Protein evolution and data-driven sequence landscapes

Matteo Bisardi

Université Paris Cité

Université Paris Cité

École Doctorale Physique en Île-de-France (ED 564)

Laboratoire de Physique de l'ENS

Protein evolution and data-driven sequence landscapes

Thèse de doctorat de Physique

Réalisée par

Matteo BISARDI

Dirigée par

Francesco ZAMPONI	Directeur
Martin WEIGT	Co-directeur

Présentée et soutenue publiquement le
29/11/2023

Devant un jury composé de:

Richard NEHER	Professeur associé	University of Basel	Rapporteur
Annalisa PASTORE	Professeur	King's College London	Rapporteuse
Alessandra CARBONE	Professeur	Sorbonne Université	Examinatrice
Guillaume ACHAZ	Professeur	Université Paris Cité	Membre invité
Francesco ZAMPONI	Professeur associé	Ecole Normale Supérieure	Directeur
Martin WEIGT	Professeur	Sorbonne Université	Co-directeur

"Le mouvement évolutif serait chose simple, nous aurions vite fait d'en déterminer la direction, si la vie décrivait une trajectoire unique, comparable à celle d'un boulet plein lancé par un canon. Mais nous avons affaire ici à un obus qui a tout de suite éclaté en fragments, lesquels, étant eux-mêmes des espèces d'obus, ont éclaté à leur tour en fragments destinés à éclater encore, et ainsi de suite pendant fort longtemps. "

Henry Bergson

ABSTRACT

Title: Protein evolution and data-driven sequence landscapes

Abstract: Thanks to the explosion of available protein sequence data, driven by next-generation sequencing, unsupervised machine learning models can be harnessed to learn protein sequence landscapes. Specifically, methods like Direct Coupling Analysis (DCA) directly consider the patterns of conservation and coevolution between protein sites. DCA has been applied to many biological problems, from predicting the fitness effects of mutations to artificial sequence generation. In this thesis, we broaden the application of DCA to study protein evolution, the dynamical process of amino acid substitutions in protein sequences driven by random mutations and natural selection. We particularly focus on the effects and implications of epistasis, which refers to the context-dependence of mutational effects. Our goal is to design and test various dynamical algorithms to navigate the data-driven sequence landscape inferred by DCA, with a focus on beta-lactamase enzyme families. Beta-lactamases are enzymes capable of degrading many commonly prescribed antibiotics posing a serious threat to health systems worldwide. Initially, we modeled two recently published neutral drift protein evolution experiments that used beta-lactamases as starting sequences. We demonstrate that our artificially generated libraries reproduce well the statistics of the experimental ones by introducing an evolutionary dynamics that relies on the fitness landscape inferred by DCA and accounts for the degeneracy of the genetic code. Computationally, we explore the influence of different experimental parameters, highlighting a trade-off between the number of experimental rounds and the sequencing depth when trying to elucidate epistatic constraints. Thanks to a simple modification of the sampling algorithm, we manage to model proteins' evolutionary dynamics across timescales, spanning from dozens of years to eons. For the first time, this approach allows us to quantitatively model the predictability of mutational effects as sequences diverge, a phenomenon driven by epistasis. Subsequently, we examine the deep mutational scanning experiments of two other beta-lactamase enzymes, VIM-2 and NDM-1. We analyze the role and prevalence of context-dependent mutability by integrating experimental data with DCA predictions and we propose that structural regions with intermediate solvent exposure exhibit the greatest variability in mutational effects. Furthermore, we computationally characterize mutational heterogeneities across the entire enzyme family. We then develop and test a straightforward approach to produce mutational paths between VIM-2 and NDM-1, successfully obtaining functional proteins with numerous mutations and uncovering the presence of unexpected epistatic interactions.

Keywords: Direct Coupling Analysis, generative models, epistasis, protein evolution, neutral drift experiment, Deep Mutational Scanning, beta-lactamases

RÉSUMÉ

Titre: Évolution des protéines et paysages de séquences guidés par les données

Résumé: Grâce à l'explosion des données disponibles sur les séquences de protéines, alimentée par le séquençage de nouvelle génération, les modèles d'apprentissage automatique non supervisé peuvent être exploités pour apprendre les paysages de séquence de protéines. Plus précisément, des méthodes comme Direct Coupling Analysis (DCA) considèrent directement les motifs de conservation et de coévolution entre les sites protéiques. La DCA a été appliquée à de nombreux problèmes biologiques, allant de la prédiction des effets de la fitness des mutations à la génération de séquences artificielles. Dans cette thèse, nous élargissons l'application de la DCA pour étudier l'évolution des protéines, le processus dynamique des substitutions d'acides aminés dans les séquences de protéines, entraîné par des mutations aléatoires et la sélection naturelle. Nous nous concentrons particulièrement sur les effets et les implications de l'épistasie, qui fait référence à la dépendance contextuelle des effets mutationnels. Notre objectif est de concevoir et de tester différents algorithmes dynamiques pour naviguer dans le paysage de séquence guidé par les données, inféré par la DCA, en mettant l'accent sur les familles d'enzymes beta-lactamase. Les bêta-lactamases sont des enzymes capables de dégrader de nombreux antibiotiques couramment prescrits, représentant une menace sérieuse pour les systèmes de santé du monde entier. Au départ, nous avons modélisé deux expériences d'évolution de protéines de dérive neutre récemment publiées qui utilisaient des bêta-lactamases comme séquences de départ. Nous démontrons que nos bibliothèques générées artificiellement reproduisent bien les statistiques des expériences en introduisant une dynamique évolutive qui repose sur le paysage de fitness inféré par la DCA et prend en compte la dégénérescence du code génétique. Computationnellement, nous explorons l'influence de différents paramètres expérimentaux, mettant en évidence un compromis entre le nombre de tours expérimentaux et la profondeur de séquençage lors de la tentative d'élucider les contraintes épistatiques. Grâce à une simple modification de l'algorithme d'échantillonnage, nous parvenons à modéliser la dynamique évolutive des protéines à travers les échelles de temps, allant de dizaines d'années à des éons. Pour la première fois, cette approche nous permet de modéliser quantitativement la prédictibilité des effets mutationnels à mesure que les séquences divergent, un phénomène entraîné par l'épistasie. Par la suite, nous examinons les expériences de deep mutational scanning de deux autres enzymes bêta-lactamase, VIM-2 et NDM-1. Nous analysons le rôle et la prévalence de la mutabilité dépendante du contexte en

intégrant les données expérimentales avec les prédictions de la DCA et nous proposons que les régions structurales avec une exposition intermédiaire au solvant présentent la plus grande variabilité des effets mutationnels. De plus, nous caractérisons computationnellement les hétérogénéités mutationnelles à travers toute la famille d'enzymes. Nous développons ensuite et testons une approche simple pour produire des voies mutationnelles entre VIM-2 et NDM-1, obtenant avec succès des protéines fonctionnelles avec de nombreuses mutations et révélant la présence d'interactions épistatiques inattendues.

Mots-clés: Direct Coupling Analysis, modèles génératifs, épistasie, évolution des protéines, expérience de dérive neutre, Deep Mutational Scanning, bêta-lactamases

Cerca di vivere armoniosamente, gli eccessi, credo,
ci allontanano vertiginosamente dalla salute.

— Mamma, 1996

ACKNOWLEDGEMENTS

Ogni sforzo umano richiede intuizione, intenzione, dedizione. Per quanto mi riguarda, la scrittura di questa tesi ha messo a dura prova ciascuno di questi aspetti, oltre quanto mi sarei mai immaginato. Eppure sono riuscito ad arrivare in fondo, e questo lo devo alle persone che mi hanno accompagnato, sostenuto, e amato durante il percorso.

Voglio ringraziare innanzitutto i miei supervisors, Francesco e Martin. Grazie per essere stati sempre presenti quando avevo bisogno di sostegno. Grazie per avermi fatto viaggiare. Grazie per avermi lasciato la libertá di lavorare come volevo, dove volevo, su ciò che volevo. Grazie ai miei colleghi e compagni di viaggio, che hanno reso leggere le giornate pesanti, interessanti le conferenze noiose e divertenti le autostrade buie[1]. Grazie Bart, Edo, Francesco, Jeanne, Kai, Luca, Maureen, Roberto. Grazie Sabrina, per aver condiviso cosí tanto in cosí poco. Grazie agli amici della magistrale con cui sono arrivato a Parigi: Checco, Chiare, Francesca, Francesco, Giulio, Matte e le nuove arrivate Camilla, Gaia e Justine. Grazie per essere stati una presenza costante, una garanzia di risate, un luogo sicuro. Grazie a Chiare e Matte per avermi ospitato quando io, di morbi, ne avevo ben due. Grazie di nuovo a Chiare per la strada percorsa assieme fino al dottorato e per essermi stata amica e vicina anche dopo, nonostante tutto. Grazie a tutti gli amici che ho conosciuto poi a Parigi, grazie per le avventure vissute assieme, le serate lunghe, la musica in compagnia. Grazie ad Aurora, Claudia, Cama, Eugenio, Filippo, François, Maura, Gabri, Ginevra, Rebecca, Silvia e tutti gli altri e le altre con cui ho passato una sola serata o un intero semestre in questi tre anni stupendi, turbolenti, esagerati. Grazie a Cris per avermi mostrato una dimensione nuova del vivere. Grazie a Chiara Fil e Lollo per essere diventati casa. Grazie a Fra per avermi insegnato che la samba cura tutti i mali. Grazie a Chiara Martini per ridere cosí tanto di te, di me, con me. Grazie agli amici di Verona, sono ció che sono anche perché siamo cresciuti assieme.

Grazie Angelica, per avermi fatto scoprire posti nuovi dentro di me e fuori nel mondo. Grazie per avermi fatto innamorare. Grazie per avermi amato anche nei miei momenti più oscuri. Grazie alla mia terapista per avermi fermamente teso una mano ogni volta che mi serviva. Grazie infine alla mia famiglia. Grazie Tommaso, grazie Papà, grazie Mamma. Il vostro amore ed il vostro sostengno non si possono misurare, e quindi nemmeno ringraziare per intero. Grazie per essere il punto di partenza e di ritorno, grazie per avermi dato i mezzi umani e materiali per arrivare fin qui, grazie per il vostro amore.

[1] a Cuba

CONTENTS

1 INTRODUCTION

In this chapter, we present an overview of the biological and theoretical concepts essential for understanding the thesis. We will introduce more specific topics individually in the subsequent chapters as they become relevant. Section 1.1 offers an elementary introduction to protein biology. Section 1.2 provides some basic concepts about protein evolution, spanning from the concept of protein space to that of sequence alignments. Section 1.3 discusses the phenomenon of antibiotic resistance in bacteria, specifically focusing on the role of beta-lactamase enzymes that confer resistance to β-lactam antibiotics. A significant portion of the research presented in this thesis deals with those enzymes as model systems. In section 1.4 one of the most central topics of my thesis is introduced: epistasis. The term refers to the dependence on the amino acid context of the mutational effects in proteins. Section 1.5 gives a brief overview of some of the most recent experimental techniques and approaches to study fitness landscapes and protein evolution. Finally, section 1.6 offers a theoretical depiction of the Potts model, which we use as the primary tool for computationally modeling protein sequence landscapes and studying evolution.

1.1 PROTEIN BIOLOGY

Proteins are fundamental biomolecules present in all living organisms, serving as the vital building blocks of life. They facilitate a wide array of essential functions necessary for the survival and operation of biological systems. Among the roles that they execute, proteins are responsible for structural support, accelerating chemical reactions, and acting as signaling agents to mediate communication between cells. In its most essential form, a protein is a linear polypeptide chain made up of smaller monomeric sub-units, also referred to as residues, which are chemically known as amino acids.

1.1.1 Amino acids

There are 20 amino acids encoded by the standard genetic code, plus 2 additional ones - selenocysteine and pyrrolysine - that can be incorporated by non-standard translation mechanisms. Each amino acid is associated with a standard letter. Chemically speaking, amino acids

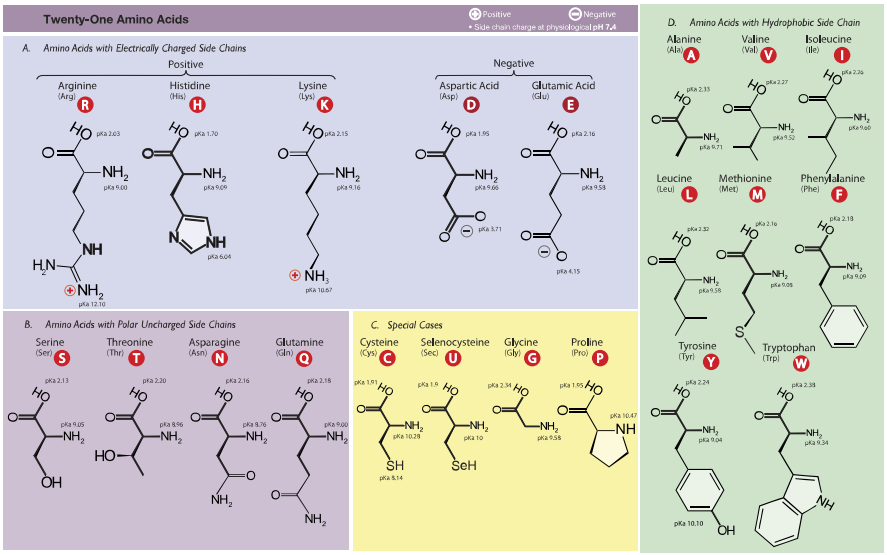


Figure 1.1 – Chemical structure of the 20 standard amino acids found in eukaryotes, plus selenocysteine. Pyrrolysine is not depicted owing to its rarity. Image by Bert Hubert, CC BY-SA 4.0 License.

are organic compounds constituted of two main elements: a fixed common core and a variable side chain.

The fixed common core of each amino acid consists of an amino group (-NH2) and an acidic carboxyl group (-COOH). These groups are identical across all amino acids, forming their basic structure. The variable side chain, on the other hand, differs among every amino acid, giving each its unique combination of physicochemical properties (i.e. charge, polarity, hydrophobicity, size, etc.), as partially shown in Figure 1.1.

1.1.2 Protein structure & function

The conventional view of protein structure often focuses on the 3D configuration of the sequence of amino acids in the folded state. However, a better way to describe proteins is through a hierarchy of structures, each level offering a progressively more comprehensive view of their organization. This in turn elucidates the relationship between structure and function which is essential to comprehend how proteins operate in the cell.

Primary structure

The primary structure refers to the linear sequence of amino acids, which constitutes the most fundamental level of a protein, as described in the previous subsection. The primary structure is intimately linked to the 3D shape of proteins. This concept is emphasized by Anfinsen's dogma [2] which states that all information required to specify the

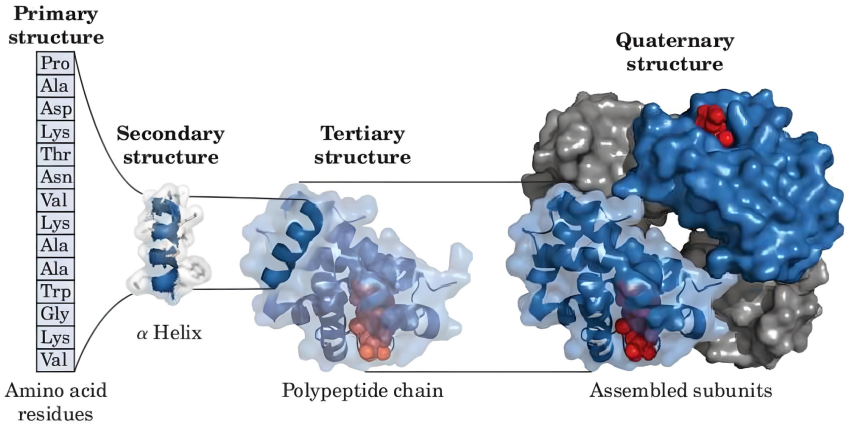


Figure 1.2 – Hierarchical organization of protein structure, illustrating the progression from the primary to the quaternary level. Image adapted from [1].

structure of a protein is encoded in its amino acid sequence. Although generally true, there are limitations. For example, in real cellular environments, additional factors such as molecular chaperones, or environment-dependent factors can influence the folding process and the final 3D structure of proteins.

Secondary structure

The secondary structure represents the local geometrical organization of the protein and can be classified into two main groups: *alpha helices*, *beta sheets*. These elements are usually connected by less structured sections, typically in the form of loops.

- *Alpha helices*: right-handed coiled strands, held together by hydrogen bonds between nearby residues. Each helix turn typically corresponds to $3 - 4$ amino acids. Alpha helices are common to most proteins and provide stability and flexibility.
- *Beta sheets*: comprising strands lying side by side, beta sheets are connected through hydrogen bonds between different parts of the protein chain. They add rigidity and form the core of many proteins. Beta-sheets can be categorized into two main types: parallel and antiparallel, depending on the relative orientation of the adjacent strands.

Tertiary structure

The tertiary structure indicates the overall 3D arrangement of a protein. This level of organization is critical, as it enables proteins to carry out their respective functions effectively. For example, enzymes - proteins dedicated to catalyzing specific chemical reactions - must have a precise 3D arrangement of their active site to fit correctly with

their substrates. This necessity is even more evident for structural proteins, which play a critical role in maintaining the integrity and shape of cells. These proteins provide support and stability to various biological structures through their unique shapes.

In conclusion, the tertiary structure is foundational to a protein's function, dictating the specific molecular interactions that determine the protein's role within the cell.

Quaternary structure

The quaternary structure involves the assembly of multiple protein subunits into a complex, essentially functioning as a large molecular machine that performs various functions. Both homomeric (identical subunits) and heteromeric (different subunits) complexes exist in nature. Hemoglobin, a homomeric protein complex present in red blood cells, is a classic example. It consists of four subunits working together to bind and transport oxygen throughout the body. The hemoglobin complex is an example of permanent interaction between proteins. The interaction can also be transient, allowing proteins to interact with multiple partners quickly in processes like signal transduction or metabolic regulation.

Quinary structure

The highest level of protein organization involves the interactions between proteins and their broader cellular context. The quinary structure of proteins includes the transient or semi-permanent interactions between proteins or protein complexes within the cellular environment. This level of organization entails interactions with nucleic acids, lipids, or other cellular components, and can lead to macroscopic phenomena such as liquid-liquid phase-separated regions.

1.1.3 Protein biosynthesis

Protein biosynthesis is the process, common to all living beings, that generates proteins by assembling amino acids based on the instructions encoded in the DNA. As stated by the central dogma of molecular biology, which elucidates the flow of genetic information within a cell, this process occurs in two main steps: *transcription* and *translation*.

Transcription

The genetic information that codes for proteins is stored within genes, which are specific segments of DNA carrying the information necessary to determine the amino acid sequence of proteins. In the first stage of protein biosynthesis, transcription, the information stored in the genes is transcribed into messenger RNA (mRNA). Inside the

cell's nucleus, the enzyme RNA polymerase binds to the DNA, reads the specific sequence of a gene, and synthesizes a complementary mRNA strand. This mRNA acts as a temporary copy of the genetic information and is later transported out of the nucleus to the cytoplasm. It is important to note that this process is specific to eukaryotes. In contrast, prokaryotes, such as bacteria, do not have a defined nucleus, and transcription occurs directly in the cytoplasm.

Translation

Following transcription, the translation phase happens. Translation takes place in the ribosome, a cellular machine found in the cytoplasm. The ribosome reads the sequence of the mRNA, and with the help of molecules called transfer RNA (tRNA), it links together amino acids in the specific order dictated by the mRNA sequence. This process forms a polypeptide chain of amino acids, the primary structure of proteins, that subsequently folds autonomously into its functional three-dimensional structure. The folding process is guided by the sequence of amino acids within the protein chain and is influenced by various factors, e.g. the hydrophobic interaction of amino acids with the surrounding water. Ultimately, the protein chain assumes the set of dynamical configurations that reflect the most energetically favorable conformations within the cellular environment.

1.1.4 The genetic code

Crucial to the process of translation is the genetic code, the set of rules that determine how a sequence of nucleotides of DNA (or mRNA) is translated into the sequence of amino acids in a protein. The genetic code consists of three nucleotide sequences called codons, each of which codes for a specific amino acid or serves as a stop signal to terminate translation. Across a vast range of organisms, from simple bacteria to complex multicellular eukaryotes, the same genetic code applies. While there are a few minor exceptions, this universality underscores the shared evolutionary ancestry of all life on Earth.

A noteworthy aspect of the genetic code is its degeneracy or redundancy. Since there are only 4 nucleotides in DNA, a total of $4^3 = 64$ possible codons exist. These must code for 20 amino acids, plus the stop signal, meaning that multiple codons often code for the same amino acid. This fact is illustrated in Figure 1.3, which presents a genetic codon table. By reading the nucleotides from the inner to the outer circle, it is possible to associate each codon with its specific amino acid. Codons that code for the same amino acid are called synonymous. The figure reveals that degeneracy often arises from variability in the third nucleotide of codons; consequently, a mutation at this position generally does not alter the coded amino acid. For

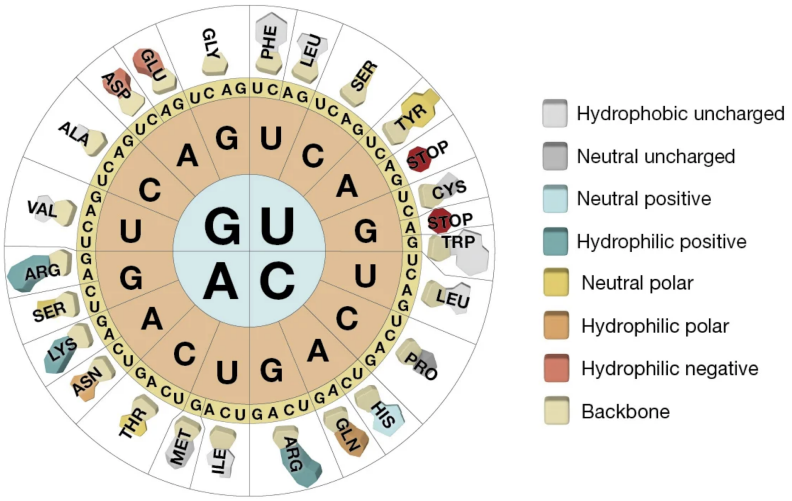


Figure 1.3 – Circular RNA codon table. Codons are read from the inner to the outer circle. Amino acids are shown next to codons, colored by their chemical properties. Image adapted from [3].

instance, Threonine (Thr) is coded by all four codons starting with AC, irrespective of the last nucleotide. However, different mappings between codons and amino acids exist. For instance, the amino acid serine (Ser) is encoded by six different codons: all codons starting with UC and the codons AGC and AGU. There is also the case of methionine (Met), which is coded by only one codon, AUG, and is also used to initiate translation. A remarkable characteristic of the genetic code is its capacity to reduce the negative impacts of non-synonymous amino acid mutations. Mutations are usually deleterious to protein structure and function. However, a mutation might not affect much a protein's function if the chemical characteristics of the replaced amino acid, such as hydrophobicity, are preserved. Interestingly, the genetic code mapping is such that non-synonymous single-nucleotide mutations often have similar chemical characteristics. For example, codons in the form of NUN (where N represents any nucleotide) always code for hydrophobic amino acids (grey color in the amino acid side chain of Figure 1.3). The degenerate nature of the genetic code adds a layer of resilience and adaptability to protein evolution. Minor changes in the nucleotide coding sequences may not lead to significant alterations in the resulting protein, ensuring that essential functions are preserved even in the face of genetic variations.

1.2 PROTEIN EVOLUTION

Life on Earth has evolved for over three billion years, beginning with the self-replication of the first primitive molecules. This process has

enabled the development and spread of a wide variety of species, creating the complex web of life we see today, much of which is still to be explored.

Evolution is defined as the heritable change in the characteristics of living beings over successive generations, guided by natural selection, which acts on phenotypic variation. One of the biggest discoveries of the 20^{th} century is the cause and nature of this phenotypic variation: random and spontaneous mutations in the genome.

Mutations

Among all types of mutations that affect the genetic material of living beings, the most well-understood affect genes, the specific subsections that code for proteins. Protein or gene evolution, therefore, is a primary and essential form of molecular evolution. Point mutations, insertions, and deletions are common types of mutations in genes. While point mutations involve a change in a single nucleotide base, insertions or deletions can have more profound effects. These latter mutations are likely to disrupt proteins unless they happen in groups of three, as this preserves the reading frame of the genetic information.

As noted in the previous section, point mutations can be synonymous or non-synonymous. Synonymous mutations result in the translation of the same protein, while non-synonymous mutations lead to the translation of proteins with one mutated amino acid. The distinction between these two types of mutations is crucial in understanding how genetic changes can either preserve or alter the function of proteins. Interestingly, even though only non-synonymous mutations are supposed to affect the synthesis of proteins, through amino acid changes, synonymous mutations can play an effect as well. In particular, it has been shown that organisms can have specific preferences for specific synonymous codons, a phenomenon known as "codon usage bias". One of the explanations that has been put forward to explain the overuse of specific codons claims that those codons are translated faster and more efficiently, conferring an evolutionary advantage to genes containing them.

Natural selection plays the role of favoring organisms with mutations that increase their fitness. These mutations make the organisms better adapted to their environment, increasing their chances of survival and reproduction. Conversely, mutations that result in less favorable traits may be wiped out over time, as they hinder the organism's ability to survive.

1.2.1 Sequence space

The number of possible protein sequences of a given length is huge. For a gene of length 1000, corresponding to a medium-length protein

of around 300 amino acids, the total combinatorial space amounts to approximately $20^{300} = 10^{390}$ unique amino acid sequences. Compared to the known universe's approximate 10^{80} atoms, the magnitude of the distinct possible sequences becomes clear. The set of all possible sequences of a given length L, composed of amino acids or nucleotides, is commonly referred to as *sequence space*. The topology of sequence space can be envisioned as a network where adjacent points denote sequences differing by just a single mutation. Navigating through sequence space thus facilitates the representation of sequences via mutations or chains of mutations along evolutionary pathways.

Despite the vastness of this space, only a minuscule fraction can be occupied by well-structured and functional proteins and even a smaller fraction corresponds to extant sequences. Random sequences have a negligible probability of functioning, and most mutations around a given sequence are deleterious. Nonetheless, as Maynard Smith emphasized in [4], a prerequisite for the existence of modern sequences is that sequence space must be mostly connected: most sequences must be linked to at least another functional sequence through a single mutation.

Protein evolution continuously expands and diversifies protein sequences, transforming them from their ancient precursors into highly specialized functional entities upon which all known biological life is based. This diversification has led to the emergence of a vast array of protein functions, each tailored to its specific biological need.

1.2.2 Protein families

Protein sequences can be systematically classified into protein families with a common evolutionary origin. Proteins in the same protein family are referred to as *homologous*, meaning that they have a common ancestry. Homologous proteins can be further subdivided into two categories: *paralogs* and *orthologs*. Paralogs are homologous proteins deriving from the same species, originating from a gene duplication event. Over time, these duplicate genes diverge, leading to variations in function while still retaining common sequence features. Orthologs, by contrast, are homologous proteins that are found in different species. They result from the incremental accumulation of mutations during and after speciation events, while retaining a similar functional role. As such, they can provide insights into evolutionary relationships and conserved biological mechanisms of genes. Homologous sequences exhibit substantial amino acid divergence, up to only 20% of the amino acid being identical between two pairs of proteins in the same family. Despite this sequence divergence, the conservation of the three-dimensional structure and function across species is remarkable - particularly for orthologs. For example, humans and the baker's yeast *S. cerevisiae* share hundreds of orthologous genes

[5]. Although the typical sequence identity between those genes is around 30%, almost half of the yeast genes can be swapped for the human version and still allow normal growth of yeast as it was shown in Ref. [6]. Notably, sequence identity was not the strongest predictor of replaceability.

Similar experiments [7, 8] have replaced essential *E. coli* genes with thousands of homologous sequences from different species and have confirmed that a sizable percentage of the divergent homologs have no to little detrimental effect on fitness. Indeed, swapping genes between different species has been a method for identifying functionally equivalent homologs for a long time [9].

Sequence alignments and HMM

The advent of next-generation sequencing technologies has greatly facilitated the study of protein families, making hundreds of millions of protein sequences available in public databases [10]. These sequences are often assembled into Multiple Sequence Alignments (MSAs), a powerful data structure used to visualize and analyze the relationships between homologous protein sequences coherently. While sequences within the same family generally have a similar length, the specific number of amino acids can vary significantly from one sequence to another. This is due to the presence of insertions and deletions which change the length of sequences. Although the structure of homologous proteins can slightly differ between distant sequences [11], the overall shape remains remarkably conserved. This is why aligning sequences [12] and comparing aligned positions often unveils a lot of biological information. Sequence alignments are instrumental for phylogeny reconstruction, structure prediction, and prediction of the function of uncharacterized proteins.

Bio-informatics tools to align sequences are necessary to deal with the huge amount of available sequences, and have a long history of development [13]. Hidden Markov Models (HMM) [14] are powerful statistical models used for aligning protein sequences and discovering new homologous ones [15]. The underlying concept of profile Hidden Markov Models (HMMs) is to capture the statistical variations in an alignment based on the frequency of amino acids and gaps in each column of the alignment. A gap is a special symbol '-' used for sequence alignments to indicate the absence of an amino acid with respect to another sequence in the alignment. Profile HMMs consist of a Markov chain based on a directed graph as depicted in figure 1.4, which can be broken down into three primary types of nodes or states:

- *Match states*: Representing the frequency of amino acids in a non-insert column of an alignment, these model the position-specific amino acid usage patterns within a family.

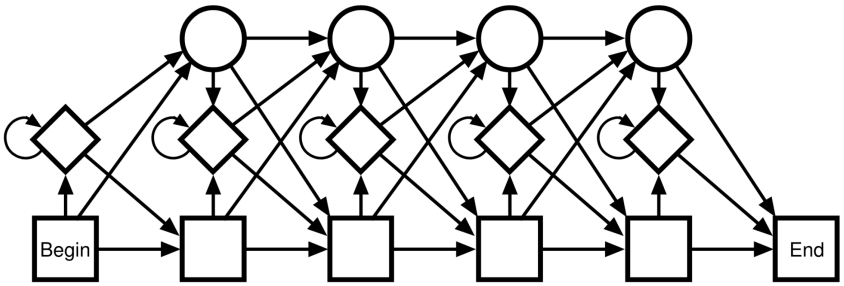


Figure 1.4 – HMM structure depicted as a directed graph consisting of three types of states: squares for match states, diamonds for insert states, and circles for deletion states. Image from [17].

- *Insertion states*: Accounting for the potential addition of amino acids in certain sequences, resulting in gaps in the MSA for other members.
- *Deletion states*: Reflecting the removal of amino acids in certain sequences, lead to a corresponding gap in the aligned sequence.

Each match and insertion state is associated with an amino acid distribution. In the case of insertion states, the MSA frequency of each amino acid is utilized, while match states rely on the distribution of residues in the corresponding column of the MSA. The Markov chain emits a residue by the corresponding distribution, or a gap symbol if a deletion state is reached. Transition probabilities exist between two "layers" of the graph, moving from left to right.

The parameters of the HMM, encompassing emission probabilities for match states and transition probabilities between different states, are learned from a seed alignment, which often is manually aligned by experts. Parameter learning is based on the seed alignment's conservation profile.

Once the HMM is trained, it can be employed to identify new members of the seed family or align new sequences. This alignment is achieved through computing the most likely corresponding path from "Begin" to "End," often utilizing algorithms such as the Viterbi algorithm. HMMer [16] is one of the most common tools used to both infer HMM models and align sequences.

1.2.3 Statistical signals in sequence alignments

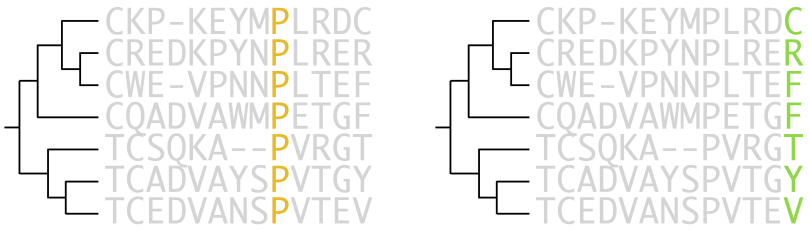
A multiple sequence alignment can be formally represented as a matrix A with M rows and L columns, where rows represent protein sequences, and each column represents one aligned position of the sequences. Each matrix element, denoted as a_{im}, represents one of the 20 standard amino acids or the gap symbol "-". Assuming that we can perfectly align an MSA, the outcome is a set of very divergent

homologous sequences that despite their differences, represent various solutions to a unique problem: specifying a functional protein. It thus becomes evident that sequences must adhere to certain common constraints to ensure proper folding and functionality. These crucial indicators are frequently found imprinted in MSAs in the form of statistical patterns. The most notable statistical indicator is the 1-point frequency of the amino acid a at position i:

$$f_i(a) = \frac{1}{M} \sum_{m=1}^{M} \delta(a_{im}, a) \tag{1.1}$$

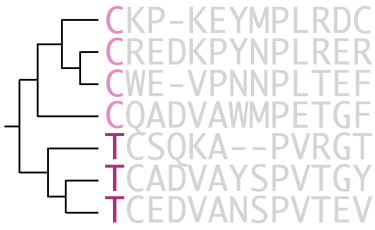
Intuitively, the distribution of amino acids in a certain column can reveal a lot about the functional role of that site. If we observe $f_i(a)$ to be close to one in site i, we expect the amino acid a in that position to be very important for the protein fold or function. More complex patterns exist as well, which involve multiple sites at the same time. These patterns can unveil not only the conserved regions but also potential sites of functional importance, as well as structural features of the proteins in question. Let us now examine some common patterns one by one to understand the intricacies they reveal about different protein sequences through the lens of multiple sequence alignment. A pictorial representation is presented in figure 1.5:

- *Conservation:* the preservation of one specific amino acid across different sequences in the alignment, likely reflecting essential functional or structural roles within the protein family. Represented in figure 1.5a.
- *Variability:* the lack of conservation of any amino acid in a specific column of the alignment. Amino acids differing among sequences potentially signify a position where substitutions do not impact function. Represented in figure 1.5b.
- *Specific conservation:* the preservation of one specific amino acid only within a subset of sequences in the alignment, usually part of the same phylogenetic branch. It possibly highlights functional requirements relevant only to a subgroup within the family. Those residues are sometimes called "specificity-determining residues". Represented in figure 1.5c.
- *Specific variability:* the lack of conservation of specific amino acids only within a subset of sequences in the alignment, usually part of the same phylogenetic branch. It possibly highlights the absence of functional requirements relevant to that subgroup. Represented in figure 1.5d.
- *Covariation:* the correlated evolution between two residues, where changes in one position may be associated with changes in another. This pattern may suggest structural or functional interac-



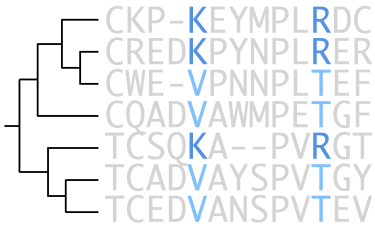
(a) *Conservation.*

(b) *Variability.*

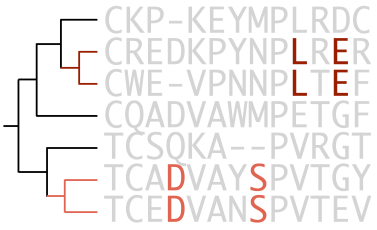


(c) *Specific conservation.*

(d) *Specific variability.*



(e) *Coevolution.*



(f) *Phylogeny.*

Figure 1.5 – Toy MSA depicting typical statistical patterns common to real MSA.

tions between the corresponding residues. Represented in figure 1.5e.

- *Phylogeny:* the correlated presence of amino acids between recently diverged sequences due to a recent evolutionary origin. It might not have functional significance and be confused with coevolution. Represented in figure 1.5f.

Beyond single-column conservation and variability patterns, the most prominent statistical signal captured by MSAs is *covariation*, which often arises due to correlated evolution occurring between residues. To further illustrate, let us refer to figure 1.5e where we observe that pairs of amino acids — specifically K with R and V with T — seem to only appear in this specific combination. We can quantify this behavior by comparing the frequency of amino acids in each column, with the frequency of amino acids in a pair of columns. If they are not independent, covariance will be non-zero, i.e.

$$c_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b) \neq 0, \tag{1.2}$$

with

$$f_{ij}(a,b) = \frac{1}{M}\sum_{m=1}^{M}\delta(a_{im},a)\delta(a_{jm},b) \tag{1.3}$$

The cause of this covariation is commonly attributed to coevolution. In this scenario, a pair of amino acids is important in either maintaining the three-dimensional structure of a protein or facilitating its proper functioning. Therefore, any mutation affecting one residue in the pair can have detrimental effects, necessitating a coevolutionary adjustment in the other residue to maintain the protein's functionality. Naturally, multiple couples of amino acids fulfill the same functional role, but those couples are often mutually exclusive. It is these interactions between sites that are thought to generate observable correlations, a topic that however remains a subject of debate as evidenced in various studies [18, 19].

1.3 ANTIBIOTIC RESISTANCE

Antibiotic resistance refers to the ability of microorganisms, such as bacteria, to counteract the effects of antibiotics that are designed to kill them or inhibit their growth. This resistance can compromise the effectiveness of antibiotic treatments, posing a critical global health concern [20], with serious repercussions [21]. Since Alexander Fleming's landmark discovery of penicillin in 1929 [22], antibiotics have been a cornerstone in the battle against bacterial infections. These drugs have significantly reduced mortality rates and have been indispensable in various medical procedures, from surgeries to chemotherapy.

However, the excessive and inappropriate use of antibiotics in both medical and agricultural settings has facilitated the emergence of antibiotic-resistant bacterial strains. In this context, bacterial populations undergo mutations in their genes, some of which grant resistance to specific antibiotics. Proteins encoded by these mutated genes can alter the target site of the antibiotic within the bacterial cell, expel the antibiotic out of the cell, or reduce the antibiotic's effectiveness [23].

Recent data suggests [24] that 1.3 million deaths globally were attributable to antibiotic-resistant infections in 2019. If the trend continues, this number is projected to escalate to 10 million by 2050. In addition, the use of antibiotics in livestock feed is a contributing factor that has exacerbated the issue. Such agricultural practices facilitate the horizontal gene transfer of resistance traits among different bacterial species, thereby accelerating the spread of antibiotic resistance [25]. The history of antibiotic resistance is characterized by a perpetual arms race between the development of new antibiotics and the emergence of new bacterial resistance mechanisms. Comprehensively understanding the mechanics, contributing factors, and ramifications of antibiotic resistance is necessary to devise effective strategies for mitigating this global health crisis.

1.3.1 β-lactam antibiotics

β-lactams represent the most commonly employed category of antibiotics and serve as a cornerstone in fighting bacterial infections [26]. Since the initial discovery of penicillin, a range of classes of β-lactams have been discovered [27]. Their widespread use in clinical and industrial settings has invariably led to the emergence and spread of bacterial resistance mechanisms against these drugs [28].

β-lactams include penicillins, cephalosporins, carbapenems, monobactams and clavams. These antibiotics possess a β-lactam ring in their molecular structure. Their primary mode of action is to hinder the creation of bacterial cell walls by targeting penicillin-binding proteins (PBPs) essential in forming and cross-linking peptidoglycan layers. This interference weakens the bacterial cell wall, leading to osmotic imbalance and eventually causing bacterial cell lysis. Mechanisms of resistance include the production of beta-lactamases, enzymes that break open the beta-lactam ring [29], and alterations of PBPs, which reduce the antibiotics' affinity for their target.

1.3.2 β-lactamase resistance

Most of the work detailed in this thesis deals with β-lactamase-resistant enzymes. Enzymes that attack β-lactams have evolved independently several times in history. The evolutionary pressure behind this convergent evolution may be the result of an arms race between

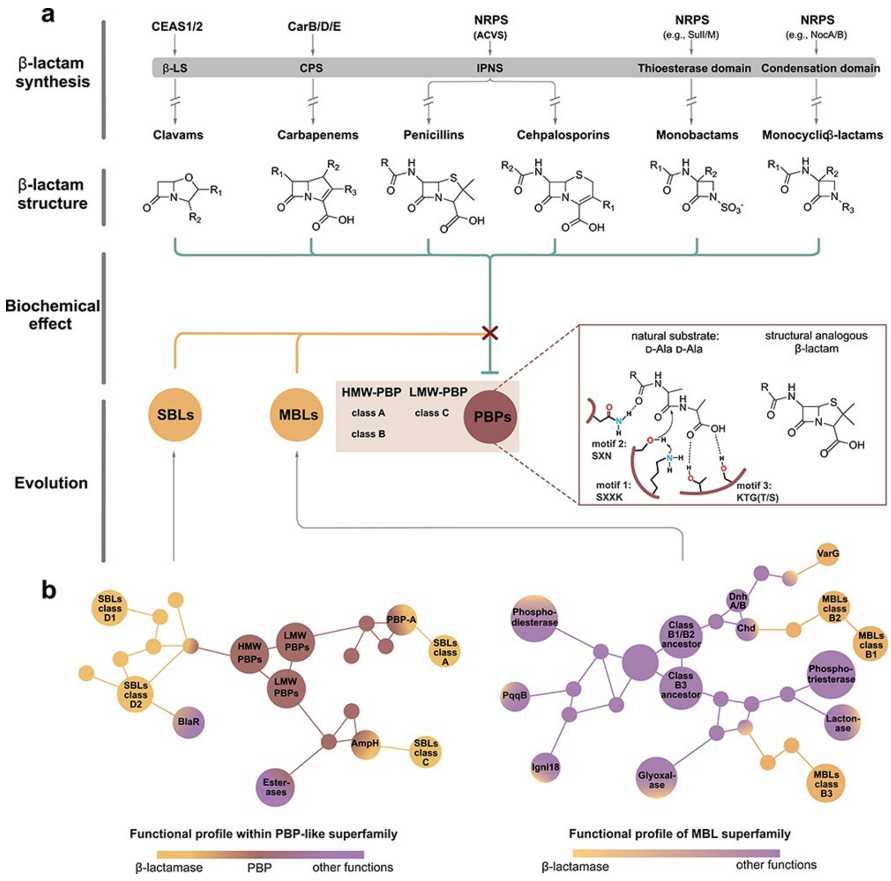


Figure 1.6 – The evolution and function of β-lactamase enzymes. a) Proceeding from top to bottom: the synthesis pathways of five distinct classes of β-lactams along with their structures; the structural analogy between the natural substrate of PBPs and β-lactams; the evolutionary relation of the SBL and MBL enzymes to their respective superfamily origins. b) Schematic representation of the sequence-function network within each enzyme superfamily. Image source: [31].

bacteria and fungi, although many uncertainties remain. While β-lactams are primarily produced by fungi, some bacteria also secrete them as secondary metabolites [30]. The most well-known source of beta-lactams is the mold genus *Penicillium*, from which the first antibiotic, penicillin, was isolated. Enzymes with β-lactamase capacity are classified into two main groups, serine-β-lactamases (SBLs) and metallo-β-lactamases (MBLs) [31].

Serine-β-lactamases

Serine-β-lactamases constitute a specialized subset of enzymes within the PBP-like superfamily. This superfamily encompasses a diverse collection of proteins, most of which are instrumental in bacterial cell wall synthesis, particularly in the formation of the peptidoglycan layer that grants structural integrity to bacterial cells. A key feature

of the β-lactam antibiotics is their β-lactam ring structure, which closely resembles the natural substrates of these cell-wall-synthesizing enzymes, D-Ala D-Ala, as depicted in panel "a" of figure 1.6. As a result of this structural similarity, the antibiotics are capable of binding to the enzyme's active site, thereby inhibiting its function. This inhibition disrupts the formation of the peptidoglycan layer, leading to bacterial cell death.

However, SBLs are no longer involved in the cell synthesis process, but they retain the β-lactam acylation mechanism, which involves the covalent binding of the β-lactam ring of the antibiotic to the enzyme's active serine residue. On top of this, they have evolved the enzymatic capability to hydrolyze this intermediate. The hydrolysis enables the detachment of the antibiotic from the enzyme, deactivating the antibiotic in the process. This means the enzyme can re-engage in successive catalytic cycles, thereby conferring resistance to β-lactam antibiotics [32]. SBLs are further divided into Classes A, C, and D, each evolved from different subclasses of the PBP superfamily, namely high molecular weight (HMW) and low molecular weight (LMW) PBPs [31], see panel "b" of figure 1.6. Each class has distinct sequence and molecular patterns, and a very diverse evolutionary origin, making each a protein family on its own.

Metallo-β-lactamases

The metallo-β-lactamase superfamily, a diverse group of proteins characterized by distinct sequences and functions, possess multiple functions, including β-lactamase activity, RNA degradation, arylsulfatase action, phosphonate metabolism, and DNA repair [33], see panel "b" of figure 1.6. Within the MBL superfamily, subclasses B1, B2, and B3 share a common trait: the capability to hydrolyze beta-lactams.

Class B β-lactamases constitute approximately 1.5% of the entire MBL superfamily with around 6000 sequences deposited on public databases [34]. Despite substantial divergence in their amino acid sequences, these enzymes share strong structural similarities. Their common structural feature is an $\alpha\beta\beta\alpha$-fold domain, which typically consists of 200-300 amino acids. An intriguing aspect of these enzymes is the presence of two conserved metal-binding motifs within this fold, which play a critical role in their catalytic activity. The intricate catalytic mechanism of MBLs in hydrolyzing β-lactams hinges on a range of factors including diverse zinc ligands and active site geometries. Generally, zinc ions facilitate the initiation of the β-lactam ring breakdown process.

Phylogenetic analysis indicates two distinct evolutionary events leading to the acquisition of β-lactamase activity within MBLs, one leading to the emergence of the B1 and B2 subclasses, while the other giving rise to the B3 subclass. These evolutionary events were characterized by alterations in the metal-binding motif, potentially

leading to adjustments in the arrangement of metal ions. Additionally, modifications in the overall configuration of the active site occurred, influencing the enzyme's ability to interact with specific substrates.

1.4 MOLECULAR EPISTASIS

The rise of antibiotic-resistant bacteria, as well as the evolution of viruses like SARS-CoV-2 and HIV, constitutes a mounting threat to public health. Commonly, even a small number of mutations in crucial proteins of emerging pathogens (for example, the spike protein in SARS-CoV-2 [35]) can substantially augment their potential to infect humans. To safeguard against these threats and respond quickly when harmful variants emerge, it is crucial to develop the capability to anticipate the emergence and the consequences of such mutations, in a way, to predict molecular evolution [36, 37]. This challenge is the dynamical version of an already very complex problem: understanding genotype-phenotype-fitness maps in proteins [38]. A precise understanding of the mutational effects in proteins is indeed the basis to tackle the problem of protein evolution [39]. However, this task is very intricate due to a phenomenon known as epistasis [40, 41]. Epistasis refers to the interaction between genetic mutations, that is when their combined effect on a protein's function is different from the sum of individual effects, i.e. non-additive. Understanding the effect of mutations in proteins and how individual mutations interact with one another [42] in an epistatic manner is vital for predicting how a pathogen might evolve or how to design effective drugs. If the relationships between genotype and phenotype were linear or additive, predicting the effect of any mutation on a protein would be straightforward [43]. A protein of length L would only require measuring the effect of all possible $19L$ mutations, to predict its behavior. However, epistasis complicates this matter, as mutational effects in proteins are sequence-dependent, making it necessary to measure a greater number of combinations of mutations.

1.4.1 Types of epistasis

Specific and non-specific epistasis

Epistasis manifests itself as a non-additivity in the effect of multiple mutations. In this context, it is possible to identify two broad types of epistasis: one is non-specific, stemming from the non-linear nature of genotype–phenotype maps and therefore applies to all possible mutation combinations; the other is specific and takes place only among certain groups of mutations. These two cases are commonly referred to as specific and non-specific - or global - epistasis. Let's

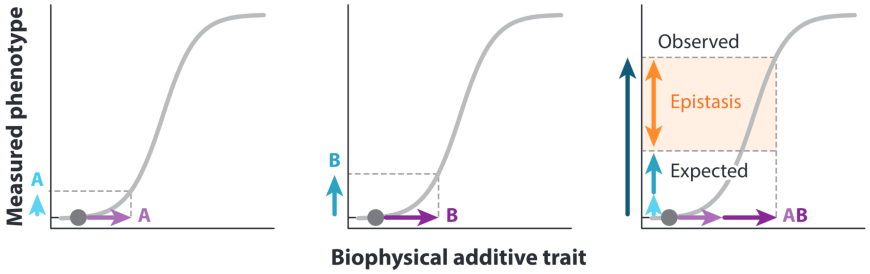


Figure 1.7 – Depiction of nonspecific epistasis. The grey non-linear curve connects the biophysical traits to phenotypic observable traits. Mutations A and B combine additively in the biophysical dimension, but their combined phenotype deviates from the additive expectation. Epistasis is the deviation from the expected effect. Image adapted from [40].

start with non-specific epistasis. In this case, the effect of mutations is independent at the molecular level. However, at the level of the biological organization that we observe, the effect is non-linear. For example, the relationship between a protein's stability and the fraction of folded protein is sigmoidal. If we consider a marginally stable protein, each single destabilizing mutation will not affect much its folding probability, but the combination of two mutations could have a dramatic effect. As a consequence, any assay that is affected by the amount of folded protein will measure a non-linear effect in the interaction of two mutations, even though their combined effect on stability was additive [44]. This is pictorially shown in figure 1.7 where the effect of mutations A and B is additive at the molecular level, but not at the level where the phenotype is measured. Specific epistasis, on the other hand, depends on the identity of the mutations involved. This type of epistasis primarily emerges from unique combinations of mutations, such as amino acid pairs in a folded protein that come into direct contact or confer specificity for a particular ligand. These interactions are often mediated by direct physical interactions among residues and can result directly in non-additive effects on various physical properties of the protein itself. In-depth case studies have shed light on some molecular mechanisms behind specific epistasis. Those include interaction between contacting residues [45], cofactor repositioning [46] and changes to the structural dynamics of proteins [47]. Specific epistasis is thought to be one of the main reasons behind the correlated evolution of residues [48, 49], as revealed by the statistical analysis of MSAs.

Magnitude and sign epistasis

When mutations interact non-additively, the consequent effect may be categorized according to both the sign and the magnitude of the

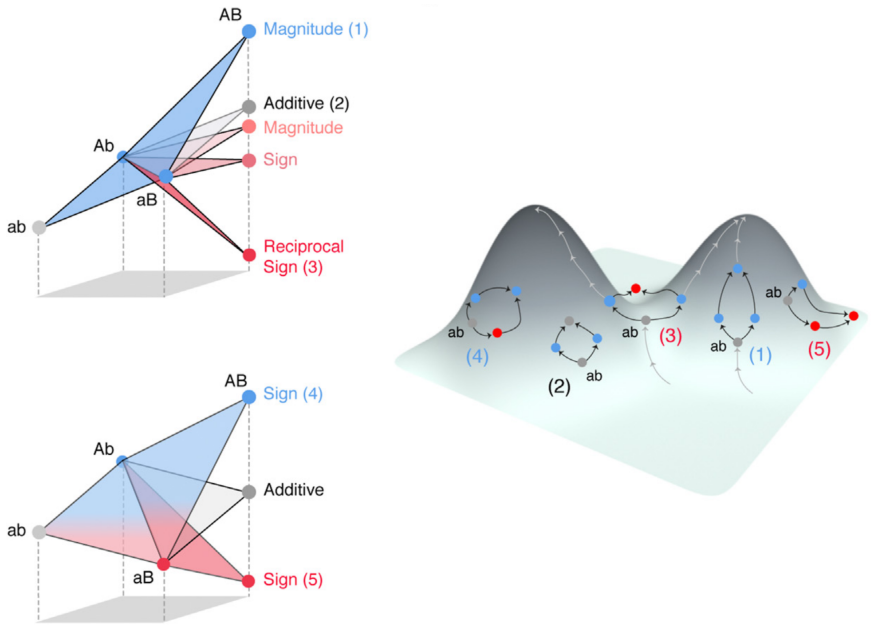


Figure 1.8 – The figure illustrates various forms of epistasis. The example illustrated an initial 'ab' hypothetical genotype transitioning to a derived 'AB' genotype through mutations a to A and b to B. Fitness levels for each genotype are indicated by colored dots. Image adapted from [42].

resultant epistatic interaction. Specifically, in the case of *magnitude* epistasis, the collective effect of the interacting mutations preserves its overall orientation, or sign, irrespective of whether the individual mutations were beneficial or detrimental. *Positive magnitude* epistasis occurs when the synergistic interaction of the two mutations enhances the combined effect beyond the summation of their contributions. Conversely, *negative magnitude* epistasis happens when the mutual influence of the two mutations attenuates their combined effect.

Distinct from magnitude epistasis, *sign* epistasis represents a unique interaction class. In this scenario, the combination between the two mutations alters the direction of their collective effect. In the case of *reciprocal sign* epistasis, the co-occurrence of both mutations reverses the sign of each mutation's impact. Figure 1.8 schematically depicts all the cases described here.

1.4.2 The consequences of epistasis

Overall, epistasis has a key influence on protein evolution [39] and complicates our ability to predict phenotypic outcomes based on genotypes alone [43]. It also shapes factors like evolvability and robustness. For example, positive epistasis enhances a protein's capacity to evolve and be robust. However, the unpredictability injected by epistasis, particularly specific epistasis, introduces a degree of historical con-

tingency in evolutionary outcomes, making them dependent on rare events [50]. This unpredictability is particularly surprising when deleterious mutations combine to improve the function of proteins, for example, multiple destabilizing mutations can allow an enzyme to change its preferred substrate for a novel one, conferring an evolutionary advantage. Beyond natural evolution, epistasis has implications for healthcare, affecting the development of treatments for rapidly evolving pathogens like influenza and HIV [51].

1.5 MEASURING SEQUENCE LANDSCAPES

A fundamental challenge in protein biology is to determine the precise function of proteins. We do not refer here to the problem of 'functional annotation,' which involves determining the broad roles of a protein within a biological system based on its amino acid sequence. We rather allude to the quantitative characterization of the function of proteins, in the sense of determining their activity level. For example, the turnover rate of an enzyme, or the binding energy of an antibody. Understanding and quantifying protein function is critical in the field of molecular biology. It helps us grasp how biological systems work at a fundamental level and has a major impact on medicine and biotechnology. Applications range from the development of novel therapeutics, like antibodies, to the optimization of industrial enzymes. The task of protein function characterization usually involves measuring the effect of mutations concerning a target wild-type (WT) protein. For example, we might be interested in knowing how resistant a particular β-lactamase protein sequence is, concerning a specific antibiotic. Usually, it is quite hard to specifically measure the function of a protein at a chemical level, particularly if many thousands of variants have to be assessed. This is why the problem is often transformed into a fitness problem, where the functional question shifts to the survival of organisms harboring the proteins under scrutiny. In general, function and fitness are related through a complicated, non-linear relationship, which might even be non-monotonous. In this manuscript, we will however restrict our discussion to cases where improved function corresponds to equal or higher fitness.

1.5.1 High-throughput sequencing

To measure the effect of mutations of a protein in vivo typically needs three components: a method for mutating genes, a strategy for selecting those mutations, and a system to correlate the survival of the cells harboring the genes with their function. The main issue is scalability. If applied serially by hand, this approach can quantify a few dozen to a few hundred variants, but it is generally inadequate

for characterizing fitness landscapes more broadly, due to the huge size of mutational spaces. For example, all single mutations of a typical gene of 300 amino acids amount to almost 6000 different sequences that need to be tested. This is precisely where the adoption and increasing affordability of next-generation sequencing (NGT) has proved to be transformative in exponentially augmenting the capabilities of experiments to measure protein fitness landscapes. NGT allows the determination of the sequence of nucleotides within a DNA or RNA molecule, a process called 'sequencing', by reading millions of sequences in parallel during a single run. This large-scale parallelization enables experimental labs to quantify a gene's fitness by analyzing its prevalence in a sequenced library. By evaluating the abundance of sequencing reads in different libraries (for example before and after a selection process), it becomes possible to infer the fitness of a huge number of variants at the same time. For instance, a frequently appearing mutant sequence indicates that this variant is more "fit" or favorable under the chosen experimental conditions. In subsequent sections, we will outline the basic concepts behind some of the most common experimental techniques employed to measure fitness landscapes in vivo, as well as experiments that mimic protein evolution in the laboratory.

1.5.2 Deep mutational scans

Deep mutational scans (DMS) are an array of tools that provide a way to systematically study how thousands of different mutations individually affect a protein's function, helping to build a detailed understanding of protein fitness landscapes in biologically relevant conditions [52]. In the following, I will describe experiments performed in live cells, like bacteria, but a similar approach can be modified to work in vitro as well. To start, a diverse library of mutated versions of the target gene is created. This is done through different mutagenesis techniques, like error-prone PCR (epPCR) or site-directed mutagenesis, based on the required needs for the experiment. The first is a technique that allows to randomly mutate the nucleotides along a gene, while the second is a targeted approach that introduces specific mutations at predetermined positions within the gene. After the creation of the mutant library, the next step is to introduce these sequences into a suitable host organism, like bacteria or yeast, through a process called transformation. This allows for the mutant proteins to be expressed within living cells, providing a living laboratory in which to study their functions. Following this, a selection or screening process is initiated. This involves applying a specific selective pressure to identify which variants are functional and which are not. This step is usually complicated because not all genes specify phenotypes that are easy to select for. Once a suitable selective assay has been found,

selection can usually be applied in batch, i.e. at the same time for all cells expressing each individual variant. In this case, selective pressure might be an environmental condition or a chemical substance that permits only cells with functional variants to survive. There exist also different techniques for selecting cells, like Fluorescent Activated Cell Sorting (FACS), which requires specialized equipment to sort and analyze cells based on specific fluorescent markers (which are linked to the desired phenotypic outcome), enabling the identification and isolation of mutants that exhibit desired traits or functionalities. Next, the gene of interest in the surviving cells is extracted in bulk and prepared for high-throughput sequencing. At this step, it is crucial to have previously analyzed the initial set of variants (before selection) to compare it with the final set of variants that have survived the selection process. This dual analysis allows researchers to get an unbiased picture of the impacts of each mutation. Finally, the data from the sequencing is analyzed. By comparing the number of each variant before and after the selection process, it is possible to determine a fitness score for each mutation, providing a measure of how each change has affected the survival of the cells, and as a consequence, the function of the protein. In specific cases, where a reliable biophysical model is available, it is possible to infer also the molecular functions of the variants, such as stability or enzyme turnover rate.

Overall, DMS studies over the last decade have illuminated the relationship between protein sequence and function at an unprecedented scale [53]. To date, more than two hundred datasets have been produced [54], each characterizing all single mutations of given proteins, as well as more complex mutational combinations.

Measuring epistasis

The most common type of DMS measures the effect of single mutations in a protein. However, to measure and quantify epistasis, it is essential to assess the effect of multiple mutations and compare them with their individual effects. To probe epistatic interactions, two main strategies can be employed: introducing the same mutations across distantly related enzymes (to probe context-dependence) or examining the effects of multiple mutations in the vicinity of the same gene (to probe for non-additive effects).

One of the first studies [55] that demonstrated mutational incompatibilities between the swap of amino acids across orthologs introduced 168 mutations from the *P. aruginosa* IMDH enzyme into the *E. coli* IMDH enzyme. Approximately one-third of the mutations resulted in deleterious effects, demonstrating widespread epistasis. The fully functional wild-type amino acids from one ortholog could single-handedly impair the function of another. Similar studies have reconstructed ancestral protein sequences and have systematically introduced one by one all the modern amino acids into the ancient

background sequence, and vice-versa [50, 56]. Also in this case strong and specific epistatic effects were detected. Many of the modern mutations were incompatible with the ancestral background, relying on interactions with preceding mutations for their viability.

Other studies have aimed to characterize the local sequence landscape of protein sequences more extensively, relying strongly on HTS and parallelization. A pivotal experiment led by Olson et al. [57] examined the effect of all double mutants in the IgG-binding domain of protein G (GB1). The team succeeded in measuring the binding affinity for over $5,000,000$ variants, revealing both prevalent positive and negative epistasis. Different works sought to investigate smaller regions of sequence space, but more in depth. Two seminal studies [58, 59] assessed the effects of every possible combination of mutations at 4 critical positions, amounting to a total of $20^4 = 160,000$ assayed mutations. This allowed for the characterization of higher-order epistasis and the study of alternative evolutionary trajectories in sequence space.

Another popular method for exploring fitness landscapes and detecting higher-order epistasis involves creating complete binary landscapes. In this case, every possible combination of the wild-type amino acids from two distinct sequences is tested, yielding a total of 2^n variants, n representing the number of differing amino acids between the two sequences. Those studies are limited by experimental constraints to a maximum of around 15 mutations, i.e. $< 100,000$ variants. These studies have revealed a strong yet surprisingly sparse network of epistatic interactions [60] in a fluorescent protein and have enabled the reconstruction of all possible evolutionary intermediates between one unmutated germline sequence [61] and two mature broadly neutralizing influenza antibodies.

Those massive datasets have begun to shed light on the complexity of protein function and interaction between mutations. However, achieving a more comprehensive understanding remains a challenging endeavor due to the difficulties in distinguishing specific from global epistasis, the necessity to focus on the correct levels of biological organization, and the complex task of generalizing findings to biological systems in the wild [62].

1.5.3 Broad mutational scans

The experimental approaches described earlier predominantly focused on analyzing the effects of all possible single mutations in a specific gene, the complete combinations of all mutations within a very restricted number of residues, or the binary landscapes between pairs of them. All of these approaches are centered on single or very few enzymes, largely ignoring the reality that protein families consist of thousands of members, all belonging to a complex - and in principle

unique - global landscape. Broad mutational scanning (BMS) [7] is a technique developed to address this limitation. BMS seeks to analyze in parallel the fitness landscape of a large set of homologs. Unlike typical DMS studies, which focus on all single mutations, BMS encompasses a broader range of mutations, offering a more expansive view of the mutational space, albeit less systematically. To achieve this, the authors first created a method capable of cheaply synthesizing long genes, called DropSynth [63]. Essentially, they improved upon the initial step of DMS, which is the creation of the variant library. Traditional techniques generally allow for either the introduction of targeted mutations into long genes ($>$ 300 base pairs) or the synthesis of many thousands of small genes ($<$ 300 bp). DropSynth overcomes these limitations by allowing the synthesis of long genes at cost-effective prices.

Here is a very simplified description of how DropSynth works. Initially, short DNA or RNA sequences called oligonucleotides are designed and synthesized to correspond to different regions of the target genes. These oligonucleotides are then placed into a microfluidic device that creates tiny droplets, each containing a different mix of oligonucleotides. The droplets are created in such a way that each one contains all the oligonucleotides necessary to assemble a single gene. Through a process called emulsion PCR, the droplets function as individual chambers where the oligonucleotides are amplified and assembled into full genes. This process can sometimes introduce errors, generating random variants around the intended sequences. As a consequence, a library of diverse genes and random mutations around them is created. The library can be transformed inside bacteria and tested for the trait of interest, like typical DMS experiments.

In this study [63] BMS was used to analyze 5775 homologs of the dihydrofolate reductase (DHFR) enzyme as well as 1152 homologs of the enzyme phosphopantetheine adenylyltransferase (PPAT). 497 of the PPAT homologs survived their experimental assay. Thanks to the random mutations generated by DropSynth, the researchers managed to assay 71,061 mutants in total around the functional sequences, studying the effect of mutations across the whole sequence space of PPAT enzymes.

1.5.4 Laboratory protein evolution

Another important tool in the study of epistasis and evolution is laboratory protein evolution. Laboratory protein evolution is the study of protein evolution in a controlled and simplified environment. The aim is to either capture and reproduce the mechanisms of protein evolution as observed in the wild or to leverage the principles of natural evolution as an engineering tool for designing molecules and organisms with specific attributes. The controlled conditions

of the laboratory facilitate the development of particular traits in organisms or molecules, although tuning the environment to achieve the desired settings can be quite challenging. Typically, but not always, evolution is conducted in vivo, necessitating the use of a host organism. This process hinges on two fundamental components: a source of mutations like epPCR (that occur at a significantly faster rate than in nature) and a reliable selection method. These two steps are usually applied iteratively, exploiting the full potential of evolution. The term 'experimental evolution' can denote different concepts depending on the specific focus of interest. In this discussion, we consistently refer to cases where a single gene undergoes mutations while the rest of the genome remains unchanged, a scenario distinct from the approach taken, for example, by the Lenski lab [64], where entire populations of one organism evolve at natural mutation rates for thousands of generations in controlled environments.

Directed evolution

One major sub-field of laboratory protein evolution is *directed evolution*. Directed evolution is a laboratory technique that simulates the principles of Darwinian evolution, aiming to artificially steer the evolution of molecules [65] towards useful and applied goals. The original motivation behind this method was the fact that nature had already created remarkable enzymes capable of accelerating chemical reactions millions of times at ambient temperatures and pressures. Thereby, simulating the same process in the lab would hopefully reach similar goals. The technique was pioneered by Nobel Prize winner Francis Arnold [66, 67, 68]. Initially, the primary application of this methodology focused on evolving proteins individually to enhance aspects such as stability and suitability for industrial environments. Over time, the technique has evolved to encompass the engineering of enzymes for novel functions not previously seen in biological systems, as well as genetic circuits and even whole genomes. To generate genetic variability, directed evolution leverages heightened rates of mutation and recombination vastly exceeding those encountered in natural processes exploiting biotechnological techniques such as epPCR. Subsequently, a strong selection pressure is applied, to allow only the best variants to survive. Then, further amplification and diversification of the genetic material are performed, as depicted in figure 1.9. Iterating this process, the hope is that enzymes gradually improve towards some peaks of the fitness landscape. However, as we have observed, sequence space is immense. Therefore, the appropriate means and intensity of selection are necessary to guide the biological systems to the desired outcome. Equally important is to iterate the process sufficiently to fully optimize the enzyme under investigation. Usually, this approach favors only the fitter or the fittest genotypes in each round, potentially leading to the risk of becoming trapped in

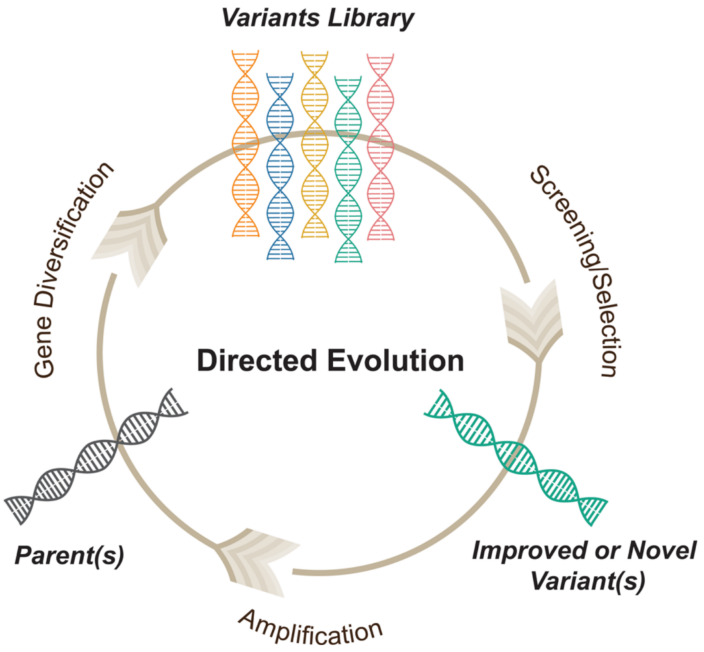


Figure 1.9 – Schematic representation of the directed evolution experimental workflow. Image adapted from [65].

local fitness or functional peaks without reaching the optimum. In other words, the strategy is usually greedy with all the limitations that this search method brings with it. Notwithstanding all those difficulties, directed evolution is currently one of the main approaches to engineering enzymes.

Neutral drift experiments

Directed evolution is fundamentally adaptive, and designed to enhance existing functions continuously. Nonetheless, we know that MSAs contain thousands of sequences with low degrees of similarity but nearly identical functions. Essentially, this suggests that a significant portion of the genetic diversity observed across species stems from the accumulation of neutral mutations that neither improve nor degrade function. Experiments involving neutral drift focus on the diversification of enzymes to understand the processes that drive natural evolution or to generate large libraries of variants for subsequent experiments. Researchers manage to introduce many neutral mutations by fine-tuning the selection strength to a level that merely maintains essential enzyme characteristics without promoting optimization, thereby encouraging divergence in the amino acid sequences [69, 70]. The primary limitation of neutral genetic drift is the necessity to maintain high diversity within a population of growing cells, such as bacteria, and to transfer this variability from one generation to the next while avoiding bottlenecks. Through neutral genetic drift

experiments, scientists have successfully explored the spontaneous emergence of a predisposition to new functions solely through random occurrences [71], created enormous libraries of orthologous sequences harboring numerous mutations to study the local fitness landscape of an antibiotic resistance gene [72], and examined phenotypic variability within populations evolving under uniform selection pressures [73].

1.6 SEQUENCE MODELS

The surge in the number of unlabeled sequences deposited in public databases [10], driven by rapid advancements in next-generation sequencing technologies, presents an ideal opportunity for leveraging statistical and machine learning methods to learn the statistics of protein families. As previously noted, sequences belonging to the same family often represent diverse solutions to the same problem: creating a functionally well-folded protein fine-tuned to perform a specific function. Therefore, the task of inferring sequence landscapes can be posed as an unsupervised machine-learning problem; its goal being the fit of the statistical distribution of the data. This is useful because it allows the approximation of the fitness landscape of protein families, with applications ranging from the study of evolution in silico to the computational engineering of enzymes. It is logical to assume that proteins within the same family, sharing the same three-dimensional structure and biological role, are subjected to similar evolutionary pressures. These shared pressures have imprinted discernible patterns in the statistics of multiple sequence alignments, suggesting that a single global model for each family could coherently describe the entirety of the available data. Therefore, sequence models utilize the evolutionary information from alignments of proteins to construct statistical representations that can produce insights about protein biology. Specifically, those models are trained to assign a high probability to functional sequences that are present in the training data, and conversely, a low probability to non-functional sequences that are supposed to be outside a target family.

1.6.1 Profile models

One of the earliest classes of sequence models, which only model the single-column statistics of MSA, are profile models or position-specific scoring matrices (PSSM). For a general reference, see [74].

In these statistical models, the probability of observing a sequence is proportional to the product of the frequencies of each of the amino acids in a corresponding MSA, denoted as $f_i(s_i)$:

$$P(\mathbf{s}) = \prod_{i=1}^{L} f_i(s_i). \tag{1.4}$$

By definition, the model is factorizable over the sites. Given that there is no interaction between the sites, this model is inherently non-epistatic. Despite their simplicity, such models have been utilized extensively for fitness predictions in proteins [75], for sequence alignment via profile HMM, for phylogeny reconstruction or in conjunction with machine learning, for the prediction of secondary structure elements [76]. However, they are unable to model some critical statistical signals of MSAs, notably covariation. Assuming that all necessary information to specify the protein fold is encoded in the amino acid sequence [2], one can investigate whether the evolutionary information encapsulated by profile models is adequate to statistically differentiate between good and bad proteins. This hypothesis was experimentally tested in 2005 [77, 78]. The research team generated and tested several artificial sequences in the WW protein domain family, comparing them to natural sequences. The non-natural sequences were fabricated based on the statistical properties of an MSA of natural sequences using three methods: the first involved generating sequences of random amino acids (R), thereby disrupting any site conservation pattern; the second shuffled the amino acids of the MSA by preserving the statistics of individual positions and eliminating covariation patterns, thereby retaining only independent conservation (IC); the third sought to replicate the pairwise amino acid frequencies found in the original MSA, achieved through coupled conservation (CC) obtained via Markov Chain Monte Carlo (MCMC) annealing procedure. Experimental results revealed that neither the R nor IC sequences achieved correct folding and thereby did not function. However, 31% of CC sequences and 67% of natural sequences were functional. Those results indicate that the sequence information contained in MSAs is sufficient to impose the structural constraints of the WW domain. It further suggested that correlations between column pairs in MSAs are sufficient to adequately define the protein fold. We see that, as previously noted, epistasis is vital in understanding evolutionary biochemistry and vice versa [79]. Drawing from these findings, we can learn an important lesson for sequence models: it is important when they are *generative*, i.e. capable of generating data statistically similar, yet distinct, to the training set. This is particularly relevant to discriminating functional sequences and going beyond the training set of natural sequences. Practically, reproducing the two-point statistics of the MSA used for training appears to be a sufficient condition.

1.6.2 Potts models

One of the first global statistical models capable of modeling the correlation between sites of MSAs was Direct Coupling Analysis (DCA) [49, 80, 81].

DCA allows us to infer a probabilistic model that leverages both the first and second-order statistics of MSAs to recapitulate the evolutionary forces influencing the sequences. It attributes a probability, denoted by $P(\mathbf{s}|\mathbf{h},\mathbf{J})$, to each possible amino acid sequence in the space. This probability distribution adopts the form of a Potts model, an extension of the renowned Ising model [82]. Potts models, an active research area on their own [83], describe a probability distribution over configurations of categorical units (in our case the L residues of a protein sequence) coupled in a pairwise manner with each other, each harboring multiple states (the q amino acids).

The parameters defining this model are termed 'fields' $\mathbf{h}$ and 'couplings' $\mathbf{J}$. Fields act on single amino acids, while couplings connect pairs of different residues. For an amino acid sequence denoted as $\mathbf{s} = (s_1, s_2, \ldots, s_L)$, where L is the length of the aligned sequences in the MSA used for training, the probability distribution of a Potts model is expressed as:

$$P(\mathbf{s}|\mathbf{h},\mathbf{J}) = \frac{1}{Z(\mathbf{h},\mathbf{J})} \exp\left(-\mathcal{H}(\mathbf{s})\right). \tag{1.5}$$

Here, $\mathcal{H}(\mathbf{s})$ stands for the Hamiltonian, or energy function, defined by

$$\mathcal{H}(\mathbf{s}) = -\sum_{i=1}^{L} h_i(s_i) - \sum_{i<j}^{L} J_{ij}(s_i, s_j). \tag{1.6}$$

In this context, $h_i(s_i)$ represents the field term pertaining to the i-th amino acid of type s_i, while $J_{ij}(s_i, s_j)$ is the coupling term for amino acids s_i and s_j. The field terms affect the propensity of individual amino acids to occur at specific positions, in particular a positive $h_i(a)$ field, favors amino acid a in positions i. The coupling terms instead model the co-evolutionary interaction between amino acid pairs at different positions within the sequence, and similarly a positive $J_{ij}(a,b)$ respectively favors the presence of a and b in positions i and j. The partition function $Z(\mathbf{h},\mathbf{J})$ guarantees a correct normalization of the probability distribution:

$$Z(\mathbf{h},\mathbf{J}) = \sum_{\mathbf{s}\in\mathcal{S}} \exp\left(-\mathcal{H}(\mathbf{s})\right) \tag{1.7}$$

where $\mathcal{S}$ is the space of all sequences. We will use the expressions 'DCA' and 'Potts model' interchangeably throughout the text.

Originally, the form of Potts models for modeling protein sequences was derived through the maximum-entropy principle [84]. f_i and f_{ij} were chosen as the relevant variables that the statistical

model had to fit, and the Potts Hamiltonian is the m that incorporates the fewest possible number of assumptions while still being able to reproduce the data statistics. Nonetheless, different models have also demonstrated proficiency in modeling MSA statistics, including Restricted Boltzmann Machines (RBM) [85] and simple autoregressive models [86] and more complex deep learning architectures. An important advantage of Potts models is their capacity to describe the interactions between sites via direct coupling $\mathbf{J}$, which can disentangle direct from indirect interactions between sites [87]. The drawback of the use of direct couplings between every pair of amino acids is that the models become fully connected. The number of couplings is therefore of the order of L^2q^2, with q representing the number of amino acids. As a consequence, Potts models can have millions of parameters even for medium size length proteins. This complexity not only makes models slow to learn but also poses challenges to the accurate determination of the parameters due to limited statistics. Potts models are the primary sequence model that we have used in this thesis to investigate sequence landscapes and study evolution.

Inference

Once we have established a form for the probability distribution of our model, the exact values of the parameters need to be inferred. The goal of the inference procedure is to adjust the parameters in such a way that the training set has the highest possible probability. In other words, we aim to maximize the likelihood of the training data awith respect to the model parameters. Since the introduction of DCA, many fast and approximate approaches have been developed [88, 89] for this task. However, we will focus here on an asymptotically exact, yet slower method: Boltzmann machine learning [90]. This is how all models mentioned in this thesis have been trained unless specified otherwise. If we assume that we have M sequences in our training MSA and that they are all independent, we can write the log-likelihood of the data as:

$$\begin{aligned}\mathcal{L}(\mathbf{h},\mathbf{J}|\{\mathbf{s}^\mu\}) &= \log\prod_{\mu=1}^{M} P(\mathbf{s}^\mu|\,\mathbf{h},\mathbf{J}) \\ &= \sum_{\mu=1}^{M}\left[\sum_{i=1}^{L} h_i s_i^\mu + \sum_{i<j}^{L} J_{ij}(s_i^\mu, s_j^\mu)\right] - M\log Z(\mathbf{h},\mathbf{J}).\end{aligned} \tag{1.8}$$

The log-likelihood is a convex function of the parameters; therefore, to find its maximum, a simple gradient ascent strategy is sufficient. Once the parameters are initialized, the update equations are as follows:

$$h_i^{t+1}(a) \leftarrow h_i^t(a) + \eta_h \frac{\partial\mathcal{L}(\{\mathbf{s}^\mu\}|\mathbf{J}^t,\mathbf{h}^t)}{\partial h_i(a)} \tag{1.9}$$

$$J_{ij}^{t+1}(a,b) \leftarrow J_{ij}^{t}(a,b) + \eta_J \frac{\partial \mathcal{L}(\{\mathbf{s}^\mu\}|\mathbf{J}^t, \mathbf{h}^t)}{\partial J_{ij}(a,b)} \tag{1.10}$$

where η_h, η_J represent the learning rates that must be chosen independently. The derivatives of the log-likelihood are easy to compute:

$$\frac{\partial \mathcal{L}(\{\mathbf{s}^\mu\}|\mathbf{J}^t, \mathbf{h}^t)}{\partial h_i(a)} = f_i(a) - p_i^t(a) \tag{1.11}$$

$$\frac{\partial \mathcal{L}(\{\mathbf{s}^\mu\}|\mathbf{J}^t, \mathbf{h}^t)}{\partial J_{ij}(a,b)} = f_{ij}(a,b) - p_{ij}^t(a,b) \tag{1.12}$$

with p_i^t, p_{ij}^t being the the marginal of the model at time t/ The learning procedure reaches a fixed point when p_i^t, p_{ij}^t match the empirical frequencies f_i, f_{ij}. A proper stopping condition needs to be set to reach the desired accuracy. A significant challenge with this approach is that the model frequencies cannot be calculated analytically, necessitating a Markov Chain Monte Carlo (MCMC) approach for the estimation. As a consequence, each iteration of the algorithm samples an artificial MSA using the model parameters at time t, and then uses this sample to estimate p_i^t, p_{ij}^t. This method is based on the fact that a sufficiently large MCMC sample allows in principle for the exchange of the ensemble average over sequence space with the empirical average over the data sample. For a detailed explanation of the algorithm's implementation, refer to [91].

A point must be raised about one of the crucial assumptions of Boltzmann learning: the independence of the training data. This condition is tclearly not satisfied in the case of MSAs due to the complex phylogenetic process that has produced our modern sequences. To address this discrepancy, a common strategy is to down-weight closely related sequences when estimating their amino acid frequencies. Empirically, we assign a weight w^m to each sequence $\mathbf{s}^m$, which is inversely proportional to the number of other sequences in its neighborhood, namely:

$$w^m = \left(1 + \sum_{n \neq m} \Theta\left(\frac{d(\mathbf{s}^m, \mathbf{s}^n)}{L} < \delta\right)\right)^{-1} \tag{1.13}$$

where $d(\mathbf{s}^m, \mathbf{s}^n)$ is the number of different amino acids, or Hamming distance, between sequences $\mathbf{s}^m$ and $\mathbf{s}^n$, Θ is the Heaviside function and δ is usually 0.2.

Another important point about the Potts model concerns its inverse temperature, denoted as β. By changing this parameter the DCA probability distribution can be skewed towards sequences with either higher or lower energy. Specifically, this inverse temperature enters the model as follows:

$$P_\beta(\mathbf{s}) = \frac{\exp\left(-\beta \mathcal{H}(\mathbf{s})\right)}{Z} \tag{1.14}$$

where $\beta = 1$ is the temperature at which the data is inferred. Altering this parameter allows for the sampling of sequences with skewed statistics. For instance, using a β lower than 1 results in sequences with higher energies, i.e. low probabilities, and vice versa. Conventionally, the temperature $T = \beta^{-1}$ is also used as a variable. Analogous to thermodynamic systems, a high temperature corresponds to random configurations, while a low temperature is indicative of low-energy, more structured configurations. In the next chapter, I will show that high T can be used to model low selection pressure, and vice versa.

Structure prediction

Predicting the structure of a protein from its sequence is one of the central goals of molecular biology. Understanding a protein's structure is crucial as it sheds light on its biological function through insights gleaned from mechanical features or resemblances to known structures. Despite proteins having an intrinsic ability to autonomously fold into their native configurations in physiological environments with remarkable precision, unraveling this intricate process has challenged scientists for years. Although direct and precise computational simulations of protein folding exist, they are often prohibitively resource-intensive, and the simplifications introduced to make the simulations feasible often fail to recapitulate the true folding dynamics, leading to the wrong structures.

Historically, determining protein structures has been primarily performed experimentally, relying on the crystallization of proteins and subsequent analysis through X-ray crystallography — a method both time-consuming and costly. In the last decade, a strategy rooted in evolutionary biology has introduced a completely new tool capable of aiding the determination of structures. Initial investigations into multiple sequence alignments (MSAs) of homologous proteins revealed a wealth of information regarding protein structure [92, 93]. Correlations between different sites in the sequence were found to be indicative of contacting residues in the protein structure. As we have explained before, this is likely due to the coevolution of the amino acids in those positions. A significant advancement over those first observations came with the introduction of DCA, which was originally utilized to predict direct contacts between protein sequence residues [49]. The idea was to leverage and interpret the values in the J_{ij} coupling matrices once the model is inferred. If positions i and j are in contact with the 3D structure and coevolving, high values of the couplings are expected to be observed to accommodate for the covariation of the sites. An effective scalar metric derived from the couplings was found to be the Frobenius norm, defined as:

$$F_{ij} = \sqrt{\sum_{a=1}^{q} \sum_{b=1}^{q} J_{ij}(a,b)^2}. \tag{1.15}$$

Utilizing this metric, it was possible to rank all possible residue pairs based on the DCA scores. The breakthrough was that most of the top L predicted contacts were true structural contacts. The high accuracy of those contacts was shown to be enough to predict protein folds correctly [94]. The method has seen enhancements, notably by the introduction of the APC correction [95].

Inspired partly by these early successes, there has been a real breakthrough in resolving the protein structure problem. Advances in computational power, algorithms, and expansion of protein structure databases have allowed supervised models to enter the game. While Potts models and related approaches are unsupervised and do not utilize any prior structural information, new deep-learning models can incorporate more information from previously resolved structures. A pivotal moment came in 2020 with Google DeepMind's introduction of AlphaFold [96] and AlphaFold2 [97]. The models combine the intuition about the relevance of evolutionary information for structure prediction put forward by DCA but also incorporate structural information for much more precise predictions. In simple terms, AlphaFold-like models train supervised deep neural networks end to end to predict the atomic positions of all amino acids in a protein structure using as input information only the MSA of the family of the sequence of interest. Interestingly, despite the complex internal architectures of these machine-learning models, models that use only single sequences as inputs, instead of the full MSA have much worse performances.

Nevertheless, the protein folding problem has still many hurdles yet to be overcome. The elusive nature of protein folding pathways continues to be a monumental challenge; the exact routes that amino acid chains take to attain their final configurations remain largely unknown. However, significant progress has been made in deducing the end states of these pathways using sequence data alone.

Mutational effect prediction

As discussed in paragraph 1.5, a pivotal question in protein biology is understanding the effects of mutations in protein sequences, a concept we will loosely refer to as fitness prediction.

Sequence models have emerged as some of the most affordable and powerful tools for predicting the fitness effects of mutations. These models primarily use information from homologous sequences to foresee the effect of mutations in protein sequences. Given a sequence model that assigns probabilities to each possible sequence, it can be employed to assess mutation effects. Currently, the most proficient models are complex sequence models that factor in epistatic interactions, outclassing profile models and setting the standard in the field. Notably, Potts models were the pioneers in successfully predicting mutational effects using only sequence data [98, 99, 100].

These models have also been extended to predict the adverse effects of mutations in human proteins, emphasizing the significant role of site couplings [101]. A method rooted in similar models has also recently achieved notable success [102].

So, how do Potts models predict mutation effects in proteins? The strategy involves utilizing the model probability as a score, specifically, the energy of the sequences which is equivalent to the negative elog-probability. A perfect testing ground for mutation effect prediction is DMS datasets, where mutation effects are usually expressed relative to those of a wild-type sequence. Here's how the DCA scoring system works:

$$\Delta\mathcal{H} = \mathcal{H}(\mathbf{s}_{wt}) - \mathcal{H}(\mathbf{s}_{mut}). \tag{1.16}$$

In this formula, if $\Delta\mathcal{H} > 0$ the mutant sequence $\mathbf{s}_{mut}$ is predicted to be deleterious with respect to the wildtype $\mathbf{s}_{wt}$, and vice versa. In the case of a single mutation occurring at position i, swapping amino acid a with amino acid b :

$$\Delta\mathcal{H}(a \to b) = h_i(b) - h_i(a) + \sum_k \left(J_{ik}(b, s_k) - J_{ik}(a, s_k)\right). \tag{1.17}$$

As it can seen from this equation, $\Delta\mathcal{H}$ explicitly couples the wt and the mutant amino acids to the sequence background. Subsequently, predictions are compared to experimental outcomes to estimate the model's predictive accuracy. This methodology has proven effective across a range of proteins found in viruses, bacteria, and humans. Typically, the correlation, calculated using the Spearman correlation coefficient to account for potential non-linearities, is around 0.5, with some proteins exceeding 0.7.

De novo sequence generation

As we previously mentioned, the ultimate test for generative sequence models is experimental validation. In other words, novel artificial sequences generated from those models must be tested in vivo to determine whether they perform their functions as efficiently as their homologous counterparts.

The Potts model gives us a probabilistic framework to assess the likelihood of sequences belonging to a specific family. This enables two major applications: first, to score mutations, as we have seen in the previous section; and second, to generate artificial sequences by sampling from the Potts distribution. To generate artificial sequences the most used and straightforward way for Potts models is MCMC sampling.

The question of whether sampled sequences from the Potts model are functional in vivo was tested with DCA on the chorismate mutase

(CM) enzyme family in 2020 [8]. CMs are essential enzymes composed of about 100 amino acids, playing a crucial role in bacteria survival due to their ability to synthesize a precursor to the essential amino acids tyrosine and phenylalanine. The authors of the work gathered an alignment of 1259 homologous sequences of the *E. coli* CM enzyme (ecCM) and substituted it with each of the homologs. Initially, a high-throughput selection screen was employed to verify whether each of the sequences in the alignment, consisting of the homologs of different species, would support growth in *E. coli*, and approximately 38% of the sequences proved successful.

Subsequently, the researchers trained a Potts model on the same alignment using Boltzmann machine learning. They created over 1000 artificial variants through MCMC sampling, aiming to generate highly diverse proteins with varying parameters. In particular, sequences were generated at different temperatures to influence their average energy. A profile model was also utilized to create sequences serving as a negative control. Each sequence was tested in vivo for its ability to substitute the ecCM enzyme.

The results were notable, as detailed below:

- Sequences generated with the profile model were ineffective.
- 3% of the sequences generated at $T = 1$ were functional.
- 31% of the sequences generated at $T = 0.66$ were functional.
- 48% of the sequences generated at $T = 0.33$ were functional.

In summary, the findings underscore three critical aspects: the inadequate generative capacity of the profile model, already verified by [78]; the remarkable ability of the Potts model to create previously unseen functional sequences; and a correlation between lower temperatures (i.e., sequences with higher probabilities according to the model) and increased functionality. An important note regarding the latter point is that the temperatures of $T = 0.33$ and $T = 0.66$ facilitated the sampling of sequences with energy comparable to that of natural sequences, albeit limiting their divergence from the training set. Sampling at low temperatures was needed due to the specifics of the regularization in the inference procedure of the Potts model in the paper. In practice, the temperature allowed to interpolate between the statistics of natural sequences (low temperature) and that of random sequences (higher temperature).

Since this pivotal study, others have replicated the results [103], although sometimes with a lower success [104], but the underlying reasons for limited success remain somewhat unclear in this latter work. Recently, researchers have leveraged the model for designing functional antibiotic-resistant proteins [105], focusing on a wild type and sampling various sequences along evolutionary trajectories.

1.6.3 Other generative models

In recent years, generative models have undergone significant development for protein sequence modeling, fitness prediction, and generation. This surge has been largely driven by advancements in machine learning algorithms and increased computational power. Predominantly, these models employ deep learning techniques capable of capturing complex interactions between amino acids in protein sequences. However, their complexity makes those models more challenging to train and interpret compared to Potts' models. A pivotal test for new architectures is their ability to generate functional artificial sequences [106]. Trinquier et al. proposed one of the simplest autoregressive models with generative capabilities [86]. While suggesting its potential to generate functional sequences, they did not furnish conclusive evidence. Autoregressive models forecast new tokens in a sequence based on previously observed elements alone, positioning them as useful frameworks for modeling protein sequences. Furthermore, these models can leverage deep-learning architectures inspired by neural language processing, such as transformers, and be trained on vast datasets comprising all known proteins, and not just single families. Madani et al. demonstrated that large language models based on transformers are capable of crafting functional protein sequences [104]. These models are very big, with over 1 billion parameters, and need substantial computational costs to be trained. The training is usually composed of hundreds of millions of sequences from all known families. This is done to leverage some kind of transfer learning, allowing the models to extrapolate and reuse information across different families. Nonetheless, further exploration of those models is needed, especially since the study revealed something interesting. To surpass the quality of Potts models (which are only trained on the family of interest) in fitness prediction tasks, it was essential to do a fine-tuning on a curated alignment of that family as well. Repecka et al. utilized a Generative Adversarial Network (GAN) approach to produce functional, highly mutated malate dehydrogenase enzymes [107]. Hawkins-Hooker et al. harnessed variational autoencoders in the design of bacterial luciferase [108].

In conclusion, generative models pave the way for a plethora of new techniques for creating artificial protein sequences and forecasting the repercussions of mutations. However, pressing questions linger, including the optimal balance between interpretability and expressivity, the relevance of various models in studying protein evolution, and the comparative efficacy of different models when trained with identical data.

2 EXPLORATION OF SEQUENCE SPACE

2.1 INTRODUCTION

This chapter discusses our work on modeling protein evolution. We initially focus on a simple setting, namely experimental evolution by neutral drift, a technique that we have discussed in section 1.5.4. Shortly, the experiments consist of multiple rounds of mutation and selection of a library of genes evolved in vivo. The approach aims at diversifying the library of genes while maintaining its function with a low selection pressure. Section 2.2 contains the article [109] that resulted from the modeling of two such experiments involving beta-lactamase genes. The article shows that our modeling framework can quantitatively reproduce the statistics of the experimental libraries, by relying on a sequence landscape inferred from distant homologs and a sampling algorithm that takes into account many biological details of the experiment. In the following two sections, we discuss some improvements to our approach. Section 2.3 discusses a refinement of the modeling approach which involves taking into account mutational biases of the nucleotide substitutions. Section 2.4 describes a major change to the sampling dynamics that allows us to reproduce the amino acid statistics of distant homologous proteins. Section 2.5 builds upon this newly developed stochastic evolutionary dynamics joining data-driven models of protein sequences with a realistic mutational process happening directly on nucleotides. The evolutionary model proposed provides a comprehensive new tool to model and investigate long evolutionary trajectories in sequence space. In particular, we investigate the emergence of epistasis-driven evolutionary timescales affecting the change in mutability of sites over time. The content of this section will be the object of a separate publication that will appear soon.

2.2 ARTICLE

Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution

Matteo Bisardi,[1,2,†] Juan Rodriguez-Rivas,[2,†] Francesco Zamponi,[1,*] and Martin Weigt [*,2]

[1]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France

[2]Biologie Computationnelle et Quantitative LCQB, Institut de Biologie Paris Seine, CNRS, Sorbonne Université, Paris, France

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: francesco.zamponi@ens.fr; martin.weigt@sorbonne-universite.fr.

Associate editor: Banu Ozkan

Abstract

During their evolution, proteins explore sequence space via an interplay between random mutations and phenotypic selection. Here, we build upon recent progress in reconstructing data-driven fitness landscapes for families of homologous proteins, to propose stochastic models of experimental protein evolution. These models predict quantitatively important features of experimentally evolved sequence libraries, like fitness distributions and position-specific mutational spectra. They also allow us to efficiently simulate sequence libraries for a vast array of combinations of experimental parameters like sequence divergence, selection strength, and library size. We showcase the potential of the approach in reanalyzing two recent experiments to determine protein structure from signals of epistasis emerging in experimental sequence libraries. To be detectable, these signals require sufficiently large and sufficiently diverged libraries. Our modeling framework offers a quantitative explanation for different outcomes of recently published experiments. Furthermore, we can forecast the outcome of time- and resource-intensive evolution experiments, opening thereby a way to computationally optimize experimental protocols.

Key words: **protein evolution, fitness landscapes, sequence space, epistasis, data-driven models.**

Introduction

In the course of evolution, biological sequences encoding proteins explore functional sequence space. The observable sequence variability between homologous sequences, that is, sequences connected by common ancestry, results from a delicate balance between mutation and selection. Mutations tend to randomize sequences, whereas natural selection prunes most of those mutations having a deleterious effect on fitness. When analyzing large databases of homologous protein families (Mistry et al. 2021), we therefore find sequences with 70–80% different amino acids, but highly conserved functional and structural properties.

In turn, it is possible to search for statistical patterns in ensembles of homologous proteins (Durbin et al. 1998), using tools borrowed from statistical inference and unsupervised machine learning, and to relate them to selective constraints acting in these proteins. The most prominent signal is conservation; a position in a protein not (or rarely) changing amino acid over extended evolutionary time scales, is likely to play an important role in the protein's function (e.g., active sites in enzymes) or for the protein's structural stability (e.g., residues buried in the protein core).

A second type of statistical signal has received a lot of attention during the last decade (De Juan et al. 2013; Levy et al. 2017; Cocco et al. 2018). The correlations between the amino acids present in pairs of residue positions can be extracted via global statistical models like those used in direct coupling analysis (DCA) (Weigt et al. 2009; Morcos et al. 2011), Gremlin (Balakrishnan et al. 2011), or PSICOV (Jones et al. 2012). This signal of residue–residue coevolution results from epistatic couplings between residues in structural contact in the folded proteins, that is, of residue pairs in direct physical interaction in the 3D structure of the protein, even if possibly located at long distance along the primary amino acid sequence. Coevolutionary methods, in particular when used as input for structurally supervised deep-learning methods like RaptorX (Xu 2019), DeepMetaPSICOV (Greener et al. 2019), AlphaFold (Senior et al. 2020), or trRosetta (Yang et al. 2020), have recently induced a revolution in protein-structure prediction, reaching unprecedented accuracy in computationally predicted structures close to the accuracy of experimentally determined structures (Jumper et al. 2021). Hundreds of previously unknown protein structures have been predicted this way (Ovchinnikov et al. 2017; Tunyasuvunakool et al. 2021).

However, coevolutionary methods rely on the availability of large alignments of homologous but diverged proteins, since they rely on statistical signatures extracted from sequence variability (Haldane and Levy 2019). Recently, two

groups have independently asked the question, if experimentally generated sequences can be used instead of natural homologs for contact prediction (Fantini et al. 2020; Stiffler et al. 2020). To this aim, they have proposed and performed similar experiments. First, starting from a given wildtype sequence, they have iterated several rounds of alternating sequence diversification via error-prone polymerase chain reaction (epPCR) (Cadwell and Joyce 1992), and selection for functionality (antibiotic resistance for both experiments). In contrast to traditional directed evolution (Arnold 1998, 2018), selection was very weak (low antibiotic concentrations), so proteins are not simply optimized for function, but may diversify their sequences while maintaining a certain level of functionality. After a certain number of rounds, the resulting sequence library was sequenced, to provide the data for statistical learning.

The resulting functional sequence libraries were quite diversified, with typical distances of 10–15% of the sequence length from the wildtype protein used as a starting point. This is much less than in natural protein families, characterized typically by average distances of 70–80% between homologs. However, the simultaneous emergence of about 10–40 mutations, and the depth of more than $10^4 - 10^5$ sequences in the experimentally evolved libraries, could make the detection of epistasis, and thus contact prediction, possible (Fantini et al. 2020; Stiffler et al. 2020).

Interestingly, both teams have run plmDCA (Ekeberg et al. 2013), or evCouplings based on plmDCA (Hopf et al. 2019), on the data—with very different results. Although the contact signal in (Fantini et al. 2020) was quite weak, and mostly concentrated to nearby positions along the sequence, (Stiffler et al. 2020) found a sufficiently accurate contact prediction to enable the subsequent construction of a precise structural model.

To understand the differences in results given the similarity in approaches, we have developed a modeling scheme, which allows us to simulate protein evolution in a data-driven sequence landscape. Comparison of simulated and experimental data of both experiments shows that our simulations reproduce quantitatively the experimental observations. Furthermore, the simulation scheme allows us to control important parameters of the experiments, like the evolutionary distance from the wildtype in the final evolved library, the sequencing depth of the library, or the strength of selection. We find that our model is able to explain the difference in contact prediction between the two experiments in terms of sequence divergence and sequencing depth.

The agreement between simulations and experiments suggests that our modeling framework allows for a quantitative analysis of important questions about protein evolution, like the mechanism underlying sequence space exploration and the emergence of signatures of epistasis with sequence divergence, compare also the related Sequence Evolution with Epistatic Contributions (SEEC) model (de la Paz et al. 2020). Beyond such basic questions in evolutionary biology, our framework has also the potential to help in optimizing experimental design. To give an example, our simulations predict that both experiments would have benefited from slightly weaker selection, represented by slightly lower antibiotic concentrations. This would have enabled a faster exploration of the neighborhood of the wildtype sequence and the occurrence of slightly more deleterious mutations, which have a better chance to be coupled by epistasis than the predominantly neutral mutations accepted at strong selection. Such predictions are very interesting, since our computational approach is efficient and can be applied to thousands of protein families, whereas the experiments are expensive in time and resources. Guiding them to increase the success probability may therefore be an impactful strategy. For instance, our approach can be used to explore different protocols, such as alternating cycles of strong and weak selection.

Results

The general procedure of our modeling approach is graphically illustrated in figure 1. In this section, we first describe the data-driven sequence landscape, which is inferred from multiple sequence alignments (MSA) of natural homologs of the experimentally studied wildtype, that is, from data unrelated to the experiment. As a first check of robustness, we show that this landscape represents well the mutational effects of single-residue substitutions when compared with a deep-mutational scanning experiment, and that the inclusion of epistatic couplings in the landscape model is essential for its accuracy. The landscape can thus be used as a proxy for the protein's fitness landscape.

Next, we present a minimal model of evolutionary dynamics, very similar to but more quantitative than SEEC. In this model, mutations appear at the level of the DNA sequence via single-nucleotide mutations, but selection acts exclusively at the protein level, that is, on the amino acid sequence translated from the DNA sequence, via the inferred sequence landscape. We will show that sequences generated in silico by this model reproduce quantitative features of the experimentally generated sequences, like mutational profiles or the fitness distribution.

Subsequently, we explore the potential of the experiments by performing simulations under variable conditions for sequence divergence, sequencing depth, or selection strength. This allows us to locate the two experiments in an exhaustively scanned parameter space, to understand the limitations of the experiments, and to propose schemes for overcoming current limitations.

An Epistatic Data-Driven Sequence Landscape Captures Mutational Effects

The basis of our approach is a computationally inferred sequence landscape, used as a proxy to quantify protein fitness and selection acting on proteins. To obtain this landscape, we first use the Pfam protein-family database (Mistry et al. 2021) to extract an MSA of diverged homologs of the wildtype protein used in the experiments. Both studies performed experiments with a member of the beta-lactamase family (Pfam accession PF13354), TEM-1 in (Fantini et al. 2020) and PSE-1 in (Stiffler et al. 2020); the latter work also studied the acetyltransferase AAC6 (PF00583). The details of the MSA

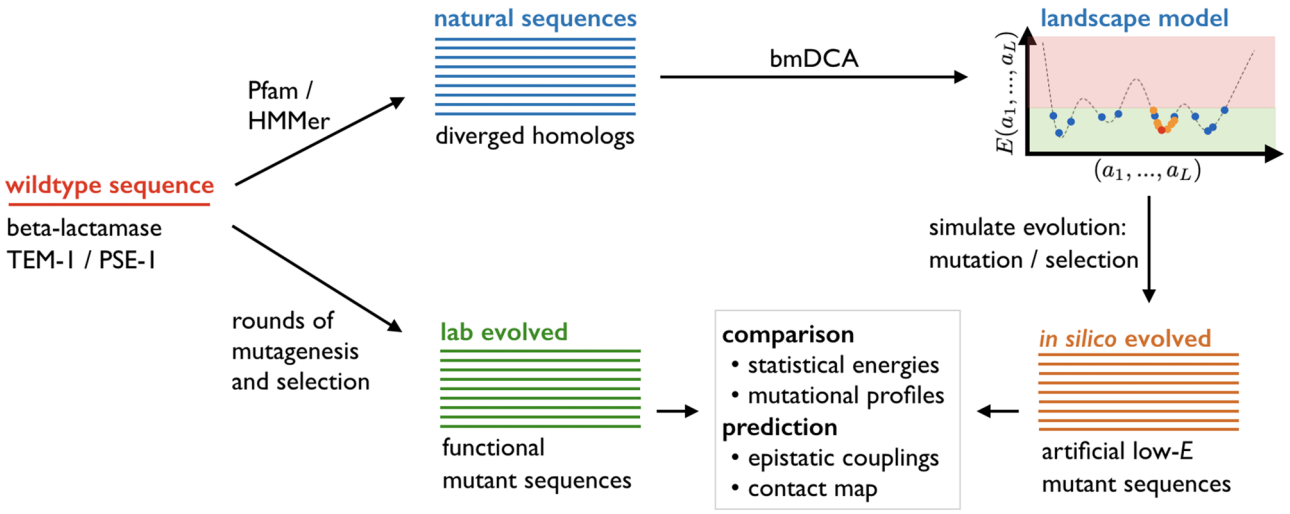


FIG. 1. Scheme of our evolutionary modeling approach: starting from a wildtype sequence (red), we collect a large multiple sequence alignment of naturally diverged homologs (blue), which are used to learn a generative landscape model using bmDCA (Figliuzzi et al. 2018). Evolution is simulated as a Markov process in this landscape, leading to simulated, or in silico evolved mutant sequences. These sequences can be compared with the results of evolution experiments (Fantini et al. 2020; Stiffler et al. 2020) (green), to assess estimated protein fitness (so-called statistical energies, compare below), mutational profiles, and DCA-based epistasis and contact prediction. The simulation scheme also allows for changing experimental control parameters like final sequence divergence, sequencing depth, and selection strength.

construction are given in Materials and Methods below; we find, for example, an MSA of 18,334 beta-lactamase sequences.

The underlying idea of our work is to represent the natural variability of this MSA via a generative statistical model $P(a_1, \ldots, a_L)$, with $(a_1, \ldots, a_L)$ representing an aligned amino acid sequence, that is, the a_i are either one of the 20 natural amino acids, or an alignment gap. Since data are limited, we need to assume some mathematical form for $P(a_1, \ldots, a_L)$. Introducing:

$$P(a_1, \ldots, a_L) = \frac{1}{Z} \exp\{-E(a_1, \ldots, a_L)\}, \qquad (1)$$

we write the "statistical energy" $E(a_1, \ldots, a_L)$, which is to be seen as a proxy for negative protein fitness (Morcos et al. 2014; Levy et al. 2017), in the form used by DCA (Weigt et al. 2009; Morcos et al. 2011; Cocco et al. 2018),

$$E(a_1, \ldots, a_L) = -\sum_i h_i(a_i) - \sum_{i<j} J_{ij}(a_i, a_j), \qquad (2)$$

as a sum over position- and amino acid-specific single-residue biases, or fields, $h_i(a_i)$ and pairwise epistatic residue–residue couplings $J_{ij}(a_i, a_j)$. This model, also known as Potts model, assigns low statistical energy E to "good/fit" sequences of high probability, and high E to "bad/unfit" nonfunctional sequences of low probability. As illustrated in figure 1, we expect to find low statistical energies for both natural and experimentally evolved sequences. The strongest couplings are known to be related to residue–residue contacts in the 3D protein structure, compare with (Morcos et al. 2011).

The model parameters are inferred by the currently most accurate version of DCA, called bmDCA (Figliuzzi et al. 2018), which maximizes the model's likelihood via Boltzmann-machine learning (Ackley et al. 1985). As is known from the literature (Sutto et al. 2015; Levy et al. 2017; Figliuzzi et al. 2018), this model is generative because sequences sampled from $P(a_1, \ldots, a_L)$ reproduce many statistical properties of the MSA of natural sequences. This does not only concern fitted quantities like one- and two-site amino acid frequencies, but also nonfitted properties like three-residue amino acid frequencies or the clustering of beta-lactamases into subfamilies in sequence space. Note that the epistatic couplings are essential for the model to be generative: a profile model having only fields $h_i(a_i)$ but no couplings $J_{ij}(a_i, a_j)$, that is, a model assuming statistical independence of all positions in the protein, is not generative in the rather strict sense discussed above (Figliuzzi et al. 2018). It misses both nontrivial second- and higher-order correlations and the clustered sequence distribution. Note also that, in a different protein family (chorismate mutase, PF01817), the same modeling approach was recently shown to artificially generate fully in vivo functional protein sequences (Russ et al. 2020).

To test the quantitative character of our landscape $E(a_1, \ldots, a_L)$, we compare the model predictions $\Delta E = E(\text{mutant}) - E(\text{wildtype})$ for the effect of mutations introduced into a wildtype sequence, with the results of a deep-mutational scan of the beta-lactamase TEM-1 (Firnberg et al. 2014). As is shown in figure 2A and B, the two are highly correlated, with a Spearman rank correlation of −0.77, compare also with (Figliuzzi et al. 2016) and (Hopf et al. 2017) and the scatter plot supplementary figure S1A, Supplementary Material online, directly comparing prediction and experiment. This correlation shows that our landscape $E(a_1, \ldots, a_L)$, even if inferred using distantly diverged TEM-1 homologs, provides quantitative information in the direct vicinity of TEM-1. As expected, low statistical energies correspond to high fitness values. To underline the importance of the epistatic couplings in

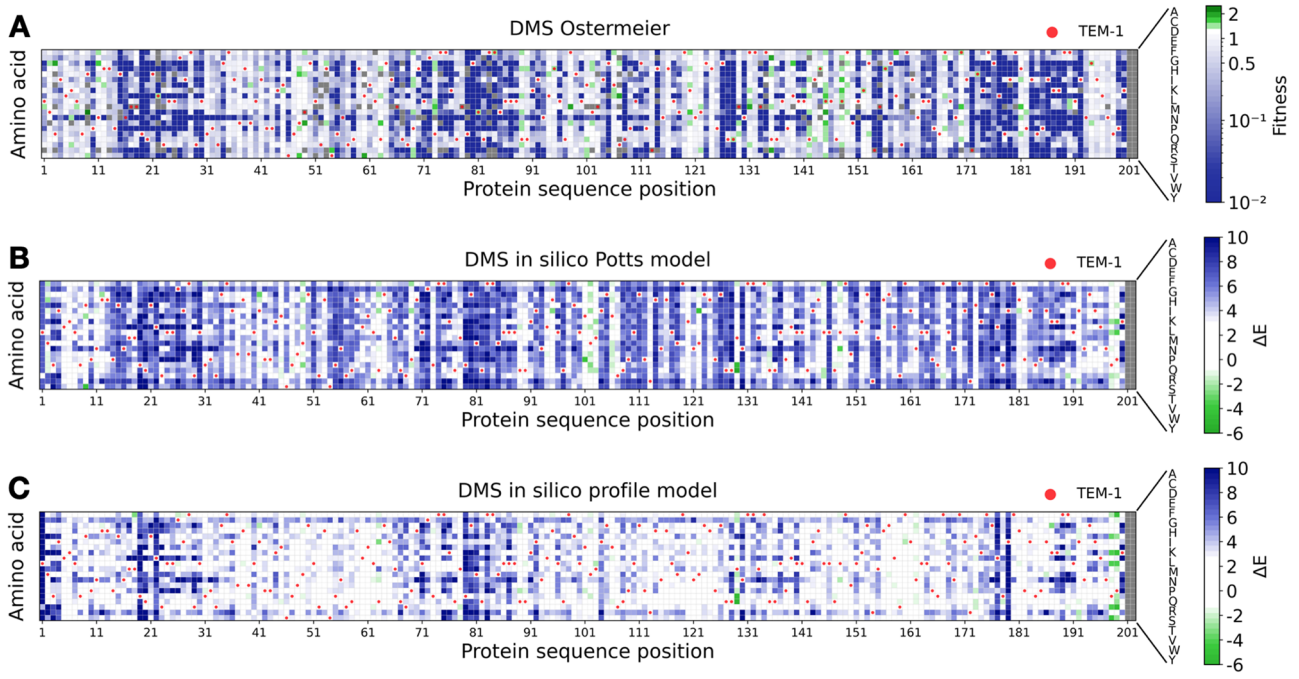


FIG. 2. Experimental and predicted mutational effects in TEM-1: panel (*A*) shows the results of the deep-mutational scanning experiment of (Firnberg et al. 2014), as compared with the computational predictions using the epistatic Potts model (*B*) and the nonepistatic profile model (*C*). Panels (*A*) and (*B*) have a Spearman rank correlation of −0.77, showing that low energies correspond to high fitness. Panels (*A*) and (*C*) have a reduced Spearman correlation of −0.6 due to the absence of epistatic couplings in the profile model.

our model, we also show in figure 2C and supplementary figure S1*B*, Supplementary Material online, the predictions of a nonepistatic profile model inferred from the same beta-lactamase MSA: the correlation with the experimental data decreases to −0.6, compare with (Figliuzzi et al. 2016).

This observation is central for our evolutionary model since the selection of sequences with few mutations with respect to the wildtype reference will be modeled by energy differences ΔE as introduced above.

A Model of Evolutionary Dynamics Reproduces Quantitative Features of Experimentally Evolved Sequences

Evolution (natural and experimental) can be seen as a stochastic process in a sequence landscape, with random mutations and phenotypic selection modeled by our statistical energy $E(a_1, \ldots, a_L)$. A minimal model realizing this idea is SEEC (de la Paz et al. 2020): a random site $i \in \{1, \ldots, L\}$ is selected, and an amino acid $b \in \{A, C, \ldots, Y\}$ is selected to substitute a_i with a probability proportional to $\exp\{-\Delta E(a_i \to b)\}$, with ΔE being the statistical-energy difference between the mutated and the unmutated sequences. A nonaccepted or synonymous mutation is characterized by $a_i{=}b$. Note that deletions and insertions are currently not considered in our model.

Although this model can be used to explore the qualitative influence of epistasis on protein sequence evolution, our analysis requires a more quantitative model taking in particular two differences into account:

- Mutations happen at the "nucleotide" level. As a consequence, not all amino acids are accessible from all amino acids via a single nucleotide mutation; and the set of accessible amino acids depends specifically on the used codon.
- The experiments allow to "vary selection strength." For TEM-1 and PSE-1, this is done by modifying the antibiotic concentration: the same mutation can be more or less strongly favored or suppressed.

To include these factors into our evolutionary model, we introduce two important modifications with respect to SEEC: first, we model evolution at the level of the nucleotide sequence $(n_{11}, n_{12}, n_{13}, \ldots, n_{i1}, n_{i2}, n_{i3}, \ldots, n_{L1}, n_{L2}, n_{L3})$ coding for the amino acid sequence $(a_1, \ldots, a_L)$, that is, the nucleotide triplet (n_{i1}, n_{i2}, n_{i3}) codes for amino acid a_i. For each possible codon $(n_1, n_2, n_3) \in \{A, C, G, T\}^3$ (with the exception of the stop codons), we introduce the set of amino acids $\mathcal{A}_{acc}(n_1, n_2, n_3) \subset \{A, \ldots, Y\}$, which are accessible from (n_1, n_2, n_3) by at most a single nucleotide mutation. Possible substitutions for a_i are now only selected from $\mathcal{A}_{acc}(n_{i1}, n_{i2}, n_{i3})$. Note that also a_i is in $\mathcal{A}_{acc}(n_{i1}, n_{i2}, n_{i3})$, accessible via its original codon and any synonymous mutation.

Second, selection strength will be regulated by a new parameter β, having the form of an inverse temperature $\beta = 1/T$ in statistical physics, which modifies the sequence probability to $P \sim \exp\{-\beta E\}$. The "low-temperature" case $\beta > 1$ ($T < 1$) corresponds to increased selection (e.g., higher antibiotic concentration, or directed evolution), in the limit $\beta \to \infty$ ($T \to 0$) only the best possible amino acid in position i

is accepted. The "high-temperature" case $\beta < 1$ ($T > 1$) corresponds to decreased selection (e.g., lower antibiotic concentration); the limit $\beta \to 0$ ($T \to \infty$) describes the case of mutation-accumulation experiments without selection.

This idea is implemented in the following three steps, which are iterated, compare with Materials and Methods for details:

(1) We randomly select a site $i \in \{1, \ldots, L\}$ to be mutated, corresponding to the codon $\mathbf{n}_i = (n_{i1}, n_{i2}, n_{i3})$ and the amino acid a_i.
(2) One of the accessible amino acids $b \in \mathcal{A}_{acc}(\mathbf{n}_i)$ is selected to substitute a_i with a probability $P(b|a_1, \ldots, a_{i-1}, a \;\; i+1, \ldots, a_L) \propto \exp\{-\beta \Delta E(a_i \to b)\}$. Due to the epistatic couplings in (equation 2), this probability depends explicitly on the sequence context $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_L)$.
(3) One out of the possible codons for amino acid b, which differs from $\mathbf{n}_i$ in at most a single nucleotide, is selected uniformly at random.

The resulting nucleotide and amino acid sequences remain thus mutually consistent.

The proposed dynamics can be efficiently implemented, and very large sequence libraries can be simulated over long times. To make these data comparable with the libraries generated by experimental evolution, we need to adapt the simulation parameters: first, the number of mutational steps in our simulation is not directly related to the number of experimental generations (because error-prone PCR may introduce multiple mutations each round); we choose it to reach the same average number of substituted amino acids in the simulated and experimental libraries. In this sense, different experimental mutation rates can be parametrized by the number of steps needed by our dynamics to reach the same number of mutations. Second, the selection strength $\beta = 1/T$ has no evident relation to the antibiotic concentration used in the experiment. We therefore tune the value of $\beta = 1/T$ such that the statistical energy $E(a_1, \ldots, a_L)$ of the simulated and the experimental sequences have the same linear slope as a function of the number of substitutions. For the case of PSE-1, shown in figure 3, we find that $T = 1.4$ is a good value, compare figure 3A for the experimental data from (Stiffler et al. 2020), and figure 3B for simulated data. This corresponds to low selection strength $\beta = 1/T < 1$. Even if we adjust only average distance and slope, we find that also the overall distribution is well reproduced. Similar observations for TEM-1 and AAC6 are shown in supplementary figures S2 and S3, Supplementary Material online.

Figure 3C shows that for strong selection $T = 0.05$ ($\beta = 20$) the sequence energy decreases with the number of substitutions, corresponding to an increasing fitness as expected in a directed-evolution scenario. Weak selection, shown in figure 3D for $T = 20$ ($\beta = 0.05$), corresponds to a sharp increase in statistical energy, and thus a loss in fitness, as expected from the accumulation of predominantly deleterious random mutations.

Figures show global measures comparing experimental and simulated sequences: the Hamming distance is the number of substitutions along the entire amino acid sequence, the energy also depends on the entire sequence. To increase our confidence in the quantitative character of our evolutionary model, we compare in figure 4 the site- and amino acid-specific mutational frequencies between experimental and simulated sequence data. To this end, we extract the quantities $f_i(a)$ describing the fraction of sequences in an MSA having amino acid a in position i. Interestingly, also this refined measure of sequence diversity is very similar for simulated and experimental sequences; we observe a high correlation of 86%, compare with supplementary figure S4, Supplementary Material online. These plots highlight the importance of working only with amino acid substitutions accessible via single-nucleotide mutations: many amino acids show zero frequency in both plots due to inaccessibility. The mutational spectrum predicted without considering the accessibility of amino acids is shown in supplementary figure S4, Supplementary Material online: we see that the mutational frequencies are more homogeneously distributed, close-to-zero frequency mutations become very rare as compared with the experimental sequences. The correlation goes down to 65% between simulated and experimental data in this case.

Based on these observations, we conclude that our evolutionary model, which combines mutations at the nucleotide level with selection at the amino acid level, is able to reproduce well the statistical features of the experimental sequences. This conclusion is also confirmed, when using TEM-1 and AAC6 as initial wildtype sequences, compare with supplementary figures S5 and S6, Supplementary Material online.

In Silico Sequence-Space Exploration, and the Emergence of Epistatic Signals

Having developed a quantitative model to simulate experimental evolution, we are now able to explore evolutionary scenarios going well beyond those realized in the experiments. We can systematically analyze the influence of the sequence divergence from wildtype, of the sequenced library depth, and of the selection strength on the accuracy of coevolution-based contact prediction. Each setting of these parameters would require long experiments and would sometimes be inaccessible due to the high number of experimental rounds or the depth of the sequenced library.

Computationally this becomes straightforward although intensive: we have performed many runs of evolutionary simulations, each producing an MSA with specific parameters, simulating the possible outcome of an evolutionary experiment, as represented in figure 5. Each square in these plots corresponds to the average over five simulation runs. Depicted is the positive predictive value (PPV), which measures the fraction of true positive contact predictions within the first 100 contact predictions, compare with Materials and Methods for details. Due to the large number of contact predictions to be performed, we used GaussDCA (Baldassi et al. 2014), a very fast, even if not the most accurate contact

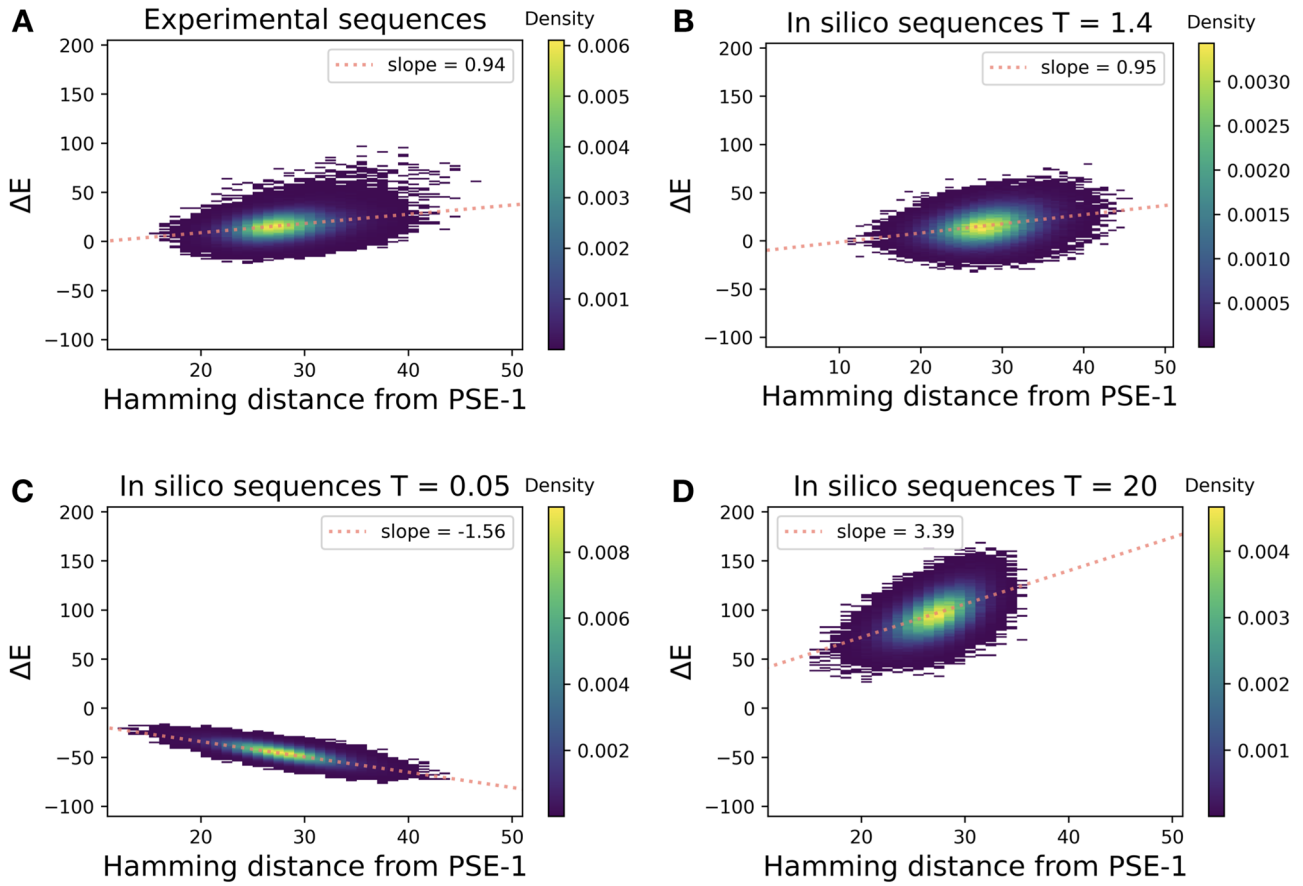


FIG. 3. Statistical energy in dependence of sequence distance from wildtype: panel (*A*) shows the statistical energies of the sequences from generation 20 in Stiffler et al., as a function of the Hamming distance (number of substituted amino acids) from the wildtype PSE-1. Panel (*B*) shows the same quantities for the in silico simulated sequences, where selection strength T and the number of simulated evolutionary steps are adjusted to reproduce the average distance and the slope from panel (*A*). Panel (*C*) shows an example of strong selection ($T \ll 1$) leading to optimized sequences having lower statistical energies/higher fitness. Panel (*D*) shows the case of very weak selection ($T \gg 1$) resulting in random, mostly deleterious substitutions strongly increasing statistical energy.

predictor. Figure 5*A* shows the plot for the selection strength used in the experiments for PSE-1. The red zone corresponds to inaccurate contact predictions, being sometimes hardly better than random (PPV $\sim$ 0.13). It is found consistently for small sequence libraries, and for sequence libraries of low divergence from wildtype. It becomes evident that we need to go to a sufficient number of simultaneous mutations to be able to detect at least a weak epistatic signal between mutations, which can be used for contact prediction. However, this signal remains weak: we need much larger sequence libraries of at least about 50,000 sequences to reach a reasonable contact prediction. However, even for the largest and most diverged library we have studied, a PPV of only 0.7–0.8 is reached, which remains below the contact prediction reached by using the MSA of natural homologs, which was used before for the inference of our sequence landscape. The latter reaches a PPV of 0.98 using GaussDCA. Figure 5*B* shows the same observables for experiments starting with the TEM-1 sequence, the overall results are very similar to PSE-1, even if some quantitative details depend on the initial wildtype sequence.

It might be speculated that better contact-prediction algorithms may shift the region of nontrivial predictions down to lower Hamming distances from wildtype, or to lower sequence numbers. Although the computational cost of plmDCA is too high to reproduce the full analysis of figure 5, we have reanalyzed two columns at average Hamming distance 41 and 65. As is shown in supplementary figure S7, Supplementary Material online, for low sequence numbers GaussDCA and plmDCA give very similar low prediction accuracies, whereas the improved accuracy of plmDCA over GaussDCA becomes visible only at sufficiently high sequence numbers. At the resolution of our analysis, no shift in the boundary is observable.

The conditions of the experiments for PSE-1 and TEM-1 are highlighted, in the two panels of figure 5. For PSE-1, 20 rounds of evolution led to an average sequence distance of 27 amino acid substitutions from wildtype, and a sequenced library of 165,000 distinct sequences (Stiffler et al. 2020). Interestingly, this point is located slightly beyond the boundary of emergence of coevolutionary signal. The predicted average PPV of 0.58 is comparable with the 0.65 obtained using

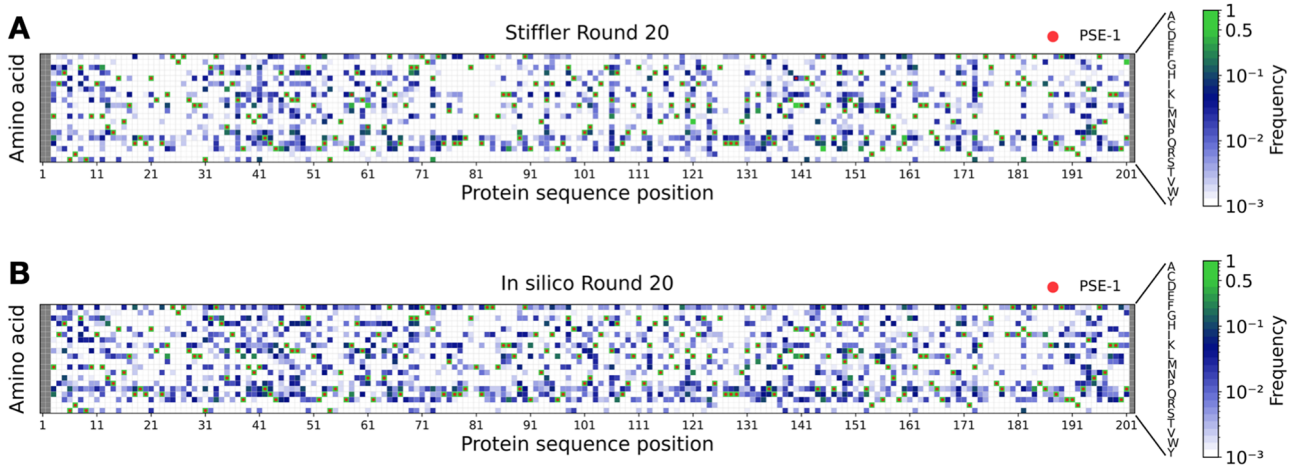


FIG. 4. Position-specific amino acid frequencies for experimental and simulated sequence libraries: panel (*A*) shows the frequencies $f_i(a)$ of usage of amino acid *a* in site *i* in round 20 of experimental PSE-1 evolution, panel (*B*) shows the same quantity for simulated evolution. The Spearman rank correlation between the two frequency spectra is 86%.

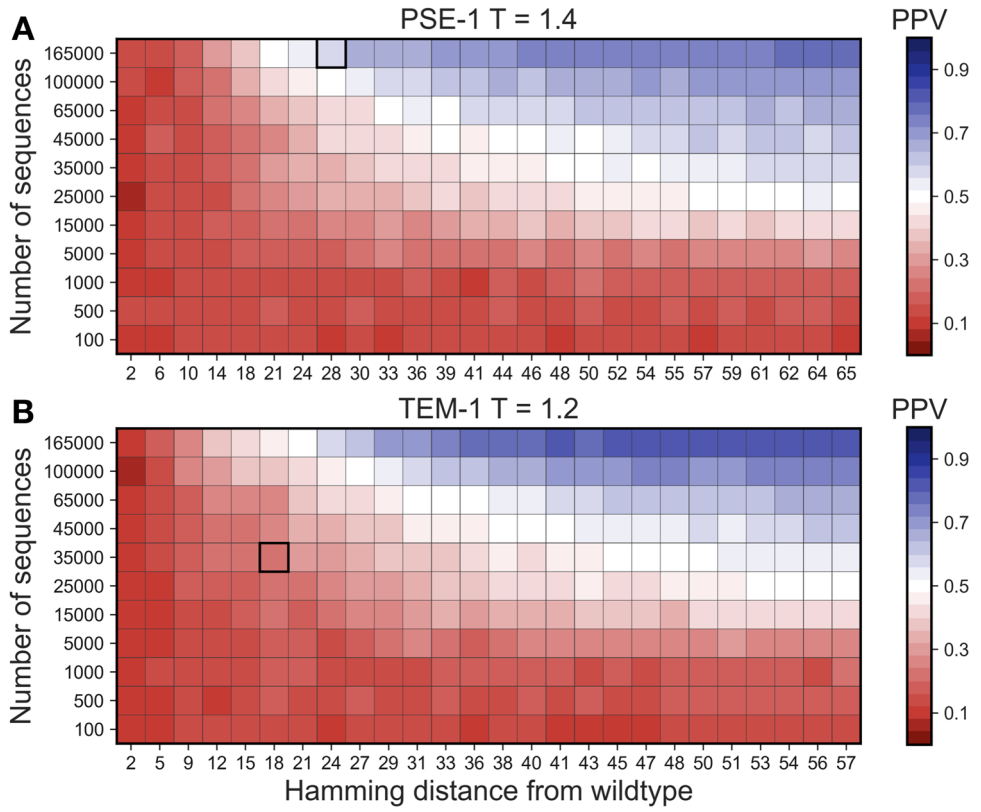


FIG. 5. Accuracy of contact prediction as a function of sequence number and sequence divergence: panel (*A*) shows the accuracy of contact prediction as a function of the average sequence divergence from wildtype PSE-1 and the depth of the sequenced library. The accuracy is measured via the PPV, that is, the fraction of true positive contact predictions in the first 100 DCA-predicted contacts, compare with Materials and Methods for details. The selection strength $T = 1.4$ corresponds to the experimental condition in (Stiffler et al. 2020). The highlighted square indicates an average Hamming distance of about 27 and a sequence library of 165,000, as realized in (Stiffler et al. 2020). Panel (*B*) shows the same quantities for wildtype TEM-1, and for the experimental conditions used in (Fantini et al. 2020).

the experimental MSA, compare with Materials and Methods section.

This is in contrast to the TEM-1 experiment of (Fantini et al. 2020), compare with figure 5*B*: the experiment was performed for fewer rounds, leading to less divergence from TEM-1, and the sequence library was less deeply sequenced. The resulting library, with an average Hamming distance of 18 from TEM-1 and with 34,431 unique sequences, is located slightly below the line of emergence of coevolution signal. This observation provides

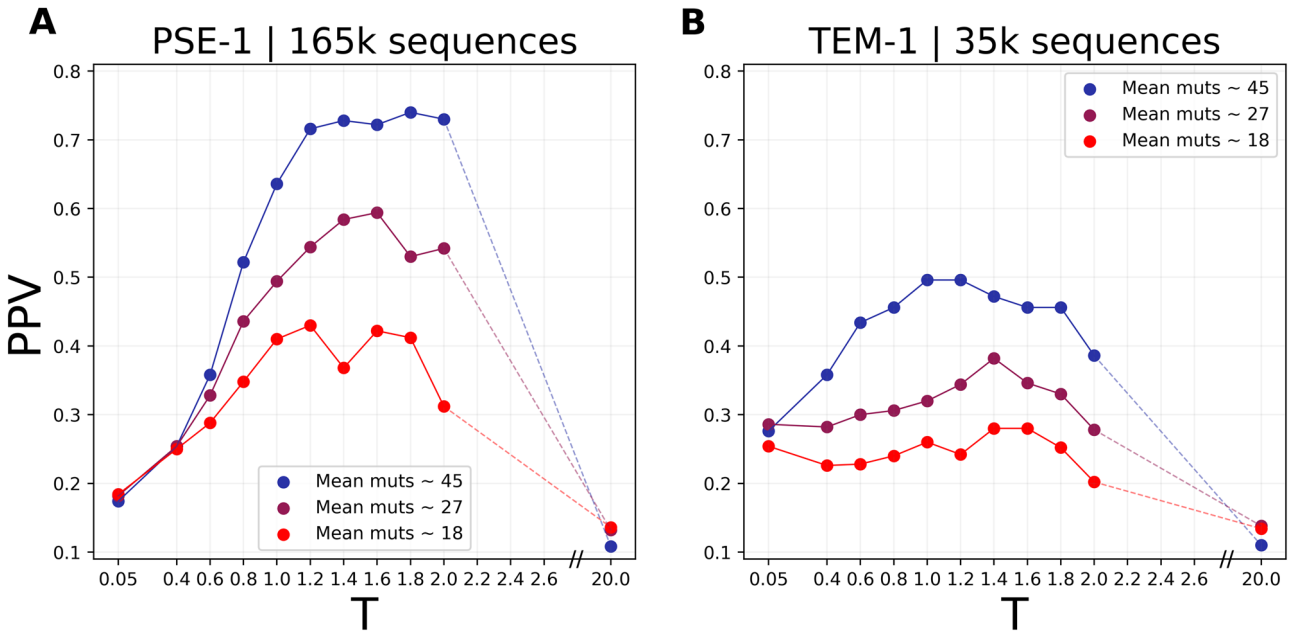


FIG. 6. Dependence of the contact-prediction accuracy on selection strength: we show the PPV (100 predicted contacts) of simulated MSAs at variable selection strength T (panel *A* for PSE-1, panel *B* for TEM-1), and for different sequence distances from the wildtype protein. We predict that, for the distances observed in the evolution experiments (27 for PSE-1, 18 for TEM-1), both experiments would have benefited from slightly lower antibiotic concentrations.

a potential explanation for the observed reduced performance in contact prediction.

The AAC6 results show that reduced sequence divergence can, at least partly, be compensated by a strong increase in the number of sequences in the evolved MSA, compare with supplementary figure S8, Supplementary Material online, which confirms original findings of (Stiffler et al. 2020). Even if having only an average Hamming distance of about 8 substitutions, the large library of more than 10^6 sequences allows for the detection of a weak contact-related signal.

The results depend substantially on the strength of selection. Supplementary figure S9, Supplementary Material online, shows the extreme cases of very strong and very weak selection discussed before. Both show inaccurate prediction. An important difference becomes visible when looking at the horizontal axes: all use the same number of simulated evolutionary steps. In the case of strong selection, sequences stay closer to the wildtype, since most mutations are deleterious and selected against, and they stay close to each other. So while being all functional, they do not accumulate sufficient sequence variability to provide a reliable epistatic signal. In the case of extremely weak selection, almost all mutations are acceptable. Sequences are found to diverge strongly from the initial PSE-1 sequence, but the absence of selection causes also an absence of coevolution.

Discussion

The aim of this work was to showcase the potential of evolutionary models in data-driven sequence landscapes. Recent progress in landscape modeling has led to advances in using sequence alignment to predict protein structure, mutational effects, and even to design non-natural but biologically functional sequences. Here we show that, equipped with a simple stochastic dynamics capturing the interplay between mutation and selection, these landscapes lead to models which are able to describe in a quantitatively accurate way the results of evolution experiments. This is not only restricted to proteins, as studied in this work, but similar evolution experiments have been performed for RNA (Zhou et al. 2018) and could therefore be analyzed in an analogous way starting from sequence landscapes for RNA families (Kalvari et al. 2021).

The applications for experimental evolution are evident: we can use our modeling to optimize experimental evolution protocols, for example, when we search for fully functional sequences but at some minimum number of mutations from a starting sequence, or when we want to explore sequence space optimally for contact prediction. In this case, we could, for example, optimize the selection strength. In the case of the beta-lactamases studied in this article, figure 6 shows that a slightly lower selection pressure (i.e., higher selection temperature) would have led to even better contact predictions. However, this potential increase is weak as compared with the one reachable by more diverged sequences.

A possible obstacle in such applications is the fact that the selection temperature T, which we use to model selective pressure, has to be fitted from experimental data via the slope of the statistical energies of the evolved sequences vs. their distance from wildtype. To understand the minimal sequence requirements for reaching robust and accurate slope estimates, we have subsampled the experimental sequence libraries of PSE-1 for rounds 10 and 20. As is shown in supplementary figure S10, Supplementary Material online, we observe: 1) that the slope

can be estimated accurately already from about 200–300 sequences, whereas the estimation error becomes large when using less than 100 sequences, and 2) that the estimates are almost equal for round 10 and round 20. We conclude that the selection temperature T can be reliably determined with moderate experimental effort (low number of sequences, few experimental rounds). Once estimated, the parameters can be used in simulations, which may guide more massive experiments evolving large sequence libraries over many rounds.

We see our current model as a starting point for more detailed evolutionary models. There is space for a substantial gain in accuracy: we can introduce biases in the mutations introduced by error-prone PCR directly into the model (Moore and Maranas 2000; Pritchard et al. 2005), the latter can be derived from data by analyzing synonymous mutations. Furthermore, we can introduce codon bias, the difference between transitions and transversions, the fact that error-prone PCR may introduce simultaneously several mutations before selection, or the emergence of phylogeny in cycles of mutation and selection.

The modeling can also benefit from experimental feedback. If sequence libraries would also be sequenced before and after the selection step, we could establish a better correspondence between statistical energies and selection, up to a gauge of statistical energies vs. antibiotic concentrations.

However, the potential of such evolutionary models in data-driven landscapes goes far beyond the application to experimental evolution. As is shown by SEEC (de la Paz et al. 2020), already the simplest nontrivial evolutionary model allows for illuminating important consequences of epistasis in evolution, like the site- and time-dependence of substitution rates. We anticipate that the proposed modeling framework may capture many of these effects in a highly quantitative way. The relatively simple modeling framework proposed in our paper might also be a starting point for more theoretical–mathematical analyses about, for example, the emergence of epistatic signals in sequence libraries. In this context, it might also be interesting to see in how far more distributed signatures of epistatic signal, possibly related to protein function rather than contacts, become visible in experimentally evolved sequence libraries, compare with (Rivoire et al. 2016), (Shimagaki and Weigt 2019), and (Tubiana et al. 2019).

Materials and Methods

Sequence Data

Sequences from Experimental Evolution

We include in our analysis the sequence data coming from the experiments of in vitro evolution by (Fantini et al. 2020) on TEM-1 and by (Stiffler et al. 2020) on PSE-1 and AAC6.

The aligned amino acid sequences from (Fantini et al. 2020) were kindly provided by the authors prior to publication, and can also be found at http://laboratoriobiologia.sns.it/supplementary-mbe-2019/ (last accessed November 17, 2021). The raw sequencing reads are available at the National Centre for Biotechnology Information Sequence Read Archive (SRA) with accession code PRJNA528665 (http://www.ncbi.nlm.nih.gov/sra/PRJNA528665, last accessed November 17, 2021). Amino acid sequences with more than six gaps were discarded as a quality control to remove sequences with lower quality.

Stiffler et al. (2020) ran two experiments using the PSE-1 beta-lactamase and the AAC6 acetyltransferase as starting wildtypes. Aligned sequencing reads from the last round of the two experiments (translated into amino acid sequences) can be found at https://github.com/sanderlab/3Dseq (last accessed November 17, 2021). The raw sequencing reads are available at the National Centre for Biotechnology Information Sequence Read Archive (SRA) with accession code PRJNA578762 (http://www.ncbi.nlm.nih.gov/sra/PRJNA578762, last accessed November 17, 2021).

Our models are built for the Pfam-annotated positions using the corresponding Pfam domains PF13354 (Beta-lactamase2) and PF00583 (Acetyltransf1). We realigned the wildtype sequence using the hmmalign command from the HMMer software suite (Eddy 2011) and profile Hidden Markov Models downloaded from Pfam (Mistry et al. 2021). We then removed from the experimental MSAs all columns corresponding to nonmatched states of the wildtype sequence.

The resulting MSAs of experimentally evolved sequences have 202 sites and 165,855 sequences for PSE-1 (round 20), and 34,431 sequences for TEM-1 (generation 12). For AAC6, we find 117 sites and 1,260,048 sequences (round 8).

Natural Homologous Sequences and Preprocessing of the Training Set

The MSAs of natural homologous sequences of the two considered protein families PF13354 (Beta-lactamase2) and PF00583 (Acetyltransf1) were generated running the hmmsearch command from the HMMer software suite (Eddy 2011) on the UniProt database (The UniProt Consortium 2021). Insertions were removed, and sequences with more than 10% gaps and duplicated sequences were excluded to improve the quality of the alignment. Any sequence closer than 80% to the wildtypes TEM-1, PSE-1, or AAC6 was excluded from the alignments to avoid the introduction of biases toward these sequences in the bmDCA learning. The resulting MSAs included 18,333 (43,576) homologous and nonidentical aligned sequences of length 202 (117) for PF13354 (PF00583).

Note that some residues, which are present in the N- and C-terminal regions of the experimental sequences, are not covered by the Pfam domains, and therefore excluded from our analyses. Extending the MSA beyond the borders of the Pfam domains would lead to the inclusion of evolutionarily less conserved positions, and thus to the inclusion of highly gapped columns into the MSA of natural data. Such columns have been previously found to compromise the accuracy of DCA landscapes (Figliuzzi et al. 2016) and are therefore left out in this study.

The natural MSA were used to train two Potts models using bmDCA (Figliuzzi et al. 2018) in the implementation of Barrat-Charlaix et al. (2021), which provides the currently most accurate DCA models.

Evolutionary Model

As already discussed in Results section, our evolutionary model combines mutations at the nucleotide level with selection at the level of aligned amino acid sequences. We therefore need to specify both the nucleotide sequence $\mathbf{n} = (n_{11}, n_{12}, n_{13}, \ldots, n_{i1}, n_{i2}, n_{i3}, \ldots, n_{L1}, n_{L2}, n_{L3})$ and the resulting amino acid sequence $\mathbf{a} = (a_1, \ldots, a_L)$, which is translated from $\mathbf{n}$ using the standard genetic code. Since we consider full-length aligned sequences of Pfam domains, stop codons are not allowed in $\mathbf{n}$. Furthermore, we have to accommodate alignment gaps possibly existing in $\mathbf{a}$: a gap in $\mathbf{a}$ is represented by a triplet of gaps in $\mathbf{n}$. Gaps are not changed during our simulations, our model does consider only single-nucleotide substitutions, but no insertions and no deletions. Note that the gray columns in figure 4 and supplementary figures S5 and S6, Supplementary Material online, correspond to gaps in the wildtype sequence, which are conserved both in the experiment and in the model.

As mentioned before, for each codon $(n_1, n_2, n_3) \in \{A, C, G, T\}^3$, we consider the set of amino acids $\mathcal{A}_{acc}(n_1, n_2, n_3) \subset \{A, \ldots, Y\}$, which are accessible from (n_1, n_2, n_3) by at most a single nucleotide mutation.

Our simulation of sequence evolution proceeds by iterating the following three steps defining a Markov chain (MC) in the space of nucleotide sequences (note that, due to the degeneracy of the genetic code, the process is "not" an MC in amino acid sequence space):

(1) A position $i \in \{1, \ldots, L\}$ is chosen uniformly at random along the amino acid sequence, corresponding to the codon $\mathbf{n}_i = (n_{i1}, n_{i2}, n_{i3})$ and the amino acid a_i. Although $a_i = "-"$, that is, a gap is chosen, we repeat the selection of the position i.
(2) Out of all accessible amino acids $b \in \mathcal{A}_{acc}(\mathbf{n}_i)$, we selected one using the conditional probability $P_\beta(b|\mathbf{a}_{-i})$, which couples the amino acid b explicitly to the sequence context $\mathbf{a}_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_L)$:

$$P_\beta(b|\mathbf{a}_{-i}) = \frac{\exp\{\beta h_i(b) + \beta \sum_{j(\neq i)} J_{ij}(b, a_j)\}}{z_i(\mathbf{a}_{-i})}, \quad (3)$$

with

$$z_i(\mathbf{a}_{-i}) = \sum_{b \in \mathcal{A}_{acc}(\mathbf{n}_i)} \exp\{\beta h_i(b) + \beta \sum_{j(\neq i)} J_{ij}(b, a_j)\} \quad (4)$$

being a normalization constant. In difference to Z in equation (1), it can be calculated efficiently by summing over the less than 20 accessible amino acids.

(3) One out of the possible codons for amino acid b, which differs from $\mathbf{n}_i$ in at most a single nucleotide, is selected uniformly at random.

The new amino acid b substitutes a_i in $\mathbf{a}$, and the new codon $\mathbf{n}_i$ in $\mathbf{n}$. We thereby conserve the coherence between nucleotide and amino acid sequence.

To simulate an entire MSA of M sequences, the process is initiated M times in the wildtype reference sequence, and M independent runs of the MC are performed. The number of steps in these MCs is chosen such that the average Hamming distance of the generated amino acid sequences reaches a target number. Note that the Hamming distances may vary from MC to MC, since $\mathcal{A}_{acc}(\mathbf{n}_i)$ contains the case $b{=}a_i$ accessible via any synonymous mutations. The Hamming distance can therefore assume any value between zero and the number of performed mutational steps.

Simulated Sequence Data for Contact Prediction

Our evolutionary algorithm has three input parameters adding to the wildtype sequence and the statistical-energy model: the number of sequences M, the number N_{MC} of steps of our evolutionary MC model, and the selection temperature T. Given this triplet of numbers it outputs an MSA obtained simulating evolution for N_{MC} iterations starting from the wildtype sequence, repeating the sampling independently M times at temperature $T = 1/\beta$.

For each wildtype sequence, we simulated the outcome of different protein evolution experiments by scanning these three input parameters within a range of interest. For MSA generated starting from TEM-1 or PSE-1 (AAC6), we varied M in the range $100 - 165,000$ ($500 - 1,250,000$), N_{MC} in the range 5–255 (4–120), and T in the range 0.05–20.

To save resources and time, given the computational cost of sampling, we opted for a scheme that would allow us to reduce the number of independent MC chains needed to simulate evolution. For each temperature T, we run 165,000 (1,250,000) independent MCs for TEM-1 and PSE-1 (AAC6) and printed MSAs at the desired number of MC steps until 255 (120) MC steps. The MSAs with less sequences were obtained by randomly subsampling without replacement from the MSA with 165,000 (1,250,000) sequences. To produce more statistics, we ran the same simulations five times.

Contact Prediction

Contact prediction was performed using GaussDCA (Baldassi et al. 2014) for all MSAs, included, for coherence, the experimental ones. GaussDCA is the computationally most efficient implementation of DCA. Its accuracy of contact prediction is slightly inferior to plmDCA or bmDCA. However, we use it: in our analysis, we had to predict contacts for a large number of partially deep simulated MSAs (cf., fig. 5) to explore multiple combinations of sampling time, sample size, and selection strengths.

The reweighting parameter was set to 0 for contact prediction of in silico MSAs, as this reduces computational time and is coherent with the independence of the simulated MCs. On the other hand, contact prediction of experimental MSAs

was performed using the default option ":auto" of GaussDCA for reweighting. These different treatments of simulated and experimental sequences are based on the fact, that simulations generate statistically independent sequences (conditioned to wildtype initialization), whereas the experiments may generate sequence ensembles having nontrivial phylogenetic effects. The pseudocount was set to 0.6 (0.5) for PSE-1 and TEM-1 (AAC6) empirically, as we found it to be a good intermediate value for MSAs with very different statistics.

Intrachain atomic distances for both families were obtained by running the single-protein mode of the code provided by Pfam Interactions (https://doi.org/10.5281/zenodo.4080947, last accessed November 17, 2021), we used the shortest distance between heavy atoms of the two amino acids among all structures of the Protein Data Bank (PDB) (Burley et al. 2021) listed in Pfam. Following standards in coevolutionary contact prediction, all pairs with distance below 8 Å and a minimum separation of 5 positions along the sequence are kept as contacts for the calculation of the PPV. For AAC6, we used a more stringent cutoff of 5.5 Å, since the structural variability across the protein family is already well represented in the PDB.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to the authors of (Fantini et al. 2020), who shared their data with us prior to publication, and in particular M. Fantini, A. Pastore, and P. de los Rios for interesting discussions about our first results. We also thank Anna Paola Muntoni for contributing to the implementation of bmDCA used in this work, and to Giancarlo Croce, Kai Shimagaki, Jeanne Trinquier, Simona Cocco, and Remi Monasson for useful discussions. This work was partially funded by the EU H2020 Research and Innovation Programme MSCA-RISE-2016 under Grant Agreement No. 734439 InferNet (M.W.), and by a grant from the Simons Foundation (No. 454955, F.Z.).

Data Availability

The sequence data underlying this article are available on Github at https://github.com/matteobisardi/DataSeqEvol. The datasets were derived from sources in the public domain (see Materials and Methods). The software implementation for the evolutionary simulations, SeqEvol, is available on Github at https://github.com/matteobisardi/SeqEvol, together with the instructions to reproduce the results of this work.

References

Ackley DH, Hinton GE, Sejnowski TJ. 1985. A learning algorithm for Boltzmann machines. *Cogn Sci*. 9(1):147–169.

Arnold FH. 1998. Design by directed evolution. *Acc Chem Res*. 31(3):125–131.

Arnold FH. 2018. Directed evolution: bringing new chemistry to life. *Angew Chem Int Ed Engl*. 57(16):4143–4148.

Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. 2011. Learning generative models for protein fold families. *Proteins* 79(4):1061–1078.

Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. 2014. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* 9(3):e92721.

Barrat-Charlaix P, Muntoni AP, Shimagaki K, Weigt M, Zamponi F. 2021. Sparse generative modeling via parameter reduction of Boltzmann machines: application to protein-sequence families. *Phys Rev E*. 104(2):024407.

Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, et al. 2021. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res*. 49(D1):D437–D451.

Cadwell RC, Joyce GF. 1992. Randomization of genes by PCR mutagenesis. *PCR Methods Appl*. 2(1):28–33.

Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M. 2018. Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys*. 81(3):032601.

De Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. *Nat Rev Genet*. 14(4):249–261.

de la Paz JA, Nartey CM, Yuvaraj M, Morcos F. 2020. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc Natl Acad Sci U S A*. 117(11):5873–5882.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, United Kingdom: Cambridge University Press.

Eddy SR. 2011. Accelerated profile hmm searches. *PLoS Comput Biol*. 7(10):e1002195.

Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys*. 87(1):012707.

Fantini M, Lisi S, De Los Rios P, Cattaneo A, Pastore A. 2020. Protein structural information and evolutionary landscape by in vitro evolution. *Mol Biol Evol*. 37(4):1179–1192.

Figliuzzi M, Barrat-Charlaix P, Weigt M. 2018. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol Biol Evol*. 35(4):1018–1027.

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol*. 33(1):268–280.

Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol*. 31(6):1581–1592.

Greener JG, Kandathil SM, Jones DT. 2019. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun*. 10(1):1–13.

Haldane A, Levy RM. 2019. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Phys Rev E*. 99(3):032405.

Hopf TA, Green AG, Schubert B, Mersmann S, Schärfe CP, Ingraham JB, Toth-Petroczy A, Brock K, Riesselman AJ, Palmedo P, et al. 2019. The evcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 35(9):1582–1584.

Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. 2017. Mutation effects predicted from sequence co-variation. *Nat Biotechnol*. 35(2):128–135.

Jones DT, Buchan DW, Cozzetto D, Pontil M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021.

Highly accurate protein structure prediction with alphafold. *Nature* 596(7873):583–589.
Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. 2021. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res*. 49(D1):D192–D200.
Levy RM, Haldane A, Flynn WF. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr Opin Struct Biol*. 43:55–62.
Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, Tosatto SC, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 49(D1):D412–D419.
Moore GL, Maranas CD. 2000. Modeling DNA mutation and recombination for directed evolution experiments. *J Theor Biol*. 205(3):483–503.
Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*. 108(49):E1293–E1301.
Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc Natl Acad Sci U S A*. 111(34):12408–12413.
Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. 2017. Protein structure determination using metagenome sequence data. *Science* 355(6322):294–298.
Pritchard L, Corne D, Kell D, Rowland J, Winson M. 2005. A general model of error-prone PCR. *J Theor Biol*. 234(4):497–509.
Rivoire O, Reynolds KA, Ranganathan R. 2016. Evolution-based functional decomposition of proteins. *PLoS Comput Biol*. 12(6):e1004817.
Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, Hilvert D, Monasson R, Cocco S, Weigt M, et al. 2020. An evolution-based model for designing chorismate mutase enzymes. *Science* 369(6502):440–445.
Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706–710.
Shimagaki K, Weigt M. 2019. Selection of sequence motifs and generative Hopfield-Potts models for protein families. *Phys Rev E*. 100(3):032128.
Stiffler MA, Poelwijk FJ, Brock KP, Stein RR, Riesselman A, Teyra J, Sidhu SS, Marks DS, Gauthier NP, Sander C. 2020. Protein structure from experimental evolution. *Cell Syst*. 10(1):15–24.
Sutto L, Marsili S, Valencia A, Gervasio FL. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci U S A*. 112(44):13567–13572.
The UniProt Consortium. 2021. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 49(D1): D480–D489.
Tubiana J, Cocco S, Monasson R. 2019. Learning protein constitutive motifs from sequence data. *Elife* 8:e39397.
Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596(7873):590–596.
Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A*. 106(1):67–72.
Xu J. 2019. Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci U S A*. 116(34):16856–16865.
Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A*. 117(3):1496–1503.
Zhou Q, Kunder N, De la Paz JA, Lasley AE, Bhat VD, Morcos F, Campbell ZT. 2018. Global pairwise RNA interaction landscapes reveal core features of protein recognition. *Nat Commun*. 9(1):1–10.

2.2.1 Limits of the evolutionary model

The modeling approach chosen to study neutral drift experiments described in this paper is based on a highly simplified dynamics. Nonetheless, we show that artificial sequences simulated with our method are in quantitative agreement with the experimental libraries. On one side, we assessed the relative importance of two experimental parameters in determining contact prediction accuracy, namely the number of sequences in the library and their divergence from the wild-type sequence; on the other side, the artificial libraries that we generated had a mutational profile statistically similar to that of the experimental libraries. We turn now to discussing the simplifications of our evolutionary dynamics, to understand its limitations in modeling neutral drift experiments and, more generally, protein evolution:

- *Mutational biases*: mutations were introduced via epPCR during the neutral drift experiments discussed in section 2.2: random nucleotide mutations were disseminated in the surviving genes at each new round of evolution. Unfortunately, these mutations were not evenly distributed and presented mutational biases. In particular, the type of nucleotide already present in a specific position biased the kind of substitutions possible. For the moment, our dynamics does not take this fact into account, as new mutations are proposed independently from the current sequence.

- *Stationary distribution*: in the long run, we would like the sequences simulated with our dynamics to reproduce the natural statistics of the beta-lactamase family. However, our sampling method does not respect detailed balance, a condition that guarantees that the stationary distribution of a given Markov chain converges to a specific probability distribution. In our case, we want the target distribution to be the sequence landscape approximated by our Potts model. This is not a problem by itself when modeling short-term evolution but it is desirable when modeling protein diversification over the time scales of natural evolution.

- *Transition probability*: when running the Markov chains that simulate the evolutionary process, there exist many different local dynamics, or transition probabilities, that satisfy the detailed balance condition. In practice, all members of this class of dynamics guarantee the convergence to the target distribution in the long run, however, on shorter timescales, there are differences in the kind and frequency of mutations sampled. We chose the Gibbs sampling dynamics due to its empirical superiority, but it is still unclear why the Metropolis-Hastings algorithm had a poorer performance. In practice, we need to explore more

sampling algorithms and understand their differences in short time frames.

- *Population genetics*: our modeling approach lacks competition between strains, i.e. each sequence evolves independently. This greatly simplifies the generation of the artificial sequences and allows for a parallel implementation. Mutations and selection act on each protein site independently from the rest of the population of sequences in the library. As a consequence, mutations are not accepted or rejected based on population sweeps, or survival of the fittest strains in the pool, but rather through an effective dynamics that selects for good enough mutations. We need to implement a version of the algorithm in which mutations are selected on a population level, by devising an appropriate fitness function.

In the next sections, we will address the first two points detailed above to better account for the experimental details in one case, and to define a sampling dynamics that can describe longer evolutionary timescales. The other points are left for future work.

2.3 INCLUSION OF MUTATIONAL BIASES

In this section, we discuss a first attempt to include mutational biases in our simulations. The results described are preliminary and further research is needed. Part of the research presented has seen the participation of two master's students: Tiziri Terkmani and Aya Elmesaoudi. To test our approach we focused on the library of beta-lactamase sequences gathered from round 20 of the experiment of Stiffler et al. [110]. We chose it for two reasons: the sequences in this library are in high number and the simulations presented in our paper show that on this dataset we have the highest predictive accuracy with the base model.

2.3.1 Quantification of the experimental bias

First, we need to quantify the mutational bias induced by the experiment. To this aim, we require the nucleotide sequences, which we will use to analyze biases in the synonymous codons. The supplementary data accompanying the original paper exclusively listed the amino acid sequences of the genes that underwent neutral genetic drift. Furthermore, the dataset of amino acid sequences was pre-processed to enhance the accuracy of subsequent contact prediction, removing hundreds of thousands of sequences. To overcome this limitation, we collected the experimental raw PacBio sequencing reads (`https://www.ncbi.nlm.nih.gov/bioproject/?term=`

PRJNA578762) in FASTQ format from the Sequence Read Archive and we cleaned them with a Matlab code adapted from the one released by the original paper. We downloaded the DNA sequence of the wt PSE-1 gene from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/api/fasta/AAA25741.1?lineLimit=1000). Upon cleaning the reads from round 20 we obtained more than 450,000 aligned nucleotide sequences of round 20 of the genetic drift experiment of [110].

From this dataset, we were able to measure the prevalence of mutational biases. A mutational bias in our context refers to a non-uniform mutational probability between nucleotides caused by the mutational process. Measuring this effect is nontrivial due to the concurrent selection process that is applied at each experimental round. Therefore, we focused on synonymous mutations, which may affect the fitness but much less than amino acid changes, and can therefore isolate biases in the mutational process from the effects of non-synonymous mutations. In particular, to compute the bias we focused on amino acids in the wt sequences of PSE-1 with a fourfold degenerate third position, i.e. NNX where NN are two specific nucleotides, and X represents any of the four nucleotides. We can compute the relative abundance of these codons compared to that in the wt to estimate the mutational bias.

As an illustrative example let's consider the four amino acids Threonine, Glycine, Valine, and Alanine, which are coded respectively by ACX, GGX, GTX, and GCX. Let's further focus on the positions in the wt-sequence PSE-1 where the wt-codon ends with A. For each of those codons, we computed the relative frequency in the alignment of the other three synonymous codons (ending with C, G, and T). In the absence of any bias, we expect them to be equally represented. Instead, we can see from figure 2.1 that there is a clear bias: the mutation A $\to$ C is disfavored, compared to A $\to$ G and A $\to$ T. This phenomenon is consistent across positions and amino acids: it is a mutational bias, and not a codon bias, or another phenomenon dictated by selection.

Using this set of four amino acids we computed the relative frequency of all possible transitions $w(n \to n')$ which we report in the left side of table of 2.1. Interestingly, when we checked the specifics of the mutational biases of the epPCR protocol employed in the experiments (see https://www.agilent.com/cs/library/usermanuals/public/200550.pdf), we found very similar transition probabilities (right side of table 2.1).

2.3.2 Re-weighted Gibbs sampling

Once established the kind and prevalence of mutational bias in the experiments, we implemented a simple modification in our Gibbs

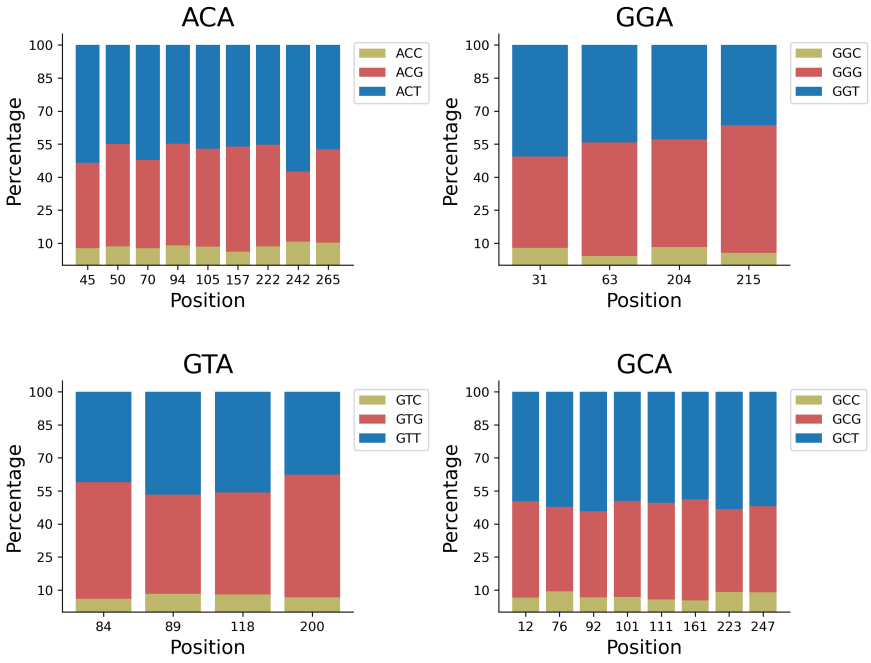


Figure 2.1 – Percentage of synonymous codons in the final library of Stiffler et al. experiments of 4 wt-codons: ACA, GGA, GTA, and GCA.

sampling strategy to take them into account. In particular, at every Monte Carlo step, the algorithm works as follows:

1. A nucleotide position i_k is randomly selected, with $i \in \{1, \ldots, L\}$ being the codon index and $k \in \{1, 2, 3\}$ being the nucleotide position inside the codon.

2. The set of 4 codons $\mathcal{C}_{i_k}$ is created. Each codon is derived by inserting into position i_k one of the 4 DNA nucleotides, i.e. $\mathcal{N} = \{\text{A, C, G, T}\}$. The codons are then translated to produce a corresponding set of amino acids, represented as $\mathcal{A}_{i_k}$.

3. For each $b(n') \in \mathcal{A}_{i_k}$, the corresponding $P(b|\mathbf{a}_{-i})$ are computed. These conditional Potts probabilities are identical to those described in the methods of section 2.2, except for the normalization, which is now computed over $\mathcal{A}_{i_k}$.

4. Each $P(b|\mathbf{a}_{-i})$ is re-weighted with a factor which depends on the nucleotide n^{o} already present in position i_k and the substituting nucleotide n' introduced in its place. In particular, the experimentally derived mutational biases $w(n^{\text{o}} \to n')$ from table 2.1 are used.

5. A new nucleotide n' is emitted from the following probability distribution:

$$P(n_{i_k} = n'|\mathbf{n}_{-i_k}) = \frac{\tilde{w}(n^{\text{o}}_{i_k} \to n')P(b(n')|\mathbf{a}_{-i})}{\sum_{n \in \mathcal{N}} \tilde{w}(n^{\text{o}}_{i_k} \to n)P(b(n)|\mathbf{a}_{-i})} \tag{2.1}$$

	A	C	G	T
A		8 ± 2	45 ± 6	48 ± 5
C	34 ± 13		7 ± 3	58 ± 13
G	56 ± 12	10 ± 3		34 ± 11
T	48 ± 9	44 ± 10	9 ± 2	

	A	C	G	T
A		9	35	56
C	32		10	58
G	58	10		32
T	56	35	9	

Table 2.1 – Left table: experimentally measured transition probabilities obtained from nucleotide sequences of round 20 of Stiffler et al, experiments with PSE-1 beta-lactamases. Right table: nominal mutation spectra induced by the epPCR protocol as reported by the manufacturer. Values are expressed as percentages, normalization of the rows might not be exact due to rounding errors.

where $\tilde{w}(n \to n')$ is a redefinition of the mutational bias transition matrix described in the following. Since Gibbs sampling contemplates re-emitting the same symbol present before the Monte Carlo step, we introduce an ad hoc value for $w(n \to n)$, namely λ. This allows us to define $\tilde{w}(n \to n')$ in the following way:

$$\tilde{w}(n \to n') = \lambda \delta_{nn'} + (1 - \lambda) w(n \to n') \tag{2.2}$$

where $\delta_{nn'}$ is the Kronecker delta, which is equal to 1 if $n = n'$ and 0 otherwise. $\lambda \in [0, 1]$ is a parameter that models explicitly a mutation rate. In particular, when $\lambda = 1$ no mutation can occur, while $\lambda < 1$, controls the average proportion of accepted mutations at each Monte Carlo step.

The first results of our new sampling strategy are encouraging. We run this dynamics starting from the wt PSE-1 nucleotide sequence with and without the mutational bias. As always we generate a number of Monte Carlo chains equal to those of the experiment and we tune the number of Monte Carlo steps to reach the same average Hamming distance as the wildtype. For the first simulations, we chose a value of $\lambda = 0.25$. To evaluate the results we compared the 1-point statistics of our simulations with that of the experiments, with and without bias. Results can be seen in figure 2.2. As expected, the new sampling strategy provides a small boost in the statistical accuracy of the sampling algorithm. This is likely due to the different accessibility of non-synonymous amino acids [111] compared to our previous model, which is more consistent with the experiments. Overall, our algorithm is able to integrate detailed nucleotide-level mutation information into a sequence model, achieving a balance between a faithful representation of the experimental mutational process and the simplicity of the implementation. Future work may explore strategies for setting λ based on the experimental data and the influence of the mutation rate on the simulations.

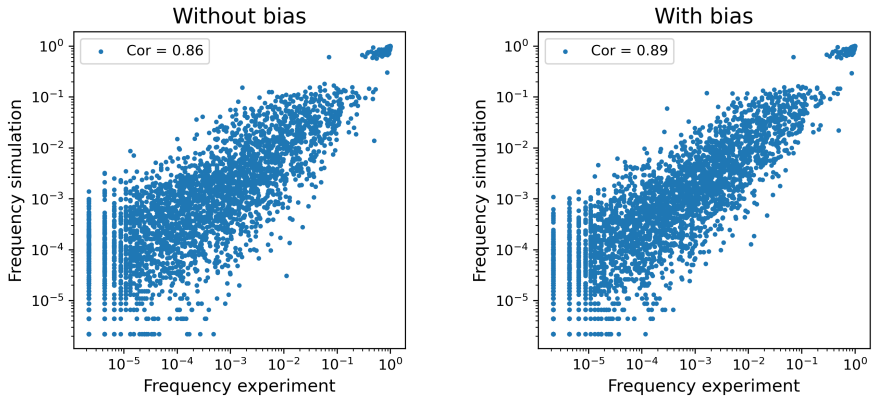


Figure 2.2 – Comparison of the 1-point amino acid statistics between experimental and in silico libraries. The left panel presents the in silico data generated without including the mutational bias. The right panel showcases data generated using the mutational bias, exhibiting a higher correlation.

2.4 INCLUSION OF DETAILED BALANCE

Sampling from high-dimensional probability distributions usually requires advanced algorithms, such as Markov Chain Monte Carlo (MCMC) sampling [112]. For spin models defined over categorical variables, like Potts models, many sampling algorithms are available. For example, Metropolis-Hastings and Gibbs Monte Carlo are both utilized as a sampling technique in the implementation [91] of Boltzmann learning that our group currently exploits to infer protein sequence landscapes. In this context, MCMC sampling generates equilibrium sequences via long simulations, ensuring the proper exploration of sequence space. However, MCMC sampling can also be used to model the dynamics of biological systems [113, 114], by interpreting the meaning of configurations drawn along the sampling. As we have shown in the article presented in section 2.2, Markov chain sampling can also be used to model short-term protein evolution in neutral genetic drift experiments. In this setting, short simulations performed with Gibbs sampling, with the addition of the constraints of the genetic code, accurately reproduced many statistical features of the experimental data. In the next sections, we discuss how to extend this approach to sample long chains that reach equilibrium, thereby modeling protein evolution over billions of years, while also accurately incorporating the amino acid accessibility dictated by the genetic code. All the results have been obtained in collaboration with Leonardo di Bari, a master's student in our group.

2.4.1 Model definition over nucleotide space

Let's consider a Potts model

$$P(\mathbf{a}) = \frac{e^{-\beta H(\mathbf{a})}}{Z} \tag{2.3}$$

whose parameters of the Hamiltonian H have been inferred from the natural sequences of a protein family. We aim to define an evolutionary dynamics using this model that takes into account the fact that evolution happens in nucleotide space. This necessitates specifying the probability distribution of $P(\mathbf{a})$ over nucleotide sequences instead of amino acid sequences. A straightforward procedure to accomplish this consists of uniformly distributing the probability associated with an amino acid sequence across all its synonymous nucleotide sequences. To formalize this concept, we define an amino acid sequence of length L as $\mathbf{a} = (a_1, a_2, ..., a_L)$ and a corresponding nucleotide sequence as $\mathbf{n} = (n_{1_1}, n_{1_2}, n_{1_3}, n_{2_1}, ..., n_{L_3})$. The function $A(\mathbf{n})$ applies the genetic code to transform the nucleotide sequence $\mathbf{n}$ into its respective amino acid sequence $\mathbf{a}$. For each amino acid sequence $\mathbf{a}$, we also define the set

$$\mathbf{\Gamma}(\mathbf{a}) = \{\mathbf{n} : A(\mathbf{n}) = \mathbf{a}\}, \tag{2.4}$$

and the function $N(\mathbf{a}) = |\mathbf{\Gamma}(\mathbf{a})|$, i.e. the cardinality of $\mathbf{\Gamma}(\mathbf{a})$. $N(\mathbf{a})$ counts the number of synonymous nucleotide sequences coding for a given amino acid sequence $\mathbf{a}$. $N(\mathbf{a})$ factorizes over the amino acids and can be rewritten as:

$$N(\mathbf{a}) = \prod_{i=1}^{L} N(a_i) \tag{2.5}$$

where $N(a)$ is the number of synonymous codons coding for amino acid a. We now have all the elements to define the probability distribution of a new model $\mathcal{P}(\mathbf{n})$ over nucleotide space:

$$\mathcal{P}(\mathbf{n}) = \frac{P(A(\mathbf{n}))}{N(A(\mathbf{n}))} = \frac{1}{N(A(\mathbf{n}))} \frac{e^{-\beta H(A(\mathbf{n}))}}{Z} = \frac{e^{-\beta \mathcal{H}(\mathbf{n})}}{Z} \tag{2.6}$$

such that the new Hamiltonian is defined as:

$$\begin{aligned} \mathcal{H}(\mathbf{n}) &= H(A(\mathbf{n})) + T \log\left[N(A(\mathbf{n}))\right] \\ &= H(\mathbf{a}) + T \sum_{i=1}^{L} \log\left[N(a_i)\right]. \end{aligned} \tag{2.7}$$

The Hamiltonian in Eq. (2.7) incorporates a novel term compared to the standard Potts Hamiltonian defined over amino acid space. Namely,

$$T \sum_{i=1}^{L} \log\left[N(a_i)\right] \tag{2.8}$$

which accounts for the entropic contribution arising from the degeneracy of synonymous DNA sequences. By assigning non-uniform weights to the codons, we can also account for the specific codon usages of different species. The term in Eq. (2.8) introduces a correction that disfavors amino acid sequences with high genetic redundancy by assigning them higher energy, i.e. lower probability, compared to sequences with low degeneracy which are assigned lower energy, i.e. higher probability. Consequently, the original Potts probability of amino acid $\mathbf{a}$ defined in Eq. (2.3) is correctly recovered after summing over all synonymous nucleotide sequences:

$$P(\mathbf{a}) = \sum_{\mathbf{n} \in \mathbf{\Gamma}(\mathbf{a})} \frac{e^{-\beta H(A(\mathbf{n}))}}{Z}. \tag{2.9}$$

2.4.2 Description of the algorithm

Detailed balance (DB) is a mathematical condition that ensures that the configurations generated by a Markov process converge to a stationary distribution. In our case, DB ensures that the Markov chains converge to the statistics of natural amino acid sequences that we used to infer the Potts model. Unfortunately, the sampling dynamics presented in the article of section 2.2 does not respect the DB condition. Adapting the Gibbs sampling algorithm described in the paper to satisfy DB is the goal of this section. DB is a statement about the transition probabilities π between configurations of our system. In the case of nucleotide sequences, for every pair of sequences $\mathbf{n}, \mathbf{n}'$, this expression must hold:

$$\pi(\mathbf{n} \to \mathbf{n}')\mathcal{P}(\mathbf{n}) = \pi(\mathbf{n}' \to \mathbf{n})\mathcal{P}(\mathbf{n}'). \tag{2.10}$$

The condition requires that, in equilibrium, the probability flux $\mathbf{n} \to \mathbf{n}'$ equals the one from $\mathbf{n}' \to \mathbf{n}$, i.e. the probabilities of $\mathbf{n}$ and $\mathbf{n}'$ do not change. A vast class of Markov Chain Monte Carlo algorithms depends on finding a suitable transition matrix that satisfies this equation. In addition to that, we want π to reflect plausible evolutionary dynamics on nucleotides. To this aim, we need to account for at least three processes:

- single-nucleotide mutations: a nucleotide is replaced by another one
- amino acid deletion: a codon is lost from the sequence, removing 3 nucleotides
- amino acid insertion: a codon is inserted in the sequence, adding 3 nucleotides

We model only triplets of nucleotide insertions or deletions because inserting or deleting single bases results in a frameshift, misaligning subsequent codons in the reading frame and leading to a non-functional sequence with high probability. On top of this, in terms of the amino acid sequence, this move would be highly non-local, thus hard to model. As a consequence, we assume that this process has zero probability and we exclude it from the dynamics. We observe that among the types of mutational processes described above, one operates at the individual nucleotide level, while the other operates on triplets of nucleotides, or codons. While this approach represents a simplification of the actual indel statistics, we contend that it maintains a foundation in biological understanding. Replication slippage [115], one of the most well-understood mechanisms generating indels in protein evolution, exemplifies this biological basis. During DNA replication, the DNA polymerase sometimes pauses and disconnects from the DNA. This pause can allow the new, growing DNA strand to momentarily detach and then accidentally reconnect to a similar sequence either ahead or behind the original spot. When the DNA polymerase restarts copying, it may either miss a section or repeat a section of the DNA, creating a deletion or an insertion in the new DNA strand. Including both indel mutations and single-nucleotide substitutions in a single MCMC framework is not trivial, since they act differently on variables and satisfaction of DB cannot be guaranteed. To overcome this issue, we suggest an approach that combines Gibbs and Metropolis sampling. We devised a mixed sampler operating as follows:

- With probability p, execute a Metropolis move simulating codon indels, i.e. operating on triplets of nucleotides.
- With probability $1 - p$, perform a Gibbs move simulating a single-nucleotide mutation.

Note that this reflects also the different nature of the two processes: our dynamics determines first which process happens, and then its outcome.

Indels: Metropolis sampling

To model codon insertions and deletions (indels), we introduce a new symbol: the triple gap "---" which we denote with c^0. This codon, equivalent to an amino acid gap, models the deletion of a codon within a nucleotide sequence. We also designate with c^k, $k \in \{1, ..., 63\}$, all other amino-acid-coding codons, but the stop codon.

Moving to the Metropolis sampling algorithm for indels, we decompose the transitions matrix $\pi(\mathbf{n} \to \mathbf{n}')$ into two components: a proposal term $p(\mathbf{n} \to \mathbf{n}')$ and an acceptance term $\alpha(\mathbf{n} \to \mathbf{n}')$. Since we exclusively focus on single codon substitutions, we only need to

consider transitions $\mathbf{n} \to \mathbf{n}'$ such that $\mathbf{n} = (c_1, c_2, \ldots, c_i, \ldots, c_L)$ and $\mathbf{n}' = (c_1, c_2, \ldots, c_i', \ldots, c_L)$. As a consequence, we will only consider the matrices $p(c \to c')$ and $\alpha(c \to c')$.

We restrict our proposal to mutations from amino-acid-coding codons towards the gap codon and vice versa, leading to the following proposal matrix $p(c \to c')$:

$$\begin{array}{c|cccccc} & c^0 & c^1 & c^2 & \cdots & c^{63} \\ \hline c^0 & \eta & \beta & \beta & \cdots & \beta \\ c^1 & \beta & \gamma & & \multirow{2}{*}{0} & \\ c^2 & \beta & & \gamma & & \\ \vdots & \vdots & \multirow{2}{*}{0} & & \ddots & \\ c^{63} & \beta & & & & \gamma \end{array} \tag{2.11}$$

This symmetric matrix accommodates insertions and deletions via the parameter β and prohibits amino acid substitutions:

$$p(c^m \to c^n) = 0 \iff m \neq n \quad \& \quad m, n > 0. \tag{2.12}$$

The parameters η and γ are used to normalize the proposal probability and correspond to "empty" moves not changing the state of the codon. To speed up the algorithm, we want to maximize β while maintaining the proposal probability normalized. As a result we get $\eta = 0$, $\beta = 1/63$ and $\gamma = 62/63$. This means that when we select an amino-acid-coding codon, in 62 out of 63 attempts we emit the same codon, and only in one case we emit the triple gap.

We are now in a position to describe how the Metropolis sampling algorithm for indels works. For each Monte Carlo step:

1. A sequence position $i \in \{1, \ldots, L\}$ is randomly selected, along with its corresponding c_i codon.

2. A new codon c_i' is proposed though the proposal matrix $p(c_i \to c_i')$

3. The acceptance probability of the proposed transition $c_i \to c_i'$ is computed according to the Metropolis prescription:

$$\alpha(c \to c') = min\left(1, \frac{\mathcal{P}(\mathbf{n}')}{\mathcal{P}(\mathbf{n})}\right) = min\left(1, e^{-\beta[\mathcal{H}(\mathbf{n}') - \mathcal{H}(\mathbf{n})]}\right) \tag{2.13}$$

4. Codon c_i' is accepted with probability $\alpha(c_i \to c_i')$ or refused with probability $1 - \alpha(c_i \to c_i')$.

Thanks to the fact that the proposal matrix in Eq. (2.11) is symmetric and that we accept codon mutations according to Eq. (2.13), detailed balance is guaranteed. Note that this part of the algorithm, modeling indels, was completely absent in the version of section 2.2, which simply did not sample them.

Point mutations: Gibbs sampling

The second type of mutation that we need to consider is single nucleotide substitutions. We resort to Gibbs Monte Carlo sampling generalizing the approach discussed in the paper, but introducing DB for accurate long-term sampling. Gibbs sampling works by iteratively emitting new variables from their conditional distribution, given the current value of the rest of the sequence. According to Eq. (2.6), we can express the conditional probability of n_{i_k} given the rest of the sequence as:

$$\begin{aligned}\mathcal{P}(n_{i_k}|\mathbf{n}_{-i_k}) &= \frac{\mathcal{P}(\mathbf{n})}{\mathcal{P}(\mathbf{n}_{-i_k})} = \frac{\mathcal{P}(\mathbf{n})}{\sum_{n_{i_k}\in\mathcal{N}}\mathcal{P}(\mathbf{n})} \\ &= \frac{e^{-\beta\mathcal{H}(n_{1_1}, n_{1_2}, ..., n_{i_k}, ..., n_{L_3})}}{\sum_{n\in\mathcal{N}} e^{-\beta\mathcal{H}(n_{1_1}, n_{1_2}, ..., n_{i_k}=n, ..., n_{L_3})}}\end{aligned} \tag{2.14}$$

where $\mathcal{N} = \{\text{A, C, G, T}\}$.

We can now present how a step of Gibbs Monte Carlo sampling works in our situation:

1. A nucleotide position i_k, $i \in \{1, \dots, L\}$ and $k \in \{1, 2, 3\}$, is randomly selected, excluding currently gapped positions.

2. The set of amino acids $\mathcal{A}_{i_k}$ is generated to explicitly compute the value of Eq. (2.14). We recall that $\mathcal{A}_{i_k}$ is the amino acid equivalent of the 4 codons derived by inserting each of the nucleotides of $\mathcal{N}$ in position i_k.

3. A new amino acid n', and the respective amino acid a', is sampled from the following probability distribution, computed exploiting the Hamiltonian of Eq. (2.7):

$$\begin{aligned}\mathcal{P}(n_{i_k} = n'|\mathbf{n}_{-i_k}) &= \frac{e^{-\beta H(a_1, ..., a_i=a', ..., a_L) - \log[N(a_1, ..., a_i=a', ..., a_L)]}}{\sum_{a\in\mathcal{A}_{i_k}} e^{-\beta H(a_1, ..., a_i=a, ..., a_L) - \log[N(a_1, ..., a_i=a, ..., a_L)]}} \\ &= \frac{e^{\beta h_i(a') + \beta\sum_j J_{ij}(a', a_j) - \log[N(a')]}}{\sum_{a\in\mathcal{A}_{i_k}} e^{\beta h_i(a) + \beta\sum_j J_{ij}(a, a_j) - \log[N(a)]}}\end{aligned} \tag{2.15}$$

4. n' is accepted unless it produces a stop codon. In either case, the procedure is restarted from point 1.

Eq. (2.15) is quick to evaluate, thanks to the many simplifications happening. In particular, the fields h and the couplings J between unmutated sites, as well as the degeneracies $N(\mathbf{n})$ of unmutated codons appear in both the numerator and the denominator, so they can be divided out from the expression, keeping only parameters containing the mutated site.

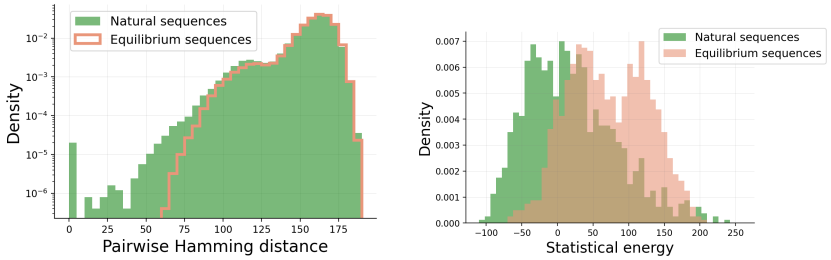
2.5 INTERMEDIATE TIME SCALES

We have developed a novel data-driven evolutionary dynamics for protein sequences, operating on nucleotide sequences and accounting for mutational effects at the level of amino acids. The dynamics takes place on an epistatic sequence landscape derived from an MSA of extant natural sequences. The underlying model we employed to fit the sequence landscape has a Potts form, and the parameters are derived via Boltzmann learning to reproduce the 1 and 2-point statistics of the natural MSA. However, any model capable of defining a probability distribution over sequence space can be adapted to our approach. We name our evolutionary model GENIE (Gene Evolution on Nucleotides Including Epistasis). GENIE can be considered an improvement of the SEEC (Sequence Evolution with Epistatic Contributions) framework presented by De La Paz et al. [116] adapted to nucleotide space. GENIE works by iteratively sampling new mutations from a given nucleotide sequence. Mutations are of two different kinds: Gibbs moves produce nucleotide substitutions, and Metropolis moves produce indels, with parameter p controlling the relative frequency between the two moves.

A significant breakthrough of GENIE is the modeling of nucleotide mutations while respecting DB. When considering extended trajectories, DB enables us to reach the statistics of natural sequences, a critical attribute when looking to model protein evolution spanning billions of years. However, we still face a challenge when attempting to incorporate mutational biases into our dynamics. The issue lies in the use of table 2.1 as a proposal matrix. Its non-symmetric structure does not guarantee DB, contrary to the proposal matrix used in Eq. (2.11). We plan to delve deeper into this issue in the future.

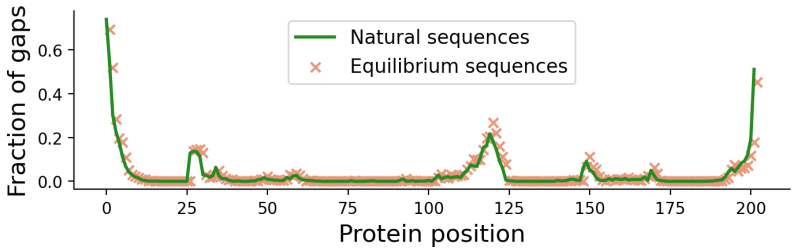
2.5.1 Equilibrium properties of long-term sampled sequences

GENIE is well suited to model the local statistics of genetic drift experiments since it is a simple modification of our previous approach. To confirm that it works on genetic drift experiments we re-run all simulations from section 2.2 and, as expected, we found identical results. More interestingly, GENIE is supposed to reproduce the statistics of natural sequences after long enough sampling. In particular, we want to verify that GENIE can reproduce the amino acid statistics of beta-lactamase sequences. To do so, we started from the wt gene of PSE-1 and we ran GENIE in parallel for 1000 sequences, for 6×10^7 Monte Carlo steps, and $p = 0.5$. The results are shown in Fig. 2.3. Fig. 2.3a shows that we recover the distribution of pairwise distances typical of natural beta-lactamases, with a peak around 160 mutations, or $160/202 \sim 80\%$ of sequence divergence. The left side of the distribution that is not reproduces is related to phylogeny, which



(a) *Pairwise Hamming distance of natural and equilibrium sequences.*

(b) *Energy of natural and equilibrium sequences.*



(c) *Fraction of gaps per position of natural and equilibrium sequences. The excess of gaps at the extremities of the alignment is due to the alignment procedure.*

Figure 2.3 – Statistical properties of 1000 sequences sampled with GENIE compared with natural sequences.

we cannot reproduce. Another important statistical test regards the energy of generated sequences. The lower the energy, the better they are according to the model. The energy of the sequences we generate at end of the sampling is close to that of natural sequences (Fig. 2.3b), although a bit higher. This is expected since the model is trained to assign very low energy to the training set, i.e. the natural sequences. We also confirm that we recover the gap statistics (Fig. 2.3c) of the training alignment, validating our Metropolis sampling of indels.

A more stringent test of the sampling capacity of GENIE regards the generative capabilities of the Potts model. We know that a good equilibrium sample obtained from the model has to respect the 1- and 2-point statistics of natural sequences, as computed in Eq. (1.1) and Eq. (1.2). Interestingly, we can track the quality of the statistics of our 1000 sequences over time, as they diverge from the PSE-1 sequence and plot it in Fig. 2.4. We see that after 10^7 MC steps the 1- and 2- point statistics are well recovered, meaning that, from the point of view of this metric, artificially sampled and natural sequences are hard to distinguish. At the end of the sampling, the correlation between f_i and c_{ij} computed from natural and sampled sequences reaches respectively the value of 0.98 and 0.87.

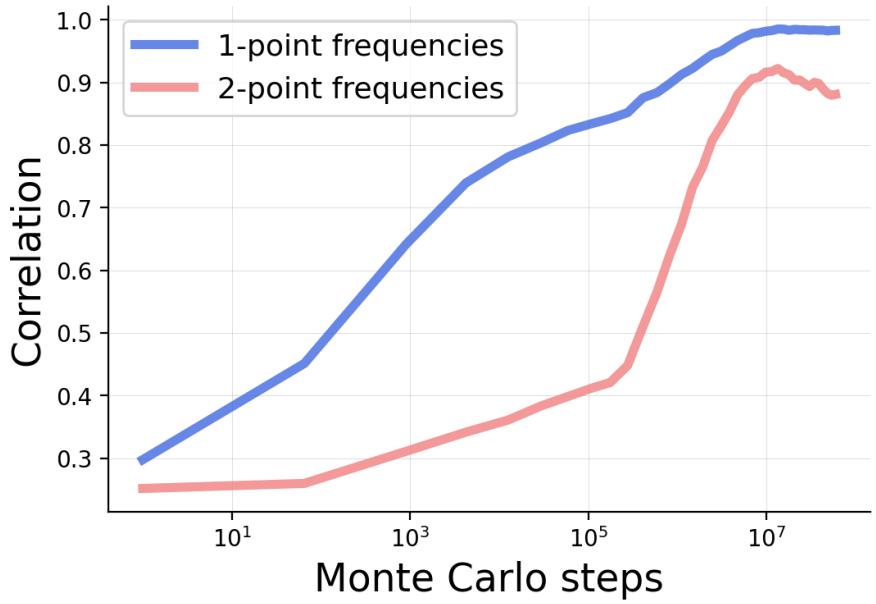


Figure 2.4 – Correlation of f_i and c_{ij} statistics between natural sequences and a set of 1000 sequences sampled with GENIE over time. The starting MSA is entirely composed of PSE-1 wt sequences.

2.5.2 Context-dependent and context-independent entropies

In Fig. 2.4 we have seen a glimpse of an interesting dynamics taking place while sequences diverge from a given point in sequence space and explore the surroundings. To quantify better this behavior we introduce two quantities, the Context-Independent Entropy (CIE) and the Context-Dependent Entropy (CDE). Those two measures will help us examine how a residue's variability is affected by its surrounding sequence context, and how this changes over time. For a residue at position i, we determine the CIE as:

$$CIE_i = -\sum_{a_i=1}^{20} f_i(a_i) \log_2 f_i(a_i) \tag{2.16}$$

Using the frequency of amino acids $f_i(a_i)$ computed from the columns of the natural alignment allows us to capture the global variability of protein sites over many different sequence contexts. CIE has a minimum of 0 in the case of complete conservation, i.e. no amino acid variability in the alignment column, and a maximum of $\log_2(20) \sim 4.3$ in the case of a completely variable column. We can also define a similar entropy, namely

$$S_i^t = -\sum_{a_i=1}^{20} f_i^t(a_i) \log_2 f_i^t(a_i) \tag{2.17}$$

where $f_i^t(a_i)$ is the amino acid frequency during the sampling with GENIE, at Monte Carlo step t. Both those entropies characterize

the site variability of a collection of sequences. When $t \to \infty$, or in practical terms when the sampling goes on for long enough, we have that $S_i^\infty = CIE_i$. We can also define a more specific context-dependent entropy, leveraging the DCA model. For every position i:

$$CDE_i(\mathbf{a}) = -\sum_{a_i=1}^{20} P(a_i|\mathbf{a}_{-i}) \log_2 P(a_i|\mathbf{a}_{-i}) \tag{2.18}$$

where

$$P(a_i|\mathbf{a}_{-i}) = \frac{\exp\left(h_i(a_i) + \sum_{j\neq i} J_{ij}(a_i, a_j)\right)}{\sum_{b=1}^{20} \exp\left(h_i(b) + \sum_{j\neq i} J_{ij}(b, a_j)\right)} \tag{2.19}$$

The CDE allows the quantification of the variability, or mutability, of a site in a *specific* context, thanks to the coupling J_{ij} which couples the amino acid a_i with the rest of the sequence. CDE has shown to be informative on the evolution and emergence of novel variants in the SARS-CoV-2 RBD spike domain [117], as well as to predict polymorphisms in *E.Coli* strains.

2.5.3 Emergence of time scales driven by epistasis

Epistasis, or the context dependence of mutations, emerges as mutations accumulate. This phenomenon is pivotal in shaping the evolutionary paths linking homologous proteins across various distances. To better quantify this process, we can use the difference between CDE and CIE that captures the impact of the sequence context on the mutability of a specific site, compared to the global natural variability. For example, a site can be globally variable, i.e. high CIE, but locally constrained, i.e. low CDE. Using PSE-1 as a reference we characterized each protein site based on its CIE and CDE variability, as illustrated in Fig. 2.5. We colored in red, green, and blue respectively three types of sites: *variable*, *conserved*, and *epistatic* sites. While the first two categories are intuitive, we define epistatic sites as those which are constrained in PSE-1 (low CDE), but very variable in the MSA of natural homologs.

Tracking the sites belonging to these three categories during the sampling of GENIE (see Fig. 2.6) reveals three types of emergent behavior:

- *Variable* sites (high CIE and CDE, red) mutate rapidly due to the absence of constraints imposed by the background.
- *Conserved* sites (low CIE and CDE, green) mutate slowly, reflecting their conservation both in the overall alignment and in the local context of PSE-1.
- *Epistatic* sites (high CIE but low CDE, blue) remain conserved when the context does not change too much, i.e. during the first

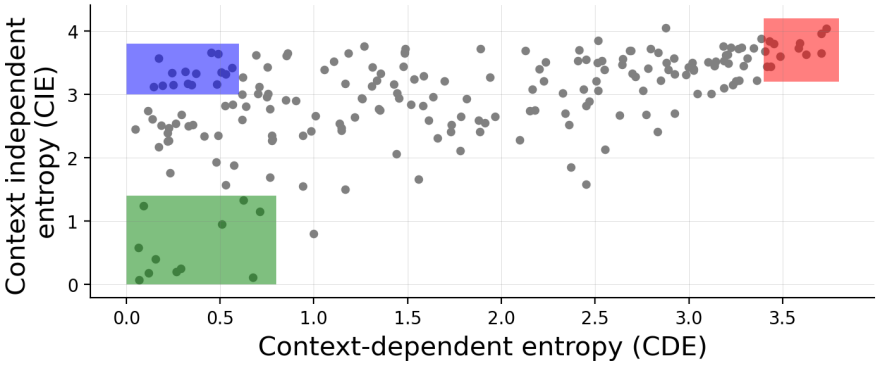


Figure 2.5 – Site classification of beta-lactamase PSE-1 residues by using the local variability (CDE) computed with the model and the global variability (CIE) computed using the natural protein family amino acid frequencies. Sites are colored according to three classes: conserved (green), mutable (red), and epistatic (blue) sites.

10^5 steps. However, they evolve rapidly when the background deviates from PSE-1, achieving at the end a high variability, as predicted by the CIE.

We recall that the final values S_i^t of the entropy of the sites along the GENIE simulations correspond to the CIE, indeed it is high for variable (red) and epistatic (blue) sites and low for conserved (green) sites.

Overall, our simulations reveal that the local context of a site is enough to predict its evolutionary dynamics over long time scales. In particular, epistatic sites start to mutate after 10^5 MC steps, or 50% sequence divergence, see Fig. 2.7 for reference. Ultimately, what we observe in Fig. 2.6 is the emergence of a time-scale separation in the mutability of amino acids during evolution:

- diversification happens very fast in a given context, mostly on sites with high CDE which can tolerate mutations,
- the context itself, on the contrary, evolves much slower, needing $30 - 50\%$ of sequence divergence to start accepting mutations, as exemplified by the entropy of the blue trajectories

In the future, we expect to define more quantitatively this separation over time scales and to compare it with experimental data from DMS in different protein families. Note also that experimental evolution never reaches the divergence needed to substantially alter the background, and mutations in the TEM-1 [72] and PSE-1 [110] experiments that we analyzed mostly happened in positions with high CDE. This may also be a reason why it was so difficult to detect contacts from experimentally evolved libraries.

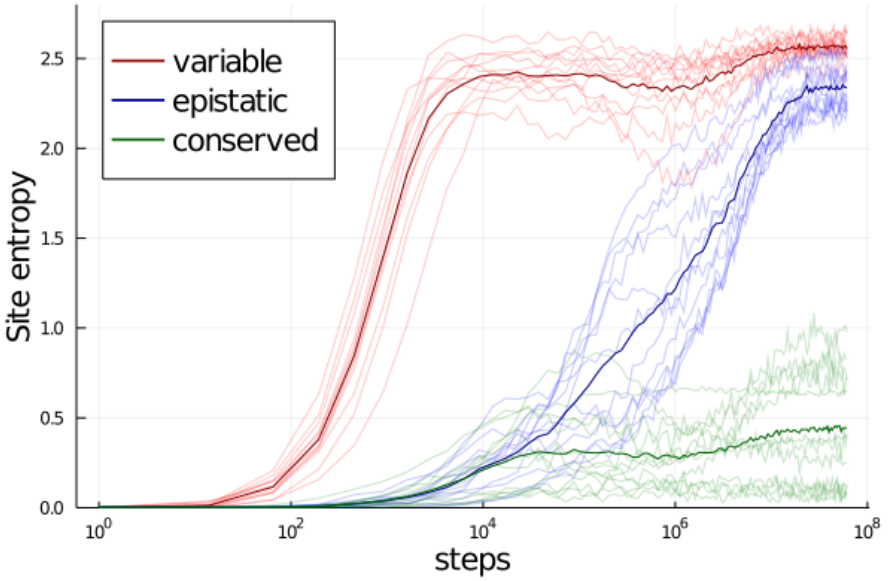


Figure 2.6 – Site entropy over 1000 GENIE chains for three different site categories, variable, epistatic, and conserved. Faded lines refer to single sites whereas thick lines represent the mean over the sites in each category.

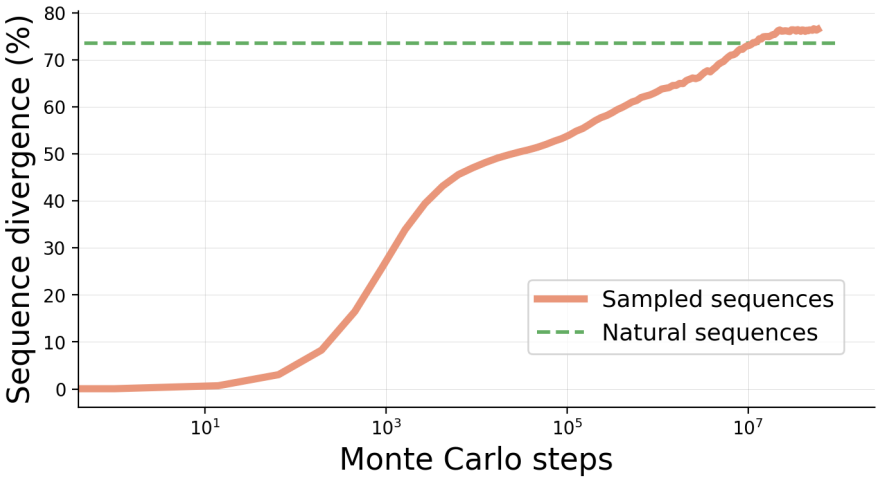


Figure 2.7 – Relation between the number of Monte Carlo steps and the average % sequence divergence from wildtype PSE-1 along 1000 evolutionary trajectories.

3 FAMILY-WIDE MUTATIONAL INCOMPATIBILIES

3.1 INTRODUCTION

The analysis of protein fitness landscapes is complicated by the presence of epistasis. Due to epistatic interaction between amino acids, identical mutations can cause very different effects on protein function when compared across different homologs. This phenomenon becomes manifest when comparing identical mutations across homologous sequences belonging to different species. Dobzhansky-Muller (DM) incompatibilities in protein evolution [118] refer to genetic changes that are neutral within one species but can be harmful in another species. These incompatibilities originate when different lineages accumulate genetic changes that, while adaptive or neutral within their respective lineages, become incompatible when brought together, typically causing degradation of function. The sites where DM incompatibilities take place are precisely the epistatic ones that we have described in the previous chapter. Those sites can be very conserved in the context of PSE-1, but very mutable in the context of different backgrounds.

One of the first studies that investigated this phenomenon in protein fitness landscapes [118] collected a set of human pathogenic missense mutations in 32 proteins to amino acid substitutions that naturally happened during evolution in other species. They found that around 10% of the mutations that are deleterious in humans, were naturally occurring in nonhuman proteins, i.e. they were compensated by interactions with other sites. Multiple studies since then, have measured the incompatibility of mutations across different sequence backgrounds on a much larger scale, using modern high-throughput biotechnological techniques like Deep Mutational Scans (DMS) to probe fitness landscapes. There are two important elements to consider when comparing identical mutations across different sequence backgrounds: the genetic distance between the backgrounds and the number of mutations compared, i.e. the coverage of the possible proximal mutations.

An interesting study from Lunzer et al. [55] tested all reversion mutations in the IMDH gene of *E. coli* towards the very distant homolog in *P. aeruginosa*. They found up to 30% of mutations with deleterious effects, although they were naturally occurring in *P. aeruginosa*. Genetic distance plays an important role in order to detect epistasis [119]. As we have seen in the previous chapter, epistasis reveals itself over

time, following the accumulation of many small effects. Studies using sequences that are too genetically close, although testing all possible single mutations in two homologs, could miss the chance of detecting epistasis [120], or finding very little effects [121]. The comprehensive saturation mutagenesis of three beta-lactamase sequences with up to 2 amino acid differences [122] confirms the high degree of correlation between mutational effects over short genetic distance, highlighting however isolated cases of strong epistatic effects.

The recent work of Park et al. [123] presented for the first time results including multiple complete DMS for homologs belonging to the same protein family, although sequence divergence was never higher than 40%. Other studies have managed to compare mutations across much more diverged sequences [124, 125, 126, 127], however never testing mutations across the full gene length. In the next section, we present the first complete comparison of all single mutants across two genes belonging to the B1 metallo-β-lactamase protein family, with more than 60% sequence divergence between each other. In the article, we show how combining these two DMS studies with the analysis at a family level coming from DCA allows us to better understand the epistatic network in B1 β-lactamases. DCA model is a good approximation of protein fitness landscapes in the local neighborhood of single sequences by predicting the effects of mutations [98, 86] over multiple protein families, predicting the emergence of novel SARS-CoV-2 variants [117] or capturing the variability of mutations within an entire specie [128]. In our previous work, we have shown how it was able to model the local landscape of two very distant genes, TEM-1 and PSE-1 [109]. In the next section, we extend this work by analyzing the predictive power and the limits of DCA to capture epistasis across a protein family.

3.2 ARTICLE

Understanding epistatic networks in B1 beta-lactamases through coevolutionary statistical modeling and deep mutational scanning

Chen, JZ[1], Bisardi M[23], Lee D[1], Cotogno S[23], Zamponi F[2], Weigt M[3] and Tokuriki, N[1].

[1]Michael Smith Laboratories, University of British Columbia
[2]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005, Paris, France
[3]Sorbonne Université, CNRS, Institut de Biologie Paris Seine, Biologie Computationnelle et Quantitative LCQB, F-75005, Paris, France

Abstract

Throughout evolution, proteins undergo substantial sequence divergence while preserving a consistent structure and function. Although most single mutations in a protein are deleterious, evolution is able to explore sequence space via interconnected networks of epistatic interactions that alleviate the harmful effect of mutations. To investigate this phenomenon, we studied the B1 class of beta-lactamase enzymes by combining the predictive power of a computational approach known as Direct Coupling Analysis (DCA) with the experimental insights from two Deep Mutational Scanning (DMS). DCA unveiled significant heterogeneity in mutational tolerance across 100 homologs belonging to the family, which was confirmed by the DMS of two clinically relevant enzymes: VIM-2 and NDM-1. The experiments show that more than half of the alignable residues from the two enzymes display significantly different mutability, indicating a widespread presence of epistasis. Moreover, around 30% of the reversion mutations are reciprocally deleterious, with no obvious compensatory mutations available among those tested. By comparing the DCA-predicted mutability with the accessible surface area in the respective amino acid positions, we found a predominance of epistatic residues among those with intermediate exposure. Analyzing the family both with DCA and through DMS experiments highlights the collective character of epistasis. This suggests a model of protein evolution where mutational effects change due to the slow buildup of small interactions between residues. Overall, our study demonstrates that the combination of DCA and DMS enables a thorough exploration and understanding of the intramolecular networks and coevolutionary signatures in B1 beta-lactamases, offering a panoramic view of epistatic networks across a protein family.

Introduction

Proteins are a fundamental component of life, and the comprehension of their sequence-structure-function relationship is essential to the understanding and application of a variety of fields, including biology, biochemistry, and protein engineering. Over the course of billions of years of evolution, many proteins have emerged to perform a specific function. Subsequently, these proteins further diversified their sequences, forming protein families through the accumulation of mutations, *i.e.*, homologous protein sequences spread among diverse organisms. Consequently, homologs often share as little as <25% amino acid sequence identity[1,2], suggesting evolution has navigated a vast sequence space while maintaining the main protein function. However, the sequence space is often riddled with 'fitness valleys'[3]. Many recent comprehensive mutant characterization studies, such as deep mutational scanning (DMS), showed that most mutations (~35-70%) in a single protein are deleterious[4–12]. While variants with these deleterious mutations would have been purged out from the evolutionary process by natural selection, many mutations that are deleterious in a given protein can be observed in other orthologous protein sequences, *i.e.*, in the context of many other mutations[13]. These observations suggest that evolution finds mutational paths that go around fitness valleys by exploring networks of intramolecular interactions (*i.e.*, epistatic networks) that can compensate and moderate the deleterious effects of some mutations. In other words, the existence of complex intramolecular networks and coevolution between residues can lead to differential mutational behavior in homologous proteins.

The evolution of natural proteins thus prompts a number of important and unanswered questions. Why can mutations that destroy function in one sequence be fixed in the sequence of homologs? How often are such heterogeneities in mutational effect encountered in a protein family? Can these patterns be modeled and predicted? Knowledge of such epistatic networks within a protein sequence is key to the understanding of the sequence-structure-function relationships, especially in predicting protein sequences that encompass functional proteins. Thanks to advances in both genomics (sequencing) and biochemical (DMS) approaches, we can obtain information on the epistatic network from both: as the coevolutionary trend in sequences within an orthologous family, or as the experimentally measured mutational behavior in each protein. Epistatic networks can be observed as coevolution of amino acid residues within a protein family. Statistical models trained on multiple sequence alignments are able to capture those patterns of coevolution and even

reproduce them by generating artificial sequences that respect the statistics of protein families[14–20]. While such data-driven computational approaches have shown to be powerful enough to design functional protein sequences that do not exist in nature[21,22], it is still unclear to what extent and accuracy such models can capture protein epistatic networks in terms of specific interactions. Experimentally, the influence of epistatic networks can be detected as mutational incompatibilities, *i.e.*, mutations that can be tolerated in one genetic background but may be deleterious in another. Such experimental insight provides an in-depth understanding of specific interactions within proteins in terms of biochemical and biophysical mechanisms. However, the experimental study of functional constraints at a family-wide level by comparing residue-level mutational epistasis across homologous sequences remains scarce. This is because only a small number of protein families have been studied[13], including a few large-scale datasets acquired through DMS[23–28]. Furthermore, most analyses are restricted to higher-level, broader trends between homologs with limited examinations of detailed mechanistic bases for epistasis.

In this study, we combine these two complementary approaches to improve our understanding of epistatic networks. We apply both approaches to the same system, the class B1 metallo-beta-lactamases (MBL), which is a family of highly diversified antibiotic degrading enzymes with a long evolutionary history[29]. Using Direct Coupling Analysis (DCA)[16], we analyze the coevolutionary signatures of the entire B1 family, arriving at a global statistical description of epistatic tendencies based on sequence data. At the same time, we perform DMS on two distantly related members of the B1 family, NDM-1 and VIM-2 (~30% sequence identity), revealing protein-specific mutational trends and incompatibilities as experimentally measured through functional characterization. By employing both methods in parallel, we have the opportunity to compare the results between them and to learn how they can strengthen and explain each other. We find a general consensus between the two methods on the prevalence and strength of epistasis and find interesting cases where complementary information provides insight beyond what can be discerned from either method individually. There appear to be trends of mutational behavior in the structure, but the exact intramolecular network appears much more complex upon further testing, as predicted by the DCA model. Overall, the combination of DCA and DMS allows us to greatly enhance our understanding of epistatic networks.

Results

Co-evolutionary model of the B1 family reveals sequence-specific mutational heterogeneities

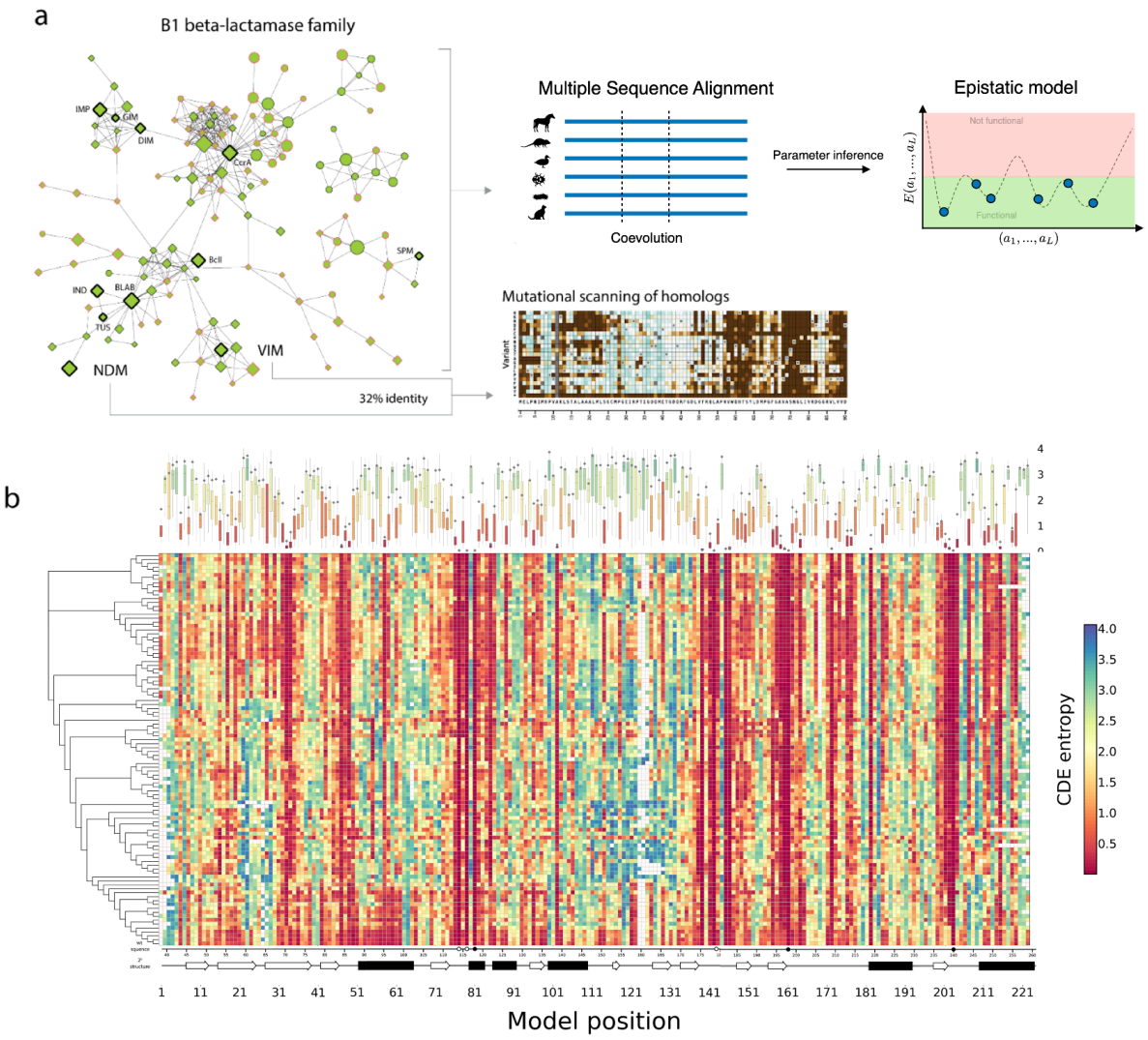


Figure 1. Family-wide residue mutability by direct coupling analysis (DCA). (a) Schematic of computational and experimental workflow. The B1 beta-lactamase enzyme family is isolated through sequence space exploration via a sequence similarity network. The entire set of sequences is used to generate a co-evolutionary model via DCA. Two highly diversified homologs within the family are selected for DMS to generate a large mutational dataset. **(b)** Each square in the heat map represents the context-dependent entropy (CDE) calculated for a given position in each of the 100 aligned homologs. Blue cells represent positions, which are highly mutable in the given context, while red cells signify context-dependent conservation. White cells represent alignment gaps. For each position, the percentiles of the distribution of CDEs are presented as a box plot on top of the heat map: 0-100% as thin lines, 25-75% as bars, and the median as a black dot. The bars are colored by the median value, with the same color scale as the heatmap. The secondary structure of the homologs, as well as the active site residues, are depicted under the heat map. The phylogenetic tree of the 100 homologs is shown on the left of the heat map. We observe that changes in CDE typically follow the evolutionary tree.

Class B1 metallo-beta-lactamases are a highly diversified family of enzymes (as low as 20% identity within the family) that degrade β-lactam antibiotics[1,30,31], a function for which there is likely a long evolutionary history[32,33]. Given the high sequence divergence, this family is an excellent model for epistatic networks, as different orthologs likely formed different sets of intramolecular interactions. To create a comprehensive dataset of all B1 sequences, we collected ~5000 sequences that belong to B1 MBLs, from the MBL domain superfamily (Interpro ID: IPR001279) along with metagenomic data from the Joint Genomics Institute (JGI) (see methods). Then, we inferred a global statistical model using DCA (see methods) on the resulting B1 MBL orthologs (~3500), which to this aim were aligned in a multiple-sequence alignment (MSA) based on a seed multiple-structure alignment of 14 diverged homologs (see methods); residues outside the structurally conserved region were excluded from the MSA and consequently from the computational analysis. The DCA model explicitly describes the coevolutionary relationships between all pairs of residues in the B1 family through direct couplings, and it models residue conservation via biases (or fields). Using these parameters, the model assigns a probability *P(sequence)* to any sequence, with high probabilities predicting viable sequences (functional MBLs), and low probabilities assigned to dysfunctional ones. The model can thus be used to score the effect of mutations in any specific sequence background. This is achieved by computing the difference in the model's statistical score, the "energy" E, as $\Delta E = -\log[P(\text{mutant})/P(\text{WT})]$, corresponding to the negative change in log-probability between the mutant sequence and the wild-type (WT). The inclusion of coevolutionary couplings into the DCA model makes this prediction directly dependent on all other WT residues in the sequence context[15]. Thanks to the global nature of DCA, it is possible to predict the effect of all possible single amino acid substitutions starting from any possible WT protein.

We calculated ΔE for all single mutants of 100 diverse homologs in the B1 family, as well as for NDM-1 and VIM-2, to probe the heterogeneous mutational behavior across the family. The homologs were chosen to minimize pairwise sequence identity. To gauge the effects of epistatic networks at the level of protein positions, we calculated the Shannon Entropy of all mutant probabilities relative to the wild type (proportional to $\exp(-\Delta E)$) at each position in each homolog, which we will refer to as the context-dependent entropy (CDE)[34,35]. The CDE can be interpreted as the (base-2 logarithm of the) effective number of tolerated mutations at a position in this specific WT sequence, such that 0 means only 1 (2^0) residue is tolerated (the WT, no mutation), and 4.3 means all 20 ($2^{4.3}$) amino acids are equally tolerated, with all other residues being kept fixed to the

specific WT context. The CDEs at each aligned position in the conserved B1 family are shown for 100 homologs as a heat map in **Fig. 1b**. The figure reveals a number of interesting aspects about the B1 family and its constituent homologs. We can see that there are regions that are highly constrained in terms of mutations, but also others that are highly tolerant. Moreover, phylogenetic proximity between sequences (represented by the phylogenetic tree) typically correlates with similarity in mutation patterns, yet distinct subtrees present different patterns. Active site residues are highly constrained across the entire family.

Computing the median and the spread of the CDEs at each position across the 100 homologs highlights substantial heterogeneity in mutational tolerance. There is a considerable spread of CDE values across homologs in most positions (bar plot in **Fig. 1b**), suggesting that positions that only allow for a few mutations in one homolog may show substantial mutability in other homologs. We quantified the spread in CDE at each position as the interquartile range (IQR), defined as the distance between the 25-75 percentile of the data. We see a median IQR spread of 0.87 ($2^{0.87}$=1.83), meaning that half of the positions have at least an almost 2-fold difference in the effective number of tolerated amino acids between homologs. The spread can be as high as 2.65, or a 6.3 fold difference in effective amino-acid number. Furthermore, the full range of CDEs across positions has a median spread of 2.67 and a max spread of 3.85, meaning half of the positions in the protein have a 6.3-14.4 fold difference in CDE between the most and the least mutationally tolerant sequences. This strong mutational heterogeneity of equivalent positions across homologs is a hallmark of epistasis and is represented collectively via the DCA couplings in our sequence model.

Heterogeneity in mutational behavior is supported by DMS data of NDM-1 and VIM-2

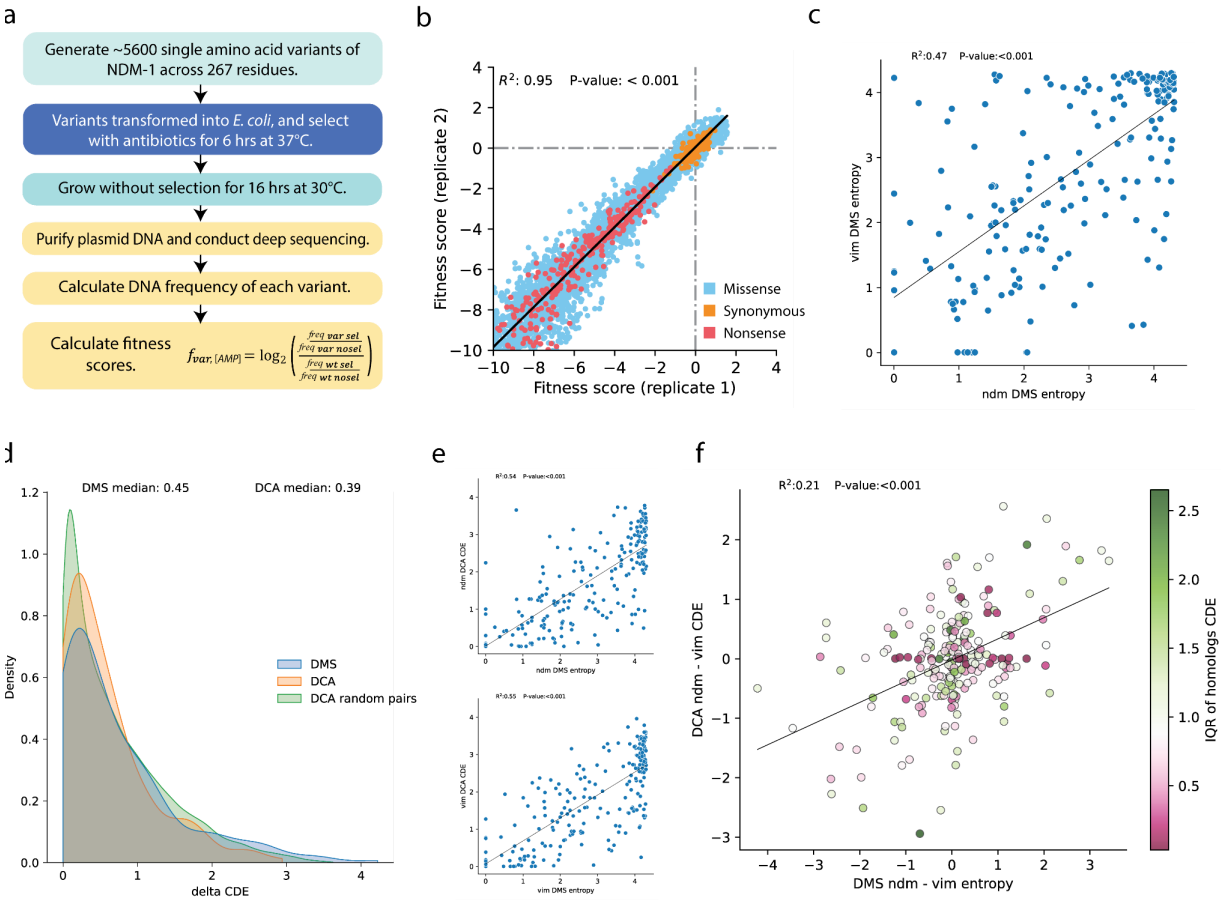


Figure 2. Overview of DMS for NDM-1 and structural similarity to VIM-2. (**a**) Workflow for DMS of NDM-1. (**b**) Correlation between replicates of the NDM-1 library selected at 256μg/mL AMP. The R^2 and P-value of a linear regression are shown at the top. **(c)** Comparisons of entropies calculated from all mutations' DMS fitness scores at each aligned position for the NDM-1 and VIM-2 experiments. The R^2 and line of best fit for a linear regression are shown. **(d)** Distribution of differences in entropy (DMS) or CDE (DCA) between NDM-1 and VIM-2 at the same aligned position. The distribution of the difference in CDE between the same aligned position in 100 random pairs of homologs CDEs is also plotted. **(e)** Comparison of entropies (DMS and CDE) for NDM-1 (top) and VIM-2 (bottom). The R^2 and line of best fit for a linear regression are shown. **(f)** Comparison of the entropy difference between NDM-1 and VIM-2 at each position to the difference in CDE. The data is colored by the spread in CDE IQR, with the color scale (scaled so that the median of the distribution is the center) shown to the right. The R^2 and line of best fit for a linear regression are shown.

To gain experimental observations for differential mutational effects between homologous enzymes, we conducted DMS to obtain all single-mutational effects on two homologs, NDM-1 and VIM-2 (~30% sequence identity). A complete DMS dataset of all single amino acid variants of VIM-2 was previously published[6,36]. To conduct a comparison, we performed DMS on NDM-1 in an identical manner to VIM-2 (**Fig. 2a**). Briefly, all single amino acid variants were generated for NDM-1 and

placed under selection under three different antibiotics: ampicillin, cefotaxime, and meropenem. The plasmid DNA was isolated after selection and sent for deep sequencing. The fitness conferred by each variant relative to wtNDM-1 was then characterized as the fitness score in **Eq. (1):**

$$f_{var} = \log_2\left(\frac{freq\ var\ sel}{freq\ var\ nosel}\right) - \log_2\left(\frac{freq\ wt\ sel}{freq\ wt\ nosel}\right) \qquad (1)$$

The experiments were conducted in duplicate on separate days, and replicates typically show good correlation (R^2 of 0.77-0.95 across conditions) (**Fig. 2b**, **Supp. Fig. 1**). As was the case for VIM-2, the NDM-1 data shows that fitness scores of -4 or lower are below the fitness of nonsense variants. As such, we also limit the fitness scores to a minimum of -4 for analyses, and scores below -4 are set to be -4 instead.

In a similar fashion as in the DCA analysis, we define the DMS entropy, a measure of the mutational tolerance of each position in NDM-1 and VIM-2, as the Shannon entropy of the mutational probability of each amino acid. For the probability, we use the normalized enrichment ratio $\frac{freq\ var\ sel}{freq\ var\ nosel}$ of the variant frequencies with and without selection; this is also proportional to the exponential of the DMS fitness (with basis 2). We find mutational behaviors in the same protein positions to be quite varied between the two homologs, where differences in certain positions can span nearly the full range of possible entropy values, *i.e.*, the same position can accept only the WT amino acid in one homolog (entropy 0) and almost all amino acids in the other homolog (entropy >4) (**Fig. 2c**). We compare the magnitude of entropy differences between VIM-2 and NDM-1 in the DMS data to differences in CDE computed from DCA, and we find strikingly similar distributions (**Fig. 2d**). The median difference between two homologs across all positions is 0.45 in DMS (0.39 in DCA), meaning that half of the positions have more than a 1.37 fold difference in entropies, ranging up to a maximum of 4.22 (18.7 fold). As a baseline for the expected behavior between any pair of homologs, we also computed the CDE difference between 100 random pairs of homologs. Wound a similar distribution (**Fig. 2d**), suggesting a common statistical trend of mutational behavior between homologs where certain positions exhibit similar tolerance (such as conserved sites) while others can greatly vary in mutation tolerance.

We further probe whether the DMS and DCA data are in agreement specifically on NDM-1 and VIM-2, by comparing DMS entropies and DCA CDEs of each position within either NDM-1 or VIM-2 (**Fig. 2e**). There is significant correlation (R^2~0.55) between the DMS and DCA values for each of NDM-1 and VIM-2, and this correlation is slightly higher than between the DMS entropies of

NDM-1 vs. VIM-2. This observation shows that DCA captures sequence-specific trends to a good degree. Despite the significant correlation, DCA predicts fewer variable sites (entropy > 3.5) compared to DMS. A possible explanation of this difference may be attributed to the fact that DMS experiments are limited to specific experimental conditions, i.e. a single antibiotic, while DCA likely reflects broader evolutionary constraints. Furthermore, the CDE differences between VIM-2 and NDM-1 computed from DCA can also capture the specific entropy differences in DMS (**Fig. 2f**). The combined analysis of mutational behaviors from DMS and DCA data also reveals some interesting observations. When the potential spread across homologs (IQR of 100 homolog CDE distribution) is overlaid on the mutational differences between NDM-1 and VIM-2 (**Fig. 2f**), we can see that positions with large differences between NDM-1 and VIM-2 also show high spread in all homologs. Furthermore, positions with low spread in all homologs tend to show lower differences in both DMS and DCA. Finally, positions that have lower differences between NDM-1 and VIM-2 in both methods can also have high spread across all homologs, further underscoring the fact that mutational behaviors can be quite varied for different homologs. Hence, the combination of methods can reveal and reinforce patterns that would not be obvious from just a single approach.

Structural basis of mutational tolerance and incompatibilities

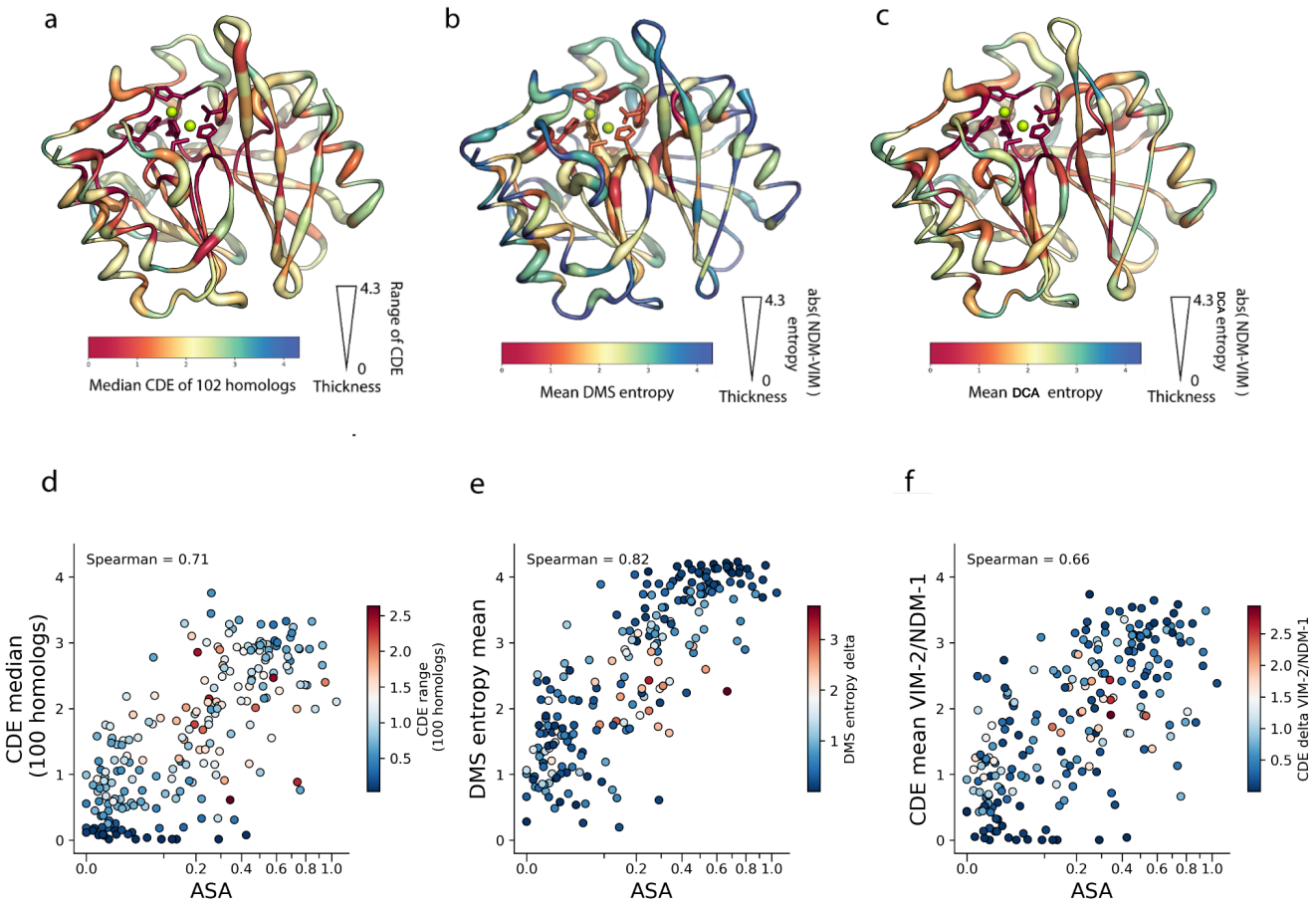


Figure 3. Structural basis of tolerance classifications. (a) CDE data overlaid on the crystal structure of VIM-2 (5yd7), with the thickness of the backbone representing the range in CDE of the 100 homologs, and colored by the median CDE of all homologs. **(b)** DMS entropy data overlaid on the crystal structure of VIM-2, with the thickness of the backbone representing the absolute entropy difference between NDM-1 and VIM-2, and colored by the average entropy. **(c)** CDE of VIM-2 and NDM-1 overlaid on the crystal structure of VIM-2, with the thickness of the backbone representing the absolute CDE difference, and colored by the average CDE. **(d)** Scatter plot of the CDE median values of 100 homologs versus the average ASA of VIM-2 and NDM-1, with positions colored by the CDE range. **(e)** Scatter plot of the average DMS entropy of NDM-1 and VIM-2 DCA CDE versus the average ASA, with the positions colored by the difference in the DMS entropy between NDM-1 and VIM-2. **(f)** Same as panel (e) but for the DCA predictions: scatter plot of the average CDE of VIM-2 and NDM-1 versus their average ASA, colored by the CDE differences.

In this section, we investigate in detail the relationship between mutational heterogeneity and structure, for the two selected wildtypes and at the family level. We proceed by analyzing the experimental datasets and the model predictions in terms of the protein structure of the B1 family, using the crystal structures of NDM-1 (PDB ID:3spu) and VIM-2 (PDB ID:5yd7) as representatives.

For VIM-2 and NDM-1, we quantify the discrepancy in mutability using the absolute difference in entropy, computed via the DMS experiments and the DCA model, while for the 100 homologs, the IQR of CDEs serves as a DCA-based indicator of variability. It should be noted that points exhibiting either very high or very low average mutability obviously have inherently restricted variability. This is evidenced by the observation that data points in **Fig.3** with an entropy lower than 1 or greater than 3 always have low spread (mainly blue).

We first visualize the DCA-derived CDEs of the 100 homologs superimposed on the structure in terms of the spread (IQR, the thickness of the backbone) and mutational tolerance (median of CDE, the color scale) (**Fig. 3a**). To compare this global predicted trend with the specific behavior of the two wild types we produce equivalent figures using the DMS entropies (**Fig. 3b**) and CDEs (**Fig. 3c**) for VIM-2 and NDM-1 only. We see similar tendencies in the three figures. Regions that are buried in the protein core, including the active-site metal-binding residues, tend to have both low mutational tolerance and spread, likely as a result of the regions being critical to folding or activity and hence mutationally constrained. The most exposed positions generally have high entropy and very low spread according both to the DMS and the DCA, even if we observe that there is a higher spread across the homologs according to DCA, which is a result of both NDM-1 and VIM-2 being completely mutationally tolerant at these positions. Finally, we also observe that some of the more buried residues have a higher entropy as computed from the DMS of NDM-1 and VIM-2 than in the DCA model. This observation is consistent with the fact that DCA typically tends to underestimate the number of neutral, or almost neutral, mutations. It can also possibly be due to specific differences in residues and spatial arrangements between the two homologs.

To better quantify the observations about the role of the structure we use the average accessible surface area[37] (ASA) of NDM-1 and VIM-2. The first pattern that emerges is a significant correlation between average ASA and the site-specific mutability, evident in both experimental data and model predictions (**Fig. 3d-f**). In particular, there is a pronounced correlation between average ASA and DMS-derived entropies (Spearman correlation = 0.82) (**Fig. 3e**). This observation has been reported before[4,8,38,39], and is possibly due to the higher prevalence of structural interactions among internally situated protein residues, thereby amplifying the potential for mutations with adverse effects and vice versa. The correlation is still large (Spearman 0.66-0.71) when we compare ASA to the DCA-derived mutability, both at the family level (**Fig. 3d**) and for the two specific homologs VIM-2 and NDM-1 (**Fig. 3f**). In this case, the capacity of DCA to identify these signals is mainly attributable to the single-column conservation patterns embedded within the MSA utilized to train the model.

Furthermore, we can use the ASA as a structural variable to distinguish three classes of residues: very buried (ASA<0.1), partially exposed (0.1<ASA<0.7), and very exposed (ASA>0.7). Let's analyze in detail how mutational heterogeneity is influenced by the three levels of residue burial. First of all, very buried residues tend to be more mutationally constrained, i.e. positions with very low ASA (ASA < 0.1) have low entropy and low spread (blue points in the bottom left of **Fig. 3d-f**). Second, the very exposed positions (ASA >0.7) typically display high mutational entropy and low spread, i.e. they are very mutable in all homologs. These first two observations are consistent with a classical picture of conservation due to structural constraints. Most importantly, we find an intermediate region of partially exposed residues (0.1<ASA<0.7) showing some residues with very high spread in mutability as expressed by CDE and DMS entropy (white and red points in **Fig. 3d-f**). This is an original finding of this work, which highlights the largest variability in mutational behaviors of moderately buried sites, possibly due to having more freedom to mutate than the fully buried positions while still forming interactions with other residues.

On top of this distinction, we observe that in the case of buried residues, some positions do exhibit a fairly large mutational variability between the two wildtypes VIM-2 and NDM-1, as highlighted by the presence of light blue-colored points in panels **d-e** of **Fig. 3**. Interestingly, this contrasts with the almost total absence of mutational heterogeneity observed between very exposed residues of VIM-2 and NDM-1 (almost all points are dark blue), especially in the DMS data.

A possible cause of this difference is the numerous intramolecular interactions occurring around buried residues, suggesting that a rich intramolecular network not only reduces the mutability of residues but also leads to homolog-specific differences of such constraints. However, it is the intermediately exposed region that exhibits the largest variability in behaviors for all datasets. This ASA range corresponds to positions that have more freedom to mutate than the fully buried positions while still being capable of forming interactions with other residues. The possibility of a mutation is therefore strongly dependent on the sequence context and is, therefore, homolog-specific. The large spread in the mutational patterns of the intermediate region (manifest from the red and light-red color of the points in the central regions of **Fig. 3d-f** is common to both the two and the one-hundred homolog analyses, confirming the dependence between epistatic networks and the structure.

Epistasis from individual variants

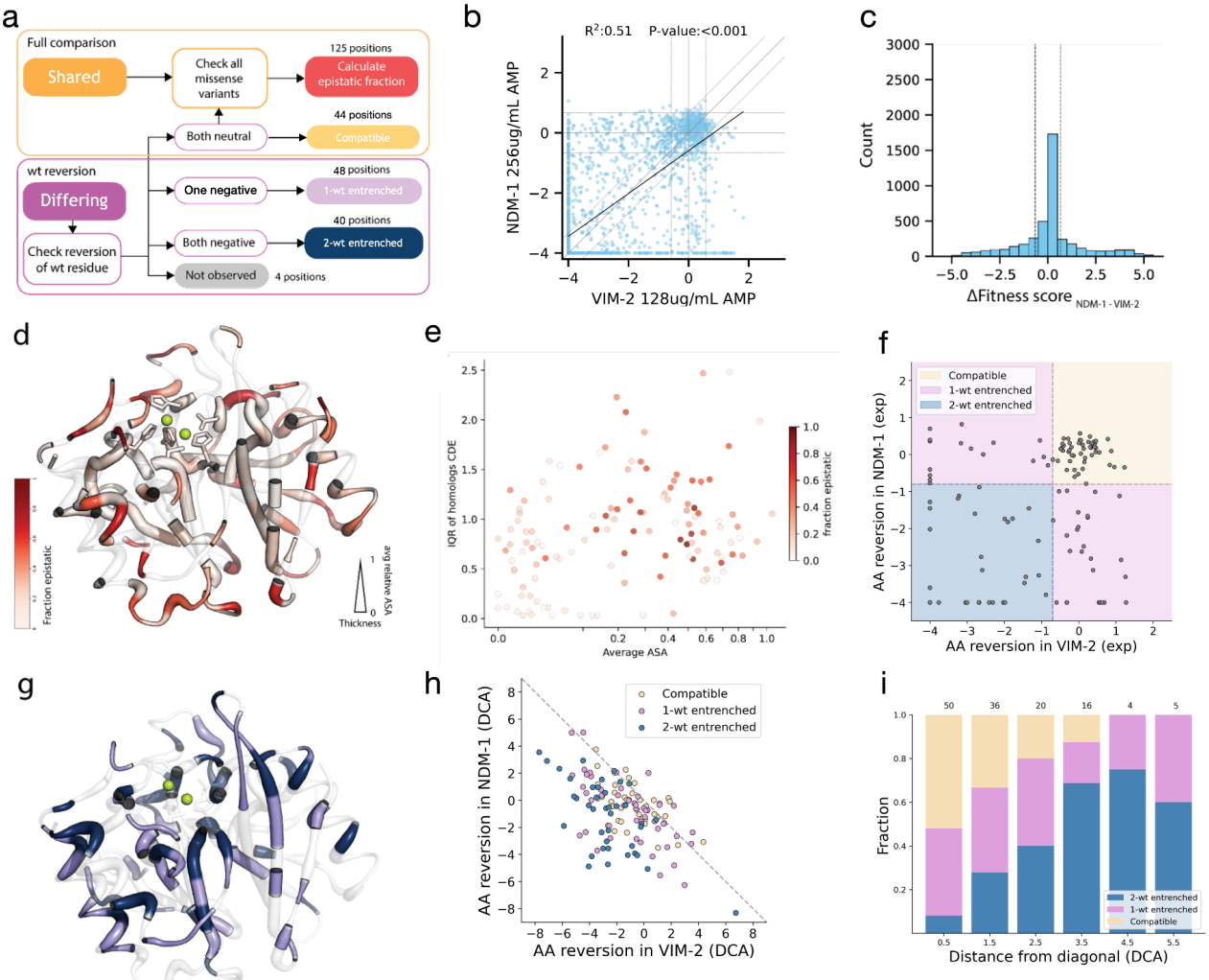


Fig. 4. Residue level epistasis between NDM-1 and VIM-2. (**a**) Flowchart of the epistasis classification method. (**b**) Correlation of DMS data between NDM-1 and VIM-2. The regions between dashed lines in each axis represent the range of neutral fitness effects for each homolog (1.96 x SD of synonymous variants). The diagonal line shares the same neutral range as the y-axis. (**c**) Distribution of fitness effect differences between NDM-1 and VIM-2. The region between vertical dashed lines represents the range of neutral fitness, equal to 1.96xSD of synonymous variants for NDM. (**d**) Positions in which mutations can be compared directly in VIM-2 and NDM-1 ('comparable'): fraction of epistatic mutations overlaid on the VIM-2 structure, and colored by the color scale to the lower left. The thickness of the structure corresponds to the average ASA of the NDM-1 and VIM-2 crystal structures. Regions outside the classification are made transparent. (**e**) Plot of the IQR in CDE of 100 homologs and the ASA, with each position colored by the fraction of mutations that are epistatic. Only positions highlighted in panel (d) are included. (**f**) Scatter plot of fitness effects for mutations of VIM-2 amino acids towards NDM-1 amino acids (x-axis) and NDM-1 amino acids towards VIM-2 amino acids (y-axis) for equivalent positions. The vertical dashed line indicates the left side of the region of neutral effects for VIM-2 based on the synonymous

variant distribution, and the horizontal dashed line shows the right side of the region of neutral effects for NDM-1. Quadrants with different behavioral classes are colored as in **Fig 4a**. (**g**) Positions that are diverged between NDM-1 and VIM-2 and have undergone epistasis analysis of WT reversion mutations. The structure thickness corresponds to the average ASA as in panel (d) and regions outside the classification are transparent. (**h**) Scatter plot of ΔE for mutations of VIM-2 amino acids towards NDM-1 amino acids (x-axis) and NDM-1 amino acids towards VIM-2 amino acids (y-axis) for equivalent positions. Points are colored according to their experimental classification as in **Fig 4c.** The dashed line represents the expected behavior without epistasis. (**i**) Bar plot showing the relative fraction of points in each epistatic class at various distances from the diagonal of **Fig. 4i.** The total number of points in each distance bin is written over the bar.

We now turn to the study of intramolecular interactions and use both experimental and model information to characterize the epistatic networks and reveal the interdependencies between residues. For this purpose, we compare the effects of individual variants in the mutational scans of VIM-2 and NDM-1 as fitness differences ($\delta f = f_{\text{NDM-1}} - f_{\text{VIM-2}}$). A large number of variants can be compared directly for positions in which VIM-2 and NDM-1 have the same WT amino acid, which we refer to as 'shared' positions. However, when the WT amino acids are not identical, that is at 'differing' positions, a direct comparison is not always possible. Therefore, we devise a systematic classification scheme to analyze mutations in 'shared' and 'differing' positions between NDM-1 and VIM-2 accordingly (**Fig. 4a**).

At 'differing' positions, the effect of the starting point of each mutation may impede a straightforward comparison and characterization of epistasis. To accommodate this, for each of these positions, we first examine the effect of swapping WT residues between NDM-1 and VIM-2 in both backgrounds. If the swap is neutral in both directions, we define the site as 'compatible' (44 positions) and proceed to the comparison of all the missense variants as if the starting points at that position were 'shared' (**Fig. 4a**). If the swap is incompatible, that is if one of the mutations is deleterious in either of the two backgrounds (88 positions) we only compare the reversion mutations and consider the position as 'entrenched' (**Fig. 4f, supp. Fig. 2**). We further classify the entrenched positions as '1-wt entrenched' if the swap is incompatible in one background (44 positions) and '2-wt entrenched' if the mutation is incompatible in both directions (40 positions). Because 4 positions lack reversion mutants in our DMS data, we exclude them from the analysis and label them as 'not observed'. The great number of entrenched positions points to a pervasive presence of epistasis: for each individual WT only ~50% of WT-swapping mutations are neutral, while the remaining mutations lead to a significant loss in fitness, even though they are occurring naturally in another homolog. Thus, a complex set of compensating epistatic effects must be considered to account for the collective presence of those mutations.

We then analyze 'shared' positions, that is positions with the same WT amino acid. Together with the 'compatible' positions described earlier, they add up to a total of 125 residues which we collectively call 'comparable'. The analysis of the DMS data for these positions allows us to compare equivalent mutations across the NDM-1 and VIM-2 backgrounds and, therefore, directly quantify context dependence. In this case, we can compare the fitness of all mutations A->B, where A is the common WT amino acid and B represents one of the 19 possible mutations (18 for 'compatible' residues). Mutational effects are strikingly different between homologs, with an overall Pearson correlation R^2 of 0.51 (**Fig. 4b**). For each mutation, we consider it epistatic if the fitness difference δf is greater than the range of neutral effects (standard deviation [SD] of the synonymous variants of NDM-1). Then, a large proportion of variants (~55%) shows a statistically significant difference in mutational effects between homologs; differences are not strongly biased in sign or effect size toward a single homolog (**Fig. 4c**). To evaluate the degree of epistasis, for each position we compute the fraction of epistatic mutations. We observe a widespread presence of epistasis (**Fig. 4d**) across the entire structure. Quantified as a distribution, we find the median fraction of epistatic mutations at each position is 0.21, meaning that half of the positions have significant epistasis in at least ~4 mutations (0.21 x 19 missense aa). The fraction of epistasis at each position can be used to compare site-specific epistasis of VIM-2 and NDM-1 with the spread in mutational tolerance across all 100 homologs and ASA (**Fig. 4e**). Strongly epistatic positions (dark red) are enriched in regions with a higher spread in CDE across homologs, and in particular in positions with intermediate ranges of ASA, as previously noticed (**Fig. 2f** and **Fig. 3d-f**). In fact, the fraction of epistasis is closely linked to the delta DMS entropy (**supp Fig. 3**). The subset of positions classified as entrenched is especially suited to study the influence of epistatic constraints in evolution. They cannot be easily linked to site-specific features or residue burial, as they are scattered across the whole protein structure (**Fig. 4g**). From a DCA point of view, the deleterious effect of the mutations is caused by the sum of many distributed negative interactions with the surrounding amino acids (**supp. Fig. 4**). This picture points to a model of protein evolution where the effect of mutations changes gradually due to the incremental accumulation of small-magnitude interactions between residues[40].

We can exploit a DCA-based analysis at the level of individual mutations to explore the role and meaning of entrenched positions in VIM-2 and NDM-1. We produce a scatterplot analogous to that of **Fig. 4f**, but using ΔE instead of fitness (**Fig. 4h**). Despite having a stretch along the diagonal, mutations mainly concentrate around the neutral center, or populate the lower half-plane defined by the diagonal. As expected, we see almost no points towards the upper half-plane, where the swap of both WTs would result in higher fitness. Ideally, in the absence of epistasis, the effect of the

mutations A->B and B->A in different backgrounds should have opposite signs, exhibiting a perfect anticorrelation of the reversion effects, indicated by the diagonal in **Fig. 4h**. Epistasis, however, causes a deviation from this simple picture. We can quantify the magnitude of epistasis between reverse mutations in different backgrounds as the deviation from the diagonal: the further away the points are from the diagonal, the bigger the difference of the mutational effect in the two sequence contexts. We can verify this idea by using the distance from the diagonal as an “entrenchment metric". We compute the fraction of mutations from each entrenchment class at different distances from the diagonal (**Fig. 4i**). We see that regions far from the diagonal are enriched by 2-wt entrenched mutations, those near the diagonal are mostly compatible, while intermediate regions have mutations with mixed classifications.

The biggest qualitative difference between the mutation reversion plots in **Fig. 4f** and **Fig. 4h** is that DCA predicts many WT-swapping mutations to be beneficial, in clear contrast with the DMS data. Since the experimental setting is not well suited to measure gain of function mutations, experimental points cannot populate the pink quadrants in **Fig. 4f** outside the regions of neutral effects. As a consequence, a potential confusion between epistatic and non-epistatic mutations arises, as illustrated in **supp. Fig. 5**. While 2-wt entrenched mutations are guaranteed by the experiment to be epistatic, this is not true for 1-wt entrenched mutations. In fact, if one of the mutations is negative, and the reversion is neutral, in principle we cannot tell whether the mutation was truly neutral, or if it was a a beneficial mutation that the experimental assay was incapable of measuring. In the absence of direct experimental data, there are however a few reasons that suggest that most of those mutations are truly neutral in one background and deleterious in the other.

First, the interpretation is supported by the distribution of effects observed in our previous EC50 analysis of VIM-2 variants (**supp. Fig. 6**). The rarity of gain-of-function mutations (constituting only 1-2% of all occurrences) makes it improbable that there would be sufficient data points to create the anti-correlation expected in the absence of epistasis. Moreover, DCA supports this interpretation as well: as we have shown in **Fig. 4i** many 1-wt entrenched positions are statistically different from compatible ones, to the extent that some reach far away from the diagonal, just like 2-wt entrenched mutations.

We argue that, once again, the model and the experiment complement each other: the model suggests that epistatic interactions identified in the experiment are sparse and pervasive and supports the interpretation of 1-wt entrenched mutations as being mainly neutral in one of the two directions. Moreover, DCA proves to be quite accurate in predicting epistasis: all of the most distant points from the diagonal are either 1- or 2-wt entrenched according to the experiments.

Experimentally probing specific epistatic interactions

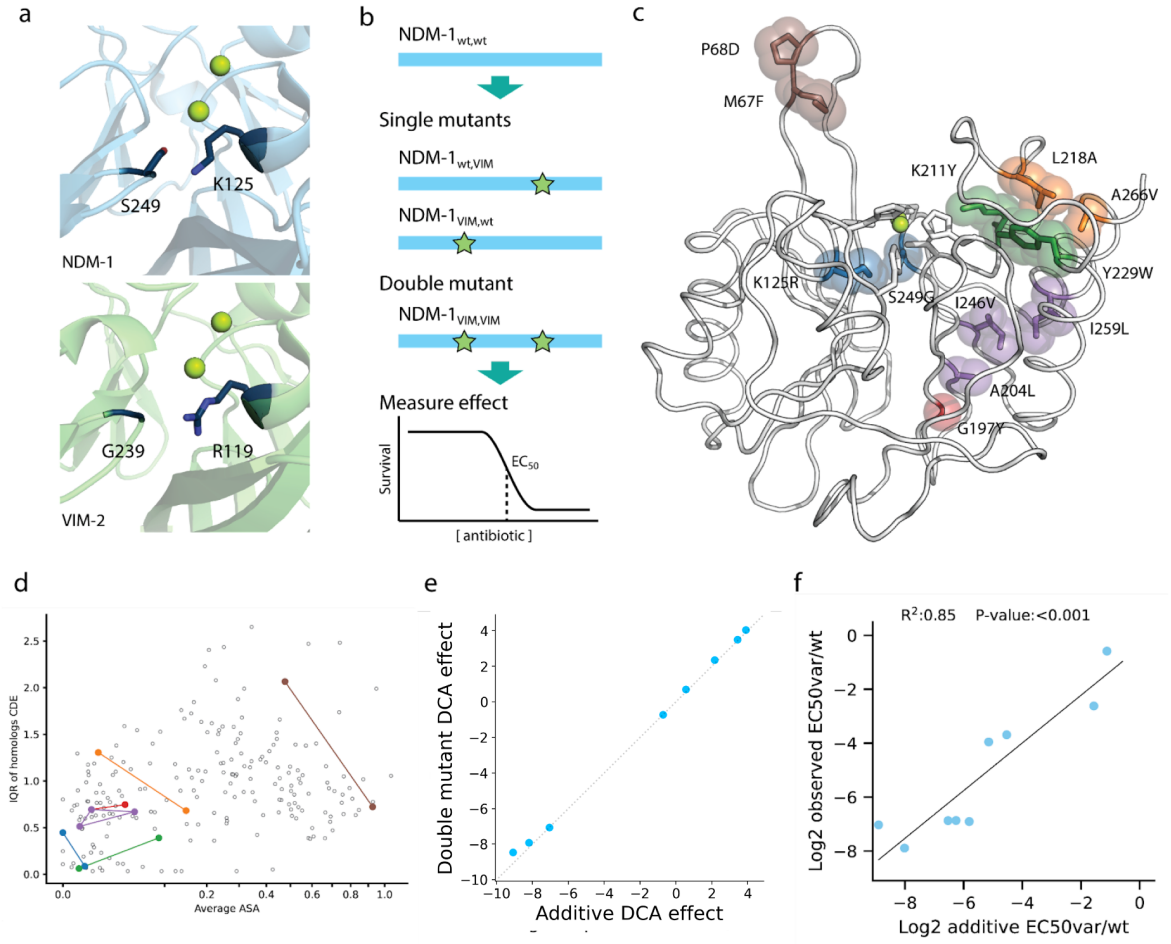


Fig. 5. Testing interactions of entrenched positions in NDM-1. (**a**) Example of potentially interacting entrenched WT positions in the crystal structures of NDM-1 (top) and VIM-2 (bottom). (**b**) Experimental scheme for testing single or combined mutational effects in the NDM-1 background. (**c**) Entrenched WT positions chosen for testing epistatic interactions. Positions with the same color are mutated together to test for compensation of entrenchment; A204L overlaps 2 sets and is also tested with G192Y (red). (**d**) Plot of IQR in CDE across 100 homologs and the average ASA of NDM-1 and VIM-2 structures, with the tested positions highlighted. Tested combinations are shown as lines. (**e**) Scatterplot of DCA energy change of all tested double (1 triple) mutants, with the expected additive single mutant effects in the x-axis, and the observed double mutant effects in the y-axis. (**f**) Scatterplot of all tested double (1 triple) mutants, with the expected additive single mutant effects in the x-axis, and the observed double mutant effects in the y-axis. Effects are calculated as fold change relative to wtNDM-1.

A subset of our data also gives us the opportunity to dissect specific epistatic interactions in the homologs. Positions that have entrenched WT provide a potential signal for residues that are functionally important and may participate in specific interactions. We sought to find surrounding interactions within the proteins by identifying pairs of positions that are in close proximity, and where the pair of positions have entrenched WT residues (**supp. Fig. 7**). One example of shape

complementarity can be found between NDM-1 positions 125 and 249 (VIM-2 positions 119 and 239) directly underneath the active site Zn^{2+} ions, where both are entrenched positions within contact distance to each other (**Fig. 5a**). In VIM-2, the Arg at position 125 is paired with Gly at position 249. In contrast, NDM bears a somewhat smaller, but still positively charged Lys at position 125, which is now paired with a larger Ser at position 249. To test if these positions possess significant interactions, we generated the single and double mutants of each position pair in the NDM-1 background by mutating to the VIM-2 WT at those positions, then measuring their phenotype (EC_{50} against AMP) (**Fig. 5b**). As the effect of VIM-2's WT is deleterious in NDM-1 as single mutants, we expect that mutating all interacting positions may lead to compensation, generating a less negative effect. We also extend this experiment to pairings in the L3 active site loop (NDM-1 positions 67+68), L10 active site loop (218+266, 211+229), and some buried positions beneath the L10 loop (197+204, and a triplet of 204+246+259) (**Fig. 5c**); the triplet was tested as all pairs of doubles and the full triplet. Combined, our selection of positions covers a variety of positions, including those with different biochemical properties (size, polarity, charge), different solvent accessibilities, and different spread in mutational behaviors across the whole family (**Fig. 5c, d**).

The effects of the single mutants validate the deleterious nature of the observed fitness scores, and we observe a sigmoidal relationship between a mutant's EC_{50} and the fitness score, which is consistent with previous observations[6]. This validates the entrenchment observed in DMS, as all selected mutations that are WT residues in VIM-2 are deleterious in the NDM-1 background. It is also notable that our selected mutations evenly span a wide range of deleterious effects. When we tested the mutants in combination, however, we did not observe significant compensation in any of the mutant combinations (**Fig. 5f**). In fact, the log-additive effects of the single mutants (null model for no epistasis) show a distinctly linear correlation with the observed double mutant effects, with an R^2 of 0.85. It appears that entrenched positions cannot be easily swapped simply by mutating other nearby entrenched residues. This scenario is confirmed by DCA where, as previously discussed, the epistatic effects can only arise as a cumulation of small contributions including a multitude of epistatically coupled positions. The double mutation effects are therefore basically additive in DCA (**Fig 5e**) as in the experiment (**Fig 5f**). It is also not the case that the positions are globally independent in their effect, as these same mutations are fixated in VIM-2 and the deleterious effects have been compensated for during evolution. Thus, these experiments suggest that the intramolecular network of residue interactions is much more complex, and even entrenched WT residues are not sufficient to dictate the epistatic networks.

Discussion

By combining the computational approach of DCA and the experimental approach of DMS, we have gained a better picture of the epistatic and mutability tendencies between the homologs within a protein family. In the B1 family, we find a prevalence of heterogeneity in mutational effects both at the level of the mutational tolerance at each position, as well as epistasis of the same mutation across different backgrounds. At a global level across all homologs, over half of the positions can exhibit a > 6-fold difference in mutational tolerance depending on the sequence background. When examined specifically in the NDM-1 and VIM-2 contexts, half of the positions with conserved WT residues have >4 mutants that are epistatic. In particular, the complementarity of the two approaches has enabled better understanding than either approach alone. The global level view provided by DCA for 100 homologs reveals the full spread in mutational behaviors, which would have been obscured for positions that behave similarly for the 2 specific homologs tested. In turn, while both approaches generally agree on the mutational behaviors of each position, the experimental results can highlight peculiarities that the statistical approach may not identify from just the extant sequence data.

To expand upon previous DMS studies involving multiple homologs, which are generally focused on higher-level statistical analyses[23–26,28,40,41], we also performed a deeper examination of the mechanisms behind the observed incompatibilities. The ASA of a protein position shows trends with various facets of epistasis. The mutability spread across homologs is lower at low and high ASA, while the epistatic behavior between two specific sequences seems most prevalent at intermediate ASA. These observations are likely underpinned by differential intramolecular networks in different homologs, as a result of gradually co-evolved epistatic networks over the course of evolution. However, directly replacing just two or three residues is not enough to compensate for evolutionarily entrenched residues, suggesting that a much more complex network of interactions is at play[42], as predicted by the DCA model.

We suspect that many of the observed behaviors are not limited to NDM-1 and VIM-2, or the B1 family. The trends with regard to structure in both DCA and DMS, which are typical of those observed in other systems[43–45], suggest that the behaviors we observe can be explained through general mechanisms such as secondary structure formation and structural packing. However, more evidence from other systems would be required to distinguish the global trends from system-specific trends. Overall, we provide a global view of epistatic networks at the protein family level that complements more detailed examinations of epistasis in specific residues[26,46].

Methods

Sequence collection for the B1 MBL family

In an effort to comprehensively isolate all B1 MBLs, we used a broad sweep approach through a sequence similarity network (SSN) using data from genomic (UniProt) and metagenomic (JGI) databases. Initially, an SSN was constructed using the EFI-EST tool using the InterPro ID IPR001279 (MBL domain superfamily) as the input. The network was analyzed at different alignment score cut-offs to find a cut-off where all clinically isolated B1 enzymes fell within the same isolated cluster (~1,800 sequences at EFI alignment score cut-off of 25). The sequences that were shorter than 220aa were removed, and the remainder (~1,300) were clustered by cd-hit to 60% identity to reduce redundancy (187 sequences final) and then used to generate an HMM, *i.e.*, sequence profile, of the B1 family. The HMM was used to search for B1 sequences in the JGI database resulting in ~2.5M sequences from JGI, with 1,859,503 non-redundant sequences determined by cd-hit clustering at 100%.

To construct the final SSN, all sequences of the MBL domain superfamily were downloaded from UniProt using the InterPro IPR001279 definition (434,623 sequences, accessed 2019). The UniProt and JGI data were combined, sequences with ambiguities were removed, and the length was restricted to between 100-1500aa. Cd-hit was used to successively cluster the sequences to 100%, 90%, 70%, and finally, 50% identity to reduce redundancy, resulting in 86,947 representative sequences for the entire MBL superfamily. A SSN was generated by performing 'all by all' BLASTP with an e-value cutoff of 1e-5 and 1000 maximum hits per sequence, giving a raw network of ~90Mil pairwise bitscore values (edges) between all sequences (nodes). The network was processed using an in-house SSN analysis pipeline (MetaSSN) to identify the lowest bitscore cutoff at which all clinically isolated B1 sequences break off into an isolated cluster (10,252 sequences). Finally, to identify sequences that are most likely to be active B1 sequences, the dataset was filtered by length to between 200-350aa (6,828) and used to build a multiple sequence alignment (MSA) using Clustal Omega on default settings. The MSA was manually curated to exclude any misaligned sequences, resulting in 6308 curated sequences. Finally, to ensure the sequences were likely to have B1-like function, only sequences aligned with the conserved B1 active-site metal binding residues(H116/H118/H196 and D120/C221/H263, BBL numbering) were kept, resulting in a final dataset of 5035 B1 family sequences, with a roughly 50/50 split between sequences from UniProt and JGI.

Direct coupling analysis of the B1 family

To generate an alignment of conserved structure, crystal structures of 14 orthologs were aligned using mTM-align (https://yanglab.nankai.edu.cn/mTM-align/)[47]. The orthologs (PDB IDs) are: NDM-1 (3spu), VIM-2 (5yd7), DIM-1 (4wd6), ECV-1 (6t5k), FIM-1 (6v3q), GIM-1 (2ynt), IMP-1(4uam), IND-7 (3l6n), MYO-1 (6t5l), TMB-1 (5mmd), VMB-1 (6jv4), bc-II (1bc2), blaB (1m2x) and cfiA (1znb). An HMM sequence profile was trained on this curated dataset by using HMMER via the *hmmbuild* command. We used a *-symfrac* value of 0.3 to control the maximum number of gaps in each alignment column. Then, the 5035 B1 sequences were then aligned to this profile via the *hmmsearch* command to produce the MSA for training the DCA model. After converting the alignment to FASTA format, all sequences exhibiting more than 10% gaps were removed from the alignment. Additionally, flank columns exhibiting more than 75% gaps were eliminated. The resulting alignment contained 222 sites. A final refinement was achieved after removing all sequences that exhibited more than 80% sequence identity to NDM-1 or VIM-2, thereby ensuring the alignment was not biased toward the sequences used for further analysis. The resulting MSA had 3655 sequences. We then inferred on this alignment a DCA model by using the standard settings of adabmDCA[48].

Library generation and deep mutational scanning of NDM-1

The procedure for library generation and DMS on NDM-1 was conducted in an identical manner to VIM-2. The wtNDM-1 sequence is encoded on an in-house pIDR5.1 plasmid, expressed under a constitutive AmpR promoter. To generate all single amino acid mutants, we used a PCR-based method (restriction-free cloning) to introduce a degenerate 'NNN' sequence at each codon in the coding sequence in separate reactions. After every single position was mutated, the library was combined into 7 groups of 39 consecutive positions each, forming 117nt long mutated regions that were fully sequenced by paired-end Illumina NextSeq reads.

The NDM libraries are then transformed into *Escherichia coli* (*E. cloni* 10G, Lucigen) and stored as glycerol stocks, with the number of colony-forming units after transformation to be >= 100,000 to ensure complete coverage of each group. To perform selection experiments, we inoculate the glycerol stocks into fresh LB (Fisher) and grow the cultures shaking overnight at 30°C for 16 hours. The cultures are diluted into fresh LB to an inoculum of $1.5x10^6$ cells/mL (targeting a 1:1000 dilution of a culture with OD_{600} of 1.5) and grown shaking at 37°C for 2 hours. Selection pressure is introduced by mixing 960uL of cell culture with 40uL of a 25x stock of each antibiotic suspended in

LB, for a final culture volume of 1mL. We tested selection at 32, 128, and 256ug/mL AMP, 2, 16, 32ug/mL CTX, and 0.063, 0.25 and 0.5ug/mL MEM. We also grew a sample of the library without selection. The culture is grown under selection while shaking at 37°C for 6 hours, then removed from selection via centrifugation and resuspension in 1mL of fresh LB, repeated 3 times. The post-selection culture is grown, shaking overnight at 30°C for 16hrs and the plasmid DNA is purified using a QiaPrep 96-column DNA purification kit (Qiagen). This procedure was conducted twice on different days to produce two separate replicates.

To deep sequence the selected library, we used primers targeting unmutated regions that directly flank the mutated region of each of the 7 library groups to amplify the DNA and attach Nextera adaptors to the amplicons. The amplicons undergo a second PCR to attach the Illumina sequencing indices and flowcell binding sites. All tested samples (all groups, conditions, replicates) were sequenced in the same Illumina NextSeq 550 run with fully overlapping paired-end reads. We also included control samples of amplicons extracted from just wtNDM-1 using the primers for each group in the NextSeq run. After deep sequencing, the forward and reverse reads are merged together using our previously published pipeline (https://github.com/johnchen93/DMS-FastQ-processing), and we discard reads with greater than 20 mismatches between forward and reverse reads or with a posterior Q score of less than 10. We use the wtNDM-1 samples as an estimate for error rates arising from the deep sequencing process, and we filter the non-selected libraries to remove variants with frequencies less than 2x of the expected frequency from sequencing noise alone, or variants with less than 5 reads. We then calculate the fitness scores for each variant in all conditions according to eq (1). All variants that pass filtering in the non-selected condition are considered to truly exist in the library, and if the same variants are not observed in conditions undergoing selection they are assumed to have been depleted by selection and are given a dummy count of 1 to simulate the lowest possible frequency.

Generation and dose-response assay of NDM single and combined mutants

The selected positions with entrenched WT behaviors were mutated in the NDM-1 background from the NDM-1 wt residue to the VIM-2 wt residue using Golden Gate cloning as single mutants (NDM-1 positions M67F, P68D, K125R, G197Y, A204L, K211Y, L218A, Y229W, I246V, I259L, A266V). Then, combinations of positions were generated by a second (or third for the triplet) round of mutations. Dose-response curves were carried out to obtain the half-maximal effective concentration (EC_{50}) under ampicillin (AMP) selection. For testing, mutants and wtNDM-1 were transformed into *E. coli* and grown overnight for 16hrs at 30°C, then diluted to a target inoculum of1.2Mil cells / mL (OD_{600}

of 0.0015) the next day. The diluted culture is grown for 1.5 hours at 37°C, then 180uL of culture is mixed with 20uL of 10x AMP stock, with final ampicillin concentrations from 1-1024ug/mL. Growth under AMP selection is done for 6 hours at 37°C, and the OD_{600} of each culture is measured after selection. For each mutant or wtNDM-1, the OD_{600} across all selected AMP concentrations is plotted as a dose-response curve and fitted using a sigmoidal equation to obtain the EC_{50}.

References

1. Socha, R. D., Chen, J. & Tokuriki, N. The Molecular Mechanisms Underlying Hidden Phenotypic Variation among Metallo-β-Lactamases. *Journal of Molecular Biology* **431**, 1172–1185 (2019).
2. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology* **307**, 1113–1143 (2001).
3. De Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* **15**, 480–490 (2014).
4. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution* **31**, 1581–1592 (2014).
5. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Comput. Biol.* **11**, e1004421 (2015).
6. Chen, J. Z., Fowler, D. M. & Tokuriki, N. Environmental selection and epistasis in an empirical phenotype–environment–fitness landscape. *Nat Ecol Evol* 1–12 (2022) doi:10.1038/s41559-022-01675-5.
7. Matreyek, K. A., Stephany, J. J., Ahler, E. & Fowler, D. M. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Med* **13**, 165 (2021).

8. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
9. Flynn, J. M. *et al.* Comprehensive fitness maps of Hsp90 show widespread environmental dependence. *eLife* **9**, e53810 (2020).
10. Thompson, S., Zhang, Y., Ingle, C., Reynolds, K. A. & Kortemme, T. Altered expression of a quality control protease in E. coli reshapes the in vivo mutational landscape of a model enzyme. *eLife* **9**, e53476 (2020).
11. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* **160**, 882–892 (2015).
12. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *Journal of Molecular Biology* **369**, 1318–1332 (2007).
13. Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive Cryptic Epistasis in Molecular Evolution. *PLOS Genetics* **6**, e1001162 (2010).
14. Fram, B. *et al. Simultaneous enhancement of multiple functional properties using evolution-informed protein design*. http://biorxiv.org/lookup/doi/10.1101/2023.05.09.539914 (2023) doi:10.1101/2023.05.09.539914.
15. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol Biol Evol* **33**, 268–280 (2016).
16. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108**, (2011).
17. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
18. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* **35**,

128–135 (2017).

19. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Molecular Biology and Evolution* **35**, 1018–1027 (2018).
20. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.* **81**, 032601 (2018).
21. Russ, W. P. *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
22. Alvarez, S. *et al.* In vivo *functional phenotypes from a computational epistatic model of evolution*. http://biorxiv.org/lookup/doi/10.1101/2023.05.24.542176 (2023) doi:10.1101/2023.05.24.542176.
23. Doud, M. B., Ashenberg, O. & Bloom, J. D. Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Mol Biol Evol* **32**, 2944–2960 (2015).
24. Chan, Y. H., Venev, S. V., Zeldovich, K. B. & Matthews, C. R. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nature Communications* **8**, 14614 (2017).
25. Lee, J. M. *et al.* Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proc Natl Acad Sci U S A* **115**, E8276–E8285 (2018).
26. Heyne, M. *et al.* Climbing Up and Down Binding Landscapes through Deep Mutational Scanning of Three Homologous Protein-Protein Complexes. *J Am Chem Soc* **143**, 17261–17275 (2021).
27. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *eLife* **7**, e34300 (2018).
28. Pokusaeva, V. O. *et al.* An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLOS Genetics* **15**, e1008079 (2019).

29. Fröhlich, C., Chen, J. Z., Gholipour, S., Erdogan, A. N. & Tokuriki, N. Evolution of beta-lactamases and enzyme promiscuity. *Protein Engineering, Design and Selection* **34**, gzab013 (2021).
30. Bush, K. Past and Present Perspectives on β-Lactamases. *Antimicrob. Agents Chemother.* **62**, (2018).
31. Berglund, F., Johnning, A., Larsson, D. G. J. & Kristiansson, E. An updated phylogeny of the metallo-β-lactamases. *J Antimicrob Chemother* doi:10.1093/jac/dkaa392.
32. Hall, B. G. & Barlow, M. Evolution of the serine β-lactamases: past, present and future. *Drug Resistance Updates* **7**, 111–123 (2004).
33. Fröhlich, C., Chen, J. Z., Gholipour, S., Erdogan, A. N. & Tokuriki, N. Evolution of β-lactamases and enzyme promiscuity. *Protein Engineering, Design and Selection* **34**, (2021).
34. Vigué, L. *et al.* Deciphering polymorphism in 61,157 Escherichia coli genomes via epistatic sequence landscapes. *Nat Commun* **13**, 4030 (2022).
35. Rodriguez-Rivas, J., Croce, G., Muscat, M. & Weigt, M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113118119 (2022).
36. Chen, J. Z., Fowler, D. M. & Tokuriki, N. Comprehensive exploration of the translocation, stability and substrate recognition requirements in VIM-2 lactamase. *eLife* **9**, e56707 (2020).
37. Chothia, C. Hydrophobic bonding and accessible surface area in proteins. *Nature* **248**, 338–339 (1974).
38. Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7159–7164 (2015).
39. Deng, Z. *et al.* Deep Sequencing of Systematic Combinatorial Libraries Reveals β-Lactamase Sequence Constraints at High Resolution. *Journal of Molecular Biology* **424**,

150–167 (2012).

40. Park, Y., Metzger, B. P. H. & Thornton, J. W. Epistatic drift causes gradual decay of predictability in protein evolution. 9 (2022).
41. Youssef, N., Susko, E., Roger, A. J. & Bielawski, J. P. Shifts in amino acid preferences as proteins evolve: A synthesis of experimental and theoretical work. *Protein Science* **30**, 2009–2028 (2021).
42. Birgy, A. *et al.* Local and Global Protein Interactions Contribute to Residue Entrenchment in Beta-Lactamase TEM-1. *Antibiotics* **11**, 652 (2022).
43. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).
44. Probing the biophysical constraints of SARS-CoV-2 spike N-terminal domain using deep mutational scanning. *SCIENCE ADVANCES* (2022).
45. Sruthi, C. K., Balaram, H. & Prakash, M. K. Toward Developing Intuitive Rules for Protein Variant Effect Prediction Using Deep Mutational Scanning Data. *ACS Omega* **5**, 29667–29677 (2020).
46. Ogbunugafor, C. B., Guerrero, R. F., Shakhnovich, E. I. & Shoulders, M. D. Epistasis meets pleiotropy in shaping biophysical protein subspaces associated with antimicrobial resistance. 2023.04.09.535490 Preprint at https://doi.org/10.1101/2023.04.09.535490 (2023).
47. Dong, R., Peng, Z., Zhang, Y. & Yang, J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* **34**, 1719–1725 (2018).
48. Muntoni, A. P., Pagnani, A., Weigt, M. & Zamponi, F. adabmDCA: adaptive Boltzmann machine learning for biological sequences. *BMC Bioinformatics* **22**, 528 (2021).

3.3 FURTHER ANALYSIS

3.3.1 Gain of function mutations

Here we briefly elaborate on the lack of gain of function mutations in our experimental system and why it confounds the detection of epistatic mutations. As mentioned, the experimental setup used to assess the fitness effects of mutations in VIM-2 and NDM-1 has a big limitation: it struggles to detect gain of function mutations. This limitation arises because mutations that allow for survival at higher antibiotic concentrations than those necessary for the growth during the experiments are irrelevant. To address this, an alternative experimental system would be required. This shortcoming complicates the understanding and measurement of epistasis due to the non-linear relationship between fitness and function; while function can continue to increase, fitness may plateau.

To visualize this, imagine an experimental system that effectively detects gains in both function and fitness. It would be straightforward to differentiate between epistatic and non-epistatic positions, as shown in Panel A of Fig. 3.1. Non-epistatic positions align with the $y = -x$ diagonal, while non-epistatic ones are located in the lower quadrant. However, without the ability to detect gain of function mutations, the data points skew towards the negative quadrant, leading to ambiguity. This distortion is evident in Panel B, where non-epistatic positions adopt an 'L' shape, causing epistatic and non-epistatic positions to mix. As we have discussed in the paper, we are still able to distinguish the two classes in the case of 2-wt entrenched mutations and we have discussed how 1-wt entrenched mutations might be discerned.

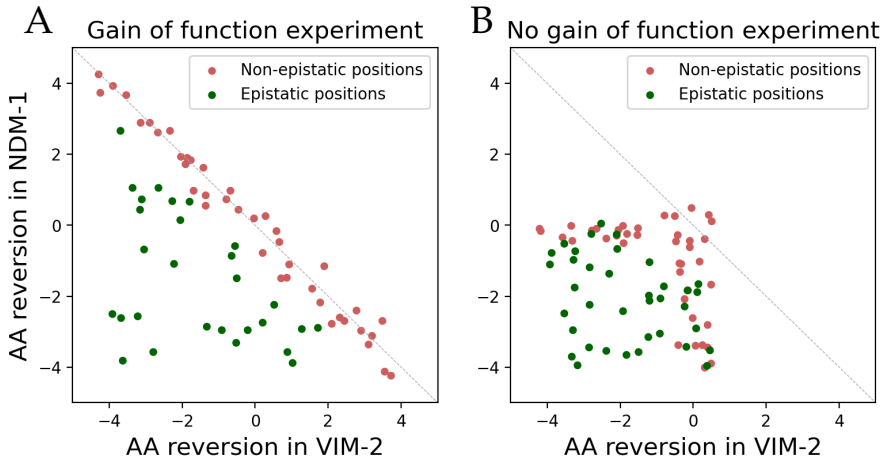


Figure 3.1 – Scatterplots illustrating two hypothetical datasets of reversion mutations between VIM-2 and NDM-1. The left panel depicts data with gain of function mutations, while the right panel, consistent with experimental data, excludes them. Both charts differentiate between non-epistatic and epistatic positions.

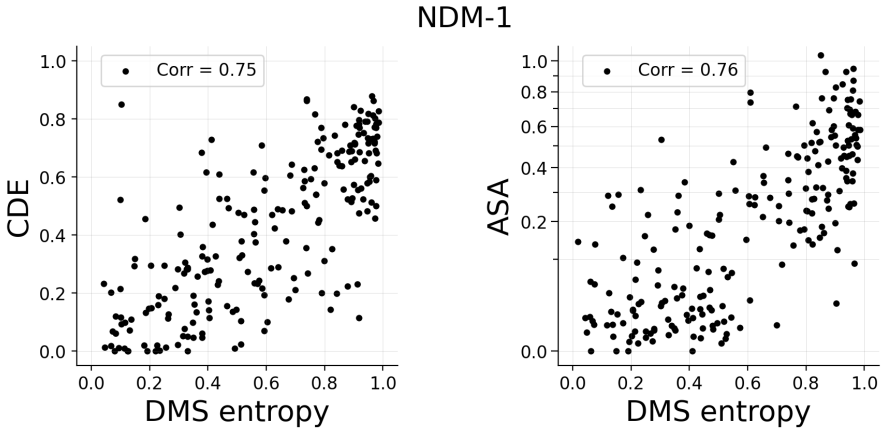


Figure 3.2 – Scatterplot between DMS entropies and CDE (left plot) or ASA (right plot) for the enzyme NDM-1.

3.3.2 Combining structural and evolutionary information

In the paper discussed in the previous section, we have obtained the mutability data of two homologous sequences from an experimental, structural, and computational point of view. Here we discuss how to combine these different quantities to create a predictive mutability score. In particular, we combine the residue accessible surface area (ASA) with the context-dependent entropy (CDE) of each site, to predict the experimental mutability of sites, computed as the entropy of the DMS fitnesses. For both NDM-1 and VIM-2 there is already a solid correlation (~ 0.75) between ASA and DMS mutability, as well as between CDE and DMS mutability, see for example Fig. 3.2.

To improve this correlation and combine the the two types of information, we trained a logistic regression model on 217 aligned positions of NDM-1 to predict the DMS entropy based on the ASA and CDE values of each position (each quantity was rescaled to lie between 0 and 1). The logistic regression model is:

$$p_{\boldsymbol{\beta}}(x_1, x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} \tag{3.1}$$

where x_1 and x_2 represent the rescaled ASA and CDE values respectively, and $\boldsymbol{\beta}$ denotes the set of 3 parameters $(\beta_0, \beta_1, \beta_2)$ learned by the model.

Once fitted, we can compare the model's predictions with the DMS values. In the case of NDM-1, we see an increase in correlation from 0.75 to 0.82. More interestingly, we can use the same $\boldsymbol{\beta}$ parameters to compute the DMS entropies for VIM-2, without re-training our model, by using the corresponding x_1 and x_2 values for VIM-2. The correlation in this case raises quite a lot, from 0.75 of CDE or ASA alone to 0.86, see Fig. 3.3. An increase in correlation by combining structural and evolutionary information has been seen before, for example [129], but

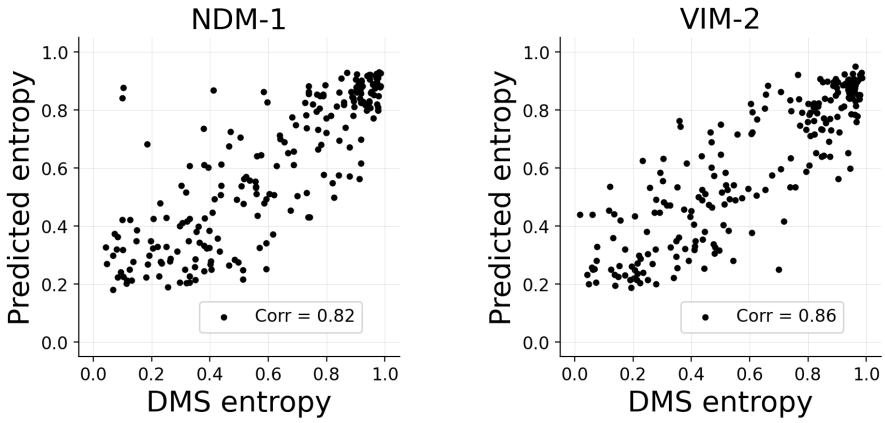


Figure 3.3 – Scatterplot between DMS entropies and predictions by a logistic regression. The logistic regression has been trained with CDE and ASA of NDM-1 and then tested on both NDM-1 (left plot) and VIM-2 (right plot) datasets.

to our knowledge is was never tested independently on a different homolog, suggesting that the learned parameters are robust. Since the model is so simple, we can also try to interpret it. First, we look at the value of the β_1 and β_2 parameters of the logistic regression. We find for each of them an almost identical value of $\beta_{1,2} \sim 2.5$ indicating an equal participation of the evolutionary and structural information to the prediction accuracy. Using $\beta_{1,2}$ in place of β_1 and β_2, we can interpret the logistic regression as a soft OR classifier:

- when ASA and CDE are concordant, i.e. they are both high or low, the prediction is respectively high and low. For example:

$$\begin{aligned} p_\beta(1,1) &= \frac{1}{1+e^{-(\beta_0+2\beta_{1,2})}} = 0.97 \\ p_\beta(0,0) &= \frac{1}{1+e^{-\beta_0}} = 0.17 \end{aligned} \tag{3.2}$$

- when ASA and CDE are discordant, i.e. one is high and the other is low, the prediction tends to be quite high. For example:

$$p_\beta(1,0) = p_\beta(0,1) = \frac{1}{1+e^{-(\beta_0+\beta_{1,2})}} \sim 0.7 \tag{3.3}$$

In practice, the prediction is intuitive when the information from both ASA and CDE is consistent, as one would naturally anticipate. However, when there is a discrepancy between the two, the site is predicted to be mutable. Indeed, we could already see from Fig. 3.2 that both ASA and CDE underestimate the mutability of the sites. The fact that most sites of NDM-1 and VIM-2 are well predicted by the logistic regression (with the notable exception of a few sites in NDM-1, see left plot of Fig. 3.3) indicates two things. Firstly, some

buried sites are mutable according to the DMS data, yet these sites can be accurately identified as mutable using the DCA-derived entropies. Secondly, there are sites that, while appearing conserved based on the CDE, are mutable according to the DMS. Notably, these sites are exposed and can therefore be predicted to be variable. Such a case might require a more in-depth examination. It is intriguing that some sites can be variable in the DMS, yet show no variability in the alignment of their homologs, i.e. having a low CDE. This inconsistency may indicate that the experimental selection used in DMS experiments might not fully capture the natural selection pressures exerted on these enzymes, at least for certain sites.

4 CONCLUSION AND OUTLOOK

Protein evolution is a fundamental biological process that underlies the adaptation and diversification of life on Earth. A deep understanding of this dynamical process can significantly enhance our ability to respond to emerging pathogens in a clinical setting [130], refine our comprehension of the history of life on Earth [131], and even aid in the creation of artificial enzymes [132]. However, despite being a pervasive phenomenon across all branches of life, protein evolution remains extremely hard to predict. One of the primary challenges in developing a predictive theory of protein evolution is epistasis, the context-dependence of mutational effects. The fact that mutations have different effects depending on the sequence background they happen in, makes it difficult to predict the outcomes of new mutations accurately. Epistasis dictates which mutations can be accepted in a specific sequence background, and can open or close evolutionary paths. Notably, proteins from different species can have the same function, i.e. catalyze the same biochemical reaction, while exhibiting great differences in their amino acid sequences. This diversity is integral to the flexibility of proteins and plays a crucial role in the development of machine learning models, such as Potts models, which aim to reproduce the probability distribution of functional proteins within sequence space. Traditionally, these models have been employed to capture the equilibrium properties of protein families and predict protein-protein interactions [49, 133], residue-residue contacts [80], or the fitness effect of mutations [98].

The research presented in this thesis shifts the focus from the static properties of protein fitness landscapes toward the dynamical properties that dictate the evolution of proteins. In section 2.2, we demonstrate how to simulate genetic drift experiments with β-lactamase enzymes [72, 110] thanks to an algorithm for sampling new mutations based on a data-driven sequence landscape while including the constraints of the genetic code. These simulations unveil the values of the experimental parameters necessary for identifying the emergence of epistasis and inferring protein structures from sequencing data. The quantitative character of our evolutionary dynamics suggests that it could be exploited to develop methods that directly infer the fitness landscape from sequence data, similar to the work of Sesta et al. [134]. In section 2.4 we further refine our stochastic dynamics to generate long trajectories that converge to the sequence statistics of the training set, i.e. natural sequences. This allows us to simulate evolution at every time scale. Our findings highlight the

pivotal role of epistasis in driving protein evolution and determining the timing of mutations' emergence. The mutability of sites is strongly dependent on a local measure of mutability that we define thanks to our model. Our approach is highly versatile, requiring only a model that can calculate the relative likelihood of two sequences, opening the applicability to more powerful machine learning models. We envisage further refinements of this model, including species-specific mutational biases and more complex mutational processes.

In section 3.2, we analyze a dataset comprising 100 homologs from the B1 β-lactamase protein family, revealing widespread variability in the mutability of sites across the protein family. We validate the predictions of our model using DMS data from two clinically relevant sequences, VIM-2 and NDM-1. We show a strong correlation between experimental and model mutability and we detect the presence of epistasis in a significant proportion of sites. This study compares mutations between two enzymes with more than 60% sequence divergence, a distance big enough to disclose mutations that are neutral in one genetic background but deleterious in another. How this transition happens, however, cannot be determined by analyzing only the snapshot given by two sequences. To better understand how epistasis changes the effect of mutations during evolution, the same mutations need to be probed along an evolutionary pathway. A recently published work [135, 136] has shown a way to sample evolutionary trajectories between two points in sequence space with sequence models. We believe that it could be possible to employ a similar strategy, inspired by Transition Path Sampling [137, 138], with Potts models, to construct and test evolutionary trajectories of β-lactamases. This would allow the creation of "anchors" protein in sequence space, that are not accessible experimentally, and that can be further characterized. We are currently developing different path-sampling algorithms to generate all single mutants directly connecting the enzymes VIM-2 and NDM-1. This means sampling the space of 2^n mutations, where $n = 141$ is the number of different residues between the two aligned sequences, searching for low-energy intermediate sequences. The designed sequences will be synthesized thanks to DropSynth (see section 1.5.3), transformed into *E.coli* and phenotyped in bulk against different antibiotics to find functional sequences. By performing Deep Mutational Scanning (DMS) experiments on multiple intermediates along these paths, we aim to unravel the functional and possibly structural mechanisms behind epistasis.

BIBLIOGRAPHY

[1] Mahmood A. Rashid, Firas Khatib, and Abdul Sattar. *Protein preliminaries and structure prediction fundamentals for computer scientists*. en. arXiv:1510.02775 [cs, q-bio]. Oct. 2015.

[2] Christian B. Anfinsen. "Principles that Govern the Folding of Protein Chains". en. In: *Science* 181.4096 (July 1973), pp. 223–230. DOI: 10.1126/science.181.4096.223.

[3] Genetic Science Learning Center. *How do cells read genes?* Mar. 2016.

[4] John Maynard Smith. "Natural Selection and the Concept of a Protein Space". en. In: *Nature* 225.5232 (Feb. 1970), pp. 563–564. DOI: 10.1038/225563a0.

[5] Maido Remm, Christian E.V. Storm, and Erik L.L. Sonnhammer. "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons". en. In: *Journal of Molecular Biology* 314.5 (Dec. 2001), pp. 1041–1052. DOI: 10.1006/jmbi.2000.5197.

[6] Aashiq H. Kachroo et al. "Systematic humanization of yeast genes reveals conserved functions and genetic modularity". en. In: *Science* 348.6237 (May 2015), pp. 921–925. DOI: 10.1126/science.aaa0769.

[7] Calin Plesa et al. "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes". en. In: *Science* 359.6373 (Jan. 2018), pp. 343–347. DOI: 10.1126/science.aao5167.

[8] William P. Russ et al. "An evolution-based model for designing chorismate mutase enzymes". en. In: *Science* 369.6502 (July 2020), pp. 440–445. DOI: 10.1126/science.aba3304.

[9] Melanie G. Lee and Paul Nurse. "Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2". en. In: *Nature* 327.6117 (May 1987), pp. 31–35. DOI: 10.1038/327031a0.

[10] The UniProt Consortium et al. "UniProt: the Universal Protein Knowledgebase in 2023". en. In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D523–D531. DOI: 10.1093/nar/gkac1052.

[11] C. Chothia and A.M. Lesk. "The relation between the divergence of sequence and structure in proteins." en. In: *The EMBO Journal* 5.4 (Apr. 1986), pp. 823–826. DOI: 10.1002/j.1460-2075.1986.tb04288.x.

[12] O Gotoh. “Multiple sequence alignment: Algorithms and applications”. en. In: *Advances in Biophysics* 36 (1999), pp. 159–206. DOI: 10.1016/S0065-227X(99)80007-0.

[13] Jiannan Chao, Furong Tang, and Lei Xu. “Developments in Algorithms for Sequence Alignment: A Review”. en. In: *Biomolecules* 12.4 (Apr. 2022), p. 546. DOI: 10.3390/biom12040546.

[14] S R Eddy. “Profile hidden Markov models.” en. In: *Bioinformatics* 14.9 (Jan. 1998), pp. 755–763. DOI: 10.1093/bioinformatics/14.9.755.

[15] Jaina Mistry et al. “Pfam: The protein families database in 2021”. en. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D412–D419. DOI: 10.1093/nar/gkaa913.

[16] Simon C Potter et al. “HMMER web server: 2018 update”. en. In: *Nucleic Acids Research* 46.W1 (July 2018), W200–W204. DOI: 10.1093/nar/gky448.

[17] Anders Krogh. “An Introduction to Hidden Markov Models for Biological Sequences”. In: *Computational Methods in Molecular Biology*. Elsevier, 1998.

[18] David Talavera, Simon C. Lovell, and Simon Whelan. “Covariation Is a Poor Measure of Molecular Coevolution”. en. In: *Molecular Biology and Evolution* 32.9 (Sept. 2015), pp. 2456–2468. DOI: 10.1093/molbev/msv109.

[19] Shilpi Chaurasia and Julien Y. Dutheil. “The Structural Determinants of Intra-Protein Compensatory Substitutions”. en. In: *Molecular Biology and Evolution* 39.4 (Apr. 2022). Ed. by Rebekah Rogers, msac063. DOI: 10.1093/molbev/msac063.

[20] EClinicalMedicine. “Antimicrobial resistance: a top ten global public health threat”. en. In: *eClinicalMedicine* 41 (Nov. 2021), p. 101221. DOI: 10.1016/j.eclinm.2021.101221.

[21] Mohammad Ahmad and Asad U. Khan. “Global economic impact of antibiotic resistance: A review”. en. In: *Journal of Global Antimicrobial Resistance* 19 (Dec. 2019), pp. 313–316. DOI: 10.1016/j.jgar.2019.05.024.

[22] Alexander Fleming. “On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae”. en. In: *British Journal of Experimental Pathologyol* (1929).

[23] Jessica M. A. Blair et al. “Molecular mechanisms of antibiotic resistance”. en. In: *Nature Reviews Microbiology* 13.1 (Jan. 2015), pp. 42–51. DOI: 10.1038/nrmicro3380.

[24] Christopher J L Murray et al. “Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis”. en. In: *The Lancet* 399.10325 (Feb. 2022), pp. 629–655. DOI: 10.1016/S0140-6736(21)02724-0.

[25] Chunming Xu et al. “A Review of Current Bacterial Resistance to Antibiotics in Food Animals”. In: *Frontiers in Microbiology* 13 (May 2022), p. 822689. DOI: 10.3389/fmicb.2022.822689.

[26] Karen Bush and Patricia A. Bradford. “Beta-Lactams and Beta-Lactamase Inhibitors: An Overview”. en. In: *Cold Spring Harbor Perspectives in Medicine* 6.8 (Aug. 2016), a025247. DOI: 10.1101/cshperspect.a025247.

[27] Arnold L. Demain and Richard P. Elander. “The beta-lactam antibiotics: past, present, and future”. In: *Antonie van Leeuwenhoek* 75.1/2 (1999), pp. 5–19. DOI: 10.1023/A:1001738823146.

[28] Louis B. Rice. “Mechanisms of Resistance and Clinical Relevance of Resistance to Beta-Lactams, Glycopeptides, and Fluoroquinolones”. en. In: *Mayo Clinic Proceedings* 87.2 (Feb. 2012), pp. 198–208. DOI: 10.1016/j.mayocp.2011.12.003.

[29] Karen Bush. “Past and Present Perspectives on beta-Lactamases”. en. In: *Antimicrobial Agents and Chemotherapy* 62.10 (Oct. 2018), e01076–18. DOI: 10.1128/AAC.01076-18.

[30] Kapil Tahlan and Susan E Jensen. “Origins of the beta-lactam rings in natural products”. en. In: *The Journal of Antibiotics* 66.7 (July 2013), pp. 401–410. DOI: 10.1038/ja.2013.24.

[31] Christopher Fröhlich et al. “Evolution of beta-lactamases and enzyme promiscuity”. en. In: *Protein Engineering, Design and Selection* 34 (Feb. 2021), gzab013. DOI: 10.1093/protein/gzab013.

[32] Catherine L. Tooke et al. “Beta-Lactamases and Beta-Lactamase Inhibitors in the 21st Century”. en. In: *Journal of Molecular Biology* 431.18 (Aug. 2019), pp. 3472–3500. DOI: 10.1016/j.jmb.2019.04.002.

[33] Florian Baier and Nobuhiko Tokuriki. “Connectivity between Catalytic Landscapes of the Metallo-Beta-Lactamase Superfamily”. en. In: *Journal of Molecular Biology* 426.13 (June 2014), pp. 2442–2456. DOI: 10.1016/j.jmb.2014.04.013.

[34] Sevan Gholipour. “Comprehensive characterization of beta-lactamase resistome”. en. PhD thesis. Univeristy of British Columbia, 2018.

[35] William T. Harvey et al. “SARS-CoV-2 variants, spike mutations and immune escape”. en. In: *Nature Reviews Microbiology* 19.7 (July 2021), pp. 409–424. DOI: 10.1038/s41579-021-00573-0.

[36] J. Arjan G.M. De Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution". en. In: *Nature Reviews Genetics* 15.7 (July 2014), pp. 480–490. DOI: 10.1038/nrg3744.

[37] Michael Lässig, Ville Mustonen, and Aleksandra M. Walczak. "Predicting evolution". en. In: *Nature Ecology & Evolution* 1.3 (Feb. 2017), p. 0077. DOI: 10.1038/s41559-017-0077.

[38] Susanna Manrubia et al. "From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics". en. In: *Physics of Life Reviews* 38 (Sept. 2021), pp. 55–106. DOI: 10.1016/j.plrev.2021.03.004.

[39] Tyler N. Starr and Joseph W. Thornton. "Epistasis in protein evolution". en. In: *Protein Science* 25.7 (July 2016), pp. 1204–1218. DOI: 10.1002/pro.2897.

[40] Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner. "The Causes and Consequences of Genetic Interactions (Epistasis)". en. In: *Annual Review of Genomics and Human Genetics* 20.1 (Aug. 2019), pp. 433–460. DOI: 10.1146/annurev-genom-083118-014857.

[41] Milo S. Johnson, Gautam Reddy, and Michael M. Desai. "Epistasis and evolution: recent advances and an outlook for prediction". en. In: *BMC Biology* 21.1 (May 2023), p. 120. DOI: 10.1186/s12915-023-01585-3.

[42] Charlotte M. Miton, Karol Buda, and Nobuhiko Tokuriki. "Epistasis and intramolecular networks in protein evolution". en. In: *Current Opinion in Structural Biology* 69 (Aug. 2021), pp. 160–168. DOI: 10.1016/j.sbi.2021.04.007.

[43] Charlotte M. Miton and Nobuhiko Tokuriki. "How mutational epistasis impairs predictability in protein evolution and design: How Epistasis Impairs Predictability in Enzyme Evolution". en. In: *Protein Science* 25.7 (July 2016), pp. 1260–1272. DOI: 10.1002/pro.2876.

[44] Jakub Otwinowski. "Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function". en. In: *Molecular Biology and Evolution* 35.10 (Oct. 2018). Ed. by Claus Wilke, pp. 2345–2354. DOI: 10.1093/molbev/msy141.

[45] Jamie T. Bridgham, Eric A. Ortlund, and Joseph W. Thornton. "An epistatic ratchet constrains the direction of glucocorticoid receptor evolution". en. In: *Nature* 461.7263 (Sept. 2009), pp. 515–519. DOI: 10.1038/nature08249.

[46] Miriam Kaltenbach et al. "Reverse evolution leads to genotypic incompatibility despite functional and active site convergence". en. In: *eLife* 4 (Aug. 2015), e06492. DOI: 10.7554/eLife.06492.

[47] Mariano M. González et al. “Optimization of Conformational Dynamics in an Epistatic Evolutionary Trajectory”. en. In: *Molecular Biology and Evolution* 33.7 (July 2016), pp. 1768–1776. DOI: 10.1093/molbev/msw052.

[48] Tanja Kortemme et al. “Computational redesign of protein-protein interaction specificity”. en. In: *Nature Structural & Molecular Biology* 11.4 (Apr. 2004), pp. 371–379. DOI: 10.1038/nsmb749.

[49] Martin Weigt et al. “Identification of direct residue contacts in protein–protein interaction by message passing”. en. In: *Proceedings of the National Academy of Sciences* 106.1 (Jan. 2009), pp. 67–72. DOI: 10.1073/pnas.0805923106.

[50] Michael J. Harms and Joseph W. Thornton. “Historical contingency and its biophysical basis in glucocorticoid receptor evolution”. en. In: *Nature* 512.7513 (Aug. 2014), pp. 203–207. DOI: 10.1038/nature13410.

[51] Anthony D. Kelleher et al. “Clustered Mutations in HIV-1 Gag Are Consistently Required for Escape from Hla-B27–Restricted Cytotoxic T Lymphocyte Responses”. en. In: *The Journal of Experimental Medicine* 193.3 (Feb. 2001), pp. 375–386. DOI: 10.1084/jem.193.3.375.

[52] Douglas M Fowler and Stanley Fields. “Deep mutational scanning: a new style of protein science”. en. In: *Nature Methods* 11.8 (Aug. 2014), pp. 801–807. DOI: 10.1038/nmeth.3027.

[53] Huijin Wei and Xianghua Li. “Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes”. en. In: *Frontiers in Genetics* 14 (Jan. 2023), p. 1087267. DOI: 10.3389/fgene.2023.1087267.

[54] Alan F Rubin et al. *MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays*. en. preprint. Genomics, Nov. 2021. DOI: 10.1101/2021.11.29.470445.

[55] Mark Lunzer, G. Brian Golding, and Antony M. Dean. “Pervasive Cryptic Epistasis in Molecular Evolution”. en. In: *PLoS Genetics* 6.10 (Oct. 2010). Ed. by David S. Guttman, e1001162. DOI: 10.1371/journal.pgen.1001162.

[56] Tyler N. Starr et al. “Pervasive contingency and entrenchment in a billion years of Hsp90 evolution”. en. In: *Proceedings of the National Academy of Sciences* 115.17 (Apr. 2018), pp. 4453–4458. DOI: 10.1073/pnas.1718133115.

[57] C. Anders Olson, Nicholas C. Wu, and Ren Sun. “A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain”. en. In: *Current Biology* 24.22 (Nov. 2014), pp. 2643–2651. DOI: 10.1016/j.cub.2014.09.072.

[58] Nicholas C Wu et al. "Adaptation in protein fitness landscapes is facilitated by indirect paths". en. In: *eLife* 5 (July 2016), e16965. DOI: 10.7554/eLife.16965.

[59] Tyler N. Starr, Lora K. Picton, and Joseph W. Thornton. "Alternative evolutionary histories in the sequence space of an ancient protein". en. In: *Nature* 549.7672 (Sept. 2017), pp. 409–413. DOI: 10.1038/nature23902.

[60] Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. "Learning the pattern of epistasis linking genotype and phenotype in a protein". en. In: *Nature Communications* 10.1 (Sept. 2019), p. 4213. DOI: 10.1038/s41467-019-12130-8.

[61] Angela M Phillips et al. "Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies". en. In: *eLife* 10 (Sept. 2021), e71393. DOI: 10.7554/eLife.71393.

[62] Claudia Bank. "Epistasis and Adaptation on Fitness Landscapes". en. In: *Annual Review of Ecology, Evolution, and Systematics* 53.1 (Nov. 2022), pp. 457–479. DOI: 10.1146/annurev-ecolsys-102320-112153.

[63] Angus M Sidore et al. "DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions". en. In: *Nucleic Acids Research* 48.16 (Sept. 2020), e95–e95. DOI: 10.1093/nar/gkaa600.

[64] Richard E Lenski. "Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations". en. In: *The ISME Journal* 11.10 (Oct. 2017), pp. 2181–2194. DOI: 10.1038/ismej.2017.69.

[65] Yajie Wang et al. "Directed Evolution: Methodologies and Applications". en. In: *Chemical Reviews* 121.20 (Oct. 2021), pp. 12384–12444. DOI: 10.1021/acs.chemrev.1c00260.

[66] Olga Kuchner and Frances H. Arnold. "Directed evolution of enzyme catalysts". en. In: *Trends in Biotechnology* 15.12 (Dec. 1997), pp. 523–530. DOI: 10.1016/S0167-7799(97)01138-4.

[67] Frances H. Arnold. "Design by Directed Evolution". en. In: *Accounts of Chemical Research* 31.3 (Mar. 1998), pp. 125–131. DOI: 10.1021/ar960017f.

[68] Philip A. Romero and Frances H. Arnold. "Exploring protein fitness landscapes by directed evolution". en. In: *Nature Reviews Molecular Cell Biology* 10.12 (Dec. 2009), pp. 866–876. DOI: 10.1038/nrm2805.

[69] Jesse D Bloom et al. "Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution". en. In: *Biology Direct* 2.1 (Dec. 2007), p. 17. DOI: 10.1186/1745-6150-2-17.

[70] Miriam Kaltenbach and Nobuhiko Tokuriki. "Generation of Effective Libraries by Neutral Drift". en. In: *Directed Evolution Library Creation*. Ed. by Elizabeth M.J. Gillam, Janine N. Copp, and David Ackerley. Vol. 1179. Series Title: Methods in Molecular Biology. New York, NY: Springer New York, 2014, pp. 69–81. DOI: 10.1007/978-1-4939-1053-3_5.

[71] Shimon Bershtein, Korina Goldin, and Dan S. Tawfik. "Intense Neutral Drifts Yield Robust and Evolvable Consensus Proteins". en. In: *Journal of Molecular Biology* 379.5 (June 2008), pp. 1029–1044. DOI: 10.1016/j.jmb.2008.04.024.

[72] Marco Fantini et al. "Protein Structural Information and Evolutionary Landscape by In Vitro Evolution". en. In: *Molecular Biology and Evolution* 37.4 (Apr. 2020). Ed. by Tal Pupko, pp. 1179–1192. DOI: 10.1093/molbev/msz256.

[73] Ayşe N. Erdoğan et al. *Neutral Drift and Threshold Selection Promote Phenotypic Variation*. en. preprint. Evolutionary Biology, Apr. 2023. DOI: 10.1101/2023.04.05.535609.

[74] Gregor Urban et al. "Protein profiles: Biases and protocols". en. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2281–2289. DOI: 10.1016/j.csbj.2020.08.015.

[75] Prateek Kumar, Steven Henikoff, and Pauline C Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm". en. In: *Nature Protocols* 4.7 (July 2009), pp. 1073–1081. DOI: 10.1038/nprot.2009.86.

[76] B Rost and C Sander. "Improved prediction of protein secondary structure by use of sequence profiles and neural networks." en. In: *Proceedings of the National Academy of Sciences* 90.16 (Aug. 1993), pp. 7558–7562. DOI: 10.1073/pnas.90.16.7558.

[77] Michael Socolich et al. "Evolutionary information for specifying a protein fold". en. In: *Nature* 437.7058 (Sept. 2005), pp. 512–518. DOI: 10.1038/nature03991.

[78] William P. Russ et al. "Natural-like function in artificial WW domains". en. In: *Nature* 437.7058 (Sept. 2005), pp. 579–583. DOI: 10.1038/nature03990.

[79] Michael J. Harms and Joseph W. Thornton. "Evolutionary biochemistry: revealing the historical and physical causes of protein properties". en. In: *Nature Reviews Genetics* 14.8 (Aug. 2013), pp. 559–571. DOI: 10.1038/nrg3540.

[80] Faruck Morcos et al. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". en. In: *Proceedings of the National Academy of Sciences* 108.49 (Dec. 2011). DOI: 10.1073/pnas.1111471108.

[81] Simona Cocco et al. “Inverse statistical physics of protein sequences: a key issues review”. en. In: *Reports on Progress in Physics* 81.3 (Mar. 2018), p. 032601. DOI: 10.1088/1361-6633/aa9965.

[82] Ernst Ising. “Beitrag zur Theorie des Ferromagnetismus”. de. In: *Zeitschrift für Physik* 31.1 (Feb. 1925), pp. 253–258. DOI: 10.1007/BF02980577.

[83] Francesco Chippari, Marco Picco, and Raoul Santachiara. *Two-dimensional Ising and Potts model with long-range bond disorder: a renormalization group approach*. arXiv:2306.01887 [cond-mat, physics:hep-th]. Aug. 2023.

[84] E. T. Jaynes. “Information Theory and Statistical Mechanics”. en. In: *Physical Review* 106.4 (May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[85] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. “Learning protein constitutive motifs from sequence data”. en. In: *eLife* 8 (Mar. 2019), e39397. DOI: 10.7554/eLife.39397.

[86] Jeanne Trinquier et al. “Efficient generative modeling of protein sequences using simple autoregressive models”. en. In: *Nature Communications* 12.1 (Oct. 2021), p. 5800. DOI: 10.1038/s41467-021-25756-4.

[87] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. “How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins?” en. In: *Molecular Biology and Evolution* 35.4 (Apr. 2018), pp. 1018–1027. DOI: 10.1093/molbev/msy007.

[88] Magnus Ekeberg et al. “Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models”. en. In: *Physical Review E* 87.1 (Jan. 2013), p. 012707. DOI: 10.1103/PhysRevE.87.012707.

[89] Carlo Baldassi et al. “Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners”. en. In: *PLoS ONE* 9.3 (Mar. 2014). Ed. by Kay Hamacher, e92721. DOI: 10.1371/journal.pone.0092721.

[90] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. “A Learning Algorithm for Boltzmann Machines*”. en. In: *Cognitive Science* 9.1 (Jan. 1985), pp. 147–169. DOI: 10.1207/s15516709cog0901_7.

[91] Anna Paola Muntoni et al. “adabmDCA: adaptive Boltzmann machine learning for biological sequences”. en. In: *BMC Bioinformatics* 22.1 (Dec. 2021), p. 528. DOI: 10.1186/s12859-021-04441-9.

[92] I.N. Shindyalov, N.A. Kolchanov, and C. Sander. "Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?" en. In: *"Protein Engineering, Design and Selection"* 7.3 (1994), pp. 349–358. DOI: 10.1093/protein/7.3.349.

[93] Alan S. Lapedes et al. "Correlated mutations in models of protein sequences: phylogenetic and structural effects". en. In: *Institute of Mathematical Statistics Lecture Notes - Monograph Series*. Hayward, CA: Institute of Mathematical Statistics, 1999, pp. 236–256. DOI: 10.1214/lnms/1215455556.

[94] Joanna I. Sułkowska et al. "Genomics-aided structure prediction". en. In: *Proceedings of the National Academy of Sciences* 109.26 (June 2012), pp. 10340–10345. DOI: 10.1073/pnas.1207864109.

[95] S.D. Dunn, L.M. Wahl, and G.B. Gloor. "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction". en. In: *Bioinformatics* 24.3 (Feb. 2008), pp. 333–340. DOI: 10.1093/bioinformatics/btm604.

[96] Andrew W. Senior et al. "Improved protein structure prediction using potentials from deep learning". en. In: *Nature* 577.7792 (Jan. 2020), pp. 706–710. DOI: 10.1038/s41586-019-1923-7.

[97] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". en. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

[98] Matteo Figliuzzi et al. "Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1". en. In: *Molecular Biology and Evolution* 33.1 (Jan. 2016), pp. 268–280. DOI: 10.1093/molbev/msv211.

[99] William F. Flynn et al. "Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease". en. In: *Molecular Biology and Evolution* 34.6 (June 2017), pp. 1291–1306. DOI: 10.1093/molbev/msx095.

[100] Thomas A Hopf et al. "Mutation effects predicted from sequence co-variation". en. In: *Nature Biotechnology* 35.2 (Feb. 2017), pp. 128–135. DOI: 10.1038/nbt.3769.

[101] Christoph Feinauer and Martin Weigt. *Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease*. arXiv:1701.07246 [q-bio]. Jan. 2017.

[102] Jonathan Frazer et al. "Disease variant prediction with deep generative models of evolutionary data". en. In: *Nature* 599.7883 (Nov. 2021), pp. 91–95. DOI: 10.1038/s41586-021-04043-8.

[103] Pengfei Tian et al. “Design of a Protein with Improved Thermal Stability by an Evolution-Based Generative Model”. en. In: *Angewandte Chemie International Edition* 61.50 (Dec. 2022), e202202711. DOI: 10.1002/anie.202202711.

[104] Ali Madani et al. “Large language models generate functional protein sequences across diverse families”. en. In: *Nature Biotechnology* 41.8 (Aug. 2023), pp. 1099–1106. DOI: 10.1038/s41587-022-01618-2.

[105] Sophia Alvarez et al. *In vivo functional phenotypes from a computational epistatic model of evolution*. en. preprint. Evolutionary Biology, May 2023. DOI: 10.1101/2023.05.24.542176.

[106] Cyril Malbranke et al. “Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies”. en. In: *Current Opinion in Structural Biology* 80 (June 2023), p. 102571. DOI: 10.1016/j.sbi.2023.102571.

[107] Donatas Repecka et al. “Expanding functional protein sequence spaces using generative adversarial networks”. en. In: *Nature Machine Intelligence* 3.4 (Mar. 2021), pp. 324–333. DOI: 10.1038/s42256-021-00310-5.

[108] Alex Hawkins-Hooker et al. “Generating functional protein variants with variational autoencoders”. en. In: *PLOS Computational Biology* 17.2 (Feb. 2021). Ed. by Christine A. Orengo, e1008736. DOI: 10.1371/journal.pcbi.1008736.

[109] Matteo Bisardi et al. “Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution”. In: *Molecular Biology and Evolution* 39.1 (Jan. 2022), msab321. DOI: 10.1093/molbev/msab321.

[110] Michael A. Stiffler et al. “Protein Structure from Experimental Evolution”. en. In: *Cell Systems* 10.1 (Jan. 2020), 15–24.e5. DOI: 10.1016/j.cels.2019.11.008.

[111] P. Alexander Gunnarsson and M. Madan Babu. “Predicting evolutionary outcomes through the probability of accessing sequence variants”. en. In: *Science Advances* 9.30 (July 2023), eade2903. DOI: 10.1126/sciadv.ade2903.

[112] Charles Geyer. “Introduction to Markov Chain Monte Carlo”. en. In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks et al. Vol. 20116022. Chapman and Hall/CRC, May 2011. DOI: 10.1201/b10905-2.

[113] Sebastien Wolf et al. “Emergence of time persistence in a data-driven neural network model”. en. In: *eLife* 12 (Mar. 2023), e79541. DOI: 10.7554/eLife.79541.

[114] Xiaowen Chen et al. *Modelling collective behavior in groups of mice housed under semi-naturalistic conditions*. en. preprint. Animal Behavior and Cognition, July 2023. DOI: 10.1101/2023.07.26.550619.

[115] Simone Savino, Tom Desmet, and Jorick Franceus. "Insertions and deletions in protein evolution and engineering". en. In: *Biotechnology Advances* 60 (Nov. 2022), p. 108010. DOI: 10.1016/j.biotechadv.2022.108010.

[116] Jose Alberto De La Paz et al. "Epistatic contributions promote the unification of incompatible models of neutral molecular evolution". en. In: *Proceedings of the National Academy of Sciences* 117.11 (Mar. 2020), pp. 5873–5882. DOI: 10.1073/pnas.1913071117.

[117] Juan Rodriguez-Rivas et al. "Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes". en. In: *Proceedings of the National Academy of Sciences* 119.4 (Jan. 2022), e2113118119. DOI: 10.1073/pnas.2113118119.

[118] Alexey S. Kondrashov, Shamil Sunyaev, and Fyodor A. Kondrashov. "Dobzhansky–Muller incompatibilities in protein evolution". en. In: *Proceedings of the National Academy of Sciences* 99.23 (Nov. 2002), pp. 14878–14883. DOI: 10.1073/pnas.232565499.

[119] Noor Youssef et al. "Shifts in amino acid preferences as proteins evolve: A synthesis of experimental and theoretical work". en. In: *Protein Science* 30.10 (Oct. 2021), pp. 2009–2028. DOI: 10.1002/pro.4161.

[120] Michael B. Doud, Orr Ashenberg, and Jesse D. Bloom. "Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs". en. In: *Molecular Biology and Evolution* 32.11 (Nov. 2015), pp. 2944–2960. DOI: 10.1093/molbev/msv167.

[121] Hugh K Haddox et al. "Mapping mutational effects along the evolutionary landscape of HIV envelope". en. In: *eLife* 7 (Mar. 2018), e34420. DOI: 10.7554/eLife.34420.

[122] Barrett Steinberg and Marc Ostermeier. "Shifting Fitness and Epistatic Landscapes Reflect Trade-offs along an Evolutionary Pathway". en. In: *Journal of Molecular Biology* 428.13 (July 2016), pp. 2730–2743. DOI: 10.1016/j.jmb.2016.04.033.

[123] Yeonwoo Park, Brian P. H. Metzger, and Joseph W. Thornton. "Epistatic drift causes gradual decay of predictability in protein evolution". en. In: *Science* 376.6595 (May 2022), pp. 823–830. DOI: 10.1126/science.abn6895.

[124] Yvonne H. Chan et al. “Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints”. en. In: *Nature Communications* 8.1 (Mar. 2017), p. 14614. DOI: 10.1038/ncomms14614.

[125] Victor H Salinas and Rama Ranganathan. “Coevolution-based inference of amino acid interactions underlying protein function”. en. In: *eLife* 7 (July 2018), e34300. DOI: 10.7554/eLife.34300.

[126] Michael Heyne et al. “Climbing Up and Down Binding Landscapes through Deep Mutational Scanning of Three Homologous Protein–Protein Complexes”. en. In: *Journal of the American Chemical Society* 143.41 (Oct. 2021), pp. 17261–17275. DOI: 10.1021/jacs.1c08707.

[127] Louisa Gonzalez Somermeyer et al. “Heterogeneity of the GFP fitness landscape and data-driven protein design”. en. In: *eLife* 11 (May 2022), e75842. DOI: 10.7554/eLife.75842.

[128] Lucile Vigué et al. “Deciphering polymorphism in 61,157 Escherichia coli genomes via epistatic sequence landscapes”. en. In: *Nature Communications* 13.1 (July 2022), p. 4030. DOI: 10.1038/s41467-022-31643-3.

[129] John Zhongshi Chen. “Mutational scanning of metallo-beta-lactamases to probe functional determinants, selection pressure dependence and homolog incompatibilities”. en. In: (2022). Publisher: University of British Columbia Version Number: 1. DOI: 10.14288/1.0417580.

[130] Michael Baym, Laura K. Stone, and Roy Kishony. “Multidrug evolutionary strategies to reverse antibiotic resistance”. en. In: *Science* 351.6268 (Jan. 2016), aad3292. DOI: 10.1126/science.aad3292.

[131] David Alvarez-Ponce. “Richard Dickerson, Molecular Clocks, and Rates of Protein Evolution”. en. In: *Journal of Molecular Evolution* 89.3 (Apr. 2021), pp. 122–126. DOI: 10.1007/s00239-020-09973-x.

[132] Wen Jun Xie, Mojgan Asadi, and Arieh Warshel. “Enhancing computational enzyme design by a maximum entropy strategy”. en. In: *Proceedings of the National Academy of Sciences* 119.7 (Feb. 2022), e2122355119. DOI: 10.1073/pnas.2122355119.

[133] Carlos A. Gandarilla-Pérez et al. “Combining phylogeny and coevolution improves the inference of interaction partners among paralogous proteins”. en. In: *PLOS Computational Biology* 19.3 (Mar. 2023). Ed. by Andrea Ciliberto, e1011010. DOI: 10.1371/journal.pcbi.1011010.

[134] Luca Sesta et al. "AMaLa: Analysis of Directed Evolution Experiments via Annealed Mutational Approximated Landscape". en. In: *International Journal of Molecular Sciences* 22.20 (Oct. 2021), p. 10908. DOI: 10.3390/ijms222010908.

[135] Eugenio Mauri, Simona Cocco, and Rémi Monasson. "Mutational Paths with Sequence-Based Models of Proteins: From Sampling to Mean-Field Characterization". en. In: *Physical Review Letters* 130.15 (Apr. 2023), p. 158402. DOI: 10.1103/PhysRevLett.130.158402.

[136] Eugenio Mauri, Simona Cocco, and Rémi Monasson. "Transition paths in Potts-like energy landscapes: General properties and application to protein sequence models". en. In: *Physical Review E* 108.2 (Aug. 2023), p. 024141. DOI: 10.1103/PhysRevE.108.024141.

[137] Christoph Dellago, Peter G. Bolhuis, and Phillip L. Geissler. "Transition Path Sampling". en. In: *Advances in Chemical Physics*. Ed. by I. Prigogine and Stuart A. Rice. 1st ed. Vol. 123. Wiley, July 2002, pp. 1–78. DOI: 10.1002/0471231509.ch1.

[138] Thierry Mora, Aleksandra M. Walczak, and Francesco Zamponi. "Transition path sampling algorithm for discrete many-body systems". en. In: *Physical Review E* 85.3 (Mar. 2012), p. 036710. DOI: 10.1103/PhysRevE.85.036710.

LONG RÉSUMÉ

Grâce à l'afflux de données de séquences protéiques disponibles grâce au séquençage de nouvelle génération, les modèles d'apprentissage automatique non supervisé peuvent désormais apprendre efficacement les paysages de séquences protéiques. Parmi les outils, Direct Coupling Analysis (DCA) se distingue car elle évalue les motifs de conservation et de coévolution entre les sites protéiques. Le DCA a été utilisé pour divers défis biologiques, allant de la prédiction des effets des mutations sur la fitness à la génération de séquences artificielles. Cette thèse étend l'utilité du DCA pour sonder l'évolution des protéines, qui fait référence aux changements continus dans les séquences d'acides aminés dus aux mutations aléatoires et à la sélection naturelle. Comprendre les subtilités de ce processus dynamique est crucial pour des défis tels que la lutte contre les pathogènes émergents, la compréhension de l'histoire de la vie sur Terre, et même pour la conception d'enzymes artificielles. Malgré sa nature omniprésente, prédire l'évolution des protéines reste un défi, surtout lorsqu'on considère l'épistasie, les effets contextuels des mutations. Notre travail vise à développer et évaluer des algorithmes pour traverser le paysage de séquences déduit par DCA, mettant l'accent sur les familles d'enzymes bêta-lactamases, notoires pour leur capacité à dégrader de nombreux antibiotiques couramment utilisés.

Le premier chapitre de recherche, intitulé " Exploration of sequence space ", se penche sur la modélisation de l'évolution des protéines. Ici, nous explorons d'abord l'évolution des protéines par évolution expérimentale à travers la dérive neutre. Cette approche implique plusieurs cycles de mutation et de sélection de bibliothèques géniques évoluées in vivo. L'objectif est de diversifier la bibliothèque génique tout en préservant sa fonction, même sous de faibles pressions de sélection. Nous présentons un premier article discutant d'un modèle évolutionnaire qui reproduit la diversification des enzymes résistantes aux antibiotiques dans des expériences de dérive génétique : "Modeling Sequence-Space Exploration and Emergence of Epistatic Signals in Protein Evolution". Les modèles reproduisent avec précision les statistiques expérimentales, en utilisant un paysage de séquences d'homologues éloignés et un algorithme d'échantillonnage qui prend en compte les contraintes du code génétique. De plus, nous analysons computationnellement l'impact des paramètres expérimentaux, identifiant un équilibre entre les tours expérimentaux et la profondeur de séquençage lors du décryptage des contraintes épistatiques. Ce travail démontre le potentiel des modèles évolutionnaires dans l'analyse des paysages de séquences basés sur des données. En utilisant des

dynamiques stochastiques simples qui capturent la relation entre mutation et sélection, ces modèles peuvent dépeindre avec précision les résultats des expériences d'évolution. Les modèles fournissent des applications pratiques pour l'évolution expérimentale, comme l'optimisation des protocoles pour obtenir les résultats de séquence désirés. Le modèle actuel est considéré comme fondamental, avec des possibilités d'amélioration comme l'incorporation de biais de mutation, de biais de codon, et d'autres subtilités biologiques.

Nous présentons ensuite des améliorations à notre méthodologie, en nous concentrant sur les biais mutationnels dans les substitutions nucléotidiques et en mettant à jour la dynamique d'échantillonnage avec un équilibre détaillé. Nous appelons ce nouveau modèle évolutionnaire pour les séquences protéiques GENIE (Gene Evolution on Nucleotides Including Epistasis), qui se concentre sur les séquences nucléotidiques tout en considérant les effets mutationnels au niveau des acides aminés. Il propose deux types de méthodes de mutation : les mouvements de Gibbs pour les substitutions nucléotidiques et les mouvements de Metropolis pour les insertions ou délétions. Cela nous permet de reproduire avec précision les statistiques d'acides aminés de protéines lointainement apparentées. Un aspect essentiel de GENIE est qu'il nous permet de générer de manière plausible des trajectoires évolutionnaires pour nous aider à comprendre les échelles de temps évolutionnaires pilotées par l'épistasie et leur effet sur la mutabilité des sites au fil du temps. Grâce à nos méthodes, nous pouvons modéliser quantitativement la prévisibilité des effets mutationnels à mesure que les séquences divergent, l'épistasie étant la force motrice. Nous capturons l'épistasie en analysant les différences de variabilité locale (Context-Dependent Entropy, CDE) et globale (Context Independent Entropy, CIE) dans les sites protéiques. Les sites sont classés comme variables, conservés ou épistatiques, avec des comportements de mutation distincts observés dans chacun pendant les simulations. Les sites variables mutent rapidement, les sites conservés lentement, et les sites épistatiques restent stables jusqu'à ce que leur contexte change considérablement, puis mutent rapidement. La recherche met en évidence une séparation à l'échelle du temps dans les taux de mutation, avec une diversification rapide dans des contextes spécifiques et des changements de contexte évolutionnaire plus lents.

Dans la section suivante, intitulée "Family-Wide Mutational Incompatibilities", nous présentons un deuxième article : "Understanding epistatic networks in B1 beta-lactamases through coevolutionary statistical modeling and deep mutational scanning ". Tout au long de l'évolution, les protéines maintiennent leur structure et leur fonction malgré des changements de séquence significatifs, grâce à des réseaux interconnectés d'interactions épistatiques qui réduisent les impacts négatifs des mutations. En analysant la classe B1 des enzymes bêta-lactamase à l'aide du DCA et du Deep Mutational Scanning

(DMS), nous identifions des clusters coévolutifs de sites protéiques qui sont interconnectés par des motifs de conservation et de coévolution. Nous interprétons ces motifs comme des contraintes structurelles et fonctionnelles, ainsi que des incompatibilités mutationnelles. Les bêta-lactamases B1 sont un groupe diversifié d'enzymes qui évoluent rapidement pour dégrader un large éventail d'antibiotiques. Cette diversité est possible grâce à des mutations bénéfiques et neutres qui parcourent les séquences et les structures protéiques. Les clusters coévolutifs identifiés sont interconnectés par des motifs de conservation et de coévolution, qui se reflètent dans les structures et les dynamiques de repliement protéique. Ces clusters sont probablement des unités fonctionnelles dans les bêta-lactamases qui travaillent ensemble pour conférer une résistance aux antibiotiques. En utilisant le DCA, nous modélisons les contraintes de séquence et de structure dans ces clusters et identifions des motifs coévolutifs qui ne sont pas immédiatement apparents dans les analyses de séquences simples. Le DMS nous fournit une carte de l'impact fonctionnel des mutations à travers les séquences de bêta-lactamase, nous montrant comment les mutations influencent la fitness et l'activité enzymatique. En combinant les résultats du DCA et du DMS, nous identifions des zones de la protéine qui sont plus tolérantes aux mutations et d'autres zones qui sont essentielles pour la fonction et la stabilité. Cette recherche fournit des informations sur les contraintes évolutionnaires dans les bêta-lactamases et peut aider à informer la conception d'inhibiteurs et d'enzymes modifiées.

La dernière section de cette thèse résume nos principaux résultats et leur impact sur le champ de la biologie évolutive. Nous soulignons comment notre recherche contribue à une meilleure compréhension de l'évolution des protéines et offre de nouvelles méthodes pour modéliser et analyser les paysages de séquences protéiques. Nous discutons également des implications pratiques de notre travail pour la conception de protéines, la pharmacologie et la médecine. En conclusion, cette thèse souligne le potentiel des méthodes basées sur le DCA pour prédire et comprendre l'évolution des protéines, et offre une vision nouvelle sur les mécanismes sous-jacents qui régissent les changements de séquence au fil du temps.

COLOPHON

This document was written with LaTeX on Mac using ArsClassica, a reworking of the classicthesis v4.6 style designed by André Miede, inspired by the masterpiece *The Elements of Typographic Style* by Robert Bringhurst.

Final Version as of October 30, 2023