Towards robust machine olfaction: debiasing GC-MS data enhances prostate cancer diagnosis from urine volatiles.

Adan Rotteveel[1], Wen-Yee Lee[2], Zoi Kountouri[3], Nikolas Stefanou[3], Howard Kivell[4], Clifford Gluck[3,5], Shuguang Zhang[3,6], Andreas Mershin[3,7,8]*

1 Hcyon Technology, Amsterdam, North Holland, The Netherlands
2 Department of Chemistry and Biochemistry, University of Texas El Paso, El Paso, Texas, United States of America
3 RealNose Inc., Arlington, Massachusetts, United States of America
4 Endless Frontiers Laboratory, New York University Stern School
5 Gluck Urology Clinic and Wellness Center, 72 Sharp St A10, Hingham, Massachusetts United States of America
6 MIT Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America
7 Osmocosm Foundation, Boston, Massachusetts, United States of America
8 MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Current Address: RealNose Inc. 626 Massachusetts Avenue, Arlington, MA 02476, United States of America
* mershin@mit.edu

Abstract

We present here a generalizable approach to de-biasing and de-trending before clustering complex datasets, using the specific example of prostate cancer diagnosis from gas chromatography-mass spectrometry ion chromatograms. We endeavor to mimic the olfactory cancer diagnostic prowess of trained dogs to establish, in perceptual space (as opposed to analytical chemistry space), a robustly recognizable signal that can be associated with the presence or absence of prostate cancer in the person who provided the urine sample. Upon performing principal component analysis, we found that using unsupervised clustering, the data had a strong tendency to cluster with respect to the urine sample source medical center, so a bias detector and baseline-drift remover had to be created as part of our pre-processing. Our machine olfaction approach marks a departure from conventional analytical chemistry sample analysis that leads to identification of the constituent compounds by name and concentration and moves us towards a generalizable, adaptable biomarker which in this case we call the emergent scent character. These scent characters are informed by measurements performed on volatile organic compounds (VOCs) but are not in themselves lists of VOCs or any of their physicochemical parameters added linearly. Ultimately, we use techniques prevalent in machine vision repurposed for machine olfaction by allowing for recognition and categorization of our scent character patterns as if they were images.

Introduction

Prostate cancer (PCa) is the second most frequent cancer globally in men in 2022, while being the first most frequent cancer for men in 118 countries [1]. It was the fifth leading cause of cancer death globally among men in 2022 [1], with over 397,000 deaths overall [2]. The incidence of PCa also increases with age, with an estimated 59% of men over 79 years old having suffered from PCa in 2015 [3]. PCa incidence is predicted to more than double from 1.4M per year in 2020 to 2.9M per year by 2040 [4]. Global mortality rates are also expected to surge from 375,000 in 2020 to 700,000 by 2040, which is an 85% increase [4].

Conventional prostate-specific antigen (PSA) blood tests and digital rectal examinations (DRE) are widely used at the initial detection stage of prostate cancer, when combined, the probability of prostate cancer evading detection, when the results are within normal levels, is only 10% [5]. However, both of these techniques do have drawbacks in terms of their accuracy and require a follow-up biopsy; DRE has a specificity of 59% and a sensitivity of 51% [6], while PSA is not tumor specific [7], and has low specificity at 20% [8]. This means that while these methods perform well at detecting abnormalities in the prostate, they often cannot differentiate between cancer and other conditions, such as benign prostate hyperplasia (BPH). Leading to over-diagnosis of PCa, unnecessary biopsies, and over-treatment [7]. Magnetic Resonance Imaging (MRI) has a high sensitivity of 93% and a lower specificity of 41% [9], requiring follow-up biopsies. Furthermore, in a study in 2017 called PROMIS, out of the 158 men tested negative during an MRI scan, 17 were found to have cancer in a biopsy [9].

Studies on using human body fluids such as sweat [10], saliva [11], and breath [12] to detect prostate cancer have been gathering interest. In a 2014 study by Taverna *et al.*, dogs were able to detect PCa in human urine through olfaction [13]. The researchers achieved this by training the dogs to pick up on certain volatile organic compounds (VOCs) in urine samples, which are a group of organic chemicals that liquids and solids release as gas [14]. The dogs presented high rates of sensitivity and specificity (dog 1: sensitivity 100%, specificity 98.7%. dog 2: sensitivity 98.6 %, specificity 97.6 %) [13]. In an effort to mimic the dog's diagnostic capabilities, researchers have been using Gas Chromatography-Mass Spectrometry (GC-MS) to obtain lists of urinary VOCs as potential biomarkers for cancer detection [15] [16] as well as used without first reducing to a list of molecular constituents [17]. Diagnostic trained animals such as rats, worms, bees and dogs have outperformed clinical tests and they did so without access to a list of constituent volatiles [18]. Currently, one of the key limitations of VOC-based diagnosis is the lack of concensus regarding which VOCs serve as reliable biomarkers for any condition such as prostate cancer. Various studies have identified different lists of VOCs that all show a significant correlation with cancer, yet these sets of VOCs often differ completely and do not generalize from one study to another [15, 17, 19, 20]. Canines never answer the question of "what is *in* this sample, by name of molecule and its concentration," only "what does this sample *smell of*." So olfactory perception, while clearly triggering on VOCs does not stop at analytical chemistry space but operates also in perceived scent character space. Here we focus on the concept of a biomarker as an emergent scent character, and propose a pipeline enabling direct classification on GC-MS data without having a list of molecules in the middle.

Materials and methods

Dataset Characteristics

Original data were generated in our research laboratory at the University of Texas El Paso collected for an ongoing research in PCa diagnostics [21]. This dataset included a total of 387 urine samples, and was collected from four different medical centers in the United States of America, namely, Duke University Medical Center, Durham, North Carolina; Eastern Virginia Medical Center, Norfolk, Virginia; Michael H. Annabi Internal Medicine Clinic, El Paso, Texas, and Massachusetts General Hospital, Boston, Massachusetts. GC-MS ion chromatograms were obtained but only 365 patients' samples were selected for use after we rejected data that was corrupted. The dataset includes urine samples from patients aged 44-93 who initially showed abnormal PSA results of 4.0 ng/ml or higher, or were found to have abnormalities during a DRE, and subsequently underwent a biopsy for further diagnosis. Based on the biopsy result, the dataset is split into three categories; High-Risk (Gleason scores 7 and higher), Low-Risk (Gleason group 6), and Control (prostate biopsy negative, including BPH samples) (S1 Table). For further information on the urine sample collection and the way these were processed for GC-MS analysis, please refer to S1 Appendix.

Methods

GC-MS ion chromatograms are typically presented as time-resolved amplitudes before being converted to a list of molecules. Algorithms find peaks in the data indicative of a molecule and then match its mass per charge ratio (M/Z) fingerprint and retention times against one of the various established reference libraries [22]. Noting that molecule(s) of interest might not yet be in a library or have overlapping retention times, potentially leaving compounds unidentified, is an under-appreciated problem in reducing a rich raw dataset into a highly information-reducing list of human-readable molecule names and concentrations [23].

Here we depart from the concept of biomarkers as a set of molecular IDs and their concentrations and instead generalize it to mean an arbitrary dimensional fingerprint (of peak amplitudes or receptor activations or anything that can resemble a time series), where an emergent "scent character" can be computed and assigned and whose detection improves with training. To visualize the dataset we operate on here, the raw GC-MS signals were minimally processed into a Total Ion Chromatogram (TIC) generating a one-dimensional time series of total ion intensity over time. Typical such traces from healthy and prostate cancer positive (as determined via biopsy) are shown in Fig 1A and Fig 1B respectively. To visualize the data, principal component analysis was performed on the TIC data Fig 1C.

Fig 1. Time series representing the total ion count over time. A: Raw total ion chromatograph obtained from a patient testing negative for prostate cancer via biopsy. B: Raw total ion chromatogram obtained from a patient testing a positive for Prostate cancer via biopsy, Gleason 9 (4+5). C: Principal component analysis shows the data splitting up into source clusters. D: Upon removing the bias using the Empirical Bayes method the data no longer clusters around source. E: The outliers come from patients EPPCa 524 (green), EPPCa 51 (green), BST875 (red), BR204_20 (orange), DK4676 (blue), DK2068 (blue), DK4186 (blue), DK4688 (blue) and DK3674 (blue).

An empirical Bayes method [24] was also used to counteract the source bias effects (Fig 1D). For modeling these batch bias effects, we used different prior distributions for each variable associated with batch effects. For instance, a gamma distribution was used to handle multiplicative batch effects and a normal distribution to account for mean shifts across batches.

To reduce source bias further from being part of a predictive model we integrated a bias removal pipeline in our machine learning framework [25], [26]. Signals transformed into images were used as input to a ResNet18 convolutional neural network, which then compressed them into a latent space [27]. On top of this latent space, two models were built: one aimed at diagnosing the disease and another at predicting the source hospital of the sample [28]. The first model on top of the latent space transforms the latent space such that it is minimally possible to detect source bias using gradient reversal techniques, and the second model, in parallel, transforms the latent space to diagnose the patient. To leverage established machine vision techniques, we transformed the GC-MS data into image representations suitable for input into pre-trained convolutional neural networks such as ResNet18. Two main image formats were considered, 3d heat maps and 3d peaks.

However, directly converting an entire sample into an acceptable image size (e.g., 256×256 pixels) for machine vision models would result in significant data compression and loss of critical information, including small but potentially significant peaks. A pipeline was introduced here too, to correct for noise in the model. This pipeline first applies a simple baseline correction, by fitting a polynomial function to the data and calculating the difference between the found baseline and original signal [29]. Following the baseline correction, a quantile correction is applied to remove any remaining noise.

Results

The data originally contains approximately 7300 timesteps and for each timestep there are 480 M/Z recordings. Therefore the first step is to efficiently transform the 480-dimensional time series into a simpler, 1-dimensional time series by computing the Total Ion Chromatogram (TIC). This process involves summing each timestep over all M/Z scores, thus creating a 1-dimensional time series representing the total ion count over time, as can be seen in Fig 1A and Fig 1B.

Upon conducting principal component analysis on the TIC data, it was observed that the data clustered by their source hospital, rather than prostate cancer risk level, as illustrated in Fig 1C. This type of problem is a pervasive and under-reported issue with many AI diagnostics, as some models end up using predicted socioeconomic status and other non-medical parameters that create the illusion of accuracy [30–33]. As shown in Fig 1C, it was almost always the case that data from the same source would naturally co-cluster, with only minimal outlying data points (Fig 1C). These outliers exhibited certain special characteristics, such as all outliers having African American or Hispanic heritages, while these groups represent 55% of the total dataset. Furthermore, six of these outliers have high PSA levels, with two showing exceptionally high values (above the 97th percentile). All samples had lower total ion counts than their batch counterparts. Following batch normalization, the mean ion count was 5.197 ($\pm$1.165 SD), with none of the outliers exceeding it (S2 Table).

To counter these source batch effects, an algorithm initially developed by Johnson *et al.* for microarray expression data was used [24]. This algorithm employs the empirical Bayes method, where the prior distribution is computed from the data itself, as opposed to from some set of prior assumptions. This method makes the implicit and empirically logical assumption that batch effects are not changing significantly from timestep to timestep in the GC-MS data, but instead follow a sample-wide distribution. We reduced the prominence of source-based clustering by fitting several prior distributions, such as a gamma distribution for multiplicative batch effects and a normal distribution for mean shifts across batches, and adjusting the data afterwards accordingly (Fig 1D).

Fig 2. 3D peaks of patient BST2000 A: 3D peaks of patient BST2000 shown for a full timeframe of 41.33 minutes. B: 3D peaks of patient BST2000 zoomed in on minute 16.0 until 17.0. Noise and small peaks show. C: Same timeframe as B but after applying the correction algorithms. Noise is removed and hidden peaks show up. D: An alternative visualization of patient BST2000's peak data by converting it to a heatmap

Recognizing the significant loss of information while converting the GC-MS data to one-dimensional TICs, we decided to transform the data into images instead. Converting the GC-MS data into images has benefits, because we can then use the spatial dimensions of an image effectively: one axis for representing the M/Z ratios, the second axis for representing time (minutes), and a third dimension or color to represent intensity levels in terms of ion count. An additional advantage of working in visual space is that there is substantial research done in image processing and computer vision. For example, pre-trained computer vision models such as ResNet have shown good results in scientific domains when fine-tuned appropriately [34]. Computer vision models like ResNet work best with 256×256 pixel resolution images, where each pixel contains three color channels, each encoded in 8 bits. Given that the GC-MS data originally contains approximately 3.5×10^6 integer values, the information lost results in a final compression ratio of:

$$\text{Compression Ratio} = \frac{\text{GC-MS Data Size in Bits}}{\text{Image Size in Bits}} = \frac{3.5 \times 10^6 \times 32}{256 \times 256 \times 3 \times 8} \approx 71.2$$

To ensure no important information was lost, we decided to create a machine learning pipeline capable of analyzing "zoomed in" images, as in Fig 2B.

When examining the 3D peak data image closely, numerous small peaks become apparent that are not visible in a zoomed-out view, such as in Fig 2A and in Fig 2D, an alternative heatmap representation of the same data. At the same time, significant noise patterns become evident, such as steady signals at certain M/Z values that persist for several minutes before disappearing (Fig 3B). These signals are unlikely to be true peaks and may be the cumulative result of device settings, column bleed, or other artefacts. The following pipeline was introduced to eliminate background noise prevalent in the data (Fig 2):

Fig 3. Total Ion Chromatograms before and after baseline drift correction A: Total Ion Chromatograms from patient samples EPPCa 490, EPPCa 506, and EPPCa 510 before baseline correction, captured between 25 and 25.4 minutes. B: Total Ion Chromatograms from patient samples EPPCa 490, EPPCa 506, and EPPCa 510 after applying baseline correction, captured between 25 and 25.4 minutes.

1. Rounding M/Z Values

Firstly, the raw M/Z data included non-integer values (e.g., 23.6 and 23.7). To prepare the data for subsequent processing, where each specfic M/Z ratio's noise patterns are processed differently, the presence of too many different variables due to non-integer values would lead to computational overload. Therefore, each M/Z value is rounded to the nearest integer using the following equation:

$$m = \text{round}(\text{M/Z}_{\text{raw}}) \tag{1}$$

2. Segmenting Time Frames

The data was divided into smaller time frames of $\Delta T = 1.5$, resulting into $N = 28$ segments per $T_{\text{total}} = 41$ minute sample. This segmentation allows the noise and baseline removal algorithms to be applied locally, meaning that the risk of miscategorizing a small peak as noise is significantly reduced.

$$\Delta T = \frac{T_{\text{total}}}{N} = \frac{41\,\text{minutes}}{28} \approx 1.5\,\text{minutes} \tag{2}$$

3. Quantile-Based Noise Correction

Within each time frame T_i and for each specific M/Z ratio m, we calculated the lowest 15th percentile of signal intensity. This quantile, $Q_{15\%}(T_i, m)$, represents a threshold below which the signal is considered baseline noise. It aims at only retaining values that are significantly higher than routine background fluctuations, eliminating insignificant peaks:

$$Q_{15\%}(T_i, m) = \text{Quantile}_{15\%}\left(\{I_{i,m,k}\}\right) \tag{3}$$

where $\{I_{i,m,k}\}$ denotes the set of intensity values for M/Z m within time frame T_i. The baseline noise threshold $Q_{15\%}(T_i, m)$ is then subtracted from all intensity values $I(t, m)$ within the corresponding time frame:

$$I_{\text{corr}}(t, m) = \max\left(I(t, m) - Q_{15\%}(T_i, m),\ 0\right) \tag{4}$$

For example, if $Q_{15\%}(T_i, m) = 400$ ions, then 400 is subtracted from each $I(t, m)$ in that time frame. Applying this technique conserves peaks' position relative to each other and effectively removes noise, as can be seen in (Fig 2C).

4. Polynomial Baseline Drift Correction

After quantile-based correction, there may still be residual baseline drift across the time frames. To help remove this drift, a polynomial baseline drift correction algorithm was applied from the Python Baseline Removal library [35]. This algorithm was made specifically for spectroscopy data and requires fitting a two-dimensional polynomial to the corrected intensity data $I_{\text{corr}}(t, m)$ and then subtracting this fitted baseline from the data to obtain the final processed, baseline drift free, signal [36].

To eliminate source bias from our image data, we integrated a bias removal pipeline into our machine learning framework. This pipeline was thoroughly tested, and resulted in lower cross-validated accuracy scores, of 75 percent accuracy, than the original model which was achieving a cross-validated accuracy score of 81 percent. The original model, achieving the accuracy score of 81 percent, did not include the bias removal part in its pipeline, indicating that the model is likely learning partly about the source bias and not the underlying prostate cancer signature conferred in the volatile-dependent GC-MS data.

While our pre-processing methods improved peak detection within individual batches, the identified peaks did not generalize well across different medical centers. For example, certain peak patterns at specific M/Z ratios and retention times, distinguished PCa positive from PCa negative samples coming from Boston. However, for other source hospitals, such as El Paso, these peaks were either not prevalent or failed to discriminate between PCa positive or PCa negative samples. This means a clustering "key" in the form of a list of peaks or molecules would not work as a generalizable handle to cleave a dataset into positive and negative for PCa outside of the dataset it was trained on. Our approach appeared to work initially since after a few epochs it was not possible to use the transformed data for recognizing the source hospital. However, if the bias remover is discontinued and the data are then analyzed by convolutional neural networks trained for additional epochs to specifically zoom in and detect the source, we could still identify the source hospital with high accuracy scores of 92 percent and above. It therefore appears that there are additional nonlinear traces within the GC-MS data indicative of the source hospital. We can imagine this method as a tug-of-war between the neural network removing the source data that allows for easy clustering, and the opposing neural network trying to keep it maximally predictive. This tug-of-war can be shown by defining the loss functions for the risk level as $\mathcal{L}_0(\theta)$ and source location as $\mathcal{L}_q(\theta)$. Then the total loss function of the neural network can be written as:

$$\mathcal{L}_{\text{total}}(\theta, \lambda) = \mathcal{L}_0(\theta) - \lambda \mathcal{L}_q(\theta) \tag{5}$$

With $\lambda \in [0, 1]$ as the hyperparameter balancing between the maximization of the source prediction loss $\mathcal{L}_q(\theta)$ and minimizing the risk prediction loss $\mathcal{L}_0(\theta)$, during the final minimization of the total loss $\mathcal{L}_{\text{total}}(\theta, \lambda)$. Similarly to the temperature parameter in Large Language Models (LLM), which indicates the amount of randomness the LLM is allowed to use in predicting the next word, it is possible to set the strength of the neural network removing the source data by adjusting the level of transformation of the data. This method has potential for further improvement by giving us the opportunity to adjust the amount of bias removal, such that it does not also remove the usable signal. Furthermore, including the transformation strength variable in a standard neural network hyperparameter search results in an automated process towards finding the correct value for each case [37].

Discussion

Debiasing

The original impetus behind the work presented here was to build out new synaesthetic methods to diagnose prostate cancer (PCa) from raw data of a variety of types [38]. We chose GC-MS as our first data type and pre-treated it by removing the source bias (Fig 1D) and the systemic noise and baseline drifts. Any training on untreated data would have yielded bias-contaminated results potentially giving a falsely inflated accuracy and eventually undermining confidence. We find that such biases and systematic errors are common issues in AI-driven cancer detection methods [32, 39] so we created the suite of debiasing tools presented here as a pipeline (S2 Fig) [26]. This was a necessary step before we could use the GC-MS data fully, and before we could let classification models learn directly from the data. Note that at no part of the process do we reduce these data to a list of molecules' names and their concentrations. Instead, we transform the data into images, enabling the development of what we call synaesthetic memory objects (SMOs). We realized that to continue developing diagnostic algorithms without first removing the source hospita name bias would lead to error-prone results.

To remove source bias, an empirical Bayes method [24] and a bias removal pipeline were used. Although these methods were successful in removing source bias to a large extent, when tested with advanced vision neural networks we found that we could still detect the source hospital. This revealed a deeply ingrained source bias, which may have occurred due to the way that the urine samples were collected and pre-processed at each hospital. This makes the source effect difficult to completely eliminate and this finding furthers our understanding of how source bias can be dealt especially critical for research on generalizable biomarkers from urine scent. Diagnostic AI developers must be aware of the severity of such biases potentially influencing their results, creating the illusion of a causal, clinically relevant relationship when it does not exist. It is important to carefully monitor the presence of source bias and to use robust bias removal methods to mitigate it by quantifying limitations on accuracy stemming from residual source and other non-medically relevant biases. After applying our pre-processing methods, we were able to identify discriminative peaks within individual batches, but they did not align across different medical centers. This suggests, once again, that location-specific effects are at play and that due to some systematic, handling, or processing reasons, a consistent pattern across all medical centers cannot be established and instead each medical center shows up as an accurate clustering label (Fig 1C). While manageable in ways such as those described here, we see that the prevalence of biases found in datasets used in predictive models will likely be widespread, by confounding variables such as location, ethnicity, age, or other, hard to tease out factors that a final prediction would be based on [40]. The presence of confounding factors in the training set is not a problem isolated to machine learning; canines being trained on a poorly chosen set of positive (foreground) and negative (background) sets of urine samples can learn to trigger on the wrong signal. Consider a biodetection canine that is asked to detect prostate cancer but all the negative controls are female urine. The clever dog soon starts indicating the targets, but is in fact simply choosing the "non-female urine" because that is an easier-to-detect olfactory handle, and in this case it happens to perfectly correlate with the presence of the target signal. In this thought experiment based on a real world case, even though the dog finds the target amongst the decoys well, it is not a dog trained on prostate cancer, it is a dog trained on finding male urine in a background of female urine. Once such a dog is challenged with decoys from healthy males its sensitivity and specificity can be expected to fall dramatically. Similar issues arise when AI is trained on biased and improperly controlled datasets (of any type).

Limitations

Besides the constraint of trace source bias, our study also had another limitation, which appeared due to our need to remove said bias: it is possible the signal – the scent of PCa – is lost due to over-applying our debiasing process. While we mitigated the loss of information by performing minimal corrections, there is a non-zero chance our methods can end up "cleaning away" the medically relevant signal too. Moreover, GC-MS as a laboratory technique may miss-identify or completely miss certain trace molecular species and while some machines can theoretically reach limits of detection in the low femtogram ranges under ideal conditions (e.g. such as when using octafluoronaphthalene as a benchmark in a near ideal carrier gas background), when outside those parameters, such as when exposed to actual urine samples which are more complex than any single molecule [41], sensitivity rates of as low as 3-4 parts per trillion can be achieved by advanced techniques that complement conventional GC-MS, such as GC-MS/MS [42]. For the purposes of our dataset we note that even the internal standard compound Mirex is undetectable in several samples, casting doubt on the reproducibility of trace compound identification in a complex background headspace such as urine. We note that biological olfactory receptors would likely be able to trigger on their cognate odorants when presented in similarly complex backgrounds and variable sample conditions. This is due to the fact that GPCRs are single-molecule detectors and could theoretically trigger a signaling cascade terminating in an olfactory sensory neuron's action potential being fired as a result of a single molecule being present in a sample, but only when the interaction time between sample and receptor is long enough for the xodorant to bind to the receptor's binding site.

Future research directions

Relevant to our goal of olfactory diagnosis of prostrate cancer we note that homologs of human olfactory receptors OR51E1 and OR51E2, are overexpressed in human prostate cancer tumour tissue. The homologs have been named PSGR (Prostate Specific G-Protein Coupled Receptor) and PSGR2 respectively [43]. The overexpression of PSGR and PSGR2 in cancerous prostate tissue cells in comparison to healthy prostate tissue suggests that these GPCRs play a role in signaling among the proliferating prostate cancer cells [44], [45]. The cognate ligands of these receptors potentially (soluble) odorants are likely also produced by the tumour tissue, using metabolic pathways that are likely modified from their healthy state. While many studies focus exclusively on receptor deorphanization to identify activating molecules, we speculate that molecules alone might not be the only useful layer of abstraction needed here. Since olfactory receptor GPCRs have co-evolved with the specific enzymes responsible for producing the odorants these GPCRs trigger on, we have directed our future research into the relationships between receptors and the metabolic machinery producing their targets, specifically employing machine learning models to predict enzymes related to a given GPCR. We suspect that any relationship between the metabolic machinery involved in producing an odorant and the receptor tuned to that odorant will be detectable at the scales of the primary nucleic and amino acid sequences involved in coding the metabolism and detection of its products.

Supporting information

S1 Table Each medical center's contribution of samples across different risk levels. Table notes: Each medical center has a different PCa risk distribution; for example, Virginia 1 contains only PCa positive samples and is therefore highly biased.

Overall, 10.1% of the samples come from Massachusetts General Hospital, Boston, Massachusetts, 7.7% come from Duke University Medical Center, Durham, North Carolina, 38.9% come from Michael H. Annabi Internal Medicine Clinic, El Paso, Texas, and 43.3% come from Eastern Virginia Medical Center (batch 1 is 24.4%, sample 2 is 18.9%).

S2 Table Outliers identified through principal component analysis of the raw dataset. The outliers are predominantly from African American and Hispanic heritages, although these groups only represent 55% of the total dataset. Furthermore, six of these outliers have high PSA levels, with two showing exceptionally high values (over the 97th percentile). All samples had lower total ion counts than their batch counterparts. Following batch normalization, the mean ion count was 5.197 ($\pm$1.165 SD), with none of the outliers exceeding it.

S1 Fig. Total Ion Chromatogram Before and After Mirex Normalization A) Total ion chromatogram of 15 high-risk samples from El Paso, between 29.4 and 29.8 minutes, showing the internal standard Mirex compound peaks. Variation in peak areas indicates the necessity for normalization. **B)** Total ion chromatogram of 15 high-risk samples from El Paso, between 29.4 and 29.8 minutes, after Mirex normalization to ensure sample comparability.

S2 Fig. Correction Pipeline A) Baseline Drift Correction **B)** Quantile Noise Correction **C)** Mirex Normalization **D)** Bias correction (Source Bias Removal)

S1 Appendix. Dataset characteristics. A total of 387 urine samples were initially collected from four different hospitals in the USA, namely: Duke University Medical Center, Durham, North Carolina; Eastern Virginia Medical Center, Norfolk, Virginia; Michael H. Annabi Internal Medicine Clinic, El Paso, Texas, and Massachusetts General Hospital, Boston, Massachusetts. GC-MS TICs were obtained but only 365 patients' samples were selected for use after we rejected TICs with corrupted data. The dataset contains measurements of volatile organic compounds (VOCs) emitted from the patient's urine when heated, measured by Gas chromatography/mass spectrometry coupled with thermal desorption. Each sample consists of ion intensity measurements across 480 mass-to-charge (m/z) ratios for a 41-minutes long period. This 41-minute period was measured using around 7,300 different time points, resulting in a data collection frequency of 3HZ. The dataset is further annotated into three risk categories based on the Gleason score:

- Control: Including samples from patients with Benign Prostatic Hyperplasia (BPH) and those biopsy-proven negative for prostate cancer. 125 total, of which 18 are BPH.
- Low-Risk: biopsy-proven patients with a Gleason score of 6. 133 total.
- High-Risk: Biopsy proven Includes patients with Gleason scores of 7, 8, and 9. 107 total.

Acknowledgments

We are thankful to Dr. Claire Guest of Medical Detection Dogs UK for enlightening conversations about biodetection canine training and testing, Katia Katsari of General Electric Corporation for alerting us on the importance of the problem of diagnostic AI

bias, Dr. Simmie Foster of the Harvard Laboratory for Hot and Cool Research and Massachusetts General Hospital for ideation and editing, Dr. Howard Soule of the Prostate Cancer Foundation for his support of our prostate cancer machine olfaction diagnostic effort, and the students of the MIT Sloan Executive Education Advanced Management and Global CEO Programs and other participants in the "Lab-to-Market the MIT Way" and "Making Sense of Scent" (MIT IAP) classes for early feedback on our findings. We also are grateful for their early and critical support to Nelson Wang, George Kung of KungHo Fund, and Armando Cañas, Robert Jursich of Mavericus Fund and Ruben Sanders.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A cancer journal for clinicians. 2024;74(3):229–263. doi:https://doi.org/10.3322/caac.21834.

2. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, et al.. Prostate Cancer Fact Sheet; 2024. Available from: `https://gco.iarc.who.int/media/globocan/factsheets/cancers/27-prostate-fact-sheet.pdf`.

3. Bell KJL, Del Mar C, Wright G, Dickinson J, Glasziou P. Prevalence of incidental prostate cancer: A systematic review of autopsy studies. International Journal of Cancer. 2015;137(7):1749–1757. doi:https://doi.org/10.1002/ijc.29538.

4. James ND, Tannock I, N'Dow J, Feng F, Gillessen S, Ali SA, et al. The Lancet Commission on prostate cancer: planning for the surge in cases. Lancet. 2024;403(10437). doi:https://doi.org/10.1016/s0140-6736(24)00651-2.

5. Bilal M, Javaid A, Amjad F, Youssif TA, Afzal S. An overview of prostate cancer (PCa) diagnosis: Potential role of miRNAs. Translational Oncology. 2022;26:101542. doi:https://doi.org/10.1016/j.tranon.2022.101542.

6. Naji L, Randhawa H, Sohani Z, Dennis B, Lautenbach D, Kavanagh O, et al. Digital Rectal Examination for Prostate Cancer Screening in Primary Care: A Systematic Review and Meta-Analysis. The Annals of Family Medicine. 2018;16(2):149–154. doi:https://doi.org/10.1370/afm.2205.

7. Porzycki P, Ciszkowicz E. Modern biomarkers in prostate cancer diagnosis. Central European Journal of Urology. 2020;73(3):300–306. doi:https://doi.org/10.5173/ceju.2020.0067R.

8. Merriel SWD, Pocock L, Gilbert E, Creavin S, Walter FM, Spencer A, et al. Systematic review and meta-analysis of the diagnostic accuracy of prostate-specific antigen (PSA) for the detection of prostate cancer in symptomatic patients. BMC medicine. 2022;20(1):54. doi:https://doi.org/10.1186/s12916-021-02230-y.

9. Brown LC, Ahmed HU, Faria R, Bosaily AES, Gabe R, Kaplan RS, et al. Multiparametric MRI to improve detection of prostate cancer compared with transrectal ultrasound-guided prostate biopsy alone: the PROMIS study. Health Technology Assessment (Winchester, England). 2018;22(39):1. doi:https://doi.org/10.3310/hta22390.

10. Monedeiro F, dos Reis RB, Peria FM, Sares CTG, De Martinis BS. Investigation of sweat VOC profiles in assessment of cancer biomarkers using HS-GC-MS. Journal of Breath Research. 2020;14(2):026009. doi:https://doi.org/10.1088/1752-7163/ab5b3c.

11. Luedemann C, Reinersmann JL, Klinger C, Degener S, Dreger NM, Roth S, et al. Prostate Cancer-Associated miRNAs in Saliva: First Steps to an Easily Accessible and Reliable Screening Tool. Biomolecules. 2022;12(10):1366–1366. doi:https://doi.org/10.3390/biom12101366.

12. Maiti KS, Fill EE, Strittmatter F, Volz Y, Sroka R, Apolonski A. Towards reliable diagnostics of prostate cancer via breath. Scientific Reports. 2021;11. doi:https://doi.org/10.1038/s41598-021-96845-z.

13. Taverna G, Tidu L, Grizzi F, Torri V, Mandressi A, Sardella P, et al. Olfactory System of Highly Trained Dogs Detects Prostate Cancer in Urine Samples. Journal of Urology. 2015;193(4):1382–1387. doi:https://doi.org/10.1016/j.juro.2014.09.099.

14. David E, Niculescu VC. Volatile Organic Compounds (VOCs) as Environmental Pollutants: Occurrence and Mitigation Using Nanomaterials. International Journal of Environmental Research and Public Health. 2021;18(24):13147. doi:https://doi.org/10.3390/ijerph182413147.

15. Gao Q, Su X, Annabi MH, Schreiter BR, Prince T, Ackerman A, et al. Application of Urinary Volatile Organic Compounds (VOCs) for the Diagnosis of Prostate Cancer. Clinical Genitourinary Cancer. 2019;17(3):183–190. doi:https://doi.org/10.1016/j.clgc.2019.02.003.

16. Quaye GE, Lee WY, Noriega Landa E, Badmos S, Holbrook KL, Su X. Urinary volatile organic compounds (VOCs) based prostate cancer diagnosis via high-dimensional classification. Journal of Applied Statistics. 2024; p. 1–18. doi:https://doi.org/10.1080/02664763.2024.2346355.

17. Guest C, Harris R, Sfanos KS, Shrestha E, Partin AW, Trock B, et al. Feasibility of integrating canine olfaction with chemical and microbial profiling of urine to detect lethal prostate cancer. PLOS ONE. 2021;16(2):e0245530. doi:https://doi.org/10.1371/journal.pone.0245530.

18. DeWeerdt S. Could rats and dogs detect disease better than the finest lab equipment? Nature. 2024;doi:https://doi.org/10.1038/d41586-024-01751-9.

19. Khalid T, Aggio R, White P, De Lacy Costello B, Persad R, Al-Kateb H, et al. Urinary Volatile Organic Compounds for the Detection of Prostate Cancer. PLOS ONE. 2015;10(11):e0143283. doi:https://doi.org/10.1371/journal.pone.0143283.

20. Lima AR, Araújo AM, Pinto J, Jerónimo C, Henrique R, Bastos MdL, et al. Discrimination between the human prostate normal and cancer cell exometabolome by GC-MS. Scientific Reports. 2018;8. doi:https://doi.org/10.1038/s41598-018-23847-9.

21. Badmos S, Noriega-Landa E, Holbrook KL, Quaye GE, Su X, Gao Q, et al. Urinary volatile organic compounds in prostate cancer biopsy pathologic risk stratification using logistic regression and multivariate analysis models. American Journal of Cancer Research. 2024;14(1):192–209. doi:https://doi.org/10.62347/rxsh1223.

22. National Institute of Standards and Technology. NIST17 Mass Spectral Library: Updates to the NIST Tandem and Electron Ionization Spectral Libraries; 2017. Available from: https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:nist17.

23. Fiehn O. Metabolomics by Gas Chromatography-Mass Spectrometry: the combination of targeted and untargeted profiling. Current protocols in molecular biology / edited by Frederick M Ausubel [et al]. 2016;114(1):30.4.1–30.4.32. doi:https://doi.org/10.1002/0471142727.mb3004s114.

24. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2006;8(1):118–127. doi:https://doi.org/10.1093/biostatistics/kxj037.

25. Ganin Y, Lempitsky V. Unsupervised Domain Adaptation by Backpropagation; 2015. Available from: https://arxiv.org/abs/1409.7495.

26. Rotteveel A, Mershin A, Stefanou N, Kountouri Z. Systems and methods for de-biased clustering of data to enhance prostate cancer diagnosis from urine volatiles in machine olfaction; 2024. U.S. Patent No. 302733.0003.

27. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition; 2015. Available from: https://arxiv.org/abs/1512.03385.

28. Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; p. 2962–2971. doi:10.1109/cvpr.2017.316.

29. Zhang ZM, Chen S, Liang YZ. Baseline correction using adaptive iteratively reweighted penalized least squares. The Analyst. 2010;135(5):1138–1146. doi:https://doi.org/10.1039/b922045c.

30. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. npj Digital Medicine. 2023;6(1). doi:https://doi.org/10.1038/s41746-023-00858-z.

31. Salinas MP, Sepúlveda J, Hidalgo L, Peirano D, Morel M, Uribe P, et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. npj Digital Medicine. 2024;7(1). doi:https://doi.org/10.1038/s41746-024-01103-x.

32. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. npj Digital Medicine. 2022;5(1). doi:https://doi.org/10.1038/s41746-021-00544-y.

33. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. Nature Medicine. 2024;30:2838–2848. doi:https://doi.org/10.1038/s41591-024-03113-4.

34. Spolaôr N, Lee HD, Mendes AI, Nogueira CV, Rafael A, Weber, et al. Fine-tuning pre-trained neural networks for medical image classification in small clinical datasets. Multimedia Tools and Applications. 2023;83(9):27305–27329. doi:https://doi.org/10.1007/s11042-023-16529-w.

35. Haque A. Feature Engineering Selection for Explainable models: A second course for Data scientists. LULU COM; 2022.

36. Lieber CA, Mahadevan-Jansen A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. Applied Spectroscopy. 2003;57(11):1363–1367. doi:https://doi.org/10.1366/000370203322554518.

37. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: Shawe-Taylor J, Zemel RS, Bartlett P, Pereira F, Weinberger KQ, editors. 25th Annual Conference on Neural Information Processing Systems (NIPS 2011). vol. 24 of Advances in Neural Information Processing Systems. Granada, Spain: Neural Information Processing Systems Foundation; 2011.Available from: `https://inria.hal.science/hal-00642998`.

38. Mershin A, Rotteveel A, Stefanou N. Machine olfaction-centric sensor fusion architectures for non-invasive sensing; 2024. U.S. Patent No. 302733.0002.

39. Trafton A. Study reveals why AI models that analyze medical images can be biased; 2024. Available from: `https://news.mit.edu/2024/study-reveals-why-ai-analyzed-medical-images-can-be-biased-0628`.

40. Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. Nature Medicine. 2024; p. 1–11. doi:https://doi.org/10.1038/s41591-024-03113-4.

41. Fialkov AB, Steiner U, Lehotay SJ, Amirav A. Sensitivity and noise in GC–MS: Achieving low limits of detection for difficult analytes. International Journal of Mass Spectrometry. 2007;260(1):31–48. doi:https://doi.org/10.1016/j.ijms.2006.07.002.

42. Justen PT, Kilpatrick ML, Soto JL, Richardson SD. Low Parts Per Trillion Detection of Iodinated Disinfection Byproducts in Drinking Water and Urine using Vacuum-Assisted Sorbent Extraction and GC–MS/MS. Environmental science technology. 2023;58(2). doi:https://doi.org/10.1021/acs.est.3c07097.

43. Chung C, Cho HJ, Lee C, Koo J. Odorant receptors in cancer. BMB Reports. 2022;55(2):72–80. doi:https://doi.org/10.5483/bmbrep.2022.55.2.010.

44. Weng J, Wang J, Cai Y, Stafford LJ, Mitchell D, Ittmann M, et al. Increased expression of prostate-specific G-protein-coupled receptor in human prostate intraepithelial neoplasia and prostate cancers. International Journal of Cancer. 2005;113(5):811–818. doi:https://doi.org/10.1002/ijc.20635.

45. Weng J, Wang J, Hu X, Wang F, Ittmann M, Liu M. PSGR2, a novel G-protein coupled receptor, is overexpressed in human prostate cancer. International Journal of Cancer. 2006;118(6):1471–1480. doi:https://doi.org/10.1002/ijc.21527.

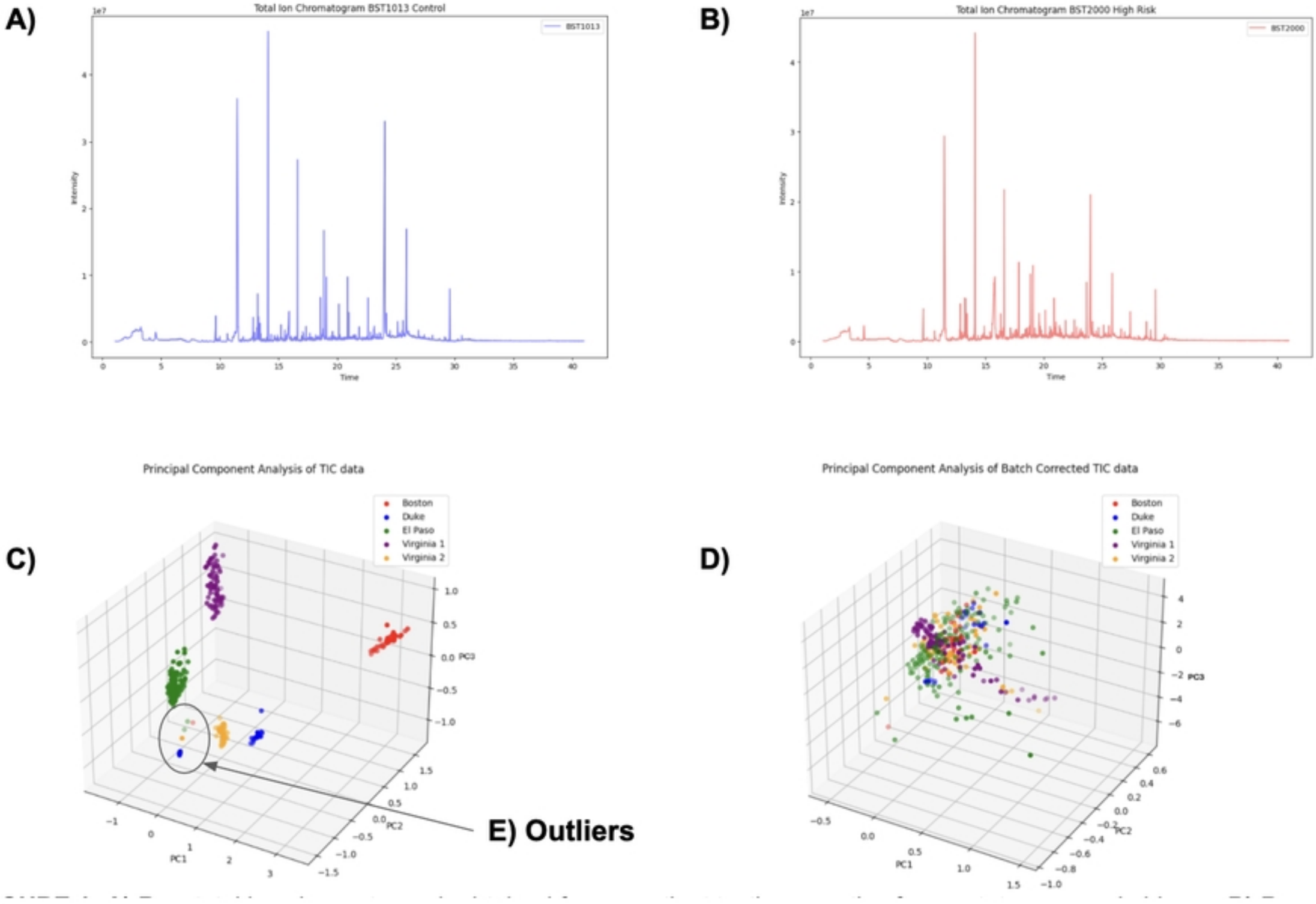
A)
Total Ion Chromatogram BST1013 Control
BST1013
Intensity
Time
B)
Total Ion Chromatogram BST2000 High Risk
BST2000
Intensity
Time
C)
Principal Component Analysis of TIC data
Boston
Duke
El Paso
Virginia 1
Virginia 2
PC1
PC2
PC3
E) Outliers
D)
Principal Component Analysis of Batch Corrected TIC data
Boston
Duke
El Paso
Virginia 1
Virginia 2
PC1
PC2
PC3

figure 1

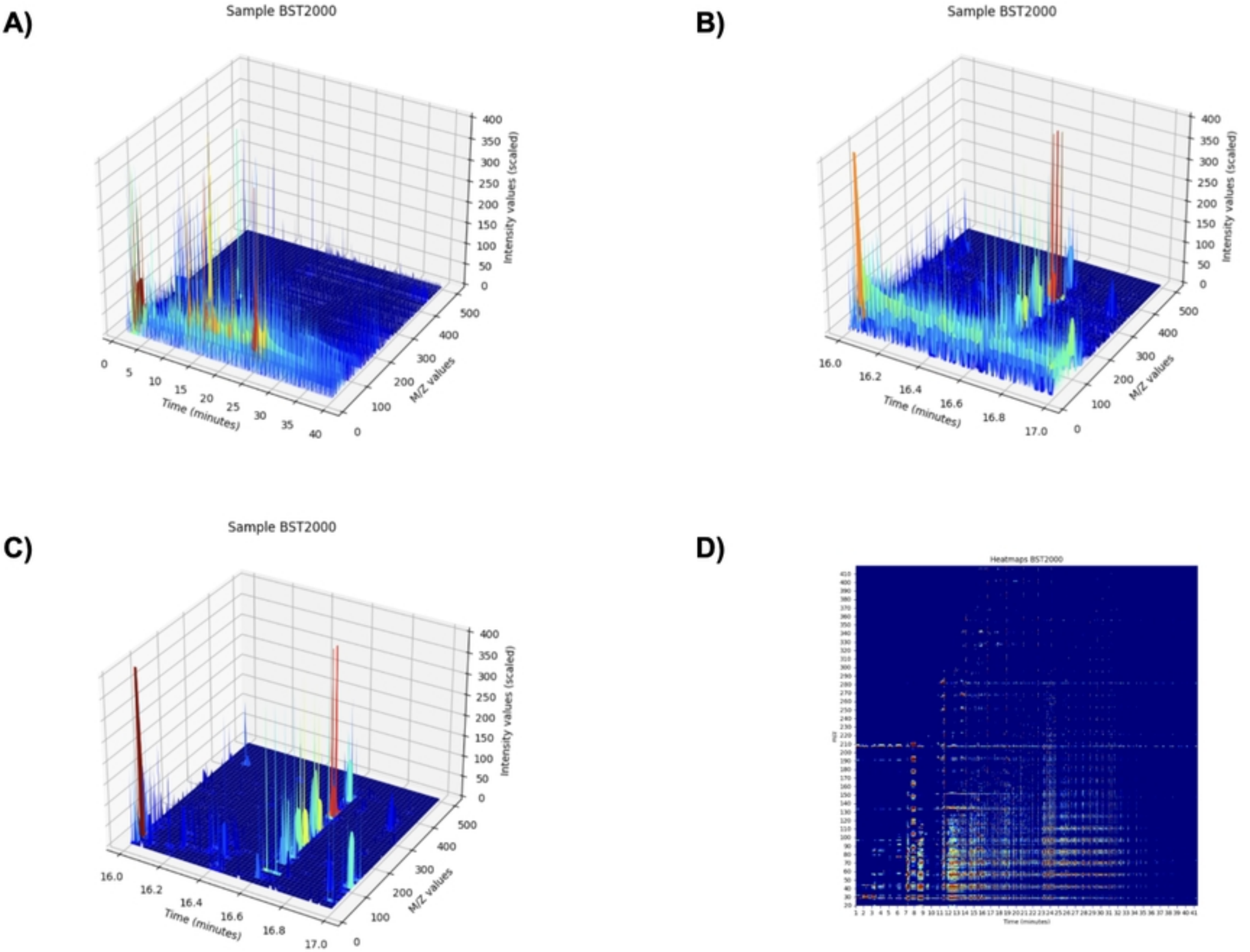
A)
Sample BST2000
Intensity values (scaled)
M/Z values
Time (minutes)
B)
Sample BST2000
Intensity values (scaled)
M/Z values
Time (minutes)
C)
Sample BST2000
Intensity values (scaled)
M/Z values
Time (minutes)
D)
Heatmaps BST2000

figure 2

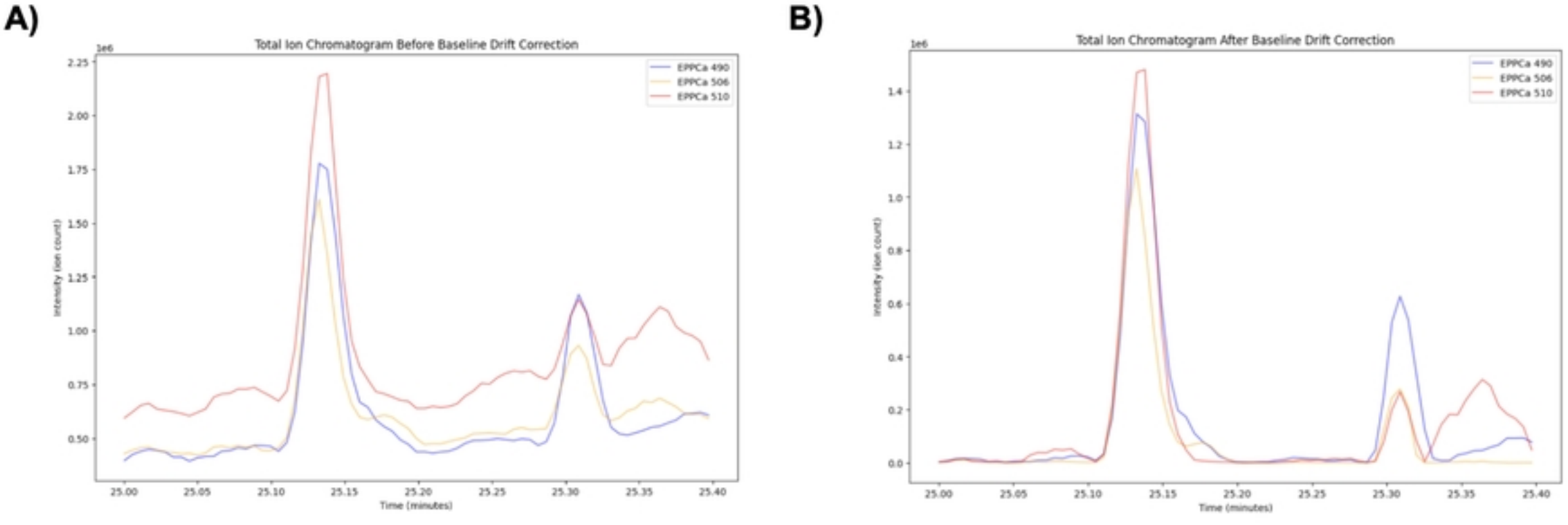
A)
Total Ion Chromatogram Before Baseline Drift Correction
1e6
Intensity (ion count)
Time (minutes)
EPPCa 490
EPPCa 506
EPPCa 510
B)
Total Ion Chromatogram After Baseline Drift Correction
1e6
Intensity (ion count)
Time (minutes)
EPPCa 490
EPPCa 506
EPPCa 510

figure 3