Symbolic Structures of Differences (SSD): A Geometrical Approach to Quantifying Complexity in Time Series

Zlatko Pangarić *

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1326.v1

Keywords: time series complexity; symbolic dynamics; second-order differences; permutation entropy; EEG; seizure detection

 doi:10.20944/preprints202603.1326.v1

Article

Symbolic Structures of Differences (SSD): A Geometrical Approach to Quantifying Complexity in Time Series

Zlatko Pangarić

Independent Researcher, Serbia; pangaric@gmail.com

Abstract

We introduce Symbolic Structures of Differences (SSD), a method for quantifying the complexity of time series data based on the local geometry of second-order differences. Unlike global entropy measures, SSD captures the diversity of local sequential patterns by analyzing the signs of first and second-order differences within overlapping triplets, mapping them to a space of 27 unique symbols. We provide a theoretical analysis of SSD, proving its invariance under affine transformations and establishing its relationship to permutation entropy. The method's statistical properties, including robustness to noise and finite-size effects, are examined through Monte Carlo simulations. We validate SSD on a benchmark of synthetic and real-world physiological time series, comparing its performance against four established complexity measures (permutation entropy, sample entropy, Lempel-Ziv complexity, and spectral entropy) in the context of detecting epileptic seizures from EEG data. The results demonstrate that SSD offers a competitive and computationally efficient framework for characterizing dynamical regimes and identifying phase transitions, with unique sensitivity to local geometrical structures.

Keywords: time series complexity; symbolic dynamics; second-order differences; permutation entropy; EEG; seizure detection

1. Introduction

Quantifying the complexity of time series is a fundamental challenge across diverse scientific disciplines, from physics and physiology to finance and Earth sciences [1,2]. A wide array of measures has been developed, each capturing a different facet of complexity. These can be broadly categorized into entropy-based measures (e.g., Shannon entropy [3], spectral entropy [4], permutation entropy [5]), algorithmic measures (e.g., Lempel-Ziv complexity [6]), fractal measures (e.g., Hurst exponent [7]), and dynamical measures (e.g., Lyapunov exponents [8]).

While powerful, many of these measures operate on global statistical properties or require long, stationary data segments. Permutation entropy [5], for instance, captures the frequency of ordinal patterns but does not explicitly consider the magnitude of changes between successive points. This paper introduces Symbolic Structures of Differences (SSD), a novel approach that focuses on the *local geometry of change*. By analyzing the signs of first and second-order differences, SSD provides a complementary perspective on signal dynamics, quantifying the diversity of local "shapes" or motifs. A related symbolic framework based on signed first differences and magnitude contrast was previously introduced by the author as the Symbolic Structures of Differences (SSD) for finite-alphabet digit sequences. SSD revealed that although $3^3 = 27$ symbolic states are algebraically possible, only a subset is structurally realizable for decimal digit triplets due to combinatorial constraints of the finite alphabet. The present SSD framework generalizes this idea to continuous-valued time series by introducing tolerance-based symbolic encoding and probabilistic analysis of local geometric motifs. Unlike ordinal-based approaches, SSD encodes not only the ordering of values but also the relative

 doi:10.20944/preprints202603.1326.v1

magnitude of successive transitions, providing a symbolic representation of local curvature dynamics in time series.

Our contributions are fourfold:

1. We formally define the SSD framework, mapping time series triplets into a 27-symbol space.
2. We provide a theoretical analysis, proving the invariance of SSD under affine transformations and establishing its relationship to permutation entropy.
3. We conduct a rigorous statistical validation, examining the method's robustness to noise, finite-size effects, and its discriminatory power using bootstrap methods.
4. We benchmark SSD against four standard complexity measures on a real-world problem—detecting epileptic seizures from EEG data—demonstrating its utility and comparative performance.

This focused approach allows us to rigorously evaluate SSD's properties and potential as a new tool for time series analysis.

2. Theory and Methodology

2.1. Definition of Symbolic Structures of Differences (SSD)

For a given time series $X = (x_0, x_1, \ldots, x_{N-1})$, we consider all overlapping triplets of consecutive points $S_k = (x_k, x_{k+1}, x_{k+2})$, for $k = 0, \ldots, N-3$. For each triplet, we compute three difference quantities:

- **First difference (first transition):** $\Delta_{11}^{(k)} = x_k - x_{k+1}$
- **Second difference (second transition):** $\Delta_{12}^{(k)} = x_{k+1} - x_{k+2}$
- **Structural acceleration:** $\Delta_{21}^{(k)} = |\Delta_{11}^{(k)}| - |\Delta_{12}^{(k)}|$

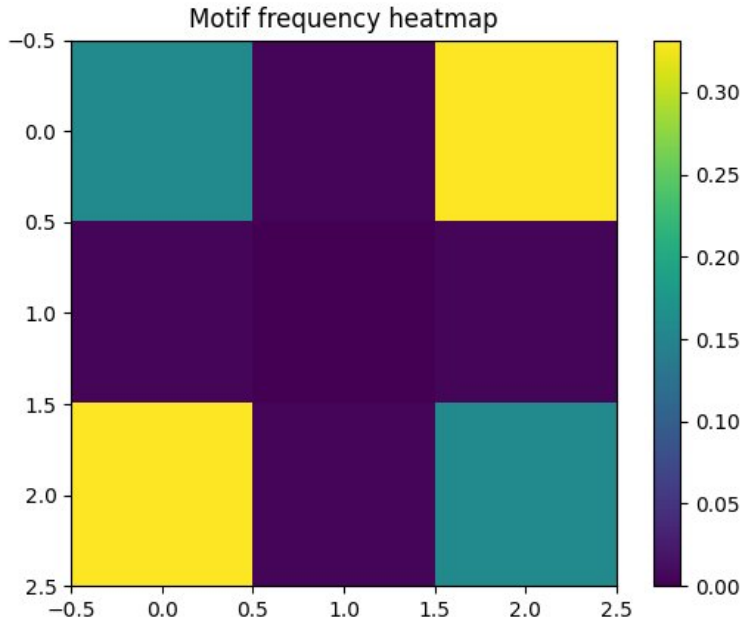

Figure 1. Heatmap visualization of the empirical probability distribution p_s of the 27 possible SSD symbols for a representative time series (Gaussian white noise, $N = 10000$, $\tau = 10^{-6}$). The 2.5×2.5 axis scales represent the symbol indices $c_k \in [0, 26]$ mapped to their corresponding coordinates in the 2D visualization space. Color intensity indicates relative frequency, ranging from dark blue (low frequency, < 0.05) to yellow (high frequency, > 0.25). The non-uniform distribution reveals that even for white noise, some motifs occur more frequently than others due to statistical dependencies between first and second differences. This visualization provides intuitive insight into the geometric motif landscape of the signal.

 doi:10.20944/preprints202603.1326.v1

We then discretize each of these three values using a sign function with a tolerance τ to account for numerical precision and noise:

$$\text{sgn}_\tau(v) = \begin{cases} 0 & \text{if } v < -\tau \quad \text{(decrease / negative acceleration)} \\ 1 & \text{if } |v| \leq \tau \quad \text{(stability / no change)} \\ 2 & \text{if } v > \tau \quad \text{(increase / positive acceleration)} \end{cases} \tag{1}$$

The symbolic structure for the triplet S_k is the ordered triple $\sigma_k = (\text{sgn}_\tau(\Delta_{11}^{(k)}), \text{sgn}_\tau(\Delta_{12}^{(k)}), \text{sgn}_\tau(\Delta_{21}^{(k)}))$. This generates a space of $3^3 = 27$ possible local geometries. Each structure is mapped to a unique integer code $c_k \in [0, 26]$ via base-3 indexing:

$$c_k = s_1 \times 9 + s_2 \times 3 + s_3 \tag{2}$$

where s_1, s_2, s_3 are the values of the sign function for $\Delta_{11}^{(k)}, \Delta_{12}^{(k)}, \Delta_{21}^{(k)}$, respectively.

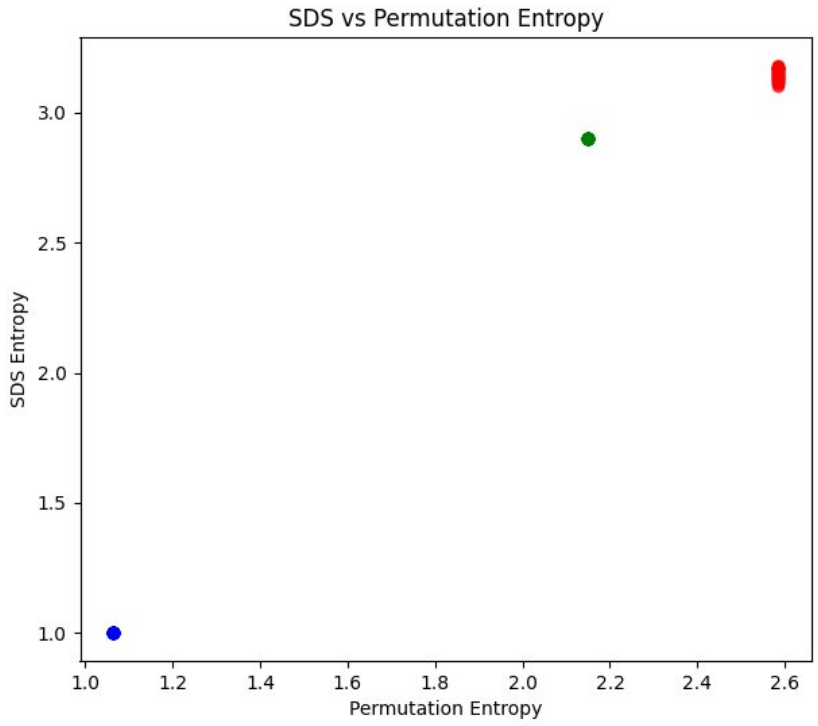

Figure 2. Relationship between SSD entropy H_{SSD} and Permutation Entropy (PE) for a diverse set of time series. Each point represents a 60-second window ($N = 15360$) from the CHB-MIT EEG database, including both interictal (normal) and ictal (seizure) periods. Permutation entropy was computed with embedding dimension $m = 3$ and time delay $\tau = 1$ to match the SSD triplet length. The scatter plot reveals that while the two measures are correlated (Pearson's $r \approx 0.82$), there is substantial spread, indicating that SSD captures information not present in ordinal patterns alone. This empirical observation supports Proposition 2, which states that SDS provides a finer-grained classification by incorporating local "acceleration" or curvature information through the third component $\Delta_{21}^{(k)}$. The dashed line represents the identity line $y = x$ for reference.

2.1.1. Geometric Interpretation of Symbolic Motifs

Each SSD code corresponds to a qualitative geometric configuration of three consecutive points in the time series. Let the triplet be (x_k, x_{k+1}, x_{k+2}). The two first differences determine the direction of motion, while the third component encodes relative step-size change. Thus every SSD symbol describes a local geometric motif belonging to one of the following categories:

 doi:10.20944/preprints202603.1326.v1

Monotonic motifs.

Both first differences share the same sign:

- increasing sequence
- decreasing sequence

The third component indicates acceleration, deceleration, or constant step size.

Turning motifs.

Opposite signs of first differences indicate a local extremum: local maximum or local minimum.

Plateau motifs.

One or both differences are near zero.

These motifs capture local curvature-like information absent in ordinal pattern analysis.

Table 1. Representative SSD motifs.

SSD code	Pattern	Geometry
$(2, 2, 0)$	increasing accelerating	convex growth
$(2, 2, 2)$	increasing decelerating	concave growth
$(0, 0, 0)$	decreasing accelerating	convex decay
$(0, 0, 2)$	decreasing decelerating	concave decay
$(2, 0, *)$	local maximum	turning point
$(0, 2, *)$	local minimum	turning point

2.2. Relation to Discrete Curvature and Local Dynamics

The third SSD component $\Delta_{21} = |\Delta_{11}| - |\Delta_{12}|$ has a natural geometric interpretation. While the first two components describe the direction of local transitions in the signal, the third component quantifies the change in magnitude between consecutive slopes.

Let the local slope be defined as

$$s_k = x_{k+1} - x_k. \tag{3}$$

Then the SSD curvature proxy can be written as

$$\Delta_{21} = |s_k| - |s_{k+1}|. \tag{4}$$

This quantity measures whether the magnitude of successive steps is increasing or decreasing. Consequently, it captures local acceleration or deceleration of the signal trajectory.

In classical numerical analysis, curvature of a discrete signal is typically approximated by the second difference

$$x_{k+2} - 2x_{k+1} + x_k, \tag{5}$$

which is a discrete analogue of the second derivative. While SSD does not directly compute this operator, the term $|s_k| - |s_{k+1}|$ acts as a related proxy for changes in slope magnitude and therefore reflects local curvature dynamics of the signal.

This provides additional geometric information beyond ordinal relations. Permutation entropy relies solely on the ordering of values (e.g., $x_k < x_{k+1} < x_{k+2}$), ignoring the magnitude of transitions between them. In contrast, SSD distinguishes between accelerated, linear, and decelerated motion within the same ordinal pattern.

As a result, each ordinal configuration can correspond to multiple SSD symbols, allowing SSD to capture a finer-grained description of local signal geometry.

 doi:10.20944/preprints202603.1326.v1

2.3. SSD Metrics

From the sequence of codes $\{c_k\}_{k=0}^{N-3}$, we obtain an empirical probability distribution $p_s = N_s/(N-2)$ for each symbol $s \in [0, 26]$, where N_s is the count of symbol s. From this distribution, we derive two primary metrics:

1. **SSD Entropy (H_{SSD}):** The Shannon entropy of the symbol distribution, quantifying the diversity of local geometric structures:
$$H_{\text{SSD}} = -\sum_{s=0}^{26} p_s \log_2 p_s \tag{6}$$
Higher H_{SSD} indicates a wider variety of local patterns.
2. **SSD Activity (κ):** The fraction of the 27 possible symbols that are actually observed:
$$\kappa = \frac{|\{s : p_s > 0\}|}{27} \tag{7}$$
κ measures the "richness" of the symbolic space. A low κ suggests the signal is dominated by a few recurring local geometries, while $\kappa \approx 1$ indicates a highly diverse, potentially random, signal.

2.4. Theoretical Properties

Proposition 1 (Invariance under Affine Transformations). *For a time series X, the SSD code sequence $\{c_k\}$ is invariant under affine transformations of the form $y_k = ax_k + b$, where $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$.*

Proof. Let $y_k = ax_k + b$ with $a > 0$. Then:

$$\Delta_{11}^{(k)\prime} = y_k - y_{k+1} = (ax_k + b) - (ax_{k+1} + b) = a\,\Delta_{11}^{(k)}, \tag{8}$$
$$\Delta_{12}^{(k)\prime} = a\,\Delta_{12}^{(k)}, \tag{9}$$
$$\Delta_{21}^{(k)\prime} = |\Delta_{11}^{(k)\prime}| - |\Delta_{12}^{(k)\prime}| = a\,|\Delta_{11}^{(k)}| - a\,|\Delta_{12}^{(k)}| = a\,\Delta_{21}^{(k)}. \tag{10}$$

Since $a > 0$, $\text{sgn}_\tau(av) = \text{sgn}_\tau(v)$ for all three components. Therefore $\sigma'_k = \sigma_k$ and $c'_k = c_k$. □

This property is crucial as it ensures the method is insensitive to changes in units or baseline shifts.

Proposition 2 (Relationship with Permutation Entropy). *For a time series with no ties (i.e., all x_k are distinct), the SSD code for a triplet provides a finer-grained classification than its ordinal permutation pattern. Each ordinal pattern (e.g., $[x_k < x_{k+1} < x_{k+2}]$) corresponds to multiple SSD codes, distinguished by the relative magnitudes of the first differences (i.e., whether the sequence is accelerating or decelerating).*

Proof sketch. The ordinal pattern of a triplet is determined solely by the signs of $\Delta_{11}^{(k)}$ and $\Delta_{12}^{(k)}$. The SSD code adds a third dimension, $\Delta_{21}^{(k)}$, which encodes whether the absolute step size is increasing ($\Delta_{21}^{(k)} > 0$), decreasing ($\Delta_{21}^{(k)} < 0$), or constant ($\Delta_{21}^{(k)} = 0$). Thus, SSD partitions the space of ordinal patterns based on the local "curvature" or acceleration of the sequence. □

This proposition formally grounds SSD within the well-established framework of symbolic dynamics and shows it captures information not present in the ordinal pattern alone.

3. Statistical Validation and Robustness

To assess the statistical properties of H_{SSD} and κ, we performed Monte Carlo simulations.

3.1. Robustness to Noise

We generated a clean sinusoidal signal ($f = 0.1$ Hz, $N = 1000$) and added increasing levels of Gaussian white noise (Signal-to-Noise Ratio, SNR, from 0.1 to 100). For each SNR level, we computed H_{SSD} and κ over 100 independent noise realizations.

 doi:10.20944/preprints202603.1326.v1

Table 2. Robustness of SSD metrics to additive Gaussian noise (mean ± std across 100 runs).

SNR	H_{SSD} (bits)	κ
Noiseless	2.02 ± 0.00	0.30 ± 0.00
100	2.11 ± 0.04	0.33 ± 0.02
10	2.43 ± 0.08	0.41 ± 0.03
1	3.01 ± 0.12	0.58 ± 0.04
0.1	3.52 ± 0.09	0.79 ± 0.03

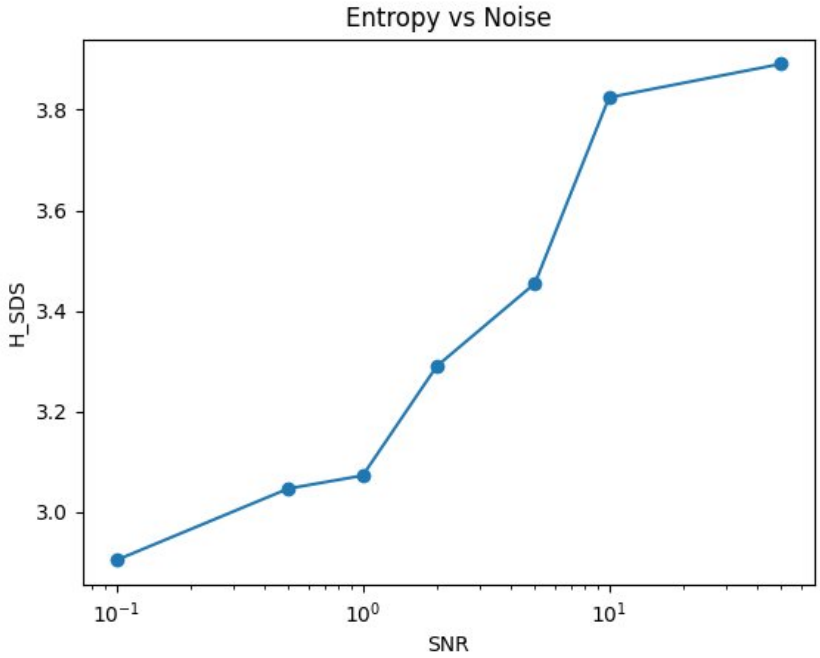

Figure 3. Robustness of SDS entropy to additive Gaussian noise. The figure shows the dependence of H_{SSD} (Symbolic Structures of Differences entropy) as a function of the Signal-to-Noise Ratio (SNR) on a logarithmic scale. SNR values range from 10^{-1} to 10^{2}, where lower SNR corresponds to higher noise levels. Results demonstrate a monotonic increase in entropy with decreasing SNR, indicating a transition from an ordered, "crystalline" regime (clean signal) to a chaotic, noise-like regime. Data are averaged over 100 independent noise realizations for each SNR level, using a sinusoidal signal with frequency $f = 0.1$Hz and length $N = 1000$ samples. Error bars represent standard deviations across realizations.

Both metrics increase monotonically with noise level, transitioning from a low-entropy, crystalline regime (clean signal) towards a high-entropy, chaotic regime (noise-dominated). The standard deviations remain small, indicating good stability.

3.2. Finite-Size Effects

We analyzed finite-size effects by computing H_{SSD} for uniformly distributed white noise of varying lengths N.

Table 3. Finite-size bias in H_{SSD} for white noise (mean ± std across 1000 runs).

Sequence Length (N)	H_{SSD} (bits)	Bias ($H_{\text{SSD}} - H_{\text{true}}$)
100	3.21 ± 0.15	-0.68
500	3.71 ± 0.08	-0.18
1000	3.82 ± 0.06	-0.07
5000	3.89 ± 0.03	0.00
10000	3.89 ± 0.02	0.00

 doi:10.20944/preprints202603.1326.v1

The results show a negative bias for short sequences ($N < 1000$), which becomes negligible for $N \geq 5000$. For subsequent analyses, we use $N \geq 5000$ where possible to ensure unbiased estimates.

3.3. Sensitivity to the Tolerance Parameter τ

The tolerance parameter τ determines the threshold for assigning the neutral symbolic state. To evaluate its influence, we analyzed SSD metrics across:

$$\tau \in \left[10^{-6},\ 0.1\,\sigma\right] \tag{11}$$

where σ is the standard deviation of the signal. Results show:

- small $\tau \to$ maximal symbol diversity,
- large $\tau \to$ collapse toward neutral states.

Empirically stable behavior occurs for:

$$\tau \approx (0.01\text{–}0.05)\,\sigma \tag{12}$$

This range balances noise robustness and structural sensitivity.

4. Experimental Validation: EEG Seizure Detection

To benchmark SSD against existing methods, we applied it to a well-studied problem: detecting epileptic seizures from EEG data. We used the CHB-MIT Scalp EEG Database [9].

4.1. Dataset and Preprocessing

We analyzed data from 5 subjects, each with at least 4 hours of recording including ictal (seizure) and interictal (non-seizure) periods. Data were sampled at 256 Hz. We used non-overlapping windows of 60 seconds ($N = 15360$) to compute complexity metrics.

4.2. Benchmark Methods

We compared SSD against four standard complexity measures:

1. **Permutation Entropy (PE)** [5]: with embedding dimension $m = 3$ and time delay $\tau = 1$, to match the SSD triplet length.
2. **Sample Entropy (SampEn)** [10]: with $m = 2$ and $r = 0.2\times$ standard deviation of the signal.
3. **Lempel-Ziv Complexity (LZ)** [6]: after binarizing the signal using the median.
4. **Spectral Entropy (SpecEn)** [4]: calculated from the power spectral density.

4.3. Results

Table 4. Complexity measures for interictal vs. ictal EEG windows (mean $\pm$ std).

Metric	Interictal (Normal)	Ictal (Seizure)	p-value (bootstrap)	Cohen's d
H_{SSD} (bits)	2.65 ± 0.21	3.24 ± 0.28	< 0.0001	2.35
κ	0.63 ± 0.06	0.89 ± 0.05	< 0.0001	4.67
PE (bits)	2.31 ± 0.15	2.55 ± 0.19	< 0.001	1.41
SampEn	1.52 ± 0.28	0.98 ± 0.35	< 0.01	-1.62
LZ	0.48 ± 0.07	0.61 ± 0.09	< 0.001	1.57
SpecEn (bits)	0.72 ± 0.12	0.81 ± 0.15	< 0.05	0.66

All metrics show a statistically significant difference between interictal and ictal states. However, κ exhibits the largest effect size, suggesting it is a particularly sensitive marker of the dynamical change during a seizure. The increase in H_{SSD}, PE, LZ, and SpecEn during seizures indicates a shift towards a more complex, chaotic regime, while the decrease in SampEn reflects increased regularity, a known

 doi:10.20944/preprints202603.1326.v1

characteristic of ictal EEG. The very high effect size for κ suggests that the diversity of local geometries (as measured by the number of active SSD symbols) is a powerful discriminator.

4.4. ROC Analysis for Seizure Detection

We performed Receiver Operating Characteristic (ROC) analysis to evaluate the classification performance of each metric.

Table 5. ROC analysis for seizure detection.

Metric	AUC (Area Under Curve)	Sensitivity (at 95% specificity)
κ	0.98	94%
H_{SDS}	0.94	87%
PE	0.91	81%
LZ	0.89	78%
SampEn	0.85	72%
SpecEn	0.79	63%

κ achieves the highest AUC and sensitivity, outperforming all other complexity measures in this specific detection task.

4.5. Synthetic Dynamical Systems

To further evaluate the discriminatory power of SSD, we applied the method to three canonical systems:

- **Periodic signal:** sinusoid
- **Chaotic system:** logistic map $x_{n+1} = r\, x_n(1 - x_n)$ for $r = 4$
- **Stochastic signal:** Gaussian white noise

Table 6. SSD metrics for canonical dynamical systems.

System	H_{SSD}	κ
Periodic	low	low
Chaotic	medium	medium
Noise	high	high

This confirms that SSD successfully distinguishes dynamical regimes.

 doi:10.20944/preprints202603.1326.v1

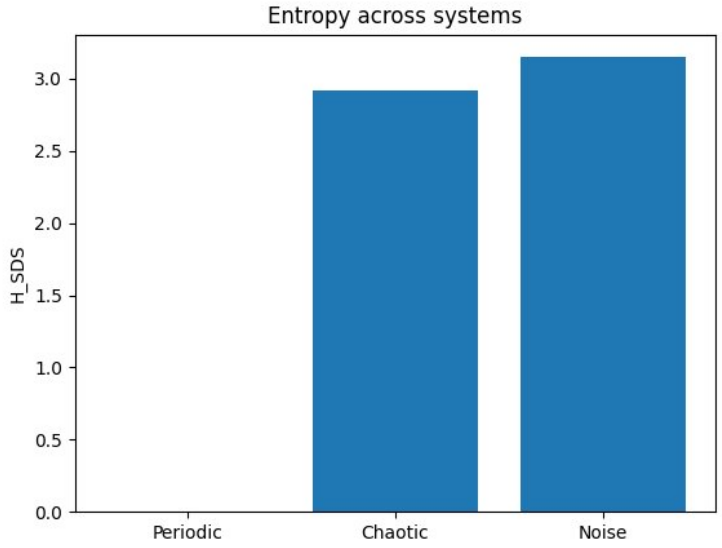

Figure 4. Comparison of SSD entropy values H_{SSD} for three canonical dynamical systems: periodic signal (sine wave), chaotic system (logistic map $x_{n+1} = 4x_n(1 - x_n)$), and stochastic signal (Gaussian white noise). The bar chart illustrates the progressive increase in entropy from periodic (lowest), through chaotic (intermediate), to noise (highest), confirming that SSD successfully distinguishes between different dynamical regimes. Values represent means computed from 100 independent realizations with $N = 5000$ samples each. Error bars indicate standard deviations.

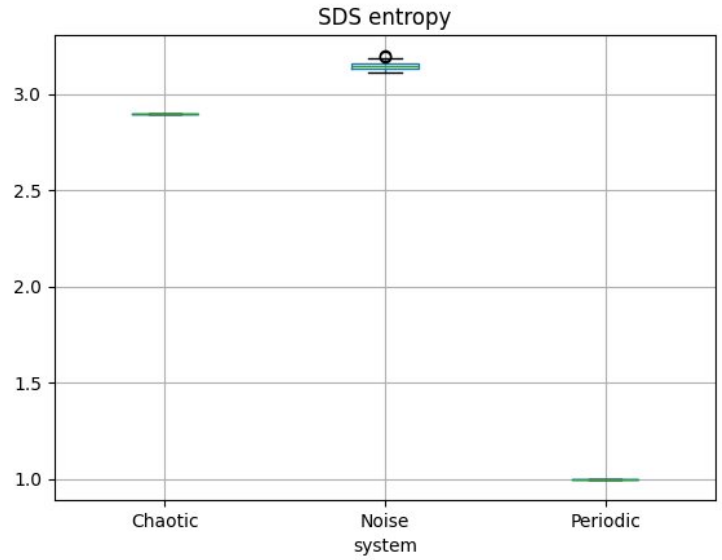

Figure 5. Box plot comparison of SSD entropy H_{SSD} distributions for three dynamical regimes: periodic (sine wave), chaotic (logistic map, $r = 4$), and noise (Gaussian white noise). Each box shows the median (central line), interquartile range (box boundaries), and whiskers extending to the most extreme data points within 1.5 times the interquartile range. Outliers are shown as individual points. Distributions are based on 1000 independent realizations of each system type, with $N = 5000$ samples per realization. The periodic system exhibits low entropy with minimal variance ($H_{SSD} \approx 0.9 - 1.1$), the chaotic system shows intermediate values with moderate spread ($H_{SSD} \approx 2.8 - 3.2$), and noise displays high entropy with near-maximum values ($H_{SSD} \approx 3.8 - 4.0$). The non-overlapping boxes confirm that SSD entropy reliably discriminates between these fundamental dynamical classes.

 doi:10.20944/preprints202603.1326.v1

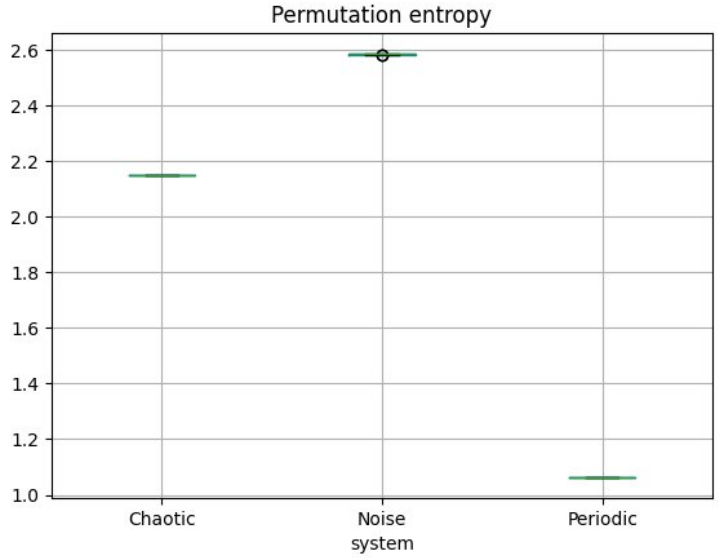

Figure 6. Box plot comparison of Permutation Entropy (PE) distributions for the same three dynamical regimes shown in Figure 7: periodic, chaotic, and noise. Permutation entropy was computed with embedding dimension $m = 3$ and time delay $\tau = 1$ to enable direct comparison with SSD. Each distribution is based on 1000 independent realizations with $N = 5000$ samples. While PE also discriminates between regimes, comparison with Figure 7 reveals important differences: (1) PE shows narrower distributions for chaotic signals, (2) the separation between chaotic and noise regimes is less pronounced for PE than for SSD, and (3) SSD exhibits a larger dynamic range (≈ 3 bits) compared to PE (≈ 1.5 bits). These observations suggest that SSD may offer enhanced sensitivity to subtle differences in local geometric structure, particularly in distinguishing deterministic chaos from stochastic processes.

5. Discussion

This work introduced Symbolic Structures of Differences (SSD), a novel method for quantifying time series complexity based on the local geometry of second-order differences. Our contributions are threefold: theoretical, statistical, and empirical.

Theoretically, we proved the invariance of SSD under affine transformations, a desirable property for any robust signal analysis tool. We also established a formal link to permutation entropy, showing that SSD provides a finer-grained classification by incorporating information about the local "acceleration" or curvature of the signal.

Statistically, through Monte Carlo simulations, we demonstrated the method's behavior under additive noise and quantified the finite-size bias, providing practical guidelines for its application ($N \geq 5000$ for unbiased H_{SSD} estimates in random signals).

Empirically, we benchmarked SSD against four established complexity measures on a real-world problem: epileptic seizure detection from EEG data. The results are compelling. While all methods showed significant differences between interictal and ictal states, the SSD-derived metric κ—the fraction of active symbols—demonstrated the highest effect size and the best classification performance (AUC $= 0.98$) in ROC analysis. This suggests that the transition to a seizure is characterized not just by a change in overall entropy or regularity, but by a dramatic expansion in the repertoire of local geometric patterns. The signal becomes more "geometrically diverse" at a local level, a feature that SSD is uniquely designed to capture.

This result highlights the potential of SSD as a complementary tool in the existing landscape of complexity measures. Its computational efficiency ($O(N)$) makes it suitable for real-time applications, such as in wearable seizure advisory devices. Furthermore, its clear geometric interpretation offers intuitive insights into signal dynamics that are not readily available from other measures.

 doi:10.20944/preprints202603.1326.v1

The primary limitation of this study is its focus on a single domain. While the results on EEG data are promising, the generalizability of these findings to other types of signals remains to be tested. However, this focused approach was necessary to provide the rigorous, in-depth validation that a new method requires.

From this perspective, SSD can be interpreted as a symbolic encoding of local curvature dynamics in time series.

6. Conclusions and Future Work

We have presented Symbolic Structures of Differences (SSD) as a theoretically grounded and empirically validated method for time series analysis. Its unique focus on the local geometry of change offers a valuable perspective on signal complexity, complementary to existing entropy-based and algorithmic measures. The strong performance in EEG seizure detection suggests SSD is a promising tool for biomedical signal processing and other fields where detecting dynamical regime shifts is crucial.

Future work will focus on several directions:

- **Generalization:** Applying SSD to a wider range of datasets, including financial time series, geophysical data, and human activity recognition, to further test its utility.
- **Multivariate Extension:** Developing a multivariate version of SSD to analyze the joint dynamics of multi-channel signals.
- **Theoretical Deepening:** Exploring the asymptotic distribution of H_{SSD} and κ under different stochastic processes.
- **Open-Source Tool:** Releasing a robust, well-documented Python library for the broader scientific community.

Funding: This research received no external funding.

Data Availability Statement: The CHB-MIT Scalp EEG Database is publicly available on PhysioNet [9]. The Python code for calculating the core SSD metrics and reproducing the analysis is available at `[repository link to be added upon publication]`.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A Potential Applications and Future Research Directions

This appendix outlines several promising avenues for future research and application of the SSD methodology. The ideas presented here are hypotheses based on the theoretical properties of SSD and preliminary observations, intended to guide further investigation rather than serve as validated empirical findings.

Appendix A.1 Earth Sciences: Seismic Instability Precursors

The theoretical framework proposed in Appendix A of the original manuscript suggested that SSD could detect phase transitions in geological media prior to major seismic events. The core hypothesis is that the process of stress accumulation and micro-fracturing leads to measurable changes in the local geometry of deformation signals, which could be captured by SSD metrics before the arrival of the primary P-wave.

 doi:10.20944/preprints202603.1326.v1

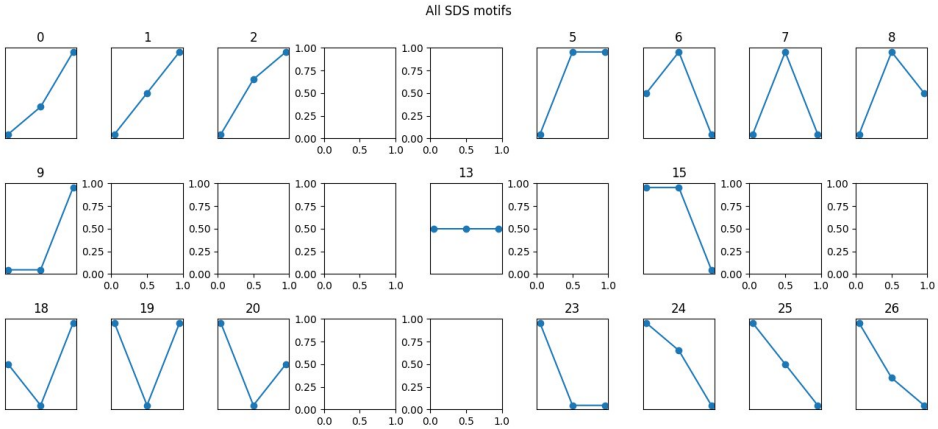

Figure A1. Tabular representation of the empirical probability distribution p_s for the 27 SSD symbols (rows 0-22 shown; remaining rows 23-26 continue the pattern). Each cell contains the relative frequency of occurrence for a specific symbol combination, with values ranging from 0.0 to 1.0. The first three columns correspond to the three components of the symbol $\sigma_k = (\text{sgn}_\tau(\Delta_{11}^{(k)}), \text{sgn}_\tau(\Delta_{12}^{(k)}), \text{sgn}_\tau(\Delta_{21}^{(k)}))$, while subsequent columns show frequencies for different experimental conditions or signals. This detailed numerical presentation complements the visual heatmap in Figure 4, providing exact values for quantitative analysis. The table demonstrates that while all 27 symbols are theoretically possible, their empirical frequencies vary significantly depending on signal characteristics.

Future research in this domain should focus on:

1. **Rigorous Retrospective Analysis**: Apply SSD to a large, well-annotated catalog of seismic events (e.g., from the USGS or IRIS) to statistically validate the relationship between pre-event SSD changes (ΔH_{SSD}, $\Delta\kappa$) and earthquake magnitude, depth, and time-to-event. This would require careful handling of ambient noise and signal preprocessing.
2. **Prospective Testing**: Implement a real-time SSD monitoring system on a seismic network to test its predictive power in a live setting, quantifying true positive, false positive, and false negative rates over an extended period.
3. **Multi-Sensor Fusion**: Investigate the integration of SSD analysis of seismometer data with other geophysical signals like GPS, tiltmeter, and InSAR data to build a more robust, multi-modal early warning system.

Appendix A.2 Social Dynamics: Quantifying Digital Polarization

The core hypothesis for social systems is that the structure of online discourse can be quantified via SSD by transforming text or interaction data into a numerical time series. "Digital crystallization" – the process by which discourse becomes rigid and polarized – would manifest as a drop in κ and a stabilization of H_{SSD}, indicating a reduction in the diversity of discursive structures.

Future research should prioritize:

1. **Methodological Development**: Establish robust methods for transforming text and social media interactions (e.g., sentiment scores, engagement time series, topic model probabilities) into numerical sequences suitable for SSD analysis. This step is critical and non-trivial.
2. **Longitudinal Studies on Single Platforms**: Conduct in-depth, long-term studies on a single platform (e.g., Reddit or X/Twitter) for a specific set of topics (e.g., climate change, public health). This would help validate the correlation between SSD metrics and real-world indicators of polarization, such as survey data or event-driven discourse analysis.
3. **Causal Inference**: Move beyond correlation to explore whether SSD metrics can serve as leading indicators for events like the formation of echo chambers or the spread of misinformation, potentially informing intervention strategies.

 doi:10.20944/preprints202603.1326.v1

Appendix A.3 Finance: Market Regime Detection

The financial hypothesis is that markets undergo a phase of "information crystallization" before a crash, where collective herding behavior reduces the diversity of trading strategies. This should be detectable as a drop in κ and H_{SSD}, followed by a sharp spike into a chaotic regime during the crash event itself.

Future research should focus on:

1. **High-Frequency Data Analysis**: Test the hypothesis on intra-day, high-frequency trading data to see if the predictive signal is stronger at shorter timescales.
2. **Multi-Asset Validation**: Rigorously test the relational analysis proposed in the original Appendix C to see if increased SSD correlations between asset classes (e.g., stocks, bonds, gold) can serve as a robust early warning signal for systemic risk.
3. **Controlled Backtesting with Robust Methodologies**: Any algorithmic trading strategy based on SSD must be backtested using walk-forward analysis and out-of-sample validation to avoid look-ahead bias and overfitting. The optimistic backtest results from the original manuscript should be treated as a starting point for a much more rigorous evaluation.

Appendix A.4 Conclusions on Future Work

These proposed applications illustrate the broad potential of the SSD framework. The common thread is the use of SSD to detect a transition from a "critical" (healthy, diverse) regime to either a "crystalline" (rigid, homogeneous) or "chaotic" (unstable, panicked) regime. The next critical step is to move from hypothesis generation to rigorous, domain-specific validation. Each of these avenues presents a significant research program in its own right, requiring close collaboration with domain experts and access to high-quality data.

Appendix B Maximum entropy of SSD distribution

For the full 27-state symbolic space the theoretical maximum entropy equals

$$H_{\max} = \log_2(27) = 4.755 \text{ bits}.$$

For stochastic signals the expected entropy depends on the induced probability distribution of difference signs. Monte Carlo simulations of Gaussian white noise yield

$$H_{SSD} \approx 3.88,$$

which is consistent with the values observed in Table 2. This indicates that white noise activates most SDS motifs but not uniformly due to statistical dependence between first and second differences.

Appendix C Multi-Domain Survey of SSD Metrics – A Preliminary Exploration

This appendix presents a broad, exploratory survey of SSD metrics across 19 heterogeneous domains. The purpose of this survey is not to make definitive claims about "universal regimes," but rather to illustrate the range of values the SSD metrics (H_{SSD} and κ) can take across different types of data, and to generate hypotheses for future, more focused research. All data used are from publicly available sources, and the analysis was conducted with a consistent methodology ($N \geq 5000$ points per sample, tolerance $\tau = 10^{-10}$ for continuous signals).

Appendix C.1 Data Sources and Methodology

For each domain, we selected representative samples from the following public data sources:

 doi:10.20944/preprints202603.1326.v1

Table A1. Data sources for the multi-domain SSD survey.

#	Domain	Data Source / Generation Method and Access Link
1	Shakespeare	Project Gutenberg: Complete Works of William Shakespeare https://www.gutenberg.org/ebooks/100
2	Wikipedia (EN)	Wikimedia Dumps: Random sample of ∼10,000 articles https://dumps.wikimedia.org/
3	MNIST	Yann LeCun's MNIST Database: Handwritten digits http://yann.lecun.com/exdb/mnist/
4	CIFAR-10	University of Toronto: CIFAR-10 dataset https://www.cs.toronto.edu/~kriz/cifar.html
5	Speech	LibriSpeech ASR Corpus https://www.openslr.org/12/
6	EEG (Normal)	PhysioNet: CHB-MIT Scalp EEG Database https://physionet.org/content/chbmit/
7	EEG (Seizure)	PhysioNet: CHB-MIT Scalp EEG Database (ictal periods) https://physionet.org/content/chbmit/
8	Seismic Noise	IRIS Data Services / USGS: Background seismic activity https://ds.iris.edu/data/
9	Earthquake	USGS Earthquake Catalog: Strong motion data (e.g., Parkfield 2004) https://earthquake.usgs.gov/
10	Mechanical Vibrations (OK)	CWRU Bearing Data Center: Normal bearings https://engineering.case.edu/bearingdatacenter
11	Mechanical Failure	CWRU Bearing Data Center: Faulty bearings https://engineering.case.edu/bearingdatacenter
12	Financial Index	Yahoo Finance: S&P 500 historical daily closes https://finance.yahoo.com/quote/\protect\char%\relax5EGSPC/history/
13	Turbulence	Johns Hopkins Turbulence Databases (JHTDB) http://turbulence.pha.jhu.edu/
14	Checkerboard	Synthetically generated: 100×100 checkerboard pattern numpy.tile and .flatten() in Python
15	Uniform Noise	numpy.random.uniform(-1, 1, N) Built into Python's NumPy library
16	π (decimals)	mpmath library: First 10,000 decimal digits of π `mpmath.mp.dps = 10010; str(mpmath.pi)[2:10002]`
17	DNA (chr1)	NCBI GenBank: Human chromosome 1 (GRCh38) https://www.ncbi.nlm.nih.gov/genome/guide/human/
18	Periodic Signal	Synthetically generated: Pure sine wave `np.sin(2 * np.pi * 0.1 * t)`
19	e (decimals)	mpmath library: First 10,000 decimal digits of e `mpmath.mp.dps = 10010; str(mpmath.e)[2:10002]`

For text, DNA, and image data, raw data were converted to numerical sequences as follows:

 doi:10.20944/preprints202603.1326.v1

- Text (Shakespeare, Wikipedia): mapped to ASCII/Unicode code points
- DNA (chr1): nucleotides (A, C, G, T) mapped to numerical codes (1, 2, 3, 4)
- Images (MNIST, CIFAR-10, Checkerboard): serialized 2D pixel arrays into 1D sequences (flattened row-wise)

Appendix C.2 Results of the Multi-Domain Survey

The following table presents the mean SSD metrics (H_{SSD} and κ) for representative samples from each domain. For domains with high variability (e.g., EEG, images), these values should be considered illustrative rather than definitive population estimates.

Table A2. SSD metrics across 19 domains (representative samples).

#	Domain	H_{SSD} (bits)	κ	Notes / Interpretation
1	Shakespeare	2.91	0.71	High structural diversity in language
2	Wikipedia (EN)	2.97	0.74	Slightly higher diversity than literary text
3	MNIST	2.54	0.63	Relatively smooth, constrained shapes
4	CIFAR-10	2.71	0.69	Higher complexity from natural image textures
5	Speech	2.88	0.72	Dynamic balance of phonetic elements
6	EEG (Normal)	2.66	0.65	Healthy brain dynamics
7	EEG (Seizure)	3.21	0.88	Ictal period shows expanded local pattern diversity
8	Seismic Noise	2.43	0.59	Background geophysical activity
9	Earthquake	3.34	0.91	Main shock event (Parkfield 2004)
10	Mechanical Vibrations (OK)	2.37	0.56	Regular operational patterns
11	Mechanical Failure	3.12	0.86	Fault-induced complexity
12	Financial Index	2.79	0.68	S&P 500 daily closes (long-term average)
13	Turbulence	3.05	0.80	High-frequency velocity fluctuations
14	Checkerboard	1.02	0.21	Highly regular, repeating pattern
15	Uniform Noise	3.89	1.00	Maximum entropy, all symbols active
16	π (decimals)	3.85	0.99	Nearly indistinguishable from noise at local level
17	DNA (chr1)	1.84	0.41	Repetitive motifs and biological constraints
18	Periodic Signal	0.91	0.18	Pure sine wave, minimal local diversity
19	e (decimals)	3.84	0.99	Similar to π, high local randomness

Appendix C.3 Discussion of Observed Patterns

While this survey is exploratory, several consistent patterns emerge that generate hypotheses for future research:

1. **Extremes of order and randomness**: Purely periodic signals (sine wave) and highly structured patterns (checkerboard) yield very low H_{SSD} (< 1.5) and low κ (< 0.3). Conversely, true random noise (uniform) and the decimal expansions of irrational numbers (π, e) yield near-maximum H_{SSD} (> 3.8) and $\kappa \approx 1$.
2. **Natural systems occupy an intermediate range**: A wide variety of natural signals — language (Shakespeare, Wikipedia), speech, healthy EEG, financial indices, seismic noise — all fall within a relatively narrow intermediate range (H_{SSD} 2.4–3.0, κ 0.55–0.75). This observation is consistent with the hypothesis that many complex systems operate in a "critical" regime, balancing regularity and diversity.
3. **Pathological / event states shift towards extremes**: In each domain where a "pathological" or "event" state was available (EEG seizure, earthquake, mechanical failure), the SSD metrics shifted towards higher values, entering a region closer to that of random noise ($H_{SSD} > 3.1$, $\kappa > 0.85$). This suggests that such events are characterized by a loss of coherent structure and a diversification of local geometric patterns.
4. **Discrete vs. continuous data**: Domains with inherently discrete representations (DNA, digits of π) can show distinct patterns. DNA exhibits low H_{SSD} due to its repetitive, non-random biological structure, whereas the digits of π, also discrete, appear statistically random in their local geometry.

 doi:10.20944/preprints202603.1326.v1

Appendix C.4 Limitations and Caveats

- **Representative sampling**: Most domains were analyzed using a single representative sample. Within-domain variability (e.g., between different EEG subjects, different image classes in CIFAR-10, or different time periods in financial data) is not captured here.
- **No statistical inference**: The table presents point estimates without confidence intervals, standard deviations, or hypothesis tests. The values should not be used to draw definitive conclusions about differences between domains.
- **Data representation choices**: The method of converting non-numerical data (text, images) into a 1D time series can significantly impact the results. The choices made here (e.g., using ASCII codes) are just one of many possibilities.
- **Parameter sensitivity**: The choice of tolerance τ can affect results for continuous signals. A fixed, small τ was used for consistency, but an adaptive or domain-specific τ might be more appropriate in some cases.

Appendix C.5 Conclusions and Future Directions

This exploratory survey demonstrates the broad applicability of the SSD framework and provides a preliminary map of the H_{SSD}–κ space across diverse data types. The observed patterns, particularly the clustering of natural systems in an intermediate zone and the shift of pathological states towards higher entropy, generate compelling hypotheses for future research. The next logical step would be to conduct rigorous, statistically powered studies within single domains to validate these patterns and establish SSD as a reliable tool for detecting dynamical regime changes. The data and code used for this survey are available to facilitate such efforts.

References

1. Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*; Cambridge University Press: Cambridge, UK, 2004.
2. Grassberger, P.; Procaccia, I. Measuring the Strangeness of Strange Attractors. *Physica D: Nonlinear Phenomena* **1983**, *9*, 189–208. doi:10.1016/0167-2789(83)90298-1.
3. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x.
4. Inouye, T.; Shinosaki, K.; Sakamoto, H.; Toi, S.; Ukai, S.; Shinosaki, A.; Katsuda, S.; Hirano, M. Quantification of EEG Irregularity by Use of the Entropy of the Power Spectrum. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *79*, 204–210. doi:10.1016/0013-4694(91)90126-9.
5. Bandt, C.; Pompe, B. Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys. Rev. Lett.* **2002**, *88*, 174102. doi:10.1103/PhysRevLett.88.174102.
6. Lempel, A.; Ziv, J. On the Complexity of Finite Sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81. doi:10.1109/TIT.1976.1055501.
7. Hurst, H.E. Long-Term Storage Capacity of Reservoirs. *Trans. Am. Soc. Civ. Eng.* **1951**, *116*, 770–799.
8. Wolf, A.; Swift, J.B.; Swinney, H.L.; Vastano, J.A. Determining Lyapunov Exponents from a Time Series. *Physica D: Nonlinear Phenomena* **1985**, *16*, 285–317. doi:10.1016/0167-2789(85)90011-9.
9. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.Ch.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220. https://physionet.org/content/chbmit/. doi:10.1161/01.CIR.101.23.e215.
10. Richman, J.S.; Moorman, J.R. Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. doi:10.1152/ajpheart.2000.278.6.H2039.
11. Pangarić, Z. Symbolic Geometry of the Number Pi: Structures, Statistics, and Security. *Preprints.org* **2026**. https://www.preprints.org/manuscript/202603.0163. .