

# Proteins: the physics of amorphous evolving matter

Jean-Pierre Eckmann <sup>1,2</sup>, Jacques Rougemont<sup>1</sup>, Tsvi Tlusty <sup>3,4</sup>

<sup>1</sup> *Département de Physique Théorique,  
Université de Genève,  
CH-1211, Geneva 4,  
Switzerland*

<sup>2</sup> *Section de Mathématiques,  
Université de Genève,  
CH-1211, Geneva 4,  
Switzerland*

<sup>3</sup> *Center for Soft and Living Matter,  
Institute for Basic Science (IBS), Ulsan 44919,  
Korea*

<sup>4</sup> *Department of Physics,  
Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919,  
Korea*

Proteins are a matter of dual nature. As a physical object, a protein molecule is a folded chain of amino acids with multifarious biochemistry. But it is also an instantiation along an evolutionary trajectory determined by the function performed by the protein within a hierarchy of interwoven interaction networks of the cell, the organism and the population. A physical theory of proteins therefore needs to unify both aspects, the biophysical and the evolutionary. Specifically, it should provide a model of how the DNA gene is mapped into the functional phenotype of the protein.

We review several physical approaches to the protein problem, focusing on a mechanical framework which treats proteins as evolvable condensed matter: Mutations introduce localized perturbations in the gene, which are translated to localized perturbations in the protein matter. A natural tool to examine how mutations shape the phenotype are Green's functions. They map the evolutionary linkage among mutations in the gene (termed epistasis) to cooperative physical interactions among the amino acids in the protein. We discuss how the mechanistic view can be applied to examine basic questions of protein evolution and design.

Sections marked with \* contain more technical material and can be omitted at first reading.

## CONTENTS

I. The protein problem: a theoretical physics perspective	2
II. Biology as a challenge to theorists	4
III. Proteins as information machines	5
A. Handling reading errors	6
B. Folding	7
IV. Mechanical views on protein evolution	8
V. Condensed-matter theory of proteins	8
A. Lattice models	9
B. The lattice Laplacian	11
C. Hooke's law	13
VI. Simulating evolution	13
VII. Green's function as a link between the theory of amorphous solids and living matter	13
A. Woodbury's formula	14
B. Dyson's formula	14
VIII. Models: Protein as an evolving machine	14
A. A model with very simple structure (Cylinder-model)	15
1. The cylindrical amino acid network	16
2. Evolution searches for a mechanical function	16
3. Rigidity propagation algorithm*	16
4. Fitness and mutations	17
5. Simulation of evolutionary dynamics	18
B. A model with more realistic interactions (HP-model)	19
1. Pinching the network	20
2. The protein backbone	20
3. Pathologies and broken networks*	21
IX. Connecting the models to biological concepts	22
A. Dimension of the solution set in the genotype and phenotype spaces	22
B. Expansion of the protein universe	24
C. Spectrum in phenotype and genotype spaces	24
1. Geometry of the genotype and phenotype solution spaces*	26
D. Stability of the mechanical phenotype under mutations	27
E. Shear modes in the amino acid network	27
1. Implementation in the case of protein structure data*	29
2. Details of shear computation*	30
F. Similarity of gene and shear	30
G. Point mutations are localized mechanical perturbations	30
H. Mechanical function emerges as a sharp transition	30
I. Correlation and alignment	31
J. Conserved amino acids	32
K. Epistasis links protein mechanics to genetic correlations	33
L. Epistasis as a sum over scattering paths	35
M. Multilocus epistasis*	37
X. Outlook	38
Acknowledgments	38
References	38

## I. THE PROTEIN PROBLEM: A THEORETICAL PHYSICS PERSPECTIVE

The macromolecules that make living matter – lipids, hydrocarbons, nucleic acids, and in particular proteins – are among the most studied objects of Nature. Proteins comprise the central nano-machinery of the cell, whose numerous functions include the formation of structural elements, catalyzing metabolic reactions and conveying biochemical signals [Alberts 1998; Fersht 1999; Goodsell 2009; Howard 2001; Whitford 2013]. For their significance in life, proteins and the genes that encode them have been

extensively investigated using various experimental methods, such as crystallography, biochemical assays, mass spectrometry, fluorescence imaging, electron microscopy, directed evolution and deep sequencing [Barrera and Robinson 2011; Chapman et al. 2011; Cohen and Chait 2001; Collins et al. 2011; Fernandez-Leiro and Scheres 2016; Ha and Tinnefeld 2012; Mandala et al. 2018; Mardis 2013; Mehmood et al. 2015; Rambo and Tainer 2013]. In parallel, sophisticated computational models, such as molecular dynamics, have been developed to predict the structure, function and folding of proteins [Adcock and McCammon 2006; Dror et al. 2018; Isralewitz et al. 2001; Karplus and Kuriyan 2005; Karplus and McCammon 2002; Scheraga et al. 2018].

These experiments and simulations provide valuable data on protein structure, dynamics and genetics. However, there remain two inherent challenges: (i) Sparsity of data – the protein is the outcome of long evolutionary search in a high-dimensional space of gene sequences, which is impossible to sample, even by high-throughput experiments. (ii) Complexity of interactions – The function of a protein arises from collective many-body interactions in the heterogeneous amino acid matter, which are hard to probe and model.

In light of these challenges, we focus in this colloquium on a complementary theoretical approach that links the protein problem to the realm of condensed matter physics. Rather than using realistic simulations predicting the dynamics and function of concrete proteins, we shall discuss minimal models that allow, under several simplifying assumptions, to examine basic questions of protein evolution, especially how the collective physical interactions within the protein direct its evolution.<sup>1</sup>

The structure of many proteins is known at a resolution of a few angstroms and there are detailed computational models of the forces between the amino acids. Here, however, the protein will be examined at a coarse grained level, in the spirit of lattice [Lau and Dill 1989; Shakhnovich et al. 1991] and network [Chennubhotla et al. 2005] models. The protein will be described as a connected network whose nodes represent the amino acids. Furthermore, from this conceptual point of view, it suffices to assume that there are only two types of amino acids instead of the usual twenty.<sup>2</sup> (for example the classical HP model [Lau and Dill 1989] used in Sec. VIII.B, in which amino acids can only be either hydrophobic (H) or polar (P)). In a real protein, forces between the amino acids are a complicated combination depending for example on their polarity, hydrophobicity, charge and shape. At the coarse grained level, it again suffices to consider instead just simple springs between pairs of neighboring sites. This is akin to using harmonic approximations in mechanics, which provide a generic understanding and a good physical insight.

With this kind of simplifications, one can translate certain questions of biology to analogous questions in the physics of amorphous networks. Among the rich set of methods in this classical subject of physics, some tools seem particularly well adapted to the protein problem. The approach is based on the dual nature of the protein; it is a physical object whose formation and physical interactions are also represented in the ‘dual’ gene, a sequence of symbols from a four-letter alphabet of the DNA bases, ‘A’, ‘C’, ‘G’, ‘T’. Evolution progresses by introducing mutations, that is, permanent modifications of this sequence. There are local mutations (nucleotide substitutions, short insertions and deletions) besides larger scale modifications (*e.g.*, translocations, inversions, duplications). A natural approach to study protein evolution is to model the effect of mutations on the physical properties of the amino acids network.

Local mutations amount to short jumps between neighboring sequences in the genotype space, differing by one letter only, while large-scale mutations are equivalent to longer jumps. Both classes of mutations can be described in terms of alterations of the mechanical properties of the amino acid network. However, we shall focus on the class of local mutations. Practically, local mutations are easy to treat with classical techniques of condensed matter, for instance via Green’s functions, since they induce localized perturbations in the spring network. More importantly, it is possible to statistically sample the genotype space with continuous trajectories progressing by consecutive local mutations. This will be the main axis of this colloquium. Along the evolutionary trajectory, mutations come in three flavors: The ones leading to some sort of functional catastrophe or significant disadvantage, and therefore get eliminated by selection; others which improve the properties of a protein and finally, the large ‘neutral’ majority which do not induce any significant change in the function of the protein [Kimura 1983; Neher and Shraiman 2011]. In this manner, the ‘learning’ evolutionary process reduces the problem of improving a protein from an exhaustive combinatorial search approach into a biased random walk. This drastically reduces the dimension of the space which one needs to explore.

The condensed-matter approach to the protein problem may be viewed as an example of a potentially general framework that may be used to examine other strongly-coupled biological systems. For example, one may analyze metabolic and genetic networks in terms of localized perturbations and Green’s functions. Such analysis may suggest common underlying principles. It might as well turn out that biology is more contingent and depends on the history of the evolutionary process, but at least the few examples we describe give us hope that a rational approach, based on the laws of physics, may be useful in some cases.

Biological molecules are far from being the spring networks we use as a model. Still, similar abstractions proved successful in many areas of physics. For example, the dynamical systems of the so-called Axiom A class [Eckmann and Ruelle 1985] are

<sup>1</sup> The main body of this colloquium is based on, and expands, ideas from papers [Dutta et al. 2018; Eckmann 2008; Tlusty 2007a; 2008b; 2010; 2016; Tlusty et al. 2017].

<sup>2</sup> 21, when counting the rare pyrrolysine [Hao et al. 2002; Srinivasan et al. 2002].

systems with a very special, yet simple, structure. And although most systems do not belong to the Axiom A class, it proved very useful to consider that they behave ‘as if’.

There is a long history of studies in similar spirit of abstraction and simplification, starting with the conformal maps of D’Arcy Thompson [Thompson 1942], through the morphogenetic studies based on the theory of catastrophes by René Thom [Thom 2018]. In the 21st century researchers have much more data available on biological systems. This allows to test hypotheses against measurements, infer from the data other questions to investigate, and suggest possible experiments to confirm or refute the theory. We close the introduction by two citations which reflect the general outlook of this colloquium:

**Misha Gromov**, in [Gromov 2013, Abstract]

When you read a textbook on molecular/cellular biology you are enchanted by the logical beauty of biological structures. You want to share your excitement with your colleagues, but... you find out you are unable to do it: there is no language in the 21st century mathematics that can express this beauty. You feel there must be a new world of mathematical structures shadowing what we see in Life, a new language we do not know yet, something in the spirit the ‘language’ of calculus we use when describing physical systems.

**Giovanni Jona-Lasinio**, in [Jona-Lasinio 2012]:

Theoretical physics was recognized as an independent field of research only at the end of the 19th century, shortly before the great conceptual revolutions of relativity and quantum mechanics. Today theoretical physics has multiple facets. I think that the time has come for a more precise characterization of the research field of theoretical biology, and for an assessment of its scope. [Translated from Italian]

We are convinced that such outlooks are important and our work should be viewed as an attempt in this general direction, in the hope that readers will be encouraged to proceed along this path.

## II. BIOLOGY AS A CHALLENGE TO THEORISTS

Biological research has been extremely active in the past decades and experimental results have flourished to vastly improve our understanding of living matter. The challenge for theorists is to find subtopics which are at a stage where theoretical abstraction can be fruitful.

Here we focus on the relation between genes and the functions of proteins: genes (in DNA) code for amino acid chains that fold into the three-dimensional configurations of functional proteins. This sequence-to-function map is hard to decrypt since it links the collective physical interactions inside the protein to the corresponding evolutionary forces acting on the genome [Dill and MacCallum 2012; Koonin et al. 2002; Liberles et al. 2012; Xia and Levitt 2004; Zeldovich and Shakhnovich 2008]. Furthermore, evolution selects the tiny fraction of functional sequences in an enormous, high-dimensional space [Keefe and Szostak 2001; Koehl and Levitt 2002; Povolotskaya and Kondrashov 2010], which implies that proteins form non-generic, information-rich matter, outside the scope of standard statistical methods. Therefore, although the structure and physical forces within a protein have been extensively studied, the fundamental question of how a functional protein originates from a linear DNA sequence still provides research challenges, in particular how functionality constrains the accessible DNA sequences.

To examine the geometry of the sequence-to-function map, we devise below a mechanical model of proteins as amorphous evolving matter.<sup>3</sup> Rather than simulating concrete proteins, we construct models which describe the hallmarks of the genotype-to-phenotype map (the translation of the gene to the protein). These models are sufficiently simple so that large-scale simulations can be performed, which allow to average over stochastic noise inherent to evolutionary dynamics. Furthermore, we restrict our approach to models in which the function of a protein arises from large-scale conformational changes, where big chunks of the protein move with respect to each other. These motions are central to certain functions [Henzler-Wildman et al. 2007; Huse and Kuriyan 2002; Koshland 1958; Savir and Tlusty 2007; 2010; 2013; Schmeing et al. 2009]. For example, allosteric proteins are a type of ‘mechanical transducers’ that transmit regulatory signals between distant sites [Ferreon et al. 2013; Goodey and Benkovic 2008; Lockless and Ranganathan 1999; Perutz 1970].

We end this section by mentioning a few papers which have dealt with similar issues, and which highlight the increasing interest in connecting biological questions with methods from solid state physics.

Common to these studies is a mechanical perspective on protein function. The motivation originates from many observations of proteins whose functions involve collective patterns of forces and coordinated displacements of their amino acids [Boehr et al.

---

<sup>3</sup> In his book “What is Life?” [Schrödinger 1944], Schrödinger uses the term ‘aperiodic crystal’ to describe material which contains genetic information. This is of course a very interesting forethought, but since the advent of quasiperiodic crystals, the term ‘amorphous’ leads to a more precise classification.

2006; Bustamante et al. 2004; Daniel et al. 2003; Eisenmesser et al. 2005; Goodey and Benkovic 2008; Hammes-Schiffer and Benkovic 2006; Henzler-Wildman et al. 2007; Huse and Kuriyan 2002; Karplus and McCammon 2002; Savir and Tlusty 2010]. In particular, the mechanisms of allosteric [Cui and Karplus 2008; Daily et al. 2008; Koshland et al. 1966; Monod et al. 1965; Motlagh et al. 2014; Perutz 1970; Thirumalai et al. 2018], induced fit [Koshland 1958], and conformational selection [Grant et al. 2010] often involve global conformational changes by hinge-like rotations, twists or shear-like sliding of protein subdomains [Gerstein et al. 1994; Mitchell and Leibler 2017; Mitchell et al. 2016].

A now-standard approach to examine the link between function and motion is to model proteins as elastic networks of amino acids connected by spring-like bonds. Early studies that apply this class of models are from the 1980s and 90s [Levitt et al. 1985; Tirion 1996], and in the last two decades the methods have been further developed and applied to many proteins [Bahar 2010; Chennubhotla et al. 2005; López-Blanco and Chacón 2016]. Decomposing the dynamics of the network into normal modes revealed that low-frequency ‘soft’ modes capture functionally relevant large-scale motion [Bahar et al. 2010; Haliloglu and Bahar 2015; Tama and Sanejouand 2001], especially in allosteric proteins [Arora and Brooks 2007; Greener and Sternberg 2015; Hawkins and McLeish 2006; Ming and Wall 2005; Tehver et al. 2009; Wrabl et al. 2011; Zheng et al. 2006].

Recent work associates the soft modes of protein conformations with the emergence of weakly connected regions as described above, but also ‘cracks’ [Miyashita et al. 2003], ‘shear bands’ or ‘channels’ [Dutta et al. 2018; Mitchell and Leibler 2017; Mitchell et al. 2016; Rocks et al. 2019; Tlusty 2016; Tlusty et al. 2017] that enable low-energy viscoelastic motion [Joseph et al. 2014; Qu and Zocchi 2013]. Such contiguous domains evolve in models of allosteric proteins [Flechsig 2017; Hemery and Rivoire 2015; Tlusty et al. 2017].

A source of inspiration for linking proteins to the physics of amorphous matter are the papers by the late Shlomo Alexander, especially [Alexander 1998; Alexander and Orbach 1982]. In these works, Alexander highlighted the essential role of ‘floppy modes’ in the mechanical spectrum of amorphous solids. Also relevant are studies by Thorpe and Phillips on constraint theory and rigidity percolation in glasses, such as [Phillips and Thorpe 1985; Thorpe 1985]. Those works highlighted the ability to control the rigidity and accessible zero-energy modes of mechanical networks by balancing the number of degrees of freedom and the number of constraints, as was noted by Maxwell in 1864 [Maxwell 1864].

The link between the dynamical spectra of proteins and amorphous matter has been further explored in a recent series of works on mechanical metamaterials. The emergence of long-range allosteric response was used in [Rocks et al. 2017] as a design principle for ‘programmable’ metamaterial made of amorphous spring networks [Rocks et al. 2018]. A similar random network approach was applied in [Yan et al. 2017] to design elastic materials with tailored mechanical response. These works suggest that tunable amorphous materials have the flexibility required to produce elaborate designs, as recently demonstrated by mimicking the cyclical conformational motion of protein motors [Flechsig and Togashi 2018]. These promising approaches to metamaterial design are discussed elsewhere, for example in [Baardink et al. 2018; Kim et al. 2018; Rocklin 2017; Ronellenfitsch et al. 2018].

The present Colloquium focuses on a different aspect: understanding fundamental properties of the protein evolution – in particular the genotype-to-phenotype map – within the framework of condensed matter theory.

### III. PROTEINS AS INFORMATION MACHINES

The building plan of a protein is determined by its corresponding gene, via the genetic code. The gene is a 1-dimensional string in an alphabet of 4 letters: the nucleotides ‘A’ (adenine), ‘C’ (cytosine), ‘G’ (guanine), and ‘T’ (thymine) (see [Alberts et al. 2002], Ch. 6). The protein is a (folded) chain of amino acids (AA) which is translated from the gene according to the genetic code: each three successive letters (each non-overlapping triplet, called a codon) maps to a single AA. In principle, this would allow for  $4^3$  possibilities, but in general there are only 20 different AA’s, making the code redundant, as we shall discuss in Sec. III.A.

We view the gene, *i.e.*, the 1-dimensional string of letters as the tape of a Turing machine [Condon et al. 2018; Herken 1992; Turing 1936]. Since any alphabet can be recoded in binary (for example, each of the 4 nucleotides can be recoded as a 2-bits number), one can always think of it as a string of ‘0’s and ‘1’s. The proteins (and the transcription-translation machinery, which is itself made of proteins) would be the computer, which is able to read and interpret the string.

This particular machine is an example of a self-reproducing Turing machine [von Neumann 1966], since the replication of the genome can be achieved by genome-encoded proteins. In addition, these machines are evolving when the genes are mutated. In other words, the machine can modify its own tape (see also [Tlusty 2016]). A further study in this direction is [Dyson 1970], but there are many more, see *e.g.*, [Freitag and Merkle 2004].

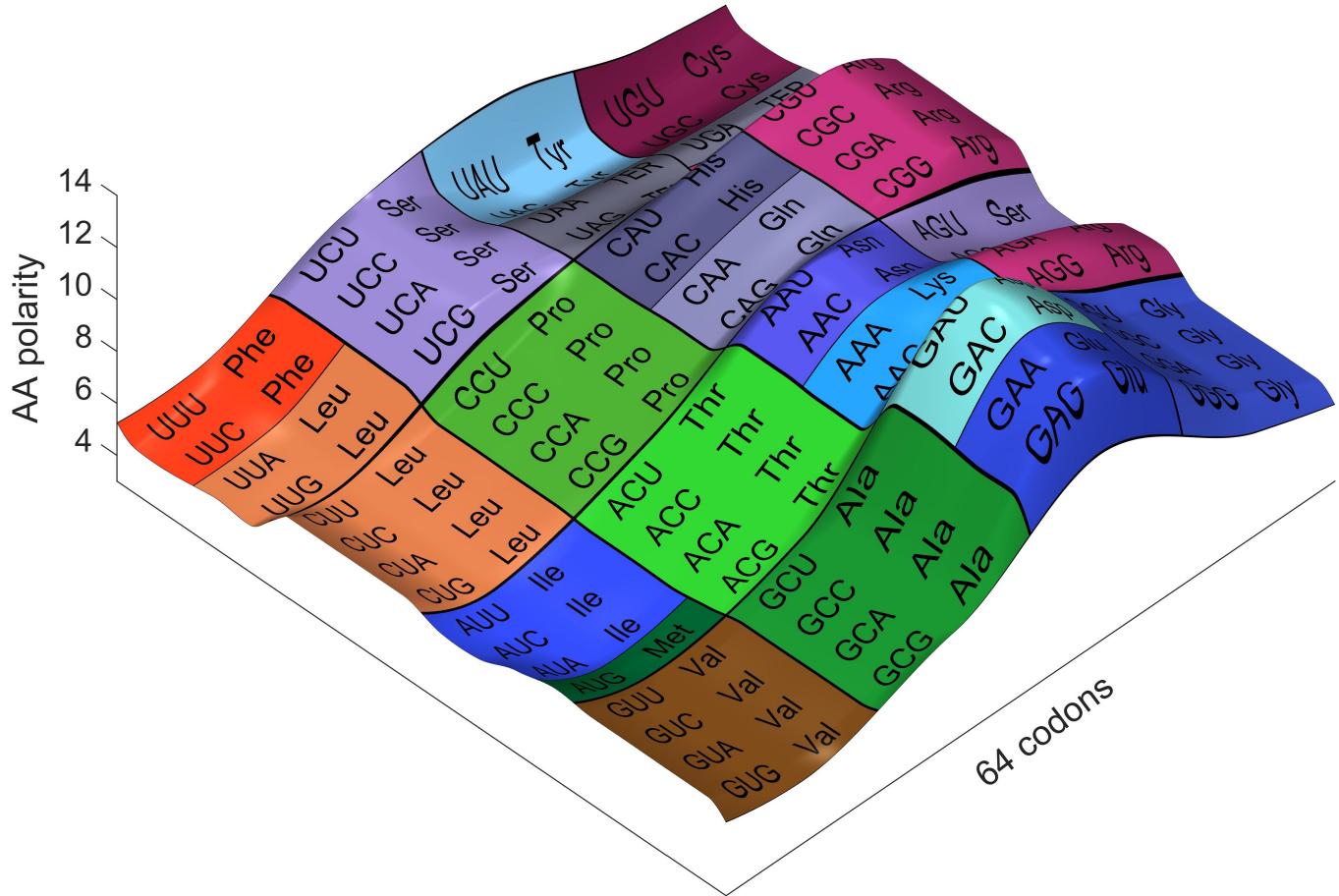


FIG. 1 A representation of the genetic code, as function of the measured polarity of each codon (values from [Haig and Hurst 1991; Woese 1965]). The smoothness of the landscape shows that moving from one AA to its neighbor does not change the polarity too abruptly.

### A. Handling reading errors

Translation of the gene into its corresponding string of amino acids requires a specialized machinery, which includes the ribosome [Alberts et al. 2002].<sup>4</sup> The translation machinery ‘reads’ the code through chemical affinity, and might therefore mis-read the tape. Most amino acids are encoded by more than one codon, and this hard-coded redundancy of the genetic code helps to reduce the impact of such misreadings (see [Tlusty 2007a; 2008a;b;c; 2010] for a theoretical study and [Eckmann 2008] for an illustration).

As noted above this system allows for  $64 = 4^3$  different codons (number of triplets from an alphabet of 4 nucleotides), but they generate only 21 different symbols.<sup>5</sup> The geometric aspects of this arrangement of 21 among 64 possibilities can be understood in graph-theoretical terms: One presents the 64 codons as the nodes of a codon-graph, and two nodes are connected by a link if the corresponding codons differ in only one symbol. Note that swapping ‘C’ and ‘T’ in the codon’s third position always results in the same AA (Fig. 1) and we can therefore reduce the graph to  $48 = 4^2 \cdot 3$  nodes.<sup>6</sup> In the codon-graph, each amino acid is coded as a simply connected region, as shown in Fig. 1, with the exception of Serine (ser) (Arginin (arg) is disconnected in the 2D table, but not in the graph). Such an arrangement minimizes the ratio of surface by area for each region. This reduces the probability of coding the wrong AA, under the assumption that most reading errors involve only one-letter differences.

Additionally, amino acids with similar chemical properties (for example polarity) tend to be neighbors in this graph. This can be visualized by plotting the measured polarity as a function of the codon, which produces a relatively smooth landscape. The

<sup>4</sup> In addition to the ribosome, the machinery includes two sets of molecules, tRNAs, which carry the amino acids, and aminoacyl-tRNA synthetases, which charge the tRNAs with amino acids. The translation is preceded by a transcription step in which the DNA gene (a segment of the genome) is copied into a mRNA (a single molecule).

<sup>5</sup> Some terminology: The individual symbols (A, C, G, T) refer to nucleotides. The triplets of 3 nucleotides form the 64 codons. The 64 codons code for 20 AA and the stop symbol (which does not generate an AA). One of the AA is Methionine (codon ATG) which marks the start of a protein.

<sup>6</sup> This graph is difficult to draw, as each node has  $8 = 3 + 3 + 2$  neighbors which differ in exactly one position. So a representation would have to be in 8 dimensions. Recall that in a cube in 3-dimensions, every corner has 3 neighbors.

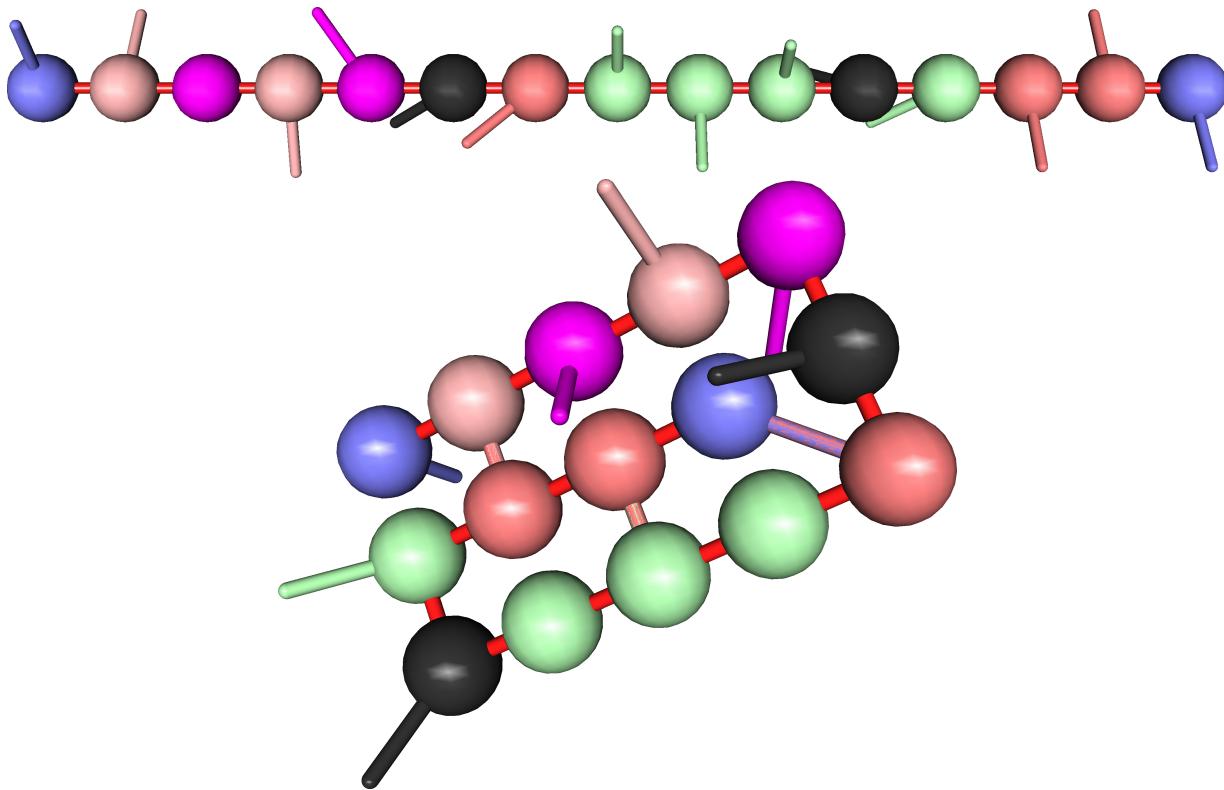


FIG. 2 Schematic illustration of protein folding: Proteins are polypeptides, linear heteropolymers of AAs (colored spheres), linked by covalent peptide bonds (red sticks), which form the protein backbone. These peptide bonds are much stronger than the non-covalent interactions among the AAs (side chains) and do not change when the protein mutates.

smoothness manifests the chemical similarity between neighboring amino acids, and implies that most misreadings change the polarity of AA only moderately. We note that, unlike the 2D landscape of Fig. 1, an ideal representation should wrap the surface so that each AA would have 8 neighbors (and can therefore be embedded only in high dimension).

For the connection between the numbers 21 and 48, an inequality can be given in terms of the genus of the codon-graph [Tlusty 2007a;b; 2010] (this uses results from [Banchoff 1965; Colin de Verdière 1993]). Without going into further detail we conclude: *The optimal code must balance contradicting needs for tolerance to errors (with the smoothness of the mapping between codons and chemical space) and chemical diversity, which is essential for the versatility of protein function.*

## B. Folding

Having translated the gene into a linear chain of amino acids (the backbone, see Fig. 2) via the genetic code (and modulo translation errors), this chain will spontaneously fold into a 3-dimensional shape which gives rise to its function. How this folding proceeds is an important and difficult question, which we shall not address here. Instead we will assume that a certain folding pattern is preserved (see [Petsko and Ringe 2004] for a discussion of these issues). This assumption is practical, as we shall be mostly interested in how the function of the protein changes under point mutations of the gene, *i.e.*, bit flips of the code in the tape. Such mutations often do not seriously affect the overall shape of the protein (see also [Bussemaker et al. 1997]).

We can next model the function of this folded amino acid chain and we will show that there is yet another level of redundancy besides the redundancy of the genetic code and the robustness of the folding. we shall see that there are many mutations which have no effect on performance. *Namely, there is high redundancy in the AA sequences that are mapped to the same or similar enough protein function.* we shall quantify this property in terms of dimension [Eckmann and Ruelle 1992; Grassberger and Procaccia 1983].

#### IV. MECHANICAL VIEWS ON PROTEIN EVOLUTION

Consider a protein interacting with a small molecule. Presence of the latter often induces a conformational change at some distance from the interaction site. One important example is the class of allosteric proteins for which an active site is regulated by binding at another site, resulting in a reconfiguration of the active site. More specifically, we shall examine the role of large-scale, functionally-relevant dynamical modes, and their link to long-range genetic correlations.

Before reviewing the literature on this issue, we illustrate such a mechanical effect on a particular example: human glucokinase (which is involved in sugar metabolism), see Fig. 3. The data were obtained from crystallographic structure of two conformations of that protein: the first (PDB<sup>7</sup> accession 1v4s) corresponds to the binding of glucose to its active site and is compared to the conformation in the absence of glucose (PDB 1v4t) [Kamata et al. 2004].

The backbone, see Sec. VIII.B.2, is shown as a light blue curled tube, and the arrows indicate the displacement from one shape to the other (any Galilean motion between the two is eliminated). The color of the arrows indicates up/down motion relative to a horizontal plane. The red coloring in the twisted tube shows the high shear region separating two low-shear domains that move as rigid bodies (shear calculated by the method of [Mitchell et al. 2016; Rougemont et al. in prep.]).

On a conceptual level, one can simplify the figure as shown in Fig. 4. The protein seems to have a central shear band and two external flaps which perform a rotating motion when a ligand attaches to the protein. This kind of mechanical phenomenology is accessible to the language of physics.

Large-scale motions take part in several basic biological functions and mechanisms. For example, in the induced fit [Koshland 1958] and conformational selection [Bahar et al. 2007; Grant et al. 2010] mechanisms, the presence of a substrate induces reshaping of the enzyme to properly align the catalytic groups in the active site. Such reshaping is a dynamic mechanism of *specific recognition* that allows the selection of a target ligand among similar competing molecules [Savir and Tlusty 2007; 2013]. In *allostery*, reconfiguration of the active site is regulated by binding at a secondary, allosteric site, often via long-range mechanical interactions [Motlagh et al. 2014; Thirumalai et al. 2018]. In this Colloquium, we describe simple physical models for the emergence of these mechanisms via evolutionary tuning of the protein's mechanical response.

Like their dynamic phenotypes, proteins' genotypes (their gene sequences), as explained in Sec. III, are remarkably collective. The history of protein evolution can be traced by gathering evolutionary related proteins in different species (homologous proteins) and aligning their sequences. Genes of these proteins sometimes display long-range correlations [de Juan et al. 2013; Göbel et al. 1994; Halabi et al. 2009; Hopf et al. 2017; Jones et al. 2012; Lockless and Ranganathan 1999; Marks et al. 2011; Poelwijk et al. 2017; Suel et al. 2003; Tesileanu et al. 2015]. The correlations indicate epistasis, the compensatory mutations that take place among residues linked by physical forces or common function. As an example [Rougemont et al. in prep.], consider again glucokinase. We aligned about 120 variants of this molecule and asked where along the gene have mutations preferentially occurred (Fig. 5).

Still, the relationship between sequence correlation, epistasis and selection pressure are not fully understood. As discussed in Sec. I, the two main challenges are the intricacy of the physical forces among the amino acids, and the high dimensionality of the genotype-to-phenotype map [Koonin et al. 2002; Liberles et al. 2012; Povolotskaya and Kondrashov 2010]. These inherent difficulties motivated the development of complementary approaches which utilized simplified coarse-grained models, such as lattice proteins [Lau and Dill 1989; Shakhnovich et al. 1991] or elastic networks [Chennubhotla et al. 2005]. Network and lattice models have been recently used to study the evolution of allostery in proteins and in biologically-inspired allosteric matter [Flechsig 2017; Hemery and Rivoire 2015; Rocks et al. 2017; Tlusty 2016; Tlusty et al. 2017; Yan et al. 2017]. Our aim here is different: to construct a simplified condensed-matter model in terms of how the mechanical interactions within the protein shape its evolution.

#### V. CONDENSED-MATTER THEORY OF PROTEINS

This section will review a theory of proteins in terms of evolvable condensed matter. we shall discuss the conceptual roots of this approach in the physics of amorphous matter (mainly glasses) and spectral theory. We will introduce the basic setting of modeling proteins as evolving amino acid networks. The emergence of function is associated with the evolution of a weakly connected region, which enables a low-energy 'floppy' mode to appear. This minimal network approach allows one to examine basic questions of protein evolution.

we shall discuss two different models in this review. One will be called the 'cylinder-model' and the other the 'HP-model'. The first model is simpler, but the second comes somewhat closer the biological reality. Before distinguishing the two models, we describe their common features.

---

<sup>7</sup> PDB = protein data bank, <https://www.rcsb.org>.

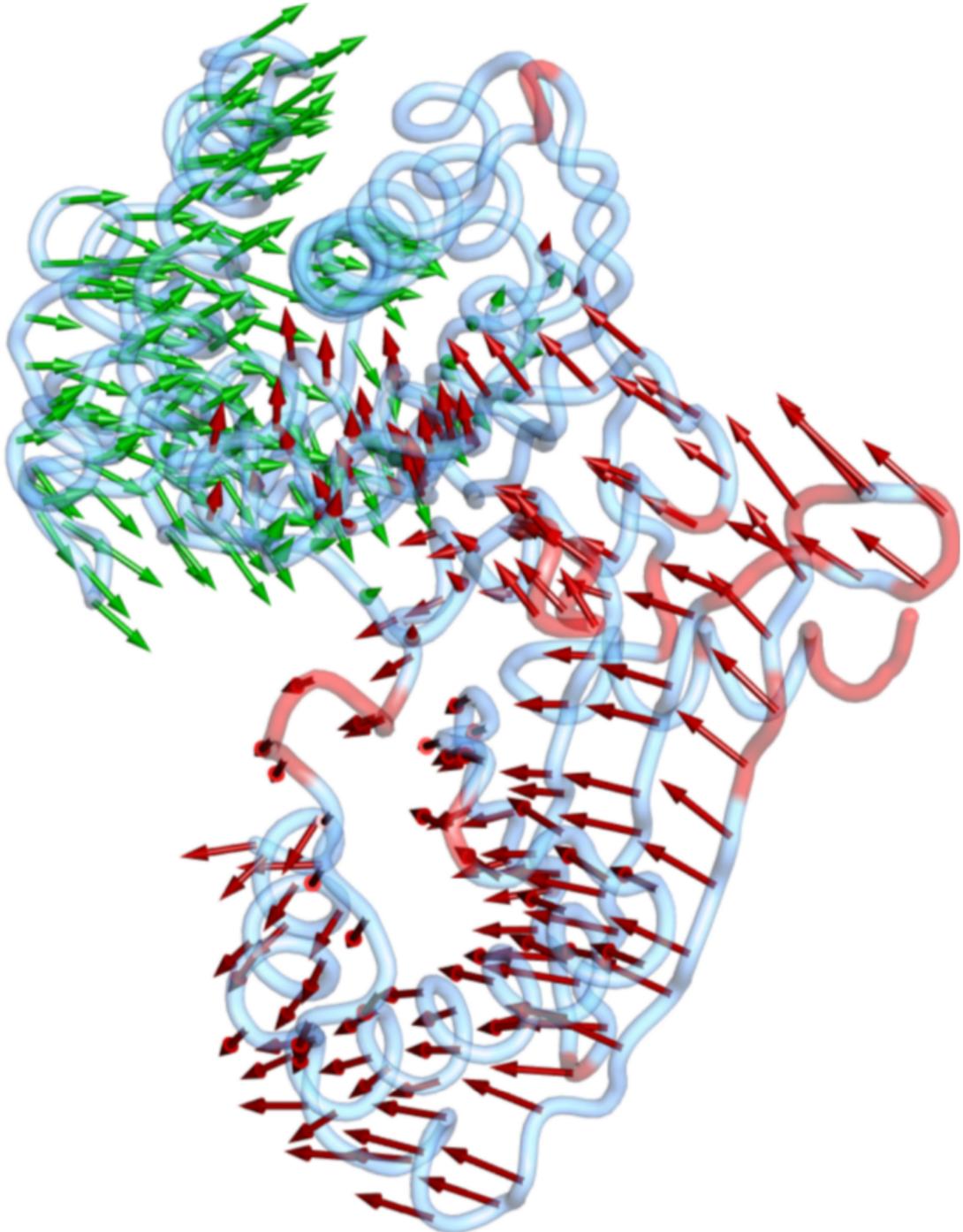


FIG. 3 The motion and deformation between two states of glucokinase in state 1v4s and 1v4t. The arrows are scaled up for better visibility. They are colored green, resp. red depending on whether they move down or up relative to a plane passing horizontally through the center of the protein. Galilean motions have been eliminated. The red coloring of the tube corresponds to concentration of shear and is the same as in the leftmost panel of Fig. 5. See Sec. IX.E.

### A. Lattice models

Our protein is modeled by a finite (regular) lattice in 2 (or 3) dimensions. We assume that the lattice forms a cylinder (periodic boundary conditions) or an open rectangle (open boundary conditions) of width  $w$  and height  $h$  (see the examples in Sec. VIII.A–VIII.B).

It is important to note that  $w$  and  $h$  are *finite* while otherwise quite arbitrary. This is so because the protein should not be viewed

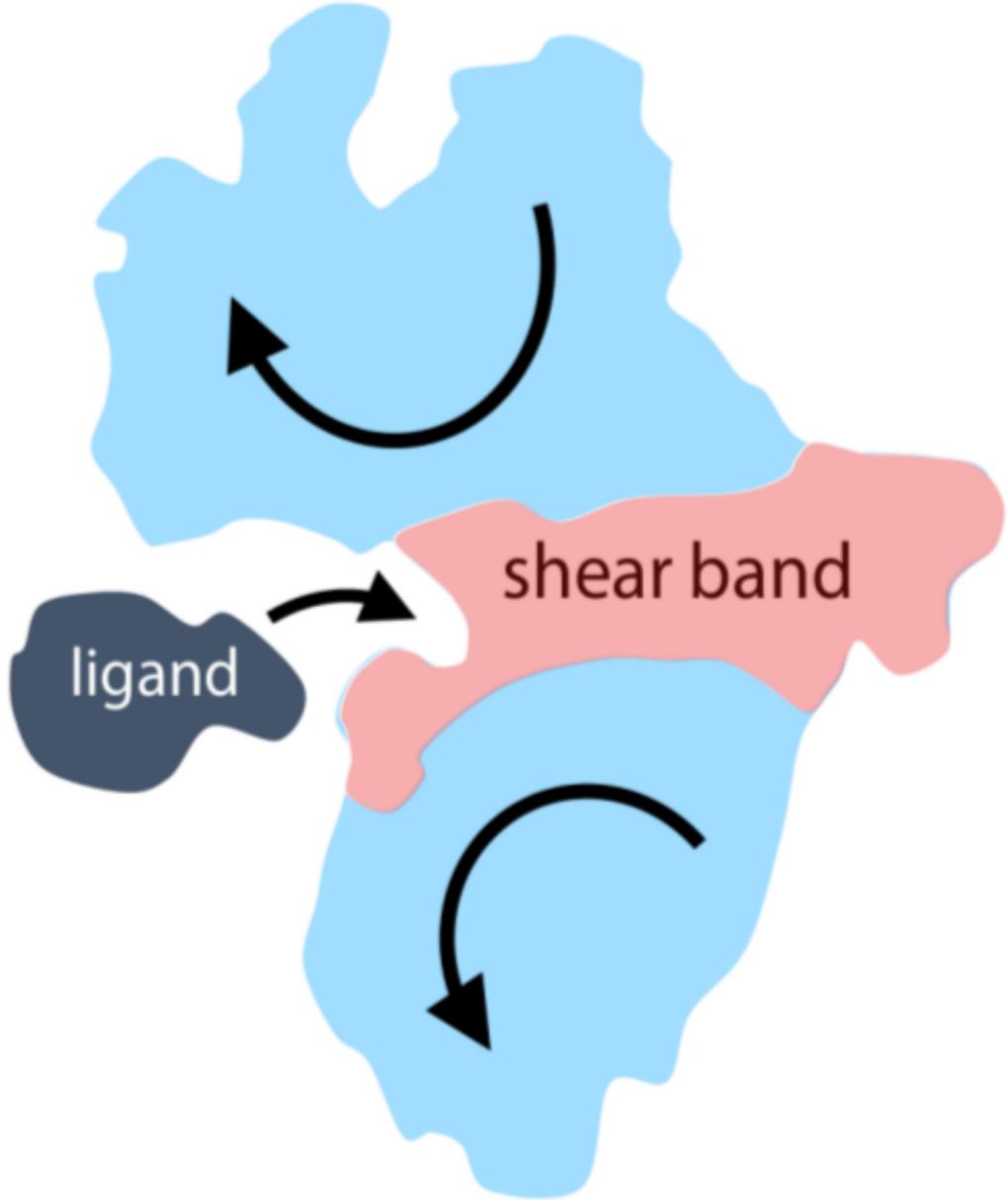
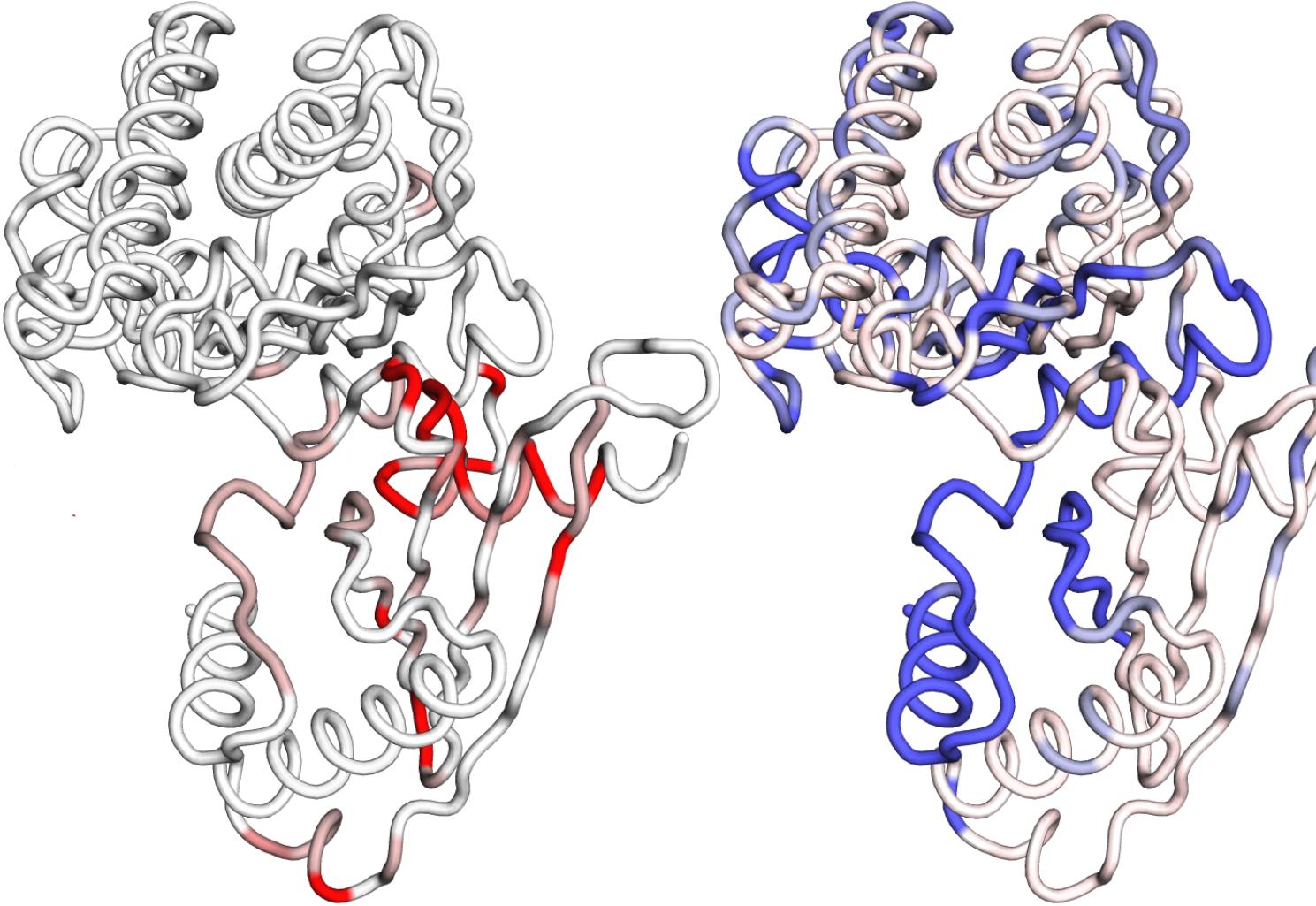


FIG. 4 A schematic interpretation of Fig. 3. Emphasis is given to the two moving pieces, with a hinge between them. This kind of hinge will be called the ‘fluid channel’ or ‘shear band’.

as a problem of thermodynamic limits, but rather in the context of small amorphous objects. This being said, other aspects of the geometry seem less important. The points may also be chosen as lying on small perturbations from the regular lattice to avoid effects of lattice symmetry on the spectrum. The number of AAs should typically be in the range 200–2000, corresponding to the typical size of the protein.

Amino acids interact via electrostatic forces, van der Waals forces, hydrogen bonds, disulfide bonds and hydrophobicity [Fersht 1999; Petsko and Ringe 2004]. All these are short range interactions, which amount to *local* coupling between lattice points. we shall therefore assume that each AA interacts with its nearest and next nearest neighbors. For example on a hexagonal lattice, with nearest and next-nearest neighbors linked, the number of connections (the node’s degree) is at most 12; all nodes in the protein interior have 12 links (*i.e.*, bonds) while those at the boundary have fewer (but at least 3), see Fig. 10.

Finally, the coupling itself is modeled by harmonic springs carried by each graph link [Alexander 1998; Born and Huang 1954; Chennubhotla et al. 2005; Tirion 1996]. Its strength is determined only by the types of AAs at each end of the link.



**FIG. 5 Evolutionary and mechanical properties of glucokinase.** Left: Shear (darker is more shear), discussed in Sec. IX.E. Center: Conservation (darker = fewer mutations), discussed in Sec. IX.J. Right: Correlation (dark red = mutations correlate), discussed in Sec. IX.K.

## B. The lattice Laplacian

The lattice and its links may be viewed as an abstract graph. This means that one can define gradients and Laplacians [Biggs 1993; Chung 1997].<sup>8</sup> In the graph, there are  $n_a = w \times h$  amino acid nodes, indexed by Roman letters, and  $n_b$  bonds, indexed by Greek letters.<sup>9</sup>

First, one endows every bond in the graph with an arbitrary but fixed orientation, and then the incidence matrix of a graph is the  $n_b \times n_a$  matrix defined by

$$\nabla_{\alpha i} = \begin{cases} -1, & \text{if } i \text{ is the initial vertex of edge } \alpha, \\ 1, & \text{if } i \text{ is the final vertex of edge } \alpha, \\ 0, & \text{if } i \text{ is not in } \alpha. \end{cases}$$

Remark that for any function  $f$  on the vertices, the map  $f \mapsto \nabla f$  is the co-boundary mapping of the graph, namely

$$\nabla f(\alpha) = f(j) - f(i),$$

where  $\alpha$  is the link connecting  $i$  to  $j$ .

<sup>8</sup> The first book is more combinatorial, and the second introduces more spectral concepts.

<sup>9</sup> If a bond  $\alpha$  connects nodes  $i$  and  $j$  ( $i \neq j$ ), we also write  $\alpha = (i, j)$ .

As in the continuum case, the Laplace operator  $\Delta$  is the product  $\Delta = \nabla^T \nabla$ , where  $T$  denotes the adjoint. The non-diagonal elements  $\Delta_{ij}$  are  $-1$  if  $i$  and  $j$  are connected and  $0$  otherwise. The diagonal part of  $\Delta$  is the degree  $\Delta_{ii} = z_i$ , *i.e.*, the number of nodes connected to  $i$ . Note that this is a discrete graph Laplacian, and no coordinates are involved so far.

We next embed the graph in a Euclidean space  $\mathbb{R}^d$  ( $d = 2$ ), by assigning positions  $r_i \in \mathbb{R}^d$  to each AA, *i.e.*, to each lattice point  $i = 1, \dots, n_a$ . This is coded as a  $n_a \times d$  real matrix  $\mathbf{r}$ . Finally, to each bond  $\alpha$  we assign a spring with constant  $k_\alpha$  which we view as the diagonal elements of an  $n_b \times n_b$  matrix  $\mathbf{K}$ :

$$\mathbf{K}_{\alpha\beta} = k_\alpha \delta_{\alpha\beta}. \quad (1)$$

This defines a deformable spring network which has an internal energy, an equilibrium configuration.

To account for the energy cost of deformations in the lattice protein, one considers the elastic tensor  $\mathbf{H}$  (or Hamiltonian) which we now describe in detail [Chung and Sternberg 1992, pp. 618–619]. The quantity  $\mathbf{H}$  is a tensor because the deformations are not scalars, but vectors in  $\mathbb{R}^d$ . We first denote by  $\mathbf{n}_\alpha$  the (normalized) direction vectors for each bond  $\alpha = (i, j)$ :  $\mathbf{n}_\alpha = (\mathbf{r}_i - \mathbf{r}_j) / |\mathbf{r}_i - \mathbf{r}_j|$ . Then, we define the ‘embedded’ gradient tensor  $\mathbf{D}$  (of size  $n_b \times n_a \times d$ ) which is obtained by multiplying each element of the graph gradient  $\nabla$  by the corresponding vector  $\mathbf{n}$ :

$$\mathbf{D}_{\alpha i} = \mathbf{n}_\alpha^T \nabla_{\alpha i}, \quad (2)$$

namely each projection of  $\mathbf{D}$  on a bond  $\alpha$ ,  $\mathbf{D}_{\alpha,:}$ , is a  $n_a \times d$  matrix containing only  $2d$  non-zero entries in rows  $i$  and  $j$ , which correspond to the components of the unit vector along the bond  $\alpha = (i, j)$ .

Let  $\mathbf{u}_i$  be the displacement vectors of each vertex from  $\mathbf{r}_i$  to  $\mathbf{r}_i + \mathbf{u}_i$ , therefore  $\mathbf{u}$  is a  $n_a \times d$  matrix. The elastic energy of such a perturbation is

$$\mathcal{E} = \frac{1}{2} \mathbf{u}^T \mathbf{H} \mathbf{u} = \frac{1}{2} \sum_{ij} \sum_{k\ell} \mathbf{u}_{ik} (\mathbf{H}_{ij})_{k\ell} \mathbf{u}_{j\ell}, \quad (3)$$

where the Hamiltonian tensor is defined as

$$\begin{aligned} \mathbf{H} &= \mathbf{D}^T \mathbf{K} \mathbf{D} \\ &= \left( \sum_{\alpha\beta} \mathbf{D}_{\alpha ik} \mathbf{K}_{\alpha\beta} \mathbf{D}_{\beta j\ell} \right)_{i,j=1,\dots,n_a; k,\ell=1,\dots,d}. \end{aligned}$$

The  $d \times d$  off-diagonal components are:

$$\begin{aligned} \mathbf{H}_{ij} &= \sum_{\alpha} \nabla_{\alpha i} \mathbf{n}_{\alpha} \mathbf{K}_{\alpha\alpha} \mathbf{n}_{\alpha}^T \nabla_{\alpha j} \\ &= \Delta_{ij} k_{(i,j)} \mathbf{n}_{(i,j)} \mathbf{n}_{(i,j)}^T, \end{aligned}$$

which we complete with the diagonal blocks ( $i = j$ ) so that rows and columns sum to zero:  $\mathbf{H}_{ii} = - \sum_{j \neq i} \mathbf{H}_{ij}$ .

In this construction, we have assumed that the equilibrium configuration of the network (described by the vectors  $\mathbf{r}_i$ ) is such that all springs are at their equilibrium length, disregarding the possibility of ‘internal stress’ [Alexander 1998], hence the initial elastic energy is 0. The extension of the theory to networks that are initially frustrated, *i.e.*, where some springs are stretched or squeezed is a difficult subject. A paper which studies the conjectures by Alexander on internally stressed networks is [Kustanovich et al. 2003]. The spring constant is therefore the derivative of the interaction at the equilibrium length of the spring.

We next consider only small deviations of the AAs from their equilibrium, *i.e.*, the linear mechanical response of the protein to an applied force. While this approximation cannot account for plastic and non-affine deformations that often occur in real proteins, it certainly simplifies the analysis, in contrast to the inherent difficulties of studying fully nonlinear systems.

Given a prescribed ‘protein fold’ (the lattice positions  $\mathbf{r}_i$ ,  $i = 1, \dots, n_a$ ), a gene first determines the spring constants via a ‘genetic code’ which maps codons to AAs on the lattice and thereby determines the interaction strength between neighbors on the lattice. This in turn defines a phenotype of the protein, namely its mechanical response under deformations. Each choice of the gene, *i.e.*, the set of codons  $\mathbf{c} = \{c_i\}$ , defines a Hamiltonian  $\mathbf{H} = \mathbf{H}(\mathbf{c})$ . Component-wise, each  $d \times d$  block  $\mathbf{H}_{ij}$  ( $i \neq j$ ) depends only on the codons  $c_i$  and  $c_j$ :  $\mathbf{H}_{ij}(\mathbf{c}) = \mathbf{H}_{ij}(c_i, c_j)$ .

In summary, note that  $\mathbf{H}$  depends on three things:

1. The position  $\mathbf{r}_i$  of each amino acid,  $i = 1, \dots, n_a$ ,
2. The type  $c_i$  of each amino acid,  $i = 1, \dots, n_a$ ,
3. The spring constants  $k(c, c')$  representing the interaction strengths between amino-acids types  $c$  and  $c'$ .

This definition is clearly versatile enough to be generalized to other systems, such as proteins made of the standard 20 AAs with specific interaction constants for each of the possible AA pair.

### C. Hooke's law

We have now a map from genes  $\mathbf{c}$  to Hamiltonians  $\mathbf{H}(\mathbf{c})$ , and we want to study the deformability of the network as a function of  $\mathbf{c}$ . In the linear regime of relatively moderate deformations, one can use Hooke's law (see *e.g.*, [Alexander 1998]) to relate a (small) deformation  $\mathbf{u}$  to the force by

$$\mathbf{f} = \mathbf{H} \mathbf{u},$$

where  $\mathbf{f}$  is a force vector field. We are interested in the inverse relation, since we want to know the deformation of the network (protein) as a function of the applied force  $\mathbf{f}$ . This inverse will be described by Green's function  $\mathbf{G}$ :

$$\mathbf{u} = \mathbf{G} \mathbf{f}. \quad (4)$$

## VI. SIMULATING EVOLUTION

Next, we consider modeling evolution, for a general ‘genetic code’. As described above, to each gene  $\mathbf{c}$  there is a natural Hamiltonian  $\mathbf{H}(\mathbf{c})$  associated with it. This is the mechanical genotype-to-phenotype map. We assume that a fitness function  $F$  is given, mapping every  $\mathbf{H}$  to its fitness score  $F(\mathbf{H}) \in \mathbb{R}$ . The observable we take later as  $F$  will be an expectation value for some components of the force field  $\mathbf{f}$ . The evolutionary process alters this fitness, by mutating individual random positions in the gene  $\mathbf{c}$  (the collection of  $c_i$ ). This is realized by a Metropolis algorithm [Metropolis et al. 1953]:

In an evolution simulation, one exchanges a randomly selected codon with another one (at the same position), while demanding that the fitness change  $\delta F$  is positive or non-negative. We call  $\delta F > 0$  a beneficial mutation, whereas  $\delta F = 0$  corresponds to a neutral one. Deleterious mutations,  $\delta F < 0$ , are generally rejected.

As in statistical physics, variants of this algorithm can be envisaged, for example, by asking for an increase of  $F$  by a minimal factor  $|F| \rightarrow |F| \cdot (1 + \varepsilon)$  with  $\varepsilon > 0$ , for a step to be accepted. Other possibilities include the introduction of ‘temperature’, *i.e.*, accepting or rejecting even deleterious mutations,  $\delta F < 0$ , with some probability. The rationale behind using these variants of Metropolis algorithms lies in the nature of natural mutations. For a review of the role of deleterious mutations, see [Kondrashov 2017]. Details of  $F$  will be given when we discuss various models in Sec. VIII.

## VII. GREEN'S FUNCTION AS A LINK BETWEEN THE THEORY OF AMORPHOUS SOLIDS AND LIVING MATTER

In the previous section, the ground was prepared for studying the connection between the genes and the mechanical properties of the proteins they code. we shall use the mapping from the gene  $\mathbf{c}$  to the Hamiltonian  $\mathbf{H}(\mathbf{c})$  introduced above. One of the questions to be examined is how the protein reacts to forces applied to it, and how this response is encoded in the gene. Such forces occur when a small ligand molecule attaches to a binding site on the protein’s surface, inducing a mechanical response in other regions of the protein.

Intuitively, this means that we are looking for a relatively strong reaction to a weak signal. Such phenomena are captured by soft modes. Such modes are given by zero eigenvalues of the Hamiltonian  $\mathbf{H}$ , and the corresponding deformations are described by the eigenvector  $\mathbf{u}$  of displacements of  $\mathbf{r}$  (corresponding to the zero eigenvalue).

Among the many approaches to the zero eigenvalue problem, we use the methods of Green's functions, which are well adapted to the emergence of soft modes in protein evolution. Green's function (also called the resolvent, matrix inverse) is useful here because of the following observation: Consider a mutation that alters just one  $c_i$ . Given the short-range nature of  $\mathbf{H}$ , this implies that only a small number of terms in Eq. (3) will change, independently of the size of the system. For example, for the hexagonal lattice in Fig. 6, no more than 12 terms change with each mutation.

Since, by Hooke's law,  $\mathbf{f} = \mathbf{H}(\mathbf{c})\mathbf{u}$ , the response of the system to an external force is given by the inverse relation  $\mathbf{u} = \mathbf{G}(\mathbf{c})\mathbf{f}$ , where  $\mathbf{G}$  is Green's function, *i.e.*, the inverse of  $\mathbf{H}$ . So,  $\mathbf{G}$  maps the genotype  $\mathbf{c}$  to the reaction of the protein to an external force  $\mathbf{f}$ . A typical example of such a stimulus appears when  $\mathbf{f}$  ‘pinches’ 2 neighboring AAs towards each other; we would like to measure the effect of the pinch on another AA pair (usually on the opposite side of the protein).

In dimension  $d=2$ , the Hamiltonian  $\mathbf{H}$  has always  $d(d-1)/2 = 3$  zero eigenvalues, owing to the rigid Galilean transformations (2 translations and 1 rotation) of the lattice as a whole. Therefore, since  $\mathbf{H}$  is bound to be singular, it lacks a proper inverse. instead, one may compute the inverse on the subspace of  $\mathbb{R}^{n_a} \times \mathbb{R}^d$  in the complement of the 3 Galilean directions. This is called the pseudo-inverse [Penrose 1955] and is usually denoted by  $\mathbf{G}(\mathbf{c}) = \mathbf{H}(\mathbf{c})^\dagger$ .

Let  $\mathbf{P}$  be the projection on the subspace spanned by the generators of the Galilean transformations, then

$$\mathbf{G}(\mathbf{c}) = ((\mathbf{1} - \mathbf{P})\mathbf{H}(\mathbf{c})(\mathbf{1} - \mathbf{P}))^{-1} \equiv \mathbf{H}(\mathbf{c})^\dagger.$$

It is easy to verify that if  $\mathbf{u}$  is orthogonal to the zero modes,  $\mathbf{u} = (1 - \mathbf{P})\mathbf{u}$ , then  $\mathbf{u} = \mathbf{G}\mathbf{H}\mathbf{u}$ . The pseudo-inverse obeys the four requirements: (i)  $\mathbf{H}\mathbf{G}\mathbf{H} = \mathbf{H}$ , (ii)  $\mathbf{G}\mathbf{H}\mathbf{G} = \mathbf{G}$ , (iii)  $(\mathbf{H}\mathbf{G})^T = \mathbf{H}\mathbf{G}$ , and (iv)  $(\mathbf{G}\mathbf{H})^T = \mathbf{G}\mathbf{H}$ .

The projection onto the complement of the 0-space commutes with the action of mutations, since changing the AA at a site does not change the Galilean invariance of the lattice. Therefore, the pseudo-inverse can be used for our purposes just like the standard inverse.

### A. Woodbury's formula

When one changes a gene  $\mathbf{c}$  to some  $\mathbf{c}'$ , then the change in the Hamiltonian is  $\delta\mathbf{H} = \mathbf{H}(\mathbf{c}') - \mathbf{H}(\mathbf{c})$  and correspondingly the changes in Green's function from  $\mathbf{G}(\mathbf{c})$  to  $\mathbf{G}(\mathbf{c}')$ . The Woodbury formula [Deng 2011; Woodbury 1950] relates  $\delta\mathbf{G} = \mathbf{G}(\mathbf{c}') - \mathbf{G}(\mathbf{c})$  to  $\delta\mathbf{H}$  as follows: First, one notes that the rank of the change tensor  $\delta\mathbf{H}$  is equal to the number  $r$  of bonds altered by the mutation.  $\delta\mathbf{H}$  can therefore be written as

$$\delta\mathbf{H} = \mathbf{MBM}^T,$$

where the  $r \times r$  diagonal matrix  $\mathbf{B}$  records the strength change of the bonds (*e.g.*, from weak to strong or vice versa).  $\mathbf{M}$  is a  $r \times n_a \times d$  tensor which is the restriction of  $\mathbf{D}$  (see Eq. (2)) to bonds which were changed. This allows one to easily calculate changes in Green's function:

$$\delta\mathbf{G} = -\mathbf{GM} (\mathbf{B}^{-1} + \mathbf{M}^T \mathbf{GM})^\dagger \mathbf{M}^T \mathbf{G}. \quad (5)$$

The reader who is not familiar with Eq. (5) can compare it to the resolvent formula (in the commutative, scalar case):

$$\frac{1}{x+y} - \frac{1}{x} = -\frac{1}{x} \left( \frac{1}{\frac{1}{x} + \frac{1}{y}} \right) \frac{1}{x},$$

with  $x^{-1}$  corresponding to  $\mathbf{G}$ , and  $y$  to  $\mathbf{B}$  (and  $\delta\mathbf{H}$ ).

The Woodbury formula is especially useful since one has to invert only square matrices of size  $r$  ( $\mathbf{B}$  and the term in brackets in Eq. (5)), instead of inverting the larger tensor  $\mathbf{H}$  of size  $d \times n_a$  [Henderson and Searle 1981]. For point mutations, this difference is dramatic, since the rank  $r$  can be at most  $z$ , the number of neighbors of the mutated AA, implying that  $r \ll n_a \times d$ . For example, in the hexagonal model Fig. 6,  $r \leq z = 12$  while  $n_a \times d = 1080$ .

### B. Dyson's formula

Another useful (and more common) identity is Dyson's formula [Abrikosov et al. 1963; Dyson 1949a,b]. It can be obtained by applying the resolvent identity to  $\mathbf{G}' = \mathbf{G} + \delta\mathbf{G}$ , leading to

$$\mathbf{G}' = \mathbf{G} - \mathbf{G} \delta\mathbf{H} \mathbf{G}'. \quad (6)$$

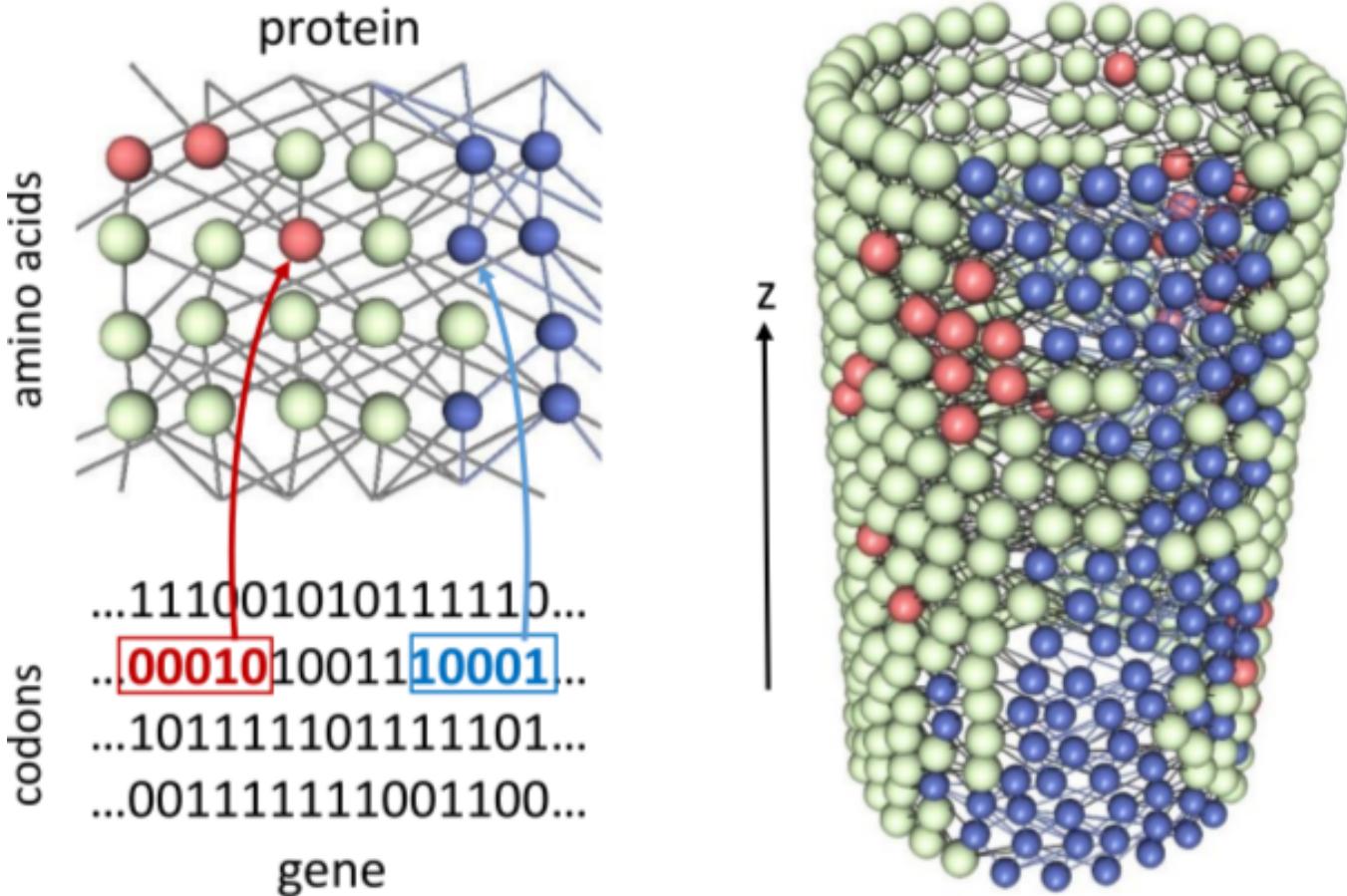
Since  $\mathbf{G}'$  appears on the r.h.s., one may successively iterate this identity to get the Dyson series,

$$\delta\mathbf{G} = \mathbf{G}' - \mathbf{G} = -\mathbf{G} \delta\mathbf{H} \mathbf{G} + \mathbf{G} \delta\mathbf{H} \mathbf{G} \delta\mathbf{H} \mathbf{G} - \dots. \quad (7)$$

The series is widely used in potential scattering, and is interpreted there as expansion in multiple scattering. The first term is usually called the Born term. We will interpret this identity in terms of multiple mutations and this will be another contact of methods known from the physics literature with questions in evolution.

## VIII. MODELS: PROTEIN AS AN EVOLVING MACHINE

After introducing the basic principles of our approach, we now discuss how to apply them in specific models. As in any simplifying model, there is an intrinsic conflict: On the one hand, one would like to keep the model as simple as possible, because the goal is to test basic principles, not specific proteins. On the other hand, there should still be some connection to real proteins. As mentioned before, one cannot apply the thermodynamic limits of standard statistical mechanics (infinite number of particles, long range potentials, and the like), since the protein boundary plays an important role. So the protein is treated as a finite, amorphous system.



**FIG. 6 The main features of the cylinder model:** Left: the mapping from the binary gene to the connectivity of the amino acid (AA) network that makes a functional protein. The color of the AAs represents their rigidity state as determined by the connectivity according to the algorithm of Sec. VIII.A.3. Each AA can be in one of three states: rigid (gray) or fluid (*i.e.*, non-rigid), which are divided between shearable (blue) and non-shearable (red).

Right: the AAs in the model protein are arranged in the shape of a cylinder, in this case with a fluid channel (blue region). Such a configuration can transduce a mechanical signal of shear or hinge-like motion along the fluid channel.

#### A. A model with very simple structure (Cylinder-model)

This model, introduced in [Tlusty et al. 2017], assumes that the coupling between nodes will only depend on one of the two AAs linked by a bond. Although we reformulate the problem differently, it is in fact equivalent to a lattice model as described above (Sec. V.A), in the limit of infinite spring constants (bonds are solid rods).

To get somewhat closer to the standard genetic code with its 20 AAs, we introduce  $2^5 = 32$  species of AAs; each AA is coded by a 5-bit codon written in a binary alphabet of 0s and 1s.<sup>10</sup> The geometry of the model is a square lattice with periodic boundary conditions in the horizontal direction, forming a cylinder. One realization is shown in Fig. 6 (right) where the blue region corresponds to the shear band. This should be compared to Fig. 3 where the shear band is between the red and green arrows (in Fig. 4 the shear band is shown in red). we shall see later that the motion around shear bands in the models is similar in nature to the one of Fig. 3.

<sup>10</sup> So there are 2 nucleotides and 32 non-redundant codons.

## 1. The cylindrical amino acid network

We now define the model in further detail: We consider a geometry with height  $h = 18$ , *i.e.*, the number of layers in the  $z$  direction, and width  $w = 30$ , *i.e.*, the circumference of the cylinder. The row and column coordinates of an AA are  $(r, q)$ , with  $r$  for the row  $(1, \dots, h)$  and  $q$  for the column  $(1, \dots, w)$ . The cylindrical periodicity is realized by taking the horizontal coordinate  $q$  modulo  $w$ ,  $q \rightarrow \text{mod}_w(q - 1) + 1$ .

Each AA in row  $r$  can connect to any of its five nearest neighbors in the next row below it. This defines  $2^5 = 32$  effective species of amino acids that differ by their ‘chemistry’, *i.e.*, by the pattern of their bonds. An AA at  $(r, q)$  is encoded in the gene as a 5-letter binary codon  $\ell_{rqk}$ ,  $k = -2, \dots, 2$ , where the  $k$ -th letter denotes the existence ( $= 1$ ) or absence ( $= 0$ ) of the  $k$ -th bond. The full gene is therefore a binary sequence of length  $2700 = 5 \cdot w \cdot h$ . Each of its  $w \cdot h = 540$  codons specifies which of the 5 bonds are present or absent. The effective size of the problem is only  $n_s = 2550 = w \cdot (h - 1) \cdot 5$  because the bonds of the bottom row are never used and do not affect the configuration of the protein and the resulting dynamical modes.

## 2. Evolution searches for a mechanical function

We next define the evolutionary fitness of the cylindrical protein as following: To become functional, the protein has to evolve a configuration of AAs and bonds that can transduce a mechanical signal from a prescribed input at the bottom of the cylinder to a prescribed output at its top. This signal is a large-scale, low-energy deformation where one domain moves rigidly with respect to another in a shear or hinge-like motion, which is facilitated by the presence of a fluidized, ‘floppy’ channel separating the rigid domains [Alexander 1998; Alexander et al. 1983; Phillips and Thorpe 1985].

The definition of the fluid channel is described in detail in Sec. VIII.A.3, but can be summarized by two features of amino acids in the channel (Fig. 6): (A) Fluidity – these AAs are not part of rigid sub-networks in the protein. Locally, this means that fluid AAs cannot be linked to too many rigid neighbors. (B) Shearability – the AAs in the channel should have enough fluid AAs around them to sustain low-energy shear motion.

The fluidity/rigidity and shearability propagate in a manner reminiscent of percolation. Note that, while the system ‘learns’, through mutations, to form a fluid channel, this learning is not by presenting it with many inputs, but by only checking the quality of the output under random mutations.

In [Tlusty 2016], Figure 8, the author imagined a feedback of the following type: a protein can evolve the ability to activate its own transcription in response to a stimulus, which is the first step towards cellular regulatory networks, see [Lee et al. 2002] for how this appears in the biological context, and [Djordjevic et al. 2003; Lässig 2007; Tlusty 2016, Fig. 8], for theoretical studies.

## 3. Rigidity propagation algorithm\*

The aim of this subsection is to define a model in which some local rigidity rules, spelled out below, are able to transmit deformability from the bottom of the cylinder to the top. There are many ways in which this can be realized, and the rules we give are a compromise between simplicity and the ability to fulfil this aim. We have tested other variants with similar outcome.

The large-scale deformations are governed by the rigidity pattern of the protein, which is determined by the connectivity of the AA network via a simple majority rule (Fig. 6, 7), as follows. These large scale deformations could in principle change the ability of the protein to bind a target, and in this way implement the response trigger in the feedback loop mentioned above. Each AA position will have two binary properties which define its state: *rigidity*  $\sigma$  (an AA is either *solid*,  $\sigma = 1$ , or *fluid*,  $\sigma = 0$ ) and *shearability*  $s$  (an AA is either *shearable*,  $s = 1$ , or *non-shearable*,  $s = 0$ ). Only 3 of the 4 possible combinations are allowed:

1. non-shearable and solid (yellow):  $\sigma = 1; s = 0$ ,
2. non-shearable and fluid (red):  $\sigma = 0; s = 0$ ,
3. shearable and fluid (blue):  $\sigma = 0; s = 1$ .

Non-shearable protein domains tend to move as rigid bodies (*i.e.*, via translation or rotation), whereas shearable regions are easy to deform. The non-shearable domains are mostly rigid, but can still have pockets of fluid AAs.

Given a fixed sequence and an input state in the bottom row of the cylinder,  $\{\sigma_{1,q}, s_{1,q}\}$  the state of the cylinder is completely determined by ‘percolation’ of the two properties, rigidity/fluidity and shearability, through the network, as follows.

In a first sweep through the rows, we establish the rigidity property  $\sigma$ . The rigidity of AAs in row  $r = 1$  are prescribed initially. In all other rows ( $r = 2, \dots, h$ ) the bonds determine the value of the rigidity of  $(r, q)$  through a majority rule:

$$\sigma_{r,q} = \theta \left( \sum_{k=-2}^2 \ell_{rqk} \sigma_{r-1,q+k} - \sigma_0 \right), \quad (8)$$

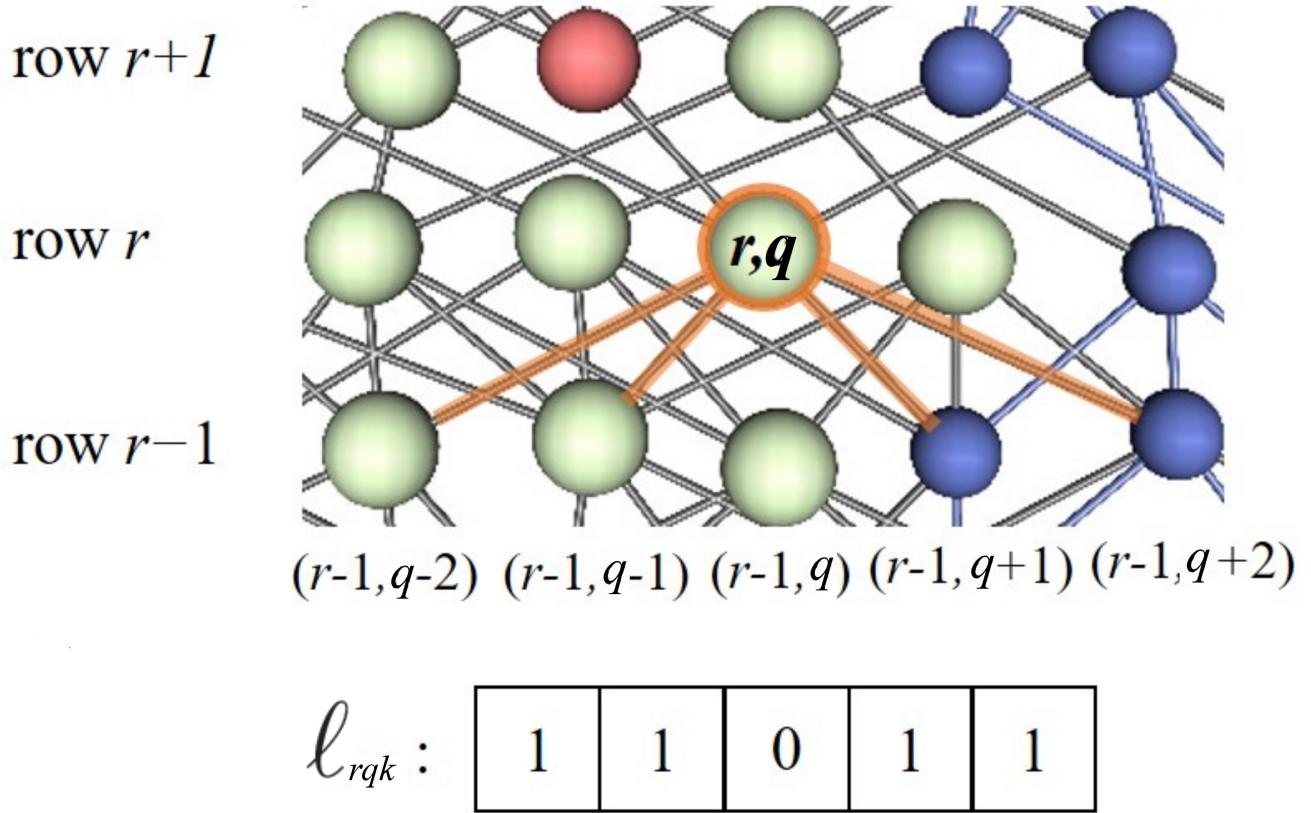


FIG. 7 Illustration of the percolation rules for shearability and rigidity states. Note that site  $(r, q)$  was turned solid because it is attached to 2 solid sites below it. Also note that the red site above it is fluid, because it is attached to less than 2 solid sites below it. But it is not shearable because it does not connect to a shearable site below it. On the other hand, the top right site is shearable and fluid, since it is attached to only one solid site (namely  $(r, q)$ ) and no others on the invisible part of the structure (as seen by its blue connections), and it is also connected to the blue site at  $(r, q + 2)$ .

where  $\theta$  is the step function ( $\theta(x \geq 0) = 1, \theta(x < 0) = 0$ ). The parameter  $\sigma_0 = 2$  is the minimum number of rigid AAs from the  $r - 1$  row required to rigidly support an AA: In 2D each AA has two coordinates which are constrained if it is connected to two or more static AAs. In this way, the rigidity property of being pinned in place propagates through the lattice as a function of the initial row and of the bonds as encoded in the gene.

We next address the shearability property  $s$  which is determined by the rigidity as follows: We assume that all fluid AAs in row  $r = 1$  are also shearable (blue:  $(\sigma = 0; s = 1)$ ). A fluid node  $(r, q)$  in row  $r$  will be shearable if any of its neighbors at  $(r - 1, q)$  or  $(r - 1, q \pm 1)$  is shearable:

$$s_{r,q} = (1 - \sigma_{r,q}) \cdot \theta \left( \sum_{k=-1}^1 s_{r-1,q+k} - s_0 \right), \quad (9)$$

where  $s_0 = 1$ . The first factor on the r.h.s. ensures that a solid AA is never shearable.

#### 4. Fitness and mutations

As explained before, evolution searches for a functional protein which can transfer forces. The simulation of this search starts from a random sequence (of 2550 codons), and from an initial state (input) in the bottom row of the cylinder. For most simulations, this initial state consisted of only solid beads except a stretch 5 consecutive shearable beads, as shown in Fig. 8.

We next define a fitness function which will direct the evolutionary process. The state with maximal fitness (*i.e.*, the ‘target’) is a chain of  $w$  values, fluid and shearable ( $\sigma = 0; s = 1$ ) or solid ( $\sigma = 1; s = 0$ ), in the top row, which the protein should yield as an output: we call it  $x^* \equiv \{\sigma_q^*, s_q^*\}_{q=1,\dots,w}$ . Given

1. a gene sequence  $c$ , which determines the connectivity  $\ell_{rqk}$ ,

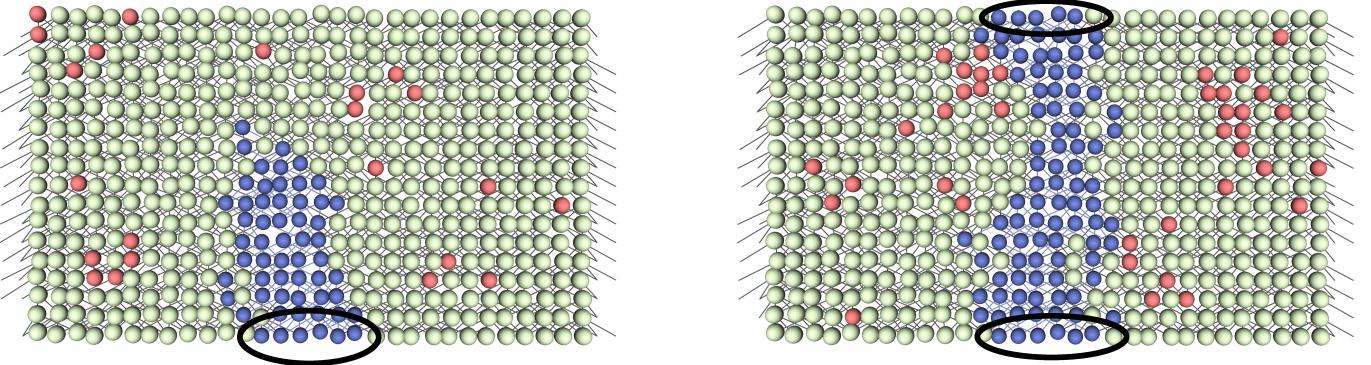


FIG. 8 **Evolution of a mechanical function:**

A configuration (left) with a prescribed input (black ellipse at bottom) and random connectivity pattern eventually evolved to form a fluid channel (right). The initial state has 6 fluid points (in black ellipse), our fitness requires 5 fluid points at the top (black ellipse).

## 2. the input state, $\{\sigma_{1,q}, s_{1,q}\}_{q=1,\dots,w}$ ,

the algorithm described above uniquely defines the output state in the top row,  $\{\sigma_{h,q}, s_{h,q}\}_{c=1,\dots,w}$ . At each step of evolution, the output state is compared to the fixed target by measuring the Hamming distance<sup>11</sup> to the target  $x^*$ :

$$F = -w + \sum_{q=1}^w (1 - |s_{h,q} - s_q^*|) \cdot (1 - |\sigma_{h,q} - \sigma_q^*|) . \quad (10)$$

This is the fitness function  $F$  of Sec. VI.

*Remark:* It is an important feature of this model that the fitness of the network is only measured at the target line. This corresponds to the biological fact that the protein can only interact with the outside world through its surface (in our case, the ends of the cylinder). One of the major outcomes of the model is that this fitness still has a strong influence on the connectivity deep inside the interior of the protein. While similar, the propagation of fluidity should not be confused with learning in neural networks: In the learning case, the system is presented with several inputs and learns to recognize others, while here there is a fixed task, and the connections are only driven by the target function  $F$ .

## 5. Simulation of evolutionary dynamics

Thanks to the simplicity of the model, one can easily perform  $10^6$  simulations in a short time, and gain much better statistical insight than is possible with typical bioinformatic data (of course, at the price of disregarding many biochemical details). We present results for one specific fitness: the input at the bottom is a fluid region of length 6 and output target at the top is a fluid region of length 5. For other variants of this model, cf. [Tlusty et al. 2017].

We study 200 independent initial states (genes), starting from a random sequence with about 90% of the bonds present at the start. Given a sequence, we sweep according to the rules of Eq. (8)–(9) through the net, and measure the Hamming distance  $F$  (Eq. (10)) between the last row and the desired target.

Solutions are then searched by successive mutations, with a Metropolis algorithm, [Metropolis et al. 1953]. At each iteration, a randomly drawn digit in the gene is flipped, that is, the values of 0 and 1 are exchanged. This corresponds to erasing or creating a randomly chosen link of a randomly chosen AA. After each flip, a sweep is performed, and the new output at the top row is again compared to the target. If  $F$  (which is negative) decreases, we backtrack and flip another randomly chosen bond. This procedure is repeated until optimal fitness is reached ( $F = 0$ ). This will happen with probability 1 if such a state exists, and typically requires  $10^3$ – $10^5$  mutations.

Although the functional sequences are extremely sparse among the  $32^{510} = 2^{2550}$  possible sequences, the small bias for getting closer to the target in configuration space directs the search rather quickly.

Once a maximal  $F$  is reached, we move away from it by further mutations and then look again for a new optimum. Reaching a state with  $F = 0$  takes around 11.2 beneficial mutations on average. Getting from an initial sequence to a maximizer is called a

<sup>11</sup> The Hamming distance between two sequences is the number of indices where they differ.

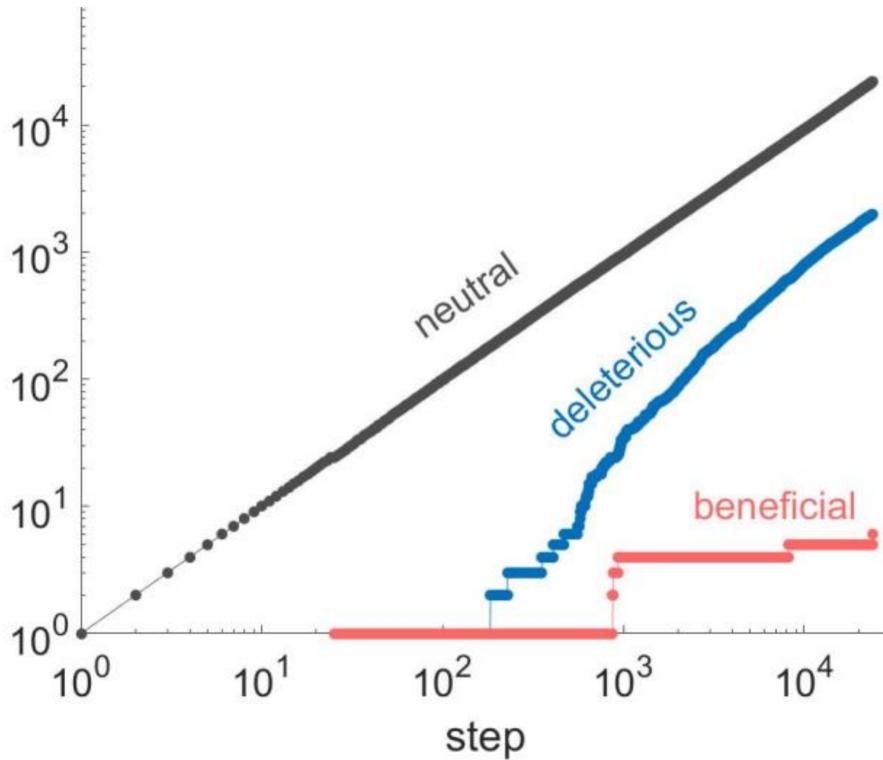


FIG. 9 Following the progress of evolution during a typical run. It is a sequence of mostly neutral steps, a fraction of deleterious ones, and rare beneficial steps. The vertical axis is the accumulated number of steps of each type.

‘generation’. For each of the 200 initial random genes, we followed 5000 generations, finding a total of  $10^6$  optima. The typical length of a generation between two maxima is about 1500 mutations (most of them neutral, see Fig. 9), similar to the time it takes starting from a totally random gene. We also simulated 1-generation paths starting from  $10^6$  random genes. The two cases are very similar, but the destruction-reconstruction simulations show some correlations between consecutive generations, which disappear after about 4 generations.

## B. A model with more realistic interactions (HP-model)

This model differs from the cylinder-model in several respects:

1. Geometry – The lattice is hexagonal with open boundary conditions,
2. Two-body interactions – the bonds depend on the nature of the AA at *both* ends of the link,
3. Amino acids species – There are only 2 species, H (hydrophobic) or P (polar).

The two amino acids species are encoded in a binary genetic alphabet and a codon size of 1: each AA chain is encoded in a gene  $\mathbf{c}$  of the same length  $n_a$ , where  $c_i = 1$  encodes H and  $c_i = 0$  encodes P,  $i = 1, \dots, n_a$  (in other words, the genetic code is the identity map).

We give next details of how the model is constructed. The lattice has width  $h = 20$  and height  $w = 10$ , and therefore  $n_a = 200$  AAs (see Sec. V.A). The bonds stretch over the 12 nearest and next-nearest neighbors of an AA (see Fig. 10, right panel, for the connectivity and any of the panels in Fig. 11 for the global arrangement of the lattice).

The strength of the springs is given by the HP model [Lau and Dill 1989] according to the rule:

$$k_{(i,j)} = \kappa_w + (\kappa_s - \kappa_w)c_i c_j ,$$

where  $c_i$  and  $c_j$  are the (binary) codons of the AA connected by bond  $\alpha = (i, j)$ , with  $c_i = 1$  corresponding to H and  $c_i = 0$  to P. This implies that a strong H–H bond has  $k_\alpha = \kappa_s$ , whereas the other bonds P–P, H–P, and P–H are weak with  $k_\alpha = \kappa_w$ . In our simulations below we used  $\kappa_s = 1$  and  $\kappa_w = 0.01$ .

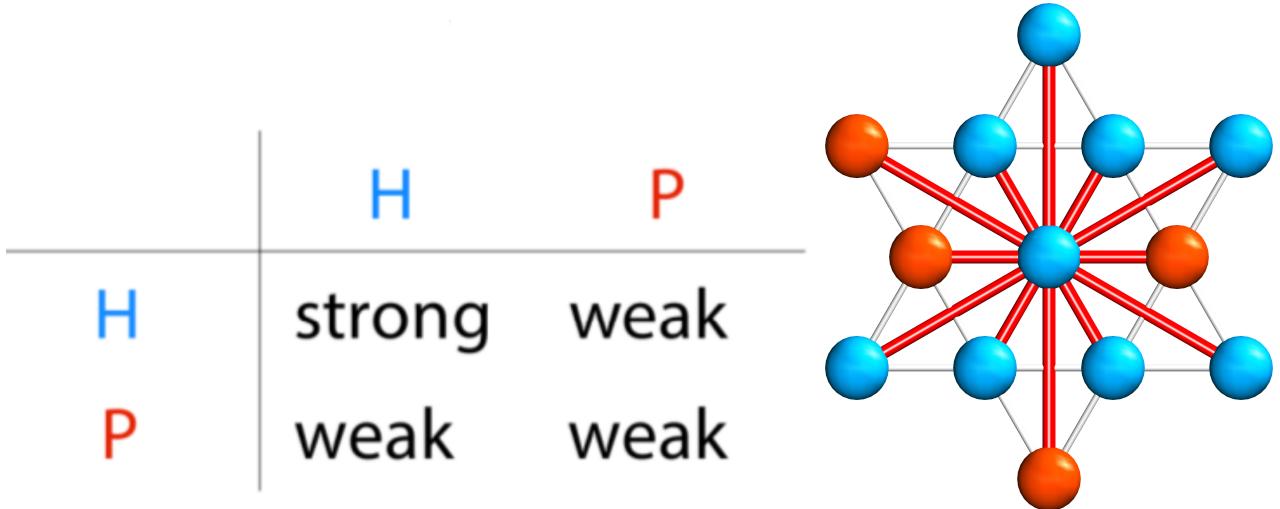


FIG. 10 The protein is made of two species of AAs, polar (P, red) and hydrophobic (H, blue) whose sequence is encoded in a gene. Each AA forms weak or strong bonds with its 12 nearest neighbors on the hexagonal lattice (right) according to the interaction rule in the table (left).

### 1. Pinching the network

The network is subjected to an external ‘pinch’ which is a localized force  $\mathbf{f}$  applied at the boundary of the network. This force acts in the complement of the subspace of Galilean invariance, *i.e.*,  $\mathbf{P}\mathbf{f} = 0$  (see Sec. VII). The pinch acts on a pair of neighboring boundary vertices,  $p'$  and  $q'$ , in the direction  $\mathbf{n}_{p'q'} = \mathbf{r}_{p'} - \mathbf{r}_{q'}$  which is parallel to the boundary (the L face of the network). Thus  $\mathbf{f}$  is a force dipole, *i.e.*, two opposing forces,  $\mathbf{f}_{q'} = -\mathbf{f}_{p'} = f \cdot \mathbf{n}_{p'q'}$ , which can be related to the deformation  $\mathbf{u}$  of the network by Green’s function Eq. (4). The pinch stimulus may represent localized interactions, for example a ligand biding at a specific binding site (Fig. 4).

The biological fitness is specified by how well the response of the network fits to a prescribed deformation vector  $\mathbf{v}$ . This vector is zero except at  $p$  and  $q$  which are on the opposite side of the network (R face). The fitness function  $F$  is therefore

$$F = \mathbf{v}^T \mathbf{u} = \mathbf{v}^T \mathbf{G} \mathbf{f} = \mathbf{v}_p \mathbf{u}_p + \mathbf{v}_q \mathbf{u}_q. \quad (11)$$

Note that Eq. (11) is a specific way of defining  $F$ , adapted to the phenomenology of building a fluid channel that can transmit allosteric interaction between two specific sites. Other choices of  $\mathbf{f}$  and  $\mathbf{u}$  could be treated similarly, such as multi-site force patterns and multi-domain dynamical modes. For example, if the response can occur at any  $(p, q)$  site at the R face, the model may describe the emergence of induced fit or conformational selection mechanisms (Sec. IV). To model the emergence of specific recognition, one sets as target strong response to a stimulus  $\mathbf{f}$ , but hardly any response to a similar ‘competitor’ stimulus  $\tilde{\mathbf{f}}$ .

In the spirit of the Metropolis algorithm used in the cylinder model, one exchanges randomly AAs between H and P while looking for changes in the fitness  $F$ . A gene  $\mathbf{c}_*$  is considered a solution if  $F$  exceeds a certain large value.<sup>12</sup> Fig. 11 illustrates the vector field  $\mathbf{u}$  of the deformation for three genes  $\mathbf{c}$ , along an evolutionary trajectory, improving the fitness value  $F$  from left to right.

### 2. The protein backbone

One hallmark of proteins is that they are made from a long chain of amino acids connected by strong covalent bonds, called a backbone (see Fig. 2). This backbone is then folded in an intricate way to form the protein, but the chain is not broken. Here, we assume that the folding process is just given and that the mutations we consider are moderate enough so that they do not change the general folding. Given this restriction, one still can ask whether the existence of the backbone affects such studies. From a conceptual point of view, having a backbone just means that some springs in the lattice are much stronger than the others, and therefore, it is not surprising that adding a backbone does not change the general picture.

<sup>12</sup> It is not reasonable to ask for  $F = \infty$ , but it suffices to look for  $F > F_{\text{crit}}$ . In our case,  $F_{\text{crit}} = 5$  is a good choice since in general, the channel will have already formed, and increasing  $F_{\text{crit}}$  will only enlarge the channel somewhat.

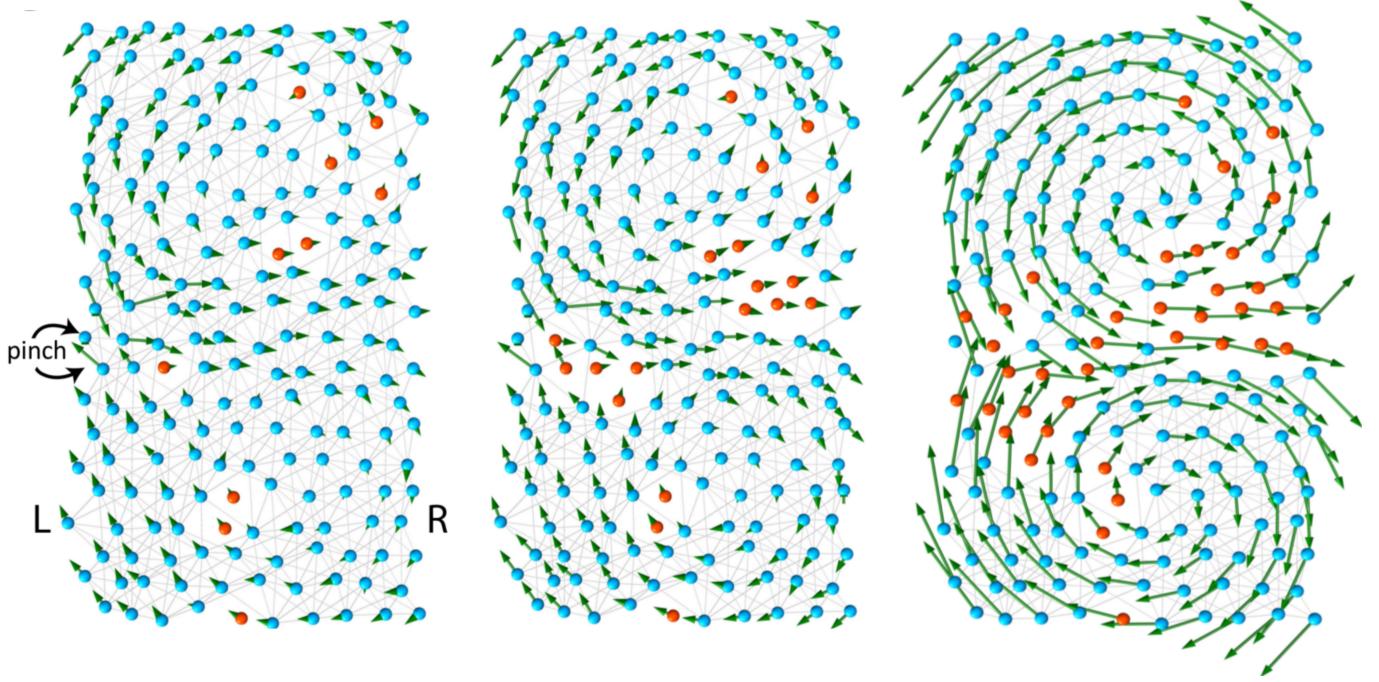


FIG. 11 Illustration of the deformation field,  $\mathbf{u}(\mathbf{c}) = \mathbf{G}(\mathbf{c})\mathbf{f}$ , Eq. (4) for three choices of  $\mathbf{c}$ . The force  $\mathbf{f} = f \cdot \mathbf{n}_{p'q'}$  is applied on the left of the lattice. The three panels show, from left to right, how the response  $\mathbf{u}$  (shown in small arrows) evolve as the network fitness  $F = \mathbf{v}^T \mathbf{u}(\mathbf{c})$ , increases. The choice of  $\mathbf{v}$  corresponds to the vertical separation of the two central points on the right.

In Fig. 12 we show two extreme cases, a serpentine backbone either parallel to the shear band or perpendicular to it. The presence of the backbone does not interfere with the emergence of a low-energy mode of the protein whose flow pattern (*i.e.*, displacement field) is similar to the backbone-less case with two eddies moving in a hinge-like fashion. In the parallel configuration, the backbone constrains the channel formation to progress along the fold (Fig. 12, left). The resulting channel is narrower than in the model without backbone (Fig. 11). In the perpendicular configuration, the evolutionary progression of the channel is much less oriented (Fig. 12, right).

We expect that, in a realistic 3D geometry, the backbone will have a weaker effect than what we observed in 2D networks, since the extra dimension adds more options to avoid the backbone constraint.

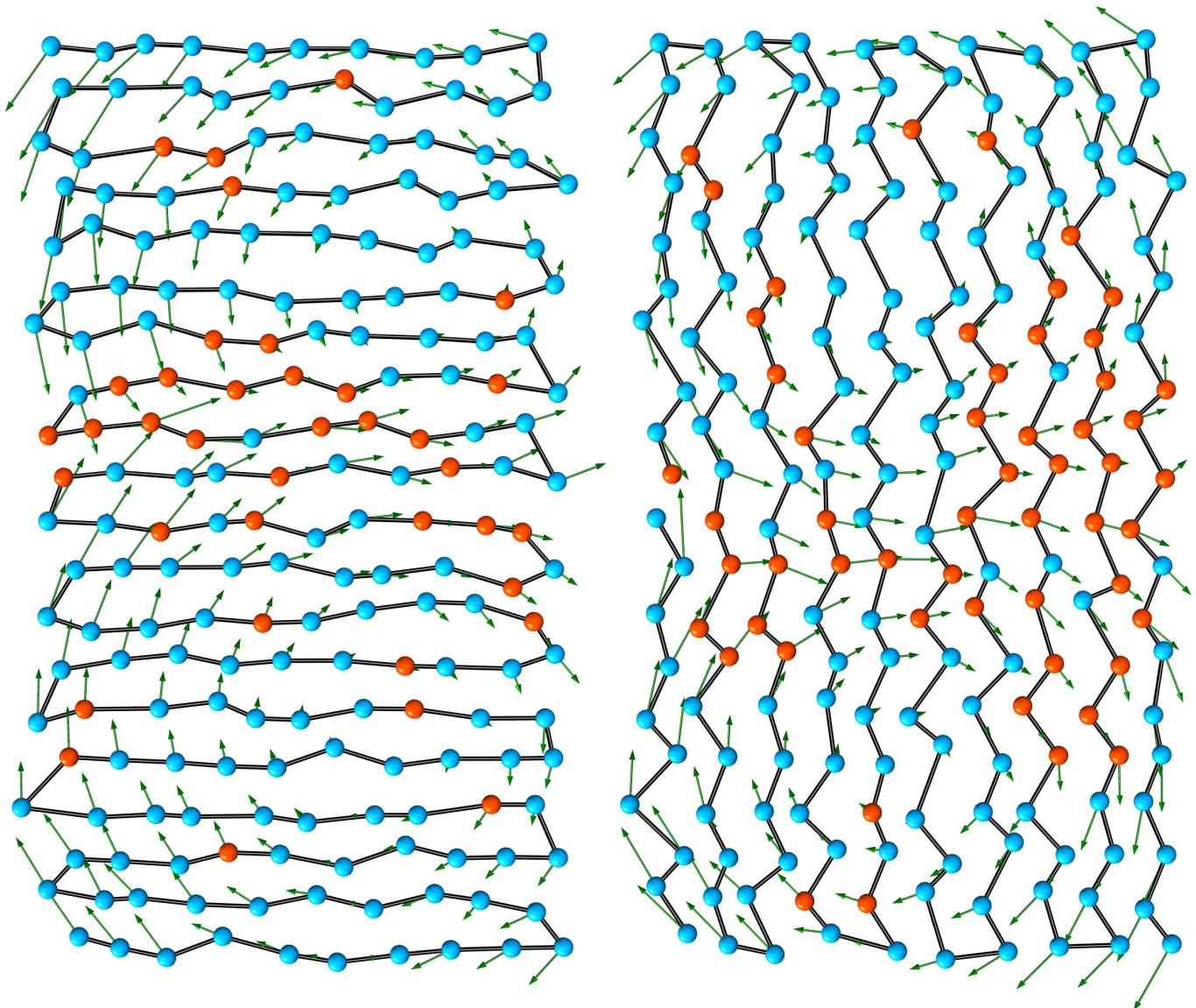
### 3. Pathologies and broken networks\*

As our criterion for evolution is the floppiness (large eigenvalue of Green's function), there is of course the trivial case where the network is just broken in two disjoint pieces or into pieces with dangling ends. Such broken networks exhibit floppy modes owing to the low energies of the relative motion of the disjoint domains with respect to each other. Any evolutionary search might end up in such non-functional unintended modes. The common pathologies one observes are:

1. isolated nodes at the boundary that become weakly connected via H→P mutations,
2. ‘sideways’ channels that terminate outside the target region (which typically include around 8–10 sites),
3. channels that start and end at the target region without connecting to the binding site.

All these are floppy modes that can vibrate independently of the pinch and cause the response to diverge ( $>F_{\text{crit}}$ ) without producing a functional mode. To avoid such pathologies, we apply the pinch force symmetrically: pinch the binding site on face L and look at responses on face R and vice versa. Thereby we not only look for the transmission of the pinch from the left to right but also from right to left. The basic algorithm is modified to accept a mutation only if it does not weaken the two-way response and enables hinge motion of the protein. This prevents the vibrations from being localized at isolated sites or unwanted channels. Of course, the presence of a backbone (see Sec. VIII.B.2 and Fig. 12) will make disconnection of the network more difficult. This is also a more realistic model.

One may also impose a stricter minimum condition,  $\delta F \geq \varepsilon F$  with a small positive  $\varepsilon$ , say 1%. An alternative, stricter criterion would be the demand that each of the terms in  $F$ ,  $\mathbf{v}_p \mathbf{u}_p$  and  $\mathbf{v}_q \mathbf{u}_q$ , increases separately.



**FIG. 12 Illustration of the backbone.** The backbone is shown as solid black serpentine curve. AAs in neighboring sites along the backbone tend to move together. We show two configurations: parallel to the channel (left) and perpendicular to the channel (right). Parallel: The backbone favors the formation of a narrow channel along the fold (compared to Fig. 11). Perpendicular: The formation of the channel is ‘dispersed’ by the backbone.

## IX. CONNECTING THE MODELS TO BIOLOGICAL CONCEPTS

The theoretical methods introduced earlier lead to a family of easily implementable numerical simulations. We discuss these simulations and show that they can explain several basic observations from the biology literature. They also suggest new connections to be explored. The main idea is that mechanical properties of the protein constrain the genetics in multiple ways.

Each subsection introduces a technique of analysis, application to one of the models described earlier, and an interpretation in terms of biological questions.

### A. Dimension of the solution set in the genotype and phenotype spaces

We describe the set of solutions for the cylinder-model of Sec. VIII.A. The (genotype) sequence space is  $\{0, 1\}^{2550}$  (see Fig. 6, bottom left). One can view this space as a 2550-dimensional hypercube, with  $2^{2550}$  corners, and any flip of a digit will move along an axis from one corner to another (the dimension of a hypercube equals the number of directions in which one can move

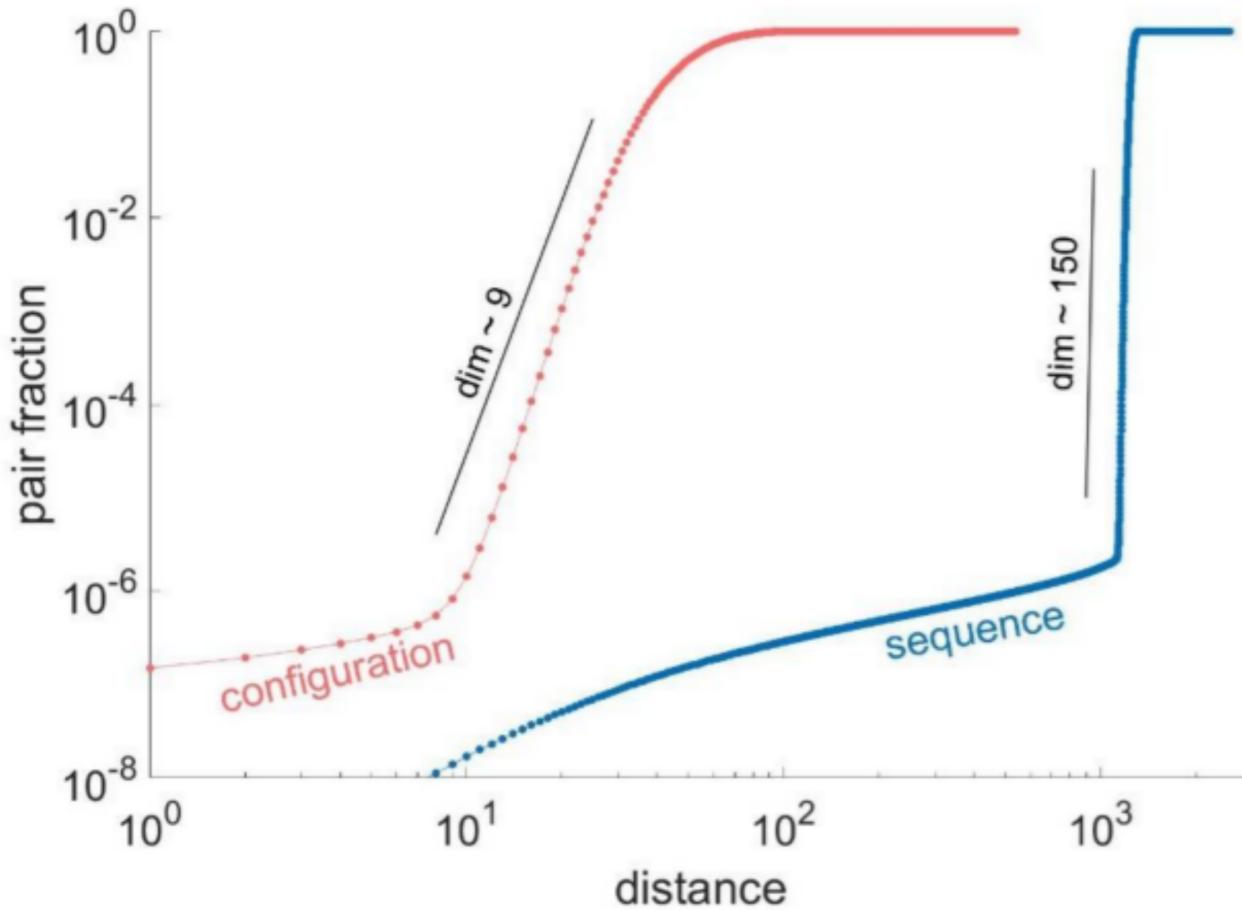


FIG. 13 Dimensional reduction of the genotype-to-phenotype map:

Dimension measurement from  $10^6$  independent configurations (phenotypes) for the cylinder-model. The dimension of the configuration space is about 9 (red curve), while in sequence space it is basically infinite (blue curve). All pairs seem to have the same distance, namely  $1275 = 2550/2$ , which is the typical distance between two random sequences.

from a corner). The space of configurations (phenotypes) is the arrangements of colors (see Fig. 6, top left), which is  $\{0, 1, 2\}^{540}$ , since there are 3 colors (red, blue, yellow).

The set  $\mathcal{S}$  of solutions to the mutation problem is a subset of this hypercube (Fig. 13). To determine the dimension of experimental data, with large sample size, it is convenient to use the box-counting algorithm [Grassberger and Procaccia 1983]. First, one counts the number  $N(\varrho)$  of pairs of points in  $\mathcal{S}$  at Hamming distances  $\leq \varrho$ , i.e., with not more than  $\varrho$  changes. One then plots  $\log N(\varrho)$  vs  $\log \varrho$  and the dimension is the slope in this log-log plot, as indicated by black lines in Fig. 13. We see that the dimension in the space of configurations (phenotypes) is about 8-9, while, in the space of sequences (genotypes), the dimension is basically ‘infinite’, namely just limited by the maximal slope one can obtain [Procaccia 1988] from the  $10^6$  simulations.<sup>13</sup> It would be interesting to discuss this problem as a special case of the problem of hitting times of small sets in hypercubes (these hitting times are usually exponentially distributed). The novelty in the current context is the use of a very small drift, namely, we do not allow steps which increase the distance to the set  $\mathcal{S}$ .<sup>14</sup>

The dramatic dimensional reduction in mapping genotypes to phenotypes stems from the different constraints that shape them [Friedlander et al. 2015; Kaneko et al. 2015; Savir and Tlusty 2013; Savir et al. 2010]. In the phenotype space, most of the protein is rigid, and only a small number of shear motions are low-energy modes, which can be described by a few degrees-of-freedom. In the genotype space, in contrast, there are many neutral mutations which do not affect the motion of the protein.

<sup>13</sup> For explanation of the flat pieces of the graph, see [Eckmann and Ruelle 1985, p. 647].

<sup>14</sup> We thank G. Ben Arous for helpful discussions on this point.

The biological interpretation is that **the gene is much more random than the phenotype of the protein it forms**. However, we shall see below, in particular in Sec. IX.F, that the gene still needs to be quite precise in certain well-defined positions. In the case of allosteric proteins these critical positions are the hinges and other locations of strong stress.

## B. Expansion of the protein universe

Here, we test the cylinder-model against the ideas of [Povolotskaya and Kondrashov 2010]. Our results will give some insight about the nature of the set of solutions, *i.e.*, genes of functional proteins. In [Povolotskaya and Kondrashov 2010], the authors consider any two solutions with gene sequences  $s_1$  and  $s_2$ . They ask how much the solution  $s_3$ , one generation after  $s_2$ , differs from  $s_1$ , and define the following observable:

Let  $x_i = (2s_{1,i} - 1) \cdot (s_{3,i} - s_{2,i})$  (since  $s_{1,i} \in \{0, 1\}$ ,  $x_i > 0$  if the change between  $s_{3,i}$  and  $s_{2,i}$  is towards  $s_1$  and  $x_i < 0$  otherwise). Finally,  $N_{\text{away}} = \#\{i : x_i < 0\}$  and  $N_{\text{towards}} = \#\{i : x_i > 0\}$ . In Fig. 14, the ratio  $N_{\text{towards}}/N_{\text{away}}$  is plotted as a function of the distance  $D$  between  $s_1$  and  $s_2$ , normalized by the diameter  $d_{\max} = 2550$ . The interested reader will notice the similarity to Fig. 3 in [Povolotskaya and Kondrashov 2010]: In their case, because of the small number of experimental samples, they only see the low- $D$  region of Fig. 14, far from the diameter of the ‘protein universe’.

The set of solutions is a very dilute, but complex, subset  $S$  of the hypercube. The search for a good gene corresponds to a slightly biased random walk along path of monotonically increasing fitness ( $\delta F \geq 0$ ). While we do not have a good mathematical description of such intricate walks, we can compare them to the null model of purely random walks. In this case, one gets a simple expression for the towards/away ratio, as a function of  $D$ , the normalized Hamming distance:  $D/(1 - D)$ , which is shown as black curve Fig. 14 (*i.e.*,  $D$  is the proportion of sites which differ between the pair of solutions  $s_1$  and  $s_2$ ). The good fit shows that the fitness-constrained evolutionary paths expand *as if one performed a random walk on the full cube*.

It is interesting to note: First, that this result must be intimately connected to the high dimension of the problem, since for low dimensional hypercubes it does not hold. Second, most samples are near the edge  $D = 1$  of the universe, where the Hamming distances among the sequences are close to the typical distance between any two random sequences. To conclude: **While maintaining functionality, the divergence of acceptable gene sequences has all aspects of a random walk (on a hypercube)**. This conclusion is close to the ‘expansion of the protein universe’ (in honor of E. Hubble), described in [Povolotskaya and Kondrashov 2010].

## C. Spectrum in phenotype and genotype spaces

Another useful method to analyze large sets of solutions is by spectral analysis in terms of Singular Value Decompositions (SVD). For the cylinder model, we have  $10^6$  binary vectors with  $n = 2550$  components each. To find the typical correlation spectrum of the solution, one forms a matrix  $W$  of size  $m \times n = 10^6 \times 2550$ . The SVD of this matrix is a generalization of the spectral decomposition of positive (semi-definite) square matrices:  $W$  is decomposed as  $U \cdot D \cdot V^T$ , where  $U$  is  $m \times m$ ,  $V$  is  $n \times n$  and  $D$  is an  $m \times n$  diagonal matrix (only the elements  $D_{ii}$  with  $i = 1, \dots, n$  are nonzero). In our case,  $m \gg n$ , which is required to obtain good statistics of the random process. The singular values  $\lambda_i^G = D_{ii}$  are in general positive and in this case the decomposition is unique. The columns of  $V$  are the (generalized) eigenvectors of  $W$ , the first few of which are shown in Fig. 15.

The singular values  $\lambda_i^G$  are the square roots of the spectrum of the covariance matrix  $W^T W$ <sup>15</sup> which has the same eigenvectors as  $W$ . Therefore, the high values correspond to the principal covariance components, the directions with maximal variation in the solution set. Mutatis mutandis, we perform the same SVD for the configurations, using the 540  $s$ -values (that is, of the shearability Eq. (9)) of vectors of the configurations.

Figure 15 illustrates the difference between the configuration space (phenotype) and the sequence space (genotype):

*Configuration space* (The eight figures on the bottom left): The first mode is proportional to the average configuration. The next modes reflect the basic deviations of the solution around this average. For example, the second mode is left-to-right shift, the third mode is expansion-contraction etc. Since, the shearable/non-shearable interface can move at most one AA sideways between consecutive rows, the modes are constrained to diamond-shaped areas in the center of the protein. This is the overlap of the influence zones of the input and output rows.

*Sequence space* (The eight figures on the bottom right): The first eigenvector is the average bond occupancy in the  $10^6$  solutions. The higher eigenvalues reflect the structure in the many-body correlations among the bonds. The typical pattern is that of ‘diffraction’ or ‘oscillations’ around the fluid channel. This pattern mirrors the biophysical constraint of constructing a rigid shell around the shearable region. Higher modes exhibit more stripes, until they become noisy, after about the tenth eigenvalue.

---

<sup>15</sup> The definition of the covariance requires to subtract the mean. Instead we project out the first eigenvalue.

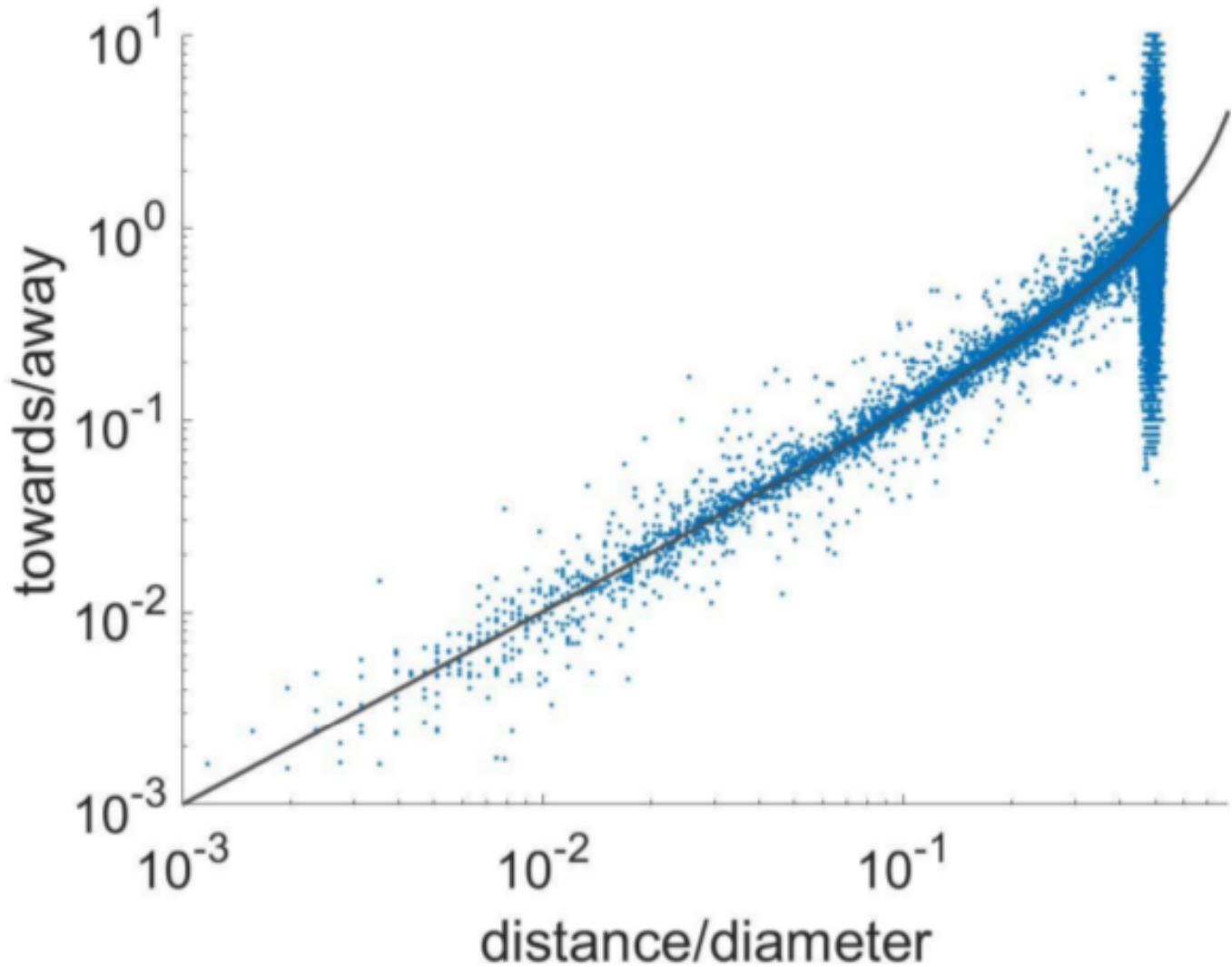


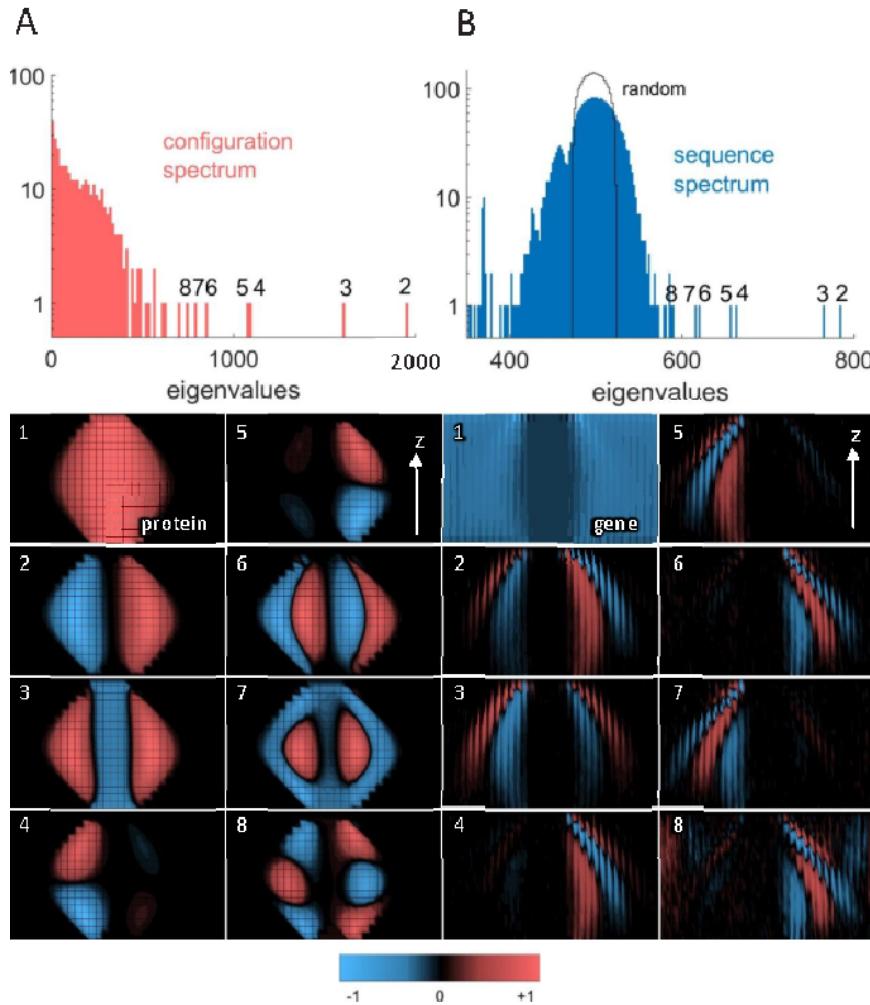
FIG. 14 Distribution of solutions in the sequence universe:

A measure for the expansion of functional genes in the sequences universe is the backward/forward ratio, the fraction of point mutations that make two sequences closer vs. the ones that increase the distance [Povolotskaya and Kondrashov 2010]. The Hamming distances  $D$  (normalized by the universe diameter  $d_{\max} = 2550$ ) show that most sequences reach the edge of the universe, where no further expansion is possible. The black curve,  $D/(1 - D)$ , is the backward/forward ratio of purely random mutations. Given the overwhelming number of samples near the maximal  $D$ , the Gaussian distribution is well visible (in the vertical direction).

The sequence-spectrum, top right in Fig. 15 has some outliers, which correspond to the localized modes shown in the eight panels below. Apart from that, the majority of the eigenvalues seem to obey the Marčenko-Pastur formula, see [Marčenko and Pastur 1967]. If the matrix is  $m \times n$ ,  $m > n$ , then the support of the spectrum is  $\frac{1}{2}(\sqrt{m} \pm \sqrt{n})$ . In our case, since we have a  $10^6 \times 2550$  matrix, one expects (if the matrices were really random) to find the spectrum at  $\frac{1}{2}(\sqrt{10^6} \pm \sqrt{2550})$ , which is close to the simulations, and confirms that most of the bonds are just randomly present or absent. The slight enlargement of the spectrum is attributed to memory effects between generations in the same branch. This corresponds to phylogenetic correlations among descendants in the same tree [Felsenstein 1985].<sup>16</sup> We conclude: **The small number of discrete eigenvalues shows that a small number of parameters characterizes both the phenotype and the non-random part of genotype of proteins.**

---

<sup>16</sup> The continuous part of the sequence spectrum, which is not quite of the standard form, could in principle be studied by taking into account the known correlations. However, even the techniques of [Guhr et al. 1998] seem difficult to implement.



**FIG. 15 Correspondence of modes in sequence (genotype) and configuration (phenotype) spaces for the cylinder-model:**

We produced the spectra by singular value decomposition of the  $10^6$  solutions.

(A) Top: the spectrum in configuration space exhibits about 8-10 eigenvalues outside the continuum (large 1<sup>st</sup> eigenvalue not shown).

Bottom: the corresponding eigenvectors describe the basic modes of the fluid channel, such as side-to-side shift (2<sup>nd</sup>) or expansion (3<sup>rd</sup>).

(B) Top: The spectrum of the solutions in sequence space is similar to that of random sequences (black line), except for about 8 to 9 high eigenvalues that are outside the continuous spectrum. (Note that the  $x$ -axis does not start at 0.)

Bottom: the first 8 eigenvectors exhibit patterns of alternating +\/- stripes – which we term correlation ‘ripples’ – around the fluid channel region. Seeing these ripples through the random evolutionary noise required at least  $10^5$  independent solutions [Teşileanu et al. 2015].

### 1. Geometry of the genotype and phenotype solution spaces\*

The  $10^6$  genotype vectors form a ‘cloud’ of points in a 2550-dimensional space. The geometry of the cloud can be explored by plotting projections along the axes defined by the eigenvectors  $\mathbf{c}_i$ ,  $i = 1, 2, \dots, 2550$ , Fig. 16. Consider for example the projection of the cloud onto the subspace spanned by  $\mathbf{c}_2$ ,  $\mathbf{c}_3$  and  $\mathbf{c}_{100}$ . The variation along the 3 axes is of comparable size. However, the equivalent projection of the 540-dimensional phenotypes along their eigenvectors  $\mathbf{u}_2$ ,  $\mathbf{u}_3$ , and  $\mathbf{u}_{100}$  shows very small variation along the vertical ( $\mathbf{u}_{100}$ ) axis, similar to the projections of a flat ellipsoid.

The projections reflect the differing shapes of the solution clouds: in the genotype space the cloud is a 2550-dimensional spheroid object, while in the phenotype space it is a flat discoid of dimension  $\sim 10$ .

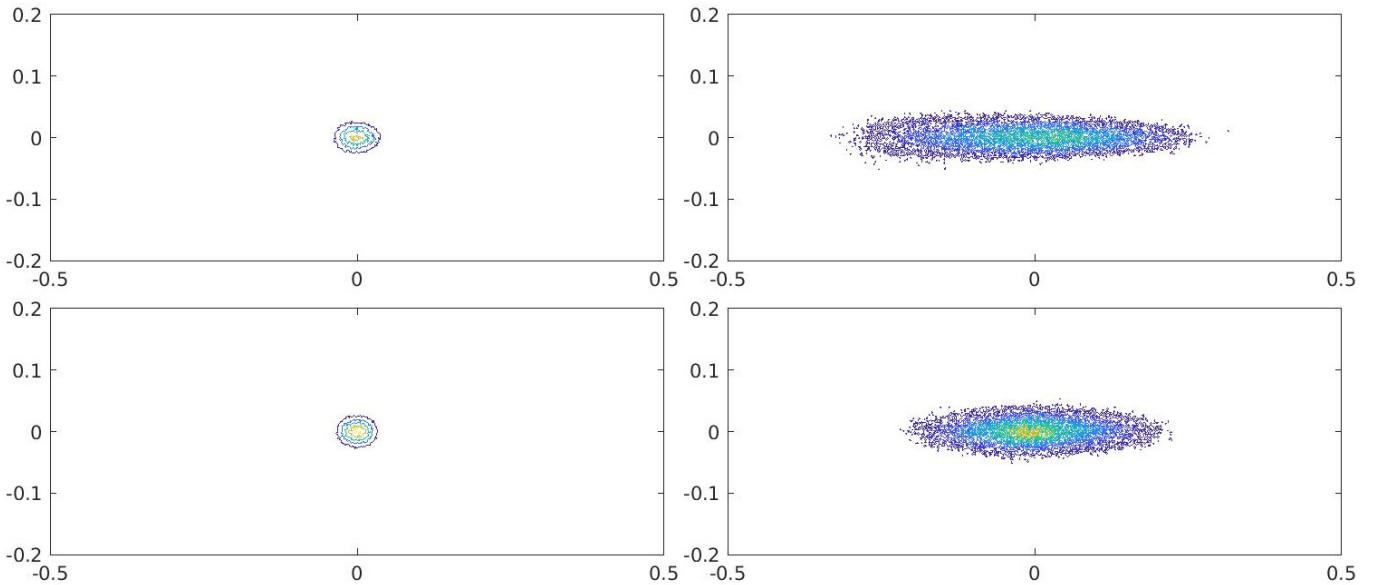


FIG. 16 The projections of the set of  $10^6$  solutions as 2550-dimensional gene sequences (left column) and corresponding 540-dimensional phenotypes (right column) onto their SVD eigenvectors  $\mathbf{c}$  and  $\mathbf{u}$ . Top row shows axes 2 (horizontal) and 100 (vertical), bottom row shows axes 3 (horizontal) and 100 (vertical). Note that the phenotypes have larger variation along their 2 and 3-components than their 100-component, unlike the genotypes which appear evenly distributed in all directions.

#### D. Stability of the mechanical phenotype under mutations

Any protein is the outcome of a long evolutionary trajectory starting from a distant ancestor. It is likely to find other descendants of the same ancestor in closely related species. In practice, one fetches the pairwise most similar proteins from a collection of species and aligns their sequences to identify homologous amino acids (all descendants of a given amino acid in the ancestor protein) [Karlin and Altschul 1993]. Once this has been accomplished, one can study mutation patterns in this multiple sequence alignment. There are two questions of interest here:

1. Which positions in the gene are *conserved*, *i.e.*, they encode the same AA?
2. Which pairs of positions are *co-varying*: a mutation at one position is frequently compensated by a mutation at the other position?

The second question, about genetic correlations, will be discussed in detail in Sec. IX.K. In this subsection, we discuss the first question, regarding AA conservation, in the context of the cylinder-model. To produce Fig. 17 one takes  $10^6$  sample solutions and mutates, for each of them, every possible position in the AA network. One then asks which mutations destroy the solution. At every position, the intensity of the blue color is proportional to the probability that a mutation at that position destroys the solution. The end of the channel is very sensitive, but also the boundaries between the channel and the bulk. This should be compared to Fig. 5 (center) where the analogous question was asked for glucokinase [Rougmont et al. in prep.], and answered by aligning 122 homologs of glucokinase. We conclude: **Sensitivity to mutations is localized near mechanically critical regions.**

#### E. Shear modes in the amino acid network

We focus here on the HP-model, although similar results hold for the cylinder-model. In Sec. VIII.B.1 we have shown a pinch stimulus  $\mathbf{f}$  leads to a deformation field  $\mathbf{u} = \mathbf{G}\mathbf{f}$ , see Eq. (4). As the fitness  $F$  of Eq. (11) improves, the system forms a fluid channel and the response field  $\mathbf{u}$  shows a hinge-like rotation, as is visible in Fig. 11. This should be compared to Fig. 3, showing the experimentally measured deformation of glucokinase. Opening a hinge in a network will not only move the two sides of the hinge but also shear the connecting bonds, especially at the hinge itself. Furthermore, remaining links near the opening of the hinge (at the opposite side of the protein) will be stretched as well. These observations can be quantified by measuring the shear. The shear  $s$  at any (lattice) point is the symmetrized derivative of the displacement field  $\mathbf{u}$ , which is computed as follows.

First, the displacement vector at the point  $\mathbf{x} \in \mathbb{R}^d$  is  $\mathbf{u}(\mathbf{x})$  with  $u_i(x) = x'_i(x) - x_i$ , the difference between the ‘new’ points



FIG. 17 The sensitivity of solutions to a single mutation, as a function of the mutation position.

$\mathbf{x}'(\mathbf{x})$  and the ‘old’ points  $\mathbf{x}$  in  $\mathbb{R}^d$ . The local deformation matrix  $\mathbf{D}(\mathbf{x})$  is then given by

$$D_{ij} = \frac{\partial x'_i(\mathbf{x})}{\partial x_j},$$

so that  $\nabla \mathbf{u}(\mathbf{x}) = \mathbf{D}(\mathbf{x}) - \mathbf{1}$ . (This matrix is also called ‘compatibility matrix’ in the review [Lubensky et al. 2015].) The shear matrix  $\boldsymbol{\varepsilon}$  is

$$\varepsilon_{ij}(\mathbf{x}) = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} + \sum_{k=1}^d \frac{\partial u_k}{\partial x_i} \cdot \frac{\partial u_k}{\partial x_j} \right).$$

In short notation

$$\begin{aligned} \boldsymbol{\varepsilon} &= \frac{1}{2} \left( (\nabla \mathbf{u}(\mathbf{x}))^\top + \nabla \mathbf{u}(\mathbf{x}) + (\nabla \mathbf{u}(\mathbf{x}))^\top \nabla \mathbf{u}(\mathbf{x}) \right) \\ &= \frac{1}{2} \left( \mathbf{C}(\mathbf{x}) - \mathbf{1} \right), \end{aligned} \tag{12}$$

with  $\mathbf{C}(\mathbf{x}) = \mathbf{D}(\mathbf{x})^\top \mathbf{D}(\mathbf{x})$ , which is the metric of the coordinate transformation.

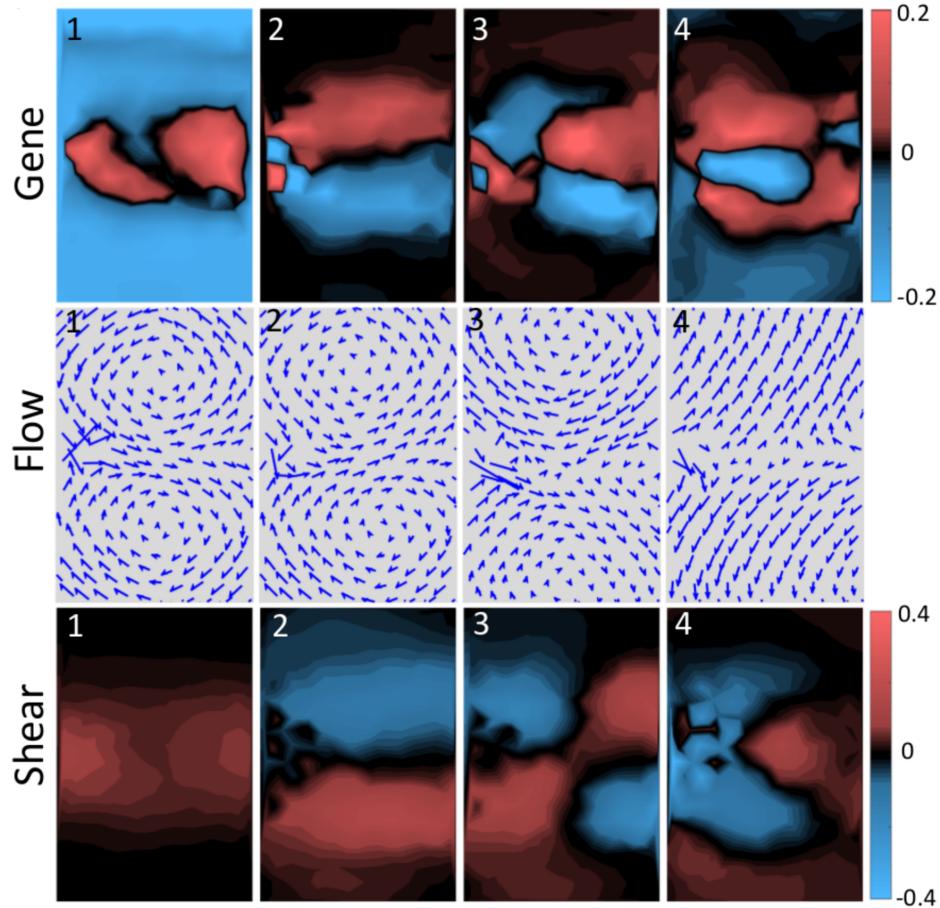
As a measure of the magnitude of the shear one can use

$$\begin{aligned} s(\mathbf{x}) &= \text{Tr}(\boldsymbol{\varepsilon}^2) - \frac{1}{d} (\text{Tr}(\boldsymbol{\varepsilon}))^2 \\ &= \text{Tr} \left( \boldsymbol{\varepsilon} - \frac{1}{d} \text{Tr}(\boldsymbol{\varepsilon}) \cdot \mathbf{1} \right)^2, \end{aligned} \tag{13}$$

which is the square of the Frobenius norm ( $L^2$ ) of the traceless part of  $\boldsymbol{\varepsilon}$ .<sup>17</sup> The trace of  $\boldsymbol{\varepsilon}$  is related to the isotropic dilation, which in the protein is a smaller effect than the shear, and therefore will not be further considered.<sup>18</sup>

<sup>17</sup> As all norms on finite dimensional spaces are equivalent, other norms amount basically just to a rescaling.

<sup>18</sup> There are many variants of the shear calculation see, e.g., [McGinty 2012–].



**FIG. 18 The vector fields for the HP-model:** The first 4 eigendirections for the three vector fields,  $k = 1, \dots, 4$ .  
 Top: The first four SVD eigenvectors of the gene  $C_k$ ,  
 Center: The corresponding displacement flow field  $U_k$ ,  
 Bottom: The corresponding shear intensity  $S_k$ .

### 1. Implementation in the case of protein structure data\*

As proteins are discrete objects, we replace the derivatives by difference operators [Gullett et al. 2008; Mitchell et al. 2016].

We consider the crystallographic data of two conformations of a given protein (for example the PDB structures 1v4s and 1v4t in Fig. 3 [Kamata et al. 2004]). To produce Fig. 3 from these data [Rougemont et al. in prep.], we take a ball of radius  $\varrho$  of about 10 Å around each atom  $X$  in the protein.<sup>19</sup> This will encompass  $m = m(X)$  other atoms, at positions  $\mathbf{r}_i \in \mathbb{R}^3$ ,  $i = 1, \dots, m$ . Let  $\mathbf{r}_0$  be the coordinates of  $X$ , and let  $\mathbf{A}(X)$  be the  $m \times 3$  matrix of the  $m$  distance vectors  $\mathbf{r}_i - \mathbf{r}_0$ , in the first configuration (*i.e.*, one of the PDB structures). Let  $\mathbf{B}(X)$  be the analogous matrix for the second configuration and compute

$$\mathbf{Q} = \frac{1}{2}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T (\mathbf{B} \mathbf{B}^T - \mathbf{A} \mathbf{A}^T) \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1}. \quad (14)$$

The matrix  $\mathbf{Q}$  is an approximation of  $\boldsymbol{\epsilon}$  and is obtained by observing that  $\mathbf{B} = \mathbf{A} \mathbf{D}^T$ :

$$\mathbf{C} = \mathbf{D}^T \mathbf{D} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{B} \mathbf{B}^T \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1}.$$

By substituting this into Eq. (12), we verify that Eq. (14) holds. With this approximate shear tensor  $\boldsymbol{\epsilon}$  one computes its magnitude  $s(\mathbf{r}_0)$  using Eq. (13).

The discrete approximation of the shear field in glucokinase is shown in the leftmost panel of Fig. 5 (with  $m$  typically around 50). The strain is found to be large in the hinge of the protein, and also at a somewhat loose outer surface.<sup>20</sup> In Fig. 18, the

<sup>19</sup> It usually suffices to look at atoms N, C, O, along the backbone, while one may also include sidechains.

<sup>20</sup> The dilatation (the trace) is much smaller.

corresponding shear magnitude fields  $s(\mathbf{x})$  are shown for the HP-model, using a statistical average over many solutions (see Sec. IX.E.2 for more details). One again observes strong shear in the hinge. We conclude: **shear is critical in the hinges among moving domains of the protein.**

## 2. Details of shear computation\*

We describe here in detail the procedure of calculating the shear in the HP-model, leading to Fig. 18. This figure shows averages over many realizations of the random process, in the following sense. One starts with  $10^6$  solutions.<sup>21</sup> Each solution  $\mathbf{c}_*$  together with the (fixed) pinch  $\mathbf{f}$ , defines 3 vectors

1. the gene of the functional protein,  $\mathbf{c}_*$ , (a vector of length  $n_a = 200$  codons),
2. the flow field (displacement),  $\mathbf{u}(\mathbf{c}_*) = \mathbf{G}(\mathbf{c}_*)\mathbf{f}$ , (a vector of length  $n_d = 400$  of the  $x$  and  $y$  velocity components),
3. the shear field  $\mathbf{s}(\mathbf{c}_*)$  (a vector of length  $n_a = 200$ ).

The  $10^6$  solutions are then written as three corresponding matrices  $W_C$ ,  $W_U$  and  $W_S$ , of size  $200 \times 10^6$  resp,  $400 \times 10^6$ , where each row of these matrices is one of  $\mathbf{c}_*$ ,  $\mathbf{u}(\mathbf{c}_*)$ , and  $\mathbf{s}(\mathbf{c}_*)$ .

Next, one calculates the singular eigenvalues and corresponding eigenvectors of the three matrices (using SVD, as in Sec. IX.C) and isolates the leading eigenvalues. The central row in Fig. 18 shows that the flow field can be decomposed into successively weaker motions  $\mathbf{U}_k$ , with the strongest being a rotating hinge motion around the fluid channel. One should note that evolution in this case did not impose this *global* rotation, but only the *localized* response to a pinch on the left side of the sample.

## F. Similarity of gene and shear

The results for the HP-model reveal a tight relation between the gene fields  $\mathbf{C}_k$  and the shear intensities  $\mathbf{S}_k$ , as shown in Fig. 18. Comparing the top and bottom rows, one observes a similar structure of the corresponding eigenfunctions.<sup>22</sup> A similar relation is visible in Fig. 15 for the cylinder-model. The functions of many proteins are known to involve large-scale motions of the amino acid network, such as hinge rotation, shear sliding, or twists [Gerstein et al. 1994]. Recently, the strain that occurs during such conformations changes was computed in several proteins by comparing structures obtained from X-ray and NMR studies [Mitchell et al. 2016]. However, the tight correspondence between the shear tensor and the genetic correlations, that we observe here (Fig. 18) has not yet been measured in real protein. In principle, one would need to follow a procedure similar to the one presented here: first, to calculate the mechanical shear using the methods of [Gullett et al. 2008; Mitchell et al. 2016], and then to compare it to the genetic correlations from sequence alignment [Rougemont et al. in prep.].

## G. Point mutations are localized mechanical perturbations

A mutation in the HP-model may vary the strength of no more than  $z = 12$  bonds around the mutated AA (Fig. 10). The corresponding perturbation of the Hamiltonian  $\delta\mathbf{H}$  is therefore localized, akin to a defect in a crystal [Elliott et al. 1974; Tewary 1973]. The mechanics of mutations can be further explored by examining perturbations of Green's function,  $\mathbf{G}' = \mathbf{G} + \delta\mathbf{G}$ . They obey the Dyson equation and the Dyson series, Eq. (6)–(7). This series has a straightforward physical interpretation as a sum over multiple scatterings (Fig. 23B): As a result of the mutation, the elastic force field is no longer balanced by the imposed force  $\mathbf{f}$ , leaving a residual force field  $\delta\mathbf{f} = \delta\mathbf{H}\mathbf{u} = \delta\mathbf{H}\mathbf{G}\mathbf{f}$ . The first scattering term in the series balances  $\delta\mathbf{f}$  by the deformation  $\delta\mathbf{u} = \mathbf{G}\delta\mathbf{f} = \mathbf{G}\delta\mathbf{H}\mathbf{G}\mathbf{f}$ . Similarly, the second scattering term accounts for further deformation induced by  $\delta\mathbf{u}$ , and so forth.<sup>23</sup> We conclude: **Standard expansions of Green's functions correspond to hierarchical organization of the effects of mutations in terms of multiple scattering.**

## H. Mechanical function emerges as a sharp transition

As the evolution reaches a solution gene  $\mathbf{c}_*$ , there emerges a new (almost) zero energy-mode,  $\mathbf{u}_*$ , in addition to the Galilean symmetry modes (which we already projected away). As the other eigenvalues of  $\mathbf{G}(\mathbf{c}_*)$  remain typically distant from this small

<sup>21</sup> The characteristics we are looking for do not show cleanly unless there are at least  $10^5$  samples.

<sup>22</sup> One could in principle measure the distance between the corresponding pairs using an  $L^2$  norm over the whole area.

<sup>23</sup> In problems of this local nature, calculating a mutated Green's function using the Woodbury formula Eq. (5) accelerates the computation by a factor of  $\sim 10^4$  as compared to standard matrix inversion.

eigenvalue  $\lambda_*$ , there will be a gap between  $\lambda_*$  and the rest of the spectrum. While we do not have a proof of that such a gap should appear, this is found to be the generic case in the models described here. The response to a pinch will be mostly through this soft mode, as we show now.

Consider a sequence of mutations  $\mathbf{c}_k$  which converges to  $\mathbf{c}_*$  (in the Hamming distance) as  $k \rightarrow k_*$ . The corresponding sequence of fitness values is  $F_k = F(\mathbf{c}_k)$ . For the HP-model with the pinch introduced earlier, the fitness is (Eq. (11)):

$$F_k = \mathbf{v}^T \mathbf{u}(\mathbf{c}_k) = \mathbf{v}^T \mathbf{G}(\mathbf{c}_k) \mathbf{f}.$$

When  $\mathbf{c}_k$  gets closer to  $\mathbf{c}_*$ , the almost-zero eigenvalue  $\lambda_k$  of  $\mathbf{H}(\mathbf{c}_k)$  will dominate Green's function,  $\mathbf{G}(\mathbf{c}_k) = \mathbf{H}(\mathbf{c}_k)^\dagger$

$$\mathbf{G}(\mathbf{c}_k) \simeq \frac{1}{\lambda_k} |\mathbf{u}(\mathbf{c}_k)\rangle \langle \mathbf{u}(\mathbf{c}_k)| \sim \frac{1}{\lambda_k} |\mathbf{u}_*\rangle \langle \mathbf{u}_*|.$$

The fitness sequence is therefore

$$F_k \simeq \frac{(\mathbf{v}^T \mathbf{u}_*) (\mathbf{u}_*^T \mathbf{f})}{\lambda_k}. \quad (15)$$

On average, the fitness increases exponentially with the number of beneficial mutations as shown in Fig. 19. The growth of the fitness follows the formation of the channel and the narrowing of the remaining rigid ‘neck’, in the middle of the channel. We lack, however, a quantitative explanation for the generic exponential dependence, which is probably related to the structure of the Hamiltonian  $\mathbf{H}$ .<sup>24</sup> In the particular instance of the two models considered here, one can argue that the mutations which improve the channel, all act multiplicatively on the fitness  $F$ . Since this discussion is ‘spectral,’ we expect it to hold for models with more colors (*i.e.*, AAs) than the HP model.

As noted in Fig. 9, beneficial mutations are rare, and are separated by long stretches of neutral mutations. One may ask where the neutral mutations take place, and this is illustrated in Fig. 20, which shows that, in most sites, the effect of mutations is practically neutral.

The vanishing of the spectral gap,  $\lambda_k \rightarrow 0$ , can further be viewed as a topological transition in the system: the AA network is being divided into two domains that can move independently of each other at low energetic cost. The relative motion of the domains defines the emergent soft mode and the collective degrees-of-freedom, for example the rotation of a hinge or the shear angle.

The soft mode appears at a dynamical transition, where the average shear in the protein jumps abruptly as the channel is formed and the protein can easily deform in response to the force probe (Fig. 21). The trajectories are plotted as a function of  $p$ , the fraction of AAs of type P. The distribution of the critical values  $p_c$  is rather wide owing to the random initial conditions and finite-size effects.

Another connection is provided by the Kirchhoff matrix-tree theorem (see *e.g.*, [Chaiken 1982; Tutte 1948]). Let  $\mathbf{M}$  be an  $n$ -by- $n$  graph Laplacian, with links  $i \leftrightarrow j$  given by  $\mathbf{M}_{ij} = -1$  and  $\sum_j \mathbf{M}_{ij} = 0$  for all  $i$ . The matrix  $\mathbf{M}$  has an eigenvalue 0, with eigenvector  $(1, 1, \dots, 1)$ . Take the submatrix  $\mathbf{M}_0$  where one row and one column are omitted. In analogy with the weighted links of the HP-model, assume now that  $\det \mathbf{M}_0 = 0$ , *i.e.*, that a second eigenvalue vanishes. Then, by the Kirchhoff theorem, the number of spanning trees of the full graph is equal to 0. In other words, the graph is disconnected, in analogy to the formation of the fluid channel.

## I. Correlation and alignment

As the shear band (fluid channel) is taking shape, the correlation among codons builds up. To see this, we align genes from the  $10^6$  simulations, in analogy to sequence alignment of real protein families [de Juan et al. 2013; Göbel et al. 1994; Halabi et al. 2009; Hopf et al. 2017; Jones et al. 2012; Lockless and Ranganathan 1999; Marks et al. 2011; Poelwijk et al. 2017; Suel et al. 2003; Teşileanu et al. 2015]. At each time step we calculate the two-codon correlation  $Q_{ij}$  between all pairs of codons  $c_i$  and  $c_j$ ,

$$Q_{ij} \equiv \langle c_i c_j \rangle - \langle c_i \rangle \langle c_j \rangle, \quad (16)$$

where brackets denote ensemble averages. One finds that most of the correlation is concentrated in the region where the channel will form. In Fig. 22 one sees that the average correlation is tenfold larger in the channel than in the whole protein. Within the channel, the correlation is long-range, and propagates from side to side in the protein (see [Dutta et al. 2018] for a figure).

---

<sup>24</sup> Note that the exponential increase is much stronger than what could be explained by the choice of the factor  $\delta F > \varepsilon F$  of Sec. VI.

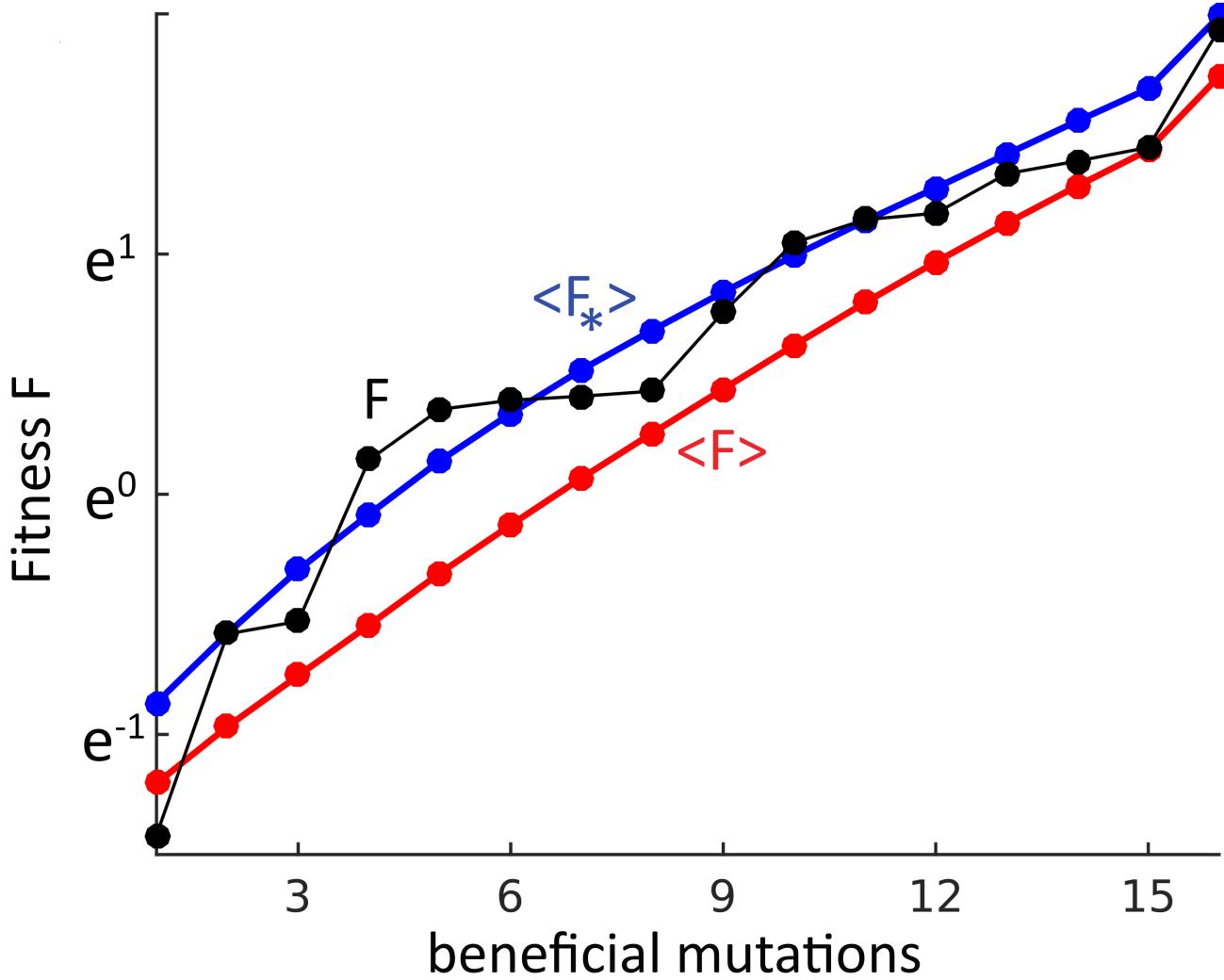


FIG. 19 Progression of the fitness  $F$  corresponding to the evolution of Fig. 11 (black). The fitness trajectory averaged over  $\sim 10^6$  runs  $\langle F \rangle$  is shown in red. Shown are the last 16 beneficial mutations towards the formation of the channel. The contribution of the emergent low-energy mode  $\langle F_k \rangle$  alone, shown in blue, dominates the fitness (according to Eq. (15)).

Analogous correlated domains containing functionally-related amino acids that co-evolve appear in real protein families [Halabi et al. 2009; Lockless and Ranganathan 1999; Suel et al. 2003; Teşileanu et al. 2015], as well as in coarse-grained models of protein allostericity [Flechsig 2017; Hemery and Rivoire 2015; Tlusty 2016; Tlusty et al. 2017] and allosteric matter [Rocks et al. 2017; Yan et al. 2017].

We conclude: **Genetic correlations are significantly larger in the mechanically important regions.**

#### J. Conserved amino acids

In this section, we discuss single mutations (and the lack thereof), while in the next, we discuss the case where one mutation is ‘compensated’ by another (this is called epistasis). Both phenomena are intimately related to those sites on the protein which matter for the function: These are the sites which are mechanically important.

In the cylinder model (Fig. 17), mutations near the top edge and the boundary of the fluid channel have the most deleterious effect on the mechanical function (dark regions in the figure). Therefore, to preserve the functionality of the protein in this model, these sensitive amino acids are also conserved more than average among the solutions.

In comparison, the center panel of Fig. 5 highlights the most conserved positions among the 122 aligned homologs of the real protein glucokinase. Similar to the model, here also conservation appears to be correlated with mechanical importance, as

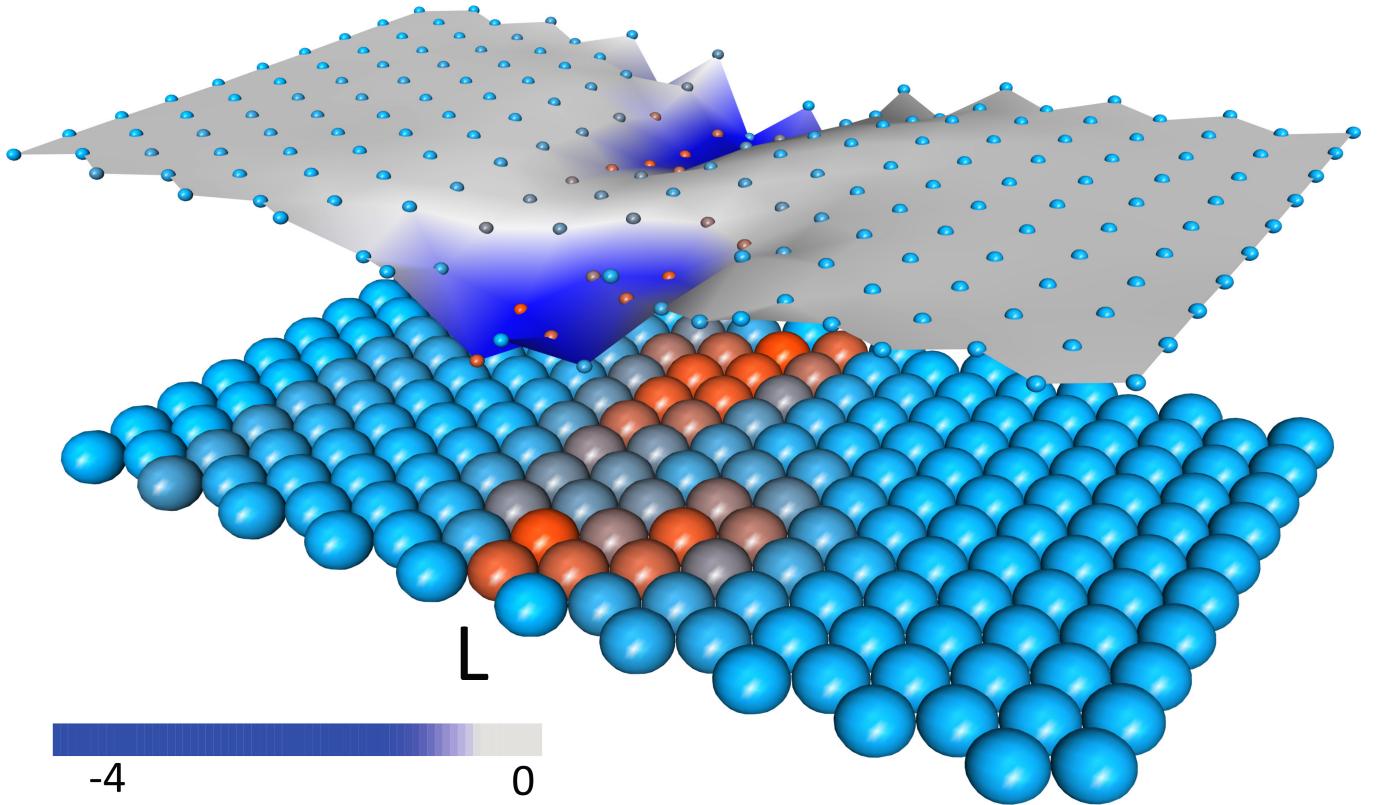


FIG. 20 Landscape of the fitness change  $\delta F = \mathbf{v}^T \delta \mathbf{G} \mathbf{f}$ , averaged over  $10^6$  solutions, for all 200 possible positions of point mutations at a solution. Underneath, the average AA configuration of the protein is shown in shades of red (P) and blue (H). In most sites, mutations are neutral, while mutations in the channel are, on average, deleterious (blue, below the flat surface).

measured by the magnitude of the shear (left panel of Fig. 5). We conclude: **Mechanically critical regions of a protein are sensitive to mutation.** If, however, a mutation does occur at such a sensitive amino acid, then it should be compensated by another one (or a few). This is dealt with in the next section.

## K. Epistasis links protein mechanics to genetic correlations

The correlations among amino acids in the gene exhibit tight correspondence to the pattern of the shear field (see Sec. IX.I and Fig. 18). We now discuss how to link these genetic correlations among mutations to the physical interaction in the amino acid network. The procedure will be similar to how the effect of a single mutation was interpreted in terms of a scattering expansion of Green's function (Sec. IX.G).

In genetics, the term epistasis refers to departure of fitness from additivity in the effect of combined mutations owing to *inter-genetic* interaction. For example, the phenotypic effect of one gene may be masked by a different gene [Cordell 2002; Mackay 2014; Phillips 2008]. In analogy, on the smaller scale of a single gene described here, *intra-genetic* epistasis is non-additivity of protein fitness owing to the non-linear interaction among its amino acids [Breen et al. 2012; Clark and Wang 1997; Harms and Thornton 2013; Ortlund et al. 2007]. For example, one mutation can be compensated by another one in order to keep the protein functional. This second mutation can be far away on the gene sequence. The use of Green's functions allows for a calculable definition of epistasis in terms of the Dyson series Eq. (6).

Algorithmically, one takes one functional solution obtained from the evolution algorithm and mutates one AA at a site  $i$ . This mutation induces a change  $\delta \mathbf{G}_i$  as the difference of the new and the old Green's function. Then,

$$\delta F_i = \mathbf{v}^T \delta \mathbf{G}_i \mathbf{f}$$

is the change of the observable fitness  $F$  (which can be computed by Eq. (5)). One can similarly perform another, independent mutation at a site  $j \neq i$ , producing a second deviation,  $\delta \mathbf{G}_j$  and  $\delta F_j$  respectively. Finally, starting again from the original solution, one mutates both  $i$  and  $j$  simultaneously, with combined effects  $\delta \mathbf{G}_{i,j}$  and  $\delta F_{i,j}$ . It is then natural to define the epistasis  $e_{i,j}$  as

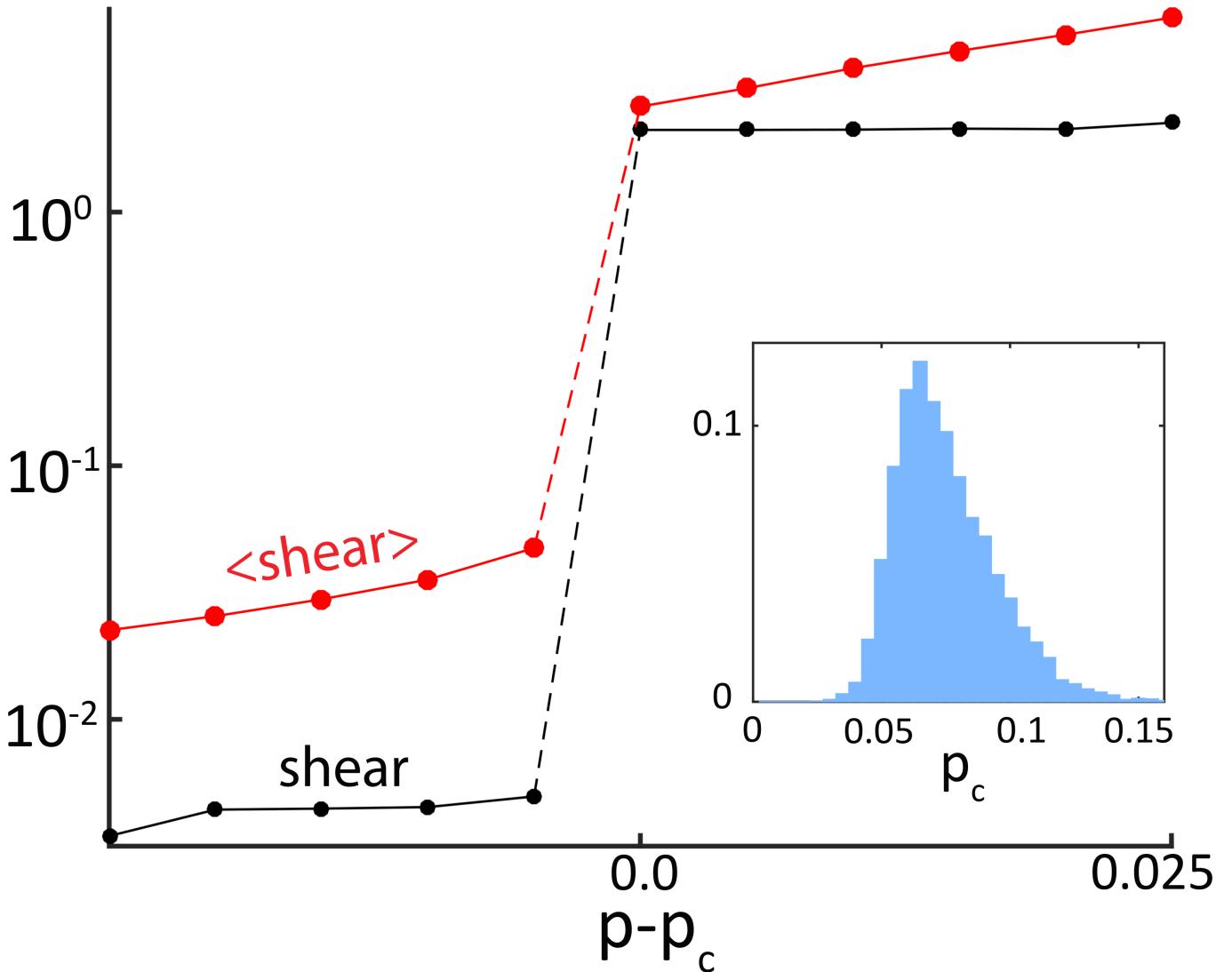


FIG. 21 The mean shear in the protein in a single run (black) and averaged over  $10^6$  samples (red) as a function of the fraction  $p$  of P-amino acids. The values of  $p$  are shifted by the position of the jump,  $p_c$ . Inset: Distribution of  $p_c$ .

the departure of the double mutation from additivity of two single mutations,

$$e_{i,j} \equiv \delta F_{i,j} - \delta F_i - \delta F_j . \quad (17)$$

The epistasis  $e_{i,j}$  is simply the inner product value of this nonlinearity with the pinch and the response,

$$e_{i,j} = \mathbf{v}^T (\delta \mathbf{G}_{i,j} - \delta \mathbf{G}_i - \delta \mathbf{G}_j) \mathbf{f} . \quad (18)$$

Eq. (18) shows how epistasis is directly related to mechanical forces among mutated AAs.

To evaluate the average epistatic interaction among amino acids in the HP-model, we perform the double mutation calculation for all  $10^6$  solutions and take the ensemble average  $E_{ij} = \langle e_{i,j} \rangle$ . Landscapes of  $E_{ij}$  show significant epistasis in the channel (Fig. 24). AAs outside the high shear region show only small epistasis, since mutations in the rigid domains hardly change the elastic response. The epistasis landscapes (Fig. 24A-C) are mostly positive since the mutations in the channel interact antagonistically [Desai et al. 2007]: after a strongly deleterious mutation, a second mutation has a smaller effect.

In the gene, epistatic interactions are manifested in codon correlations [Hopf et al. 2017; Poelwijk et al. 2017] shown in Fig. 24D, which depicts two-codon correlations  $Q_{ij}$  of Eq. (16) from the alignment of  $10^6$  functional genes  $\mathbf{c}_*$ . We find a tight correspondence between the mean epistasis  $E_{ij} = \langle e_{i,j} \rangle$  and the codon correlations  $Q_{ij}$ . Both patterns exhibit strong correlations in the channel region with a period equal to channel's length, 10 AAs. The similarity in the patterns of  $Q_{ij}$  and  $E_{ij}$  indicates

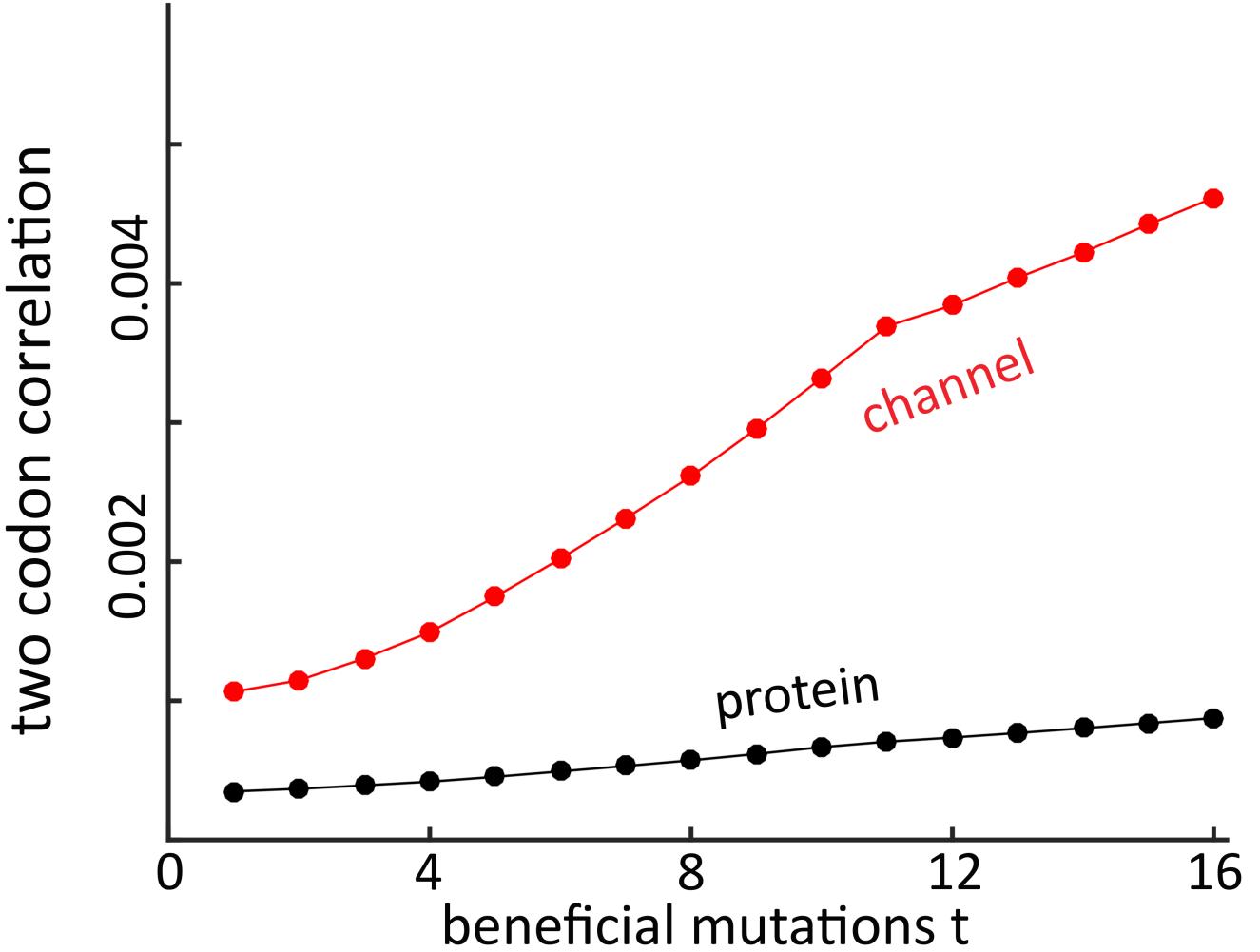


FIG. 22 The average magnitude of the two-codon correlation  $|Q_{ij}|$  Eq. (16) as a function of the number of beneficial mutations,  $t$ . The red curve shows  $|Q_{ij}|$  in the shear band (AAs in rows 7–13, of Fig. 11). The black curve shows  $|Q_{ij}|$  for the whole protein. The correlations in the channel are clearly larger.

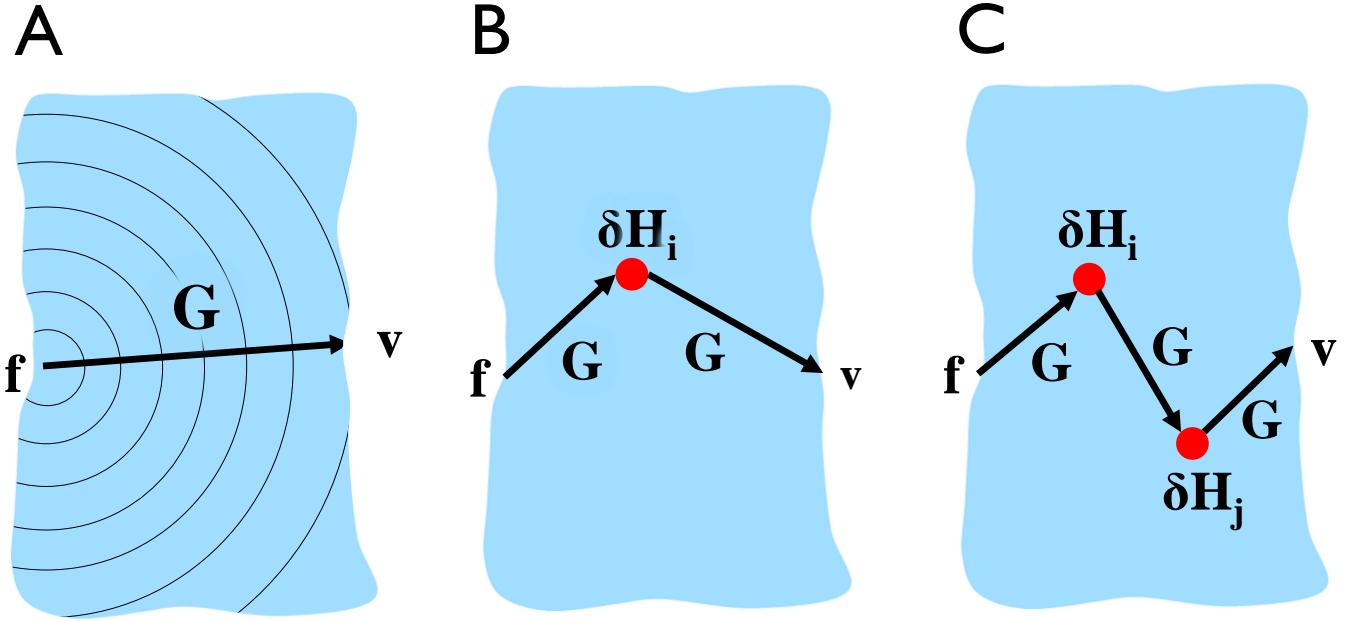
that a major contribution to the long-range, strong correlations observed among aligned protein sequences stems from the mechanical interactions propagating through the amino acid network.

## L. Epistasis as a sum over scattering paths

One can classify epistasis according to the interaction range. Neighboring AAs exhibit *contact epistasis* [Göbel et al. 1994; Marks et al. 2011], because two adjacent perturbations,  $\delta\mathbf{H}_i$  and  $\delta\mathbf{H}_j$ , interact nonlinearly via the ‘and’ gate of the interaction table of Fig. 10,  $\delta^2\mathbf{H}_{i,j} \equiv \delta\mathbf{H}_{i,j} - \delta\mathbf{H}_i - \delta\mathbf{H}_j \neq 0$  (where  $\delta\mathbf{H}_{i,j}$  is the perturbation by both mutations). In the case of contact epistasis, the leading term in the Dyson series Eq. (6) of  $\delta^2\mathbf{G}_{i,j}$  is a single scattering from an effective perturbation with an energy  $\delta^2\mathbf{H}_{i,j}$ , which yields the epistasis

$$e_{i,j} = -\mathbf{v}^T (\mathbf{G} \delta^2\mathbf{H}_{i,j} \mathbf{G}) \mathbf{f} + \dots .$$

*Long-range epistasis* among non-adjacent, non-interacting perturbations ( $\delta^2\mathbf{H}_{i,j} = 0$ ) is observed along the channel (Fig. 24). In this case, Eq. (6) expresses the nonlinearity  $\delta^2\mathbf{G}_{i,j}$  as a sum over multiple scattering paths which include both  $i$  and  $j$



**FIG. 23 Force propagation, mutations and epistasis.** (A) Green's function  $G$  measures the propagation the mechanical signal across the protein (blue) from the force source  $f$  (pinch) to the response site  $v$ , depicted as a ‘diffractio wave’. (B) A mutation  $\delta\mathbf{H}_i$  deflects the propagation of force. The effect of the mutation on the propagator  $\delta G$  can be described as a series of multiple scattering paths Eq. (6). The diagram shows the first scattering path,  $\mathbf{G} \delta\mathbf{H}_i \mathbf{G}$ . (C) Epistasis is the departure from additivity of the combined fitness change of two mutations. The epistasis between two mutations,  $\delta\mathbf{H}_i$  and  $\delta\mathbf{H}_j$ , is equivalent to a series of multiple scattering paths Eq. (19). The diagram shows the path  $\mathbf{G} \delta\mathbf{H}_i \mathbf{G} \delta\mathbf{H}_j \mathbf{G}$ .

(Fig. 23C),

$$e_{i,j} = \mathbf{v}^T (\mathbf{G} \delta\mathbf{H}_i \mathbf{G} \delta\mathbf{H}_j \mathbf{G} + \mathbf{G} \delta\mathbf{H}_j \mathbf{G} \delta\mathbf{H}_i \mathbf{G}) \mathbf{f} - \dots . \quad (19)$$

The perturbation expansion further links long-range epistasis to shear deformation: Near the transition at which the function emerges, Green's function is dominated by the single soft mode,  $\mathbf{G} \simeq \mathbf{u}_* \mathbf{u}_*^T / \lambda_*$ , with fitness  $F$  given by Eq. (15). From Eq. (6) and Eq. (18), one deduces a simple expression for the mechanical epistasis as a function of the shear,

$$e_{i,j} \simeq F \cdot \left( \frac{h_i}{1+h_i} + \frac{h_j}{1+h_j} - \frac{h_i+h_j}{1+h_i+h_j} \right) . \quad (20)$$

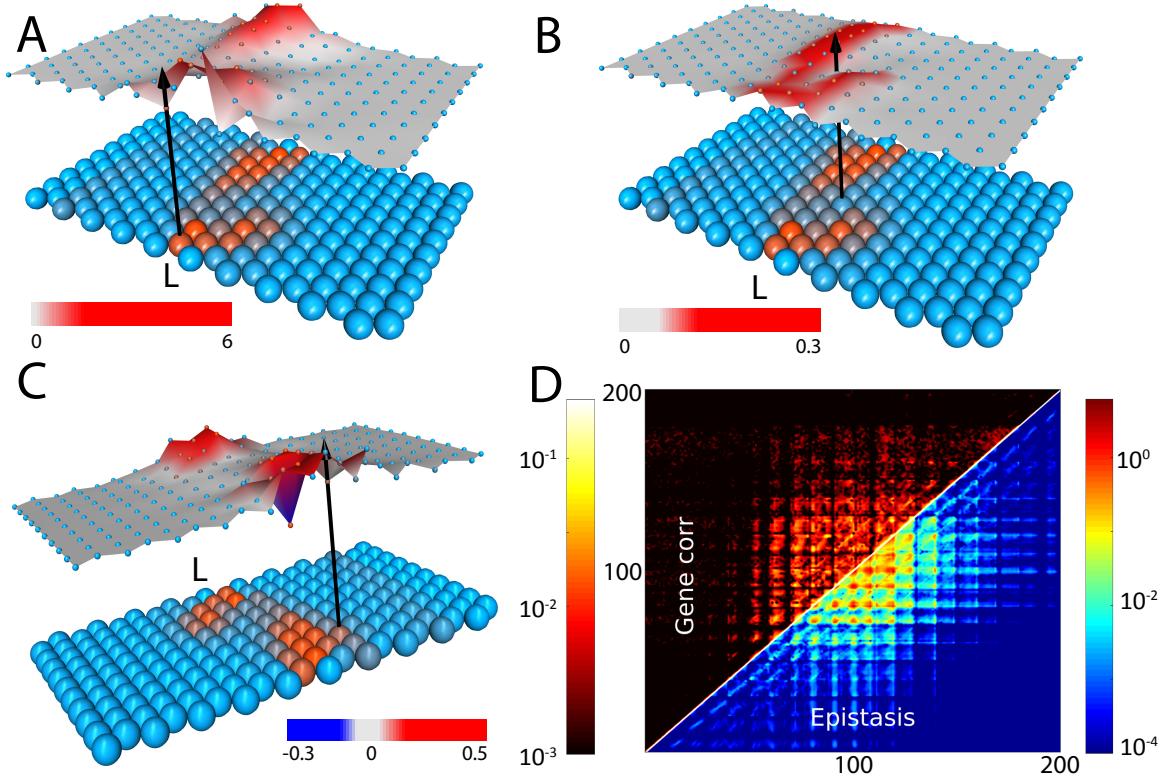
The factor  $h_i \equiv \mathbf{u}_*^T \delta\mathbf{H}_i \mathbf{u}_* / \lambda_*$  in Eq. (20) is the ratio of the change in the shear energy due to mutation at  $i$  (the expectation value of  $\delta\mathbf{H}_i$ ) and the energy  $\lambda_*$  of the soft mode, and similarly for  $h_j$ . Thus,  $h_i$  and  $h_j$  are significant only in and around the shear band, where the bonds varied by the perturbations are deformed by the soft mode.

When both sites are outside the channel,  $h_i, h_j \ll 1$ , the epistasis Eq. (20) is small,  $e_{i,j} \simeq 2h_i h_j F$ . It remains negligible even if one of the mutations,  $i$ , is in the channel,  $h_j \ll 1 \ll h_i$ , and  $e_{i,j} \simeq h_j F$ . Epistasis can only be long-ranged along the channel when both mutations are significant,  $h_i \gg 1$  and  $h_j \gg 1$ , and  $e_{i,j} \simeq 1$ . It follows that Eq. (20) can be roughly approximated as

$$e_{i,j} \simeq F \cdot \min(1, h_i) \cdot \min(1, h_j) . \quad (21)$$

We conclude that epistasis is maximal when both sites are at the start or end of the channel, as illustrated in Fig. 24. The nonlinearity of the fitness function gives rise to antagonistic epistasis since the combined effect of two deleterious mutations is non-additive as either mutation is enough to diminish the fitness.

As evident from Eq. (21), the epistasis matrix Eq. (20) is approximately a rank-one tensor  $e_{i,j} \sim |\mathbf{e}\rangle \langle \mathbf{e}|$ , with a single dominant eigenvector,  $\mathbf{e}_i \sim \min(1, h_i)$ . The eigenvector  $|\mathbf{e}\rangle$  is localized in and around the shear band. As a result, the epistasis matrix exhibits a ‘checkered’ pattern visible in Fig. 24D. The rank-one nature of the  $e_{i,j}$  is verified numerically by spectral decomposition of the epistasis matrix obtained from the simulation. Interestingly, the genetic correlation matrix (Eq. (16)) is also approximately a rank-one tensor,  $Q_{ij} \sim |\mathbf{q}\rangle \langle \mathbf{q}|$ , with a dominant eigenvector  $|\mathbf{q}\rangle$  localized in the channel. This explains the striking similarity of the genetic correlation  $Q_{ij}$  and the epistasis  $e_{i,j}$  in Fig. 24D.



**FIG. 24 Mechanical Epistasis.** The epistasis of Eq. (17), averaged over  $10^6$  solutions  $E_{ij} = \langle e_{i,j} \rangle$ , between a fixed AA at position  $i$  (black arrow) and all other positions  $j$ . Here,  $i$  is located at (A) the binding site, (B) the center of the channel, and (C) slightly off the channel. Underneath, the average AA configuration of the protein is drawn in shades of red (P) and blue (H). Significant epistasis mostly occurs along the P-rich channel, where mechanical interactions are long ranged. Though epistasis is predominantly positive, negative values also occur, mostly at the boundary of the channel (C). (D) The two-codon correlation function  $Q_{ij}$  of Eq. (16) measures the coupling between mutations at positions  $i$  and  $j$ . The epistasis  $E_{ij}$  and the gene correlation  $Q_{ij}$  show similar patterns. Axes are the positions of  $i$  and  $j$  loci. Significant correlations and epistasis occur mostly in and around the channel region (positions 70–130, rows 7–13).

Again, comparing to the real protein glucokinase, the rightmost panel of Fig. 5 shows that the correlation of mutations is concentrated in the mechanically critical regions of the protein (left panel). Mutations away from these spots seem more independent and need not be corrected for other mutations. We conclude: **mutations correlate near mechanically critical positions.**

### M. Multilocus epistasis\*

So far, we examined the interaction between two mutations in terms of the non-linearity of the double-mutation fitness function  $e_{i,j}$  Eq. (17). This two-body interaction can be seen as the change in the effect of mutation  $j$  in the presence of another mutation  $i$ . As an isolated mutation,  $j$  has a fitness effect,  $\delta F_j$ , whereas in the presence of  $i$  the effect of  $j$  is  $\delta F_{i,j} - \delta F_i$ , and the difference defines  $e_{i,j} \equiv (\delta F_{i,j} - \delta F_i) - \delta F_j$ .

Higher-order epistasis, involving more than two mutations, has a significant role in shaping the fitness landscape [Poelwijk et al. 2017; Weinreich et al. 2013]. This motivates us to generalize the methodology of Sec. IX.L to many-body interactions. For example, the three-loci epistasis,  $e_{i,j,k}$ , measures the change in the two-loci epistasis  $e_{i,j}$  of the double  $i, j$  mutation, induced by the presence of a third mutation,  $k$  [Horovitz and Fersht 1990]:

$$e_{i,j,k} \equiv \delta F_{i,j,k} - (\delta F_{i,j} + \delta F_{j,k} + \delta F_{i,k}) + \delta F_i + \delta F_j + \delta F_k , \quad (22)$$

where  $\delta F_{i,j,k}$  is the phenotypic effect of a triple  $i, j, k$  mutation.

In a similar fashion, one derives the general  $N^{\text{th}}$ -order epistasis, among mutations at positions  $i_1, \dots, i_N$ ,

$$e_{i_1, i_2, \dots, i_N} \equiv \sum_{q=1}^N (-1)^{N-q} \sum_{i_1 < \dots < i_q} \delta F_{i_1, \dots, i_q} , \quad (23)$$

where  $\delta\mathbf{F}_{i_1, \dots, i_q}$  is the fitness effect of the  $q$ -site mutation at positions  $i_1, \dots, i_q$ . Eq. (17) and Eq. (22) are the second- and third-order epistasis terms ( $N = 2, 3$ ), while the first-order epistasis ( $N = 1$ ) is the mutation effect itself,  $e_i \equiv \delta F_i$ . Summing over all orders of epistasis interactions (Eq. (23)) up to order  $N$ , one obtains the  $N$ -site mutation effect

$$\delta\mathbf{F}_{i_1, i_2, \dots, i_N} = \sum_{q=1}^N \sum_{i_1 < \dots < i_q} e_{i_1, \dots, i_q} .$$

To link the multi-locus epistasis to protein mechanics and deformation, we follow the derivation of Eq. (20). Near the transition at which the function emerges, we use the Dyson series Eq. (6), and the resulting  $N$ -site mechanical epistasis is:

$$e_{i_1, \dots, i_N} = -F \cdot \sum_{q=1}^N (-1)^{N-q} \sum_{i_1 < \dots < i_q} \frac{\sum_{p=1}^q h_{i_p}}{1 + \sum_{p=1}^q h_{i_p}} , \quad (24)$$

where elastic factor  $h_{i_p} \equiv \mathbf{u}_*^\top \delta\mathbf{H}_{i_p} \mathbf{u}_*/\lambda_*$  is the ratio of the change in the shear energy due to mutation at  $i_p$  and the energy  $\lambda_*$  of the soft mode.

One concludes from Eq. (24) that the  $N$ -order epistasis is significant only within and around the shear band, where the bonds are stretched and compressed by the soft mode. In this region, where all the elastic factors are large. *i.e.*,  $h_{i_p} \gg 1$ , all orders of epistasis are relevant and are of the same magnitude,

$$e_{i_1, i_2, \dots, i_N} \simeq F \cdot (-1)^N .$$

We conclude: **the mechanically critical regions are strongly coupled with many-body epistatic interactions among the mutations.**

## X. OUTLOOK

This colloquium has described a method which relates biological questions and concepts regarding protein evolution to the techniques of theoretical physics. Our purpose is to make this approach accessible to a wide community. While we made an effort to cite some of the current literature, there are certainly works which we have incompletely cited. We hope that this colloquium will encourage others to build bridges between other biological questions and the long tradition of physics and mathematics.

## ACKNOWLEDGMENTS

We thank Albert Libchaber for inspiring discussions and his essential participation in our work on protein. We thank Sandipan Dutta who participated in the work on Green's functions. We thank Stanislas Leibler, Michael R. Mitchell, Elisha Moses, Giovanni Zocchi, and Olivier Rivoire for helpful discussions and encouragement. We are grateful to Karsten Kruse, Alberto Mor-pourgo, Pierre Collet, and the referees, for constructive comments on the manuscript. JPE was supported by an ERC advanced grant 'Bridges', and TT by the Institute for Basic Science IBS-R020 and the Simons Center for Systems Biology of the Institute for Advanced Study, Princeton.

## REFERENCES

- Abrikosov, A., Gorkov, L., and Dzyaloshinski, I. (1963). *Methods of Quantum Field Theory in Statistical Physics*. Parentice Hall, Englewood Cliffs, New Jersey, USA.
- Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–1615.
- Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–294.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science, New York, 4 edition.
- Alexander, S. (1998). Amorphous solids: Their structure, lattice dynamics and elasticity. *Phys. Rep.*, 296(2-4):65–236.
- Alexander, S. and Orbach, R. (1982). Density of states on fractals : “fractons”. *Journal de Physique Lettres*, 43(17):625–631.
- Alexander, S., Laermans, C., Orbach, R., and Rosenberg, H. M. (1983). Fraction interpretation of vibrational properties of cross-linked polymers, glasses, and irradiated quartz. *Phys. Rev. B*, 28(8):4615–4619.
- Arora, K. and Brooks, C. L. (2007). Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, 104(47):18496–18501.

- Baardink, G., Souslov, A., Paulose, J., and Vitelli, V. (2018). Localizing softness and stress along loops in 3d topological metamaterials. *Proc Natl Acad Sci USA*, 115(3):489.
- Bahar, I. (2010). On the functional significance of soft modes predicted by coarse-grained models for membrane proteins. *J. Gen. Physiol.*, 135(6):563–573.
- Bahar, I., Chennubhotla, C., and Tobi, D. (2007). Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Current Opinion in Structural Biology*, 17(6):633–640.
- Bahar, I., Lezon, T. R., Yang, L.-W., and Eyal, E. (2010). Global dynamics of proteins: Bridging between structure and function. *Annu. Rev. Biophys.*, 39(1):23–42.
- Banchoff, T. F. (1965). Tightly embedded 2-dimensional polyhedral manifolds. *Amer. J. Math.*, 87:462–472.
- Barrera, N. P. and Robinson, C. V. (2011). Advances in the mass spectrometry of membrane proteins: from individual proteins to intact complexes. *Annual review of biochemistry*, 80:247–271.
- Biggs, N. (1993). *Algebraic graph theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, second edition.
- Boehr, D. D., McElheny, D., Dyson, H. J., and Wright, P. E. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313(5793):1638–1642.
- Born, M. and Huang, K. (1954). *Dynamical theory of crystal lattices*. The International series of monographs on physics. Clarendon Press, Oxford,.
- Breen, M. S., Kemeny, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538.
- Bussemaker, H., Thirumalai, D., and Bhattacharjee, J. (1997). Thermodynamic stability of folded proteins against mutations. *Physical Review Letters*, 79(18):3530–3533.
- Bustamante, C., Chemla, Y. R., Forde, N. R., and Izhaky, D. (2004). Mechanical processes in biochemistry. *Annu. Rev. Biochem.*, 73:705–748.
- Chaiken, S. (1982). A combinatorial proof of the all minors matrix tree theorem. *SIAM J. Algebraic Discrete Methods*, 3(3):319–329.
- Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., and Weierstall, U. (2011). Femtosecond x-ray protein nanocrystallography. *Nature*, 470(7332):73.
- Chennubhotla, C., Rader, A. J., Yang, L. W., and Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies. *Phys. Biol.*, 2(4):S173–S180.
- Chung, F. R. K. (1997). *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC.
- Chung, F. R. K. and Sternberg, S. (1992). Laplacian and vibrational-spectra for homogeneous graphs. *Journal of Graph Theory*, 16(6):605–627.
- Clark, A. G. and Wang, L. (1997). Epistasis in measured genotypes: Drosophila p-element insertions. *Genetics*, 147(1):157–163.
- Cohen, S. L. and Chait, B. T. (2001). Mass spectrometry as a tool for protein crystallography. *Annual review of biophysics and biomolecular structure*, 30(1):67–85.
- Colin de Verdière, Y. (1993). Multiplicités des valeurs propres. Laplaciens discrets et laplaciens continus. *Rend. Mat. Appl.* (7), 13(3):433–460.
- Collins, M. D., Kim, C. U., and Gruner, S. M. (2011). High-pressure protein crystallography and nmr to explore protein conformations. *Annual review of biophysics*, 40:81–98.
- Condon, A., Kirchner, H., Larivire, D., Marshall, W., Noireaux, V., Tlusty, T., and Fourmentin, E. (2018). Will biologists become computer scientists? *EMBO Rep.*, 19(9):e46628.
- Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, 11(20):2463–2468.
- Cui, Q. and Karplus, M. (2008). Allostery and cooperativity revisited. *Protein Sci.*, 17(8):1295–1307.
- Daily, M. D., Upadhyaya, T. J., and Gray, J. J. (2008). Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins-Structure Function and Bioinformatics*, 71(1):455–466.
- Daniel, R. M., Dunn, R. V., Finney, J. L., and Smith, J. C. (2003). The role of dynamics in enzyme activity. *Annu. Rev. Biophys. Biomol. Struct.*, 32:69–92.
- de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nat. Rev. Genet.*, 14:249.
- Deng, C. Y. (2011). A generalization of the sherman–morrison–woodbury formula. *Appl. Math. Lett.*, 24(9):1561–1564.
- Desai, M. M., Weissman, D., and Feldman, M. W. (2007). Evolution can favor antagonistic epistasis. *Genetics*, 177(2):1001.
- Dill, K. A. and MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046.
- Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome research*, 13(11):2381–90.
- Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2018). Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophys.*, 41(1):429–452.
- Dutta, S., Eckmann, J.-P., Libchaber, A., and Tlusty, T. (2018). Green function of correlated genes in a minimal mechanical model of protein evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 115(20):E4559–E4568.
- Dyson, F. J. (1949a). The  $S$  matrix in quantum electrodynamics. *Phys. Rev.*, 75(11):1736–1755.
- Dyson, F. J. (1949b). The radiation theories of Tomonaga, Schwinger, and Feynman. *Phys. Rev.*, 75(3):486–502.
- Dyson, F. J. (1970). The twenty-first century. Vanuxem Lecture delivered at Princeton University, 26 February 1970, Revised and reprinted as Chapter 18, “Thought Experiments,” in Freeman J. Dyson, Disturbing the Universe, Harper and Row Publishers, New York, 1979, pp. 194–204.
- Eckmann, J.-P. (2008). Trading codes for errors. *Proc. Natl. Acad. Sci. U.S.A.*, 105(24):8165–8166.
- Eckmann, J.-P. and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617–656.
- Eckmann, J.-P. and Ruelle, D. (1992). Fundamental limitations for estimating dimensions and lyapunov exponents in dynamical systems. *Physica D*, 56(2):185–187.

- Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., and Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438:117.
- Elliott, R. J., Krumhansl, J. A., and Leath, P. L. (1974). The theory and properties of randomly disordered crystals and related physical systems. *Rev. Mod. Phys.*, 46:465–543.
- Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15.
- Fernandez-Leiro, R. and Scheres, S. H. W. (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature*, 537(7620):339.
- Ferreon, A. C. M., Ferreon, J. C., Wright, P. E., and Deniz, A. A. (2013). Modulation of allostery by protein intrinsic disorder. *Nature*, 498(7454):390–394.
- Fersht, A. (1999). *Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding*. W.H. Freeman, New York.
- Flechsig, H. (2017). Design of elastic networks with evolutionary optimized long-range communication as mechanical models of allosteric proteins. *Biophys. J.*, 113(3):558–571.
- Flechsig, H. and Togashi, Y. (2018). Designed elastic networks: Models of complex protein machinery. *International Journal of Molecular Sciences*, 19(10).
- Freitag, R. A. and Merkle, R. C. (2004). *Kinematic Self-Replicating Machines*. CRC Press. <http://www.MolecularAssembler.com/KSRM.htm>.
- Friedlander, T., Mayo, A. E., Tlusty, T., and Alon, U. (2015). Evolution of bow-tie architectures in biology. *PLoS Comput. Biol.*, 11(3):e1004055.
- Gerstein, M., Lesk, A. M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry (Mosc.)*, 33(22):6739–6749.
- Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct., Funct., Bioinf.*, 18(4):309–317.
- Goodey, N. M. and Benkovic, S. J. (2008). Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.*, 4(8):474–482.
- Goodsell, D. S. (2009). *The machinery of life*. Springer Science & Business Media.
- Grant, B. J., Gorfe, A. A., and McCammon, J. A. (2010). Large conformational changes in proteins: signaling and other functions. *Curr. Opin. Struct. Biol.*, 20(2):142–147.
- Grassberger, P. and Procaccia, I. (1983). Characterization of strange attractors. *Phys. Rev. Lett.*, 50(5):346–349.
- Greener, J. G. and Sternberg, M. J. E. (2015). Allopred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC Bioinformatics*, 16(1):335.
- Gromov, M. (2013). In a search for a structure, Part 1: On entropy. In: *European Congress of Mathematics*, pages 51–78. Eur. Math. Soc., Zürich.
- Guhr, T., Müller-Groeling, A., and Weidenmüller, H. A. (1998). Random-matrix theories in quantum physics: common concepts. *Phys. Rep.*, 299(4–6):189–425.
- Gullett, P. M., Horstemeyer, M. F., Baskes, M. I., and Fang, H. (2008). A deformation gradient tensor and strain tensors for atomistic simulations. *Modell. Simul. Mater. Sci. Eng.*, 16(1).
- Ha, T. and Tinnefeld, P. (2012). Photophysics of fluorescent probes for single-molecule biophysics and super-resolution imaging. *Annual review of physical chemistry*, 63.
- Haig, D. and Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. *Journal of Molecular Evolution*, 33(5):412–417.
- Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–86.
- Haliloglu, T. and Bahar, I. (2015). Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr. Opin. Struct. Biol.*, 35:17–23.
- Hammes-Schiffer, S. and Benkovic, S. J. (2006). Relating protein motion to catalysis. *Annu. Rev. Biochem.*, 75:519–541.
- Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A., and Chan, M. K. (2002). A new uag-encoded residue in the structure of a methanogen methyltransferase. *Science*, 296(5572):1462–1466.
- Harms, M. J. and Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, 14:559.
- Hawkins, R. J. and McLeish, T. C. B. (2006). Coupling of global and local vibrational modes in dynamic allostery of proteins. *Biophys. J.*, 91(6):2055–2062.
- Henmyer, M. and Rivoire, O. (2015). Evolution of sparsity and modularity in a model of protein allostery. *Physical Review E*, 91(4).
- Henderson, H. and Searle, S. (1981). On deriving the inverse of a sum of matrices. *SIAM Rev.*, 23(1):53–60.
- Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. A., Petsko, G. A., Karplus, M., Hubner, C. G., and Kern, D. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–U13.
- Herken, R. (1992). *The Universal Turing Machine: A Half-Century Survey*. Oxford University Press, Inc., New York, NY, USA.
- Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schrefe, C. P. I., Springer, M., Sander, C., and Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35:128.
- Horovitz, A. and Fersht, A. R. (1990). Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *Journal of Molecular Biology*, 214(3):613–617.
- Howard, J. (2001). *Mechanics of motor proteins and the cytoskeleton*. Sinauer associates Sunderland, MA.
- Huse, M. and Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell*, 109(3):275–282.
- Isralewitz, B., Gao, M., and Schulter, K. (2001). Steered molecular dynamics and mechanical functions of proteins. *Current opinion in structural biology*, 11(2):224–230.
- Jona-Lasinio, G. (2012). Modelli e linguaggi matematici nello studio dei problemi biologici. *Lettera Matematica Pristem*, 83(1):14–20.
- Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190.

- Joseph, C., Tseng, C. Y., Zocchi, G., and Tlusty, T. (2014). Asymmetric effect of mechanical stress on the forward and reverse reaction catalyzed by an enzyme. *PLoS One*, 9(7).
- Kamata, K., Mitsuya, M., Nishimura, T., Eiki, J.-I., and Nagata, Y. (2004). Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase. *Structure (London, England : 1993)*, 12(3):429–38.
- Kaneko, K., Furusawa, C., and Yomo, T. (2015). Universal relationship in gene-expression changes for cells in steady-growth state. *Phys. Rev. X*, 5(1):011014.
- Karlin, S. and Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 90(12):5873–5877.
- Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.*, 102(19):6679–6685.
- Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, 9:646.
- Keefe, A. D. and Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, 410(6829):715–718.
- Kim, J. Z., Lu, Z., Strogatz, S. H., and Bassett, D. S. (2018). Conformational control of mechanical networks. *arXiv preprint arXiv:1804.00173*.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.
- Koehl, P. and Levitt, M. (2002). Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 99(3):1280–1285.
- Kondrashov, A. S. (2017). *Crumbling Genome: The Impact of Deleterious Mutations on Humans*. Wiley-Blackwell.
- Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223.
- Koshland, D. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, 44(2):98–104.
- Koshland, D. E., Nmethyl, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits\*. *Biochemistry*, 5(1):365–385.
- Kustanovich, T., Rabin, Y., and Olami, Z. (2003). Organization of atomic bond tensions in model glasses. *Phys. Rev. B*, 67:104206.
- Lässig, M. (2007). From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC bioinformatics*, 8 Suppl 6:S7.
- Lau, K. F. and Dill, K. A. (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997.
- Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804.
- Levitt, M., Sander, C., and Stern, P. S. (1985). Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, 181(3):423–447.
- Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., de Koning, A. P. J., Dokholyan, N. V., Echave, J., Elofsson, A., Gerloff, D. L., Goldstein, R. A., Grahn, J. A., Holder, M. T., Lakner, C., Lartillot, N., Lovell, S. C., Naylor, G., Perica, T., Pollock, D. D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovich, E., Sjölander, K., Sunyaev, S., Teufel, A. I., Thorne, J. L., Thornton, J. W., Weinreich, D. M., and Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci.*, 21(6):769–785.
- Lockless, S. W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299.
- López-Blanco, J. R. and Chacón, P. (2016). New generation of elastic network models. *Curr. Opin. Struct. Biol.*, 37:46–53.
- Lubensky, T. C., Kane, C. L., Mao, X. M., Souslov, A., and Sun, K. (2015). Phonons and elasticity in critically coordinated lattices. *Rep. Prog. Phys.*, 78(7).
- Mackay, T. F. C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, 15(1):22–33.
- Mandala, V. S., Williams, J. K., and Hong, M. (2018). Structure and dynamics of membrane proteins from solid-state nmr. *Annual review of biophysics*, 47:201–222.
- Marčenko, V. A. and Pastur, L. A. (1967). The spectrum of random matrices. *Teor. Funkcií Funkcional. Anal. i Priložen. Vyp.*, 4:122–145.
- Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303.
- Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, 6(12):e28766.
- Maxwell, J. C. (1864). L. on the calculation of the equilibrium and stiffness of frames. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 27(182):294–299.
- McGinty, R. (2012–). Continuum Mechanics. <http://www.continuummechanics.org/greenstrain.html>.
- Mehmood, S., Allison, T. M., and Robinson, C. V. (2015). Mass spectrometry of protein complexes: from origins to applications. *Annual review of physical chemistry*, 66:453–474.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092.
- Ming, D. and Wall, M. E. (2005). Allostery in a coarse-grained model of protein dynamics. *Phys. Rev. Lett.*, 95(19):198103.
- Mitchell, M. R. and Leibler, S. (2017). Elastic strain and twist analysis of protein structural data and allostery of the transmembrane channel kcsa. *Phys. Biol.*
- Mitchell, M. R., Tlusty, T., and Leibler, S. (2016). Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings. *Proc. Natl. Acad. Sci. U.S.A.*, 113(40):E5847–E5855.
- Miyashita, O., Onuchic, J. N., and Wolynes, P. G. (2003). Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 100(22):12570–12575.
- Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.*, 12(1):88–118.
- Motlagh, H. N., Wrabl, J. O., Li, J., and Hilser, V. J. (2014). The ensemble nature of allostery. *Nature*, 508(7496):331–339.
- Neher, R. A. and Shraiman, B. I. (2011). Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics*, 83(4).

- Ortlund, E. A., Bridgman, J. T., Redinbo, M. R., and Thornton, J. W. (2007). Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science*, 317(5844):1544–1548.
- Penrose, R. (1955). A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.*, 51(3):406–413.
- Perutz, M. F. (1970). Stereochemistry of cooperative effects in haemoglobin: Haem-haem interaction and the problem of allostery. *Nature*, 228(5273):726–734.
- Petsko, G. A. and Ringe, D. (2004). *Protein structure and function*. New Science Press.
- Phillips, J. C. and Thorpe, M. F. (1985). Constraint theory, vector percolation and glass-formation. *Solid State Commun.*, 53(8):699–702.
- Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, 9(11):855–867.
- Poelwijk, F. J., Socolich, M., and Ranganathan, R. (2017). Learning the pattern of epistasis linking genotype and phenotype in a protein. *bioRxiv*.
- Povolotskaya, I. S. and Kondrashov, F. A. (2010). Sequence space and the ongoing expansion of the protein universe. *Nature*, 465(7300):922–926.
- Procaccia, I. (1988). Complex or just complicated? *Nature*, 333(6173):498–499.
- Qu, H. and Zocchi, G. (2013). How enzymes work: A look through the perspective of molecular viscoelastic properties. *Phys. Rev. X*, 3(1).
- Rambo, R. P. and Tainer, J. A. (2013). Super-resolution in solution x-ray scattering and its applications to structural systems biology. *Annual review of biophysics*, 42:415–441.
- Rocklin, D. Z. (2017). Directional mechanical response in the bulk of topological metamaterials. *New Journal of Physics*, 19(6):065004.
- Rocks, J. W., Pashine, N., Bischofberger, I., Goodrich, C. P., Liu, A. J., and Nagel, S. R. (2017). Designing allostery-inspired response in mechanical networks. *Proc. Natl. Acad. Sci. U.S.A.*, 114(10):2520–2525.
- Rocks, J. W., Ronellenfitsch, H., Liu, A. J., Nagel, S. R., and Katifori, E. (2018). The limits of multifunctionality in tunable networks. *arXiv preprint arXiv:1805.00504*.
- Rocks, J. W., Liu, A. J., and Katifori, E. (2019). The topological basis of function in flow networks. *ArXiv*, page 1901.00822.
- Ronellenfitsch, H., Stoop, N., Forrow, A., and Dunkel, J. (2018). Designing spectral bandgaps in phononic networks. *arXiv preprint arXiv:1802.07214*.
- Rougemont, J., Eckmann, J.-P., and Tlusty, T. (in prep.). To be published.
- Savir, Y. and Tlusty, T. (2007). Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition. *PLoS One*, 2:e468.
- Savir, Y. and Tlusty, T. (2010). Reca-mediated homology search as a nearly optimal signal detection system. *Mol. Cell*, 40(3):388–396.
- Savir, Y. and Tlusty, T. (2013). The ribosome as an optimal decoder: A lesson in molecular recognition. *Cell*, 153(2):471–479.
- Savir, Y., Noor, E., Milo, R., and Tlusty, T. (2010). Cross-species analysis traces adaptation of rubisco toward optimality in a low-dimensional landscape. *Proc. Natl. Acad. Sci. U.S.A.*, 107(8):3475–3480.
- Scheraga, H. A., Khalili, M., and Liwo, A. (2018). Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.*, 58(1):57–83.
- Schmeing, T. M., Voorhees, R. M., Kelley, A. C., Gao, Y. G., Murphy, F. V. t., Weir, J. R., and Ramakrishnan, V. (2009). The crystal structure of the ribosome bound to ef-tu and aminoacyl-trna. *Science*, 326(5953):688–94.
- Schrödinger, E. (1944). *What is life?* Cambridge University Press.
- Shakhnovich, E., Farztdinov, G., Gutin, A. M., and Karplus, M. (1991). Protein folding bottlenecks: A lattice monte carlo simulation. *Phys. Rev. Lett.*, 67(12):1665–1668.
- Srinivasan, G., James, C. M., and Krzycki, J. A. (2002). Pyrrolysine encoded by uag in archaea: Charging of a uag-decoding specialized trna. *Science*, 296(5572):1459–1462.
- Suel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10(1):59–69.
- Tama, F. and Sanejouand, Y.-H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng. Des. Sel.*, 14(1):1–6.
- Tehver, R., Chen, J., and Thirumalai, D. (2009). Allostery wiring diagrams in the transitions that drive the groel reaction cycle. *Journal of Molecular Biology*, 387(2):390–406.
- Teşileanu, T., Colwell, L. J., and Leibler, S. (2015). Protein sectors: Statistical coupling analysis versus conservation. *PLoS Comput. Biol.*, 11(2):e1004091.
- Tewary, V. K. (1973). Green-function method for lattice statics. *Adv. Phys.*, 22(6):757–810.
- Thirumalai, D., Hyeon, C., Zhuravlev, P. I., and Lorimer, G. H. (2018). Symmetry, rigidity, and allosteric signaling: From monomeric proteins to molecular machines. *arXiv preprint arXiv:1812.04969*.
- Thom, R. (2018). *Structural stability and morphogenesis*. CRC Press.
- Thompson, D. W. (1942). On growth and form. *On growth and form*.
- Thorpe, M. F. (1985). Rigidity percolation in glassy structures. *J. Non-Cryst. Solids*, 76(1):109–116.
- Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77(9):1905–1908.
- Tlusty, T. (2007a). A model for the emergence of the genetic code as a transition in a noisy information channel. *J. Theor. Biol.*, 249(2):331–342.
- Tlusty, T. (2007b). A relation between the multiplicity of the second eigenvalue of a graph laplacian, courant's nodal line theorem and the substantial dimension of tight polyhedral surfaces. *Electronic Journal of Linear Algebra*, 16:315–324.
- Tlusty, T. (2008a). A simple model for the evolution of molecular codes driven by the interplay of accuracy, diversity and cost. *Physical Biology*, 5(1):016001.
- Tlusty, T. (2008b). Rate-distortion scenario for the emergence and evolution of noisy molecular codes. *PRL*, 100(4):048101.
- Tlusty, T. (2008c). Casting polymer nets to optimize noisy molecular codes. *Proc Natl Acad Sci USA*, 105(24):8238.
- Tlusty, T. (2010). A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. *Phys. Life Rev.*, 7(3):362–376.

- Tlusty, T. (2016). Self-referring DNA and protein: a remark on physical and geometrical aspects. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2063):20150070.
- Tlusty, T., Libchaber, A., and Eckmann, J.-P. (2017). Physical model of the genotype-to-phenotype map of proteins. *Phys. Rev. X*, 7(2):021037.
- Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proc. London Math. Soc.* (2), 42(3):230–265.
- Tutte, W. T. (1948). The dissection of equilateral triangles into equilateral triangles. *Proc. Cambridge Philos. Soc.*, 44:463–482.
- von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. University of Illinois Press, Champaign, IL, USA.
- Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23(6):700–707.
- Whitford, D. (2013). *Proteins: structure and function*. John Wiley & Sons.
- Woese, C. R. (1965). Order in the genetic code. *Proc. Natl. Acad. Sci. U.S.A.*, 54(1):71–75.
- Woodbury, M. A. (1950). *Inverting modified matrices*. Statistical Research Group, Memo. Rep. no. 42. Princeton University, Princeton, N. J.
- Wrabl, J. O., Gu, J., Liu, T., Schrank, T. P., Whitten, S. T., and Hilser, V. J. (2011). The role of protein conformational fluctuations in allostery, function, and evolution. *Biophysical Chemistry*, 159(1):129–141.
- Xia, Y. and Levitt, M. (2004). Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.*, 14(2):202–207.
- Yan, L., Ravasio, R., Brito, C., and Wyart, M. (2017). Architecture and coevolution of allosteric materials. *Proc. Natl. Acad. Sci. U.S.A.*, 114(10):2526–2531.
- Zeldovich, K. B. and Shakhnovich, E. I. (2008). Understanding protein evolution: from protein physics to darwinian selection. *Annu. Rev. Phys. Chem.*, 59:105–127.
- Zheng, W. J., Brooks, B. R., and Thirumalai, D. (2006). Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U.S.A.*, 103(20):7664–7669.