

IMPACT OF EPISTASIS ON PHYLOGENETIC TREES

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE

by
Mina Mahboubi
December 2025

Thesis Approvals:

Vincenzo Carnevale, Thesis Advisor, Biology
Shohreh Amini, Biology
Tonia Hsieh, Biology

ABSTRACT

The reconstruction of phylogenetic trees from molecular sequences is a central task in evolutionary biology. These trees are typically inferred from pairwise sequence distances, under the assumption that sequence similarity reflects shared evolutionary history. However, this task often relies on models that treat sites in a sequence as evolving independently of one another. While this simplification enables tractable inference, it does not account for the effects of epistasis (interactions among sites), which, by constraining the accessible sequence space, distort the distribution of pairwise distances. This model misspecification can introduce biases in tree topology, such as artificial hierarchies or compressed branches, even under neutral evolution and in the absence of phylogenetic relatedness.

This thesis investigates the impact of epistasis on phylogenetic tree structure through a multi-scale approach, combining theoretical modeling, numerical simulations, and natural protein sequence analysis. We introduce a two-state model of protein evolution based on an Ising-like model with tunable pairwise correlations, allowing precise control over epistatic strength and structure. We show that epistatic constraints restrict the dynamics to a subset of sequence space. By analyzing the distribution of distances between evolved sequences, we show that sequences lie on a low-dimensional hypersurface, which we call the Neutral Evolution Manifold (NEM). We then demonstrate that the dimensionality of this manifold is controlled by the strength of epistasis and significantly affects the shape of the reconstructed phylogenetic trees. Thus, we derive several useful analytical results and validate our approximations with extensive numerical simulations.

We then extend this framework to real protein sequence data. Using multiple methods, including linear autoencoders, geodesic graph analysis, and discrete metric-based estimators, we show that real proteins exhibit significantly reduced intrinsic dimensionality compared to shuffled controls. This supports the hypothesis that

epistatic constraints are a dominant factor shaping the observed sequence landscape. To isolate the impact of epistasis on tree topology, we generate synthetic MSAs from variational autoencoders trained on real protein data and assess tree structure using lineage-through-time plots and cherry proportion metrics. Finally, we apply statistical tests, including a likelihood ratio test, to quantify the dependence of tree shape on epistasis strength.

Our results reveal that even modest epistatic interactions can bias phylogenetic inference, leading to trees that suggest evolutionary structure where none exists. We conclude that a more accurate understanding of sequence evolution is essential for reliable phylogenetic reconstruction, especially in the presence of site dependencies. This work lays the foundation for dimensionality-aware models of sequence evolution and offers a geometric perspective on the relationship between sequence space constraints and tree topology.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Vincenzo Carnevale, for his guidance throughout this research. I am also grateful to my committee members, Professor Shohreh Amini and Professor Tonia Hsieh, for their support, encouragement, and helpful feedback on my thesis.

I would like to thank my lab members, and in particular Dr. Pietro Chiarantoni, for his help.

I am deeply thankful to my family for their constant support, and especially to my husband, Dr. Meisam Zaferani, whose encouragement and inspiration made this journey possible.

TABLE OF CONTENTS

	Page
ABSTRACT	ii
ACKNOWLEDGMENTS	iv
LIST OF FIGURES	viii
LIST OF ABBREVIATIONS	x
CHAPTER	
1 INTRODUCTION	1
1.1 The Problem of Gene Family Phylogenetic Inference	1
1.2 Potential Role of Epistasis in Shaping Phylogenetic Trees	3
1.3 Advances in Generative Models for Molecular Evolution	4
1.4 Fitness Landscapes and the Geometry of Protein Sequence Space	7
2 RESULTS 1: A MINIMALIST EPISTATIC MODEL OF SEQUENCE EVOLUTION	10
2.1 Distance Estimation and the Geometric Consequences of Correlation	12
2.2 Constructing a Minimal Model of Epistasis	13
2.3 Simulating Sequence Evolution Using Markov Chain Monte Carlo	16
2.4 Geometric Analysis of Distance Distributions	18
2.5 Mutational Dynamics and Phylogenetic Tree Shape	21
2.5.1 Autocorrelation and Convergence to Equilibrium	21
2.5.2 Recurrence Time and Dimensional Reduction	24
2.5.3 Lineage-Through-Time (LTT) Curves and Tree Topology	27
3 RESULTS 2: DIMENSIONAL COMPRESSION AND TREE INFERENCE IN NATURAL SEQUENCES	31
3.1 Method I: Linear Autoencoders	33

3.2	Method II: Geodesic Graph-Based Estimation of Intrinsic Dimensionality	35
3.3	Method III: Intrinsic Dimensionality Estimation in Discrete Sequence Spaces	37
3.4	Epistasis as the Source of Dimensional Compression	40
3.4.1	Column Shuffling as a Tool to Disrupt Epistasis	42
3.4.2	Quantifying the Effect: Intrinsic Dimensionality vs. Epistasis Strength	43
3.5	Epistasis and the Topology of Phylogenetic Trees	45
3.5.1	Tree Balance and the Number of Cherries	46
3.5.2	Likelihood Ratio Test Against the Star Model	47
4	DISCUSSION	49
4.1	Summary of Findings	49
4.2	Interpretation of Theoretical Models	49
4.3	Evidence from Natural Protein Sequences	50
4.4	Consequences for Phylogenetic Inference	51
4.5	Broader Implications	51
4.6	Limitations	52
4.7	Future Directions	52
4.8	Conclusions	53
	BIBLIOGRAPHY	54
	APPENDICES	61
A	DERIVATION OF THE DISTRIBUTION OF GEODESIC DISTANCES ON A $(D-1)$ -SPHERE	62
B	DERIVATION OF THE DISTRIBUTION OF EUCLIDEAN DISTANCES ON THE SURFACE OF A HYPERSPHERE	66
C	MAPPING BETWEEN HAMMING AND EUCLIDEAN DISTANCES	69

D	CONVERGENCE OF DISTANCE DISTRIBUTIONS ON THE L -CUBE AND THE $(L-1)$ -SPHERE	72
E	NORMALIZED HAMMING DISTANCE OF BINARY SEQUENCES . . .	78
F	BLOCKWISE 1D ISING MODEL	79

LIST OF FIGURES

Figure		Page
1	Phylogenetic tree reconstruction for sequences evolved from the same founder	6
2	True Phylogenetic Tree vs Inferred One	11
3	Blockwise Correlation in Ising Model	15
4	Correlation's Effect on Geometry of Sequence Space	20
5	Autocorrelation Time and Convergence to Equilibrium	23
6	Fraction of Recurrent Pairs $F_r(t)$ For Radius $r = 1$ Across Block Sizes.	25
7	Normalized Recurrence as a Linear Function of ID	26
8	Sequence Extraction Process and LTT Test	28
9	Inflection Point and Its slope of Sigmoidal Fit To LTT	29
10	Connection Between the Inflection Point and ID	30
11	Reconstruction Accuracy of a LAE Trained on Local Clusters of Protein Sequences	34
12	RMSD Between the Empirical Distribution of Geodesic Distances and the Theoretical Distribution on a D -dimensional Hypersphere	37
13	Posterior Distribution over Dimension	40
14	Epistatic Couplings	41
15	Epistasis Modulation via Shuffling MSA	42

16	Epistasis–dimensionality Relationship	44
17	Effect of Epistasis on Number of Cherries	47
18	Log-Likelihood Ratio Test (LLRT) as a Function of Epistasis Strength	48
19	Variance of Normalized Hamming Distance $\text{Var}(\hat{d}_H)$ as a Function of Block Size b for Different Temperatures, with $J = 1$	86

LIST OF ABBREVIATIONS

I3D	Intrinsic Dimension for Discrete Dataset
ID	Intrinsic Dimension
ILS	Incomplete Lineage Sorting
LAE	Linear Autoencoder
LLR	Likelihood Ratio
LTT	Lineage Through Time
MCMC	Markov Chain Monte Carlo
MSA	Multi Sequence Alignment
NEM	Neutral Evolution Manifold
NJ	Neighbor Joining
RMSD	Root Mean Square Deviation
TVD	Total Variation Distance
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
VAE	Variational Autoencoder

CHAPTER 1

INTRODUCTION

The core problem of molecular phylogenetics is uncovering how present-day genes evolved from their ancestors by reconstructing the branching events that shaped their history. This reconstruction is obtained exclusively from the sequences we observe today, with no direct records of extinct intermediates. Each gene sequence, therefore, serves as a clue—an evolutionary “trace” of past mutations, divergences, and shared ancestries. By comparing these traces across multiple organisms or genes, we piece together plausible histories of divergence, building a tree-like representation that illustrates how extant lineages likely branched off from common ancestors over evolutionary time.

1.1 The Problem of Gene Family Phylogenetic Inference

Gene families are groups of genes that come from a shared ancestral gene and diversify over time into orthologs and paralogs [1, 2, 3]. Orthologs, which arise from speciation events, typically preserve their ancestral biological roles across divergent species and thus serve as reliable indicators of conserved functions. For example, the *FOXP2* gene, linked to neural development and vocalization, is common among mammals and highly conserved, reflecting its critical and preserved function across species [4]. In contrast, paralogs emerge from gene duplication events within the same genome and often diverge functionally, contributing to phenotypic innovation. A well-known example is the globin gene family, whose members have evolved distinct oxygen-binding properties tailored to different tissues and developmental stages, such as embryonic, fetal, and adult hemoglobins [5].

Studying gene families sheds light on the evolutionary history of genomes and the mechanisms that drive functional diversification [6]. After a duplication

event, genes can either acquire new functions (neofunctionalization) [7] or divide their original functions (subfunctionalization), contributing to the emergence of novel traits and biological specialization. Phylogenetic trees constructed from gene families are key to identifying these duplication events and elucidating the evolutionary pathways leading to novel traits and the specialization of biological functions [8, 9].

Despite offering crucial insights into molecular evolution, phylogenetic reconstruction of gene families remains a complex and error-prone task. A key challenge lies in correctly distinguishing between orthologous and paralogous relationships, which often requires identifying and interpreting discrepancies between gene-family trees and the overarching species tree [10, 11]. While gene duplication events are expected to produce specific topological patterns in the tree [1], these signals are frequently masked by gene loss, where a gene present in an ancestral species is lost in one or more descendant lineages [12]. This results in missing branches and ambiguous relationships within the inferred trees.

Furthermore, incomplete lineage sorting (ILS) adds another layer of complexity. ILS occurs when ancestral gene variants persist across multiple speciation events and are sorted randomly among descendant species [13], sometimes producing gene trees that conflict with the actual species relationships [11]. Additional biological processes—such as horizontal gene transfer [14], hybridization [15], introgression [16], recombination [17], and convergent molecular evolution [18, 19]—can similarly distort tree topologies and lead to incorrect inferences if not properly accounted for. All these factors highlight a major problem in molecular phylogenetics: the inferred shape and structure of gene-family trees often do not straightforwardly reflect evolutionary history. This brings us to a key unresolved question: what underlying factors shape the topology of phylogenetic trees, and how much of the observed complexity in gene-family trees reflects true evolutionary processes? Understanding this is essential for accurate phylogenetic inference.

Several computational frameworks have been developed to disentangle the contributions of gene duplication, loss, ILS, and other evolutionary processes to observed gene-family tree topologies [10, 20, 21, 22]. These models aim to reconcile discordances between gene and species trees by inferring the most likely sequence of evolutionary events. However, a common limitation of most such approaches is their reliance on site-independent substitution models [23, 24], which assume that each site in a sequence evolves independently of others. This assumption simplifies computation but overlooks a biologically significant factor: epistasis.

1.2 Potential Role of Epistasis in Shaping Phylogenetic Trees

Epistasis occurs when the effect—or even the probability—of a mutation at one site depends on the genetic context, i.e., the presence or absence of variants at other sites within the same sequence [25]. In protein-coding genes, where structure and function are often shaped by inter-residue interactions, such dependencies are especially prominent. Ignoring epistasis can distort evolutionary inferences by misrepresenting the constraints that shape sequence evolution.

Recent simulation-based studies underscore this concern. Nasrallah et al. (2011) showed that neglecting pairwise epistatic interactions reduces the accuracy of phylogenetic inference [26]. Similarly, Magee et al. (2021) found that while the degradation in inference accuracy due to epistasis may not always be catastrophic, it is consistent and measurable across a range of evolutionary scenarios [27]. These results highlight a critical gap in current models: the shape of phylogenetic trees may be influenced not only by evolutionary events like duplication and ILS, but also by constraints imposed by epistasis.

This raises an important question that motivates the present work: To what extent can epistasis induce distortions in inferred tree topology, and how can these distortions be quantified or corrected? In the following chapters, we address this

question by first constructing theoretical models that incorporate epistatic constraints and then extending these insights to real protein sequences.

1.3 Advances in Generative Models for Molecular Evolution

Recent developments in generative modeling have transformed how protein sequence evolution is studied. Machine learning models such as Variational Autoencoders (VAEs) [28, 29, 30, 31, 32], Transformers [33, 34], Diffusion Models [35], and Potts models [36] now generate synthetic protein sequences that closely reflect the statistical and structural properties of natural sequences. These models capture complex mutational patterns by reproducing high-order correlations arising from intramolecular epistasis. The Potts model, inferred from multiple sequence alignments, estimates residue-residue couplings that reflect coevolutionary constraints. These couplings encode not only local interaction patterns but also long-range dependencies important for structural stability and function. Similarly, deep generative models like VAEs and diffusion-based frameworks learn latent manifolds that support the generation of viable sequence variants within biologically meaningful regions of the sequence space.

These innovations enable controlled simulations of protein evolution under realistic conditions, where sequence diversity and constraint patterns are preserved. The resulting synthetic datasets provide a new lens through which to examine fundamental questions in molecular evolution, such as the distribution of sequence distances, the emergence of conserved sites, and the statistical structure of phylogenetic trees. This paradigm shift—from empirical sequence collection to mechanistic generative modeling—offers a testbed for evaluating evolutionary hypotheses *in silico* under known ground truth conditions [37, 38, 39, 40, 41, 42, 43, 44, 45].

In this work, we leverage this framework to investigate how epistasis influences the structure of phylogenetic trees inferred from synthetic protein sequence alignments. A primary observation motivating this work arises from a comparison

of phylogenetic trees inferred from sequences generated by different models using the multiple sequence alignment of the PFAM family PF00520. In a controlled experiment, we simulated the neutral evolution of 1,000 sequences from a shared ancestral state under two conditions: an independent site model and an epistatic Potts model. Both models were calibrated to reproduce identical marginal amino acid distributions, using the same founder sequence and evolving in parallel over 200 generations. Despite these shared initial conditions and equivalent constraints at the single-site level, the resulting phylogenetic trees differ markedly in structure. To quantify these structural differences, we analyzed the distribution of cherries as a measure of tree balance. In phylogenetics, a cherry refers to a pair of leaves (tips) that share a direct common ancestor—i.e., they form a two-taxon clade. The cherry proportion quantifies the number of such leaf pairs in a tree, normalized by the maximum possible number for a given number of tips. This metric serves as a proxy for tree balance, with higher values indicating more symmetric topologies [46].

Figure 1 compares the outcomes of two models: Panel A shows results from the independent site model, while Panel B shows results from the epistatic Potts model. In each panel, the left side displays a representative phylogenetic tree reconstructed from one of the simulations, and the right side shows the distribution of normalized cherry proportions across all 50 repetitions. In Panel A, the tree appears balanced, and the cherry proportions are generally higher. In Panel B, the tree is more asymmetric, and the cherry proportions are consistently lower. The result suggests that correlations among sites can introduce biases in inferred tree topologies, even under neutral evolution. This finding prompts a broader question: to what extent does sequence similarity reliably reflect evolutionary history?

While sequence similarity—and by extension, tree topology—is often interpreted as evidence of shared ancestry, this assumption does not always hold. Similar sequences can also arise between unrelated proteins through random coincidence or

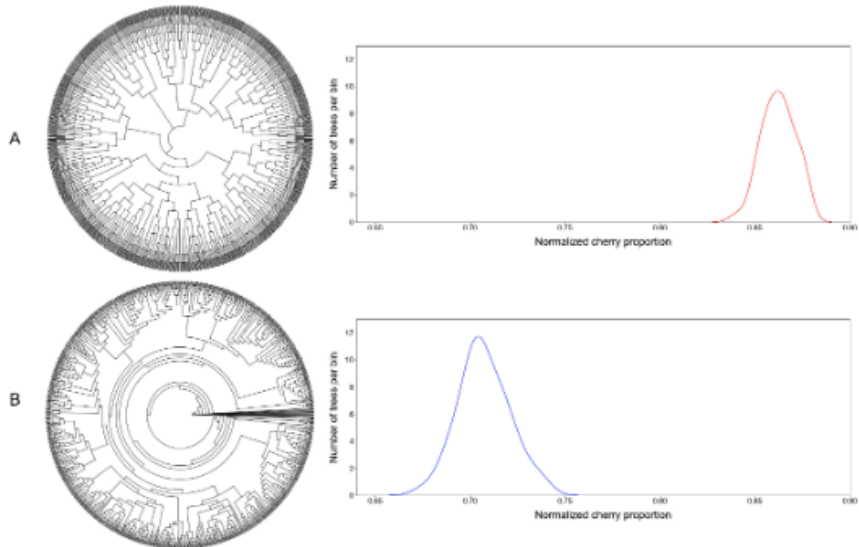


Figure 1: **Phylogenetic tree reconstruction for sequences evolved from the same founder.** Panel A: sequences generated by an independent site model. Panel B: sequences generated by a Potts model. Despite identical evolutionary conditions (200 generations, neutral drift), the Potts model produces a tree with a strong hierarchy and fewer cherries. This highlights the spurious phylogenetic structure introduced by epistasis.

convergent evolution driven by functional constraints [47]. Our simulation results reflect this concern: even under neutral evolution, epistatic interactions alone can produce tree structures with apparent hierarchy. When residues co-evolve, evolution becomes confined to a limited set of accessible directions in sequence space. This confinement imposes a geometric structure on the distribution of observed pairwise distances between sequences and suggests that tree topology may not solely reflect a record of branching events, but may also bear the geometric imprint of the underlying sequence space. To understand how such constraints influence tree topology, it is essential to examine the geometry of protein sequence space and its organization by the fitness landscape. The next section introduces these foundational concepts and formalizes the idea of the Neutral Evolution Manifold (NEM)—the effective subset of sequence space explored under neutral and epistatically constrained dynamics.

1.4 Fitness Landscapes and the Geometry of Protein Sequence Space

The fitness landscape is a mathematical representation that assigns a scalar value to each protein sequence, quantifying its biological efficacy. Formally, it is a function $f : \mathcal{S} \rightarrow \mathbb{R}$, where \mathcal{S} denotes the space of all possible protein sequences of a given length, and $f(s)$ is the fitness value of sequence $s \in \mathcal{S}$. This value reflects the sequence's ability to fold into a stable structure, perform a specific function, or contribute to the reproductive success of the organism.

In the simplest models, the fitness of a sequence is additive: each site contributes independently to the total fitness. Such models result in smooth landscapes with a single global optimum [48]. However, proteins are subject to a variety of physical and biochemical constraints, including epistasis. Epistasis introduces ruggedness into the landscape, characterized by multiple local optima and large regions of near-neutral fitness [48, 49, 50].

According to Kimura's neutral theory, most mutations that become fixed in a population are selectively neutral — they do not significantly alter the organism's fitness. As a result, molecular sequences can drift through sequence space, accumulating mutations, as long as the resulting fitness remains within a nearly similar range [51]. The dynamics in this regime can be approximated as a random walk restricted to sequences of roughly equal fitness [52].

Protein Sequence Space and Geometric Embedding. Protein sequences of fixed length L , composed from an alphabet of size q , can be represented as points in a discrete, high-dimensional space, where each axis corresponds to a residue position and each coordinate represents a possible amino acid. Mathematically, the full sequence space, for sequences of length L over an alphabet of size q , contains q^L possible configurations. In binary models ($q = 2$), sequences are often represented as vectors $s = (s_1, s_2, \dots, s_L) \in \{-1, +1\}^L$, forming the vertices of an L -dimensional hypercube

[53]. In such a space, two sequences are connected by an edge if they differ at exactly one position (a single mutation), establishing a natural graph structure via Hamming distance [54].

Epistasis Makes Only a Subset of Sequence Space Accessible. While the ambient dimension of this space is $q * L$, not all directions are equally explored under realistic evolutionary scenarios. Epistasis limits the accessibility of many regions in sequence space, thereby concentrating the distribution of viable or observed sequences onto a lower-dimensional structure. Formally, a finite set of sequences $A \subset \{q\}^L$ has no intrinsic geometry beyond its cardinality. However, once this set is embedded into a metric space, it becomes meaningful to ask whether the points approximately lie on a lower-dimensional surface. The key question then becomes: can the accessible sequence ensemble be approximated by a manifold of reduced dimension $D \ll q * L$?

The Neutral Evolution Manifold. We define the Neutral Evolution Manifold (NEM) as the effective geometric support of the sequence ensemble evolving under neutral dynamics with epistatic constraints. The NEM is not an explicit geometric object defined by coordinates, but rather an abstract manifold $\mathcal{M} \subset \mathbb{R}^{q*L}$ of dimension D , such that the empirical distribution of accessible sequences is concentrated near \mathcal{M} . This notion captures the idea that although the full sequence space is combinatorially large, neutral evolution in the presence of epistasis explores only a structured, low-dimensional subset of that space, which in turn affects the statistical properties of sequence ensembles and the topology of inferred phylogenetic trees [55]. Understanding the geometry of the NEM is crucial for accurately modeling evolutionary processes. In the following chapters, we formalize this concept and analyze how different types of epistatic interactions affect the shape and dimension of the NEM.

Investigating the NEM’s Geometry. This thesis is organized to build a conceptual and computational framework for understanding how epistasis influences the geometry of sequence space and, in turn, reshapes the topology of phylogenetic trees. Following this introductory chapter, the subsequent chapters progressively develop and test this framework, beginning with minimal models and ending with analyses of real protein data. The next chapter lays the theoretical foundation by examining simplified two-state sequence models, where sequences are encoded as binary strings evolving under either independent-site dynamics or correlated dynamics induced by epistatic couplings. This minimal model enables the derivation of exact results and offers intuitive geometric interpretations of how epistasis reduces the effective dimensionality of the space. Through both analytical calculations and numerical simulations, the chapter establishes a direct link between the strength of site-site interactions and the emergence of hierarchical structure in the phylogenetic tree. Building on this foundation, the third chapter transitions to real protein sequence datasets and biologically motivated generative models. Here, the focus shifts to assessing whether the dimensionality-reducing effects observed in simplified models persist in natural systems. Various techniques—including shuffling experiments, generative modeling, and topological tests on inferred phylogenies—are applied to disentangle the role of epistasis from true relatedness among sequences.

Together, these chapters develop and validate a new perspective on sequence evolution: one that treats epistasis not merely as a biochemical constraint but as a geometric force capable of reorganizing the space of possible sequences and distorting the inferred evolutionary history.

CHAPTER 2

RESULTS 1: A MINIMALIST EPISTATIC MODEL OF SEQUENCE EVOLUTION

To motivate the central question of this chapter, we begin with a conceptual schematic that illustrates how distance-based phylogenetic inference can distort the true evolutionary relationships among sequences. Figure 2 presents two scenarios involving five binary sequences composed of letters A and B. Each panel compares the true evolutionary history, the distribution of Hamming distances, and the inferred phylogenetic tree derived from those distances. Panel a represents a case in which the sequences share an evolutionary history: they have diverged from a common ancestor through a series of branching events. The left subfigure shows the true phylogenetic tree, where two representative sequences are highlighted, along with their patristic distance—the sum of branch lengths connecting them along the tree—and their Hamming distance, which simply counts the number of differing characters. In the center, we show the distribution of all pairwise Hamming distances among the sequences. On the right is the inferred tree constructed from these distances. While the Hamming metric approximates the patristic distances to some extent, it may inherently not capture the true evolutionary relatedness. Panel b depicts a contrasting scenario in which the sequences are not evolutionarily related—they are equally distant from one another and do not descend through a branching process. This corresponds to a star phylogeny in which all sequences are equidistant from a common founder. The true tree (left) reflects this structure, where all patristic distances are equal. However, due to random fluctuations and local correlations, the pairwise Hamming distances (center) vary, leading to an inferred tree (right) that displays artificial hierarchy, despite the absence of genuine evolutionary relatedness. This schematic encapsulates the major problem we aim to explore: distance-based inference methods, such as those using Hamming distance, can produce spurious phylogeny.

This occurs when the metric used for inference fails to accurately reflect the underlying evolutionary process, often due to the presence of correlations or epistasis. The remainder of this chapter develops a theoretical framework to investigate how these distortions arise.

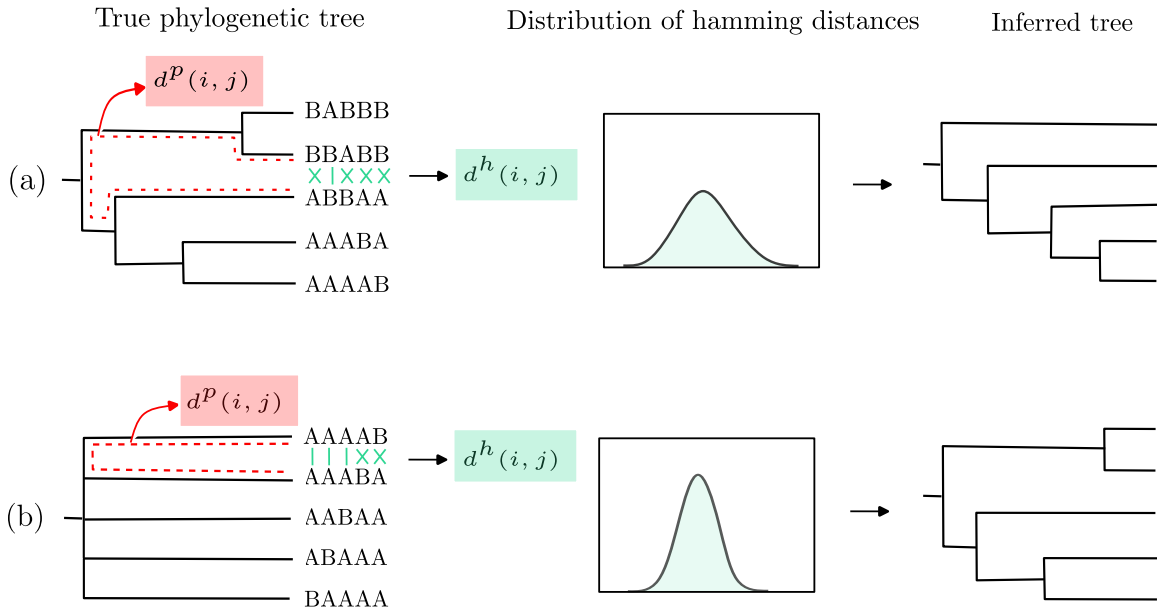


Figure 2: **True Phylogenetic Tree vs Inferred One.** (a) When the true history is hierarchical, Hamming distances can approximate the underlying structure. (b) When the true history is star-like, correlations among sites cause sequences with closer Hamming distances to cluster, producing a spurious hierarchy in the inferred tree.

To investigate the distortions illustrated above, we require models that are both analytically tractable and capable of modulating epistatic interactions. In this chapter, we introduce a minimalist model based on correlated binary characters, which allows us to explore how local epistatic interactions shape the structure of sequence space and the geometry of phylogenetic inference. By simplifying protein sequences to two-state systems and implementing blockwise correlations inspired by the Ising model, we construct a framework that isolates the contribution of epistasis to patterns of sequence divergence. This controlled setting enables direct comparison between independent and epistatic models, facilitates theoretical analysis of distance distributions, and manifold dimensionality. The simplicity of the model offers both

computational efficiency and conceptual clarity, making it an ideal starting point for investigating the geometric and topological consequences of epistasis in sequence evolution.

2.1 Distance Estimation and the Geometric Consequences of Correlation

In distance-based phylogenetic methods, the structure of a tree is inferred from pairwise distances between sequences. Ideally, the dissimilarity between two taxa should reflect their total evolutionary divergence. This is quantified by the patristic distance, which is defined as the sum of branch lengths separating two leaves in a phylogenetic tree. The patristic distance between two sequences $s^{(1)}$ and $s^{(2)}$ is denoted $d_P(s^{(1)}, s^{(2)})$ and represents the total number of substitutions (or expected substitutions under a model) accumulated since their last common ancestor [56].

However, patristic distances are not directly observable from sequence data. Instead, empirical distances are often estimated from aligned sequences using simple metrics such as the Hamming distance. For two aligned sequences of length L , $s^{(1)} = (s_1^{(1)}, \dots, s_L^{(1)})$ and $s^{(2)} = (s_1^{(2)}, \dots, s_L^{(2)})$, the Hamming distance is defined as:

$$d_H(s^{(1)}, s^{(2)}) = \sum_{i=1}^L \delta(s_i^{(1)} \neq s_i^{(2)}), \quad (2.1)$$

where $\delta(\cdot)$ is the indicator function. The normalized Hamming distance is:

$$\hat{d}_H(s^{(1)}, s^{(2)}) = \frac{1}{L} d_H(s^{(1)}, s^{(2)}). \quad (2.2)$$

In the binary $\{-1, +1\}$ encoding used in this chapter, the normalized Hamming distance can also be expressed as:

$$\hat{d}_H(s^{(1)}, s^{(2)}) = \frac{1}{2} \left(1 - \frac{1}{L} \sum_{i=1}^L s_i^{(1)} s_i^{(2)} \right). \quad (2.3)$$

This form emphasizes the geometric interpretation: the inner product $\sum_i s_i^{(1)} s_i^{(2)}$ measures angular similarity in Euclidean space after normalization (derivation of equation 2.3 can be found in Appendix E).

2.2 Constructing a Minimal Model of Epistasis

To explore the effects of site correlations explicitly, we adopt the Ising model—a canonical model from statistical physics used to study systems of binary variables with pairwise interactions [57]. A sequence $s = (s_1, \dots, s_L)$ with $s_i \in \{-1, +1\}$ is assigned an energy:

$$E(s) = - \sum_{\langle i,j \rangle} J_{ij} s_i s_j, \quad (2.4)$$

where J_{ij} denotes the coupling between sites i and j . Positive J_{ij} favors alignment ($s_i = s_j$), while negative J_{ij} favors anti-alignment.

The equilibrium distribution over sequences is given by the Boltzmann distribution:

$$P(s) = \frac{1}{Z} \exp(-\beta E(s)), \quad (2.5)$$

where β is the inverse temperature and Z is the partition function. The partition function is defined as

$$Z = \sum_{s \in \{-1,+1\}^L} \exp(-\beta E(s)), \quad (2.6)$$

that is, a weighted sum over all possible sequences in the space. It ensures normalization of the probability distribution, $\sum_s P(s) = 1$, and encapsulates the full statistical structure of the system: quantities such as average energy, entropy, and fluctuations can all be derived from Z or its logarithm. In particular, $\log Z$ corresponds to the free energy, which measures the effective number of states accessible under the given interaction structure.

At high temperatures ($\beta \rightarrow 0$), all configurations contribute nearly equally,

so $Z \approx 2^L$ and the distribution approaches the independent-site model. At low temperatures, only low-energy configurations contribute significantly, so Z effectively counts a much smaller set of states. Thus, the partition function provides a natural way to quantify how epistasis restricts accessible configurations: stronger correlations reduce the effective number of degrees of freedom, collapsing the enormous space of 2^L sequences into a much smaller manifold.

In this framework, epistasis is not added as a perturbation but built into the generative model itself. By varying the coupling structure (e.g., blockwise, random, sparse), we can simulate sequence ensembles with different correlation strengths and directly study how these affect the distribution of distances and tree topology.

Modeling Epistasis via Structured Local Correlations To explore how localized epistatic interactions influence sequence variability, we simulate correlations using a blockwise coupling scheme. Instead of coupling all sites in a fully connected network, we divide the sequence into non-overlapping blocks of fixed size b and introduce pairwise interactions only within each block. This approach mimics modular structures in real proteins, such as structural motifs or domains, where residues within a module co-evolve more tightly than those in separate regions [58].

The coupling matrix J_{ij} for the spin system is defined as:

$$J_{ij} = \begin{cases} J & \text{if } i \text{ and } j \text{ belong to the same block,} \\ 0 & \text{otherwise.} \end{cases} \quad (2.7)$$

This structured coupling introduces intra-block correlations while maintaining independence across blocks. Each block resembles a ferromagnetic domain: local interactions within the block encourage spin alignment. As temperature decreases, this results in collective behavior where entire blocks tend to stabilize in uniform spin states, thereby reducing the number of effective degrees of freedom.

Figure 3 illustrates this mechanism for sequences with different block sizes: $b = 1, 2, 4,$ and 8 . The color scale reflects correlation strength, with blue indicating weak or no correlation ($b = 1$), and red indicating strong correlations for larger block sizes. As b increases, regions of coordinated evolution expand, and the color shifts from blue to red, visually encoding the growing epistatic influence.

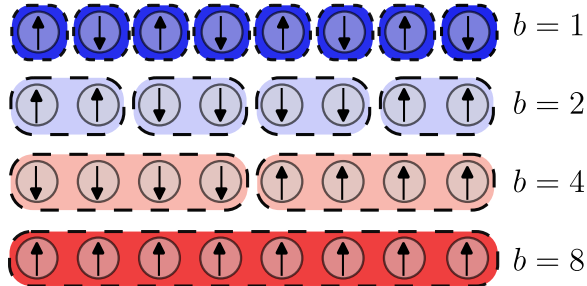


Figure 3: **Blockwise Correlation in Ising Model.** Each row shows a spin sequence under a different block size: $b = 1$ (independent sites), $b = 2$, $b = 4$, and $b = 8$. Color denotes correlation strength: blue for weak or no correlations, red for strong intra-block correlations. As block size increases, sites within each block evolve as coordinated units, introducing stronger epistasis.

The parameter b effectively controls the correlation length in the model. For $b = 1$, the model behaves like an independent-site model, where all positions mutate independently. For larger b , blocks of residues evolve in a correlated fashion, leading to sequence manifolds with fewer effective degrees of freedom. This reduction is reflected directly in the partition function. As shown in Appendix E, the blockwise Ising model has an effective partition function that scales approximately as

$$Z \sim 2^{L/b}, \quad (2.8)$$

rather than 2^L as in the independent case. In other words, correlations collapse the L degrees of freedom into L/b effective ones, since each block of size b behaves like a single coupled unit at low temperature. This scaling provides a natural measure of dimensional reduction: increasing block size compresses the accessible state space exponentially, with direct consequences for equilibrium distributions and dynamical

quantities. This framework provides a controlled setting to modulate the strength of epistasis and investigate its impact on geometric and evolutionary properties of sequences. In the next section, we define the generative model using an Ising-like Hamiltonian and describe the simulation procedure used to generate these synthetic sequences.

2.3 Simulating Sequence Evolution Using Markov Chain Monte Carlo

To investigate how structured epistatic interactions affect the distribution of pairwise distances and the topology of inferred phylogenetic trees, we simulate binary sequences sampled from the equilibrium distribution of the blockwise Ising model introduced in the previous section. The goal is to generate sequence ensembles that reflect the thermodynamic constraints imposed by the model, while retaining a mutation mechanism that resembles biological evolution.

We employ a Markov Chain Monte Carlo (MCMC) technique based on Glauber dynamics to sample from the model’s Boltzmann distribution. MCMC is a class of algorithms that constructs a stochastic process whose equilibrium distribution matches the desired target distribution, enabling efficient sampling from complex, high-dimensional spaces [59]. Glauber dynamics is a type of MCMC update rule where, at each step, a single site in the sequence is randomly selected and its state is updated probabilistically based on its local energy change [60]. This method allows us to model biologically plausible evolutionary trajectories while ensuring convergence to the equilibrium distribution.

MCMC Sampling via Glauber Dynamics Glauber dynamics proceeds by selecting a site i at random and proposing a spin flip, $s_i \rightarrow -s_i$. The proposed mutation is accepted with probability:

$$P_{\text{accept}} = \frac{1}{1 + \exp(\beta\Delta E)}, \quad (2.9)$$

where $\Delta E = E(s_{\text{new}}) - E(s_{\text{current}})$ is the change in energy due to the spin flip, and $\beta = 1/T$ is the inverse temperature. In all simulations, we set $T = 0.5$, balancing exploration and constraint.

This probabilistic rule ensures that energetically favorable changes are usually accepted, while unfavorable changes are occasionally accepted, preserving ergodicity and enabling the system to explore the full configuration space. Glauber dynamics is ideal for mimicking biological mutation because:

- It models single-point mutations.
- It satisfies detailed balance (which guarantees that the Markov chain’s stationary distribution is exactly the target distribution) and converges to the Boltzmann distribution.

Star Phylogeny Design: Independent Walkers To focus solely on the effect of epistasis without confounding effects from shared ancestry, we simulate a star phylogeny. All sequences originate from a single common ancestor and evolve independently. We simulate $N = 100$ independent trajectories (walkers), each starting from the same root sequence and subject to identical model parameters.

This setup ensures that any observed structure in the inferred phylogenetic trees arises purely from convergence in sequence space due to epistatic constraints, not from actual shared evolutionary history.

Simulation Parameters and Equilibrium We fix sequence length $L = 100$ and vary block size b to control epistasis strength:

$$b \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50, 70, 100\}.$$

Here, $b = 1$ denotes the independent-site model, and $b = 100$ indicates a fully correlated system.

Each walker is evolved for a number of MCMC steps ranging from 5×10^4 to 10^6 , depending on b , selected based on empirical autocorrelation analysis to ensure equilibrium. Larger b leads to stronger correlations and longer autocorrelation times, requiring more iterations for convergence.

The resulting ensemble of 100 sequences per b serves as the input for downstream analyses, including distance distribution measurements, intrinsic dimensionality estimation, and phylogenetic tree inference. By modulating b , we construct a continuum of models spanning from uncorrelated to highly epistatic systems, enabling investigation of how epistasis forms the geometry of sequence space and biases phylogenetic inference.

2.4 Geometric Analysis of Distance Distributions

After generating sequences using the blockwise Ising model, we analyze the geometry of the resulting sequence ensembles. Specifically, we examine how correlations among sites reshape the distribution of pairwise distances and how this can be interpreted in terms of effective dimensionality. The analytical expression for the variance of pairwise distances under this model, as a function of block size, is derived in Appendix F. This analysis is grounded in a geometric framework that maps binary sequences onto the surface of a hypersphere, allowing us to leverage results from high-dimensional geometry to quantify epistatic compression.

Each binary sequence $s = (s_1, s_2, \dots, s_L)$, with $s_i \in \{-1, +1\}$, can be viewed as a vertex of the L -dimensional hypercube [53]. To facilitate geometric analysis, we normalize each sequence by scaling it by $1/\sqrt{L}$:

$$x = \frac{1}{\sqrt{L}}s. \tag{2.10}$$

The resulting vector $x \in \mathbb{R}^L$ has unit norm and thus lies on the surface of the $(L - 1)$ -dimensional unit hypersphere S^{L-1} . This embedding enables direct comparison to theoretical distance distributions derived for uniformly sampled points on the hypersphere. The convergence between the distribution of vertices on the L -cube and that on the $(L - 1)$ -sphere is formally derived in Appendix D.

To connect this geometric representation to Hamming distance, we consider two normalized sequences $x^{(1)}$ and $x^{(2)}$:

$$d_H(s^{(1)}, s^{(2)}) = \sum_{i=1}^L \delta(s_i^{(1)} \neq s_i^{(2)}), \quad (2.11)$$

We embed the discrete Hamming space into the surface of the $(L - 1)$ -sphere. In this embedding, the Hamming distance:

$$d_{\text{Euc}}(x^{(1)}, x^{(2)}) = \|x^{(1)} - x^{(2)}\| = 2\sqrt{\hat{d}_H(s^{(1)}, s^{(2)})}. \quad (2.12)$$

Thus, through the map $s \mapsto x$, we pass from the discrete metric space $(\{\pm 1\}^L, d_H)$ to the continuous manifold $(S^{L-1}, d_{\text{Euc}})$, allowing us to leverage tools of continuous geometry to analyze sequence-space structure (derived in Appendix C).

To benchmark these empirical distances, we compare them to the theoretical distribution of chord lengths between points sampled from a uniform distribution on a hypersphere. This distribution is [61] :

$$f_D(d) = \frac{d^{D-2}}{B\left(\frac{D-1}{2}, \frac{1}{2}\right)} \left(1 - \frac{d^2}{4}\right)^{\frac{D-3}{2}}, \quad (2.13)$$

where $B(a, b)$ is the Beta function and D is the dimension (derived in Appendix B). As D increases, the distribution concentrates around its mean, reflecting the concentration of measure [62].

To estimate the intrinsic dimension, ID, we minimize the total variation dis-

tance (TVD) between the empirical and theoretical distance distributions:

$$\text{TVD}(D) = \frac{1}{2} \int_0^2 |f_{\text{empirical}}(d) - f_D(d)| dd, \quad (2.14)$$

$$ID = \arg \min_D \text{TVD}(D). \quad (2.15)$$

Figure 4b illustrates this procedure: the gray area between the empirical and theoretical curves represents the TVD, and the dimension minimizing this area is selected as ID.

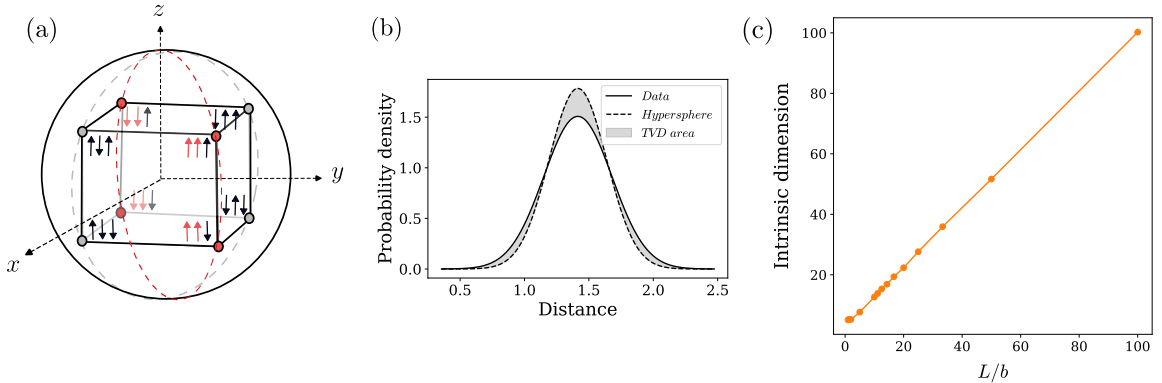


Figure 4: **Correlation's Effect on Geometry of Sequence Space.** (a) Schematic of dimensional reduction in a three-spin system. Each vertex of the cube corresponds to a configuration of three binary variables (± 1), lying on the surface of a 2D sphere in \mathbb{R}^3 . Without correlations, all eight vertices (gray) are accessible. Introducing a ferromagnetic coupling between first two spins, leaving four states on a great circle (red) accessible. (b) ID estimation by minimizing the TVD (gray area) between the empirical distance distribution and the theoretical chord distribution on S^{D-1} . (c) Estimated ID across block sizes b , plotted against L/b . Larger blocks correspond to stronger correlations and lower dimensionality.

To build intuition, Figure 4a shows a simple three-spin system. Without correlations, all eight binary configurations lie on the surface of a 2D sphere in \mathbb{R}^3 . Introducing a ferromagnetic coupling between the first two spins constrains them to align, reducing the accessible states to four points on a great circle—effectively a one-dimensional submanifold. This demonstrates how correlations compress the configuration space.

An alternative estimate can be derived from the variance of Hamming distances,

$$\sigma_H^2 = \frac{1}{4L}, \quad ID = \frac{1}{4\sigma_H^2}, \quad (2.16)$$

which provides a computationally efficient proxy when sequences are approximately uniform on a hypersphere. Empirically, as shown in Figure 4c, ID decreases with increasing block size b . Expressing the estimates as a function of L/b reveals a monotonic trend: larger blocks impose stronger correlations, reducing independent variation and collapsing the sequence space onto a lower-dimensional manifold.

2.5 Mutational Dynamics and Phylogenetic Tree Shape

The geometric structure of sequence space profoundly influences not only static properties such as the distribution of pairwise distances but also the temporal dynamics of sequence evolution. Correlations among sites, introduced via epistasis, constrain mutational pathways and slow the exploration of sequence space. This reduced dynamical freedom shapes the diversity of sequences observed over time and manifests as structural changes in inferred phylogenetic trees.

2.5.1 *Autocorrelation and Convergence to Equilibrium*

Since our simulations of sequence evolution are based on a Markov chain (via Glauber dynamics), it is important to understand how quickly this chain explores sequence space. Two complementary quantities help us characterize this: the autocorrelation function, which measures temporal dependence along the chain, and the total variation distance (TVD), which measures how close the ensemble distribution is to equilibrium.

Autocorrelation time. In a Markov chain, successive states are not independent: each new sequence is generated from the previous one by a single-site mutation. The autocorrelation function quantifies the dependency between the system’s state at time t and its state at a later time $t + \Delta t$. For an observable $O(t)$, it is defined as

$$C(\Delta t) = \frac{\langle O(t + \Delta t) O(t) \rangle - \langle O \rangle^2}{\langle O^2 \rangle - \langle O \rangle^2}. \quad (2.17)$$

If $C(\Delta t)$ is close to one, the system remains highly correlated with its past; if it is close to zero, the system is almost independent from its initial state. The rate of this decay is summarized by the autocorrelation time τ_{auto} , which may be defined either as

$$\tau_{\text{auto}} = \sum_{\Delta t=0}^{\infty} C(\Delta t), \quad (2.18)$$

or as the time at which $C(\Delta t)$ drops to $1/e$. These definitions coincide when correlations decay exponentially. Intuitively, τ_{auto} tells us how many Monte Carlo steps are required before successive samples can be treated as approximately independent [63]. Longer autocorrelation times indicate stronger constraints on sequence dynamics.

Figure 5a shows that τ_{auto} grows with block size. As epistasis strengthens, mutations within a block are no longer independent, and the chain requires more steps to decorrelate. Thus, autocorrelation time provides a direct measure of how intrablock interactions slow down exploration of sequence space.

Total variation distance. Earlier in this chapter, we introduced TVD as a way of comparing empirical distance distributions to the theoretical hypersphere distribution, thereby estimating the intrinsic dimension of sequence data. Here, we use the same concept in a different way: to quantify how far the evolving ensemble is from its stationary distribution. Specifically, the TVD between the empirical distribution

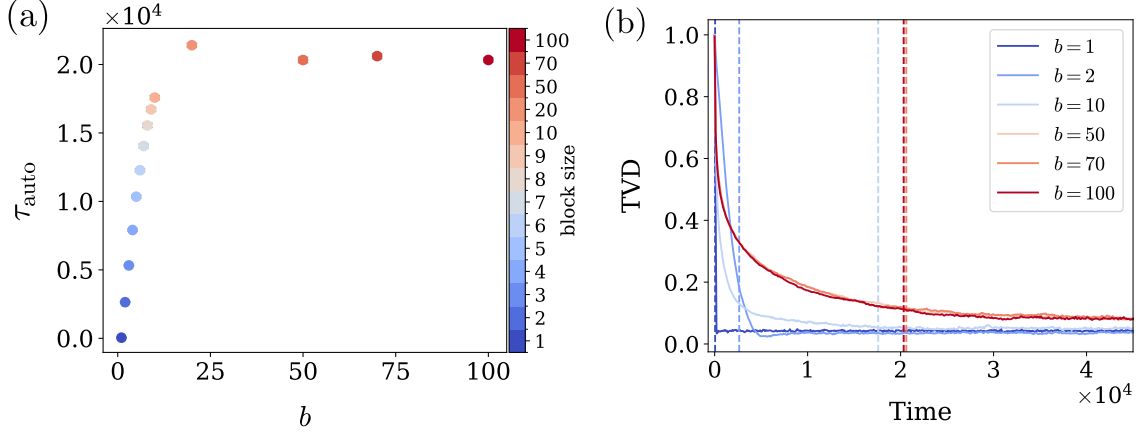


Figure 5: **Autocorrelation Time and Convergence to Equilibrium.** (a) τ_{auto} vs block size: Larger block sizes (stronger epistasis) produce longer decorrelation times. (b) $\text{TVD}(t)$ vs time t for the different block sizes. Vertical dashed lines mark $t = \tau_{\text{auto}}(b)$ from panel (a). Stronger epistasis slows both local decorrelation and global convergence to equilibrium, with τ_{auto} setting the characteristic timescale for TVD decay.

of pairwise distances at time t and the stationary distribution f_∞ is

$$\text{TVD}(t) = \frac{1}{2} \sum_d |f_t(d) - f_\infty(d)|. \quad (2.19)$$

TVD takes values between 0 and 1: it is 0 if the chain has reached equilibrium, and 1 if the current distribution is completely disjoint from the stationary one. Thus, TVD provides a rigorous way of quantifying global convergence in the Markov chain.

Figure 5b shows that TVD decreases over time as the chain converges. Stronger epistasis (larger block sizes) slows this decay, consistent with the longer autocorrelation times observed in Figure 5a. Vertical dashed lines mark the autocorrelation times, illustrating how the local persistence of correlations sets the characteristic timescale for global convergence. Together, autocorrelation and TVD give a complementary view: τ_{auto} captures dependence along trajectories, while TVD tracks the ensemble's approach to equilibrium.

Having characterized both local convergence (via autocorrelation) and global

convergence (via TVD), we now ask how often the system revisits previously explored neighborhoods of sequence space. This quantity provides a direct probe of dimensional reduction and extends our analysis from temporal persistence to spatial return dynamics.

2.5.2 *Recurrence Time and Dimensional Reduction*

While autocorrelation measures how quickly correlations between states decay, recurrence addresses this question: how often does the system revisit already explored neighborhoods of sequence space? Correlations reduce the number of effectively accessible configurations, thereby increasing the likelihood of recurrence. This makes recurrence time a natural probe of dimensional reduction: the more constrained the system, the faster typical neighborhoods are revisited.

Formally, recurrence time measures the expected number of steps for a stochastic process to return to a specified set $A \subseteq \Omega$. For an ergodic stationary process $(X_t)_{t \geq 0}$, the first return time to A is

$$R_A = \inf\{t \geq 1 : X_t \in A \mid X_0 \in A\}.$$

Kac's lemma [64] states that the mean recurrence time is the reciprocal of the stationary probability of the set,

$$\mathbb{E}[R_A] = \frac{1}{P(A)}. \quad (2.20)$$

Rare states or small neighborhoods therefore, have long recurrence times, while constrained or low-dimensional spaces yield rapid recurrence.

This probability can be expressed in statistical-mechanics form as a ratio of partition functions,

$$P(A) = \frac{Z_A}{Z}, \quad \mathbb{E}[R_A] = \frac{Z}{Z_A}, \quad (2.21)$$

where Z_A sums over states in A and Z over the full state space. In the presence of blockwise correlations, the effective number of degrees of freedom is reduced from L to $D \approx L/b$. Embedding this effective state space into a $(D - 1)$ -sphere, the neighborhood of radius r corresponds to a Hamming ball with radius r (B_r) which can be approximated by a spherical cap with volume scaling as αr^D . Thus,

$$\mathbb{E}[R_{B_r}] \approx \frac{2^D}{\alpha r^D}, \quad (2.22)$$

capturing the exponential growth of the full state space with dimension D versus the polynomial growth of neighborhoods with radius r .

Simulation results confirm these predictions. Fixing $r = 1$, we tracked the fraction of recurrent pairs $F_r(t)$ over time. As shown in Figure 6, larger block sizes lead to higher stationary recurrence fractions, consistent with reduced effective dimensionality. Fitting the exponential relaxation of $F_r(t)$ yields estimates of recurrence time τ_{rec} .

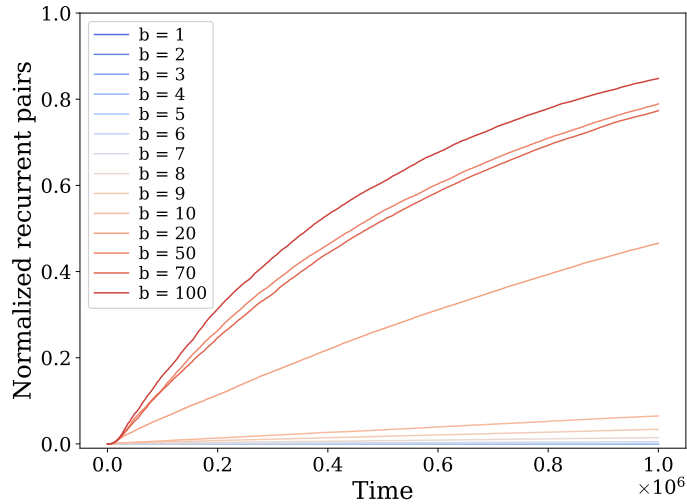


Figure 6: **Fraction of Recurrent Pairs $F_r(t)$ For Radius $r = 1$ Across Block Sizes.** Larger block sizes yield higher stationary recurrence fractions, consistent with dimensional reduction under correlations. Averages are taken over 10 independent simulations.

Then, we normalize the recurrence time τ_{rec} by the autocorrelation time τ_{auto} ,

which is the characteristic timescale of the Markov chain. from Eq. (2.22) there is a linear relation between $\log(\tau_{rec})$ and ID:

$$\log(\tau_{rec}) \approx -\log(\alpha) + \log\left(\frac{2}{r}\right) D \quad (2.23)$$

where \log denotes the natural logarithm. Thus, for fixed neighborhood radius r , the quantity $\log(\tau_{rec}/\tau_{auto})$ increases linearly with D with slope $\log(2/r)$. Smaller D implies shorter recurrence (higher revisit probability), while larger D expands the effective state space and raises τ_{rec} . As shown in Figure 7, simulation results follow this predicted linear dependence: lower intrinsic dimensions yield reduced recurrence times, an effect further amplified by increasing r , whereas higher dimensions and smaller radii correspond to markedly longer recurrence times.

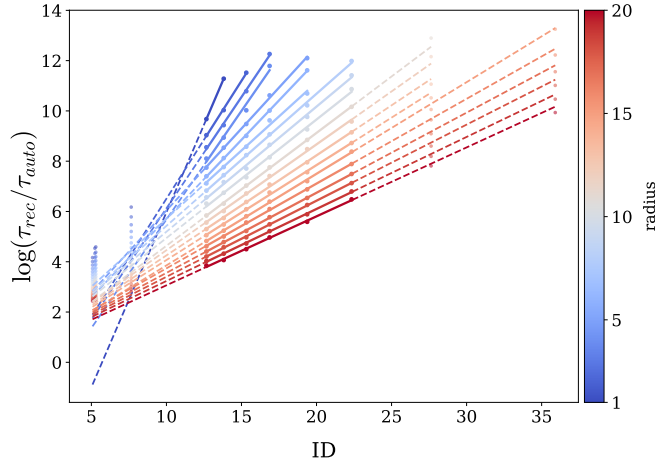


Figure 7: **Normalized Recurrence as a Linear Function of ID.** Log normalized recurrence time, $\log(\tau_{rec}/\tau_{auto})$, plotted against ID for multiple neighborhood radii r . For each r , a straight line is fit (solid) over the empirical linear regime, with dashed extensions indicating the model’s extrapolation outside the fitted range. Lower ID (stronger epistasis) yields shorter recurrence times; increasing r further elevates revisit probability (downward shift and flatter slope), whereas higher ID or smaller r lengthen recurrence.

As shown, epistasis slows convergence to equilibrium, lengthens autocorrelation times, and accelerates recurrence by constraining the accessible sequence space

to a lower-dimensional manifold. These effects, controlled by ID, directly shape the branching and diversification of evolutionary trajectories, as we explore in the next section.

2.5.3 Lineage-Through-Time (LTT) Curves and Tree Topology

To examine how epistasis alters tree topology, we simulate evolution under a star phylogeny. At fixed time intervals, we extract the current sequences from each lineage and construct a phylogenetic tree using the UPGMA algorithm [65]. This distance-based clustering method assumes uniform mutation rates and reflects our symmetric simulation design.

From each tree, we compute a Lineage-Through-Time (LTT) curve that tracks the cumulative number of branching events versus depth from the root. While these branches do not reflect true speciation events, they represent structure emerging from sequence similarity. To quantify LTT curves, we fit a sigmoid function and extract two features:

- X_{in} : Inflection point, indicating the time of rapid diversification.
- S_{in} : Slope at the inflection point, quantifying diversification rate.

Figure 8a depicts the sampling procedure and sequence-to-tree workflow. Panels b and c compare star-like and hierarchical topologies, respectively. In star-like trees (panel b), branching happens early and saturates quickly, leading to steep slopes. In more hierarchical trees (panel c), diversification is gradual and spread over time, resulting in shallower slopes.

Figure 9 illustrates the temporal evolution of two features derived from the LTT analysis under varying levels of epistasis, modeled by different block sizes. Panel (a) shows the inflection point X_{in} , which captures when branching activity is most pronounced. Panel (b) displays the slope S_{in} , quantifying the rate of lineage formation

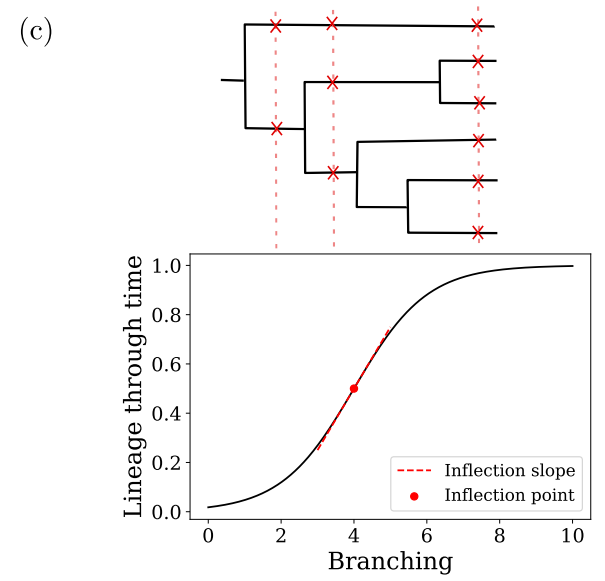
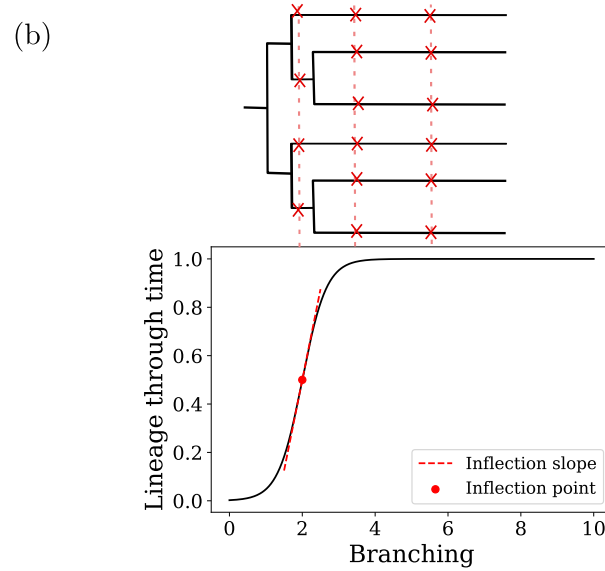
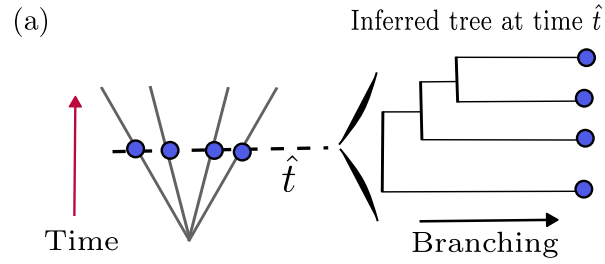


Figure 8: **Sequence Extraction Process and LTT Test** (a) Sampling scheme and inference process. Sequences are extracted at fixed time steps and clustered via UPGMA. (b) Star-like tree with early saturation in the LTT curve and steep slope. (c) Hierarchical tree with gradual branching and shallower slope at inflection.

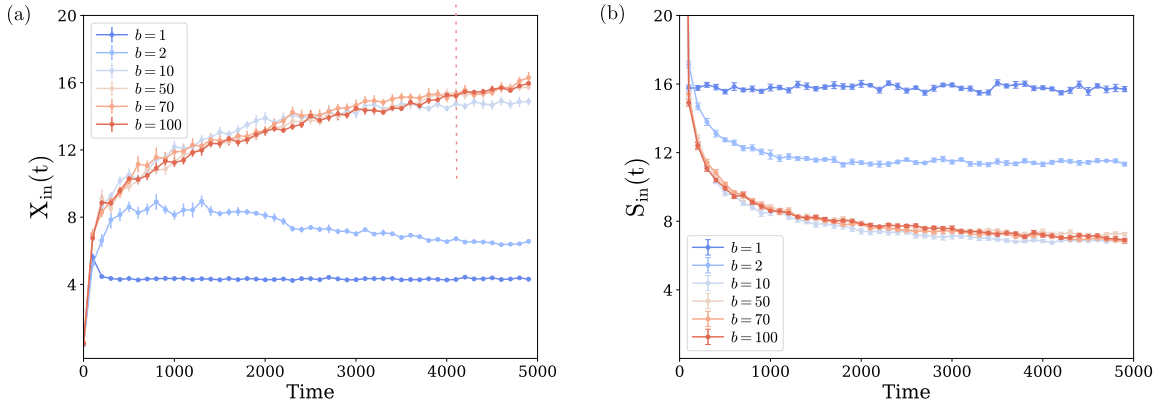


Figure 9: **Inflection Point and Its slope of Sigmoidal Fit To LTT** Inflection point (a) and slope (b) over simulation time for different block sizes. Larger blocks (stronger epistasis) delay diversification and reduce branching speed.

through time. For all block sizes, both metrics exhibit rapid changes at early times, reflecting an initial phase of accelerated diversification. As evolution progresses, these changes gradually slow down and saturate, indicating a stabilization in the branching dynamics. Importantly, increasing the block size increases X_{in} , suggesting that stronger epistatic constraints postpone the emergence of diversification. At the same time, the slope S_{in} decreases with larger block size, indicating a slower accumulation of branching events over time. This effect is also influenced by the increase in the autocorrelation time of the underlying Monte Carlo process, as stronger epistasis slows down the decorrelation of sequence states, delaying diversification events along the phylogeny.

These trends reveal a clear shift in tree structure with increasing epistasis: smaller block sizes yield more star-like trees, marked by early inflection points and steeper slopes in the LTT curve. In contrast, larger block sizes produce more hierarchical trees, where diversification is delayed and unfolds more gradually over time. This reflects how stronger epistatic constraints reduce the rate of lineage expansion and reshape the overall topology of the phylogenetic tree.

Finally, Figure 10 illustrates the relationship between the intrinsic geometry of

sequence space and the structure of the resulting phylogenetic trees. Here, we plot the inflection X_{in} and S_{in} at convergence —i.e., after the system has reached its stationary distribution—as functions of the estimated intrinsic dimensionality (inferred via TVD minimization). The results reveal a clear geometric-tree correspondence: as intrinsic dimensionality decreases, X_{in} increases and S_{in} decreases.

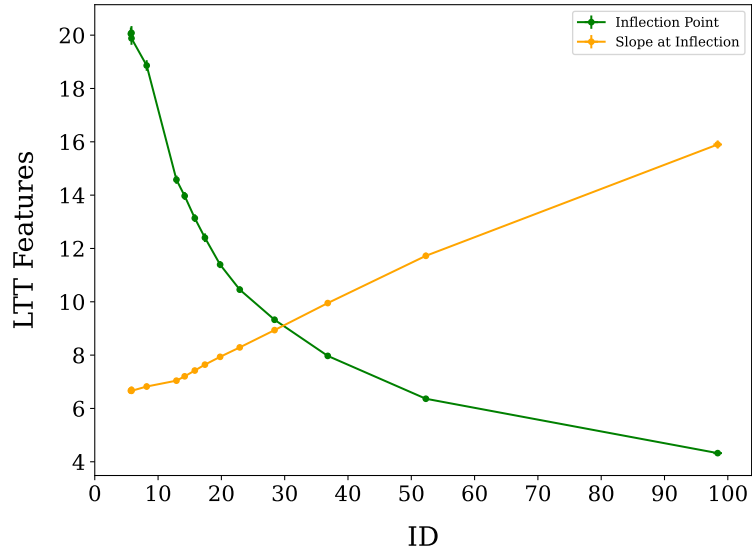


Figure 10: **Connection Between the Inflection Point and ID** Inflection index X_{in} and slope S_{in} at equilibrium as functions of intrinsic dimensionality estimated via TVD minimization. Lower dimensionality correlates with later inflection points and shallower slopes, indicating that epistatic compression leads to more hierarchical tree structures.

In other words, lower-dimensional evolutionary manifolds are associated with delayed diversification and more gradual lineage accumulation. This finding demonstrates that, even without shared ancestry, strong epistatic interactions can induce tree-like structures in inferred phylogenies. These emergent hierarchies reflect the geometry of constrained sequence space and highlight the risk of interpreting tree shape solely as evidence of historical divergence.

CHAPTER 3

RESULTS 2: DIMENSIONAL COMPRESSION AND TREE INFERENCE IN NATURAL SEQUENCES

From Models to Molecules In the previous chapter, we developed a theoretical framework to examine how epistatic interactions influence the structure of phylogenetic trees. Using simplified two-state models with blockwise Ising-like correlations, we demonstrated that epistasis compresses the accessible sequence space, reduces its intrinsic dimensionality, and induces spurious hierarchy in inferred phylogenies, even under neutral evolution. These findings suggest that deviations from star-like topologies, often interpreted as signatures of historical divergence or selection, can instead arise as geometric artifacts of constrained evolution within a low-dimensional manifold.

Building on these theoretical insights, this chapter investigates whether similar phenomena are detectable in real protein sequence data. Specifically, we test the central hypothesis that epistasis constrains the set of viable protein sequences to a lower-dimensional subset of the full sequence space, and that this dimensional compression distorts phylogenetic inference.

Protein Family The dataset used throughout this study is derived from the PFAM protein family PF00520, which corresponds to the Ion Transporter (IT) family. This family encompasses a broad class of membrane proteins responsible for facilitating the movement of ions across cellular membranes. This family includes diverse members such as sodium and calcium ion transporters, many of which share conserved structural motifs that are essential for their functional integrity. These proteins often exhibit co-evolving residue networks, which makes this family particularly suitable for analyzing the impact of correlated mutations.

To evaluate our hypothesis, we analyze multiple sequence alignments (MSAs)

introduced above and apply three distinct methods to estimate the intrinsic dimensionality of these datasets. Each method draws on a different conceptual foundation: a linear autoencoder that identifies minimal latent representations supporting accurate reconstruction; a graph-based geodesic distance estimator that infers dimension from connectivity patterns in a neighborhood graph; and a discrete metric-based approach rooted in probabilistic modeling of distances. Despite their differing assumptions and methodologies, these estimators consistently reveal that natural protein sequences occupy a significantly lower-dimensional space than expected under independent-site models.

We then explore the role of epistasis in driving this dimensional reduction. Under the assumption that correlated sites constrain the number of effective degrees of freedom, we repeatedly disrupt epistatic couplings by shuffling MSA columns and re-estimating the intrinsic dimensionality. As expected, increasing the number of shuffled columns leads to an increase in dimensionality, supporting the notion that epistasis is responsible for compressing the sequence space.

To isolate the effects of epistasis from those of shared evolutionary ancestry, we introduce a generative approach based on variational autoencoders (VAEs). By training VAEs on MSAs with varying levels of shuffled columns, we generate synthetic sequences that reproduce local statistical properties of the data while eliminating phylogenetic relatedness. The dimensionality of these generated sequences follows the same increasing trend with shuffling, confirming that the underlying correlations—and not shared descent—drive the observed compression.

Finally, we assess how these geometric constraints influence tree topology. We employ two statistical tests: a cherry test, which measures the abundance of cherry structures (pairs of closely related leaves) in reconstructed trees, and a likelihood ratio test (LLRT), which compares the fit of a star-like tree to that of a more hierarchical alternative. Both tests reveal that stronger epistatic constraints—quantified via de-

creased shuffling or higher correlation—lead to deviations from star phylogenies, even in the absence of true lineage divergence.

Together, these results demonstrate that the geometric consequences of epistasis are not confined to synthetic models but are evident in empirical protein sequence data. They underscore the need for dimensionality-aware phylogenetic inference and motivate a shift from purely historical interpretations of tree structure to ones that account for the topological imprint of the sequence space itself.

3.1 Method I: Linear Autoencoders

Here, we estimate the intrinsic dimensionality of the NEM using a Linear Autoencoder (LAE). This approach is motivated by the principle of local linearization, which suggests that a complex high-dimensional manifold can be approximated locally by linear subspaces [66, 67]. According to the manifold hypothesis, these data points, although embedded in a high-dimensional space, lie on an intrinsically low-dimensional manifold [68]. Manifold learning techniques aim to uncover this underlying structure by identifying a compact set of latent variables that preserve the essential features of the data. Our objective is to determine the intrinsic dimension of this linearized manifold representation of the sequence space.

LAE is an unsupervised neural network architecture commonly used for dimensionality reduction and feature extraction [69]. Unlike nonlinear autoencoders, LAEs rely solely on linear transformations in both the encoding and decoding stages. The network is trained to minimize the reconstruction error between the input sequence \mathbf{x} and its reconstruction $\hat{\mathbf{x}}$, defined as:

$$\min_{W,V} \|\mathbf{x} - VW\mathbf{x}\|^2 \tag{3.1}$$

where $W \in \mathbb{R}^{d \times L}$ is the encoding matrix, $V \in \mathbb{R}^{L \times d}$ is the decoding matrix, L

is the original dimension, and $d \ll L$ is the latent (intrinsic) dimension. By optimizing this reconstruction loss, the LAE identifies the minimal number of directions necessary to accurately represent the data, thereby providing an estimate of the manifold’s intrinsic dimensionality.

To better capture the local geometry of the sequence space—consistent with the concept of a curved NEM—we restricted the analysis to local clusters of sequences. Specifically, sequences from an MSA were grouped using a normalized Hamming distance threshold of 0.08, corresponding to approximately 15 amino acid substitutions. A separate linear autoencoder was trained within each cluster to evaluate reconstruction accuracy as a function of latent dimensionality.

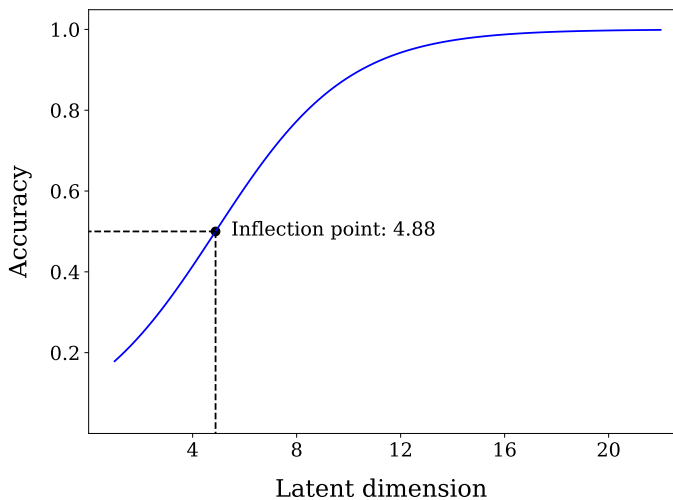


Figure 11: **Reconstruction Accuracy of a LAE Trained on Local Clusters of Protein Sequences.** Sequences were grouped using a normalized Hamming distance threshold of 0.08 (approximately 15 substitutions). The latent dimension d is varied from 1 to 30. The fitted sigmoidal curve shows an inflection point at $d = 4.88$, which is interpreted as the local ID of the NEM.

The latent dimension d was varied from 1 to 30. For each value of d , we trained the LAE on a training subset and evaluated its performance on a held-out test set. The resulting reconstruction accuracy curve is fitted with a sigmoidal function as shown in Figure 11.

The figure supports the conclusion that the reconstruction accuracy is higher

when the latent dimensions is low. The inflection point was estimated to be $d = 4.88$. We interpret this value as the intrinsic dimensionality of the sequence data for the following reason: it marks the transition from rapidly improving reconstruction (when each added dimension contributes significantly to variance explained) to a regime of diminishing returns. In differential geometry, the tangent space to a manifold is the best linear approximation at a point. Similarly, the inflection point in this curve corresponds to the dimension at which a linear subspace provides a locally sufficient representation of the underlying nonlinear structure—in this case, the NEM.

This result aligns with our findings from synthetic models in Chapter 2, where blockwise correlations induced low-dimensional sequence manifolds. Here, the same compression is observed in empirical data, suggesting that epistasis reduces the number of degrees of freedom required to describe functional sequence diversity.

3.2 Method II: Geodesic Graph-Based Estimation of Intrinsic Dimensionality

To validate and complement the linear autoencoder results, we adopt a second method that estimates intrinsic dimensionality based on geodesic distances derived from a neighborhood graph. This approach, developed by Granata and Carnevale [70], is particularly well-suited for analyzing high-dimensional biological data because it captures the global geometric structure of the underlying manifold.

The core idea of the method is to approximate geodesic distances between points on the data manifold using the shortest paths through a k -nearest neighbor (k-NN) graph. Unlike Euclidean distances, which may cut across the curved manifold and underestimate true separation, geodesic distances respect the intrinsic geometry of the data. The statistical distribution of these distances contains information about the dimensionality of the space in which the data is embedded.

Let $\mathcal{D} = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^L$ be the set of n protein sequences encoded as

binary vectors. We begin by constructing a k -nearest neighbor graph: each node corresponds to a sequence, and edges connect it to its k closest neighbors in normalized Hamming distance. The graph provides a local approximation of the manifold’s geometry. To compute global distances, we use Dijkstra’s algorithm to find the shortest path length between all pairs of nodes in the graph, which serves as a proxy for the geodesic distance.

To estimate the intrinsic dimension D , we compare the empirical distribution of these geodesic distances with the theoretical distribution of pairwise distances on the surface of a unit hypersphere in \mathbb{R}^D . Let $\theta \in [0, \pi]$ denote the angle between two points uniformly distributed on the surface of an $(D - 1)$ -dimensional unit hypersphere. The probability density function is given by:

$$f(\theta; D) = k \cdot \sin^{D-2}(\theta), \tag{3.2}$$

where the normalization constant k is:

$$k = \frac{\Gamma\left(\frac{D}{2}\right)}{\sqrt{\pi} \cdot \Gamma\left(\frac{D}{2} - \frac{1}{2}\right)}, \tag{3.3}$$

and $\Gamma(\cdot)$ denotes the Gamma function (See Appendix A for derivation).

For each value of D , we compute the root mean square deviation (RMSD) between the empirical geodesic distribution and the theoretical distribution $f_D(\theta)$. The dimension that minimizes the RMSD is taken as the estimated ID.

In Figure 12, we present the results of this analysis for the MSA dataset. The dataset was filtered at 80% sequence identity, and pairwise distances were computed using normalized Hamming distance.

The figure shows that the minimum RMSD occurs at $D = 5$, in close agreement with the inflection point obtained from the linear autoencoder analysis (Section 3.1). This convergence between two conceptually distinct methods—one local

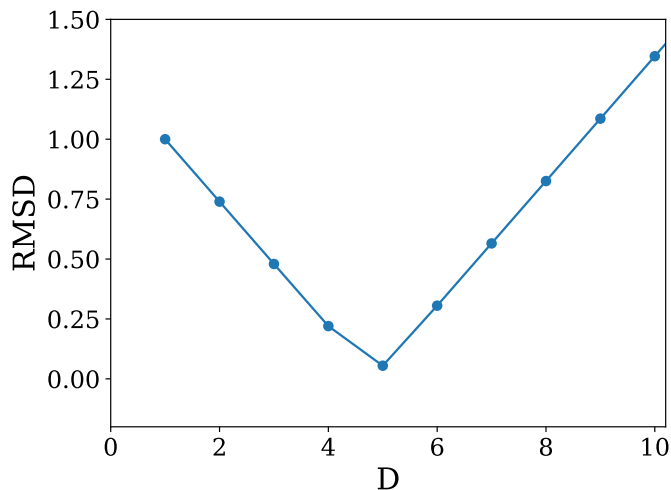


Figure 12: **RMSD Between the Empirical Distribution of Geodesic Distances and the Theoretical Distribution on a D -dimensional Hypersphere.** The minimum RMSD in this example occurs at dimension $D=5$. This result represents one of five independent repetitions, with an average estimated dimension of 4.7 across all runs.

and linear, the other global and graph-based—provides strong evidence that the accessible space of real protein sequences is effectively embedded in a low-dimensional manifold. Therefore, only a small number of collective variables govern the diversity of protein sequences. This supports the NEM hypothesis as a guiding framework for understanding the geometry of sequence space.

3.3 Method III: Intrinsic Dimensionality Estimation in Discrete Sequence Spaces

In this section, we present a statistically grounded method to estimate the intrinsic dimensionality of discrete datasets developed by Macocco et. al [71]. The method, known as **I3D** (Intrinsic Dimension for Discrete Datasets), is implemented in the DADapy Python library and is well-suited for data such as protein sequences, which reside in Hamming space [72].

Volume in Discrete Sequence Space Consider sequences of length L compared by Hamming distance. Under the assumption of an intrinsic dimension d , the number of points within Hamming radius t is given by Ehrhart theory:

$$V(t, d) = \binom{d+t}{d} {}_2F_1(-d, -t; -d-t; -1), \quad (3.4)$$

where ${}_2F_1$ is the Gauss hypergeometric function. Consider that the number of neighbors within radius t_1 is n_i and the number of neighbors within radius t_2 is k_i . so the shell volum of the nested radii $t_1 < t_2$ is:

$$x = \frac{V(t_1, d)}{V(t_2, d)}. \quad (3.5)$$

Modeling Neighbor Counts as a Binomial Process Given k_i neighbors in the outer shell, the count n_i in the inner shell is

$$P(n_i | k_i, d) = \binom{k_i}{n_i} x^{n_i} (1-x)^{k_i-n_i}. \quad (3.6)$$

Hence the joint likelihood over all N points is:

$$\mathcal{L}(d | \{n_i, k_i\}) = \prod_{i=1}^N \binom{k_i}{n_i} x^{n_i} (1-x)^{k_i-n_i}, \quad x = \frac{V(t_1, d)}{V(t_2, d)}. \quad (3.7)$$

Bayesian Inference and Posterior Distribution We choose a uniform prior on x :

$$x \sim \text{Beta}(\alpha_0 = 1, \beta_0 = 1) \quad (3.8)$$

Combining with the binomial likelihood yields the posterior:

$$x | \{n_i, k_i\} \sim \text{Beta}\left(\alpha = 1 + \sum_{i=1}^N n_i, \beta = 1 + \sum_{i=1}^N (k_i - n_i)\right) \quad (3.9)$$

Since x and d are related by $x = p(d) \equiv V(t_1, d)/V(t_2, d)$, the posterior density on d is obtained by:

$$P(d) = \text{Beta}(p(d); \alpha, \beta) \times |p'(d)|, \quad (3.10)$$

where $|p'(d)| = \left| \frac{d}{dd} p(d) \right|$ is the Jacobian. Scanning d over its domain gives the full posterior $P(d)$, whose mean and variance yield the Bayesian estimate and uncertainty of the intrinsic dimension.

Practical Implementation The I3D method proceeds as follows:

1. Encode the sequences numerically (one-hot encoding).
2. Compute all pairwise Hamming distances.
3. For each point, count the number of neighbors n_i and k_i within radii t_1 and t_2 .
4. Fit the intrinsic dimension d using Bayesian inference.

Applying this methodology to our protein dataset (filtered at 80% sequence identity, as in previous analyses), we compute the posterior distribution over the intrinsic dimension D , using the discrete-metric estimator introduced by Macocco et al. [71] The resulting posterior curve is shown in Figure 13, and the maximum a posteriori (MAP) estimate is extracted as our inferred dimension.

As shown in Figure 13, the posterior is sharply peaked near $D = 5$, indicating a strong and unambiguous estimate of intrinsic dimensionality. This finding is consistent with our previous estimates from the linear autoencoder (Section 3.1) and the geodesic graph-based method (Section 3.2), all of which independently converge near $D \approx 5$.

Together, these complementary approaches—linear projection, graph-based geometry, and discrete statistical modeling—lead to the same conclusion: natural protein sequences are not scattered uniformly throughout the high-dimensional space,

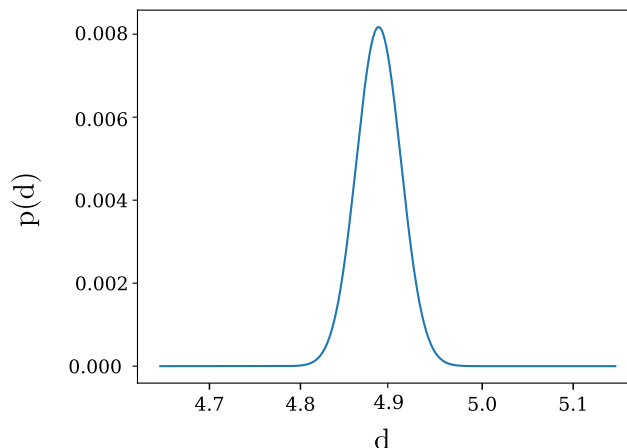


Figure 13: **Posterior Distribution over Dimension** Posterior distribution $P(d \mid \{n_i, k_i\})$ over dimension d , computed for the protein sequence dataset using the discrete-metric estimator of Macocco et al. This narrow peak reflects a strong, statistically supported estimate of the dataset’s intrinsic dimension.

but are instead confined to a low-dimensional NEM, with an effective dimension of approximately $D \approx 5$.

3.4 Epistasis as the Source of Dimensional Compression

Having established in Sections 3.1–3.3 that the NEM for protein sequences resides in a low-dimensional subspace, we now examine the hypothesis that epistasis is a primary driver of this dimensional reduction. Epistasis refers to statistical dependencies between residues, where the presence of a specific amino acid at one site influences the distribution of residues at another site. These interdependencies reduce the number of effectively independent positions in a sequence and thereby constrain the accessible configuration space. To understand how site correlations contribute to dimensional compression, we begin with a schematic example shown in Figure 14. In the left panel, we highlight two columns of an MSA, positions i and j , which show a strong pattern of covariation. For example, when Arginine (R) at position i is replaced by Lysine (K), Aspartic acid (D) at position j changes to Glutamic acid

(E); when Lysine mutates to Tryptophan (W), Glutamic acid shifts to Valine (V). Correlated amino acid substitutions often reflect structural or functional constraints in the protein’s 3D fold. Conservative changes like Arginine to Lysine and Aspartic acid to Glutamic acid preserve charge and bonding capacity, maintaining local electrostatic interactions. In contrast, coordinated non-conservative changes—such as Lysine to Tryptophan and Glutamic acid to Valine—suggest compensatory mutations that preserve hydrophobic packing or structural stability. These patterns indicate that epistasis arises from the need to maintain compatibility within the folded structure.

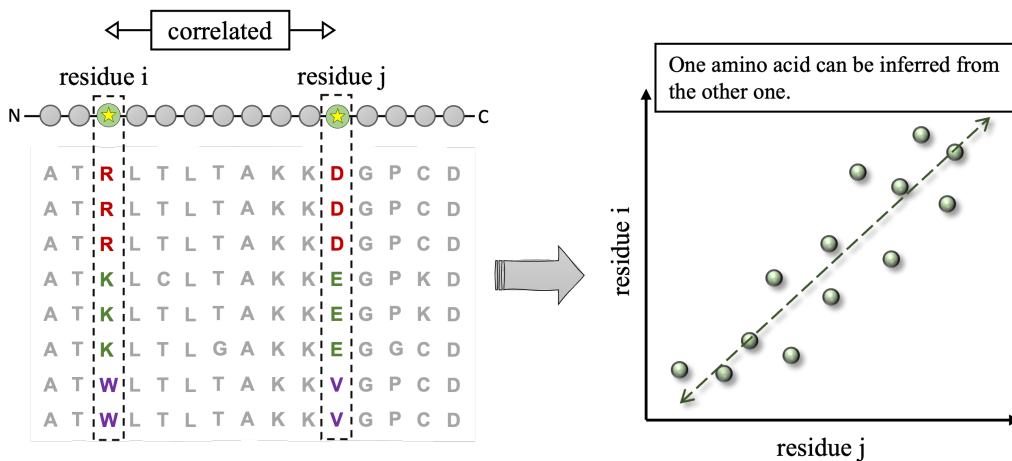


Figure 14: **Epistatic Couplings.** **Left:** Example from an MSA highlighting two positions, i and j , that undergo coordinated substitutions. Amino acid changes such as $R \rightarrow K$ with $D \rightarrow E$, or $K \rightarrow W$ with $E \rightarrow V$, illustrate joint constraints imposed by epistasis. **Right:** Scatter plot of residue identities at positions j (x-axis) and i (y-axis) across sequences. The linear relationship indicates that the state of one site predicts the other.

This relationship is formalized by the covariance matrix C , which quantifies how amino acid identities at two positions co-vary across an alignment. Let $f_i(a)$ be the marginal frequency of amino acid a at site i , and $f_{ij}(a, b)$ the joint frequency of observing a at site i and b at site j . Then the covariance is:

$$C_{ij}(a, b) = f_{ij}(a, b) - f_i(a)f_j(b).$$

If sites are independent, then $f_{ij}(a, b) = f_i(a)f_j(b)$, and $C_{ij}(a, b) = 0$. Epistasis manifests as nonzero covariances—off-diagonal structure in the matrix C .

3.4.1 Column Shuffling as a Tool to Disrupt Epistasis

To test whether such correlations are responsible for the observed low dimensionality of protein sequence space, we perturb epistasis directly. This is done by shuffling entire columns of the MSA independently, one at a time. Shuffling destroys inter-site correlations while preserving intra-column information, such as conservation.

Figure 15 illustrates this procedure. In the original alignment (left), positions i and j contain coordinated amino acid substitutions. After shuffling column j (right), the marginal distribution of amino acids at j is unchanged, but the joint distribution $f_{ij}(a, b)$ is randomized, effectively eliminating correlations:

$$f_{ij}^{\text{shuffled}}(a, b) \approx f_i(a)f_j(b).$$

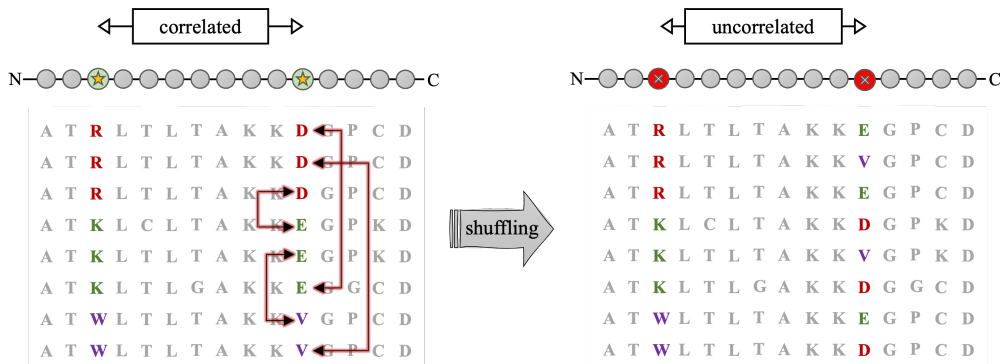


Figure 15: **Epistasis Modulation via Shuffling MSA.** **Left:** correlated positions i and j in the original MSA. **Right:** column j is shuffled across sequences, preserving its amino acid composition (i.e., marginal frequencies) but eliminating joint correlations with column i .

3.4.2 Quantifying the Effect: Intrinsic Dimensionality vs. Epistasis Strength

To quantify how epistasis influences the dimensionality of the sequence space, we incrementally shuffle an increasing number of columns and compute the intrinsic dimensionality for each modified MSA using the geodesic graph and discrete metric-based estimators. We also compute the strength of inter-site coupling using the Frobenius norm of the covariance matrix:

$$\|C\|_F = \sqrt{\sum_{i,j} \sum_{a,b} C_{ij}(a,b)^2}.$$

This norm aggregates the total magnitude of covariation in the alignment. As more columns are shuffled, we expect $\|C\|_F$ to decrease, reflecting reduced epistatic constraint.

Figure 16 presents these results. Panel a shows that the ID increases as more columns are shuffled. When only a few columns are shuffled, correlations remain intact and dimensionality stays low (around five), whereas extensive shuffling disrupts correlations and drives the dimension toward higher values. The colorbar quantifies epistatic strength as $1 - \text{shuffled columns}/L$, providing a numerical measure of how much correlation is preserved at each level of shuffling. Panel b shows the same trend in terms of the Frobenius norm of the covariance matrix: datasets with stronger coupling (higher $\|C\|_F$) have lower dimensionality, while those with weaker coupling exhibit higher values. Together, Panels a and b demonstrate an inverse relationship between coupling strength and dimensionality, showing that epistasis effectively reduces the number of accessible directions in sequence space.

To confirm that these trends are not artifacts of shared evolutionary ancestry (phylogeny), we use variational autoencoders (VAEs) to generate synthetic MSAs with different levels of epistasis. VAEs consist of an encoder–decoder architecture

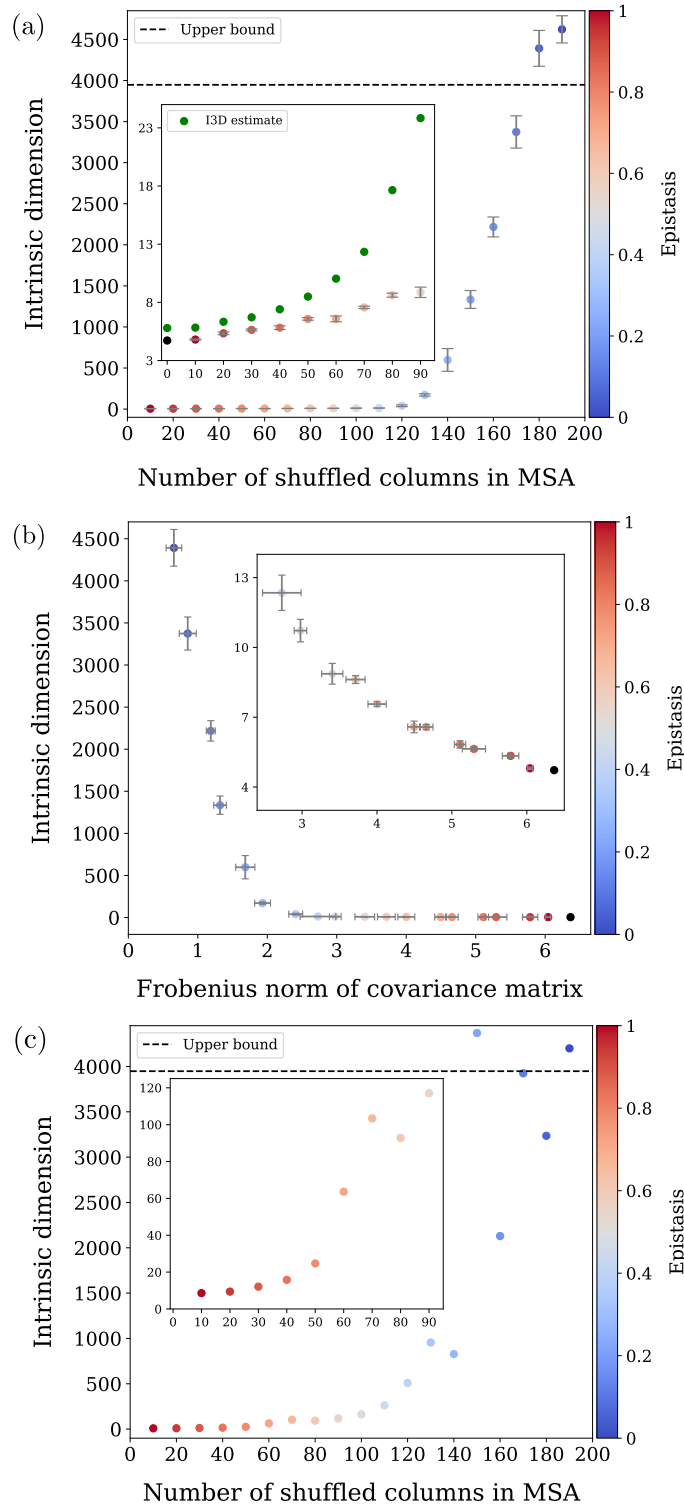


Figure 16: **Epistasis–dimensionality Relationship.** (a) Estimated ID vs. number of shuffled columns in the MSA. I3D estimates are shown as green dots; the inset highlights the low-shuffling regime. The dashed line indicates the theoretical maximum dimension, set at $D_{\max} = q \times L$. (b) ID vs. Frobenius norm of the covariance matrix. Higher coupling strength corresponds to lower dimension. (c) Same analysis as (a) with sequences generated by VAEs trained on MSAs with different levels of shuffling.

trained to reconstruct sequences while compressing them into a latent space. The VAE optimizes a loss function that balances reconstruction accuracy and prior regularization:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \text{KL}[q(z|x) \parallel p(z)],$$

where $q(z|x)$ is the encoder, $p(x|z)$ is the decoder, and $p(z)$ is a Gaussian prior. By training VAEs on MSAs with increasing numbers of shuffled columns, we control the amount of epistasis present in the training data.

Panel c shows the ID of VAE-generated sequences as a function of the number of shuffled columns in the training set. Even though these sequences lack any phylogenetic relatedness, they reproduce the same pattern: more epistasis leads to lower dimensionality.

These results collectively support the interpretation that epistatic interactions—not phylogeny—are responsible for the low ID of protein sequence space. By restricting mutational freedom through inter-site constraints, epistasis compresses the NEM into a structured, low-dimensional manifold. Having established this connection between sequence-level correlations and the structure of the NEM, we now turn to investigating how these geometric constraints influence phylogenetic inference and the resulting tree topologies.

3.5 Epistasis and the Topology of Phylogenetic Trees

Having established that the dimensionality of the NEM is tightly linked to epistasis, we now ask whether this dimensional constraint influences the structure of phylogenetic trees inferred from protein sequences. Phylogenetic trees, which encode evolutionary relationships, often exhibit either hierarchical or star-like structures depending on the distribution of pairwise distances between sequences. A broad distance distribution tends to yield trees with diverse branch lengths and nested clades, while

a narrow distribution leads to star-like topologies where all tips are approximately equidistant from the root.

In Section 3.4, we demonstrated that high epistasis results in lower intrinsic dimensionality. Here, we test whether this reduced dimensionality correlates with increased tree hierarchy. We analyze trees reconstructed from both natural and synthetic MSAs with varying levels of epistasis using two metrics: the number of cherries and a log-likelihood ratio test (LLRT).

3.5.1 Tree Balance and the Number of Cherries

The number of cherries serves as a useful proxy for assessing tree balance: star-like topologies typically contain many cherries, while more hierarchical trees contain fewer. To examine how epistasis influences tree structure, we reconstruct phylogenies using the Neighbor-Joining (NJ) method, a widely used distance-based algorithm that builds trees by iteratively joining pairs of taxa to minimize total branch length [73].

We applied this method across a spectrum of datasets varying in their level of epistatic signal: from natural protein MSAs (unshuffled), to sequences generated by VAEs trained on these MSAs, to VAE-generated sequences with increasing levels of column-wise shuffling, and finally to sequences evolved under an independent site model. A clear monotonic trend emerges (Figure 17): the number of cherries increases as epistatic constraints are progressively disrupted. Natural sequences yield the most hierarchical trees, with the fewest cherries, followed by unshuffled VAE-generated sequences. As more columns are shuffled, the cherry count steadily rises, resulting in the highest values for trees inferred from independent-site simulations.

This continuous transition supports the interpretation that epistasis restricts accessible sequence configurations to a low-dimensional NEM, leading to hierarchical phylogenies even under neutral evolution.

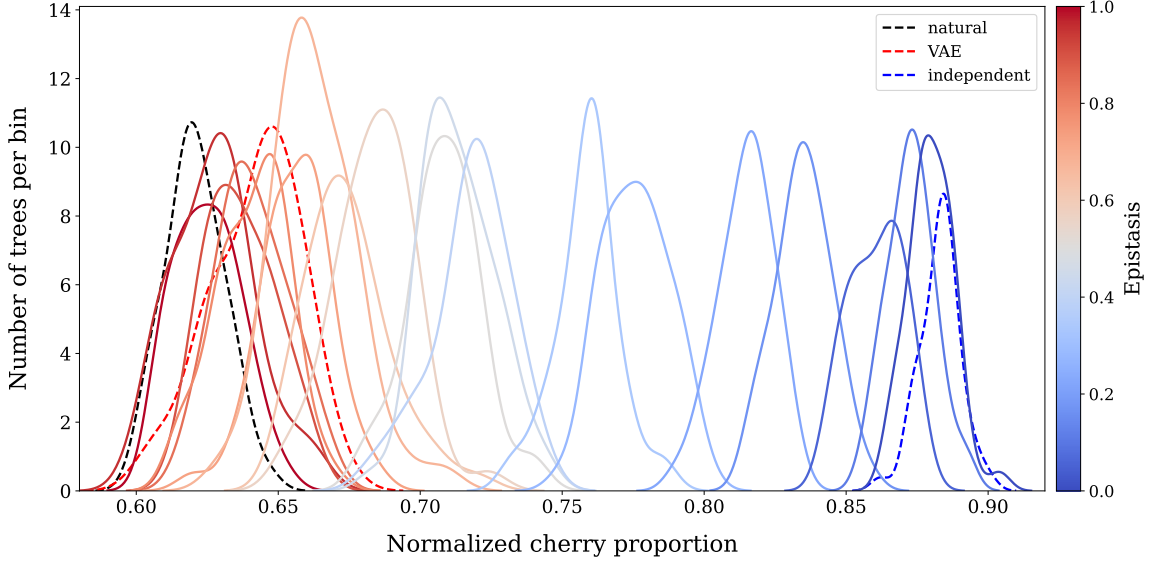


Figure 17: **Effect of Epistasis on Number of Cherries.** Each point corresponds to a tree reconstructed from a dataset with a specific level of shuffling. Red points correspond to unshuffled or highly epistatic data; blue points correspond to heavily shuffled or independently evolving datasets. The trend confirms that increasing epistasis reduces cherry count, supporting the emergence of hierarchy in tree shape.

3.5.2 Likelihood Ratio Test Against the Star Model

To quantify this effect further, we apply a Log-Likelihood Ratio Test (LLRT) comparing the fit of each inferred tree to a null star phylogeny [74, 75]. The test evaluates whether the observed pairwise distances significantly deviate from the expectations under a star model. Formally, the LLRT statistic is given by:

$$\Delta(t_1, t_0 | x) = 2 [\ln \mathcal{L}(t_1 | x) - \ln \mathcal{L}(t_0 | x)] \quad (3.11)$$

where x is the observed pairwise distance data, t_1 is the inferred tree (alternative hypothesis), t_0 is the star tree (null hypothesis), and $\mathcal{L}(t | x)$ is the likelihood of the data under tree t . This formulation compares how well each tree explains the observed data. A large Δ indicates that the tree t_1 fits the data significantly better than the null star tree.

As shown in Figure 18, datasets with higher epistasis (and thus lower intrinsic

dimension) yield strong rejection of the null model. As epistasis diminishes through shuffling, the LLRT statistic drops, approaching insignificance—indicating that the inferred trees become increasingly compatible with the star topology.

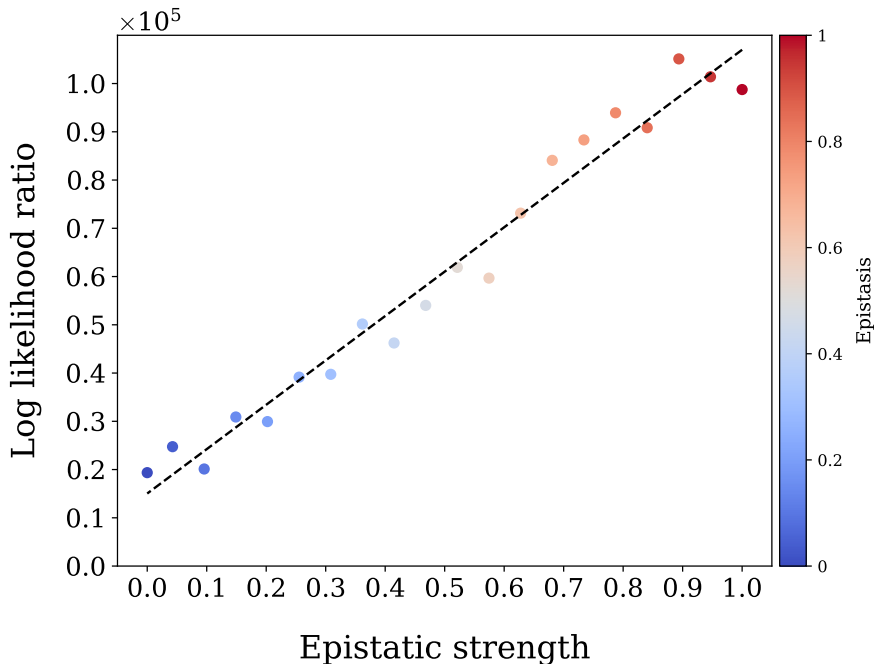


Figure 18: **Log-Likelihood Ratio Test (LLRT) as a Function of Epistasis Strength.** The y -axis shows the computed Δ for each dataset. Datasets with strong epistatic structure deviate significantly from the star tree model, while weakly epistatic datasets align more closely with the star model. This demonstrates that epistasis not only shapes the geometry of sequence space but also alters the topology of the inferred phylogeny.

Together, these findings confirm that the shape of phylogenetic trees is not solely a consequence of shared ancestry but also reflects the geometric constraints imposed by epistasis. Reduced dimensionality of the NEM—driven by correlated evolution—leads to more hierarchical tree structures. Ignoring this effect could result in overinterpreting hierarchical signal as evidence of evolutionary divergence rather than structural constraints. Accounting for NEM geometry thus offers a principled way to improve phylogenetic inference.

CHAPTER 4

DISCUSSION

4.1 Summary of Findings

This thesis set out to investigate how epistatic interactions influence the reconstruction of phylogenetic trees, with particular emphasis on the geometry of sequence space and the dimensionality of accessible sequence ensembles. By combining theoretical modeling, numerical simulation, and empirical protein data, we established a framework that connects epistasis to distortions in inferred tree topologies. The main result is that correlated mutations compress the effective dimensionality of sequence space into what we defined as the NEM, and that this compression alters the distribution of distances in ways that bias phylogenetic inference.

Across synthetic and real systems, the findings converge on the same conclusion: even modest levels of epistasis produce measurable reductions in intrinsic dimensionality, which in turn generate tree-like structures with strong hierarchy. These structures can emerge independently of any genuine branching history, underscoring the risk of attributing topological patterns to evolutionary divergence. This manifold thus provides a unifying concept for interpreting how genetic dependencies reshape sequence landscapes and their phylogenetic representations.

4.2 Interpretation of Theoretical Models

The minimalist models presented in Chapter 2 demonstrated that even simple blockwise correlations, when embedded into an Ising-like framework, are sufficient to reconfigure the geometry of sequence space. Independent-site models produce ensembles that approximate a uniform distribution on a high-dimensional hypersphere, consistent with star-like tree topologies. By contrast, blockwise correlated models col-

lapse the space into lower-dimensional manifolds, narrowing the range of accessible directions and amplifying apparent hierarchy in reconstructed trees.

These results highlight a purely geometric mechanism by which phylogenetic inference can be misled. In distance-based approaches, a wide spread of pairwise distances is interpreted as evidence of nested branching. Yet in our simulations, this structure emerged only from correlations constraining mutational freedom. The implication is clear: dimensionality itself becomes a hidden parameter in phylogenetic inference, and ignoring it can lead to overestimation of historical structure.

4.3 Evidence from Natural Protein Sequences

Analyses of empirical sequence alignments confirmed that the dimensional compression observed in minimal models persists in real proteins. Three methods—linear autoencoders, geodesic graph-based estimation, and discrete metric-based Bayesian inference—consistently identified intrinsic dimensionalities near five, far lower than the length of the sequences. This finding reinforces the notion that functional and structural constraints confine protein diversity into restricted regions of sequence space.

Importantly, shuffling experiments demonstrated that this low dimensionality is a direct consequence of inter-site correlations. Disrupting epistasis by permuting alignment columns led to increases in estimated dimensionality, and VAE-generated sequences trained on shuffled data reproduced the same pattern in the absence of shared ancestry. Together, these results provide strong evidence that epistasis, rather than phylogeny, is the primary driver of dimensional compression in protein sequence space.

4.4 Consequences for Phylogenetic Inference

The impact of epistasis extends beyond abstract geometry to the practical interpretation of phylogenetic trees. Hierarchical topologies, delayed diversification in lineage-through-time curves, and reduced cherry counts all appeared in simulations and empirical analyses under high correlation. Likelihood ratio tests further confirmed that correlated datasets reject star-like models, favoring more structured alternatives even when no branching history is present.

These results highlight the risk of assuming that tree balance or hierarchy must reflect biological processes. Instead, they may also encode the geometric imprint of epistatic constraints on sequence space. In this sense, phylogenetic inference becomes sensitive not only to historical processes but also to the statistical dependencies governing mutational accessibility. Recognizing and correcting for this influence is essential if trees are to be interpreted as records of ancestry rather than artifacts of correlation.

4.5 Broader Implications

Beyond their technical implications, these findings contribute to a broader rethinking of how molecular evolution is modeled. Traditional substitution models treat sites as independent, sacrificing realism for tractability. The evidence presented here shows that such simplifications underestimate the role of epistasis, which acts as a pervasive organizing force in sequence space. Accounting for this structure opens the door to dimensionality-aware models that better capture both the geometry and the dynamics of protein evolution.

Moreover, the NEM framework offers a conceptual bridge between molecular evolution and disciplines such as statistical physics and machine learning. Generative models, whether Potts-based or neural-network driven, provide practical tools for ex-

ploring constrained sequence spaces and testing how geometric properties translate into tree topology. By framing epistasis as a geometric constraint rather than a complication, this thesis lays the groundwork for unifying diverse approaches to sequence analysis under a common perspective.

4.6 Limitations

Several limitations of the present work should be acknowledged. The minimalist models employed binary alphabets and blockwise interactions, choices made for analytical tractability rather than biological realism. While these simplifications revealed essential principles, they cannot capture the full diversity of epistatic patterns in natural proteins. Similarly, intrinsic dimensionality estimators, though consistent across methods, rely on assumptions about uniform sampling and metric embeddings that may not hold universally.

In addition, while empirical datasets were carefully controlled with shuffling and generative modeling, they remain influenced by evolutionary processes not explicitly modeled here, such as selection, recombination. These factors may interact with epistasis in complex ways that were not disentangled in this study. Thus, while the results strongly support the role of epistasis in dimensional compression, they should be viewed as a foundation for more comprehensive models rather than a complete description of molecular evolution.

4.7 Future Directions

Future research can build on this framework in several directions. On the methodological side, developing phylogenetic inference algorithms that explicitly incorporate dimensionality estimates would allow tree reconstruction to account for the NEM. Extending models to capture higher-order and heterogeneous epistatic in-

teractions could also reveal how complex dependency structures further compress accessible sequence space.

On the empirical side, applying these methods across diverse protein families and integrating them with experimental fitness landscapes will be crucial for validating the generality of the findings. Advances in generative modeling, including diffusion frameworks and transformer-based architectures, provide new opportunities to probe the relationship between sequence space geometry and evolutionary inference. Ultimately, the integration of geometric, statistical, and biological perspectives may yield a new generation of evolutionary models that are both computationally feasible and biologically faithful.

4.8 Conclusions

This thesis has shown that epistasis is not merely a complicating factor but a fundamental determinant of sequence space geometry and phylogenetic inference. By compressing sequence ensembles into low-dimensional manifolds, epistatic interactions generate tree structures that can mimic historical branching even under neutral evolution. Recognizing this effect shifts our interpretation of phylogenetic trees, highlighting that hierarchy does not arise from ancestry alone.

The introduction of the NEM provides a conceptual and analytical framework for capturing these effects. By bridging theoretical models, simulations, and real protein data, this work highlights the necessity of dimensionality-aware approaches to molecular evolution. This work provides a foundation for future studies that seek to reconstruct history and understand the geometric and statistical forces that shape the landscape of life.

BIBLIOGRAPHY

- [1] Michael Lynch and John S Conery. The evolutionary fate and consequences of duplicate genes. *science*, 290(5494):1151–1155, 2000.
- [2] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39(1):309–338, 2005.
- [3] Walter M Fitch. Homology: a personal view on some of the problems. *Trends in genetics*, 16(5):227–231, 2000.
- [4] Wolfgang Enard, Molly Przeworski, Simon E Fisher, Cecilia SL Lai, Victor Wiebe, Takashi Kitano, Anthony P Monaco, and Svante Pääbo. Molecular evolution of *foxp2*, a gene involved in speech and language. *Nature*, 418(6900):869–872, 2002.
- [5] Ross Hardison. Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *Journal of Experimental Biology*, 201(8):1099–1117, 1998.
- [6] Jonathan A Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome research*, 8(3):163–167, 1998.
- [7] Gavin C Conant and Kenneth H Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, 2008.
- [8] Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 2013.
- [9] Allan Force, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.
- [10] Matthew D Rasmussen and Manolis Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research*, 22(4):755–765, 2012.
- [11] Pekka Pamilo and Masatoshi Nei. Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583, 1988.
- [12] Yoshihito Niimura and Masatoshi Nei. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PloS one*, 2(8):e708, 2007.
- [13] Olena Meleshko, Michael D Martin, Thorfinn Sand Korneliussen, Christian Schröck, Paul Lamkowski, Jeremy Schmutz, Adam Healey, Bryan T Piatkowski,

- A Jonathan Shaw, David J Weston, et al. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Molecular biology and evolution*, 38(7):2750–2766, 2021.
- [14] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, 2015.
- [15] Tao Sang and Yang Zhong. Testing hybridization hypotheses based on incongruent gene trees. *Systematic Biology*, 49(3):422–434, 2000.
- [16] Dan Vanderpool, Bui Quang Minh, Robert Lanfear, Daniel Hughes, Shwetha Murali, R Alan Harris, Muthuswamy Raveendran, Donna M Muzny, Mark S Hibbins, Robert J Williamson, et al. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. *PLoS biology*, 18(12):e3000954, 2020.
- [17] Hayley C Lanier and L Lacey Knowles. Is recombination a problem for species-tree analyses? *Systematic Biology*, 61(4):691–701, 2012.
- [18] Timothy B Sackton and Nathan Clark. Convergent evolution in the genomics era: new insights and directions, 2019.
- [19] Ying Li, Zhen Liu, Peng Shi, and Jianzhi Zhang. The hearing gene prestin unites echolocating bats and whales. *Current Biology*, 20(2):R55–R56, 2010.
- [20] Jeffrey C Oliver. Augist: inferring species trees while accommodating gene tree uncertainty. *Bioinformatics*, 24(24):2932–2933, 2008.
- [21] Benoit Morel, Paul Schade, Sarah Lutteropp, Tom A Williams, Gergely J Szöllősi, and Alexandros Stamatakis. Speciesrax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *Molecular biology and evolution*, 39(2):msab365, 2022.
- [22] Mozes PK Blom, Jason G Bragg, Sally Potter, and Craig Moritz. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of australian lizards. *Systematic Biology*, 66(3):352–366, 2017.
- [23] Nick Goldman and Ziheng Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5):725–736, 1994.
- [24] Jesse D Bloom. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular biology and evolution*, 31(8):1956–1978,

2014.

- [25] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.
- [26] Chris A Nasrallah, David H Mathews, and John P Huelsenbeck. Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Systematic Biology*, 60(1):60–73, 2011.
- [27] Andrew F Magee, Sarah K Hilton, and William S DeWitt. Robustness of phylogenetic inference to model misspecification caused by pairwise epistasis. *Molecular biology and evolution*, 38(10):4603–4615, 2021.
- [28] Francisco McGee, Sandro Hauri, Quentin Novinger, Slobodan Vucetic, Ronald M Levy, Vincenzo Carnevale, and Allan Haldane. The generative capacity of probabilistic protein sequence models. *Nature communications*, 12(1):6302, 2021.
- [29] Xinqiang Ding, Zhengting Zou, and Charles L Brooks III. Deciphering protein evolution and fitness landscapes with latent space models. *Nature communications*, 10(1):5644, 2019.
- [30] Cheyenne Ziegler, Jonathan Martin, Claude Sinner, and Faruck Morcos. Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications*, 14(1):2222, 2023.
- [31] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.
- [32] Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. *arXiv preprint arXiv:1712.03346*, 2017.
- [33] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- [34] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.
- [35] John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable gen-

- erative model. *Nature*, 623(7989):1070–1078, 2023.
- [36] Ronald M Levy, Allan Haldane, and William F Flynn. Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current opinion in structural biology*, 43:55–62, 2017.
- [37] Jeanne Trinquier. *Data-driven generative modeling of protein sequence landscapes and beyond*. PhD thesis, Sorbonne Université, 2023.
- [38] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holtom, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Deep neural language modeling enables functional protein generation across families. *BioRxiv*, pages 2021–07, 2021.
- [39] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature communications*, 12(1):5800, 2021.
- [40] Sergio Romero-Romero, Sebastian Lindner, and Noelia Ferruz. Exploring the protein sequence space with global generative models. *Cold Spring Harbor Perspectives in Biology*, 15(11):a041471, 2023.
- [41] Adam Winnifrieth, Carlos Outeiral, and Brian L. Hie. Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology*, 86:102794, 2024.
- [42] Mehrsa Mardikoraem, Zirui Wang, Nathaniel Pascual, and Daniel Woldring. Generative models for protein sequence modeling: recent advances and future directions. *Briefings in Bioinformatics*, 24(6):bbad358, 2023.
- [43] Chloe Hsu, Clara Fannjiang, and Jennifer Listgarten. Generative models for protein structures and sequences. *nature biotechnology*, 42(2):196–199, 2024.
- [44] Timothy Truong Jr and Tristan Bepler. Poet: A generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems*, 36:77379–77415, 2023.
- [45] Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.
- [46] Hassan W Kayondo, Alfred Ssekagiri, Grace Nabakooza, Nicholas Bbosa, Degratius Ssemwanga, Pontiano Kaleebu, Samuel Mwalili, John M Mango, Andrew J Leigh Brown, Roberto A Saenz, et al. Employing phylogenetic tree shape statistics to resolve the underlying host population structure. *BMC bioinformat-*

- ics*, 22:1–20, 2021.
- [47] Russell F Doolittle. Similar amino acid sequences: chance or common ancestry? *Science*, 214(4517):149–159, 1981.
- [48] Dmitry A Kondrashov and Fyodor A Kondrashov. Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*, 31(1):24–33, 2015.
- [49] Kristina Crona, Devin Greene, and Miriam Barlow. The peaks and geometry of fitness landscapes. *Journal of theoretical biology*, 317:1–10, 2013.
- [50] Frank J Poelwijk, Sorin Tănase-Nicola, Daniel J Kiviet, and Sander J Tans. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of theoretical biology*, 272(1):141–144, 2011.
- [51] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge university press, 1985.
- [52] Inês Fragata, Alexandre Blanckaert, Marco António Dias Louro, David A Liberles, and Claudia Bank. Evolution in the light of fitness landscape theory. *Trends in ecology & evolution*, 34(1):69–82, 2019.
- [53] Sergey Gavrilets and Janko Gravner. Percolation on the fitness hypercube and the evolution of reproductive isolation. *Journal of theoretical biology*, 184(1):51–64, 1997.
- [54] Andries E Brouwer and Willem H Haemers. *Spectra of graphs*. Springer Science & Business Media, 2011.
- [55] Philippe Huneman. Neutral spaces and topological explanations in evolutionary biology: lessons from some landscapes and mappings. *Philosophy of Science*, 85(5):969–983, 2018.
- [56] Tod F Stuessy and Christiane König. Patrocladistic classification. *Taxon*, 57(2):594–601, 2008.
- [57] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.
- [58] L Dib and A Carbone. Co-evolution of blocks of residues and sectors in protein structures. *Journées Ouvertes en Biologie Informatique Mathématiques*, page 153.
- [59] Stephen Brooks. Markov chain monte carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, 47(1):69–100, 1998.

- [60] Roy J Glauber. Time-dependent statistics of the ising model. *Journal of mathematical physics*, 4(2):294–307, 1963.
- [61] Panagiotis Sidiropoulos. N-sphere chord length distribution. *arXiv preprint arXiv:1411.5639*, 2014.
- [62] Assaf Naor. Concentration of measure, 2008.
- [63] Madeleine B Thompson. A comparison of methods for computing autocorrelation time. *arXiv preprint arXiv:1011.0175*, 2010.
- [64] Mark Kac. On the notion of recurrence in discrete stochastic processes. 1947.
- [65] SOKAL RR. A statistical method for evaluating systematic relationships. *Univ Kans sci bull*, 38:1409–1438, 1958.
- [66] Lawrence K Saul and Sam T Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research*, 4(Jun):119–155, 2003.
- [67] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.
- [68] Lawrence Cayton et al. *Algorithms for manifold learning*. eScholarship, University of California, 2008.
- [69] Elad Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*, 2018.
- [70] Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific reports*, 6(1):31377, 2016.
- [71] Iuri Macocco, Aldo Glielmo, Jacopo Grilli, and Alessandro Laio. Intrinsic dimension estimation for discrete metrics. *Physical Review Letters*, 130(6):067401, 2023.
- [72] Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Dadapy: Distance-based analysis of data-manifolds in python. *Patterns*, 3(10), 2022.
- [73] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

- [74] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.
- [75] B de Finetti. Edwards awf likelihood. 1972.
- [76] Tony Cai, Jianqing Fan, and Tiefeng Jiang. Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research*, 14(1):1837–1864, 2013.
- [77] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

APPENDICES

APPENDIX A

DERIVATION OF THE DISTRIBUTION OF GEODESIC DISTANCES ON A $(D-1)$ -SPHERE

In this section, our goal is to derive the probability distribution of geodesic distances between points on the $(D - 1)$ -dimensional unit hypersphere $S^{D-1} \subset \mathbb{R}^D$. The geodesic distance between two points on the sphere is defined as the length of the shortest arc connecting them along the surface. On a unit sphere, this geodesic distance is equal to the angle θ between the two vectors originating from the origin to each point. Therefore, deriving the distribution of geodesic distances is equivalent to deriving the distribution of angles between two independent and uniformly sampled vectors on S^{D-1} [76].

$$\text{Geodesic distance} = \theta = \cos^{-1}(\mathbf{x} \cdot \mathbf{y}). \quad (\text{A.1})$$

Uniform Sampling and Rotational Symmetry Let $\mathbf{x}, \mathbf{y} \in S^{D-1}$ be two independent, uniformly distributed unit vectors. The angle θ between them is defined via the dot product:

$$\cos \theta = \mathbf{x} \cdot \mathbf{y}, \quad \theta \in [0, \pi]. \quad (\text{A.2})$$

Due to the rotational symmetry of the sphere, we can fix one vector without loss of generality. Let us set

$$\mathbf{x} = (1, 0, 0, \dots, 0), \quad (\text{A.3})$$

and sample \mathbf{y} uniformly from S^{D-1} . The angle θ is then determined entirely by the first coordinate of \mathbf{y} :

$$\cos \theta = y_1 \quad \Rightarrow \quad \theta = \cos^{-1}(y_1). \quad (\text{A.4})$$

Therefore, the distribution of θ can be derived from the marginal distribution

of the first coordinate y_1 of a uniform vector on the sphere.

Hyperspherical Coordinates and the Surface Measure To describe points on the surface of the sphere, we introduce hyperspherical coordinates, which generalize the familiar polar and spherical coordinate systems from 2D and 3D to D dimensions. A point $\mathbf{y} \in S^{D-1}$ can be represented as:

$$\begin{aligned}
 y_1 &= \cos \theta_1, \\
 y_2 &= \sin \theta_1 \cos \theta_2, \\
 y_3 &= \sin \theta_1 \sin \theta_2 \cos \theta_3, \\
 &\vdots \\
 y_{D-1} &= \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{D-2} \cos \phi, \\
 y_D &= \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{D-2} \sin \phi,
 \end{aligned} \tag{A.5}$$

where the angular variables vary over:

$$\theta_1, \dots, \theta_{D-2} \in [0, \pi], \quad \phi \in [0, 2\pi]. \tag{A.6}$$

This coordinate system ensures that all points lie on the unit sphere. Importantly, the angle θ we care about corresponds exactly to θ_1 , since $\cos \theta = y_1 = \cos \theta_1$.

The infinitesimal surface area element in hyperspherical coordinates is determined by the Jacobian of the transformation from Cartesian to angular variables. This Jacobian reflects how the coordinates are arranged on the sphere. In these coordinates, each point on the sphere is described by a set of angles. Here, we change variables from (y_1, y_2, \dots, y_D) to $(\theta_1, \theta_2, \dots, \theta_{D-2}, \phi)$ which leads to:

$$dS = \left(\prod_{j=1}^{D-2} \sin^{D-1-j} \theta_j \right) d\theta_1 \cdots d\theta_{D-2} d\phi. \tag{A.7}$$

In particular, the first angle θ_1 — which corresponds to the angle between a sampled point and the fixed reference axis — appears in the form $\sin^{D-2} \theta_1$. This factor reflects the surface area of the $(D - 2)$ -dimensional subsphere (a "band" or "slice") at angular height θ_1 , and it determines the relative number of points on the sphere at that angle. Therefore, the marginal distribution of the angle $\theta = \theta_1$ between two random vectors is directly proportional to:

$$\sin^{D-2} \theta. \tag{A.8}$$

Deriving the Angle Distribution Function $F(\theta)$ We now derive the density function $F(\theta)$ for the angle θ between two independent random vectors on S^{D-1} . As discussed, the number of such pairs forming an angle in the infinitesimal range $[\theta, \theta + d\theta]$ is proportional to the surface area at that angle, which scales like:

$$F(\theta) \propto \sin^{D-2} \theta. \tag{A.9}$$

To turn this into a proper probability density function, we normalize it over the interval $[0, \pi]$. That is, we define:

$$F(\theta) = \frac{1}{Z_D} \sin^{D-2} \theta, \tag{A.10}$$

where the normalizing constant Z_D is:

$$Z_D = \int_0^\pi \sin^{D-2} \theta \, d\theta. \tag{A.11}$$

This integral has a known closed-form expression involving the Gamma function:

$$Z_D = \sqrt{\pi} \cdot \frac{\Gamma\left(\frac{D-1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)}. \tag{A.12}$$

Thus, the final expression for the probability density function of the angle θ is:

$$F(\theta) = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma\left(\frac{D}{2}\right)}{\Gamma\left(\frac{D-1}{2}\right)} \cdot \sin^{D-2} \theta, \quad \theta \in [0, \pi]. \quad (\text{A.13})$$

High-Dimensional Behavior As the dimension D increases, the function $\sin^{D-2}(\theta)$ becomes sharply peaked around $\theta = \pi/2$. In the limit of large D , nearly all randomly sampled vectors are almost orthogonal, and the angle distribution approaches a normal distribution centered at $\pi/2$ with variance close to zero.

APPENDIX B

DERIVATION OF THE DISTRIBUTION OF EUCLIDEAN DISTANCES ON THE SURFACE OF A HYPERSPHERE

This appendix derives the probability distribution of Euclidean (chord) distances between points uniformly sampled on the surface of a $(D-1)$ -dimensional sphere (hypersphere) [61]. The results offer geometric insight into the structure of high-dimensional spaces.

Preliminaries Let $\mathbb{S}^{(D-1)} \subset \mathbb{R}^D$ denote the surface of a hypersphere of radius R , defined as:

$$\mathbb{S}^{(D-1)} = \{\mathbf{x} \in \mathbb{R}^D \mid \|\mathbf{x}\| = R\}. \quad (\text{B.1})$$

Let $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ be two independent, uniformly distributed points on $\mathbb{S}^{(D-1)}$. The Euclidean (chord) distance d between them is given by:

$$d = \|\mathbf{s}^{(1)} - \mathbf{s}^{(2)}\| = \sqrt{2R^2 - 2R^2 \cos \theta} \quad (\text{B.2})$$

where $\theta \in [0, \pi]$ is the central angle between the two vectors. Then using half-angle identity $1 - \cos \theta = 2 \sin^2 \left(\frac{\theta}{2}\right)$, we can rewrite it as:

$$d = \|\mathbf{s}^{(1)} - \mathbf{s}^{(2)}\| = 2R \sin \left(\frac{\theta}{2}\right), \quad (\text{B.3})$$

Cumulative Distribution Function via Spherical Cap Geometry To find the distribution of chord lengths, we first compute the cumulative distribution function (CDF). For a fixed point $\mathbf{s}^{(1)}$, the set of points within Euclidean distance ℓ from it forms a hyperspherical cap on $\mathbb{S}^{(D-1)}$. The CDF is the probability that a randomly

chosen second point $\mathbf{s}^{(2)}$ lies within this cap:

$$F(d) = \mathbb{P} (\|\mathbf{s}^{(1)} - \mathbf{s}^{(2)}\| \leq d). \quad (\text{B.4})$$

The surface area of a hyperspherical cap of angular radius α is given by integrating the surface area of a $(D-2)$ -sphere of radius $R \sin \theta$ [77]:

$$\begin{aligned} A_{(D-1),\text{cap}}(R) &= R \cdot \int_0^\alpha A_{D-2}(R \sin \theta) R d\theta \\ &= R A_{D-2}(R) \int_0^\alpha \sin^{D-2} \theta d\theta, \end{aligned} \quad (\text{B.5})$$

where $A_{D-2}(R)$ is the surface area of a $(D-2)$ -sphere of radius R .

The total surface area of $\mathbb{S}^{(D-1)}$ is:

$$A_{D-1}(R) = \frac{2\pi^{D/2} R^{D-1}}{\Gamma(\frac{D}{2})}. \quad (\text{B.6})$$

Thus, the CDF of the chord length is:

$$F_{D-1}(d) = \frac{A_{\text{cap}}(R, \alpha)}{A_{D-1}(R)} = \frac{A_{D-2}(R)}{A_{D-1}(R)} \int_0^\alpha \sin^{D-2} \theta d\theta. \quad (\text{B.7})$$

Probability Density Function (PDF) To obtain the probability density function of the chord length d , we differentiate the CDF with respect to d . This yields:

$$f_d(d) = \frac{d}{R^2 B(\frac{D-1}{2}, \frac{1}{2})} \left(\frac{d^2}{R^2} - \frac{d^4}{4R^4} \right)^{\frac{D-3}{2}}, \quad 0 \leq d \leq 2R, \quad (\text{B.8})$$

where $B(a, b)$ is the Beta function:

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

This PDF describes the distribution of Euclidean distances between uniformly

chosen pairs of points on a hypersphere.

Mean and Variance of the Euclidean Distance As shown by Sidiropoulos (2014) in the analysis of chord-length distributions on N -spheres, explicit expressions for the raw moments—including the mean and variance—are available [61]. The exact expressions for the mean and variance are as follows:

$$\mu = \mathbb{E}[d] = \frac{B\left(\frac{D}{2}, \frac{D}{2}\right)}{B\left(D - \frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{2^{D-1}}{R}, \quad (\text{B.9})$$

where $B(a, b)$ is the Beta function. The variance can also be written as:

$$\sigma^2 = \text{Var}(d) = \left(2 - \frac{B^2\left(\frac{D}{2}, \frac{D}{2}\right)}{B^2\left(D - \frac{1}{2}, \frac{1}{2}\right)} \cdot 2^{2D-2}\right) R^2, \quad (\text{B.10})$$

Using properties of the Beta and Gamma functions, this can also be written as:

$$\sigma^2 = \left(2 - \frac{\Gamma^4\left(\frac{D}{2}\right)}{\pi\Gamma^2\left(D - \frac{1}{2}\right)} \cdot 2^{2D-2}\right) R^2. \quad (\text{B.11})$$

Asymptotic Behavior: As the dimension $D \rightarrow \infty$, these expressions simplify due to properties of the Gamma function and concentration of measure [62] on high-dimensional spheres:

$$\mathbb{E}[d] \rightarrow R\sqrt{2}, \quad \text{Var}(d) \rightarrow 0.$$

In this limit, the Euclidean distance d between two randomly chosen points on the hypersphere $\mathbb{S}^{(N-1)}$ concentrates sharply around $R\sqrt{2}$. This behavior reflects the concentration of measure phenomenon: in high dimensions, most points lie on the equator, resulting in a near-constant pairwise distance. As a result, the variance of the distance distribution tends to zero.

APPENDIX C

MAPPING BETWEEN HAMMING AND EUCLIDEAN DISTANCES

This section proves the relationship between Hamming and Euclidean distances for binary sequences. The Euclidean distance between rescaled binary sequences is proportional to the square root of their normalized H distance, connecting discrete and continuous measures of similarity

For two binary sequences $s^{(1)}$ and $s^{(2)}$ of length L , the Hamming distance is:

$$d_H(s^{(1)}, s^{(2)}) = \sum_{i=1}^L \delta(s^{(1)}(i) \neq s^{(2)}(i)), \quad (\text{C.1})$$

and the normalized Hamming distance is:

$$d_{H, \text{norm}}(s^{(1)}, s^{(2)}) = \frac{d_H(s^{(1)}, s^{(2)})}{L}. \quad (\text{C.2})$$

Binary sequences are mapped to vectors on the hypersphere using:

$$\phi(s) = \left(\frac{2s_1 - 1}{\sqrt{L}}, \dots, \frac{2s_L - 1}{\sqrt{L}} \right). \quad (\text{C.3})$$

This transformation maps each binary element s_i such that:

$$s_i = 1 \Rightarrow \frac{2s_i - 1}{\sqrt{L}} = \frac{1}{\sqrt{L}}, \quad s_i = 0 \Rightarrow \frac{2s_i - 1}{\sqrt{L}} = -\frac{1}{\sqrt{L}}. \quad (\text{C.4})$$

The resulting vector $\phi(s)$ is centered around the origin and lies on the surface of a hypersphere of radius $R = 1$.

Let $s^{(1)}$ and $s^{(2)}$ be two binary sequences of length L , and let $\phi(s^{(1)})$ and $\phi(s^{(2)})$ be their corresponding rescaled vectors. The Euclidean distance between $\phi(s^{(1)})$ and $\phi(s^{(2)})$ is given by:

$$d_{\text{Euc}}(s^{(1)}, s^{(2)}) = \sqrt{\sum_{i=1}^L (\phi(s^{(1)}(i)) - \phi(s^{(2)}(i)))^2}, \quad (\text{C.5})$$

Expanding the expression:

$$d_{\text{Euc}}(s^{(1)}, s^{(2)}) = \sqrt{\sum_{i=1}^L \left(\frac{2s^{(1)}(i) - 1}{\sqrt{L}} - \frac{2s^{(2)}(i) - 1}{\sqrt{L}} \right)^2}, \quad (\text{C.6})$$

Simplifying each term under the square root:

$$\frac{2s^{(1)}(i) - 1}{\sqrt{L}} - \frac{2s^{(2)}(i) - 1}{\sqrt{L}} = \frac{2(s^{(1)}(i) - s^{(2)}(i))}{\sqrt{L}}, \quad (\text{C.7})$$

Thus, the Euclidean distance becomes:

$$d_{\text{Euc}}(s^{(1)}, s^{(2)}) = \sqrt{\sum_{i=1}^L \left(\frac{2(s^{(1)}(i) - s^{(2)}(i))}{\sqrt{L}} \right)^2} = \frac{2}{\sqrt{L}} \sqrt{\sum_{i=1}^L (s^{(1)}(i) - s^{(2)}(i))^2}, \quad (\text{C.8})$$

Since $(s^{(1)}(i) - s^{(2)}(i))^2 = 1$ if $s^{(1)}(i) \neq s^{(2)}(i)$ and 0 otherwise, the summation $\sum_{i=1}^L (s^{(1)}(i) - s^{(2)}(i))^2$ is simply the H distance $d_{\text{H}}(s^{(1)}, s^{(2)})$. Therefore:

$$d_{\text{Euc}}(s^{(1)}, s^{(2)}) = \frac{2}{\sqrt{L}} \sqrt{d_{\text{H}}(s^{(1)}, s^{(2)})}, \quad (\text{C.9})$$

The normalized Hamming distance is given by:

$$d_{\text{H, norm}}(s^{(1)}, s^{(2)}) = \frac{d_{\text{H}}(s^{(1)}, s^{(2)})}{L}, \quad (\text{C.10})$$

Substituting $d_{\text{H}}(s^{(1)}, s^{(2)}) = m$, where m is the number of mismatches, we have:

$$d_{\text{Euc}}(s^{(1)}, s^{(2)}) = \frac{2}{\sqrt{L}} \sqrt{m} = 2\sqrt{\frac{m}{L}} = 2\sqrt{d_{\text{H, norm}}(s^{(1)}, s^{(2)})}. \quad (\text{C.11})$$

The Euclidean distance between rescaled binary sequences is proportional to the square root of their normalized H distance, connecting discrete and continuous measures of similarity.

APPENDIX D

CONVERGENCE OF DISTANCE DISTRIBUTIONS ON THE L -CUBE AND THE $(L-1)$ -SPHERE

Consider a sequence of length L , where each position can assume one of two possible values: A or B . We define an encoding that maps these characters to real numbers:

$$A \mapsto \frac{1}{\sqrt{L}}, \quad B \mapsto -\frac{1}{\sqrt{L}}. \quad (\text{D.1})$$

Using this encoding, each sequence is represented as a vector in \mathbb{R}^L , where the i -th component corresponds to the value at position i in the sequence.

This representation has several important geometric properties:

- Each encoded sequence corresponds to a vertex of an L -dimensional hypercube with side length $\sqrt{2L}$.
- Since each vector contains components of magnitude $\frac{1}{\sqrt{L}}$ or $-\frac{1}{\sqrt{L}}$, the Euclidean norm of every vector is:

$$\|\mathbf{x}\| = \sqrt{L \cdot \left(\frac{1}{\sqrt{L}}\right)^2} = \sqrt{L \cdot \frac{1}{L}} = 1, \quad (\text{D.2})$$

meaning that all such vectors lie on the surface of the $(L-1)$ -sphere of radius 1 embedded in \mathbb{R}^L .

Thus, this encoding defines a discrete set of points lying both on the unit sphere and at the vertices of a hypercube in \mathbb{R}^L .

As L grows, the vertices of the L -cube become uniformly dense on the surface of the L -sphere. Hence, one can approximate the distribution of distances between

these vertices using the distribution of chord lengths between points on the $(L - 1)$ -sphere.

Distance Distribution on the L-Cube

Consider a vector \vec{x} uniformly distributed on the vertices of an L -cube of side $\sqrt{2L}$. The Cartesian components of this vector satisfy $x_i = \pm \frac{1}{\sqrt{L}}$. The scalar product between two independent vectors $\vec{s}^{(1)}$ and $\vec{s}^{(2)}$ is:

$$\vec{s}^{(1)} \cdot \vec{s}^{(2)} = \sum_{i=1}^L s_i^{(1)} s_i^{(2)} = \sum_{i=1}^L \pm \frac{1}{L}. \quad (\text{D.3})$$

This scalar product is a random variable equal to the sum of L i.i.d. terms with mean 0 and variance $\frac{1}{L}$. By the Central Limit Theorem, it converges to a Gaussian distribution as $L \rightarrow \infty$. The squared Euclidean distance between two such vectors is:

$$d_{\text{cube},L}^2 = 2 - 2\vec{s}^{(1)} \cdot \vec{s}^{(2)}. \quad (\text{D.4})$$

This is distributed as a Gaussian centered at 2 with variance $\frac{4}{L}$. For large L , the corresponding probability density function is:

$$P(d_{\text{cube},L}^2) = \frac{1}{\sqrt{8\pi/L}} \exp\left(-\frac{(d_{\text{cube},L}^2 - 2)^2}{8/L}\right). \quad (\text{D.5})$$

Alternatively, the scalar product can be expressed in terms of a binomial variable:

$$\vec{s}^{(1)} \cdot \vec{s}^{(2)} = \frac{2k}{L} - 1, \quad d_{\text{cube},L}^2 = 4 \left(1 - \frac{k}{L}\right). \quad (\text{D.6})$$

The corresponding characteristic function is:

$$\phi_{\text{cube},L}(t) = \frac{e^{4it}}{2^L} \left(1 + e^{-\frac{4it}{L}}\right)^L. \quad (\text{D.7})$$

By computing the derivative:

$$\frac{\partial}{\partial t} \phi_{\text{cube},L} = 4i \phi_{\text{cube},L} \left(1 - \frac{1}{1 + e^{\frac{4it}{L}}} \right), \quad (\text{D.8})$$

and expanding for $L \rightarrow \infty$:

$$\frac{\partial}{\partial t} \phi_{\text{cube},L} = \phi_{\text{cube},L} \left(2i - \frac{4t}{L} + \frac{12it^2}{L^2} + o\left(\frac{1}{L^3}\right) \right). \quad (\text{D.9})$$

It is clear that eq. D.9 has the Gaussian solution in eq. D.5 if approximated at the order $o(\frac{1}{d})$. We will use the previous expression to prove that the convergence between the distribution on the cube and on the sphere is limited to the order $o(\frac{1}{d})$. For higher orders, both the distributions converge to zero, but with different rates.

Distance Distribution on the $(L-1)$ -Sphere

Consider two vectors $\vec{s}^{(1)}$ and $\vec{s}^{(2)}$ uniformly distributed on the surface of a unit $(L-1)$ -sphere. In Appendix B, we derived the distribution of distances for such a sphere. Using a change of variable, the distribution of the squared distance $d_{\text{sphere},L}^2 = \|\vec{s}^{(1)} - \vec{s}^{(2)}\|^2$ satisfies:

$$P(d_{\text{sphere},L}^2) \propto \left(d_{\text{sphere},L}^2 - \frac{d_{\text{sphere},L}^4}{4} \right)^{(L-3)/2}. \quad (\text{D.10})$$

Introducing the change of variable $x = \frac{d_{\text{sphere},L}^2}{4}$, this becomes a beta distribution with parameters $\alpha = \beta = \frac{L-1}{2}$:

$$P(x) = \frac{\Gamma(L-1)}{\Gamma^2\left(\frac{L-1}{2}\right)} x^{\frac{L-3}{2}} (1-x)^{\frac{L-3}{2}}. \quad (\text{D.11})$$

Note that because the maximum distance between points on a unitary $(L-1)$ -sphere is always 2, x is defined in the range $[0, 1]$.

The characteristic function of a beta distribution with parameters $\alpha = \beta = \frac{L-1}{2}$ can be written in terms of the confluent hypergeometric function ${}_1F_1(\alpha, 2\alpha, it)$ and the modified Bessel functions I_α . Rescaling the variable $t \rightarrow 4t$, we obtain:

$$\phi(t) = e^{2it} {}_1F_1(\alpha, 2\alpha, it) = e^{2it} \frac{1}{\Gamma(\alpha)} \int_0^1 e^{itx} x^{\alpha-1} (1-x)^{\alpha-1} dx. \quad (\text{D.12})$$

Alternatively, using the modified Bessel functions, the characteristic function can be expressed as:

$$\phi(t) = e^{2it} I_\alpha(2it), \quad (\text{D.13})$$

where I_α is the modified Bessel function of the first kind.

$$\phi_{\text{sphere},L}(t) = e^{2it} \sum_{m=0}^{\infty} \frac{\Gamma\left(\frac{L-1}{2}\right)}{m! \Gamma\left(m + \frac{L-1}{2}\right)} (-t^2)^m. \quad (\text{D.14})$$

Convergence of Distributions as $L \rightarrow \infty$

Consider the beta distribution with equal parameters $\alpha = \beta = \frac{L-1}{2}$, as defined in equation (D.10). Recall that the variable $x = \frac{d_{\text{sphere},L}^2}{4}$, is defined between 0 and 1. The average of this distribution is equal to $\frac{1}{2}$, while the variance is $\frac{1}{4L}$.

If we standardize x with the change of variable $y = 2\sqrt{L}\left(x - \frac{1}{2}\right)$, and compute the corresponding Jacobian, we get:

$$\begin{aligned}
P(y) &= \frac{1}{2\sqrt{L}} P\left(\frac{y}{2\sqrt{L}} + \frac{1}{2}\right) \\
&= \frac{1}{2\sqrt{L}} \frac{\Gamma(L-1)}{\Gamma^2\left(\frac{L-1}{2}\right)} \left(\frac{y}{2\sqrt{L}} + \frac{1}{2}\right)^{\frac{L-3}{2}} \left(1 - \left(\frac{y}{2\sqrt{L}} + \frac{1}{2}\right)\right)^{\frac{L-3}{2}} \\
&= \frac{1}{2\sqrt{L}} \frac{\Gamma(L-1)}{\Gamma^2\left(\frac{L-1}{2}\right)} \left(\frac{1}{2} + \frac{y}{2\sqrt{L}}\right)^{\frac{L-3}{2}} \left(\frac{1}{2} - \frac{y}{2\sqrt{L}}\right)^{\frac{L-3}{2}} \\
&= \frac{1}{2\sqrt{L}} \frac{\Gamma(L-1)}{\Gamma^2\left(\frac{L-1}{2}\right)} \left(\frac{1}{4} - \frac{y^2}{4L}\right)^{\frac{L-3}{2}}.
\end{aligned} \tag{D.15}$$

We can now use Stirling's formula for the Gamma function which given by $\Gamma(z) \sim \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left[1 + o\left(\frac{1}{z}\right)\right]$, and the fact that $L \rightarrow \infty$, to write:

$$\begin{aligned}
P(y) &= \frac{1}{2\sqrt{L}} \frac{\sqrt{2\pi(L-1)} \left(\frac{L-1}{e}\right)^{L-1}}{\pi(L-1) \left(\frac{L-1}{2e}\right)^{L-1}} \left(\frac{1}{4} - \frac{y^2}{4L}\right)^{\frac{L-3}{2}} \\
&= \frac{L-1}{2L-1} \frac{2^{L-1}}{\sqrt{2\pi}} \left(\frac{1}{4} - \frac{y^2}{4L}\right)^{\frac{L-3}{2}} \left[1 + o\left(\frac{1}{L}\right)\right] \\
&= \frac{1}{\sqrt{2\pi}} \left(\frac{1}{4} - \frac{y^2}{4L}\right)^{\frac{L-3}{2}} \left[1 + o\left(\frac{1}{L}\right)\right] \\
&= \frac{1}{\sqrt{2\pi}} \left(1 - \frac{y^2}{L}\right)^{\frac{L-3}{2}} \left[1 + o\left(\frac{1}{L}\right)\right] \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \left[1 + o\left(\frac{1}{L}\right)\right].
\end{aligned} \tag{D.16}$$

which is the same distribution in D.5.

This proves that the random variable y is distributed as a standard normal, up to corrections of the order $o\left(\frac{1}{L}\right)$. Returning to the squared distance with the change $d_{\text{sphere},L}^2 = 4x = \frac{y}{\sqrt{L}} + 2$, we find:

$$d_{\text{sphere},L}^2 \sim N\left(2, \frac{4}{L}\right), \tag{D.17}$$

which matches the Gaussian distribution derived for $d_{\text{cube},L}^2$ in equation D.5.

The distance distributions on the L -cube and $(L-1)$ -sphere converge to the

same Gaussian distribution as $L \rightarrow \infty$, with corrections of order $o\left(\frac{1}{L}\right)$. This demonstrates the equivalence of the two distributions in the high-dimensional limit.

APPENDIX E

NORMALIZED HAMMING DISTANCE OF BINARY SEQUENCES

Here, we derive normalized hamming distances between binary sequences. Each site i along the chain can take one of two values, commonly denoted as $+1$ or -1 . These values can represent simplified sequence elements, such as the presence or absence of a feature. We denote each site by $s_i \in \{-1, +1\}$, and a full sequence of length L by the vector $s = (s_1, s_2, \dots, s_L)$.

The normalized Hamming distance between two sequences $s^{(1)}$ and $s^{(2)}$, which measures the fraction of sites at which the sequences differ, is defined as:

$$\hat{d}_H(s^{(1)}, s^{(2)}) = \frac{1}{L} \sum_{i=1}^L \delta(s_i^{(1)}, s_i^{(2)}), \quad (\text{E.1})$$

where $\delta(a, b)$ is the Kronecker delta:

$$\delta(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases} \quad (\text{E.2})$$

For binary values in $\{-1, +1\}$, we can express the Kronecker delta using the identity:

$$\delta(a, b) = \frac{1 - ab}{2}. \quad (\text{E.3})$$

Substituting this identity into the expression for \hat{d}_H , we obtain:

$$\begin{aligned} \hat{d}_H &= \frac{1}{L} \sum_{i=1}^L \frac{1 - s_i^{(1)} s_i^{(2)}}{2} \\ &= \frac{1}{2} - \frac{1}{2L} \sum_{i=1}^L s_i^{(1)} s_i^{(2)}. \end{aligned} \quad (\text{E.4})$$

APPENDIX F

BLOCKWISE 1D ISING MODEL

We consider binary sequences of length L generated from an **independent block Ising model**. The sequence is partitioned into $n = L/b$ independent blocks, each containing b sites. Inside each block, spins interact according to the one-dimensional nearest-neighbor Ising Hamiltonian:

$$H_b(\mathbf{s}) = -J \sum_{i=1}^{b-1} s_i s_{i+1}, \quad (\text{F.1})$$

where $s_i \in \{-1, +1\}$, and $J > 0$. There is no interaction between different blocks.

This energy function favors aligned spins, with the lowest energy achieved when all spins are identical within the block (i.e., fully ferromagnetic).

Partition Function of Block-Wise Correlated Sequences

The partition function is the weighted sum over all configurations in the block, where each configuration is weighted by its Boltzmann factor:

$$Z_b(\beta) = \sum_{\mathbf{s} \in \{-1, +1\}^b} \exp \left(\beta \sum_{i=1}^{b-1} s_i s_{i+1} \right). \quad (\text{F.2})$$

As $\beta \rightarrow \infty$ (i.e., low temperature), the system becomes increasingly biased toward low-energy configurations (all spins aligned).

Block Size $b = 1$: When $b = 1$, there are no interactions between spins:

$$H_1(s_1) = 0 \quad \Rightarrow \quad Z_1(\beta) = \sum_{s_1 = \pm 1} 1 = 2. \quad (\text{F.3})$$

All configurations are equally probable, independent of temperature. This corresponds to a model with no epistasis.

Block Size $b = 2$: With two spins, the Hamiltonian contains one interaction:

$$H_2(s_1, s_2) = -s_1 s_2. \quad (\text{F.4})$$

Thus, the partition function becomes:

$$\begin{aligned} Z_2(\beta) &= \sum_{s_1, s_2 = \pm 1} e^{\beta s_1 s_2} \\ &= 2e^\beta + 2e^{-\beta} = 4 \cosh(\beta). \end{aligned} \quad (\text{F.5})$$

General Case: For larger blocks, we apply the transfer matrix method. Define the transfer matrix T by:

$$T_{s,s'} = e^{\beta s s'}, \quad s, s' \in \{-1, +1\}. \quad (\text{F.6})$$

This yields:

$$T = \begin{bmatrix} e^\beta & e^{-\beta} \\ e^{-\beta} & e^\beta \end{bmatrix}. \quad (\text{F.7})$$

Let $\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Then the partition function for block size b is given by:

$$Z_b(\beta) = \mathbf{v}^\top T^{b-1} \mathbf{v}. \quad (\text{F.8})$$

The eigenvalues of T are:

$$\lambda_+ = 2 \cosh(\beta), \quad \lambda_- = 2 \sinh(\beta), \quad (\text{F.9})$$

and the matrix is diagonalizable. Therefore, the partition function simplifies to:

$$Z_b(\beta) = \lambda_+^{b-1} + \lambda_-^{b-1} = (2 \cosh(\beta))^{b-1} + (2 \sinh(\beta))^{b-1}. \quad (\text{F.10})$$

Block size b	Number of interactions	Partition function $Z_b(\beta)$
1	0	2
2	1	$4 \cosh(\beta)$
≥ 3	$b - 1$	$(2 \cosh(\beta))^{b-1} + (2 \sinh(\beta))^{b-1}$

Table F.1: **Summary** Partition function $Z_b(\beta)$ for various block sizes

Full Partition Function For a sequence of length L divided into blocks of size b , the full partition function is the product of block contributions,

$$Z_{\text{seq}}(\beta) = \prod_{m=1}^{L/b} Z_b(\beta) = [Z_b(\beta)]^{L/b}. \quad (\text{F.11})$$

Using the block expression derived above, this becomes

$$Z_{\text{seq}}(\beta) = \left[(2 \cosh \beta)^{b-1} + (2 \sinh \beta)^{b-1} \right]^{L/b}. \quad (\text{F.12})$$

High temperature ($\beta \rightarrow 0$). When β is small, $\cosh \beta \approx 1$, $\sinh \beta \approx \beta$, so each block has $Z_b(0) = 2^{b-1}$. Thus

$$\lim_{\beta \rightarrow 0} Z_{\text{seq}}(\beta) = (2^{b-1})^{L/b} \sim 2^L, \quad (\text{F.13})$$

corresponding to all 2^L configurations being equally likely.

Low temperature ($\beta \rightarrow \infty$). At large β , only the ground states contribute. Each block has two ferromagnetic ground states, so the total degeneracy is $2^{L/b}$. Including the Boltzmann weight of the ground-state energy,

$$Z_{\text{seq}}(\beta) \sim 2^{L/b} e^{\beta(L-L/b)}. \quad (\text{F.14})$$

Thus,

$$\lim_{\beta \rightarrow \infty} e^{-\beta(L-L/b)} Z_{\text{seq}}(\beta) = 2^{L/b}. \quad (\text{F.15})$$

As block size increases, the partition function becomes increasingly dominated by low-energy (ferromagnetically aligned) configurations, especially at low temperatures (high β). This reflects stronger epistatic constraints within each block. When $b = 1$, the model reduces to an independent site model with no correlation. When $b = L$, the entire sequence behaves as a single Ising chain, corresponding to maximal correlation (global epistasis).

Expected Value and Variance of Hamming Distance

For two independent sequences $s^{(1)}$ and $s^{(2)}$, we define the normalized Hamming distance:

$$\hat{d}_H = \frac{1}{L} \sum_{i=1}^L X_i, \quad (\text{F.16})$$

where $X_i = \frac{1 - s_i^{(1)} s_i^{(2)}}{2}$.

Expected value. The mean of the normalized Hamming distance remains unchanged even in the presence of intra-block correlations. Since the 1D Ising model without an external field is symmetric under global spin flip, each spin has zero mean, i.e., $\mathbb{E}[s_i] = 0$. For two independently sampled sequences from this model, the expectation $\mathbb{E}[s_i^{(1)} s_i^{(2)}] = \mathbb{E}[s_i]^2 = 0$ still holds. Consequently, the expected value of the normalized Hamming distance is $\mathbb{E}[d_H] = \frac{1}{2}$, regardless of block size or interaction strength within blocks.

Variance. As defined in Appendix E, the variance of normalized Hamming distance is given by:

$$\text{Var}(\hat{d}_H) = \frac{1}{L^2} \sum_{i=1}^L \sum_{j=1}^L \text{Cov}(X_i, X_j). \quad (\text{F.17})$$

Since blocks are independent, covariance terms vanish between different blocks,

so:

$$\text{Var}(\hat{d}_H) = \frac{n}{L^2} \sum_{i=1}^b \sum_{j=1}^b \text{Cov}(X_i, X_j). \quad (\text{F.18})$$

Using $n = L/b$, this simplifies to:

$$\text{Var}(\hat{d}_H) = \frac{1}{Lb} \sum_{i=1}^b \sum_{j=1}^b \text{Cov}(X_i, X_j). \quad (\text{F.19})$$

We first compute $\mathbb{E}[X_i]$. Since $\mathbb{E}[s_i] = 0$, we have:

$$\mathbb{E}[X_i] = \frac{1}{2}. \quad (\text{F.20})$$

Next, we expand:

$$X_i X_j = \frac{1}{4} \left(1 - s_i^{(1)} s_i^{(2)} - s_j^{(1)} s_j^{(2)} + s_i^{(1)} s_i^{(2)} s_j^{(1)} s_j^{(2)} \right). \quad (\text{F.21})$$

Taking expectation and using independence between sequences:

$$\mathbb{E}[X_i X_j] = \frac{1}{4} (1 + C_{ij}^2), \quad (\text{F.22})$$

where $C_{ij} = \mathbb{E}[s_i s_j]$ is the two-point correlation function within a block.

Thus, the covariance is:

$$\text{Cov}(X_i, X_j) = \frac{1}{4} C_{ij}^2. \quad (\text{F.23})$$

We obtain:

$$\text{Var}(\hat{d}_H) = \frac{1}{4Lb} \sum_{i=1}^b \sum_{j=1}^b C_{ij}^2. \quad (\text{F.24})$$

For the 1D Ising model without an external field, the correlation function is:

$$C_{ij} = \tanh^{|i-j|}(\beta J), \quad (\text{F.25})$$

where $\beta = 1/T$. Thus:

$$\text{Var}(\hat{d}_H) = \frac{1}{4Lb} \sum_{i=1}^b \sum_{j=1}^b \tanh^{2|i-j|}(\beta J). \quad (\text{F.26})$$

Let $x = \tanh^2(\beta J)$. The double sum can be rewritten as:

$$\sum_{i=1}^b \sum_{j=1}^b x^{|i-j|} = \sum_{k=0}^{b-1} (b-k)x^k. \quad (\text{F.27})$$

Hence,

$$\text{Var}(\hat{d}_H) = \frac{1}{4Lb} \sum_{k=0}^{b-1} (b-k)x^k. \quad (\text{F.28})$$

This sum can be rewritten as:

$$\sum_{k=0}^{b-1} (b-k)x^k = b \sum_{k=0}^{b-1} x^k - \sum_{k=0}^{b-1} kx^k. \quad (\text{F.29})$$

The closed-form expressions for the sums are:

$$\sum_{k=0}^{b-1} x^k = \frac{1-x^b}{1-x}, \quad (\text{F.30})$$

and

$$\sum_{k=0}^{b-1} kx^k = \frac{x(1-x^b(1+b(1-x)))}{(1-x)^2}. \quad (\text{F.31})$$

Thus, we obtain the final formula:

$$\text{Var}(\hat{d}_H) = \frac{1}{4Lb} \left[b \cdot \frac{1-x^b}{1-x} - \frac{x(1-x^b(1+b(1-x)))}{(1-x)^2} \right]. \quad (\text{F.32})$$

High temperature limit At high temperature ($T \gg 1$), $\tanh(\beta J) \approx \beta J \ll 1$, so $x \rightarrow 0$. Then:

$$\text{Var}(\hat{d}_H) \approx 0. \quad (\text{F.33})$$

This corresponds to the independent site model.

Low temperature limit At low temperature ($T \rightarrow 0$), $\tanh(\beta J) \rightarrow 1$, so $x \rightarrow 1$.

Then,

$$\sum_{k=0}^{b-1} (b-k) \rightarrow \frac{b(b+1)}{2}, \quad (\text{F.34})$$

which gives:

$$\text{Var}(\hat{d}_H) \approx \frac{b+1}{8L}. \quad (\text{F.35})$$

Numerical Evaluation In Figure 19, we show the variance $\text{Var}(d_H)$ as a function of block size for different temperatures. It illustrates how the variance of the normalized Hamming distance depends on both block size b and temperature T . At low temperature, increasing the block size leads to stronger intra-block correlations, resulting in larger variance due to collective block-level fluctuations. This behavior saturates according to the analytical low-temperature approximation $\text{Var}(\hat{d}_H) \approx (b+1)/(8L)$. In contrast, at high temperature, where spin interactions are weak, the variance approaches its uncorrelated limit of which is approximately 0, independent of the block structure.

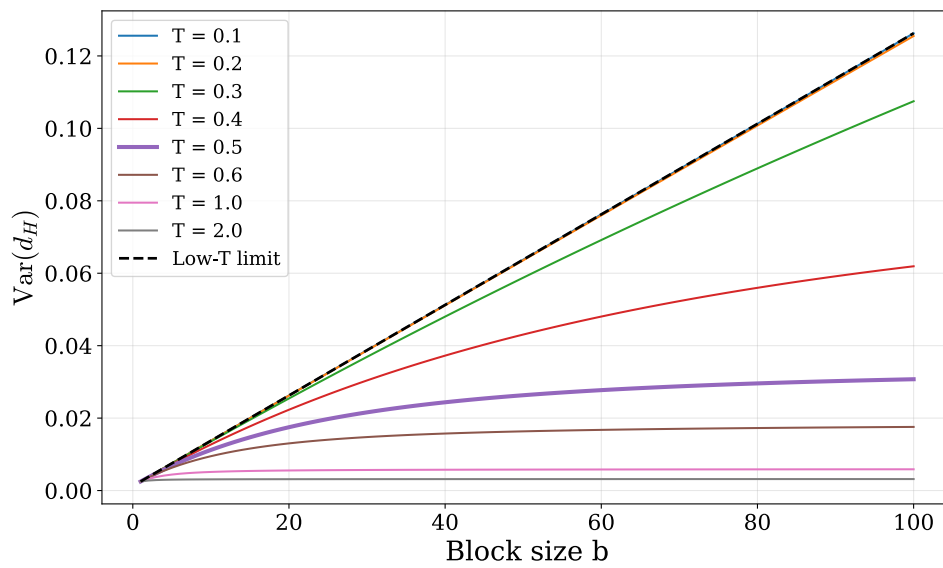


Figure 19: Variance of Normalized Hamming Distance $\text{Var}(d_H)$ as a Function of Block Size b for Different Temperatures, with $J = 1$. The thick purple curve corresponds to $T = 0.5$. The dashed line is the low-temperature limit $\frac{b+1}{8L}$.