

# Hypothesis Testing of Multivariate Biomechanical Responses using Statistical Parametric Mapping and Arc-Length Re-Parameterization

Devon C. Hartlen<sup>1</sup> & Duane S. Cronin<sup>\*,2</sup>

[1] [Devon.Hartlen@UWaterloo.ca](mailto:Devon.Hartlen@UWaterloo.ca). ORCID: 0000-0001-8170-7021

Department of Mechanical and Mechatronic Engineering, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada, N2L 3G1

[2] [Duane.Cronin@UWaterloo.ca](mailto:Duane.Cronin@UWaterloo.ca). ORCID: 0000-0001-7538-0560

Department of Mechanical and Mechatronic Engineering, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada, N2L 3G1

\* Corresponding Author

## Abstract

Detection of differences between experimental biomechanical data sets is critical to quantify effects and their significance. Many forms of biomechanical data are continuous and multivariate in nature, yet contemporary statistical analysis and hypothesis testing most often utilize single-value scalar metrics. However, reducing continuous responses to single-value scalar metrics can introduce bias and eliminate much of the physical context of a response. This study proposes a methodology to perform hypothesis testing directly on continuous multivariate experimental data sets. The methodology couples arc-length re-parameterization with statistical parametric mapping (SPM) to provide a general framework that can be applied to many of the response types found in biomechanics, including sets of responses that do not terminate at a common coordinate or are hysteretic, such as load-unload data. The arc-length-based SPM methodology was applied to three literature data sets representing a cross-section of the types of responses encountered in biomechanics. In each case, the arc-length-based SPM methodology produced results that agreed with contemporary statistical techniques while providing quantification and identification of statistically significant differences between the data sets. The proposed method provided important contextual information and a deeper understanding of the underlying behaviour of a dataset that would otherwise be missed using contemporary single-value scalar metric statistical techniques, such as highlighting specific response features that drive differences between datasets.

## 1 Introduction and Background

Inferring differences between datasets or test conditions is a key challenge in biomechanics [1], complicated by the fact that most experimental data is continuous in nature, such as time-series kinematics or the force-displacement response of materials or components. While characteristic averages and response corridors have been commonly used to capture the mean and variability of a particular set of responses [2,3], corridors do not provide a statistically valid way to infer differences between sets of responses. As such, continuous responses are often reduced to single-value scalar metrics such as peak acceleration or maximum force to be analyzed with statistical techniques such as Student's *t*-test or ANOVA. However, care must be taken when extracting scalar metrics to ensure that the metrics are appropriate for discerning differences without introducing bias by focusing on a specific feature of a set of responses [4]. Ideally, statistical analysis of continuous responses should also be continuous, eliminating the need to select scalar value metrics.

One approach that removes the need to reduce responses to scalar metrics is statistical parametric mapping (SPM), which enables statistical hypothesis testing on continuous experimental data. Originally developed for neuroimaging [5,6], SPM has seen some adoption in biomechanics, particularly for gait analysis [7]. However, current implementations of SPM are limited to datasets that can be mapped to a common sampling variable, such as a known brain atlas in neuroimaging [8], or responses with shared start and end points that can be cycle-normalized, such as gait data [4]. Many types of responses common in biomechanics do not have common termination coordinates or cannot be cycle normalized, limiting the application of SPM.

One solution arises from Hartlen and Cronin [2], who presented a general arc-length-based methodology to calculate a characteristic average and response corridors capable of accommodating a wide range of datasets, including responses that do not terminate at common coordinates, experience significant oscillatory behaviour, and even hysteretic load-unload data. While the resulting corridors could not be used to perform statistical inferences, the approach utilized arc-length re-parameterization to redefine responses with respect to a common sampling variable, specifically normalized arc-length, regardless of the shape of the inputted responses. Further, signal registration aligned shared features of responses with respect to normalized arc-length without impacting the underlying shape and variability of the responses. While studies in the literature have investigated applications of arc-length or similar measures for shape analysis (e.g. [9]), they have been focused on mapping responses to a target or atlas, which is a limitation when considering biomechanical responses where a template or representative response often does not exist or cannot be objectively selected.

The current study proposes an arc-length-based SPM methodology for statistical hypothesis testing to compare families of multivariate continuous responses, moving beyond scalar

metrics. The methodology enables two or more sets of continuous responses to be resampled to a common sampling variable while aligning shared features using arc-length re-parameterization and signal registration. SPM is applied using multivariate statistical tests to perform feature-based statistical hypothesis testing considering variability in both measurement axes simultaneously. Three exemplar datasets representing a cross-section of the types of continuous responses common in biomechanics were selected from the literature to serve as test cases to demonstrate the flexibility and limitations of the arc-length-based SPM approach.

## 2 Methods and Datasets

The proposed arc-length-based SPM methodology represents the integration of two analysis techniques: arc-length re-parameterization and statistical parametric mapping to perform statistical inferences on two or more sets of continuous experimental responses (Fig. 1). The proposed methodology can be applied to compare two or more sets of responses to identify if a statistically significant difference in responses exists and to identify regions that contribute to such differences. However, it is up to the analyst to interpret the meaning of these differences in the context of the data set and quantities, which may vary between data sets.

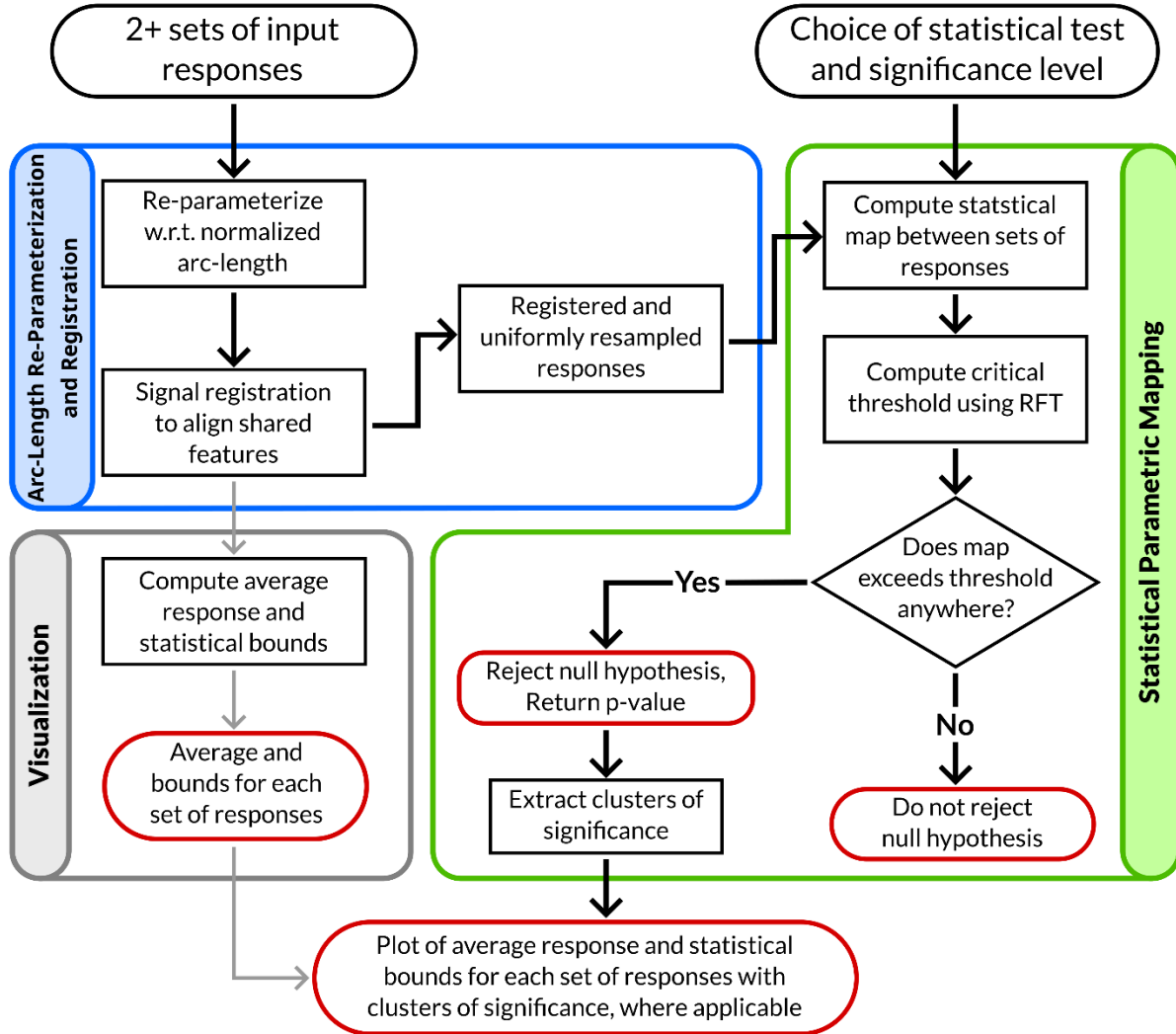


Figure 1: A flowchart of the arc-length-based statistical parametric mapping methodology. Experimental responses, a choice of test statistic, and a significance level were inputs. Hypothesis testing was performed between two or more sets of responses, and the null hypothesis was accepted or rejected globally. Additionally, the method identified the specific regions of the experimental responses that led to the rejection of the null hypothesis.

The first step of the proposed methodology was the re-parameterization and resampling of all experimental responses with respect to normalized arc-length. Normalized arc-length was computed for each response by first determining the arc-length of each response, then dividing by the total arc-length of the response, creating a common, monotonically increasing sampling variable bounded on  $[0,1]$ , regardless of the shape or monotonicity of the original responses. Once the normalized arc-length of each response was calculated, all responses were uniformly resampled to ensure that each response had the same number of points equally spaced with respect to arc-length [2].

Nonlinear, continuous signal registration [10] was applied to align features shared between all responses with respect to normalized arc-length. While arc-length re-parameterization generally located features shared between responses at similar normalized arc-lengths, intrinsic

variability between sets of responses could cause shared features to occur at different normalized arc-lengths. Signal registration only changed the arc-length assigned to shared features, not the coordinates of the feature in either the ordinate or abscissa, ensuring that statistical inferences accurately resolved variation in both measurement axes of two or more sets of responses. Signal registration was applied to all responses across all sets simultaneously to ensure that shared features were aligned for statistical testing in later steps. Not simultaneously aligning all signals could lead to instances where features shared between sets were not properly registered with respect to normalized arc-length. In contrast with other registration techniques suggested in the literature, the signal registration approach used in this methodology did not require a template or target to register to (such as an MRI atlas [9] or representative response [11]).

Monotone cubic splines were used as warping functions to perform signal registration, building on previous work by Hartlen and Cronin [2]. This registration method did not require the use or selection of a template or representative response, which often do not exist or could not be objectively selected for many types of biomechanical data. Using monotone cubic splines necessitated two user-defined parameters: the number of interior warping control points (or simply warping points) and a penalty factor. While the selection of these parameters is somewhat subjective, a rule of thumb proposed by Hartlen and Cronin was extended for the proposed arc-length-based SPM methodology. The number of warping points should be defined as one plus the number of prominent features (such as shared peaks, valleys, or changes in inflection) shared between the loading conditions, and the warping penalty factor should generally fall between  $10^{-3}$  and  $10^{-2}$ . The effect of under- and over-registration was discussed in detail in Hartlen & Cronin [2].

Re-parameterization and registration of all signals was an important prerequisite to performing SPM, which required a common sampling variable across all inputted signals (normalized arc-length, in this instance). The re-parameterized and registered signals were also used to calculate the average response and statistical bounds (also known as response corridors) for each set of signals. While the average response and bounds could not be used for statistical testing alone, these simplified representations of a larger set of individual signals were used to help visualize differences between data sets found during the SPM process. Average responses and corridors provided less visual clutter compared to many individual signals.

Statistical testing between sets of re-parameterized and registered responses was undertaken using SPM [7–9]. The methodology in the current study directly utilizes the well-established mathematical underpinnings of SPM, as presented in [5,9,12,13]. These established SPM approaches are coupled with arc-length re-parameterization to enable statistical testing on a wider range of data types than can be accommodated with current SPM approaches. The explanation of the arc-length-based SPM methodology is presented below with the goal of providing a conceptual understanding to the reader. Detailed coverage of the established mathematical foundations of SPM can be found in [5,9,12,13]. Like traditional hypothesis

testing, SPM provides the means to accept or reject a null hypothesis. However, in addition to a binary decision to accept or reject the null hypothesis, SPM provides a framework to visualize where differences occurred between sets of continuous responses.

The SPM process began with creating what is known as a statistical map, which effectively encoded both the variability of the inputted sets of responses and how much the data sets differ from one another. Statistical maps can be considered as analogous to computing a test statistic in traditional scalar statistics. The statistical map was generated by evaluating the chosen test statistic at each point of the common sampling variable, in this case, normalized arc-length. As arc-length re-parameterization and registration aligned both the ordinate and abscissa with respect to normalized arc-length, two multivariate test statistics were adopted for arc-length-based SPM. A two-sample Hotelling's T-squared ( $T^2$ ) test statistic was adopted to discriminate differences between the two sets of responses. Similarly, the multiple analysis of variance (MANOVA) test statistic was adopted for omnibus testing to determine if a difference existed between three or more sets of responses. These two test statistics can be considered multivariate extensions of the more commonly known Student's t and ANOVA tests, respectively.

Next, a critical threshold for the statistical map was computed to perform statistical inferences at a desired significance level,  $\alpha$ . The process was conceptually similar to how critical values for scalar statistical analysis are calculated, wherein an appropriate probability distribution is chosen and a critical value calculated based on the desired significance level, the number of samples, and properties such as mean and variance. SPM extended the concept for continuous responses using random field theory and assumed that variability could be modelled as Gaussian random fields (effectively a multi-dimensional extension of the well-known Gaussian distribution for scalar value statistics). The key benefit of SPM was a critical threshold that intrinsically captured the spatial correlation between neighbouring points in the statistical map. Regions of the statistical map that exceeded the critical threshold, known as clusters of significance or simply clusters, indicated differences between sets of responses would not occur due to random chance for a desired significance level.

The final step of the arc-length-based SPM methodology was interpretation. If a cluster of significance was found, the null hypothesis could be rejected, and it could be inferred that a statistically significant difference existed between the sets of responses over the region of the cluster. Two quantitative measures of significance are also produced during an SPM inference: a  $p$ -value for each cluster of significance and a  $p$ -value for the entire dataset. Cluster-level  $p$ -values describe the probability of a cluster of the same size and extent arising from chance. The dataset level  $p$ -value describes the probability of the same number and size of clusters arising anywhere in the dataset from chance. If only one cluster of significance were found, the cluster-level and dataset-level  $p$ -values would be the same. However, if more than one cluster were found, the dataset level  $p$ -value would generally be much lower, as there would be a much lower chance of two clusters arising from random chance [5].

Clusters of significance were visualized on top of the experimental data by highlighting the corresponding common regions of the average response across inputted test conditions. Because the clusters were defined with respect to normalized arc-length and variation in both the ordinate and abscissa was considered simultaneous, the highlighted regions encompassed shared features that differed between sets, regardless of the exact position of the feature in either axis. When interpreting these results, it is important to note that the statistical inference does not indicate whether the significant difference between groups arises from a particular measurement axis. Rather, users should report the general location of the cluster and the associated significance level. Contextual information can be provided to describe the underlying physics that may be driving the difference, but it is important to note that such information is interpreted based on one's knowledge of the situation and not quantitative results from the arc-length-based methodology. For example, one might report:

"A cluster of significance ( $p = 0.021$ ) was found that encompassed the initial rise and first peak of the signals between  $p$  and  $q$  milliseconds. This cluster of significance likely corresponds to differences in stiffness between the test conditions."

## 2.1 Exemplar Datasets for Analysis

Three datasets from the literature were analyzed using the arc-length-based SPM methodology. A range of response types common in biomechanics was selected, including time-based kinematics, monotonically increasing responses that do not terminate at a common coordinate, and hysteretic load-unload data.

### 2.1.1 Dataset 1: Head Kinematics during Frontal Acceleration

Ewing and Thomas [14] measured the head kinematics of human volunteers during a range of frontal acceleration pulses and reported planar head kinematics collected at 2000 Hz. The x-axis was aligned with the direction of the acceleration pulse. Kinematic data was retrieved from the online database of the National Highway Traffic Safety Administration [15]. The arc-length-based SPM methodology was applied to determine if a difference in response between head acceleration in the x-direction could be inferred between 12 g and 14 g acceleration pulses (Fig. 2). It is important to note that while variability in both acceleration and time were considered when applying the arc-length-based SPM methodology, this did not mean that there was variability in the measured time used to define the time series data. Rather, the intent of applying the arc-length-based SPM methodology is to capture the temporal response of the volunteers across the two acceleration pulses, simultaneous with the magnitude of head acceleration.

To contrast the arc-length-based approach against contemporary techniques, statistical hypothesis testing was also undertaken by performing a scalar t-test on peak acceleration between 12 g and 14 g, as well as performing time-based SPM using the package "spm1d" [7]. Whereas the arc-length-based SPM approach utilized a multivariate Hotelling's  $T$ -squared test, time-based SPM testing was performed using a  $t$ -test considering variation in acceleration.

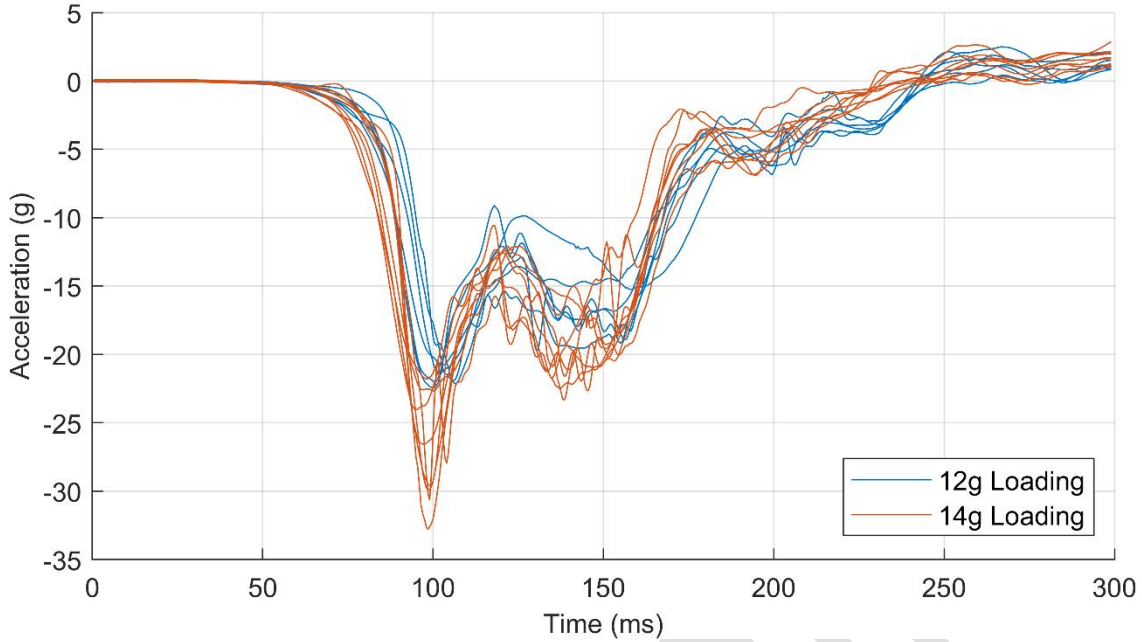


Figure 2: Head acceleration in the x-direction of human volunteers subject to a 12 g and 14 g deceleration pulse

No filtering or temporal alignment was performed on the raw head acceleration data. Signal registration for the arc-length-based SPM methodology used four interior warping control points and a penalty factor,  $\lambda$ , of  $10^{-2}$  to align the timing and magnitude of features shared between the 12 g and 14 g test conditions with respect to normalized arc-length. A significance level,  $\alpha$ , of 0.05 was selected for statistical testing.

#### 2.1.2 Dataset 2: Effect of Storage Condition on Porcine Brain Tissue

Singh et al. [16] studied the effect of storage conditions on the quasi-static compressive behaviour of porcine brain tissue. Test conditions included storing tissues at room temperature for 24 and 48 hours and storing tissues frozen for 24 and 48 hours before thawing to room temperature for testing. Fresh tissue was used as a control (Fig. 3). Compressive testing was performed at room temperature, with tests terminating at the onset of specimen damage (denoted by in drop load). Raw experimental data was provided by the authors.

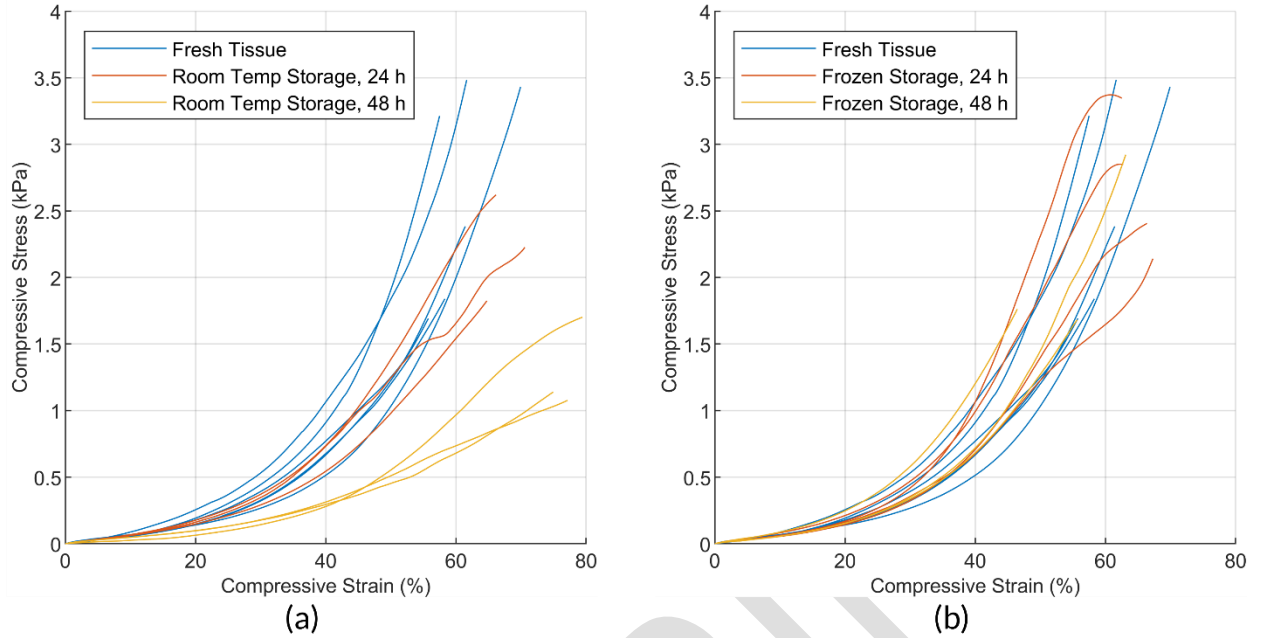


Figure 3: The compressive stress-strain response of porcine brain tissues that have been (a) stored at room temperature and (b) stored frozen and thawed before testing. Fresh tissue (blue) is considered baseline behaviour and is presented on both plots.

A challenge faced by Singh et al. when analyzing the mechanical behaviour of porcine brain tissue was that the compressive stress-strain curves terminated at various strain levels owing to experimental variation in the compressive failure strain of the material. To avoid the issue of differing failure strains, Singh et al. cropped all responses within a test condition to the lowest failure strain. Additionally, statistical testing was performed by fitting an Ogden hyperelastic model to each cropped response and then performing inferences on the fitted coefficients to detect differences between storage conditions. Using constitutive model parameters for statistical testing assumed the Ogden model accurately represented material behaviour and that the fitted constitutive parameters were valid and representative metrics for statistical testing. Further, contemporary SPM approaches could not be applied as all responses did not terminate at the same strain level, despite responses accurately reflecting variability in strain to failure.

The effects of storage time on tissue response at room temperature and frozen were investigated separately with the arc-length-based SPM methodology. For both room temperature and frozen tests, signal registration was undertaken with one warping control point and a penalty factor of  $10^{-2}$ . For both storage conditions, two levels of statistical testing were adopted. First, an omnibus MANOVA test was performed at a significance level of  $\alpha = 0.05$  to determine any differences between the three storage times (fresh, 24 hours, and 48 hours). If differences were detected with the MANOVA test, pairwise post hoc testing was undertaken with Hotelling's  $T$ -squared test to determine which conditions contributed to the differences. For all post hoc testing, the significance level was adjusted using the Šidák correction to  $\alpha = 0.017$  to control for family-wise error rates.

### 2.1.3 Dataset 3: Lateral Thoracic Impacts

Viano et al. [17] studied the thoracic response of post-mortem human subjects (PMHS) impacted with a 23 kg cylinder at velocities of 4.5, 6.7, and 9.4 m/s. The force-displacement responses were digitized from the published manuscript.

A two-level statistical testing approach was applied. First, an omnibus MANOVA test was applied to determine if significant differences could be resolved between the three impact speeds at a significance level of  $\alpha = 0.05$ . If a difference was detected, pairwise post hoc testing using a Hotelling's  $T$ -squared test was undertaken using a corrected significance level of  $\alpha = 0.017$  to control for family-wise error rates. Arc-length re-parameterization was critical to analyze the hysteresis observed in the force-displacement results. Signal registration with three interior warping points and a penalty factor of  $10^{-2}$  was used to align shared features with respect to normalized arc-length.

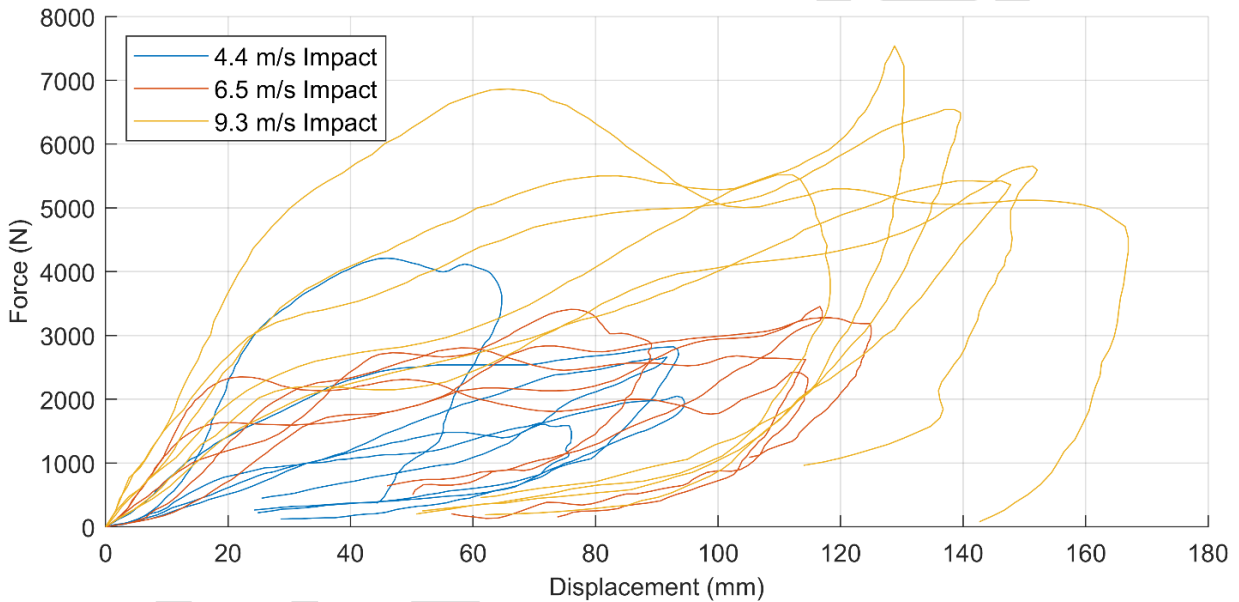


Figure 4: Force-displacement response of a human thorax subject to lateral pendulum impact at three velocities.

## 3 Results

### 3.1 Dataset 1: Head Kinematics during Frontal Acceleration

A two-sample scalar  $t$ -test on peak negative x-acceleration at a significance level of 0.05 detected a significant difference between the 12 g and 14 g sled pulses ( $p=0.004$ ) (Table 1).

Table 1: Peak negative acceleration statistics used to perform a two-sample t-test between 12g and 14g acceleration pulses

| Sled Acceleration | Mean     | St. Dev. | Repeats |
|-------------------|----------|----------|---------|
| 12 g              | -21.78 g | 0.5908 g | 6       |
| 14 g              | -27.44 g | 3.805 g  | 8       |

Arc-length-based SPM was carried out with a Hotelling's T-squared test ( $\alpha = 0.05$ ) to determine if there was a difference in response between 12 g and 14 g acceleration pulses considering variation in time and acceleration simultaneously. One cluster of significance was found (Fig. 5a), allowing the null hypothesis to be rejected ( $p_{set} < 0.001$ ) and indicating a significant difference between the two sets of responses. The cluster of significance encompassed much of the initial slope up to and including the peak negative acceleration of the responses between 80-110 ms, suggesting that the only significant difference between the sled accelerations stemmed from the initial period of head acceleration. Both differences in peak magnitude and timing likely drove this difference.

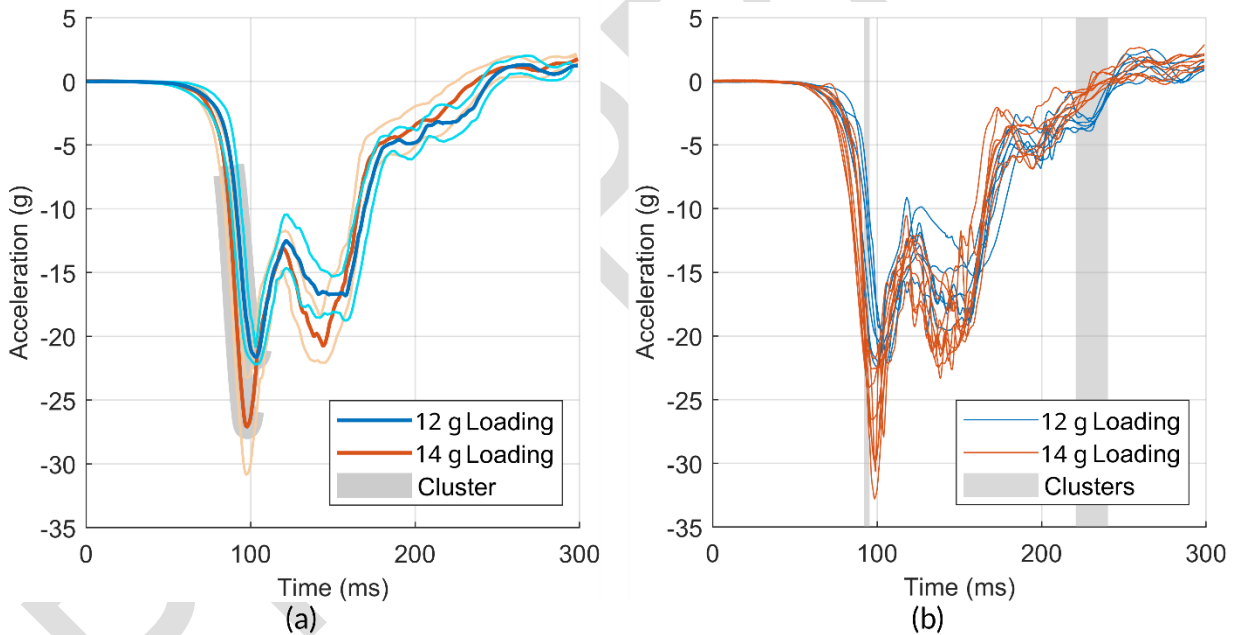


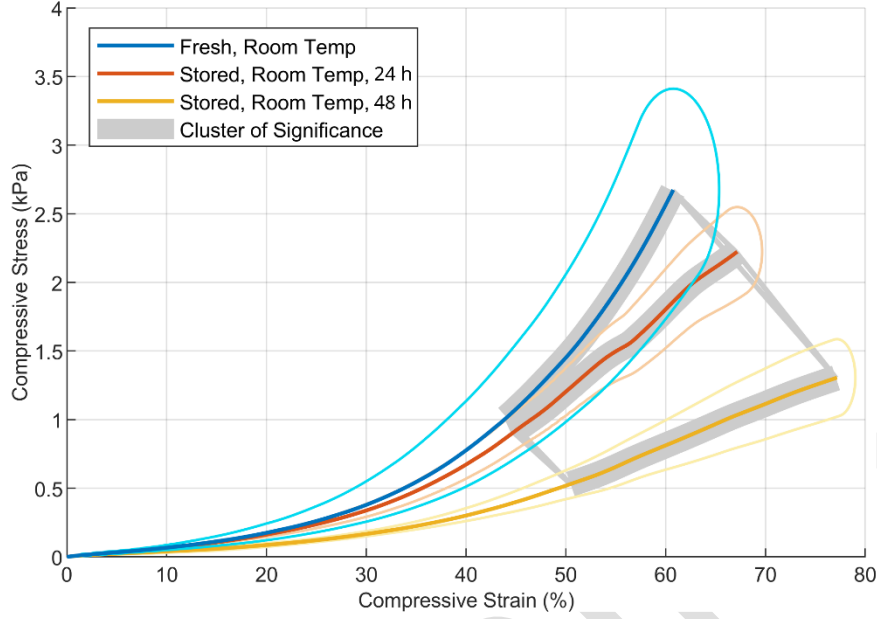
Figure 5: Results of the (a) arc-length-based SPM methodology in comparison to (b) time-based SPM ( $\alpha = 0.05$ ). Both approaches found significant differences between the responses (grey clusters), allowing the null hypothesis to be rejected. However, the position and extent of clusters varied between techniques. The lighter-coloured outlines in (a) represent  $\pm 1$  standard deviation corridors about the mean response.

Time-based SPM was conducted using a Student's t-test to determine whether head acceleration differed in the x-direction between 12 g and 14 g acceleration pulses. Variability in the timing of features could not be considered simultaneously. Two clusters of significance were found (Fig. 5b), allowing the null hypothesis to be rejected ( $p_{set} < 0.001$ ) and indicating a significant difference existed between the acceleration response of the two sets of responses.

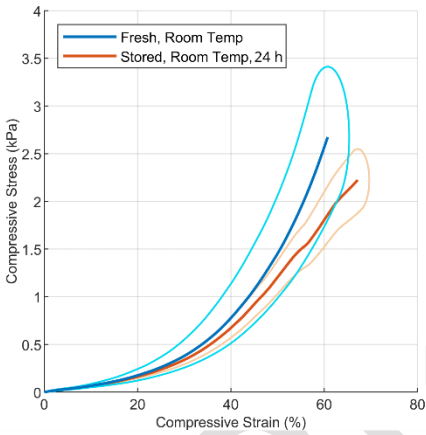
A larger cluster occurred between 220 and 240 ms, with a much smaller cluster encompassing a small part of the first slope between 92 and 95 ms.

### 3.2 Dataset 2: Effect of Storage Condition on Porcine Brain Tissue

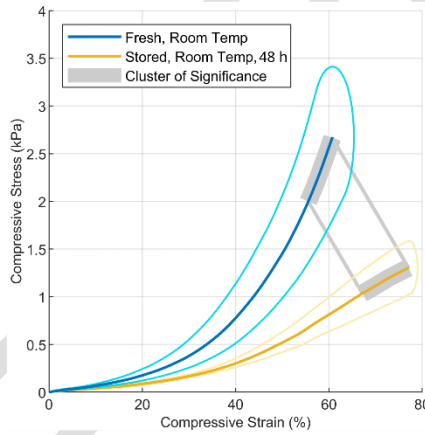
Statistical testing first considered the effect of room temperature storage on the compressive response of porcine brain tissue. An omnibus MANOVA test ( $\alpha = 0.05$ ) to determine if there was an effect on the mechanical response of porcine brain tissue with time stored at room temperature found a single cluster of significance that encompassed much of the large deformation behaviour above 43% strain (Fig. 6a), indicating that the sets of responses did not share the same mean.



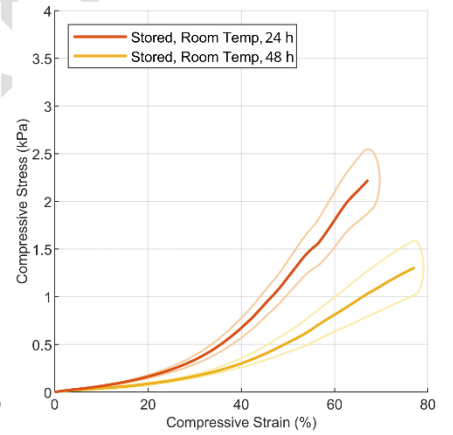
(a) Omnibus MANOVA between all storage conditions ( $\alpha = 0.05$ )



(b) Pairwise Hotelling's  $T^2$  between fresh and 24h at room temp ( $\alpha=0.017$ )



(c) Pairwise Hotelling's  $T^2$  between fresh and 48h at room temp ( $\alpha=0.017$ )



(d) Pairwise Hotelling's  $T^2$  between 24h and 48h at room temp ( $\alpha=0.017$ )

Figure 6: Results of hypothesis testing performed with the arc-length-based SPM methodology on porcine brain tissue tested after storage at room temperature. (a) MANOVA testing ( $\alpha = 0.05$ ) found a single cluster at large deformation response where test conditions are significantly different, indicating a significant difference between all three conditions. However, post hoc testing with a Hotelling's T-squared test ( $\alpha = 0.017$ ) only found a significant difference between tissue tested fresh versus after 48 hours at room temperature (c) despite the corridors of the 24 and 48-hour responses exhibiting visually distinct behaviour (d).

Pairwise post hoc testing with Hotelling's T-squared test ( $\alpha = 0.017$ ) detected differences between the tissues tested fresh and those tested after being stored for 48 hours at room temperature ( $p_{set} = 0.003$ , Fig. 6c), with a single cluster of significance encompassing strains above 56%. The identified cluster of significance did not have the same extent as that found during MANOVA testing. Despite a visual separation in the average response and statistical bounds between tissues stored for 24 and 48 hours at room temperature (Fig. 6d), the arc-length-based SPM methodology did not report a statistically significant difference between the two conditions.

Applying the arc-length-based SPM methodology to brain tissue that was tested after being frozen for 24 and 48 hours, an omnibus MANOVA test did not discriminate a difference in mean response between previously frozen tissue and fresh tissue with no clusters of significance detected ( $p_{set} > \alpha$ ). While post hoc testing was unnecessary, pairwise Hotelling's  $T$ -squared testing supported the conclusions of the omnibus test (Fig. 7b-d).

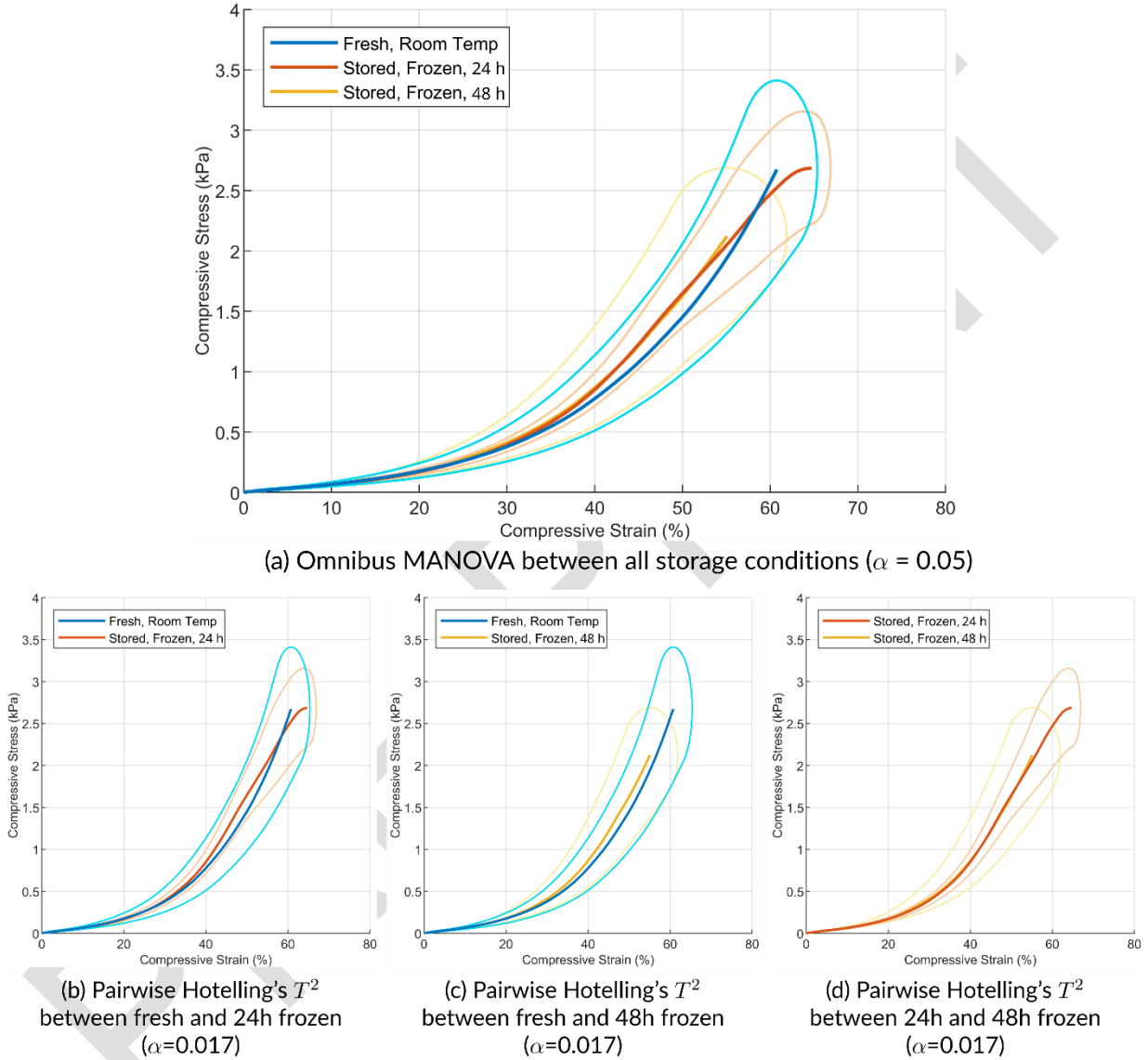
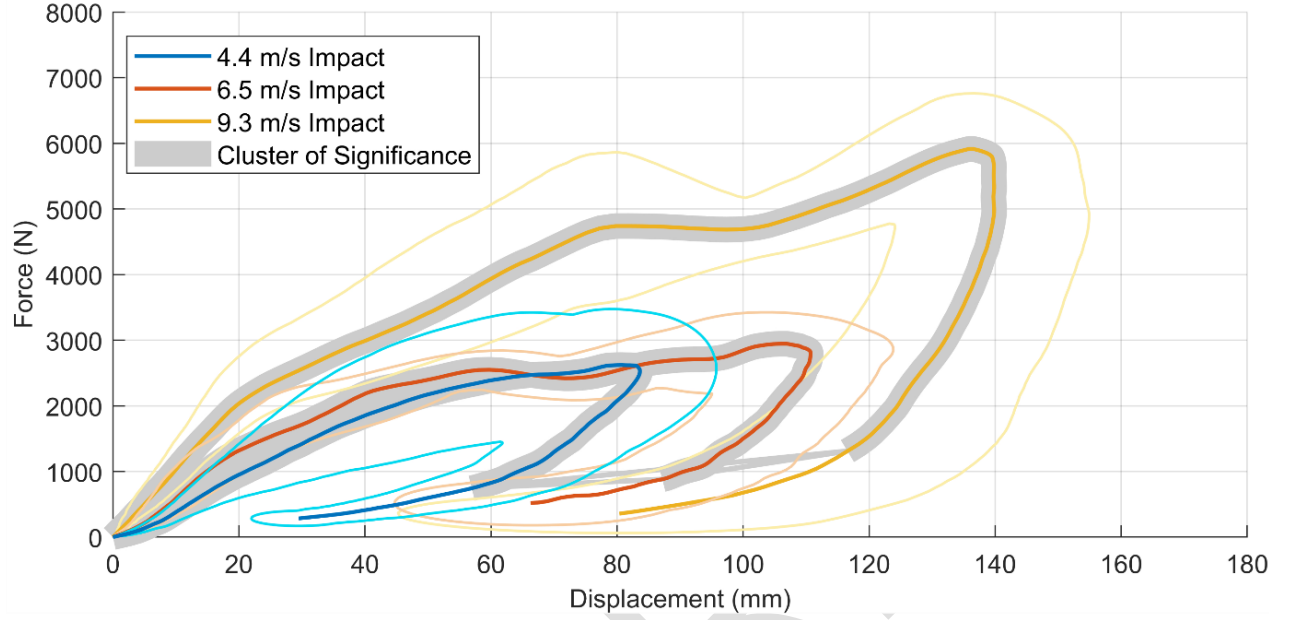


Figure 7: The arc-length-based SPM methodology applied to porcine brain tissue tested after being frozen for two lengths of time. (a) Omnibus MANOVA testing found no significant difference ( $\alpha = 0.05$ ) between the two storage times and fresh tissue. (b-d) While unnecessary in this case, post hoc Hotelling's  $T$ -squared tests also did not report any differences between pairs of test conditions

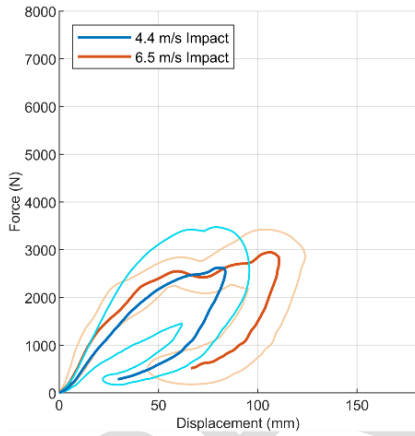
### 3.3 Dataset 3: Thoracic Response under Lateral Impact

An arc-length-based MANOVA test ( $\alpha=0.05$ ) found a single cluster of significance (Fig. 8a), indicating that mean responses between the three impact conditions were significantly

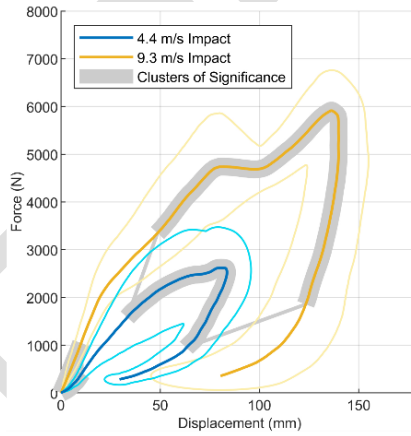
different ( $p_{set} < 0.001$ ). The cluster of significance encompassed much of the force-displacement responses except for the final portion of the unloading stage.



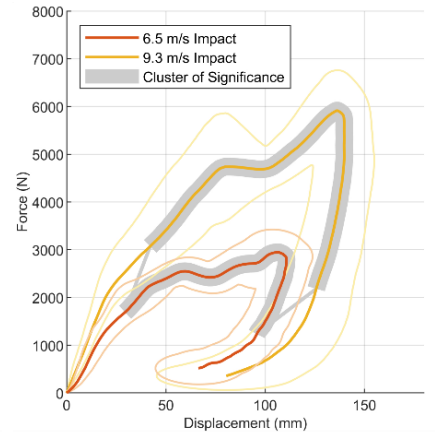
(a) Omnibus MANOVA between all impact velocities ( $\alpha = 0.05$ )



(b) Pairwise Hotelling's  $T^2$  between 4.4 and 6.5 m/s impacts ( $\alpha=0.017$ )



(c) Pairwise Hotelling's  $T^2$  between 4.4 and 9.3 m/s impacts ( $\alpha=0.017$ )



(d) Pairwise Hotelling's  $T^2$  between 6.5 and 9.3 m/s impacts ( $\alpha=0.017$ )

Figure 8: The arc-length-based SPM methodology applied to the lateral thoracic impact response of PMHS under different impact velocities. (a) An omnibus MANOVA test ( $\alpha = 0.05$ ) reported one large cluster spanning much of the responses between conditions. Pairwise post hoc Hotelling's  $T$ -squared tests ( $\alpha = 0.017$ ) found differences between the (c) 4.4 and 9.3 m/s impacts and (d) 6.5 and 9.3 m/s impacts, with clusters of significance indicating the differences in plateau, peak, and initial unload response. No statistical difference was found between the (b) 4.4 and 6.5 m/s impacts.

Pairwise post hoc testing with a Hotelling's  $T$ -squared test ( $\alpha = 0.017$ ) detected differences between the 4.4 m/s and 9.3 m/s impacts ( $p_{set} < 0.001$ ) and the 6.5 m/s and 9.3 m/s impacts ( $p_{set} < 0.001$ ). In both cases, prominent clusters were found that encompassed much of the plateau region, peak, and initial unload response of the force-displacement responses (Fig. 8c-

d). No statistical difference in response was found for the thoracic responses at impact velocities of 4.4 m/s and 6.5 m/s.

## 4 Discussion

Discussion of the arc-length-based SPM methodology is divided into three sections, beginning with a discussion related to implementing the arc-length-based SPM approach and things to consider when interpreting the results of statistical testing. Next, the application of the arc-length-based SPM approach to the three datasets is discussed. Finally, limitations of the arc-length-based approach are presented.

### 4.1 Implementation and Use

When performing arc-length re-parameterization on a set of responses, an important assumption was that shared features occurred at approximately the same normalized arc-length between responses in that set. However, intrinsic variability within a population can cause features to occur at different normalized arc-lengths. The misalignment of shared features with respect to normalized arc-length would prevent effective usage of statistical parametric mapping. Signal registration provided the means to adjust the normalized arc-length assigned to a feature without changing the coordinates of the feature with respect to either measurement axis. As the signal registration technique used in this work aligned signals to each other without a template or reference, all signals across all test conditions must be registered simultaneously to ensure that features are aligned.

While arc-length re-parameterization and signal registration enabled the use of SPM across the wide range of data types encountered in biomechanical data, the signals must reflect the same physics or have the same general shape. For example, comparing the mechanical responses of tissues tested to failure would generally involve signals that terminate at different strengths and deformations, reflecting variability in the underlying materials. However, if time-varying oscillatory signals were terminated at different times such that one or more major oscillations were removed, the same physics would not be reflected across all responses, affecting registration and warping outcomes. Therefore, the user must appropriately understand and curate the input signals to ensure consistent and objective results.

There has been published literature using signal registration in concert with SPM, but these attempts have been limited to time-series data where features are registered with respect to time [18]. The downside of that approach was that by registering features temporally, the underlying shape and context of the responses were changed, making it more difficult to resolve differences between sets of responses that might arise from temporal features such as phase lead or lag. While the authors did propose a two-level statistical testing scheme to account for variation in the timing of features by performing statistical testing on the amount of warping required to register features, the approach created a degree of abstraction between the statistical analysis and the physical context of the data. In contrast, the arc-length-based

SPM methodology captured variability in both axes simultaneously without changing the coordinates and shape of features with respect to either axis.

It is important to note that the selection of user-defined registration parameters could impact the inferences performed with arc-length-based SPM. Without sufficient registration (stemming from too few registration points or too high a penalty factor), features will not be properly aligned with respect to arc-length and SPM could fail to find statistically significant differences or find clusters of significance that are smaller than expected. Generally speaking, increasing the number of warping points can lead to over-registration, producing many smaller clusters rather than one larger cluster as less prominent features are registered together. A rule of thumb proposed previously in this paper for the appropriate selection of registration parameters is recommended, and one should refer to Hartlen & Cronin [2] for strategies to identify the appropriate registration parameters.

When performing statistical tests using SPM, one important consideration is that the null hypothesis differs from traditional scalar value statistical tests. Consider a statistical inference on two sets of stress-strain responses. Using single-value scalar metrics, the null hypothesis might be “there is no difference in peak stress between the two sets of responses.” The necessity of choosing a scalar metric results in a “directed” hypothesis that can introduce bias into the statistical analysis. In contrast, an SPM-based approach would have a null hypothesis of “there is no difference in stress between the two sets of responses.” This hypothesis is not dependent on a specific metric or response feature, which leads to SPM-based analysis being referred to as a “non-directed” method [4]. As a result, there is a reduced chance of bias as continuous responses do not need to be reduced to a single scalar value to discriminate differences between responses. It also indicates that any difference found between two sets of responses using SPM, regardless of the size or position of the difference in either axis, would prompt the rejection of the null hypothesis.

In addition to providing a framework to accept or reject a null hypothesis between sets of responses, the arc-length-based SPM methodology produced clusters of significance that highlighted the specific areas that lead to significant differences between sets of responses. While more qualitative than binary hypothesis testing, observing the specific features where sets of responses differ can provide an informed user with important information about where and how variability occurs.

## 4.2 Results of the Arc-Length-Based SPM Approach

The head kinematics dataset (dataset 1) provided the opportunity to compare arc-length-based SPM to published time-based SPM approaches, as all responses were uniformly sampled and terminated at the same time. Furthermore, this dataset displayed remarkably low variability in the timing of signal features. All three statistical methods (scalar value  $t$ -tests, time-based SPM, and arc-length-based SPM) suggested that the null hypothesis could be rejected, providing confidence in the validity of the arc-length-based approach. However, the differences between

clusters of significance found between the time-based and arc-length-based SPM approaches were illustrative of the advantages provided by the arc-length-based methodology.

The largest difference between the two approaches occurred around the initial rise and peak negative acceleration. While time-based SPM identified a very small cluster of significance in this region, the arc-length-based SPM approach identified a large cluster encompassing the entire rise and peak acceleration, owing to the ability of arc-length-based SPM to capture variation in both timing and magnitude of features. Another difference between the approaches was in the rebound phase of the acceleration data around 170-240 ms. Time-based SPM identified a cluster between 220-240 ms. However, given the oscillatory nature of the rebound phase and the difference in timing between the two load conditions, this cluster of significance likely arose from the out-of-phase alignment of conditions between responses. Arc-length-based SPM did not find a cluster in this region as signal registration aligned the oscillatory behaviour between loading conditions with respect to arc-length. Once aligned, simultaneous testing timing and magnitude between 12 g and 14 g loading conditions using a Hotelling's *T*-square test did not find that differences in the conditions was significant. The outcome can be observed quantitatively by observing the overlap in mean and  $\pm 1$  standard deviation corridors in Fig. 5a. While not statistically rigorous, observations using corridors can often provide a useful visual for assessing differences between groups.

The Singh porcine brain tissue dataset (dataset 2) was selected for this study because the stress-strain response of each sample did not terminate at a common coordinate, making contemporary SPM techniques difficult to apply. The published average response and corridors were computed by truncating responses at a shared strain well before failure and computing a pointwise mean [16] and the published statistical analysis was performed on fitted functions rather than directly on experimental data. In contrast, the arc-length-based SPM methodology did not require any a priori assumptions of material behaviour and used uncropped stress-strain responses without a shared termination strain. However, one important drawback of the Singh dataset was the limited number of repeats per test condition. While there were a reasonable number of repeats for fresh tissue ( $n = 6$ ), each storage condition had a much lower number of repeats. Except for tissues tested after being frozen for 24 hours ( $n = 4$ ), each storage condition only had three repeats ( $n = 3$ ). One limitation of SPM shared with traditional statistical hypothesis testing is that small sample sizes will produce higher critical values or thresholds for significance due to the greater uncertainty in variability, potentially hampering the ability to draw inferences between sets.

Omnibus MANOVA testing for three test conditions (room temperature fresh, 24 hour and 48 hour storage) indicated that the three conditions did not share the same mean response, which made intuitive sense when observing the average and variability of the responses (Fig. 6a). Pairwise post hoc testing supported the MANOVA analysis, detecting differences between the response of fresh tissue and tissue stored at room temperature for 48 hours (Fig. 6c). However, there were two possibly unintuitive findings. First, the cluster of significance found using a

Hotelling's  $T$ -squared test did not encompass the same extent of the responses as the cluster detected using MANOVA. Second, the arc-length-based SPM methodology did not discern between room temperature storage for 24 and 48 hours, despite clear visual separation between responses (Fig. 6d). These discrepancies stemmed from the limited number of test repeats. While there were six repeats for the fresh tissue, there were only three for the 24 and 48 hour storage conditions. Without a larger number of repeats, the critical threshold for significance computed during the SPM process was high, limiting the statistical power of the test.

Another motivating factor for investigating the Singh et al. [16] dataset was to compare the arc-length-based SPM approach to the published statistical analysis based on fitted material model parameters. Singh et al. performed pairwise t-tests on the small deformation shear moduli of Ogden hyperelastic models fitted to each material response to compare each test condition against the response of fresh tissue. However, unlike the arc-length-based approach used here, Singh et al. could not infer a difference between fresh tissue and tissues stored under any storage condition. While the small number of repeats likely hampered the statistical tests performed by Singh et al., using initial shear moduli as the scalar parameter for hypothesis testing limited the types of differences between responses that could be resolved. Arc-length-based SPM did not require abstraction of the experimental data with analytical functions and parameters and could resolve differences between conditions by assessing the entire response simultaneously.

The lateral thoracic impact response dataset (Dataset 3) was selected to demonstrate how the arc-length-based SPM methodology could be applied to complex datasets, such as hysteretic force-displacement responses. Both MANOVA and post hoc Hotelling  $T$ -squared testing using the arc-length-based SPM approach demonstrated how clusters of significance were associated with the response features. However, it is important to note that the arc-length-based SPM approach did not provide information on why differences existed. For that, deeper investigation may be needed, such as the necropsy performed in the original study [17], where physiological factors such as rib fractures were used to explain differences between impact conditions.

The Viano dataset also provides an opportunity to study how an outlier signal affected the application of arc-length-based SPM with the post hoc Hotelling's  $T$ -squared test between the 4.4 and 6.5 m/s impact velocities (Fig. 8b). The 4.4 m/s condition had a prominent outlier that had a much higher peak force compared to the other responses (Fig. 4). While arc-length re-parameterization and signal registration helped minimize the effect of the identified outlier on the average response and statistical bounds, the added variability did not allow for clusters of significance to be identified, and hence, the null hypothesis was not rejected.

#### 4.3 Limitations and Considerations

It should be emphasized that while the arc-length-based SPM approach developed in this study is applicable for statistical hypothesis testing across a wide array of datasets, the technique

operated agnostic to the physical interpretation of the data being analyzed, which is also true of traditional scalar statistics. While the arc-length-based SPM approach can be applied to a wide range of data, assigning causality to a difference requires that the situations being compared share similar boundary conditions across sets of input responses. For example, in Dataset 3 of this paper, the same experimental setup was used across test conditions, with only the impactor speed changing, meaning impact velocities all share the same underlying boundary conditions. For Dataset 2, the same compressive loading condition was applied to all materials, with only the storage condition being varied.

One limitation of the arc-length SPM methodology was that a  $p$ -value could not be calculated if a significant difference was not found. This limitation arose due to the need for a cluster of significance to exist to calculate a  $p$ -value for both the cluster and the entire dataset. Generally, if a cluster of significance is not found, such a finding should be reported as not having found a significant difference and reporting  $p > \alpha$ . Another important consideration when interpreting  $p$ -values is that the dataset-level  $p$ -value will generally be lower than an individual cluster-level  $p$ -value, as noted previously, due to the dataset-level  $p$ -value calculation accounting for the number and extent of the clusters that are present. If discerning global differences between conditions, rather than resolving differences between specific features, then the dataset-level  $p$ -value is the more appropriate value to report.

A second limitation of arc-length-based SPM, and SPM as a method in general, is that post hoc testing is not as rigorous as the schemes developed for contemporary statistical testing [18]. While post hoc testing using the Šidák correction to control family-wise error rates was effective in this study, there are cases where post hoc tests could conflict with the results of a higher-level test. One example of this limitation would be the difference in the extent of clusters found for omnibus testing of porcine brain tissue stored at room temperature (Fig. 6a) and the cluster found between fresh tissue and tissues stored for 48 hours (Fig. 6c). The cluster found using a post hoc Hotelling  $T$ -squared test was smaller than that of the omnibus MANOVA test when a cluster of the same size would have been expected. In this case, the discrepancy arises from the limited number of test repeats per condition but highlights the effect of the simplistic nature of post hoc testing with SPM-based approaches.

As highlighted in Sections 3.2 and 4.2, SPM is generally more conservative than traditional scalar statistics, particularly for small sample sizes. Small sample sizes will generally produce higher critical thresholds, which can limit the resolving power of a statistical test. While Hotelling's  $T$ -squared and MANOVA tests are extensions of scalar tests suitable for small sample sizes (the Student's  $t$ -test and ANOVA), the ability of SPM to account for correlation between neighbouring points produces a naturally more conservative threshold, similar to how a Bonferroni correction is applied to deal with the multiple comparison problem. While some extensions to SPM have been presented that may provide more resolving power, such as threshold-free cluster enhancement [19], evaluation of these techniques within the context of biomechanics data is ongoing.

It is important to note that the statistical tests presented in this study are relatively simple compared to the current state-of-the-art statistical tests performed on scalar values, such as regression analysis, random effects analysis, and linear mixed effects models. By contrast, Hotelling's  $T$ -squared and MANOVA tests presented in this work are comparatively simple and can only detect differences in means between groups without accounting for interactions or random effects. While more advanced statistical techniques have been developed for SPM in neuroimaging, applying these techniques is often challenging owing to the complexity of extending probability distributions and techniques to multiple dimensions [5]. Nevertheless, the extension of current SPM approaches in biomechanics to include more sophisticated analysis approaches to biomechanical data will be investigated in future research.

An important consideration when applying the arc-length-based SPM methodology stems from the underlying assumption on which SPM is based. The mathematics underpinning SPM are based on the assumption that data is continuous and that variability can be described with Gaussian random fields. In the case of the first point, most biomechanical signals can generally be considered continuous so long as those signals do not include discontinuities associated with the underlying physics. For example, the fracture during material testing would create a discontinuity in a force-displacement response where the behaviour before fracture would not be considered continuous with post-fracture behaviour. However, a rib fracture during PMHS testing would not create a discontinuity, as a rib fracture was an expected outcome in the context of the larger system. Researchers should be careful to pre-process data to ensure that the underlying physics encompassed by inputted signals is continuous.

Checking that variability is appropriately described with Gaussian random fields is a more challenging endeavour, as these fields encompass both spatial variation (such as noise within a signal) and variability between signals. Unlike traditional scalar statistics, there are limited methods to test whether the variability of a family of responses is appropriately described with Gaussian random fields. However, as a rule of thumb, the noise in a set of responses should be consistent across all responses [20]. Filtering or smoothing of responses as a pre-processing step has been shown to impact how accurately a Gaussian random field describes the variability of a set of signals, with more aggressive smoothing generally increasing the accuracy of the assumption at the cost of potentially impacting the resolution of the signals [17]. Another factor that can impact the assumption of Gaussian random fields is if signals are highly trended, such as signals like force-displacement responses that originate from a common coordinate. In these cases, variability is not generally uniform across the entire length of the response. In some cases, SPM can result in overly conservative inferences. While nonparametric methods have been proposed for SPM to compute a critical threshold without assuming a Gaussian response, such as permutation testing [8,21], a relatively large number of repeats per test condition is needed. Nonparametric methods were explored in this work but did not yield an improvement compared to standard parametric SPM, often owing to the relatively low number of repeats per condition.

## 5 Conclusions

An arc-length-based statistical parametric mapping (SPM) methodology was developed, integrating arc-length re-parameterization and signal registration techniques to extend the capabilities of contemporary SPM techniques to multivariate continuous responses common in biomechanics such as those that do not terminate at a common coordinate, time-series data, highly oscillatory or variable datasets, and hysteretic load-unload responses. The arc-length-based SPM method allowed statistical hypothesis testing on continuous responses without extracting single-value scalar metrics by simultaneously capturing variability in both measurement axes. In addition to providing a binary indicator of whether two or more sets of responses were significantly different, the arc-length-based SPM methodology provided important contextual information in the form of the specific regions of a set of responses that drove differences between sets.

The arc-length-based SPM methodology was applied to three literature datasets to demonstrate that a range of response types can be used with this approach. In each case, the arc-length-based SPM methodology produced hypothesis tests that largely agreed with the published literature analysis without reducing responses to a single scalar value statistic and providing important information on the specific feature that drove differences. The SPM analysis also highlighted a limitation with low numbers of repeat tests per condition. In one case, the published literature dataset did not have a sufficient number of repeats per test condition to infer differences between datasets despite visually distinct responses. To that end, it is recommended that researchers developing experimental test programs, especially in areas where a large degree of variability is expected, prioritize the number of repeats per test condition over the total number of test conditions in a study wherever possible.

## Acknowledgements

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada. This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

## Data Availability Statement

Datasets will be made available upon request.

## CRediT Statement

D.C. Hartlen: Conceptualization, Methodology, Software, Writing – Original Draft, Writing – Review & Editing. D.S. Cronin: Conceptualization, Supervision, Funding acquisition, Writing – Review & Editing.

## Conflict of Interest

The authors declare that research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- [1] K.U. Schmitt, P.F. Niederer, D.S. Cronin, M.H. Muser, F. Walz, Trauma Biomechanics: an introduction to injury biomechanics, *Trauma Biomechanics: An Introduction to Injury Biomechanics* 1 (2014) 1–243. <https://doi.org/10.1007/978-3-642-53920-6>.
- [2] D.C. Hartlen, D.S. Cronin, Arc-Length Re-Parametrization and Signal Registration to Determine a Characteristic Average and Statistical Response Corridors of Biomechanical Data, *Frontiers in Bioengineering and Biotechnology* 10 (2022) 357. <https://doi.org/10.3389/fbioe.2022.843148>.
- [3] D. Lessley, J. Crandall, G. Shaw, R. Kent, J. Funk, A normalization technique for developing corridors from individual subject responses, in: *SAE Technical Papers*, SAE International, 2004. <https://doi.org/10.4271/2004-01-0288>.
- [4] T.C. Pataky, M.A. Robinson, J. Vanrenterghem, Vector field statistical analysis of kinematic and force trajectories, *Journal of Biomechanics* 46 (2013) 2394–2401. <https://doi.org/10.1016/j.jbiomech.2013.07.031>.
- [5] K. Friston, J. Ashburner, S. Kiebel, T.E. Nichols, W. Penny, eds., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, 1st ed., Academic Press, 2006.
- [6] K.J. Friston, A.P. Holmes, K.J. Worsley, J.-P. Poline, C.D. Frith, R.S.J. Frackowiak, Statistical parametric maps in functional imaging: A general linear approach, *Human Brain Mapping* 2 (1994) 189–210. <https://doi.org/10.1002/hbm.460020402>.
- [7] T.C. Pataky, Generalized n-dimensional biomechanical field analysis using statistical parametric mapping, *Journal of Biomechanics* 43 (2010) 1976–1982. <https://doi.org/10.1016/j.jbiomech.2010.03.008>.
- [8] R.S.J. Frackowiak, K.J. Friston, C.D. Frith, R.J. Dolan, C.J. Price, S. Zeki, J.T. Ashburner, W.D. Penny, eds., *Human Brain Function*, 2nd ed., Academic Press, 2004. <https://www.sciencedirect.com/book/9780122648410/human-brain-function>.
- [9] J.E. Taylor, K.J. Worsley, Random fields of multivariate test statistics, with applications to shape analysis, *The Annals of Statistics* 36 (2008) 1–27. <https://doi.org/10.1214/0090536070000000406>.
- [10] J. Ramsay, Curve registration, in: F. Ferraty, Y. Romain (Eds.), *The Oxford Handbook of Functional Data Analysis*, Oxford University Press, 2018. <https://doi.org/10.1093/oxfordhb/9780199568444.013.9>.
- [11] T.C. Pataky, M. Yagi, N. Ichihashi, P.G. Cox, Landmark-free, parametric hypothesis tests regarding two-dimensional contour shapes using coherent point drift registration and statistical parametric mapping, *PeerJ Comput. Sci.* 7 (2021) e542. <https://doi.org/10.7717/peerj-cs.542>.
- [12] R.J. Adler, J.E. Taylor, *Random Fields and Geometry*, 1st ed., Springer-Verlag, New York, 2007. <https://link.springer.com/book/10.1007/978-0-387-48116-6> (accessed February 9, 2023).

- [13] M. Brett, W. Penny, S. Keibel, An Introduction to Random Field Theory, in: J. Ashburner, K. Friston, W. Penny (Eds.), *Human Brain Function*, Academic Press, Amsterdam, 2003.
- [14] C.L. Ewing, D.J. Thomas, *Human Head and Neck Response to Impact Acceleration.*, Navel Aerospace Medical Research Lab, Pensacola, FL, 1972.
- [15] National Highway Traffic Safety Administration, Biomechanics Test Database, (2017). <https://www.nhtsa.gov/research-data/databases-and-software> (accessed December 10, 2021).
- [16] D. Singh, S. Boakye-Yiadom, D.S. Cronin, Comparison of porcine brain mechanical properties to potential tissue simulant materials in quasi-static and sinusoidal compression, *Journal of Biomechanics* 92 (2019) 84–91. <https://doi.org/10.1016/j.jbiomech.2019.05.033>.
- [17] D.C. Viano, I.V. Lau, C. Asbury, A.I. King, P. Begeman, Biomechanics of the human chest, abdomen, and pelvis in lateral impact, *Accident Analysis & Prevention* 21 (1989) 553–574. [https://doi.org/10.1016/0001-4575\(89\)90070-5](https://doi.org/10.1016/0001-4575(89)90070-5).
- [18] T.C. Pataky, M.A. Robinson, J. Vanrenterghem, C.J.W. Donnelly, Simultaneously assessing amplitude and temporal effects in biomechanical trajectories using nonlinear registration and statistical nonparametric mapping, *Journal of Biomechanics* 136 (2022) 111049. <https://doi.org/10.1016/j.jbiomech.2022.111049>.
- [19] S.M. Smith, T.E. Nichols, Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference, *NeuroImage* 44 (2009) 83–98. <https://doi.org/10.1016/j.neuroimage.2008.03.061>.
- [20] S.J. Kiebel, J.-B. Poline, K.J. Friston, A.P. Holmes, K.J. Worsley, Robust Smoothness Estimation in Statistical Parametric Maps Using Standardized Residuals from the General Linear Model, *NeuroImage* 10 (1999) 756–766. <https://doi.org/10.1006/nimg.1999.0508>.
- [21] T.E. Nichols, A.P. Holmes, Nonparametric permutation tests for functional neuroimaging: A primer with examples, *Human Brain Mapping* 15 (2002) 1–25. <https://doi.org/10.1002/HBM.1058>.