Biophysical Journal
Article

Biophysical Society

Protein dynamics provide mechanistic insights about epistasis among common missense polymorphisms

Nicholas J. Ose,[1] Paul Campitelli,[1] Ravi Patel,[2,3] Sudhir Kumar,[2,3,4,*] and S. Banu Ozkan[1,*]
[1]Department of Physics and Center for Biological Physics, Arizona State University, Tempe, Arizona; [2]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania; [3]Department of Biology, Temple University, Philadelphia, Pennsylvania; and [4]Center for Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT Sequencing of the protein coding genome has revealed many different missense mutations of human proteins and different population frequencies of corresponding haplotypes, which consist of different sets of those mutations. Here, we present evidence for pairwise intramolecular epistasis (i.e., nonadditive interactions) between many such mutations through an analysis of protein dynamics. We suggest that functional compensation for conserving protein dynamics is a likely evolutionary mechanism that maintains high-frequency mutations that are individually nonneutral but epistatically compensating within proteins. This analysis is the first of its type to look at human proteins with specific high population frequency mutations and examine the relationship between mutations that make up that observed high-frequency protein haplotype. Importantly, protein dynamics revealed a separation between high and low frequency haplotypes within a target protein cytochrome P450 2A7, with the high-frequency haplotypes showing behavior closer to the wild-type protein. Common protein haplotypes containing two mutations display dynamic compensation in which one mutation can correct for the dynamic effects of the other. We also utilize a dynamics-based metric, EpiScore, that evaluates the epistatic interactions and allows us to see dynamic compensation within many other proteins.

SIGNIFICANCE Compensatory mutations provide a clear and widely studied example of adaptation in proteins. A large number of these compensatory mutations are found within humans at sites associated with genetic disease. Experimental studies suggest mutations that compensate for the effects of a disease causing mutation. Therefore an approach is needed to determine the functional effects and mechanisms underlying such compensatory interactions. Here, we show that compensatory mutations are distinguishable from noncompensatory mutations using a novel protein dynamics analysis. EpiScore may be used to determine the nature of interactions between two residues regardless of distance, allowing for increased accuracy in predicting compensatory sites.

INTRODUCTION

Numerous population-level interrogations of human protein variation have revealed tens of thousands of missense amino acid mutations resulting from single nucleotide polymorphisms that segregate in human populations at an appreciable frequency (>1%) (1). Many more missense mutations have been observed, but they occur with rather low frequencies. For our analysis of either case, we obtained a data set containing the population frequencies of known missense polymorphisms within 114 different enzymes for which we were also able to obtain crystal structures. We focused only on these missense polymorphisms and their consequences. Therefore any mention of polymorphisms, the amino acid mutations (or variants) that they cause, different alleles as defined by those mutations, and haplotypes consisting of different sets of those alleles as found within the human population all refer only to missense polymorphisms. From this data set, we observe that some common variants (8% of common minor alleles) are almost always observed with another common variant in the same protein in individuals of the population. These combinations are high-frequency double variants (HD) (Fig. 1 *A*). In contrast, many common variants never co-occur with other

Submitted October 12, 2022, and accepted for publication January 26, 2023.

*Correspondence: s.kumar@temple.edu or banu.ozkan@asu.edu

Editor: Chris Chipot.
https://doi.org/10.1016/j.bpj.2023.01.037

common variants in the same haplotype. We refer to them as high-frequency single variants (HS) (Fig. 1 *B*). They make up 19% of common minor alleles, a term that refers to the second most observed allele in human populations. In terms of population, both HD and HS alleles would likely be considered benign, as they exist within a substantial portion of all humans. This follows from the neutral theory of evolution (2), which suggests that amino acid substitutions are permitted so long as they have no adverse affects on a species' survival. The fact that certain haplotypes containing either HD or HS alleles are found at very low population frequencies implies that they are disease-associated haplotypes that are expected to be removed from a population through natural selection.

HD and HS variants are chosen as very clear examples of epistasis, which we define as a nonadditive outcome from two or more amino acid changes within a protein (3), such as two mutant alleles that may each increase the flexibility of a certain region but end up decreasing the flexibility of that region when they are found together. Upward of 90% of all amino acid substitutions have experienced some form of epistasis, as shown by Kondrashov et al. (4). Further computational and theoretical work expanding on that of Kondrashov (4) has shown that of this 90%, a further 30%–40% can be accounted for by the nonlinear dependence of fitness to folding stability (5). Epistasis is a common phenomenon, but HD and HS variants represent two opposite extremes, with two disease mutations leading to a functional variant and two functional mutations leading to disease, respectively. How is it possible that two wrongs make a right or that two rights make a wrong?

For this question, we hypothesized that HD haplotypes might be direct results of compensatory epistasis, where a prior mutation may accommodate the permissibility of a second mutation. Simulated evolution (6–9) and in vitro mutation (10) studies have explored the concept of epistasis as emergent sets of cooperative alleles. They have observed that each protein mutation may change the effect of subsequent mutations. Currently, metrics using the Potts model can predict such interactions (11–14); however, these models become less effective when the spatial distance between the two positions increases. Other studies have attempted to predict epistatic effects (15,16). These effects can be negative, reducing the fitness of a protein (10,17), or compensatory, where the subsequent mutation helps maintain the functional state (6,9,18). In this sense, HS alleles will be used as an example of negative epistasis.

In the past, others have used sequence-based methods to examine the epistasis taking place within a protein (11,12,19). Although they can determine residue contacts with a high degree of accuracy, more insight is needed in regard to longer range interactions and global protein behavior. More recent studies have explored long-range epistasis using various techniques, such as statistical models (20), machine learning (21), nuclear magnetic resonance spectroscopy (22), and the dynamic analysis of molecular dynamics (MD) simulations (23,24). We aim to utilize a similar protein structural dynamics-based approach to provide additional insight on a set of unique epistatic cases, our HD and HS mutations.

Within this study, we test our hypothesis from a biophysics perspective and explore the epistasis of HD and HS haplotypes using protein structural dynamics to provide additional mechanistic insights, just as such tools have already provided insight into protein evolution. The diversity of protein functions has grown over time via molecular evolution. Analyses of protein families indicate that proteins evolve for different functions through sequence variation while conserving their 3D structures (25,26). Evolution may instead produce changes in protein dynamics such as stability, flexibility, and allosteric dynamic coupling (27–34). By measuring such changes using MD simulations, we can infer overall protein function and predict how functionality may be lost or maintained (25,35–40).

MATERIALS AND METHODS

Dynamic flexibility index

The dynamic flexibility index (DFI) functions by measuring how a single residue responds to perturbations from each residue within a protein. The perturbation response scanning technique (41) is used to find the dynamic response profile of a given position through a combination of equilibrium dynamics and linear response theory. The perturbation response scanning approach can make use of the elastic network model (ENM) (41,42) to find correlated dynamics of positions in native equilibrium. In ENM, a protein is modeled as an arrangement of Cα atoms in which each atom pair is connected via a harmonic spring potential. A random Brownian kick is applied to each Cα atom sequentially to provide a perturbation to the elastic network. The goal of this perturbation is to simulate the forces exerted on a protein in a crowded cellular environment. When one residue is perturbed, the effect propagates through the rest of the network and causes the entire protein to respond. That response profile is obtained using linear response theory and given by the equation

$$[\Delta R]_{3N\times 1} = [H]^{-1}_{3Nx3N}[F]_{3N\times 1}, \quad (1)$$

where H is the Hessian, a $3N \times 3N$ matrix that can be built from 3D atomic coordinate information and is composed of the second derivatives of the harmonic potential with respect to the components of the position's vectors of length N. F is the external force vector applied at N residues in the protein, and ΔR is the response fluctuation of a residue position upon external force. In order to give an isotropic measure of response, the force is applied in several directions at each residue, and the magnitude of the response profile is averaged.

ENM is a coarse-grained model, which in this study was used for the analysis of the 114 enzyme set. In order to improve the accuracy of this model and allow sensitivity to mutations, the Hessian inverse can be replaced with the covariance matrices obtained from MD simulations. MD simulations were used in this study only for the analysis of cytochrome p450 CYP2A7 variants.

$$[\Delta R]_{3N\times 1} = [G]_{3N\times 3N}[F]_{3N\times 1} \quad (2)$$

Here, G is the covariance matrix containing the dynamic properties of the system. The covariance matrix contains the data for long-range interactions, solvation effects, and biochemical specificities of all types of interactions.

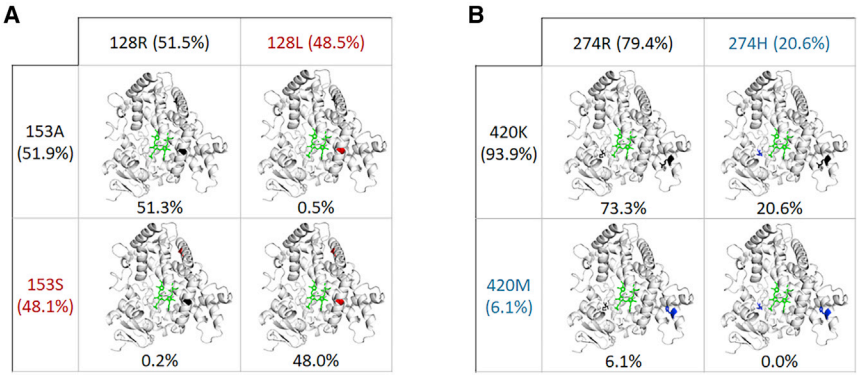

FIGURE 1 Population frequencies (given as percent values) of various allele combinations within cytochrome P450 CYP2A7. Two alleles are given for each locus position (residue index on the top and left sides). One major and one minor allele are given. The values next to single-letter amino acid codes denote the overall frequency at which the allele is observed. Other values denote the frequency of haplotypes containing variants directly upward and to the left. (*A*) For alleles 131A and 153S, we observe HD behavior, where both are found together in around half of all sequences. Yet neither is found at an appreciable frequency with the major allele present at the other position. Within our data set, 8% of common minor alleles exhibit similar behavior. (B) For alleles 368H and 274T, we observe HS behavior, where both are often found with a major allele present at the other position but are never observed together with an HS allele at both positions. Within our data set, 19% of common minor alleles exhibit this behavior. Both pairs shown here are also shown on the structure in Fig. S1.

In this work, the covariance matrices of wild-type cytochrome p450 CYP2A7 and all variants were constructed using previously obtained 800-ns all-atom explicit water MD simulations (see molecular dynamics simulations for details).

To calculate DFI, we apply unit isotropic perturbations to each individual residue, one by one, and obtain the residue fluctuation response profile of each position upon perturbing a specific position using Eq. 2. This process is repeated until we obtain the perturbation response matrix that contains the residue response profiles for all positions in a protein.

$$[A]_{3N\times 3N} = \begin{bmatrix} \left|\Delta R^1\right|_1 & \cdots & \left|\Delta R^N\right|_1 \\ \vdots & \ddots & \vdots \\ \left|\Delta R^1\right|_N & \cdots & \left|\Delta R^N\right|_N \end{bmatrix} \quad (3)$$

Where $\left|\Delta R^j\right|_i = \sqrt{\left(\Delta R_i^j\right)^2}$ is the magnitude of response at site i due to the perturbation at site j.

The DFI value of position i is then treated as the displacement response of position i relative to the net displacement response of the entire protein, which is calculated by sequentially perturbing each position in the structure:

$$DFI_i = \frac{\sum_{j=1}^{N}\left|\Delta R^j\right|_i}{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\Delta R^j\right|_i} \quad (4)$$

Therefore, the greater the DFI score at position i is, the more flexible that site will be, and the lower the score is, the more rigid that site will be, meaning it has less of a response to perturbations in the protein.

Dynamic coupling index

The dynamic coupling index (DCI) is an extension of DFI, developed to capture dynamic coupling between any given residue and functionally critical residues such as binding sites. This metric can locate sites that impact active site dynamics even at a distance through dynamic allosteric coupling.

As defined, *DCI* is the ratio of the sum of the mean-square fluctuation response of the residue i upon functional site perturbations (i.e., catalytic residues) to the response of residue i upon perturbations on all residues:

$$DCI_{ij} = \frac{\sum_{j=1}^{N_{functional}}\left|\Delta R^j\right|_i / N_{functional}}{\sum_{j=1}^{N}\left|\Delta R^j\right|_i / N} \quad (5)$$

Where $\left|\Delta R^j\right|_i$ is the response fluctuation profile of residue i upon perturbation of residue j. The numerator is the average mean-square fluctuation response obtained over the perturbation of the functionally critical residues $N_{functional}$, and the denominator is the average mean-square fluctuation response over all residues (43).

In this study, we use an additional tool, EpiScore, based off of DCI. EpiScore can capture nonadditive behavior between a given pair of residues using the ratio of the normalized perturbation response to a position k when a force is applied at two residues i and j simultaneously versus the average additive perturbation response when each residue i, j, is perturbed individually (Fig. 6 *A*). An EpiScore value less than 1 at a given position indicates that additive perturbations of positions i and j generate a greater response than a simultaneous perturbation at that position. Likewise, EpiScores greater than 1 indicate less of a response from additive perturbations than from a simultaneous perturbation. As EpiScore is a linear scale, the further the value is from 1, the greater is the effect described above (44).

Molecular dynamics simulations

To find the DFI and DCI scores of cytochrome p450 CYP2A7, we used MD to find the native conformational ensemble of each variant. The crystal structure was found using AlphaFold and based on the protein structure for CYP2A6 from the Protein Data Bank (PDB: 1Z10 (45)). The most common isoform of CYP2A7 shares a 93.72% sequence similarity with CYP2A6. To obtain structures for the cytochrome P450 CYP2A7 variants used, mutagenesis was performed using the PyMol (46) Mutagenesis Wizard. Topology files for all structures were prepared using the AMBER TLEAP program with the ff14SB force field (47), which added hydrogen atoms and surrounded each structure with a 14.0-Å cubic box of water molecules using the TIP3P (48) water model. Na^+ and Cl^- atoms were added for neutralization. The SANDER module of AMBER 14 (49,50) was used to ensure that the system reached a local energetic minimum and remove any unfavorable torsional angles or steric clashes.

During the simulation proper, the protein was first kept fixed with harmonic restraints, using a force constant of 10 kcal/mol(Å)^2, to allow surrounding water molecules and ions to relax. A second minimization phase followed afterward, in which the restraints were removed, and the protein-solution was further minimized. Both minimization phases employed the method of steepest descent followed by conjugate gradient. During the first phase, the solvent underwent 25,000 steepest descent cycles with a maximum of 50,000 minimization cycles, and during the second phase, the entire solution underwent 50,000 steepest descent cycles with a maximum of 100,000 minimization cycles The systems were then heated from 0 to 300 K over 250 ps using the GPU-accelerated PMEMD module of AMBER 14 (49), at which point long-range electrostatic interactions were calculated using the particle mesh Ewald method (51,52). Direct-sum,

nonbonded interactions were cut off at distances of 9.0 Å or greater. A Langevin thermostat (53) was used to control the temperature at 300 K and a Berendsen barostat (54) to adjust the pressure at 1 bar. The systems were then simulated using MD at constant temperature and pressure with 2-fs time steps for 800 ns. During these simulations, periodic boundary conditions were used, and the bond length of all covalent hydrogen bonds were constrained using SHAKE (50).

This study uses a protocol for the convergence of protein dynamics that has been established and used in prior studies (37,39,40,55). By using the Hessian in our DFI and DCI calculations, we restrict ourselves to a harmonic potential, and as such, we assume that data are sampled from a Gaussian distribution. This assumption is appropriate, provided that ergodicity is fulfilled in both simulation time as well as the time windows used in covariance matrix calculations. These requirements result in two basic conditions. 1) All conformations must be sampled from the same distribution. 2) The covariance matrices obtained ought to be independent of the initial atomic coordinates in order to eliminate global motions and accurately capture equilibrium coordinate information. To satisfy these conditions, we calculated covariance matrices using 50-ns moving windows that overlap by 25 ns over the last 300 ns of the trajectory of each simulation. The final average DFI profiles will be independent of the window size, so that averaging DFI profiles from different time window sizes (for example, 75 ns rather than 50 ns) will give similar results, and the calculated covariance matrices extracted from different times of trajectories within the last 300 ns should also result in similar DFI profiles.

RESULTS AND DISCUSSION

We first explore whether 3D structure alone provides insights about HD and HS variants. Because the protein sequence of cytochrome P450 CYP2A7 is nearly identical (93.72%) to CYP2A6, the structure of the proteins should be nearly identical as well. Therefore, the x-ray structure of cytochrome P450 CYP2A6 (PDB: 1Z10 (45)) was used to find the pairwise distances of HS or HD sets as well as their distance separation from the active site residues. Contrary to expectations, no evidence was found of any patterns differentiating HS sets from HD sets. For example, one could speculate that HD variant pairs, which always occur together, would always be interacting residues (i.e., in contact). Thus when one is mutated, the other needs to be mutated to retain the interactions necessary for the 3D fold. However, the pairwise distance distribution of HD pairs does not support this expectation, as we find many cases of HD pairs occurring spatially far apart (Fig. 2).

Rather than having only one native structure, we now know that proteins have an ensemble of states that accurately represent the native state. Proteins undergo many different local changes, which lead to a variety of conformations. In fact, this variety of conformations is unique to proteins compared with other natural macromolecules, and the conformational dynamics of proteins play a complex role in how those proteins evolve (32,35,56–59). Laboratory-directed evolution studies have shown that function-altering mutations require stabilizing mutations to compensate for a loss in stability (60–62). In addition, experimental and computational studies have shown that the timescales and motions of enzymatic activity can be widely different among enzyme homologs of different species, indicating

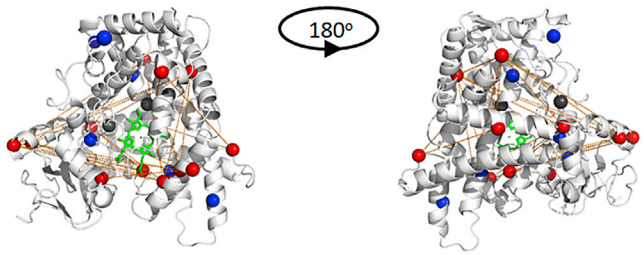

FIGURE 2 Ribbon diagrams of cytochrome P450 CYP2A7 in two different orientations. Important binding sites are shown in dark gray. HS variants and HD variants are represented by blue and red spheres, respectively. Dotted yellow lines connect pairs of red spheres, showing which HD variants exist together. The average distance between HD sites is 26.00 ± 2.84, and for HS sites, it is 28.46 ± 3.01. There is no significant difference (*t*-test, $p = 0.57$) between the distance distributions.

that these enzymes possess fundamentally different conformational dynamics while maintaining similar folds (36,63). Certain key regions were revealed in these studies, which exhibit differences in flexibility that may be mostly responsible for functional divergence. The functional specificity among structural homologs usually arises due to the evolution of intermediate frequency modes of structural dynamics, which are governed by the motion of local regions within a protein structure (32,35,56–59). Therefore, there should be a connection between the variation in conformational dynamics of specific positions and evolutionary rates. In support of that theory, studies involving specific protein families and subsets of enzymes have shown that residues that act as hinges (i.e., sites with low flexibility) are generally more evolutionarily conserved than other positions for specific protein families or a subset of enzymes (36). Although the above analyses suggest that dynamics play a role in evolution, we understand that, ultimately, evolution of proteins involves a continuous accumulation of amino acid substitution at different positions, some of which are functionally critical. For this reason, we attempted to determine the difference between HD and HS substitution types using a protein structural dynamics-oriented approach.

Particularly, we utilized a position-specific metric, the DFI. These DFI profiles estimate the role each protein residue plays in mediating structure-encoded dynamics and have been previously utilized in several important studies. In addition to being a powerful method to understand the global dynamics of a protein, DFI profiles have been used to gain several important insights into the roles that dynamics play in protein function. For example, DFI-based studies have shown that 1) preservation of dynamic of residues (i.e., flexibility) within 3D structures is critical for the maintenance of biological function (36); 2) alteration in structural dynamics during evolution leads to the emergence of new or altered functions in a diverse set of protein families including green fluorescent proteins (35), beta-lactamase inhibitors (37), and nuclear receptors (38); 3)

enzymatic function is regulated by dynamically coupled residues that form an allosteric communication network with the active site (64); and 4) many disease-associated mutations often trigger a global loss in dynamic coupling interactions, which in turn disrupts dynamic communication networks and ultimately leads to losses or gains in function (39). Given these successes, it stands to reason that DFIs are powerful tools that may provide fundamental mechanistic insights and shed light on the differences between HD and HS mutations.

The protein ensemble used for this study contains 114 different enzymes with known crystal structures. However, we chose to focus on cytochrome P450 CYP2A7 as it harbors many known missense variants and HD pairs. Cytochrome P450 proteins are catalysts for many reactions involved in drug metabolism and the synthesis of lipids (65). CYP2A7 in particular competes with its more widely studied isoform, CYP2A6, for miR-126* binding (66). Functionally, this downregulates the expression of CYP2A6 within the liver.

To compute the DFI profiles of each variant, we first performed MD simulations to obtain each variant's native ensemble and then applied our DFI analysis (see materials and methods). DFI quantifies the dynamic stability of a given position. It measures the resilience of a position to perturbations initiated at positions in the protein distal to the residue in question, but to which it is linked via structurally encoded global dynamics. Therefore, DFI profiles provide important information about protein function. Namely, residues that exhibit very low DFI*s* do not exhibit large amplitude fluctuations in response to random Brownian kicks but rather transfer the perturbation energies throughout the chain in a cascade fashion; examples of low DFI residues are those in hinge regions of proteins that control the motion critical for function (i.e., hinges in proteins look like hinges on a door; they do not move much but exert control over large-scale motions). On the other hand, sites with very high DFI are prone to perturbations to the protein backbone; they are structurally flexible sites and therefore play an important role in binding, signaling, or product release during enzymatic function. Particularly, the change in DFI profile upon amino acid substitutions captures changes in function, whereas similar DFI profiles yield similar biophysical properties and functions (59). To measure the similarities among DFI profiles of HD and HS variants, we use principal component analysis by performing singular value decompositions on the DFI profiles for each variant type. Fig. 3 shows their separation in the top three principal component axes.

Indeed, a relationship can be seen between the observed population frequency of variants and a 3D clustering of DFI values, made by analyzing the principal components of the DFI profiles of various mutant cytochrome P450 CYP2A7 proteins (Fig. 3). Common polymorphisms are generally clustered together on one side, whereas the rare

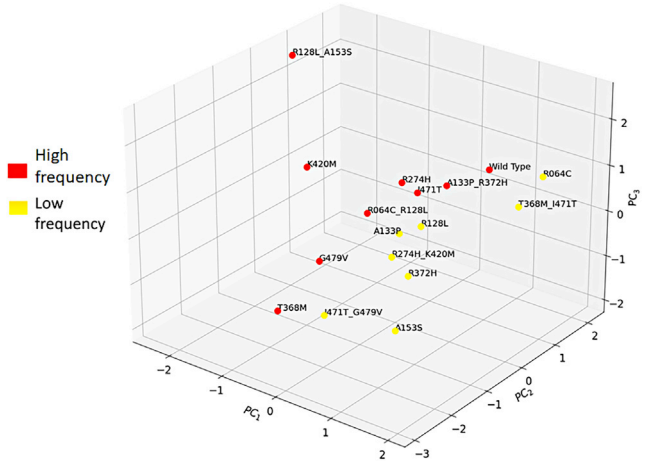

FIGURE 3 3D scatter plot presenting the clustering of the different variants of cytochrome P450 CYP2A7 based on principle component analysis of their DFI profiles. High-frequency mutations are generally clustered on one side of the plot along with the wild-type, suggesting that they exhibit similar biophysical properties.

(and therefore most likely function-damaging) variants are clustered on the other, suggesting that they exhibit significantly different flexibility profiles leading to drastic changes in their biophysical properties. Moreover, the DFI profiles of high population frequency variants are clustered together with that of the wild-type, demonstrating the role of epistatic relations. Particularly, variants lead to change in the flexibility profiles of different positions (even positions distal from the substitution sites in the 3D structure). Thus, the impact of the substitution can only be evaluated by computing the impact on the whole network of interactions (i.e., the given amino acid background composition), with special consideration to how these substitutions modulate the dynamics of functionally critical sites.

Indeed, as discussed above for each of the tested substitutions, differences in the DFI profiles were observed, often at locations distant to the mutation site, suggesting that changes in long-range dynamic coupling may be responsible for the altered flexibilities. To further analyze the specific interactions with the mutation sites, particularly with respect to functionally critical active sites, we employ the DCI. DCI is a parameter that captures the strength of a displacement response for position i upon perturbation of position j, relative to the average fluctuation response of position i to all other positions in the protein. In this way, DCI can show the degree of dynamic coupling between i and j. In this study, the DCI results are presented as %DCI, a percentile rank of the DCI range observed with values ranging from 0 to 1. It should be noted that the DFI and DCI are not the same: DFI measures flexibility, and DCI measures coupling. Furthermore, DCI specifically quantifies the coupling between individual positions, and as such, DCI values depend explicitly upon the positions selected for analysis. In any given protein, every amino acid position

has a unique network of direct, local interactions that gives rise to a unique network of highly coupled partner positions. Across the protein structure, this gives rise to an inhomogeneous, overall 3D network (39).

It is known from previous research that residues that are highly coupled to the active sites are more likely to be pathogenic upon mutation because of dynamic allosteric regulation (36,37,59). Highly coupled residues can dynamically affect one another, even at a distance, through a network of direct interactions. Through these allosteric effects, residues that are highly coupled to the active site may impact the flexibility profiles at the active site. For all mutation sites in either set of pairs, we calculated the maximum %DCI with any active site residue, because even a single active site residue may be essential to enzyme activity. These DCI values show a much more distinctly bimodal distribution in HD sites and have an almost exclusively high DCI in HS sites (Fig. 4 *A*). The bimodal distribution suggests one of the pair of variant sites is highly coupled, whereas the other is not, so the different residues in an HD pair may play one of two roles: the main communicator, which has a high DCI with the active site, or the communication modulator, which has a lower DCI (Fig. 4 *B*).

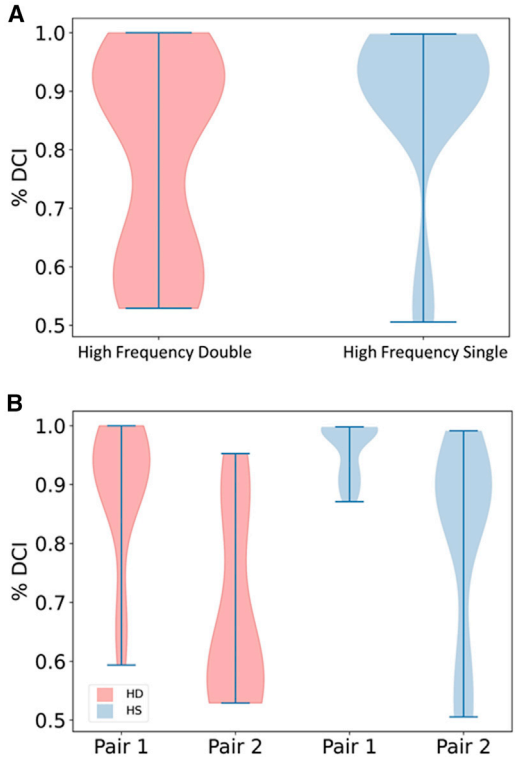

FIGURE 4 (*A*) Violin plots showing the distribution of %DCI with the active sites for sites of HD ($n = 38$) and HS ($n = 68$) pairs within cytochrome P450 CYP2A7. HS variant sites have almost entirely high DCI values. HD variant sites on the other hand display two different peaks within their distribution, suggesting that HD pairs have a broader range of acceptable values. (*B*) HD and HS distributions from (*A*) have been split in half. Each pair is split into a higher %DCI site, pair 1, and a lower %DCI site, pair 2. These pair 1 and pair 2 sites form their own distributions. Usually, HD pairs contain one high %DCI site and one low %DCI site, suggesting a pattern for compensatory mutations.

Alone, these variants cause problems within the protein, but together, they can compensate for the harmful effects of one another, canceling out more extreme dynamic shifts of the active site. On the other hand, the single peak around the high coupling region (0.9 %DCI) for HS pairs suggest that double mutants of HS pairs may drastically impact the active site dynamics. Because the HD mutations are observed in high frequencies together in populations, we would expect compensation that we should not see in HS pairs. The HD mutations provide an example of compensatory epistasis, as they are dependent on one another to function. Knowing that they compensate for one another in how they affect the active site specifically, that region can be further examined through the lens of DFI.

We particularly focus on the A117, N297, T305, I366, and F480 of the active site residues. These residues directly interact with the substrate, so the change in their flexibility from that of the wild-type could inhibit binding interaction by way of mechanically interfering with the function. We expect that the high-frequency variants should have only a minor effect on the DFI of the active site, thus allowing the dynamics of the active site to behave similarly to those of the WT. Indeed, it is seen that higher frequency variants do result in active site flexibilities closer to those of the wild-type. Particularly, we compared the DFI profiles of the double HS and HD variants with the average DFI profile of the corresponding single variants. Each single mutant from a given mutation pair, on its own, affects the flexibility of the entire protein through a complex network of interactions. Sometimes regions become more flexible, and sometimes they become more rigid. When the flexibility changes associated with a single mutation from a given mutation pair are averaged with the flexibility changes from the other of the pair, one ends up with DFI profiles that are closer to the wild-type than either single mutation on its own (Fig. 5 *A* and *B*). When we examine how double mutant models of the HD pairs affect the DFI profile, we observed that both mutations together result in a profile that is much like the average of single mutants. On the other hand, when we repeated the same analysis for HS pairs, we observed that their dynamic effects are not necessarily an average of their two individual effects. One may observe a large dynamic shift that is not present in a single variant due to some interference between the targeted residues. This interference seems to have an adverse effect on protein function. Only HS mutation pairs were found with active site flexibility that differed highly from the mean of either individual mutation. HD mutation pairs on the other hand showed active site DFI values that were close to both the wild-type protein and the average DFI of either individual mutation.

The observation that active sites are affected in approximately the same way by both the average of HD mutations and HD mutations together implies a degree of additivity

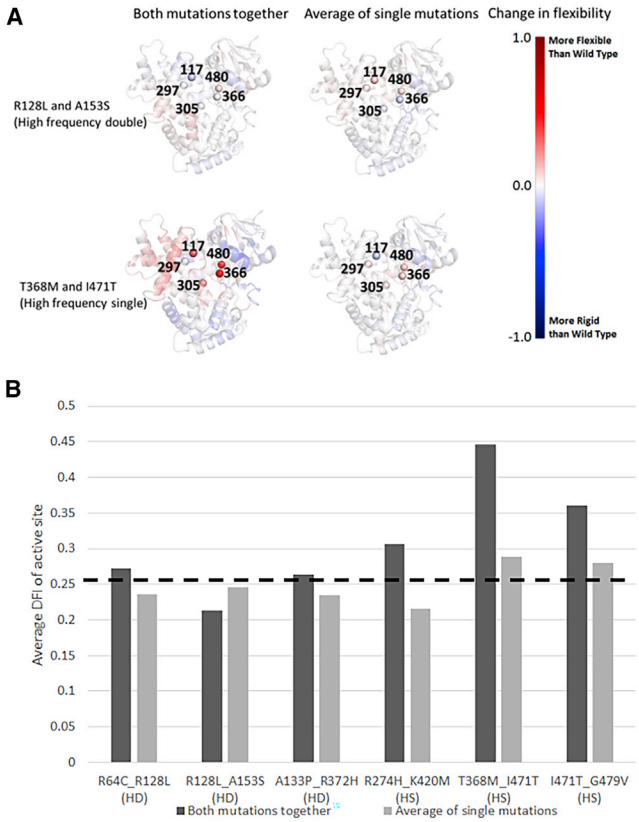

FIGURE 5 (*A*) Ribbon diagrams colored to show a change in flexibility of the mutants with respect to the wild-type using the dynamic flexibility index ($\Delta DFI = DFI_{mutant} - DFI_{wt}$). Spheres indicate active site residues. (*B*) Bar plot showing how the flexibility of the active site changes when the DFI profiles of the double mutants are analyzed (*in black*) and compared with the average DFI profiles using each DFI profile of single mutants (*in gray*).

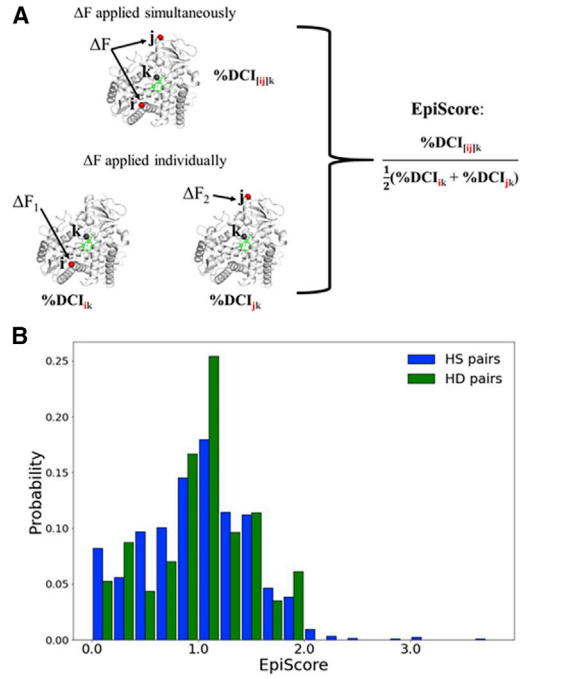

FIGURE 6 (*A*) Schematic describing the calculation of EpiScore. The numerator is the %dynamic coupling index (%DCI) value at position k upon a simultaneous perturbation to positions i and j divided by the average %DCI value at position k when positions i and j are perturbed individually. (*B*) Histogram showing the EpiScore distribution of commonly found together HD mutations (38 pairs total) and rarely found together HS mutations (151 pairs total). EpiScores of HD mutations are more strongly peaked around 1, indicating an additive interaction at the active site. EpiScores of HS mutations also peaked around 1 but are more widely distributed. Cytochrome P450 CYP2A7 was excluded from this figure, as the abundance of HD mutations would crowd out samples from the rest of the ensemble. The distribution with cytochrome P450 CYP2A7 included may be found in Fig. S2.

within these HD mutations. Simply put, the effects of mutation 1 plus the effects of mutation 2 equal the effects of mutations 1 and 2. This additivity is not observed in HS mutations, where the effects of example mutations 1 and 2 together can be quite unexpected. The HD mutations should be more compensatory and additive depending on one another for function. To help predict the epistatic effects of a pair or residues, we use EpiScore (25,44).

EpiScore (Fig. 6 *A*) is a metric designed to predict the effects of double mutations on a target residue compared with their respective single mutations. Using EpiScore, interactions are modeled as the fluctuation response strength of simultaneous versus individual, additive perturbations between epistatic residue pairs and important catalytic sites. A lower EpiScore at a position suggests that the two residues generate less of a response at that location, whereas an EpiScore near 1 suggests that the two residues generate an equal response as the average of their single mutations at that location, so they are additive.

To expand how general is the distinction between the epistatic relationships of frequently observed HD versus HS variants, we use an ensemble of 114 enzymes harboring 38 HD and 151 HS variant pairs. For this large set, we computed the EpiScores of the mutational sites and generated distributions (Fig. 6 *B*) based on their observed frequency in the populations. A distribution centered near 1 suggests that most mutation sites behave additively. We find that the HD mutation distribution is more highly peaked around 1 with a mean of 1.03 and standard deviation of .61, and we see HS mutations more frequently at higher EpiScore of value greater than 1 (a mean of 1.17 and standard deviation of 3.70). This implies that a majority of mutation pairs will behave additively with one another, providing the role of negative selection. More importantly, the HD pair distribution suggests that the requirements for the HD mutation phenotypes, in terms of their epistatic interplay, are stricter than for HS mutations. Particularly, the HD mutations, which are only observed when both mutations are present, interact in a compensatory manner in which the effects of one mutation are counterbalanced by the other mutation such that the overall stability and dynamical behavior of the protein for a given

HD sequence allows the mutation pair to be jointly permissible.

CONCLUSION

In our study, we took a sample of 114 enzymes and examined recorded variants from human populations. A number of these are HD variants, which occur almost exclusively in pairs. Comparing these with HS variants, which are almost always found alone, we find significant differences in their structural dynamic profiles. Previous studies have found that the closer the DFI profile of a variant is to the wild-type, the more likely it is that the variant does not negatively impact protein function (38,67). These findings are supported by the results of this study as well, which shows that within cytochrome p450 CYP2A7, pairs of HD mutations are closer to the wild-type than pairs of HS mutations. Analysis of DFI profiles of HD or HS variants on their own further reveals how they differ. Taking an average of pairs of DFI profiles from single mutations, we generally find that changes brought about by the single mutations are corrected for, especially near the active site. However, when both mutations are present in a single protein, HS pairs display nonadditive effects in that their flexibility profiles vary significantly from the average flexibility profiles of individual single variants and therefore from that of the wild-type. This suggests that epistatic relationships play a crucial role in allowing double versus single variants observed in human populations. To further analyze this on a large scale of 114 enzymes, we computed EpiScore values for residue pairs with respect to active sites positions, which is a more direct measure of the additivity using structural dynamics. EpiScores of HD mutations are near 1 on average, indicating that additivity is an important quality for these mutations.

DCI also reveals dynamic mechanisms behind HD mutations. We notice that within cytochrome p450 CYP2A7, a vast majority of HS mutations are highly coupled to the active site, whereas only around half of HD mutations are highly coupled to the active site. Because of this, we theorize that HD mutations pairs contain a residue that is highly coupled to the active site and a residue that is not. Together, these residues interact in a way that has little negative impact on the protein. In summary, epistatic HD mutations leave unique dynamic signatures that affect the active site differently than other more typical mutations. Perhaps in future works, these signatures may be used to predict the locations of epistatic residues.

As additional insight on this phenomenon, we observed that although the alleles studied here are often observed in humans, an analysis of homologous species marks them as unlikely to occur through evolution (1). These evolutionarily unlikely alleles have been found to alter protein function at a higher rate than expected. According to our analysis, HD and HS mutations are found to have a low evolutionary probability (EP) (less than .05) at rates 1.33 times and 1.16 times higher, respectively, than other alleles occurring in the same range of frequencies. Fig. S3 shows that indeed a majority of HS and HD alleles are low EP. These low-EP, high-frequency mutations would be expected in the case of special divergence and therefore may represent uniquely human adaptations.

A lack of these alleles in inferred ancestral sequences indicates negative selective pressure, and high rates of observation in humans indicate that this is no longer the case. The reason for this change is presumed to be a continued buildup of neutral mutations that change the protein's dynamic landscape in such a way that allows these previously forbidden alleles. In this way, the very nature of low-EP, high-population alleles suggests some epistatic interactions. More specifically to our studied pairs, the landscape may change in such a way that forces the interaction of two residues. The fact that low-EP alleles are most common among HD mutations may indicate that pairwise interactions are a common mechanism through which this landscape change occurs.

As a final thought, we note that the dynamic behaviors of high-EP alleles show similar patterns to their low-EP counterparts. We find in cytochrome P450 CYP2A7 two high-frequency, high-EP polymorphisms that are often observed alone (i.e., HS pairs). They are neutral according to population data and expected to be neutral according to their high EP values, but they are rarely observed together. Like the low-EP HS pairs, these single variants have the opposite behavior compared with HD pairs: the average EpiScore at active sites is 1.21 (ranging in value from 0.37 to 1.62), indicating a degree of dynamic nonadditivity. Indeed, when we compare the dynamic flexibility (DFI) of the double mutant with the average DFI profiles of the single mutants, we see that double mutant significantly alters the flexibility of the active site positions and is drastically different from the average additive profile of the single mutants. The high-EP pair in Fig. S4 may be compared with the low-EP pair in Fig. 5 *A*. In either case, although they are observed frequently in human populations, two neutral mutations significantly alter dynamics when they are presented together, leading to a negative epistatic effect on function.

Although it has been shown before that protein dynamics have been able to predict the effects of single mutations, shown here for the first time to our knowledge, dynamics allows us to distinguish compensatory variants from noncompensatory variants. In particular, the concept of dynamic epistasis is explored with the metric of EpiScore, which uses the concept of dynamic additivity. Our findings suggest that the effects of more enigmatic mutations, which are almost never found separately, may be additive and compensate for one another to maintain the function of the wild-type.

DATA AVAILABILITY STATEMENT

The code to perform DFI and DCI analysis is available at https://github.com/SBOZKAN/DFI-DCI. MD data are

available upon request. The enzyme structure list, catalytic sites, mutation sites, allele frequencies, and EP values are contained in the Supporting Material files as "Supplementary_protein_info.csv".

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2023.01.037.

AUTHOR CONTRIBUTIONS

N.J.O. and S.B.O. conceptualized and designed the project. S.B.O. supervised the project. R.P. obtained genomic data. N.J.O. performed and analyzed simulations. P.C. wrote the EpiScore code. P.C. and S.K. provided feedback on results. The manuscript was written by N.J.O. and S.B.O. with input from all other authors.

ACKNOWLEDGMENTS

Funding was provided to N.J.O. and P.C. by the Gordon and Betty Moore Foundation (award number AWD00034439) and to S.B.O. by the National Science Foundation (award numbers: 1715591 and 1901709) and the National Institutes of Health R01GM147635-01. SK acknowledges fundings: National Science Foundation (GCR 1934848) and the National Institutes of Health (GM139540).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Liu, L., K. Tamura, …, S. Kumar. 2016. A molecular evolutionary reference for the human variome. *Mol. Biol. Evol.* 33:245–254.
2. Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.
3. Starr, T. N., and J. W. Thornton. 2016. Epistasis in protein evolution. *Protein Sci.* 25:1204–1218.
4. Breen, M. S., C. Kemena, …, F. A. Kondrashov. 2012. Epistasis as the primary factor in molecular evolution. *Nature.* 490:535–538.
5. Dasmeh, P., and A. W. R. Serohijos. 2018. Estimating the contribution of folding stability to nonspecific epistasis in protein evolution. *Proteins.* 86:1242–1250.
6. de la Paz, J. A., C. M. Nartey, …, F. Morcos. 2020. Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. USA.* 117:5873–5882.
7. Bisardi, M., J. Rodriguez-Rivas, …, M. Weigt. 2022. Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. *Mol. Biol. Evol.* 39:msab321.
8. Peters, A. D., and C. M. Lively. 1999. The red queen and fluctuating epistasis: a population genetic analysis of antagonistic Coevolution. *Am. Nat.* 154:393–405.
9. Otten, R., L. Liu, …, J. S. Fraser. 2018. Rescue of conformational dynamics in enzyme catalysis by directed evolution. *Nat. Commun.* 9:1314.
10. Bershtein, S., M. Segal, …, D. S. Tawfik. 2006. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature.* 444:929–932.
11. Ekeberg, M., C. Lövkvist, …, E. Aurell. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E - Stat. Nonlinear Soft Matter Phys.* 87:012707.
12. Levy, R. M., A. Haldane, and W. F. Flynn. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* 43:55–62.
13. Rizzato, F., A. Coucke, …, S. Cocco. 2020. Inference of compressed Potts graphical models. *Phys. Rev. E.* 101:012309.
14. Shimagaki, K., and M. Weigt. 2019. Selection of sequence motifs and generative Hopfield-Potts models for protein families. *Phys. Rev. E.* 100:032128.
15. Collins, S. R., M. Schuldiner, …, J. S. Weissman. 2006. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol.* 7:R63.
16. Harrigan, P., H. D. Madhani, and H. El-Samad. 2018. Real-time genetic compensation defines the dynamic demands of feedback control. *Cell.* 175:877–886.e10.
17. Yang, Q. E., C. MacLean, …, T. R. Walsh. 2020. Compensatory mutations modulate the competitiveness and dynamics of plasmid-mediated colistin resistance in Escherichia coli clones. *ISME J.* 14:861–865.
18. Rojas Echenique, J. I., S. Kryazhimskiy, …, M. M. Desai. 2019. Modular epistasis and the compensatory evolution of gene deletion mutants. *PLoS Genet.* 15:e1007958.
19. Marks, D. S., L. J. Colwell, …, C. Sander. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One.* 6:e28766.
20. Barnes, J. E., C. R. Miller, and F. M. Ytreberg. 2022. Searching for a mechanistic description of pairwise epistasis in protein systems. *Proteins.* 90:1474–1485.
21. Li, G., Y. Qin, …, M. T. Reetz. 2021. Machine learning enables selection of epistatic enzyme mutants for stability against unfolding and detrimental aggregation. *Chembiochem.* 22:904–914.
22. Kim, I., A. Dubrow, …, J.-H. Cho. 2022. Energy landscape reshaped by strain-specific mutations underlies epistasis in NS1 evolution of influenza A virus. *Nat. Commun.* 13:5775.
23. Yu, H., and P. A. Dalby. 2018. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proc. Natl. Acad. Sci. USA.* 115:E11043–E11052.
24. Yu, H., and P. A. Dalby. 2018. Exploiting correlated molecular-dynamics networks to counteract enzyme activity–stability trade-off. *Proc. Natl. Acad. Sci. USA.* 115:E12192–E12200.
25. Modi, T., P. Campitelli, …, S. B. Ozkan. 2021. Protein folding stability and binding interactions through the lens of evolution: a dynamical perspective. *Curr. Opin. Struct. Biol.* 66:207–215.
26. McLeish, T. C. B., T. L. Rodgers, and M. R. Wilson. 2013. Allostery without conformation change: modelling protein dynamics at multiple scales. *Phys. Biol.* 10:056004.
27. Bhabha, G., D. C. Ekiert, …, P. E. Wright. 2013. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat. Struct. Mol. Biol.* 20:1243–1249.
28. Keskin, O., I. Bahar, …, D. G. Covell. 2000. Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure. *Anti Cancer Drug Des.* 15:79–98.
29. Kuzmanic, A., G. R. Bowman, …, F. L. Gervasio. 2020. Investigating Cryptic binding sites by molecular dynamics simulations. *Acc. Chem. Res.* 53:654–661.
30. Nussinov, R., and C.-J. Tsai. 2013. Allostery in disease and in drug discovery. *Cell.* 153:293–305.
31. Swint-Kruse, L., K. S. Matthews, …, B. M. Pettitt. 1998. Comparison of simulated and experimentally determined dynamics for a variant of the LacI DNA-binding domain, nlac-P. *Biophys. J.* 74:413–421.
32. Campbell, E., M. Kaltenbach, …, C. J. Jackson. 2016. The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* 12:944–950.

33. Ma, B., and R. Nussinov. 2016. Conformational footprints. *Nat. Chem. Biol.* 12:890–891.
34. Saavedra, H. G., J. O. Wrabl, …, V. J. Hilser. 2018. Dynamic allostery can drive cold adaptation in enzymes. *Nature.* 558:324–328.
35. Kim, H., T. Zou, …, R. Wachter. 2015. A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. *Structure.* 23:34–43.
36. Campitelli, P., T. Modi, …, S. B. Ozkan. 2020a. The role of conformational dynamics and allostery in modulating protein evolution. *Annu. Rev. Biophys.* 49:267–288.
37. Modi, T., V. A. Risso, …, S. Banu Ozkan. 2021. Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity. *Nat. Commun.* 12:1852.
38. Kumar, A., T. J. Glembo, and S. B. Ozkan. 2015b. The role of conformational dynamics and allostery in the disease development of human ferritin. *Biophys. J.* 109:1273–1281.
39. Campitelli, P., L. Swint-Kruse, and S. B. Ozkan. 2021. Substitutions at nonconserved rheostat positions modulate function by rewiring long-range, dynamic interactions. *Mol. Biol. Evol.* 38:201–214.
40. Ose, N. J., B. M. Butler, …, S. B. Ozkan. 2022. Dynamic coupling of residues within proteins as a mechanistic foundation of many enigmatic pathogenic missense variants. *PLoS Comput. Biol.* 18:e1010006.
41. Atilgan, C., and A. R. Atilgan. 2009. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.* 5:e1000544.
42. Atilgan, A. R., S. R. Durell, …, I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–515.
43. Campitelli, P., J. Guo, …, S. B. Ozkan. 2018. Hinge-shift mechanism modulates allosteric regulations in human Pin1. *J. Phys. Chem. B.* 122:5623–5629.
44. Campitelli, P., and S. B. Ozkan. 2020. Allostery and epistasis: emergent properties of anisotropic networks. *Entropy.* 22:667.
45. Yano, J. K., M.-H. Hsu, …, E. F. Johnson. 2005. Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen. *Nat. Struct. Mol. Biol.* 12:822–823.
46. Schrodinger 2015. The PyMOL Molecular Graphics System Version 2.0.4. .
47. Maier, J. A., C. Martinez, …, C. Simmerling. 2015. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theor. Comput.* 11:3696–3713.
48. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.
49. Salomon-Ferrer, R., A. W. Götz, …, R. C. Walker. 2013. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theor. Comput.* 9:3878–3888.
50. Pearlman, D. A., D. A. Case, …, P. Kollman. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91:1–41.
51. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.
52. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.
53. Hünenberger, P. H. 2005. Thermostat algorithms for molecular dynamics simulations. *In* Advanced Computer Simulation. C. Holm and K. P. D. Kremer, eds Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 105–149.
54. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.
55. Campitelli, P., J. Lu, and S. B. Ozkan. 2022. Dynamic allostery highlights the evolutionary differences between the CoV-1 and CoV-2 main proteases. *Biophys. J.* 121:1483–1492.
56. Glembo, T. J., D. W. Farrell, …, S. B. Ozkan. 2012. Collective dynamics differentiates functional divergence in protein evolution. *PLoS Comput. Biol.* 8:e1002428.
57. Bhabha, G., D. C. Ekiert, …, P. E. Wright. 2013. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat. Struct. Mol. Biol.* 20:1243–1249.
58. Zou, T., V. A. Risso, …, S. B. Ozkan. 2015. Evolution of conformational dynamics determines the Conversion of a promiscuous generalist into a specialist enzyme. *Mol. Biol. Evol.* 32:132–143.
59. Modi, T., J. Huihui, …, S. B. Ozkan. 2018. Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 373:20170184.
60. Alber, T. 1989. Mutational effects on protein stability. *Annu. Rev. Biochem.* 58:765–798.
61. Guerois, R., J. E. Nielsen, and L. Serrano. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320:369–387.
62. Yue, P., Z. Li, and J. Moult. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353:459–473.
63. Butler, B. M., I. C. Kazan, …, S. B. Ozkan. 2018. Coevolving residues inform protein dynamics profiles and disease susceptibility of nSNVs. *PLoS Comput. Biol.* 14:e1006626.
64. Kumar, A., B. M. Butler, …, S. B. Ozkan. 2015a. Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine. *Curr. Opin. Struct. Biol.* 35:135–142.
65. Bethesda (MD): National Library of Medicine (US); National Center for Biotechnology Information. 2004. CYP2A7 [Internet].
66. Nakano, M., Y. Fukushima, …, M. Nakajima. 2015. CYP2A7 pseudogene transcript affects CYP2A6 expression in human liver by acting as a decoy for miR-126. *Drug Metab. Dispos.* 43:703–712.
67. Nevin Gerek, Z., S. Kumar, and S. Banu Ozkan. 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. *Evol. Appl.* 6:423–433.