http://pubs.acs.org/journal/acsodf

Article

Insights into the Effects of Violating Statistical Assumptions for Dimensionality Reduction for Chemical "-omics" Data with Multiple Explanatory Variables

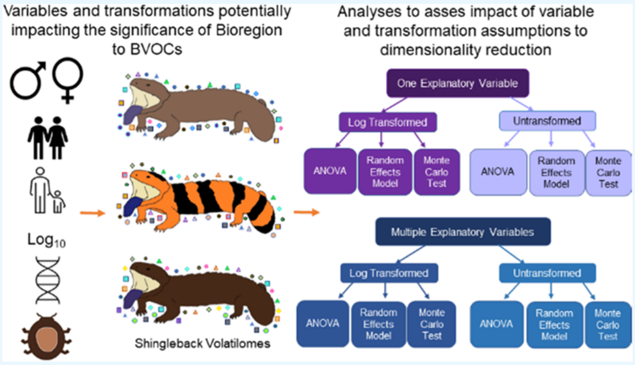
Amber O. Brown,* Peter J. Green, Greta J. Frankham, Barbara H. Stuart, and Maiken Ueland

Cite This: *ACS Omega* 2023, 8, 22042−22054

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Biological volatilome analysis is inherently complex due to the considerable number of compounds (i.e., dimensions) and differences in peak areas by orders of magnitude, between and within compounds found within datasets. Traditional volatilome analysis relies on dimensionality reduction techniques which aid in the selection of compounds that are considered relevant to respective research questions prior to further analysis. Currently, compounds of interest are identified using either supervised or unsupervised statistical methods which assume the data residuals are normally distributed and exhibit linearity. However, biological data often violate the statistical assumptions of these models related to normality and the presence of multiple explanatory variables which are innate to biological samples. In an attempt to address deviations from normality, volatilome data can be log transformed. However, whether the effects of each assessed variable are additive or multiplicative should be considered prior to transformation, as this will impact the effect of each variable on the data. If assumptions of normality and variable effects are not investigated prior to dimensionality reduction, ineffective or erroneous compound dimensionality reduction can impact downstream analyses. It is the aim of this manuscript to assess the impact of single and multivariable statistical models with and without the log transformation to volatilome dimensionality reduction prior to any supervised or unsupervised classification analysis. As a proof of concept, Shingleback lizard (*Tiliqua rugosa*) volatilomes were collected across their species distribution and from captivity and were assessed. Shingleback volatilomes are suspected to be influenced by multiple explanatory variables related to habitat (Bioregion), sex, parasite presence, total body volume, and captive status. This work determined that the exclusion of relevant multiple explanatory variables from analysis overestimates the effect of Bioregion and the identification of significant compounds. The log transformation increased the number of compounds that were identified as significant, as did analyses that assumed that residuals were normally distributed. Among the methods considered in this work, the most conservative form of dimensionality reduction was achieved through analyzing untransformed data using Monte Carlo tests with multiple explanatory variables.



1. INTRODUCTION

Volatilomics is the study of cumulative biogenic volatile organic compounds (BVOCs) which are produced by living organisms. Endogenous BVOCs originate from various metabolic, genetic, and chemical processes.[1−5] BVOCs can also be produced exogenously by organisms through their diets,[6,7] secretions,[8] or skin and gut microbiomes.[9,10] After production, BVOCs diffuse into their surrounding environments, where they can be collected and analyzed for various scientific interests such as health or physiology assessments,[11,12] species identifications,[13] and other forensic detection purposes.[14,15]

Traditionally, BVOCs have been analyzed using thermal desorption gas chromatography (GC) coupled with various forms of mass spectrometry due to their sensitivities and abilities to separate compounds.[16] One-dimensional gas chromatography has recently been replaced by two-dimensional gas chromatography (GC × GC; i.e., two-dimensional (2D)) as a separation method when analyzing biological volatilomes.[17] This is in part due to the increased ability of 2D GC to accurately detect more compounds through its enhanced separation ability and the increased peak capacity when analyzing complex samples.[17−20] The resultant data are highly dimensional with hundreds of compounds being tentatively identified per sample.[13,17]

Received: March 29, 2023
Accepted: May 25, 2023
Published: June 9, 2023

https://doi.org/10.1021/acsomega.3c01613

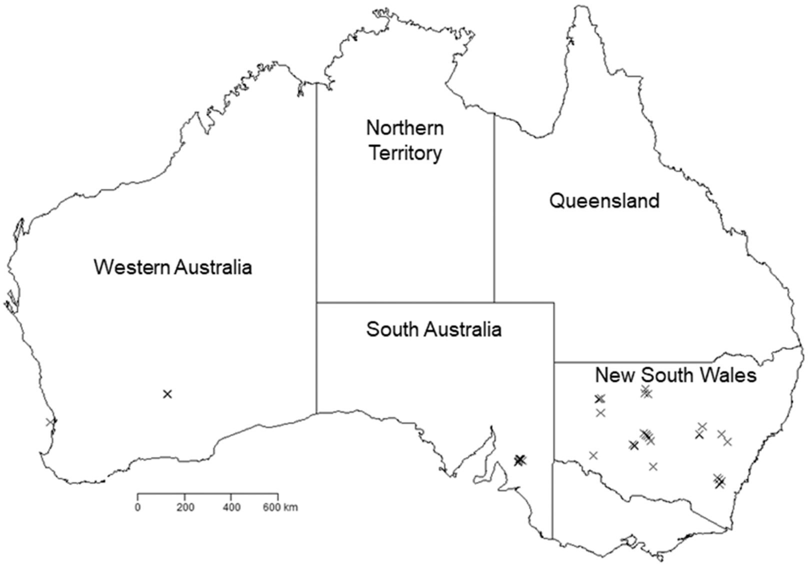


Figure 1. Locations from where wild shingleback volatilome samples were taken across mainland Australia. The GPS location of paired shinglebacks is represented by one cross.

There is interest in determining how BVOCs differ between or are impacted by different explanatory variables, such as species identity, sex, disease status, and individualization. This can be evaluated through the identification of compounds for which particular factors of interest have a significant effect. Various statistical techniques, such as compound variability assessments (e.g., Fisher ratio filtering, *t*-tests, partial least squares) or unsupervised screening techniques (e.g., principal component analysis (PCA)), are generally employed as dimensionality reduction tools that aid in the identification of significant compounds prior to further multivariate analysis.[21−24] Compound reduction can be complex for biological samples as the number of corresponding explanatory variables, which may impact compound presence or relative abundance, are increased with the expansion of the dataset.[25] Furthermore, many explanatory biological variables with differing intralevel variabilities innately exist within biological datasets. Thus, there is the potential for each of these variables to have an effect on the distribution of data, despite whether or not these variables are recorded and subsequently analyzed. At minimum, it has been determined that biological volatilomes are influenced by biological and environmental variables, including age class,[26,27] sex,[28−30] individual genetics,[28,31] and reproductive status.[30,32] Additionally, the biological variables can be correlated and their effects can combine additively, multiplicatively, or in other ways.[33]

In some instances, one-dimensional variable-level comparisons of biological samples may be appropriate for determining the effect of significance for a single variable. For example, one variable may be analyzed in scenarios where samples are taken from one individual and compared against control samples.[34] Otherwise, in order to make similar comparisons, biological studies have to be completely balanced for each variable level found within the dataset (e.g., same number of males of a certain age, same number of females of a certain disease status).[7,35] Biological observational studies often do not exhibit this balance due to various project limitations.[13,19] As such, targeting one explanatory variable for compound dimensionality reduction is typically inappropriate in studies where volatilome analysis expands across many individuals with differing biological variables. Targeting one explanatory variable may lead to erroneous compound selection due to Simpson's paradox. This paradox demonstrates that associations between two variables change when sample populations are partitioned into smaller denominations,[36] which is exhibited in numerous real-world examples.[37,38] The innate multivariate and multilevel variability and high dimensionality found within each biological volatilome dataset will likely prevent accurate dimensionality reduction through traditional uni-variate techniques. Without the careful consideration of multiple variables and their residual distributions, nontargeted "-omic" studies are at a greater risk of unintended bias or false biomarker denotations. These inadvertent consequences are generated due to inappropriate selections of statistical models, oversimplifications, and overfittings of datasets.[39]

The assumptions of the chosen statistical analysis and the dataset characteristics should be understood prior to selecting the respective method when determining the effect of explanatory variable significance to the response variables. This includes the assessment of whether the distributions of the data can be modeled parametrically (i.e., analyst assumes the data distributions follow an algebraic expression) or nonparametrically. This also includes an assessment of data transformation, as the peak area values that are compared in these types of analyses can differ by orders of magnitude both within and between compounds.[40] In order for the data modeling to adequately address the research question, the data should be analyzed using the most appropriate approach. The most widely used transformation to attempt to convert skewed data to normal or near normal distributions for biological datasets is the log transformation[41] and has been applied to volatilome data generated by GC × GC analyses.[42,43] The appropriateness of potential log transformations to data also relies on the assumptions of how explanatory variables may interact and their effect on the data on a reciprocal scale. Subject-specific knowledge will guide the analyst in determining whether or not the scale of the variable effect is additive or multiplicative. For example, in some research areas, it has been demonstrated that variables act independently of each other on an appropriate scale. In these cases, the data are best suited to be analyzed on that scale and not to be log transformed. However, in other fields, it has been demonstrated that the impacts of each explanatory variable are multiplicative on an appropriate scale. Here, the data are better suited to be log transformed. Biological variables in volatilome data are suspected to exhibit both additive and multiplicative qualities, though the scale of which is unknown. This is because some BVOCs are generated through independent pathways, while others are generated at differing stages of the same pathway.[44] Regardless, overlooking or assuming multiplicity can lead to an exaggeration of the effect of the analyzed explanatory variable.[45]

A relevant way to assess the impacts of data transformations and statistical violations in a chemical "-omics" context is through the comparison of outcomes from a variety of statistical tests with varying assumptions. Due to the complex nature of biologic volatilomes, both "fixed effects" and "random effects" models should be considered. Fixed effects models assume that all levels of a factor (e.g., "diseased" and "healthy" are levels of the variable "disease status") that are of interest are observed within the data.[46] Random effects models assume that only a random sample of possible levels is observed within the dataset.[46]

Currently, there are no formal statistical methodologies or guidelines for dimensionality reduction or for the determination of the effects of multiple explanatory variables in the field of volatilomics. Recent manuscripts have called for the validation of analytical methods selected for two-dimensional data and the identification of more appropriate multidimensional models.[23] The aims of this manuscript are to assess the impacts of data transformation and statistical analysis (i.e., single and multiple explanatory variables, fixed and random effects models) and to assess the statistical significance of biological variables to the presence and relative abundance of BVOCs. This proof of concept will use volatilome data collected from Shingleback lizards (*Tiliqua rugosa*), which are one of the most highly illegally trafficked animals in Australia.

2. MATERIALS AND METHODS

2.1. Sampling Areas. Shingleback volatilomes were collected from captive animals at Featherdale Wildlife Park (Sydney, Australia) and from wild caught animals across New South Wales (NSW), South Australia (SA), and Western Australia (WA) (Figure 1 and Table S1). Wild Shinglebacks were visually located on roadways[47] while driving (speed ~10 km/h) or by foot and were hand caught once observed. Due to the limitations associated with restricted travel and border closures imposed during the COVID-19 pandemic, this work was unable to collect representative samples from Bioregions in Victoria, and specific areas of WA and SA. GPS coordinates from sampling locations were used to determine the Bioregion of each collected sample using the SEED Central Resource for Sharing and Enabling Environmental Data in NSW.[48] In total, 11 variable levels for the explanatory variable Bioregion were created; 10 of which were defined from discrete Bioregions[48] and one which was defined for captive animals.

2.2. Morphometric Analysis. Shinglebacks were captured by hand and morphometric measurements, including snout-to-vent lengths, vent-to-tail lengths, mid-body and tail circumferences were taken in order to calculate approximate total volumes. The total volumes of Shinglebacks were converted into discrete levels with the following cut-offs ($\leq$800 000 mm^3, <935 000 mm^3, $\leq$1 095 000 mm^3, and >1 095 000 mm^3). Cloacal inspections were used to determine the sex of the Shinglebacks. When the reproductive organs could not be detected, Shinglebacks were labeled as "unknown". All Shinglebacks were visually assessed for ticks, and the presence or absence of these parasites was recorded (Table S1).

2.3. Volatilome Collection. Volatilomes were collected through optimized methods as described in Brown et al.[40] using preconditioned dual sorbent tubes (Tenax and Carbograph 5DT Markes International Ltd., UK; parts number C2-AAXX-5149).

2.4. Volatilome Analysis. Shingleback volatilome samples were desorbed with a Markes Unity 2 Thermal Desorber and Series 2 ULTRA multi-tube autosampler (Markes International Ltd., UK) with a Markes General Purpose Carbon C4/5-C30/32 cold trap (parts number U-T11GPC-2S) using optimized methods outlined in Brown et al.[40] Samples were analyzed using comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry (Pegasus 4D GC × GC-ToFMS (LECO Australia Pty Ltd., Australia)) with optimized reptile volatilome columns and separation methods detailed in (Brown et al).

2.5. Chromatogram Alignment. Chromatograms were aligned and analyzed with ChromaTOF (version 4.51.6.0; LECO) using the same parameters as Brown et al.[40] Compounds were identified with a minimum similarity match criteria of 75% in The National Institute of Standards and Technology (NIST) Mass Spectral Library. Peaks were aligned across samples from the same Bioregion using a 60% match threshold.

2.6. Internal Standard Normalization and Data Processing. Volatilome samples have compounds with different molecular weights, polarities, and elemental compositions. These differences will lead to differing ionization efficiencies per compound and will result in disparate and relative quantifiable detections by the mass spectrometer.[49,50] To address this phenomenon, untargeted chemical "-omic" studies use limited panels of compounds, or single compounds, as internal standards to assess analytical reproducibility and to reduce variability in compound signal detection.[51−53] The inclusion of an internal standard allows for semiquantitative analysis where relative signal intensities from each compound can be compared within the total profile after normalization to the peak area of the internal standard.[53] Prior to analysis, 0.2 μL solution of 10 ppm d_5-chlorobenzene (CAS number 3114-55-4; Merck, AUS) was injected into each sorbent tube to serve as an internal standard. All compound peak areas were normalized to the corresponding peak area of the internal standard. Following normalization, all data were exported to Microsoft Excel (2010) where the internal standard and solvent were removed. For each sample taken, the compounds found in the container blanks were compared with associated Shingleback replicates and removed using a 50% threshold as described in Brown et al.[40] All unidentified analytes were kept for further analysis.

2.7. Tentative Compound Classification. A thorough literature search was performed by which each identified compound name was searched to determine if other studies had detected these compounds during analysis. This was used as an indirect assessment to determine potential origins (e.g., from a biological source, including flora, fauna, meat, or decomposition products, or from an industrial source, such as from petrol, pesticides, or lab contaminants) of the identified BVOCs from Shingleback samples. Based on the citations

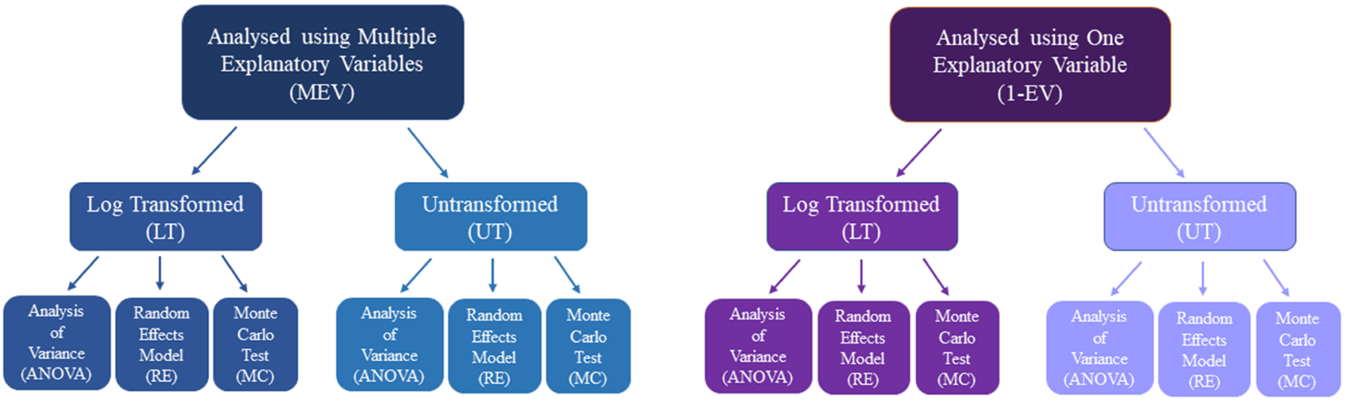

Analysed using Multiple Explanatory Variables (MEV)
Log Transformed (LT)
Untransformed (UT)
Analysis of Variance (ANOVA)
Random Effects Model (RE)
Monte Carlo Test (MC)
Analysed using One Explanatory Variable (1-EV)

Figure 2. Project plan by which Shingleback volatilome data were analyzed. The results of the 12 potential variable, transformation, and model analyses were compared.

found during the literature search, compounds were classed into four categories. The first was defined "biologically unstable" compounds in which the citations for the compound were related to pheromones or hormones, related to atmospheric compounds, or related to faces, urine, or exhaled breath. The "unnatural" category was defined as compounds of which citations were only related to petrol, plastics, or the GC columns or compounds which had originated from the sanitizing alcohols which were sprayed on the Shingleback's tails prior to blood draws. The "natural" category included compounds that had at least one citation related to living organisms or organic products. Finally, the "unknown" category was defined as compounds that had citations from numerous different origins, compounds that had no citations, or compounds that were unidentified analytes. All compounds in the "natural" and "unknown" categories were used for further analysis. This was justified, as the primary aim of the manuscript was to determine for which compounds Bioregion had a significant effect. As compounds found in the "biologically unstable" category are likely to be artifacts of the Shingleback's metabolism or response to stimulus, these compounds were excluded. All "unnatural" compounds were also excluded, as they would not naturally occur in Shingleback volatilomes. These compounds could be representative of the sampling condition of a particular day, as opposed to significant compounds related to elements of the Bioregion.

2.8. Statistical Analysis. All analyses were conducted using RStudio (version 4.1.0) using code that was written in-house (using multiple explanatory variables; Supporting Information 1, 1-EV; Supporting Information 2). In accordance with standard scientific practice, all references to statistical significance should be read as "at the 5% level", that is, to the results of tests where the p-value is ≤ 0.05 The data consist of the response variable (i.e., compound peak area on either the natural scale or log transformed) by which different explanatory factors, factorial structures, and linear models were applied. The data were represented as a matrix by which the explanatory factors were the column labels (columns 1−16; Table S1) and response variables were indexed by compound chemical names in alphabetical order (columns 17−1170; Table S1). In describing the results, the term "significant compound" is used in place of "compound for which the effect of Bioregion was statistically significant at the 5% level" for brevity. The term "not-significant compound" is used in the same manner. The dataset was analyzed using the following analyses and data combinations, found in Figure 2.

2.8.1. Log Transformations. Both log transformed and untransformed forms of the generated data were used for each statistical test using each variable combination (multiple explanatory variables and one explanatory variable; Figure 2). The peak areas from each compound were transformed using the common log ($\log_{10}$). Any peak area that was found to be below the detection limit or removed through the container blank threshold was assigned the conventional value 0 under the log transformation.

2.8.2. Multiple Linear Regression. The effects of significance of multiple explanatory variables (MEV) that are biological (i.e., sex, tick presence, genetic lineage, age class, total volume, captive versus wild status) or environmental (i.e., Bioregion) were also assessed through the use of multiple linear regressions of each of the below statistical analyses. These analyses were conducted to determine the effect of significance due to Bioregion despite the presence of other known explanatory variables.

2.8.3. Alternative Model Assumptions and Analyses. Three different linear model formulations (i.e., analysis of variance, random effects model, Monte Carlo test) with alternative data assumptions were used for analysis (Figure 2). Both log transformed and untransformed data were used for each data analysis. Throughout this analysis, the focus was to assess the statistical significance of the effect of Bioregion on compound peak area, whether other factors were also fitted or not, and whether or not the log transformation was used. Each tested statistical model with log transformed or untransformed data makes different assumptions about the data distributions.

2.8.3.1. Analysis of Variance. Analysis of variance ("ANOVA") is one of the most routine analyses for assessing the statistical significance of explanatory factors in linear models throughout biological and chemical research.[54] It assumes a linear model and involves computations based on the sum of squares of the sample deviations (in this case, compound peak areas) from their respective means. The assumptions validating ANOVA tests are that the residuals of the dependent variables from their fitted values are normally distributed and that the samples are independent of each other and that all populations have a common variance.[35] For the ANOVA to be appropriately used, the sources of compound variability must be attributed to the explanatory variables as opposed to the variability that could otherwise be attributed to other potential explanatory variables.

2.8.3.2. Monte Carlo Randomization Testing. When the assumption that data residuals from a linear model are not normally distributed is not justified, alternative parametric linear models could be considered for analysis. This is especially pertinent when models that are justified by the theory are available for analysis. Otherwise, a linear nonparametric model could be selected. One very general class of nonparametric methods that can be applied to factorial linear models is that of randomization tests.

Instead of following the standard theory underlying ANOVA based on treating the response variables as (normally distributed) random variables, randomization tests follow a completely different approach. The logic behind randomization hypothesis testing is that if a null hypothesis is true (so that the specified explanatory variable has no effect and is thus irrelevant to predicting the response variable), then the recorded values of the explanatory variable could be permuted without materially changing the information in the data. If we were to consider all possible permutations of the values of this explanatory factor, then the one true dataset should not stand out as unique among all of the artificially permuted datasets. In particular, for any test statistic quantifying the apparent size of the effect of the explanatory variable, then all possible permutations of the values of the test statistic calculated on the true and randomized datasets will be equally likely (under the null hypothesis). A *p*-value can then be directly calculated as the proportion of the randomized data test statistics that exceed the true data test statistic.

In practice, even for moderate sample sizes, there are far too many possible permutations of the explanatory variable values to follow this procedure literally. Instead, statisticians use Monte Carlo randomization tests,[55,56] where the complete enumeration of all possible permutations is replaced by a fixed number of random permutations. The procedure is very general and flexible and applies to many data structures and any test statistic.[57,58] However, due to the random number generation used, the *p*-value will vary slightly over runs. For some authors, there is a concern that the Monte Carlo *p*-value is therefore open subject to misinterpretation.[59] However, this is not a serious concern, as unless a very small number of permutations are used, the variation in the calculated *p*-values will not be large enough to alter the scientific interpretation. In this study, 5000 permutations were used.

2.8.3.3. Random Effects Modeling. The ANOVA and Monte Carlo analyses described above treat all factors as having fixed effects. Some datasets require that some, or all, factors are treated as having random effects (refer to the Introduction for the distinction between fixed and random effects). In the case of random effects, the parameters to be estimated are just the parameters of the random effects distribution (typically zero-mean normal). For this type of modeling, individual-level effects can no longer be estimated, but this is acceptable if these are not of prime interest. Models with both fixed and random effects are called "mixed models". This is a rich class of models that in full generality can allow for example (a) variability related to a hierarchical structure within the data,[60] (b) dependence between samples (e.g., lack complete independence, repeated measures),[61] and (c) groups of observations whose distributions will be more related to one another than they are to other samples (e.g., "Litter Effect"[62]).

In the case of this work, most Shinglebacks had volatilome replicates, which should be accounted for in statistical analysis by including the factor AnimalID in the analysis. All replicates from a singular shingleback were taken from the same Bioregion, which partially confounds the two variables Bioregion and AnimalID (Table S1). Because these variables are cofounded, their effects cannot be separately estimated when using fixed effects analyses. Treating AnimalID as a random effect not only captures the reality that the replicates will be more similar than observations taken on different animals but also reduces the number of parameters to be estimated and eliminates confounding. To date, there is no suitable nonparametric test (e.g., Monte Carlo randomization tests) that accounts for random effects models for dimensionality reduction, so this work was limited to tests with Gaussian assumptions.

2.8.3.4. Simple Linear Regression. In accordance with the previous literature in which chemical "-omic" data were analyzed using a single variable, all analyses presented above were also conducted using one explanatory variable (Figure 2). In these instances, the sole explanatory variable was Bioregion.

2.9. Data Visualization. *2.9.1. p-Value Distributions.* The total number of compounds for which Bioregion had a significant effect was compared for all statistical analyses and transformations. Both the total number of compounds where there was a significant effect, along with those where there were significant effects that were shared between analyses, were calculated. The *p*-values generated from each statistical analysis were plotted against every statistical combination using the "pairs" function in R for visual comparison of *p*-value distributions.

2.9.2. Schweder and Spjøtvoll p-Value Plots. When conducting tests of multiple null hypotheses (in this case, hypotheses are that Bioregion has no effect for each compound) within the same dataset, even when all null hypotheses are true, the expectation is that some null hypotheses will be rejected (specific to this study, $p \leq 0.05$). All statistical decisions are subject to errors, in which a hypothesis is falsely accepted or rejected, which can be worrisome in large datasets. Schweder and Spjøtvoll[63] developed a graphical representation of the classification of *p*-values, to aid in the determination of falsely accepted or rejected null hypotheses that set the foundation for the later development of false discovery rate analysis. Here, the *p*-values from each compound are ranked and plotted on a graph where the *y*-axis represents the number of *p*-values greater than $p(N_p)$ and the *x*-axis represents (1-*p*). The slope of the graph provides an estimate of the number of true hypotheses.[63] For this work, Schweder and Spjøtvoll *p*-value plots (S&S plots) were utilized to compare results from all statistical tests and transformations.

2.10. False Discovery Rate. When many null hypotheses are tested on the same dataset (in this work, each compound), it is expected that a large number of these hypotheses are rejected (or considered "significant"; these results are called "discoveries"). Some of these discoveries will be true (that the hypothesis in question was correctly rejected) and the others will be false (the hypothesis should not have been rejected). The traditional approach to hypothesis testing implies that the number of false discoveries will be approximately a proportion $\mathbf{p}$ of the number of true null hypotheses, where $\mathbf{p}$ is the threshold *p*-value. In other words, when testing at significance level 0.05, 5% of the hypotheses tested will be declared "discoveries", even if all null hypotheses are true. This traditional approach could be called "controlling the familywise error rate".

Benjamini and Hochberg[64] proposed a solution to this undesirable outcome, by devising a method of multiple

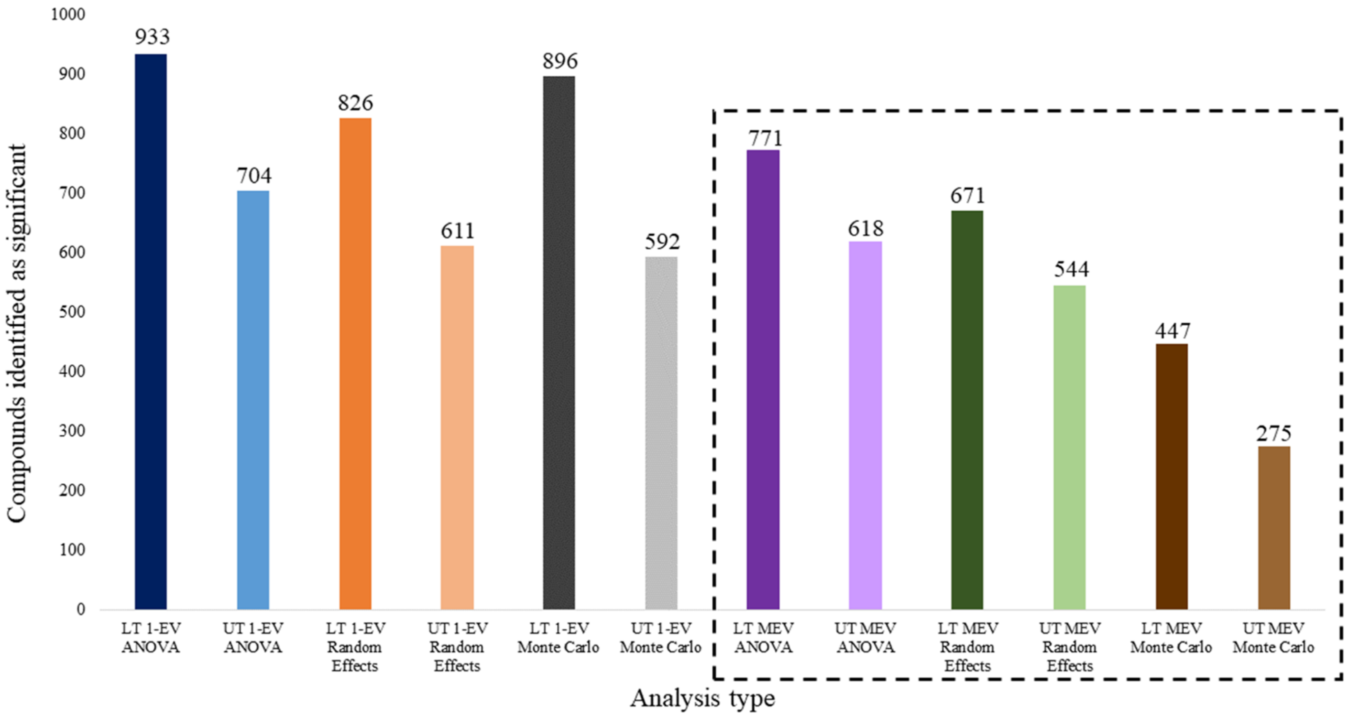


Figure 3. Number of compounds with a significant effect ($p \leq 0.05$) associated with Bioregion generated per statistical test. MEV represents "multiple explanatory variables", 1-EV represents "one explanatory variable", LT represents "log transformed", and UT represents "untransformed". The dashed square represents all MEV analyses.

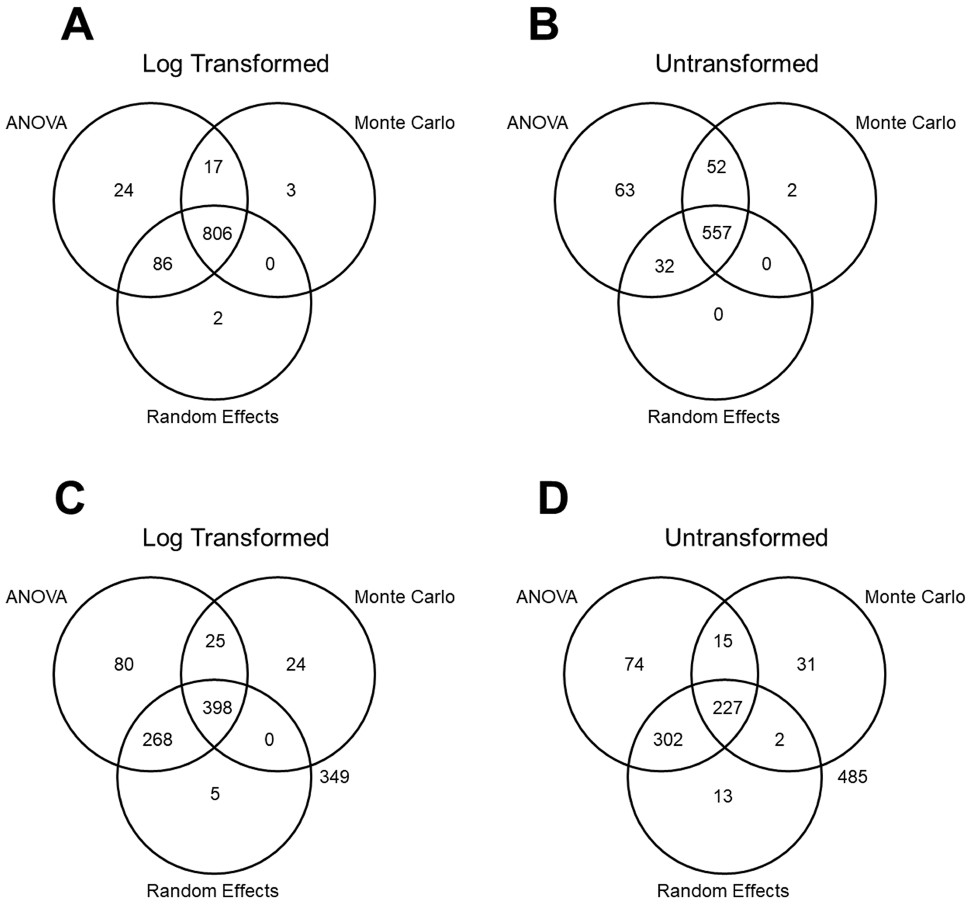


Figure 4. Venn diagram comparing the number of ($p \leq 0.05$) compounds that had significant effects associated with Bioregion shared across all tests. (A, B) Results from statistical analyses with one explanatory variable. (C, D) Results from multiple explanatory variables.

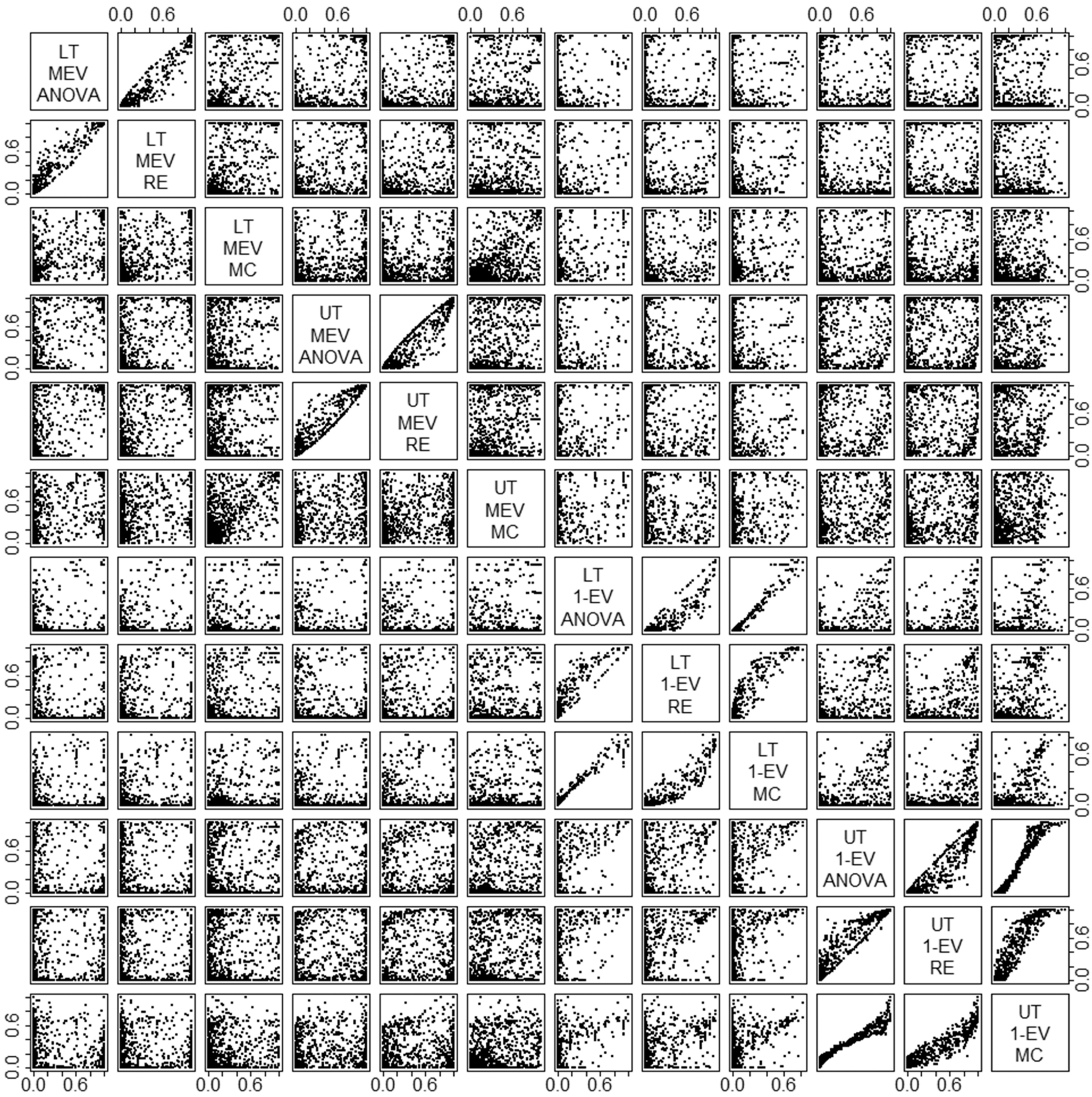


Figure 5. Pairs plot showing pairwise comparisons between *p*-values according to different analyses across all compounds. Each axis represents the *p*-values from each respective test. "LT" represents "log transformed", "UT" represents "untransformed", "MEV" represents "multiple explanatory variables", and "1-EV" represents "one explanatory variable". Each individual panel has a scale (0,1) for both axes.

hypothesis testing where the number of false discoveries is instead (approximately) a prefixed proportion α of the total number of discoveries. This is achieved by post-processing the entire collection of individual test *p*-values ranked in numerical order, and sequentially applying successively stricter criteria for declaration as a discovery. This approach is called "controlling the false discovery rate". False discovery rate (FDR) analysis is often used instead of family-wise error rates for chemical and genetic "-omic" studies as this method is considered less conservative.[64,65]

FDR has become a more routinely implemented analysis as larger datasets have been produced by high-throughput technologies (e.g., GC × GC-ToFMS). Additionally, FDR is useful when analyzing high-dimensional data (e.g., number of compounds) which have large numbers of explanatory variables and limited sample sizes. FDR ultimately introduces more conservative estimations of the effects from explanatory variables and also facilitates the identification of compounds of interest for further analysis or experimentation. Controlling the number of false discoveries is beneficial when conducting multiple tests simultaneously, especially in novel "-omics" type work. Though routinely used in genomic work,[66] FDR has only recently been performed on volatilome datasets.[67] For this work, we defined α = 0.05. In doing so, we have bound our false discovery rate to be 5%.

3. RESULTS AND DISCUSSION

3.1. Compound Classification. Prior to compound classification, 1407 discrete compounds were tentatively identified across all samples. After an assessment of the literature, it was determined that 131 compounds were classified as "biologically unstable" and 122 compounds were classified as "unnatural" and were subsequently removed. The remaining 1154 compounds were classified as "natural" (n = 725) or "unknown" (n = 429). Five additional compounds (Table S1, compound numbers 615, 633, 658, 742, 1025) which were only identified once, or did not exhibit scalar differences, were not included in the analysis as variances could not be calculated. In total, 1149 compounds were assessed by all models.

3.2. Impacts of Log Transformation. The use of the log transformation led to an increase in the number of compounds for which the effect of the Bioregion factor was significant for each statistical analysis compared to their untransformed counterparts (Figure 3). Statistical models with similar assumptions (i.e., residuals that were normally distributed, single or multiple explanatory variables) and with the same transformations (e.g., log transformed) shared comparable outcomes for all three assessments, including the number of significant compounds (Figure 4), the distribution of *p*-values (Figure 5) and the slope of the S&S plots (Figure 6). In some instances, the number and distribution of significant compounds (see Section 2.8 for clarification of a "significant compound") were more impacted by the log transformation as opposed to the selection of fixed effects or random effects models. For example, the results from the log transformed multiple explanatory variable (MEV) ANOVA were more similar to the results from the log transformed MEV RE model as compared to the results generated from the MEV ANOVA whose data were untransformed (Figures 4 and 5). In this case, the data from the log transformed MEV ANOVA and MEV RE models shared a similar number of compounds which were considered significant due to Bioregions (Figure 3), *p*-value distributions (i.e., concentrated to the diagonal; Figure 5) and S&S plots (i.e., shape and slope; Figure 6) compared to their untransformed counterparts.

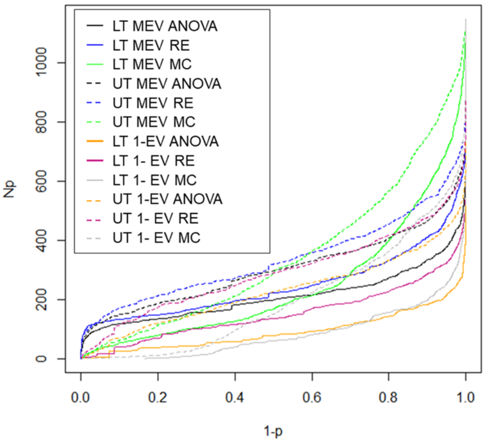


Figure 6. Schweder and Spjøtvoll plot of ordered *p*-values for all log transformed and untransformed statistical tests. The *y*-axis represents the number of *p*-values greater than *p*, and the *x*-axis is a uniform distribution of (1-*p*) scaled to (0,1).

3.3. Impacts of Multiple Explanatory Variables. Throughout this work, it was determined that many of the defined multiple explanatory variables had significant effects on compound peak area (data not shown). Excluding relevant explanatory variables from the analysis will bias the estimates of the effects of the variables that are included. For this work, this occurred for all tests when solely using the variable, Bioregion. When conducting this analysis, the estimated effect of Bioregion was exaggerated and hence the number of compounds for which Bioregion shows a significant effect was also exaggerated. However, for the purposes of comparison, all 1-EV analyses were still included in this work.

3.4. Interpretation of Statistical Results. *3.4.1. ANOVA Versus Monte Carlo.* Overall, the log transformation had a noticeable impact on the number of significant compounds found for both the ANOVA and MC tests. The ANOVA with one explanatory variable (1-EV) analyzed using log transformed (LT) data produced the highest number of significant compounds of any conducted statistical test and had over 200 more significant compounds than its untransformed (UT) counterpart (Figure 3). The LT 1-EV MC simulation generated the second highest number of significant compounds of all statistical analyses or transformations, despite its differing assumptions of normality (Figure 3). By contrast, the UT 1-EV MC analysis generated the least number of significant compounds (n = 592) of all UT 1-EV analyses (Figure 3). Both the LT and UT 1-EV ANOVA results shared the greatest number of significant compounds with all other LT and UT 1-EV analyses (Figure 4A,B). Unlike the LT analyses, the UT 1-EV ANOVA shared the second most compounds with the UT 1-EV MC, despite having the same assumption of Gaussian residual distribution with the UT 1-EV RE model (Figure 4B). Although the UT 1-EV ANOVA shared more significant compounds solely with the UT 1-EV MC analysis (i.e., 52 as opposed to 17; Figure 4A,B), the distribution of *p*-values generated from the LT 1-EV ANOVA was more linear with the distribution generated from the LT 1-EV MC analyses (Figure 5). This may be explained as overall, the LT 1-EV ANOVA and MC shared a larger total number of significant compounds in comparison to their UT counterparts (823 vs 609; Figure 4) or that the log transformation made the residuals more normally distributed. The slope of the LT 1-EV ANOVA, LT 1-EV MC, and UT 1-EV MC analyses on the S&S plots were zero toward the left side, indicating that there were no "true null hypotheses" for those compounds (Figure 6). The notable exception is the LT 1-EV MC S&S plot, where the (1-*p*) values generated from this analysis did not begin at the origin as with the rest of analyses, but instead at ∼0.17. Despite this, the slope is approximately 0, suggesting that there were no true null hypotheses in this analysis (Figure 6). It is likely that the assumption that there are no true null hypotheses for these compounds is incorrect, as it was demonstrated that other explanatory variables had significant effects on compound peak area.

As with the LT 1-EV methods, the LT MEV ANOVA shared the most significant compounds with all other log transformed, MEV analyses (i.e., 398, Figure 1C). The UT MEV MC analysis generated the lowest number of significant compounds (n = 275) of any statistical analysis with or without a log transformation (Figure 3). This type of analysis is likely the most conservative, as MC analyses make the least assumptions about the data, and the data has not been transformed which will alter the distribution of residuals. This analysis also shared the least number of significant compounds with other analyses (n = 244) of any other statistical test comparison (Figure 4D). Unlike the LT 1-EV ANOVA, there was little correlation between *p*-values with the LT MEV ANOVA and any other log transformed untransformed analysis (Figure 5). The UT MEV MC analysis did not show a clear pattern of *p*-value distributions with any other analysis but did demonstrate clustering toward lower *p*-values with the LT MEV MC analysis, where LT MEV MC *p*-values trended to the lower *p*-values (Figure 5). This indicates that the effect of Bioregion on compound peak area is largely affected by the log transformation when all other elements of data analysis are constant. The slope of the S&S *p*-value plot for the LT and UT MEV ANOVA were infinite on the left side (Figure 6), leaving these slopes to be uninterpretable by methods outlined by Schweder and Spjøtvoll.[63] The biological interpretation of this slope directly contradicts the p-value plots generated by the LT 1-EV ANOVA, LT 1-EV MC, and UT 1-EV MC models. Again, this supports the proposal that excluding multiple explanatory variables from analysis exaggerates the effect of significance of the explanatory variable Bioregion.

3.4.2. ANOVA vs Random Effects Model. The log transformation also affected how closely related the results of the two models with the same Gaussian assumptions were to each other. The 1-EV ANOVA and RE with the same transformation shared more significant compounds with each other than their untransformed counterparts (Figure 3). The LT 1-EV ANOVA shared the greatest number of compounds (n = 806) with all other LT 1-EV analyses, followed by the LT 1-EV RE model (n = 86) (Figure 4A), contradicting its UT counterpart (Figure 4B). For both transformations, the 1-EV ANOVA generated smaller p-values than the 1-EV RE models (Figure 4). The biological interpretations of the 1-EV results are directly contradictory, depending on whether or not the log transformation was applied (Figure 6).

As with the 1-EV models, the number of significant compounds generated by the LT MEV ANOVA was more similar to the LT MEV RE model than its counterpart which was untransformed (Figure 3). The LT MEV RE model generated the least number of compounds that were discrete to this analysis (n = 5), which was likely due to the large number of significant compounds shared with the LT MEV ANOVA (n = 268; Figure 4C). The p-value distribution from the LT MEV RE model most closely resembled the LT MEV ANOVA and was quite different from that of any other analysis (Figure 4). The p-value distribution of the UT MEV RE analysis did not share a similar distribution pattern with any other analysis except the UT MEV ANOVA (Figure 5). Here, the p-values of the UT MEV RE analysis were higher than its fixed effect counterpart, except for some clustering seen between 0 and 0.2 (Figure 5). This may indicate that the additional explanatory variable, AnimalID, being analyzed also impacts inference about the effect of significance of Bioregion. The slopes of both LT and UT MEV analyses on the S&S p-value plot from this analysis were infinity on the left-hand side, preventing interpretation (Figure 6). These slopes are biologically contradictory to the S&S p-value plots generated by their 1-EV counterparts, where the LT slopes were 0 and the UT slopes were interpretable (Figure 6).

3.4.3. Random Effects vs Monte Carlo Simulation. The number of significant compounds generated from both the LT and UT 1-EV RE analyses differed less between the MC analyses than the ANOVA analyses (Figure 3). Although these two analyses shared a similar number of significant compounds, the distribution of those p-values, along with true null hypotheses, noticeably differs. For both 1-EV transformations, the RE and MC models did not share any significant compounds (Figure 4A,B).

The LT MEV MC model produced the least number of statistically significant compounds of any LT model and the second least number of statistically significant compounds overall (Figure 3). The MC analyses have less data assumptions and are less likely to exaggerate the effect of significance of the explanatory variable Bioregion. As both LT and UT MEV MC analyses produced the least number of compounds with significant effects associated with Bioregion, these analyses are likely the most conservative data analysis models. As both the ANOVA and RE models share the assumption of a Gaussian distribution, the unequal distribution of shared significant compounds between the MEV RE and MEV MC models may be associated with a different data assumption (Figure 4). It is likely that the replicates found within the dataset, or the inclusion of the variable AnimalID (replicate samples), impacts the effect of Bioregion on some compounds. The p-value distributions of the LT MEV MC model did not compare well with any other statistical test (Figure 5). Of all of the S&S p-value plots, the LT MEV MC plot was the most linear of the LT statistical models and the UT MEV MC plot was the most linear overall (Figure 6). As with the 1-EV counterparts, the slope of the S&S p-value plots between the UT and LT MEV MC and RE models was biologically contradictory (Figure 6).

3.4.4. False Discovery Rate. When using the Benjamini and Hochberg, procedure,[64] the number of discoveries was the lowest for the UT MEV Monte Carlo analysis. All tests that had an assumption of a Gaussian distribution for the residuals had 3−11 times more discoveries than the corresponding Monte Carlo analyses when the selected p-value was <0.05 (Table 1). For this work, the log transformation increased the number of discoveries between 29 and 461% in comparison to the untransformed counterparts (Table 1). There was a similar proportion of false discovery (in terms of Type 1 error)

Table 1. Number of Identified Rejected Hypotheses vs the Number of Discoveries after FDR Analysis for Each Statistical Test for Log Transformed and Untransformed Data

	log transformed		
statistical method	original number of rejected hypotheses (p-value ≤ 0.05)	maximum p-value that declares a discovery	number of discoveries (α = 0.05)
1-EV ANOVA	933	0.0400	921
1-EV random effects	826	0.0347	798
1-EV Monte Carlo	896	0.0318	870
MEV ANOVA	771	0.0312	733
MEV random effects	671	0.0263	606
MEV Monte Carlo	447	0.0092	213
	untransformed		
statistical method	original number of rejected hypotheses (p-value ≤ 0.05)	maximum p-value that declares a discovery	number of discoveries (α = 0.05)
1-EV ANOVA	704	0.0293	673
1-EV random effects	611	0.0245	571
1-EV Monte Carlo	592	0.0226	522
MEV ANOVA	618	0.0234	533
MEV random effects	544	0.0202	470
MEV Monte Carlo	275	0.0016	38

between transformed and untransformed datasets (Table 1). The two analyses most affected by the false discovery, in terms of number of discoveries, were the MEV MC analyses. The reduction in the number of the identified compounds of interest is beneficial from a conservative downstream data analysis perspective. As the cost of running volatilome analysis is high, and the current sample sizes are limited, more conservative assumptions are beneficial as they lead to lower Type 1 errors.

4. CONCLUSIONS

The aim of this manuscript was to compare the results of multiple statistical analyses for identifying compounds for which Bioregions had a significant effect. The identification of these compounds can be used for further analyses, such as for further BVOC biomarker investigation or for identifying candidate compounds for classification models (e.g., random forest classifier). The comparison of statistical outputs from linear models with both single and multiple explanatory variables, with or without log transformations, was intended to improve the selection of dimensionality reduction techniques when assessing biological volatilome data with many underlying variables. Historically, the most employed volatilome dimensionality reduction methods have relied on comparing level-to-level variance to intralevel variance that include Student *t*-test filtering[19] and Fisher ratio filtering.[13,68,69] As biological data inherently contain numerous underlying variables and relationships, it is critical to investigate the impacts of statistical assumption violations and their relationship to the significance of analyzed variables. This work supports that traditional dimensionality reduction analyses are not appropriate for Shingleback volatilome dimensionality reduction. These methods are either limited in ability to assess multivariate data (e.g., *t*-test filtering, Fisher ratio filtering), data assumptions (e.g., presence of linear correlations and assumptions of orthogonality with PCA analysis) or variables with different units (e.g., if units are not transformed in PCA). The work conducted here required less data and unit transformations which lead to a more direct interpretation of the generated results.

The assessment of significant effects due to a target variable for biological datasets is inherently complex. This is in part because the effect of biological variables can be independent of each other, by which the effect would be additive on an appropriate scale, or can interact, making their effects multiplicative to a certain scale, or a combination of the two.[33] The distribution for each variable can also be Gaussian or non-normal, which affects both transformations and model statistical assumptions. Because of this, each statistical analysis conducted in this study could not be conducted in a manner that all assumptions could be satisfied. For example, it is not reasonable to ignore the replication of individuals in this dataset or the extreme skewness of peak area distributions that would necessitate a log transformation.

All analyses demonstrated that at least 38 compounds had significant effects attributed to Bioregion, and in one case as many as 921. This indicates that Bioregion likely has a significant effect on the compound peak area present in Shingleback BVOCs. However, this effect is not quantifiable, as the extent to which this effect is significant differs based on data assumptions and transformations. Overall, the number of compounds that had a significant effect due to Bioregion differed due to the log transformation and the addition of other explanatory variables. Specifically, the log transformation increased the indicative significance of Bioregion for all statistical analyses. The log transformation also increased the incidence of rejected null hypotheses for all statistical analyses. From a conservative data perspective, this is likely an inappropriate transformation for Shingleback volatilome data when the selected p-value ≤ 0.05. Most of the untransformed p-value plots had steeper slopes than their logged counterparts. This indicates that there are more true hypotheses, or that there are alternatives other than Bioregion that may be causing the effect of significance for these compounds. In this work, it was demonstrated that multiple other explanatory variables also had a significant effect on compound peak area. This indicates that the exclusion of multiple explanatory factors will lead to an overexaggeration in the number of compounds with a significant effect due to Bioregion. Thus, all models analyzed with one explanatory variable are inappropriate for accurate Shingleback volatilome dimensionality reduction.

The most appropriate model for Shingleback volatilome dimensionality reduction for the purpose of classification analysis is likely the MEV MC analyses. The strength of the MC analysis is that it makes no distributional assumptions about the data. The slope of the p-value plots for both the LT and UT MEV MC analyses are very similar and indicate the potential for other true null hypotheses for those compounds. It is likely that other variables that were not included in this study, including reproductive status or time of day in which the animal was sampled, may have also influenced the total volatilome profile. From a conservative data perspective, the inclusion of untested alternative variables also supports MC analyses as the most appropriate data reduction tool for this Shingleback volatilome dataset. The UT MEV MC analysis had the lowest number of rejected null hypotheses in comparison to the other statistical tests by the largest margin. It may be that this combination of transformations, variables, and model is the most appropriate for further classification analysis. The MC analysis is still limited in that it assumes that all variables are independent and is thus limited with repeated measured sampling. Further analysis is required to determine the influence of repeated measures or individual variation on Shingleback dimensionality reduction.

Although this is the largest animal volatilome database to date, this sample size is still limited in a statistical sense, in view of the large number of explanatory variables and the very large number of compounds. It is likely that the effects of some explanatory variables are additive, while others are multiplicative. To handle this properly would require going beyond linear models, and hence computational methods not routinely found in statistical packages. The continued building of this database would be beneficial so that higher-order hierarchical analyses can be conducted.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c01613.

Additional experimental details, including R code for models analyzed with one and multiple explanatory variables (PDF)

Raw data analyzed (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Amber O. Brown − *Australian Museum Research Institute, Australian Museum, Sydney 2001 NSW, Australia; Centre for Forensic Science, University of Technology Sydney, Ultimo 2007 NSW, Australia;* orcid.org/0000-0002-6761-2170; Email: Aosingabrown@gmail.com

Authors

Peter J. Green − *University of Bristol, Bristol BS8 1UG, U.K.; University of Technology Sydney, Ultimo 2007 NSW, Australia*

Greta J. Frankham − *Australian Museum Research Institute, Australian Museum, Sydney 2001 NSW, Australia; Centre for Forensic Science, University of Technology Sydney, Ultimo 2007 NSW, Australia*

Barbara H. Stuart − *Australian Museum Research Institute, Australian Museum, Sydney 2001 NSW, Australia*

Maiken Ueland − *Australian Museum Research Institute, Australian Museum, Sydney 2001 NSW, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c01613

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

All volatilome collection methods were conducted using procedures approved by the UTS Animal Ethics Committee (ETH19-3420) and Flinders University Animal Ethics Committee (E454/17). Shingleback volatilome samples taken from wild animals were approved through the following permits: NSW Scientific License (SL102281), South Australia (A23436-27), Western Australia through the Authorisation to take or disturb threatened species (TFA 2020-0062), Fauna taking licence (FO25000284), and the Rottnest Island Research Permit (2020/374679). This work was funded through the RSPCA Alan White Scholarship for Animal Welfare Research, the Society of Wildlife Forensic Science Small Grants Fund, The Holsworth Wildlife Research Endowment & The Ecological Society of Australia, the UTS International Research Scholarship, UTS President's Scholarship and the Australian Research Council DECRA Fellowship (DE210100949). The authors thank Andrew V. Brown for financial and sampling assistance; the staff at Featherdale Wildlife Park including Chad Staples, Christina Kingsford, Sam Hermann, and Rebecca Gallagher for sampling assistance; Samara Garrett-Rickman, Alisha Deo, Nicola Hubert-Smith, Dr. Gerrut Norval, Dr. Jessa Clayton, Dr. Aaron Fenner, and Dr. Paul Thomas for sampling assistance; and Sandali Alahakone for assistance in sample analysis. The authors also thank the NSW and WA National Parks services and the Rottnest Island Authority for accommodation.

■ REFERENCES

(1) Caissard, J. C.; Joly, C.; Bergougnoux, V.; Hugueney, P.; Mauriat, M.; Baudino, S. Secretion mechanisms of volatile organic compounds in specialized cells of aromatic plants. *Recent Res. Dev. Cell Biol.* **2004**, *2*, 1−15.

(2) Niinemets, Ü.; Fares, S.; Harley, P.; Jardine, K. J. Bidirectional exchange of biogenic volatiles with vegetation: emission sources, reactions, breakdown and deposition. *Plant Cell Environ.* **2014**, *37*, 1790−1809.

(3) Leguet, A.; Gibernau, M.; Shintu, L.; Caldarelli, S.; Moja, S.; Baudino, S.; Caissard, J. C. Evidence for early intracellular accumulation of volatile compounds during spadix development in *Arum italicum L.* and preliminary data on some tropical Aroids. *Naturwissenschaften* **2014**, *101*, 623−635.

(4) Buljubasic, F.; Buchbauer, G. The scent of human diseases: A review on specific volatile organic compounds as diagnostic biomarkers. *Flavour Fragrance J.* **2015**, *30*, 5−25.

(5) Watson, S. B.; Monis, P.; Baker, P.; Giglio, S. Biochemistry and genetics of taste-and odor-producing cyanobacteria. *Harmful Algae* **2016**, *54*, 112−127.

(6) Kistler, M.; Szymczak, W.; Fedrigo, M.; Fiamoncini, J.; Höllriegl, V.; Hoeschen, C.; Klingenspor, M.; de Angelis, M. H.; Rozman, J. Effects of diet-matrix on volatile organic compounds in breath in diet-induced obese mice. *J. Breath Res.* **2014**, *8*, No. 016004.

(7) Fischer, S.; Bergmann, A.; Steffens, M.; Trefz, P.; Ziller, M.; Miekisch, W.; Schubert, J. S.; Köhler, H.; Reinhold, P. Impact of food intake on in vivo VOC concentrations in exhaled breath assessed in a caprine animal model. *J. Breath Res.* **2015**, *9*, No. 047113.

(8) Noonan, M. J.; Tinnesand, H. V.; Müller, C. T.; Rosell, F.; Macdonald, D. W.; Buesching, C. D. Knowing me, knowing you: anal gland secretion of European badgers (*Meles meles*) codes for individuality, sex and social group membership. *J. Chem. Ecol.* **2019**, *45*, 823−837.

(9) Verhulst, N. O.; Beijleveld, H.; Knols, B. G.; Takken, W.; Schraa, G.; Bouwmeester, H. J.; Smallegange, R. C. Cultured skin microbiota attracts malaria mosquitoes. *Malar. J.* **2009**, *8*, No. 302.

(10) Sagar, N. M.; Cree, I. A.; Covington, J. A.; Arasaradnam, R. P. The interplay of the gut microbiome, bile acids, and volatile organic compounds. *Gastroenterol. Res. Pract.* **2015**, *2015*, No. 398585.

(11) Kasal-Slavik, T.; Eschweiler, J.; Kleist, E.; Mumm, R.; Goldbach, H. E.; Schouten, A.; Wildt, J. Early biotic stress detection in tomato (*Solanum lycopersicum*) by BVOC emissions. *Phytochemistry* **2017**, *144*, 180−188.

(12) Lawson, C. A.; Possell, M.; Seymour, J. R.; Raina, J. B.; Suggett, D. J. Coral endosymbionts (*Symbiodiniaceae*) emit species-specific volatilomes that shift when exposed to thermal stress. *Sci. Rep.* **2019**, *9*, No. 17395.

(13) Ueland, M.; Brown, A.; Bartos, C.; Frankham, G. J.; Johnson, R. N.; Forbes, S. L. Profiling volatilomes: A novel forensic method for identification of confiscated illegal wildlife items. *Separations* **2020**, *7*, 5.

(14) Braun, B. Wildlife Detector Dogs - A Guideline on the Training of Dogs to Detect Wildlife in Trade WWF Germany: Berlin, 2013; pp 1−16.

(15) Furton, K. G.; Caraballo, N. I.; Cerreta, M. M.; Holness, H. K. Advances in the use of odour as forensic evidence through optimizing and standardizing instruments and canines. *Philos. Trans. R. Soc., B* **2015**, *370*, No. 20140262.

(16) Dewulf, J. O.; Van Langenhove, H.; Wittmann, G. Analysis of volatile organic compounds using gas chromatography. *TrAC, Trends Anal. Chem.* **2002**, *21*, 637−646.

(17) Olander, A.; Lawson, C. A.; Possell, M.; Raina, J. B.; Ueland, M.; Suggett, D. J. Comparative volatilomics of coral endosymbionts from one-and comprehensive two-dimensional gas chromatography approaches. *Mar. Biol.* **2021**, *168*, 76.

(18) Stadler, S.; Stefanuto, P. H.; Brokl, M.; Forbes, S. L.; Focant, J. F. Characterization of volatile organic compounds from human analogue decomposition using thermal desorption coupled to comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. *Anal. Chem.* **2013**, *85*, 998−1005.

(19) Perrault, K. A.; Stefanuto, P. H.; Stuart, B. H.; Rai, T.; Focant, J. F.; Forbes, S. L. Reducing variation in decomposition odour profiling using comprehensive two-dimensional gas chromatography. *J. Sep. Sci.* **2015**, *38*, 73−80.

(20) Zanella, D.; Focant, J. F.; Franchina, F. A. 30th Anniversary of comprehensive two-dimensional gas chromatography: Latest advances. *Anal. Sci. Adv.* **2021**, *2*, 213−224.

(21) Pierce, K. M.; Hoggard, J. C.; Hope, J. L.; Rainey, P. M.; Hoofnagle, A. N.; Jack, R. M.; Wright, B. W.; Synovec, R. E. Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts. *Anal. Chem.* **2006**, *78*, 5068−5075.

(22) Heim, J.Utilization of statistical compare software and fisher ratios prior to multivariate analysis for complex GCxGC-TOFMS data in order to define statistical variation between the small molecule metabolite profiles of different fish species *Appl. Notes LECO* 2010, p 14.

(23) Stefanuto, P. H.; Smolinska, A.; Focant, J. F. Advanced chemometric and data handling tools for GC × GC-TOF-MS. *TrAC, Trends Anal. Chem.* **2021**, *139*, No. 116251.

(24) Trinklein, T. J.; Cain, C. N.; Ochoa, G. S.; Schöneich, S.; Mikaliunaite, L.; Synovec, R. E. Recent advances in GC × GC and Chemometrics to addressing emerging challenges in nontargeted analyses. *Anal. Chem.* **2023**, *95*, 264−286.

(25) Verleysen, M.; François, D. The Curse of Dimensionality in Data Mining and Time Series Prediction. In *Computational Intelligence and Bioinspired Systems: 8th International Work-Conference on Artificial Neural Networks*, IWANN 2005, Vilanova i la Geltrú, Barcelona, Spain. Proceedings 8; Springer: Berlin Heidelberg, 2005; pp 758−770.

(26) Fischer, S.; Trefz, P.; Bergmann, A.; Steffens, M.; Ziller, M.; Miekisch, W.; Schubert, J. S.; Köhler, H.; Reinhold, P. Physiological variability in volatile organic compounds (VOCs) in exhaled breath and released from faeces due to nutrition and somatic growth in a standardized caprine animal model. *J. Breath Res.* **2015**, *9*, No. 027108.

(27) Conte, M.; Conte, G.; Salvioli, S. VOCs profile can discriminate biological age. *Aging* **2021**, *13*, 9156.

(28) Zhang, J. X.; Soini, H. A.; Bruce, K. E.; Wiesler, D.; Woodley, S. K.; Baum, M. J.; Novotny, M. V. Putative chemosignals of the ferret (*Mustela furo*) associated with individual and gender recognition. *Chem. Senses* **2005**, *30*, 727−737.

(29) Curran, A. M.; Ramirez, C. F.; Schoon, A. A.; Furton, K. G. The frequency of occurrence and discriminatory power of compounds found in human scent across a population determined by SPME-GC/MS. *J. Chromatogr. B* **2007**, *846*, 86−97.

(30) Kean, E. F.; Müller, C. T.; Chadwick, E. A. Otter scent signals age, sex, and reproductive status. *Chem. Senses* **2011**, *36*, 555−564.

(31) Woidtke, L.; Dreßler, J.; Babian, C. Individual human scent as a forensic identifier using mantrailing. *Forensic Sci. Int.* **2018**, *282*, 111−121.

(32) Buesching, C. D.; Waterhouse, J. S.; Macdonald, D. W. Gas-chromatographic analyses of the subcaudal gland secretion of the European badger (*Meles meles*) part I: chemical differences related to individual parameters. *J. Chem. Ecol.* **2002**, *28*, 41−56.

(33) Siemiatycki, J.; Thomas, D. C. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.* **1981**, *10*, 383−387.

(34) Rust, L.; Nizio, K. D.; Forbes, S. L. The influence of ageing and surface type on the odour profile of blood-detection dog training aids. *Anal. Bioanal. Chem.* **2016**, *408*, 6349−6360.

(35) Fisher, R. A. *Theory of Statistical Estimation*, Mathematical Proceedings of the Cambridge Philosophical Society; Cambridge University Press, 1995; pp 700−725.

(36) Blyth, C. R. On Simpson's paradox and the sure-thing principle. *J. Am. Stat. Assoc.* **1972**, *67*, 364−366.

(37) Wagner, C. H. Simpson's paradox in real life. *Am. Stat.* **1982**, *36*, 46−48.

(38) Kievit, R. A.; Frankenhuis, W. E.; Waldorp, L.; Borsboom, D. Simpson's paradox in psychological science: A practical guide. *Front. Psychol.* **2013**, *4*, 513.

(39) Broadhurst, D. I.; Kell, D. B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2007**, *2*, 171−196.

(40) Brown, A. O.; Frankham, G. J.; Stuart, B. H.; Ueland, M. Reptile volatilome profiling optimisation: A pathway towards forensic applications. *Forensic Sci. Int.: Animal Environ.* **2021**, *1*, No. 100024.

(41) Changyong, F.; Hongyue, W.; Naiji, L. U.; Tian, C. H.; Hua, H. E.; Ying, L. U. Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* **2014**, *26*, 105.

(42) Schoeman, J. C.; du Preez, I.; Du Preez, I. A comparison of four sputum pre-extraction preparation methods for identifying and characterising *Mycobacterium tuberculosis* using GCxGC-TOFMS metabolomics. *J. Microbiol. Methods* **2012**, *91*, 301−311.

(43) Purcaro, G.; Stefanuto, P. H.; Franchina, F. A.; Beccaria, M.; Wieland-Alter, W. F.; Wright, P. F.; Hill, J. E. SPME-GC × GC-TOF MS fingerprint of virally-infected cell culture: Sample preparation optimization and data processing evaluation. *Anal. Chim. Acta* **2018**, *1027*, 158−167.

(44) Laothawornkitkul, J.; Taylor, J. E.; Paul, N. D.; Hewitt, C. N. Biogenic volatile organic compounds in the Earth system. *New Phytol.* **2009**, *183*, 27−51.

(45) Pocock, S. J.; Hughes, M. D.; Lee, R. J. Statistical problems in the reporting of clinical trials. *N. Engl. J. Med.* **1987**, *317*, 426−432.

(46) Nickell, S. Biases in dynamic models with fixed effects. *Econometrica* **1981**, *49*, 1417−1426.

(47) Norval, G.; Gardner, M. G. The natural history of the sleepy lizard, *Tiliqua rugosa* (Gray, 1825)−Insight from chance observations and long-term research on a common Australian skink species. *Austral Ecol.* **2020**, *45*, 410−417.

(48) SEED The Central Resource for Sharing and Enabling Environmental Data in NSW 2021 https://geo.seed.nsw.gov.au/Public_Viewer/index.html?viewer=Public_Viewer&locale=en-AU&runWorkflow=AppendLayerCatalog&CatalogLayer=SEED_Catalog.140.IBRA7%20regions.

(49) Morrison, J. D.; Nicholson, A. J. C. Studies of ionization efficiency. Part II. The ionization potentials of some organic molecules. *J. Chem. Phys.* **1952**, *20*, 1021−1023.

(50) Choi, S. S.; Lee, H. M.; Jang, S.; Shin, J. Comparison of ionization behaviors of ring and linear carbohydrates in MALDI-TOFMS. *Int. J. Mass Spectrom.* **2009**, *279*, 53−58.

(51) Bicchi, C.; Maffei, M. The Plant Volatilome: Methods of Analysis. In *High-Throughput Phenotyping in Plants*; Humana Press: Totowa, NJ, 2012; pp 289−310.

(52) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. Breaking with trends in pre-processing? *TrAC, Trends Anal. Chem.* **2013**, *50*, 96−106.

(53) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. Untargeted metabolomics strategies-challenges and emerging directions. *J. Am. Soc. Mass Spectrom.* **2016**, *27*, 1897−1905.

(54) Kaufmann, J.; Schering, A. G. *Analysis of Variance ANOVA*; Wiley Encyclopedia of Clinical Trials, 2007.

(55) Besag, J.; Diggle, P. J. Simple Monte Carlo tests for spatial pattern. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1977**, *26*, 327−333.

(56) Manly, B. F. *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd ed.; Chapman and Hall/CRC, 2018.

(57) Ferkingstad, E.; Holden, L.; Sandve, G. K. Monte Carlo null models for genomic data. *Stat. Sci.* **2015**, *30*, 59−71.

(58) Sinclair, E.; Walton-Doyle, C.; Sarkar, D.; Hollywood, K. A.; Milne, J.; Lim, S. H.; Kunath, T.; Rijs, A. M.; De Bie, R. M.; Silverdale, M.; Trivedi, D. K. Validating differential volatilome profiles in Parkinson's disease. *ACS Cent. Sci.* **2021**, *7*, 300−306.

(59) Harwell, M. R. Summarizing Monte Carlo results in methodological research. *J. Educ. Stat.* **1992**, *17*, 297−313.

(60) Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinburgh* **1919**, *52*, 399−433.

(61) Bryk, A. S.; Raudenbush, S. W. Application of hierarchical linear models to assessing change. *Psychol. Bull.* **1987**, *101*, 147−158.

(62) Kupper, L. L. Litter Effect *Wiley StatsRef: Statistics Reference Online* 2014.

(63) Schweder, T.; Spjøtvoll, E. Plots of p-values to evaluate many tests simultaneously. *Biometrika* **1982**, *69*, 493−502.

(64) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **1995**, *57*, 289−300.

(65) Sabatti, C.; Service, S.; Freimer, N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **2003**, *164*, 829−833.

(66) Bickel, D. R. *Genomics Data Analysis: False Discovery Rates and Empirical Bayes Methods*; CRC Press, 2019.

(67) Pinto, J.; Amaro, F.; Lima, A. R.; Carvalho-Maia, C.; Jerónimo, C.; Henrique, R.; Bastos, M. D. L.; Carvalho, M.; Guedes de Pinho, P. Urinary volatilomics unveils a candidate biomarker panel for noninvasive detection of clear cell renal cell carcinoma. *J. Proteome Res.* **2021**, *20*, 3068−3077.

(68) Stefanuto, P. H.; Perrault, K. A.; Stadler, S.; Pesesse, R.; LeBlanc, H. N.; Forbes, S. L.; Focant, J. F. GC × GC−TOFMS and supervised multivariate approaches to study human cadaveric decomposition olfactive signatures. *Anal. Bioanal. Chem.* **2015**, *407*, 4767−4778.

(69) Berrier, K. L.; Prebihalo, S. E.; Synovec, R. E. Advanced Data Handling in Comprehensive Two-dimensional Gas Chromatography. In *Separation Science and Technology*; Academic Press, 2020; Vol. *12*, pp 229−268.