

R0: Foundations of the sportsfeatures Analytical Framework

Mohammad Abbas

2026-08-25

Table of contents

1	Overview	1
1.1	Core Premise	2
1.1.1	Framework Structure	2
2	Pillar 1: Variance Structure & Hierarchical Non-Constant Spread	4
2.1	Baseline Illustration	4
2.2	Hierarchical Variance Decomposition	7
2.2.1	Modelling Implications	7
3	Pillar 2: Missing Not At Random (MNAR) & MICE Imputation	7
3.1	Wearable Device Dropout Mechanics	8
3.2	Imputation Workflow	8
3.3	Diagnostic Verification: Iterative Chain Convergence	8
3.4	Interpretation	9
4	Pillar 3: Derived Variables & Physiological Constructs	10
5	What Next? From Framework to Analytical Discovery	12

1 Overview

The **sportsfeatures analytical framework** introduces a rule-based simulation environment designed to meet the challenges of modern statistical training. It models **real-world athlete performance** through structured, interpretable data generation that includes temporal variation, hierarchical dependencies, and realistic missingness. Thereby, **sportsfeatures** is represents a viable option for advance statistical modelling and data science training.

1.1 Core Premise

- **Stable Subject Baselines:** Each athlete maintains, stable physiological characteristics.
- **Dynamic Session Evolution:** Training sessions vary across environmental, physiological, and personal conditions.
- **Controlled Realism:** Incorporates domain rules, controlled randomness, and a deliberate *Missing Not At Random (MNAR)* mechanism tied to device usage.

1.1.1 Framework Structure

1. Phase 1 – Simulation Pipeline

- Conceptual idea and physiological hypothesis
- Rule-based simulation engine
- Core dataset: 500 observations $\times$ 30 variables

2. Phase 2 – Feature Engineering & Data Enhancement

- Enhanced dataset: 500 $\times$ 42 variables
- Key analytical measures: efficiency, workload normalisation, baselines, variability

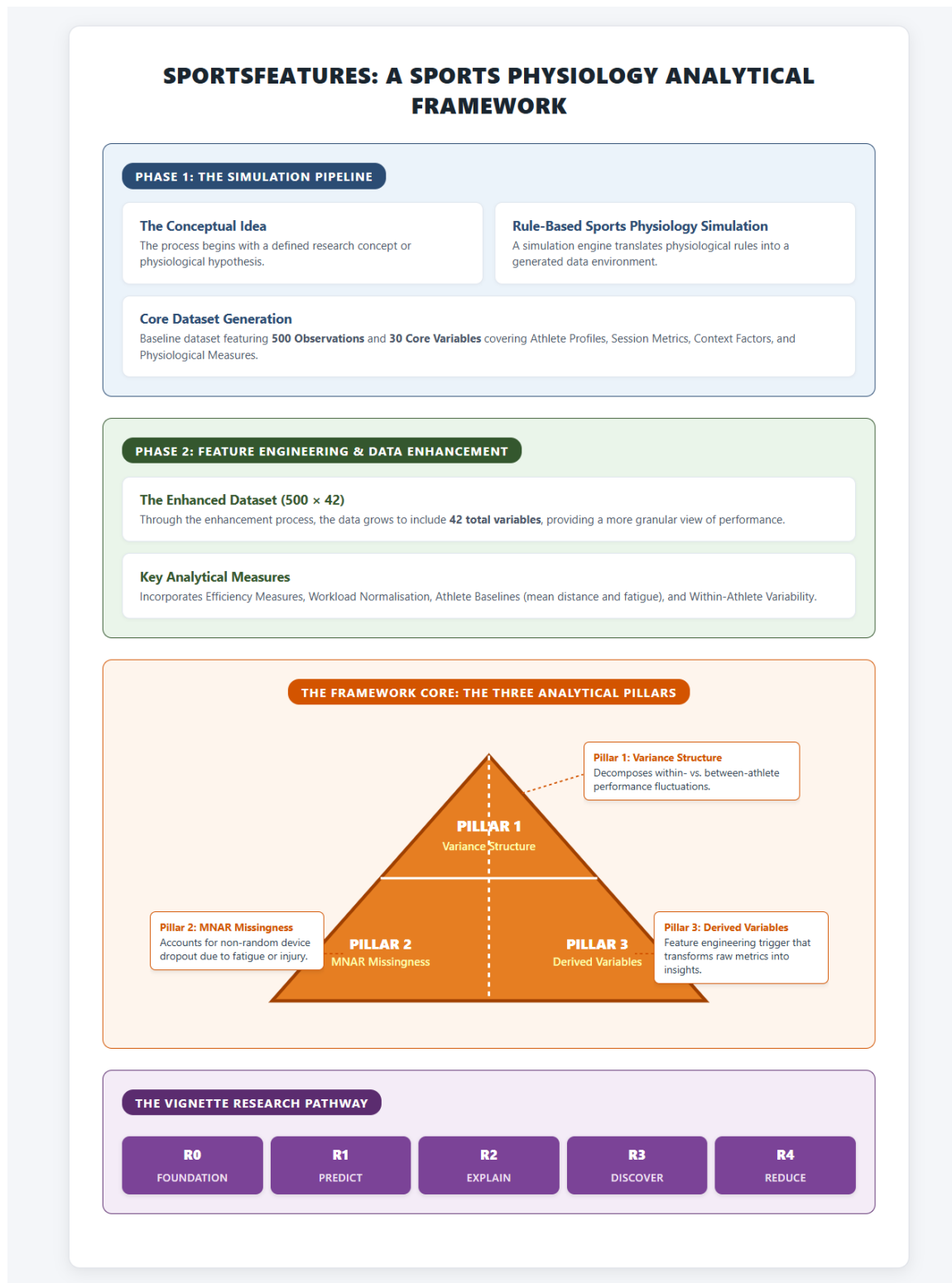
3. Framework Core – Three Analytical Pillars

- **Pillar 1:** Variance Structure – within- vs between-athlete fluctuations
- **Pillar 2:** MNAR Missingness – non-random dropout due to fatigue/injury
- **Pillar 3:** Derived Variables – feature engineering transforming metrics into insights

4. Vignette Research Pathway

R0 – Foundation $\rightarrow$ R1 – Predict $\rightarrow$ R2 – Explain $\rightarrow$ R3 – Discover $\rightarrow$ R4 – Reduce

Figure 1: Conceptual Overview of the R0 Analytical Framework



As shown in Figure 1, the multi-phase structure of the *sportsfeatures* environment illustrates how rule-based simulation, data enhancement, and analytical pillars integrate to form the foundation for the R-series vignettes.

2 Pillar 1: Variance Structure & Hierarchical Non-Constant Spread

Human physiology and athletic behaviour are inherently dynamic and non-linear. When performance metrics are analysed using traditional linear models, non-constant variance (heteroscedasticity) naturally emerges — a reflection of biological complexity rather than a modelling flaw.

2.1 Baseline Illustration

A simple Ordinary Least Squares (OLS) model relating energy expenditure to session volume demonstrates this phenomenon:

$$\text{Calories Burned} = \beta_0 + \beta_1(\text{Distance KM}) + \varepsilon$$

```
df_lm <- df_sports %>%  
  filter(!is.na(calories_burned), !is.na(distance_km))  
  
model_simple <- lm(calories_burned ~ distance_km, data = df_lm)
```

Figure 2: Bivariate Scatterplot of Calories Burned vs Distance KM

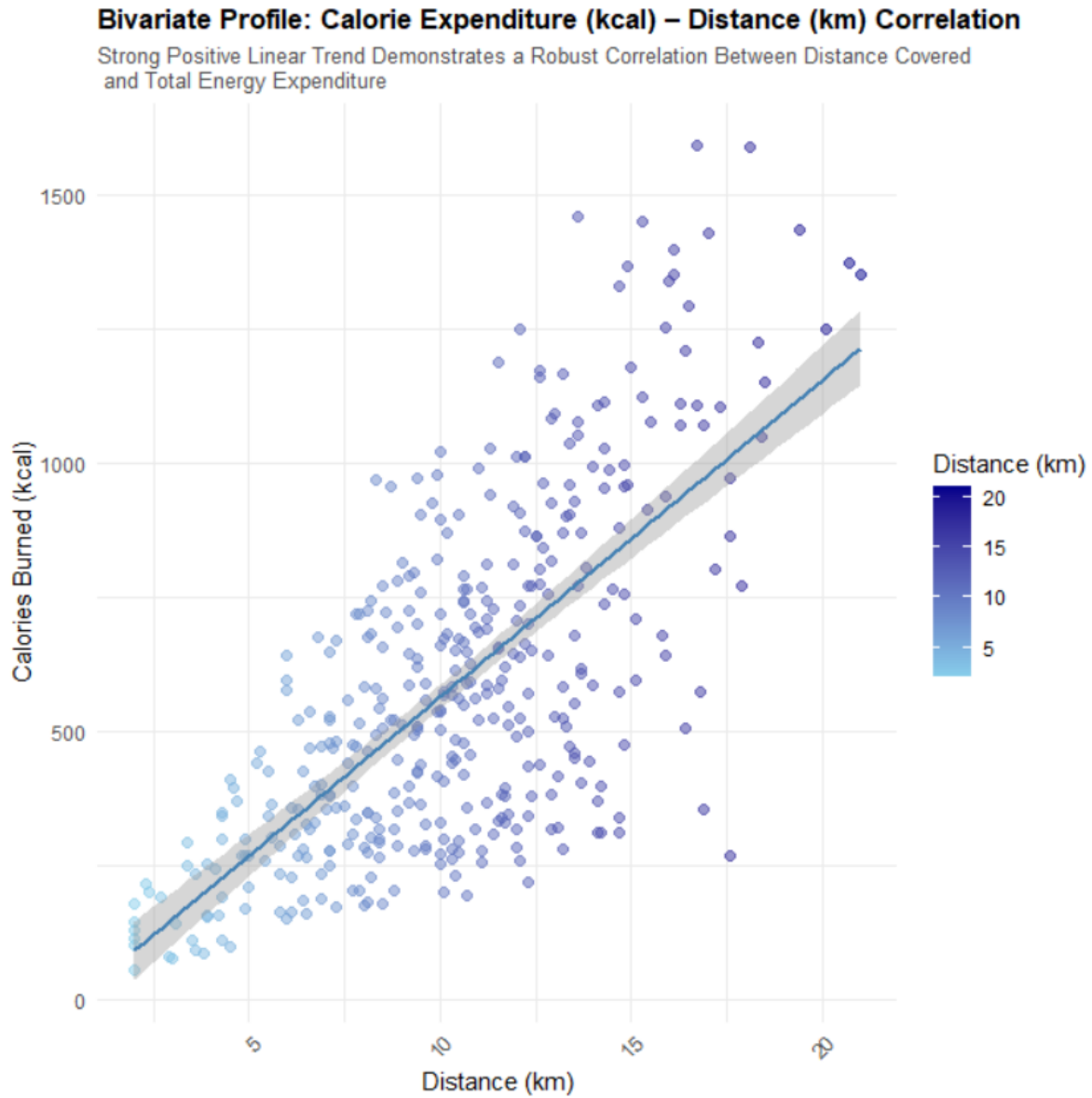
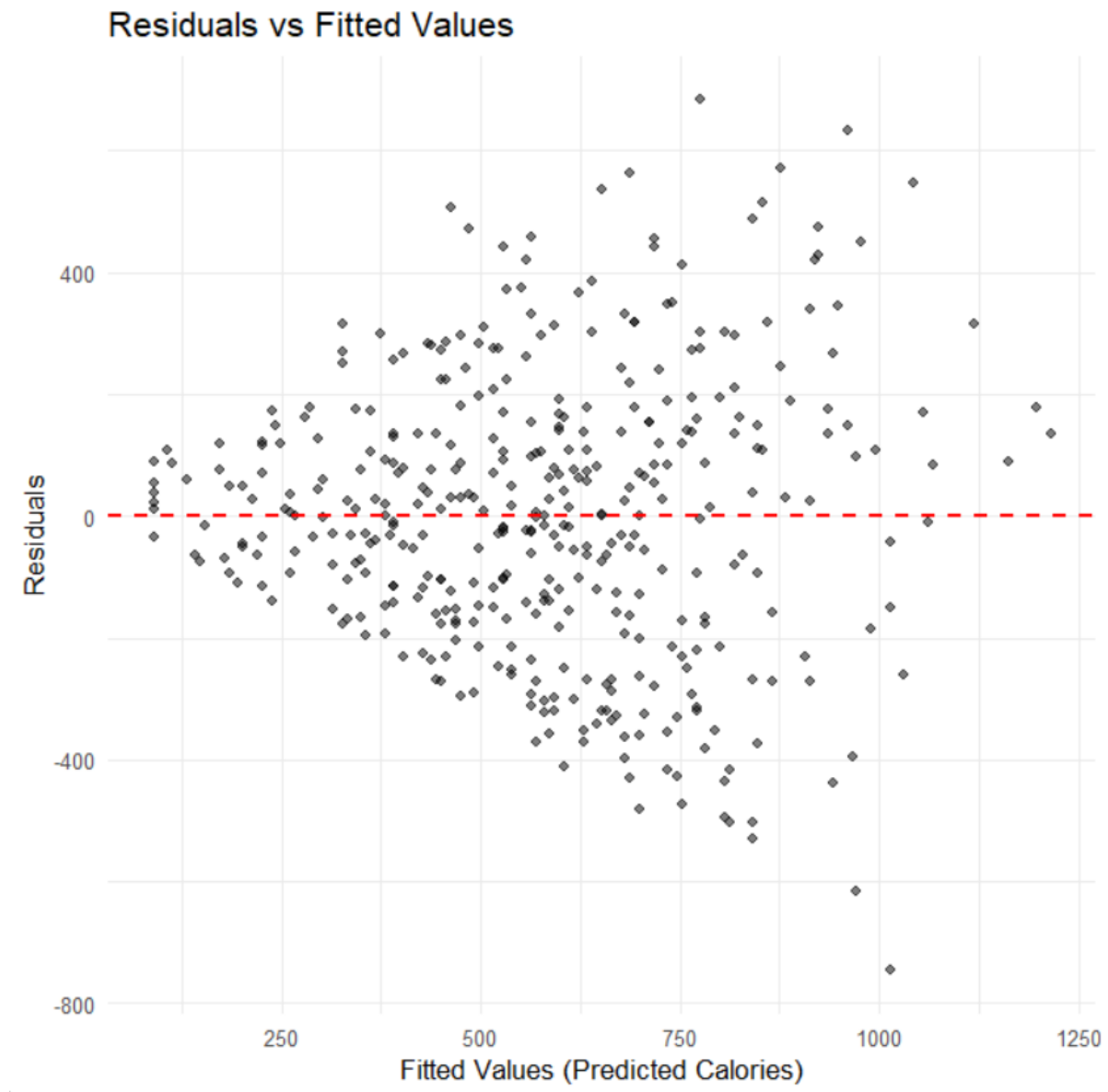


Figure 2 demonstrates a relatively strong positive trend between total energy expenditure (kcal) and distance covered (km). However, Looking closely on the spread of data points, it can be seen the spread is fanning out suggesting the presence of non-constant variance (Heteroscedasticity).

Figure 3 below demonstrates the presence of non-constant variance. The plot also reveals that as training volume increases, the spread of observations widens — signalling the influence of

physiological, behavioural, and environmental factors.

Figure 3: Residual Diagnostics for Simple OLS Model



Based on the residual diagnostics plot (Figure 3), The following is worth noting:

2.2 Hierarchical Variance Decomposition

In the *sportsfeatures* framework, this non-constant spread arises from nested sources of variability:

- **Within-Athlete vs Between-Athlete Variation:** Repeated sessions within individuals differ from those across athletes with distinct fitness baselines and metabolic efficiencies.
- **Contextual & Environmental Drivers:** Terrain, weather, fatigue, hydration, and readiness states introduce additional layers of variance.

Rather than treating these fluctuations as random noise, *sportsfeatures* partitions them through hierarchical and mixed-effects modelling (see R1–R2), enabling structured interpretation of nested dependencies.

This aligns with Figure 1, where *Variance Structure* forms the apex of the framework’s three analytical pillars.

2.2.1 Modelling Implications

Non-constant variance in *sportsfeatures* is not a defect but a hallmark of realism.

Traditional approaches often suppress heteroscedasticity through transformations; here, it is embraced as evidence of biological and behavioural richness.

The framework’s layered design mirrors human performance itself — extendable, hierarchical, and never fully complete - which in turns imply that data are expressions of living systems, not static values.

3 Pillar 2: Missing Not At Random (MNAR) & MICE Imputation

Missing data is an inherent feature of real-world sports telemetry. The *sportsfeatures* framework deliberately embeds a deterministic **Missing Not At Random (MNAR)** mechanism to mirror realistic sensor dropout and athlete fatigue patterns.

3.1 Wearable Device Dropout Mechanics

Whenever `device_type == "none"`, key physiological streams—`heart_rate_avg`, `hydration_status`, and `exhaustion_level`—are absent by design.

This configuration replicates common wearable data gaps caused by device failure, non-compliance, or extreme fatigue.

Across the full dataset, this mechanism introduces **97 missing values (19.4%)** per affected variable, creating a controlled environment for testing imputation strategies.

3.2 Imputation Workflow

To preserve the multivariate coherence of athletic performance data, missing entries are handled using **Multivariate Imputation by Chained Equations (MICE)** with predictive mean matching (PMM):

```
mice_vars <- df_sports %>%  
  select(calories_burned, distance_km, duration_min,  
         heart_rate_avg, exhaustion_level, hydration_status, device_type)  
  
mice_model <- mice(mice_vars, m = 5, maxit = 20,  
                  method = "pmm", seed = 123, quiet = TRUE)
```

This iterative process models each variable conditionally, drawing plausible replacements that maintain physiological realism and statistical consistency.

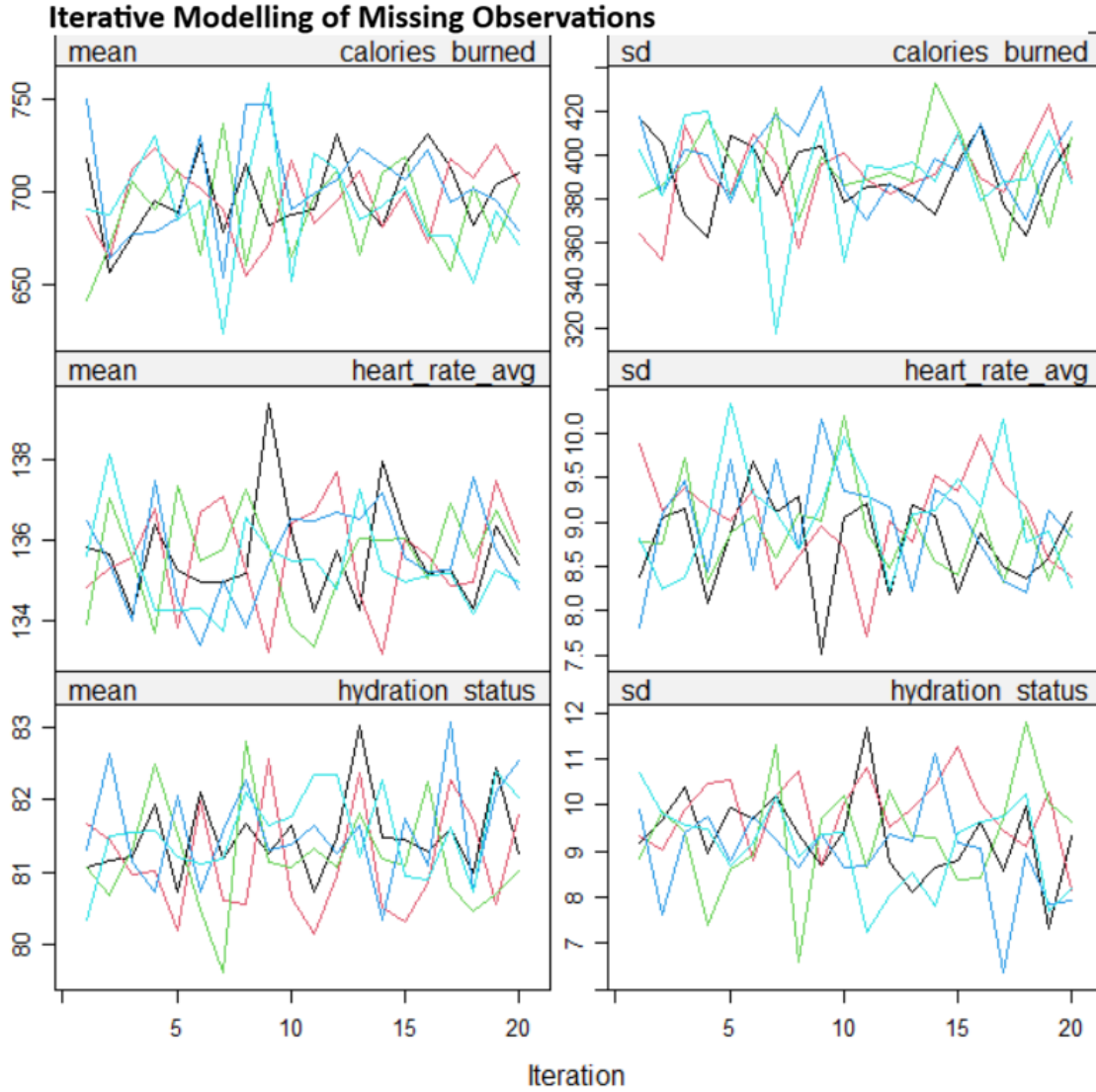
3.3 Diagnostic Verification: Iterative Chain Convergence

The convergence plot (Figure 4) illustrates how the imputation chains stabilize across 20 iterations for key metrics—`calories_burned`, `heart_rate_avg`, and `hydration_status`.

Left panels show mean trajectories; right panels show standard deviations.

Consistent crossover among independent chains indicates healthy mixing and convergence toward a stationary imputation distribution.

Figure 4: Convergence Diagnostics for Mice Imputation Chains



3.4 Interpretation

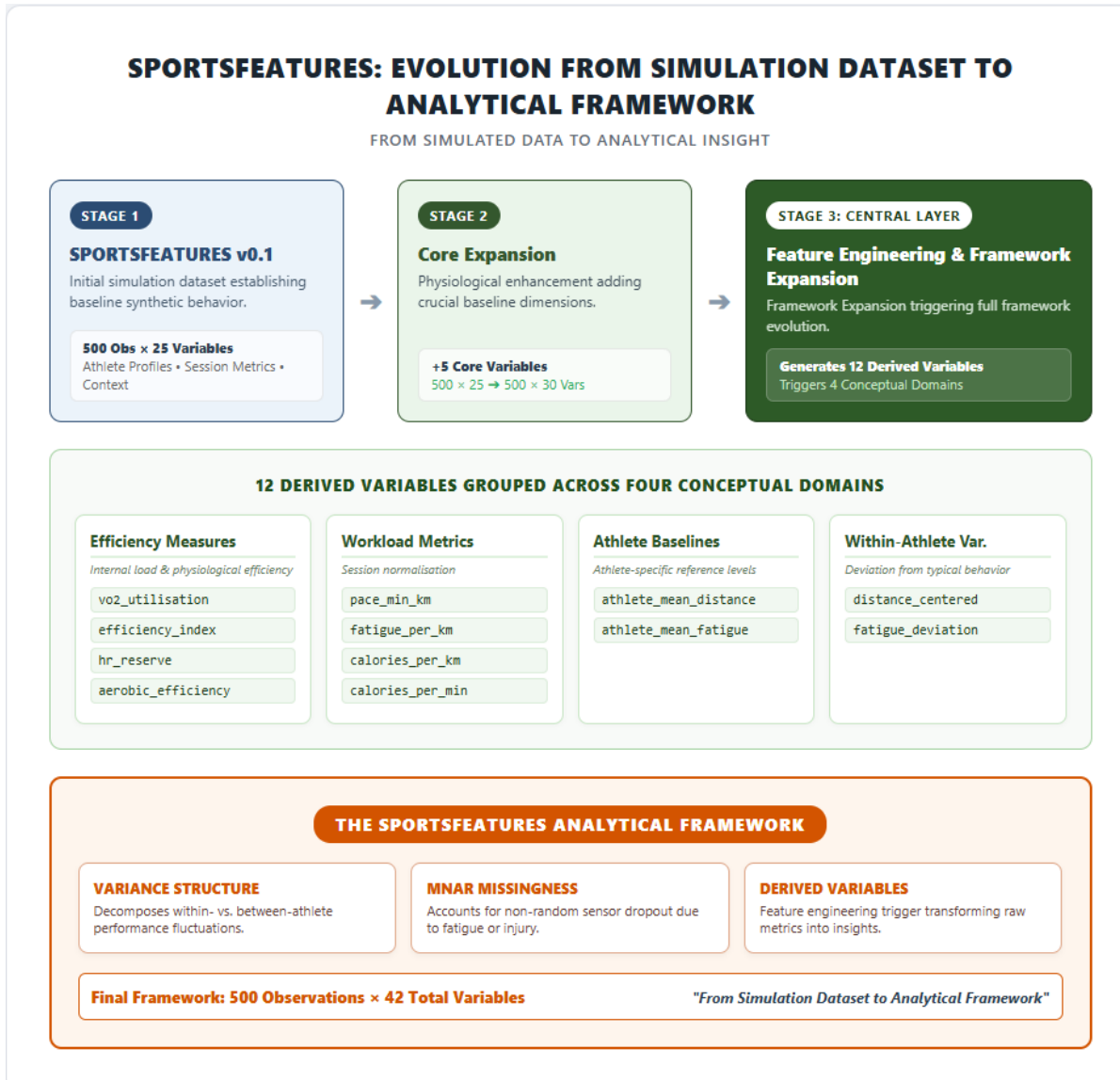
- **Mean & Variance Stabilization:** The imputed streams interweave smoothly, confirming robust chain mixing and absence of systemic drift.
- **Physiological Coherence:** The completed data vectors remain statistically valid and biologically plausible, ensuring unbiased downstream modelling.

- **Framework Integration:** This pillar complements the *Variance Structure* (Pillar 1) by addressing data incompleteness—reinforcing the realism of the *sportsfeatures* simulation pipeline.

4 Pillar 3: Derived Variables & Physiological Constructs

Derived variables form the dynamic layer of the *sportsfeatures* framework. While core variables describe observable training outputs—such as total distance, duration, or calories burned—they do not explain why performance fluctuates. Derived variables bridge this gap by translating raw metrics into interpretable physiological signals.

Figure 5: Evolution from Simulation to Feature Engineering



As illustrated in Figure 5, the framework evolves through three stages:

- Simulation Dataset (Stage 1) establishes baseline synthetic behaviour.
- Core Expansion (Stage 2) adds physiological depth.
- Feature Engineering (Stage 3) introduces twelve derived variables grouped across four conceptual domains—Efficiency Measures, Workload Metrics, Athlete Baselines, and Within-Athlete Variation.

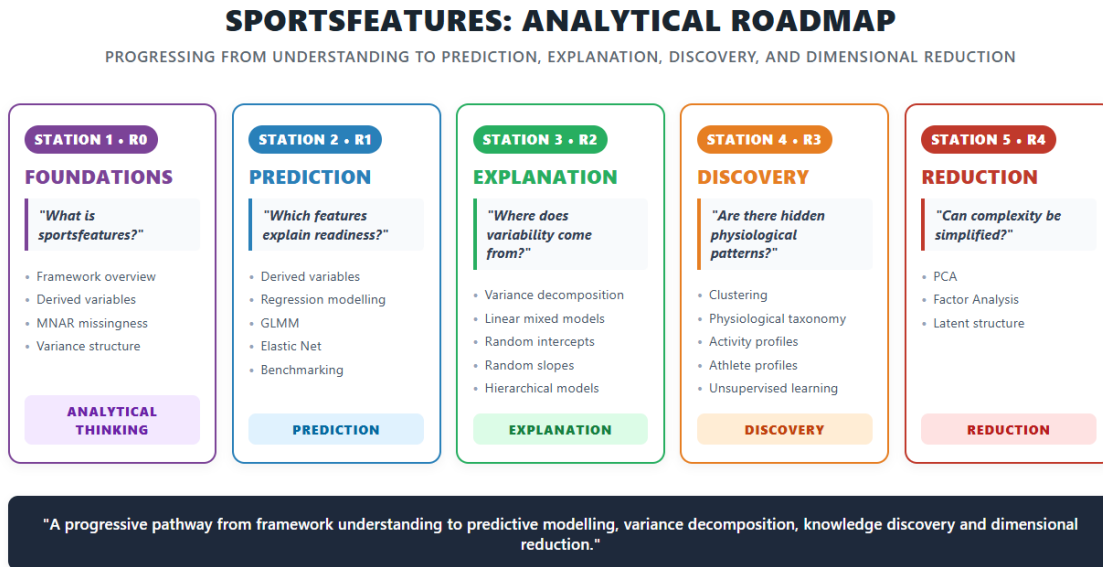
These constructs—such as Heart Rate Reserve, Fatigue Deviation, VO Utilisation, and EfficiencyIndex — act as mechanism generators. They reveal the underlying processes driving variability in human performance, allowing the framework to decompose, model, and interpret fluctuations rather than treat them as noise.

By embedding these derived features, **sportsfeatures** transitions from a dataset to a full analytical framework. Derived variables enable hierarchical modelling, interaction profiling, and advanced statistical exploration across vignettes R1 to R4, completing the triad of pillars that define the system’s analytical architecture. This sets the stage for a four-part empirical progression across **R1 through R4**, as given below

5 What Next? From Framework to Analytical Discovery

The **sportsfeatures** framework establishes a structured foundation that transforms raw athletic telemetry into biological signals. With individual baseline variance partitioned, missingness appropriately mapped, and 12 feature domains fully established, **sportsfeatures** evolves from a dataet into an extensible analytical framework.

Figure 6: Framework to analytical discovery



This sets the stage for a four-part empirical progression across **R1 through R4**, as illustrated in Figure 6:

- **R1 (Predict):** Evaluating mechanical efficiency against athlete readiness using predictive multilevel frameworks.
- **R2 (Explain):** Deconstructing cardiorespiratory cost and partitioning within- versus between-athlete variance.
- **R3 (Discover):** Uncovering hidden physiological taxonomies through unsupervised profiling.
- **R4 (Reduce):** Isolating key latent dimensions across dynamic training variables using factor reduction.