UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea Magistrale in Biologia Evoluzionistica

TESI DI LAUREA

Genetic Drift and Structural Constraints: The Determinants of Compensatory Mutations in Mammalian Proteins

Relatore
Prof. Alessandro Grapputo
Dipartimento di Biologia
Correlatore
Prof. Dr. Julien Y. Dutheil
Max Planck Institute for Evolutionary Biology

Laureando: *Lorenzo Penone*

ANNO ACCADEMICO 2024/2025

Riassunto

In presenza di continue mutazioni, le proteine devono preservare la loro integrità strutturale per funzionare correttamente. Un meccanismo fondamentale per il mantenimento della funzionalità proteica è rappresentato dalle mutazioni compensatorie. Queste costituiscono una forma di coevoluzione intra-proteica, in cui una mutazione secondaria in un sito distinto ripristina la fitness ridotta da una precedente mutazione deleteria. Sebbene le basi biofisiche di tale processo siano note, le dinamiche di genetica di popolazioni che ne determinano la prevalenza nell'albero della vita rimangono ancora poco comprese.

Il presente studio affronta questa lacuna testando l'ipotesi centrale che la dimensione effettiva della popolazione (N_e) sia un determinante primario del tasso di evoluzione compensatoria. Questo processo si colloca al centro di un apparente paradosso teorico. Esso richiede, infatti, due eventi con dipendenze opposte dalla N_e: (i) la fissazione iniziale di una mutazione leggermente deleteria, un evento più probabile in piccole popolazioni dove domina la deriva genetica; (ii) la successiva fissazione di una mutazione compensatoria vantaggiosa, un processo più efficiente in grandi popolazioni dove la selezione naturale è più efficace.

Per verificare questa ipotesi, è stato sviluppato un complesso quadro analitico che integra dati genomici, filogenetici e strutturali relativi a 78 specie di mammiferi. In primo luogo, è stato validato l'uso del rapporto tra sostituzioni non-sinonime e sinonime (d_N/d_S) specifico per ramo filogenetico come proxy della N_e, confermandone la correlazione sia con stime dirette di N_e sia con la massa corporea. Successivamente, sono stati identificati gli eventi compensatori per tre diverse proprietà biochimiche (polarità, volume e carica) mediante il programma CoMap. Per localizzare tali eventi su rami specifici dell'albero filogenetico, è stato inoltre sviluppato un nuovo test di bootstrap parametrico, atto a superare i limiti dei test globali. Infine, questi dati sono stati integrati con strutture proteiche ad alta risoluzione, sia sperimentali (PDB) sia predette (AlphaFold), al fine di fornire un contesto biofisico.

I nostri risultati stabiliscono una correlazione forte, positiva e statisticamente significativa tra d_N/d_S e l'accumulo di mutazioni compensatorie. Questa scoperta, supportata da analisi di controllo che impiegano stime dirette di N_e e dati di massa corporea, indica che le linee evolutive caratterizzate da dimensioni effettive della popolazione più piccole (e quindi soggette a una maggiore deriva genetica) accumulano un numero significativamente maggiore di mutazioni compensatorie.

L'analisi strutturale ha confermato l'effetto della N_e quale predittore primario del tasso di mutazioni compensatorie. Ha inoltre rivelato che il nucleo strutturale

compatto (il core) delle proteine è maggiormente interessato da questi eventi. Un secondo modello, costruito a livello di gruppi coevolutivi, ha ulteriormente analizzato le proprietà di questi residui—che si trovano tipicamente in prossimità spaziale (mediana 16-20Å) ma non in contatto fisico diretto—rivelando interazioni significative tra l'accessibilità media al solvente, la distanza interna al gruppo e l'effetto della deriva genetica.

Questo studio fornisce la prima evidenza genomica su larga scala del ruolo della deriva genetica come motore chiave dell'evoluzione compensatoria. I nostri risultati suggeriscono fortemente che la fissazione iniziale della mutazione leggermente deleteria, facilitata dalla deriva, costituisca il passo limitante dell'intero processo.

Questo lavoro offre un plausibile meccanismo molecolare per la componente di "valley crossing" della Teoria della Shifting Balance di Sewall Wright. Fornisce inoltre una base di genetica delle popolazioni fondamentale per la comprensione dell'evoluzione della resistenza antimicrobica e dei segnali coevolutivi che sono alla base della moderna predizione della struttura proteica.

Abstract

Proteins must maintain function integrity in the face of relentless mutational challenges. One key mechanism is compensatory mutation, a form of intra-protein coevolution where a subsequent mutation at a different site restores fitness lost to a prior deleterious mutation. While the biophysical basis of this process is well-studied, the population genetic forces that dictate its prevalence across the tree of life remain poorly understood. This thesis confronts this gap by testing the central hypothesis that effective population size (N_e) is a primary determinant of the rate of compensatory mutation, a process with conflicting theoretical dependencies on population size.

To test this hypothesis, we developed a comprehensive analytical framework integrating genomic, phylogenetic, and structural data from 78 mammalian species. We first validated the branch-specific ratio of non-synonymous to synonymous substitutions (d_N/d_S) as a robust proxy for N_e. Compensatory events for three key biochemical properties (polarity, volume, and charge) were then identified using CoMap. A novel branch-specific parametric bootstrap test was developed to map these events to specific lineages, allowing for a direct correlation with our N_e proxy. Finally, we integrated these findings with high-resolution experimental and predicted protein structures to provide a mechanistic, biophysical context.

Our results establish a strong, positive, and highly significant correlation between d_N/d_S and the accumulation of compensatory mutations. This finding—supported by control analyses using direct N_e estimates and body mass data—indicates that lineages with smaller effective population sizes accumulate significantly more compensatory mutations. Structural analysis confirmed that this N_e effect is a primary predictor of compensation and further revealed that the compact structural core of proteins is a hotspot for such events.

This study provides the first large-scale, genomic evidence that genetic drift is a key driver of compensatory evolution. Our findings strongly suggest that the initial fixation of a slightly deleterious mutation, a process facilitated by drift, is the rate-limiting step. This work offers a plausible molecular mechanism for the "valley crossing" component of Sewall Wright's Shifting Balance Theory and provides a fundamental population-genetic basis for understanding antimicrobial resistance and protein structure prediction.

Acknowledgements

First and foremost, I wish to extend my sincere gratitude to my supervisor, Prof. Julien Y. Dutheil, for his constant guidance and invaluable support throughout the entire development of this work.

My gratitude also extends to my university advisors, Prof. Alessandro Grapputo and Prof. Gil Guastoni Rosenthal, for their precious advice and constructive criticism, which have significantly enriched this thesis.

Special thanks go to Raşit Durak for his valuable collaboration in supervising the project.

I would also like to express my appreciation to the Erasmus+ programme, which offered me the extraordinary opportunity to carry out this research activity abroad, and to the Max Planck Institute for the warm hospitality with which they welcomed me.

A special thanks to all the other members of the Molecular Systems Evolution group (Alicem Susam, Chongyang Wang, Fernanda Fontes Trancoso, Dr. Jinyang Liang, and Moya Lieffering) and to all the colleagues at the institute with whom I shared a fantastic experience.

A special thought goes to my girlfriend, Francesca, for being a fundamental presence, and for always supporting and encouraging me.

To my sister, Sara, for her constant presence and support.

Finally, I want to dedicate the most important and heartfelt acknowledgement to my parents. Every step I have taken and every goal I have achieved in these years rests on their sacrifices. They believed in me and supported me with love and hard work, allowing me to get here and complete my studies. It is only thanks to their unconditional support, which never wavered, that I have had the privilege of completing this academic journey.

Ringraziamenti

Desidero in primo luogo rivolgere un sentito ringraziamento al mio supervisor, il Prof. Julien Y. Dutheil, per la sua guida costante e il prezioso supporto fornito durante l'intero svolgimento di questo lavoro.

La mia gratitudine si estende ai miei relatori universitari, il Prof. Alessandro Grapputo e il Prof. Gil Gustoni Rosenthal, per i loro preziosi consigli e le loro critiche costruttive, che hanno significativamente arricchito questa tesi.

Un ringraziamento particolare va anche a Raşit Durak, per la sua preziosa collaborazione nella supervisione del progetto.

Desidero inoltre esprimere la mia riconoscenza al programma Erasmus+, che mi ha offerto la straordinaria opportunità di svolgere questa attività di ricerca all'estero, e all'Istituto Max Planck per la calorosa ospitalità con cui mi ha accolto.

Un ringraziamento speciale a tutti gli altri membri del gruppo Molecular Systems Evolution (Alicem Susam, Chongyang Wang, Fernanda Fontes Trancoso, Dr. Jinyang Liang e Moya Lieffering) e a tutti i colleghi dell'istituto con cui ho condiviso una fantastica esperienza.

Un pensiero speciale va alla mia ragazza, Francesca, per essere stata una presenza fondamentale, per avermi sempre supportato e incoraggiato.

A mia sorella, Sara, per la sua costante presenza e il suo sostegno.

Infine, voglio dedicare il ringraziamento più importante e sentito ai miei genitori. Ogni passo che ho fatto e ogni obiettivo raggiunto in questi anni poggia sui loro sacrifici. Hanno creduto in me e mi hanno sostenuto con amore e fatica, permettendomi di arrivare fin qui e di completare i miei studi. È solo grazie al loro sostegno incondizionato, che non è mai venuto meno, che ho avuto il privilegio di portare a termine questo percorso accademico.

Contents

Abstract **iii**

Acknowledgements **v**

1 Introduction **1**
- 1.1 The Principle of Coevolution from Organisms to Molecules 1
- 1.2 A Theoretical Impasse and a Molecular Solution 3
- 1.3 The Broader Importance of Compensatory Evolution 4
- 1.4 Hypothesis and Objectives 5
- 1.5 A Proxy for The Effective Population Size 6
- 1.6 A Framework for Detecting Compensatory Events 9

2 Methods **13**
- 2.1 Data Curation and Preparation 13
- 2.2 Phylogenetic and Evolutionary Rate Estimations 16
- 2.3 Detection and Mapping of Compensatory Mutations 20
- 2.4 Correlative Analyses: Linking Coevolution to Population Size and Structure 24

3 Results **31**
- 3.1 Validation of d_N/d_S as a Proxy for Effective Population Size . . . 31
- 3.2 Correlation Between Compensatory Mutations and d_N/d_S 32
- 3.3 Structural Determination of Drivers of Compensatory Mutations . 36

4 Discussion **45**
- 4.1 A Robust Framework for Inferring Coevolution 45
- 4.2 The Efficacy of Selection: Validating a Key Evolutionary Proxy . . 46
- 4.3 Genetic Drift as a Driver of Compensatory Evolution 46
- 4.4 A Mechanistic Link: How Structural Constraints Modulate Coevolution 47
- 4.5 Implications 48
- 4.6 Limitations 49

Supplementary Materials **59**
- A.1 Implementation and Data Availability 59
- A.2 Supplementary Tables 59
- A.3 Supplementary Figures 60

Introduction

The functional integrity of proteins is relentlessly challenged by mutations, many of which can destabilize structure and reduce fitness. A key evolutionary rescue mechanism is compensatory mutation: a form of intra-protein coevolution where a subsequent mutation at a different site restores stability and function. While the biophysical basis of this process is well-studied, the macro-evolutionary forces that dictate its prevalence across the tree of life remain poorly understood. This thesis confronts this gap by investigating the primary determinants of compensatory evolution in mammals, proposing that its rate is shaped by a fundamental interplay between population genetics and biophysical constraints.

This chapter builds the foundation for this argument. We will first trace the principle of coevolution from classic organismal examples down to the molecular level, establishing the theoretical context. We then explore how compensatory mutations provide a molecular solution to a long-standing theoretical impasse in population genetics. Finally, we will outline our central hypothesis—that genetic drift is a primary driver of compensatory mutations—and describe the analytical framework developed to test it using the d_N/d_S ratio as a proxy for effective population size (N_e).

1.1 The Principle of Coevolution from Organisms to Molecules

Coevolution, the process of reciprocal adaptation between interacting species, is a fundamental driver of biodiversity. A classic example is the mutualism between Morgan's sphinx moth (*Xanthopan morganii*) and Darwin's orchid (*Angraecum sesquipedale*), where a tightly coupled evolutionary response produced the moth's exceptionally long proboscis and the orchid's correspondingly deep nectar spur (Arditti et al., 2012). This principle, however, is not limited to mutualism; it also fuels the perpetual arms races of antagonistic relationships, from predator-prey systems to parasite-host conflicts (Van Valen, 1973).

This dynamic of reciprocal adaptation cascades across all biological scales, from the cellular-level conflict between a host's immune system and an invading virus to an even finer resolution: within individual molecules (Woolhouse et al., 2002). Correlated evolution has been documented between interacting nucleotides in tRNA

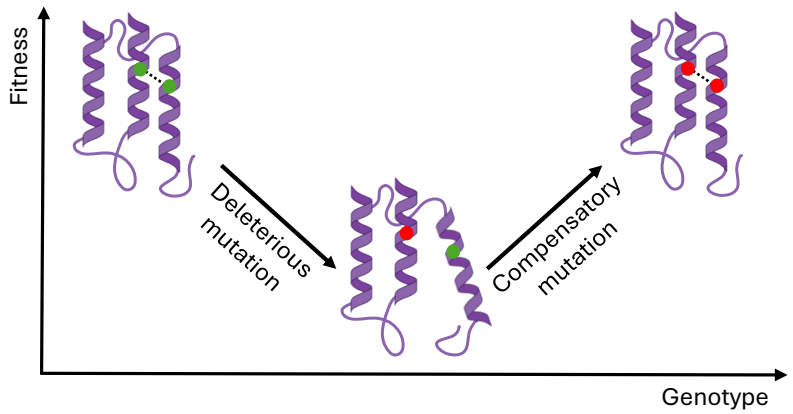

Figure 1.1. The molecular mechanism of intra-protein compensatory mutation. (Left) A stable, wild-type protein with a high-fitness conformation, maintained by a favorable intramolecular interaction (green residues). (Bottom) A first, deleterious mutation disrupts this stabilizing contact, causing a loss of function and fitness. (Right) A subsequent, compensatory mutation restores a favorable interaction (red residues), rescuing protein stability.

and, crucially, between amino acid residues within a single protein (Dutheil et al., 2010; Shindyalov et al., 1994). This latter phenomenon, known as intra-protein coevolution, arises to preserve a protein's structural and functional integrity. It is the central focus of this thesis.

1.1.1 The Mechanism of Compensatory Mutations

A primary mechanism driving intra-protein coevolution is compensatory mutations, a process involving a subsequent substitution that is positively selected to counteract the deleterious effects of a prior mutation (Davis et al., 2009). For instance, a first mutation might destabilize a protein's structure, but a second, compensatory mutation at another site can restore its stability, thereby partially or completely rescuing its function.

While this simple pairwise model provides a useful conceptual framework, the underlying dynamics are often more complex. Coevolutionary interactions can involve networks of multiple residues that collectively buffer a mutation's destabilizing effects (Chaurasia and Dutheil, 2022). Moreover, the temporal order of substitutions is critical. An initially neutral "permissive" mutation can create a structural context that accommodates a subsequent, otherwise deleterious, change. Fundamentally, all such interactions are a form of sign epistasis, where the fitness effect of a given mutation is contingent upon its genetic background.

1.2 A Theoretical Impasse and a Molecular Solution

1.2.1 The Shifting Balance Theory and the Problem of Valley Crossing

The concept of compensatory mutations, rooted in the biophysics of protein structure, provides a powerful molecular lens through which to examine one of population genetics' most influential and debated ideas: Sewall Wright's Shifting Balance Theory (SBT) (Coyne et al., 1997). Proposed in the 1930s, the SBT uses the metaphor of a "fitness landscape" to explain how a population might escape a local adaptive optimum to discover novel, potentially superior ones (Joshi, 1999). To cross an intervening fitness "valley," Wright proposed a three-phase process requiring a species to be subdivided into small, semi-isolated demes: first, genetic drift allows a small deme to fix a new combination of alleles, even if it is temporarily maladaptive; second, mass selection drives the deme up the slope of a new adaptive peak; and third, inter-demic gene flow allows this new, superior adaptation to spread throughout the species (Coyne et al., 1997; Joshi, 1999).

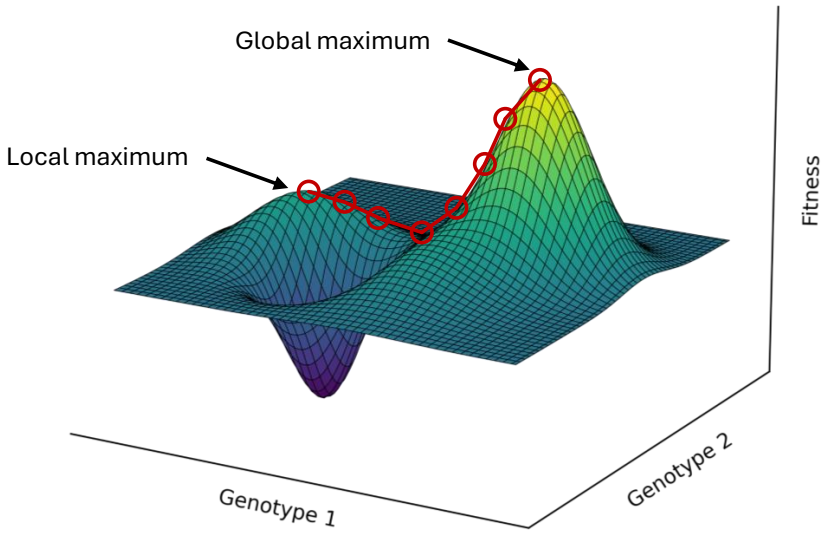

Figure 1.2. A visual representation of the "valley crossing" problem central to Sewall Wright's Shifting Balance Theory. The evolutionary trajectory (red path) shows a population descending from a local adaptive peak (left) and traversing a fitness valley—a process requiring genetic drift—to subsequently ascend a superior, global adaptive peak (right) via selection.

The theory's primary weakness, and a long-standing source of criticism, is the profound difficulty of its first phase. The sequential fixation of a deleterious

mutation followed by a compensatory one offers a plausible molecular mechanism for this valley-crossing event. In this scenario, a slightly deleterious mutation first fixes in a small deme due to genetic drift, temporarily lowering mean fitness. Subsequently, a second, compensatory mutation arises in an individual carrying the first. This new, doubly-mutant genotype is strongly favored by selection, allowing the population to reach a new adaptive peak and complete the valley crossing (Phillips, 1996; Stephan, 1996).

1.2.2 The Waiting Time Problem: A Central Dilemma

Despite this conceptual elegance, this molecular scenario faces a significant obstacle: the "waiting time problem". The time required for the necessary deleterious and then compensatory mutations to arise and fix can be extraordinarily long, posing a major barrier to the theory's realization. The problem lies in the conflicting demands of population size (Coyne et al., 1997; Phillips, 1996). In large populations, selection is highly efficient. The initial deleterious mutation is likely to be purged before a compensatory one has a chance to occur. In small populations, where drift allows the deleterious mutation to persist, the low mutation supply rate means the wait for the specific compensatory mutation to appear can be prohibitively long.

This dilemma leaves a critical gap in our understanding of how effective population size truly governs the rate of compensatory evolution—a gap this thesis aims to address directly.

1.3 The Broader Importance of Compensatory Evolution

1.3.1 Antimicrobial Resistance

Beyond its theoretical importance, compensatory evolution is a central actor in the urgent public health crisis of antimicrobial resistance (AMR). While resistance-conferring mutations allow a pathogen to survive drug treatment, they often alter essential cellular machinery (such as ribosomes) and impose a severe fitness cost in drug-free environments (Melnyk et al., 2015). Compensatory mutations provide a powerful evolutionary solution by restoring this lost fitness without sacrificing resistance. This process is critical for the stabilization and long-term persistence of resistant strains in pathogen populations and has been widely documented in *Escherichia coli*, *Salmonella typhimurium*, and HIV (Levin et al., 2000; Sander et al., 2002; Schulz zur Wiesch et al., 2010). A deeper understanding of its population genetic drivers is therefore essential for predicting and managing the evolution of

AMR.

1.3.2 Protein Structure Prediction

Furthermore, these same evolutionary forces leave an indelible signature on the genome: covariation. The process of compensatory evolution weaves a complex web of statistical dependencies between amino acid positions (Chaurasia and Dutheil, 2022). This coevolutionary signal is the predictive engine behind state-of-the-art tools like AlphaFold, which leverage this information from multiple sequence alignments to generate remarkably accurate predictions of protein 3D structure (Jumper et al., 2021; Mohan et al., 2022). The success of these methods has fundamentally transformed structural biology, underscoring that intra-protein coevolution is a key source of information for deciphering protein architecture and function (Burger and van Nimwegen, 2010).

1.4 Hypothesis and Objectives

Despite the centrality of compensatory mutations to evolutionary theory and applied biology, its underlying population genetic dynamics remain poorly understood. While a substantial body of theoretical and experimental work exists (Piskol and Stephan, 2011; Whitlock et al., 2003; Zhang and Watson, 2009), our comprehension is limited by a scarcity of large-scale, empirical research investigating its evolutionary drivers at a genomic scale.

The central hypothesis of this thesis is that the effective population size (N_e) is a primary determinant of the rate of accumulation of compensatory mutations. However, the direction of this effect is not obvious a priori. This uncertainty arises because compensatory evolution requires at least two events with opposing dependencies on population size (Davis et al., 2009; Sjödin et al., 2005): first, an initial deleterious mutation must spread within the population, a scenario more probable in small populations where genetic drift can overwhelm weak negative selection. Second, a subsequent advantageous compensatory mutation must arise and be efficiently selected, a process significantly more effective in large populations where selection is potent.

Accordingly, the main goal of this research is to provide the first large-scale, genomic evidence of a direct correlation between effective population size and the rate of compensatory mutations. By exploring the direction of this relationship, we aim to better understand the primary evolutionary forces driving this pattern. Finally, we will integrate protein structure data to provide a deeper, mechanistic context for our primary findings.

To rigorously test this central hypothesis, however, requires overcoming two significant methodological hurdles. First, we must establish a reliable proxy for effective population size that can be applied to ancestral branches in the phylogeny. Second, we must employ a statistical framework powerful enough to detect true compensatory events and distinguish their signal from the confounding noise of shared ancestry. The remainder of this introduction will therefore establish the theoretical and analytical foundations for the approaches we adopted to solve these two challenges, respectively.

1.5 A Proxy for The Effective Population Size

1.5.1 The Challenge of Estimating Ancestral Population Size

The central aim of this research is to connect the rate of compensatory mutations to population size, the fundamental parameter modulating the strength of genetic drift. In idealized theoretical frameworks, such as the Wright-Fisher model (Fisher, 1930; Wright, 1931), the magnitude of drift is a simple inverse function of the census population size (N). However, real populations rarely meet these ideal assumptions, exhibiting complex factors like fluctuating population sizes, overlapping generations, and non-random mating.

To bridge this gap between theory and reality, Sewall Wright introduced the concept of effective population size (N_e). N_e is formally defined as the size of an idealized Wright-Fisher population that would experience the same amount of genetic drift as the actual population under study (Charlesworth, 2009; Wright, 1931). A small N_e indicates that genetic drift is a powerful evolutionary force, capable of overwhelming weaker selective pressures.

N_e is almost always smaller than the census count N, often by orders of magnitude. This distinction is critical, as different analytical approaches may be proxies for one or the other. For instance, life-history traits like body mass are expected to correlate with the census size N (Damuth, 1981), while proxies derived from molecular data, such as the d_N/d_S ratio, are expected to reflect the long-term N_e that directly governs the fate of mutations (Figuet et al., 2016; Ohta, 1992).

Testing this relationship across deep time presents a major challenge: we require robust estimates of population size for the entire phylogeny, including ancestral lineages. Direct genomic methods for estimating historical N_e are limited to extant species at the tips of the tree (Wilder et al., 2023) and cannot provide estimates for these internal branches. Therefore, to test our central hypothesis across a broad evolutionary history, a reliable, branch-specific proxy for long-term effective population size is essential. This study focuses specifically on mammals, as this

clade offers the extensive genomic resources and well-characterized patterns of molecular evolution necessary to develop and validate such a proxy.

1.5.2 The Nearly Neutral Theory: A Theoretical Foundation

The theoretical foundation for such a proxy lies in the nearly neutral theory of molecular evolution (NNT), which builds upon the foundational principles of Motoo Kimura's neutral theory. Proposed in the late 1960s, Kimura's theory challenged the prevailing selectionist view by positing that the majority of genetic variation is selectively neutral, with its evolutionary fate governed primarily by random genetic drift rather than natural selection (Kimura, 1968, 1983). While this provided a powerful null model, it was soon refined by Tomoko Ohta's NNT. This crucial extension incorporated the large class of mutations that are not strictly neutral but instead have slight fitness effects. The NNT provides a more nuanced framework by explicitly connecting the efficacy of selection against these mutations to the effective population size (Ohta, 1973, 1992).

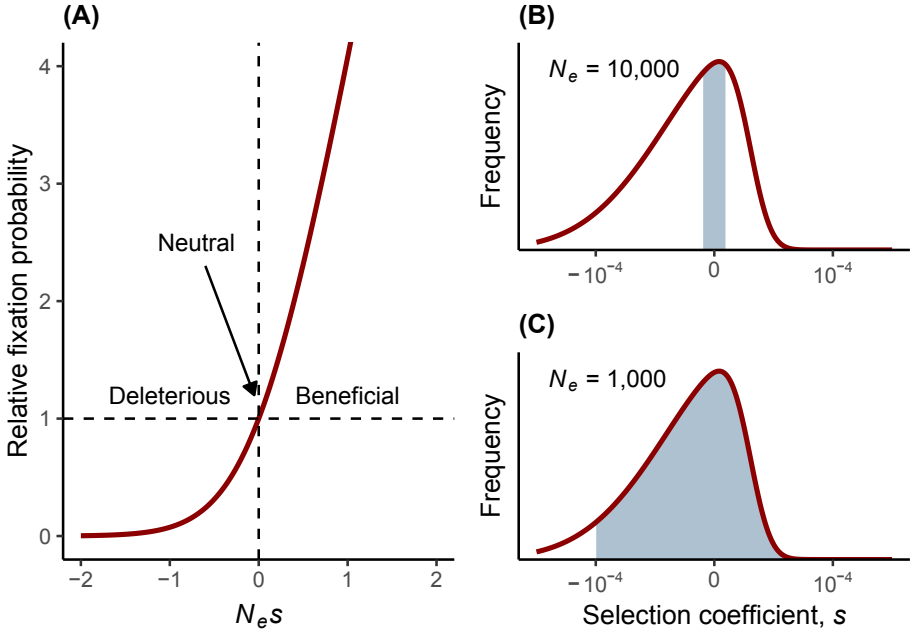

Figure 1.3. Principles of the Nearly Neutral Theory. (A) The fixation probability of a new mutation, relative to a neutral one (dashed line), as a function of the product of effective population size (N_e) and the selection coefficient (s). Selection is effective when $|N_e s| > 1$, whereas genetic drift dominates when $|N_e s| < 1$. (B) and (C) The same Distribution of Fitness Effects (DFE; red curve) shown for populations of different sizes. The shaded area represents the "effectively neutral" zone ($|s| \leq 1/(2N_e)$), where the fate of mutations is determined by drift rather than selection. (B) In a large population ($N_e = 10,000$), selection is highly efficient. The effectively neutral zone is narrow, and most slightly deleterious mutations are purged. (C) In a small population ($N_e = 1,000$), genetic drift is strong. The neutral zone is 10-fold wider, encompassing a broad range of mutations (including slightly deleterious ones) whose fixation is now governed by chance.

The NNT posits that the fate of a new mutation is determined by the interplay between its selection coefficient (s) and the power of genetic drift, which is proportional to ($1/N_e$). The theory assumes that most new non-synonymous mutations are slightly deleterious, with their impacts described by a distribution of fitness effects (DFE) (Eyre-Walker and Keightley, 2007; Ohta, 1973). A mutation is considered "effectively neutral" if the force of drift is greater than the force of selection. Based on the mathematics of fixation probability (Kimura, 1962), this occurs when the magnitude of the selection coefficient is less than the inverse of the effective population ($|s| < 1/N_e$). In this regime, the mutation is primarily governed by drift and can fix by chance. Conversely, if $|s| > 1/N_e$, selection becomes the dominant force, leading to its efficient removal (Ohta, 1992).

1.5.3 The d_N/d_S Ratio as a Measure of Selection Efficacy

To understand how this dynamic creates a proxy for N_e, we must consider two classes of mutations in protein-coding sequences. Synonymous mutations alter a codon but do not change the encoded amino acid. As they are often presumed to have minimal fitness effects, their fixation is thought to be driven primarily by drift, making their substitution rate (d_S) a baseline for neutral evolution (Ohta, 1995). In contrast, non-synonymous mutations change the encoded amino acid. Such alterations can affect a protein's structure and function and are therefore exposed to natural selection. While a small fraction may be advantageous, the majority are weakly deleterious (Eyre-Walker and Keightley, 2007).

The ratio of these two rates (d_N/d_S) thus provides a powerful measure of the efficacy of purifying selection (Yang and Bielawski, 2000). The NNT predicts a direct, negative correlation between N_e and the genome-wide average d_N/d_S. In species with a small long-term N_e, genetic drift is strong. A larger proportion of slightly deleterious mutations fall into the nearly neutral range ($|s| < 1/N_e$), allowing them to fix by chance and elevating the rate of non-synonymous substitution (d_N). In species with a large N_e, selection is potent; even weakly deleterious mutations are efficiently purged, suppressing the dN rate. The d_N/d_S ratio is therefore expected to be higher in species with smaller effective population sizes, making it a powerful proxy for long-term N_e (Brevet and Lartillot, 2021; Figuet et al., 2016).

1.5.4 Empirical Validation in Mammals

While this expected correlation assumes that the balance between selection and drift remains stable over evolutionary time, a substantial body of empirical research in mammals has provided strong support for the theory's predictions. Early studies established an inverse relationship between the efficacy of purifying selection and

life-history traits—known to correlate with small N_e—such as large body mass and long generation times (Figuet et al., 2016). More recent work has reinforced these findings by demonstrating a direct negative correlation between genome-wide d_N/d_S and explicit estimates of N_e, validating the utility of this proxy across diverse mammalian lineages (Brevet and Lartillot, 2021; James and Eyre-Walker, 2020).

1.6 A Framework for Detecting Compensatory Events

Detecting intra-protein coevolution requires robust methods that can distinguish true functional co-dependence from statistical noise. Early, tree-independent approaches like mutual information (MI) (Korber et al., 1993) assessed correlations directly from a sequence alignment but were fundamentally flawed. Their major weakness is a high rate of false positives caused by phylogenetic inertia—a strong correlation signal arising from shared ancestry rather than coevolution (Ashenberg and Laub, 2013; Dutheil, 2012).

To overcome this limitation, modern phylogeny-aware techniques were developed. These fall into two main categories: corrected statistical methods, which mathematically adjust metrics like MI to filter out phylogenetic noise (Achaz and Dutheil, 2021), and the more robust model-based phylogenetic methods (Achaz and Dutheil, 2021; Pollock et al., 1999; Yeang and Haussler, 2007). The latter explicitly incorporates the evolutionary tree to reconstruct ancestral states and identify co-substitutions—instances where mutations at different sites occur on the same branch. This approach effectively separates the signal of functional linkage from the noise of phylogeny, providing a powerful framework for detecting coevolution.

1.6.1 CoMap: A Model-Based Approach for Detecting Coevolution

The primary tool employed in this thesis is CoMap (cosubstitutions Mapping), a state-of-the-art program designed to identify groups of coevolving positions by analyzing sequence alignments within their phylogenetic context (Dutheil and Galtier, 2007; Dutheil et al., 2005). Its methodology is built on the principle that cosubstitutions must be studied within an evolutionary framework to distinguish functional correlation from the confounding signal of phylogenetic inertia (Dutheil, 2012).

CoMap introduces several key innovations to improve the detection of coevolutionary signals. The method utilizes Weighted Substitution Mapping, which

considers the biochemical properties of amino acids—such as polarity, volume, and charge—to weight different substitution events. Unlike many pairwise methods, CoMap employs a hierarchical clustering approach to detect coevolving groups of any size, a critical feature for understanding complex compensatory interactions. Furthermore, it distinguishes between different coevolutionary mechanisms through two distinct statistical tests: one based on a correlation coefficient for general co-substitutions, and another using a compensation index to pinpoint compensatory changes that maintain a global biochemical property (Dutheil and Galtier, 2007).

To ensure the reliability of its findings, CoMap implements a parametric bootstrap procedure for statistical significance testing. This involves simulating datasets under a null hypothesis of independent evolution to create a null distribution that is conditioned on site variability, thereby avoiding known analytical biases (Dutheil and Galtier, 2007; Dutheil et al., 2005).

1.6.2 Quantifying Compensation: The Compensation Index

The power of CoMap lies in its precise mathematical formulation for detecting compensatory changes. The core of this approach is the compensation index, which is calculated by modeling amino acid substitutions as vectors along each branch of the phylogeny.

The process begins with substitution mapping, where the estimated number of substitutions per site for every branch is stored in a substitution vector, V. This estimation uses an empirical Bayesian approach that incorporates branch length, amino acid substitution rates, and the reconstructed ancestral states at the tree's nodes. To account for the inherent uncertainty in ancestral state reconstruction, the substitution counts are averaged over all possible ancestral scenarios.

This vector can then be weighted according to a specific biochemical property of interest (e.g., charge), resulting in a weighted substitution vector, V^*. Each entry k in vector V_i^* represents the estimated amount of biochemical change at site i on branch k. Crucially, by assigning opposite weights to reciprocal changes (e.g., X $\rightarrow$ Y vs. Y $\rightarrow$ X), these signed, weighted vectors can track the direction of change. This allows for a direct test of the hypothesis that a global chemical property of a coevolving group of sites is conserved, even as the properties of individual sites change. From this, we can define the compensation index, C, for a pair of sites i and j:

$$C_{ij} = 1 - \frac{|\tilde{V}_i + \tilde{V}_j|}{|\tilde{V}_i| + |\tilde{V}_j|}, \tag{1.1}$$

where $|V|$ is the length of vectors, that is the L2-norm:

$$|\tilde{V}_i| = \sqrt{\sum_b v_{ib}^2}. \tag{1.2}$$

When C approaches 1, the numerator is close to zero, indicating that the changes are nearly perfectly opposite and symmetric. This is a strong signal of compensation. In contrast, if C approaches 0, the numerator and denominator are nearly equal, which occurs when the change vectors point in the same direction. This indicates parallel evolution, not compensation. This pairwise index can be generalized to assess compensation within a group of s sites by summing all vectors in the numerator and denominator:

$$C = 1 - \frac{\left|\sum_{i=1}^{s} \tilde{V}_i\right|}{\sum_{i=1}^{s} |\tilde{V}_i|}. \tag{1.3}$$

Ultimately, a high compensation index for a group of sites suggests that natural selection has acted to conserve a global biochemical property by coordinating substitutions among them, even as individual residues have mutated.

Methods

2.1 Data Curation and Preparation

To test the relationship between effective population size (N_e) and intra-protein compensatory mutation rate, we first assembled a curated database of mammalian orthologs. Since rigorous hypothesis testing requires both comprehensive taxonomic representation and data accuracy, our approach prioritized breadth of coverage while minimizing potential sources of error. This section describes the construction of our mammalian ortholog database, from the initial download of sequences to the filtering criteria we applied to ensure data quality and suitability for downstream analyses (Fig. 2.1).

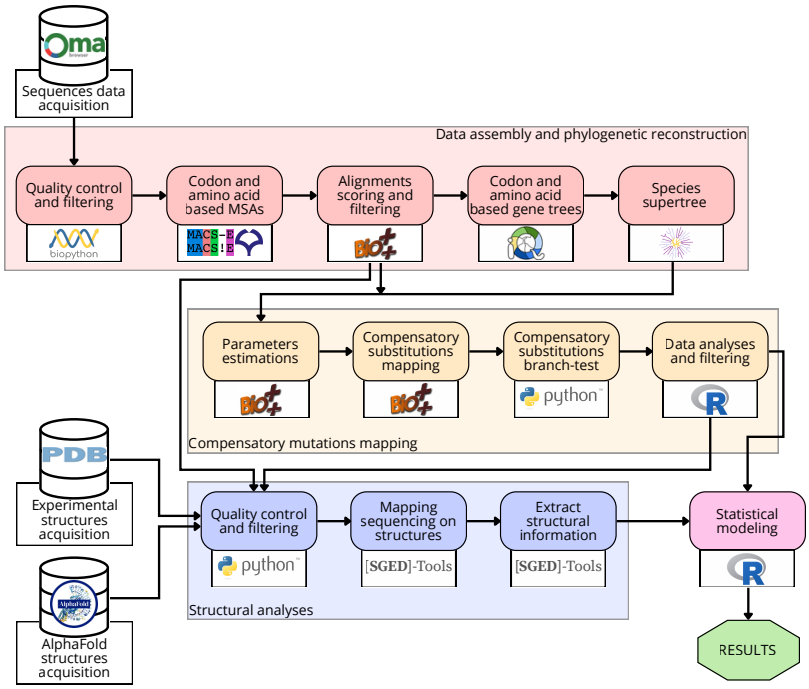

Figure 2.1. Overview of the bioinformatic pipeline. The workflow is divided into three main modules: (i) data assembly and phylogenetic reconstruction (red box), (ii) mapping of compensatory substitutions (yellow box), and (iii) structural analysis (blue box). Primary inputs are sequences from the OMA database and structures from PDB and AlphaFold.

2.1.1 Mammalian Ortholog Database Construction

We used the OMA (Orthologous MAtrix) database (Supplementary Materials Table A.1) to build a local database of mammalian orthologous genes. OMA is a database that provides orthology inferences across more than 2,900 fully sequenced genomes. In particular, we used OMA groups (OGs), sets of protein sequences that are inferred to be strictly orthologous (Altenhoff et al., 2023). The "All.Jul2024" release of the database contains 78 mammalian species.

We downloaded the complete set of protein and complementary DNA (cDNA) sequences from OMA and identified OGs corresponding to mammalian species. OGs to be investigated were selected based on the following criteria: (i) a maximum cDNA length of 4,500 nucleotides, to exclude genes with extreme length, which are often difficult to align accurately; (ii) an ambiguous-nucleotide (N) content of no more than 5 %; (iii) a minimum of 60 sequences per group, to have enough statistical power for coevolution analyses (Dutheil, 2012); (iv) the presence of at least one species from the order Monotremata to serve as a consistent outgroup, thereby ensuring all inferred phylogenies could be reliably rooted for subsequent topological comparisons. Of the 110,056 OGs with mammalian sequences present in the database, 9,467 matched our criteria. Compliance with these thresholds was verified after each filtering step.

To exclude genes that do not translate properly (e.g., potentially misannotated genes) we translated cDNA sequences into amino acid sequences using the Python package `Biopython` (v1.85) (Cock et al., 2009) with the standard genetic code. Only sequences whose translations matched the corresponding protein sequences were retained.

To further support the sequence information with taxonomic data, we obtained the full taxonomy of the included species from the NCBI Taxonomy database (Schoch et al., 2020) (Supplementary Materials Table A.1).

2.1.2 Sequence Alignment and Quality Control

To reduce potential methodological biases in downstream analyses, we generated multiple sequence alignments (MSAs) using two distinct algorithms and later combined their outputs.

Codon-based Alignment

Codon-level MSAs of cDNA sequences were generated using MACSE (Multiple Alignment of Coding SEquences) (v2.07), a codon-aware multiple sequence aligner that preserves reading frames and accounts for frameshifts, thereby ensuring accurate alignment of coding DNA, even in error-prone datasets (Ranwez et al., 2018).

Alignments were performed using the alignSequences module with default parameters, followed by refinement using the refineAlignment module, also with default settings, to improve the alignment quality—following the authors' recommendation.

MACSE outputs two parallel MSAs: one aligning the nucleotide sequences and another with their translated amino acid sequences, enabling comparative analyses at both the DNA and protein levels. We excluded from the dataset all sequences in which frameshifts were detected—i.e., insertions or deletions that shift the triplet reading frame—as the evolutionary models used do not account for frameshift events, which may produce noise in downstream analyses (Yang, 2006). A total of 221 OGs were excluded, resulting in 9,256 OGs.

For the next alignment stage, we used MACSE's protein alignment output, removed all gap characters to obtain dealigned (gap-stripped) sequences, and supplied these to the following alignment tool. This gap-stripping step also minimized DNA-protein inconsistencies present in some OMA entries, which often appear as stretches of unknown residues (X) at sequence termini and would otherwise cause mismatches.

Amino Acid-based Alignment

In addition to the codon-based alignments, amino acid-level MSAs were generated with MAFFT (v7.526), which utilizes fast Fourier transform for rapid pairwise scoring, using the `--auto` preset (Katoh and Standley, 2013).

Alignment Scoring and Filtering

We filtered MSAs using the Column Score (CS) as defined by (Thompson et al., 1999), which compares a test alignment with a reference one and assigns 1 to each column that matches and 0 otherwise; the final CS is therefore the proportion of matching columns (Thompson et al., 1999).

We computed CS with BppAlnScore, part of the Bio++ Program Suite (v2.4.1), a suite of programs for phylogenetic and sequence analysis (Guéguen et al., 2013). We treated the MACSE MSAs as the test and the MAFFT MSAs as the reference. Columns with CS of 0 or with more than 10 % gaps were trimmed from the test alignment using BppSeqMan, also part of the same suite. We chose a 10% gap threshold as a standard compromise between aggressively removing poorly aligned regions and retaining a sufficient number of informative sites. Trimmed positions were then mapped back onto MACSE's codon-level alignment via custom Python scripts, producing a nucleotide-level MSA consistent with the filtered protein sequences. Stop codons were removed, and codon-protein consistency was verified.

The final output consists of two filtered alignment sets: one at the amino acid-level and one at the nucleotide-level, in an attempt to compare them and ensure

the accuracy of the final MSAs.

2.2 Phylogenetic and Evolutionary Rate Estimations

With a curated set of orthologous genes, we first reconstructed the phylogenetic relationships among all species and next we estimated branch-specific rates of protein evolution (d_N/d_S), in order to generate our proxy for (N_e).

2.2.1 Phylogenetic Tree Reconstruction

We estimated two different phylogenies from both codon-level and amino acid-level alignment sets, with the aim of comparing their topologies and assessing consistency.

Gene Trees Inference

We estimated the maximum likelihood (ML) tree for each OG with IQ-TREE (v2.4.0), which uses a fast and effective stochastic algorithm to reconstruct trees by ML from MSAs (Minh et al., 2020). For protein MSAs we applied the Le and Gascuel's four-matrix substitution model with a four-class Gamma rate distribution (Le et al., 2012). We set 100 bootstrap replicates, a number that provides reasonable estimates of node support while remaining computationally feasible for a dataset of this scale. For codon-level MSAs, we used the empirical codon model (ECMK07) from Kosiol et al. (2007), which estimates substitution rates based on observed data rather than relying on a predefined mathematical model. In this case, performing bootstrap replicates was not feasible due to the dimensionality of the substitution matrix and the consequently high computational time required, given the size of the dataset.

The workflow produced two sets of 9,256 unrooted gene trees: one based on protein MSAs and one based on codon MSAs.

Species Trees Inference

To accommodate our large dataset and account for potential gene tree heterogeneity, we inferred a species phylogeny using a supertree approach. This method integrates multiple individual gene trees to produce a species tree. We inferred a supertree for each of the two gene-tree sets. By combining many smaller phylogenies into a single, comprehensive tree, supertrees provide broader taxonomic coverage and improved topological accuracy (Bininda-Emonds, 2004).

Supertrees were estimated with the Accurate Species Tree EstimatoR (ASTER), a family of optimization algorithms for species tree inference. In particular, we adopted the Accurate Species Tree ALgorithm (ASTRAL) algorithm, an approach that estimates an unrooted species tree given a set of gene trees, which is consistent under the multi-species coalescent model and accounts for incomplete lineage sorting. ASTRAL identifies the species tree that maximizes the number of quartet topologies shared with the input gene trees, restricting the search to tripartitions present in at least one gene tree (Mirarab et al., 2014).

We obtained the codon-based supertree with the ASTRAL-IV (v1.23.4.6), a tool that implements ASTRAL algorithm in a scalable framework, suitable for large datasets and gene trees with missing species (Zhang et al., 2018).

To infer the amino acid-based supertree, we adopted Weighted ASTRAL (wASTRAL) (v1.23.3.7), which implements in the ASTRAL algorithm a down-weight for poorly supported quartets based on bootstrap values, to improve inference accuracy (Zhang and Mirarab, 2022).

We compared the two supertrees with a custom Python script that uses `ete3` (v3.1.3), a library to read and manipulate phylogenetic trees (Huerta-Cepas et al., 2016). Tree similarity was measured with the Robinson-Foulds (RF) distance, a simple metric that counts the number of branch rearrangements needed to convert one topology into the other (Robinson and Foulds, 1981). The resulting RF distance was 2, indicating a single conflicting node (Supplementary Materials Fig. A.2): the common ancestor of Scandentia, Primates, and Glires; a well-known point of uncertainty in mammalian phylogeny (Brady et al., 2024). We collapsed that node—merging the descendant lineages into a single trifurcation—and generated a final consensus supertree that best represents the relationships among the species in our dataset (Fig. 2.2).

2.2.2 Estimation of d_N/d_S ratio

To estimate the genome-wide d_N/d_S ratio across the whole phylogeny (including internal branches), we adopted a delete-d jackknife variant, a generalization of the classical leave-one-out jackknife (Wu, 1986).

The jackknife builds pseudo-datasets by systematically omitting observations from the original sample; the dispersion of the statistics across these replicates provides bias and variance estimates. In the delete-d variant, each replicate removes a fixed number d of data points—here, codon sites. This procedure constitutes the first-order (linear) approximation to the non-parametric bootstrap, recovering the leading terms of its bias and variance expansion at much lower computational cost (Efron and Tibshirani, 1993; Wu, 1986).

Using a custom Python script with the `Biopython` library, we concatenated

all codon alignments, filling any missing sequences with N nucleotides—ignored by the software. For each of 100 jackknife replicates, we randomly sampled 10,000 codon positions. This sampling depth was large enough to be representative of the full concatenated alignment. We then estimated d_N/d_S for every replicate with BppML from the Bio++ program suite (Guéguen et al., 2013), which fits the ratio in a ML framework using the non-homogeneous branch model derived from PAML (Yang, 2007), with one d_N/d_S parameter per branch. We used the Yang & Nielsen (YN98) codon substitution model (Yang et al., 1998). Of the 300 replicates initiated, 63 failed to converge due to computational errors, yet 237 replicates were sufficient to obtain stable variance estimates for the d_N/d_S parameter. To read the BppML output, we implemented a custom parser in Python. Finally, we averaged the branch-specific d_N/d_S estimates across the 237 replicates. Fig. 2.2 shows the final result plotted on the phylogenetic tree. This sampling procedure was necessary since the full concatenated alignment contains 13,481,062 sites, making both standard maximum-likelihood analysis and full bootstrap computations intractable.

BppML also returns branch-length estimates, which we used in downstream analyses.

2.2.3 Validation of d_N/d_S as a Proxy for N_e

To validate the use of the branch-specific ratio of non-synonymous to synonymous substitution rates (dN/dS) as a proxy for effective population size (N_e), we tested its correlation with two independent variables at the terminal branches of our phylogeny. First, we used published, genome-derived estimates of N_e. Second, we used body mass, a life-history trait known to correlate inversely with N_e in mammals (Damuth, 1981; White et al., 2007).

Estimates of N_e for 45 species were obtained from Wilder et al. (2023). In that study, the authors inferred historical N_e using the Pairwise Sequentially Markovian Coalescent (PSMC) model, rescaled the output with species-specific generation times and mutation rates, and calculated the harmonic mean of N_e over time, excluding the last 10,000 years. Body mass data for 77 species were compiled from the Animal Ageing and Longevity Database (AnAge) (De Magalhães and Costa, 2009), PanTHERIA (Jones et al., 2009), Animal Diversity Web (ADW, (Myers et al., 2025)), and PHYLACINE 1.2 (Faurby et al., 2018), with missing values supplemented from (Feeney, 2012) (Supplementary Materials Table A.1).

To account for the statistical non-independence of traits among related species, we performed all regression analyses using Phylogenetic Generalized Least Squares (PGLS) in R with the `nlme` package (v3.1-168). This approach models the covariance of residuals based on the phylogenetic tree, providing an unbiased estimate of the regression coefficients (Martins and Hansen, 1997). We specified the phylogenetic

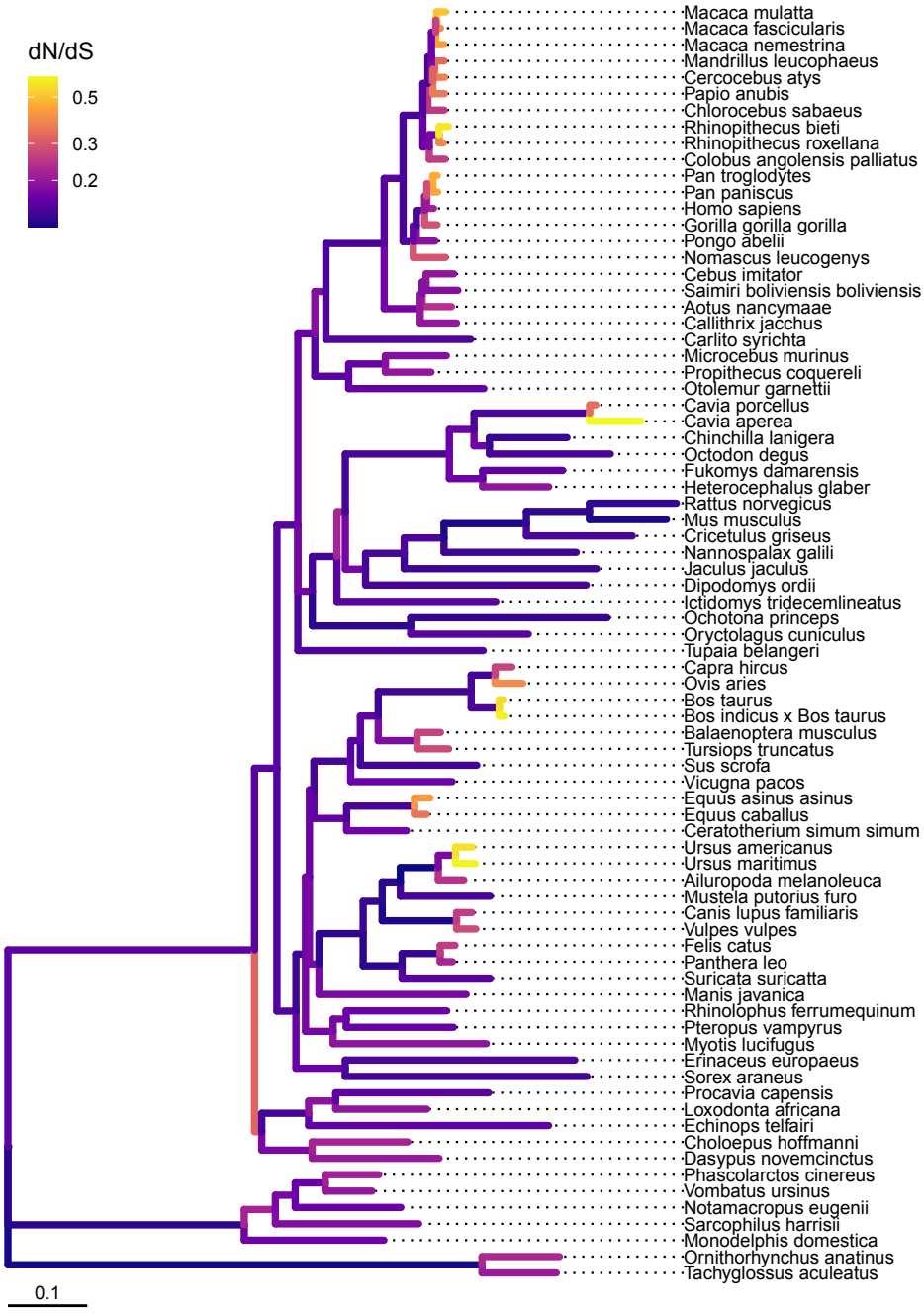

Figure 2.2. Mammalian supertree with estimated genome-wide d_N/d_S ratios. Branch lengths are proportional to substitutions per site (see scale bar). Branch color indicates the d_N/d_S ratio, as shown in the key.

correlation structure using the `corMartins` function from the ape package. To ensure the robustness of our findings, we also performed an additional analysis using the alternative correction proposed by Grafen (Grafen, 1989) (`corGrafen` function), which yielded qualitatively similar results. These analyses allowed us to test the relationships between d_N/d_S and N_e, d_N/d_S and body mass, and body mass and N_e while controlling for the confounding effect of shared ancestry. These models take the general form detailed in Equation (2.1):

$$\log(d_N/d_S) = \beta_0 + \beta_1 \cdot \log(Trait) + \epsilon, \tag{2.1}$$

where $Trait$ represents either N_e or Body Mass. The critical distinction from a standard linear model is that the error term ϵ is assumed to covary according to the shared evolutionary history, modeled as $\epsilon \sim \mathcal{N}(0, \sigma^2 V)$, where V is the phylogenetic covariance matrix derived from the species tree.

2.3 Detection and Mapping of Compensatory Mutations

In parallel with estimating branch-specific d_N/d_S ratios, we quantified our second key variable: the occurrence of compensatory mutations along each lineage on the phylogeny. Since coevolutionary methods detect correlated substitutions between amino acid sites, our approach proceeded in two steps. First, we identified statistically significant clusters of coevolving residues. Second, we examined the mutations within those clusters to pinpoint compensatory events. To achieve this, we implemented a pipeline that integrates coevolutionary signal detection with a novel branch-specific significance test. This test, which localizes compensatory mutations to individual phylogenetic branches, makes a direct comparison possible between the distribution of compensatory events and our branch-specific d_N/d_S estimates.

2.3.1 Identifying Coevolving Residues Clusters

Model Parameter Estimation

Before mapping compensatory mutations, we re-estimated the substitution model parameters for each alignment. This step was necessary to ensure methodological consistency, as the software used for phylogenetic reconstruction (IQ-TREE) and substitution mapping (CoMap) differ in their available models and likelihood calculation implementations. We therefore used BppML, which is based on the same Bio++ libraries as CoMap, to fit the Γ-shape parameter (α) under a specified model,

branch lengths and site-specific relative substitution rates (Guéguen et al., 2013).

We fitted the Le & Gascuel (LG08) (Le and Gascuel, 2008) amino acid substitution model with four discrete Γ rate categories ($k = 4$). Under this discrete-Γ model, rates across sites are assumed to follow a Γ distribution (mean = 1) that is approximated by k equal-probability categories; the shape parameter α is estimated from the data, and each site's likelihood is averaged over the k category mean rates.

We discarded OGs that met any of the following criteria: (i) $\alpha \leq 0.05$ or $\alpha \geq 3$; (ii) total tree length < 0.1; (iii) at least one branch > 1. Values of α or branch lengths outside this range indicate unreliable parameter estimates or highly conserved sequences—extreme α values suggest over- or under-dispersion; unusually long/short branches often reflect misalignment or insufficient evolutionary signal. At this stage, we excluded 417 OGs and retained 8,839.

Probabilistic Substitution Mapping with CoMap

For substitution mapping, we performed one run for each biochemical property (polarity, volume, and charge), resulting in three runs per analysis.

Substitution mapping and the detection of compensatory mutations were carried out using CoMap (v1.5.5). CoMap computes probabilistic substitution maps weighted by biochemical properties and derives compensation-coevolution statistics for clusters of sites. Cluster significance was assessed with the accompanying R script, which implements a parametric bootstrap procedure with corrections for multiple testing (Dutheil and Galtier, 2007).

Filtered amino acid MSAs and their corresponding phylogenetic trees with previously estimated branch lengths were provided as input to CoMap. We used the LG08 substitution model with discrete-Γ among-site rate variation ($k = 4$), fixing the shape parameter α to its previously estimated value. Gapped and constant sites were excluded, while sites containing ambiguous residues were retained to minimize data loss. To identify groups of coevolving sites, we ran CoMap in clustering mode using the compensation-based distance and complete-linkage hierarchical clustering, under the same LG08 + Γ ($k = 4$) model and marginal ancestral-state probabilities described above. We capped cluster size at five sites to limit the accumulation of uncertainty from substitution mapping, which tends to make larger clusters noisy and less reliable. To assess significance, we enabled the parametric null procedure with 10,000 bootstrap replicates, generating a null distribution under site independence against which the observed clusters were evaluated.

Across three runs, the weight matrix was set to: `GranthamPolarity`, `GranthamVolume`, or `Charge`. The first two are Grantham scales—per-residue numerical indices capturing side-chain polarity and molecular volume; the charge

matrix encodes residue charge with values −1 (negatively charged; Asp, Glu), +1 (positively charged; Lys, Arg, His), and 0 (neutral, other amino acids) (Dutheil and Galtier, 2007; Grantham, 1974).

At the end of the analysis, we detected 38,643 significant clusters in 7,845 OGs for polarity, 42,513 significant clusters in 8,103 OGs for volume, and 20,349 significant clusters in 6,917 OGs for charge.

2.3.2 Branch-Specific Significance Testing

A critical limitation of CoMap is that the significance test is global; it can identify a cluster of coevolving sites but it cannot localize where along the phylogeny compensatory evolution occurred. This impedes the testing of hypotheses that relate compensatory events to branch-specific properties, such as d_N/d_S.

To overcome this limitation, we developed and implemented a novel, branch-specific parametric bootstrap procedure as a post-hoc test. This method tests for significant compensation within a pre-identified cluster on each individual branch of the tree, allowing us to statistically assign coevolutionary events to specific lineages. Our approach consists of three main steps: simulating a null distribution of biochemical change, calculating an observed compensation index for each branch, and comparing the observed value to the null to derive a p-value.

Simulation of the Null Distribution

The parametric-bootstrap procedure is based on the assumption that the original dataset is a random sample of a specific parametric distribution; we therefore fitted the model to the data and then simulated random datasets from it. The resulting replicates provide an empirical null distribution against which the observed compensation statistic can be compared (Efron and Tibshirani, 1993).

Using the site-specific substitution rates previously estimated with BppML, we simulated protein sequences for each site in a significant cluster. Each simulation contained at least 10,000 non-constant sites, yielding the same number of bootstrap replicates. We used BppSeqGen, part of the Bio++ Program Suite (Guéguen et al., 2013), to simulate sequences given a phylogenetic tree and a substitution rate. Because neither BppSeqGen nor CoMap accepts a site-specific rate parameter directly, we incorporated the rate by scaling the tree: for each site we multiplied every branch length by its corresponding substitution rate. We then ran BppSeqGen under a constant-rate LG08 model (rate-distribution option set to `Constant`) and supplied the same settings to CoMap. The tree topology and the weighted substitution matrix were identical to those used in the earlier analyses. The simulated sequences and the rescaled tree were finally passed to CoMap to

generate substitution maps.

Calculation of the Compensation Index and p-value

We calculated the compensation index (C) as defined by Chaurasia and Dutheil (2022), which quantifies biochemical compensation within each cluster of sites, given their individual variation. For a cluster G that contains n sites, the index is:

$$C(G) \;=\; 1 - \frac{\left|\sum_{i=1}^{n} X_i\right|}{\sum_{i=1}^{n} |X_i|}, \tag{2.2}$$

where, in the original work by Dutheil and Galtier (2007), X_i is a vector that corresponds to the mean changes on each branch at site i and its absolute value $|X_i|$ is the amount of change independently of the direction. This statistic was redefined in (Chaurasia and Dutheil, 2022) to compute the statistic for each branch of the phylogeny, allowing us to measure the compensation on a specific branch. If the directions of change within the cluster cancel out, the numerator approaches zero and C approaches 1—perfect compensation. Conversely, changes that act in the same direction make the numerator and denominator similar, so C is close to zero.

We then calculated the compensation index (C) for each significant cluster of sites and derived its null distribution by recomputing the statistic on the simulated corresponding sites, using 10,000 samples from the constant-rate alignments. We estimated the p-value as a Monte Carlo (parametric-bootstrap) p-value with plus-one (Phipson-Smyth) correction (Phipson and Smyth, 2010) as follows:

$$p_{\text{val}} = \frac{\sum_{i=1}^{m} \mathbf{1}(S_i \geq T_{\text{obs}}) + 1}{m + 1}, \tag{2.3}$$

where $S_i, \ldots, S_m$ are the simulated statistics, T_{obs} is the observed value, and $\mathbf{1}\{\cdot\}$ the indicator function. This definition yields a one-sided (right-tailed) p-value. To perform these calculations, we developed a custom Python script.

Given the large number of tests, a standard $p \leq 0.05$ threshold would be overly liberal. We therefore inspected the empirical p-value distribution (Fig. A.1) to establish a more rigorous data-driven cutoff. The histogram shows a sharp spike of low p-values, characteristic of a strong signal, which begins to level off into a more uniform distribution. We chose a conservative cutoff of $p \leq 5 \times 10^{-4}$, which corresponds to the "elbow" of this distribution, to stringently separate branches with strong evidence of coevolution from background noise.

2.4 Correlative Analyses: Linking Coevolution to Population Size and Structure

Having generated two comprehensive, branch-specific datasets: one for effective population size (via d_N/d_S) and one for the frequency of compensatory events, we were finally able to test our central hypothesis. This section details the statistical models used to correlate these two variables while controlling for known confounders like branch length. Furthermore, to explore the underlying biophysical mechanisms, we integrated protein structural information to investigate whether compensatory mutations preferentially occur in specific structural positions.

2.4.1 Correlation between Compensatory Mutations and Effective Population Size

For each alignment position i, the CoMap model considers every substitution that could have occurred on the branch, weighted by its posterior probability from the ancestral-state reconstruction. Let d_{ij} be the magnitude of the biochemical change—e.g., Grantham polarity distance—for the j^{th} possible substitution at that site, and let p_{ij} be the probability that this substitution actually happened. We define:

$$X_i \;=\; \sum_j p_{ij}\, d_{ij}, \tag{2.4}$$

Thus X_i is the expected biochemical change at site i. If a site is almost certainly conserved, the "no-change" term dominates the sum. For instance, with:

$$p_{\text{no change}} = 0.999, \qquad d_{\text{no change}} = 0$$

$$p_{\text{substitution}} = 0.001, \qquad d_{\text{substitution}} = 3$$

we obtain:

$$X_i \;=\; (0.999 \times 0) \;+\; (0.001 \times 3) \;=\; 0.003,$$

a near-zero contribution. Sites with higher probabilities of substitution contribute proportionally more, and the sum $\sum |X_i|$ across a cluster reflects its total expected biochemical change (Dutheil et al., 2005).

Because it is unclear whether minimal biochemical changes represent mere noise or meaningful biological signal, we imposed a conservative threshold on the total biochemical change per group, $\sum |X_i|$ (Supplementary Materials Fig. A.3).

The threshold was derived from the Grantham matrix (Grantham, 1974): for each biochemical property we took 80% of the smallest non-zero Grantham distance between any two amino acids (i.e., $\Delta_{\min} \times 0.8$), thereby allowing a 20% margin of uncertainty.

We repeated all analyses on both the threshold-filtered dataset and the full, unfiltered dataset. This dual approach ensures that any overarching trends are retained regardless of how these small changes are interpreted. The key correlations were qualitatively identical under both conditions, indicating that our results are robust to the treatment of minimal changes.

To estimate the branch-specific compensatory rate per site, the number of compensatory events on each branch was divided by the corresponding branch length—since longer branches naturally are expected to accumulate more substitutions—and subsequently divided by the total number of sites we tested.

A known confounder in this analyses is the correlation between branch length and d_N/d_S. We first confirmed this in our dataset, finding a significant negative correlation (Kendall's $\tau = -0.37$, $p = 6.36 \times 10^{-11}$) between branch length and d_N/d_S ratio, a pattern consist with previous reports (Mugal et al., 2014). To account for this effect, we employed to distinct statistical approaches. First, we used a non-parametric partial-correlation analysis (Kendall's method) with the R library `ppcor` (v1.1) to correlate the compensatory mutation rate with d_N/d_S while statistically controlling for branch length. Second, to better model this relationship, we also fitted a linear model (Equation (2.5)) with the compensatory mutation rate as the response variable and both d_N/d_S and branch length as explanatory variables. Based on the distributions of our data, we square-root-transformed the variables and fitted a model of the form detailed in Equation (2.5):

$$\sqrt{R_{comp}} = \beta_0 + \beta_1 \cdot \sqrt{d_N/d_S} + \beta_2 \cdot \sqrt{L} + \beta_3 \cdot \left(\sqrt{d_N/d_S} \cdot \sqrt{L}\right) + \epsilon, \qquad (2.5)$$

where R_{comp} is the compensatory mutation rate, d_N/d_S is the ratio of non-synonymous to synonymous substitution rates, L is the branch length, β_0 is the intercept, β_1 and β_2 are the coefficients for the main effects, β_3 is the coefficient for the interaction term, and ϵ is the residual error.

We performed additional control analyses on the terminal branches using species-specific traits (N_e and body mass). These traits present a different confounder: phylogenetic inertia. We therefore used a dual approach to test them. First, as a non-parametric test, we used a partial Kendall's correlation for N_e (controlling for branch length) and a standard Kendall's correlation for body mass, which serves as a branch-length-independent proxy. Second, to explicitly and more robustly control for phylogenetic non-independence, we also fitted separate PGLS models

(Equation (2.1)) for both traits, using the methods described in Section 2.2.3.

2.4.2 Analysis of Structural Determinants

To investigate the structural determinants of compensatory evolution, we incorporated protein structure information using a two-pronged strategy. First, we conducted analyses using only high-confidence, experimentally resolved protein structures. Second, to identify the most suitable computational method for our data, we generated and evaluated two parallel datasets of predicted structures from state-of-the-art deep learning models. This approach allowed us to ground our findings in empirical data while also systematically selecting the highest-quality predicted structures for a broader analysis. These models take the general form:

Dataset Assembly from Experimental and Predicted Structures

Using the cross-references provided by the OMA database and UniProt (The UniProt Consortium, 2025), we linked our sequences to entries in the RCSB Protein Data Bank (RCSB PDB) database (Berman et al., 2000) (see Supplementary Materials Table A.1). We then downloaded all available experimentally determined structures in mmCIF format, obtaining 13,987 structure files that map to 1,782 OGs. For every OG we kept only structures whose experimental resolution was ≤ 5 Å. If an OG was associated with more than five qualifying structures, we retained the five with the largest residue counts; 1,367 OGs were associated with suitable structures files.

Next, to generate the two computational datasets for comparison, we used two leading prediction methods employing distinct strategies. The first set was compiled from the AlphaFold Protein Structure Database(Jumper et al., 2021; Varadi et al., 2024) (Supplementary Materials Table A.1), from which we downloaded the single best-quality model available for each OG irrespective of the species, based on the highest confidence score. This approach yielded 5,760 structure files.

For the second set, we predicted one 3D protein model for each orthologous group with Meta's ESMFold (fair-esm v2.0.0). This tool uses a protein-language model, which removes the need for traditional multiple-sequence alignments. It generates a structure from a single sequence and then refines the atomic geometry (Lin et al., 2023). We adopted the *Homo sapiens* sequence as the reference, because a human ortholog is present in the vast majority of groups in our dataset, guaranteeing uniform input length and annotation quality. We successfully predicted 6,554 protein structures.

Mapping Residues and Extracting Structural Features

We first mapped the residues of each significant cluster back onto the original multiple-sequence alignment using the site indices produced by BppSeqMan. CoMap outputs results as TSV files in the Site-Group-Extended-Data (Dutheil et al., 2024). To add structural information to SGED files, we used SGEDtools, a set of Python programs to manipulate SGED files.

The coordinate translation was carried out with `sged-translate-coords.py`, part of the SGEDtools. Next, we created an index that links alignment positions to residue numbers in the available 3D structures. This was done with `sged-create-structure-index.py` (SGEDtools), to which we added the `–coverage` option so that only structures covering a user-defined fraction of the alignment are retained, we set this option to 80%. The program accepts multiple PDB/mmCIF files, aligns every OG sequence to each structure, and selects the best pairwise match. In this way we automatically choose, for every orthologous group, the structure that best fits the corresponding sequence and meets the minimum-coverage criterion.

At this step, we successfully generated indexes to map sequence residues onto their corresponding protein structures. The quantitative results of this process are summarized in Table 2.1. The table details the overall mapping coverage and provides a further breakdown for significant sites categorized by their biochemical properties (polarity, volume, and charge). The partial mapping coverage shown in the table is an expected outcome, as many experimentally determined protein structures are incomplete, often representing only a single domain or chain rather than the full-length sequence.

Table 2.1. Summary of sequence residues mapped to protein structures. The "General Mapping Summary" details the total number of structural indexes created, the total residues in the dataset, and the number (and percentage) of those residues successfully mapped. The "Coverage of Significant Sites" section details the mapping percentage specifically for significant compensatory sites, broken down by biochemical property.

General Mapping Summary			
Statistic		Value	
Created Indexes			1,253
Total Sequence Residues			678,493
Mapped Sites on Structures			381,698 (56.26%)
Coverage of Significant Sites	Polarity	Volume	Charge
% in OGs with structures	66%	53%	53%
% of total sites	37%	36%	27%

Using `sged-structure-infos.py` (SGEDtools), we extracted two per-residue structural properties from each mmCIF file: degree of contact, the count of

residues whose C-α atoms lie within 5 Å of the focal residue (covalently bonded neighbours included), and solvent accessibility, the relative solvent-accessible surface area (RSA) of the residue. For every significant compensatory cluster we also calculated the mean pairwise C-α distance among its member residues, providing a quantitative measure of their spatial proximity.

Selection of the Best Predicted Dataset

To quantitatively evaluate the two predicted datasets—AlphaFold and ESMFold—, we benchmarked them against our high-confidence experimental structures. For each orthologous group where structures from all three sources were available, we first identified the common set of residues. Then, we assessed the accuracy of each predicted dataset by measuring the Pearson's correlation (r) between the values derived from the predicted structures and those from the corresponding experimental structures. This analysis was conducted separately for RSA and degree of contact. The computational dataset exhibiting the highest correlation with the experimental data was selected for all subsequent structural analyses. The AlphaFold dataset consistently performed better (RSA: $r = 0.781$; Degree of Contact: $r = 0.837$) than the ESMFold dataset (RSA: $r = 0.775$; Degree of Contact: $r = 0.825$) and was therefore carried forward. With AlphaFold we were able to map onto structures a total of 3,353,556 sites.

2.4.3 Branch-Level Analysis of the Interplay between Structure and Selection

With the experimental and AlphaFold datasets finalized, we investigated how structural context modulates the relationship between N_e and compensatory evolution. We performed parallel analyses for each of the three biochemical properties (polarity, volume, and charge) using both the experimental and AlphaFold structural datasets, resulting in six independent analyses that each followed the same statistical pipeline.

Analysis of Site-Level Structural Properties: RSA and Contact Degree

For each of the six analyses, we first stratified all relevant compensatory sites by their structural environment. Each site was categorized into one of five equal-sized quintiles based on its contact degree and one of five equal-sized quintiles based on its Relative Solvent Accessibility (RSA). This categorization, performed using the `cut2()` function from the `Hmisc` R library (v5.2.3), generated 25 distinct structural categories (5 contact quintiles $\times$ 5 RSA quintiles).

Next, for each branch of the phylogeny, we counted the number of compensatory sites (for the specific property) that fell into each of these 25 structural categories. This yielded a frequency count for each combination of branch and structural category. These frequencies were then divided by the corresponding branch length and the total number of sites tested to calculate the branch-specific compensatory mutation rate for that specific structural context.

We then fitted a standard linear model (`lm()`) to each of the six datasets. The predictors for the structural categories were not the quintile labels (e.g., 1-5) but rather the mean continuous value (e.g., mean contact degree) for all sites within that specific quintile.

Each of the six models tested how the log-transformed rate of compensatory mutations was predicted by the main effects and all two-way interactions between the following predictors: the standardized branch-specific d_N/d_S ratio, the standardized mean contact degree for the quintile, and the standardized mean RSA for the quintile.

All three predictor variables were standardized (z-scored) using the `scale()` R function to improve interpretability and model stability. This full model is detailed in Equation (2.6):

$$\begin{aligned}\log(R_{comp}) = \beta_0 &+ \beta_1 \cdot (d_N/d_s)_z + \beta_2 \cdot RSA_z + \beta_3 \cdot NbContact_z \\ &+ \beta_4 \cdot ((d_N/d_s)_z \cdot RSA_z) \\ &+ \beta_5 \cdot ((d_N/d_s)_z \cdot NbContact_z) \\ &+ \beta_6 \cdot (RSA_z \cdot NbContact_z) + \epsilon,\end{aligned} \tag{2.6}$$

where $\log(R_{comp})$ is the log-transformed compensatory rate; $(d_N/d_s)_z$, RSA_z, and $NbContact_z$ are the standardized d_N/d_S ratio, mean RSA, and mean contact degree, respectively; β_{1-3} are the main effect coefficients; β_{4-6} are the interaction term coefficients; and ϵ is the error term.

Each full model was then simplified using the `step()` function, which performs a stepwise backward selection to find the model with the minimum AIC (Akaike Information Criterion) value. Finally, we checked for multicollinearity issues among the predictors in each final model using the `vif()` function (with the `type="predictor"` argument) from the `car` package (v3.1.3).

Analysis of Group-Level Structural Properties: RSA and C_α-C_α Distance

We applied a similar framework to analyze the structural properties of compensatory groups rather than individual sites. This analysis used two group-level structural predictors: the mean RSA of all residues within a significant compensatory group,

and the mean pairwise C_α-C_α distance between all residues within that same group.

As in the site-level analysis, we stratified all significant compensatory groups into 25 categories based on the quintiles of these two predictors (5 mean group RSA quintiles $\times$ 5 mean intragroup distance quintiles). This stratification and the subsequent modeling were performed in parallel for the six datasets (three properties $\times$ two structure sources).

For each branch, we counted the number of compensatory groups that fell into each of the 25 structural categories. This frequency was divided by the branch length and the total number of sites tested to calculate the branch-specific rate of compensatory groups for that structural context.

We then fitted a linear model (lm()) to each dataset. The predictors were the standardized d_N/d_S ratio and the standardized mean continuous values for the group RSA and intragroup distance quintiles. The full model, including all main effects and two-way interactions, is detailed in Equation (2.7):

$$\begin{aligned} \log(R_{comp}) = \beta_0 &+ \beta_1 \cdot (d_N/d_S)_z + \beta_2 \cdot GroupRSA_z + \beta_3 \cdot CalphaDist_z \\ &+ \beta_4 \cdot ((d_N/d_S)_z \cdot GroupRSA_z) \\ &+ \beta_5 \cdot ((d_N/d_S)_z \cdot CalphaDist_z) \\ &+ \beta_6 \cdot (GroupRSA_z \cdot CalphaDist_z) + \epsilon, \end{aligned} \tag{2.7}$$

where $\log(R_{comp})$ is the log-transformed compensatory rate; $(d_N/d_S)_z$ is the standardized d_N/d_S ratio; $GroupRSA_z$ is the standardized mean group RSA; and $CalphaDist_z$ is the standardized mean intragroup C_α-C_α distance. The β terms represent the coefficients for main effects (β_{1-3}) and interactions (β_{4-6}), and ϵ is the residual error. This model was simplified and validated for multicollinearity using the same AIC-based stepwise selection (`step()` function) and `vif()` procedures described in Section 2.4.3.

Results

We present our findings on the link between effective population size (N_e) and the accumulation of intra-protein compensatory mutations. We begin by presenting the validation of the branch-specific ratio of non-synonymous to synonymous substitutions (d_N/d_S) as a robust proxy for N_e. We then address our central hypothesis by quantifying the correlation between this proxy and the rate of compensatory mutation across the mammalian phylogeny. Finally, we explore how the structural properties impact the effect of N_e on the compensatory mutation rate.

3.1 Validation of d_N/d_S as a Proxy for Effective Population Size

A prerequisite for testing our central hypothesis was to validate that branch-specific d_N/d_S serves as a reliable proxy for N_e. We used Phylogenetic Generalized Least Squares (PGLS) to model the relationships between our genome-wide d_N/d_S estimates, available body mass data (a life-history trait known to be inversely correlated with N_e) and published species-specific N_e estimates (Wilder et al., 2023). This phylogenetic approach was essential for correcting for the non-independence of data points due to shared ancestry. All variables were log-transformed prior to analysis to better meet the assumptions of linear modeling. To ensure robustness, we fitted each model with two distinct phylogenetic correction parameters: Martins's α and Grafen's ρ.

The PGLS models confirmed all theoretically expected relationships. First, we

Table 3.1. Summary of Phylogenetic Generalized Least Squares (PGLS) model results. Predictor variables (N_e, body mass) were log-transformed prior to analysis. For each model, β_1 represents the regression coefficient, accompanied by its standard error (SE). The phylogenetic signal was estimated using both Martins' α and Grafen's ρ parameters.

Model	Corr.	$\beta_1 \pm$ SE	p	Phylo. Sig.
$\log(d_N/d_S) \sim \log(Body\ Mass)$	Martins	0.048 ± 0.019	0.017	$\alpha = 57.9$
	Grafen	0.063 ± 0.020	0.004	$\rho = 0.104$
$\log(d_N/d_S) \sim \log(N_e)$	Martins	-0.285 ± 0.063	< 0.001	$\alpha = 88.1$
	Grafen	-0.267 ± 0.072	< 0.001	$\rho = 0.341$
$\log(N_e) \sim \log(Body\ Mass)$	Martins	-0.126 ± 0.031	< 0.001	$\alpha = 40.7$
	Grafen	-0.114 ± 0.035	0.002	$\rho = 0.460$

verified the expected negative relationship between N_e and body mass. Furthermore, as predicted by the nearly neutral theory of molecular evolution, our estimated d_N/d_S was significantly negatively correlated with published N_e estimates for the 45 species for which data were available. As a direct consequence of these two patterns, d_N/d_S showed significant positive correlation with body mass (Table 3.1, Fig. 3.1). Taken together, these results provide strong support for the use of branch-specific d_N/d_S as a proxy for N_e in our subsequent analyses.

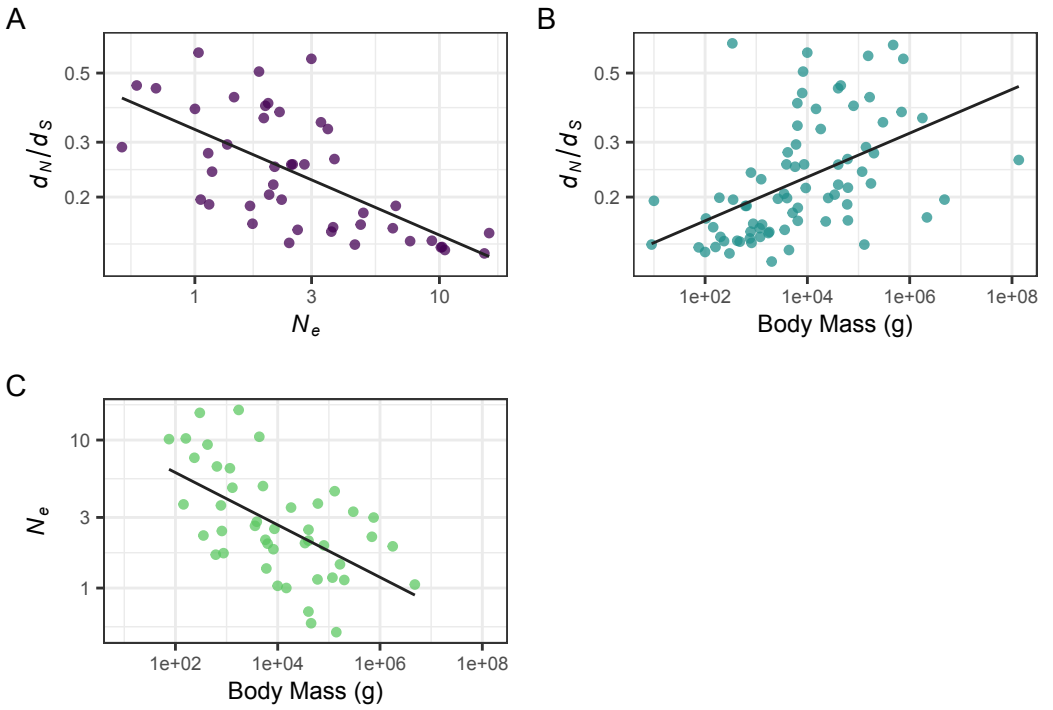

Figure 3.1. Relationships between d_N/d_S, N_e, and body mass. These plots show the relationships tested using Phylogenetic Generalized Least Squares (PGLS). The solid line in each panel is the regression line, and all axes are log-scaled. The panels show: (A) the significant negative correlation between d_N/d_S and N_e; (B) the significant positive correlation between d_N/d_S and body mass; and (C) the significant negative correlation between body mass and N_e.

3.2 Correlation Between Compensatory Mutations and d_N/d_S

Having validated d_N/d_S as a proxy for effective population size, we now address the central hypothesis of this study: that a lower N_e drives an increased rate of compensatory evolution. To test this, we analyzed the correlation between our branch-specific d_N/d_S estimates and the rate of compensatory mutation. The following results provide, to our knowledge, the first evidence for this relationship across a broad phylogenetic scale, advancing our understanding of the mechanisms

governing intra-protein coevolution.

3.2.1 Lineages with Smaller Effective Population Sizes Accumulate More Compensatory Mutations

Our coevolution analyses identified thousands of significant compensatory events for all three biochemical properties. Consistent with the removal of lower-impact substitutions, the filtered dataset produced a more conservative count of events than the unfiltered dataset. The specific counts are summarized in Table 3.2.

Table 3.2. Summary of compensatory event datasets. The table shows the total number of compensatory events and the number of orthologous groups (OGs) containing at least one event. Data is shown for both filtered and unfiltered datasets across the three biochemical properties.

	Dataset	Compensatory Events	OMA groups
Polarity	Filtered	5,620	3,363
	Unfiltered	11,995	3,843
Volume	Filtered	6,476	3,706
	Unfiltered	13,441	4,310
Charge	Filtered	7,405	3,943
	Unfiltered	10,310	4,368

Our primary hypothesis posits that lineages with smaller N_e accumulate more compensatory mutations due to the increased fixation of slightly deleterious mutations. To test this, we performed a non-parametric partial correlation (Kendall's τ) to assess the relationship between branch-specific d_N/d_S and the rate of compensatory mutation, while controlling for the effect of branch length. The analysis revealed a strong, positive, and highly significant correlation between d_N/d_S and the accumulation of compensatory mutations for all three biochemical properties examined.

To ensure this result was not an artifact of our filtering criteria, we performed the analysis on both the primary dataset, which excludes substitutions with minimal biochemical impact, and the complete unfiltered dataset. The correlation remained highly significant in both cases. For the filtered dataset, Kendall's τ ranged from 0.540 to 0.568 (all $p < 0.001$, Table 3.3 and Fig. 3.2). While moderately weaker in the unfiltered data (τ = 0.421 to 0.523), the relationship remained unambiguous (all $p < 0.001$, Table 3.3).

The linear models, which parametrically test this relationship while explicitly accounting for branch length as a covariate, confirmed this finding. In all models, the square-root-transformed d_N/d_S ratio was a highly significant positive predictor of the square-root-transformed compensatory mutation rate. For the filtered dataset, the β coefficients for $\sqrt{d_N/d_S}$ were consistently positive and significant, ranging

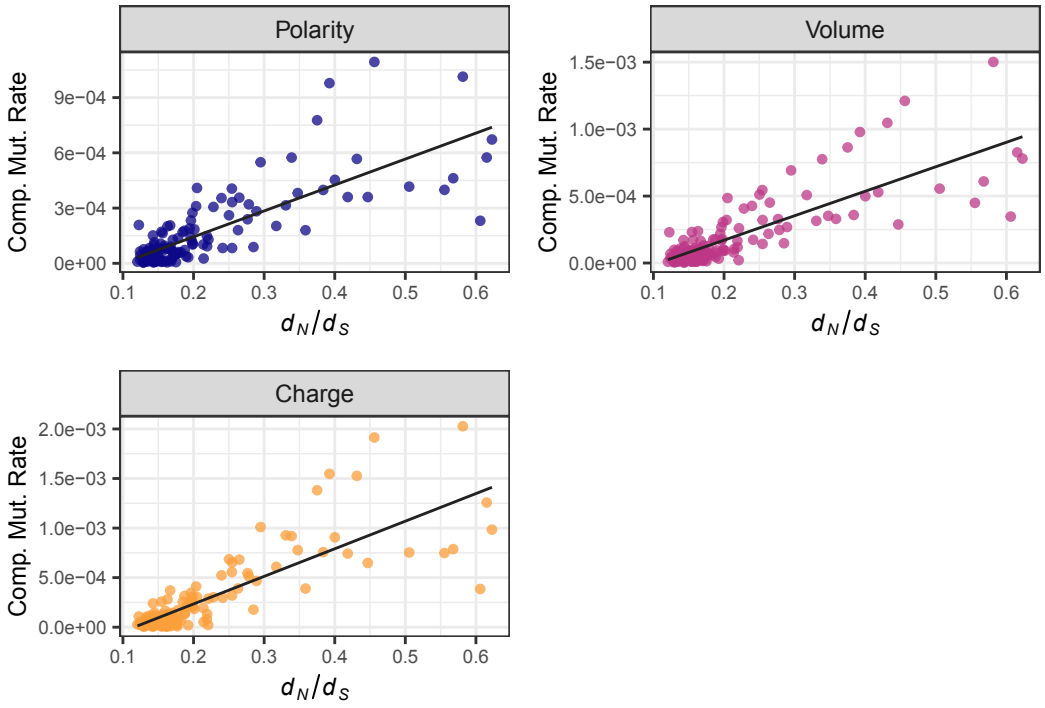

Figure 3.2. Relationship between the branch-specific d_N/d_S ratio and the normalized frequency of compensatory mutations. Each panel corresponds to a different biochemical property. The positive trend indicates that lineages with low N_e (higher d_N/d_S) accumulate more compensatory mutations.

Table 3.3. Correlations between compensatory mutation rate and key evolutionary parameters. The table reports Kendall's τ correlation coefficients. The relationships between the compensatory rate and d_N/d_S or N_e were tested using partial correlation to control for branch length. The body mass correlation was performed on terminal branches. Analyses are shown for both filtered and unfiltered datasets.

		Filtered Data		Unfiltered Data	
	Correlate	τ	p	τ	p
Polarity	d_N/d_S	0.540	< 0.001	0.421	< 0.001
	N_e	−0.257	0.014	−0.343	0.001
	Body mass	0.250	< 0.001	0.346	< 0.001
Volume	d_N/d_S	0.552	< 0.001	0.456	< 0.001
	N_e	−0.322	0.002	−0.356	< 0.001
	Body mass	0.282	< 0.001	0.363	< 0.001
Charge	d_N/d_S	0.568	< 0.001	0.523	< 0.001
	N_e	−0.301	0.004	−0.325	0.002
	Body mass	0.315	< 0.001	0.352	< 0.001

from 0.056 to 0.080. This pattern held for the unfiltered dataset as well, with β ranging from 0.044 to 0.078. These models demonstrated strong predictive power, explaining between 56% and 73% of the variance ($R^2 = 0.557 - 0.729$). The detailed

results of all models are summarized in Table 3.4.

This robust, dual approach (non-parametric and parametric) provides the first large-scale evidence that lineages with a higher d_N/d_S (indicative of smaller N_e) systematically accumulate more compensatory mutations.

Table 3.4. Summary of linear model results testing the effect of the d_N/d_S ratio ($\sqrt{d_N/d_S}$) on the compensatory mutation rate ($\sqrt{R_{comp}}$). The full models also included terms for square-root-transformed branch length ($\sqrt{L}$) and its interaction with $\sqrt{d_N/d_S}$. As these terms were not significant in any model, they are omitted from the table for clarity. The table reports the coefficients ($\beta \pm$ SE) for the intercept and the significant d_N/d_S predictor, along with the overall model fit (Adj. R^2), grouped by property and dataset.

	Dataset	Term	$\beta \pm$ SE	p	Adj. R^2
Polarity	Filtered	(Intercept)	-0.018 ± 0.003	< 0.001	0.666
		$\sqrt{d_N/d_S}$	0.056 ± 0.007	< 0.001	
	Unfiltered	(Intercept)	0.000 ± 0.003	0.912	0.557
		$\sqrt{d_N/d_S}$	0.044 ± 0.008	< 0.001	
Volume	Filtered	(Intercept)	-0.022 ± 0.003	< 0.001	0.711
		$\sqrt{d_N/d_S}$	0.066 ± 0.007	< 0.001	
	Unfiltered	(Intercept)	-0.003 ± 0.003	0.316	0.626
		$\sqrt{d_N/d_S}$	0.052 ± 0.007	< 0.001	
Charge	Filtered	(Intercept)	-0.025 ± 0.004	< 0.001	0.729
		$\sqrt{d_N/d_S}$	0.080 ± 0.009	< 0.001	
	Unfiltered	(Intercept)	-0.020 ± 0.004	< 0.001	0.711
		$\sqrt{d_N/d_S}$	0.078 ± 0.008	< 0.001	

3.2.2 Validation using Direct N_e Estimates and a Branch-Length-Independent Proxy

To independently validate this central finding, we conducted two control analyses on the terminal branches of the phylogeny, for which direct species-specific data are available. First, we replaced our d_N/d_S proxy with published estimates of N_e. As a non-parametric test, a partial correlation controlling for branch length revealed a significant negative relationship between N_e and the compensatory mutation rate for all biochemical properties. For the filtered dataset, Kendall's τ ranged from −0.257 to −0.315 (all $p \leq 0.014$, Table 3.3 and Fig. 3.3), and this held for the unfiltered data as well (τ from −0.325 to −0.356, all $p \leq 0.002$, Table 3.3). To more robustly test this relationship while accounting for phylogenetic inertia, we also fitted PGLS models (Equation (2.1)). These models confirmed the finding, showing a strong, significant negative relationship across all datasets (Table 3.5). The β coefficients ranged from −0.442 to -0.665 (all $p \leq 0.002$), reinforcing that species

with smaller effective population sizes accumulate more compensatory mutations (see Supplementary Materials Fig. A.5).

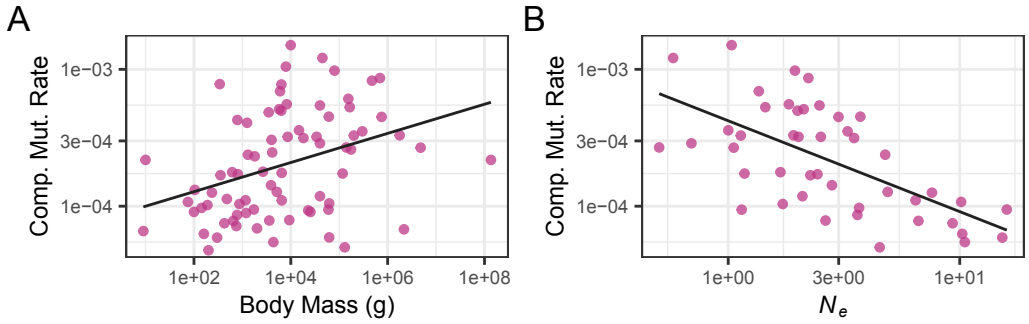

Figure 3.3. Compensatory mutation rate vs. proxies for effective population size. The plots, shown for the Volume dataset as a representative example, illustrate the key correlations found. All axes are log-scaled. The rate of compensatory mutation is plotted against (A) body mass, and (B) direct N_e estimates.

Second, to ensure the trend was not a methodological artifact of branch length effects, we used body mass as a proxy for N_e that is entirely independent of the phylogenetic model. A non-parametric correlation confirmed a significant positive relationship between body mass and the rate of compensatory mutation. This was true for both filtered (τ from 0.250 to 0.315, all $p < 0.001$, Table 3.3 and Fig. 3.3) and unfiltered data (τ from 0.352 to 0.363, all $p < 0.001$). The corresponding PGLS models, controlling for phylogeny, also supported this positive relationship (Table 3.6). All β coefficients were positive, ranging from 0.075 to 0.104 (all $p < 0.05$). This is consistent with our hypothesis, as larger body mass is a strong indicator of smaller N_e (Supplementary Materials Fig. A.4).

3.3 Structural Determination of Drivers of Compensatory Mutations

Having established that smaller N_e drives the accumulation of compensatory mutations, we next sought to identify the structural determinants of this process. Based on previously published results (Chaurasia and Dutheil, 2022), we hypothesized that compensatory mutations would occur more frequently in structurally constrained regions, where mutations are more likely to be deleterious and thus require compensation. We also investigated the separate, more complex question of whether this structural environment modulates the effect of N_e. To test both the main effect of structure and its potential N_e, we stratified coevolving sites into five quintiles ("bands") based on two correlated metrics of structural constraint: degree of structural contact—from isolated to highly connected—and Relative Solvent

Table 3.5. Summary of Phylogenetic Generalized Least Squares (PGLS) results testing the relationship between the compensatory mutation rate ($\log(R_{comp})$) and effective population size ($\log(N_e)$). Models were fitted separately for each biochemical property and dataset (Filtered and Unfiltered), using two different phylogenetic correction structures (Martins's α and Grafen's ρ). The table reports the regression coefficients ($\beta \pm$ SE), p-values, and the estimated phylogenetic parameter.

	Dataset	Corr.	$\beta \pm$ SE	p	Phylo. Param.
Polarity	Filtered	Martins	-0.543 ± 0.147	< 0.001	$\alpha = 38.112$
		Grafen	-0.526 ± 0.156	0.002	$\rho = 0.068$
	Unfiltered	Martins	-0.442 ± 0.104	< 0.001	$\alpha = 12.059$
		Grafen	-0.497 ± 0.099	< 0.001	$\rho = 0.219$
Volume	Filtered	Martins	-0.639 ± 0.136	< 0.001	$\alpha = 45.646$
		Grafen	-0.596 ± 0.147	< 0.001	$\rho = 0.086$
	Unfiltered	Martins	-0.477 ± 0.102	< 0.001	$\alpha = 15.799$
		Grafen	-0.516 ± 0.102	< 0.001	$\rho = 0.229$
Charge	Filtered	Martins	-0.665 ± 0.142	< 0.001	$\alpha = 25.387$
		Grafen	-0.635 ± 0.153	< 0.001	$\rho = 0.174$
	Unfiltered	Martins	-0.617 ± 0.130	< 0.001	$\alpha = 19.585$
		Grafen	-0.613 ± 0.136	< 0.001	$\rho = 0.218$

Table 3.6. Summary of Phylogenetic Generalized Least Squares (PGLS) results testing the relationship between the compensatory mutation rate ($\log(R_{comp})$) and body mass ($\log(\text{body mass})$). Models were fitted separately for each biochemical property and dataset (Filtered and Unfiltered), using two different phylogenetic correction structures (Martins's α and Grafen's ρ). The table reports the regression coefficients ($\beta \pm$ SE), p-values, and the estimated phylogenetic parameter.

	Dataset	Model	$\beta \pm$ SE	p	Phylo. Param.
Polarity	Filtered	Martins	0.080 ± 0.032	0.015	$\alpha = 58.059$
		Grafen	0.075 ± 0.035	0.039	$\rho = 0.057$
	Unfiltered	Martins	0.081 ± 0.026	0.003	$\alpha = 13.835$
		Grafen	0.090 ± 0.025	< 0.001	$\rho = 0.374$
Volume	Filtered	Martins	0.084 ± 0.032	0.012	$\alpha = 31.134$
		Grafen	0.081 ± 0.036	0.025	$\rho = 0.120$
	Unfiltered	Martins	0.086 ± 0.026	0.001	$\alpha = 12.686$
		Grafen	0.102 ± 0.025	< 0.001	$\rho = 0.370$
Charge	Filtered	Martins	0.104 ± 0.034	0.003	$\alpha = 35.567$
		Grafen	0.084 ± 0.037	0.027	$\rho = 0.149$
	Unfiltered	Martins	0.095 ± 0.030	0.002	$\alpha = 18.826$
		Grafen	0.091 ± 0.032	0.006	$\rho = 0.235$

Accessibility (RSA)—from buried to exposed.

We modeled the compensatory mutation rate as a function of d_N/d_S (our proxy for N_e) and the structural bands, including main effects for all variables as well as the interaction terms. Although the two structural metrics are strongly correlated

(Pearson's $r = -0.73$, $p < 0.001$), we performed collinearity diagnostics and confirmed that including both band predictors in the same model was statistically sound, as all VIF coefficients were below 5 (Hair et al., 1995). This unified model allows us to assess the relative contribution of each structural property while accounting for the other.

3.3.1 Effect of Relative Solvent Accessibility and Number of Contacts

Our analysis of 1,782 experimentally determined and 5,760 AlphaFold protein structures allowed us to model the impact of effective population size and structural constraints on the compensatory mutation rate. The full model results are summarized in Table 3.7.

Two findings were strongly significant and consistent across all biochemical properties and both datasets. First, the standardized d_N/d_S ratio, our proxy for N_e, showed a large, significant positive effect in all six models (Experimental: β from 0.330 to 0.349, all $p < 0.001$; AlphaFold: β from 0.316 to 0.364, all $p < 0.001$). Second, we found a strong, negative interaction effect between the number of contacts and RSA ($NbContact_z$:RSA_z) in all models (Experimental: β from -0.091 to -0.158, all $p < 0.001$, Fig. 3.4; AlphaFold: β from -0.225 to -0.244, all $p < 0.001$, see Supplementary Materials Fig. A.6).

The high significance of the d_N/d_S ratio and the $NbContact_z : RSA_z$ interaction—both of which were retained following model simplification—strongly suggests they represent a true biological signal rather than a statistical artifact. The d_N/d_S result confirms the main finding of this thesis, while the $NbContact_z : RSA_z$ interaction indicates that the effects of being buried (low RSA) and being highly connected (high contacts) are not merely additive.

In contrast to the clear main effect of d_N/d_S, the main effects of the structural variables (RSA_z and $NbContact_z$) are not readily interpretable on their own. This is expected, given the strong and consistent negative interaction we found between them ($NbContact_z : RSA_z$). This interaction reveals a complex biophysical relationship: the effect of solvent exposure on the compensatory mutation rate is highly dependent on the site's contact level. As shown in Fig. 3.4, the analysis reveals a clear crossover interaction: at buried sites (low RSA), a high contact degree (more contacts) is associated with a higher compensatory mutation rate. Conversely, at exposed sites (high RSA), a low contact degree (fewer contacts) is associated with a higher compensatory mutation rate.

Crucially, the interaction terms between d_N/d_S and the structural properties—our key test for how N_e is modulated by structure—yielded ambiguous results. While some terms were nominally significant in individual models (e.g., $(d_N/d_s)_z$:

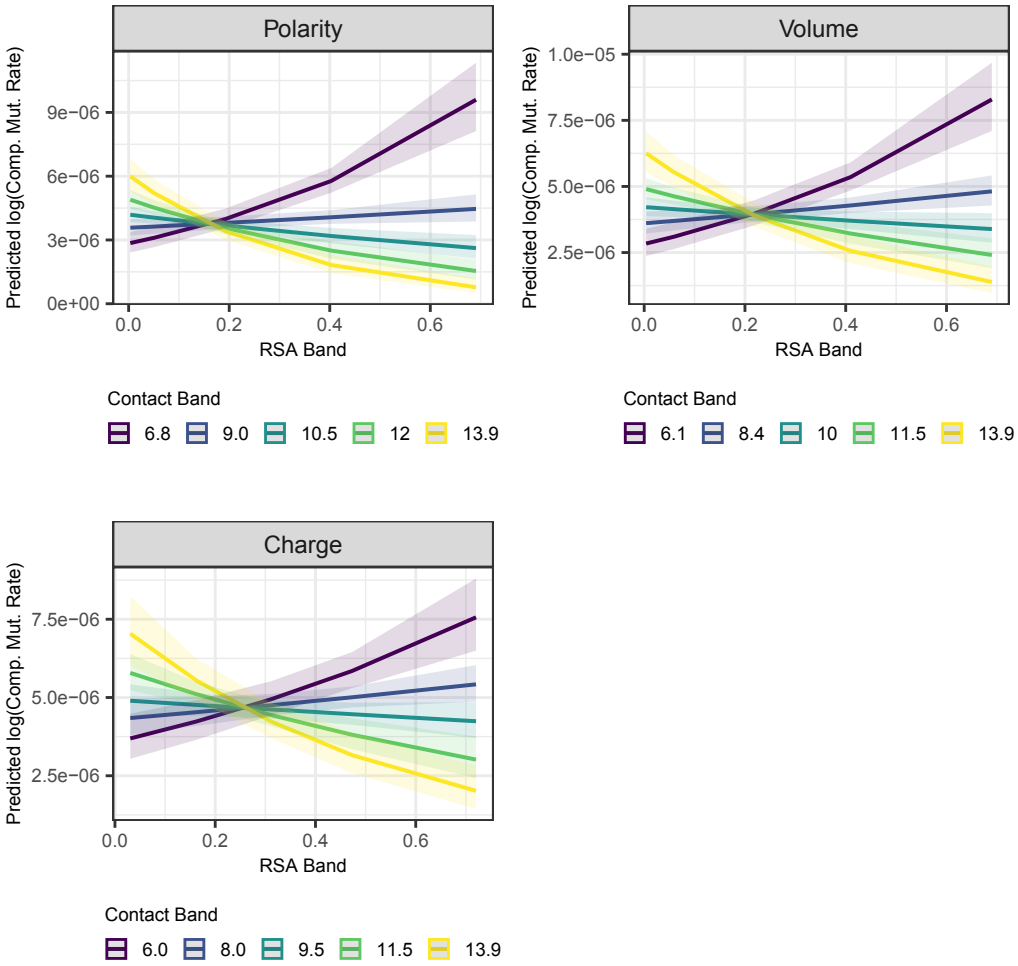

Figure 3.4. Interaction effect between RSA and number of contact on the predicted compensatory mutation rate in the experimental dataset. The plot is divided into three panels showing the results from models for polarity, volume, and charge. Each panel visualizes the predicted log(Comp. Mut. Rate) (Y-axis) as a function of RSA Band (X-axis) at different levels of Contact Band (colored lines). The plot reveals a strong crossover interaction, which is consistent across all three models. At low RSA values (e.g., 0.00), sites with a high Contact Band (bright yellow lines) are predicted to have a higher compensatory mutation rate. This relationship inverts at high RSA values (e.g., 0.75), where sites with a low Contact Band (dark purple lines) show the highest predicted compensatory mutation rate. This suggests that the effect of solvent exposure (RSA) on the compensatory mutation rate is highly dependent on the site's contact level.

$NbContact_z$ for Volume, $p < 0.001$), others were non-significant (e.g., $(d_N/d_s)_z$: $NbContact_z$ for Polarity, $p = 0.149$), only marginally significant ($p = 0.059$), or removed entirely by model simplification. This inconsistency across models, and the fact that these p-values were not corrected for multiple testing, prevents a clear interpretation of how structural constraints modulate the effect of N_e.

Finally, all models demonstrated strong predictive power, with adjusted R^2 values ranging from 0.468 to 0.524, indicating that our predictors explain between 46.8% and 52.4% of the variance in the compensatory mutation rate.

Table 3.7. Summary of site-level linear model (lm) results. The dependent variable is the log-transformed compensatory mutation rate ($\log(R_{comp})$). Predictors are standardized (z-scored): $(d_N/d_s)_z$ is the d_N/d_S ratio, RSA_z is the site Relative Solvent Accessibility, and $NbContact_z$ is the site contact degree. Blank cells indicate non-significant terms removed by model simplification.

		Experimental		AlphaFold	
	Term	$\beta \pm$ SE	p	$\beta \pm$ SE	p
Polarity	$(d_N/d_s)_z$	0.333 ± 0.013	< 0.001	0.316 ± 0.010	< 0.001
	RSA_z	-0.043 ± 0.017	0.013	-0.047 ± 0.012	< 0.001
	$NbContact_z$	-0.047 ± 0.016	0.003	-0.020 ± 0.011	0.084
	$(d_N/d_s)_z$:RSA_z	-0.024 ± 0.016	0.117		
	$(d_N/d_s)_z$:$NbContact_z$	-0.022 ± 0.015	0.149	0.018 ± 0.010	0.059
	RSA_z:$NbContact_z$	-0.158 ± 0.015	< 0.001	-0.244 ± 0.010	< 0.001
	Adj. R^2	0.524		0.505	
Volume	$(d_N/d_s)_z$	0.330 ± 0.012	< 0.001	0.320 ± 0.010	< 0.001
	$NbContact_z$	-0.019 ± 0.015	0.190	-0.029 ± 0.013	0.023
	RSA_z	-0.026 ± 0.015	0.090	-0.040 ± 0.013	0.002
	$(d_N/d_s)_z$:$NbContact_z$	0.039 ± 0.011	< 0.001		
	$(d_N/d_s)_z$:RSA_z			-0.023 ± 0.010	0.028
	$NbContact_z$:RSA_z	-0.121 ± 0.013	< 0.001	-0.225 ± 0.011	< 0.001
	Adj. R^2	0.523		0.449	
Charge	$(d_N/d_s)_z$	0.349 ± 0.012	< 0.001	0.364 ± 0.012	< 0.001
	$NbContact_z$	-0.031 ± 0.015	0.036	-0.018 ± 0.013	0.174
	RSA_z	-0.019 ± 0.015	0.190	-0.013 ± 0.013	0.332
	$(d_N/d_s)_z$:$NbContact_z$			0.026 ± 0.012	0.027
	$NbContact_z$:RSA_z	-0.091 ± 0.014	< 0.001	-0.225 ± 0.013	< 0.001
	Adj. R^2	0.488		0.468	

3.3.2 Structural Distances Between Residues in Compensatory Groups

To assess the spatial arrangement of co-evolving sites, we analyzed the mean pairwise distances between residues within each significant compensatory group. For this analysis, we pooled the results from all three biochemical properties. In the experimental structures, the median of these mean intragroup distances was 16.04Å (Interquartile Range [IQR]: 17.96Å). A similar, consistent pattern was observed in the AlphaFold-predicted dataset, which yielded a slightly larger median distance of 20.18Å (IQR: 27.96Å) (Table 3.8).

These results demonstrate that residues within a compensatory group are typically in close spatial proximity but are not in direct physical contact, as they far exceed the standard 5Å threshold for contact (Anishchenko et al., 2017; Ovchinnikov et al., 2017) (Fig. 3.5). This suggests that the co-evolutionary mechanism is often mediated by local structural stability or short-range allosteric communication rather

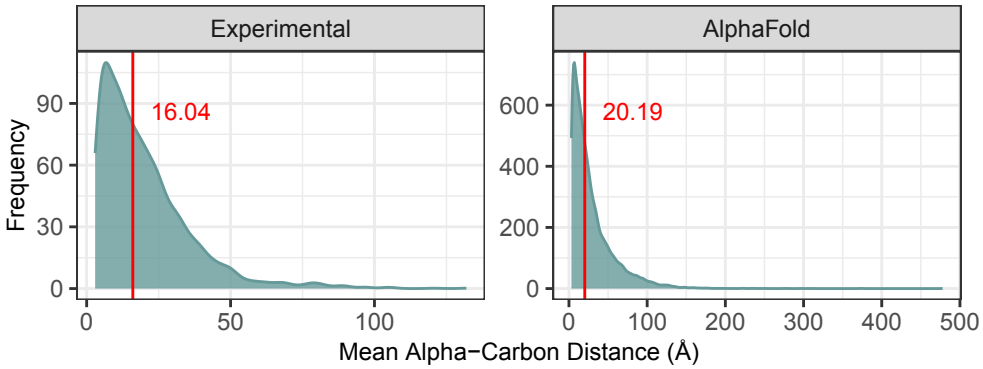

Figure 3.5. Distribution of mean pairwise distances within compensatory groups. The plots show the distribution of mean distances (Å) between co-evolving residues for the Experimental and AlphaFold datasets. The red vertical line in each panel marks the median.

than direct side-chain interactions.

Table 3.8. Summary of mean pairwise distances within compensatory residue groups. The table presents the median, mean, and Interquartile Range (IQR) for distances calculated from the experimental structures dataset and AlphaFold-predicted dataset.

	Experimental Dataset			AlphaFold Dataset		
Property	Median	Mean	IQR	Median	Mean	IQR
Polarity	15.255	19.721	16.951	19.520	28.580	27.380
Volume	16.335	20.083	17.655	20.788	29.600	28.622
Charge	16.830	20.723	19.707	20.223	28.913	27.768
Total	16.038	20.173	17.962	20.185	29.052	27.959

3.3.3 Interplay of d_N/d_S, Group RSA, and Intragroup Distance

Having characterized the typical spatial arrangement of compensatory groups, we next tested our full group-level model. This model aimed to predict the rate of compensatory groups as a function of N_e, mean group RSA, and mean intragroup C_α-$C_\alpha$distance.

The results from this analysis (Table 3.9) strongly reinforce our primary findings. The standardized d_N/d_S ratio was, once again, the most powerful and consistent positive predictor of the compensatory group rate across all six models (β ranging from 0.317 to 0.350, all $p < 0.001$). This indicates that the strong link between genetic drift and the accumulation of compensatory events is robust, holding true whether the unit of analysis is the individual site or the coevolving group.

In contrast to this clear main effect, the structural predictors and their interactions were more complex, with notable differences between the experimental and the AlphaFold datasets. In the AlphaFold models, we identified two strong

interaction patterns. First, a significant negative interaction between d_N/d_S and intragroup distance emerged for Polarity ($\beta = -0.046, p < 0.001$) and Charge ($\beta = -0.053, p < 0.001$). This suggests that the effect of genetic drift is strongest for compensatory groups that are already spatially compact. Second, a consistent positive interaction was found between mean group RSA and intragroup distance (RSA_z:$Dist_z$) in all three AlphaFold models ($p < 0.001$). The main effects of the structural variables, and nearly all terms in the smaller experimental dataset, were inconsistent or non-significant.

Table 3.9. Summary of group-level linear model (lm) results. The dependent variable is the log-transformed compensatory group rate (labeled $\log(R_{comp})$ for consistency). Predictors are standardized (z-scored): $(d_N/d_s)_z$ is the d_N/d_S ratio, $GroupRSA_z$ is the mean group RSA, and $CalphaDist_z$ is the mean intragroup C_α-C_α distance. Blank cells indicate non-significant terms removed by model simplification.

		Experimental		AlphaFold	
	Term	$\beta \pm$ SE	p	$\beta \pm$ SE	p
Polarity	$(d_N/d_s)_z$	0.333 ± 0.016	< 0.001	0.317 ± 0.009	< 0.001
	$GroupRSA_z$	−0.022 ± 0.017	0.213	−0.029 ± 0.009	0.001
	$CalphaDist_z$	−0.015 ± 0.017	0.389	−0.012 ± 0.009	0.207
	$(d_N/d_s)_z$:$GroupRSA_z$	0.012 ± 0.017	0.482		
	$(d_N/d_s)_z$:$CalphaDist_z$	−0.006 ± 0.018	0.722	−0.046 ± 0.009	< 0.001
	$GroupRSA_z$:$CalphaDist_z$	0.003 ± 0.017	0.841	0.096 ± 0.009	< 0.001
	Adj. R^2	0.534		0.554	
Volume	$(d_N/d_s)_z$	0.325 ± 0.015	< 0.001	0.334 ± 0.009	< 0.001
	$CalphaDist_z$	−0.027 ± 0.016	0.092	−0.015 ± 0.009	0.096
	$GroupRSA_z$	−0.025 ± 0.016	0.120	−0.024 ± 0.009	0.010
	$(d_N/d_s)_z$:$CalphaDist_z$	0.001 ± 0.016	0.947		
	$(d_N/d_s)_z$:$GroupRSA_z$			−0.021 ± 0.009	0.014
	$CalphaDist_z$:$GroupRSA_z$	0.001 ± 0.016	0.953	0.108 ± 0.009	< 0.001
	Adj. R^2	0.538		0.564	
Charge	$(d_N/d_s)_z$	0.317 ± 0.015	< 0.001	0.350 ± 0.008	< 0.001
	$CalphaDist_z$	−0.017 ± 0.016	0.300	0.001 ± 0.009	0.928
	$GroupRSA_z$	−0.002 ± 0.016	0.884	−0.018 ± 0.008	0.034
	$(d_N/d_s)_z$:$CalphaDist_z$			−0.053 ± 0.009	< 0.001
	$CalphaDist_z$:$GroupRSA_z$	0.020 ± 0.016	0.205	0.094 ± 0.009	< 0.001
	Adj. R^2	0.477		0.569	

In summary, the findings of this chapter establish a multi-layered case for the role of effective population size in shaping intra-protein coevolution. We first validated that the d_N/d_S ratio serves as a robust proxy for N_e in our mammalian dataset. Using this proxy, we then provided the first large-scale evidence that a reduction in N_e is significantly correlated with an increased accumulation of compensatory

mutations. This primary finding was robustly supported as a strong main effect of d_N/d_S across all biochemical properties in our subsequent structural models.

Our final analyses, which dissected this relationship in a structural context, revealed a complex picture. Our models consistently failed to find a clear, generalizable interaction between d_N/d_S and structural constraint; this was true at both the site-level and the group-level. However, the analyses did confirm strong interactions between the structural properties themselves, such as between contact number and RSA, as well as between intragroup distance and RSA. Collectively, these findings forge a clear link between macro-evolutionary population genetics and the micro-evolutionary dynamics of protein evolution, confirming the central role of N_e.

Discussion

The evolution of proteins is driven by the interplay of three fundamental forces: mutation, selection, and genetic drift. A key question in this field is how proteins maintain function in the face of potentially destabilizing mutations. One proposed answer is compensatory evolution, yet the population genetic forces that dictate its prevalence have remained poorly understood (Piskol and Stephan, 2011; Whitlock et al., 2003). This thesis confronts this uncertainty by testing the central hypothesis that effective population size (N_e) is a primary determinant.

By integrating phylogenetic, genomic, and structural data across 78 mammalian species, our research makes three key contributions. First, we validate the branch-specific d_N/d_S ratio as a robust proxy for long-term N_e in mammals. Second, we establish a strong, positive correlation between d_N/d_S and the accumulation of compensatory mutations. Third, our structural analysis investigated the role of biophysical constraints. We demonstrate complex and significant interactions between structural properties themselves, but we found that the hypothesized interaction between N_e and structural constraint was ambiguous and inconsistent across our models.

Collectively, these results offer the first large-scale, genomic evidence that genetic drift is a key driver of compensatory evolution, providing powerful insight into a long-standing theoretical question in population genetics.

4.1 A Robust Framework for Inferring Coevolution

A core principle of this study was the adoption of a methodologically conservative approach, designed to minimize Type I errors (false positive) at the expense of potentially increasing Type II error (false negative). Our primary goal was not to identify all likely coevolving positions, but rather to build a high-confidence dataset of unequivocally compensatory mutations. This stringent philosophy was implemented at every stage of our pipeline.

First we built our dataset using OMA Groups from the Orthologous MAtrix (OMA) database, which employs a notably strict algorithm for orthology inference (Altenhoff et al., 2023). This was followed by a multi-step filtering protocol to ensure data quality. Second, we generated alignments using two distinct, state-

of-the-art methods and retained only positions with a high concordance score, ensuring alignment robustness. Third, our phylogenetic reconstruction also combined multiple approaches, and the resulting species supertree was validated against the current consensus in mammalian phylogenetics (Brady et al., 2024). Finally, this dual-source validation was extended to our structural analyses, where we integrated the highest-resolution experimental structures with predictions from the leading computational model, AlphaFold (Jumper et al., 2021). By demanding that our main conclusions be supported across these multiple, independent lines of evidence and controls, we have ensured that the results presented in this thesis are exceptionally robust. To apply this rigorous framework, we first had to validate our foundational proxy for population size.

4.2 The Efficacy of Selection: Validating a Key Evolutionary Proxy

Although a substantial body of theoretical and empirical work supports the use of d_N/d_S ratio as a proxy for N_e (Brevet and Lartillot, 2021; Figuet et al., 2016; James and Eyre-Walker, 2020), we acknowledge that this approach is not without criticism. Given that this proxy is central in our study's main hypothesis, we took deliberate steps to validate its use within our specific dataset. Our analyses confirmed theoretically expected relationship: we demonstrated a significant positive correlation between d_N/d_S and body mass, a well-established life history trait correlate of N_e, and a significant negative correlation between d_N/d_S and published, species-specific estimates of N_e itself (Damuth, 1981).

While acknowledging the limitations imposed by the sparse availability of body mass and direct N_e data across species, the consistency of these results provides a strong justification for our use of d_N/d_S as reliable, branch-specific proxy for long-term N_e. To further ensure the robustness of these validation analyses, we implemented two distinct methods for phylogenetic correction, accepting correlations as valid only when they remained significant under both statistical framework. With this key proxy confidently established, we could proceed to test our central hypothesis.

4.3 Genetic Drift as a Driver of Compensatory Evolution

Our central finding—that lineages with a smaller N_e and a correspondingly higher d_N/d_S ratio accumulate more compensatory mutations—directly addresses the

theoretical impasse of the "waiting time" problem (Phillips, 1996; Stephan, 1996). The evolution of a compensated genotype requires two events with opposing dependencies on population size: first, the initial fixation of a slightly deleterious mutation, which is more probable under strong genetic drift in small populations (Ohta, 1992); and second, the subsequent fixation of an advantageous compensatory mutation, a process more efficient under strong selection in large populations. Our results strongly suggest that the former step—the persistence of the deleterious intermediate—is the rate-limiting factor in this process. In lineages with a small N_e, the increased power of genetic drift allows a greater number of slightly deleterious substitutions to fix, creating a larger substrate of destabilized proteins upon which compensatory mutations can subsequently arise and be selected.

This finding provides empirical support for a key component of Sewall Wright's Shifting Balance Theory (SBT) (Wright, 1931), a concept debated for nearly a century (Coyne et al., 1997). Our work illustrates a plausible molecular mechanism for "valley crossing" on a fitness landscape, positioning genetic drift not merely as a stochastic, non-adaptive force, but as a creative one that can open new evolutionary trajectories. By allowing populations to fix otherwise inaccessible intermediate genotypes, drift can unlock novel adaptive peaks. However, it is crucial to note that while our findings support the first phase of SBT (drift enabling a shift to a new state), this study does not address the other required phases, such as interdemic selection and the spread of the new adaptive combination, which rely on a specific metapopulation structure not tested here.

4.4 A Mechanistic Link: How Structural Constraints Modulate Coevolution

Beyond establishing that genetic drift drives compensatory mutations, our study sought to understand how this process is mediated by the biophysical constraints of protein structure. Our structural models, which incorporated data from both experimental and AlphaFold structures, provided several crucial mechanistic insights.

First, these models confirm the robustness of our central finding: the d_N/d_S ratio remained the most powerful positive predictor of the compensatory mutation rate across all tested conditions. This indicates that the macro-evolutionary signal of N_e is not an artifact, but persists even when controlling for fine-scale structural properties.

Second, our hypothesis–that the effect of N_e would be "intensified" in structurally constrained regions—was not clearly supported by our models. As detailed in the Results (Section 3.3.1), the site-level interaction terms between d_N/d_S and structural predictors were inconsistent across datasets and biochemical properties,

yielding ambiguous results. Our subsequent group-level analysis (Section 3.3.2) likewise failed to find a consistent signal; while a significant interaction between d_N/d_S and intragroup distance appeared in the larger AlphaFold dataset for two properties, it was not present in the experimental dataset or for the Volume property. Thus, a clear, generalizable interaction between N_e and structural constraint remains elusive in our data. However, while this specific interaction was unclear, the structural analyses did uncover two other, more robust findings. Our models consistently identified a strong, negative interaction between the number of contacts and solvent accessibility ($NbContact_z : RSA_z$). This highly significant finding implies that the effects of being buried and being highly connected are not merely additive. This negative coefficient statistically distinguishes between two classes of buried residues: (i) the compact structural core (low RSA, high contacts), which is associated with a higher rate of compensation, and (ii) buried pockets or cavities (low RSA, low contacts), which are associated with a lower rate. This latter group likely represents functionally critical, and thus more highly conserved, regions such as active sites.

Finally, our analysis of the spatial arrangement of coevolving sites supports a mechanism of local, rather than direct, stabilization. We found that residues within a compensatory group are typically in close spatial proximity (median distances of 16.0Å-20.2Å) but are not in direct physical contact (i.e., > 5Å). This suggests that compensation operates not by simple one-to-one side-chain replacement, but by restoring local structural stability or mediating short-range allosteric communication. This interpretation is strongly supported by the consistent, positive interaction we found between mean group RSA and intragroup distance ($RSA_z : Dist_z$) in our group-level model (Table 3.9). The high rate of compensation observed at both low-RSA/low-Distance (buried and close) and high-RSA/high-Distance (exposed and distant) sites strongly points to these two distinct mechanisms: local core stabilization and long-range surface allostery.

Collectively, these latter findings provide a powerful mechanistic explanation for how and where proteins accommodate the destabilizing mutations that fix more readily in small populations.

4.5 Implications

The implications of this work extend beyond theoretical population genetics. Although our study focuses on mammals, the principles linking population bottlenecks to compensatory evolution are broadly applicable and offer a potential explanation for the stubborn persistence of antimicrobial resistance (AMR) in pathogens. While resistance-conferring mutations often impose a fitness cost (Melnyk et al., 2015),

our results highlight a powerful counter-mechanism. The severe genetic bottlenecks induced by antibiotic treatment create ideal conditions—a transient period of strong genetic drift—for the fixation of costly resistance alleles. Subsequently, compensatory mutations that restore fitness can arise and sweep through the recovering population (Levin et al., 2000; Schulz zur Wiesch et al., 2010), creating resistant lineages that are evolutionarily robust and difficult to eradicate. This dynamic provides a critical insight for managing the AMR crisis.

This work also reinforces the foundational principles of modern protein structure prediction. Tools like AlphaFold achieve their remarkable accuracy by leveraging the very coevolutionary signals we analyzed (Jumper et al., 2021). Our research provides a population-genetic basis for the origin of these signals, confirming that the most predictive information for determining a protein's 3D structure comes from regions under the most intense structural and functional constraint.

4.6 Limitations

Despite the robustness of our findings, this study has several limitations that should be acknowledged. Our dataset is confined to 78 mammalian species and is taxonomically biased—for example, with an over-representation of primates. Consequently, while our results are consistent within this group, their generalizability to all mammals—let alone other taxonomic clades—cannot be directly inferred and awaits further investigation. Another source of uncertainty lies in our use of the d_N/d_S ratio as a proxy for long-term effective population size (N_e) is inherently indirect. This metric averages evolutionary dynamics over millions of years, which can obscure the effects of more recent or short-term population changes. Furthermore, d_N/d_S can be influenced by factors other than genetic drift, such as fluctuating selective pressures over time. The detection of compensatory mutations, while using a state-of-the-art method (CoMap), is fundamentally dependent on the accuracy of ancestral sequence reconstruction, which carries inherent uncertainty. Additionally, our model does not account for "super-compensation", a known phenomenon where a compensated genotype achieves higher fitness than the original, unmutated ancestor. Furthermore, our structural analyses were performed on static 3D models. This approach does not capture the dynamic nature of proteins. Crucial factors like conformational flexibility and long-range allostery, which may play a significant role in mediating compensatory effects, are therefore not accounted for in our analysis.

Finally, while our results are consistent with a causal relationship, the study is correlational. Our interpretation is that strong genetic drift drives the accumulation of compensatory mutations. However, an alternative hypothesis could involve

a confounding factor. For example, a shift in the distribution of fitness effects (DFE) (Eyre-Walker and Keightley, 2007) within certain lineages could theoretically increase both the fixation of slightly deleterious mutations (raising d_N/d_S) and the opportunity for compensation. Although the broad scale of our genome-wide analysis makes this alternative less probable, we emphasize that our findings represent a robust statistical association rather than a direct demonstration of causality.

These limitations naturally point toward several avenues for future research. As methods for inferring ancestral population sizes become more refined, future studies could revisit this question with more direct and time-resolved estimates of N_e. Expanding this analytical framework to other major clades, would reveal whether the observed dynamic is a universal principle of molecular evolution. On a finer scale, integrating molecular dynamics simulations could provide a more dynamic view of how compensatory mutations restore protein stability and function. Ultimately, the hypotheses generated by this large-scale correlative study could be tested directly through laboratory-based directed evolution experiments, where population size can be precisely controlled to observe the rate and nature of compensatory evolution in real time.

In conclusion, this thesis forges a direct, evidence-based link between a macro-evolutionary parameter (N_e) and a micro-evolutionary molecular process (compensatory mutation). By demonstrating that genetic drift fuels the accumulation of compensatory changes, particularly within the protein core, we provide a comprehensive, multi-scale explanation for a fundamental evolutionary dynamic.

Bibliography

Achaz, G. and Dutheil, J. (2021). Correlated evolution: Models and methods.

Altenhoff, A. M., Warwick Vesztrocy, A., Bernard, C., Train, C.-M., Nicheperovich, A., Prieto Baños, S., Julca, I., Moi, D., Nevers, Y., Majidian, S., Dessimoz, C., and Glover, N. M. (2023). Oma orthology in 2024: improved prokaryote coverage, ancestral and extant go enrichment, a revamped synteny viewer and more in the oma ecosystem. *Nucleic Acids Research*, 52(D1):D513–D521.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences*, 114(34):9122–9127.

Arditti, Joseph., Elliot, John., Kitching, I. J., and Wasserthal, L. T. (2012). 'Good Heavens what insect can suck it' – Charles Darwin, *Angraecum Sesquipedale* and *Xanthopa Morganii Praedicta*. *Botanical Journal of the Linnean Society*, 169(3):403–432.

Ashenberg, O. and Laub, M. T. (2013). Chapter Nine - Using Analyses of Amino Acid Coevolution to Understand Protein Structure and Function. In Keating, A. E., editor, *Methods in Enzymology*, volume 523 of *Methods in Protein Design*, pages 191–212. Academic Press.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

Bininda-Emonds, O. R. P. (2004). The evolution of supertrees. *Trends in Ecology & Evolution*, 19(6):315–322.

Brady, P. L., Castrellon Arteaga, A., López-Torres, S., and Springer, M. S. (2024). The effects of ordered multistate morphological characters on phylogenetic analyses of eutherian mammals. *Journal of Mammalian Evolution*, 31(3):28.

Brevet, M. and Lartillot, N. (2021). Reconstructing the History of Variation in Effective Population Size along Phylogenies. *Genome Biology and Evolution*, 13(8):evab150.

Burger, L. and van Nimwegen, E. (2010). Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLOS Computational Biology*, 6(1):e1000633.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205.

Chaurasia, S. and Dutheil, J. Y. (2022). The Structural Determinants of Intra-Protein Compensatory Substitutions. *Molecular Biology and Evolution*, 39(4):msac063.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Coyne, J. A., Barton, N. H., and Turelli, M. (1997). Perspective: A Critique of Sewall Wright's Shifting Balance Theory of Evolution. *Evolution*, 51(3):643–671.

Damuth, J. (1981). Population density and body size in mammals. *Nature*, 290(5808):699–700.

Davis, B. H., Poon, A. F., and Whitlock, M. C. (2009). Compensatory mutations are repeatable and clustered within proteins. *Proceedings of the Royal Society B: Biological Sciences*, 276(1663):1823–1827.

De Magalhães, J. P. and Costa, J. (2009). A database of vertebrate longevity records and their relation to other life-history traits. *Journal of Evolutionary Biology*, 22(8):1770–1774.

Dutheil, J. and Galtier, N. (2007). Detecting groups of coevolving positions in a molecule: A clustering approach. *BMC Evolutionary Biology*, 7(1):242.

Dutheil, J., Pupko, T., Jean-Marie, A., and Galtier, N. (2005). A model-based approach for detecting coevolving positions in a molecule. *Molecular Biology and Evolution*, 22(9):1919–1928.

Dutheil, J. Y. (2012). Detecting coevolving positions in a molecule: Why and how to account for phylogeny. *Briefings in Bioinformatics*, 13(2):228–243.

Dutheil, J. Y., Hamidi, D., and Pajot, B. (2024). The Site/Group Extended Data Format and Tools. *Genome Biology and Evolution*, 16(2):evae011.

Dutheil, J. Y., Jossinet, F., and Westhof, E. (2010). Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Molecular Biology and Evolution*, 27(8):1868–1876.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618.

Faurby, S., Davis, M., Pedersen, R. Ø., Schowanek, S. D., Antonelli1, A., and Svenning, J.-C. (2018). PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology. *Ecology*, 99(11):2626–2626.

Feeney, W. P. (2012). Chapter 35 - The Chinese or Striped-Back Hamster. In Suckow, M. A., Stevens, K. A., and Wilson, R. P., editors, *The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents*, American College of Laboratory Animal Medicine, pages 907–922. Academic Press, Boston.

Figuet, E., Nabholz, B., Bonneau, M., Mas Carrio, E., Nadachowska-Brzyska, K., Ellegren, H., and Galtier, N. (2016). Life History Traits, Protein Evolution, and the Nearly Neutral Theory in Amniotes. *Molecular Biology and Evolution*, 33(6):1517–1527.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 326(1233):119–157.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, 185(4154):862–864.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C., Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K., and Dutheil, J. Y. (2013). Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution. 30(8):1745–1750.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1995). *Multivariate data analysis (4th ed.): with readings*. Prentice-Hall, Inc., USA.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638.

James, J. and Eyre-Walker, A. (2020). Mitochondrial DNA Sequence Diversity in Mammals: A Correlation between the Effective and Census Population Sizes. *Genome Biology and Evolution*, 12(12):2441–2449.

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib, M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., Teacher, A., Bininda-Emonds, O. R. P., Gittleman, J. L., Mace, G. M., and Purvis, A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, 90(9):2648–2648.

Joshi, A. (1999). The Shifting Balance Theory of Evolution. *Resonance*, 4(12):66–75.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetics Research*, 11(3):247–270.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15):7176–7180.

Kosiol, C., Holmes, I., and Goldman, N. (2007). An Empirical Codon Model for Protein Sequence Evolution. *Molecular Biology and Evolution*, 24(7):1464–1479.

Le, S. Q., Dang, C. C., and Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution*, 29(10):2921–2936.

Le, S. Q. and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.

Levin, B. R., Perrot, V., and Walker, N. (2000). Compensatory Mutations, Antibiotic Resistance and the Population Genetics of Adaptive Evolution in Bacteria. *Genetics*, 154(3):985–997.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Martins, E. P. and Hansen, T. F. (1997). Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *The American Naturalist*, 149(4):646–667.

Melnyk, A. H., Wong, A., and Kassen, R. (2015). The fitness costs of antibiotic resistance mutations. *Evolutionary Applications*, 8(3):273–283.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Mirarab, S., Reaz, R., Bayzid, Md. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Mohan, S., Ozer, H. G., and Ray, W. C. (2022). The Importance of Weakly Co-Evolving Residue Networks in Proteins is Revealed by Visual Analytics. *Frontiers in Bioinformatics*, 2.

Mugal, C. F., Wolf, J. B., and Kaj, I. (2014). Why Time Matters: Codon Evolution and the Temporal Dynamics of dN/dS. *Molecular Biology and Evolution*, 31(1):212–231.

Myers, P., Espinosa, R., Parr, C. S., Jones, T., Hammond, G. S., and Dewey, T. A. (2025). The animal diversity web. `https://animaldiversity.org`. Accessed 2025-06-25.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–98.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, 23:263–286.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution*, 40(1):56–63.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C., and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298.

Phillips, P. C. (1996). Waiting for a compensatory mutation: phase zero of the shifting-balance process. *Genetics Research*, 67(3):271–283.

Phipson, B. and Smyth, G. K. (2010). Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Article39.

Piskol, R. and Stephan, W. (2011). The Role of the Effective Population Size in Compensatory Evolution. *Genome Biology and Evolution*, 3:528–538.

Pollock, D. D., Taylor, W. R., and Goldman, N. (1999). Coevolving protein residues: Maximum likelihood identification and relationship to structure1. *Journal of Molecular Biology*, 287(1):187–198.

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). Macse v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, 35(10):2582–2584.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Sander, P., Springer, B., Prammananan, T., Sturmfels, A., Kappler, M., Pletschette, M., and Böttger, E. C. (2002). Fitness Cost of Chromosomal Drug Resistance-Conferring Mutations. *Antimicrobial Agents and Chemotherapy*, 46(5):1204–1211.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., and Karsch-Mizrachi, I. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, 2020:baaa062.

Schulz zur Wiesch, P., Engelstädter, J., and Bonhoeffer, S. (2010). Compensation of Fitness Costs and Reversibility of Antibiotic Resistance Mutations. *Antimicrobial Agents and Chemotherapy*, 54(5):2085–2095.

Shindyalov, I. N., Kolchanov, N. A., and Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering*, 7(3):349–358.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. (2005). On the Meaning and Existence of an Effective Population Size. *Genetics*, 169(2):1061–1070.

Stephan, W. (1996). The Rate of Compensatory Evolution. *Genetics*, 144(1):419–426.

The UniProt Consortium (2025). UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617.

Thompson, J. D., Plewniak, F., and Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic acids research*, 27(13):2682–2690.

Van Valen, L. (1973). A new evolutionary law. *Foundations of Macroecology; University of Chicago Press: Chicago, IL, USA*, pages 284–314.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., and Velankar, S. (2024). AlphaFold Protein Structure Database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375.

White, E. P., Ernest, S. K. M., Kerkhoff, A. J., and Enquist, B. J. (2007). Relationships between body size and abundance in ecology. *Trends in Ecology & Evolution*, 22(6):323–330.

Whitlock, M. C., Griswold, C. K., and Peters, A. D. (2003). Compensating for the meltdown: The critical effective size of a population with deleterious and compensatory mutations. *Annales Zoologici Fennici*, 40(2):169–183.

Wilder, A. P., Supple, M. A., Subramanian, A., Mudide, A., Swofford, R., Serres-Armero, A., Steiner, C., Koepfli, K.-P., Genereux, D. P., Karlsson, E. K., Lindblad-Toh, K., Marques-Bonet, T., Munoz Fuentes, V., Foley, K., Meyer, W. K., Zoonomia Consortium, Ryder, O. A., and Shapiro, B. (2023). The contribution of historical processes to contemporary extinction risk in placental mammals. *Science*, 380(6643):eabn5856.

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., and Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nature Genetics*, 32(4):569–577.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, 16(2):97–159.

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261–1295.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503.

Yang, Z., Nielsen, R., and Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, 15(12):1600–1611.

Yeang, C.-H. and Haussler, D. (2007). Detecting Coevolution in and among Protein Domains. *PLOS Computational Biology*, 3(11):e211.

Zhang, C. and Mirarab, S. (2022). Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Molecular Biology and Evolution*, 39(12):msac215.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. 19(6):153.

Zhang, L. and Watson, L. T. (2009). Analysis of the fitness effect of compensatory mutations. *HFSP Journal*, 3(1):47–54.

Supplementary Materials

A.1 Implementation and Data Availability

Data manipulation, parsing, and the execution of other software tools were performed using Python (v3.12.2) and Bash (v5.2.21) scripts. Data visualization and statistical analyses were carried out using R (v4.4.2).

Computationally intensive analyses were performed on the Raven and Viper HPC clusters at the Max Planck Computing and Data Facility (MPCDF; `https://www.mpcdf.mpg.de/services/supercomputing/`).

To ensure the reproducibility of the entire workflow, all programs, dependencies, and the specific Python and R environments are available and distributed as Apptainer containers.

The complete code and documentation are available in the following repository: `https://github.com/lorenzopenone/coevolution`.

A.2 Supplementary Tables

Table A.1. Summary of the databases utilized in this study, including version numbers or access dates and corresponding URL.

Database	Version / Accession Data	URL
OMA DB	All.Jul2024 release	`https://omabrowser.org/oma/home/`
NCBI Taxonomy	Acc. 27 May 2025	`https://www.ncbi.nlm.nih.gov/taxonomy/`
AnAge	Acc. 16 June 2025	`https://genomics.senescence.info/species/index.html`
PanTHERIA	v1.0_WR05_Aug208	`https://esapubs.org/archive/ecol/E090/184/`
PHYLACINE 1.2	Acc. 16 June 2025	`https://megapast2future.github.io/PHYLACINE_1.2/`
Animal Diversity Web	Acc. 16 June 2025	`https://animaldiversity.org/`
UniProt	Release 2025_03	`https://www.uniprot.org/`
RCSB Protein Data Bank	Acc. 17 July 2025	`https://www.rcsb.org/`
AlphaFold Protein Structure Database	Acc. 2 September 2025	`https://alphafold.ebi.ac.uk/`

A.3 Supplementary Figures

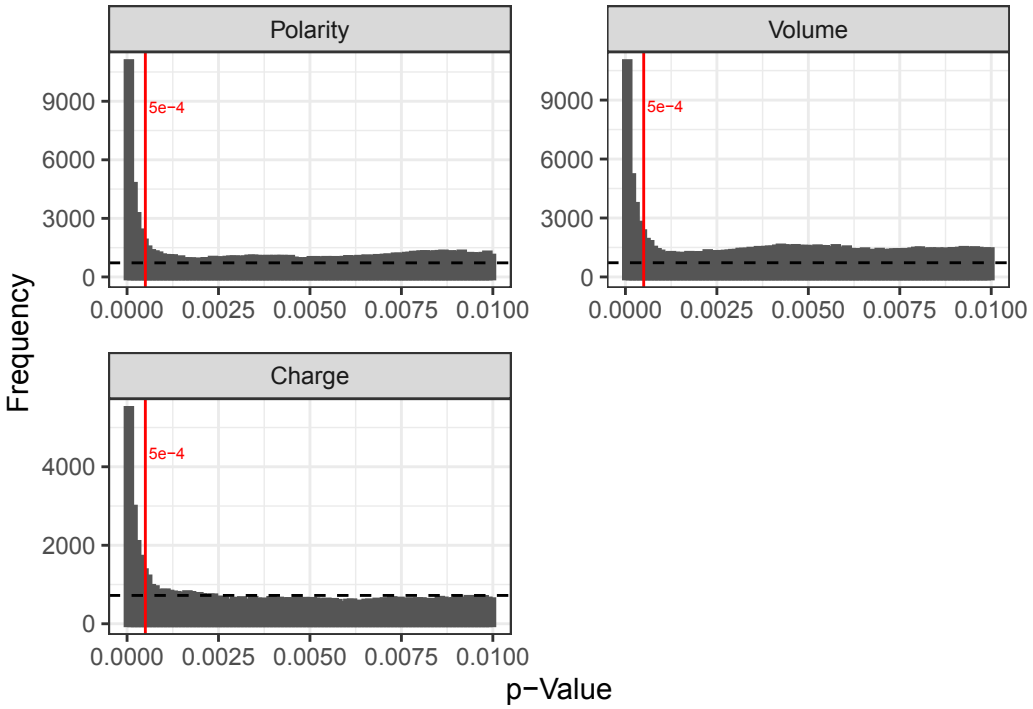

Figure A.1. Empirical p-value distributions for the three biochemical properties (polarity, volume, charge). Red vertical lines mark the data-driven significance cutoffs described in the text.

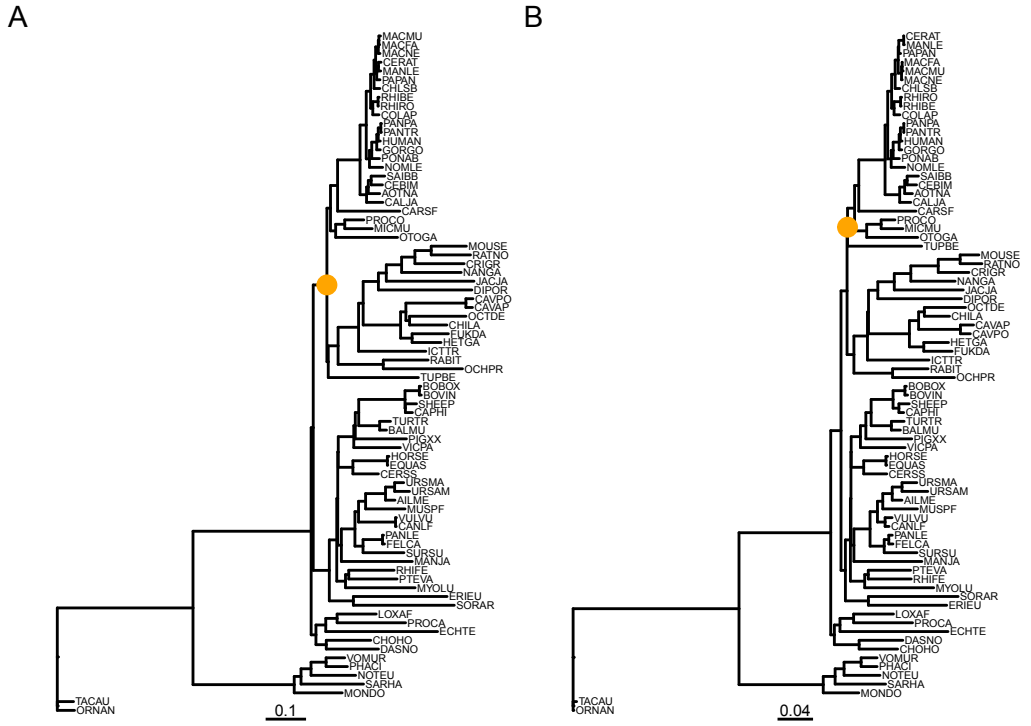

Figure A.2. Comparison of species supertrees inferred from codon and amino acid Data. The figure displays the two mammalian supertrees reconstructed for this study. (A) The supertree inferred from the concatenation of codon-based gene trees. (B) The supertree inferred from the concatenation of amino acid-based gene trees. The two topologies are highly congruent, with a Robinson-Foulds distance of 2, indicating a single point of conflict. This node, representing the common ancestor of Scandentia, Primates, and Glires, is highlighted with an orange circle. This single incongruence was resolved by collapsing the node into a trifurcation to generate the consensus species tree used for all downstream analyses.

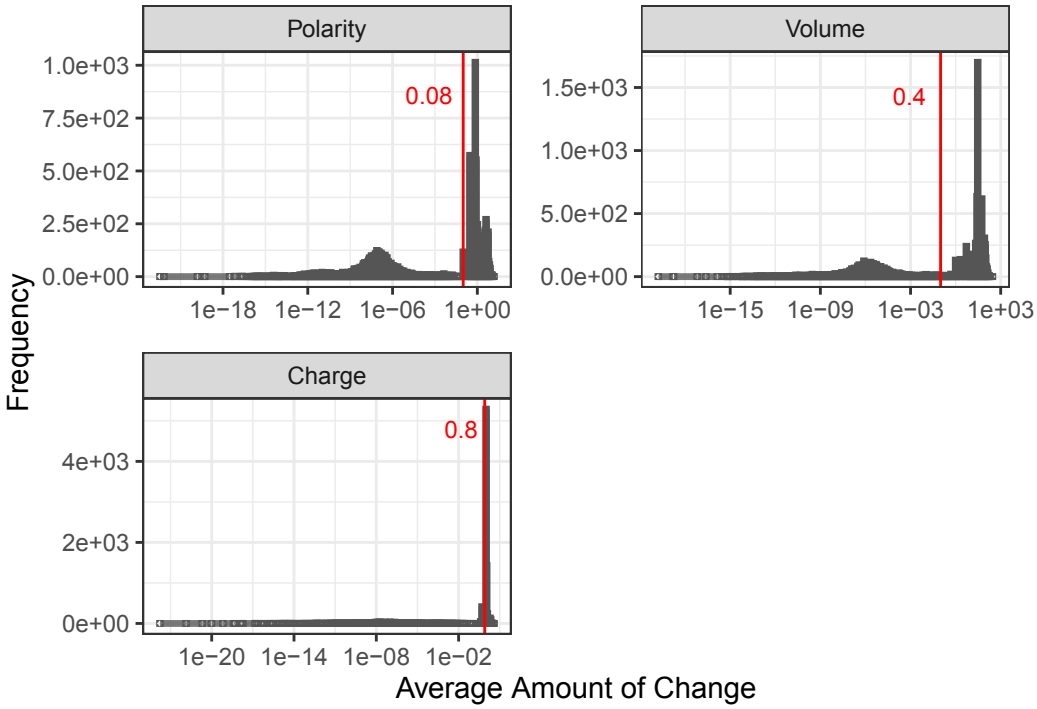

Figure A.3. Distribution of the average change for three biochemical properties (polarity, volume, and charge) in the unfiltered dataset. The red vertical line indicates the threshold for a significant change. This threshold was calculated as 80% of the smallest non-zero Grantham distance between any two amino acids, accounting for a 20% uncertainty.

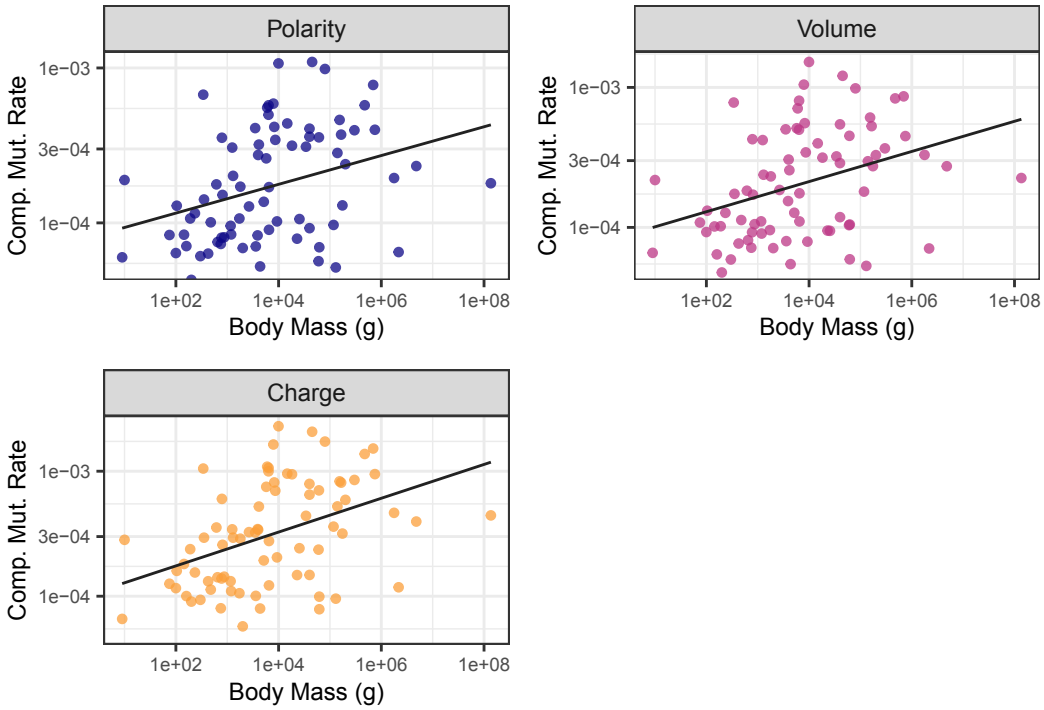

Figure A.4. Relationship between body mass and the normalized frequency of compensatory mutations for extant species. Each panel corresponds to a different biochemical property. All axes are log10-scaled. The positive trend confirms that species with larger body mass (a proxy for smaller N_e) accumulate more compensatory mutations.

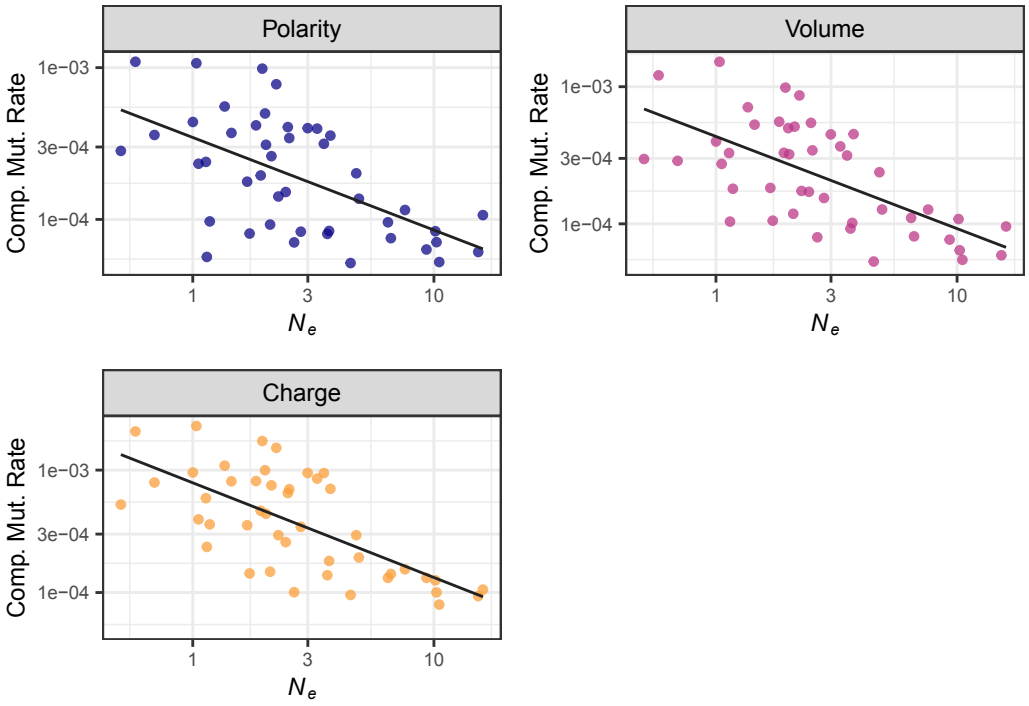

Figure A.5. Relationship between effective population size (N_e) and the normalized frequency of compensatory mutations. Each panel corresponds to a different biochemical property. All axes are log10-scaled. The negative trend provides direct evidence that smaller populations accumulate more compensatory mutations.

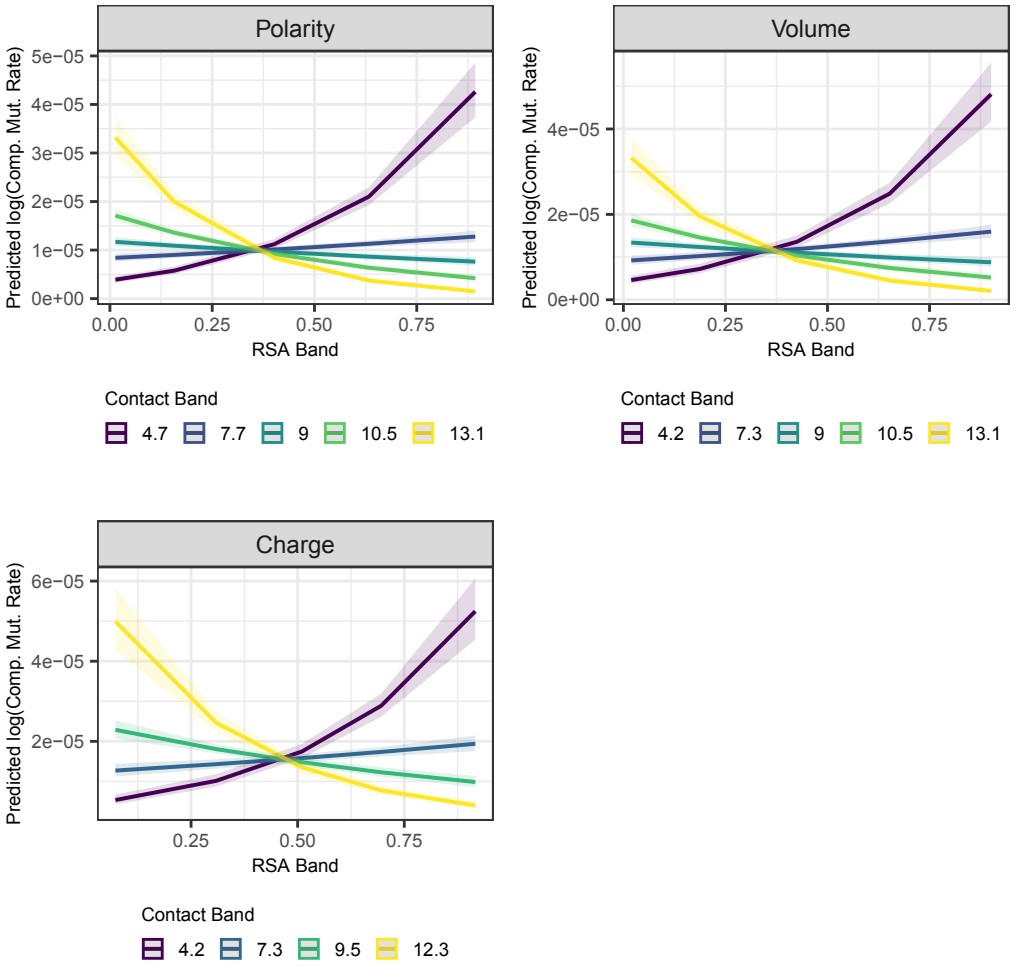

Figure A.6. Interaction effect between RSA and number of contact on the predicted compensatory mutation rate in the AlphaFold dataset. The plot is divided into three panels showing the results from models for polarity , volume , and charge. Each panel visualizes the predicted log(Comp. Mut. Rate) (Y-axis) as a function of RSA Band (X-axis) at different levels of Contact Band (colored lines). The plot reveals a strong crossover interaction, which is consistent across all three models. At low RSA values (e.g., 0.00), sites with a high Contact Band (bright yellow lines) are predicted to have a higher compensatory mutation rate. This relationship inverts at high RSA values (e.g., 0.75), where sites with a low Contact Band (dark purple lines) show the highest predicted compensatory mutation rate. This suggests that the effect of solvent exposure (RSA) on the mutation rate is highly dependent on the site's contact level.