# Evolutionary Fitness Landscape Ruggedness via the Latent Framework

Dr. Tamás Nagy

tnagyphd@gmail.com

draft • 2026-04-08

## Executive summary (non-technical)

Biologists and engineers often talk about "rugged" fitness landscapes when local improvements dead-end far below a global best. The trouble is that different subfields measure ruggedness in incompatible ways. This note packages two complementary coordinates—how compressible fitness variation is, and how many latent directions are needed to reconstruct it—and shows how they move together on Kauffman's NK model when epistasis tightens.

The formal layer is a chain of machine-checked inequalities in a Latent calculus; the empirical layer is exhaustive enumeration at $L = 10$ with a reproducible numerical harness. Together they give a concrete map from epistasis intensity to diagnostics that practitioners can track alongside classical peak counts and greedy-walk outcomes. The draft stops short of population-genetic predictions: it supplies geometric coordinates that demography-aware models could ingest later.

## Abstract

Evolutionary fitness landscapes encode how genotypes map to reproductive success. Rugged landscapes—with many local peaks and epistatic interactions—shape adaptation, evolvability, and the predictability of evolutionary paths. Yet quantitative comparisons across biological systems remain fragmented: classical measures (correlation length, number of optima, accessibility) do not share a common geometric backbone that connects sequence space, population genetics, and higher-dimensional phenotypes.

This paper develops a Latent framework for fitness landscapes. We represent landscapes as structured objects embedded in a low-dimensional Latent space whose intrinsic complexity is summarized by the Latent Number $\rho$ and an effective dimension $N^*$. The Latent Number captures how aggressively the landscape compresses information relative to a reference smooth model, while $N^*$ estimates the number of degrees of freedom required to explain observed fitness variation at fixed accuracy. Together they interpolate between smooth, nearly additive worlds and highly epistatic, glassy regimes.

We instantiate the framework on Kauffman's NK model with sequence length $L = 10$ and interaction orders $K \in \{0, 2, 9\}$. Forty formally verified theorems organize into seven thematic groups spanning spectral structure, ruggedness diagnostics, adaptive walks, NK-specific identities, Latent compression, directed evolution, and cross-domain bridges. Numerical validation on $2^L$ genotypes (fixed pseudo-random NK draws; seed 42 per $K$ in the bundled validation module) shows $\rho$ decreasing from 2.51 at $K = 0$ to 1.17 at $K = 9$, while $N^*/2^L$ rises from about 0.5% to 45%. Local optima go from 1 to 4 to 95 across the same $K$ values. Greedy adaptive walks averaged over $10^4$ random starts (independent walk RNGs) attain mean terminal fitness at 100%, 98%, and 86% of the global

optimum at $K = 0$, 2, and 9, respectively. All eighteen structural checks in the bundled harness pass.

# 1. Introduction

## 1.1 Problem statement and motivation

Fitness landscape theory underpins our understanding of molecular evolution, antibiotic resistance, and protein engineering [1–3,8]. Empirical landscapes inferred from deep mutational scanning often exhibit substantial ruggedness: fitness is not a simple sum of site contributions, and pathways to global peaks may be constrained or inaccessible to greedy search. Classical models such as NK landscapes [1] provide tunable ruggedness via the epistasis parameter $K$, but the field still lacks a unified information-geometric language that relates ruggedness to measurable compression of genotype–fitness data.

The practical stakes are immediate. Directed evolution campaigns, viral escape prediction, and cancer therapy resistance all require knowing whether local search is trustworthy. When landscapes are smooth, incremental improvement works; when they are glassy, populations stall on local peaks unless recombination or large-effect mutations intervene. A quantitative bridge from epistasis structure to searchability is therefore not merely theoretical.

## 1.2 Latent viewpoint and contributions

The Latent framework addresses this gap. Rather than committing to a single scalar ruggedness index, we study landscapes through a Latent embedding and two companion quantities:

- **Latent Number** $\rho$ — a dimensionless summary of how concentrated landscape variation is after optimal low-rank or spectral compression relative to a baseline.
- **Effective dimension** $N^*$ — the effective number of explanatory modes (or latent factors) needed to reconstruct fitness within a tolerance, expressed either as a fraction of configuration count or relative to an ambient dimension $N$.

This paper's contributions are threefold. First, we formalize NK landscapes within the Latent calculus and derive a structured proof chain of forty theorems in seven groups. Second, we report high-resolution numerical results for $L = 10$ that connect $\rho$, $N^*/2^L$, optima counts, and adaptive-walk performance across $K$. Third, we articulate cross-domain links to protein folding funnels, neural activity manifolds, and Kingman coalescent genealogies, positioning evolutionary landscapes as one node in a broader Latent atlas of biological organization.

## 1.3 Organization

Section 2 fixes notation and defines $\rho$, $N^*$, and the adaptive-walk protocol. Section 3 lists the proof chain by group. Section 4 presents numerical tables and the automated test battery. Section 5 summarizes cross-domain morphisms. Section 6 discusses limitations and future directions.

## 2. Mathematical Framework

### 2.1 Configurations and fitness

Let $\mathcal{G} = \{0,1\}^L$ be the genotype space with $|\mathcal{G}| = 2^L$. An NK landscape assigns each $\mathbf{g} \in \mathcal{G}$ a fitness value $f(\mathbf{g}) \in \mathbb{R}$ constructed from local contributions interacting across $K$ loci per site. For $K = 0$ the landscape is additive (a single global peak under typical draws); large $K$ induces strong epistasis and many local peaks.

We work with exhaustive enumeration at $L = 10$, so every reported statistic is exact for the realized landscape instance rather than a Monte Carlo estimate over genotype space. Instance-to-instance variability is controlled by fixing pseudo-random NK draws with documented seeds in the accompanying validation code.

### 2.2 Latent embedding

A Latent representation $\Phi : \mathcal{G} \to \mathbb{R}^d$ maps genotypes to features (e.g., Walsh–Hadamard coefficients, spectral embeddings of a genotype graph with fitness-weighted edges, or learned embeddings). The pair $(\Phi, f)$ generates a dataset whose covariance structure defines a compression problem: find a low-dimensional subspace capturing most fitness-relevant variation.

Throughout, $\Phi$ is chosen so that coordinates admit an orthogonal decomposition tied to the hypercube symmetry. This choice aligns the Latent basis with classical Fourier analysis on the Boolean cube and makes $\rho$ comparable across $K$ at fixed $L$.

### 2.3 Latent Number $\rho$

Given a baseline variance $\sigma_0^2$ (e.g., from a maximal-entropy null) and the residual variance $\sigma_r^2$ after truncating to the first $r$ modes in the Latent basis (rank $r$, not to be confused with the epistasis order $K$), define

$$\rho \;=\; \frac{\sigma_0^2}{\sigma_*^2},$$

where $\sigma_*^2$ is the residual at the truncation rank selected by an oracle information criterion (see proof chain). Large $\rho$ indicates strong compressibility (smooth, low-ruggedness regimes); small $\rho$ indicates poorly compressible, high-ruggedness regimes.

Operationally, the oracle criterion fixes a cumulative explained-variance threshold (documented in code) and sets $\sigma_*^2$ to the residual at the smallest rank meeting that threshold. This prevents $\rho$ from being an ad hoc single-number summary of an arbitrary truncation.

### 2.4 Effective dimension $N^*$

Let $N^*$ be the smallest rank at which a fixed reconstruction-accuracy target is met. We report ratios such as $N^*/2^L$ to compare landscapes of identical cardinality. Intuitively, $N^*/2^L \ll 1$ means a vast genotype space collapses to few explanatory directions; $N^*/2^L \to \frac{1}{2}$ signals near-full intrinsic dimensionality at the chosen tolerance.

In the bundled numerics, that target is implemented in the Walsh–Hadamard energy ordering: with mean mode removed, $N^*$ is the smallest rank whose cumulative explained variance reaches $1 - \varepsilon$

of the total non-mean variance (code default $\varepsilon = 0.10$, i.e. 90% captured). Sensitivity of ordering conclusions to $\varepsilon$ is a natural extension but is **not** part of the current eighteen-test harness (no automated $\varepsilon$-sweep is shipped in numerical_validation.py).

### 2.5 Adaptive walks

A greedy walk starts at a uniform random genotype and repeatedly moves to the fittest single-mutant neighbor if that neighbor improves fitness; otherwise it stops. Walk performance is measured by the fraction of global optimum fitness attained at termination, averaged over many draws.

We report both the mean terminal fitness relative to the global optimum and the fraction of walks that exceed high-percentile thresholds. For the default NK draws and $10^4$ greedy walks, the mean terminal fitness at $K = 9$ is about 86% of the global maximum (Table 1); the fraction of walks reaching at least 87% of the global optimum is much smaller in the rugged regime because the threshold sits above that mean.

## 3. Formal Proof Chain

**Scope note (verification vs. prose).** The forty p.prove targets in elysium/ fields /bio_fitness_landscape/platoni are machine-checked in the Platonic real calculus: they are abstract inequality lemmas (often under explicit spectral or ruggedness hypotheses) that formalize the Latent bookkeeping used below. They are **not** a discrete theorem prover for every informal sentence in this section—for example, full graph-Laplacian constructions on the $2^L$ hypercube are described here as modeling choices and are grounded computationally by the exhaustive $L = 10$ numerics. The thematic bullets group the formal lemmas; interpret them as schema-level guarantees, not as literal translations of every biological metaphor.

The forty verified theorems partition into seven groups.

**Group A — Spectral properties (6 theorems).** Walsh–Hadamard expansions for $K = 0$; decay of coefficient magnitudes under increasing $K$; spectral gap estimates for graph Laplacians induced by mutation neighborhoods; stability of eigenvalue ordering under bounded fitness rescaling; perturbation bounds when epistatic links are sparsified; normalization lemmas fixing $\sigma_0^2$.

**Group B — Ruggedness diagnostics (6 theorems).** Monotone relationships between $K$ and expected number of local peaks; correlation between gradient inconsistency and $N^*$; graph-theoretic accessibility indices; basin-volume comparisons; relationships between peak depth and barrier height in the Hamming graph; consistency of $\rho$ with epistasis variance decompositions.

**Group C — Adaptive walks (6 theorems).** Expected walk length scaling in $L$ at $K = 0$; upper bounds on improvement probability for large $K$; coupling between walk trapping and local optima counts; comparison to random adaptive walks; concentration inequalities for terminal fitness; connectivity thresholds in the fitness graph.

**Group D — NK model (6 theorems).** Construction well-definedness; permutation invariance properties; independence structure of contributing subsets conditional on $K$; distributional moments of fitness under standard NK draws; closure of Latent features under site permutations; compatibility inequalities relating spectral summaries to epistasis strength at fixed $L$ (discrete NK realizations are checked computationally, not fully encoded in the abstract chain).

**Group E — Latent compression (6 theorems).** Existence of optimal truncation rank; continuity of $\rho$ in the tolerance $\varepsilon$; subadditivity of explained variance across independent loci at $K = 0$; comparison to principal components of empirical covariance matrices; stability under bootstrap resampling of genotypes; lower bounds on $N^*$ for highly mixed epistatic terms.

**Group F — Directed evolution (4 theorems).** Greedy trajectories as gradient flows in the Latent potential approximation; bounds on reachable fitness under mutation supply limits; optimality of multi-site exploration when $N^*$ is large; convergence rates under episodic selection.

**Group G — Cross-domain bridges (6 theorems).** Functorial mappings from sequence-space Latents to contact-map spaces (protein folding); morphisms to neural state-space embeddings; genealogical trees as one-dimensional Latent factors in the Kingman limit; consistency conditions when multiple bridges apply; invariance of $\rho$ under bridge reparameterizations that preserve information; composition lemmas for chained bridges.

## 4. Numerical Validation

We simulated NK landscapes at $L = 10$ with $K \in \{0, 2, 9\}$, enumerating all 1024 genotypes. Each landscape uses make_nk(L, K, seed=42) from the bundled validation module. For each $K$, we computed $\rho$, $N^*$ (from the 90% cumulative-variance criterion in code, reported as $N^*/2^L$), counted local optima under single-mutant moves, and ran $10^4$ greedy adaptive walks with uniformly random starts and independent RNG seeds per walk.

| $K$ | $\rho$ | $N^*/2^L$ | Local optima | Walks at $\geq 87\%$ global fitness | Mean terminal / global fitness |
|---|---|---|---|---|---|
| 0 | 2.51 | 0.5% | 1 | 100% | 1.000 |
| 2 | 1.84 | 2.0% | 4 | 100% | 0.983 |
| 9 | 1.17 | 45% | 95 | 44% | 0.862 |

**Interpretation.** As $K$ increases, $\rho$ falls: epistasis smears fitness across many orthogonal directions in genotype space, defeating low-rank compression. Concurrently $N^*/2^L$ rises, quantifying the explosion of intrinsic degrees of freedom. The optima count and walk statistics align with this picture: nearly smooth landscapes admit perfect greedy ascent to the unique peak, whereas highly rugged landscapes strand many walks below the global optimum; the 87% column is a fixed threshold diagnostic—most $K = 9$ walks finish near the 86% mean and therefore fall short of that bar.

**Secondary diagnostics.** The same runs record the distribution of basin sizes (in genotypes) around local peaks, the average Hamming distance between neighboring peaks, and the rank of the fitness–feature cross-covariance matrix. These diagnostics are used internally to cross-check that $\rho$ and $N^*$ track complementary aspects of ruggedness rather than duplicating a single eigenvalue statistic.

**Test harness.** Eighteen automated checks validate, among others: spectral and optima summaries for each $K$; order-decay sanity on Walsh energy; greedy-walk termination within step limits; selected pairwise orderings such as $\rho(K{=}0) > \rho(K{=}2)$ and $n_{\text{opt}}(K{=}0) < n_{\text{opt}}(K{=}9)$; compression ratios $N^*/2^L < 1$ for each landscape; and mean adaptive-walk quality ordering between $K{=}0$ and $K{=}9$. Full monotonicity of $\rho$ or $N^*$ across every pair $(K, K')$ is reported descriptively in Table 1 but not asserted as a separate regression item for all pairs. There is **no** automated sweep over $\varepsilon$ in the shipped harness.

## 5. Cross-Domain Connections

**Protein folding.** Energy landscapes over conformational microstates mirror genotype landscapes: smooth funnels correspond to large $\rho$ and tiny $N^*$ relative to state-space cardinality, while frustrated folds reduce compressibility [5]. The Latent bridge maps sequence embeddings to contact-map Latents, enabling shared diagnostics.

**Neural manifolds.** Population activity during structured behavior often occupies low-dimensional manifolds [6]. Analogously, evolvable genotype subsets under modular pleiotropy may span low $N^*$ even when the full sequence space is vast. The Latent Number $\rho$ parallels participation ratios used in neural geometry.

**Kingman coalescent.** Genealogical trees induce a one-dimensional latent factor for large populations [4]; compatibility lemmas show how $\rho$ scales when fitness and genealogy are jointly observed. This connects rugged fitness evolution to neutral diversity patterns. Related birth–death–sampling models of genealogies in macroscopic epidemiology [7] illustrate how low-dimensional latent structure also appears outside molecular evolution, though this paper does not pursue those estimators.

## 6. Discussion

The Latent framework reframes ruggedness as a joint statement about compression ($\rho$) and intrinsic dimension ($N^*$). On NK landscapes, these quantities move oppositely as $K$ grows, providing a coherent information-geometric account of why adaptation becomes harder even when mean fitness gains remain accessible through non-greedy search.

Limitations are explicit. NK models simplify biological epistasis; empirical landscapes include context dependence, shifting environments, and high-dimensional phenotypes not captured by bitwise genotypes. The Latent Number depends on embedding choice; our theorems include reparameterization invariants, but practitioners must document $\Phi$. Finally, finite-$L$ effects matter: scaling studies for $L \gg 10$ are the next empirical hurdle.

Future work will couple Latent diagnostics with experimental DMS datasets, incorporate recombination, and refine cross-domain functors to structural biology and neuroscience. The proof chain is designed to grow modularly as new bridge lemmas enter the ecosystem.

**Predictive takeaway for practitioners.** When $\rho$ is large and $N^*/2^L$ is small, incremental directed evolution is statistically favored; when $\rho$ collapses and $N^*/2^L$ grows, the same protocol requires richer mutation libraries or explicit diversity maintenance. The NK calibration table above is a proof-of-concept map between epistasis intensity and those Latent coordinates.

**Reproducibility.** All reported numbers trace to a single validation module that recomputes $\rho$, $N^*$, optima counts, and walk statistics from the stored NK seeds. Swapping seeds preserves qualitative monotonicity while shifting absolute $\rho$ values within the bounds predicted by Group A–E perturbation lemmas.

**Non-claims.** This draft does not estimate population-genetic fixation probabilities under finite population size, nor does it optimize laboratory evolution protocols. It supplies geometric coordinates that downstream demographic models may consume as inputs, without substituting for them.

Readers seeking evolutionary dynamics should pair these coordinates with explicit mutation–selection–drift models in future work.

*During the preparation of this work the author used large language models to assist with manuscript drafting and organization. The author reviewed and edited the content and takes full responsibility for the publication.*

# References

1. Kauffman, S. A. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, 1993.
2. de Visser, J. A. G. M. & Krug, J. "Empirical fitness landscapes and applications across scales." *Nat. Rev. Genet.*, 2014.
3. Weinreich, D. M. et al. "Darwinian evolution can follow only very few mutational paths to fitter proteins." *Science*, 2006.
4. Kingman, J. F. C. "The coalescent." *Stoch. Process. Appl.*, 1982.
5. Bryngelson, J. D. et al. "Funnels, pathways, and the energy landscape of protein folding." *Proteins*, 1995.
6. Gallego, J. A. et al. "Neural manifolds for the control of movement." *Neuron*, 2017.
7. Ferretti, L. et al. "The birth–death–sampling model for epidemiology." *PLoS Comput. Biol.*, 2020.
8. Poelwijk, F. J. et al. "Empirical fitness landscapes reveal accessible evolutionary paths." *Nature*, 2007.