Topological Invariants of Sequence-Derived Constraint Structures:
A Hypothesis for Predicting Protein Function Without Structural Simulation

Richard Griffiths
Katiya McKerry
Independent Researchers

December 2025 – Version 2.1

Abstract

Predicting protein function typically requires either full 3D structure prediction or comparison with known homologues. Both approaches treat function as emergent from folding. This paper presents a testable hypothesis: *protein functional class is encoded in topological invariants of the constraint structure that sequences specify*—not in the sequence string itself, but in the pattern of coevolutionary couplings, contact requirements, and folding pathway restrictions that any viable sequence must satisfy. If correct, functional classification could be performed without structural simulation, offering significant computational savings. We provide concrete definitions of candidate invariants, clarify what is meant by "constraint structure," address known biological challenges (sequence-divergent homologues, intrinsically disordered proteins, convergent evolution), propose operational experimental criteria, and compare against modern baselines including transformer-based methods. The objective is a falsifiable framework suitable for computational biology groups to evaluate.

1 Introduction

Contemporary protein function prediction relies on two paradigms:

(i) inferring function *via structure*, using methods such as AlphaFold [1] or ESMFold [2]; or

(ii) inferring function via sequence homology or learned embeddings [3].

Both approaches have achieved remarkable success. Structure-based methods now approach experimental accuracy; transformer-based sequence models match or exceed homology methods on many functional classification tasks. Yet both treat function as information that becomes accessible only after geometric inference or high-dimensional embedding.

This paper explores whether a *third* approach is viable: extracting topological invariants directly from sequence-derived constraint structures, bypassing both explicit structure prediction and black-box embedding.

1.1 The Central Reframing

An earlier version of this hypothesis proposed that invariants exist “at sequence level.” This framing is inadequate. Protein families such as TIM barrels exhibit conserved structure and function despite sequence identity below 10% [4]. If invariants lived in the sequence string, such divergence would destroy them.

The refined hypothesis is that invariants live not in the sequence itself, but in the **constraint structure** the sequence specifies: the pattern of “what must interact with what” that any viable sequence for a given fold or function must satisfy. Two sequences with 10% identity can encode *isomorphic constraint graphs.* The letters diverged; the topology of constraints did not.

Operational definition. Because constraint structure is not directly observable, we define it through empirical proxies: coevolutionary coupling matrices, predicted contact maps, k-mer adjacency complexes, or attention-derived interaction graphs. The hypothesis does not rely on any specific extraction method; it concerns the invariants preserved across them.

2 Candidate Invariants

We propose four candidate invariant families, each computable from sequence-derived objects without explicit 3D structure prediction.

2.1 Coevolutionary Coupling Topology

Direct Coupling Analysis (DCA) [5] extracts pairwise dependencies among residue positions. The coupling matrix J_{ij} captures compatibility constraints needed to preserve function.

Invariant candidates:

- Persistent homology of the thresholded coupling graph
- Betti numbers β_0, β_1, β_2
- Spectral properties of the coupling Laplacian

Prediction: TIM barrel sequences, despite <10% sequence identity, should exhibit quasi-isometric coupling graphs.

2.2 k-mer Simplicial Complex

Construct a simplicial complex where vertices represent codons or k-mers, and simplices reflect co-occurrence or adjacency patterns.

Invariant candidates:

- Persistent homology barcodes
- Euler characteristic
- Clique dimension distribution

This approach captures “local grammar” independent of global alignment.

2.3 Predicted Contact Map Topology

Coevolution-derived contact predictions [6] can be treated as graphs prior to geometric embedding.

Invariant candidates:

- Graph homology
- Degree and clustering spectra
- Modularity and community structure

2.4 Attention Topology from Learned Models

Transformer models such as ESM-2 [3] induce attention matrices that implicitly encode constraints.

Invariant candidates:

- Persistent homology of attention graphs across layers
- Information-theoretic measures of attention flow

If ESM-2's performance reflects its capacity to learn constraint topology, making these invariants explicit should explain its strengths and weaknesses.

3 Addressing Biological Challenges

3.1 Sequence-Divergent Homologues

Challenge: TIM barrels share fold and function despite extreme sequence divergence [4].

Response: This is *expected* under the constraint-topology framing. Any sequence that satisfies the barrel's interaction constraints will fold appropriately. Constraint topology, not letters, is conserved.

Test: Coupling topology across divergent TIM barrels should cluster tightly.

3.2 Intrinsically Disordered Proteins

Challenge: Approximately 30% of eukaryotic proteins are intrinsically disordered [7].

Response: IDPs should form a distinct invariant class with sparse, high-entropy constraint structures. Their separation in invariant space supports, rather than contradicts, the framework.

3.3 Convergent Evolution

Challenge: Different global folds can yield the same catalytic function [8].

Response: Global invariants will fail in these cases. Constraint topology is more suited to functions tied to global architecture (e.g., channels, barrels, coiled-coils). For convergent enzymes, local invariants around active-site neighborhoods may be more appropriate.

Explicit limitation: The hypothesis is not expected to capture functions realizable via diverse global architectures.

4 Functorial Framing

We model the biological expression chain as:

$$\mathcal{S} \xrightarrow{F_1} \mathcal{M} \xrightarrow{F_2} \mathcal{A} \xrightarrow{F_3} \mathcal{P} \xrightarrow{F_4} \mathcal{Q},$$

where $\mathcal{S}$ = sequences, $\mathcal{M}$ = mRNA, $\mathcal{A}$ = amino acid chains, $\mathcal{P}$ = folded structures, and $\mathcal{Q}$ = quaternary assemblies.

An invariant I satisfies $I(F_i(x)) = I(x)$ for all x.

4.1 The Kernel of Folding

Two sequences lie in the kernel of the composed functor $F = F_4 \circ F_3 \circ F_2 \circ F_1$ if they fold to functionally equivalent proteins. Neutral evolution provides natural samples of $\ker(F)$.

Concrete example: TIM barrel sequences with 90% divergence but preserved fold inhabit the same kernel class. Their constraint structures should therefore exhibit topological equivalence.

4.2 Adjoint-Like Behavior

If an approximate left adjoint to F exists, it would correspond to "free" generation of sequences from functional specifications—consistent with the success of inverse folding models such as ProteinMPNN [9].

5 Experimental Program

5.1 Phase 1: Algebraic Structure Discovery

Train a regularised sequence-to-function model. Extract learned operations and test for:

- closure,
- associativity,
- existence of a small generating set.

5.2 Phase 2: Invariant Extraction and Clustering

Compute topological features from:

1. DCA coupling topology,
2. k-mer complexes,
3. attention-derived graphs,
4. delay embeddings.

Validate by consistency, discriminability, and mutation stability.

5.3 Phase 3: Predictive Validation

Compare invariant-based classification against:

- random baseline,
- CNN + BLAST,
- ESM-2 embeddings,
- AF2-derived structure-based annotation.

Explainability criterion: Correlating invariants with where ESM-2 succeeds or fails supports the claim even without exceeding ESM-2 accuracy.

5.4 Phase 4: Synthetic Biology Validation

Synthetic proteins offer a controlled environment. Divergence between synthetic and natural families illuminates what biological constraints the model captures.

6 Falsification Criteria

The hypothesis is unsupported if:

1. learned operations do not compress into an algebraic structure,
2. topological features do not cluster by function,
3. invariants are unstable under neutral mutations,
4. predictive performance does not exceed CNN + BLAST,
5. no relationship with ESM-2's internal structure emerges.

Each failure mode yields scientifically meaningful information.

7 Relationship to Existing Work

The hypothesis connects to:

- structure prediction (AlphaFold, ESMFold),
- transformer embeddings (ESM-2),
- topological data analysis [11],
- coevolutionary methods [5],
- algebraic frameworks in biology [12].

Recent Related Work

Recent advances in sequence-to-function prediction further motivate the search for interpretable, topology-aware approaches. Several 2024–2025 methods integrate structural or contact-derived information to improve functional classification. DPFunc [13] incorporates domain-guided structural features and demonstrates that purely sequence-based models remain limited in capturing functionally relevant constraints. Similarly, TAWFN [14] and related deep-learning frameworks show improved performance when auxiliary structural priors are introduced.

Meanwhile, contact prediction from sequence alone continues to advance through graph-based and attention-based architectures such as PCP-GC-LM [15] and Attention-UNet models [16], which infer residue–residue interaction patterns without requiring solved structures. These developments support the central premise of this work: that *sequence-derived constraint information carries significant functional signal*.

However, none of these approaches attempt to extract or characterise the *topological invariants* of these constraint structures. The present hypothesis differs in proposing that such invariants may be both function-determining and preserved across extreme sequence divergence, offering a conceptually unified and potentially more interpretable route to functional prediction.

8 Implications if Validated

8.1 Computational Efficiency

Invariant extraction is $O(L)$ or $O(\text{MSA size})$, avoiding expensive geometric inference.

8.2 Interpretability

Topological invariants are mathematically characterisable, enhancing transparency.

8.3 Protein Design

Understanding the invariant–function map enables rational design by specifying desired invariants.

8.4 Generalisation

The constraint-topology lens may apply to regulatory networks, metabolic pathways, developmental constraints, and neural circuits.

9 Conclusion

We present a refined, testable hypothesis: protein functional class is encoded in topological invariants of sequence-derived constraint structures. These invariants can be preserved across extreme sequence divergence and may explain the success of coevolutionary and transformer-based models. Regardless of outcome, investigating this hypothesis clarifies the role of topology in biological information processing and may inform new computational strategies for protein science.

Acknowledgments

This version benefited from critiques that strengthened invariant definitions, clarified biological limitations, and sharpened the experimental program.

References

[1] Jumper, J., et al. “Highly accurate protein structure prediction with AlphaFold.” *Nature*, 596:583–589, 2021.

[2] Lin, Z., et al. “Evolutionary-scale prediction of atomic-level protein structure with a language model.” *Science*, 379:1123–1130, 2023.

[3] Lin, Z., et al. “Language models of protein sequences at the scale of evolution enable accurate structure prediction.” *bioRxiv*, 2022.

[4] Nagano, N., Orengo, C.A., & Thornton, J.M. “One fold with many functions: the evolutionary relationships between TIM barrel families.” *J. Mol. Biol.*, 321:741–765, 2002.

[5] Morcos, F., et al. “Direct-coupling analysis of residue coevolution captures native contacts across many protein families.” *PNAS*, 108:E1293–E1301, 2011.

[6] Jones, D.T., et al. “MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding.” *Bioinformatics*, 31:999–1006, 2015.

[7] Dunker, A.K., et al. “Intrinsically disordered protein.” *J. Mol. Graph. Model.*, 19:26–59, 2001.

[8] Doolittle, R.F. “Convergent evolution: the need to be explicit.” *Trends Biochem. Sci.*, 19:15–18, 1994.

[9] Dauparas, J., et al. “Robust deep learning-based protein sequence design using ProteinMPNN.” *Science*, 378:49–56, 2022.

[10] Bauer, U. “Ripser: efficient computation of Vietoris-Rips persistence barcodes.” *J. Appl. Comput. Topol.*, 5:391–423, 2021.

[11] Carlsson, G. “Topology and data.” *Bull. Amer. Math. Soc.*, 46:255–308, 2009.

[12] Baez, J.C. & Pollard, B.S. “A compositional framework for reaction networks.” *Rev. Math. Phys.*, 29:1750028, 2017.

[13] Wang, X., et al. “DPFunc: accurately predicting protein function via deep learning with domain-guided structure information.” *Nature Communications*, 2025.

[14] Meng, Z., et al. “TAWFN: a deep learning framework for protein function prediction.” *Bioinformatics*, 40(10):btae571, 2024.

[15] Ouyang, Q., et al. “PCP-GC-LM: Single-sequence-based protein contact prediction via graph-convolutional language models.” *BMC Bioinformatics*, 25:448, 2024.

[16] Zhang, Y., et al. “Using Attention-UNet Models to Predict Protein Contact Maps.” *Journal of Computational Biology*, 2024.