Open access • Posted Content • DOI:10.1101/069039

Physical model of the sequence-to-function map of proteins — Source link

Tsvi Tlusty, Albert Libchaber, Jean-Pierre Eckmann

Institutions: Rockefeller University, University of Geneva

Published on: 12 Aug 2016 - bioRxiv (Cold Spring Harbor Labs Journals)

Topics: Sequence space (evolution)

Physical model of the genotype-to-phenotype map of proteins

Tsvi Tlusty,[1,2,3] Albert Libchaber,[4] and Jean-Pierre Eckmann[5]

[1]*Center for Soft and Living Matter, Institute for Basic Science (IBS), Ulsan 44919, Korea*
[2]*Department of Physics, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea*
[3]*Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA*
[4]*The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA*
[5]*Département de Physique Théorique and Section de Mathématiques, Université de Genève, CH-1211, Geneva 4, Switzerland*

How DNA is mapped to functional proteins is a basic question of living matter. We introduce and study a physical model of protein evolution which suggests a mechanical basis for this map. Many proteins rely on large-scale motion to function. We therefore treat protein as learning amorphous matter that evolves towards such a mechanical function: Genes are binary sequences that encode the connectivity of the amino acid network that makes a protein. The gene is evolved until the network forms a shear band across the protein, which allows for long-range, soft modes required for protein function. The evolution reduces the high-dimensional sequence space to a low-dimensional space of mechanical modes, in accord with the observed dimensional reduction between genotype and phenotype of proteins. Spectral analysis of the space of 10^6 solutions shows a strong correspondence between localization around the shear band of both mechanical modes and the sequence structure. Specifically, our model shows how mutations are correlated among amino acids whose interactions determine the functional mode.

PACS numbers: 87.14.E-, 87.15.-v, 87.10.-e

I. INTRODUCTION: PROTEINS AND THE QUESTION OF THE GENOTYPE-TO-PHENOTYPE MAP

DNA genes code for the three-dimensional configurations of amino acids that make 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 gene [1–5]. Furthermore, evolution has to select the tiny fraction of functional sequences in an enormous, high-dimensional space [6–8], which implies that protein is a 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 as to how a functional protein originates from a linear DNA sequence is still open, in particular, how the functionality constrains the accessible DNA sequences.

To examine the geometry of the sequence-to-function map, we devise a mechanical model of proteins as amorphous learning matter. Rather than simulating concrete proteins, we construct a model which captures the hallmarks of the genotype-to-phenotype map. The model is simple enough to be efficiently simulated to gain statistics and insight into the geometry of the map. We base our model on the growing evidence that large-scale conformational changes – where big chunks of the protein move with respect to each other – are central to function [9–15]. In particular, allosteric proteins can be viewed as 'mechanical transducers' that transmit regulatory signals between distant sites [16–19].

Dynamics is essential to protein function, but it is hard to measure and simulate due to the challenging spatial and temporal scales. Nevertheless, recent studies suggest a physical picture of the functionally-relevant conformational changes within the protein: Nanorheological measurements showed low-frequency viscoelastic flow within enzymes [20], with mechanical stress affecting catalysis [21]. Computation of amino acid displacement, by analysis of structural data, demonstrated that the strain is localized in 2D bands across allosteric enzymes [22]. We therefore take as a target function to be evolved in our protein such a large-scale dynamical mode. Other important functional constraints, such as specific chemical interactions at binding sites, are disregarded here because they are confined to a small fraction of the protein. We focus on this mechanical function whose large scale, collective nature leads to long-range correlation patterns in the gene.

Our model includes essential elements of the genotype-to-phenotype map: the target mechanical mode is evolved by mutating the 'gene' that determines the connectivity in the amino acid network. During the simulated 'evolution', mutations eventually divide the protein into rigid and 'floppy' domains, and this division enables large-scale motion in the protein [23]. This provides a concrete map between sequence, configuration, and function of the protein. The computational simplicity allows for a massive survey of the sequence universe, which reveals a strong signature of the protein's structure and function within correlation 'ripples' that appear in the space of DNA sequences.

II. MODEL AND RESULTS

We give here a summary and interpretation of our results, The appendix contains further details and explains choices we made in designing the model as close as possible to real proteins.

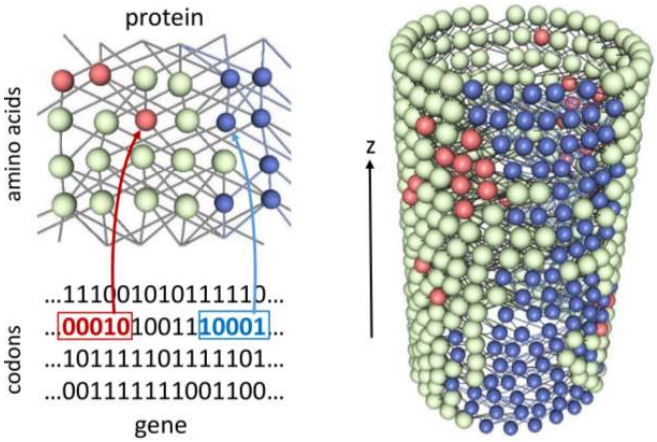

FIG. 1. **The main features of the physical model**:
(left) The mapping from the binary gene to the connectivity of the amino acid (AA) network that makes a functional protein. AAs are beads and links are bonds. The color of the AAs represents their rigidity state as determined by the connectivity according to the algorithm of Sect. 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 motion along the fluid channel.

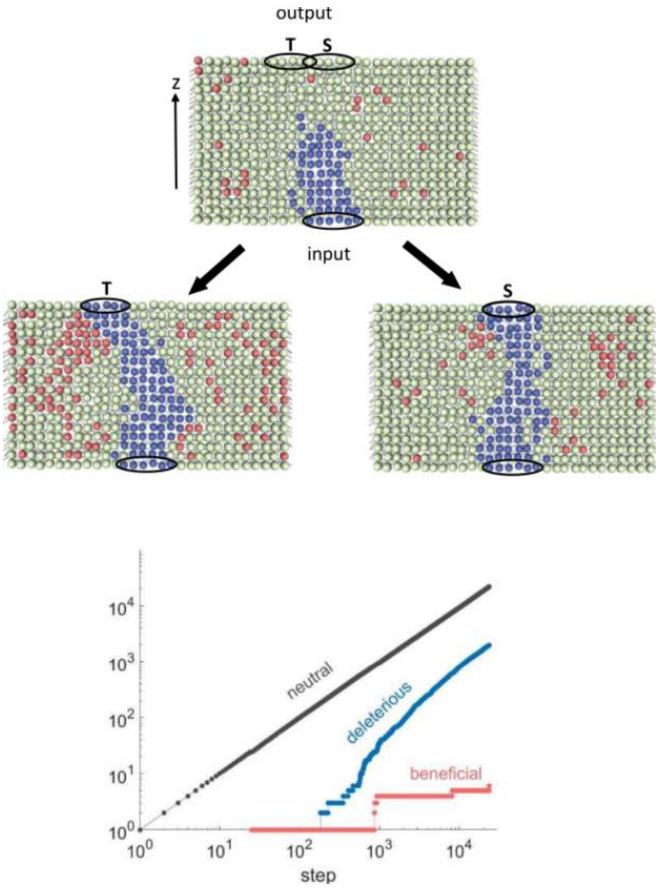

FIG. 2. **Evolution of mechanical function:**
(top) An initial configuration with a given input (black ellipse at bottom) and a random sequence is required to evolve into a straight fluid channel (S) or a tilted one (T).
(bottom) Following the success of evolution. In each generation, a randomly drawn bit (a letter in the 5-bit codon) is flipped, and this 'point mutation' is changing one bond (similar to point mutations that change one base in a codon). A typical run is a sequence of mostly neutral steps, a fraction of deleterious ones, and rare beneficial steps. Note that the 'fitness' of the configuration is only measured at the top, not in the interior of the cylinder.

A. Mechanical model of protein evolution

Our model is based on two structures: a *gene*, and a *protein*, which are coupled by the genotype-to-phenotype map. The coarse-grained protein is an aggregate of amino acids (AAs), modeled as beads, with short-range interactions given as bonds (Fig. 1). A typical protein is made of several hundred AAs, and we take $N = 540$. We layer the AAs on a cylinder, 18 high 30 wide, similar to dimensions of globular proteins. The cylindrical configuration allows for fast calculation of the low energy modes, and thereby fast evolution of the protein. Each AA may connect to the nearest five AAs in the layer below, so that we get $2^5 = 32$ effective AA species, which are encoded as 5-letter binary *codons* [1]. These codons specify the bonds in the protein in a 2550-long *sequence* of the *gene*($5 \times 30 \times (18-1)$, because the lowest layer is connected only upwards).

To become functional, we want the protein to evolve to 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 [2]. The solution we search turns out to be a large-scale, low-energy deformation where one domain moves rigidly with respect to another in a shear or hinge motion, which is facilitated by the presence of a fluidized, 'floppy' channel separating the rigid domains [25–27].

These large-scale deformations are governed by the rigidity pattern of the configuration, which is determined by the connectivity of the AA network via a simple majority rule (Fig. 1) which we detail in Sect. A 3. The basic idea is that each AA can be either rigid or fluidized and that this rigidity state propagates upwards: Depending on the number of bonds and the state of other AAs in its immediate neighborhood, an AA will be rigidly connected, 'shearable', *i.e.*, loosely connected, or in

[1] In our model, the AA species is determined by the bonds, while in real proteins the bonds are determined by the chemical nature and position of the AA (see also Sect. II G).

[2] Note that in this simulation, we do not take as evolutionary criterion the mechanical signal itself, but require that the protein forms a fluid channel with a prescribed configuration. We show that this configuration *facilitates* the sought-after mechanical shear motion in Sections II E and B 3. (In [24] we take the mechanical modes themselves as the target function.)

a pocket of less connected AAs within a rigid neighborhood [3]. As the sequence and hence the connections mutate, the model protein adapts to the desired input-output relation specified by the extremities of the separating fluid channel (Fig. 1(right)).

The model is easy to simulate: We start from a random gene of 2550 bits, and at each time step we flip a randomly drawn bit, thus adding or deleting a bond. In a zero-temperature Metropolis fashion, we keep only mutations which do not increase the distance from the target function, *i.e.*, the number of errors between the state in the top row and the prescribed outcome. Note that, following the logics of biological evolution, the 'fitness' of the protein is only measured at its functional *surface* (*e.g.*, where a substrate binds to an enzyme) but not in its interior.

Typically, after 10^3-10^5 mutations this input-output problem is solved (Fig. 2). Although the functional sequences are extremely sparse among the 2^{2550} possible sequences, the small bias for getting closer to the target in configuration space directs the search rather quickly. Therefore, we could calculate as much as 10^6 runs of the simulation which gave 10^6 independent solutions of the evolutionary task.

B. Dimensional reduction in the phenotype-to-genotype map

Thanks to the large number of simulations, we can explore vast regions of the genetic universe. That the sampling is well-distributed can be seen from the typical inter-sequences distance, which is comparable with the universe diameter (Fig. 4). This also indicates that the dimension of the solution set is high. Indeed, the observed dimension of *sequence* space, as estimated following [28, 29], is practically infinite (~ 150) [4]. This shows that the bonds are chosen basically at random, although we only consider functional sequences.

On the other hand, very few among the 2^{540} *configurations* are solutions, owing to the physical constraints of contiguous rigid and shearable domains. As a result, when mapped to the configuration space, the solutions exhibit a dramatic reduction to a dimension of about 8-10 [30]. This reduction between 'genotype' (sequence) and 'phenotype' (configuration, function) [31, 32] is the outcome of physical constraints on the mechanical transduction problem. In the nearly random background of sequence space, these constraints are also manifested in long-range correlations among AAs on the boundary of the shearable region (Fig. 5 and Sect. B 4).

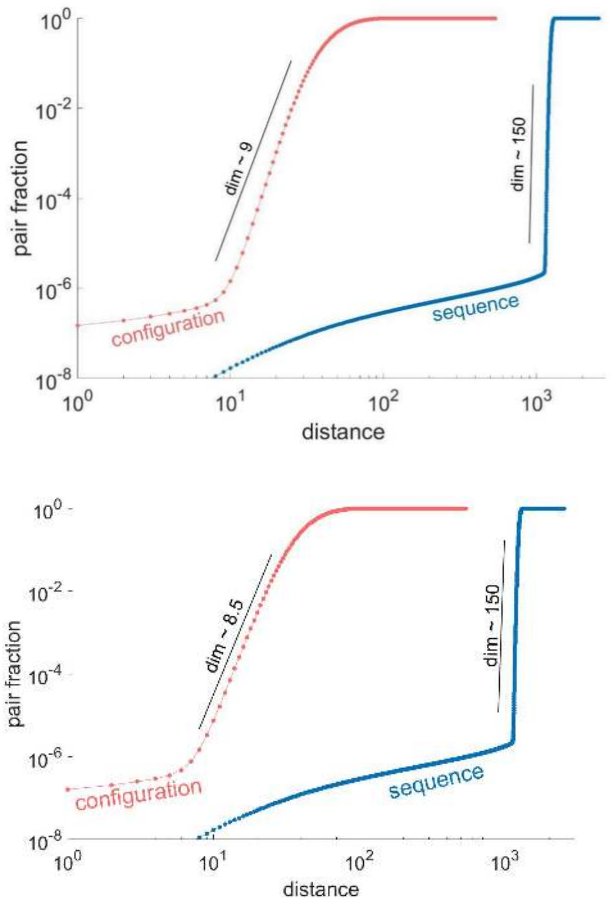

FIG. 3. **Dimensional reduction of the genotype-to-phenotype map:**
Dimension measurement for the straight (S, top) and tilted (T, bottom) cases. 10^6 independent functional configurations were found for the input-output problem. An estimate for the dimension of the solutions is the correlation length, the slope of the cumulative fraction of solution pairs as a function of distance. In configuration space (red), the distance is the number of AAs (out of 540) with a different rigidity state. The estimated dimension from $10^{12}/2$ distances is about 9 (black line) for problem S and 8.5 in problem T. The sequence space is a 2550-dimensional hypercube with 32^{510} sequences. Most distances are close to the typical distance between two random sequences (2550/2 = 1275), indicating a high-dimensional solution space. An estimate for the dimension is $\sim$ 150 (black line) for both S and T problems. The similarity of the dimensions in both cases suggests that these numbers are not specific to the problem.

C. Spectral analysis reveals correspondence of genotype and phenotype spaces

Spectral analysis of the solution set in both sequence and configuration spaces provides further information on the sequence-to-function map (Fig. 6). The sequence spectrum is obtained by singular value decomposition (SVD) of a $10^6 \times 2550$ matrix, whose rows are the binary genes of the solution set. The first few eigenvectors (EVs) with the larger eigenval-

[3] The propagation of rigidity is effectively a "double" percolation problem in which both fluid (blue) and rigid (gray) regions are continuous (see Sect. A 3).

[4] We lack sufficient data to determine such high dimensions precisely, and 150 is a lower bound.

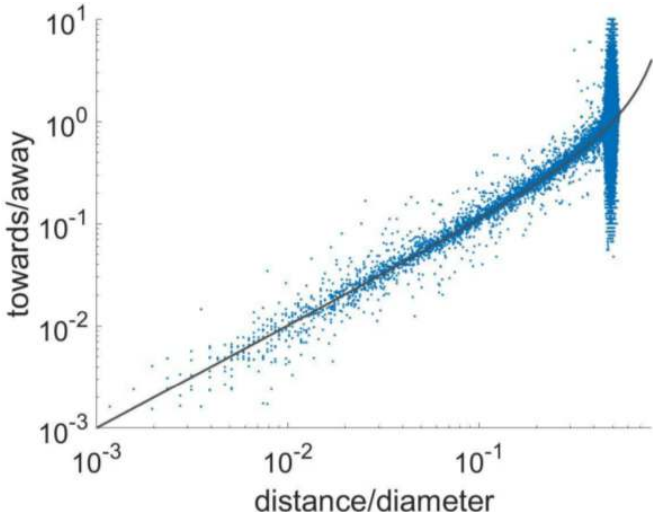

FIG. 4. **Distribution of solutions in the sequence universe:** A measure for the expansion in the functional sequence universe is the backward/forward ratio, the fraction of point mutations that make two sequences closer vs. the ones that increase the distance [6]. The Hamming distances D (normalized 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 from purely random mutations.

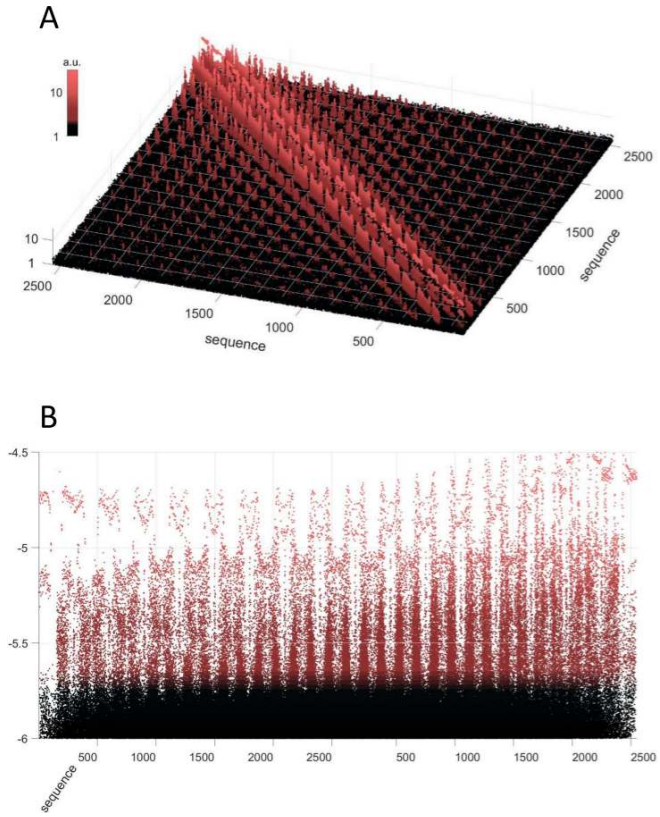

FIG. 5. **Long-range genetic correlations:** (top) The sequence correlation matrix across the 10^6 examples shows long-range correlations among the bits (codons) at the rigid/fluid boundary, and short-range correlations in the rigid domains. (bottom) A cross section perpendicular to the diagonal axis.

ues capture most of the genetic variation among the solutions, and are therefore the *collective degrees-of-freedom of protein evolution* (Fig. 6B). The 1st EV is the average sequence, and the next EVs highlight positions in the gene that tend to mutate together to create the fluid channel.

The spectrum of the configuration space is calculated in a similar fashion by the SVD of a $10^6 \times 540$ matrix, whose rows are the configurations of the solutions set (Fig. 6A). In the configuration spectrum, there are 8-10 EVs which stand out from the continuous spectrum, corresponding to the dimension 8 shown in Fig. 3. Although the dimension of the sequence space is high (~ 150), there are again only 8-9 eigenvalues outside the continuous random spectrum.

These isolated EVs distill beautifully the non-random components within the mostly-random functional sequences. The EVs of both sequence and configuration are localized around the interface between the shearable and rigid domains. The similarity in number and in spatial localization of the EVs reveals the tight correspondence between the configuration and sequence spaces.

This duality is the outcome of the sequence-to-function map defined by our simple model: The geometric constraints of forming a shearable band within a rigid shell, required for inducing long-range modes, are mirrored in long-range correlations among the codons (bits) in sequence space. The corresponding sequence EVs may be viewed as weak 'ripples' of information over a sea of random sequences, as only about 8 out of 2550 modes are non-random (0.3%). These information ripples also reflect the self-reference of proteins and DNA via the feedback loops of the cell circuitry [34].

It is instructive to note similarities and differences between the spectra. While the spectra of the configuration space and of the sequence space have a similar form — with a continuous, more or less random, part and a few isolated eigenvalues above it — the *location* of the random part is different: In the configuration case it is close to zero while in the sequence case it is concentrated at large values around 500.

The geometric interpretation is that the cloud of solution points looks like an 8-9 dimensional flat disk in the configuration case, while in the sequence space, it looks like a high-dimensional almost-spherical ellipsoid. The few directions slightly more pronounced of this ellipsoid correspond to the non-random components of the sequence. The slight eccentricity of the ellipsoid corresponds to the weak non-random signal above the random background. This also illustrates that the dimension of the sequence space is practically infinite, while in the configuration space it is comparable to the number of isolated eigenvalues.

We verified that the dimensional reduction and the spectral correspondence depend very little on the details of the models. For example, we examined a model with 16 AA species

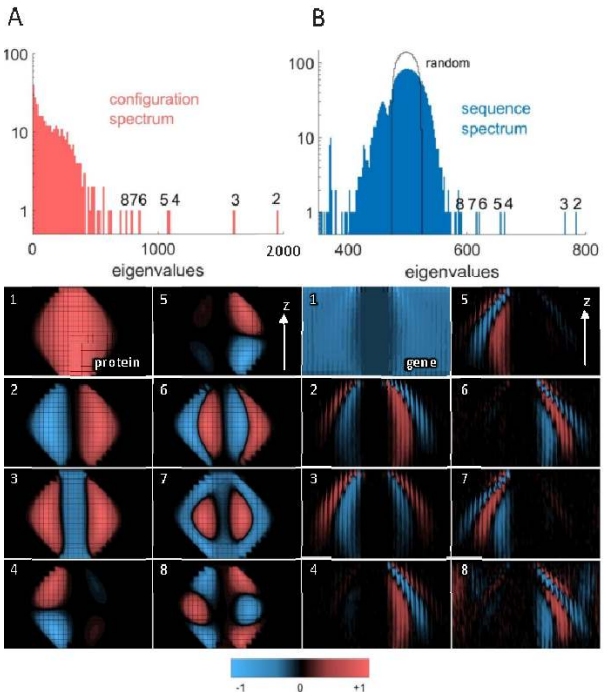

FIG. 6. **Correspondence of modes in sequence and configuration spaces:**
We produced the spectra by singular value decomposition of the 10^6 solutions of problem S. The corresponding spectra for the T case are shown in Fig. 7.
(A) Top: the spectrum in configuration space exhibits about 8-10 eigenvalues outside the continuum (large 1st eigenvalue not shown). Bottom: the corresponding eigenvectors describe the basic modes of the fluid channel, such as side-to-side shift (2nd) or expansion (3rd).
(B) Top: The spectrum of the solutions in sequence space is similar to that of random sequences (black line), except for about 8-9 high eigenvalues that are outside the continuous spectrum.
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 [33].

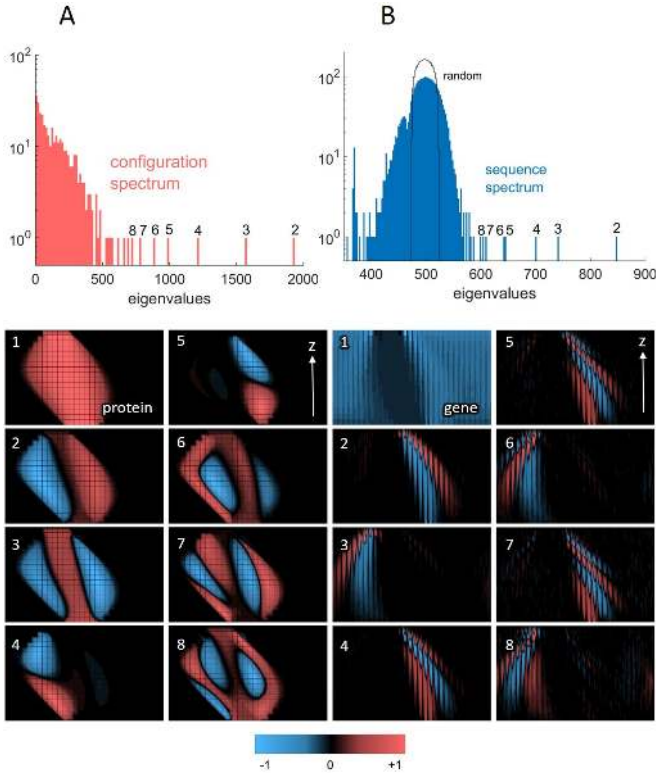

FIG. 7. **Spectra and eigenfunctions for the tilted example (T)**: Note the similarity with Fig. 6, and also how the tilt is manifested not only in the protein modes , but also in the gene modes. This demonstrates that the gene and the protein *share* common features.
(A) The configuration spectrum and eigenfunctions.
(B) The sequence spectrum and eigenfunctions.

instead of 32.[5] We found that the dimension of the phenotype space was ~ 9.1, while a lower bound on the genotype dimension was ~ 150, very similar to the dimensions of the 32 AA model (compare to Fig. 3). The spectra and the eigenmodes of both configuration and sequence spaces were also similar (not shown).

D. Stability of the mechanical phenotype under mutations

First, we determine how many mutations lead to a destruction of the solution (Fig. 8A). About 10% of all solutions are destroyed by just one random mutation. The exponentially decaying probability of surviving m mutations signals that these mutations act quite independently. Fig. 8B which shows the location of these destructive mutations around the shearable channel [6].

We have also studied the loci where two *interacting* mutations will destroy a solution (*i.e.*, none of the two is by itself destructive). In most cases, the two mutations are close to each other, acting on the same site. The channel is less vulnerable to such mutations, but the twin mutations are evenly distributed over the whole rigid network (Fig. 8C).

E. Fluid channel supports low-energy shear modes

The evolved rigidity pattern supports low-energy modes with strain localized in the floppy, fluid channel. We tested

[5] The natural genetic code with its 20 AAs is therefore an intermediate case.

[6] The natural genetic code is redundant, *i.e.* several codons encode the same AA and are therefore synonymous. Such redundancy reduces the fraction of destructive mutations, since mutations that exchange synonymous codons do not change the encoded AA and are theretofore bound to be neutral. A case of redundant code is examined in Sect. II G.

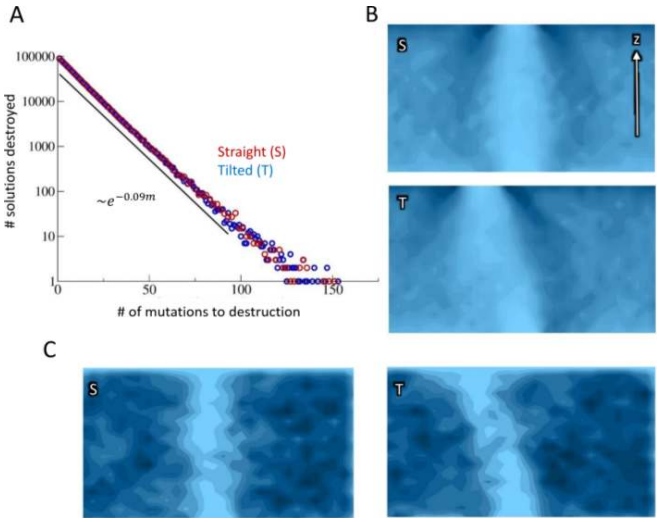

FIG. 8. **Stability of the mechanical phenotype to mutations:** Mutations at sensitive positions of the sequence move the output away from the prescribed solution.
(A) Fraction of runs (among 10^6) destroyed by the m-th mutation. A single mutation destroyed about 9% of solutions. The proportion decays exponentially like $\exp(-0.09m)$.
(B) The density map of such mutations for problems S and T (Fig. 2) shows accumulation around the fluid channel and at the top layer (dark regions).
(C) The double mutations are evenly distributed in the rigid regions.

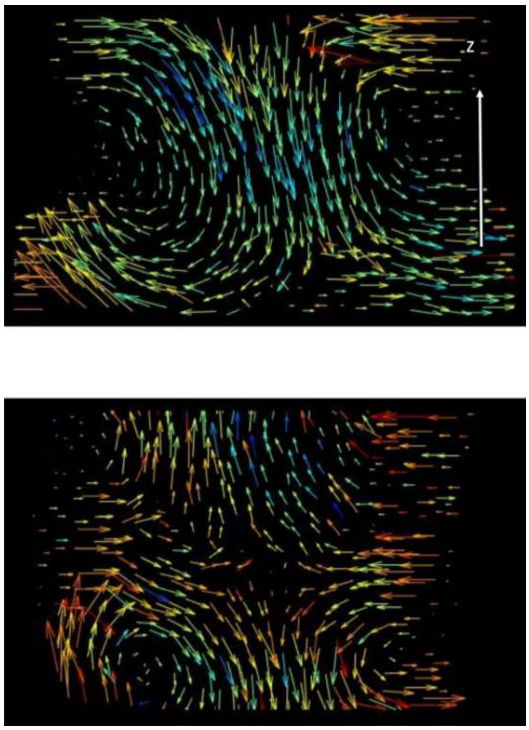

FIG. 9. **Mechanical shear modes:** Displacement and strain fields for the tilted solution T for two low eigenvalues. The vectors show the direction of the displacement and the color code denotes the strain (*i.e.*, the local change in the vector field as a function of position, maximal stress is red).

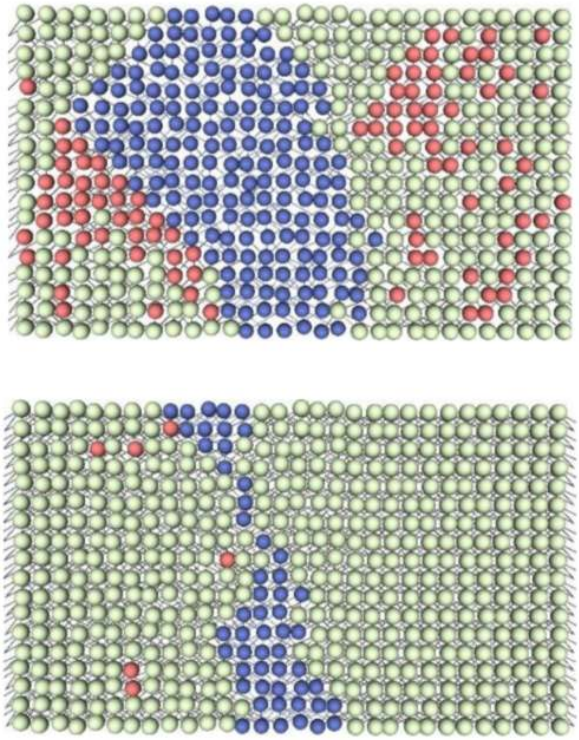

FIG. 10. **Adaptation of thermal stability:** Extreme configurations, with low (50%, left) and high (95%, right) bond density, solve problem T.

whether the evolved AA network indeed induces such modes (Fig. 9), by calculating the mechanical spectrum of a spring network in which bonds are substituted by harmonic springs. The shear motion of the network is characterized by the modes of $\mathcal{H}$, its elastic tensor. $\mathcal{H}$ is the $2N$x$2N$ curvature matrix in the harmonic expansion of the elastic energy $E \simeq \frac{1}{2}\delta\mathbf{r}^T\mathcal{H}\,\delta\mathbf{r}$, where $\delta\mathbf{r}$ is the $2N$-vector of the 2D displacements of the N AAs. $\mathcal{H}$ has the structure of the network Laplacian multiplied by the 2x2 tensors of directional derivatives (see Sect. B 3, which is derived from [35, pp. 618–9]).

We traced the mechanical spectrum of the protein during the evolution of the fluidized channel (a shear band). We found that the formation of a continuous channel of less connected amino acids indeed facilitates the emergence of low-energy modes of shear or hinge deformations (Fig. 9). The energy of such low modes nearly vanishes as the channel is close to completion. Similar deformations, where the strain is localized in a rather narrow channel, occur in real proteins, as shown in recent analysis of structural data [22].

F. Proteins can adapt simultaneously to multiple tasks

Our models were designed to trace the evolution of a mechanical function and show how it constrains the genotype-to-phenotype map, as shown above. Real proteins also evolve towards other essential functions, such as binding affinity and

biochemical catalysis at specific binding sites. Here, we examine another important molecular trait, stability.

Many studies examine the energetic stability of the protein, as measured by its overall free energy (ΔG) [4, 5, 8]. In the present model, this free energy is given by the number of bonds, which represent chemical and physical interactions among the amino acids. The higher the number of bonds the more stable and less flexible is the protein. By tuning stability, organisms adapt to their environment. Thermophiles that live in hotter places, such as hydrothermal vents, evolve more stable proteins to withstand the heat. Cryophiles that reside in colder niches have more flexible proteins [36].

We simulated the evolution of the two phenotypes, our specific dynamical mode together with an energetic state (*i.e.*, a given bond density). We find that the large solution set of the mechanical problem allows the protein to select a subset with a specific energetic state. Thus, the evolutionary dynamics could find solutions to the same mechanical function when we imposed extreme values of bond density (Fig. 10). This demonstrates the capacity of the protein to search in parallel for the solutions of several biological tasks. Evolving a specific binding site is expected to be an easier task, since such sites are confined to a small fraction of the protein.

G. Amino acid interactions

In the model described so far, the bonds were determined by the AA species alone, while in real proteins, it is the interaction between *pairs* of AA which determines the formation of bonds [7]. This raises the question as to how much our results are sensitive to the fine details of the interaction model. As we show, a more realistic interaction model does not change the main results, which demonstrates the robustness of our approach.

To model two-body AA interactions we consider a set of three AA species, which we call A_0, A_1 and A_2. Whether a bond is formed or not is determined by a symmetric binary relation $b(A_i, A_j)$, which we write as a 3×3 interaction matrix,

	A_0	A_1	A_2
A_0	1	1	1
A_1	1	1	0
A_2	1	0	0

TABLE I. The interaction $b(A_i, A_j)$ among the three AAs. The formation of a bond by the pair A_i-A_j is the denoted by a '1', while '0' denoted the absence of a bond.

This variant of the model is reminiscent of the HP model with its two species of AAs [37]. The interaction range is kept identical to that our standard model, namely an AA can form a bond the 5 nearest neighbors in the adjacent rows.

[7] At least two AAs. There may be also higher order terms of three-body interactions etc.

The 'gene' in this variant of the model is a sequence of $18 \times 30 = 540$ two-letter binary codons, g_i, each representing an AA, such that the overall length of the gene is 1080 bits. The genetic code is a map C from codons to AAs, $C : g_i \rightarrow A_i$. Since there are four codons and only three AAs, there is a 25% *redundancy* in the 'genetic code'. This is reminiscent of the (higher) redundancy of the natural genetic code in which 20 AAs are encoded by 61 codons [38–40] (out of the $4^3 = 64$ codons 3 are 'stop' codons). In our 4-codon genetic code, the redundant AA is chosen to be A_0, $C(00) = C(01) = A_0$, and the two other AA are encoded as $C(10) = A_1, C(11) = A_2$. For a given gene, the bond pattern is determined by looking at all AA pairs within the interaction range and calculating their coupling according to the interaction matrix (Table I), $b(C(g_i), C(g_j)) = b(A_i, A_j)$. Once the bond network is determined from the gene, the rigidity pattern, rigid, fluid or 'trapped', is calculated as in the standard model (Sect. II A).

In the simulations, at each step we flip one letter in a randomly selected codon. A quarter of the mutations are synonymous, since they exchange '00' and '01'. The other three quarters add or cut bonds, and we check, as before, whether the connectivity change moves the rigidity pattern closer to a pattern that allows for a low-energy floppy mode. A small number of beneficial mutations eventually resolve the mechanical transduction problem, typically after $10^3 - 10^4$ mutations.

In Fig. 11 we present some data (obtained from $4 \cdot 10^5$ solutions) to illustrate the robustness of the results relative to model changes. We find that, despite having changed the connectivity model, our main conclusions regarding the geometry of the phenotype-to-genotype map remain intact: A huge reduction from a high-dimensional genotype space ($\dim > 100$) to a low-dimensional phenotype space ($\dim \sim 10$), similar to the dimensions in Fig. 3. It is noteworthy that the configuration eigenvectors are very similar to those of simpler model (as in Fig. 6), although they are determined by very different bonding interactions. This is evident in the (non-random) bond eigenvectors which are similar in number to those of the pervious model but differ in pattern owing to the different bonding rules of Table I. The robustness of the results manifests the universality of the dimensional reduction which originates from the continuity of the mechanical transduction.

III. CONCLUSIONS

Our models of the genotype-to-phenotype map put forward a new physical picture of protein evolution. Our thesis is that rather than structure itself, it is the *dynamics* that governs protein fitness. Our method considers proteins as evolving amorphous matter with a mechanical function, a specific low-energy conformational change. The rigidity/shearability pattern of the protein, and hence its dynamical modes, are determined by the connectivity of the amino acid interaction network. The model explains how the spatially-extended modes appear as the gene mutates and changes the amino acid network. These modes are shear and hinge motions where the

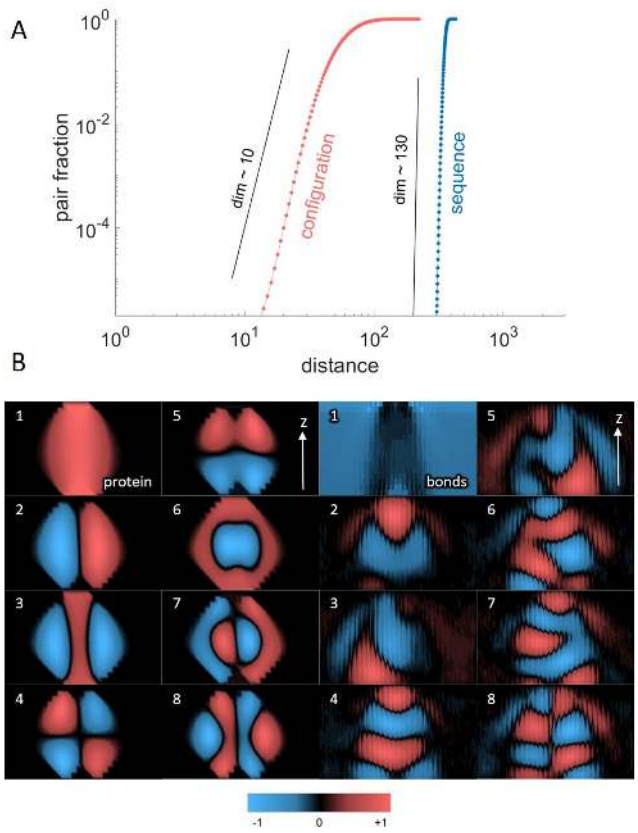

FIG. 11. (A) **The AA interaction model.** (A) Dimension of the genotype and phenotype spaces are similar to the standard model (Fig. 3). (B) Left: The first few eigenfunctions for the configuration. Right: the same for the bond patterns.

strain is localized in the shearable channel and where the surrounding domains translate or rotate as rigid bodies (Fig. 9).

A main insight from our model is that requiring the protein to have 'floppy' modes puts strong constraints on the space of mechanical phenotypes. As a consequence there is a huge dimensional reduction when mapping genotypes to phenotypes. We find that the collective mechanical interactions among the amino acids are mirrored in corresponding modes of sequence correlation in the genes. These main results do not depend on details of the model and have been reproduced in versions with (i) a different number of AA species (16 instead of 32), (ii) bonds that depend on pairwise interactions, and (iii) harmonic spring network [24]. All these suggest that the results are generic and apply to a wide range of realizations.

Our models are distilled to their simplest physical-mathematical schemes, but have concrete, experimentally testable predictions. In the functional protein, the least random, strongly correlated sites are concentrated in a rigid shell that envelops the shearable channel [22]. Our model therefore predicts that these sites are also the most vulnerable to mutations (Fig. 8B), which distort the low-frequency modes and thus hamper the biological function. These effects can be examined by combining mutation surveys, biochemical assays of the function, and physical measurements of the low-frequency spectrum, especially in allosteric proteins.

To that end, one may take an enzyme with a known shear band (via analysis similar to [22]) and mutate amino acids within and around the band. We expect the mutation of these amino acids to have a significant impact on the dynamics and biochemical function of the protein, as compared to other mutations in the rigid subdomains. By sequence alignment methods [33, 41–43], it is possible to test whether these sensitive positions in the protein exhibit strong correlations in the gene, as predicted by the model. One may also search for the dimensional reduction predicted by the model in high resolution maps of molecular fitness landscapes [44–47].

Past studies have shown that the motion of proteins [48–51] and their hydrophobicity patterns [52] may often be approximated by a few normal modes, while others have demonstrated that the variation in aligned sequences may be characterized by a few correlation modes [33, 41–43]. The present study links the genotype and phenotype spaces, and explains the dimensional reduction as the outcome of a non-linear mapping between genes and patterns of mechanical forces: We characterize the emergent functional mode to be a soft, 'floppy' mode, localized around a fluidized channel (a shear band), a region of lower connectivity which is therefore easier to deform. The contiguity of this rigidity pattern implies that it can be described by a few collective degrees of freedom, implying a vast dimensional reduction of configuration space.

The concrete genotype-to-phenotype map in our simple models demonstrates that most of the gene records random evolution, while only a small non-random fraction is constrained by the biophysical function. This drastic dimensional reduction is the origin of the flexibility and evolvability in the functional solution set.

Appendix A: The protein evolution model

1. The cylindrical amino acid network

We model the protein as an aggregate of amino acids (AAs) with short range interactions. In our coarse grained model, beads represent the AAs and bonds their interactions with neighboring AAs (Fig. 1). We consider a simplified cylindrical geometry, where the AAs are layered on the surface of a cylinder at randomized positions, to represent the non-crystalline packing of this amorphous matter. Throughout this study, we examine 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. When the cylinder is shown as a flat 2D surface (such as in Fig. 2), there are still periodic boundary conditions in the horizontal w direction. The row and column coordinates of an AA are (r, c), with r for the row $(1, \ldots, h)$ and c for the column $(1, \ldots, w)$. The cylindrical periodicity is accounted for by taking the horizontal coordinate c modulo $w = 30$, $c \to \mod_w(c-1)+1$.

Each AA in row r can connect to any of its five nearest neighbors in the next row below, $r-1$. This defines $2^5 = 32$ effective species of amino acids that differ by their 'chemistry', *i.e.*, by the pattern of their bonds. Therefore, in the

gene, each AA at (r, c) is encoded as a 5-letter binary *codon*, ℓ_{rck}, where the k-th letter denotes the existence $(= 1)$ or absence $(= 0)$ of the k-th bond. The gene is the sequence of $N_{AA} = w \cdot h = 540$ codons which represent the AAs of the protein. This means that each codon just specifies which of the 5 bonds are present or absent. Therefore, the codons are a genetic *sequence* of $2700 = w \cdot h \cdot 5$ digits 0 or 1. Each of these numbers determines whether or not a *bond* connects two positions of the grid. Since the bonds from the bottom row do not affect the configuration of the protein and the resulting dynamical modes, the relevant length of the gene is somewhat smaller, $N_S = 2550 = w \cdot (h-1) \cdot 5$.

2. Evolution searches for a mechanical function

We now *define* the target of evolution as finding a functional protein, in the following specific sense: 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 motion, which is facilitated by the presence of a fluidized, 'floppy' channel separating the rigid domains [25–27].

3. Rigidity propagation algorithm

The large-scale deformations are governed by the rigidity pattern of the configuration, which is determined by the connectivity of the AA network via a simple majority rule (Fig. 1). The details of this majority rule are as follows (Fig. 12): Each AA position will have two binary properties, which define its state:

- The *rigidity* σ: This property can be *rigid* $(\sigma = 1)$ or *fluid* $(\sigma = 0)$.
- The *shearability* s: This property can be *shearable* $(s = 1)$ or *non-shearable* $(s = 0)$. As shown below, a non-shearable AA can be either rigid or fluid within a rigid domain of the protein. Non-shearable domains tend to move as a rigid body (*i.e.*, via translation or rotation), whereas shearable regions are easy to deform.

Only 3 of the 4 possible combinations are allowed :

1. Non-shearable and solid AA (yellow): $(\sigma = 1; s = 0)$.
2. Non-shearable and fluid AA (red): $(\sigma = 0; s = 0)$.
3. Shearable and fluid AA (blue): $(\sigma = 0; s = 1)$.
4. Shearable solid is forbidden.

Given a fixed sequence, and an *input* state in the bottom row of the cylinder, $\{\sigma_{1,c}, s_{1,c}\}$ the state of the cylinder is completely determined as follows: The three states percolate through the network, from row r to row $r+1$ (see Fig. 12). This propagation is directed by the presence of bonds, with a maximum of 5 bonds ending in each AA (of rows $r = 2$ to h; the state of the first row is given as input). These bonds can be *present*(=1) or *absent*(=0). according to the codon ℓ_{rck}, $k = -2, \ldots, 2$ when they point to the AA with coordinate (r, c) coming from the AA $(r-1, c+k)$.

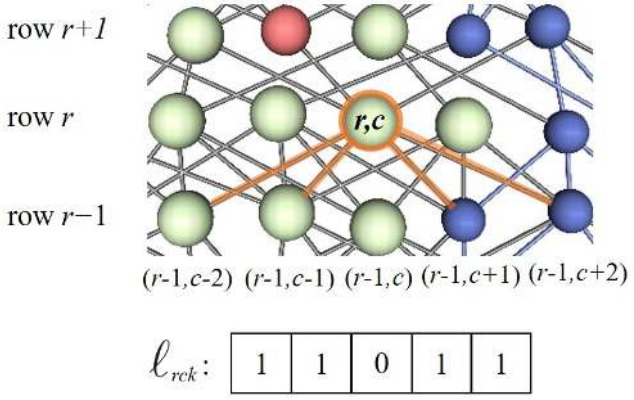

FIG. 12. Illustration of the percolation rules for shearability and fluid/solid states. Note that site (r, c) was turned solid because it is attached to 2 solid sites below it. Also note that the red site above it is fluid, because it is attached to less than 2 solid sites below it. But it is not shearable 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, c)) 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 $(r, c+2)$.

In a first sweep through the rows, we deal with the *rigidity* property σ. In row $r = 1$ each of the w AAs is in a rigidity state rigid $(\sigma = 1)$ or fluid $(\sigma = 0)$. In all other rows, $r = 2$ to h, the 5 bonds determine the value of the rigidity of (r, c) through a majority rule:

$$\sigma_{r,c} = \theta\left(\sum_{k=-2}^{2} \ell_{rck}\sigma_{r-1,c+k} - \sigma_0\right), \qquad \text{(A1)}$$

where θ 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 that are required to rigidly support 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 the choice of the bonds which are present as encoded in the gene.

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

$$s_{r,c} = (1 - \sigma_{r,c}) \cdot \theta\left(\sum_{k=-1}^{1} s_{r-1,c+k} - s_0\right), \qquad \text{(A2)}$$

where $s_0 = 1$. The first term on the lhs ensures that a solid AA can never become shearable. This completes the definition of the map from the sequence to the state.

4. Fitness and mutations

As we explained before, the aim is to find a functional protein which can transfer forces. To find such a protein, we start from a random sequences (of 2550 codons), and from an initial state (*input*) in the bottom row of the cylinder. This initial state is just made from rigid and fluid beads, as shown *e.g.*, in Fig. 2. For most simulations, we just took 5 consecutive fluid beads among the remaining solid beads.

We next define the *target*. It 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*: $\{\sigma_c^*, s_c^*\}_{c=1,\ldots,w}$. Given (i) a gene sequence, which determines the connectivity ℓ_{rck} and (ii) the *input* state, $\{\sigma_{1,c}, s_{1,c}\}_{c=1,\ldots,w}$, the algorithm described above uniquely defines the output state in the top row, $\{\sigma_{h,c}, s_{h,c}\}_{c=1,\ldots,w}$. At each step of evolution, the output state is compared to the fixed, given target, by measuring the Hamming distance, the number of positions where the output differs from the target:

$$F = \sum_{c=1}^{w} [1 - (|s_{h,c} - s_c^*| - 1) \cdot (|\sigma_{h,c} - \sigma_c^*| - 1)]. \quad \text{(A3)}$$

In the biological convention $-F$ is the *fitness* that should increase towards a maximum value of $-F = 0$, when the input-output problem is *solved*.

Solutions are found by *mutations*. 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. A mutation is kept *only if the Hamming distance is not increased as compared to the value before the mutation* (equivalently the fitness is not allowed to decrease). This procedure is repeated until a solution ($F = 0$) is found. This will happen with probability 1, perhaps after very many flips, if the problem has a solution at all. This is really the Metropolis algorithm [53] algorithm (at 0 temperature).

Remark: *It is an important feature of our model that the quality of a 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 is surface (in our case, the ends of the cylinder). One of the surprising outcomes of our study is that this requirement has a strong influence on what happens in the interior of the protein. Also, the propagation of fluidity should not be confused with learning in neural networks, but is rather of the percolation type.*

5. Simulation of evolutionary dynamics

All simulations are done on the $30 \times 18 = 540$ playground, as described above. We have done simulations for many variants of the model, and many targets, but we present only two specific problems, for which the most extensive study was done: In the first, the fluid regions of the input and the target are opposite and of length 6 at the bottom and length 5 at the top. In the second run, top and bottom are the same, but the top is shifted sideways by 5 units. We will call these two examples *straight* and *tilted*, denoted as S and T. We have also studied examples in which the position of the target (relative to the input) is left free, but here we only discuss the results for the 'S' and 'T' case. This serves to illustrate that the results are largely independent of the details of the model. We have studied many other variants, and in all cases, the main results are qualitatively unchanged.

Remark: *We view this as an important outcome of our theory, namely that it illustrates a close connection between gene and protein which goes way beyond the simple model we consider here.*

For both, S and T, we study 200 independent *branches*, starting from a random sequence with about 90% of the bonds present at the start. Given any fixed sequence, we sweep according to the rules of Eq(A1)-(A2) through the net, and measure the Hamming distance F (Eq(A3)) between the last row and the desired target. When this Hamming distance is 0, we consider the problem as solved. If not, we flip randomly a bond (exchanging 0 with 1) and recalculate the Hamming distance. We view this flip as a *mutation* of the sequence, equivalent to mutating one nucleic base in a gene. If the Hamming distance decreases or remains unchanged, we keep the flip, otherwise we backtrack and flip another randomly chosen bond. This is repeated until a solution is found. (This is really a Metropolis algorithm [53] at zero temperature.) Typically, after 10^3-10^5 mutations this input-output problem is solved. Although the functional sequences are extremely sparse among the 2^{2550} possible sequences, the small bias for getting closer to the target in configuration space directs the search rather quickly.

Once a solution is found, we destroy it by further mutations and then look for a new solution, as before, starting from the destroyed state. This we call a *generation*. For each of the 200 branches, we followed 5000 generations, leading to a total of 10^6 solutions. The time to recover from a destroyed state is about 1500 flips per error in that state, which is similar to time it takes to find a solution starting from a random gene. A destruction takes around 11.2 mutations on average.

We also did another 10^6 simulations starting each time from another random configuration. The statistics in both cases are very similar, but the destruction-reconstruction simulations obviously show some correlations between a generation and the next. This effect disappears after about 4 generations.

Appendix B: Results, analysis and interpretation

1. Dimension of solution set

Dimension of a space measures the number of directions in which one can move from a point. In the case of our model, since from any sequence in sequence space one can move along $N_S = 2550$ axes by flipping just one bit, we see that the sequence space has dimension 2550, and the number of different elements in this space is a hypercube with $2^{2550} \sim 10^{768}$ elements.

The set of solutions which we find, has however much smaller dimension, as we show in Fig. 3 for the straight and tilted example. In the case of experimental data, as ours, the dimension is most conveniently determined by the box-counting (Grassberger-Procaccia [30]) algorithm. This is obtained by just counting the number $N(\varrho)$ of pairs at distances $\leq \varrho$, and then finding the slope in a log-log plot. This is indicated by the black lines in Fig. 3 we see that, clearly, the dimension in the space of configurations is about 8-9, while, in the space of sequences, the dimension is basically 'infinite', namely just limited by the maximal slope one can obtain [28].

2. Spectrum in phenotype and genotype spaces

We compute spectra for both the sequences and the configurations, for the 10^6 solutions. Let us detail this for the case of sequences: We have 10^6 binary vectors with $N_S = 2550$ components each, and we want to know the 'typical' spectrum of such vectors. This is conveniently found with the Singular Value Decomposition (SVD), in which one forms a matrix W of size $m \times n = 10^6 \times 2550$. This matrix can be written as $U \cdot D \cdot V^*$, where U is $m \times m$, V is $n \times n$ and D is an $m \times n$ matrix which is diagonal in the sense that only the elements D_{ii} with $i = 1, \ldots, n$ are nonzero. (We assume here that we are in the case $m > n$.) The D_{ii} are in general > 0 and in this case the singular value decomposition is unique. We call the set of the $\lambda_i^G = D_{ii}$ the spectrum of the sequences, and the vectors in V the eigenvectors of the SVD. It is the first few of those which are shown in Fig. 6.

Note that the SVD eigenvalues λ_i^G are the square roots of the spectrum of the covariance matrix $W^{\mathrm{T}}W$ which has the same eigenvectors as W. Therefore the high SVD eigenvalues correspond to the *principal components*, the directions with maximal variation in the solution set.

Mutatis mutandis, we perform the same SVD for the case of the configurations, using the s-values (that is, of the shearability) of vectors of the configurations. (This is reasonable, because, in general, there are very few non-shearable and fluid AAs.)

Apart from the numerical findings, which are shown in Fig. 6 for the straight (S) example and in Fig. 7 for the tilted (T) one, some comments are in order:

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 modes 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 joint effect 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.

The bond-spectrum, top right in Figs. 6 and 7, 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 [54]. 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 they were really random) to find the spectrum at $\frac{1}{2}(\sqrt{10^6} \pm \sqrt{2550})$, which is close to the experiment, and confirms that most of the bonds are just randomly present or absent. We attribute the slight enlargement of the spectrum to memory effects between generation in the same branch. This corresponds to the well-known phylogenetic correlations among descendants in the same tree.

It is tempting to also study the continuous part of this spectrum, which is not quite of the standard form. While in principle, this could be done by taking into account the known correlations, even the techniques of [55] seem difficult to implement. We thank T. Guhr for helpful discussions on his issue.

3. Shear modes in the amino acid network

Consider now either of the two examples, straight or tilted (S and T). A solution of such an example is given by a set of bonds, and this set of bonds defines a graph on the $N_{AA} = h \cdot w = 540$ AAs. This graph is embedded in 2D where $\vec{x}_{r,c}$ are the positions of the AAs, which are connected by straight bonds. We now extend the scope of our study somewhat, by assuming that the bonds are not totally rigid, but given by harmonic springs (see also [24]). This allows us to study mechanical properties which would be too stiff if we only worked with bonds which are rigid sticks.

In this case, the calculations are straightforward, if somewhat complex, and they are, *e.g.*, well explained in [35, pp. 618–619]. We thus consider the elastic tensor, $\mathcal{H}$, which is the tensor product of the network Laplacian with the 2 by 2 tensor of directional derivatives.

For the reader who is unfamiliar with [35], we describe what this means component-wise. The playground $\Omega \subset \mathbf{Z}^2$ has size h in the z-direction and size w in the x direction, with periodic boundary condition in the x direction. All bonds go from some (r, c) to $(r+1, c)$, $(r+1, c\pm1)$, $(r+1, c\pm2)$, again with periodic boundary conditions in the c-direction. Each such bond defines a direction vector (d_z, d_x) in $\mathbf{R}^2$ which we normalize to $d_x^2 + d_z^2 = 1$. Note that this vector depends on both the origin and the target of the bond.

If we imagine harmonic springs between the nodes connected by bonds (all with the same spring constant), then we can define the (symmetric) tensor matrix of deformation energies in the x and y direction by

$$\mathcal{H}'_{km} = M(k, m) \text{ , with } k, m \in \Omega \text{ ,}$$

and where each element of $\mathcal{H}'_{km}$ is—when k and m are connected by a bond—the 2 by 2 matrix (indexed by $i, j \in \{1, 2\}$)

$$M(k,m) = (d_x(k,m), d_z(k,m))^{\mathrm{T}} \otimes (d_x(k,m), d_z(k,m)) = \begin{pmatrix} d_x^2 & d_x d_z \\ d_x d_z & d_z^2 \end{pmatrix} .$$

If k and m are not connected, then $M(k,m)$ is the 0 matrix. The elements of $M(k,m)$ are denoted $M(k,m)_{ij}$.

Finally we complete the $2N \times 2N$ matrix $\mathcal{H}'$ to a 'Laplacian' $\mathcal{H}$ by adding diagonal elements to it, so that the row (and column) sums are 0. In components, this means that we require, for each $k \in \Omega$ and each $i, j \in \{1, 2\}$, the sums

$$\sum_{\ell} (\mathcal{H}_{km})_{ij}$$

to vanish. Other properties of A are described in [35].

Since we take periodic boundary conditions in the x direction, there will always be a (simple) 0 eigenvalue of $\mathcal{H}$ in this direction. Other 0 eigenvalues correspond to translation in the z direction or rotation in the $x - z$ plane. Another type of (double) 0 eigenvalues are associated with any patch of nodes which is totally disconnected from the rest of the lattice. Since the density ϱ of bonds is about $1/2$ and otherwise quite random, and there are twice 5 bonds at each interior node we expect (assuming random distribution of bonds) there to be about $N \cdot 2^{-10} \sim 0.001N$ isolated nodes, *i.e.*, isolated singletons, and even fewer patches of greater size.

Further zero modes come from nodes which can oscillate sideways without first order effects. This will happen if a node is only connected by one bond. Since $\varrho \sim 1/2$, the probability of finding such a node is about

$$N \frac{\binom{10}{1}}{2^{10}} \sim 0.01N .$$

Thus, we show in Figures 6 and 7 the eigenfunctions only for the first eigenvalues after the trivial ones. Due to the tensorial nature of the problem, the eigenvectors have two components, which we show as 2D shear-flow.

4. Genetic correlation matrix

In Fig. 5, we study the correlations among the 10^6 solutions in sequence space. Given the matrix W_{ij}, of all sequences, with $i = 1, \ldots, N = 10^6$, $j = 1, \ldots, 2550$ (of binary digits), we compute the means $\langle W_{\cdot\, j} \rangle = \sum_{i=1}^{N} W_{ij}/N$ and the standard deviations $\mathrm{std}_j = \left(\sum_i |W_{ij} - \langle W_{\cdot\, j} \rangle|^2 \right)^{1/2}$. Then, in the usual way, we form $M_{ij} = W_{ij} - \langle W_{\cdot\, j} \rangle$ and

$$C_{j,j'} = \frac{(M^* M)_{j,j'}}{\mathrm{std}_j \mathrm{std}_{j'}} .$$

Fig. 5 then shows $\log(|C_{j,j'}|)$, with the autocorrelation C_{jj} omitted.

Note that both, the means and the variances depend very weakly on j. Fig. 5 reveals and reinforces several observations also made in other calculations of this paper. First, looking onto the axis $j = j'$ in the figure one sees a periodicity of the patterns corresponding to the 17 gaps between the 18 rows of the configuration space. This reflects the necessity to maintain a *connected* liquid channel. Also, as seen in Fig. 5, the correlations grow somewhat towards the ends, especially toward the upper ($j = 2550$) end. This is because of the mechanical constraint which forces the channel to become more precise towards the ends, in analogy with Fig. 8B.

The periodic patterns all over the square reflect not only the natural periodicity of 150 ($= 5 \cdot w$) elements in the sequence, but also show that the boundaries of the channel form a special shell (with *two* peaks per row).

5. Survival under mutations

Here, we ask how robust the solutions are as further mutations take place. First, we determine how many mutations lead to a destruction of the solution. The statistics of this is shown in Fig. 8. We note that about 10% of all solutions are destroyed by just one mutation, while there is an exponential decay of survival of m mutations. This signals that the mutations act independently.

One can also ask *where* the critical mutations take place. This is illustrated in Fig. 8B, and was discussed in the main text. We have also studied the places where exactly *two* mutations will kill a solution (and none of the 2 is a single site 'killer') (Fig. 8C) and in these cases, one finds that the two mutations are generally close to each other, acting on the same site. Again, the channel is less vulnerable to mutations but now the mutations are evenly distributed over the rest of the network.

6. Expansion of the protein universe

Let us explain in further detail how Fig. 4 was obtained. Here, we test our model against the ideas of [6]. Our results will give some insight about the nature of the graph of solutions. First, we describe the question as it is found in [6]. Take any two solutions and consider their gene sequences s_1 and s_2. They will have a Hamming distance $d(s_1, s_2)$, which we normalize by dividing by 2550 (the number of elements in $s_i, i = 1, 2$), which we call the protein universe diameter. The question is how much the solution following one generation after s_2 differs from s_1. If we call that solution s_3, then the observed quantity is defined as follows: Let $w_i = 1$ if $s_{1,i} = 1$ and -1 if $s_{1,i} = 0$, for $i = 1, \ldots 2550$. Then for each i let $x_i = w_i \cdot (s_{3,i} - s_{2,i})$. Note that $x_i > 0$ if the change between $s_{3,i}$ and $s_{3,i}$ is *towards* s_1 and < 0 if it is *away from* s_1. Finally, $N_{\text{away}} = \sum_{i:w_i<0} 1$ and $N_{\text{towards}} = \sum_{i:w_i>0} 1$, and we plot in Fig. 4 $N_{\text{towards}}/N_{\text{away}}$ as a function of D.

In Fig. 4 we show the results for data set S, (the plot for set T looks similar). The black curve is nothing but $D/(1 - D)$,

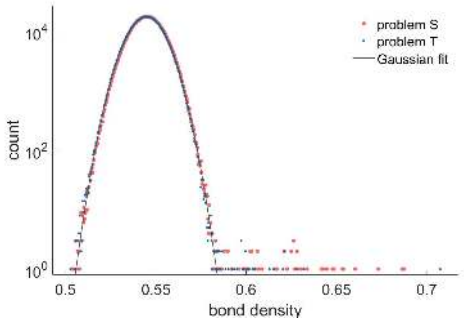

FIG. 13. The distributions of the bond densities for the 10^6 solutions. Note that these densities are just like random Gaussian variables, except for the outliers.

where D is the normalized Hamming distance, *i.e.*, the proportion of sites which are different between s_1 and s_2. The fit to this curve tells us an important aspect about the set of possible solutions. Note that the set of all possible s forms a hypercube of dimension 2550 with 2^{2550} corners. The set of solutions is a very small subset of this hypercube, where all corners which are not solutions have been taken away, including the bonds leading to these corners. This leads to a very complicated sub-graph of the hypercube. While we do not have a good mathematical description of how it looks, the good fit shows that the comparisons between s_1, s_2, and s_3 are *as if one performed a random walk on the full cube*. (Note that such a result must be intimately connected to the high dimension of the problem, since for low dimensional hypercubes it does not hold.) Almost all solutions are at the edge of the universe, where the typical Hamming distances among the sequences are close to the typical distance between random sequences,

7. Flexibility of solutions: thermal stability

The histogram of the density of links for the 10^6 solutions is shown in Fig. 13. These distributions are obtained for simulations in which links are flipped randomly in a symmetric fashion. One can easily push these densities somewhat up or down, by favoring/restricting the flips of links towards 1. However, much more extreme solutions can be found by deterministic procedures which turn as many links to 1 resp. 0. In these cases, we have obtained densities of as high as 0.96 and as low as 0.14, that is, 2452/2550 links, resp. 372/2550 links. Two such extreme cases are illustrated in Fig. 10. This shows that the model, if needed, can be adapted to questions of temperature dependence of the protein, for example, by giving more or less weight to the number of bonds, something like a chemical potential in statistical mechanics.

ACKNOWLEDGMENTS

We thank Stanislas Leibler, Michael R. Mitchell, Elisha Moses and Giovanni Zocchi for essential discussions and encouragement. JPE is 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.

[1] Eugene V. Koonin, Yuri I. Wolf, and Georgy P. Karev, "The structure of the protein universe and genome evolution," Nature **420**, 218–223 (2002).
[2] Y. Xia and M. Levitt, "Simulating protein evolution in sequence and structure space," Curr Opin Struct Biol **14**, 202–207 (2004).
[3] Ken A. Dill and Justin L. MacCallum, "The protein-folding problem, 50 years on," Science **338**, 1042–1046 (2012).
[4] Konstantin B Zeldovich and Eugene I Shakhnovich, "Understanding protein evolution: from protein physics to darwinian selection," Annu Rev Phys Chem **59**, 105–127 (2008).
[5] David A. Liberles, Sarah A. Teichmann, Ivet Bahar, Ugo Bastolla, Jesse Bloom, Erich Bornberg-Bauer, Lucy J. Colwell, A. P. Jason de Koning, Nikolay V. Dokholyan, Julian Echave, Arne Elofsson, Dietlind L. Gerloff, Richard A. Goldstein, Johan A. Grahnen, Mark T. Holder, Clemens Lakner, Nicholas Lartillot, Simon C. Lovell, Gavin Naylor, Tina Perica, David D. Pollock, Tal Pupko, Lynne Regan, Andrew Roger, Nimrod Rubinstein, Eugene Shakhnovich, Kimmen Sjölander, Shamil Sunyaev, Ashley I. Teufel, Jeffrey L. Thorne, Joseph W. Thornton, Daniel M. Weinreich, and Simon Whelan, "The interface of protein structure, protein biophysics, and molecular evolution," Protein Sci **21**, 769–785 (2012).
[6] Inna S. Povolotskaya and Fyodor A. Kondrashov, "Sequence space and the ongoing expansion of the protein universe," Nature **465**, 922–926 (2010).
[7] A. D. Keefe and J. W. Szostak, "Functional proteins from a random-sequence library," Nature **410**, 715–718 (2001).
[8] Patrice Koehl and Michael Levitt, "Protein topology and stability define the space of allowed sequences," Proceedings of the National Academy of Sciences **99**, 1280–1285 (2002).
[9] DE Koshland, "Application of a theory of enzyme specificity to protein synthesis," Proc Natl Acad Sci U S A **44**, 98–104 (1958).
[10] K. A. Henzler-Wildman, V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hubner, and D. Kern, "Intrinsic motions along an enzymatic reaction trajectory," Nature **450**, 838–U13 (2007).
[11] Y. Savir and T. Tlusty, "Conformational proofreading: the impact of conformational changes on the specificity of molecular recognition," PLoS One **2**, e468 (2007).
[12] T. M. Schmeing, R. M. Voorhees, A. C. Kelley, Y. G. Gao, F. V. th Murphy, J. R. Weir, and V. Ramakrishnan, "The crystal structure of the ribosome bound to ef-tu and aminoacyl-trna," Science **326**, 688–94 (2009).
[13] Y. Savir and T. Tlusty, "Reca-mediated homology search as a nearly optimal signal detection system," Mol Cell **40**, 388–396 (2010).
[14] Morgan Huse and John Kuriyan, "The conformational plasticity of protein kinases," Cell **109**, 275–282 (2002).

[15] Y. Savir and T. Tlusty, "The ribosome as an optimal decoder: A lesson in molecular recognition," Cell **153**, 471–479 (2013).

[16] M. F. Perutz, "Stereochemistry of cooperative effects in haemoglobin: Haem-haem interaction and the problem of allostery," Nature **228**, 726–734 (1970).

[17] N. M. Goodey and S. J. Benkovic, "Allosteric regulation and catalysis emerge via a common route," Nat Chem Biol **4**, 474–482 (2008).

[18] Steve W. Lockless and Rama Ranganathan, "Evolutionarily conserved pathways of energetic connectivity in protein families," Science **286**, 295–299 (1999).

[19] Allan Chris M. Ferreon, Josephine C. Ferreon, Peter E. Wright, and Ashok A. Deniz, "Modulation of allostery by protein intrinsic disorder," Nature **498**, 390–394 (2013).

[20] H. Qu and G. Zocchi, "How enzymes work: A look through the perspective of molecular viscoelastic properties," Phys Rev X **3** (2013).

[21] C. Joseph, C. Y. Tseng, G. Zocchi, and T. Tlusty, "Asymmetric effect of mechanical stress on the forward and reverse reaction catalyzed by an enzyme," PLoS One **9** (2014), e101442.

[22] Michael R. Mitchell, Tsvi Tlusty, and Stanislas Leibler, "Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings," Proc Natl Acad Sci U S A **113**, E5847–E5855 (2016).

[23] M. Gerstein, A. M. Lesk, and C. Chothia, "Structural mechanisms for domain movements in proteins," Biochemistry **33**, 6739–6749 (1994).

[24] S. Dutta, J-P. Eckmann, and T. Tlusty, To be published.

[25] S. Alexander, C. Laermans, R. Orbach, and H. M. Rosenberg, "Fraction interpretation of vibrational properties of cross-linked polymers, glasses, and irradiated quartz," Phys Rev B **28**, 4615–4619 (1983).

[26] J. C. Phillips and M. F. Thorpe, "Constraint theory, vector percolation and glass-formation," Solid State Commun **53**, 699–702 (1985).

[27] S. Alexander, "Amorphous solids: Their structure, lattice dynamics and elasticity," Phys Rep **296**, 65–236 (1998).

[28] Itamar Procaccia, "Complex or just complicated?" Nature **333**, 498–499 (1988).

[29] J.-P. Eckmann and D. Ruelle, "Fundamental limitations for estimating dimensions and lyapunov exponents in dynamical systems," Physica D **56**, 185–187 (1992).

[30] Peter Grassberger and Itamar Procaccia, "Characterization of strange attractors," Phys Rev Lett **50**, 346–349 (1983).

[31] Y. Savir, E. Noor, R. Milo, and T. Tlusty, "Cross-species analysis traces adaptation of rubisco toward optimality in a low-dimensional landscape," Proc Natl Acad Sci U S A **107**, 3475–3480 (2010).

[32] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, "Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space," Science **336**, 1157–1160 (2012).

[33] Tiberiu Teşileanu, Lucy J. Colwell, and Stanislas Leibler, "Protein sectors: Statistical coupling analysis versus conservation," PLoS Comput Biol **11**, e1004091 (2015).

[34] Tsvi Tlusty, "Self-referring dna and protein: a remark on physical and geometrical aspects," Phil. Trans. Roy. Soc. A **374** (2016).

[35] F. R. K. Chung and S. Sternberg, "Laplacian and vibrational-spectra for homogeneous graphs," Journal of Graph Theory **16**, 605–627 (1992).

[36] Rainer Jaenicke and Gerald Böhm, "The stability of proteins in extreme environments," Curr Opin Struct Biol **8**, 738–748 (1998).

[37] Ken A. Dill, "Theory for the folding and stability of globular proteins," Biochemistry (Mosc) **24**, 1501–1509 (1985).

[38] Tsvi Tlusty, "A model for the emergence of the genetic code as a transition in a noisy information channel," J Theor Biol **249**, 331–342 (2007).

[39] Jean-Pierre Eckmann, "Trading codes for errors," Proceedings of the National Academy of Sciences **105**, 8165–8166 (2008).

[40] Tsvi Tlusty, "A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes," Phys Life Rev **7**, 362–376 (2010).

[41] G. Casari, C. Sander, and A. Valencia, "A method to predict functional residues in proteins," Nat Struct Biol **2**, 171–8 (1995).

[42] Arhonda Gogos, Derek Jantz, Sema Sentürker, Delwood Richardson, Miral Dizdaroglu, and Neil D. Clarke, "Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: An experimental test using dna glycosylase homologs," Proteins: Struct , Funct , Bioinf **40**, 98–105 (2000).

[43] O. Rivoire, K. A. Reynolds, and R. Ranganathan, "Evolution-based functional decomposition of proteins," PLoS Comput Biol **12** (2016), ARTN e1004817 10.1371/journal.pcbi.1004817.

[44] Hervé Jacquier, André Birgy, Hervé Le Nagard, Yves Mechulam, Emmanuelle Schmitt, Jérémy Glodt, Beatrice Bercot, Emmanuelle Petit, Julie Poulain, Guilne Barnaud, Pierre-Alexis Gros, and Olivier Tenaillon, "Capturing the mutational landscape of the beta-lactamase tem-1," Proc. Nat. Acad. Sci. USA **110**, 13067–13072 (2013).

[45] Benjamin P. Roscoe, Kelly M. Thayer, Konstantin B. Zeldovich, David Fushman, and Daniel N. A. Bolon, "Analyses of the effects of all ubiquitin point mutants on yeast growth rate," J Mol Biol **425**, 1363–1377 (2013).

[46] Elad Firnberg, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier, "A comprehensive, high-resolution map of a genes fitness landscape," Mol Biol Evol **31**, 1581–1592 (2014).

[47] Dmitry A. Kondrashov and Fyodor A. Kondrashov, "Topological features of rugged fitness landscapes in sequence space," Trends Genet **31**, 24–33 (2015).

[48] Michael Levitt, Christian Sander, and Peter S. Stern, "Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme," J Mol Biol **181**, 423–447 (1985).

[49] M. M. Tirion and D. Benavraham, "Normal mode analysis of g-actin," J Mol Biol **230**, 186–195 (1993).

[50] I. Bahar and A. J. Rader, "Coarse-grained normal mode analysis in structural biology," Curr Opin Struct Biol **15**, 586–592 (2005).

[51] W. J. Zheng, B. R. Brooks, and D. Thirumalai, "Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations," Proc Natl Acad Sci U S A **103**, 7664–7669 (2006).

[52] JeremyL England, "Allostery in protein domains reflects a balance of steric and hydrophobic effects," Structure **19**, 967–975 (2011).

[53] N Metropolis, AW Rosenbluth, MN Rosenbluth, AH Teller, and E Teller, "Equation of state calculations by fast computing machines," J Chem Phys **21**, 1087–1092 (1953).

[54] V. A. Marčenko and L. A. Pastur, "The spectrum of random matrices," Teor. Funkciĭ Funkcional. Anal. i Priložen. Vyp. **4**, 122–145 (1967).

[55] Thomas Guhr, Axel Müller-Groeling, and Hans A. Weidenmüller, "Random-matrix theories in quantum physics: common concepts," Phys Rep **299**, 189–425 (1998).