], and the overexpression of ATP citrate lyase has been reported as RCC progresses[46]. ATP citrate lyase links carbohydrate metabolism to fatty acid biosynthesis *via* the conversion of citrate to acetyl-CoA. In addition, elevated pyruvate levels, another evidence of dysregulated glucose metabolism, were reported in a urine metabolomics study that compared benign RCC tumors with malignant RCC[44]. The lower levels of glycine abundance in advanced RCC urine samples is in agreement with the role of glycine in rapid cancer cell proliferation[47-50]. In a study that used mass spectrometry to measure the consumption and release of metabolites in media of NCI-60 cancer cell lines, glycine consumption and the expression of mitochondrial glycine biosynthetic pathway correlated with proliferation of the cancer cell lines[47]. This is because glycine can contribute to both purine and pyrimidine biosynthesis[48-49], playing a pivotal role in sustaining cancer cell growth[50]. Indeed, urinary glycine has been shown to decrease in response to RCC cancer

development when benign tumors are compared to malignant tumors[13]. The higher relative abundance of choline in advanced RCC might be due to increased levels of choline-containing compounds that have been reported in tumors[51]. These compounds are a major component of cell membranes required for cell proliferation[51]. Magnetic resonance spectroscopic imaging has been used to show that total choline is associated with the aggressiveness of breast cancer[52] and prostate cancer[53]. In addition, it had also been used in the detection and grading of brain tumors[54-55]. The increased levels of acetone in advanced RCC in the panel can be explained in the light of the increase in the level of ketone bodies associated with some cancers[56]. And indeed, an increase in the level of acetoacetate, another ketone, detected and quantified with NMR, is reported in our study. A higher abundance of ketone bodies might be due to the Warburg effect, which leads to an accumulation of Acetyl-CoA, and in turn, increased production of ketone bodies[56]. In an *in vitro* metabolomics study, ketones were found to increase significantly in the exometabolome of RCC cells compared to a non-tumor cell line[57].

Overall, our study reveals metabolites associated with RCC progression, with a panel of metabolites discriminating between early RCC and advanced RCC with high accuracy. While our results demonstrate the potential of a urine metabolomic approach to identify biomarkers for RCC stage stratification, further validation of the results, especially in larger and independent cohorts, is necessary.

3.7 References

1. Padala, S. A.; Barsouk, A.; Thandra, K. C.; Saginala, K.; Mohammed, A.; Vakiti, A.; Rawla, P.; Barsouk, A., Epidemiology of Renal Cell Carcinoma. *World J Oncol* **2020,** *11* (3), 79-87.

2. Siegel, R. L.; Miller, K. D.; Fuchs, H. E.; Jemal, A., Cancer Statistics, 2021. *CA Cancer J Clin* **2021,** *71* (1), 7-33.

3. Escudier, B.; Porta, C.; Schmidinger, M.; Rioux-Leclercq, N.; Bex, A.; Khoo, V.; Grunwald, V.; Gillessen, S.; Horwich, A.; clinicalguidelines@esmo.org, E. G. C. E. a., Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-updagger. *Ann Oncol* **2019,** *30* (5), 706-720.

4. Faubert, B.; Solmonson, A.; DeBerardinis, R. J., Metabolic reprogramming and cancer progression. *Science* **2020,** *368* (6487).

5. Lameirinhas, A.; Miranda-Goncalves, V.; Henrique, R.; Jeronimo, C., The Complex Interplay between Metabolic Reprogramming and Epigenetic Alterations in Renal Cell Carcinoma. *Genes (Basel)* **2019,** *10* (4).

6. Wettersten, H. I.; Aboud, O. A.; Lara, P. N., Jr.; Weiss, R. H., Metabolic reprogramming in clear cell renal cell carcinoma. *Nat Rev Nephrol* **2017,** *13* (7), 410-419.

7. Pomyen, Y.; Wanichthanarak, K.; Poungsombat, P.; Fahrmann, J.; Grapov, D.; Khoomrung, S., Deep metabolome: Applications of deep learning in metabolomics. *Comput Struct Biotechnol J* **2020,** *18*, 2818-2825.

8. Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M., Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020,** *10* (6).

9. Cuperlovic-Culf, M., Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. *Metabolites* **2018,** *8* (1).

10. Mitchell, T. M., *Machine Learning*. McGraw-Hill: New York, 1997; p xvii, 414 p.

11. Khan, S. R.; Manialawy, Y.; Wheeler, M. B.; Cox, B. J., Unbiased data analytic strategies to improve biomarker discovery in precision medicine. *Drug Discov Today* **2019,** *24* (9), 1735-1748.

12. Liu, X.; Zhang, M.; Liu, X.; Sun, H.; Guo, Z.; Tang, X.; Wang, Z.; Li, J.; Li, H.; Sun, W.; Zhang, Y., Urine Metabolomics for Renal Cell Carcinoma (RCC) Prediction: Tryptophan Metabolism as an Important Pathway in RCC. *Front Oncol* **2019,** *9*, 663.

13. Falegan, O. S.; Ball, M. W.; Shaykhutdinov, R. A.; Pieroraio, P. M.; Farshidfar, F.; Vogel, H. J.; Allaf, M. E.; Hyndman, M. E., Urine and Serum Metabolomics Analyses May Distinguish between Stages of Renal Cell Carcinoma. *Metabolites* **2017,** *7* (1).

14. Manzi, M.; Palazzo, M.; Knott, M. E.; Beauseroy, P.; Yankilevich, P.; Gimenez, M. I.; Monge, M. E., Coupled Mass-Spectrometry-Based Lipidomics Machine Learning Approach for Early Detection of Clear Cell Renal Cell Carcinoma. *J Proteome Res* **2021,** *20* (1), 841-857.

15. Fabian Pedregosa, G. V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay, Scikit-learn: Machine Learning in Python. *JMLR* **2011,** *12* (85), 2825−2830.

16. Nembrini, S.; Konig, I. R.; Wright, M. N., The revival of the Gini importance? *Bioinformatics* **2018,** *34* (21), 3711-3718.

17. McKinney, W., Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* **2010,** *445*, 56-61.

18. Hunter, J. D., Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007,** *9* (3), 90-95.

19. Stéfan van der Walt, C. C., Gaël Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **2011,** *13* (2), 22-30.

20. Oliphant, T. E., Python for Scientific Computing. *Computing in Science & Engineering* **2007,** *9* (3), 10-20.

21. Seabold Skipper, J. P., Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference* **2010**.

22. Fernando Pérez, B. E. G., IPython: A System for Interactive Scientific Computing. *Computing in Science and Engineering* **2007,** *9* (3), 21-29.

23. Hu, J.; Locasale, J. W.; Bielas, J. H.; O'Sullivan, J.; Sheahan, K.; Cantley, L. C.; Vander Heiden, M. G.; Vitkup, D., Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nat Biotechnol* **2013,** *31* (6), 522-9.

24. Vander Heiden, M. G.; Cantley, L. C.; Thompson, C. B., Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **2009,** *324* (5930), 1029-33.

25. Jin, L.; Zhou, Y., Crucial role of the pentose phosphate pathway in malignant tumors. *Oncol Lett* **2019,** *17* (5), 4213-4221.

26. Seidel, A.; Brunner, S.; Seidel, P.; Fritz, G. I.; Herbarth, O., Modified nucleosides: an accurate tumour marker for clinical diagnosis of cancer, early detection and therapy control. *Br J Cancer* **2006,** *94* (11), 1726-33.

27. Harada, F.; Nishimura, S., Possible anticodon sequences of tRNA His , tRNA Asm , and tRNA Asp from Escherichia coli B. Universal presence of nucleoside Q in the first postion of the anticondons of these transfer ribonucleic acids. *Biochemistry* **1972,** *11* (2), 301-8.

28. Pathak, C.; Jaiswal, Y. K.; Vinayak, M., Hypomodification of transfer RNA in cancer with respect to queuosine. *RNA Biol* **2005,** *2* (4), 143-8.

29. Ma, Q.; He, J., Enhanced expression of queuine tRNA-ribosyltransferase 1 (QTRT1) predicts poor prognosis in lung adenocarcinoma. *Ann Transl Med* **2020,** *8* (24), 1658.

30. Emmerich, B.; Zubrod, E.; Weber, H.; Maubach, P. A.; Kersten, H.; Kersten, W., Relationship of queuine-lacking transfer RNA to the grade of malignancy in human leukemias and lymphomas. *Cancer Res* **1985,** *45* (9), 4308-14.

31. Baranowski, W.; Dirheimer, G.; Jakowicki, J. A.; Keith, G., Deficiency of queuine, a highly modified purine base, in transfer RNAs from primary and metastatic ovarian malignant tumors in women. *Cancer Res* **1994,** *54* (16), 4468-71.

32. Zhang, J.; Lu, R.; Zhang, Y.; Matuszek, Z.; Zhang, W.; Xia, Y.; Pan, T.; Sun, J., tRNA Queuosine Modification Enzyme Modulates the Growth and Microbiome Recruitment to Breast Tumors. *Cancers (Basel)* **2020,** *12* (3).

33. Teulings, F. A.; Mulder-Kooy, G. E.; Peters, H. A.; Fokkens, W.; Van Der Werf-Messing, B., The excretion of 3-hydroxyanthranilic acid in patients with bladder and kidney carcinoma. *Acta Vitaminol Enzymol* **1975,** *29* (1-6), 108-12.

34. Hornigold, N.; Dunn, K. R.; Craven, R. A.; Zougman, A.; Trainor, S.; Shreeve, R.; Brown, J.; Sewell, H.; Shires, M.; Knowles, M.; Fukuwatari, T.; Maher, E. R.; Burns, J.;

Bhattarai, S.; Menon, M.; Brazma, A.; Scelo, G.; Feulner, L.; Riazalhosseini, Y.; Lathrop, M.; Harris, A.; Selby, P. J.; Banks, R. E.; Vasudev, N. S., Dysregulation at multiple points of the kynurenine pathway is a ubiquitous feature of renal cancer: implications for tumour immune evasion. *Br J Cancer* **2020,** *123* (1), 137-147.

35. Badawy, A. A., Kynurenine Pathway of Tryptophan Metabolism: Regulatory and Functional Aspects. *Int J Tryptophan Res* **2017,** *10*, 1178646917691938.

36. Fallarino, F.; Grohmann, U.; Vacca, C.; Bianchi, R.; Orabona, C.; Spreca, A.; Fioretti, M. C.; Puccetti, P., T cell apoptosis by tryptophan catabolism. *Cell Death Differ* **2002,** *9* (10), 1069-77.

37. Hakimi, A. A.; Reznik, E.; Lee, C. H.; Creighton, C. J.; Brannon, A. R.; Luna, A.; Aksoy, B. A.; Liu, E. M.; Shen, R.; Lee, W.; Chen, Y.; Stirdivant, S. M.; Russo, P.; Chen, Y. B.; Tickoo, S. K.; Reuter, V. E.; Cheng, E. H.; Sander, C.; Hsieh, J. J., An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancer Cell* **2016,** *29* (1), 104-116.

38. Commisso, C.; Davidson, S. M.; Soydaner-Azeloglu, R. G.; Parker, S. J.; Kamphorst, J. J.; Hackett, S.; Grabocka, E.; Nofal, M.; Drebin, J. A.; Thompson, C. B.; Rabinowitz, J. D.; Metallo, C. M.; Vander Heiden, M. G.; Bar-Sagi, D., Macropinocytosis of protein is an amino acid supply route in Ras-transformed cells. *Nature* **2013,** *497* (7451), 633-7.

39. Mizushima, N.; Komatsu, M., Autophagy: renovation of cells and tissues. *Cell* **2011,** *147* (4), 728-41.

40. Vanholder, R.; De Smet, R.; Glorieux, G.; Argiles, A.; Baurmeister, U.; Brunet, P.; Clark, W.; Cohen, G.; De Deyn, P. P.; Deppisch, R.; Descamps-Latscha, B.; Henle, T.; Jorres, A.; Lemke, H. D.; Massy, Z. A.; Passlick-Deetjen, J.; Rodriguez, M.; Stegmayr, B.;

Stenvinkel, P.; Tetta, C.; Wanner, C.; Zidek, W.; European Uremic Toxin Work, G., Review on uremic toxins: classification, concentration, and interindividual variability. *Kidney Int* **2003,** *63* (5), 1934-43.

41. Sweedman, M. C.; Tizzotti, M. J.; Schafer, C.; Gilbert, R. G., Structure and physicochemical properties of octenyl succinic anhydride modified starches: a review. *Carbohydr Polym* **2013,** *92* (1), 905-20.

42. Shuch, B.; Linehan, W. M.; Srinivasan, R., Aerobic glycolysis: a novel target in kidney cancer. *Expert Rev Anticancer Ther* **2013,** *13* (6), 711-9.

43. Ragone, R.; Sallustio, F.; Piccinonna, S.; Rutigliano, M.; Vanessa, G.; Palazzo, S.; Lucarelli, G.; Ditonno, P.; Battaglia, M.; Fanizzi, F. P.; Schena, F. P., Renal Cell Carcinoma: A Study through NMR-Based Metabolomics Combined with Transcriptomics. *Diseases* **2016,** *4* (1).

44. Falegan, O. S.; Arnold Egloff, S. A.; Zijlstra, A.; Hyndman, M. E.; Vogel, H. J., Urinary Metabolomics Validates Metabolic Differentiation Between Renal Cell Carcinoma Stages and Reveals a Unique Metabolic Profile for Oncocytomas. *Metabolites* **2019,** *9* (8).

45. Icard, P.; Poulain, L.; Lincet, H., Understanding the central role of citrate in the metabolism of cancer cells. *Biochim Biophys Acta* **2012,** *1825* (1), 111-6.

46. Teng, L.; Chen, Y.; Cao, Y.; Wang, W.; Xu, Y.; Wang, Y.; Lv, J.; Li, C.; Su, Y., Overexpression of ATP citrate lyase in renal cell carcinoma tissues and its effect on the human renal carcinoma cells in vitro. *Oncol Lett* **2018,** *15* (5), 6967-6974.

47. Jain, M.; Nilsson, R.; Sharma, S.; Madhusudhan, N.; Kitami, T.; Souza, A. L.; Kafri, R.; Kirschner, M. W.; Clish, C. B.; Mootha, V. K., Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* **2012,** *336* (6084), 1040-4.

48. Villa, E.; Ali, E. S.; Sahu, U.; Ben-Sahra, I., Cancer Cells Tune the Signaling Pathways to Empower de Novo Synthesis of Nucleotides. *Cancers (Basel)* **2019,** *11* (5).

49. Zhang, W. C.; Shyh-Chang, N.; Yang, H.; Rai, A.; Umashankar, S.; Ma, S.; Soh, B. S.; Sun, L. L.; Tai, B. C.; Nga, M. E.; Bhakoo, K. K.; Jayapal, S. R.; Nichane, M.; Yu, Q.; Ahmed, D. A.; Tan, C.; Sing, W. P.; Tam, J.; Thirugananam, A.; Noghabi, M. S.; Pang, Y. H.; Ang, H. S.; Mitchell, W.; Robson, P.; Kaldis, P.; Soo, R. A.; Swarup, S.; Lim, E. H.; Lim, B., Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell* **2012,** *148* (1-2), 259-72.

50. Reina-Campos, M.; Diaz-Meco, M. T.; Moscat, J., The complexity of the serine glycine one-carbon pathway in cancer. *J Cell Biol* **2020,** *219* (1).

51. Glunde, K.; Bhujwalla, Z. M.; Ronen, S. M., Choline metabolism in malignant transformation. *Nat Rev Cancer* **2011,** *11* (12), 835-48.

52. Chen, J. H.; Mehta, R. S.; Baek, H. M.; Nie, K.; Liu, H.; Lin, M. Q.; Yu, H. J.; Nalcioglu, O.; Su, M. Y., Clinical characteristics and biomarkers of breast cancer associated with choline concentration measured by 1H MRS. *NMR Biomed* **2011,** *24* (3), 316-24.

53. Scheenen, T. W.; Futterer, J.; Weiland, E.; van Hecke, P.; Lemort, M.; Zechmann, C.; Schlemmer, H. P.; Broome, D.; Villeirs, G.; Lu, J.; Barentsz, J.; Roell, S.; Heerschap, A., Discriminating cancer from noncancer tissue in the prostate by 3-dimensional proton magnetic resonance spectroscopic imaging: a prospective multicenter validation study. *Invest Radiol* **2011,** *46* (1), 25-33.

54. McKnight, T. R.; Lamborn, K. R.; Love, T. D.; Berger, M. S.; Chang, S.; Dillon, W. P.; Bollen, A.; Nelson, S. J., Correlation of magnetic resonance spectroscopic and growth characteristics within Grades II and III gliomas. *J Neurosurg* **2007,** *106* (4), 660-6.

55. Zeng, Q.; Liu, H.; Zhang, K.; Li, C.; Zhou, G., Noninvasive evaluation of cerebral glioma grade by using multivoxel 3D proton MR spectroscopy. *Magn Reson Imaging* **2011,** *29* (1), 25-31.

56. Janfaza, S.; Khorsand, B.; Nikkhah, M.; Zahiri, J., Digging deeper into volatile organic compounds associated with cancer. *Biol Methods Protoc* **2019,** *4* (1), bpz014.

57. Amaro, F.; Pinto, J.; Rocha, S.; Araujo, A. M.; Miranda-Goncalves, V.; Jeronimo, C.; Henrique, R.; de Lourdes Bastos, M.; Carvalho, M.; de Pinho, P. G., Volatilomics Reveals Potential Biomarkers for Identification of Renal Cell Carcinoma: An In Vitro Approach. *Metabolites* **2020,** *10* (5).

CHAPTER 4

APPLYING TREE-BASED SHAPLEY ADDITIVE EXPLANATIONS TO METABOLOMICS DATASETS[3]

[3] Olatomiwa O. Bifarin. Submitted to *Metabolites*, 06/17/2021

Contributing Author

Olatomiwa O. Bifarin conducted all computational works.

4.1 Abstract

Partial least square discriminant analysis (PLS-DA) and its variants are used widely in metabolomics, primarily due to the model's interpretability with the Variable Influence in Projection (VIP) scores. As such, unexplainable (black box) models with potentially higher accuracy are used less in metabolomics studies. Shapley Additive explanations (SHAP), a machine learning method grounded in a game theory, can explain ML models with local explanations of individual samples. This study shows that metabolomics scientists can use Tree SHAP as a model machine learning interpretable algorithm using tree-based models. The tree-based algorithm used includes random forests and extreme gradient boosting (XGBoost). Machine learning experiments (binary classification) were conducted with four published metabolomics datasets using the python programming language. It was observed that PLS-DA is less accurate than tree-based models as a classification algorithm after feature size reduction. The tree-based model was explained with Tree SHAP using plots like the summary plot for global explanations, waterfall plot for local explanations, and dependence plots for feature effects. Thus, metabolomics scientists should not be restricted to PLS-DA solely due to model interpretability.

4.2 Introduction

Metabolomics aims to report a global snapshot of a biological system's metabolic status.[1] The adequate tools to answer this question give rise to large datasets inherently. Therefore, the need to engage in classical and modern statistics, with machine learning being an

example of the latter. Classical statistics aim to draw inferences about the population from a sample, while machine learning (ML) finds a generalizable predictive pattern in a dataset without explicit instructions.[2-3] ML is used in experimental metabolomics workflows to find predictive patterns in data, for example, for biomarker discoveries.[4-5] However, the interpretation of these predictive models has mainly been limited to the use of partial least squares discriminant analysis (PLS-DA) in metabolomics.[6-7]

PLS-DA, also known as a projection on latent structures, combines features from principal component analysis (PCA) and multiple linear regression.[8] It extracts latent variables, the best predictors, from the independent variables and project results to a lower-dimensional space, as in PCA. PLS has been the standard multivariate analysis algorithm used in metabolomics for two main reasons. One, given the structure of the metabolomics dataset, a large number of features *vs.* smaller sample sizes, latent variables' projection onto a smaller dimensional space allows for its utility as a feature selection algorithm. Two, the linear regression structure inherent with the algorithm makes for an interpretability method, the variable importance in projection (VIP) scores.[9] As such, PLS-DA is an interpretable ML algorithm that models the linear latent covariance with the feature's matrix ($\boldsymbol{X}$) and the response matrix ($\boldsymbol{Y}$). Despite the popularity of PLS-DA in metabolomics, the linear relationship assumption and the global interpretability it affords are apparent limitations.

Some of the best-performing machine learning methods are notoriously black box models[10-11] i.e., models that are not interpretable. Linear models like linear regression and PLS-DA are interpretable because of their linear assumptions. In general, intrinsically interpretable models are so because of their simple structures, lending themselves to

features such as the weights in linear models and the learned tree structure in decision trees. However, biological data can have non-linear relationships,[12] which might require models to learn more complex relationships in such datasets for best performance. One approach to explain complex, black-box models is to apply interpretation methods after machine learning modeling. These methods are called post-hoc explainable artificial intelligence (XAI) methods, and they tend to be model agnostic.[13]

Additionally, despite the interpretability property of intrinsically interpretable models, they are usually limited to global explanations – explaining the entire model behavior, rather than local explanations that explain individual predictions. In general, a ML interpretable method with local and global interpretations (high representativeness) is indeed desirable. Other desirable properties of explanation methods include high expressive power (the *language* of expression), low algorithmic complexity, and high fidelity (the accuracy of the interpretation).[14] There are several XAI method like partial dependence plot (PDP),[15] individual conditional expectation (ICE),[16] accumulated local effects (ALE),[17] permutation feature importance,[18-19] and local interpretable model-agnostic explanations (LIME).[20] However, only SHapley Additive exPlanations (SHAP) gives a solution that satisfies the quality of high representativeness, fidelity, and high expressive power.[21-22] Thus, it was the method of choice in the study. The method has been empirically verified,[23] and it has been applied to many fields of study, including medicine,[23-24] cheminformatics,[25-26] and ecology.[27]

In this chapter, 1) Shapley values and SHAP for tree-based models (TreeSHAP) were introduced, 2) the classification performance of PLS-DA and tree-based models (random forest and XGBoost) were compared using four clinical metabolomics data sets, and

finally, 3) the most accurate tree ensemble model was explained with the aid of Tree SHAP.[22]

4.3 Results

4.3.1 Shapley Additive Explanations

4.3.1.1 Shapley Values

In a standard metabolomics experimental workflow,[28] after metabolic measurements and appropriate spectral and data processing, the resultant data matrix can be used for modeling using machine learning (**Figure 4-1a**). Shapley Additive exPlanations (SHAP) allows for the local interpretations of predictions by showing each metabolomic feature importance score for each sample. Additionally, SHAP is used to derive an accurate global interpretation of the model, giving rise to its high representativeness as a post-hoc ML interpretability method (**Figure 4-1b**). SHAP is based primarily on Shapley values, a cooperative game theory method.[21] Developed by Lloyd Shapley, it is a fair and axiomatically unique method of attributing rewards from a cooperative game.[29] Where a *game* is a machine learning model, each metabolomic feature values are *players* in a *game*, and the predicted class membership of the sample is the *outcome* of the *game*; Shapley value gives a unique solution to fairly attribute the contributions of each *player* to the outcome of the *game*. The Shapley value defines the feature importance of feature value i in the equation below and **Figure 4-1c**:

$$\emptyset_i(val) = \frac{1}{N!}\sum_{S\subseteq\{x_1,\ldots,x_N\}\setminus\{x_i\}} |S|!\,(|N| - |S| - 1)!\,[val(S \cup \{x_i\}) - val(S)] \quad (4.1)$$

Where S is a subset of features in the model, $val(S)$ corresponds to the model output for S, N is the total number of features, and x is the feature values for the sample to be

explained, that is $x \triangleq \{x_1, \ldots, x_N\} \in \mathbb{R}^N$. In brief, the marginal contribution of x_i is given by $val(S \cup \{x_i\}) - val(S)$. Weights are assigned to these marginal contributions by the different ways the sub-set could be formed before the addition of x_i: $|S|!$ and after the addition of x_i: $(|N| - |S| - 1)!$. The summation over all the possible sets S is conducted, followed by the average $\frac{1}{N!}$.

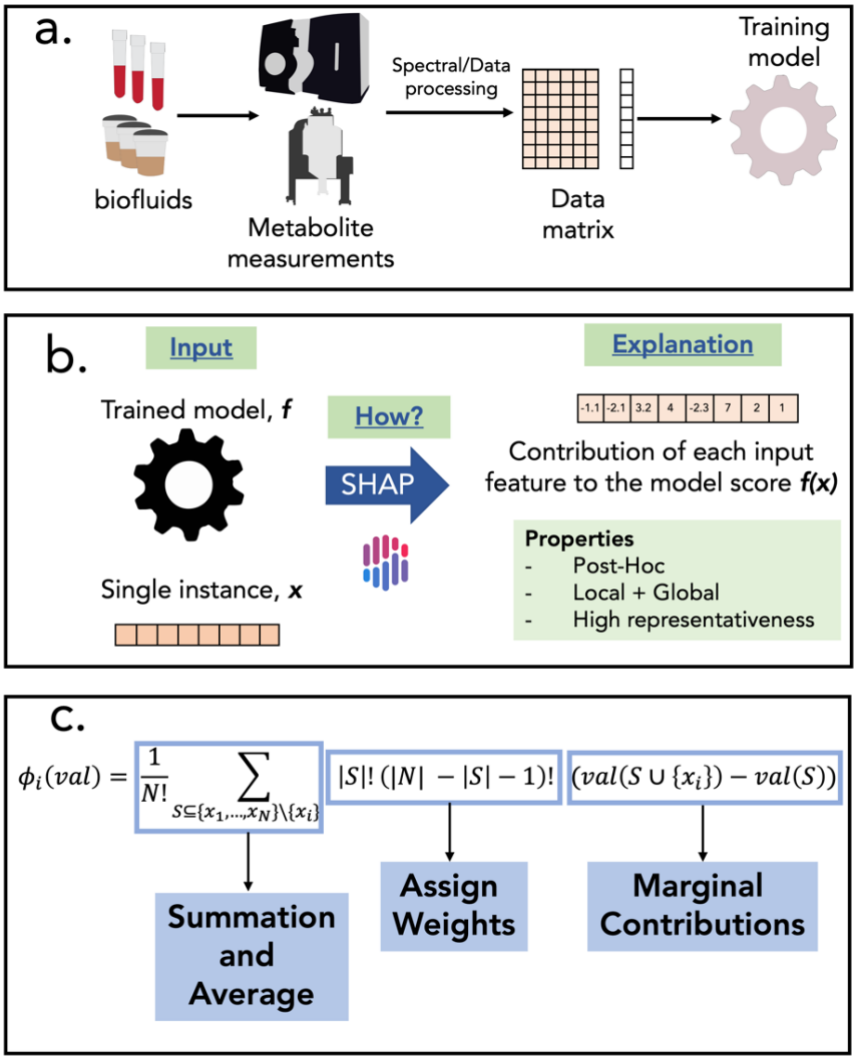


Figure 4-1. Metabolomics Workflow and SHAP Methodology.

a) A metabolomics workflow that culminates with model training for predictive or regression purposes. b) SHAP allows for local and global interpretations of model

predictions. Explanations are made locally, and because of the additivity property of Shapley values, the methods allow for global interpretations. c) A sample calculation of Shapley values of a feature x_i.

The Shapley value is a unique solution because it satisfies the axioms of symmetry (or consistency), dummy (or null effect), and additivity (or local accuracy) [29]. Symmetry implies that if the marginal contribution of the metabolomic feature values x_z and x_k is the same, the Shapley value attributed to each feature value will also be the same.

$$\begin{gathered} val(S \cup \{x_z\}) - val(S) \\ = val(S \cup \{x_k\}) - val(S)\ \forall\ S \subseteq \{x_1, \dots, x_N\} \backslash \{x_z, x_k\} \\ \Longrightarrow \emptyset_z(val) = \emptyset_k(val) \end{gathered} \tag{4.2}$$

Dummy implies that if a feature value x_z do not impact a model, the Shapley value attributed will be zero.

$$val(S \cup \{x_z\}) - val(S) = 0\ \forall\ S \subseteq \{x_1, \dots, x_N\} \Longrightarrow \emptyset_k(val) = 0 \tag{4.3}$$

Finally, local accuracy means that the summation of the Shapley values of all feature values in the model equals the model output. As such, the total contributions of all feature values will equal the impact of all feature values on the model output minus the impact with no feature value, mathematically expressed below:

$$\sum_{i \in N} \emptyset_i(val) = val(N) - val(\{\}) \tag{4.4}$$

4.3.1.2 Tree SHAP

The choice of tree-based SHAP ML interpretations for this study was based on the relatively low computational complexity of computing Shapley values and the exact

Shapley value that result from the computation.[22] However, other methods for approximately Shapley values for non-tree-based methods exist,[21] making it a model agnostic ML interpretability method. Two problems emanate from the computation of naïve Shapley values, namely, 1) handling of missing features when computing marginal contributions and 2) exponential computational time and algorithmic complexity. First, when computing Shapley values for a feature value, the absence of the feature must be defined while computing the marginal contributions. This is not straightforward in machine learning as opposed to game theory. The problem has been addressed in kernel SHAP[21] by simulating missing features *via* random sampling by replacing the missing value with a fixed value. This invariably creates a sampling-based estimation variance problem; on the other hand, kernel SHAP can be applied to any ML model. In Tree SHAP, exact Shapley values are calculated by ignoring the decision paths of the missing features in the tree, getting rid of sampling-based estimation variance in the process. Second, because Tree SHAP computes Shapley values by keeping track of tree transversals to prevent repetition[22], it gives rise to an efficient algorithm by reducing algorithmic complexity from exponential time ($O(TL2^N)$) to polynomial time ($O(TLD^2)$). Where T is the number of trees, L is the maximum number of leaves in any tree, N is the number of features, and D is the maximum tree depth.

4.3.2. Machine Learning Pipeline and Performance

ML models were built using four metabolomics datasets, as summarized in **Table 4-1** (See materials and methods for details). All problems were binary classification problems, and

since the goal of the study is to show the utility of SHAP, a simplified machine learning workflow was used (**Figure 4-2**).

Table 4-1. Metabolomics Datasets Used in the Interpretable Machine Learning Study.

Dataset	**MTBLS404**	**MTBLS547**	**ST000369**	**MTBLS161**
Analytical platform	LC-MS	LC-MS	GC-MS	NMR
Sample Type	Urine	Caecal	Serum	Serum
Sample Size	184	97	80	59
Subject of classification	Gender	High-fat diet	Adenocarcinoma	Chronic fatigue syndrome
Classes (size)	Male/Female (101/83)	Case/Control (46/51)	Case/Control (49/31)	Case/Control (34/25)
Metabolomic feature number	184	42	69	29
Publication	Thevenot et al. 2015[30]	Zheng et al. 2017[31]	Fahrmann et al. 2015[32]	Armstrong et al. 2015[33]

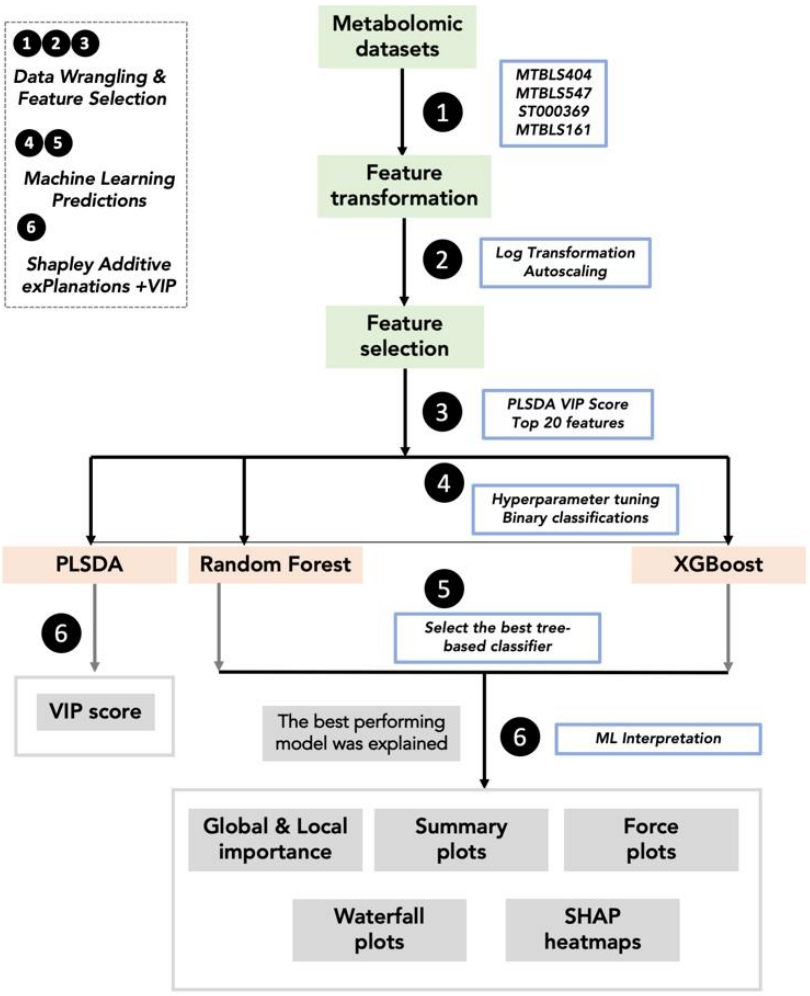


Figure 4- 2. Machine Learning Pipeline for Machine Learning Explanations.

PLS-DA: Partial Least Square Discriminant Analysis; XGBoost: Extreme Gradient Boosting; VIP: Variable Importance in Projection.

The selected metabolomic datasets were prepared for classification tasks *via* a log transformation followed by autoscaling. Samples were split into train and test set with 5/6 and 1/6 of the sample size respectively in each set, and PLS-DA algorithm was used to select the best-performing 20 features *via* its VIP score. Afterward, the selected features were used to build machine learning models using PLS-DA, random forest, and XGBoost.

Appropriate hyperparameters were tuned (See materials and methods and **Table 4-3**), and the best tree-based classifier was used for shapley additive explanations. The machine learning performance results of both the baseline and tuned ML method are presented in **Table 4-2**. Random forests gave the best predictive scores on all the datasets (test set) with an AUC of 0.90 for MTBLS404, 0.88 for MTBLS547, 0.74 for ST000369, and 0.67 for MTBLS161. An XGBoost model performed as well as random forests on the MTBLS547 dataset with an AUC of 0.88, while PLS-DA was as predictive as random forests on the MTBLS161 dataset with an AUC of 0.67. As such, the MTBLS404 dataset will be used to illustrate the utility of SHAP in this study because it has the highest AUC score. The MTBLS404 dataset is from a urine metabolomics study that investigated human adult urine for variations in age, BMI, and gender.[30] In this analysis, the dataset has been used for gender classifications.

Table 4-2. Machine Learning Performance for the Interpretable Machine Learning Study.

Model	**MTBLS404**	**MTBLS547**	**ST000369**	**MTBLS161**
PLS-DA	0.85±0.08	0.83±0.14	0.80±0.16	0.83±0.16
Baseline	(0.80)	(0.82)	(0.69)	**(0.67)**
PLS-DA	0.85±0.08	0.87±0.14	0.80±0.16	0.81±0.16
Tuned	(0.80)	(0.82)	(0.69)	**(0.67)**
RF	0.93±0.06	0.94±0.09	0.91±0.08	0.91±0.17
Baseline	(0.87)	(0.82)	(0.63)	**(0.67)**
RF	0.93±0.06	0.92±0.08	0.92±0.08	0.91±0.17
Tuned	**(0.90)**	**(0.88)**	**(0.74)**	**(0.67)**
XGB	0.90±0.08	0.93±0.09	0.79±0.19	0.83±0.17
Baseline	(0.77)	**(0.88)**	(0.69)	(0.62)
XGB	0.93±0.06	0.95±0.08	0.84±0.22	0.87±0.16
Tuned	(0.84)	(0.82)	(0.69)	(0.62)

Baseline models use the default hyperparameters in the sci-kit learn library for the python programming language. The tuned models underwent hyperparameter tuning as described in the materials and methods. Predictive scores are the Area Under the Receiver Operating Characteristic Curve (ROC AUC). The training score reports the mean±standard deviation of ROC AUC under 10-fold cross-validation conditions. The test set performance scores are reported in brackets below the train set scores, and the highest test set scores for each dataset are shown in bold texts and underlined. PLS-DA: Partial Least Squares-Discriminant Analysis; RF: Random Forests; XGB: Extreme Gradient Boosting.

4.3.3 Model Interpretations with SHAP

SHAP explanations were computed for the test set of the MTBLS404 dataset trained using random forest with an AUC of 0.90, and these explanations are presented in this section. The PLS-DA VIP score plot, a global interpretation of the PLS-DA model, is shown in **Figure 4-3a**. Likewise, **Figure 4-3b** shows the SHAP bar plot, displaying the mean of the absolute SHAP values (mean[|SHAP value|]), that is, the average impact of the feature on the model output magnitude. Testosterone glucuronide and *p*-anisic acid have the same feature importance rank on both plots. In addition, the Pearson correlation coefficient of the VIP score and the mean[|SHAP value|] is 0.84 (**Figure 4-3c**), indicating high explanation similarities. On the other hand, the Pearson correlation coefficient of Gini importance, and mean[|SHAP value|] is 0.99 (**Figure 4-3d**), confirming the fidelity of Tree SHAP in computing the feature importance. Gini importance is an intrinsic global interpretation of random forests; therefore, it gives the metabolomic feature importance attribute for the ML algorithm. Finally, to confirm that SHAP identified features that are

important for the predictive accuracy of the model, features were eliminated in the order of feature ranking, first, one at a time (**Figure 4-3e**) then, four features at a time (**Figure 4-3f**). Afterward, the random forest model was used to predict the test set using the reduced feature sets. When features are removed one at a time, in general, there was a downward trend of AUC, while the intervening undulating trend might be partly due to the closeness of SHAP values of some of the metabolomic features. However, an uninterrupted downward trend is reported when features are removed in fours.

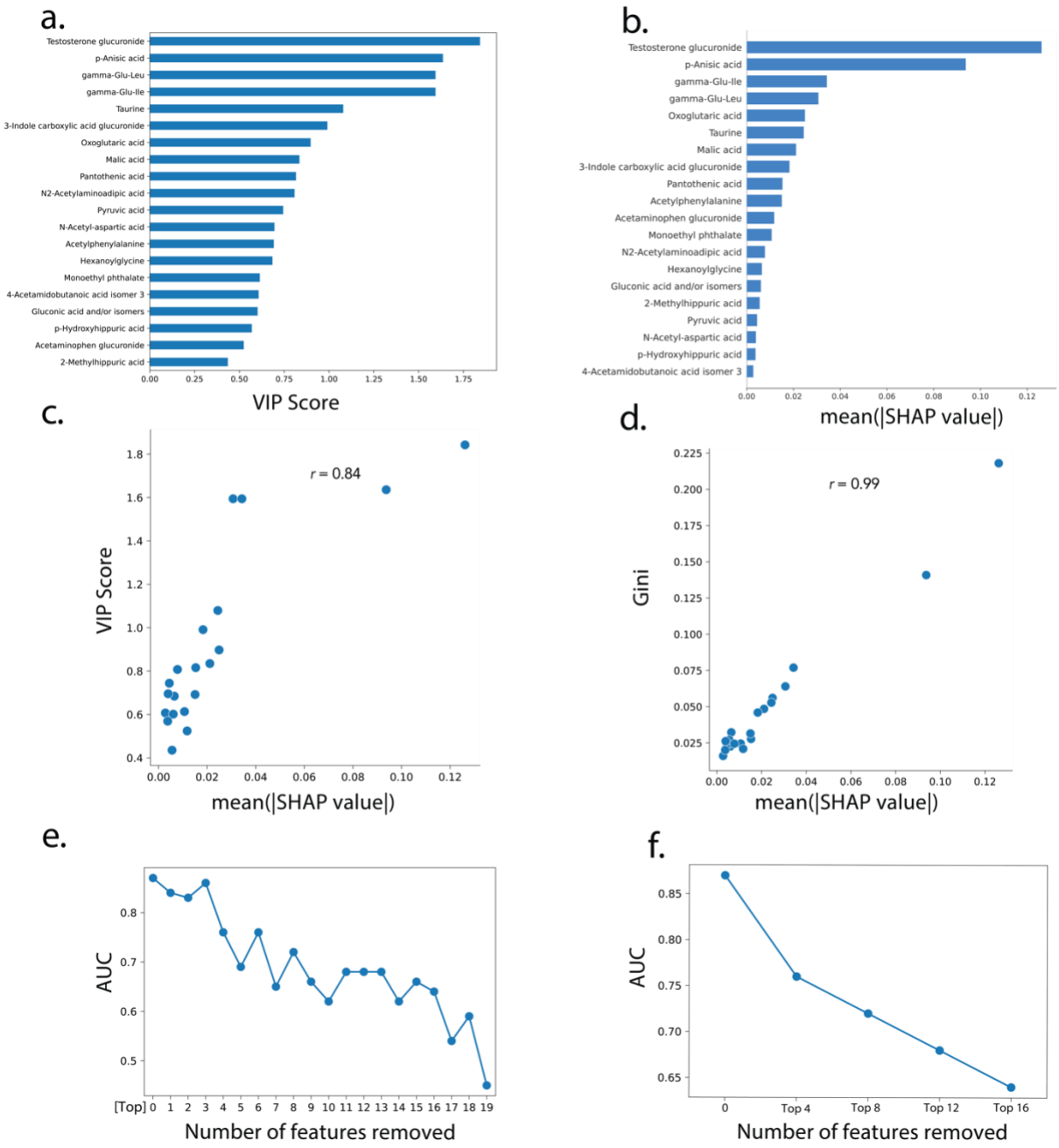


Figure 4-3. Global Feature Importance and Feature Importance Correlations.

a) PLS-DA VIP score plot. b) SHAP bar plot. c) Scatterplot of the VIP score and the mean(|SHAP value|) with a Pearson's correlation coefficient of 0.84. d) Scatterplot of the Gini importance score and the mean(|SHAP value|) with a Pearson's correlation coefficient of 0.99. e) AUC of the test set after the removal of the top rank features sequentially as

ranked by the mean absolute SHAP values. f) AUC of the test set after the removal of the top 4, 8, 12, and 16 features as ranked by the mean absolute SHAP values.

In addition, because of the local representativeness property of SHAP, it can give both the global importance score and an explanation of individual predictions in the SHAP summary plot (also called the beeswarm plot), enabling a richer visual summarization, as shown in **Figure 4-4a**. In the plot, metabolomic features are arranged in a descending order based on relative importance $\sum_{j=1}^{N} |\emptyset_i^{(j)}|$, where $\emptyset_i$ is the Shapley value of feature i, j is a sample, and N is the total number of samples. Each dot in the summary plot represents a sample plotted against its impact on the model output $\emptyset_i^{(j)}$. The color of each sample represents the relative abundance of the metabolites, ranging from low (blue) to high (red). **Figure 4-4b** displays the most important metabolite in the panel – testosterone glucuronide. High feature values of the metabolite tend to have positive SHAP values, which drives the model to predict males; on the other hand, low feature values of testosterone glucuronide tend to have negative SHAP values, which drives the model to predict the female class. *p*-anisic acid has an opposite trend, as shown in **Figure 4-4c**.

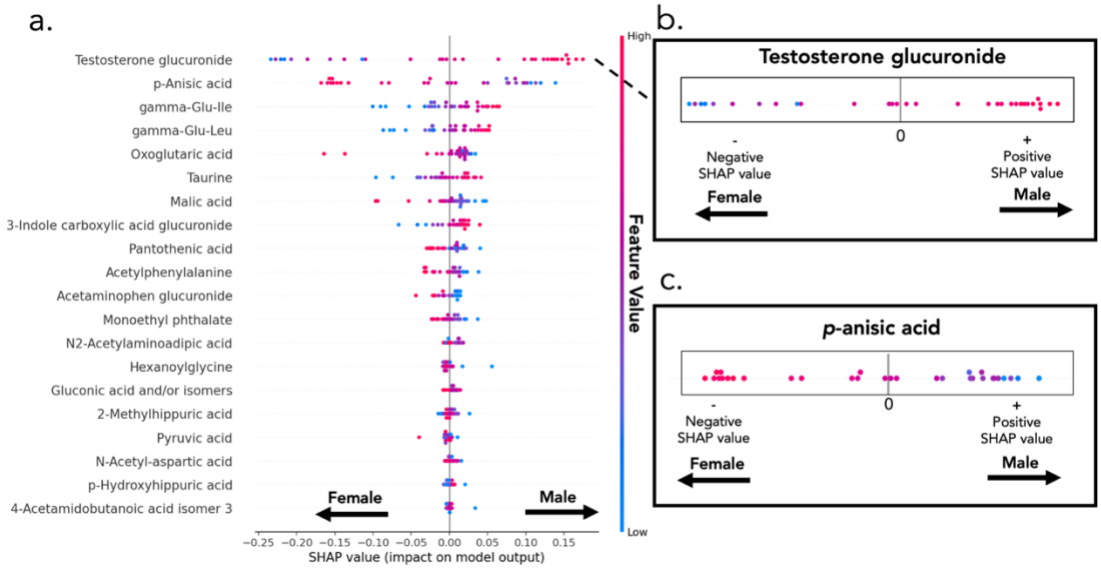


Figure 4-4. SHAP Summary Plot for Explaining the Gender Classification.

a) SHAP summary plot. b) Illustration with testosterone glucuronide. c) Illustration with *p*-Anisic acid.

Unsupervised clustering techniques are widely used in metabolomics studies to identify groups of classes that cluster together. Because such analyses rely on raw data, which are only processed by a simple standardization process, it is impossible to cluster samples based on a prediction outcome. On the other hand, SHAP values can be used to generate supervised clustering, where samples are clustered based on the same prediction outcome. For example, in the heatmap shown in **figure 4-5**, the hierarchical supervised clustering analysis clustered samples based on the decreasing order of the model output, that is, from male to female predictions. The heat map shows that testosterone glucuronide has the highest absolute SHAP values associated to each sample in the test set, when compared to other metabolomic features. Hence, justifying why testosterone glucuronide is the most important metabolomic feature.

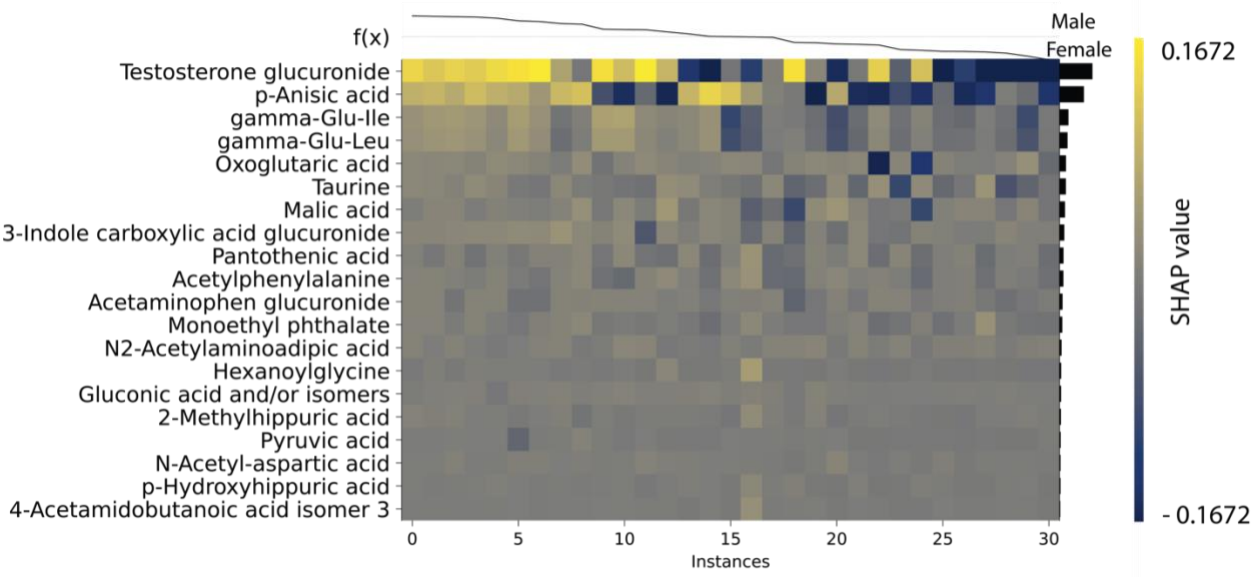


Figure 4-5. Supervised Interpretable Hierarchical Clustering of SHAP values for Explaining the Gender Classification.

Samples are displayed on the x-axis, while features are arranged in ascending order of importance on the y-axis. *f(x)* indicates the prediction outcome, with the line plot over the dotted line indicating male predictions, while the line plot below the dotted line indicates female prediction. The bar plot represents the mean absolute SHAP value, the average impact on the model output.

Just like the partial dependence plot (PDP) shows the marginal effect that one or two features have on a model output with a line plot,[15] SHAP values can be used to create a better alternative graph called the SHAP dependence plot. The plot shows how the relative abundance of a metabolite changes with the respective impact on the model output (**Figure 4-6**). In addition, dots representing the samples can be colored by the relative

abundance of another metabolite to capture the interaction effect if it exists (**Figure 4-6b**). **Figure 4-6a** displays the relationship between testosterone glucuronide and the SHAP values for the metabolite (the impact of the metabolite on model output), while **Figure 4-6b** adds γ-glu-leu to the plot by coloring the samples by the relative abundance of γ-glu-leu. The S-like curve shows that higher feature values of testosterone glucuronide and γ-glu-leu, tend to lead to a male prediction.

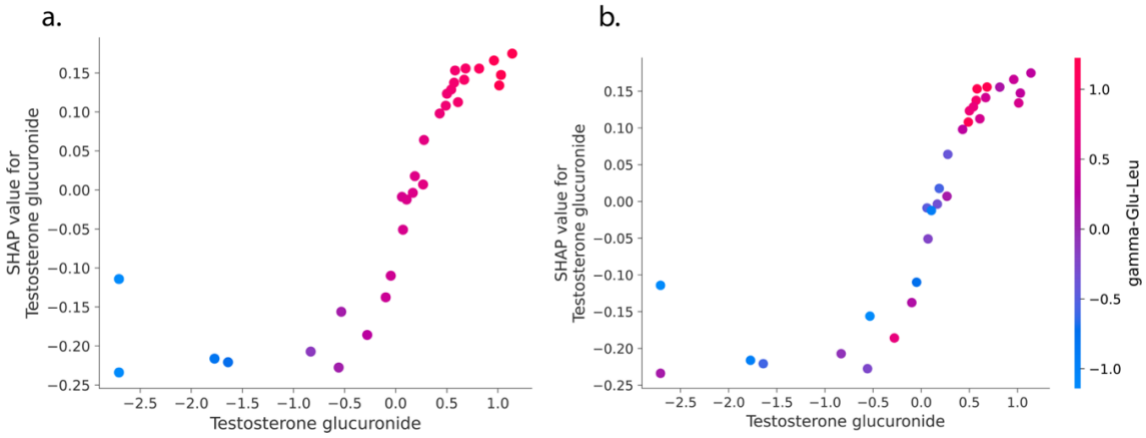


Figure 4-6. SHAP Dependence Plot of Testosterone Glucuronide and γ-glu-leu.
a) testosterone glucuronide. b) testosterone glucuronide and γ-glu-leu.

Finally, SHAP provides many avenues for local explanations (individual sample explanations), such as the waterfall plot and the force plot visualizers. An example of waterfall plot and force plot are shown in **Figures 4-7a** and 4-**7b,** respectively. In the waterfall plot (**Figure 4-7a**), the x-axis represents the probability of a sample being classified as male, while the y-axis shows the metabolomic features and the respective feature values for the sample. Waterfall plots begin with the expected value of the model output on the x-axis ($E[f(X)] = 0.55$). The base value, 0.55, is the average prediction probability over the test set. The plot displays the impact of the metabolomic features on

the model output. The combination of the positive contributions (in red) and the negative contributions (in blue) moves the expected value output to the final model output ($f(x) = 0.65$). Positive SHAP values increase the probability of the sample being classified as male, while negative SHAP values decrease the probability of the sample being classified as male. Similarly, the force plot helps visualize each feature's SHAP values for a given sample with an additive force layout (**Figure 4-7b**).

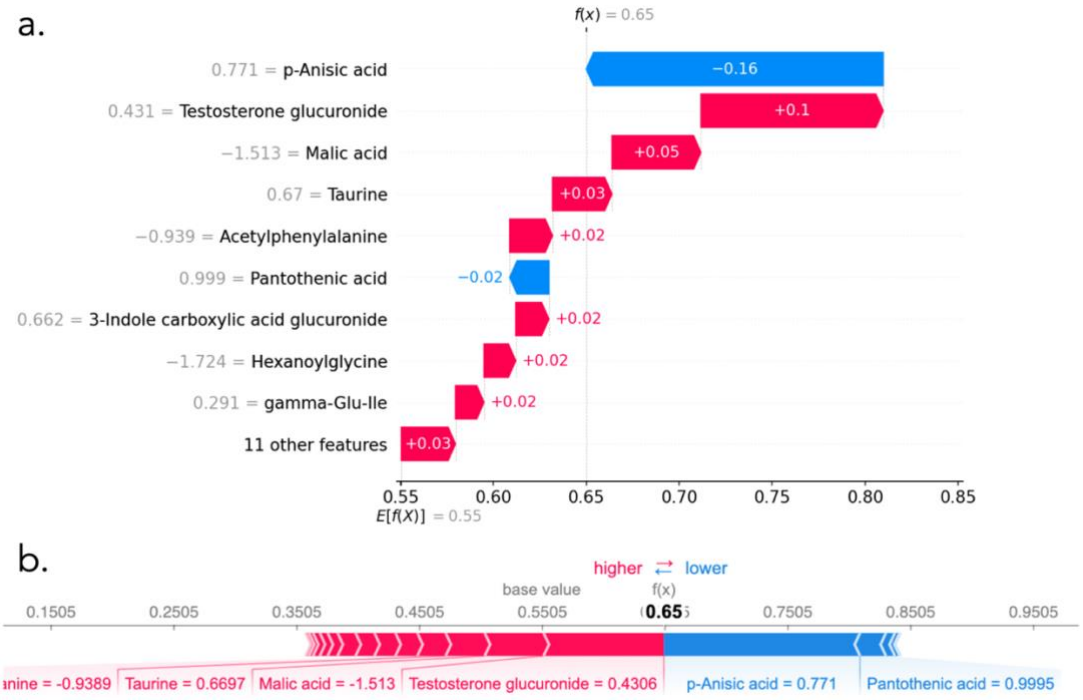


Figure 4-7. Local Explanations of a Representative Sample Predicted as Male.
a) Waterfall plot. b) Force plot.

One of the utilities of the SHAP local explanations is the error analysis of individual predictions. The confusion matrix of test set predictions is shown in **Figure 4-8a**, with a true negative rate of 38.71%, a false positive rate of 6.45%, a false negative rate of 3.23%, and a true positive rate of 51.61%. The waterfall plot of two representative true positive

samples (male samples successfully classified as males) are shown in **Figures 4-8b & c** with a probability output of 0.98 and 0.88, respectively. As in the global explanation, the importance rank of testosterone glucuronide, *p*-anisic acid, γ-Glu-Ile, and γ-Glu-Leu, are conserved. However, the only false negative sample (male sample wrongly classified as female) in the test cohort can be seen to be missing testosterone glucuronide (**Figure 4-8d**), a highly ranked metabolomic feature of importance in the true positives, providing a plausible explanation for the model's wrong classification. In that vein, two representative true negative samples (female successfully classified as female) are shown in **Figures 4-8e & f**. A low relative abundance of testosterone glucuronide is the most important factor driving the samples' classification to the female category. Likewise, for the two false-positive samples (female sample wrongly classified as a male sample), one of such samples (**Figure 4-8g**) has a high relative abundance of testosterone glucuronide, contributing to the wrong classification. On the other hand, for the second false positive sample (**Figure 4-8h**), which has a low relative abundance of the testosterone metabolite, the absolute SHAP value was not high enough to drive the classification to the female category below the 0.5 probability cut-off (the model output was 0.56).

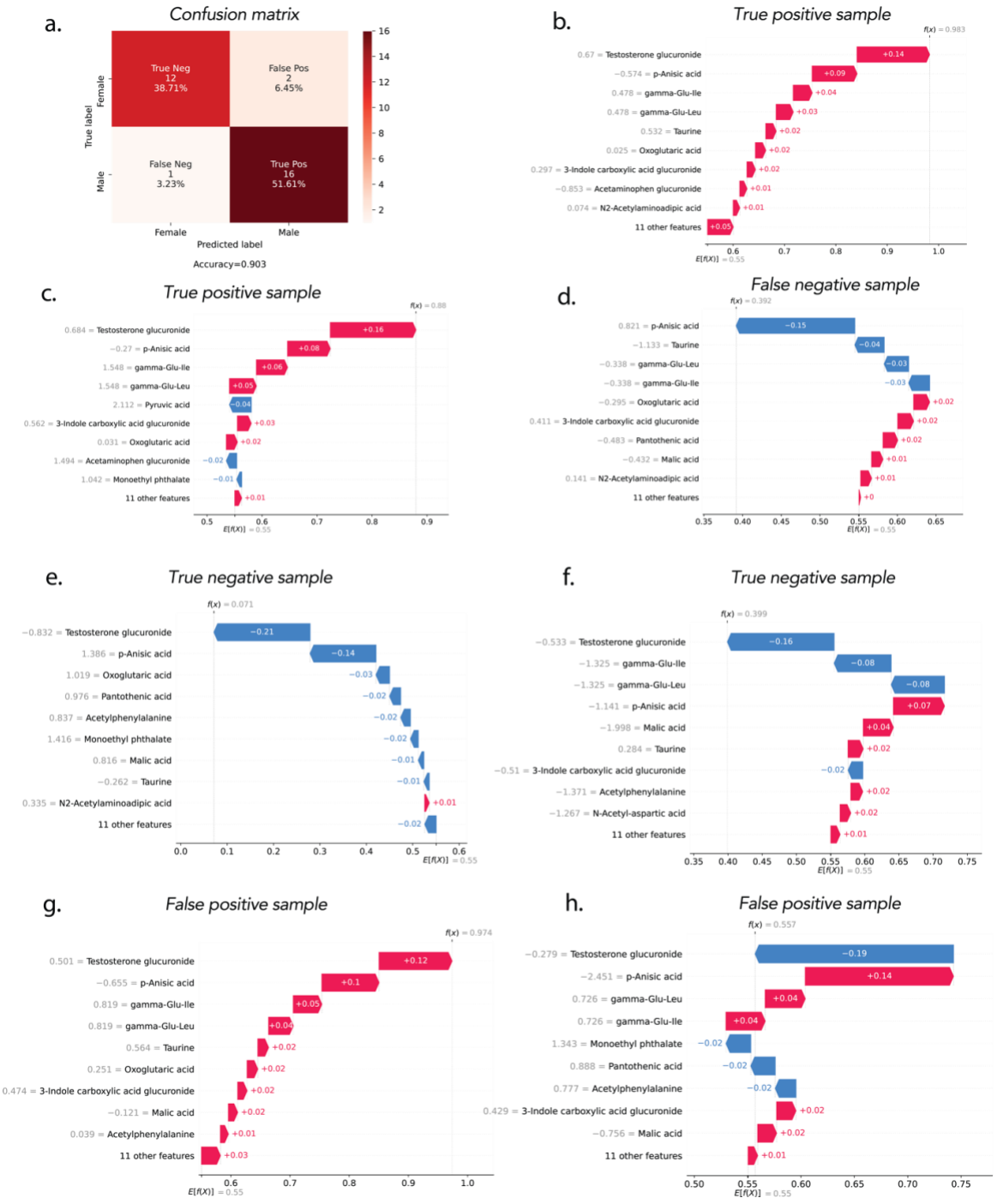


Figure 4-8. SHAP for Error Analysis of the Gender Classification.

a) Confusion matrix of the test set. **b) & c)** Waterfall plot of true positive representative samples. **d)** Waterfall plot of the only false negative sample. **e) & f)** Waterfall plot of the

true negative representative samples. **g) & h)** Waterfall plot of the two false positive samples.

4.4 Discussion

One underlying hypothesis in this study is that non-linear models could have higher predictive power compared to linear models. In our study, after feature selection with PLS-DA, we compared random forest and xgboost with the most widely used ML algorithm in metabolomics (PLS-DA) using publicly available metabolomics datasets to test the hypothesis. The hypothesis was confirmed suggesting that tree models should be more applied in metabolomics for classification purposes. However, it has been showed that PLS-DA outperformed random forest using the same datasets in this study before feature selection.[34] This might be because tree-based models like random forest perform poorly when the fraction of relevant features is small compared to a large number of features for a relatively small sample size.[35] Furthermore, this paper has shown that there are added benefits to use a tree-based model for classification tasks.

In addition to the compatibility of PLS-DA methods to metabolomics datasets, the interpretability of PLS-DA *via* its VIP score is useful to scientist in the metabolomics field for explaining the model.[7] As such, PLS-DA is appealing because it is intrinsically interpretable. However, over the past several years, there had been great stride of achievements in the field of explainable AI, especially on model agnostic methods,[36] one of such methods is local interpretable model-agnostic explanations (LIME).[20] The key idea of LIME is that it selects an intrinsically interpretable class of model, and then used that to approximate a black box model locally, therefore interpretations are not globally faithful. This class of models are called local surrogate models. SHAP unifies LIME and other

interpretable machine learning methods with Shapley values to explain both local and global properties leveraging on the unique properties of Shapley values.[21] This SHAP computation is called kernel SHAP, and it is computationally expensive and does not calculate exact Shapley values hence the choice of Tree SHAP for the study. On the other hand, Tree SHAP calculates exact Shapley values *via* its conditional expectation method, also it is relatively computationally inexpensive in comparison with kernel SHAP.[22]

The VIP score and the mean absolute SHAP values (as shown in SHAP bar plot) of the MTLB404 dataset indicates a consensus in the top two most important metabolomic features – testosterone glucuronide and *p*-anisic acid. In addition, γ-glu-ile, and γ-glu-leu follows in both explanations albeit with different ranks. In general, there is no corresponding match between both global explanations. This is expected because the models have different predictive performances: PLS-DA with an AUC of 0.80 *vs.* random forest with an AUC of 0.90. However, a high explanation similarity exists with a Pearson's correlation score of 0.84 between VIP score and mean absolute SHAP values of the MTLB404 test set. Also, importantly the same predictive performances might not always imply the same explanations, due to the Rashomon effect.[37] Consistency in explanations should only be desirable if the models rely on the same relationships to make predictions. Going beyond a bar plot of global importance, the summary plot allows a more detailed explanations of the global importance by showing the impact of the abundance levels of metabolites on the model output. In addition, the supervised interpretable hierarchical clustering analysis clustered the model output, allowing for the isolation of testosterone glucuronide as an important marker for the male gender. Which is indeed correct as the metabolite is an endogenous, urinary metabolite of testosterone, the principal male sex

hormone.[38] In the original study that investigated the impact of age, body mass index, and gender on human urinary metabolome, testosterone glucuronide was identified to be strongly associated with the male sex.[30] This observation is further corroborated in this study *via* the error analysis enabled by the SHAP local interpretation plots, with the only false negative sample (male classified wrongly as female) in the test set conspicuously missing testosterone glucuronide as an important feature responsible for the wrong female classification. In addition, testosterone glucuronide was also implicated in the false positive samples.

As shown in this study, SHAP is a desirable post-hoc interpretability method that will enhance the interpretations of metabolomics data analysis, given its local and global interpretations, high expressive power, and high fidelity.[14] While SHAP explains predictions, it is important to note that explanation does not imply causation. This is a general limitation of models and data.

4.5 Materials and Methods

4.5.1 Metabolomics Datasets

All the datasets used in this study were deposited on *Metabolomics Workbench* (www.metabolomicsworkbench.org)[39] and *Metabolights* (www.ebi.ac.uk/metabolights) [40-41] repositories. MTBLS161, MTBLS404, and MTBLS547 datasets are from *Metabolomics Workbench*, while ST000369 is from *Metabolights*. Mendez et al 2019 has used these datasets in a computational metabolomics study, where generalized predictive ability of machine learning algorithms was compared.[34] The datasets had been converted to tiny data format in the study for ease of computational analysis. As shown in **Table 4-1**, biofluids used in the study selected include urine, caecal, and serum. The sample size ranges from

59 to 84, and the metabolic features range from 29 to 184. All machine learning problems are binary classification, and the subject of classification include gender classification (male *vs.* female),[30] impact of high fat diet (case *vs.* control),[31] adenocarcinoma (case *vs.* control),[32] and chronic fatigue syndrome (case *vs.* control).[33]

4.5.2 Machine Learning Models

Partial least squares (PLS) regression models a linear covariance between the feature matrix $\boldsymbol{X}$ and the response matrix $\boldsymbol{Y}$**.** The goal of the algorithm is to predict dependent variables using predictors, and it does so by projecting the data points into a lower dimensional space such that the covariance between response groups are maximized. The projection can be represented mathematically as a PLS coefficient value vector (B_{PLS}) where predictions are made by $\hat{Y} = XB_{PLS}$. Which is, in essence, a multiple linear regression equation. PLS discriminant analysis (PLS-DA) is the PLS method for binary classification, where $\hat{Y} < 0.5$ is attributed to a negative classification and $\hat{Y} > 0.5$ is a positive classification.

Decision trees (DT) are the base learner for tree-based machine learning algorithms. DTs are inverted trees with the root node at the top, the leaves at the bottom, and the internal nodes in between. Root and internal nodes are assigned metabolomic features used for splitting, while the leaves at the end of the tree signify the final predictions. In brief, decision trees split data into branches until the algorithm attain the highest accuracy. This property of DT makes it particularly prone to overfitting, and one of the robust solutions to this problem is the aggregation of decision trees into ensembles such as random forests and extreme gradient boosting (XGBoost).

Random forest is a bootstrap aggregation (bagging) of decision trees. Each decision trees make its predictions, and the result of the forest of decision trees are made using the majority rules. In the end, bootstrapping increases diversity, while aggregation reduces variance. Despite the strength of random forests, they could be limited by individual trees. If all the trees in the forest make the same mistake, the random forest, in turn, makes that mistake. However, boosting improves on this limitation by making the trees learn from the mistake of the preceding trees; it is this class of tree-based models XG-Boost belongs.

XGBoost is an advanced variant of gradient boosting that transforms weak learners into strong ones via the summation of trees' residuals. XGBoost is designed primarily for speed with its speed-enhancing capabilities like approximate and sparsity-aware split-finding, block compression and sharding, parallel computing, and cache-aware computing.[42] Also, there can be accuracy gains to using it over gradient boosting and random forests because of the inclusion of regularization in the loss function.[42] The regularization parameters help to penalize complexity and prevent overfitting.

4.5.3 Hyperparameter Fine-Tuning

To attempt to improve predictive performance and find a balance between bias and variance, hyper-parameter fine-tuning was carried out. As opposed to parameters learned by the machine learning algorithm, hyper-parameters are set by the user. A linear search for a single hyperparameter or a grid search for two or greater hyperparameters was carried out under 10-fold cross-validation conditions. In PLS-DA, a linear search was conducted for the number of latent variables (*n_components*), the only hyperparameter. For random forest, hyperparameters considered for tuning include the number of trees in the forest

(*n_estimators*), the maximum length of trees (*max_depth*), the maximum number of features to choose from when making a split (*max_features*), the number of samples required before the split can occur (*min_samples_split*), and the minimum number of samples required for a node to be a leaf (*min_samples_leaf*). For XGBoost, hyperparameter considered include maximum depth of tree (*max_depth*), that is the number of branches a tree has; learning rate (*learning_rate*), which limits overfitting by reducing the weight ascribed to each tree to a given percentage; number of boosted trees in the model (*n_estimators*); *subsample*, which limits the percentage of training samples for each boosting round, and the minimum sum of weight required for a node to split into a child (*min_child_weight*).

Table 4-3. Hyperparameters Tuned for PLS-DA, Random Forest, and XGBoost for the Interpretable Machine Learning Study.

Technique	Parameter	Search Space
PLS-DA	*n_components*	[1, 2, 3, 4, 5, 6]
Random Forest	*n_estimators*	[50, 100, 150, 200]
	max_depth	[10, 20, 30]
	max_features	[‘auto’, ‘sqrt’, ‘log2’]
	min_samples_split	[2, 4, 6, 8]
	min_samples_leaf	[1, 2, 3, 4, 5]
XGBoost	*subsample*	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
	min_child_weight	[1, 2, 3, 4, 5]
	learning_rate	[0.1, 0.2, 0.3, 0.4, 0.5]
	max_depth	[1, 2, 3, 4, 5, None]
	n_estimators	[2, 25, 50, 75, 100]

4.5.4 Computational Libraries

Programming was carried out using the Python 3.7.0 programming language. SHAP 0.39.0 library was used for SHAP explanations.[21, 23-24] Pandas 1.1.5 was used for data handling.[43] Matplotlib 3.3.4 and Seaborn 0.11.1 was used for data visualization.[44-45] Numpy 1.19.2 was used for numerical computation and Sci-kit learn 0.24.2 was used for machine learning. Jupyter notebook was used as the integrated development environment (IDE).

4.6 Explaining ML predictions for RCC Detection Study

In Chapter 2, a seven-biomarker panel for RCC detection was presented. A new random forest model was trained in this session, and the predictions were explained using Tree SHAP. In the RCC detection study presented in chapter 2, the small size of the training dataset (healthy controls=31, RCC=31) *vs.* a relatively large test cohort (healthy controls=143, RCC=51) was used because of the patient selection constraint. This data distribution is not optimal for machine learning experiments, given the property that a machine learning model should learn more with experience (*i.e.*, more data). Therefore, the performance scores are likely to be conservative.

However, since we have shown that the urine collection location does not affect the robustness of the selected markers, the following ML strategy was deployed: 1) The model and the test cohort were merged into a single data frame. 2) Samples were randomly split into 80% training set (n=204) and 20% hold-out test set (n=52). 3) Random forests hyperparameters were tuned using a grid search (See **Table 4-4** for hyperparameters tuned), and models were trained under stratified 5-fold cross-validation conditions. 4) Classification was carried out on the test set. 5) Tree SHAP was used to explain the model predictions using the test set.

Table 4-4. Initial Distribution and Optimized Random Forest Hyperparameters for the RCC Detection Explanation Study

Parameters	Initial distribution	Optimized
Max_depth	10, 20, 30	10
Max_features	'auto', 'sqrt', 'log2'	'auto'
Min_samples_leaf	1, 2, 3, 4, 5	1
Min_samples_split	2, 4, 6, 8	2
N_estimators	50, 100, 150, 200	100

Table 4-5. Random Forests Performance Scores for the RCC Detection Explanation Study.

AUC	0.98±0.02 (0.95)
Accuracy	0.96±0.02 (0.92)
Sensitivity	0.89±0.08 (0.83)
Specificity	0.99±0.01 (0.97)

Training scores are presented as mean±standard deviation, while the hold-out test set scores are shown in brackets.

Figure 4-9 shows the barplot (a), summary plot (b), and the supervised hierarchical clustering of the explanatory SHAP values. The global importance plots show that the dipeptide lys-ile/lys-leu as the most important metabolite driving the differences between RCC and healthy control groups. In addition, the supervised clustering indicates that – more often than not – an RCC prediction is associated with positive SHAP values of lys-ile/lys-leu (high relative abundance of lys-ile/lys-leu) and vice versa.

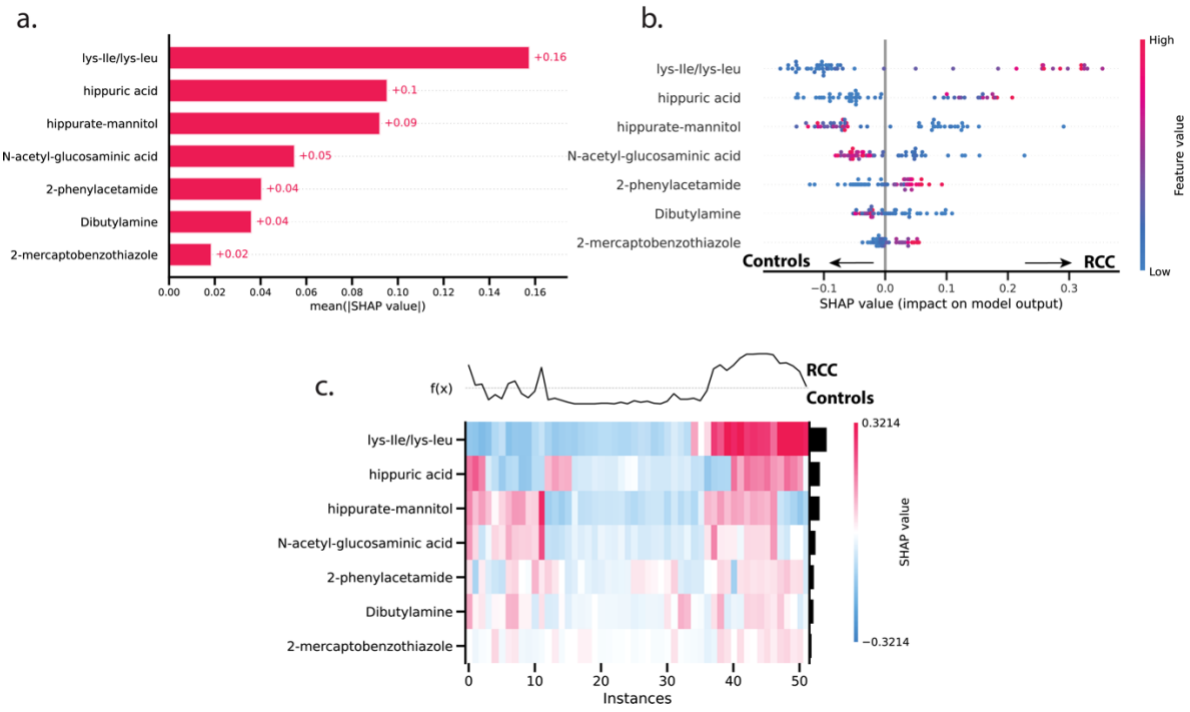


Figure 4-9. **Global Explanations of Metabolomic-based RCC detection using SHAP.**

Figure 4-10 displays the confusion matrix of the test set predictions with a true negative rate of 63.46%, a false positive rate of 1.92%, a false negative rate of 5.77%, and a true positive rate of 28.85%. To investigate the basis of the false-negative errors – the largest class of error in the model, the local explanation of representative true positive samples was compared to the false-negative samples (**Figure 4-11**).

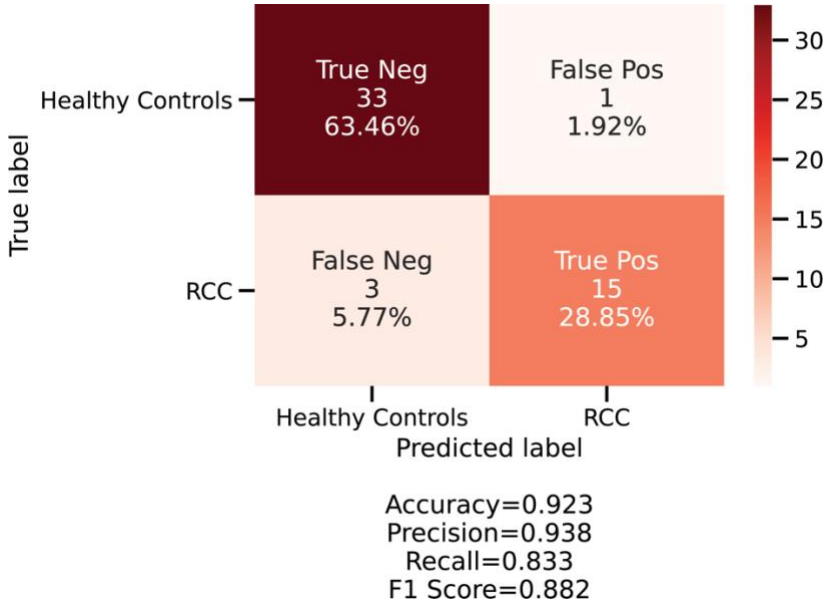


Figure 4-10. Confusion Matrix for the Random Forest Prediction on the Test Set for the RCC Detection Dataset.

The waterfall plots support the lys-ile/lys-leu importance hypothesis. The true positive samples (correctly predicted RCC samples) have lys-ile/lys-leu as the most important metabolite for successful predictions – increasing the probability of an RCC prediction. This is not the case for all false-negative samples (wrongly predicted RCC).

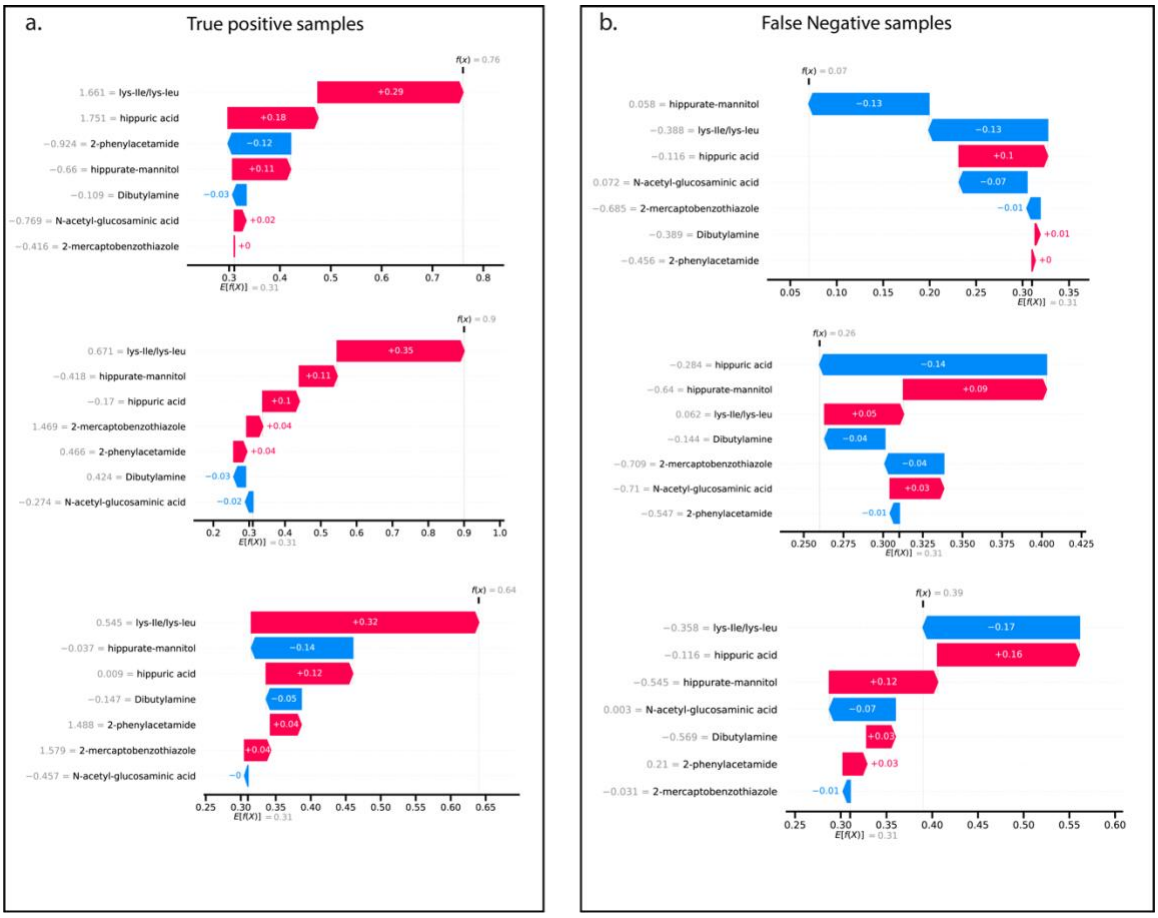


Figure 4-11. Error Analysis with SHAP for RCC detection.

Waterfall plots of true positive samples *vs.* false negative samples.

Supplementary Materials: All Jupyter notebooks used in this study can be found here: https://github.com/artedison/shap-metabolomics

4.7 References

1. Nicholson, J. K.; Lindon, J. C., Systems biology: Metabonomics. *Nature* **2008,** *455* (7216), 1054-6.

2. Bzdok, D.; Altman, N.; Krzywinski, M., Statistics versus machine learning. *Nat Methods* **2018,** *15* (4), 233-234.

3. Bzdok, D.; Krzywinski, M.; Altman, N., Points of Significance: Machine learning: a primer. *Nat Methods* **2017,** *14* (12), 1119-1120.

4. Smolinska, A.; Blanchet, L.; Buydens, L. M.; Wijmenga, S. S., NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal Chim Acta* **2012,** *750*, 82-97.

5. Grissa, D.; Petera, M.; Brandolini, M.; Napoli, A.; Comte, B.; Pujos-Guillot, E., Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Front Mol Biosci* **2016,** *3*, 30.

6. Worley, B.; Powers, R., Multivariate Analysis in Metabolomics. *Curr Metabolomics* **2013,** *1* (1), 92-107.

7. Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R., A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. *Anal Chim Acta* **2015,** *879*, 10-23.

8. Ruiz-Perez, D.; Guan, H.; Madhivanan, P.; Mathee, K.; Narasimhan, G., So you think you can PLS-DA? *BMC Bioinformatics* **2020,** *21* (Suppl 1), 2.

9. Galindo-Prieto, B.; Eriksson, L.; Trygg, J., Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J. Chemom.* **2014,** *28*, 623-632.

10. Wu, L.; Huang, R.; Tetko, I. V.; Xia, Z.; Xu, J.; Tong, W., Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chem Res Toxicol* **2021,** *34* (2), 541-549.

11. London, A. J., Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep* **2019,** *49* (1), 15-21.

12. Mosconi, F.; Julou, T.; Desprat, N.; Sinha, D. K.; Allemand, J.-F.; Croquette, V.; Bensimon, D., Some nonlinear challenges in biology. *Nonlinearity* **2008,** *21* (8), 131-147.

13. Ribeiro, M. T.; Singh, S.; Guestrin, C., Model-Agnostic Interpretability of Machine Learning. *arXiv* **2016**, arXiv:1606.05386.

14. Molnar, C., Interpretable machine learning. A Guide for Making Black Box Models Explainable. **2019**.

15. Friedman, J. H., Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001,** *29* (5), 1189-1232.

16. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E., Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015,** *24* (1), 44-65.

17. Apley, D. W.; Zhu, J., Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv* **2019**, arXiv:1612.08468.

18. Breiman, L., Random Forests. *Machine Learning* **2001,** *45*, 5-32.

19. Fisher, A.; Rudin, C.; Dominici, F., All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *arXiv* **2019**, arXiv:1801.01489v5.

20. Ribeiro, M. T.; Singh, S.; Guestrin, C., "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv* **2016**, arXiv:1602.04938v3.

21. Lundberg, S. M.; Lee, S. I., A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874v2.

22. Lundberg, S. M.; Erion, G.; Lee, S. I., Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2018**, arXiv:1802.03888v3.

23. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. I., From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2020,** *2* (1), 56-67.

24. Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.; Newman, S. F.; Kim, J.; Lee, S. I., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* **2018,** *2* (10), 749-760.

25. Rodriguez-Perez, R.; Bajorath, J., Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J Comput Aided Mol Des* **2020,** *34* (10), 1013-1026.

26. Rodriguez-Perez, R.; Bajorath, J., Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J Med Chem* **2020,** *63* (16), 8761-8777.

27. Cha, Y.; Shin, J.; Go, B.; Lee, D. S.; Kim, Y.; Kim, T.; Park, Y. S., An interpretable machine learning method for supporting ecosystem management: Application to species distribution models of freshwater macroinvertebrates. *J Environ Manage* **2021,** *291*, 112719.

28. Liu, X.; Locasale, J. W., Metabolomics: A Primer. *Trends Biochem Sci* **2017,** *42* (4), 274-284.

29. Shapley, L. S., A value for n-person games. *Contributions to the Theory of Games* **1953,** *2* (28), 307-317.

30. Thevenot, E. A.; Roux, A.; Xu, Y.; Ezan, E.; Junot, C., Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by

Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. *J Proteome Res* **2015,** *14* (8), 3322-35.

31. Zheng, X.; Huang, F.; Zhao, A.; Lei, S.; Zhang, Y.; Xie, G.; Chen, T.; Qu, C.; Rajani, C.; Dong, B.; Li, D.; Jia, W., Bile acid is a significant host factor shaping the gut microbiome of diet-induced obese mice. *BMC Biol.* **2017,** *15* (1).

32. Fahrmann, J. F.; Kim, K.; DeFelice, B. C.; Taylor, S. L.; Gandara, D. R.; Yoneda, K. Y.; Cooke, D. T.; Fiehn, O.; Kelly, K.; Miyamoto, S., Investigation of metabolomic blood biomarkers for detection of adenocarcinoma lung cancer. *Cancer Epidemiol Biomarkers Prev* **2015,** *24* (11), 1716-23.

33. Armstrong, C. W.; McGregor, N. R.; Lewis, D. P.; Butt, H. L.; Gooley, P. R., Metabolic profiling reveals anomalous energy metabolism and oxidative stress pathways in chronic fatigue syndrome patients. *Metabolomics* **2015,** *11*, 1626-1639.

34. Mendez, K. M.; Reinke, S. N.; Broadhurst, D. I., A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **2019,** *15* (12), 150.

35. Hastie, T.; Tibshirani, R.; Friedman, J., *The elements of statistical learning (2nd ed.)*. New York: Springer: 2009.

36. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S., Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel)* **2020,** *23* (1).

37. Breiman, L., Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* **2001,** *16* (3), 199-231.

38. Zumoff, B.; Rosenfeld, R. S.; Friedman, M.; Byers, S. O.; Rosenman, R. H.; Hellman, L., Elevated daytime urinary excretion of testosterone glucuronide in men with the type A behavior pattern. *Psychosom Med* **1984,** *46* (3), 223-5.

39. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S., Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* **2016,** *44* (D1), D463-70.

40. Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; Gonzalez-Beltran, A.; Sansone, S. A.; Griffin, J. L.; Steinbeck, C., MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* **2013,** *41* (Database issue), D781-6.

41. Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C., MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* **2020,** *48* (D1), D440-D444.

42. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**.

43. McKinney, W., Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* **2010,** *445*, 56-61.

44. Hunter, J. D., Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007,** *9* (3), 90-95.

45. Waskom, M. L., Seaborn: statistical data visualization. *Journal of Open Source Software* **2021,** *6* (60).

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

Urine has played a sustained role in human medicine since the ancient world. The physicians from Babylon and Sumeria were reported to have recorded the assessment of urine on clay tablets[1]. The middle ages saw the introduction of the Matula – a vessel made of light, thin glass that allows physicians to get a better view of the properties of urine, such as clarity and color, to aid diagnosis[2-3]. In the renaissance period, urinalysis became popular and imbued with optimism that the diagnostic fluid could not keep up with – as physicians were convinced that seeing the patient is superfluous and all ailments can be solved by observing the urine[4]. This later led to the implosion of uroscopy[5].

Today, analytical tools like LC-MS and NMR allow for identifying metabolites for disease biomarker discovery. In this dissertation, both analytical tools were employed for urine metabolite profiling. I successfully showed that it is possible to detect and stage renal cell carcinoma with urinary metabolites. This was achieved by incorporating advances in machine learning and data science to mine the datasets generated by the metabolomics platforms. Some of the metabolic pathways that are thought to be impacted based on the altered metabolites include hexosamine biosynthetic pathway, phenylalanine metabolism, nucleotide metabolism, aerobic glycolysis, tryptophan metabolism, and biotin metabolism. In addition, chemical exposome – the totality of chemical exposure to an individual – was retained in the biomarker studies. Furthermore, a recent advancement in explainable AI

method – Shapley additive explanations – was applied, first to a published metabolomics dataset for gender discrimination for validation purposes, and then to explain the predictions of the RCC detection using the seven biomarkers discovered.

5.1 Detecting Renal Cell Carcinoma *via* Urinary Metabolites

After conducting metabolic profiling with NMR and MS, I applied a compendium of machine learning techniques to narrow down thousands of metabolomic features detected by LC-MS and NMR. These methods rely on different inductive biases allowing for varied decision functions, and a voting methodology was employed to select the top, overlapping metabolites as biomarkers. The selected biomarker panel was then used for the classification task of discriminating RCC from healthy controls. Again, here using ML algorithms that afford different induction biases. For example, *k*-NN classifier relies on the decision functions that neighboring samples belong to the same class, while the SVM algorithm argues that binary classes are separable by a hyperplane.

A ten-metabolite panel and a five-metabolite panel (consisting of only upregulated metabolites in RCC) were presented in the study. However, only seven metabolites were identified, leading to the seven-metabolite panel presented in the study. Apart from hippurate[6-7], all of the metabolites in the panel are presented as a urinary metabolomic biomarker for RCC detection for the first time. However, some of the biomarkers identified have had other metabolites in the same metabolic pathway reported, such as alpha-N-phenylacetyl-L-glutamine[6, 8] and 4-hydroxyphenylacetate[6] in the phenylalanine metabolic pathway. In addition, aspartyl-phenylalanine and glutamyl-threonine have been reported for dipeptide metabolism in RCC urine metabolomics[9]. N-acetyl-D-glucosaminic acid, another metabolite in the panel, has been identified to play a central role in tumorigenesis,

not limited to RCC[10]. In addition, two exposure metabolites were selected as markers – 2-mercaptobenzothiazole and dibutylamine – suggesting a potential role of the exogenous metabolites in renal cancer development.

5.2 Staging Renal Cell Carcinoma *via* Urinary Metabolites

For the first time, I showed that it is possible to predict the primary tumor size of RCC using urinary metabolites with a R^2 score of 0.58. Metabolites used for the prediction include cytosine dimer, dihydrouridine, asparaginyl-hydroxyproline, and an unidentified metabolite. Predictions were conducted using a voting ensemble regressor consisting of elastic net, ridge, and support vector regressors. In addition, using similar machine learning strategies as in chapter 2, a 16-metabolite panel was initially selected to discriminate between early RCC and advanced RCC. However, the performance scores were slightly improved *via* the addition of the metabolites selected for RCC tumor prediction and NMR-derived metabolites with $p<0.05$ (Student *t*-test). This gave rise to the 24-metabolite RCC staging panel presented in chapter 3. Of the 24-metabolite panel, 16 metabolites were identified, and they suggest alterations in metabolic pathways like nucleotide metabolism, fatty acid metabolism, and protein degradation and re-utilization. These pathways have been reported to be altered in tissue-based metabolomics studies in RCC progression[11].

The voting ensemble learning technique was used for both the regression and classification tasks in chapter 3. The voting classifier works by aggregating the predictions of its base learners, which sometimes results in a superior prediction, as was the case in the RCC staging and primary tumor predictions. This success is because of the independence of the base learners. Independent learners are more likely to make different mistakes, compensating for each other in an ensemble, and the independence is derived from

different induction biases. The best improvement from using a voting ensemble in the study is reported in the discrimination of early RCC from advanced RCC, where AUC values include: random forests = 0.89, AdaBoost = 0.92, SVM-RBF=0.90, and logistic regression=0.94. The voting classifier ensemble gave an AUC of 0.96.

5.3 Explaining Metabolomics Machine Learning Models with SHAP

In chapter 4, a state-of-the-art explainable AI method – Shapley additive explanations – was applied for machine learning prediction interpretations. The metabolomics field utilizes PLS-DA for ML interpretability with the limitation of its restrictions to global explanations. SHAP, on the other hand, has high representativeness of explanations. I applied Tree SHAP to publicly available metabolomics datasets, specifically explaining a machine learning model used in gender predictions. The pre-eminent role of testosterone glucuronide was discovered through a series of explanation methodologies such as the summary plot (or bee swarm plot), waterfall plot, force plots, and supervised hierarchical clustering with SHAP values. This gave validation for the applicability of Tree SHAP, as testosterone glucuronide is a metabolite of the most important male sex hormone – testosterone.

Furthermore, the RCC detection dataset was trained on random forests and explained using Tree SHAP. Lysyl-isoleucine was identified as the most important metabolite with global interpretations. In addition, the comparison of the local interpretations of the true positive samples (correctly classified RCC) and false-negative samples (RCC classified as healthy controls) indicates the importance of lysyl-isoleucine for accurately predicting RCC. In conclusion, SHAP can be used to investigate the metabolic uniqueness of RCC subtypes – if there exist any. This will require a much larger

datasets that are equally representative of the various RCC subtypes, sufficient for the ML algorithms to extract useful patterns.

5.4 Four Concluding Thoughts

The importance of the exposome: We live in an increasingly industrial world that invents faster than it can feasibly assess risks associated with such inventions. The implications for such behavior have been increasingly evident. For example, phthalates – a group of compounds used to make plastic more durable – have been reported to reduce testosterone biosynthesis in males[12]. In fact, of the several thousands of chemicals registered for commercial use in the US, less than 1% of them are tested for toxicity[13]. Therefore, biomarker studies of the kind presented in this thesis provide the opportunity to identify exposure metabolites associated with disease conditions. 2-mercaptobenzothiazole and dibutylamine are selected as markers for RCC, while succinic anhydride was selected in RCC staging. This is the first time these associations were discovered. Further studies are required to show the constitutive presence of the compounds in RCC urine samples – as exposure markers for RCC – and, importantly, their potential roles in the development of renal cell cancer, if there are any.

Causal inference and biomarker discovery: There are unlimited potential confounders that can impact the validity of any biomarker discovery study. This is because biomarker discoveries, such as those conducted in this thesis, are based on observational studies. However, causal claims are only permissible under well-conducted randomized controlled trials. And yet, randomizing RCC is impossible (and unethical). In this study, I have taken approaches such as propensity score matching to limit selection bias – as such, causal claims cannot be made, only provisional causality. To this end, massive,

independent studies with diversity in race and geographical locations are required to validate biomarkers, which will implicitly answer the causality question.

Specificity of urine metabolomic markers: Biomarkers such as lactate and hippurate selected in the RCC metabolic panel are by no means exclusive to RCC. This, at a cursory look, connotes a negative result. However, this need not be the case because overlapping urine tumor markers in several cancer types can potentially screen for the onset of tumorigeneses in patients. Afterward, further tests can be conducted to identify the specific organ affected based on the patients' clinical history. What is being proposed here is a universal cancer screening[14], and such urine markers do not exist today. It is through the analysis of several specific cancer types in comparison with healthy controls can this be identified. In addition, large comparative studies of biomarkers of different cancer types are essential to identify the uniqueness of markers — for example, renal carcinoma *vs.* bladder cancer.

Predictive results for RCC detection and staging: There are currently no urine RCC metabolomic markers in use in the clinic, nor are there any RCC tumor markers[15]. As such, there is no urine metabolomic marker standard for comparison. For the RCC detection study, a sensitivity of 94% and specificity of 85% were reported, and for RCC staging, sensitivity = 80% and specificity = 93%. These are promising predictive scores that can be used for screening and monitoring the progression of RCC, pending validation of markers in more diverse populations, as described above. As a manner of comparison, one of the promising urine tumor marker panels currently in development is the UROSEEK. The UROSEEK detects alterations in 11 genes in bladder and upper tract urothelial cancers. It

has reported having a specificity of 99.5% and sensitivity of 83% in one of its latest studies[16].

5.5 References

1. Armstrong, J. A., Urinalysis in Western culture: a brief history. *Kidney Int* **2007,** *71* (5), 384-7.

2. Harvey, R., The judgement of urines. *CMAJ* **1998,** *159* (12), 1482-4.

3. Viswanathan, S., Urine bag as a modern day matula. *ISRN Nephrol* **2013,** *2013*, 215690.

4. Voswinckel, P., From uroscopy to urinalysis. *Clin Chim Acta* **2000,** *297* (1-2), 5-16.

5. Gardner, K. D., Jr., The art and gentle science of Pisse-Prophecy. *Hawaii Med J* **1971,** *30* (3), 166-9.

6. Monteiro, M. S.; Barros, A. S.; Pinto, J.; Carvalho, M.; Pires-Luis, A. S.; Henrique, R.; Jeronimo, C.; Bastos, M. L.; Gil, A. M.; Guedes de Pinho, P., Nuclear Magnetic Resonance metabolomics reveals an excretory metabolic signature of renal cell carcinoma. *Sci Rep* **2016,** *6*, 37275.

7. Ragone, R.; Sallustio, F.; Piccinonna, S.; Rutigliano, M.; Vanessa, G.; Palazzo, S.; Lucarelli, G.; Ditonno, P.; Battaglia, M.; Fanizzi, F. P.; Schena, F. P., Renal Cell Carcinoma: A Study through NMR-Based Metabolomics Combined with Transcriptomics. *Diseases* **2016,** *4* (1).

8. Zhang, M.; Liu, X.; Liu, X.; Li, H.; Sun, W.; Zhang, Y., A pilot investigation of a urinary metabolic biomarker discovery in renal cell carcinoma. *Int Urol Nephrol* **2020,** *52* (3), 437-446.

9. Liu, X.; Zhang, M.; Liu, X.; Sun, H.; Guo, Z.; Tang, X.; Wang, Z.; Li, J.; Li, H.; Sun, W.; Zhang, Y., Urine Metabolomics for Renal Cell Carcinoma (RCC) Prediction: Tryptophan Metabolism as an Important Pathway in RCC. *Front Oncol* **2019,** *9*, 663.

10. Ma, Z.; Vosseller, K., Cancer metabolism and elevated O-GlcNAc in oncogenic signaling. *J Biol Chem* **2014,** *289* (50), 34457-65.

11. Hakimi, A. A.; Reznik, E.; Lee, C. H.; Creighton, C. J.; Brannon, A. R.; Luna, A.; Aksoy, B. A.; Liu, E. M.; Shen, R.; Lee, W.; Chen, Y.; Stirdivant, S. M.; Russo, P.; Chen, Y. B.; Tickoo, S. K.; Reuter, V. E.; Cheng, E. H.; Sander, C.; Hsieh, J. J., An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancer Cell* **2016,** *29* (1), 104-116.

12. Chang, W. H.; Li, S. S.; Wu, M. H.; Pan, H. A.; Lee, C. C., Phthalates might interfere with testicular function by reducing testosterone and insulin-like factor 3 levels. *Hum Reprod* **2015,** *30* (11), 2658-70.

13. *Blocking Smoke: Chemical Companies Say "Trust Us," But Environmental and Workplace Safety Violations Belie Their Rhetoric*; Center for Effective Government: 2015.

14. Ahlquist, D. A., Universal cancer screening: revolutionary, rational, and realizable. *NPJ Precis Oncol* **2018,** *2*, 23.

15. Tumor Markers in Common Use. https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-list (accessed June 3, 2021).

16. Springer, S. U.; Chen, C. H.; Rodriguez Pena, M. D. C.; Li, L.; Douville, C.; Wang, Y.; Cohen, J. D.; Taheri, D.; Silliman, N.; Schaefer, J.; Ptak, J.; Dobbyn, L.; Papoli, M.; Kinde, I.; Afsari, B.; Tregnago, A. C.; Bezerra, S. M.; VandenBussche, C.; Fujita, K.; Ertoy, D.; Cunha, I. W.; Yu, L.; Bivalacqua, T. J.; Grollman, A. P.; Diaz, L. A.; Karchin, R.; Danilova, L.; Huang, C. Y.; Shun, C. T.; Turesky, R. J.; Yun, B. H.; Rosenquist, T. A.;

Pu, Y. S.; Hruban, R. H.; Tomasetti, C.; Papadopoulos, N.; Kinzler, K. W.; Vogelstein, B.; Dickman, K. G.; Netto, G. J., Non-invasive detection of urothelial cancer through the analysis of driver gene mutations and aneuploidy. *Elife* **2018,** *7*.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

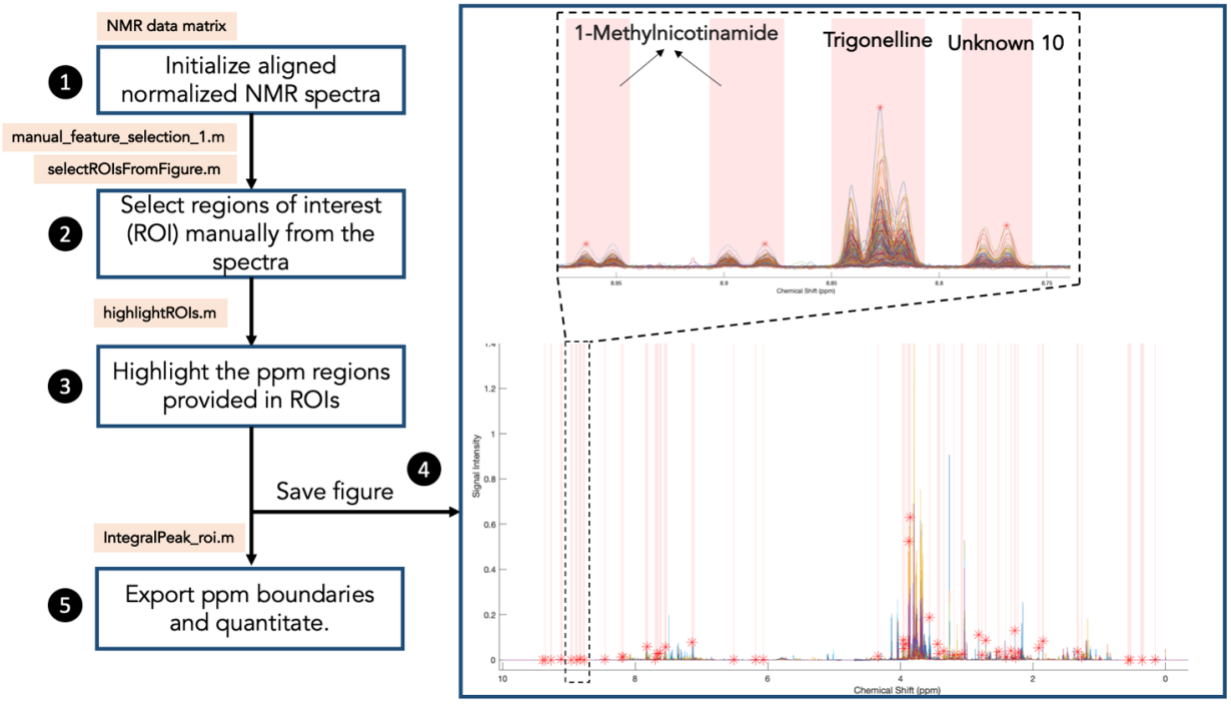
Section A-1. NMR quality assurance and quality control and Spectral Binning

All study cohort urine samples were randomized prior to running NMR experiments. Twelve NMR buffer blanks, 3 per run, were added to ensure there is no carry-over of samples between urine samples. One buffer blank was added at the beginning, middle, and end of each run. Thirty-two external controls (Nicotine, Ethanol & Drug Free Human Urine, Female; Golden West Diagnostics, LLC), 8 in each of the four NMR racks (run) used, were added to the study to ascertain the reliability of the sample acquisition process particularly for comparing across runs. One external control was added after the first blank at the beginning the run, and one before the last blank at the end of the run, while the remaining six external controls were randomized with the study's urine samples. The external pooled controls served its purpose as the controls were clustered tightly together on a PCA plot (data not shown). Also, 27 µL were taken out of each study urine sample for two internal pooled controls. 1D ^{1}H NMR experiment was conducted on all samples, while HSQC was carried out on one internal pooled control sample, and HSQC-TOCSY was carried out on the other. The NMR data for the internal pooled samples were used for metabolite annotation using AssureNMR and COLMARm[1] as described in the manuscript. Fifty metabolomic features in the aligned and normalized 1D ^{1}H NMR spectra were manually binned and quantified by taking spectral areas for integration in regions without

overlap, and combined with MS features for downstream analysis. A manual binning workflow in the Edison laboratory in-house metabolomics toolbox MATLAB scripts was used for binning (https://github.com/artedison/Edison_Lab_Shared_Metabolomics_UGA).

Scheme A-1 shows the steps involved in this process: aligned normalized NMR spectra were initialized, the functions manual_feature_selection1.m and selectROIsFromFigure.m allowed us to manually select multiple regions in the displayed spectra, where a rectangle is drawn around the region of interest (ROI) for binning in an interactive fashion. The highlighROIs.m function highlights the ppm regions provided in ROI, and the binned, highlighted NMR spectra figure is saved. Finally, ppm boundaries are exported, and IntegralPeak_roi.m function was used to integrate metabolomic features in the binned spectra.

Scheme A-1. NMR Peak Picking Methods



Section A-2. Model evaluation metrics

The following metrics were used for model evaluation, where TP is true positive, TN is true negative, FP is false positive, and FN is false negative:

Accuracy measures the percentage of all correctly predicted samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity (or recall) measures the percentage of correctly predicted RCC patients out of the total RCC samples.

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity measures the percentage of correctly predicted controls out of the total control samples.

$$Specificity = \frac{TN}{TN + FP}$$

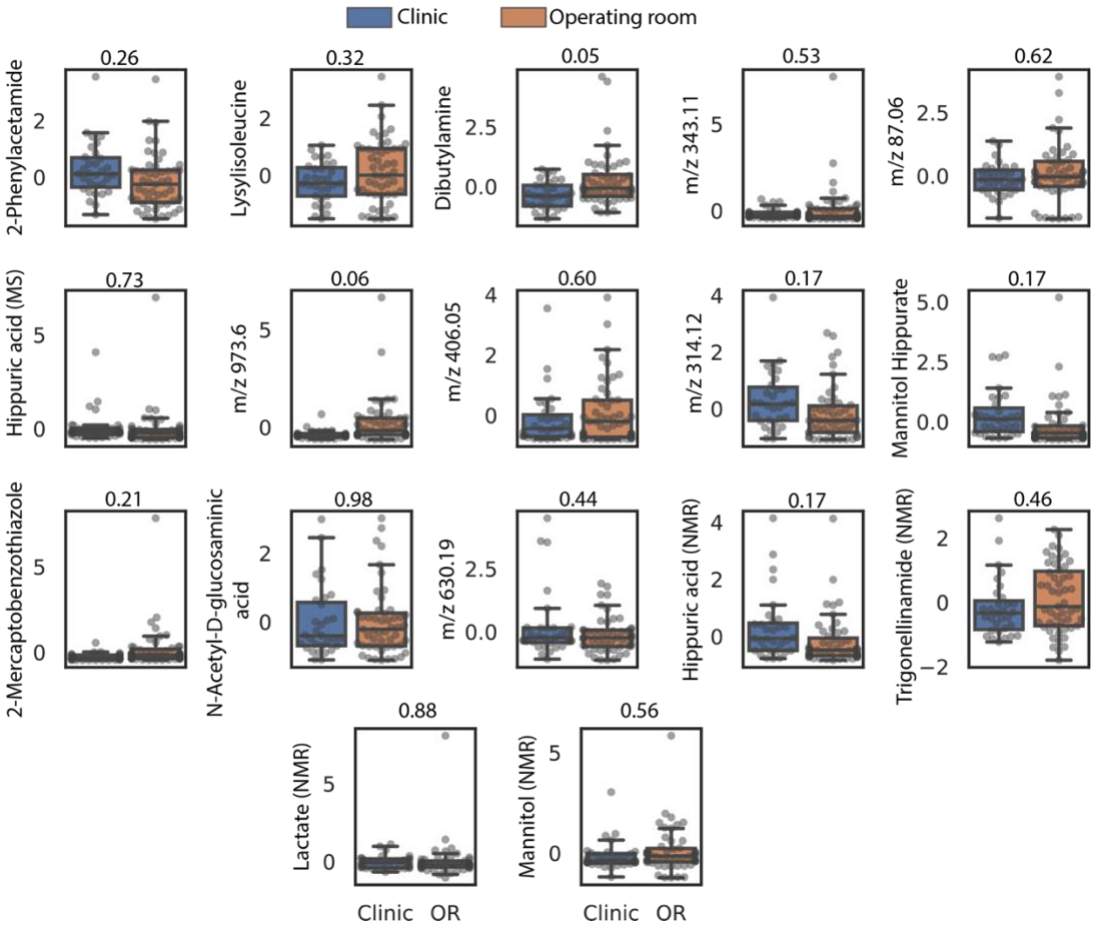


Figure A-1. Relative Quantification of all Discriminating Metabolomic Features Identified in the Study, for RCC Samples Collected in the Clinic *vs.* Operating Room.

q-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test. All features were statistically insignificant. Raw data were transformed *via* autoscaling for visualization.

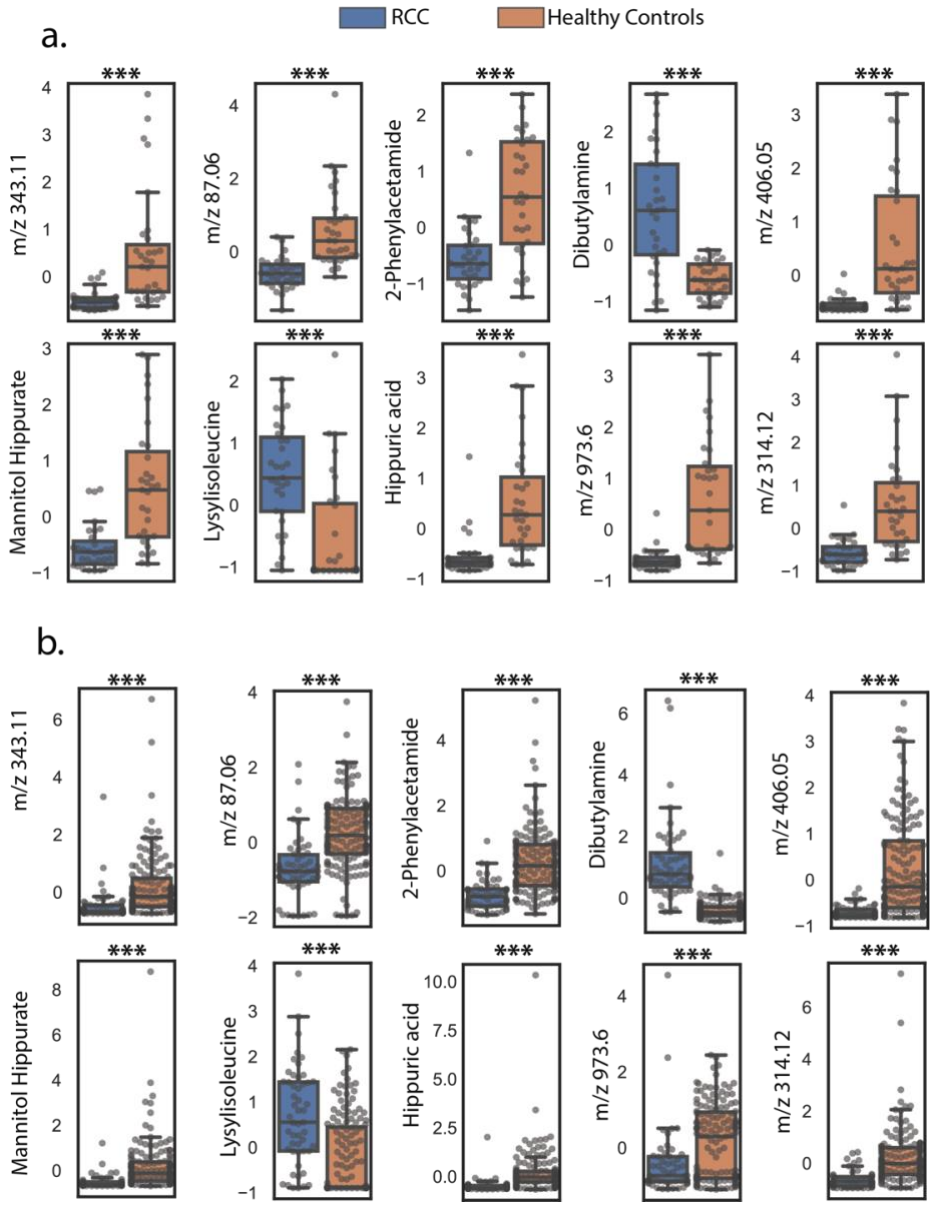


Figure A-2. Relative Quantification of the 10 Metabolite Panel for RCC Detection.

(a) in the model cohort. After selecting features with greater than a one-fold change between control and RCC groups, *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test. ($* \, q \leq 0.05$, $** \, q \leq 0.01$, $*** \, q \leq 0.001$). (b) Test cohort, *p*-values from the Welch t-test were reported (unequal sample size). ($* \, p \leq 0.05$, $** \, p \leq 0.01$, $*** \, p \leq 0.001$). Raw data were transformed *via* autoscaling for visualization

a.

ID	Metabolite ID	Mode	RT [min]	FC	*q*-value
2102	dibutylamine, N-butylisobutylamine, diisobutylamine (isomer)	positive	3.449	1.57	7.87E-05
3872	3872	positive	4.049	-2.58	1.90E-04
3675	3675	positive	1.184	-1.08	2.62E-04
3757	3757	positive	4.003	-2.39	3.26E-04
6261	6261	negative	2.591	-1.88	3.68E-04
5383	5383	negative	4.067	-2.62	3.85E-04
720	2-phenylacetamide	positive	2.562	-1.07	3.86E-04
4401	4401	positive	4.039	-2.69	4.81E-04
6276	6276	negative	2.636	-1.65	4.81E-04

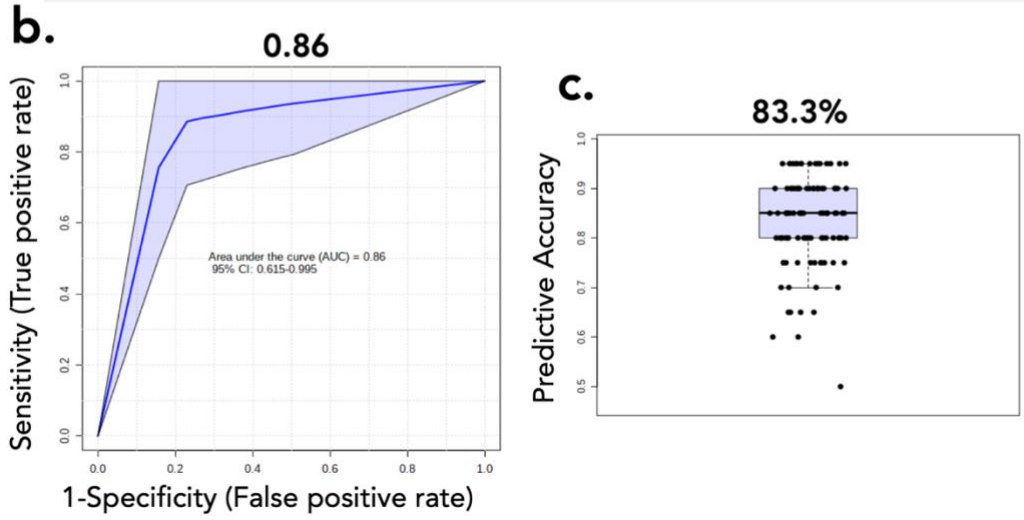


Figure A-3. Selection of Metabolomic Features with *q*-values and Classifying with Logistic Regression using the Metaboanalyst 5.0 Biomarker Analysis Platform.

(a) Metabolomic features with the top ten highest *q*-values from univariate analysis. (b) ROC-AUC (c) Predictive accuracy. Analysis was carried out using the model cohort.

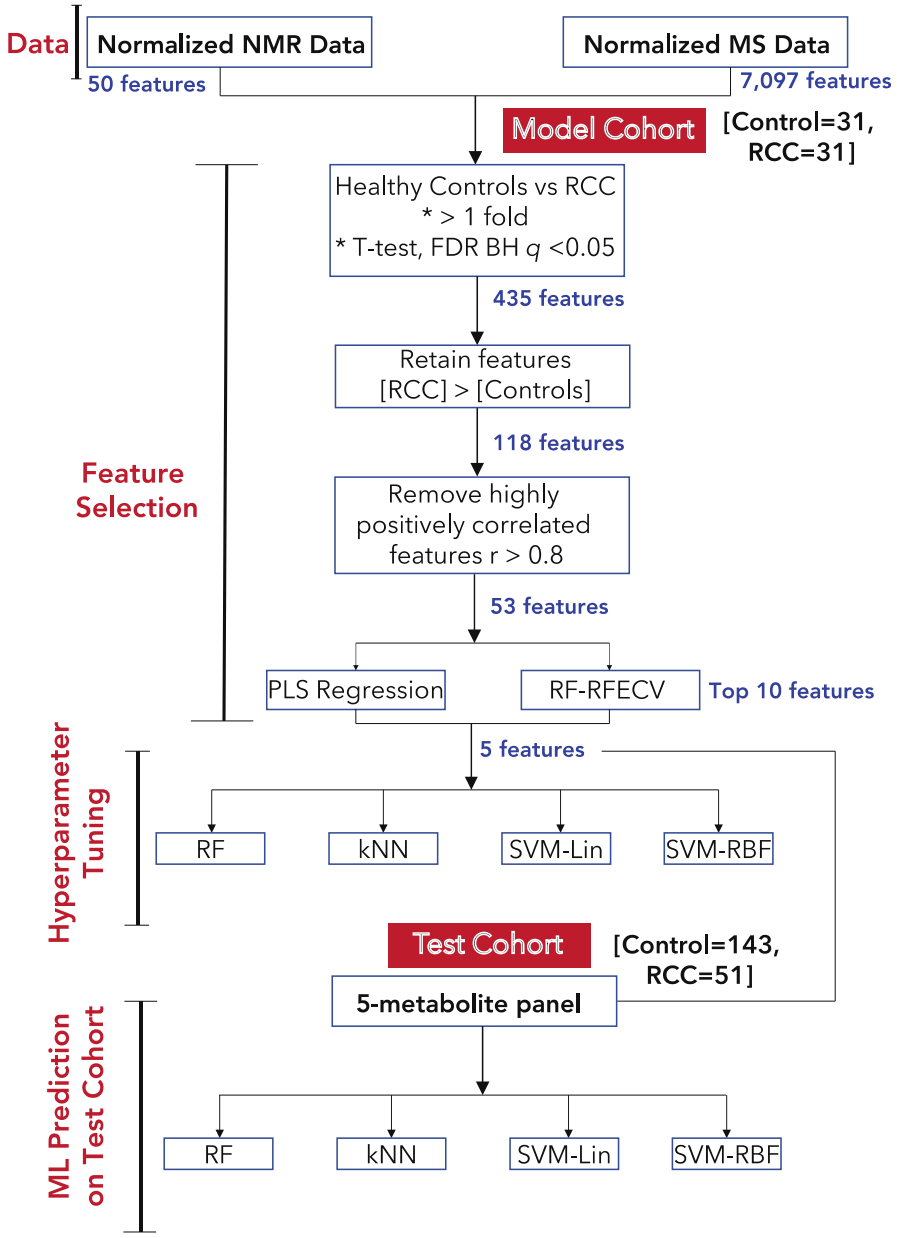


Figure A-4. Machine Learning Pipeline Focused on Upregulated Features in RCC *vs.* Controls.

PLS: partial least squares, **RF- RFECV:** random forest recursive feature elimination – cross validation, **FDR-BH:** false discovery rate Benjamini-Hochberg procedure, ***k*-NN:** *k*-nearest neighbors.

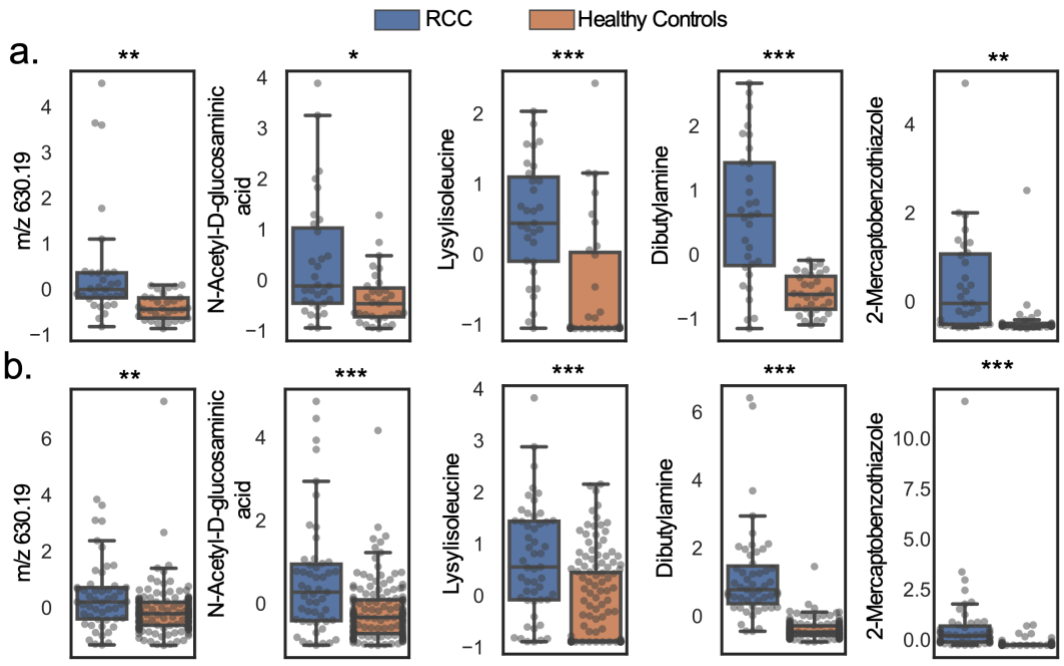


Figure A-5. Relative Abundances for the Panel of Upregulated Metabolites in RCC.

(a) Model cohort. *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test. ($* \, q \leq 0.05, \; ** \, q \leq 0.01, \; *** \, q \leq 0.001$). (b) In the test cohort, *p*-values from the Welch t-test were reported (unequal sample size). ($* \, p \leq 0.05, \; ** \, p \leq 0.01, \; *** \, p \leq 0.001$). Raw data were transformed *via* autoscaling for visualization

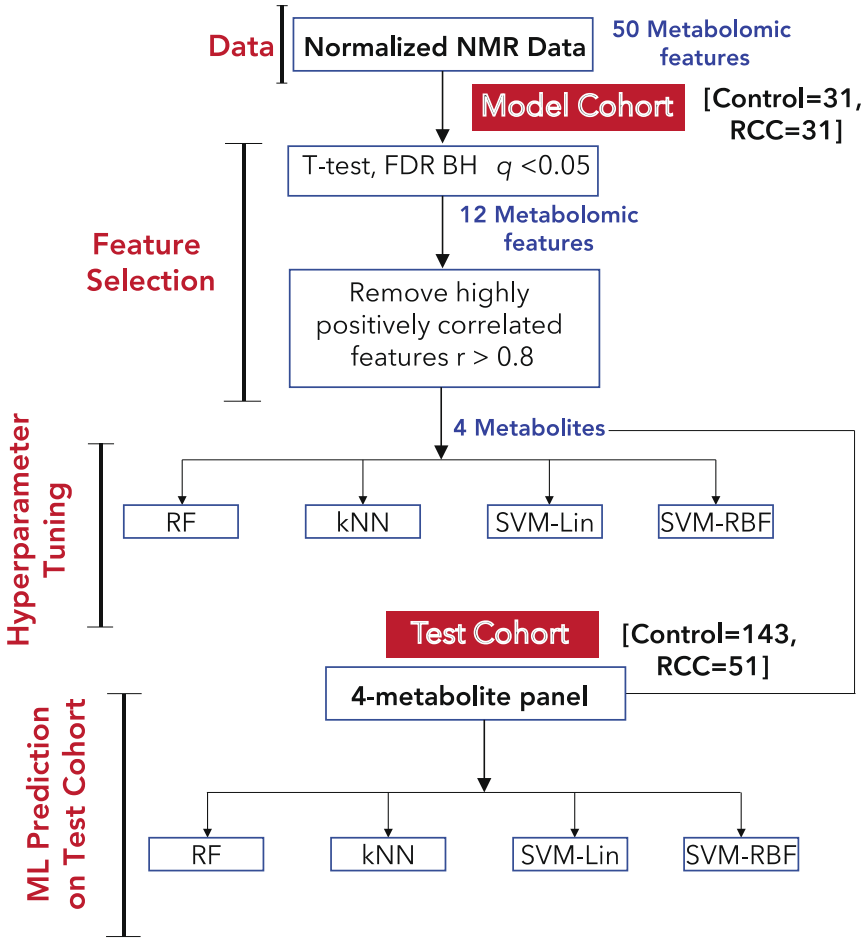


Figure A-6. Machine Learning Pipeline Focused Only on NMR features for RCC Detection.

Using the model cohort, all NMR features were subjected to feature selection strategies culminating in four selected metabolites (hippurate, trigonellinamide, lactate, and mannitol). Hyperparameters for four different machine learning models were tuned using the model cohort and the 4-metabolite panel. Final predictions were made using the test cohort under cross-validated conditions. **PLS:** partial least squares, **RF- RFECV:** random forest recursive feature elimination – cross validation, **FDR-BH:** false discovery rate

Benjamini-Hochberg procedure, ***k*-NN:** *k*-nearest neighbors, **SVM:** support vector machines (Lin: linear, RBF: radial basis function).

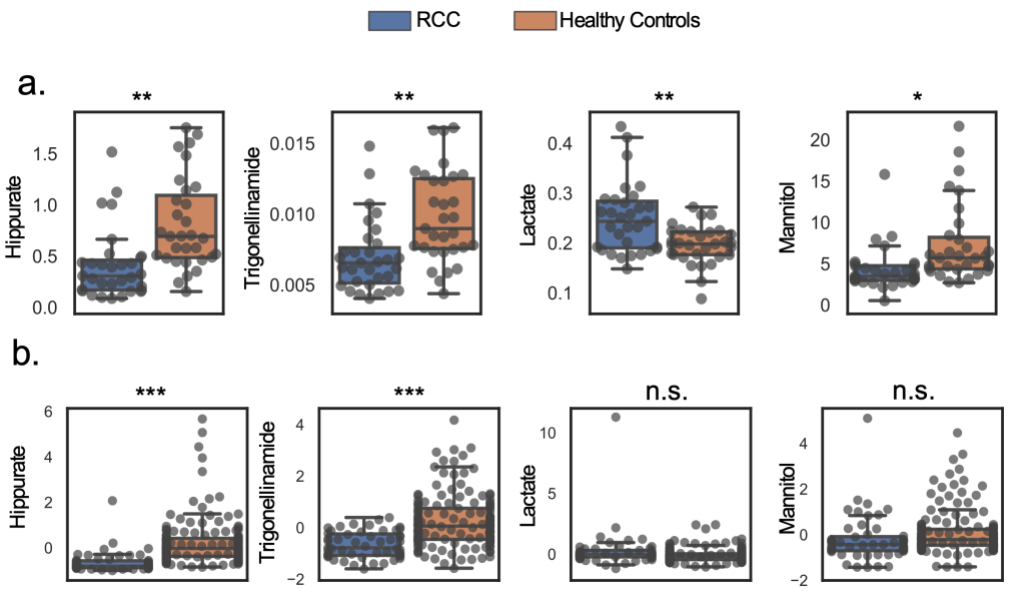


Figure A-7. Relative Quantification of Features in the NMR RCC Metabolic Panel.

(a) Model cohort. *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test. ($* \ q \leq 0.05, \ ** \ q \leq 0.01, \ *** \ q \leq 0.001$). (b) In the test cohort, *p*-values from the Welch t-test were reported (unequal sample size). (n.s. not significant $* \ p \leq 0.05, \ ** \ p \leq 0.01, \ *** \ p \leq 0.001$). Raw data were transformed *via* autoscaling for visualization.

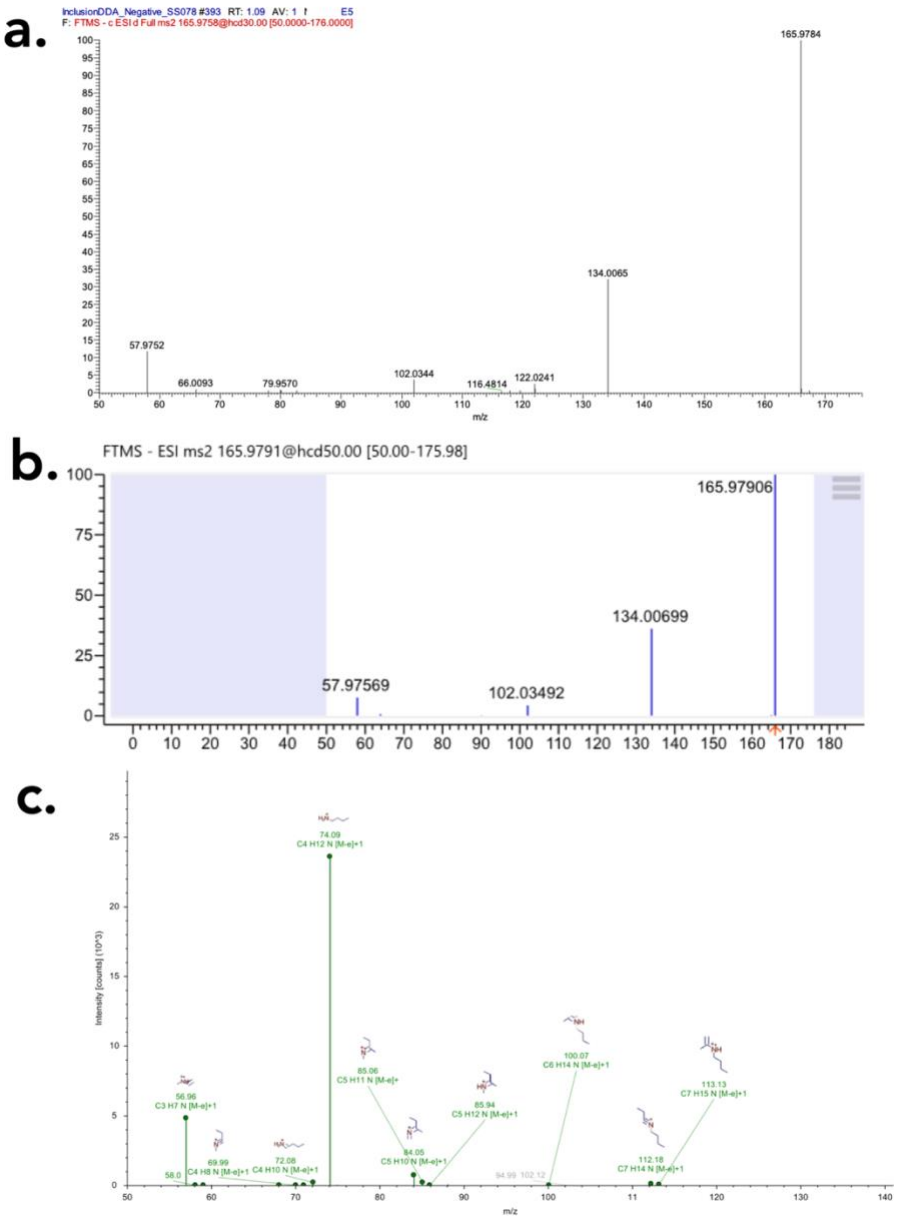


Figure A-8. **MS/MS Annotation of 2-mercaptobenzothiazole and Dibutylamine/ N-butylisobutylamine/Disobutylamine.**

a) Experimental spectra of 2-mercaptobenzothiazole b) MS/MS of spectrum of 2-mercaptobenzothiazole from mzCloud database. c) Annotated MS/MS spectrum of feature identified as dibutylamine/n-butylisobutylamine/disobutylamine. Annotations are obtained from in silico fragmentation in Compound Discoverer (Thermo Fisher).

Table A-1. Propensity Score Matching and Model Cohort Characteristics for the RCC Detection Study.

	Pre-match Groups			**Post-match Groups**		
Characteristic	**Controls**	**RCC**	***p*-Value**	**Controls**	**RCC**	***p*-Value**
No of Urine Samples	174	31		31	31	
Mean Age± SD	54.4±10.3	59.5±12.4	0.03	58.0±13.0	59.5±12.1	0.64
BMI	27.3±4.5	29.1±5.8	0.11	26.5±4.9	29.1±5.8	0.06
Race						
Caucasian	162 (93.1%)	21 (67.7%)		25 (80.6%)	21 (67.7%)	
Black/African American	5 (4.0%)	9 (29.0%)		3 (9.7%)	9 (29.0%)	
Others	7 (2.8%)	1 (3.2%)		3 (9.7%)	1 (3.2%)	
Smoker						
Never	131 (75.3%)	19 (61.3%)		17 (54.8%)	19 (61.3%)	
Former/Current	43 (24.7%)	12 (38.7%)		14 (45.2%)	12 (38.7%)	
Gender						
Male	145 (83.3%)	14 (45.2%)		14 (45.2%)	14 (45.2%)	
Female	29 (16.6%)	17 (54.8%)		17 (54.8%)	17 (54.8%)	

p-Values were calculated using Welch and Student *t*-test, for unequal and equal sample sizes, respectively. Race: unknown/missing (6), mixed (1), and Asian (1) were all classified as others (8); Smoker: former (46) and current (9) were all classified as former/current (55).

RCC: Renal Cell Carcinoma.

Table A-2. RCC Patients' Cohort Characteristics of the Model Cohort for the RCC Detection Study.

Characteristic	Frequency	Percentage
Metastasis		
Yes	7	22.6%
No	24	77.4%
Histological Subtypes		
Pure Clear Cell	23	71%
Papillary	3	9.7%
Chromophobe	2	6.5%
Clear Cell Papillary	2	6.5%
Unclassified	2	6.5%
[a] *Nuclear Grade*		
1	0	0%
2	10	33.3%
3	8	26.7%
4	12	40%
T-Stage		
T1a	11	35.4%
T1b	5	16.1%
T2a	2	6.4%
T2b	1	3.2%
T3a	11	35.4%
T4	1	3.2%
M-Stage		
M0	24	77.4%
M1	7	22.6%
N-Stage		
N0	25	80.6%
N1	4	12.9%
NX	2	6.5%
[b] *RCC Stage*		
I	13	44.8%
II	3	10.3%
III	6	20.7%
IV	7	24.1%

[a] One nuclear grade missing. [b] Two individuals cancer stages are not reported due to inconclusive TNM staging.

Table A-3. Test Cohort Characteristics for the RCC Detection Study.

Characteristic	Controls	RCC	*p*-Value
No. of Urine Samples	143	51	
Mean Age ± SD	53.6 ± 9.5	61.7 ± 13.7	0.0002
BMI	27.5 ± 4.4	29.1 ± 5.6	0.03
Race			
Caucasian	136(95.1%)	35(70.0%)	
Black/African American	5(3.4%)	11(22.0%)	
Others	2(1.4%)	4(8.0%)	
Smoker			
Never	113 (79%)	32 (62.7%)	
Former/Current	30 (21%)	19 (37.3%)	
Gender			
Male	131 (91.6%)	31 (62.0%)	
Female	12 (8.4%)	19 (38.0%)	
Histological Subtypes			
Pure Clear Cell		35 (68.6%)	
Papillary		7 (13.7%)	
Clear Cell Papillary		4 (7.8%)	
Chromophobe		3 (5.9%)	
Unclassified		2 (3.9%)	
Metastasis			
No		41 (80.4%)	
Yes		10 (19.6%)	
[a] *Nuclear Grade*			
1		0 (0%)	
2		20 (41.7%)	
3		21 (43.8%)	
4		7 (14.6%)	
[b] *RCC Stage*			
I		20 (48.8%)	
II		5 (12.2%)	
III		9 (22.0%)	
IV		7 (17.1%)	

p-Values were calculated using the Welch *t*-test. **[a]** For nuclear grades, three samples were missing. **[b]** Ten samples have missing RCC staging because of inconclusive TNM staging.

Table A-4. Quantified NMR Features. ppm Values, Confidence Score, Fold Changes, and *q*-values.

Metabolite/Features	^{1}H (ppm)	^{13}C (ppm)	Peak patterns	Confidence Score	Fold Change	*q*-value
unknown 1	0.15	-	(s)	-	0.01	0.957
unknown 2	0.36	-	(m)	-	0.42	0.050
·**bile acid 1	0.53	-	(s)	1	0.32	0.119
·**bile acid 2	0.56	-	(s)	1	0.12	0.560
3-hydroxyisovaleric acid	1.26	30.84	(s)	3	0.07	0.704
lactate	1.31	22.97	(d)	4	0.34	0.003
unknown 3	1.85	-	(s)	-	0.59	0.560
acetate	1.90	26.04	(s)	3	0.57	0.196
acetone	2.23	32.40	(s)	3	-0.12	0.196
unknown 4	2.26	-	(s)	-	-0.06	0.704
acetoacetate	2.27	32.19	(s)	3	-0.07	0.634
unknown 5	2.33	-	(s)	-	-0.03	0.860
¨pyruvate	2.41	-	(s)	2	0.05	0.560
citrate	2.53	48.52	(d)	3	-0.05	0.811
dimethylamine (DMA)	2.71	37.5	(s)	3	0.22	0.119
unknown 6	2.77	-	(s)	-	0.05	0.827
methylguanidine	2.82	30.21	(s)	3	0.16	0.256
unknown 7	3.08	-	(t)	-	-0.34	0.126
choline	3.19	56.69	(s)	3	-0.06	0.686
[a]scyllo-inositol	3.35	76.4	(s)	3	-1.14	0.002
taurine	3.42	38.07	(t)	4	-0.03	0.811
acetoacetate	3.44	56.22	(s)	3	0.23	0.368
4-hydroxyphenylacetate (4-HPA)	3.44	46.34	(s)	4	0.23	0.368
glycine	3.56	44.18	(s)	3	0.48	0.368
mannitol	3.86	65.94	(d)	4	-0.78	0.012
mannitol	3.88	65.94	(d)	4	-0.68	0.022
creatine	3.92	-	(s)	3	0.07	0.811
[a]glycolate	3.94	64.32	(s)	3	0.07	0.663
hippurate	3.96	46.46	(d)	4	-0.68	0.004
4-hydroxyhippuric acid	3.96	46.58	(d)	3	-0.68	0.004

tartrate	4.34	76.55	(s)	3	0.13	0.728
unknown 8	6.07	-	(s)	-	0.09	0.415
unknown 9	6.18	-	(s)	-	0.59	0.267
fumarate	6.52	-	(s)	2	0.17	0.560
4-hydroxyphenylacetate (4-HPA)	7.13	133.15	(d)	4	0.83	0.168
hippurate	7.55	131.50	(t)	4	-0.98	0.002
hippurate	7.65	134.92	(m)	4	-0.96	0.002
[a]4-aminohippuric acid	7.67	133.02	(d)	3	-1.64	0.002
indoxyl sulfate (IS)	7.70	127.07	(d)	3	0.18	0.492
hippurate	7.83	129.85	(dd)	4	-0.91	0.003
hypoxanthine	8.18	148.27	(s)	3	0.21	0.811
hypoxanthine	8.20	144.75	(s)	3	0.86	0.368
formate	8.45	173.71	(s)	3	0.22	0.488
unknown 10	8.77	-	(d)	-	0.19	0.791
trigonelline	8.83	147.46	(t)	3	-0.2	0.686
trigonellinamide	8.89	-	(d)	2	-0.49	0.002
trigonellinamide	8.97	-	(d)	2	-0.49	0.002
trigonelline	9.11	148.50	(s)	3	-0.33	0.524
trigonellinamide	9.27	-	(s)	2	-0.51	0.009
unknown 11	9.36	-	(s)	-	-0.14	0.686

[a] Quantification may be unreliable because of spectral overlaps. Tentative assignment (Monteiro *et al* 2016) s=singlet, d=doublet, dd=doublet of doublet, m=multiplet. Fold change (FC) was calculated as the base 2 logarithm of the average integral ratios between RCC and controls samples. Positive FC values indicate increased abundance in RCC, while negative values indicate higher abundance in control samples. *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test.

Table A-5. Chemical Information for the 10-Metabolite Panel for RCC Detection.

ID no.	Retention Time (min)	m/z		Adduct Type	Mass error (ppm)	Elemental Formula	Metabolite Identity
		Theoretical	Experimental				
720	5.68	136.0757	136.0755	$[M+H]^+$	-1.47	C_8H_9NO	2-phenylacetamide
1481	8.83	260.1969	260.1969	$[M+H]^+$	0.00	$C_{12}H_{25}N_3O_3$	Lys-Ile
2102	4.39	130.1590	130.1591	$[M+H]^+$	0.77	$C_8H_{19}N$	dibutylamine, N-butylisobutylamine, diisobutylamine (isomer)
3141	2.27	343.1135	343.1134	$[M+H]^+$	-0.20	$C_{14}H_{18}N_2O_8$	--
3675	1.18	--	87.0641	$[M+H]^+$	--	--	--
3804	2.59	202.0474	202.0478	$[M+H]^+$	1.70	$C_4H_{12}NO_6P$	hippuric acid
3872	4.05	973.6038	973.6027	$[M+2H]^{2+}$	-1.13	$C_{100}H_{158}N_{19}O_{20}$	--
4080	0.82	406.0597	406.0594	$[M+H]^+$	-0.78	$C_{10}H_{21}N_3O_8P_2S$	
6261	2.59	314.1248	314.1244	$[M-H]^-$	-1.27	$C_9H_{18}N_9O_2P$	--
6262	2.67	376.1249, 358.1143	376.1246, 358.1147	$[M+H_2O-H]^-$ $[M-H]$	-0.68	$C_{15}H_{21}NO_9$	hippurate-mannitol derivative

Table A-6. Machine Learning Hyperparameters used for Binary Classification using the MS-based 10-Metabolite Panel for RCC Detection.

Parameters	Initial distribution	Optimized
Random Forest		
Max_depth	10, 20, 30	10
Max_features	'auto', 'sqrt', 'log2'	'auto'
Min_samples_leaf	1, 2, 3, 4, 5	1
Min_samples_split	2, 4, 6, 8	2
N_estimators	50, 100, 150, 200	100
SVM-RBF		
C	0.1, 1, 10, 100	10
gamma	0.01, 0.03, 0.1, 0.3, 1.0	0.1
Lin-SVM		
C	0.001, 0.01, 0.1, 1, 5, 10	0.1
k-NN		
Number of neighbors	2 - 30	4
Distance Measure	Manhattan, Euclidean	Manhattan

Table A-7. Machine Learning Performance using the MS-based 10-Metabolite Panel for RCC Detection.

Algorithm	***RF***	***K-NN***	***SVM-RBF***	***Linear SVM***
AUC	1.0 +/- 0.0 (0.95)	0.96 +/- 0.04 (**0.96**)	0.99 +/- 0.01 (0.94)	1.0 +/- 0.0 (0.97)
Accuracy	0.95 +/- 0.04 (80%)	0.95 +/- 0.07 (**87%**)	0.93 +/- 0.06 (82%)	0.95 +/- 0.07 (81%)
Sensitivity	0.94 +/- 0.08 (100%)	0.93 +/- 0.13 (**96%**)	0.93 +/- 0.08 (84%)	0.97 +/- 0.07 (100%)
Specificity	0.97 +/- 0.07 (73%)	0.97 +/- 0.07 (**83%**)	0.93 +/- 0.08 (81%)	0.93 +/- 0.13 (75%)

Table A-8. Compound Annotation and Identification for the Panel of Five Metabolites Upregulated in RCC for RCC Detection.

ID no.	Retention Time (min)	m/z		Adduct Type	Mass error (ppm)	Elemental Formula	Metabolite Identity
		Theoretical	*Experimental*				
1481	8.83	260.1969	260.1969	$[M+H]^+$	0.00	$C_{12}H_{25}N_3O_3$	Lys-Ile
2102	4.39	130.1590	130.1591	$[M+H]^+$	0.77	$C_8H_{19}N$	dibutylamine, N-butylisobutylamine, diisobutylamine (isomer)
6578	1.09	165.9790	165.9784	$[M-H]^-$	-3.61	$C_7H_5NS_2$	2-mercaptobenzothiazole
6594	6.89	236.0776	236.0777	$[M+H]^+$	0.42	$C_8H_{15}NO_7$	N-acetyl-glucosaminic acid
5698	3.38	630.1909	630.1895	$[M-H]^-$	-2.64	$C_{24}H_{43}NO_{12}P_2S$	--

Table A-9. Hyperparameters Tuned for Machine Learning Methods used for Binary Classification for the Upregulated RCC Biomarkers in the RCC Detection Study.

Parameters	Initial distribution	Optimized
Random Forest		
Max_depth	10, 20, 30	10
Max_features	'auto', 'sqrt', 'log2'	'auto'
Min_samples_leaf	1, 2, 3, 4, 5	1
Min_samples_split	2, 4, 6, 8	2
N_estimators	50, 100, 150, 200	150
SVM-RBF		
C	0.1, 1, 10, 100	10
gamma	0.01, 0.03, 0.1, 0.3, 1.0	0.3
Lin-SVM		
C	0.001, 0.01, 0.1, 1, 5, 10	10
k-NN		
Number of neighbors	2 - 30	11
Distance Measure	Manhattan, Euclidean	Euclidean

Table A-10. Machine Learning Performance using the Upregulated RCC Biomarkers in the RCC Detection Study.

Algorithm	**RF**	**K-NN**	**SVM-RBF**	**Linear SVM**
AUC	0.95 +/- 0.06 (0.97)	0.92 +/- 0.1 (**0.92**)	0.9 +/- 0.08 (0.80)	0.9 +/- 0.08 (0.91)
Accuracy	0.87 +/- 0.04 (70%)	0.77 +/- 0.14 (**81%**)	0.89 +/- 0.1 (61%)	0.82 +/- 0.07 (71%)
Sensitivity	0.87 +/- 0.12 (98%)	0.7 +/- 0.27 (**86%**)	0.87 +/- 0.16 (96%)	0.8 +/- 0.12 (94%)
Specificity	0.87 +/- 0.12 (59%)	0.83 +/- 0.18 (**79%**)	0.9 +/- 0.13 (49%)	0.83 +/- 0.11 (62%)

Table A-11. Detailed MS/MS Information for the 7-Metabolite Panel that Distinguishes RCC from Control Samples.

Feature ID no.	CE (eV), Mode	Fragment ion m/z	Metabolite identification level	Match details	Metabolite Name
720	10,30, 45 (+)	**136.0755**, **119.0489, 118.0648**, 107.0490, **91.0541,** 95.0602, 101.2184, 87.466, 70.9133, **65.0383**	2	Fragmentation consistent with spectrum (HMDB)	2-phenylacetamide
1481	10,30, 45 (+)	260.1977**,** 171.1495, 144.1021, 100.0757, 84.0808**,** 72.0444, 65.8753, 54.7500	2	Fragmentation consistent with structure	lys-ile or lys-leu
2102	10,30, 45 (+)	130.1592, 84.0445, 74.0965, 57.0700	2	Fragmentation consistent with structure	dibutylamine (alkyl chain branching not determined)
3804	---	**180.0880**, **105.0339, 162.0771**, 95.0497, 53.0395, 110.0345,120.0812, 138.0556	2	Fragmentation consistent with spectrum (m/z cloud)	hippuric acid
6262	10,30, 45 (-)	342.2521, 310.2751, 280.0701, **194.0449,** 181.0375**, 150.0594**, **148.0392**, 138.0297, **121.0285**, **124.0061**, **113.0230**, **93.0332, 85.0281, 89.0240, 71.0134, 73.0290**	2	Fragmentation consistent with spectrum (m/z cloud)	hippurate-mannitol derivative
6578	10,30, 45 (-)	**165.9784**, **134.0065**, 122.01241, 117.9193, **102.0344**, 79.9570, 66.0093, **57.9752**	2	Fragmentation consistent with spectrum (m/z cloud)	2-mercaptobenzothiazole

6594	10,30, 45 (+)	236.0773, 230.1947, 208.9668, 149.0452, 131.0346, 119.0349, 113.0241, 104.0350, 98.9556, 89.0240, 86.0244, 85.0291, 71.0134, 59.0134,	2	Fragmentation consistent with structure	N-acetyl-glucosaminic acid

The m/z of fragment ions were obtained from DDA experiments. The corresponding collision energy (CE) is also listed in the table. Selected precursor ions are underlined. Fragment ions that matched to literature spectra or were consistent with potential structures are in bold. Metabolite identification level was assigned based on the following criteria: 1) exact mass, isotopic pattern, retention time, and MS/MS spectrum of standard matched to the feature. 2) exact mass, isotopic pattern, and MS/MS spectrum matched with literature spectra or fragmentation ions observed are consistent with the proposed structure. 3) tentative ID assignment based on elemental formula matches with literature. 4) unknowns.

Table A-12. Machine Learning Hyperparameters Tuned for Binary Classification using the 7-Metabolite Panel for RCC Detection.

Parameters	Initial distribution	Optimized
Random Forest		
Max_depth	10, 20, 30	10
Max_features	'auto', 'sqrt', 'log2'	'auto'
Min_samples_leaf	1, 2, 3, 4, 5	5
Min_samples_split	2, 4, 6, 8	2
N_estimators	50, 100, 150, 200	50
SVM-RBF		
C	0.1, 1, 10, 100	1
gamma	0.01, 0.03, 0.1, 0.3, 1.0	0.03
Lin-SVM		
C	0.001, 0.01, 0.1, 1, 5, 10	0.1
k-NN		
Number of neighbors	2 - 30	7
Distance Measure	Manhattan, Euclidean	Manhattan

Table A-13. Machine Learning Hyperparameters Tuned for Binary Classification using NMR-derived Metabolites for RCC Detection.

Parameters	Initial distribution	Optimized
Random Forest		
Max_depth	10, 20, 30	10
Max_features	'auto', 'sqrt', 'log2'	'auto'
Min_samples_leaf	1, 2, 3, 4, 5	3
Min_samples_split	2, 4, 6, 8	2
N_estimators	50, 100, 150, 200	50
SVM-RBF		
C	0.1, 1, 10, 100	10
gamma	0.01, 0.03, 0.1, 0.3, 1.0	0.3
Lin-SVM		
C	0.001, 0.01, 0.1, 1, 5, 10	1
k-NN		
Number of neighbors	2 - 30	5
Distance Measure	Manhattan, Euclidean	Manhattan

Table A-14. Machine Learning Performance using NMR-derived Metabolites for RCC Detection.

Algorithm	***RF***	***K-NN***	***SVM-RBF***	***Linear SVM***
AUC	0.95 +/- 0.03 (0.89)	0.94 +/- 0.05 (0.88)	0.94 +/- 0.06 (**0.89**)	0.89 +/- 0.08 (0.87)
Accuracy	0.84 +/- 0.05 (0.76)	0.87 +/- 0.1 (0.77)	0.86 +/- 0.06 (**0.78**)	0.81 +/- 0.08 (0.74)
Sensitivity	0.81 +/- 0.12 (0.86)	0.87 +/- 0.06 (0.84)	0.9 +/- 0.08 (**0.86**)	0.83 +/- 0.11 (0.84)
Specificity	0.87 +/- 0.19 (0.72)	0.87 +/- 0.19 (0.75)	0.81 +/- 0.12 (**0.76**)	0.77 +/- 0.23 (0.71)

Table A-15. Metabolomic Features with *q*-values < 0.05 and > 1-Fold Change in the Model Cohort for the RCC Detection Study.

(Fold change (FC) was calculated as the base 2 logarithm of the average intensity ratios between RCC and controls samples). Positive FC values indicate increased abundance in RCC, while negative values indicate higher abundance in control samples. *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after an independent *t*-test. 433 features from LC-MS are included in the table, and 2 features from NMR (*scyllo*-Inositol and aminohippurate) are in Table A-4.

ID	Metabolite ID	Mode	RT [min]	FC	FDR *p*-value
1	1	positive	2.317	-1.4885	0.045959
9	5-acetylamino-6-amino-3-methyluracil	positive	1.901	-1.22771	0.008743
95	3-(1H-1,2,4-Triazol-3-yl)alanine	positive	1.898	-1.23304	0.005531
147	147	positive	2.58	-1.20166	0.007443
163	venlafaxine	positive	2.654	-1.27957	0.018961
170	170	positive	2.313	-1.56798	0.017014
173	173	positive	2.605	-1.23927	0.004285
245	245	positive	4	-2.33614	0.00089
250	250	positive	2.581	-2.11065	0.009913
260	leupeptin	positive	3.989	-2.30823	0.000917
278	278	positive	3.057	-1.65576	0.002262
293	293	positive	2.617	-1.64467	0.020423
312	312	positive	4.523	1.113604	0.022578
314	314	positive	2.592	-2.15533	0.012163
332	N-(3-amino-4-methyl-5-nitrophenyl)acetamide	positive	1.591	-2.21385	0.035889
347	6-methylquinoline	positive	2.322	-1.64636	0.008933
363	363	positive	1.903	-1.38184	0.003166
429	429	positive	4.525	1.279205	0.018937
435	paraxanthine	positive	1.12	-1.17746	0.006372
474	474	positive	1.917	-1.29632	0.009562
479	phenylacetaldehyde	positive	2.404	-1.95368	0.011539
523	523	positive	2.183	-1.06828	0.017172
562	6-methylquinoline	positive	2.594	-1.95682	0.010866
595	595	positive	2.67	-2.75395	0.0418
610	610	positive	2.952	-1.95838	0.005316
640	1_5-anhydro-mannitol	positive	1.895	-1.22667	0.00129

643	643	positive	2.869	-2.0903	0.03379
666	666	positive	2.961	-1.67651	0.009093
672	moxaverine	positive	2.66	-1.15696	0.039852
688	688	positive	4.009	1.099602	0.023352
712	3-dehydrocarnitine	positive	1.449	-1.19664	0.02724
720	2-phenylacetamide	positive	2.562	-1.06888	0.000386
726	726	positive	6.009	1.043296	0.04006
790	2-acetolactate	positive	1.914	-1.03801	0.001699
798	indole;1-benzazole	positive	2.653	-1.46462	0.000526
800	800	positive	4.125	3.41514	0.033921
819	4-dehydropantoate	positive	1.896	-1.19163	0.00116
825	3-methyldioxyindole	positive	1.942	-1.20393	0.002882
880	880	positive	4.528	1.182288	0.032715
900	900	positive	4.532	1.172157	0.02614
926	926	positive	2.25	-1.29961	0.015545
954	954	positive	4.019	-2.6166	0.000917
958	958	positive	0.579	-2.49872	0.032106
960	960	positive	0.58	-2.4511	0.034206
995	995	positive	0.882	-1.5801	0.035476
1035	zeatin	positive	3.628	1.022119	0.039696
1047	1047	positive	2.631	-2.17057	0.012641
1066	1066	positive	2.591	-2.28759	0.005362
1098	4-imidazolone-5-propanoate	positive	3.729	-1.0197	0.027361
1153	1153	positive	4.533	1.222509	0.017014
1163	1163	positive	2.45	-1.39804	0.00089
1214	1214	positive	1.242	-1.57143	0.020423
1261	1261	positive	1.855	1.046534	0.04275
1262	1262	positive	1.9	-1.49775	0.003903
1307	1307	positive	1.147	-5.46823	0.001452
1356	delta-guanidinovalericacid	positive	3.929	-1.93217	0.043481
1365	1365	positive	4.524	1.098536	0.009093
1391	N- glucosylarylamine	positive	3.97	1.803794	0.04275
1481	lys-Ile	positive	6.29	1.447371	0.001342
1550	1550	positive	2.611	-2.04338	0.012124
1566	1566	positive	3.909	-1.29952	0.035716
1579	1579	positive	5.937	1.358104	0.041545
1587	1587	positive	4.509	1.376935	0.024
1662	1662	positive	2.502	2.944322	0.047485
1673	1673	positive	0.804	-3.74992	0.001296
1676	1676	positive	2.648	-1.80767	0.01703
1689	1689	positive	4.715	4.292918	0.017731

1691	1691	positive	3.418	-1.91163	0.040973
1701	4-(2-aminoethyl)benzenesulfonylfluoride	positive	1.807	-1.93725	0.000526
1723	1723	positive	2.574	-1.7161	0.00597
1741	PEG monolaurate n5	positive	3.738	-2.74589	0.001453
1760	4-(2-aminoethyl)benzenesulfonylfluoride	positive	1.704	-2.14754	0.000826
1771	1771	positive	3.001	1.262088	0.030398
1776	N-acetylleucylleucine	positive	2.587	-1.89519	0.021927
1820	1820	positive	1.953	-2.10726	0.004044
1839	2_3-diaminopropionicacid	positive	2.6	-2.15507	0.008703
1858	1858	positive	3.883	-3.72213	0.006498
1861	1861	positive	4.81	-2.79138	0.019817
1870	1870	positive	3.824	-3.63177	0.010482
1880	1880	positive	4.528	1.332245	0.017731
1931	anisole	positive	2.618	-1.12591	0.009093
1942	1942	positive	3.603	-1.29462	0.049502
1961	1961	positive	2.434	-1.87936	0.000515
1969	1969	positive	2.583	-1.36431	0.009913
1991	1991	positive	3.247	-3.24599	0.015568
2005	2005	positive	4.876	-2.9057	0.016077
2056	2056	positive	2.649	-1.55177	0.008562
2082	2082	positive	2.594	-1.30732	0.008588
2102	dibutylamine, N-butylisobutylamine, diisobutylamine (isomer)	positive	3.449	1.567367	7.87E-05
2138	2138	positive	1.12	-1.15134	0.014564
2158	2158	positive	4.528	1.256293	0.021194
2196	2196	positive	0.814	-2.89012	0.00089
2207	2207	positive	1.097	-3.82236	0.003166
2218	2218	positive	2.432	-1.70991	0.000586
2230	2230	positive	2.59	-2.19533	0.006379
2239	2239	positive	1.101	-3.85416	0.002882
2240	2240	positive	2.886	1.631405	0.00451
2241	2241	positive	4.838	-1.06192	0.015545
2242	2242	positive	2.587	-2.31188	0.006372
2259	2259	positive	3.706	-2.85038	0.032369
2267	beta-Ionone	positive	2.371	-2.11197	0.011891
2306	2306	positive	2.623	-1.92523	0.009819
2320	2320	positive	0.866	-2.38259	0.000912
2321	2321	positive	1.138	-3.78215	0.001432
2348	2348	positive	3.89	-1.05941	0.02614
2353	leucinamide	positive	1.118	-3.10452	0.003468
2359	2359	positive	6.637	1.794645	0.023123

2385	2385	positive	4.513	1.09136	0.01182
2403	2403	positive	4.53	1.352297	0.012999
2418	2418	positive	4.477	-3.95489	0.027524
2426	2426	positive	2.851	1.572021	0.035289
2448	2448	positive	4.712	6.772872	0.024075
2455	2455	positive	1.14	-1.32081	0.00129
2462	2462	positive	4.536	1.262401	0.005734
2476	2476	positive	4.534	1.481945	0.027785
2496	2496	positive	2.581	-2.27665	0.004492
2499	2499	positive	4.508	1.286938	0.027563
2524	2524	positive	1.299	1.044756	0.048626
2540	2540	positive	4.732	7.106785	0.020956
2568	2568	positive	3.711	-1.24199	0.013193
2571	2571	positive	3.122	-1.9284	0.035889
2577	pipecolicacid	positive	1.457	-1.52805	0.013548
2601	2601	positive	2.658	-2.07536	0.009562
2621	2621	positive	3.989	-2.63727	0.001453
2625	2625	positive	1.869	-1.023	0.022102
2652	2652	positive	4.544	1.009831	0.04275
2653	2653	positive	2.602	-1.53429	0.016708
2668	cyclo(leucylprolyl)	positive	1.067	-1.04975	0.048489
2702	2702	positive	4.518	1.167157	0.018248
2709	N-acetyl- glucosaminate	positive	3.857	1.090917	0.037911
2731	2731	positive	4.714	4.920091	0.007029
2732	2732	positive	4.653	4.191832	0.017527
2749	2_3-dimethylmalate	positive	2.894	1.338155	0.043239
2803	2803	positive	4.539	1.220539	0.005355
2804	2804	positive	1	-3.42199	0.009073
2809	2809	positive	3.775	-1.14183	0.016836
2815	2815	positive	4.557	1.935817	0.047151
2821	2821	positive	4.504	1.247197	0.033979
2829	2829	positive	0.823	-1.68661	0.001711
2840	2840	positive	3.317	-2.53883	0.005362
2850	2850	positive	3.157	-2.40116	0.001896
2852	2852	positive	4.529	1.310407	0.021409
2853	2853	positive	0.923	-3.21639	0.003894
2860	2860	positive	4.531	1.181536	0.030398
2905	2905	positive	2.579	-1.84755	0.013293
2914	2914	positive	5.316	-3.13431	0.00451
2924	2924	positive	4.727	5.847712	0.023309
2926	2926	positive	4.809	-2.25885	0.018961

2932	2932	positive	2.649	-1.82929	0.018961
2971	2971	positive	4.531	1.228874	0.013964
2973	2973	positive	2.385	-1.31894	0.004285
2986	2986	positive	0.904	-2.80027	0.023728
2992	2992	positive	3.323	-4.44632	0.014442
3025	3025	positive	4.7	3.344933	0.032349
3033	3033	positive	4.52	1.085292	0.013193
3035	3035	positive	4.53	1.226327	0.022578
3043	3043	positive	4.532	1.480025	0.016585
3074	3074	positive	4.539	1.62393	0.017014
3082	3082	positive	1.951	-1.10581	0.004436
3116	3-hydroxyaminophenol	positive	1.107	1.515954	0.040104
3127	3127	positive	1.232	-1.43787	0.02724
3141	3141	positive	1.133	-2.42642	0.001216
3148	3148	positive	0.812	-3.09632	0.000826
3154	3154	positive	3.884	-2.69949	0.000828
3159	3159	positive	1.095	-1.68751	0.011539
3160	3160	positive	0.927	-1.16062	0.008933
3169	2-hydroxyphenethylamine	positive	4.094	2.210126	0.042089
3171	3171	positive	1.131	-3.39246	0.011164
3175	3175	positive	2.581	-1.89816	0.003896
3200	3200	positive	3.487	1.035815	0.038541
3208	3208	positive	1.098	-5.19983	0.00451
3234	3234	positive	1.147	-2.41971	0.016388
3260	3260	positive	3.056	1.146707	0.012962
3262	3262	positive	4.751	3.143639	0.048424
3283	3283	positive	2.138	-1.42525	0.006117
3297	3297	positive	4.706	6.646371	0.016585
3301	3301	positive	4.531	1.001697	0.016003
3309	3309	positive	0.934	-2.04523	0.003896
3353	3353	positive	4.509	1.111479	0.043452
3362	3362	positive	4.073	-5.08514	0.001646
3370	3370	positive	4.703	6.087944	0.014648
3371	3371	positive	0.922	-3.51338	0.003894
3385	3385	positive	3.329	-2.60407	0.008088
3390	3390	positive	2.633	-1.751	0.006372
3409	3409	positive	1.857	-1.26767	0.014577
3415	3415	positive	4.534	1.165491	0.009093
3427	3427	positive	2.631	-2.44854	0.013293
3441	2-hydroxyphenethylamine	positive	3.894	2.55209	0.033619
3446	deoxycytidine	positive	3.531	1.085392	0.0313

3449	3449	positive	1.861	-1.22268	0.008986
3492	cathinone	positive	4.181	1.918288	0.022548
3514	3514	positive	4.533	1.39721	0.016455
3517	pyridafenthion	positive	0.812	-3.76591	0.00132
3526	3526	positive	1.914	-1.09411	0.002947
3528	3528	positive	2.258	-1.87687	0.015545
3545	3545	positive	0.826	-1.63608	0.00144
3546	3546	positive	0.827	-1.63797	0.00144
3552	3552	positive	4.531	1.223489	0.013193
3558	3558	positive	4.54	1.239042	0.022102
3564	3564	positive	0.848	2.507408	0.019028
3582	3582	positive	1.214	-1.5372	0.011539
3586	3586	positive	4.159	-2.72666	0.014442
3596	beta-ionone	positive	2.632	-1.75123	0.003896
3613	3613	positive	3.957	-2.40024	0.000586
3624	3624	positive	1.069	-1.34416	0.021839
3626	3626	positive	4.523	1.198337	0.01772
3632	3632	positive	4.555	1.298889	0.013193
3657	3657	positive	1.293	-1.85846	0.009093
3675	3675	positive	1.184	-1.08046	0.000262
3701	3701	positive	0.819	-1.69784	0.00105
3757	3757	positive	4.003	-2.38649	0.000326
3763	3763	positive	4.009	-2.27464	0.001528
3764	3764	positive	3.954	-2.33779	0.00089
3777	3777	positive	3.639	-2.09303	0.00144
3791	3791	positive	3.69	-2.04221	0.002947
3797	3797	positive	3.938	-2.28156	0.001221
3799	3799	positive	3.621	-2.11755	0.001216
3804	hippuric acid	positive	2.595	-2.02465	0.000526
3820	3820	positive	3.979	-2.26378	0.000579
3823	3823	positive	3.924	-2.42884	0.000734
3829	3829	positive	3.894	-2.35241	0.001496
3842	3842	positive	3.673	-2.06698	0.001737
3856	3856	positive	4.016	-2.4126	0.000734
3863	3863	positive	3.373	-2.05737	0.043366
3871	3871	positive	1.021	-2.67385	0.002122
3872	3872	positive	4.049	-2.5791	0.00019
3873	3873	positive	3.698	-2.19666	0.004044
3876	3876	positive	3.998	-2.39598	0.000515
3893	3893	positive	3.727	-2.47178	0.004492
3905	3905	positive	3.87	-2.37325	0.000826
3906	3906	positive	3.848	-2.54952	0.003254

3907	3907	positive	3.779	-2.70752	0.004285
3909	3909	positive	3.579	-2.32192	0.00089
3923	3923	positive	4.016	-3.24391	0.00089
3925	3925	positive	3.751	-2.62574	0.004245
3939	3939	positive	3.304	-2.57864	0.028679
3944	3944	positive	3.035	1.066588	0.048739
3960	3960	positive	3.996	-2.13732	0.001054
3963	3963	positive	2.184	1.099552	0.008986
3968	3968	positive	4.566	-2.64268	0.01657
3976	3976	positive	4.73	6.193516	0.022548
3987	3987	positive	3.971	-2.30352	0.000679
3991	leucinamide	positive	1.074	-3.08451	0.002279
3992	3992	positive	3.954	-2.7693	0.001256
4001	4001	positive	1.099	-4.10801	0.025007
4025	4025	positive	3.439	1.279099	0.008317
4042	4042	positive	3.883	-2.52495	0.000734
4075	4075	positive	3.416	-2.03456	0.008933
4080	4080	positive	0.821	-3.4898	0.000515
4133	4133	positive	4.036	-2.72562	0.004666
4162	4162	positive	3.455	-2.11785	0.005543
4179	4179	positive	3.602	-1.00487	0.021297
4180	4180	positive	3.622	-1.00845	0.018248
4189	4189	positive	1.281	-1.58871	0.0418
4195	4195	positive	1.114	-1.26086	0.008469
4210	4210	positive	3.555	-2.31772	0.001701
4218	4218	positive	3.48	-2.25802	0.00129
4250	4250	positive	4.057	-2.558	0.00132
4258	5-hydroxy-tryptophan	positive	3.033	1.202781	0.04448
4265	4265	positive	4.102	-2.60288	0.00129
4267	4267	positive	3.758	-1.26957	0.04047
4278	4278	positive	3.072	1.28831	0.011416
4279	4279	positive	6.949	2.098916	0.046946
4281	4281	positive	4.05	-2.76374	0.00089
4283	4283	positive	4.687	2.857637	0.017731
4288	4288	positive	4.612	1.304232	0.030454
4303	4303	positive	2.623	-1.02546	0.00597
4318	4318	positive	4.037	-2.72294	0.00073
4323	4323	positive	4.122	-2.34523	0.001873
4328	4328	positive	3.96	-2.35681	0.002367
4340	4340	positive	5.076	-1.9714	0.02593
4352	leucinamide	positive	0.919	-2.16241	0.00132

4355	4355	positive	4.057	-2.73285	0.00129
4367	pregabalin	positive	1.06	-2.97568	0.001496
4370	4370	positive	4.113	-2.29843	0.002367
4381	4381	positive	4.646	1.122786	0.030398
4382	4382	positive	4.541	-3.18497	0.004544
4384	4384	positive	4.073	-2.25311	0.001565
4392	4392	positive	3.941	-2.10559	0.000734
4401	4401	positive	4.039	-2.69474	0.000481
4408	1-aAmino-1-deoxy-scyllo-inositol	positive	3.867	-1.12923	0.011113
4413	4413	positive	4.052	-2.37203	0.001699
4428	4428	positive	4.057	-2.22436	0.002596
4444	4444	positive	3.998	-2.43206	0.000912
4447	4447	positive	4.061	-2.50909	0.000579
4490	4490	positive	3.94	-2.3298	0.001721
4553	4553	positive	3.85	-2.1915	0.002694
4587	4587	positive	4.035	-2.55891	0.002251
4616	4616	positive	0.847	1.303658	0.008088
4632	1,7-dimethyluric acid	negative	2.337	-1.11272	0.013263
4659	hippuric acid	negative	2.622	-1.12485	0.003896
4670	cinnamoylglycine	negative	2.651	-1.55137	0.012441
4672	cinnamoylglycine	negative	2.616	-1.814	0.017993
4673	7-methylxanthine	negative	1.221	-1.13939	0.009031
4685	hippuric acid	negative	4.039	1.414079	0.032791
4704	2-furoylglycine	negative	2.58	1.951921	0.04693
4706	2-furoylglycine	negative	2.512	2.68069	0.030454
4719	1-methyluric acid	negative	2.545	-1.157	0.011539
4739	5-hydroxyindole	negative	2.626	-1.0839	0.004436
4740	5-hydroxyindole	negative	2.581	-1.15681	0.005204
4766	theophylline	negative	1.432	-2.1935	0.001975
4775	4775	negative	0.869	-2.28845	0.027288
4791	4791	negative	4.511	1.244399	0.016348
4801	4801	negative	4.554	1.085615	0.016222
4844	4844	negative	3.257	2.311035	0.036851
4870	3-[-5-oxo-7-oxabicyclo[4.1.0]hept-2-yl]-alanine	negative	3.24	-1.41362	0.015568
4953	7,8-dihydro-8-oxoguanine	negative	1.198	-1.23311	0.008986
4958	4958	negative	3.679	-1.25361	0.048489
5010	5010	negative	4.637	2.953354	0.003254
5029	5029	negative	4.152	-2.73447	0.008511
5038	5038	negative	2.632	2.080506	0.037266
5065	5065	negative	0.855	-1.47704	0.003254
5077	5077	negative	1.965	1.045916	0.012965

5083	5083	negative	2.564	1.75927	0.04448
5106	5106	negative	3.534	1.315406	0.025624
5110	1-O-[5-(3,4-dihydroxyphenyl)-4-hydroxypentanoyl]-beta-glucopyranuronic acid	negative	3.5	-1.96301	0.0151
5129	5129	negative	2.64	-1.94809	0.00129
5133	5133	negative	3.707	-2.21718	0.032369
5137	5137	negative	2.517	-1.15218	0.023309
5153	5153	negative	1.421	-1.48097	0.007752
5216	5216	negative	0.863	-1.85939	0.005858
5239	5239	negative	3.529	-1.09635	0.043366
5241	butopyronoxyl	negative	0.908	1.206619	0.038863
5248	5248	negative	4.531	1.057774	0.012999
5285	5285	negative	4.644	3.023183	0.009913
5310	5310	negative	0.87	2.988694	0.017548
5341	5341	negative	0.706	-1.09358	0.03992
5352	5352	negative	2.355	-1.00687	0.04099
5358	5358	negative	3.897	1.636177	0.021839
5379	5379	negative	0.863	-1.86478	0.009093
5381	5381	negative	3.395	1.698478	0.004044
5383	5383	negative	4.067	-2.62497	0.000385
5393	5393	negative	0.862	-1.87627	0.006038
5423	5423	negative	3.745	-1.83453	0.030529
5463	N-acetyl-methionine	negative	3.758	-1.85466	0.033538
5470	sulfurol acetate	negative	3.902	-2.52855	0.034206
5482	5482	negative	0.858	-1.17721	0.034734
5507	5507	negative	3.132	-1.06713	0.01418
5514	5514	negative	2.517	-1.48437	0.004245
5553	5553	negative	2.562	-1.25185	0.03108
5576	5576	negative	0.728	-1.56108	0.008088
5604	5604	negative	3.891	-3.78571	0.003896
5612	5612	negative	3.805	-1.02889	0.043481
5647	5647	negative	3.993	-2.70499	0.000679
5648	5648	negative	3.966	-2.75565	0.006117
5683	5683	negative	4.089	-3.72301	0.001247
5698	5698	negative	3.381	1.402511	0.011539
5712	5712	negative	4.376	-1.96165	0.029132
5724	5724	negative	4.531	1.170521	0.015527
5728	5728	negative	2.931	-1.11438	0.039852
5737	gly-Lys	negative	3.995	1.486072	0.017371
5770	5770	negative	0.855	-1.45399	0.00451
5796	pidotimod	negative	2.951	1.72657	0.040795

5799	5799	negative	4	-2.4648	0.038921
5818	5818	negative	0.948	1.355602	0.048525
5820	5820	negative	4.663	1.182431	0.020072
5849	5849	negative	0.708	-1.10963	0.039503
5868	5868	negative	4.66	1.321084	0.010661
5887	5887	negative	4.532	-1.65544	0.001001
5899	5899	negative	3.5	1.151716	0.032924
5911	5911	negative	3.966	-2.51989	0.001092
5925	5925	negative	2.671	-1.32396	0.006117
5931	5931	negative	2.571	-1.02026	0.010661
5941	5941	negative	1.148	1.901496	0.032924
5942	5942	negative	2.631	-1.18178	0.002324
5994	5994	negative	2.625	-1.08762	0.00451
6001	6001	negative	2.782	-1.94029	0.006743
6007	6007	negative	0.678	-1.21015	0.041623
6014	N-acetyl-tyrosine	negative	2.632	-1.5866	0.009093
6021	6021	negative	0.708	-1.36188	0.003896
6057	6057	negative	0.947	1.642112	0.031478
6094	5-acetylamino-6-formylamino-3-methyluracil	negative	0.727	-1.59148	0.004603
6095	chiro-inositol	negative	1.118	-1.37163	0.00129
6101	6101	negative	4.052	-1.52997	0.045816
6111	6111	negative	2.577	2.205537	0.04782
6148	6148	negative	2.587	1.33016	0.028806
6161	6161	negative	0.711	-1.44766	0.005882
6190	6190	negative	2.594	-1.31465	0.027361
6212	glaucarubin	negative	3.361	-2.88626	0.025629
6233	6233	negative	1.515	-1.15693	0.015545
6236	6236	negative	0.625	1.384644	0.025156
6261	6261	negative	2.591	-1.8809	0.000368
6262	hippurate-mannitol derivative	negative	2.667	-1.80394	0.000515
6267	6267	negative	2.581	2.159467	0.032369
6276	6276	negative	2.636	-1.65054	0.000481
6286	2_5_6-trihydroxy-5_6-dihydroquinoline	negative	2.622	-1.30303	0.005058
6314	6314	negative	2.625	-1.08453	0.004492
6322	6322	negative	2.623	-1.04139	0.00451
6325	6325	negative	2.64	-1.97981	0.001564
6327	6327	negative	2.637	-1.5487	0.028679
6337	6337	negative	0.862	-1.69061	0.002457
6348	6348	negative	2.993	-2.21045	0.017548
6349	6349	negative	0.93	-1.14826	0.021927
6361	6361	negative	2.619	-1.1059	0.001844

6375	6375	negative	2.775	-1.81499	0.009031
6385	6385	negative	2.655	-1.00016	0.008703
6389	6389	negative	2.516	-1.13448	0.011046
6390	6390	negative	2.627	-1.00038	0.004603
6392	6392	negative	0.75	-1.01094	0.012242
6396	6396	negative	0.884	-1.32351	0.027563
6406	6406	negative	0.903	-2.03474	0.04006
6425	leupeptin	negative	4.004	-2.47724	0.00119
6447	6447	negative	2.421	3.406793	0.024455
6496	6496	negative	1.25	-1.16132	0.008933
6504	6504	negative	2.582	2.311681	0.032369
6508	6508	negative	2.569	-1.00789	0.004285
6534	6534	negative	3.989	-2.43582	0.001737
6544	6544	negative	0.719	-1.5791	0.004603
6545	6545	negative	4.005	-2.41651	0.001342
6565	6565	negative	2.935	-1.08416	0.041344
6569	4-[(-2-amino-1-hydroxyethyl]-2-hydroxyphenyl hydrogen sulfate	negative	1.181	1.368161	0.047987
6578	2-mercaptobenzothiazole	negative	0.832	2.229249	0.009064
6594	N-acetyl-glucosaminate	negative	3.871	1.156155	0.021542
6628	6628	negative	2.908	-1.35524	0.022578
6637	6637	negative	4.866	1.31038	0.01503
6662	6662	negative	3.528	-2.06367	0.011539
6668	6668	negative	2.612	-1.00887	0.004285
6676	6676	negative	3.184	-2.34598	0.002651
6683	6683	negative	3.025	-2.14478	0.012441
6687	6687	negative	0.866	-1.68528	0.003166
6731	6731	negative	0.859	-1.79499	0.002105
6762	6762	negative	3.751	-2.80798	0.003896
6802	6802	negative	1.504	-1.71402	0.030398
6819	6819	negative	1.109	-3.47044	0.008588
6882	6882	negative	3.692	-3.0581	0.011543
6885	6885	negative	2.615	-1.03401	0.002324
6939	6939	negative	2.586	-1.33936	0.043481
6956	6956	negative	2.715	1.925851	0.017014
6972	6972	negative	1.112	-3.3917	0.008088
6990	6990	negative	2.647	-1.67258	0.011539
6996	6996	negative	2.535	-1.04791	0.021194
7001	7001	negative	0.869	-1.71768	0.003166
7087	7087	negative	2.399	-1.19546	0.007612

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

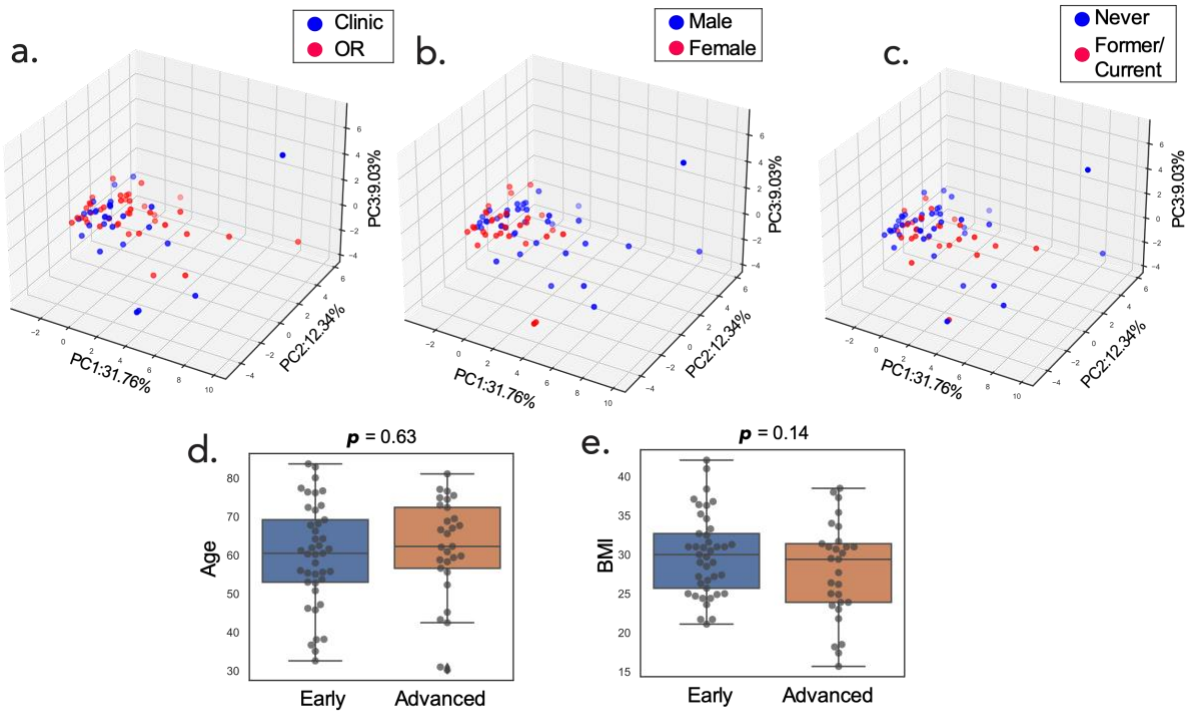


Figure B-1. Potential Confounder Analysis for RCC Stage Stratification.

PCA was conducted using the 24-metabolite panel as features. PCA shows that collection method (a), gender (b), and smoking history (c) are not discriminated by the selected biomarker panel. Age (d) and BMI (e) in the cohort are statistically insignificant between early and advanced RCC patients (Student's *t*-Test)

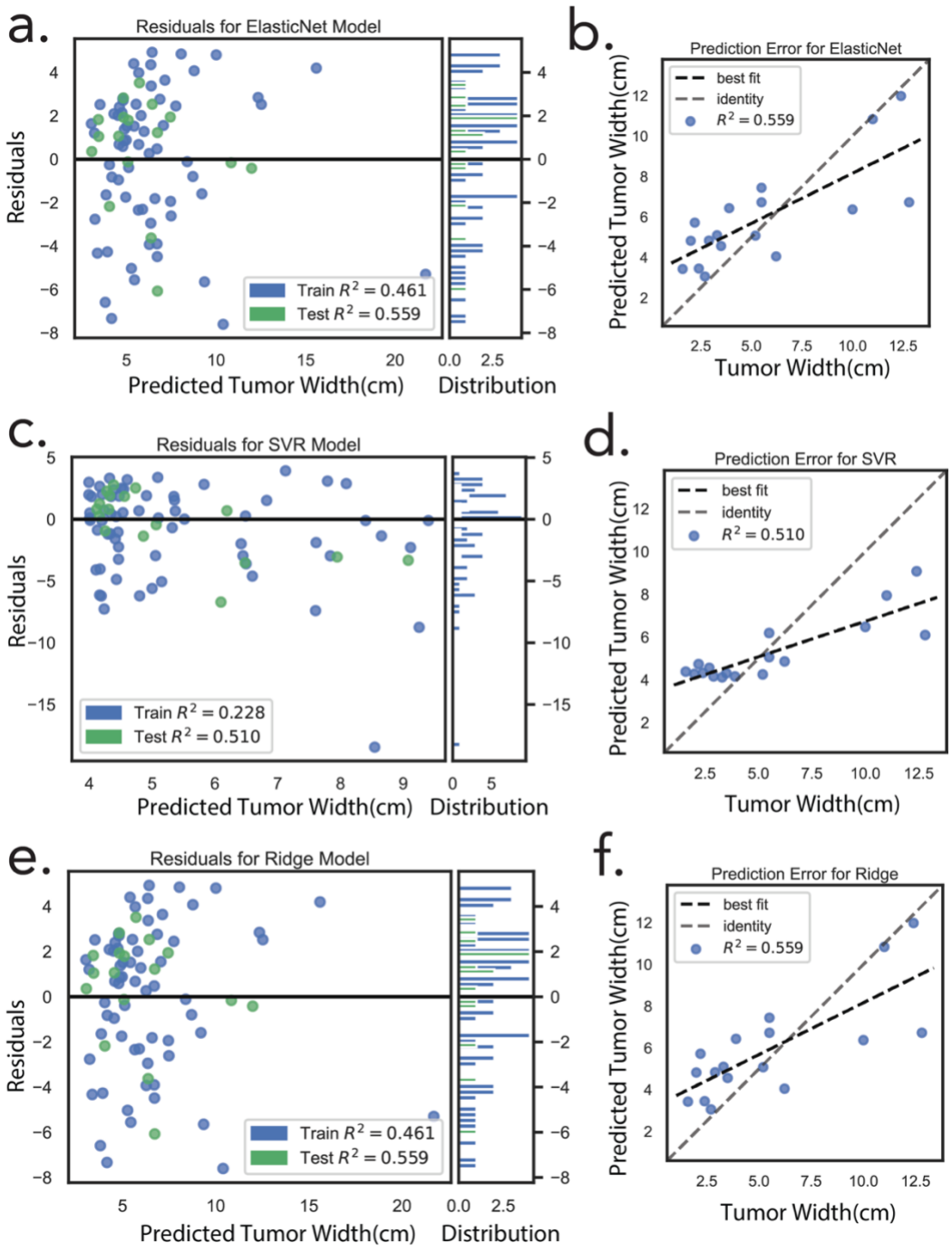


Figure B-2. RCC Primary Tumor Size Predictions.

(a) Elastic net model residual plot. **(b)** Elastic net model prediction error plot. **(c)** Support vector regression model residual plot. **(d)** Support vector regression model prediction error plot. **(e)** Ridge model residual plot. **(f)** Ridge model prediction error plot.

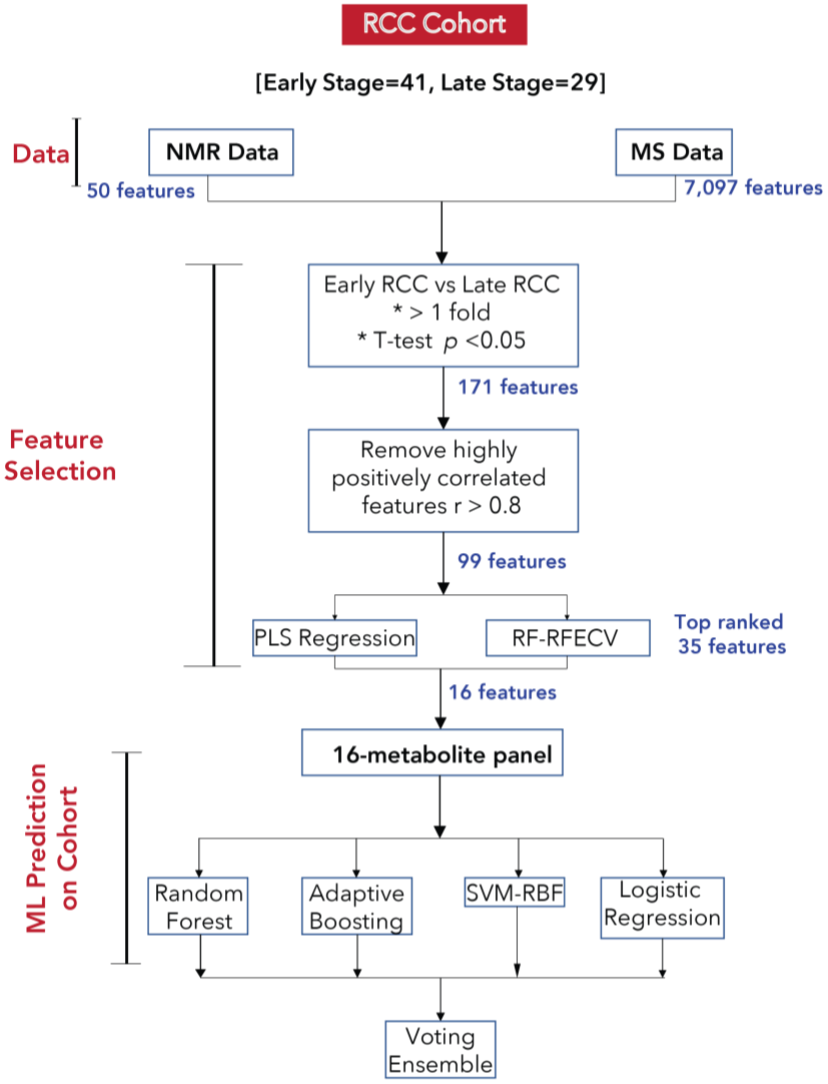


Figure B-3. Machine Learning Pipeline for the Biomarker Selection for RCC Stage Stratification.

All NMR and MS features were subjected to a hybrid method of feature selection resulting in a 16-metabolite panel. Machine learning predictions were carried out by four different algorithms and a voting ensemble. PLS: partial least squares. RF- RFECV: random forest recursive feature elimination – cross validation. SVM-RBF: support vector machines radial basis function.

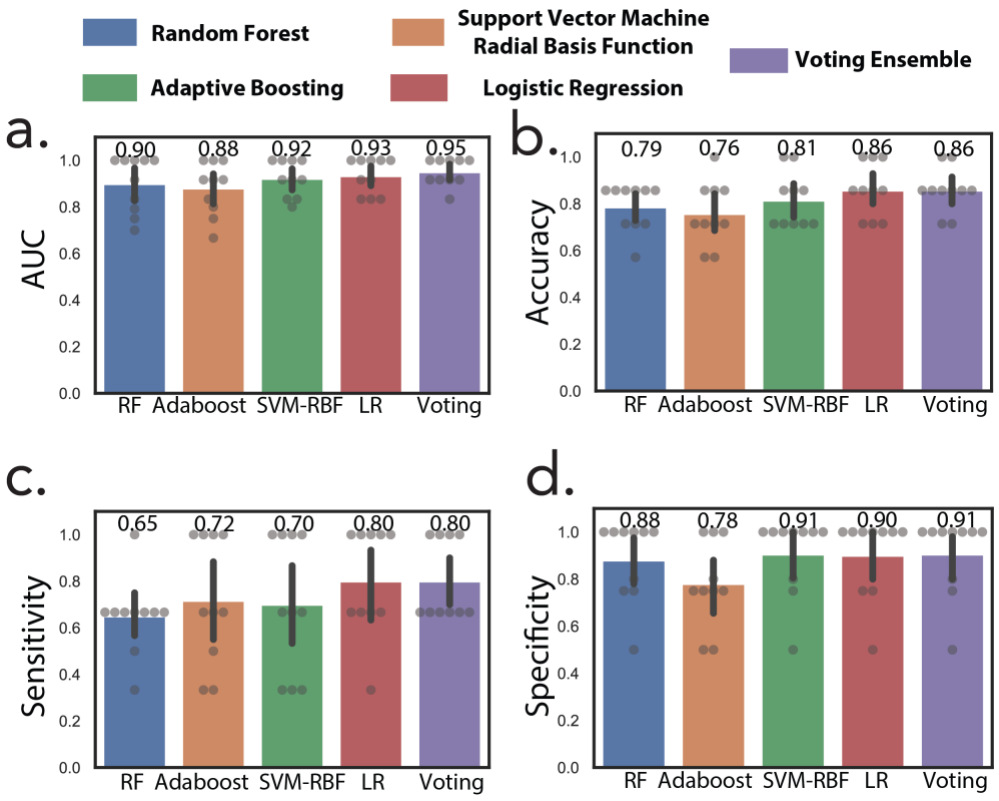


Figure B-4. Machine Learning Predictions for the RCC Stage Stratification using the 16-Metabolic Panel.

Table B-1. NMR Metabolomic Features

Metabolite /Features	^{1}H (ppm)	^{13}C (ppm)	Peak patterns	Confidence Score	Fold Change	*p*-value
unknown 1	0.15	-	(s)	-	0.13	0.604
unknown 2	0.36	-	(m)	-	-0.1	0.483
**bile acid 1	0.53	-	(s)	1	-0.11	0.473
**bile acid 2	0.56	-	(s)	1	0.04	0.78
3-hydroxyisovaleric acid	1.26	30.84	(s)	3	-0.17	0.213
lactate	1.31	22.97	(d)	4	0.18	0.426
unknown 3	1.85	-	(s)	-	0.63	0.277
acetate	1.90	26.04	(s)	3	-0.22	0.456
acetone	2.23	32.40	(s)	3	0.49	0.029

unknown 4	2.26	-	(s)	-	-0.16	0.202
acetoacetate	2.27	32.19	(s)	3	0.96	0.056
unknown 5	2.33	-	(s)	-	-0.03	0.842
¨pyruvate	2.41	-	(s)	2	0.31	0.028
citrate	2.53	48.52	(d)	3	-0.54	0.003
dimethylamine (DMA)	2.71	37.5	(s)	3	0	0.989
unknown 6	2.77	-	(s)	-	-0.2	0.141
methylguanidine	2.82	30.21	(s)	3	-0.9	0.194
unknown 7	3.08	-	(t)	-	-0.03	0.826
choline	3.19	56.69	(s)	3	0.22	0.026
[a]scyllo-inositol	3.35	76.4	(s)	3	0.05	0.726
taurine	3.42	38.07	(t)	4	0.28	0.081
acetoacetate	3.44	56.22	(s)	3	0.6	0.059
4-hydroxyphenylacetate (4-HPA)	3.44	46.34	(s)	4	0.6	0.059
glycine	3.56	44.18	(s)	3	-0.66	0.032
mannitol	3.86	65.94	(d)	4	0.07	0.74
mannitol	3.88	65.94	(d)	4	0.05	0.812
creatine	3.92	-	(s)	3	-0.12	0.644
[a]glycolate	3.94	64.32	(s)	3	-0.17	0.105
hippurate	3.96	46.46	(d)	4	-0.23	0.245
4-hydroxyhippuric acid	3.96	46.58	(d)	3	-0.23	0.245
tartrate	4.34	76.55	(s)	3	-0.15	0.507
unknown 8	6.07	-	(s)	-	-0.06	0.424
unknown 9	6.18	-	(s)	-	-0.27	0.404
fumarate	6.52	-	(s)	2	-0.16	0.256
4-hydroxyphenylacetate (4-HPA)	7.13	133.15	(d)	4	-0.14	0.85
hippurate	7.55	131.50	(t)	4	-0.37	0.275
hippurate	7.65	134.92	(m)	4	-0.35	0.276
[a]4-aminohippuric acid	7.67	133.02	(d)	3	0.09	0.472
indoxyl sulfate (IS)	7.70	127.07	(d)	3	-0.25	0.245
hippurate	7.83	129.85	(dd)	4	-0.33	0.335
hypoxanthine	8.18	148.27	(s)	3	-0.63	0.195
hypoxanthine	8.20	144.75	(s)	3	-0.1	0.849
formate	8.45	173.71	(s)	3	0.12	0.528
unknown 10	8.77	-	(d)	-	0.61	0.212
trigonelline	8.83	147.46	(t)	3	-0.14	0.748
trigonellinamide	8.89	-	(d)	2	-0.21	0.118
trigonellinamide	8.97	-	(d)	2	-0.2	0.15

trigonelline	9.11	148.50	(s)	3	-0.09	0.849
trigonellinamide	9.27	-	(s)	2	-0.23	0.165
unknown 11	9.36	-	(s)	-	-0.05	0.819

[a]Quantification may be unreliable because of spectral overlaps. Tentative assignment (Monteiro *et al* 2016) s=singlet, d=doublet, dd=doublet of doublet, m=multiplet. Fold change (FC) was calculated as the base 2 logarithm of the mean integral ratios between advanced RCC and early RCC samples. Positive FC values indicate increased abundance in advanced RCC, while negative values indicate higher abundance in early RCC. *p*-values were calculated using the Student *T*-test, while *q*-values were computed by taking the FDR correction (Benjamini-Hochberg) after a Student *T*-test. Confidence score: (1) putatively characterized compound classes or annotated compounds, (2) matches from 1D NMR to literature and/or 1D BBiorefcode compound (AssureNMR) or other database libraries such as Biological Magnetic Resonance Bank (BMRB) and Human Metabolome Database (HMDB), (3) matched to Heteronuclear Single Quantum Coherence (HSQC), (4) matched to HSQC-TOCSY.

Table B-2. RCC Patient Cohort Characteristics for the 82 Subjects used for Tumor Size Predictions.

Characteristic	Number
No of Urine Samples	82
Mean Age ± SD	60.9 ± 13.1
Mean BMI± SD	29.3 ± 5.7
Race	
Caucasian	56
Black/African American	20
American-Indian/Alaskan-Native	1
Asian	1
Mixed	1
Unknown/Missing	3

Smoker	
Never	51
Former/Current	31
Gender	
Male	45
Female	36
Not Reported	1
Histological Subtypes	
Pure Clear Cell	57
Papillary	10
Clear Cell Papillary	6
Chromophobe	5
Unclassified	4
Nuclear Grade	
1	-
2	30
3	29
4	19
Unclassified	4
RCC Stage	
I	33
II	15
III	14
IV	8
Unclassified	12

Table B-3. MS Metabolomic Features used in RCC Stage Stratification with *p*-values < 0.05 and > 1-Fold Change.

ID	Metabolite ID	Formula	Mode	RT [min]	FC	T-test *p*-value
50	Betaine	C5 H11 N O2	positive	3.784	-1.04	0.02
227	O-desmethyltramadol	C15 H23 N O2	positive	3.393	-4.94	0.08
248	248		positive	5.127	2.31	< 0.001
368	oxybenzone	C14 H12 O3	positive	1.483	-2.85	0.078
628	capuride	C9 H18 N2 O2	positive	1.66	-1.74	0.039
643	643	C11 H24 N4 O3 P2 S2	positive	2.869	1.08	0.005
776	776	C8 H24 N7 O8 P	positive	1.733	1.12	0.006
782	782	C17 H31 Br N2 S	positive	1.753	1.14	0.004
877	877		positive	3.671	-5.99	0.069
919	919	C13 H30 N3 O6 P	positive	2.993	1.13	0.017

963	5-hydroxy-4-oxo-10-propyl-6,7,8,9-tetrahydro-4H-benzo[g]chromene-2-carboxylic acid	C17 H18 O5	positive	3.023	-1.02	0.058
1077	(R)-1-aminopropan-2-ol	C3 H9 N O	positive	4.526	1.51	0.026
1125	1125		positive	3.794	-1.16	0.024
1168	N-acetylneuraminic acid	C11 H19 N O9	positive	3.384	1.07	0.001
1202	1202	C20 H33 N O10	positive	0.898	1.17	0.028
1279	N,N'-1,6-hexanediyldiacetamide	C10 H20 N2 O2	positive	1.392	-1.91	0.02
1372	4-guanidinobutanoate	C5 H11 N3 O2	positive	3.941	-1.12	0.004
1536	1536	C9 H21 N4 O5 P	positive	1.171	-1.81	0.015
1542	5-acetylamino-6-formylamino-3-methyluracil	C8 H10 N4 O4	positive	1.749	1.36	0.004
1543	1543		positive	1.751	1.15	0.014
1635	1635		positive	4.289	1.1	0.004
1673	1673	C15 H18 N3 O5 P S	positive	0.804	1.15	0.017
1689	1689	C5 H5 N2 O P3	positive	4.715	-1.36	0.042
1718	1718	C18 H29 N4 O9 P	positive	1.696	1.74	0.007
1723	1723	C6 H14 N6 S	positive	2.574	1.07	0.018
1746	1746	C18 H22 N2 O8	positive	2.607	1.77	0.016
1805	1805	C9 H28 N9 O3 P	positive	3.317	5.02	0.003
1904	7-aminomethyl-7-carbaguanine	C7 H9 N5 O	positive	4.004	1.38	0.001
1918	1918	C17 H33 O13 P	positive	1.411	2.06	0.012
1985	1985		positive	3.312	-3.99	0.086
2009	chlortoluron	C10 H13 Cl N2 O	positive	0.929	1.26	0.013
2069	2069	C21 H34 N3 O7 P	positive	3.249	1.14	0.009
2085	2085	C11 H16 N4 O5 S	positive	1.77	-1.07	0.049
2113	2113	C12 H28 N5 O6 P	positive	0.903	1.18	0.027
2122	Nalpha_Nalpha-dimethyl-L-histidine	C8 H13 N3 O2	positive	1.209	1.12	0.001
2176	5-methyldeoxycytidine	C10 H15 N3 O4	positive	0.761	1.14	0.013
2178	1-isothiocyanatobutane	C5 H9 N S	positive	1.268	-2.01	0.043
2230	2230		positive	2.59	1.09	0.016
2242	2242		positive	2.587	1.07	0.016
2281	2281		positive	1.302	-1.44	0.054
2291	2-amino-4-oxo-1,4,5,6,7,8-hexahydro-6-pteridinecarboxylic acid	C7 H9 N5 O3	positive	1.387	1.01	0.013
2313	2313	C10 H24 N O5 P3	positive	3.342	1.04	0.002
2317	diethyl2-methyl-3-oxosuccinate	C9 H14 O5	positive	0.892	1.51	0.019
2329	gabapentin	C9 H17 N O2	positive	1.078	-1.48	0.05
2339	2339	C4 H9 N6 O3 P	positive	2.073	-1.21	0.071

2377	2377	C25 H36 O15	positive	1.722	-1.07	0.049
2381	coumarin	C9 H6 O2	positive	1.184	-2.21	0.061
2417	2417	C11 H21 N2 O12 P3 S2	positive	4.637	-1.17	0.053
2440	2440	C10 H27 N8 O4 P	positive	3.015	1.13	0.011
2465	3-hydroxyanthranilic acid	C7 H7 N O3	positive	0.893	1.41	0.005
2532	2532	C9 H24 N5 O P	positive	1.127	-5.18	0.081
2553	2553	C14 H29 N2 O3 P	positive	3.455	-1.14	0.064
2558	2558	C9 H18 O P2 S	positive	3.255	1.12	0.006
2583	2583	C10 H15 N8 P	positive	1.159	-2.06	0.013
2601	2601	C4 H10 N O P	positive	2.658	1.05	0.028
2618	cyclamic acid	C6 H13 N O3 S	positive	4.026	1.43	0.006
2663	2663	C4 H13 N3 O4 P2	positive	2.667	1.49	0.013
2738	2738	C13 H27 N2 O3 P S	positive	2.209	-1.11	0.072
2817	2817	C9 H18 N5 O2 P	positive	0.59	1.77	0.022
2877	2877	C23 H25 N8 P S	positive	1.613	1.52	0.012
2905	2905		positive	2.579	1.24	0.006
2932	2932		positive	2.649	1.24	0.009
2934	2934		positive	3.808	-1.3	0.001
3001	3001	C17 H40 N2 O10 P2	positive	0.893	1.39	0.018
3093	3093	C9 H9 N3 O5	positive	3.208	1.09	0.005
3109	clavulanic acid	C8 H9 N O5	positive	1.103	1.48	0.004
3149	3149	C18 H36 N5 O9 P S2	positive	1.483	-1.78	0.049
3163	3163	C5 H15 N10 O2 P	positive	3.53	1.53	< 0.001
3191	3191	C6 H6 Cl2 N8 O17 P2	positive	4.791	-1.96	0.07
3193	3193	C46 H78 N3 O3 P3 S	positive	0.797	1.75	0.017
3257	3257		positive	3.675	-1.94	0.029
3262	3262	C10 H14 Cl N2 O16 P3 S	positive	4.751	-1.88	0.014
3297	3297	C11 H6 N8 O6 P2 S2	positive	4.706	-1.28	0.056
3306	3306	C12 H26 N7 O10 P	positive	2.694	1.3	0.011
3370	3370	C8 H13 N8 O11 P3 S2	positive	4.703	-1.23	0.062
3405	3405	C8 H13 N5 O5	positive	1.132	-1.29	0.016
3498	3498	C12 H16 N2 O2 P2	positive	0.762	-1.26	0.07
3574	3574	C6 H14 N7 O2 P S	positive	0.721	-1.34	0.048
3597	3597		positive	3.235	1.61	0.014
3602	4-ethylguaiacol	C9 H12 O2	positive	0.886	1.01	0.022
3719	3719	C18 H31 N7 O2 P2 S	positive	2.498	1.04	0.022
3746	momilactoneA	C20 H26 O3	positive	0.915	1.13	0.01
3766	apo-[3-methylcrotonoyl-CoA:carbon-dioxide ligase (ADP-forming)]	C7 H15 N3 O2	positive	3.633	1.04	0.001
3857	3857	C11 H22 N7 O6 P	positive	1.855	1.07	0.014

3934	8-azabicyclo[3.2.1]octan-3-ol	C7 H13 N O	positive	4.194	1.45	0.023
3943	2-amino-6-[(2,5-dihydroxy-2-oxido-1,3,2-dioxaphosphinan-4-yl)(hydroxy)methyl]-4a,5,8,8a-tetrahydro-4(3H)-pteridinone	C10 H16 N5 O7 P	positive	0.91	1.57	0.008
4050	4050	C4 H11 N8 O8 P3 S3	positive	4.72	-1.18	0.061
4090	4090	C5 H14 N5 O4 P3	positive	1.138	-1.8	0.035
4097	4097		positive	3.107	1.55	0.018
4116	4116		positive	3.799	-1.24	0.001
4120	4120	C14 H30 N6 O2 P2	positive	1.041	-1.19	0.076
4127	Dyphylline	C10 H14 N4 O4	positive	3.812	-1.49	0.064
4145	N_N-dihydroxy-L-tyrosine	C9 H11 N O5	positive	3.529	-2.45	0.007
4259	4259	C17 H27 N6 O16 P3 S4	positive	4.622	-1.11	0.01
4287	4287	C35 H54 N8 O18	positive	3.594	1.14	0.02
4291	4291		positive	4.027	1.29	0.006
4391	4391	C H N9	positive	3.806	-1.49	0.002
4393	4393	C8 H19 N6 O4 P	positive	3.243	1.53	0.012
4460	4460	C13 H25 N7 O5 P2 S	positive	3.314	-2.92	0.018
4467	Glycine	C2 H5 N O2	positive	3.787	-1.02	0.041
4569	4569	C5 H15 N2 O3 P	positive	4.719	-1.13	0.06
4702	4-[(2-cyclohex-1-enylethyl)amino]-4-oxobut-2-enoic acid	C12 H17 N O3	negative	2.567	-1.02	0.062
4836	4836		negative	4.022	-1.07	0.069
4902	4902	C10 H19 N8 O5 P S	negative	0.894	-1.48	0.068
4938	4938	C6 H10 N9 P S	negative	0.934	1.36	0.005
4947	2-hydroxymethylserine	C4 H9 N O4	negative	2.368	-1.22	0.02
4948	4948	C5 H10 N8 O3	negative	0.692	1.17	0.014
4992	4992	C47 H68 N10 O14 P2	negative	3.7	1.61	0.01
5045	5045	C7 H9 N O5 S	negative	3.496	1.03	0.002
5065	5065	C11 H15 Cl N5 O7 P S	negative	0.855	1.58	< 0.001
5087	5087	C3 H10 N3 O6 P	negative	3.824	-1.78	0.028
5127	5127		negative	0.939	1.32	0.008
5192	5192	C12 H20 N2 O5	negative	3.256	1.36	0.007
5206	5206	C36 H65 N2 O16 P3	negative	3.608	1.15	0.022
5226	5226	C4 H5 N3 O5	negative	0.931	1.17	0.023
5249	5249	C10 H13 N5 O5 S2	negative	3.525	1.51	0.002
5255	5255	C7 H7 N5 O5	negative	3.202	-1.9	0.005
5379	5379	C12 H14 N O3 P S	negative	0.863	1.15	0.001
5406	5406	C12 H24 O6 P2 S	negative	3.191	1.43	0.026

5408	1-(beta-D-ribofuranosyl)-1,2-dihydropyrimidine	C9 H14 N2 O4	negative	3.076	1.17	< 0.001
5409	5409	C17 H34 Cl2 N5 O5 P S3	negative	0.867	1.28	0.003
5417	5417	C4 H11 O P3	negative	3.362	-1.11	0.047
5420	5420	C4 H12 N6 P2	negative	3.38	1.73	0.003
5437	5437		negative	0.764	2.18	< 0.001
5448	5448	C11 H21 N O13 S	negative	3.497	1.09	0.001
5481	5481		negative	3.515	1.34	0.002
5482	5482	C23 H29 N O9 P2 S2	negative	0.858	1.7	0.001
5485	5485	C13 H22 N2 O5 S	negative	3.497	-1.3	0.047
5511	5511	C6 H16 N5 P3	negative	3.507	1.33	0.005
5518	(1R_2S)-1-hydroxypropane-1_2_3-tricarboxylate	C6 H8 O7	negative	3.819	-2.08	0.034
5546	5546		negative	3.828	-2.13	0.033
5580	5580	C8 H12 N6 O P2 S2	negative	3.386	-2.83	0.067
5626	5626		negative	3.433	1.13	0.009
5636	5636	C6 H20 N6 O9 P2	negative	0.946	-1.66	0.045
5666	5666	C7 H12 N3 P3 S2	negative	0.848	-3.21	0.072
5680	5680	C9 H20 N3 O10 P	negative	3.981	1.06	0.012
5713	5713	C11 H18 N2 O8	negative	1.236	1.02	0.016
5729	5729	C6 H21 N6 O18 P S	negative	3.824	-1.7	0.017
5737	gly-Lys	C8 H17 N3 O3	negative	3.995	1.14	0.001
5785	5785	C15 H34 N6 O7 P2	negative	3.42	1.61	0.013
5813	5813	C9 H8 N O2 P	negative	0.927	1.17	0.008
5825	5825	C4 H9 N7 O6 P2	negative	3.83	-1.49	0.019
5871	5871	C6 H10 N9 O13 P3	negative	3.835	-1.28	0.027
5876	5876	C6 H6 Cl N5	negative	0.642	1.24	0.004
5898	5898	C13 H27 N8 O P3	negative	4.243	1.52	0.005
5912	5912	C11 H17 N O5	negative	3.464	1.21	0.008
5985	succinic anhydride		negative	0.944	1.53	0.006
6069	6069		negative	2.05	-1.24	0.019
6089	tribenuron-methyl [ANSI]	C15 H17 N5 O6 S	negative	0.862	-1.83	0.063
6124	6124		negative	0.931	1.07	0.019
6256	metsulfuron-methyl	C14 H15 N5 O6 S	negative	0.889	-1.77	0.034
6337	6337	C10 H12 N9 O3 P S	negative	0.862	1.34	< 0.001
6351	6351		negative	0.934	1.12	0.02
6396	6396	C11 H18 N3 O7 P S	negative	0.884	1.3	< 0.001
6428	6428		negative	3.812	-1.2	0.011
6458	6458		negative	0.93	1.06	0.028
6466	ecgoninemethylester	C10 H17 N O3	negative	2.657	-1.06	0.077
6683	6683	C11 H9 N O S	negative	3.025	-1.1	0.027

6687	6687	C6 H14 N10 O5 S2	negative	0.866	1.33	< 0.001
6694	6694	C2 H5 N9 O16	negative	3.827	-1.3	0.019
6719	6719	C10 H17 N7 P2 S	negative	2.713	1.44	0.025
6731	6731		negative	0.859	1.1	0.004
6738	6738		negative	0.926	1.17	0.007
6787	6787		negative	0.939	1.07	0.01
6835	6835	C9 H15 N O5 S	negative	3.522	1.08	0.004
6912	6912		negative	3.817	-1.63	< 0.001
6933	6933	C4 H2 N8 O2	negative	3.794	-1.17	0.038
6952	6952	C15 H26 N9 O3 P	negative	3.694	-5.43	0.073
6980	6980	C10 H21 N8 O6 P S	negative	0.884	-1.54	0.057
6994	6994	C22 H33 N O9	negative	3.668	-7.01	0.073
7001	7001	C11 H15 N7 O2 P2 S	negative	0.869	1.35	< 0.001

BIOGRAPHICAL SKETCH

Olatomiwa Bifarin was born and raised in Nigeria, where he got his primary, secondary, and college education. He graduated with a degree in Microbiology at Obafemi Awolowo University in 2012. Afterward, he taught biology in a high school in Nigeria. Then, he came to the United States for his master's degree in Biotechnology at the Catholic University of America (CUA), where he graduated in 2015. At CUA, he worked in the Choy Lab to investigate the metabolic signals with DNA damage response in Saccharomyces cerevisiae. In 2015 he joined the University of Georgia ILS program and proceeded to join the Biochemistry Ph.D. program and the Edison Lab in 2016. At the Edison Lab, he studied xenobiotic metabolism in *C. elegans* and used metabolomics and machine learning to detect renal cell carcinoma in urine. In the future, he hopes to continue to work at the intersection of biomedicine and machine learning.