Proteins as Statistical Languages: Information-Theoretic Signatures of Proteomes Across the Tree of Life

Enso O. Torres Alegre
ORCID: 0000-0002-6798-8776
Pontifical Catholic University of Chile, Santiago, Chile
onill@uc.cl

Abstract

Protein sequences are commonly interpreted through biochemical and evolutionary lenses, emphasizing structure–function relationships and selection in sequence space. Here we develop a complementary viewpoint: proteins as *statistical languages*—strings over a finite alphabet generated by constrained stochastic processes. We formalize intrinsic informational descriptors of protein ensembles, including composition entropy H_1, adjacent mutual information I_1, and separation-dependent information profiles I_d. A null-model ladder (uniform, composition-matched i.i.d., and Markov-1) separates compositional effects from genuine positional dependence. We then evaluate these descriptors empirically across 20 UniProt reference proteomes spanning major clades, using protein-level bootstrap resampling and matched synthetic controls. Real proteomes consistently depart from composition-matched i.i.d. baselines and exhibit information profiles that remain elevated beyond the decay expected under first-order Markov surrogates, indicating dependencies beyond local transition statistics. Finally, a compressibility proxy (gzip) provides an orthogonal signature of redundancy relative to i.i.d. controls at matched composition. Together, these results support the view of proteomes as constrained statistical languages and provide model-agnostic fingerprints for comparing sequence ensembles.These signatures provide a lightweight diagnostic layer for comparing proteomes prior to mechanistic modeling

Keywords: protein sequences; information theory; entropy; mutual information; maximum entropy; compressibility; statistical language; proteomes

1 Introduction

Proteins are linear polymers over an alphabet of 20 amino acids that fold into three-dimensional structures and realize biological functions. Standard approaches interpret sequences through physicochemical constraints (e.g. stability and foldability) and evolutionary mechanisms operating in sequence space [2, 9, 12, 14]. Alongside these mechanistic views, protein sequences also display robust statistical regularities: non-uniform residue frequencies, context-dependent patterns, and correlations at multiple length scales [3, 5, 6, 16].

These properties suggest that proteins—viewed purely as strings—occupy a small and structured subset of the space of all possible strings.

Information theory and statistical linguistics quantify structure in symbolic sequences using entropy, mutual information, and compressibility [1, 8, 15]. Importantly, these tools detect constraints without requiring semantic interpretation. This motivates the central hypothesis explored here:

> *Protein sequences can be treated as realizations of a statistical language: a constrained generative process over a finite alphabet.*

This hypothesis does *not* claim proteins are encrypted messages. Rather, it frames compositional bias and context dependence as information-theoretic signatures of constrained generation, enabling principled comparison between ensembles.

Contributions. This paper (i) formalizes proteins as statistical objects and defines intrinsic informational measures; (ii) introduces a hierarchy of null models that separate composition from higher-order structure; (iii) defines information profiles across sequence separations; and (iv) proves a general entropy-deficit bound showing that rarity/viability constraints necessarily reduce achievable entropy. To demonstrate that these descriptors are not merely formal, we perform a cross-proteome analysis on 20 UniProt reference proteomes [17], spanning 10 biological categories (two proteomes each), using protein-level bootstrap resampling and two matched null controls (composition-matched i.i.d. and Markov-1). This empirical component tests whether real proteomes occupy distinct regions in (H_1, I_1) space and whether their information profiles I_d systematically exceed those of finite-memory surrogates.

Scope. Our goal is deliberately model-agnostic: to quantify statistical structure in proteome sequence ensembles without requiring structural annotation, functional labels, or family alignments. The resulting signatures should be viewed as descriptive fingerprints rather than mechanistic explanations.

2 Protein Sequences as Statistical Objects

Let $\mathcal{A}$ denote the amino acid alphabet with $|\mathcal{A}| = 20$. A sequence of length L is

$$S = (a_1, a_2, \ldots, a_L) \in \mathcal{A}^L,$$

and the space of all finite-length sequences is $\mathcal{S} = \bigcup_{L \geq 1} \mathcal{A}^L$. We treat an ensemble of sequences as samples from a probability distribution $\mathbb{P}$ over $\mathcal{S}$. For fixed L, write $\mathbb{P}_L$ for the induced distribution on $\mathcal{A}^L$.

Definition 2.1 (Composition-matched i.i.d. baseline)**.** *Given a one-site marginal p_1 over $\mathcal{A}$, the composition-matched i.i.d. model on $\mathcal{A}^L$ is*

$$\mathbb{P}_{\text{iid}}(S) = \prod_{i=1}^{L} p_1(a_i).$$

This baseline isolates structure beyond composition: any dependency between positions reduces entropy relative to the maximum-entropy distribution with the same one-site marginals [1, 11].

3 Amino Acid Usage and Entropy

3.1 Composition

Given an ensemble, define the marginal frequency

$$p_1(x) = \mathbb{P}(a_i = x), \qquad x \in \mathcal{A},$$

estimated by pooled counts across sequences. The compositional entropy is

$$\mathsf{H}_1 = -\sum_{x\in\mathcal{A}} p_1(x)\log_2 p_1(x). \tag{1}$$

To compare composition against a reference q (e.g. uniform or a background), use Kullback–Leibler divergence

$$\mathsf{D}_{\mathrm{KL}}(p_1\|q) = \sum_{x\in\mathcal{A}} p_1(x)\log\frac{p_1(x)}{q(x)}. \tag{2}$$

3.2 Sequence entropy and an extremal principle

Let $\mathsf{H}(S)$ denote the Shannon entropy of $\mathbb{P}_L$ on $\mathcal{A}^L$:

$$\mathsf{H}(S) = -\sum_{S\in\mathcal{A}^L} \mathbb{P}_L(S)\log_2 \mathbb{P}_L(S).$$

Proposition 3.1 (Maximum entropy at fixed one-site marginals)**.** *Among all distributions on $\mathcal{A}^L$ with fixed one-site marginals p_1, the composition-matched i.i.d. model $\mathbb{P}_{\mathrm{iid}}(S) = \prod_{i=1}^L p_1(a_i)$ maximizes $\mathsf{H}(S)$.*

Proof. Fixing one-site marginals defines a convex set of admissible distributions. The product measure is the unique maximizer because any statistical dependency introduces additional constraints that reduce entropy [1, 11]. □

Proposition 3.1 separates (i) compositional constraints from (ii) higher-order constraints encoded in reduced conditional entropies and nonzero mutual information.

4 Correlations and Information Profiles

4.1 Local (second-order) structure

Define adjacent joint frequencies

$$p_2(x,y) = \mathbb{P}(a_i = x, a_{i+1} = y).$$

The conditional entropy is

$$\mathsf{H}(a_{i+1} \mid a_i) = -\sum_{x,y\in\mathcal{A}} p_2(x,y)\log_2\frac{p_2(x,y)}{p_1(x)}. \tag{3}$$

The adjacent mutual information is

$$\mathsf{I}_1 = \mathsf{I}(a_i; a_{i+1}) = \mathsf{H}_1 - \mathsf{H}(a_{i+1} \mid a_i)\,. \tag{4}$$

If residues are i.i.d. (given p_1), then $\mathsf{I}_1 = 0$.

4.2 Length-scale dependence

For separation $d \geq 1$, define

$$p^{(d)}(x, y) = \mathbb{P}(a_i = x, a_{i+d} = y)$$

and mutual information

$$\mathsf{I}_d = \sum_{x,y \in \mathcal{A}} p^{(d)}(x, y) \log_2 \frac{p^{(d)}(x, y)}{p_1(x)p_1(y)}. \tag{5}$$

Definition 4.1 (Information profile)**.** *The mapping $d \mapsto \mathsf{I}_d$ is the* information profile *of an ensemble. Its decay characterizes the range of dependencies (short-range vs. long-range).*

Remark 4.1. *For a finite-order Markov source, dependencies are predominantly local; empirically, one often observes rapid decay of I_d with d once the relevant memory scale is exceeded. Slow decay or plateaus can arise from global constraints, repeats, or mixture structure, and do not by themselves imply semantic "meaning."*

Lemma 4.1 (Independence and vanishing information profile)**.** *If $\mathsf{I}_d = 0$ for all $d \geq 1$ and the process is stationary with well-defined one-site marginal p_1, then all pairs factorize:*

$$\mathbb{P}(a_i = x, a_j = y) = p_1(x)p_1(y) \quad \textit{for all } i \neq j,$$

i.e. the process is pairwise independent. In particular, any departure from i.i.d. behavior must yield $\mathsf{I}_d > 0$ for at least one separation d.

Proof. By definition, $\mathsf{I}_d = 0$ implies $p^{(d)}(x, y) = p_1(x)p_1(y)$ for all $x, y \in \mathcal{A}$ at that separation. If this holds for all $d \geq 1$, then every pair of positions factorizes into the product of marginals, yielding pairwise independence. □

5 Proteins as Statistical Languages

We formalize "language" in a statistical (not semantic) sense: a constrained set of admissible strings together with a generative distribution.

Definition 5.1 (Statistical protein language)**.** *A* statistical protein language *is a pair $(\mathcal{L}, \mathbb{P})$ where $\mathcal{L} \subseteq \mathcal{S}$ is a set of admissible sequences (constraints) and $\mathbb{P}$ is a probability measure supported on $\mathcal{L}$.*

5.1 Null-model ladder

A principled comparison ladder is:

1. **Uniform random (UR)**: $\mathbb{P}_{\mathrm{UR}}(S) = 20^{-L}$ on $\mathcal{A}^L$.
2. **Composition-matched i.i.d. (IID)**: $\mathbb{P}_{\mathrm{iid}}(S) = \prod_i p_1(a_i)$.
3. **Finite-order Markov (Mk)**: matches k-mer statistics (here $k = 1$).
4. **Maximum-entropy (ME)**: matches selected constraints while maximizing entropy [7].

Each rung provides a benchmark; deviations quantify additional structure.

5.2 Maximum-entropy grammars

Following Jaynes [7], maximum-entropy models provide a canonical "grammar" induced by constraints. Given constraint functions $f_m(S)$ with target averages, the ME distribution takes the exponential-family form

$$\mathbb{P}_{\mathrm{ME}}(S) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\Big(\sum_{m=1}^{M} \lambda_m f_m(S)\Big) \mathbf{1}_{\mathcal{L}}(S), \tag{6}$$

where $Z(\boldsymbol{\lambda})$ normalizes.

Proposition 5.1 (Constraints induce dependencies)**.** *If* (6) *includes only one-site constraints, then* $\mathbb{P}_{\mathrm{ME}}$ *is composition-matched i.i.d. and* $\mathsf{I}_d = 0$ *for all* d*. Introducing multi-site constraints generically yields* $\mathsf{I}_d > 0$ *at corresponding separations.*

Remark 5.1. *Pairwise maximum-entropy models are closely related to Potts/DCA formulations used to capture covariation in protein families [3, 6, 18].*

6 Empirical Data and Methods

6.1 Proteomes and preprocessing

We analyzed 20 proteomes downloaded from UniProt via the UniProt REST API, using UniProt reference proteome identifiers spanning 10 broad biological categories (two proteomes per category): mammals, birds, fish, amphibians, insects, plants, bacteria, archaea, filamentous fungi, and yeasts [17]. Protein sequences were parsed from FASTA format, filtered to retain only canonical amino acids (20-letter alphabet), and truncated by removing sequences shorter than a minimum length (here $L_{\min} = 80$) to reduce instability from extremely short proteins. For computational feasibility, large proteomes were subsampled using a fixed maximum number of sequences per proteome; all analyses below should thus be interpreted as estimates on a representative proteome subsample.

6.2 Bootstrap protocol

For each proteome, we performed protein-level bootstrap resampling. Each bootstrap replicate consists of sampling proteins with replacement from the filtered proteome (sample size $n = 200$ proteins per replicate), preserving the empirical length distribution within that replicate. We used 150 bootstrap replicates per proteome. All reported summary statistics are means and standard deviations across bootstrap replicates.

6.3 Information-theoretic estimators

One-site marginals p_1 were estimated by pooled residue counts. Adjacent joint frequencies p_2 and separated joint frequencies $p^{(d)}$ were estimated by direct counting. Mutual informations I_1 and I_d were computed using plug-in estimators (base-2 logarithms; units are bits). The information profile was evaluated for separations $d = 1, \ldots, 25$.

6.4 Matched null controls

For each bootstrap replicate, two controls were generated:

1. **IID control:** sequences generated by sampling residues i.i.d. according to the replicate-specific empirical p_1, using the same length distribution as the real replicate.

2. **Markov-1 control:** sequences generated by a first-order Markov chain with initial distribution p_1 and transition matrix estimated from the real replicate dipeptide counts p_2, again using the same length distribution.

These controls isolate compositional effects from positional dependence and test whether observed dependencies extend beyond first-order transition statistics.

6.5 Compressibility proxy

As an orthogonal signature of redundancy, we computed an overhead-corrected gzip code length per character on a long concatenation of sampled sequences (with a fixed total amino-acid budget). This measure is a practical proxy for algorithmic redundancy rather than a formal entropy-rate estimator [4, 10].

7 Entropy Deficit Under Viability Constraints

We state a general result requiring no explicit folding or functional model.

Definition 7.1 (Viable set and rarity fraction)**.** *Let $\mathcal{V}_L \subseteq \mathcal{A}^L$ be the set of "viable" sequences of length L under some broad constraint. Define the rarity fraction*

$$f_L = \frac{|\mathcal{V}_L|}{|\mathcal{A}^L|} = \frac{|\mathcal{V}_L|}{20^L}.$$

Theorem 7.1 (Entropy deficit bound)**.** *For any distribution $\mathbb{P}$ supported on $\mathcal{V}_L$,*

$$\mathsf{H}(S) \leq \log_2 |\mathcal{V}_L| = L \log_2 20 + \log_2 f_L.$$

Equivalently,

$$\frac{1}{L}\mathsf{H}(S) \leq \log_2 20 + \frac{1}{L}\log_2 f_L.$$

Proof. Entropy is maximized by the uniform distribution on the support. The uniform distribution on $\mathcal{V}_L$ has entropy $\log_2 |\mathcal{V}_L|$, which upper-bounds any supported distribution [1]. □

Corollary 7.1.1 (Rarity implies unavoidable structure)**.** *If f_L decays exponentially with L (e.g. $f_L \sim 2^{-\alpha L}$ with $\alpha > 0$), then no viable-sequence ensemble can be simultaneously close to uniform composition* and *close to i.i.d. ordering in the sense that both $\mathsf{H}_1 \approx \log_2 20$ and $\mathsf{I}_d \approx 0$ for all d hold asymptotically.*

Remark 7.1. *Corollary 7.1.1 provides a clean conceptual link: if viability is rare, statistical structure is unavoidable and must appear either as compositional bias, correlations, or both.*

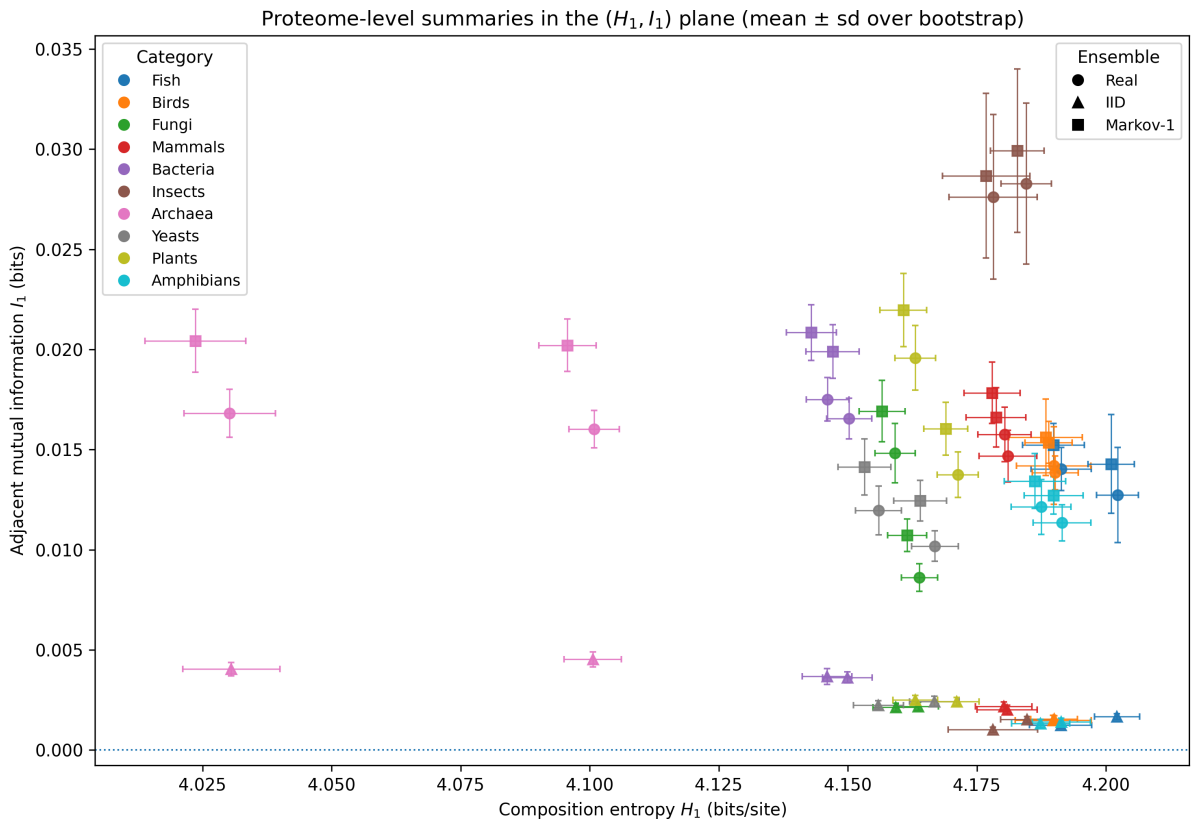


Figure 1: Proteome-level summaries in the (H_1, I_1) plane (mean ± s.d. across bootstrap replicates). Colors indicate broad biological categories and marker shapes indicate the ensemble (Real, IID, Markov-1). IID controls preserve one-site composition p_1 while destroying positional dependence, yielding $I_1 \approx 0$. Markov-1 controls preserve adjacent dipeptide statistics and reproduce local dependence by construction. Real proteomes are systematically displaced from IID baselines, indicating nontrivial ordering constraints beyond composition alone.

8 Results: Cross-Proteome Information Signatures

8.1 Real proteomes depart from composition-matched i.i.d. baselines

Figure 1 compares real proteomes to matched IID and Markov-1 controls in the (H_1, I_1) plane using bootstrap replicates. By construction, the IID control preserves H_1 but destroys positional dependence, yielding I_1 near zero. Across the 20 proteomes analyzed, the mean adjacent mutual information increased from

$$\langle I_1 \rangle_{\text{IID}} \approx 0.0023 \text{ bits} \quad \text{to} \quad \langle I_1 \rangle_{\text{Real}} \approx 0.0155 \text{ bits},$$

indicating robust short-range dependence beyond composition alone. The mean composition entropy across real proteomes was $\langle H_1 \rangle_{\text{Real}} \approx 4.164$ bits/site (as expected, closely matched by IID controls built from the same p_1).

8.2 Information profiles remain elevated beyond Markov-1 decay

To probe dependence across scales, we computed information profiles I_d for separations $d = 1, \ldots, 25$. While Markov-1 controls reproduce $d = 1$ dependence by construction, real

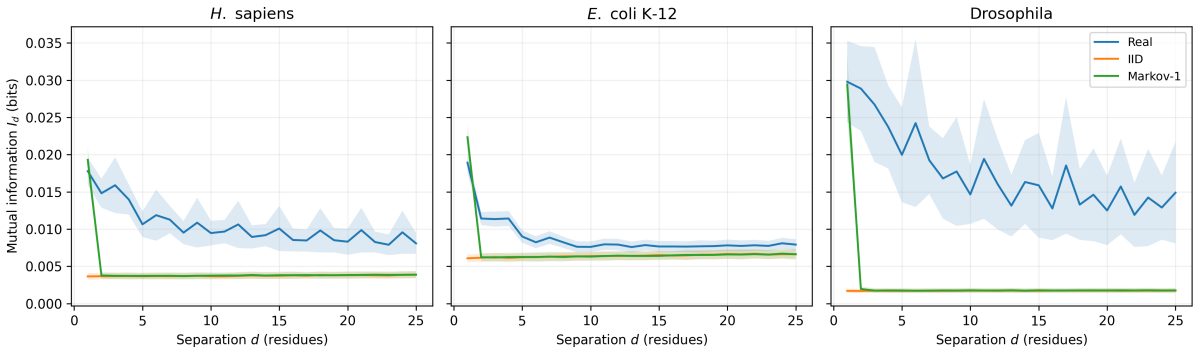


Figure 2: Information profiles I_d (mean ± s.d. across bootstrap replicates) for three representative proteomes. IID controls collapse toward a near-zero baseline for all d. Markov-1 controls reproduce the $d = 1$ dependence by construction and decay rapidly thereafter. Real proteomes often exhibit elevated I_d over multiple separations, indicating dependencies beyond a first-order Markov source.

proteomes frequently display elevated I_d for $d > 1$ relative to both IID and Markov-1 surrogates. Summarizing dependence beyond adjacency by averaging over $d = 2, \ldots, 25$, we observed:

$$\langle I_{d\geq 2}\rangle_{\text{Real}} \approx 0.00887 \text{ bits} \quad \text{vs.} \quad \langle I_{d\geq 2}\rangle_{\text{Markov-1}} \approx 0.00392 \text{ bits.}$$

These results indicate dependencies that are not fully explained by a first-order Markov model.

Figure 2 shows representative I_d profiles for three proteomes (human, *E. coli* K-12, and *Drosophila*). Full per-proteome profiles for all 20 datasets can be provided as supplementary figure S1.

8.3 Category-level fingerprints

To summarize cross-clade patterns, Table 1 reports category-averaged signatures: adjacent dependence (I_1), non-adjacent dependence averaged over $d = 2..25$ ($I_{d\geq 2}$), and a compressibility gap relative to IID controls. Categories exhibit distinct statistical fingerprints, consistent with heterogeneous constraints across the tree of life. Importantly, these statistics are descriptive and can reflect a combination of factors including heterogeneous domain composition, repeats, mixture structure, and global constraints rather than any single mechanistic cause.

8.4 Compressibility provides an orthogonal redundancy signature

We computed an overhead-corrected gzip bits/character estimate on long concatenations of sampled proteins. At matched composition, IID sequences are expected to be least compressible; deviations indicate redundant patterns and ordering constraints. Across the 20 proteomes, the mean compressibility gap was

$$\langle \Delta C_{\text{gzip}}\rangle = \langle C_{\text{IID}} - C_{\text{Real}}\rangle \approx 0.174 \text{ bits/char},$$

with pronounced gaps in several eukaryotic proteomes. Figure **??** summarizes bootstrap-level gaps.

Table 1: Category-level information signatures (means over two proteomes per category). $I_{d\geq 2}$ denotes the mean of I_d over $d = 2..25$. "gzip gap" is $\Delta C_{\text{gzip}} = C_{\text{IID}} - C_{\text{Real}}$ (bits/char; positive implies real sequences are more compressible than IID controls).

Category	n	I_1 Real	I_1 IID	$I_{d\geq 2}$ Real	$I_{d\geq 2}$ Mk1	gzip gap
Amphibians	2	0.0117	0.0014	0.0078	0.0023	0.251
Archaea	2	0.0164	0.0043	0.0111	0.0075	0.202
Bacteria	2	0.0170	0.0037	0.0086	0.0064	0.140
Birds	2	0.0140	0.0015	0.0079	0.0026	0.157
Fish	2	0.0134	0.0015	0.0079	0.0025	0.220
Fungi	2	0.0117	0.0022	0.0073	0.0037	0.100
Insects	2	0.0280	0.0013	0.0139	0.0022	0.308
Mammals	2	0.0152	0.0021	0.0095	0.0036	0.131
Plants	2	0.0167	0.0025	0.0077	0.0043	0.117
Yeasts	2	0.0111	0.0023	0.0071	0.0040	0.109

8.5 Entropy-rate reduction from local correlations

Adjacent dependence reduces per-residue uncertainty relative to a composition-only baseline. Using $H^{(1)}_{\text{rate}} = H_1 - I_1$, Fig. **??** compares the first-order Markov entropy rate to H_1 across bootstrap replicates. The systematic gap between H_1 and $H^{(1)}_{\text{rate}}$ provides an operational measure of local ordering induced by adjacent correlations.

9 Discussion and Scope

This work provides a rigorous, model-agnostic language for describing proteomes as constrained stochastic sources over a finite alphabet. The empirical results support three concrete claims.

(1) Proteomes are not i.i.d. at matched composition. Across 20 proteomes, I_1 increased from ~ 0.0023 bits (IID controls) to ~ 0.0155 bits (real), indicating robust local dependence beyond amino-acid composition.

(2) Many proteomes show dependencies beyond first-order Markov structure. Information profiles I_d frequently remained elevated for $d > 1$ relative to Markov-1 surrogates. This is consistent with constraints not fully captured by adjacent transition statistics. However, such patterns can arise from multiple sources including mixture structure (heterogeneous protein classes), repeats, and domain architecture; therefore, I_d should be interpreted as a descriptive dependence signature rather than direct evidence of a particular mechanistic grammar.

(3) Compressibility provides an orthogonal redundancy signal. Real proteomes were more compressible than IID controls at matched composition, with mean gap ~ 0.174 bits/char. This supports the presence of redundant patterns not explained by one-site composition alone.

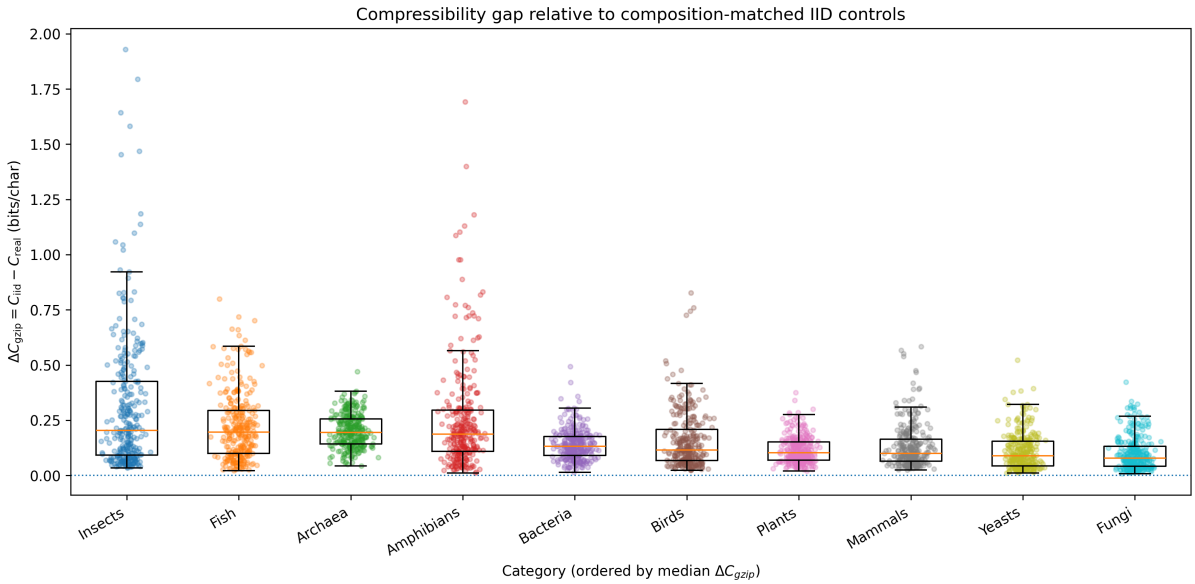


Figure 3: Compressibility gap relative to composition-matched i.i.d. controls, $\Delta C_{\text{gzip}} = C_{\text{iid}} - C_{\text{real}}$. Boxplots summarize bootstrap replicates aggregated by broad biological category; points show individual bootstrap replicates. Positive values indicate that real proteome sequences are systematically more compressible than i.i.d. surrogates at matched composition, consistent with non-random ordering constraints.

9.1 Relation to maximum-entropy and covariation models

The null-model ladder used here is compatible with a maximum-entropy view: one-site constraints produce IID models, pairwise constraints induce short-range dependencies, and richer constraint sets generically induce longer-range mutual information [3, 6, 7, 18]. The present analysis does not attempt to infer explicit couplings (as in Potts/DCA); instead, it provides a lightweight diagnostic layer for comparing ensembles before committing to mechanistic models.

9.2 Limitations

Estimation bias. Plug-in estimators for entropy and mutual information can be biased, especially when counts are sparse [13]. Bootstrap variability partially addresses sampling uncertainty but does not eliminate estimator bias. Future work can incorporate bias-corrected or Bayesian estimators.

Proteome subsampling and filtering. For computational feasibility, large proteomes were subsampled and sequences shorter than $L_{\min} = 80$ were excluded. These choices can shift estimated statistics by changing the mixture of proteins contributing to pooled counts. Sensitivity analyses (varying caps and $L_{\min}$) are recommended for high-stakes comparisons.

Interpretation of long-range dependence. Elevated I_d for $d > 1$ does not uniquely identify a biological mechanism. It can reflect repeats, domain architecture, compositional segmentation, or heterogeneous mixtures of protein families. Disentangling these contributions requires stratified analyses (e.g. by subcellular localization, domain composition,

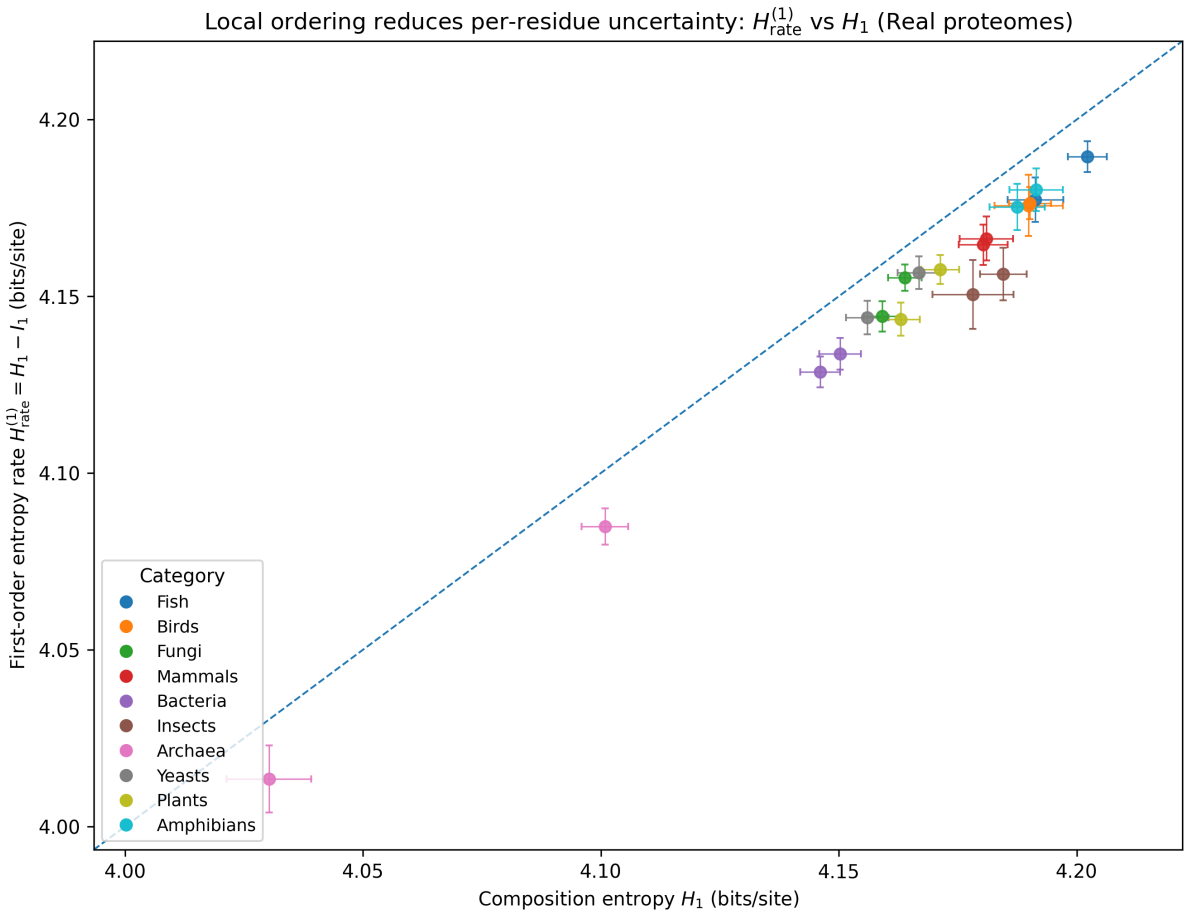


Figure 4: First-order (Markov-1) entropy rate $H_{\text{rate}}^{(1)} = H_1 - I_1$ versus composition entropy H_1 for real proteomes (points: proteome-level mean $\pm$ s.d. across bootstrap replicates). The diagonal $H_{\text{rate}}^{(1)} = H_1$ indicates vanishing adjacent dependence ($I_1 = 0$); systematic displacement below the diagonal quantifies the reduction in per-residue uncertainty induced by local correlations. Dotted lines mark the uniform-alphabet maximum $\log_2 20$.

or functional classes).

10 Conclusion

We developed a theoretical and empirical framework for viewing proteomes as statistical languages: ensembles of strings generated by constrained stochastic processes. We defined intrinsic descriptors (composition entropy H_1, adjacent mutual information I_1, information profiles I_d, and a compressibility proxy), proposed a principled null-model ladder, and proved a general entropy-deficit bound linking rarity constraints to unavoidable statistical structure. Empirically, across 20 UniProt reference proteomes, real sequences depart from composition-matched i.i.d. controls and frequently exhibit dependencies beyond first-order Markov surrogates, yielding distinct cross-clade statistical fingerprints. This framework provides a clean, reproducible foundation for comparing proteome-level sequence structure without requiring structural annotation or semantic assumptions.

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2 edition, 2006.

[2] Ken A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 1985.

[3] Matteo Figliuzzi, Hugues Jacquier, Alexander Schug, Olivier Tenaillon, and Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in proteins. *Molecular Biology and Evolution*, 33(1):268–280, 2016.

[4] Miruna Ganesh et al. Compression and complexity of protein sequences. *Bioinformatics*, 28(9):1181–1186, 2012.

[5] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[6] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Mark Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35:128–135, 2017.

[7] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, 1957.

[8] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson, Upper Saddle River, NJ, 2 edition, 2009.

[9] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.

[10] Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, New York, 3 edition, 2008.

[11] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

[12] John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

[13] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

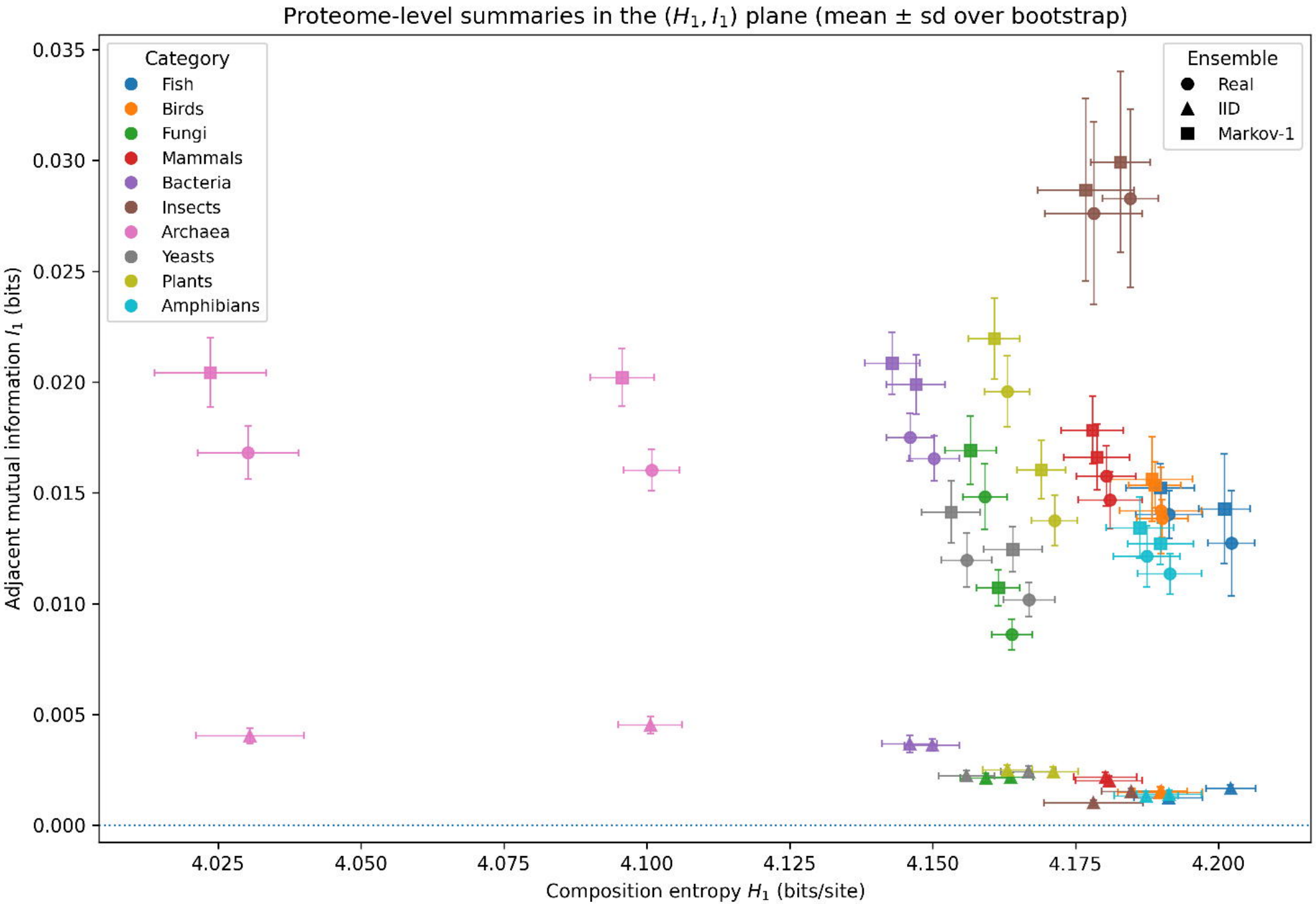
[14] Eugene I. Shakhnovich. How does a protein fold? *Nature*, 389:96–98, 1997.

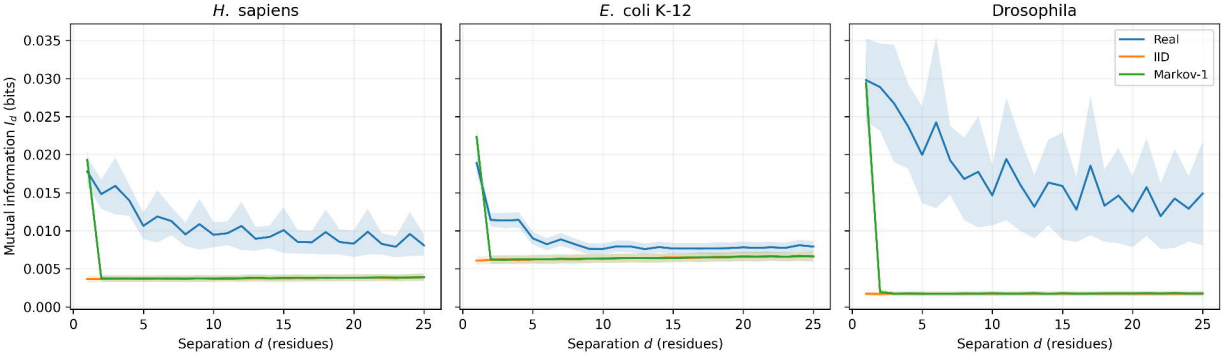
[15] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

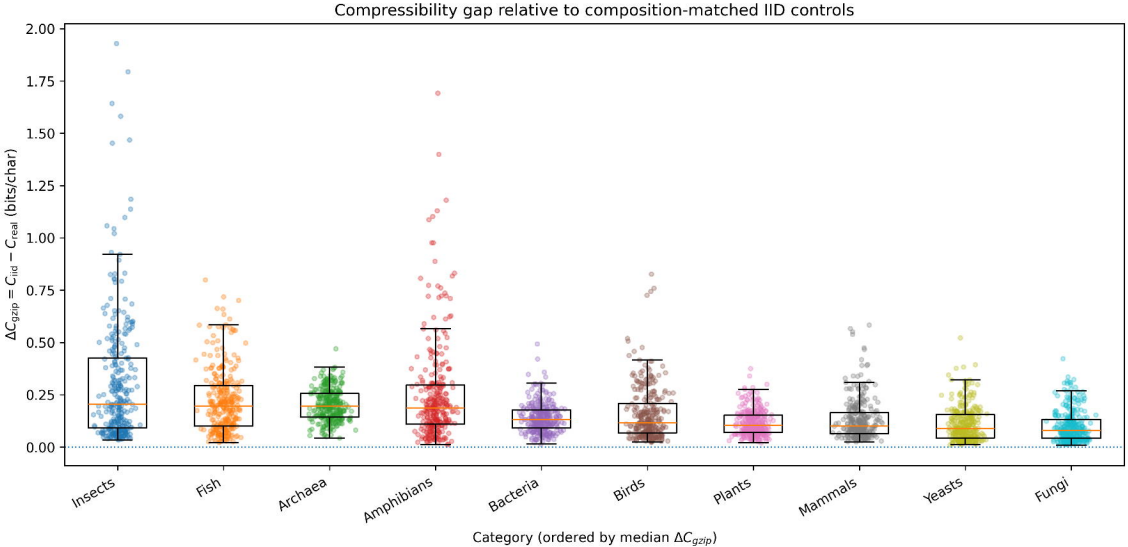
[16] Yong Tao et al. Information content of protein sequences. *Journal of Theoretical Biology*, 264(3):703–712, 2010.

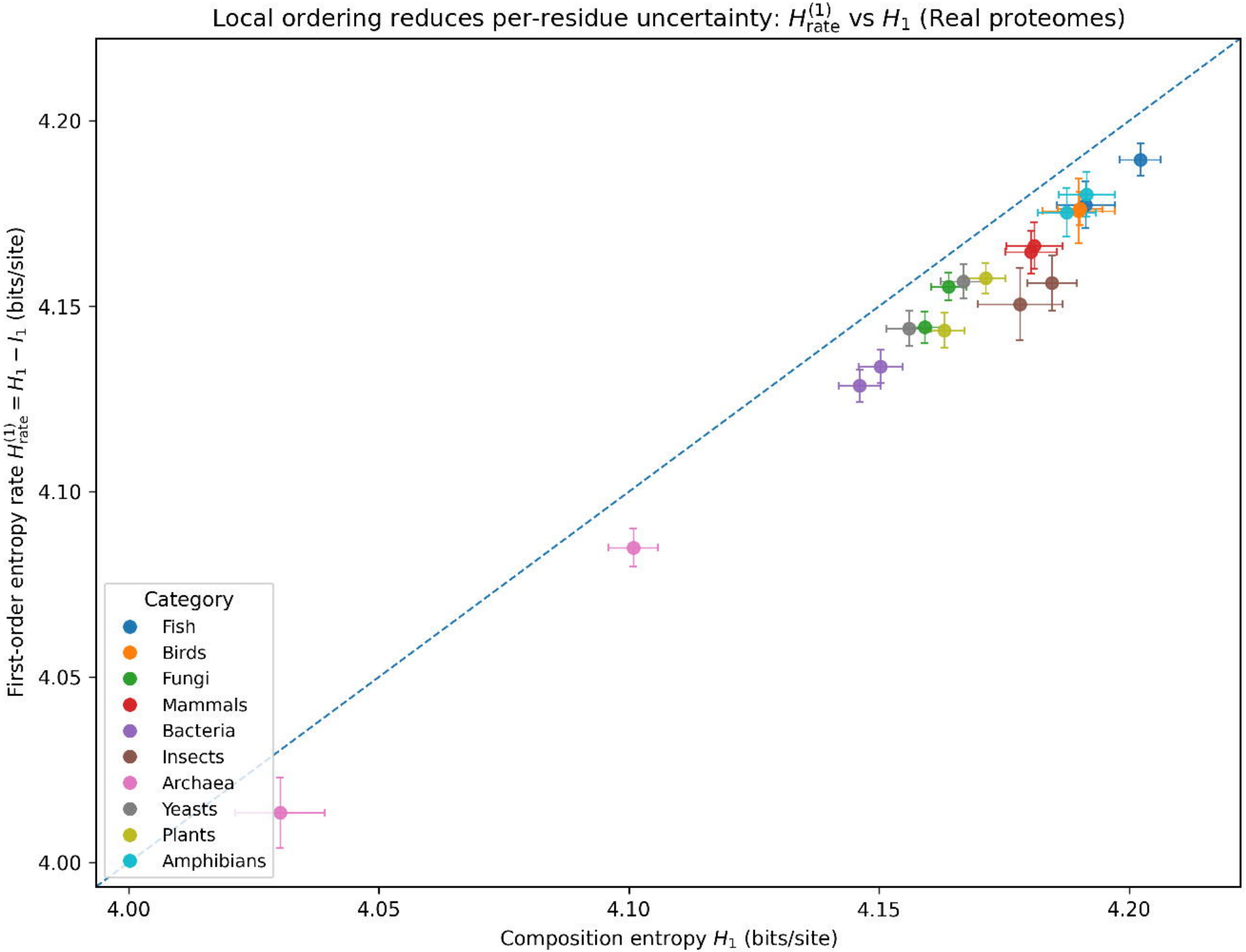
[17] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

[18] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.


Proteome-level summaries in the (H_1, I_1) plane (mean ± sd over bootstrap)
Category
Fish
Birds
Fungi
Mammals
Bacteria
Insects
Archaea
Yeasts
Plants
Amphibians
Ensemble
Real
IID
Markov-1
Adjacent mutual information I_1 (bits)
Composition entropy H_1 (bits/site)
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
4.025
4.050
4.075
4.100
4.125
4.150
4.175
4.200


H. sapiens
E. coli K-12
Drosophila
Mutual information I_d (bits)
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
5
10
15
20
25
Separation d (residues)
Real
IID
Markov-1


Compressibility gap relative to composition-matched IID controls
$\Delta C_{gzip} = C_{iid} - C_{real}$ (bits/char)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Insects
Fish
Archaea
Amphibians
Bacteria
Birds
Plants
Mammals
Yeasts
Fungi
Category (ordered by median ΔC_{gzip})


Local ordering reduces per-residue uncertainty: $H_{\text{rate}}^{(1)}$ vs H_1 (Real proteomes)
First-order entropy rate $H_{\text{rate}}^{(1)} = H_1 - I_1$ (bits/site)
Composition entropy H_1 (bits/site)
4.00
4.05
4.10
4.15
4.20
Category
Fish
Birds
Fungi
Mammals
Bacteria
Insects
Archaea
Yeasts
Plants
Amphibians