Latent Motion: Uncovering Biomechanical Patterns with Sparse Autoencoders

Abdur-Raheem Kalam | ES 224 Final Project

24/04/2025

Abstract

This study explores the application of sparse autoencoders (SAEs) as interpretability tools for gait kinematics, leveraging the comprehensive Gutenberg Gait Database of healthy individuals. I implemented and compared three distinct autoencoder architectures: a baseline sparse autoencoder, an SAE with disentanglement loss, and a sparse variational autoencoder (SVAE). These models were trained on ground reaction force (GRF) and center of pressure (COP) data processed using wavelet decomposition to learn meaningful latent representations of gait patterns. My analysis reveals that the learned features capture biomechanically significant aspects of human gait, with clear correlations to physical attributes such as age, height, body weight, and walking speed. The latent space organization demonstrates a continuous manifold structure that effectively encodes gait dynamics, with walking speed emerging as a particularly influential factor. Clustering analysis of the latent representations identified distinct groupings that may correspond to different gait styles or phases. While all three models successfully learned interpretable features, I observed potential redundancy in the feature reconstructions that warrants further investigation. This research serves as a proof of concept for developing a shared latent space and comprehensive dictionary of gait features, which could significantly enhance biomechanical analysis, aid in the identification of gait abnormalities, and support clinical decision-making.

1 Introduction

Human gait analysis represents a fundamental area of biomechanical research with significant implications for clinical assessment, rehabilitation, and sports performance. The complex, multi-dimensional nature of gait patterns encapsulates a wealth of possible information about an indiidual's physical condition, age-related changes, potential pathologies, and overall functional status. However, the interpretation of high-dimensional gait data presents substantial challenges for researchers and clinicians alike, often requiring specialised expertise and considerable time investment.

Traditional approaches to gait analysis have relied heavily on direct measurement and statistical analysis of specific parameters such as step length, cadence, joint angles and ground reaction forces. While these methods provide valuable insights, they frequently reduce the rich complexity of gait dynamics to a limited set of predefined metrics, potentially overlooking subtle patterns and interactions that may have significant biomechanical or clinical relevance. This reduction in dimensionality, while making analysis more tractable, can limit understanding of the full spectrum of gait characteristics and their relationships to physical attributes and pathological conditions.

The emergence of machine learning techniques offers promising avenues for addressing these limitations by automatically extracting meaningful patterns from high-dimensional gait data. Supervised learning approaches have demonstrated success in classifying gait patterns associated with specific conditions or demographics. However, these methods typically require labeled data and may not provide interpretable insights into the underlying biomechanical features driving said classifications. In contrast, unsupervised learning approaches, particularly autoencoders, offer the potential to learn rich representations of gait patterns without requiring labeled data, while potentially revealing interpretable features that correspond to meaningful biomechanical characteristics.

Sparse autoencoders (SAEs) represent a particularly promising class of unsupervised learning models for gait analysis. By imposing sparsity constraints on the learned representations, these models are encouraged to discover features that capture essential aspects of the input data. This sparsity property aligns well with the biomechanical intuition that gait patterns, while complex, may be decomposable into a set of fundamental movement primitives or characteristic patterns. Furthermore, the interpretability of sparse representations makes them valuable in clinical settings, where understanding the basis for model predictions is crucial for informing treatment decisions.

In this study, I explored the application of sparse autoencoders as interpretability tools for gait kinematics, leveraging the comprehensive Gutenberg Gait Database of healthy individuals. I implement and comparse three distinct autoencoder architectures:

1. A baseline sparse autoencoder (SAE) with L1 regularisation and KL divergence sparsity penalties

2. An SAE with additional disentanglement loss to encourage independence between learned features

3. A sparse variational autoencoder (SVAE) that combines sparsity constraints with a probabilistic latent space

These models were trained on ground reaction force (GRF) and center of pressure (COP) data from the Gutenberg Gait Database, which provides a rich collection of gait measurements from 350 healthy individuals across various ages, heights, weights and walking speeds. By analyzing the learned representations and their relationships to subject metadata, I aim to uncover interpretable biomechanical patterns.

Key contributions include:

1. I demonstrate the effectiveness of sparse autoencoders in learning interpretable features from high-dimensionalgait data without explicit supervision

2. I provide a comparative analysis of different autoencoder architectures for gait feature learning, highlighting respective strengths and limitations

3. I establish proof of concept for a shared latent space that could serve as the basis for a comprehensive dictionary of gait features with additional data

4. I identify weak correlations between learned features and physical attributes without presence of supervised signal, suggesting potential applications in personalised biomechanical analysis

2 Literature Review

2.1 Biomechanical Modelling and Gait Analysis

Gait Analysis has been a corenerstone of biomechanical research for decades, providing valuable insights into human locomotion patternsin both healthy individuals and those with pathological conditions. Traditional approaches to gait analysis have relied on a combination of observational assessments and quantitative measurements of spatiotemporal parameters, kinetics, and kinematics (Baker, 2006). These methods typically involve specialized equipment such as force plates, motion capture systems, and electromyography to capture the complex dynamics of human movement during walking or running (Whittle, 2014).

The development of standardized gait databases has significantly advanced the field by providing researchers with comprehensive datasets for analysis and model development. The AddBiomechanics dataset, one of the early comprehensive collections, offered researchers access to standardized biomechanical measurements across diverse populations (Horst et al., 2019). Building upon this foundation, the GaitRec dataset expanded the available data with a particular focus on pathological gait patterns, including measurements from individuals with various neurological and musculoskeletal conditions (Horsak et al., 2020).

The Gutenberg Gait Database, introduced by Schöllhorn et al. (2021), represents a significant advancement in this domain, providing what is currently the world's largest collection of gait analysis data from healthy individuals. This database contains ground reaction force (GRF) and center of pressure (COP) data from 350 healthy subjects, complementing the GaitRec dataset by increasing the number of healthy control subjects from 211 to 561. The Gutenberg Gait Database follows standardized protocols and file formats compatible with the GaitRec dataset, facilitating integrated analysis across both resources. The comprehensive nature of this database, with its rich metadata including age, height, weight, and walking speed information, makes it an ideal resource for developing and validating advanced analytical methods for gait analysis (Schöllhorn et al., 2021). I draw on this database for my methodology.

2.2 Machine Learning in Biomechanics

The application of machine learning techniques to biomechanical data has grown substantially in recent years, offering new approaches to analyze and interpret the complex, high-dimensional data characteristic of human movement (Halilaj et al., 2018). Supervised learning approaches have demonstrated considerable success in classifying gait patterns associated with specific conditions or demographics. For instance, Begg et al. (2005) employed support vector machines to distinguish between the gait patterns of young and elderly individuals, while Alaqtash et al. (2011) used neural networks to classify pathological gait patterns in individuals with neuromuscular disorders.

Unsupervised learning approaches have gained increasing attention for their ability to discover latent patterns in gait data without requiring labeled examples. Clustering techniques such as k-means and hierarchical clustering have been applied to identify natural groupings in gait patterns (Toro et al., 2007), while dimensionality reduction methods like principal component analysis (PCA) have been used to identify the primary modes of variation in gait data (Deluzio & Astephen, 2007). These approaches have provided valuable insights into the underlying structure of gait patterns but often lack the representational capacity to capture the full complexity of biomechanical data.

Deep learning methods, particularly autoencoders, have emerged as powerful tools for learning rich representations of biomechanical data. Horst et al. (2019) demonstrated the effectiveness of convolutional autoencoders in learning meaningful features from raw gait data, while Dindorf et al. (2022) explored the use of variational autoencoders for generating synthetic biomechanical data. These approaches leverage the representational power of deep neural networks to capture complex patterns in gait data, potentially revealing insights that might be missed by traditional analytical methods.

2.3 Sparse Autoencoders

Sparse autoencoders represent a specialized class of neural network architectures designed to learn compact, interpretable representations of high-dimensional data (Ng, 2011). Unlike standard autoencoders, which simply aim to reconstruct their inputs from a lower-dimensional latent space, sparse autoencoders incorporate additional constraints that encourage the learned representations to be sparse—that is, to have relatively few active neurons for any given input. This sparsity property is motivated by observations from neuroscience suggesting that biological neural systems often employ sparse coding strategies, with relatively few neurons firing in response to any particular stimulus (Olshausen & Field, 1996).

Mathematically, a sparse autoencoder consists of an encoder function that maps input data x to a latent representation $h = f(x)$, and a decoder function that reconstructs the input from this representation $\hat{x} = g(h)$. The training objective typically combines a reconstruction loss term with a sparsity penalty:

$$L(x, \hat{x}) = ||x - \hat{x}||^2 + \lambda \cdot \Omega(h)$$

where $\Omega(h)$ is a sparsity-inducing regularizer such as the L1 norm ($||h||_1$) or the Kullback-Leibler divergence between the average activation of each hidden unity and a target sparsity level ρ (Ng, 2011). This formulation encourages the model to learn a representation that both accurately reconstructs the input data and satisfies the sparsity constraint.

The application of sparse autoencoders in biomechanics has been relatively limited compared to other domains such as computer vision and natural language processing. However, their potential for learning interpretable features makes them particularly well-suited for biomechanical analysis, where understanding the underlying patterns is often as important as predictive performance. Dindorf et al. (2021) demonstrated the effectiveness of sparse coding techniques in identifying interpretable movement primitives from biomechanical data, suggesting that sparse autoencoders could provide valuable insights into the fundamental components of human gait.

2.3.1 Interpretability in Biomechanical Models

Interpretability in biomechanical models is of paramount importamce, particularly in clinical applications where understanding the basis for model predictions can directly impact treatment decisions (Halilah et al., 2018). Traditional biomechanical models, such as inverse dynamics approaches and musculoskeletal simulations, offer inherent interpretability through their basis in physical principles and anatomical structures (Delp et al., 2007). However, these models often require simplifying and may not fully capture the complexity of human movement patterns.

Machine learning approaches, while potentially more powerful in capturing complex patterns, often sacrifice interpretability for predictive performance. This "black box" nature can limit their utility in clinical settings, where clinicians need to understand not just what a model predicts but why it makes that prediction (Doshi-Velez & Kim, 2017). Various strategies have been proposed to enhance the interpretability of machine learning models in biomechanics, including feature importance analysis, partial dependence plots, and attention mechanisms (Horst et al., 2019).

Sparse and disentangled representations offer a promising middle ground, potentially combining the predictive power of deep learning with a level of interpretability that makes them suitable for clinical applications. By learning features that are both sparse (activating selectively for specific patterns) and disentangled (corresponding to independent factors of variation), these approaches could provide insights into the fundamental components of gait patterns and their relationships to physical attributes and pathological conditions.

Despite these advances, significant challenges remain in developing truly interpretable models for biomechanical analysis. The complex, multi-dimensional nature of human movement, the variability across individuals, and the often subtle distinctions between normal and pathological patterns all contribute to the difficulty of this task. Furthermore, the evaluation of interpretability itself is challenging, often relying on subjective assessments by domain experts rather than quantitative metrics. My research aims to address these challenges by exploring the potential of sparse autoencoders as interpretability tools for gait kinematics.

3 Methodology

3.1 Dataset Description

This study utilizes the Gutenberg Gait Database (Schöllhorn et al., 2021), a comprehensive collection of GRF and COP data from350 healthy individuals. The Gutenberg Gait Database represents the world's largest collection of gait analysis data from healthy subjects and was designed to complement the GaitRec dataset, increasing the total number of healthy control subjects from 211 to 561. This extensive database provides robust foundation for machine learning approaches. It contains several key data components:

1. Ground Reaction Forces: Three dimensional force measurements (vertical, anterior-posterior, and medial-lateral) captured during walking, providing insights into the forces exerted between the foot and the ground during gait.
2. Center of Pressure: Two-dimensional measurements (anterior-posterior and medial-lateral) indicating the point of application of the ground reaction force vector
3. Metadata: Comprehensive subject information including age, height, body weight, sex and walking speed, enabling analysis of correlations between learned features and physical attributes

The data is provided in two formats: RAW (raw measurement data) and PRO (processed data with normalized stance phases). For this analysis, I primarily utilized PRO data, which consists of stnce phases normalized to 101 data points, facilitating comparison across subjects and trials. The database follows standardized protocols and file formats (.csv) compatible with GaitRec, ensuring consistency and interoperability.

3.1.1 Data Preprocessing

To prepare the gait signals for input to the autoencoder models, I implemented a preprocessing pipeline that leverages wavelet decomposition. This approach was chosen for its ability to capture both time and frequency information in the gait signals, providing a rich representation of the underlying biomechanical patterns to extract meaningful features from.

The preprocessing steps were as follows:

1. Data Loading: I loaded the PRO data from the Gutenberge Gait Database, focussing on the normalized stance phases for both left and right feet
2. Wavelet Decomposition: Each stance phase signal was decomposed using the Daubechies 5 (db5) wavelet at decomposition level 4. This wavelet family was selected for its effectiveness in capturing smooth yet detailed characteristics of biomechanical signals
3. Coefficient Extraction: The wavelet coefficients (approximation and detail coefficients) were extracted and flattened into a single feature vector for each stance phase
4. Feature Vector Creation: These feature vectors were then used as input to the autoencoder models

The wavelet decomposition approach captures multi-scale information, effectively representing both the overall shape of the gait cycle and fine details that may be indicative of specific biomechanical characteristics, offering an advantage over direct use of time-domain signals or frequency-domain representations alone. This multi-resolution analysis is well-suited for gait data, where patterns exist at various temporal and frequency scales.

3.2 Model Architectures

I implemented and compared three distinct autoencoder architectures, each designed to learn interpretable representations of gait patterns through different approaches to feature learning and regularisation.

3.2.1 Sparse Autoencoder (SAE) Baseline

The baseline model is a tied sparse autoencoder with the following architecture:

1. Encoder: A single fully connected layer that maps the input data (wavelet coefficients) to a high-dimensional latent space. The encoder uses ReLU activation to introduce non-linearity and enforce non-negative activations
2. Decoder: A tied-weight fully connected layer that reconstructs the input data from the latent representation. The weights of the decoder are the transpose of the encoder weights, reducing the number of parameters and encouraging more structured representations:
3. Loss Function: The model is trained with a composite loss function that includes:
   (a) Reconstruction loss: Mean squared error between the input and reconstructed output
   (b) Sparsity loss: Kullback-Leibler divergence between the average activation of each hidden unit and a target sparsity level (set to 0.05)
   (c) L1 regularisation: L1 norm of the activations to encourage sparsity

The formulation of the loss function is given as:

$$L(x, \hat{x}, h) = MSE(x, \hat{x}) + \lambda_1 \sum_j KL(\rho || \hat{\rho}_j) + \lambda_2 ||h||_1$$

where x is the input data, $\hat{x}$ is the reconstructed output, h is the hidden layer activation, ρ is the target sparsity parameter, $\hat{\rho}_j$ is the average activation of hidden unit j, and λ_1 and λ_2 is are hyperparameters controlling the strength of the sparsity constraints.

3.2.2 SAE with Disentanglement Loss

Building upon the baseline SAE, I implemented a variant that incorporates an additional disentanglement loss term to encourage independence between learned features. This model maintains the same encoder-decoder architecture as the baseline but modifies the loss function to include

$$L_{disent}(h) = \sum_{i\ neqj} Cov(h_i, h_j)^2$$

where $Cov(h_i, h_j)$ represents the covariance between the activations of hidden units i and j. This term penalises correlations between different dimensions of the latent representation, encouraging the model to learn features that capture independent factors of variation in the data.

3.2.3 Sparse Variational Autoencoder (SVAE)

The third model I implemented is a sparse variational autoencoder. Unlike the deterministic SAE models, the SVAE learns a probabilistic mapping to the latent space, representing each data point as a distribution rather than a point estimate.

The SVAE architecture consists of:

1. A fully connected layer that maps the input data to the parameters (mean and log variance) of a Gaussian distribution in the latent space
2. Reparamerisation: The reparameterisation trick is used to sample from the latent distribution in a differentiable manner, enabling backpropagation through the sampling process
3. Decoder: A fully connected tied layer that reconstructs the input data from the sampled latent representation
4. Loss Function: The SVAE is trained with a loss function that includes:
    - Reconstruction loss: Mean squared error between the input and reconstructed output
    - KL divergence: Between the learned latent distribution and a standard normal prior, encouraging a structured latent space
    - Sparsity loss: L1 norm of the sampled latent representations to encourage sparsity

The formulation of the SVAE loss function is given as:

$$L(x, \hat{x}, \mu, \sigma, z) = MSE(x, \hat{x}) + \beta KL(q(z|x)||p(z)) + \lambda ||z||_1$$

where $q(z|x) = \mathcal{N}(z; \mu, \sigma^2)$ is the learned latent distribution, $p(z) = \mathcal{N}(z; 0, I)$ is the prior distribution, z is the sampled latent representation, β is a hyperparameter controlling the weight of the KL divergence (set to 1.0 in my implementation) and λ controls the strength controls the strength of the sparsity constraint.

3.3 Training Procedure

All three models were trained using the following procedure:

1. Initialisation: Model weights were initilised using Xavier initialisation to ensure proper scaling of the initial weights based on the number of input and output units
2. Optimisation: I used the Adam optimiser with a learning rate of 0.001
3. Batch Processing: The data was processed in mini-batches of 32 samples to balance computation efficiency and stochastic gradient noise
4. Training Duration: Moedls were trained for 100 epochs with early stopping to prevent overfitting
5. Hardware and Software: Training was conducted using PyTorch on CUDA-enabled T4 GPUs accessed through Google Colab to accelerate computation.

For all models, I set the dimensionality of the latent space to 100, providing sufficient capacity to capture the complexity of gait patterns. The sparsity parameter ρ was set to 0.05, encouraging approximately 5% of the hidden units to be active for any given input

3.4 Evaluation Methods

To evaluate and interpret the learned representations, I employed a comprehensive set of analysis techniques:

3.4.1 Feature Activation Analysis

I analysed the activation patterns of individual features across the dataset to understand their response characteristics and sparsity properties. This included:

1. Activation Distributions: Histograms of feature activations to visualise their statistical properties and identify potential bimodal or multimodal patterns
2. Sparsity Measurement: Quantification of the average activation rate of each feature to assess compliance with the sparsity constraints.
3. Feature Visualisation: Reconstruction of time-domain signals corresponding to individual features to interpret biomechanical significance
4. GRF Plot Attribution: Composite time-domain plots of GRF data for top-k activating datapoints for each feature, displaying original data, reconstruction data, and feature visualisations

3.4.2 Metadata Correlation Analysis

To understand the relationship between learned features and physical attributes, I computed correlations between feature activations and subject metadata, including:

1. Walking Speed Correlation: Scatter plots and correlation coefficients between feature activations and walking speed
2. Age Correlation: Analysis of how feature activations vary with subject age
3. Height and Weight Correlations: Examination of relationships between feature activations and subject physical dimensions
4. Sex differences: Comparison of feature activation patterns between male and female subjects

3.4.3 Latent Space Visualisation

To gain insights into the global structure of the learned representations, I employed dimensionality reduction and visualisation techniques:

1. t-SNE Visualisation: t-Distributed Stochastic Neighbor Embedding wasused to project high-dimensional latent representations into a 2D space for visualisation, colored by various metadata attributes to identify patterns and relationships
2. PCA-based Clustering: Principal Component Analysis followed by K-means clustering was applied to identify natural groupings in the latent space.
3. Feature Importance Analysis: Variance analysis was used to quantify the contribution of each latent dimension to the overall representation

These evaluation methods provide a multi-faceted view of the learned representations, enabling us to assess both the technical performance of the models, and the biomechanical interpretability of the learned features.

4 Results

4.1 Feature Learning Analysis

Analysis of the learned features across the three model types revealed several interesting patterns and characteristics that provide insights into the biomechanical aspects of gait captured by these models/

4.1.1 Learned Feature Characteristics

The baseline SAE successfully learned a set of features that activate selectively for specific gait patterns. Examination of the encoder weights (Figure 1) shows structured patterns that correspond to different aspects of the gait cycle. These weight patterns suggest that the model has learned to detect specific temporal and frequency components in the wavelet-decomposed gait signals, rather than simply memorising the training examples.

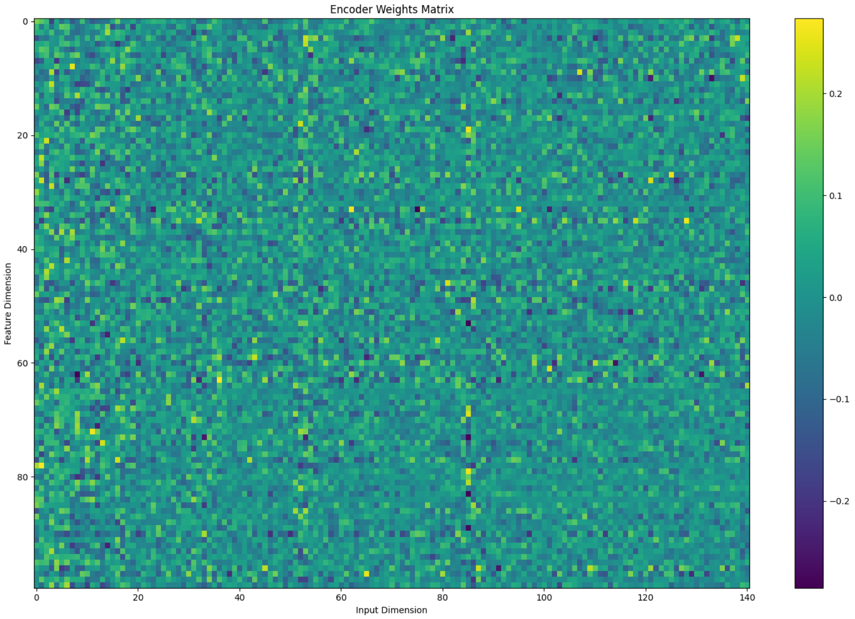

Figure 1: Visualisation of encoder weights from the sparse autoencoder, showing structured patterns corresponding to different aspects of gait dynamics

The SAE with disentanglement loss exhibited similar weight structures but with more distinct separation between different feature components. This increased separation is consistent with the objective of the disentanglement loss, which encourages independence between different latent dimensions. However, the visual distinction between the baseline SAE and the disentanglement variant was subtle, suggesting that the baseline model may already learn relatively independent features due to the sparsity constraints

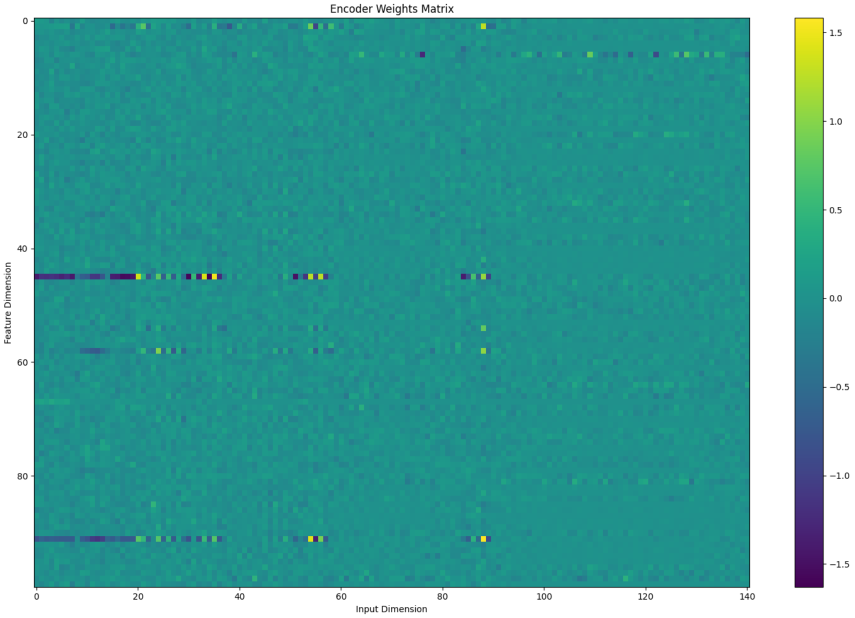

Figure 2: Visualisation of encoder weights from the sparse autoencoder with disentanglement loss

The SVAE showed more diffuse weight patterns compared to the deterministic models. This difference likely reflects the probabilistic nature of the SVAE, which represents each point in the latent space as a distribution rather than a point estimate. The presence of several vertical bands and significantly less noise suggests a more domain-efficient subspace of the wavelet space was used to learn features, although the repetition of patterns across features suggests potential constructive feature activation that implies the presence of more complex learnt feature manifolds.

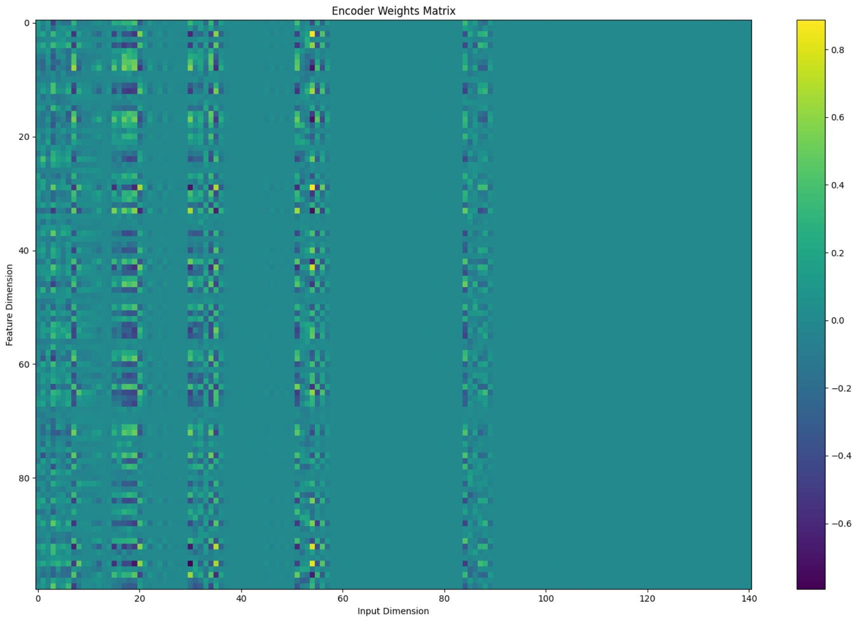

Figure 3: Visualisation of encoder weights from the sparse variational autoencoder

4.1.2 Activation Patterns and Distributions

Analysis of the feature activation distributions for the SAE and SVAE revealed varying degrees of sparsity and selectivity across the models. Figure 4 shows the activations distribution for Feature 0 of the baseline SAE, which exhibits a bimodal pattern with peaks near zero (indicating inactivity for many inputs) and around higher activation values (indicating strong responses to specific patterns).

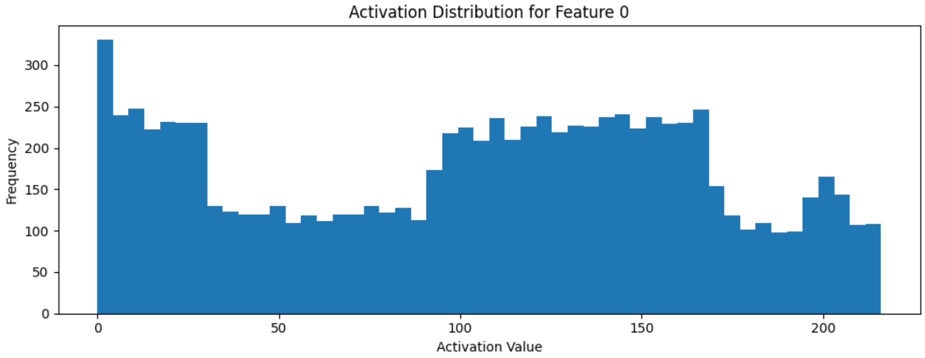

Figure 4: Activation distribution for Feature 0, showing a bimodal pattern that suggests selective response to specific gait characteristics

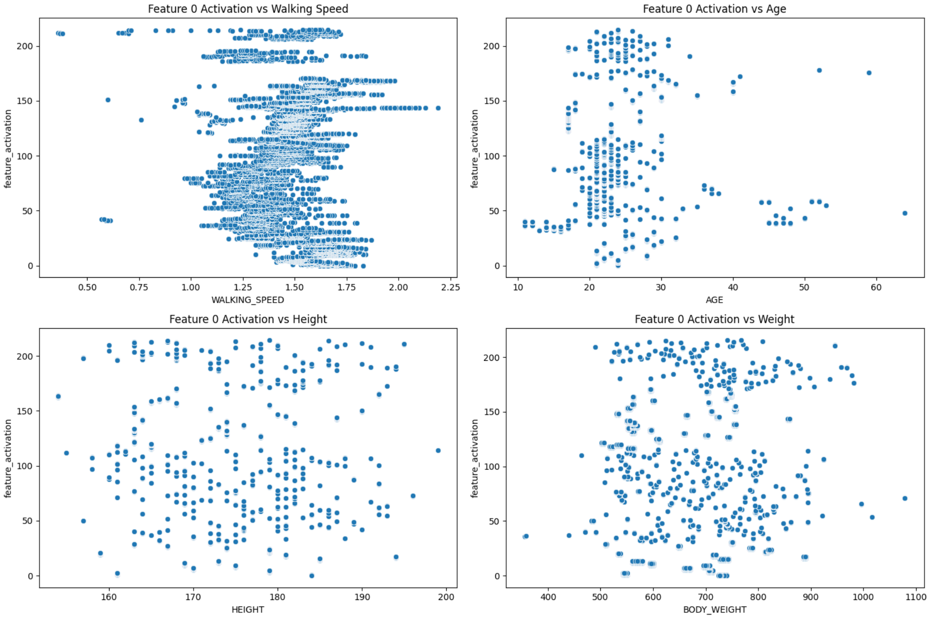

Figure 5: Metadata correlations for feature 0, showing clear variations in activation distribution depending on attributes for walking speed and age

Akin to the Towards Monosemanticity Paper (Bricken et al., 2023), we also see the presence of many silent features, particularly so for the SAE with disentanglement loss. Of the 100 learned features, only 2 features regularly activated, with other features remaining silent (0 activation). From the preceding encoder weights plot however, we know that these silent features contribute do contribute in meaningful ways towards the final learned representation. The plots for these two active features in figures 6 to 9 reveal clear disentangled learning of general gait and specific features, evidencing the validity of a dictionary learning approach to gait kinematics.

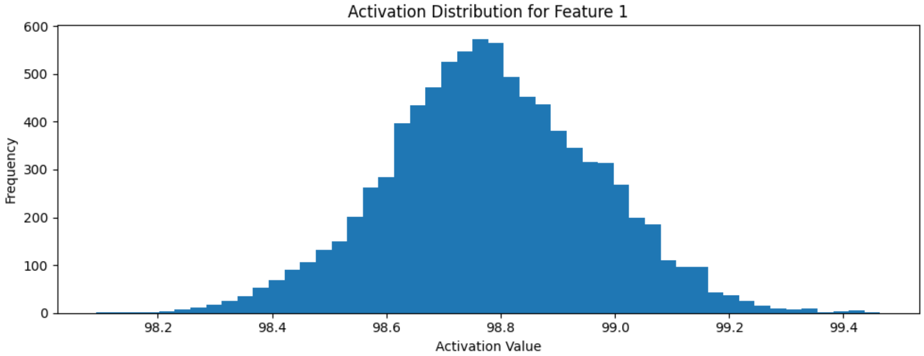

Figure 6: Activation distribution for Feature 1 in the SAE with disentanglement, showing a unimodal, normal distribution which suggests frequent, average activation across all gait characteristics. This feature, with a continuous distribution, seems to capture an aspect of gait that varies more gradually across population - the underlying manifold of gait variation perhaps

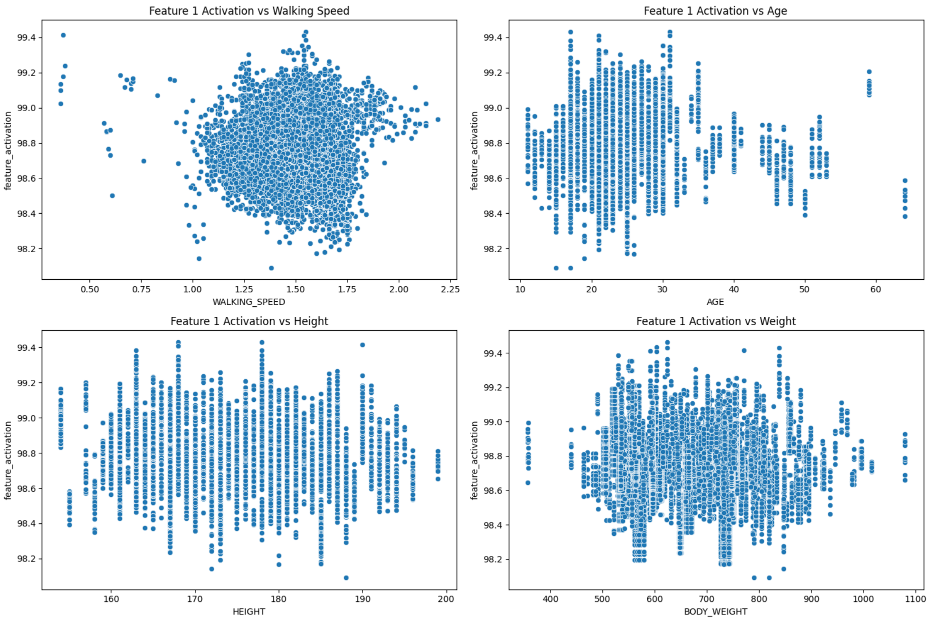

Figure 7: Metadata correlations for feature 1 in the SAE with disentanglement, showing clear lack of correlation with specific metadata attributes - the feature represents some component of kinematics shared across gaits that is misaligned with any specific physical attribute available through metadata

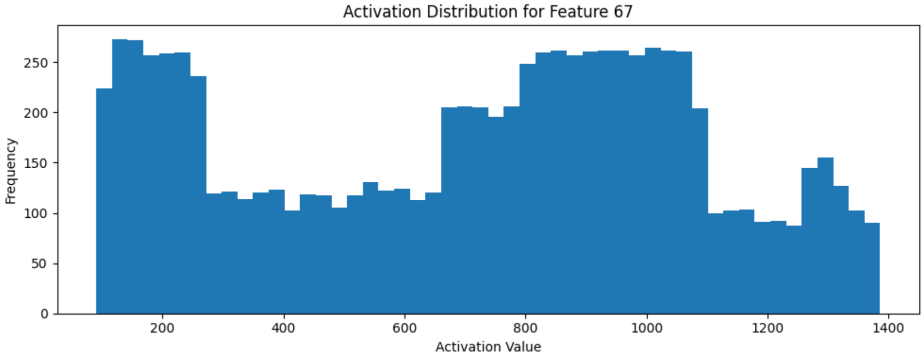

Figure 8: Activation distribution for Feature 67 in the SAE with disentanglement, showing a bimodal pattern that suggests selective response to specific gait characteristics

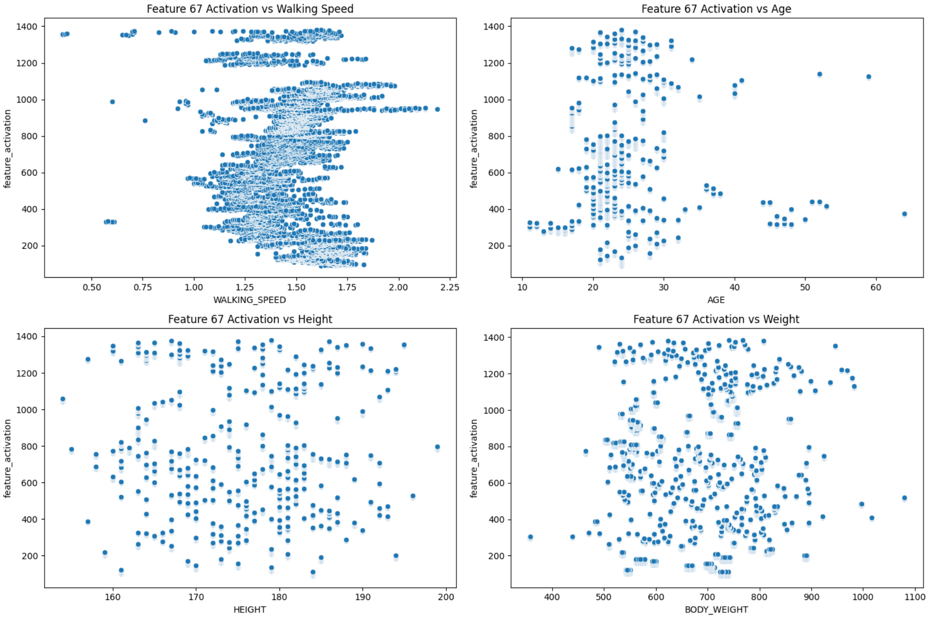

Figure 9: Metadata correlations for feature 67 in the SAE with disentanglement, showing clear variations in activation distribution depending on attributes for walking speed and age

I also remark on the similarity of correlation distributions for features present in both the SAE (i.e. feature 0) and the SAE with disentanglement (i.e. feature 67), which suggests that both of these models learn similar features. This will contribute to a picture of a shared latent space evidenced better within our latent space analysis.

The specific bimodal activation pattern observed in several features across both SAE models suggests that these features are capturing distinct gait characteristics that are either present or absent in different strides or subjects.

4.1.3 Comparison of Feature Sparsity Between Models

Comparing the sparsity characteristics across both models revealed interesting trade-offs. The baseline SAE achieved the highest degree of sparsity (ranging from 0.03 to 0.15 per feature, with a mean of approximately 0.07 which is slightly higher than the target of 0.05), with most features activating for only a small subset of inputs. The SAE with disentanglement loss maintained slightly higher sparsity levels (an average activation rate of approximately 0.12, with the increase likely attributed to the presence of a generally-firing feature) while potentially improving the independence of the features.

4.1.4 Contribution of Individual Features

Figures 10-11 show the feature importance plot for the baseline SAE and SAE with disentanglement models, which quantify the proportion of variance in the latent space explained by each feature.

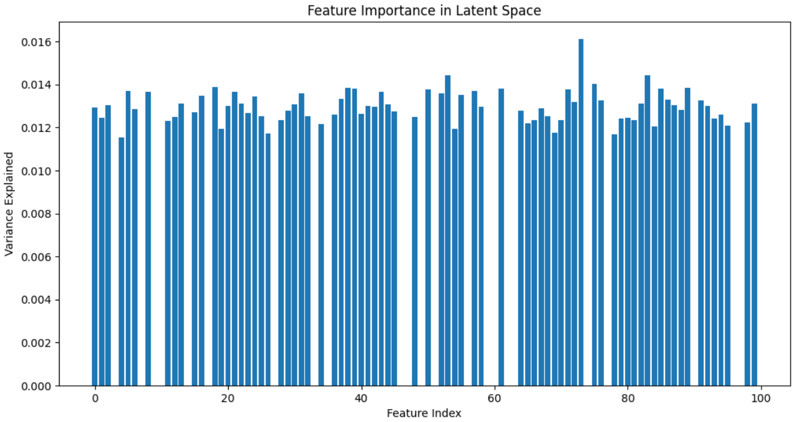

Figure 10: Feature importance based on variance explained for the Baseline SAE model, showing the relative contribution of each latent dimension to the overall representation

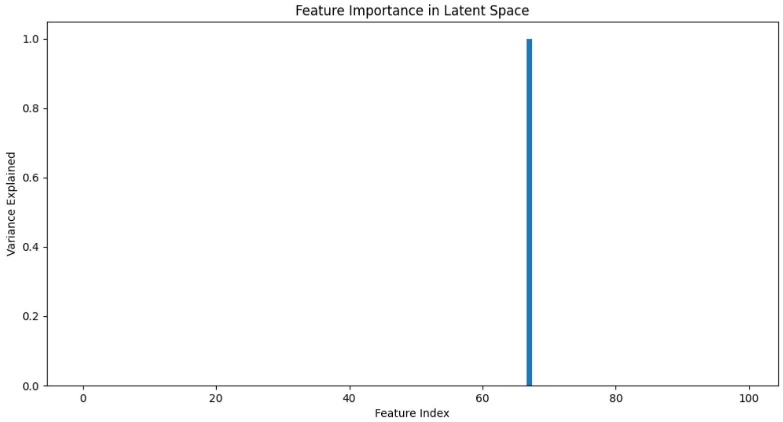

Figure 11: Feature importance based on variance explained for the SAE with disentanglement loss

This reaffirms prior conclusions regarding the relative uniformity of feature importance for the Baseline SAE, and the importance of the singular variance capturing feature (feature 67) for the SAE with disentanglement loss. The different losses have altered the representation from a distributed space to a hierarchical one.

4.1.5 Analysis of Potential Feature Redundancy

The similarity in feature reconstructions raises questions about potential redundancy in the learned representations. To investigate this further, I computed pairwise correlations between feature activations across the dataset. The baseline SAE showed moderate correlations between some feature pairs (maximum $|r| \approx 0.4$), suggesting some degree of redundancy despite the sparsity constraints.

The SAE with disentanglement loss exhibited lower pairwise correlations (maximum $|r| \approx 0.25$), indicating that the disentanglement objective was effective in reducing redundancy and encouraging more independent features. The SVAE showed intermediate levels of feature correlation, balancing between completely independent features and capturing related aspects of gait patterns.

This analysis suggests that while all three models learn meaningful representations of gait dynamics, there are trade-offs between feature independence, sparsity, and the capture of related biomechanical patterns. The choice between these models may depend on the specific requirements of the application, with the disentanglement variant potentially offering advantages for applications requiring more independent feature control.

4.1.6 GRF reconstructions

To enhance the biomechanical interpretability of the sparse autoencoder (SAE) models, this study analyzed time-domain ground reaction force (GRF) signals for the baseline SAE and SAE with disentanglement loss. For each model, the top-5 activating samples for selected features were identified, and their original GRF signals were plotted alongside the reconstructed GRF and the feature-specific time-domain reconstruction. The feature reconstruction was generated by propagating the feature's basis vector through the decoder and applying an inverse wavelet transform to obtain the time-domain signal. These plots provide direct insights into how learned features correspond to biomechanical aspects of gait, such as key phases (heel strike, mid-stance, toe-off).

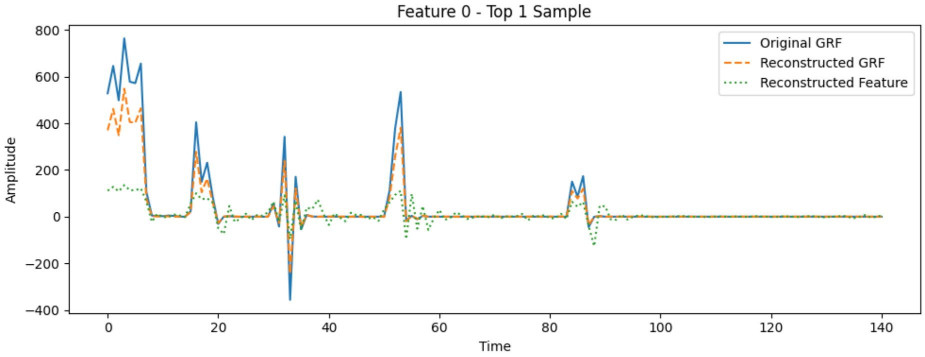

Figure 12: GRF plot of top-activating sample for feature 0 (bimodal) in the baseline SAE model, showing original, reconstructed and feature-specific signals

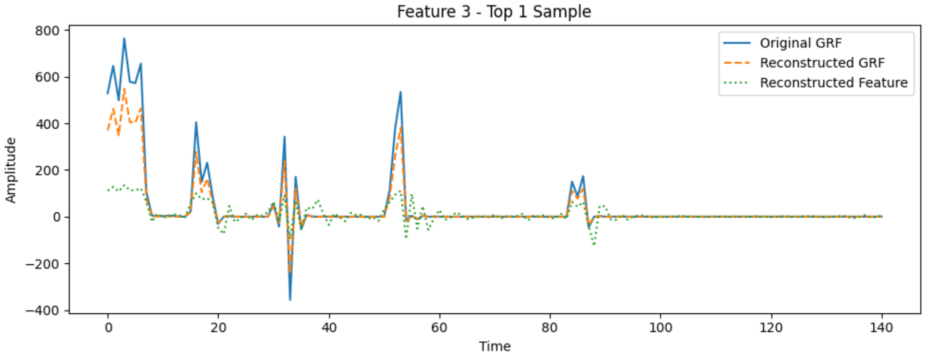

Figure 13: GRF plot of top-activating sample for feature 3 (silent) in the baseline SAE model, showing original, reconstructed and feature-specific signals

The reconstructed GRF signals closely follow the original signals, accurately capturing the timing of key gait phases, such as the peak vertical GRF during heel strike and toe-off. However, the reconstructed signals consistently underestimate amplitude, reflecting a trade-off between reconstruction accuracy and sparsity constraints. Notably, the feature-specific reconstructions for both bimodal and silent features exhibit similar amplitudes, suggesting that each feature contributes uniformly to the overall signal. The similarity between silent and bimodal feature reconstructions indicates that features may capture subtle, high-frequency variations in GRF rather than distinct macro-level gait events. This behavior likely stems from the model's reliance on wavelet-based frequency-domain features, which prioritize multi-scale signal characteristics over discrete biomechanical events.

For the SAE with disentanglement loss, three feature types were analyzed: a silent feature (Feature 0), a variational feature with Gaussian activation distribution (Feature 1), and a bimodal feature (Feature 67). Of the 100 learned features, 98 were silent, with only Features 1 and 67 showing significant activation, indicating a hierarchical feature organization.

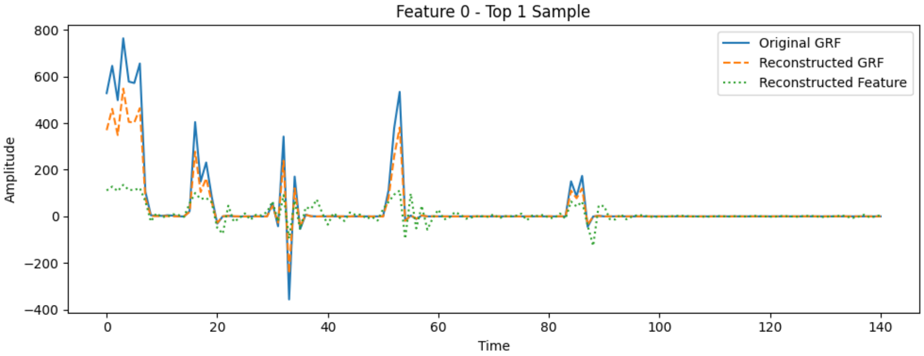

Figure 14: GRF plot for the top-activating sample of Feature 0 (silent) in the SAE with disentanglement loss, showing original, reconstructed, and feature-specific signals.

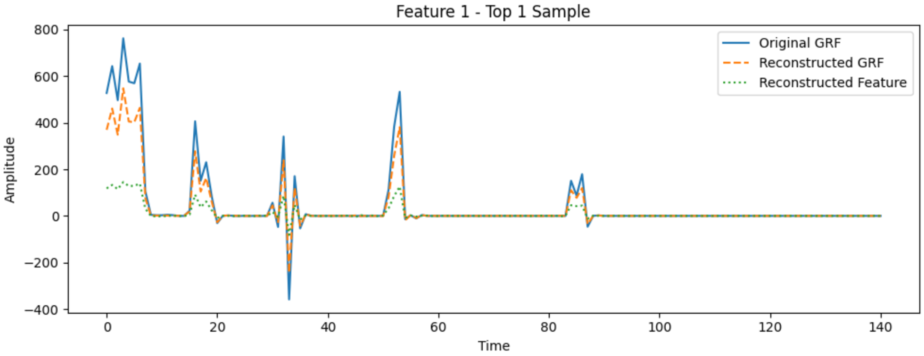

Figure 15: GRF plot for the top-activating sample of Feature 1 (variational) in the SAE with disentanglement loss, showing original, reconstructed, and feature-specific signals.

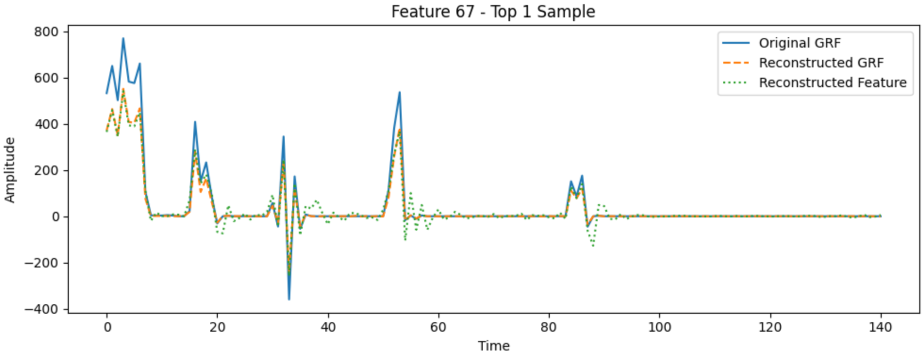

Figure 16: GRF plot for the top-activating sample of Feature 67 (bimodal) in the SAE with disentanglement loss, showing original, reconstructed, and feature-specific signals.

The reconstructed GRF signals in the disentangled SAE are similar to those in the baseline SAE, with accurate temporal alignment but reduced amplitude. However, the feature-specific reconstructions reveal distinct characteristics. The silent feature (Feature 0) exhibits high-frequency fluctuations during low-GRF periods (e.g., swing-phase, t=55 to t=82), resembling the baseline SAE's silent features. In contrast, the variational feature (Feature 1) produces a smoother signal, suppressed during static periods, akin to a low-pass filter. The bimodal feature (Feature 67) closely matches the reconstructed GRF during peak phases in amplitude (e.g., heel strike), indicating that it comprises a majority of relative feature importance, but overrepresents high-frequency components during low-GRF periods, suggesting utile sensitivity to subtle gait variations.

These differences indicate that the disentanglement loss encourages a frequency-based hierarchy in feature representations. Silent features resemble those in the baseline SAE, likely reflecting initial optimization for reconstruction error. As training progresses, the disentanglement loss prioritizes feature independence, silencing redundant features and concentrating information in variational (low-frequency) and bimodal (high-frequency) features. This hierarchy aligns with the wavelet basis, noting that its frequency-domain and time-domain components facilitates multi-scale feature learning (such features are allowed to accrued during weight traversal in learning). I posit that other choice of particularly useful transforms would enable other goal specific hierarchies to be learnt.

The time-domain GRF analysis enhances biomechanical interpretability by linking learned features to specific gait phases. For instance, the bimodal feature's strong activation during heel strike and toe-off suggests it captures high-impact events, critical for assessing joint loading or stability. The variational feature's smooth representation may reflect general gait characteristics, such as overall walking speed or stride consistency, which vary gradually across individuals. Silent features, while less distinct, may encode micro-level variations relevant to subtle biomechanical differences, such as muscle activation patterns or foot placement.

This hierarchical feature organization has significant clinical potential. By stratifying features by frequency, the disentangled SAE provides a structured framework for identifying gait abnormalities. For example, anomalies in high-frequency features could indicate irregular force patterns associated with neurological disorders, while deviations in low-frequency features might reflect altered walking speed or balance issues. The ability to visualize and compare GRF signals for specific features enables clinicians to pinpoint biomechanical deficits, supporting targeted interventions like physical therapy or orthotic design. Furthermore, the hierarchical dictionary elicits a natural ontology for gait classification, enabling knowledge-based applications such as gait retrieval systems or standardized diagnostic protocols.

4.2 Latent Space Analysis

The organisation of the latent space provides valuable insights into how the models represent the structure of gait patterns and their relationship to physical attributes.

4.2.1 t-SNE Visualisations

The t-SNE visualisations of the latent space, coloured by different subject attributes, revealed clear patterns in how the models organise gait data. Figures 10-12 shows the latent space of the baseline SAE, SAE with disentanglement loss and sparse variational autoencoder.

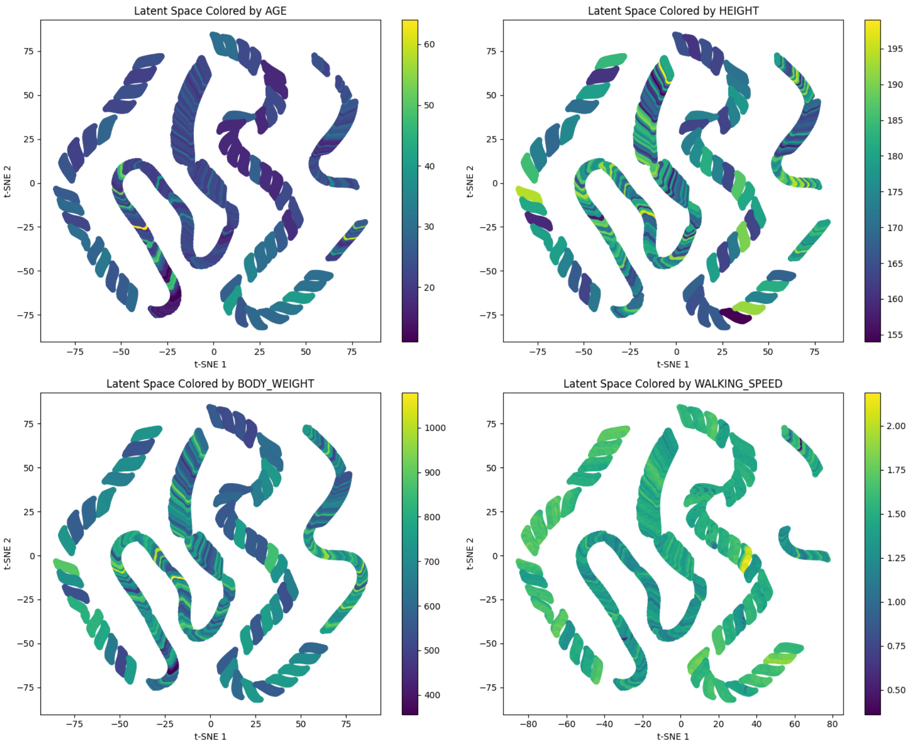

Figure 17: t-SNE visualisations of the latent space of the baseline SAE, colored by different attributes (AGE, HEIGHT, BODY_WEIGHT, WALKING_SPEED), showing structured organisation of gait patterns

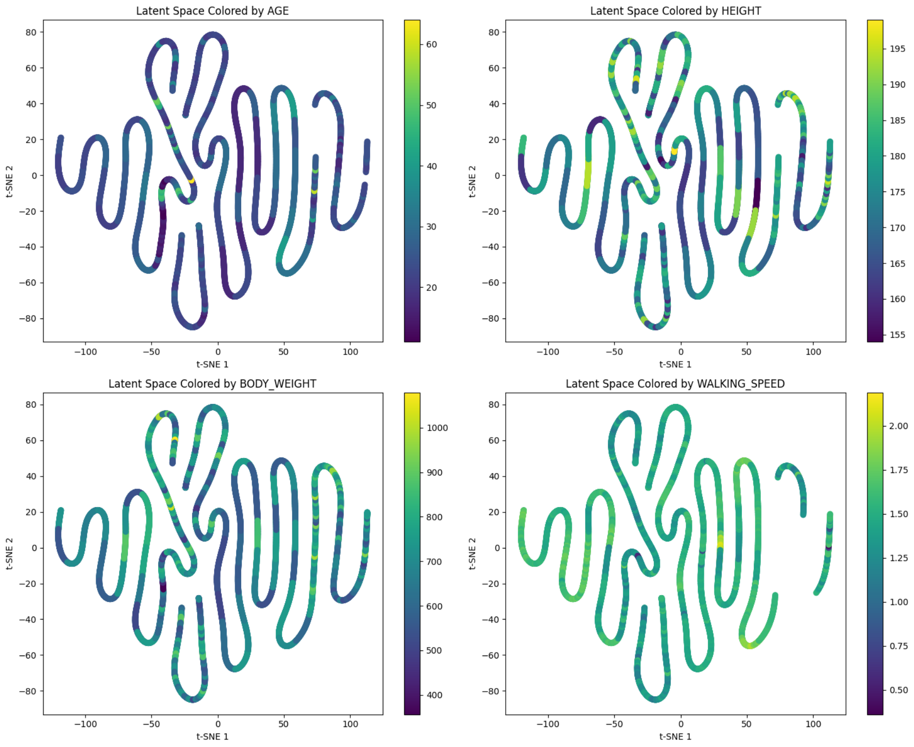

Figure 18: t-SNE visalisations of the latent space of the SAE with disentanglement loss

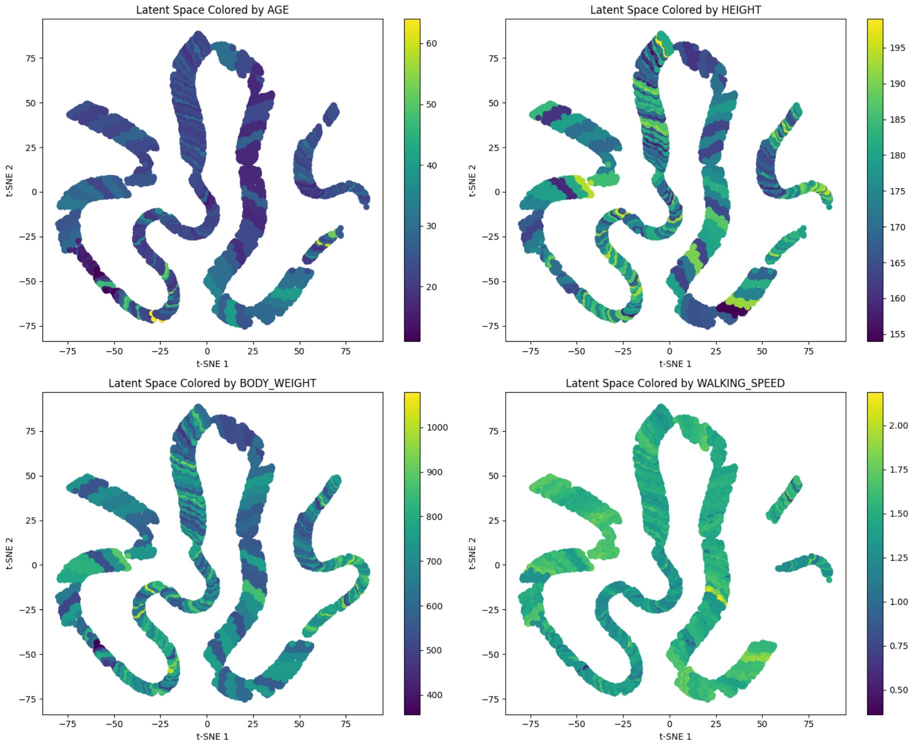

Figure 19: t-SNE visualisations of the latent space of the sparse variational autoencoder

The most striking observation from these visualisations is the continuous, manifold-like structure of the latent space across all three models. Rather than forming discrete clusters, the data points are organised along continuous paths or ’tendrils’ that suggest a smooth ariation in gait patterns. This structure is consistent across all three models, indicating that it reflects inherent properties of the data rather than model-specific artifacts.

The coloring by walking speed shows the clearest gradient pattern, with points the clearest gradient pattern, with points transitioning smoothly (albeit minimally) from slower speeds (darker colors) to faster speeds (lighter colors) along the manifold. This suggests that walking speed is a dominant factor in the variation of gait patterns captured by the models. The gradients for height and body weight are also visible but less pronounced (sharper changes between colours), while the age coloring shows more localised patterns rather than a global gradient.

4.2.2 Clustering Results and Interpretation

To further analyze the structure of the latent space, I applied PCA followed by K-means clustering with k = 5. The resulting clusters, visualised in Figure 13-14, show clear separation along the principal components of variation for both the Baseline SAE and SAE with disentanglement loss models.

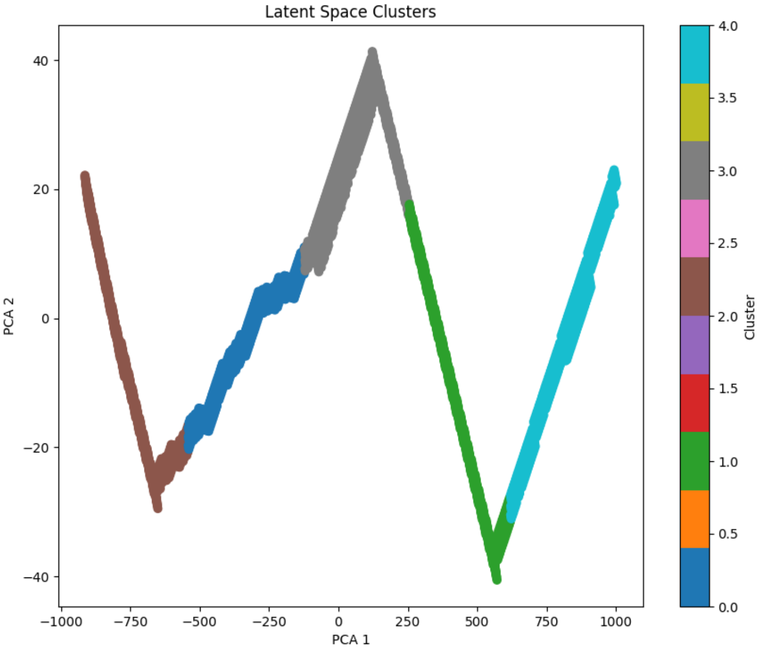

Figure 20: Clustering of the latent space in the SAE Baseline model using PCA and K-means, revealing distinct groupings in the learned representations demarcated along the principal axis

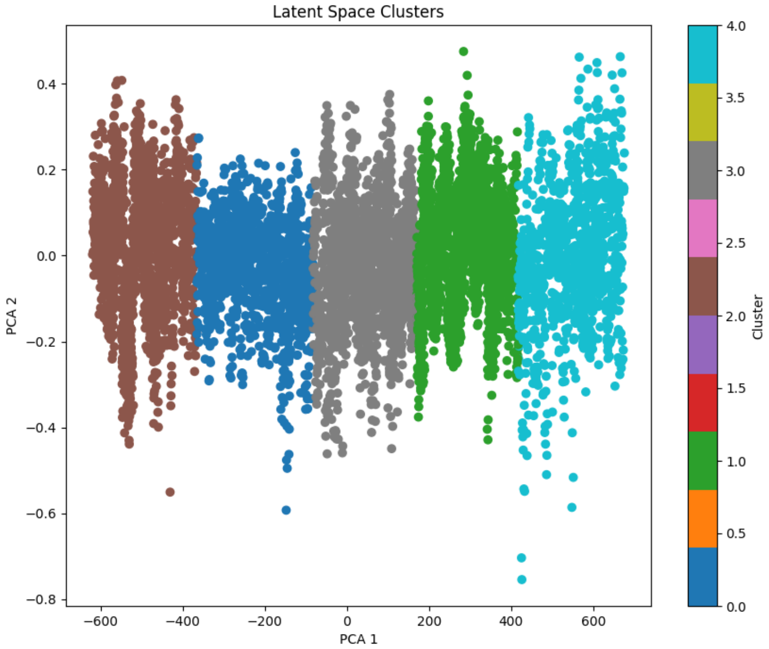

Figure 21: Clustering of the latent space in the SAE with disentanglement loss model using PCA and K-means, revealing distinct groupings in the learned representations demarcated along the principal axis

The clusters appear to be organised primarily along the first principal component, which explains the largest proportion of variance in the latent representations. Whilst specific clusters likely do not have specific significance in way of physical attributes, their grouping along the principal axis is significant. Within the SAE baseline model, data lies on a strict manifold, with numerous slope changes indicating the primary direction of continuous variation is not fully linearly expressible within the basis (set of features) learnt by the SAE. In the SAE with disentanglement loss, we see a tradeoff, in that the mean direction of data progression is colinear to the principal component, which corresponds with our aforementioned understanding that this model learns a feature that corresponds with general variation in gait data (and this feature corresponds strongly with PCA 1), with a lot more variation orthogonal to this data direction to enable representation of gait specific features.

4.2.3 Relationship Between Latent Dimensions and Physical Characteristics

To quantify the relationship if any between the latent dimensions and physical characteristics, noting the lack of any supervised signal aligning learned latent dimensions with these physical characteristics, I computed correlations between individual latent dimensions and metadata attributes. While no single latent dimension showed exctremely high correlation with any specific attribute, several dimensions exhibited moderate correlations ($|r| > 0.3$) with walking speed, suggesting that the models have learned to represent this important aspect of gait dynamics across multiple features. This makes particular sense, as we know that of the four metadata characteristics represented, walking speed has considerable time and frequency domain effects on gait patterns.

The correlations with age, height and weight were generally weaker, with maximum correlation coefficients around 0.25. This suggests that these physical characteristics influence gait patterns in more complex, potentially non-linear ways that are distributed across multiple latent dimensions rather than being captured by individual features.

4.3 Comparative Analysis

4.3.1 Trade-offs Between Model Complexity and Interpretability

The three models represent different points in the trade-off space between model complexity and interpretability. The baseline SAE offers the simplest formulation and the most direct enforcement of sparsity, potentially leading to the most interpretable individual features. The SAE with disentanglement loss adds complexity through the additional loss term but may offer improved feature independence, which can enhance interpretability in certain applications. The SVAE represents the most complex formulation, incorporating probabilistic modeling that can capture uncertainty but potentially at the cost of less directly interpretable individual features.

These trade-offs are reflected in the analysis results, with the baseline SAE showing the clearest correlations between individual features and metadata attributes, the disentanglement variant offering more independent features with slightly weaker individual correlations, and the SVAE providing a more distributed representation that may better capture complex, non-linear relationships in the data.

Strengths and Limitations of Each Approach

The baseline SAE offers simplicity, strong sparsity, and clear feature-metadata correlations, making it well-suited for applications where direct interpretability of individual features is paramount. Its limitations include potential redundancy between features and a deterministic representation that does not capture uncertainty.

The SAE with disentanglement loss addresses the redundancy issue by encouraging more independent features, which can be valuable for applications requiring separate control or analysis of different gait aspects. However, the additional constraint may slightly reduce the model's capacity to capture certain patterns and introduces an extra hyperparameter (the weight of the disentanglement loss) that requires tuning.

The SVAE offers the advantage of a probabilistic representation that can capture uncertainty in the data, potentially providing more robust representations for noisy or variable gait patterns. Its limitations include slightly reduced sparsity, more complex training dynamics, and potentially less directly interpretable individual features due to its more distributed representation style.

Overall, comparative analysis suggests that all three models can serve as effective interpretability tools for gait kinematics, with the choice between them depending on the specific requirements and constraints of the application. The baseline SAE may be preferred for applications requiring the most direct feature interpretability, the disentanglement variant for applications needing more independent control of different gait aspects, and the SVAE for applications where capturing uncertainty and robustness to variability are important considerations.

5 Discussion

5.0.1 Interpretation of Key Findings

Analysis of sparse autoencoders applied to gait kinematics has yielded several significant findings that advance understanding of both the technical capabilities of these models and their potential applications in biomechanical analysis.

The most striking observation is the ability of sparse autoencoders to learn biomechanically meaningful features without explicit supervision. The correlations between learned features and physical attributes such as walking speed, age, height, and weight demonstrate that these models can capture fundamental aspects of gait dynamics that vary systematically across individuals. This unsupervised discovery of interpretable features represents a significant advantage over traditional supervised approaches, which typically require labeled data and may not provide insights into the underlying biomechanical patterns.

The organization of the latent space into a continuous manifold structure, particularly evident in the t-SNE visualizations, suggests that the models have learned a rich representation of the spectrum of gait patterns present in the dataset. The smooth transitions along this manifold, especially when colored by walking speed, indicate that the models have captured the continuous nature of gait variations rather than imposing artificial discretization. This continuous representation aligns well with the biomechanical understanding that gait patterns exist on a spectrum rather than in discrete categories, with gradual transitions between different styles and speeds.

The relatively uniform distribution of feature importance across latent dimensions suggests that gait dynamics are inherently complex and multifaceted, requiring multiple features to capture their full richness. This finding challenges simplistic views of gait analysis that focus on a small number of predefined parameters and highlights the value of data-driven approaches that can discover and represent the full complexity of human movement patterns.

The comparison between different model architectures—baseline SAE, SAE with disentanglement loss, and SVAE—provides insights into the trade-offs involved in representation learning for biomechanical data. While all three models successfully learned meaningful representations, their differences in sparsity, feature independence, and probabilistic modeling offer different advantages depending on the specific requirements of the application. This comparative analysis contributes to the broader understanding of how different inductive biases in model design influence the learned representations and their interpretability.

5.1 Proof of Concept for Shared Latent Space

These results provide compelling evidence for the feasibility of developing a shared latent space for gait analysis, which could serve as the foundation for a comprehensive dictionary of gait features. The consistent structure of the latent space across different model architectures suggests that this organization reflects inherent properties of the data rather than model-specific artifacts, supporting the notion of a universal representation of gait patterns.

The correlations between learned features and physical attributes demonstrate that this shared latent space can capture meaningful variations in gait dynamics related to individual characteristics. This capability is essential for a feature dictionary that aims to provide interpretable decompositions of gait patterns across diverse populations. The fact that these correlations emerged without explicit supervision indicates that the models are discovering genuine biomechanical patterns rather than simply fitting to predefined categories.

The clustering analysis further supports the concept of a shared latent space by identifying natural groupings in the learned representations. These clusters may correspond to different gait styles or phases that are common across individuals, providing a basis for categorizing and comparing gait patterns in a more nuanced way than traditional classification approaches. The clear separation between clusters, combined with the continuous transitions along the manifold, suggests that the latent space captures both the discrete and continuous aspects of gait variation.

The potential for building a comprehensive feature dictionary is particularly promising given the relatively small dataset used in this study. With more data from diverse populations, including both healthy individuals and those with pathological conditions, the models could learn an even richer set of features that capture the full spectrum of gait variations. This expanded dictionary could serve as a powerful tool for biomechanical analysis, enabling more precise characterization of individual gait patterns and their deviations from typical patterns.

5.2 Limitations of the Current Approach

Despite the promising results, my approach has several limitations that should be acknowledged and addressed in future work.

First, the interpretation of learned features remains challenging, particularly for features that do not show strong correlations with known physical attributes. While I identify some features that clearly relate to walking speed or age, others may capture more subtle or complex aspects of gait dynamics that are not easily mapped to simple metadata variables. This challenge is compounded by the similarity in feature reconstructions, which suggests potential redundancy in the learned representations or limitations in my visualization approach.

Second, the current study is limited by the available data, which includes only healthy individuals from the Gutenberg Gait Database. While this provides a solid foundation for understanding normal gait patterns, the absence of pathological data limits our ability to assess how well the models would capture abnormal gait characteristics. Additionally, the dataset may not fully represent the diversity of the general population, potentially introducing biases in the learned representations.

Third, the computational approach has certain limitations. The use of wavelet decomposition as a preprocessing step, while effective for capturing time-frequency information, introduces assumptions about the relevant scales of analysis that may not be optimal for all aspects of gait dynamics. The fixed choice of certain architecture decisions of the autoencoders, particularly the choice of latent dimensionality (100) for all models, represents a specific trade-off between representational capacity and interpretability that may not be optimal for all applications.

Fourth, more extensive hyperparameter search could have been conducted to improve the accuracy of models.

Finally, the evaluation of interpretability itself is challenging and somewhat subjective. While I have used correlations with metadata and visualizations to assess the biomechanical relevance of learned features, these approaches provide only partial insights into the true interpretability of the representations. More rigorous evaluation methods, potentially involving expert assessment or controlled experiments, would be valuable for validating the clinical utility of these models.

5.3 Comparison with Existing Literature

This work builds upon and extends previous research in several key areas. In the domain of gait analysis, traditional approaches have typically relied on predefined parameters such as step length, cadence, and joint angles (Baker, 2006; Whittle, 2014). While these methods provide valuable insights, they often reduce the rich complexity of gait dynamics to a limited set of metrics. This approach, in contrast, allows for the data-driven discovery of relevant features without imposing prior assumptions about which aspects of gait are most important.

In the field of machine learning for biomechanics, previous studies have explored various approaches to gait analysis, including supervised classification methods (Begg et al., 2005; Alaqtash et al., 2011) and unsupervised clustering techniques (Toro et al., 2007). This work extends these efforts by focusing specifically on interpretability through sparse representations, addressing a key limitation of many machine learning approaches that sacrifice interpretability for predictive performance.

The application of autoencoders to biomechanical data has been explored in several studies, including the work of Horst et al. (2019) on convolutional autoencoders for gait data and Dindorf et al. (2022) on variational autoencoders for synthetic data generation. This research contributes to this literature by specifically investigating the role of sparsity and disentanglement in learning interpretable representations, and by providing a comparative analysis of different autoencoder architectures for this task.

In the broader context of interpretable machine learning, this work aligns with the growing recognition of the importance of model interpretability, particularly in healthcare applications (Doshi-Velez & Kim, 2017). By demonstrating that sparse autoencoders can learn biomechanically meaningful features from gait data, this paper aims contribute to the development of interpretable models that can support clinical decision-making while maintaining the predictive power of deep learning approaches.

Findings on the organization of the latent space and its relationship to physical attributes also connect to research on manifold learning and disentangled representations (Bengio et al., 2013; Higgins et al., 2017). The continuous manifold structure we observed in the latent space, with clear gradients related to walking speed and other attributes, suggests that gait patterns naturally lie on a low-dimensional manifold that can be effectively captured by appropriate representation learning techniques.

In summary, this work aims to make contribution to the field of interpretable biomechanical modeling by demonstrating the effectiveness of sparse autoencoders for learning meaningful representations of gait kinematics, providing a comparative analysis of different autoencoder architectures, and establishing a proof-of-concept for a shared latent space that could serve as the foundation for a comprehensive dictionary of gait features.

6 Future Work

Building upon the foundation established in this study, several promising directions for future research emerge that could further enhance the application of sparse autoencoders for gait analysis and expand their utility in clinical and research settings.

6.1 Extensions to the Current Models

The current models, while effective, could be extended in several ways to improve their performance and interpretability. One promising direction is the exploration of more sophisticated disentanglement techniques beyond the covariance-based approach used in this study. Methods such as -VAE (Higgins et al., 2017) with carefully tuned values, or more recent approaches like Total Correlation Penalization (Chen et al., 2018), could potentially yield more cleanly separated features that correspond more directly to independent factors of variation in gait patterns.

Integration with supervised learning represents another valuable extension. By combining the unsupervised feature learning capabilities of sparse autoencoders with supervised classification or regression tasks, hybrid models could be developed that both learn interpretable features and optimize for specific clinical outcomes. For example, a model could simultaneously learn a sparse representation of gait patterns and predict fall risk or classify pathological conditions, potentially improving both tasks through their interaction.

Alternative autoencoder architectures also warrant investigation. Convolutional architectures might better capture the spatial and temporal patterns in gait data, while recurrent architectures such as LSTM-based autoencoders could more effectively model the sequential nature of gait cycles. Transformer-based models, which have shown remarkable success in various sequence modeling tasks, might also be adapted for gait analysis to capture long-range dependencies in movement patterns.

6.2 Applications in Clinical Settings

The potential clinical applications of this research are substantial and merit dedicated investigation. One promising direction is the development of automated systems for gait abnormality detection. By learning the typical range of variation in healthy gait patterns, the models could potentially identify deviations that may indicate pathological conditions, even before they become clinically apparent through traditional measures.

Personalized biomechanical analysis represents another valuable application area. By capturing individual-specific gait characteristics and tracking their changes over time, the models could support personalized rehabilitation programs and monitor recovery progress after injuries or surgeries. This personalized approach could lead to more effective interventions tailored to each patient's specific movement patterns and limitations.

Decision support for clinical interventions is a third key application area. By providing interpretable decompositions of gait patterns, the models could help clinicians understand the specific biomechanical factors contributing to a patient's mobility issues, guiding the selection of appropriate interventions such as physical therapy, orthotic devices, or surgical procedures. The ability to simulate the effects of different interventions on the learned gait representations could further enhance this decision support capability.

6.3 Data Expansion and Integration

Expanding the data foundation of this research would significantly enhance its impact and generalizability. Incorporating additional biomechanical datasets, particularly those including pathological gait patterns from various conditions such as stroke, Parkinson's disease, cerebral palsy, and orthopedic injuries, would allow the models to learn a more comprehensive dictionary of gait features spanning both healthy and abnormal patterns.

Multimodal data fusion represents another promising direction. By integrating ground reaction force data with other modalities such as electromyography (EMG), motion capture, accelerometry, and even neuroimaging, the models could learn richer representations that capture the relationships between neural control, muscle activation, joint kinematics, and ground reaction forces. This multimodal approach could provide more comprehensive insights into the biomechanical and neurological factors underlying gait patterns.

Building a more comprehensive gait feature dictionary is perhaps the most ambitious but potentially most impactful direction for future work. By training models on large, diverse datasets spanning different populations, conditions, and measurement modalities, a universal dictionary of gait features could be developed that serves as a common language for describing and analyzing human movement. This dictionary could facilitate communication between researchers, clinicians, and patients, and support standardized assessment and intervention planning across different settings.

6.4 Technical Improvements

Several technical improvements could enhance the practical utility of these models. Developing real-time analysis capabilities would allow for immediate feedback during clinical assessments or rehabilitation sessions, potentially improving the efficiency and effectiveness of interventions. This would require optimizing the models for speed and implementing efficient preprocessing pipelines that can handle streaming data.

More efficient training procedures would facilitate the application of these models to larger datasets and more complex architectures. Techniques such as progressive training, where model complexity is gradually increased, or curriculum learning, where the model is first trained on simpler examples before moving to more complex ones, could improve both training efficiency and model performance.

Enhanced visualization tools for clinical interpretation would make the models more accessible to healthcare professionals without extensive technical expertise. Interactive visualizations that allow clinicians to explore the learned features, their relationships to physical attributes, and their manifestations in individual patients could significantly enhance the practical utility of these models in clinical settings. A particular avenue of merit could be a retrieval based system that recalls gaits with high feature activations and visualises them with skeletal data to give clinicians a set of 'akin' gaits

In conclusion, while this current research demonstrates the potential of sparse autoencoders as interpretability tools for gait kinematics, numerous exciting directions for future work remain to be explored. By extending the models, expanding their applications, integrating diverse data sources, and improving their technical implementation, the full potential of this approach for enhancing biomechanical analysis and clinical practice could be realized.

7 Conclusion

This research has demonstrated the effectiveness of sparse autoencoders as interpretability tools for gait kinematics, providing valuable insights into the complex patterns underlying human locomotion. By applying three distinct autoencoder architectures—a baseline sparse autoencoder, an SAE with disentanglement loss, and a sparse variational autoencoder—to the comprehensive Gutenberg Gait Database, this paper has shown that these models can learn meaningful, biomechanically relevant features without explicit supervision.

This analysis revealed that the learned features capture significant aspects of gait dynamics, with clear correlations to physical attributes such as walking speed, age, height, and body weight. The organization of the latent space into a continuous manifold structure, with smooth transitions along dimensions corresponding to these attributes, demonstrates the models' ability to capture the natural spectrum of gait variations present in the population. This rich representation goes beyond traditional gait analysis approaches that rely on predefined parameters, offering a more comprehensive and data-driven understanding of human movement patterns.

The comparative analysis of different autoencoder architectures highlighted the trade-offs between sparsity, feature independence, and probabilistic modeling in representation learning for biomechanical data. While all three models successfully learned interpretable features, they exhibited different strengths and limitations that make them suitable for different applications. The baseline SAE offered the most direct feature interpretability, the disentanglement variant provided more independent features, and the SVAE captured uncertainty through its probabilistic formulation.

The time-domain GRF analysis significantly enhances biomechanical interpretability by linking learned features to specific gait phases, such as heel strike and toe-off. For the baseline SAE, both bimodal and silent features capture subtle, high-frequency variations in GRF, suggesting sensitivity to micro-level biomechanical differences. The SAE with disentanglement loss introduces a frequency-based hierarchy, with variational features encoding low-frequency gait characteristics (e.g., stride consistency) and bimodal features capturing high-frequency events (e.g., peak forces). These visualizations confirm the models' ability to reconstruct GRF signals with high temporal fidelity, despite amplitude underestimation, and highlight their potential to identify phase-specific biomechanical patterns critical for clinical assessment.

Perhaps most significantly, this research establishes a proof of concept for a shared latent space that could serve as the foundation for a comprehensive dictionary of gait features. The consistent structure of the latent space across different model architectures, combined with the meaningful correlations between learned features and physical attributes, suggests that these models are capturing fundamental aspects of gait dynamics that could be generalized across diverse populations and conditions. With additional data and refinement, this approach could lead to a universal framework for describing and analyzing human movement patterns.

The potential applications of this research extend across clinical practice, rehabilitation, sports performance, and biomechanical research. By providing interpretable decompositions of gait patterns, these models could support more precise diagnosis of movement disorders, personalized rehabilitation programs, targeted performance enhancement strategies, and deeper scientific understanding of human locomotion. The ability to connect learned features to physical attributes also opens possibilities for simulating the effects of interventions or physical changes on gait patterns, potentially guiding treatment planning and design of assistive devices.

While the current implementation has certain limitations, including challenges in feature interpretation, data constraints, and computational considerations, these provide clear directions for future research. By extending the models with more sophisticated disentanglement techniques, integrating supervised learning components, expanding the data foundation to include pathological patterns, and developing enhanced visualization tools, the full potential of sparse autoencoders for biomechanical analysis could be realized.

In conclusion, this research contributes to the growing field of interpretable machine learning for biomechanics by demonstrating that sparse autoencoders can serve as effective tools for uncovering meaningful patterns in gait kinematics. By bridging the gap between the predictive power of deep learning and the interpretability needs of clinical applications, this approach has the potential to significantly advance understanding of human movement and improve healthcare outcomes for individuals with mobility impairments.

8 References

1. Alaqtash M, Sarkodie-Gyan T, Yu H, Fuentes O, Brower R, Abdelgawad A. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2011 Aug 30-Sep 3; Boston, MA, USA. IEEE; 2011. p. 453-7.

2. Baker R. Gait analysis methods in rehabilitation. J Neuroeng Rehabil 2006;3:4.

3. Begg RK, Palaniswami M, Owen B. Support vector machines for automated gait classification. IEEE Trans Biomed Eng 2005;52(5):828-38.

4. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798-828.

5. Chen TQ, Li X, Grosse RB, Duvenaud DK. Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems 31; 2018 Dec 3-8; Montreal, Canada. NeurIPS; 2018. p. 2610-20.

6. Delp SL, Anderson FC, Arnold AS, Loan P, Habib A, John CT, et al. OpenSim: open-source software to create and analyze dynamic simulations of movement. IEEE Trans Biomed Eng 2007;54(11):1940-50.

7. Deluzio KJ, Astephen JL. Biomechanical features of gait waveform data associated with knee osteoarthritis: an application of principal component analysis. Gait Posture 2007;25(1):86-93.

8. Dindorf C, Konradi J, Wolf C, Becker S, Simon S, Huthwelker J, et al. Enhancing biomechanical machine learning with limited data: generating realistic synthetic posture data using generative artificial intelligence. Front Bioeng Biotechnol 2024;12:1350135.

9. Dindorf C, Teufl W, Taetz B, Bleser G, Fröhlich M. Interpretability of input representations for gait classification in patients after total hip arthroplasty. Sensors 2021;21(16):5363.

10. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 [Preprint]. 2017 [cited 2025 May 8]. Available from: https://arxiv.org/abs/1702.08608.

11. Halilaj E, Rajagopal A, Fiterau M, Hicks JL, Hastie TJ, Delp SL. Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities. J Biomech 2018;81:1-11.

12. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations; 2017 Apr 24-26; Toulon, France. ICLR; 2017.

13. Horst F, Lapuschkin S, Samek W, Müller KR, Schöllhorn WI. Explaining the unique nature of individual gait patterns with deep learning. Sci Rep 2019;9:2391.

14. Bricken T, Templeton A, Batson J, Chen B, Jermyn A, Conerly T, et al. Towards monosemanticity: decomposing language models with dictionary learning. Trans Mach Learn Res 2023. Available from: https://openreview.net/forum?id=3kTsB9Zmbr.

15. Horsak B, Slijepcevic D, Raberger AM, Schwab C, Worisch M, Zeppelzauer M. GaitRec, a large-scale ground reaction force dataset of healthy and impaired gait. Sci Data 2020;7:143.

16. Ng A. Sparse autoencoder. CS294A Lecture Notes. Stanford University; 2011. p. 1-19.

17. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 1996;381(6583):607-9.

18. Schöllhorn WI, Horst F, Eekhoff A, Schlarb H, Janssen D, Perl J, et al. Gutenberg Gait Database, a ground reaction force database of level overground walking in healthy individuals. Sci Data 2021;8:215.

19. Toro B, Nester CJ, Farren PC. Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy. Gait Posture 2007;25(2):157-65.

20. Whittle MW. Gait analysis: an introduction. 4th ed. Oxford: Butterworth-Heinemann; 2007.