# Green function of correlated genes in a minimal mechanical model of protein evolution

Sandipan Dutta[a], Jean-Pierre Eckmann[b], Albert Libchaber[c,1], and Tsvi Tlusty[a,d,1]

[a]Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, Korea; [b]Département de Physique Théorique and Section de Mathématiques, Université de Genève, CH-1211 Geneva 4, Switzerland; [c]Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10021; and [d]Department of Physics, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea

Contributed by Albert Libchaber, November 20, 2017 (sent for review September 25, 2017; reviewed by Mukund Thattai and Massimo Vergassola)

**The function of proteins arises from cooperative interactions and rearrangements of their amino acids, which exhibit large-scale dynamical modes. Long-range correlations have also been revealed in protein sequences, and this has motivated the search for physical links between the observed genetic and dynamic cooperativity. We outline here a simplified theory of protein, which relates sequence correlations to physical interactions and to the emergence of mechanical function. Our protein is modeled as a strongly coupled amino acid network with interactions and motions that are captured by the mechanical propagator, the Green function. The propagator describes how the gene determines the connectivity of the amino acids and thereby, the transmission of forces. Mutations introduce localized perturbations to the propagator that scatter the force field. The emergence of function is manifested by a topological transition when a band of such perturbations divides the protein into subdomains. We find that epistasis—the interaction among mutations in the gene—is related to the nonlinearity of the Green function, which can be interpreted as a sum over multiple scattering paths. We apply this mechanical framework to simulations of protein evolution and observe long-range epistasis, which facilitates collective functional modes.**

protein evolution | epistasis | genotype-to-phenotype map | Green function | dimensional reduction

A common physical basis for the diverse biological functions of proteins is the emergence of collective patterns of forces and coordinated displacements of their amino acids (1–13). In particular, the mechanisms of allostery (14–18) and induced fit (19) often involve global conformational changes by hinge-like rotations, twists, or shear-like sliding of protein subdomains (20–22). An approach to examine the link between function and motion is to model proteins as elastic networks (23–26). Decomposing the dynamics of the network into normal modes revealed that low-frequency "soft" modes capture functionally relevant large-scale motion (27–30), especially in allosteric proteins (31–33). Recent works associate these soft modes with the emergence of weakly connected regions in the protein (Fig. 1 *A* and *B*)—"cracks," "shear bands," or "channels" (21, 22, 34–36)—that enable viscoelastic motion (37, 38). Such patterns of "floppy" modes (39–42) emerge in models of allosteric proteins (36, 43–45) and networks (46–48).

Like their dynamic phenotypes, proteins' genotypes are remarkably collective. When aligned, sequences of protein families show long-range correlations among the amino acids (49–61). The correlations indicate epistasis, the interaction among mutations that takes place among residues linked by physical forces or common function. By inducing nonlinear effects, epistasis shapes the protein's fitness landscape (62–68). Provided with sufficiently large data, analysis of sequence variation can predict the 3D structure of proteins (50–52), allosteric pathways (53–55), epistatic interactions (56, 57), and coevolving subsets of amino acids (58–60, 69).

Still, the mapping between sequence correlation and collective dynamics—and in particular, the underlying epistasis—is not fully understood. Experiments and simulations provide valuable information on protein dynamics, and extensive sequencing accumulates databases required for reliable analysis; however, there remain inherent challenges: the complexity of the physical interactions and the sparsity of the data. The genotype-to-phenotype map of proteins connects spaces of huge dimension, which are hard to sample, even by high-throughput experiments or natural evolution (70–72). A complementary approach is the application of simplified coarse-grained models, such as lattice proteins (73–75) or elastic networks (24), which allow one to extensively survey the map and examine basic questions of protein evolution. Such models have been recently used to study allosteric proteins (35, 36, 43–45) and in networks (46–48). Our aim here is different: to construct a simplified model of how the collective dynamics of functional proteins directs their evolution and in particular, to give a mechanical interpretation of epistasis.

This paper introduces a coarse-grained theory that treats protein as an evolving amino acid network with topology that is encoded in the gene. Mutations that substitute one amino acid with another tweak the interactions, allowing the network to evolve toward a specific mechanical function: in response to a localized force, the protein will undergo a large-scale conformational change (Fig. 1 *C* and *D*). We show that the application of a Green function (76, 77) is a natural way to understand the protein's collective dynamics. The Green function measures how the protein responds to a localized impulse via propagation of forces and motion. The propagation of mechanical response across the protein defines its fitness and directs the evolutionary search.

## Significance

**Many protein functions involve large-scale motion of their amino acids, while alignment of their sequences shows long-range correlations. This has motivated search for physical links between genetic and phenotypic collective behaviors. The major challenge is the complex nature of protein: nonrandom heteropolymers made of 20 species of amino acids that fold into a strongly coupled network. In light of this complexity, simplified models are useful. Our model describes protein in terms of the Green function, which directly links the gene to force propagation and collective dynamics in the protein. This allows for derivation of basic determinants of evolution, such as fitness landscape and epistasis, which are often hard to calculate.**

Author contributions: S.D., J.-P.E., A.L., and T.T. designed and performed research and wrote the paper.

Reviewers: M.T., Tata Institute of Fundamental Research, National Center for Biological Sciences; and M.V., University of California, San Diego.

The authors declare no conflict of interest.



[1]To whom correspondence may be addressed. Email: libchbr@rockefeller.edu or tsvitlusty@gmail.com.

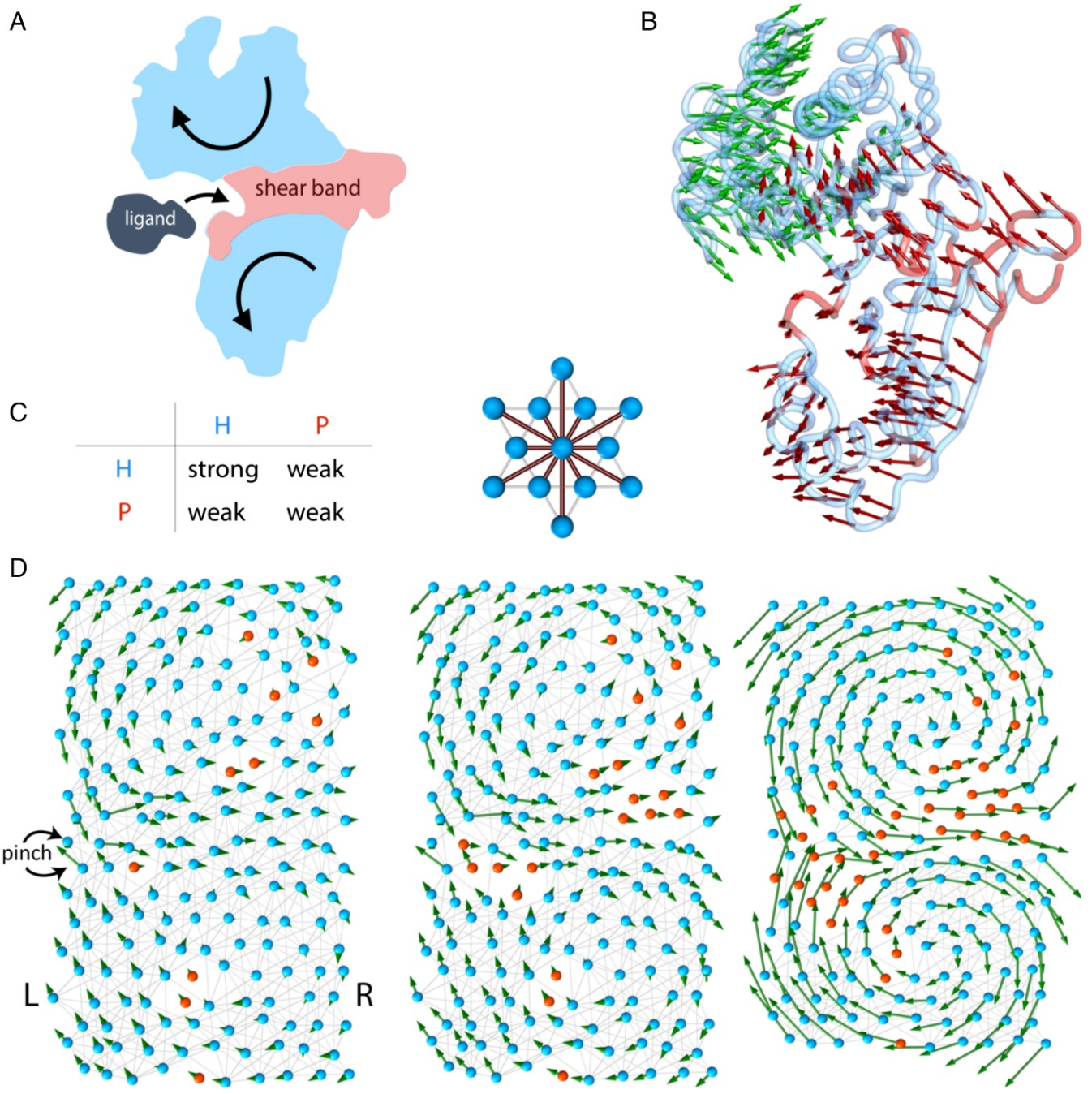


**Fig. 1.** Protein as an evolving machine and propagation of mechanical forces. (*A*) Formation of a softer shear band (red) separating the protein into two rigid subdomains (light blue). When a ligand binds, the biochemical function involves a low-energy hinge-like or shear motion (arrows). (*B*) Shear band and large-scale motion in a real protein: the arrows show the displacement of all amino acids in human glucokinase when it binds glucose (Protein Data Bank ID codes 1v4s and 1v4t). The coloring shows a high-shear region (red) separating two low-shear domains that move as rigid bodies (shear calculated as in refs. 21 and 36). (*C*) The mechanical model. The protein is made of two species of amino acids, polar (P; red) and hydrophobic (H; blue), with a sequence that is encoded in a gene. Each amino acid forms weak or strong bonds with its 12 near neighbors (*Right*) according to the interaction rule in the table (*Left*). (*D*) The protein is made of $10 \times 20 = 200$ amino acids with positions that are randomized from a regular triangular lattice. Strong bonds are shown as gray lines. Evolution begins from a random configuration (*Left*) and evolves by mutating one amino acid at each step, switching between H and P. The fitness is the mechanical response to a localized force probe (pinch) (2). After $\sim 10^3$ mutations (*Center*; intermediate stage), the evolution reaches a solution (*Right*). The green arrows show the mechanical response: a hinge-like, low-energy motion with a shear band starting at the probe and traversing the protein, qualitatively similar to *B*. L, left; R, right.

Thus, the Green function explicitly defines the map: gene → amino acid network → protein dynamics → function. We use this map to examine the effects of mutations and epistasis. A mutation perturbs the Green function and scatters the propagation of force through the protein (Fig. 2). We quantify epistasis in terms of "multiple scattering" pathways. These indirect physical interactions appear as long-range correlations in the co-evolving genes.

Using a Metropolis-type evolution algorithm, solutions are quickly found, typically after $\sim 10^3$ steps. Mutations add localized

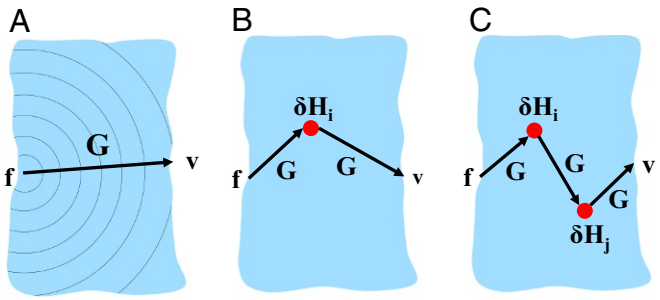


**Fig. 2.** Force propagation, mutations, and epistasis. (*A*) The Green function **G** measures the propagation of the mechanical signal, depicted as a "diffraction wave," across the protein (blue) from the force source **f** (pinch) to the response site **v**. (*B*) A mutation $\delta\mathbf{H}_i$ deflects the propagation of force. The effect of the mutation on the propagator $\delta\mathbf{G}$ can be described as a series of multiple scattering paths (6). (*C*) The epistasis between two mutations, $\delta\mathbf{H}_i$ and $\delta\mathbf{H}_j$, is equivalent to a series of multiple scattering paths (9).

perturbations to the amino acid network, which are eventually arranged by evolution into a continuous shear band. Protein function is signaled by a topological transition, which occurs when a shearable band of weakly connected amino acids separates the protein into rigid subdomains. The set of solutions is sparse: there is a huge reduction of dimension between the space of genes to the spaces of force and displacement fields. We find a tight correspondence between correlations in the genotype and phenotype. Owing to its mechanical origin, epistasis becomes long ranged along the high-shear region of the channel.

## Model: Protein as an Evolving Machine

**The Amino Acid Network and Its Green Function.** We use a coarse-grained description in terms of an elastic network (23–27, 39) with connectivity and interactions that are encoded in a gene (Fig. 1 *C* and *D*). Similar vector elasticity models were considered in refs. 35 and 36 (app. B3 therein). The protein is a chain of $n_a = 200$ amino acids: $a_i$ $(i = 1, \ldots, n_a)$ folded into a $10 \times 20$ 2D hexagonal lattice $(d = 2)$. We follow the HP model (73, 74) with its two species of amino acids, hydrophobic $(a_i = \mathrm{H})$ and polar $(a_i = \mathrm{P})$. The amino acid chain is encoded in a gene **c**, a sequence of 200 binary codons, where $c_i = 1$ encodes an H amino acid and $c_i = 0$ encodes a P amino acid.

We consider a constant fold, and therefore, any particular codon $c_i$ in the gene encodes an amino acid $a_i$ at a certain constant position $\mathbf{r}_i$ in the protein. The positions $\mathbf{r}_i$ are randomized to make the network amorphous. These $n_d = d \cdot n_a = 400$ dfs are stored in a vector **r**. Except the ones at the boundaries, every amino acid is connected by harmonic springs to $z = 12$ nearest and next nearest neighbors. There are two flavors of bonds according to the chemical interaction, which is defined as an AND gate: a strong $\mathrm{H}-\mathrm{H}$ bond and weak $\mathrm{H}-\mathrm{P}$ and $\mathrm{P}-\mathrm{P}$ bonds. The strength of the bonds determines the mechanical response of the network to a displacement field **u**, when the amino acids are displaced as $\mathbf{r}_i \to \mathbf{r}_i + \mathbf{u}_i$. The response is captured by Hooke's law that gives the force field **f** induced by a displacement field, $\mathbf{f} = \mathbf{H}(\mathbf{c})\,\mathbf{u}$. The analogue of the spring constant is the Hamiltonian $\mathbf{H}(\mathbf{c})$, a $n_d \times n_d$ matrix, which records the connectivity of the network and the strength of the bonds. $\mathbf{H}(\mathbf{c})$ is a nonlinear function of the gene **c**, reflecting the amino acid interaction rules of Fig. 1*C* (Eq. **11**, *Materials and Methods*).

Evolution searches for a protein that will respond by a prescribed large-scale motion to a given localized force **f** ("pinch"). In induced fit, for example, specific binding of a substrate should induce global deformation of an enzyme. The response **u** is determined by the Green function **G** (76):

$$\mathbf{u} = \mathbf{G}(\mathbf{c})\,\mathbf{f}. \qquad [1]$$

**G** is the mechanical propagator that measures the transmission of signals from the force source **f** across the protein (Fig. 2*A*). Eq. **1** constitutes an explicit genotype-to-phenotype map from the genotype **c** to the mechanical phenotype **u**: $\mathbf{c} \to \mathbf{u}(\mathbf{c}) = \mathbf{G}(\mathbf{c})\mathbf{f}$. This reflects the dual nature of the Green function **G**: in the phenotype space, it is the linear mechanical propagator that turns a force into motion, $\mathbf{u} = \mathbf{G}\,\mathbf{f}$, whereas it is also the nonlinear function that maps the gene into a propagator, $\mathbf{c} \to \mathbf{G}(\mathbf{c})$.

When the protein is moved as a rigid body, the lengths of the bonds do not change, and the elastic energy cost vanishes. A 2D protein has $n_0 = 3$ such zero modes (Galilean symmetries), two translations, and one rotation, and **H** is, therefore, always singular. Hence, Hooke's law and [**1**] imply that **G** is the pseudoinverse of the Hamiltonian, $\mathbf{G}(\mathbf{c}) = \mathbf{H}(\mathbf{c})^{+}$ (78, 79), which amounts to inversion of **H** in the nonsingular subspace of the $n_d - n_0 = 397$ nonzero modes (*Materials and Methods*). A related quantity is the resolvent, $\mathbf{G}(\omega) = (\omega - \mathbf{H})^{-1}$, with poles at the energy levels of **H**, $\omega = \lambda_k$.

The fitness function rewards strong mechanical response to a localized probe (pinch in Fig. 1*D*): a force dipole at two neighboring amino acids $p'$ and $q'$ on the left side of the protein (L in Fig. 1*D*), $\mathbf{f}_{q'} = -\mathbf{f}_{p'}$. The prescribed motion is specified by a displacement vector **v**, with a dipolar response, $\mathbf{v}_q = -\mathbf{v}_p$, on the right side of the protein (R in Fig. 1*D*). The protein is fitter if the pinch **f** produces a large deformation in the direction specified by **v**. To this end, we evolve the amino acid network to increase a fitness function $F$, which is the projection of the displacement $\mathbf{u} = \mathbf{G}\mathbf{f}$ on the prescribed response **v**:

$$F(\mathbf{c}) = \mathbf{v}^{\mathsf{T}}\mathbf{u} = \mathbf{v}^{\mathsf{T}}\mathbf{G}(\mathbf{c})\,\mathbf{f}. \qquad [2]$$

Eq. **2** defines the fitness landscape $F(\mathbf{c})$. Here, we examine particular examples for a localized pinch **f** and prescribed response **v**, which drive the emergence of a hinge-like mode. This approach is general and can as well treat more complex patterns of force and motion.

**Evolution Searches in the Mechanical Fitness Landscape.** Our simulations search for a prescribed response **v** induced by a force **f** applied at a specific site on the left side (pinch). The prescribed dipolar response may occur at any of the sites on the right side. This gives rise to a wider shear band that allows the protein to perform general mechanical tasks (unlike specific allostery tasks of communicating between specified sites on L and R). We define the fitness as the maximum of $F$ [**2**] over all potential locations of the channel's output (typically 8–10 sites) (*Materials and Methods*). The protein is evolved via a point mutation process where, at each step, we flip a randomly selected codon between zero and one. This corresponds to exchanging H and P at a random position in the protein, thereby changing the bond pattern and the elastic response by softening or stiffening the amino acid network.

Evolution starts from a random protein configuration encoded in a random gene. Typically, we take a small fraction of amino acid of type P (about 5%) randomly positioned within a majority of H (Fig. 1*D*, *Left*). The high fraction of strong bonds renders the protein stiff and therefore, of low initial fitness $F \simeq 0$. At each step, we calculate the change in the Green function $\delta\mathbf{G}$ (by a method explained below) and use it to evaluate from [**2**] the fitness change $\delta F$:

$$\delta F = \mathbf{v}^{\mathsf{T}}\delta\mathbf{G}\,\mathbf{f}. \qquad [3]$$

The fitness change $\delta F$ determines the fate of the mutation: we accept the mutation if $\delta F \geq 0$; otherwise, the mutation is

rejected. Since fitness is measured by the criterion of strong mechanical response, it induces softening of the amino acid network.

The typical evolution trajectory lasts about $10^3$ steps. Most are neutral mutations ($\delta F \simeq 0$) and deleterious ones ($\delta F < 0$); the latter are rejected. About a dozen or so beneficial mutations ($\delta F > 0$) drive the protein toward the solution (Fig. 3*A*). The increase in the fitness reflects the gradual formation of the channel, while the jump in the shear signals the emergence of the soft mode. The first few beneficial mutations tend to form weakly bonded P-enriched regions near the pinch site on the left side and close to the right boundary of the protein. The following ones join these regions into a floppy channel (a shear band), which traverses the protein from left to right. We stop the simulation when the fitness reaches a large positive value $F_{\rm m} \sim 5$. The corresponding gene $\mathbf{c}_*$ encodes the functional protein. The ad hoc value $F_{\rm m} \sim 5$ signals slowing down of the fitness curve toward saturation at $F > F_{\rm m}$, as the channel has formed and now only continues to slightly widen. In this regime, even a tiny pinch will easily excite a large-scale motion with a distinct high-shear band (Fig. 1*D*, *Right*).

## Results

**Mechanical Function Emerges at a Topological Transition.** The hallmark of evolution reaching a solution gene $\mathbf{c}_*$ is the emergence of a new zero-energy mode, $\mathbf{u}_*$, in addition to the three Galilean symmetry modes. Near the solution, the energy of this mode $\lambda_*$ almost closes the spectral gap, $\lambda_* \to 0$, and $\mathbf{G}(\omega)$ has a pole at $\omega \approx 0$. As a result, the emergent mode dominates the Green function, $\mathbf{G} \simeq \mathbf{u}_*\mathbf{u}_*^{\intercal}/\lambda_*$. The response to a pinch will be mostly through this soft mode, and the fitness [**2**] will diverge as

$$F(\mathbf{c}_*) \simeq F_* = \frac{(\mathbf{v}^{\intercal}\mathbf{u}_*)(\mathbf{u}_*^{\intercal}\mathbf{f})}{\lambda_*}. \qquad [4]$$

On average, we find that the fitness increases exponentially with the number of beneficial mutations (Fig. 3*A*). However, beneficial mutations are rare and are separated by long stretches

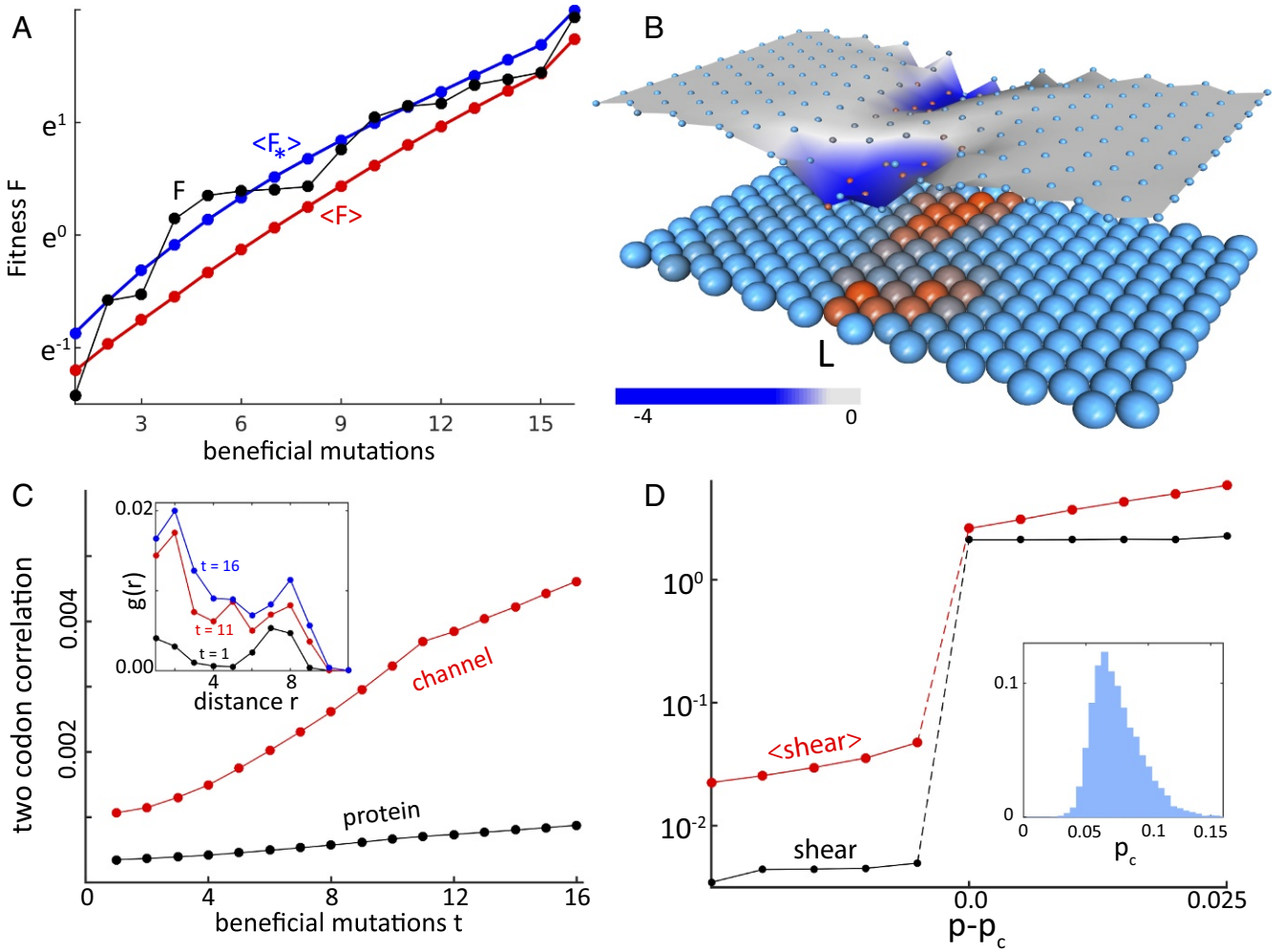


**Fig. 3.** The mechanical Green function and the emergence of protein function. (*A*) Progression of the fitness *F* during the evolution run shown in Fig. 1*D* (black) together with the fitness trajectory averaged over $\sim 10^6$ runs $\langle F \rangle$ (red). Shown are the last 16 beneficial mutations toward the formation of the channel. The contribution of the emergent low-energy mode $\langle F_* \rangle$ (blue) dominates the fitness [**4**]. (*B*) Landscape of the fitness change $\delta F$ [**3**] averaged over $\sim 10^6$ solutions for all 200 possible positions of point mutations at a solution. Underneath, the average amino acid 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 deleterious on average. L, left. (*C*) The average magnitude of the two-codon correlation $|Q_{ij}|$ [**5**] in the shear band (amino acids in rows 7–13; red) and in the whole protein (black) as a function of the number of beneficial mutations, *t*. (*Inset*) Profile of the spatial correlation $g(r)$ within the shear band (after $t = 1$, 11, 16 beneficial mutations). (*D*) The mean shear in the protein in a single run (black) and averaged over $\sim 10^6$ solutions (red) as a function of the fraction of P amino acids, *p*. The values of *p* are shifted by the position of the jump, $p_c$. (*Inset*) Distribution of $p_c$.

of neutral mutations. This is evident from the fitness landscape (Fig. 3*B*), which shows that, in most sites, the effect of mutations is practically neutral. The vanishing of the spectral gap, $\lambda_* \to 0$, manifests as a topological change in the system: the amino acid network is now divided into two domains that can move independently of each other at low energetic cost. The soft mode appears at a dynamical phase 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. 3*D*).

As the shear band is taking shape, the correlation among codons builds up. To see this, we align genes from the $\sim 10^6$ simulations in analogy to sequence alignment of real protein families (49–61), albeit without the phylogenetic correlation that hampers the analysis of real sequences. At each time step, we calculate the two-codon correlation $Q_{ij}$ between all pairs of codons $c_i$ and $c_j$:

$$Q_{ij} \equiv \langle c_i c_j \rangle - \langle c_i \rangle \langle c_j \rangle, \qquad [5]$$

where brackets denote ensemble averages. We find that most of the correlation is concentrated in the forming channel (Fig. 3*C*), where it is 10-fold larger than in the whole protein. In the channel, there is significant long-range correlation shown in the spatial profile of the correlation $g(r)$ (Fig. 3*C*, *Inset*). Analogous regions of coevolving residues appear in real protein families (53–55, 58–60) as well as in coarse-grained models of protein allostery (35, 36, 43, 44) and allosteric networks (46, 47).

**Point Mutations Are Localized Mechanical Perturbations.** A mutation may vary the strength of no more than $z = 12$ bonds around the mutated amino acid (Fig. 2*B*). The corresponding perturbation of the Hamiltonian $\delta\mathbf{H}$ is, therefore, localized, akin to a defect in a crystal (80, 81). The mechanics of mutations can be further explored by examining the perturbed Green function, $\mathbf{G}' = \mathbf{G} + \delta\mathbf{G}$, which obeys the Dyson equation (77, 82) (*Materials and Methods*):

$$\mathbf{G}' = \mathbf{G} - \mathbf{G}\,\delta\mathbf{H}\,\mathbf{G}'. \qquad [6]$$

The latter can be iterated into an infinite series

$$\delta\mathbf{G} = \mathbf{G}' - \mathbf{G} = -\mathbf{G}\,\delta\mathbf{H}\,\mathbf{G} + \mathbf{G}\,\delta\mathbf{H}\,\mathbf{G}\,\delta\mathbf{H}\,\mathbf{G} - \cdots.$$

This series has a straightforward physical interpretation as a sum over multiple scatterings: 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. In practice, we calculate the mutated Green function using the Woodbury formula [**12**], which exploits the localized nature of the perturbation to accelerate the computation by a factor of $\sim 10^4$ (*Materials and Methods*).

**Epistasis Links Protein Mechanics to Genetic Correlations.** Our model provides a calculable definition of epistasis, the nonlinearity of the fitness effect of interacting mutations (Fig. 2*C*). We take a functional protein obtained from the evolution algorithm and mutate an amino acid at a site $i$. This mutation induces a change in the Green function $\delta\mathbf{G}_i$ (calculated by [**12**]) and hence, in the fitness function $\delta F_i$ [**3**]. One can similarly perform another independent mutation at a site $j$, producing a second deviation, $\delta\mathbf{G}_j$ and $\delta F_j$. Finally, starting again from the original solution, one mutates both $i$ and $j$ simultaneously, with a combined effect $\delta\mathbf{G}_{i,j}$ and $\delta F_{i,j}$. The epistasis $e_{ij}$ measures the departure of the double mutation from additivity of two single mutations:

$$e_{ij} \equiv \delta F_{i,j} - \delta F_i - \delta F_j. \qquad [7]$$

To evaluate the average epistatic interaction among amino acids, we perform the double-mutation calculation for all $10^6$ solutions and take the ensemble average $E_{ij} = \langle e_{ij} \rangle$. Landscapes of $E_{ij}$ show significant epistasis in the channel (Fig. 4). Amino acids outside the high-shear region show only small epistasis, since mutations in the rigid domains hardly change the elastic response. The epistasis landscapes (Fig. 4 *A*–*C*) are mostly positive, since the mutations in the channel interact antagonistically (83): after a strongly deleterious mutation, a second mutation has a smaller effect.

Definition [**7**] is a direct link between epistasis and protein mechanics: the nonlinearity ("curvature") of the Green function measures the deviation of the mechanical response from additivity of the combined effect of isolated mutations at $i$ and $j$, $\Delta\mathbf{G}_{i,j} \equiv \delta\mathbf{G}_{i,j} - \delta\mathbf{G}_i - \delta\mathbf{G}_j$. The epistasis $e_{ij}$ is simply the inner product value of this nonlinearity with the pinch and the response:

$$e_{ij} = \mathbf{v}^\intercal\,\Delta\mathbf{G}_{i,j}\,\mathbf{f}. \qquad [8]$$

Relation [**8**] shows how epistasis originates from mechanical forces among mutated amino acids.

In the gene, epistatic interactions are manifested in codon correlations (56, 57) shown in Fig. 4*D*, which depicts two-codon correlations $Q_{ij}$ from the alignment of $\sim 10^6$ functional genes $\mathbf{c}_*$ [**5**]. We find a tight correspondence between the mean epistasis $E_{ij} = \langle e_{ij} \rangle$ and the codon correlations $Q_{ij}$. Both patterns exhibit strong correlations in the channel region with a period equal to channel's length: 10 amino acids. The similarity in the patterns of $Q_{ij}$ and $E_{ij}$ indicates that a major contribution to the long-range correlations observed among aligned protein sequences stems from the mechanical interactions propagating through the amino acid network.

**Epistasis as a Sum over Scattering Paths.** One can classify epistasis according to the interaction range. Neighboring amino acids exhibit contact epistasis (49–51), because two adjacent perturbations, $\delta\mathbf{H}_i$ and $\delta\mathbf{H}_j$, interact nonlinearly via the AND gate of the interaction table (Fig. 1*C*), $\Delta\mathbf{H}_{i,j} \equiv \delta\mathbf{H}_{i,j} - \delta\mathbf{H}_i - \delta\mathbf{H}_j \neq 0$ (where $\delta\mathbf{H}_{i,j}$ is the perturbation by both mutations). The leading term in the Dyson series [**6**] of $\Delta\mathbf{G}_{i,j}$ is a single scattering from an effective perturbation with an energy $\Delta\mathbf{H}_{i,j}$, which yields the epistasis

$$e_{ij} = -\mathbf{v}^\intercal\,[\mathbf{G}\,\Delta\mathbf{H}_{i,j}\mathbf{G}]\mathbf{f} + \cdots.$$

Long-range epistasis among nonadjacent, noninteracting perturbations ($\Delta\mathbf{H}_{i,j} = 0$) is observed along the channel (Fig. 4). In this case, [**6**] expresses the nonlinearity $\Delta\mathbf{G}_{i,j}$ as a sum over multiple scattering paths, which include both $i$ and $j$ (Fig. 2*C*):

$$e_{ij} = \mathbf{v}^\intercal\,[\mathbf{G}\,\delta\mathbf{H}_i\mathbf{G}\,\delta\mathbf{H}_j\mathbf{G} + \mathbf{G}\,\delta\mathbf{H}_j\mathbf{G}\,\delta\mathbf{H}_i\mathbf{G}]\mathbf{f} - \cdots. \qquad [9]$$

The perturbation expansion directly links long-range epistasis to shear deformation: near the transition, the Green function is dominated by the soft mode, $\mathbf{G} \simeq \mathbf{u}_*\mathbf{u}_*^\intercal/\lambda_*$, with fitness $F$ given by [**4**]. From [**6**] and [**8**], we find a simple expression for the mechanical epistasis as a function of the shear:

$$e_{ij} \simeq F \cdot \left[\frac{h_i}{1+h_i} + \frac{h_j}{1+h_j} - \frac{h_i+h_j}{1+h_i+h_j}\right]. \qquad [10]$$

The factor $h_i \equiv \mathbf{u}_*^\intercal\delta\mathbf{H}_i\mathbf{u}_*/\lambda_*$ in [**10**] 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 it is similar 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 [**10**] is small, $e_{ij} \simeq 2h_ih_jF$. It remains negligible even if one of the mutations, $i$, is in the channel, $h_j \ll 1 \ll h_i$, and $e_{ij} \simeq h_jF$. Epistasis can only be

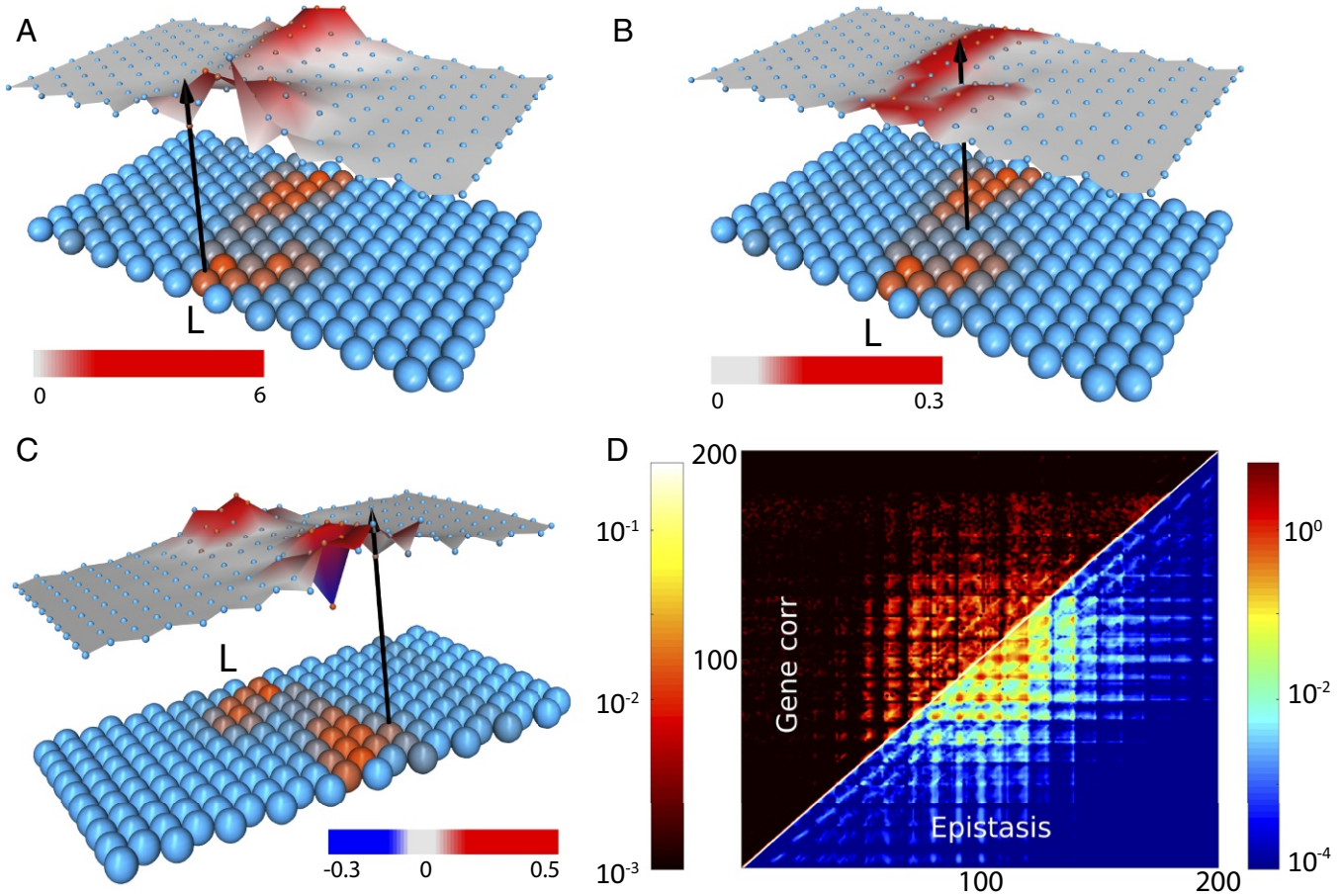


**Fig. 4.** Mechanical epistasis. The epistasis [7], averaged over $\sim 10^6$ solutions $E_{ij} = \langle e_{ij} \rangle$, between a fixed amino acid at position *i* (black arrow) and all other positions *j*. Here, *i* is located at (*A*) the binding site, (*B*) the center of the channel, and (*C*) slightly off the channel. Underneath, the average amino acid 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. Although epistasis is predominantly positive, negative values also occur, mostly at the boundary of the channel (*C*). Landscapes are plotted for specific output site at right. L, left. (*D*) The two-codon correlation function $Q_{ij}$ [5] measures the coupling between mutations at positions *i* and *j* [5]. 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).

long ranged along the channel when both mutations are significant, $h_i, h_j \gg 1$, and $e_{ij} \simeq F\left[1 - h_i^{-1} - h_j^{-1} + (h_i + h_j)^{-1}\right] \simeq F[1 - 1/\min(h_i, h_j)]$. We conclude that epistasis is maximal when both sites are at the start or end of the channel as illustrated in Fig. 4. The nonlinearity of the fitness function gives rise to antagonistic epistasis.

**Geometry of Fitness Landscape and Gene-to-Function Map.** With our mechanical evolution algorithm, we can swiftly explore the fitness landscape to examine its geometry. The genotype space is a 200D hypercube with vertices that are all possible genes $\mathbf{c}$. The phenotypes reside in a 400D space of all possible mechanical responses $\mathbf{u}$. The Green function provides the genotype-to-phenotype map [1]. A functional protein is encoded by a gene $\mathbf{c}_*$ with fitness that exceeds a large threshold, $F(\mathbf{c}_*) \geq F_{\rm m} \simeq 5$, and the functional phenotype is dominated by the emergent zero-energy mode, $\mathbf{u}(\mathbf{c}_*) \simeq \mathbf{u}_*$ (Fig. 3*A*). We also characterize the phenotype by the shear field $\mathbf{s}_*$ (*Materials and Methods*).

The singular value decomposition (SVD) of the $10^6$ solutions returns a set of eigenvectors with ordered eigenvalues that show their significance in capturing the data (*Materials and Methods*). The SVD spectra reveal strong correspondence between the genotype $\mathbf{c}_*$ and the phenotype, $\mathbf{u}$ and $\mathbf{s}_*$ (Fig. 5). In all three datasets, the largest eigenvalues are discrete and stand out from the bulk continuous spectrum. These are the collective dfs, which show loci in the gene and positions in the "flow" (i.e., displacement) and shear fields that tend to vary together.

We examine the correspondence among three sets of eigenvectors: $\{\mathbf{U}_k\}$ of the flow, $\{\mathbf{C}_k\}$ of the gene, and $\{\mathbf{S}_k\}$ of the shear. The first eigenvector of the flow, $\mathbf{U}_1$, is the hinge motion caused by the pinch, with two eddies rotating in opposite directions (Fig. 5*A*). The next two modes, shear ($\mathbf{U}_2$) and breathing ($\mathbf{U}_3$), also occur in real proteins, such as glucokinase (Fig. 1*B*). The first eigenvectors of the shear $\mathbf{S}_1$ and of the gene sequence $\mathbf{C}_1$ show that the high-shear region is mirrored as a P-rich region, where a mechanical signal may cause local rearrangement of the amino acids by deforming weak bonds. In the rest of the protein, the H-rich regions move as rigid bodies with insignificant shear. The higher gene eigenvectors, $\mathbf{C}_k$ ($k > 1$), capture patterns of correlated genetic variations. The striking similarity between the sequence correlation patterns $\mathbf{C}_k$ and the shear eigenvectors $\mathbf{S}_k$ shows a tight genotype-to-phenotype map, as is further shown in the likeness of the correlation matrices of the amino acid and shear flow (Fig. 5*C*).

In the phenotype space, we represent the displacement field $\mathbf{u}$ in the SVD basis, $\{\mathbf{U}_k\}$ (Fig. 5*B*). Since $\sim 90\%$ of the data are explained by the first $\sim 15$ $\mathbf{U}_k$, we can compress the displacement field without much loss into the first 15 coordinates. This implies

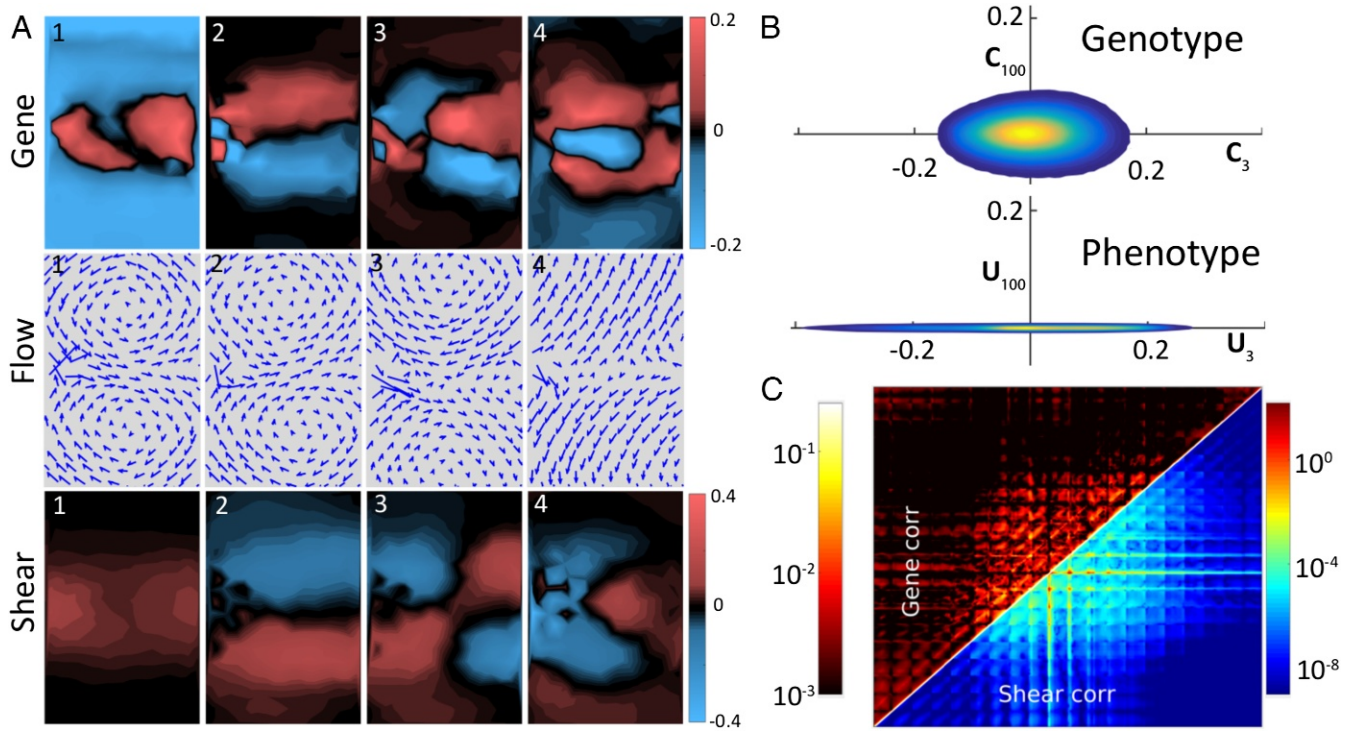


**Fig. 5.** From gene to mechanical function: spectra and dimensions. (*A*) The first four SVD eigenvectors (in the text) of the gene $\mathbf{C}_k$ (*Top*), the displacement flow field $\mathbf{U}_k$ (*Middle*), and the shear $\mathbf{S}_k$ (*Bottom*). (*B*) Cross-sections through the set of solutions in the genotype space (*Upper*) and the phenotype space (*Lower*). Density of solutions is color coded. The genotype cross-section is the plane defined by the eigenvectors $\mathbf{C}_3 - \mathbf{C}_{100}$, and in the phenotype space, it is defined by the eigenvectors $\mathbf{U}_3 - \mathbf{U}_{100}$ (in the text). The dimensional reduction is manifested by the discoid geometry of the phenotype cloud compared with the spheroid shape of the genotype cloud. (*C*) Genetic correlations $Q_{ij}$ show similarity to correlations in the shear field, $\mathbf{s}_*$ (color coded in log scale). Corr, correlation.

that the set of solutions is a 15D discoid, which is flat in most directions. In contrast, representation of the genes $\mathbf{c}_*$ in the SVD frame of reference (with the $\{\mathbf{C}_k\}$ basis) reveals that, in genotype space, the solution set is an incompressible 200D spheroid (Fig. 5*B*). The dramatic dimensional reduction in mapping genotypes to phenotypes stems from the different constraints that shape them (36, 84–89).

## Discussion

Theories of protein need to combine the many-body physics of the amino acid matter with the evolution of genetic information, which together, give rise to protein function. We introduced a simplified theory of protein, where the mapping between genotype and phenotype takes an explicit form in terms of mechanical propagators (Green functions), which can be efficiently calculated. As a functional phenotype, we take cooperative motion and force transmission through the protein [**2**]. This allows us to map genetic mutations to mechanical perturbations, which scatter the force field and deflect its propagation [**3** and **6**] (Fig. 2). The evolutionary process amounts to solving the inverse scattering problem: given prescribed functional modes, one looks for network configurations that yield this low end of the dynamical spectrum. Epistasis, the interaction among loci in the gene, corresponds to a sum over all multiple scattering trajectories or equivalently, the nonlinearity of the Green function [**7** and **8**]. We find that long-range epistasis signals the emergence of a collective functional mode in the protein. The results of this theory (in particular, the expressions for epistasis) follow from the basic geometry of the amino acid network and the localized mutations and are, therefore, applicable to general tasks and fitness functions with multiple inputs and responses.

## Materials and Methods

**The Mechanical Model of Protein.** We model the protein as an elastic network made of harmonic springs (23, 24, 39, 90). The connectivity of the network is described by a hexagonal lattice with vertices that are amino acids and edges that correspond to bonds. There are $n_a = 10 \times 20 = 200$ amino acids indexed by Roman letters and $n_b$ bonds indexed by Greek letters. We use the HP model (73) with two amino acid species, hydrophobic ($a_i = \mathrm{H}$) and polar ($a_i = \mathrm{P}$). The amino acid chain is encoded in a gene $\mathbf{c}$, where $c_i = 1$ encodes H and $c_i = 0$ encodes P, $i = 1, \ldots, n_a$. The degree $z_i$ of each amino acid is the number of amino acids to which it is connected by bonds. In our model, most amino acids have the maximal degree, which is $z = 12$, while amino acids at the boundary have fewer neighbors, $z < 12$ (Fig. 1*C*). The connectivity of the graph is recorded by the adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij} = 1$ if there is a bond from $j$ to $i$ and $\mathbf{A}_{ij} = 0$ otherwise. The gradient operator $\nabla$ relates the spaces of bonds and vertices (and is, therefore, of size $n_b \times n_a$): if vertices $i$ and $j$ are connected by a bond $\alpha$, then $\nabla_{\alpha i} = +1$ and $\nabla_{\alpha j} = -1$. As in the continuum case, the Laplace operator $\Delta$ is the product $\Delta = \nabla^{\mathsf{T}}\nabla$. The nondiagonal elements $\Delta_{ij}$ are $-1$ if $i$ and $j$ are connected and 0 otherwise. The diagonal part of $\Delta$ is the degree $\Delta_{ii} = z_i$. Hence, we can write the Laplacian as $\Delta = \mathbf{Z} - \mathbf{A}$, where $\mathbf{Z}$ is the diagonal matrix of the degrees $z_i$.

We embed the graph in Euclidean space $\mathbb{E}^d$ ($d = 2$) by assigning positions $\mathbf{r}_i \in \mathbb{E}^d$ to each amino acid. We concatenate all positions in a vector $\mathbf{r}$ of length $n_a \cdot d \equiv n_d$. Finally, to each bond, we assign a spring with constant $k_\alpha$, which we keep in a diagonal $n_b \times n_b$ matrix $\mathbf{K}$. The strength of the spring is determined by the AND rule of the HP model's interaction table (Fig. 1*C*), $k_\alpha = k_w + (k_s - k_w)c_i c_j$, where $c_i$ and $c_j$ are the codons of the amino acid connected by bond $\alpha$. This implies that a strong $\mathrm{H}-\mathrm{H}$ bond has $k_\alpha = k_s$, whereas the weak bonds $\mathrm{H}-\mathrm{P}$ and $\mathrm{P}-\mathrm{H}$ have $k_\alpha = k_w$. We usually take $k_s = 1$ and $k_w = 0.01$. This determines a spring network. We also assume that the initial configuration is such that all springs are at their equilibrium length, disregarding the possibility of "internal stresses" (39), so that the initial elastic energy is $\mathcal{E}_0 = 0$.

We define the "embedded" gradient operator $\mathbf{D}$ (of size $n_b \times n_d$), which is obtained by taking the graph gradient $\nabla$ and multiplying each nonzero element ($\pm 1$) by the corresponding direction vector $\mathbf{n}_{ij} = (\mathbf{r}_i - \mathbf{r}_j)/|\mathbf{r}_i - \mathbf{r}_j|$.

Thus, $\mathbf{D}$ is a tensor ($\mathbf{D} = \nabla_{\alpha i}\mathbf{n}_{ij}$), which we store as a matrix ($\alpha$ is the bond connecting vertices $i$ and $j$). In each row vector of $\mathbf{D}$, which we denote as $\mathbf{m}_\alpha \equiv \mathbf{D}_{\alpha,:}$, there are only $2d$ nonzero elements. To calculate the elastic response of the network, we deform it by applying a force field $\mathbf{f}$, which leads to the displacement of each vertex by $\mathbf{u}_i$ to a new position $\mathbf{r}_i + \mathbf{u}_i$ (39). For small displacements, the linear response of the network is given by Hooke's law, $\mathbf{f} = \mathbf{H}\mathbf{u}$. The elastic energy is $\mathcal{E} = \mathbf{u}^\mathsf{T}\mathbf{H}\mathbf{u}/2$, and the Hamiltonian, $\mathbf{H} = \mathbf{D}^\mathsf{T}\mathbf{K}\mathbf{D}$, is the Hessian of the elastic energy $\mathcal{E}$, $\mathbf{H}_{ij} = \delta^2\mathcal{E}/(\delta\mathbf{u}_i\delta\mathbf{u}_j)$. By rescaling, $\mathbf{D} \to \mathbf{K}^{1/2}\mathbf{D}$, which amounts to scaling all distances by $1/\sqrt{k_\alpha}$, we obtain $\mathbf{H} = \mathbf{D}^\mathsf{T}\mathbf{D}$. It follows that the Hamiltonian is a function of the gene $\mathbf{H}(\mathbf{c})$, which has the structure of the Laplacian $\Delta$ multiplied by the tensor product of the direction vectors. Each $d \times d$ block $\mathbf{H}_{ij}$ ($i \neq j$) is a function of the codons $c_i$ and $c_j$:

$$\begin{aligned}\mathbf{H}_{ij}(c_i, c_j) &= \Delta_{ij}\mathbf{n}_{ij}\mathbf{n}_{ij}^\mathsf{T} \\ &= -\mathbf{A}_{ij}\left[k_w + (k_s - k_w)c_i c_j\right]\mathbf{n}_{ij}\mathbf{n}_{ij}^\mathsf{T}.\end{aligned} \qquad [11]$$

The diagonal blocks complete the row and column sums to zero, $\mathbf{H}_{ii} = -\sum_{j\neq i}\mathbf{H}_{ij}$.

**The Inverse Problem: Green Function and Its Spectrum.** The Green function $\mathbf{G}$ is defined by the inverse relation to Hooke's law, $\mathbf{u} = \mathbf{G}\mathbf{f}$ [1]. If $\mathbf{H}$ were invertible (nonsingular),$\mathbf{G}$ would have been just $\mathbf{G} = \mathbf{H}^{-1}$. However, $\mathbf{H}$ is always singular owing to the zero-energy (Galilean) modes of translation and rotation. Therefore, one needs to define $\mathbf{G}$ as the Moore–Penrose pseudoinverse (78, 79), $\mathbf{G} = \mathbf{H}^+$, on the complement of the space of Galilean transformations. The pseudoinverse can be understood in terms of the spectrum of $\mathbf{H}$. There are at least $n_0 = d(d+1)/2$ zero modes: $d$ translation modes and $d(d-1)/2$ rotation modes. These modes are irrelevant and will be projected out of the calculation (note that these modes do not come from missing connectivity of the graph $\Delta$ itself but from its embedding in $\mathbb{E}^d$). $\mathbf{H}$ is singular but is still diagonalizable (since it has a basis of dimension $n_d$), and it can be written as the spectral decomposition, $\mathbf{H} = \sum_{k=1}^{n_d}\lambda_k\mathbf{u}_k\mathbf{u}_k^\mathsf{T}$, where $\{\lambda_k\}$ is the set of eigenvalues and $\{\mathbf{u}_k\}$ are the corresponding eigenvectors (note that $k$ denotes the index of the eigenvalue, while $i$ and $j$ denote amino acid positions). For a nonsingular matrix, one may calculate the inverse simply as $\mathbf{H}^{-1} = \sum_{k=1}^{n_d}\lambda_k^{-1}\mathbf{u}_k\mathbf{u}_k^\mathsf{T}$. Since $\mathbf{H}$ is singular, we leave out the zero modes and get the pseudoinverse $\mathbf{H}^+$, $\mathbf{G} = \mathbf{H}^+ = \sum_{k=n_0+1}^{n_d}\lambda_k^{-1}\mathbf{u}_k\mathbf{u}_k^\mathsf{T}$. It is easy to verify that, if $\mathbf{u}$ is orthogonal to the zero modes, then $\mathbf{u} = \mathbf{G}\mathbf{H}\mathbf{u}$. The pseudoinverse obeys the four requirements (78): (*i*) $\mathbf{H}\mathbf{G}\mathbf{H} = \mathbf{H}$, (*ii*) $\mathbf{G}\mathbf{H}\mathbf{G} = \mathbf{G}$, (*iii*) $(\mathbf{H}\mathbf{G})^\mathsf{T} = \mathbf{H}\mathbf{G}$, and (*iv*) $(\mathbf{G}\mathbf{H})^\mathsf{T} = \mathbf{G}\mathbf{H}$. In practice, as the projection commutes with the mutations, the pseudoinverse has most virtues of a proper inverse. The reader might prefer to link $\mathbf{G}$ and $\mathbf{H}$ through the heat kernel, $\mathcal{K}(t) = \sum_k e^{\lambda_k t}\mathbf{u}_k\mathbf{u}_k^\mathsf{T}$. Then, $\mathbf{G} = \int_0^\infty dt\ \mathcal{K}(t)$ and $\mathbf{H} = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}|_{t=0}$.

**Pinching the Network.** A pinch is given as a localized force applied at the boundary of the "protein." We usually apply the force on a pair of neighboring boundary vertices, $p'$ and $q'$. It seems reasonable to apply a force dipole (i.e., two opposing forces $\mathbf{f}_{q'} = -\mathbf{f}_{p'}$), since a net force will move the center of mass. This pinch is, therefore, specified by the force vector $\mathbf{f}$ (of size $n_d$), with the only $2d$ nonzero entries being $\mathbf{f}_{q'} = -\mathbf{f}_{p'}$. Hence, it has the same structure as a bond vector $\mathbf{m}_\alpha$ of a "pseudobond" connecting $p'$ and $q'$ A normal pinch $\mathbf{f}$ has a force dipole directed along the $\mathbf{r}_{p'} - \mathbf{r}_{q'}$ line (the $\mathbf{n}_{p'q'}$ direction). Such a pinch is expected to induce a hinge motion. A shear pinch will be in a perpendicular direction $\perp \mathbf{n}_{p'q'}$ and is expected to induce a shear motion.

Evolution tunes the spring network to exhibit a low-energy mode, in which the protein is divided into two subdomains moving like rigid bodies. This large-scale mode can be detected by examining the relative motion of two neighboring vertices, $p$ and $q$, at another location at the boundary (usually at the opposite side). Such a desired response at the other side of the protein is specified by a response vector $\mathbf{v}$, and the only nonzero entries correspond to the directions of the response at $p$ and $q$. Again, we usually consider a "dipole" response $\mathbf{v}_q = -\mathbf{v}_p$.

**Evolution and Mutation.** The quality of the response (i.e., the biological fitness) is specified by how well the response follows the prescribed one $\mathbf{v}$. In the context of our model, we chose the (scalar) observable $F$ as $F = \mathbf{v}^\mathsf{T}\mathbf{u} = \mathbf{v}_p\cdot\mathbf{u}_p + \mathbf{v}_q\cdot\mathbf{u}_q = \mathbf{v}^\mathsf{T}\mathbf{G}\mathbf{f}$ [2]. In an evolution simulation, one would exchange amino acids between H and P, while demanding that the fitness change $\delta F$ is positive or nonnegative. By this, we mean $\delta F > 0$ is thanks to a beneficial mutation, whereas $\delta F = 0$ corresponds to a neutral one. Deleterious mutations $\delta F < 0$ are generally rejected. A version that accepts mildly deleterious mutations (a finite temperature Metropolis algorithm) gave similar results. We may 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 $F$, $\mathbf{v}_p\cdot\mathbf{u}_p$ and $\mathbf{v}_q\cdot\mathbf{u}_q$, increases separately. The evolution is stopped when $F \geq F_\mathrm{m} \sim 5$, which signals the formation of a shear band. When simulations ensue beyond $F_\mathrm{m} \sim 5$, the band slightly widens, and the fitness slows down and converges at a maximal value, typically $F_\mathrm{max} \sim 8$.

**Evolving the Green Function Using the Dyson and Woodbury Formulas.** The Dyson formula follows from the identity $\delta\mathbf{H} \equiv \mathbf{H}' - \mathbf{H} = \mathbf{G}'^{+} - \mathbf{G}^{+}$, which is multiplied by $\mathbf{G}$ on the left and $\mathbf{G}'$ on the right to yield [6]. The formula remains valid for the pseudoinverses in the nonsingular subspace. One can calculate the change in fitness by evaluating the effect of a mutation on the Green function, $\mathbf{G}' = \mathbf{G} + \delta\mathbf{G}$, and then examining the change, $\delta F = \mathbf{v}^\mathsf{T}\delta\mathbf{G}\mathbf{f}$ [3]. Using [6] to calculate the mutated Green function $\mathbf{G}'$ is an impractical method, as it amounts to inverting at each step a large $n_d \times n_d$ matrix. However, the mutation of an amino acid at $i$ has a localized effect. It may change only up to $z = 12$ bonds among the bonds $\alpha(i)$ with the neighboring amino acids. Thanks to the localized nature of the mutation, the corresponding defect Hamiltonian $\delta\mathbf{H}_i$ is, therefore, of a small rank, $r \leq z = 12$, equal to the number of switched bonds (the average $r$ is about 9.3). $\delta\mathbf{H}_i$ can be decomposed into a product $\delta\mathbf{H}_i = \mathbf{M}\mathbf{B}\mathbf{M}^\mathsf{T}$. The diagonal $r \times r$ matrix $\mathbf{B}$ records whether a bond $\alpha(i)$ is switched from weak to strong ($\mathbf{B}_{\alpha\alpha} = k_s - k_w = +0.99$) or vice versa ($\mathbf{B}_{\alpha\alpha} = -0.99$), and $\mathbf{M}$ is a $n_d \times r$ matrix with $r$ columns that are the bond vectors $\mathbf{m}_\alpha$ for the switched bonds $\alpha(i)$. This allows one to calculate changes in the Green function more efficiently using the Woodbury formula (91, 92):

$$\delta\mathbf{G} = -\mathbf{G}\mathbf{M}\left(\mathbf{B}^{-1} + \mathbf{M}^\mathsf{T}\mathbf{G}\mathbf{M}\right)^{+}\mathbf{M}^\mathsf{T}\mathbf{G}. \qquad [12]$$

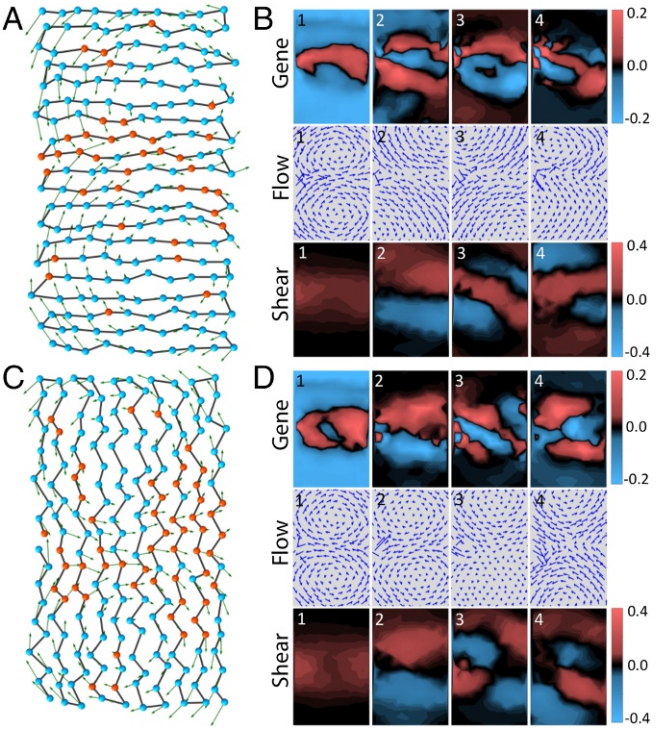


**Fig. 6.** The effect of the backbone on evolution of mechanical function. The backbone induces long-range mechanical correlations, which influence protein evolution. We examine two configurations: parallel (*A* and *B*) and perpendicular (*C* and *D*) to the channel. (*A* and *B*) Parallel. (*A*) The backbone directs the formation of a narrow channel along the fold (compared with Fig. 5*A*). (*B*) The first four SVD eigenvectors of the gene $\mathbf{C}_k$ (*Top*), the flow $\mathbf{U}_k$ (*Middle*), and the shear $\mathbf{S}_k$ (*Bottom*). (*C* and *D*) Perpendicular. (*C*) The formation of the channel is "dispersed" by the backbone. (*D*) The first four SVD eigenvectors of $\mathbf{C}_k$ (*Top*), $\mathbf{U}_k$ (*Middle*), and shear $\mathbf{S}_k$ (*Bottom*).

The two expressions for the mutation impact $\delta\mathbf{G}$, [**6** and **12**], are equivalent, and one may get the scattering series of [**6**] by a series expansion of the pseudoinverse in [**12**]. The practical advantage of [**12**] is that the only (pseudo-)inversion that it requires is of a low-rank tensor (the term in parentheses). This accelerates our simulations by a factor of $(n_a/r)^3 \simeq 10^4$.

**Pathologies and Broken Networks.** A network broken into disjoint components exhibits floppy modes owing to the low energies of the relative motion of the components with respect to each other. The evolutionary search might end up in such nonfunctional unintended modes. The common pathologies that we observed are (*i*) isolated nodes at the boundary that become weakly connected via H → P mutations, (*ii*) "sideways" channels that terminate outside the target region (which typically includes around 8–10 sites), and (*iii*) channels that start and end at the target region without connecting to the binding site. All of these are some easy to understand floppy modes, which can vibrate independent of the location of the pinch and cause the response to diverge ($> F_{\mathrm{m}}$) without producing a functional mode. We avoid such pathologies by applying the pinch force to the protein network symmetrically: pinch the binding site on face left, look at responses on face right, 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 responses and enables hinge motion of the protein. This prevents the vibrations from being localized at isolated sites or unwanted channels.

**Dimension and SVD.** To examine the geometry of the fitness landscape and the genotype-to-phenotype map, we looked at the correlation among numerous solutions, typically $N_{\mathrm{sol}} \sim 10^6$. Each solution is characterized by three vectors: (*i*) the gene of the functional protein, $\mathbf{c}_*$ (a vector of length $n_a = 200$ codons); (*ii*) the flow field (displacement), $\mathbf{u}(\mathbf{c}_*) = \mathbf{G}(\mathbf{c}_*)\mathbf{f}$ (a vector of length $n_d = 400$ of $x$ and $y$ velocity components); and (*iii*) the shear field $\mathbf{s}_*$ (a vector of length $n_a = 200$). We compute the shear as the symmetrized derivative of the displacement field using the method in ref. 21. The value of the $\mathbf{s}_*$ field is the sum of squares of the traceless part of the strain tensor (Frobenius norm). These three types of vectors are stored along the rows of three matrices $W_C$, $W_U$, and $W_S$. We calculate the eigenvectors of these matrices, $\mathbf{C}_k$, $\mathbf{U}_k$, and $\mathbf{S}_k$, via SVD (as in ref. 36). The corresponding SVD eigenvalues are the square roots of the eigenvalues of the covariance matrix $W^{\mathsf{T}}W$, while the eigenvectors are the same. In typical spectra, most eigenvalues reside in a continuum bulk region that resembles the spectra of random matrices. A few larger outliers, typically around a dozen or so, carry the nonrandom correlation information.

**The Protein Backbone.** A question may arise as to what extent the protein's backbone might affect the results described so far. Proteins are polypeptides, linear heteropolymers of amino acids, linked by covalent peptide bonds, which form the protein backbone. The peptide bonds are much stronger than the noncovalent interactions among the amino acids and do not change when the protein mutates. We, therefore, augmented our model with a "backbone": a linear path of conserved strong bonds that passes once through all amino acids. We focused on two extreme cases: a serpentine backbone either parallel to the shear band or perpendicular to it (Fig. 6).

The presence of the backbone does not interfere with the emergence of a low-energy mode of the protein with a flow pattern (i.e., displacement field) that is similar to the backboneless 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. 6*A*). The resulting channel is narrower than in the model without backbone (Figs. 1*D* and 5). In the perpendicular configuration, the evolutionary progression of the channel is much less oriented (Fig. 6*C*). While the flow patterns are similar, closer inspection shows noticeable differences, as can be seen in the flow eigenvectors $\mathbf{U}_k$ (Fig. 6 *B* and *D*). The shear eigenvectors $\mathbf{S}_k$ represent the derivative of the flow and therefore, highlight more distinctly these differences.

As for the correspondence between gene eigenvectors $\mathbf{C}_k$ and shear eigenvectors $\mathbf{S}_k$, the backbone affects the shape of the channel in concert with the sequence correlations around it. Transmission of mechanical signals seems to be easier along the orientation of the fold (parallel configuration) (Fig. 6*A*). Transmission across the fold (perpendicular configuration) necessitates significant deformation of the backbone and leads to "dispersion" of the signal at the output (Fig. 6*C*). We propose that the shear band will be roughly oriented with the direction of the fold, but this requires further analysis of structural data. Overall, we conclude from our examination that the backbone adds certain features to patterns of the field and sequence correlation without changing the basic results of our model. The presence of the backbone might constrain the evolutionary search, but this has no significant effect on the fast convergence of the search and on the long-range correlations among solutions.

**ACKNOWLEDGMENTS.** We thank Jacques Rougemont for calculations of shear in glucokinase (Fig. 1*B*) and for helpful discussions. We thank Stanislas Leibler, Michael R. Mitchell, Elisha Moses, Giovanni Zocchi, and Olivier Rivoire for helpful discussions and encouragement. We also thank Alex Petroff, Steve Granick, Le Yan, and Matthieu Wyart for valuable comments on the manuscript. J.-P.E. is supported by European Research Council Advanced Grant Bridges, and T.T. is supported by Institute for Basic Science Grant IBS-R020 and the Simons Center for Systems Biology of the Institute for Advanced Study, Princeton.

1. Daniel RM, Dunn RV, Finney JL, Smith JC (2003) The role of dynamics in enzyme activity. *Annu Rev Biophys Biomol Struct* 32:69–92.
2. Bustamante C, Chemla YR, Forde NR, Izhaky D (2004) Mechanical processes in biochemistry. *Annu Rev Biochem* 73:705–748.
3. Hammes-Schiffer S, Benkovic SJ (2006) Relating protein motion to catalysis. *Annu Rev Biochem* 75:519–541.
4. Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313:1638–1642.
5. Bahar I, Lezon TR, Yang LW, Eyal E (2010) Global dynamics of proteins: Bridging between structure and function. *Annu Rev Biophys* 39:23–42.
6. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9:646–652, and erratum (2002) 9:788.
7. Henzler-Wildman KA, et al. (2007) Intrinsic motions along an enzymatic reaction trajectory. *Nature* 450:838–844.
8. Huse M, Kuriyan J (2002) The conformational plasticity of protein kinases. *Cell* 109:275–282.
9. Eisenmesser EZ, et al. (2005) Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438:117–121.
10. Savir Y, Tlusty T (2007) Conformational proofreading: The impact of conformational changes on the specificity of molecular recognition. *PLoS One* 2:e468.
11. Goodey NM, Benkovic SJ (2008) Allosteric regulation and catalysis emerge via a common route. *Nat Chem Biol* 4:474–482.
12. Savir Y, Tlusty T (2010) Reca-mediated homology search as a nearly optimal signal detection system. *Mol Cell* 40:388–396.
13. Grant BJ, Gorfe AA, McCammon JA (2010) Large conformational changes in proteins: Signaling and other functions. *Curr Opin Struct Biol* 20:142–147.
14. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: A plausible model. *J Mol Biol* 12:88–118.
15. Perutz MF (1970) Stereochemistry of cooperative effects in haemoglobin: Haem-haem interaction and the problem of allostery. *Nature* 228:726–734.
16. Cui Q, Karplus M (2008) Allostery and cooperativity revisited. *Protein Sci* 17:1295–1307.
17. Daily MD, Upadhyaya TJ, Gray JJ (2008) Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins* 71:455–466.
18. Motlagh HN, Wrabl JO, Li J, Hilser VJ (2014) The ensemble nature of allostery. *Nature* 508:331–339.
19. Koshland D (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA* 44:98–104.
20. Gerstein M, Lesk AM, Chothia C (1994) Structural mechanisms for domain movements in proteins. *Biochemistry (Mosc)* 33:6739–6749.
21. Mitchell MR, Tlusty T, Leibler S (2016) Strain analysis of protein structures and low dimensionality of mechanical allosteric couplings. *Proc Natl Acad Sci USA* 113:E5847–E5855.
22. Mitchell MR, Leibler S (2018) Elastic strain and twist analysis of protein structural data and allostery of the transmembrane channel kcsa. *Phys Biol* 5:036004.
23. Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 77:1905–1908.
24. Chennubhotla C, Rader AJ, Yang LW, Bahar I (2005) Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies. *Phys Biol* 2:S173–S180.
25. Bahar I (2010) On the functional significance of soft modes predicted by coarse-grained models for membrane proteins. *J Gen Physiol* 135:563–573.
26. Lòpez-Blanco JR, Chacòn P (2016) New generation of elastic network models. *Curr Opin Struct Biol* 37:46–53.
27. Levitt M, Sander C, Stern PS (1985) Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol* 181:423–447.
28. Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng Des Sel* 14:1–6.
29. Bahar I, Rader AJ (2005) Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Biol* 15:586–592.
30. Haliloglu T, Bahar I (2015) Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr Opin Struct Biol* 35:17–23.
31. Ming D, Wall ME (2005) Allostery in a coarse-grained model of protein dynamics. *Phys Rev Lett* 95:198103.

32. Zheng WJ, Brooks BR, Thirumalai D (2006) Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc Natl Acad Sci USA* 103:7664–7669.
33. Arora K, Brooks CL (2007) Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc Natl Acad Sci USA* 104:18496–18501.
34. Miyashita O, Onuchic JN, Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci USA* 100:12570–12575.
35. Tlusty T (2016) Self-referring DNA and protein: A remark on physical and geometrical aspects. *Philos Trans A Math Phys Eng Sci* 374:20150070.
36. Tlusty T, Libchaber A, Eckmann JP (2017) Physical model of the genotype-to-phenotype map of proteins. *Phys Rev X* 7:021037.
37. Qu H, Zocchi G (2013) How enzymes work: A look through the perspective of molecular viscoelastic properties. *Phys Rev X* 3:011009.
38. Joseph C, Tseng CY, Zocchi G, Tlusty T (2014) Asymmetric effect of mechanical stress on the forward and reverse reaction catalyzed by an enzyme. *PLoS One* 9:e101442.
39. Alexander S (1998) Amorphous solids: Their structure, lattice dynamics and elasticity. *Phys Rep* 296:65–236.
40. Alexander S, Orbach R (1982) Density of states on fractals: "Fractons." *J Phys Lett* 43:625–631.
41. Phillips JC, Thorpe MF (1985) Constraint theory, vector percolation and glass-formation. *Solid State Commun* 53:699–702.
42. Thorpe MF, Lei M, Rader AJ, Jacobs DJ, Kuhn LA (2001) Protein flexibility and dynamics using constraint theory. *J Mol Graphics Model* 19:60–69.
43. Hemery M, Rivoire O (2015) Evolution of sparsity and modularity in a model of protein allostery. *Phys Rev E* 91:042704.
44. Flechsig H (2017) Design of elastic networks with evolutionary optimized long-range communication as mechanical models of allosteric proteins. *Biophys J* 113:558–571.
45. Thirumalai D, Hyeon C (2017) Signaling networks and dynamics of allosteric transitions in bacterial chaperonin GroEL: Implications for iterative annealing of misfolded proteins. arXiv:171007981.
46. Rocks JW, et al. (2017) Designing allostery-inspired response in mechanical networks. *Proc Natl Acad Sci USA* 114:2520–2525.
47. Yan L, Ravasio R, Brito C, Wyart M (2017) Architecture and coevolution of allosteric materials. *Proc Natl Acad Sci USA* 114:2526–2531.
48. Yan L, Ravasio R, Brito C, Wyart M (2017) Principles for optimal cooperativity in allosteric materials. arXiv:170801820.
49. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18:309–317.
50. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6:e28766.
51. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30:1072–1080.
52. Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) Psicov: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28:184–190.
53. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
54. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10:59–69.
55. Reynolds K, McLaughlin R, Ranganathan R (2011) Hot spots for allosteric regulation on protein surfaces. *Cell* 147:1564–1575.
56. Hopf TA, et al. (2017) Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 35:128–135.
57. Poelwijk FJ, Socolich M, Ranganathan R (2017) Learning the pattern of epistasis linking genotype and phenotype in a protein. bioRxiv:10.1101/213835.
58. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 138:774–786.
59. Rivoire O, Reynolds KA, Ranganathan R (2016) Evolution-based functional decomposition of proteins. *PLoS Comput Biol* 12:e1004817.
60. Teşileanu T, Colwell LJ, Leibler S (2015) Protein sectors: Statistical coupling analysis versus conservation. *PLoS Comput Biol* 11:e1004091.
61. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. *Nat Rev Genet* 14:249–261.
62. Cordell HJ (2002) Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11:2463–2468.
63. Phillips PC (2008) Epistasis–The essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9:855–867.
64. Mackay TFC (2014) Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat Rev Genet* 15:22–33.
65. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW (2007) Crystal structure of an ancient protein: Evolution by conformational epistasis. *Science* 317:1544–1548.
66. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. *Nature* 490:535–538.
67. Gong LI, Suchard MA, Bloom JD (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2:e00631.
68. Miton CM, Tokuriki N (2016) How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci* 25:1260–1272.
69. Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet* 4:579–593.
70. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. *Nature* 420:218–223.
71. Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. *Nature* 465:922–926.
72. Liberles DA, et al. (2012) The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci* 21:769–785.
73. Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry (Mosc)* 24:1501–1509.
74. Lau KF, Dill KA (1989) A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986–3997.
75. Shakhnovich E, Farztdinov G, Gutin AM, Karplus M (1991) Protein folding bottlenecks: A lattice Monte Carlo simulation. *Phys Rev Lett* 67:1665–1668.
76. Green G (1828) *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism* (T. Wheelhouse, Nottingham, England).
77. Abrikosov A, Gorkov L, Dzyaloshinski I (1963) *Methods of Quantum Field Theory in Statistical Physics* (Prentice Hall, Englewood Cliffs, NJ).
78. Penrose R (1955) A generalized inverse for matrices. *Math Proc Cambridge Philos Soc* 51:406–413.
79. Ben-Israel A, Greville TN (2003) *Generalized Inverses: Theory and Applications* (Springer Science & Business Media, Springer-Verlag, New York), Vol 15.
80. Tewary VK (1973) Green-function method for lattice statics. *Adv Phys* 22:757–810.
81. Elliott RJ, Krumhansl JA, Leath PL (1974) The theory and properties of randomly disordered crystals and related physical systems. *Rev Mod Phys* 46:465–543.
82. Dyson FJ (1949) The *s* matrix in quantum electrodynamics. *Phys Rev* 75:1736–1755.
83. Desai MM, Weissman D, Feldman MW (2007) Evolution can favor antagonistic epistasis. *Genetics* 177:1001–1010.
84. Savir Y, Noor E, Milo R, Tlusty T (2010) Cross-species analysis traces adaptation of rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci USA* 107:3475–3480.
85. Savir Y, Tlusty T (2013) The ribosome as an optimal decoder: A lesson in molecular recognition. *Cell* 153:471–479.
86. Kaneko K, Furusawa C, Yomo T (2015) Universal relationship in gene-expression changes for cells in steady-growth state. *Phys Rev X* 5:011014.
87. Friedlander T, Mayo AE, Tlusty T, Alon U (2015) Evolution of bow-tie architectures in biology. *PLoS Comput Biol* 11:e1004055.
88. Furusawa C, Kaneko K (2017) Formation of dominant mode by evolution in biological systems. arXiv:170401751.
89. Tlusty T (2010) A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes. *Phys Life Rev* 7:362–376.
90. Born M, Huang K (1954) *Dynamical Theory of Crystal Lattices*, The International Series of Monographs on Physics (Clarendon, Oxford).
91. Woodbury MA (1950) *Inverting Modified Matrices* (Princeton Univ, Princeton), Statistical Research Group Memo Report 42, p 4.
92. Deng CY (2011) A generalization of the Sherman–Morrison–Woodbury formula. *Appl Math Lett* 24:1561–1564.