Comparative Symbolic Grammar Analysis of Fungal Electrical Networks: Methods and Data

Zubair Chowdhury Independent Researcher Wimbledon, England January 4, 2026

Abstract This paper presents a comparative preprocessing, symbolic encoding, and grammar-based compression methodology for the analysis of fungal electrical recordings across multiple species. Electrophysiological data from four species recorded by Adamatzky (2021) are integrated with a 137-hour high-resolution Schizophyllum commune recording from the author's laboratory. All datasets are z-score normalised, resampled to a common length of 50,000 samples, and quantised into a five-symbol alphabet (A–E) to enable fair cross-species comparison. Shannon entropy of the resulting symbol distributions reveals striking differences in syntactic richness: three Adamatzky species show near-degenerate entropies (0.0003–0.016 bits), Flammulina velutipes shows moderate diversity (0.50 bits), and the fast-spike S. commune dataset reaches maximum information density (2.32 bits). A Sequitur-style grammar induction algorithm is then applied to each symbol sequence to infer hierarchical structure and compute grammar-derived metrics (rule counts, rule depth, compression ratio, and motif statistics) that quantify syntactic complexity beyond simple symbol frequencies, as visualised in Figures 1 and 2.

1. Data Sources Electrophysiological recordings from four fungal species were obtained from the open Zenodo repository "Recordings of electrical activity of four species of fungi" (record 5790768, Adamatzky 2021):

Cordyceps militaris (1,900,145 samples; 108.76 MB)

Flammulina velutipes (Enoki mushroom; 1,210,938 samples; 57.35 MB)

Omphalotus nidiformis (Ghost fungi; 3,279,569 samples; 156.42 MB)

Schizophyllum commune (4-species dataset; 263,959 samples; 13.54 MB)

In addition, a fast-spike window of 3,534 samples extracted from a 137-hour continuous Schizophyllum commune recording acquired in the author's laboratory (Chowdhury 2025 dataset) was included as an independent high-resolution control (0.13 MB).

All differential electrode channels were exported as tab-delimited text files and imported into Python 3.14 (pandas 2.3, NumPy 2.4) for processing.

2. Signal Preprocessing and Normalisation For each dataset, only numeric columns (voltage measurements in millivolts) were retained, and the first numeric channel was selected as the reference signal for comparative analysis to ensure consistency across species with different numbers of electrode pairs.

Each selected channel was z-score normalised according to

$$z(t) = (V(t) - \mu) / \sigma$$

where $V(t)$ is the instantaneous voltage, μ is the channel mean, and σ is the standard deviation (computed with Bessel's correction, $N - 1$ denominator). This normalisation centres and scales each dataset to zero mean and unit variance, removing offset and amplitude bias introduced by different electrode geometries or amplifier gains.

3. Temporal Resampling to Common Length Because the Adamatzky datasets span 1.2–3.3 million samples and the Chowdhury fast-spike window contains only 3,534 samples, direct comparison of raw time series is not feasible. To enable fair cross-species analysis, each normalised trace was resampled to a common length of 50,000 samples by linear interpolation using NumPy's interp function:

$$V\_resampled'(i) = interp(x'(i), x\_old, V\_z), \text{ for } i = 1, \ldots, 50{,}000$$

where x_old represents the original time indices (normalised to), x' denotes the new equispaced grid, and V_z is the z-scored signal. This approach preserves the shape and gross dynamics of each time series while bringing all datasets into a common temporal reference frame for downstream comparison.

4. Symbolic Quantisation into Five-Symbol Alphabet The resampled, normalised traces were discretised into a five-symbol alphabet (A, B, C, D, E) using empirical quintiles of the amplitude distribution. For each dataset, the z-scored values were divided into five equal-probability bins defined by the 20th, 40th, 60th, and 80th percentiles:

Bin 1 ($z < Q20$): A

Bin 2 ($Q20 \leq z < Q40$): B

Bin 3 ($Q40 \leq z < Q60$): C

Bin 4 ($Q60 \leq z < Q80$): D

Bin 5 ($z \geq Q80$): E

Each continuous sample was assigned the corresponding symbol, yielding one discrete sequence of length 50,000 per dataset.

The choice of five symbols reflects a balance between coarse-graining (loss of waveform detail) and alphabet size (combinatorial explosion), and the quintile-based quantisation ensures that each symbol would appear with equal expected frequency in a uniform distribution, so differences in symbol entropy reflect genuine structure rather than arbitrary bin widths.

5. Syntactic Richness via Shannon Entropy As a first-pass proxy for the syntactic complexity and information content of each symbolic sequence, the Shannon entropy of the symbol distribution was computed as

$$H = - \Sigma\_\{s \in \{A,B,C,D,E\}\} \; p(s) \cdot \log_2 p(s)$$

where p(s) is the relative frequency of symbol s in the sequence. The theoretical minimum is H = 0 bits (all samples map to a single symbol), and the theoretical maximum for a five-symbol alphabet is $H = \log_2(5) \approx 2.322$ bits, achieved when all symbols appear with equal probability.

The entropy results are summarised in Table 1:

Dataset	Original length	Resampled length	Symbols used	Entropy (bits)	% of max
Cordyceps militaris (Adamatzky 2021)	1,900,145	50,000	1–2	0.00064	0.03%
Flammulina velutipes (Adamatzky 2021)	1,210,938	50,000	4–5	0.49883	21.5%
Omphalotus nidiformis (Adamatzky 2021)	3,279,569	50,000	1–2	0.00034	0.01%
Schizophyllum commune (Adamatzky 2021)	263,959	50,000	2–3	0.01586	0.68%
Schizophyllum commune (Chowdhury 2025, fast-spike)	3,534	50,000	4–5	2.32193	99.9%

Three of the four Adamatzky species—C. militaris, O. nidiformis, and the Adamatzky S. commune—exhibit extremely low entropy, indicating that the resampled quantised sequences are dominated by one or two symbols and are syntactically degenerate at the five-symbol level. Flammulina velutipes shows moderate diversity (21.5% of maximum), with four or five symbols present, while the Chowdhury fast-spike S. commune dataset achieves near-maximal entropy, with all five symbols appearing at nearly equal frequency.

6. Grammar-Based Compression and Hierarchical Structure To probe structure beyond symbol counts, each 50,000-symbol sequence was subjected to grammar-based compression using a Sequitur-style context-free grammar induction algorithm. Sequitur infers a hierarchical representation of a sequence by replacing repeated digrams with nonterminal symbols while enforcing two constraints: (i) digram uniqueness (no pair of adjacent symbols appears more than once) and (ii) rule utility (every nonterminal is used at least twice).

6.1. Implementation Details For each dataset, the corresponding symbol sequence was written to 03_results/grammar_sequences/_symseq.txt as a single line of characters. A Python module 02_analysis/grammar_extraction.py implements a minimal Sequitur-style inducer that:

Initialises a grammar with a start rule S.

Ingests the symbol sequence one symbol at a time, appending to the right-hand side of S.

Whenever a repeated digram is detected, either reuses an existing rule whose right-hand side matches that digram or introduces a new nonterminal rule R_k → x y, replacing all occurrences of the digram with R_k.

Enforces rule utility by inlining any nonterminal that is used fewer than two times and then rebuilding the digram index.

The implementation is linear in the sequence length and uses cycle-safe computations when deriving depth and motif statistics, ensuring that rare recursive structures do not cause divergence in the analysis.

6.2. Grammar-Derived Metrics From each induced grammar, the following metrics are computed and stored in 03_results/grammar_metrics.json:

Grammar size and compression ratio

grammar_size: total number of symbols on the right-hand side of all rules.

compression_ratio: original length (50,000) divided by grammar_size, with larger values indicating more compression.

Rule counts and hierarchy

n_rules_total: total number of rules, including the start rule.

n_nonterminals: number of nonterminal rules (excluding S).

depth_mean, depth_max, depth_std: statistics of the maximum expansion depth of each nonterminal, defined as the maximum number of nonterminal expansions needed to reach terminals, computed with explicit cycle detection.

Motif statistics

span_length_mean, span_length_max: mean and maximum number of terminals in the full expansion of each nonterminal, interpreted as characteristic motif lengths.

usage_mean, usage_max: frequency with which each nonterminal appears on the right-hand side of any rule, summarising the dominance of the most frequently reused motifs.

These metrics provide a compact description of hierarchical structure and motif reuse that can be compared directly across species and recording regimes, as shown in Figures 1 and 2.

7. Results: Cross-Species Grammar Complexity Figure 1 plots symbol entropy against the number of nonterminals for all five datasets, revealing a non-monotonic relationship between symbol-level diversity and hierarchical grammar size. Cordyceps militaris and Omphalotus nidiformis occupy the lower left (near-zero entropy, minimal grammar), Flammulina velutipes sits at intermediate entropy with a small grammar, the Adamatzky Schizophyllum commune recording displays very low entropy yet the largest grammar (89 nonterminals), and the fast-spike S. commune window achieves maximal entropy but only 5 nonterminals.

Figure 2 illustrates the tradeoff between hierarchical depth and motif dominance (defined as usage_max / seq_length). C. militaris and O. nidiformis combine depth 1 with motif dominance near 0.67, indicating a single shallow motif repeated throughout.

F. velutipes shows depth 2 with dominance around 0.48, reflecting one strong motif plus minor patterns. The Adamatzky S. commune achieves the deepest hierarchy (depth 3) with the lowest motif dominance (~0.09), consistent with many distinct motifs reused at moderate frequencies, while the fast-spike window has depth 1 and dominance ~0.13, reflecting high entropy but minimal hierarchical organisation.

8. Interpretation and Intended Usage The combination of entropy and grammar-based metrics is designed to separate symbol-level diversity from higher-order syntactic organisation. Low entropy with minimal grammars is indicative of strongly stereotyped activity, whereas low entropy with many nonterminals and nontrivial rule depth suggests structured reuse of motifs under a highly skewed symbol distribution. Conversely, high entropy with shallow grammars may reflect locally irregular fluctuations that are not organised into deep hierarchical structure.

The present document focuses on the reproducible methods: data acquisition, preprocessing, symbolic encoding, and grammar extraction. Biological and theoretical interpretations, including links to error-correction and potential consciousness correlates, are treated in a separate results manuscript that uses the metrics defined here to construct cross-species comparisons and hypothesis-driven analyses.

References Adamatzky, A. (2021). Recordings of electrical activity of four species of fungi. Zenodo. https://doi.org/10.5281/zenodo.5790768.

Nevill-Manning, C. G., & Witten, I. H. (1997). Identifying hierarchical structure in sequences: A linear-time algorithm. Journal of Artificial Intelligence Research, 7, 67–82.

Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A., Chen, C., & Frankenstein, S. (2015). Time series anomaly discovery with grammar-based compression. Proceedings of EDBT 2015.

Yamamoto, H., et al. (2012). Developing JSequitur to study the hierarchical structure of biological sequences in a grammatical inference framework of string compression algorithms.

NumPy Developers. (2024). NumPy 2.4. https://numpy.org/.

pandas Development Team. (2024). pandas 2.3. https://pandas.pydata.org/.

Chowdhury, Z. (2025). 137-hour continuous Schizophyllum commune fast-spike recording. Unpublished laboratory data.

Figures

Figure 1: Entropy vs Grammar Size (Nonterminals) – scatter plot showing non-monotonic relationship.

Figure 2: Hierarchy vs Motif Dominance – scatter plot of maximum rule depth vs usage_max / seq_length.