# Supporting Information

## Untargeted Urine Metabolite Profiling by Mass Spectrometry Aided by Multivariate Statistical Analysis to Predict Prostate Cancer Treatment Outcome

Yiwei Ma, [a] Zhaoyu Zheng,[b] Sihang Xu,[b] Athula Attygalle,[b] Isaac Yi Kim,[c] Henry Du [a]

[a] Department of Chemical Engineering and Materials Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA

[b] Department of Chemistry and Chemical Biology, Stevens Institute of Technology, Hoboken, NJ 07030, USA

[c] Section of Urologic Oncology, Rutgers Cancer Institute of New Jersey and Division of Urology, Rutgers Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA

**Table of content**

## Control and optimization of LDTD-APCI parameters

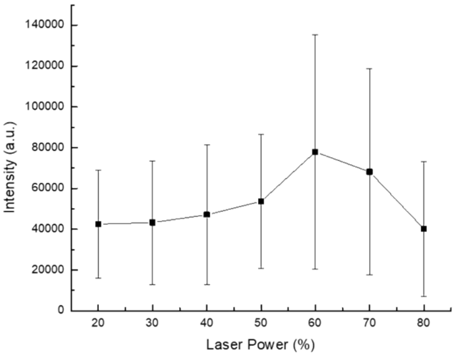


**Figure S1.** Chronogram signal intensities recorded by increasing laser power in PI (positive ionization) mode. Maximum intensity for five different urine samples is observed at laser power 60%. Vertical bars present standard deviation of signal intensity.

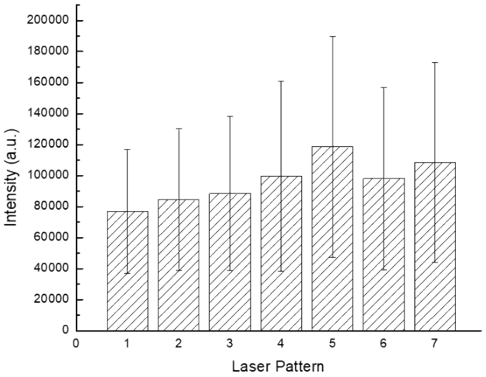


**Figure S2.** Chronogram Signal intensities obtained from five urine samples under seven different laser irradiation profiles described in Table S1 in PI mode. Maximum intensity for five different urine samples is observed at pattern 5. Vertical bards represent standard deviation.

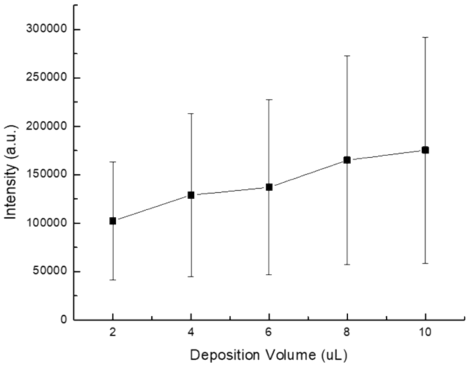


**Figure S3.** Chronogram signal intensities in different urine sample volume in PI mode. Maximum intensity for five different urine samples is observed at 10 µL (2, 4, ,6, 8, 10 µL). Vertical bards represent standard deviation.

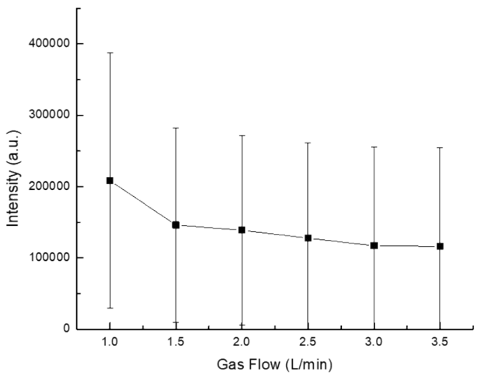


**Figure S4.** Chronogram signal intensities in different carrier gas flow rate in PI mode. Carrier gas flow rate is optimal between 1.0 and 3.5 L/min and gives significantly highest intensity response at 1.0 L/min in PI mode.

## Metabolic fingerprinting of urine

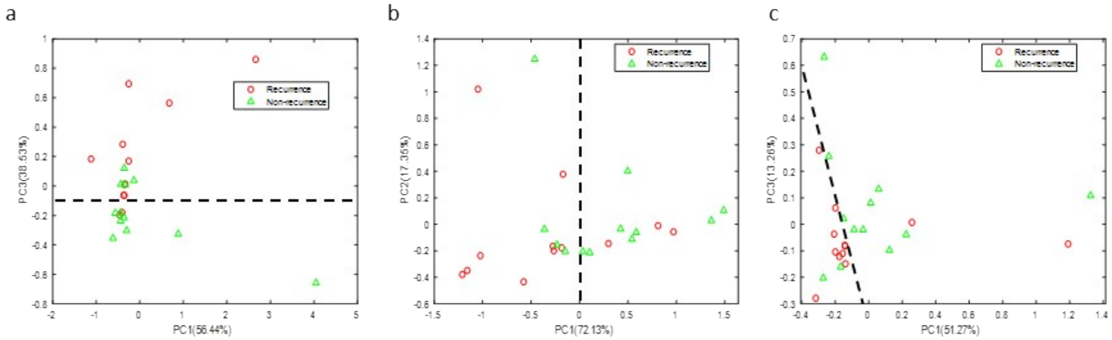


**Figure S5.** Scattering plots of posterior probabilities of PCA for (a) $H_2O$-processed urine MS dataset, two samples in Re cohort were misclassified in NRe cohort, three samples in NRe cohort were misclassified in Re cohort. (b) MeOH-processed urine MS dataset, three samples in Re cohort were misclassified in NRe cohort, four samples in NRe cohort were misclassified in Re cohort (c) ACN-processed urine MS dataset, two samples in Re cohort were misclassified in NRe cohort, two samples in NRe cohort were misclassified in Re cohort. Each point represents a single urine sample and is colored by its treatment outcomes. Re cohort is colored by red, NRe cohort is colored by green.

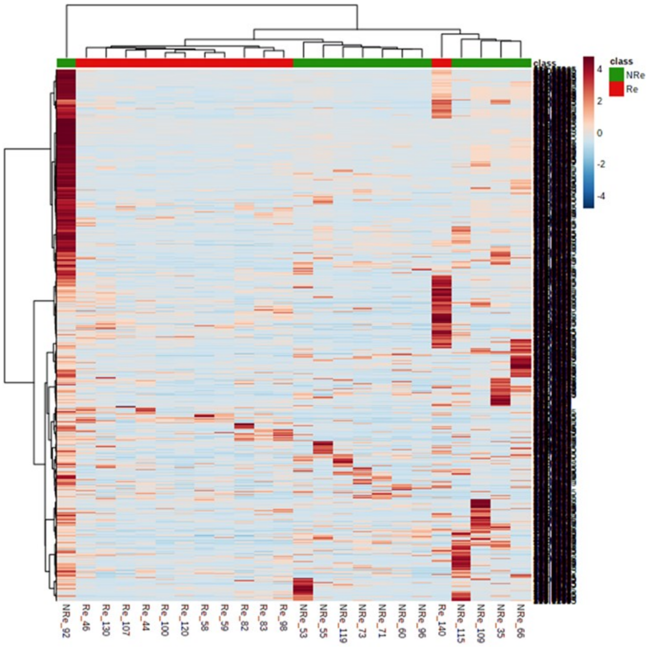


**Figure S6.** Overall hierarchical clustering heat map analysis represents the urinary global metabolome landscape of urine specimens collected before surgery between non-recurrent versus recurrent patients.  Re cohort is red, NRe cohort is green.

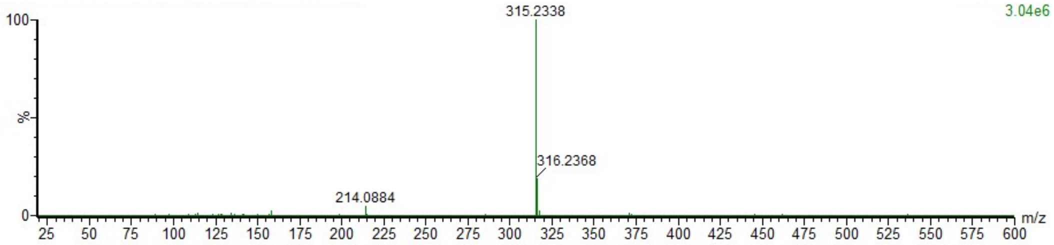


**Figure S7.** The spectrum of protonated progesterone (PROG) generated by LDTD/APCI source. The concentration of PROG solution is 10 ppm. The peak *m/z* = 214.08 is a background because of source impurity.

**Table S1.** Urine samples clinical information.

| Sample # | Gleason Score | Biochemical Recurrence | Prostate Cancer Grade |
|---|---|---|---|
| **44** | 3+3 | Recurrence | Low |
| **46** | 4+4 | Recurrence | High |
| **58** | 4+3 | Recurrence | Intermediate |
| **59** | 3+3 | Recurrence | Low |
| **107** | 4+3 | Recurrence | Intermediate |
| **82** | 3+4 | Recurrence | Intermediate |
| **83** | 4+5 | Recurrence | High |
| **98** | 4+4 | Recurrence | High |
| **100** | 4+5 | Recurrence | High |
| **120** | 4+3 | Recurrence | Intermediate |
| **131** | 4+4 | Recurrence | High |
| **140** | 4+3 | Recurrence | Intermediate |
| **60** | 3+3 | Non-recurrence | Low |
| **66** | 3+3 | Non-recurrence | Low |
| **71** | 3+4 | Non-recurrence | Intermediate |
| **73** | 3+4 | Non-recurrence | Intermediate |
| **109** | 4+3 | Non-recurrence | Intermediate |
| **119** | 4+3 | Non-recurrence | Intermediate |
| **53** | 4+3 | Non-recurrence | Intermediate |
| **96** | 3+4 | Non-recurrence | Intermediate |
| **55** | 4+4 | Non-recurrence | High |
| **92** | 4+5 | Non-recurrence | High |
| **35** | 4+5 | Non-recurrence | High |
| **115** | 3+5 | Non-recurrence | High |

**Table S2.** Seven different laser patterns are used and compared for signal intensity in PI mode

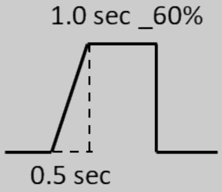
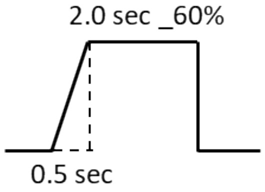
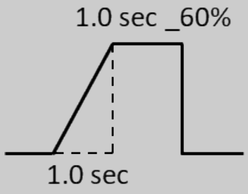
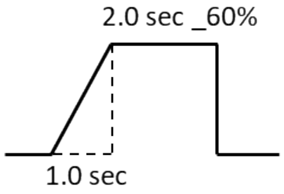
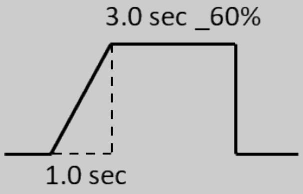
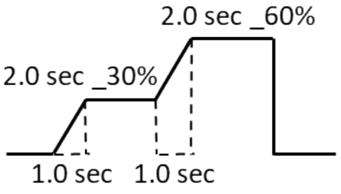
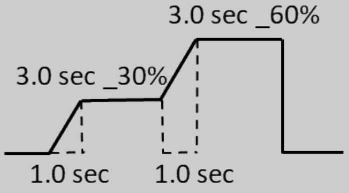
| Number | Laser Pattern |
|---|---|
| **1** |  |
| **2** |  |
| **3** |  |
| **4** |  |
| **5** |  |
| **6** |  |
| **7** |  |

**Table S3.** Standard deviation of three dominant peaks obtained by LDTD-APCI-MS for twenty-four urine samples in three different days.

| *m/z* | Standard Deviation |
|---|---|
| 114.0667 | 0.0003 |
| 313.2740 | 0.0005 |
| 341.3054 | 0.0003 |

**Table S4.** Exact masses of characteristic ions recorded from metabolites in urine (huge error is appear at uric acid, the peak at *m/z* 169.0095 is heterogenous and composited of uric acids and other metabolites).

|  | $[M+H]^+$ | Exact Mass (u) | Accurate Mass (u) | Error (ppm) |
|---|---|---|---|---|
| **Urea** | $CH_5N_2O^+$ | 61.0402 | 61.0418 | 26.21 |
| **Creatinine** | $C_4H_8N_3O^+$ | 114.0667 | 114.0668 | 0.88 |
| **Uric acid** | $C_5H_5N_4O_3^+$ | 169.0362 | 169.0995 | 374.48 |

Matlab code for PCA-LDA analysis

```
clear all;
clc;
species = readcell('AccurateMass_20210120_normalized by PROG.xlsx','Sheet','ACN_Normalized mz','Range','D2:D25');
data = readmatrix('AccurateMass_20210120_normalized by PROG.xlsx','Sheet','ACN_Normalized mz','Range','F2:VN25');
mz = readmatrix('AccurateMass_20210120_normalized by PROG.xlsx','Sheet','ACN_Normalized mz','Range','F1:VN1')';
resp=strcmp(species,'Recurrence');
[PCAcoeff,PCAscore,PCAlatent,tsquared,explained,mu] = pca(data);
figure;
PCA1 = gscatter(PCAscore(:,1),PCAscore(:,2),species,'rg','o^');xlabel('PC1');ylabel('PC2');
Y = resp;
figure;
plot(mz,PCAcoeff(:,1));
% LDA classification
sum_explained = 0;
idx = 0;
while sum_explained < 95
   idx = idx + 1;
   sum_explained = sum_explained + explained(idx);
end
idx
X = PCAscore(:,1:idx);
MdlLinear = fitcdiscr(PCAscore(:,1:idx),Y);
[~,score] = resubPredict(MdlLinear);
[x,y,t,auc] = perfcurve(resp,score(:,MdlLinear.ClassNames),'true');
figure;
plot(x,y);
legend('AUC=0.8750','Location','Best');
xlabel('False Positive Rate');ylabel('True Positive Rate');
title('ROC Curves for Recurrence Classification');
hold off;
figure;
xgrid = [1:24]';
ygrid = score(:,1);
```

```
gscatter(xgrid,ygrid,species,'rg','o^');
xlabel('Number of Spectrum');ylabel('Posterior Probability');
hold off;
figure;
plot(mz,PCAcoeff(:,1:idx));
% LOOCV;
cp = classperf(Y); %leave-one-out cross validation
for i = 1:24;
    [train,test] = crossvalind('leaveMOut',Y,1);
    mdl=fitcdiscr(X(train,:),Y(train));
    predictions = predict(mdl,X(test,:));
    classperf(cp,predictions,test);
end;
cp;
cp.CorrectRate;
```