PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Dynamic coupling of residues within proteins as a mechanistic foundation of many enigmatic pathogenic missense variants

Nicholas J. Ose[1], Brandon M. Butler[1☯], Avishek Kumar[1☯], I. Can Kazan[1], Maxwell Sanderford[2,3], Sudhir Kumar[2,3,4]*, S. Banu Ozkan[1]*

1 Department of Physics and Center for Biological Physics, Arizona State University, Tempe, Arizona, United States of America, **2** Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, United States of America, **3** Department of Biology, Temple University, Philadelphia, Pennsylvania, United States of America, **4** Center for Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

☯ These authors contributed equally to this work.
* s.kumar@temple.edu (SK); Banu.Ozkan@asu.edu (SBO)

Citation: Ose NJ, Butler BM, Kumar A, Kazan IC, Sanderford M, Kumar S, et al. (2022) Dynamic coupling of residues within proteins as a mechanistic foundation of many enigmatic pathogenic missense variants. PLoS Comput Biol 18(4): e1010006. https://doi.org/10.1371/journal.pcbi.1010006

Editor: Anders Wallqvist, US Army Medical Research and Materiel Command: US Army Medical Research and Development Command, UNITED STATES

Received: September 20, 2021

Accepted: March 9, 2022

Published: April 7, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010006

Abstract

Many pathogenic missense mutations are found in protein positions that are neither well-conserved nor fall in any known functional domains. Consequently, we lack any mechanistic underpinning of dysfunction caused by such mutations. We explored the disruption of allosteric dynamic coupling between these positions and the known functional sites as a possible mechanism for pathogenesis. In this study, we present an analysis of 591 pathogenic missense variants in 144 human enzymes that suggests that allosteric dynamic coupling of mutated positions with known active sites is a plausible biophysical mechanism and evidence of their functional importance. We illustrate this mechanism in a case study of *β-Glucocerebrosidase* (GCase) in which a vast majority of 94 sites harboring Gaucher disease-associated missense variants are located some distance away from the active site. An analysis of the conformational dynamics of GCase suggests that mutations on these distal sites cause changes in the flexibility of active site residues despite their distance, indicating a dynamic communication network throughout the protein. The disruption of the long-distance dynamic coupling caused by missense mutations may provide a plausible general mechanistic explanation for biological dysfunction and disease.

Author summary

Genetic diseases often occur when mutations in proteins cause gain/loss of functions. Although several methods based on conservation and protein biochemistry exist to predict genetic mutations that may impact function, many disease-associated mutations remain unexplained by these metrics. In this study, we sought a mechanistic explanation for such disease-associated mutations. In order to function, important regions of a protein must be able to exhibit collective motion. Through computer simulations, we observed that mutation of even a single amino acid position within a protein can change the protein

Data Availability Statement: The code to perform DFI and DCI analysis is available at https://github.com/SBOZKAN/DFI-DCI. Molecular Dynamics data are available at https://github.com/SBOZKAN/GCase. The enzyme structure list, mutation sites, and catalytic sites are contained in the Supporting information files as "S1 Table.csv". GCase disease mutation sites are contained in the Supporting information files as "S2 Table.csv". Neural net input features are contained in the Supporting information files as "S1 Data.csv". Molecular Dynamics input data are contained in the Supporting information files as "S1 Files.rar".

Funding: S.B.O. and N.J.O. were supported by the National Science Foundation Division of Molecular and Cellular Biosciences (award 1715591) (https://www.nsf.gov/) and the Gordon and Betty Moore Foundation Award #8415 (https://www.moore.org/) This research was supported by grants from the U.S. National Science Foundation to S.K. (GCR-1934848) (https://www.nsf.gov/) and the U.S. National Institutes of Health to S.K. (GM-139504-01) (https://www.nih.gov/) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

motion. We found that disease-associated mutations tend to alter the motion of regions critical to protein function, even though these mutations occur far from these critical regions. In addition, we examined the degree to which two amino acid positions within a protein may be "coupled," i.e., the extent to which motion in one position affects the other. We found that positions highly coupled to the active site of a protein are more likely to result in disease when mutated, thereby offering a new tool for predicting pathogenesis of new mutations by incorporating internal protein dynamics.

Introduction

Our understanding of factors responsible for the pathogenesis of disease-association variants (DAVs) in proteins continues to evolve. From a biophysics perspective, it has been shown that DAVs could alter the stability of a protein [1–3]. But, only one-third of over 2,000 mutations led to a decrease in protein stability, a high-throughput functional assay revealed [4]. Rather than affecting stability, a large fraction of DAVs seems to impair specific protein-ligand function or enzymatic activity [5–8]. Furthermore, studies combining evolutionary approaches with the biochemistry of protein design have revealed that DAVs at non-conserved sites can involve complex and frequently poorly understood mechanisms [5,9–11].

Through sequencing efforts, a large catalog of missense variants in thousands of human proteins has been assembled, including those implicated in diseases (Fig 1A) [5,11–13]. However, many DAVs occur at positions that are neither evolutionary well-conserved nor a part of any known functional domain (Fig 1C). Regardless of biochemical similarity, amino acid substitutions at non-conserved sites lead to a wide range of outcomes, increasing or decreasing functional activity at up to three orders of magnitude (i.e., the rheostatic pattern of change) [14]. These enigmatic mutations are frequently misdiagnosed because neither evolutionary nor static structural features are informative. In fact, many rare missense variants occur at fast-evolving positions that do not have functional annotations (Fig 1B), which adversely impacts the prediction accuracy of commonly used methods because they run counter to expectations. In Fig 1D, we see that EvoD is able to exceed the prediction accuracy of other contemporary sequence-based metrics by accounting for additional evolutionary properties [15].

Here we explore the mechanistic role of dynamic allosteric coupling of sites carrying DAVs with the catalytic sites important for enzymatic activity. Our exploration is based on the premise that many mutations alter conformational dynamics of proteins, shifting the distribution of the ensemble and protein function, including the emergence of new functions [10,17–21], adaption to different environments [22], and dysfunction [12,23].

We use the *dynamic coupling index* (DCI) to identify sites strongly coupled to active sites critical for function [24,25]. We refer to them as dynamic allosteric residue coupling (DARC) sites. A mutation at a DARC site is likely to influence conformational dynamics and allosteric regulation, making individuals carrying mutants of these sites highly susceptible to disease phenotypes.

Firstly, in order to elucidate this allosteric mechanism, we used Molecular Dynamics (MD) simulations to examine GCase, a signature human enzyme consisting of 497 amino acids and at least 94 amino acids with observed DAVs implicated in Gaucher disease (GD) [26], which is characterized by a dangerous buildup of lipids in certain organs. Genetic changes in GCase can lead to other health conditions as well, including Parkinson's disease [27–30] and Dementia with Lewy bodies [30,31]. We investigated the mechanistic impact of these mutations on conformational dynamics and allosteric regulation [32]. In the following, we report that GD

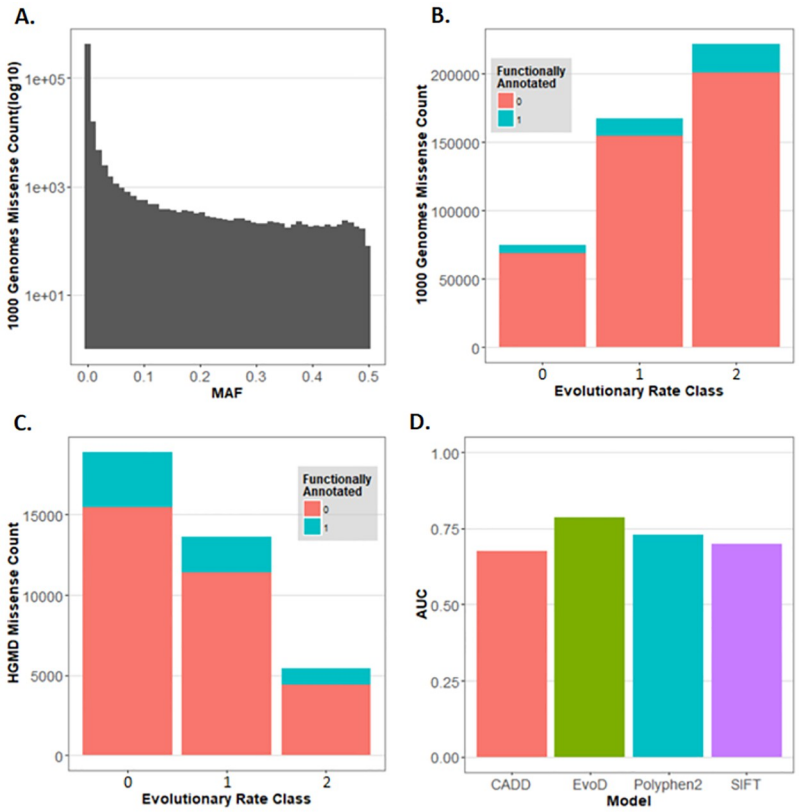

Fig 1. Frequency, evolutionary conservation, and rates of misdiagnosis of missense variants. (**a**) Histogram of minor allele frequencies (MAF) of missense variants in the 1000 Genomes data set. (**b**) Counts of these missense variants according to evolutionary conservation and the Uniprot functional annotation of their domain of residence. Evolutionary rate classes are from Kumar et al. [15] with class 0 sites containing no substitutions, class 1 sites exhibiting 0–1 substitutions per billion years, and class 2 sites exhibiting greater than 1 substitution per billion years. (**c**) Histogram of evolutionary conservation of sites containing only known pathogenic missense variants found in the Human Gene Mutation Database (HGMD) [16], with and without functional annotation in the Uniprot database. (**d**) Performance of four different missense diagnosis tools, quantified by their area under the receiver operation curve (AUC), which measures their ability to discriminate between putatively neutral (1000 Genomes missense variants with MAF>1%) and disease associated variants (DAVs) found in fast evolving positions (evolutionary rate class of 2). DAVs with MAF > 0.01% were excluded from these analyses.

https://doi.org/10.1371/journal.pcbi.1010006.g001

mutations disrupt allosteric regulation due to changes in dynamic flexibility around the catalytic sites, altering enzymatic activity essential for homeostasis. The positions harboring DAVs can be thought of as key DARC sites.

While all atom MD simulations are sensitive enough to investigate how mutations at specific distal sites (usually diagnosed as benign by conventional *in silico* tools) can modulate the overall dynamics of functionally critical sites, therefore allosterically impacting function, MD approaches are often time consuming and computationally expensive. To explore the role of conformational dynamics and allostery in missense variants of many proteins with different 3-D structures at a broader scale, we utilized a more efficient coarse-grain approach, the Elastic Network Model (ENM), to conduct a proteome-wide analysis of allosteric coupling for a set of enzymes. These analyses suggest that pathogenic variants are most abundant at DARC sites. We also present an analysis of DCI asymmetry, which measures the degree of symmetry in the dynamic coupling between two sites, revealing that mutations are likely to result in a loss of function if they occur at distal sites controlled by the active site, resulting in pathogenesis.

Results

Disease-associated mutations modify dynamics throughout the protein

GCase is a member of the family of glycoside hydrolases that use glutamates for hydrolyzing glucocerebrosidase into glucose and ceramide. Many amino acid variants of this enzyme are reported to cause Parkinson's disease [27–30], Dementia with lewy bodies [30,31], and GD [33]. Using the crystal structure of GCase (PDB ID: 1ogs) [34] and 94 sites with DAVs [35], we calculated the Euclidean distance between the mutation site and the active site (e.g., residues 235 and 340). A vast majority (87.5%) of GD pathogenic variants occur further than 10 Å from the nearest active site residue, making direct interactions implausible. This suggests the existence of a network of indirect interactions through which a mutation at a distal site can induce dynamic changes at other regions of the protein and, by extension, impact protein function. The behavior of residues within this network can be examined by using a structural dynamics-oriented approach.

We illustrate the approach using structural dynamics by considering the example of a single mutant *N370S* that is present with a high frequency (~70%) in the Ashkenazi Jewish population and studied extensively [27,32]. We first calculated the structural flexibility profiles of residues using a position-specific dynamic flexibility index, or DFI (see the Methods section). A comparison between DFI profiles of the wild-type GCase protein with the one that contains *N370S* is shown in Fig 2A. The DFI profile provides an estimate for the role of each residue in

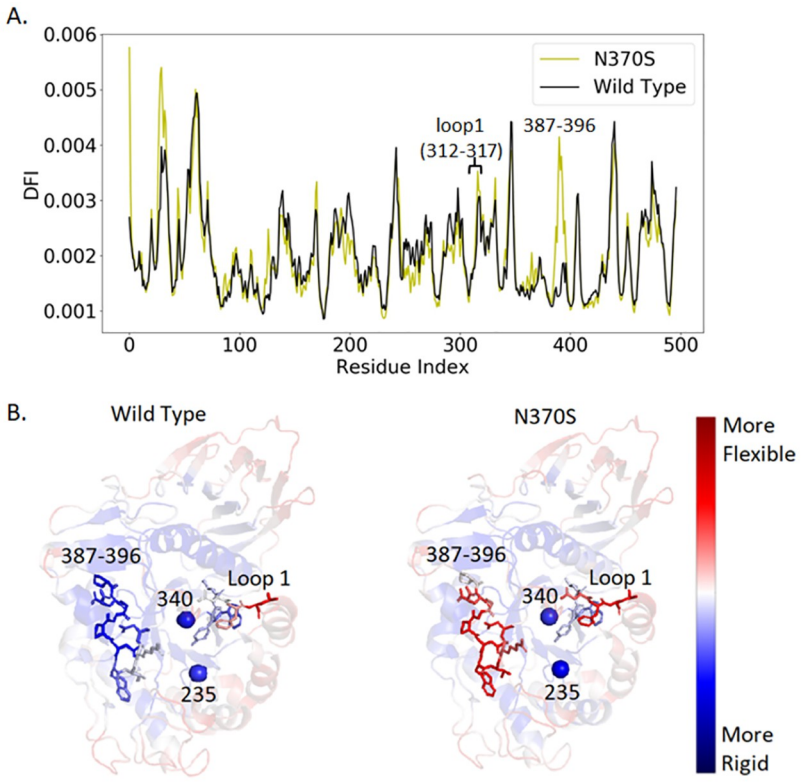

Fig 2. A comparison of DFI profiles of wild-type GCase and N370S mutant protein. (**a**) The %DFI profile of the mutant protein (*N370S*, yellow) is contrasted with that of the wild-type (black). Dissimilarities in the two profiles demonstrate how a single point mutation (*N370S*) can induce changes in the flexibility profile of a different region of the protein. (**b**) Ribbon diagrams showing DFI as a color-coded spectrum from red-white-blue; red and blue indicate the highest and lowest flexibility, respectively. The regions with the most significant changes in dynamic flexibility are highlighted.

https://doi.org/10.1371/journal.pcbi.1010006.g002

mediating structure-encoded dynamics. As for GCase, the DFI profile indicates significant shifts in dynamics caused by *N370S*. Regions of the protein that should be rigid are now flexible and vice versa (Fig 2B). Hinges in the protein have moved or disappeared, and new hinges have appeared elsewhere. As reported in previous studies, these hinge shifts suggest a major change in dynamics and thus protein function [19,24,25,36].

Among the five loops surrounding the active site, we observe that loop 1 (residues 312–317) exhibits an increase in DFI scores (Fig 2A), suggesting that increased flexibility of this loop could contribute to the decrease in enzymatic activity by hindering the accessibility of the ligand to the active site as reported previously [37]. This variant displays a small change in flexibility near loop 1. Changes in DFI within loop 1 for other studied mutations are shown in the supplementary figure (S1 Fig). Additionally, the protein with *N370S* shows a very large shift in flexibility between residues 387 to 396, which overlaps with loop 3 (residues 394–399); within the overlap is the *R395* residue, which orients differently in the active and inactive states of the enzyme [23] (Fig 2A).

Mutations at distal sites dynamically-coupled to the active site alter long-range communication

Although the only sequence difference between the wild-type and mutant GCase is a single residue, DFI changes across the protein. This behavior suggests that changes in long-range dynamic coupling may be responsible for the altered flexibility profiles. The dynamic coupling index (DCI) captures the strength of the displacement response for site *i* upon perturbation of site *j*, relative to the average fluctuation response of site *i* to all other sites in the protein. In this way, DCI can reveal the degree of dynamic coupling between *i* and *j*.

Here, we present DCI as a percentile rank of the DCI range observed with values ranging from 0 to 1 (%DCI). Importantly, DFI and DCI are distinct in that DFI measures the flexibility of a position. In contrast, DCI measures the pairwise coupling of one position with another. Furthermore, DCI estimates are conditional on the functional position selected for analysis. Every amino acid position in any given protein has a unique network of direct, local interactions that give rise to a unique network of highly coupled pair positions. Across the protein structure, this gives rise to an inhomogeneous 3D interaction network. Using DCI to explore this network can be insightful when considering active sites, because it is known that even far away positions may disrupt function through the mechanism of allostery [36]. Residues that are distant enough from the active site to likely have no direct interaction (>10 Å) yet are highly coupled to them (%DCI > 60 implying a greater than average response fluctuation when active site residues are perturbed) can play an important role in protein function.

In the example of GCase, around half (52.6%) of the studied pathogenic variants, including *N370S*, occur at DARC sites (Fig 3A). In fact, according to our list of disease mutation sites [38], approximately 28% of DARC sites are associated with GD, compared to ~15% of non-DARC sites throughout the entire protein. Also, the %DCI values of DAV sites are significantly different ($P < .001$) from those of non-disease sites, as seen in Fig 3D. This suggests that variants at DARC sites are more likely to lead to genetic disease. Moreover, a comparison of DCI values of DAV sites with all other protein sites supports the same observation: mutations at DARC sites, distal sites that exhibit high coupling (i.e., high DCI), are predisposed to impact function [24,25]. Such sites may be observed in a variety of different regions and structures across a protein, as seen in Fig 3B.

Using MD simulations, we obtained the dynamic features of 20 DAVs, 2 neutral variants, and the wild-type protein. These variants were chosen because experimental data on their function was also available, by the study of Liou et al. [35]. When comparing DCI profiles of

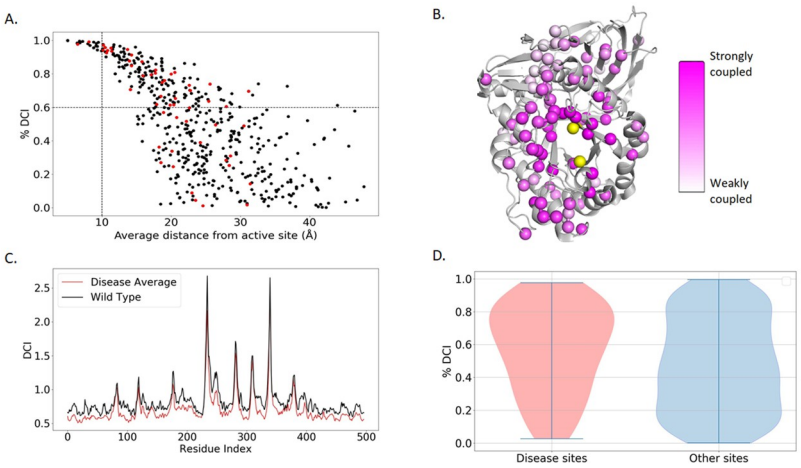

Fig 3. DCI of GCase sites. (**a**) Scatter plot of all GCase residues with dividers at %DCI = 60 and distance = 10 Å. The upper right quadrant contains DARC sites which can affect the active site without direct interaction. Red dots indicate severe DAVs, which have a significantly higher DCI ($P < .001$) than other sites. (**b**) A ribbon diagram showing known mutation sites of GCase (represented as pink-colored dots) and the degree of coupling to the active site delineated by the color gradient, where darker and lighter shades correspond to strongly coupled and weakly coupled, respectively. (**c**) Average DCI profile of 20 different DAVs compared to the wild-type. In general, we observe a global loss of coupling to the active site. (**d**) Violin plots showing that DAVs are generally located at sites that have higher DCI with the active site.

https://doi.org/10.1371/journal.pcbi.1010006.g003

the active site for the wild-type and proteins with DAVs, fluctuations in DCI occur at certain sites, while mostly decreasing in GCase sites with DAVs (Fig 3C). These changes in DCI imply that the long-distance communication pathways cannot follow typical channels to the active site. This communication breakdown is presumed to be a consequence of altered dynamics. Losing rigidity in a functionally critical hinge region impairs the dynamic allosteric residue coupling, leading to a dysfunctional protein [36]. Our data also suggests a link between DCI and the severity of disease mutations. The median %DCI for DAV sites for Gaucher disease marked as "severe" was 69.6%. In comparison, mutations marked as "mild" had a median of 56.6% ($P < .045$). This further supports the idea that positions exhibiting higher dynamic coupling to the active site have a greater impact on protein function.

Principal component analysis of DFI aligns with experimentally determined catalytic activity

As explained above, DFI profiles provide information about the dynamic function of residues throughout the protein. At the same time, DARC sites are coupled with the active site despite having no direct contact. We clustered the DFI values of DARC sites for each simulated GCase variant (Fig 4) using principal component analysis (see Methods). We found that the wild type and neutral variants (functional enzymes based on in-vitro assays) are grouped, and many of the tested proteins creating "dead enzymes" (i.e., total loss of function) are grouped as well. Liou et al. [35] used the specific activity of cross-reacting immunological material (CRIM_SA) values to estimate the catalytic rate constants (k_{cat}), thereby giving experiment-based estimates on the functionality of these variants. The fact that variants with higher CRIM_SA values are clustered together, as are variants with low CRIM_SA values, suggests a direct correlation between DFI profiles and CRIM_SA and, therefore, a direct correlation between DFI and protein function.

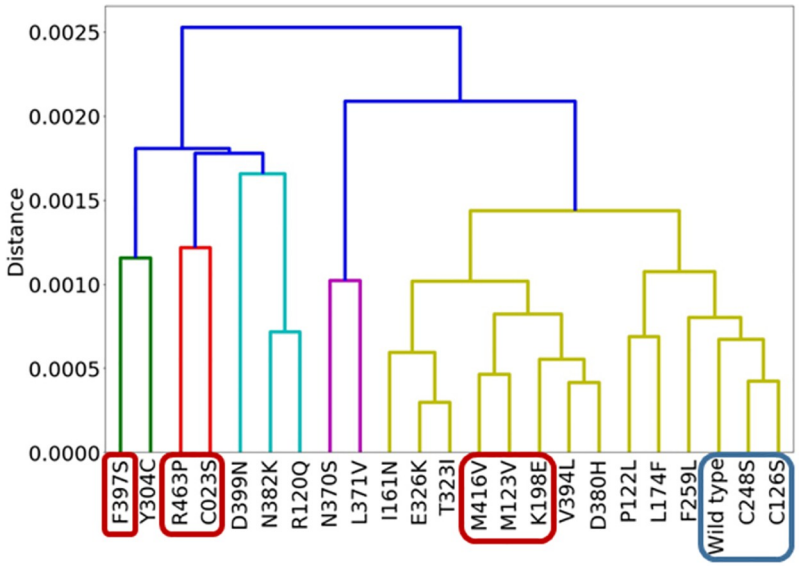

Fig 4. Dendrograph showing clusters of GCase variants based on the DFI of DARC sites. The variants for which experimental data is available show dead enzymes and fully functional (i.e. neutral) enzymes clustered within their own groups. Variants with CRIM_SA values of 0.3 to 1.0 are shown in blue, while variants with CRIM_SA values of 0.06 to 0.1 are shown in red. For other variants shown here, CRIM_SA values are between .1 and .3. These variants have reduced function compared to the wild type, but are still somewhat functional. Higher CRIM_SA values suggest superior enzyme function.

https://doi.org/10.1371/journal.pcbi.1010006.g004

A proteome-wide analysis reveals disease-associated mutations are abundant at DARC sites

After investigating GCase, we used ENM models to expand our study to include 144 human enzymes containing a total of 1024 amino acid variants (433 neutral and 591 DAVs). The ENM is a coarse grained approach and allowed us to study the dynamics of different folds efficiently. This dataset was also used in our previous work [39] incorporating the HumVar data set [40] and sequences with both a high query coverage (>80%) and sequence identity (>80%) selecting only the proteins available in the protein data bank [41]. Additionally, these protein structures had already been modeled, including any missing residues, using the Modeller software package [42].

As illustrated in Fig 5A, the DCI distribution of DAV sites shows a trend opposite to that of sites with neutral variants, exhibiting a significantly different distribution with $P < .001$. Generally, DAVs are more likely to occur at sites highly-coupled to the active site. In contrast, neutral mutations are more likely to occur at sites that are less coupled. Of the variants in this ensemble, 82% occur farther than 10 Å from the active site, suggesting that allosteric communication through 3-D network of interactions modulate the dynamics of the active site, thus impacting the function.

DCI specifically quantifies the coupling between individual positions and, as such, DCI values depend explicitly upon the positions selected for analysis. However, these pairwise interactions are not always symmetric. An interaction network may be formed such that residue perturbations may be felt more strongly in one direction than the other. If we find the difference in the DCI values between two residue positions that are not directly interacting (i.e., in spatial contact), we get a better understanding of the dynamic allostery relationship between two residues. This difference, called DCI asymmetry, provides directionality to long-distance

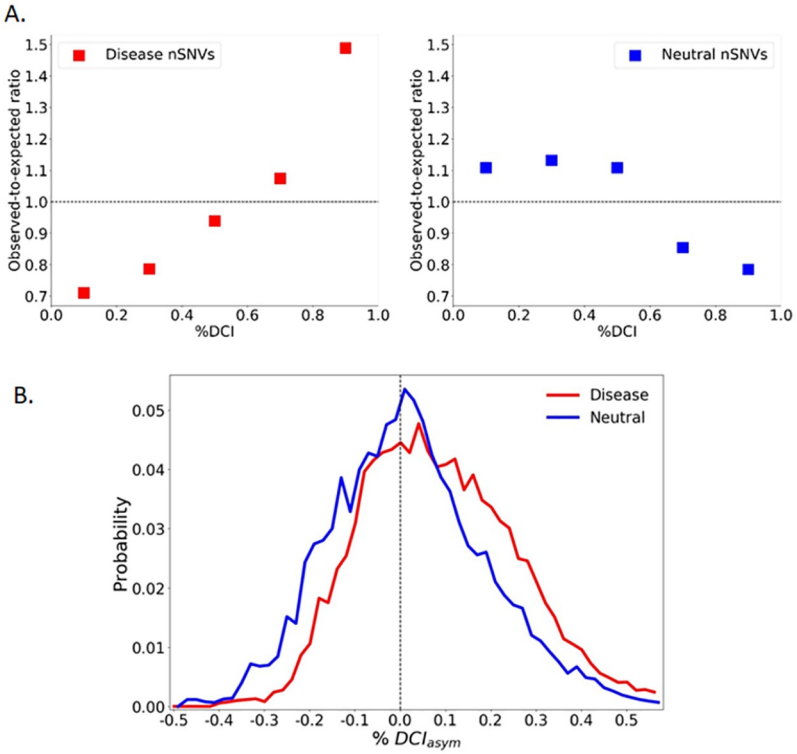

Fig 5. %DCI and asymmetry for 144 protein ensemble. (**a**) Throughout 144 proteins and 1024 variants' sites within those proteins, %DCI values were determined. These distributions were compared with the expected null distribution that %DCI values would be equally distributed over all investigates sites. Observed-to-expected ratios reveal that there are more DAVs than expected having high %DCI, whereas fewer neutral variants than expected are observed in high % DCI categories. Above the ratio equal to 1, the DAV or neutral variants occurs more often than the null expectation. Below the ratio of 1, the mutation does not occur as often as expected. (**b**) Comparison of $\%DCI_{asym}$ of sites associated with neutral variants and DAVs. The distributions show a contrast as DAV sites tend to exhibit more positive values ($P < .001$), suggesting that the active site dominates the coupling. Neutral sites on the other hand tend to give more negative asymmetry values, suggesting that the mutation site dominates. A moving average was used to visually smooth the distribution.

https://doi.org/10.1371/journal.pcbi.1010006.g005

coupling, thereby suggesting a causal relationship In any given protein, every amino acid position has a unique network of direct, local interactions that give rise to a unique network of highly coupled partner positions [9,25,43] and heterogeneity in a 3-D network of interactions. Thus, for a particular pair of coupled amino acids (*i* and *j*), their unique network constraints differentiate the coupling of *i* to *j* from the coupling of *j* to *i*. Thus, we used the wild-type structures of our enzymes to calculate i) $\%DCI_{ij}$, how strongly the position of each mutation is coupled to each active site position, ii) $\%DCI_{ji}$, how strongly each active site position is coupled to the position of each mutation. From these, we calculated iii) "$\%DCI_{asym}$" from ($\%DCI_{ij}$–$\%DCI_{ji}$) to assess the asymmetry in coupling.

Among our protein ensemble, we see a slight pattern emerge, where the interaction between disease mutation sites and active sites is generally more dominated by the active site. In contrast, the interaction between neutral mutation sites and active sites is usually dominated by the mutation site. (Fig 5B). This is indeed in agreement with our earlier findings of LacI variants [9], in which substitutions at sites where functional sites dominate the communication most often end up with a function loss.

A neural network trained on dynamic characteristics offers superior performance at highly evolved sites

Many different methods exist to predict the effect of missense variants on protein function. Some contemporary methods focus on evolutionary considerations alongside structural information to improve the accuracy of predictions [44–47]. As one example, PolyPhen-2 uses solvent accessibility, secondary structure propensities, and crystallographic B-factors to classify mutational sites [44]. Many other approaches consider change in polarity, volume, and charge due to mutant amino acid. A number of phenotypic prediction studies use solvent accessibility, which has proven to be a useful attribute in disease prediction [46]. Other methods utilize residue–residue interaction networks of protein structures to identify functionally important residues through network topology parameters [47,48]. Evolution-based methods generally offer better performance than methods that only use structural features, yet evolution-based methods have true positive rates less than 50% for known DAVs at less-conserved positions [5,15]. In addition, their rate of correct diagnosis of true negative (benign) mutations at highly conserved positions is less than 50% [11].

Like DCI, the DFI of mutation sites can indicate an effect on protein function [36,49]. Previously, our group has shown that DFI can predict pathogenicity of protein interface sites more accurately than the accessible surface area, a commonly used metric [6]. In this study, we extended this comparison to a variety of different metrics, a larger number of missense variants, and by adding DCI in the predictive model. Using DFI, DCI, and asymmetry from our protein ensemble, we trained a neural network to predict whether certain missense variants would be neutral or not (see Methods). When used to predict the pathogenicity of random subsets of our data (90% training, 10% testing; 10-fold cross-validation), this neural network reaches the upper end of performance for established predictive software in the metrics of accuracy, precision, recall, and area under the curve (AUC) evaluated for the receiver operating characteristic (ROC) curve (Fig 6). Of particular interest is the performance of our neural network at highly evolving sites (see Methods). The evolution-based metrics tend to overestimate the rate of neutral mutations at highly-evolving sites [11,15], leading to significantly lower recall scores. However, our dynamics based approach outperforms all the other methods. This is because our method accounts for enigmatic sites—allosteric DARC sites which seem to appear neutral from an evolutionary perspective. We don't expect as many neutral mutations at those sites because we don't use any sequence information related to conservation.

Improvement is also shown over another non-evolutionary metric, Rhapsody [50,51], which is a dynamics based approach. Rhapsody utilizes a Gaussian Network Model (GNM) (a 1-D version of ENM). Thus, the major difference between the GNM based approach and our approach is that we simulate perturbation forces in three dimensions, whereas Rhapsody uses one-dimensional pairwise interactions.

Discussion

Allostery was proposed as an important biophysical mechanism for protein function, which has led some to proclaim that allostery constitutes "the second secret of life," with the genetic code constituting "the first secret of life" [52].

Laboratory-directed evolutionary studies also highlight the emergence of mutations far from the active site [25,53,54]. These distal sites play a critical role in functional evolution, particularly in the emergence of novel functions. Yet, these distal mutation sites challenge enzyme design, as it is difficult to predict them in advance [25,55–57]. Likewise, resurrected ancestral protein studies also reveal that mutations distal from the active site are necessary for functional

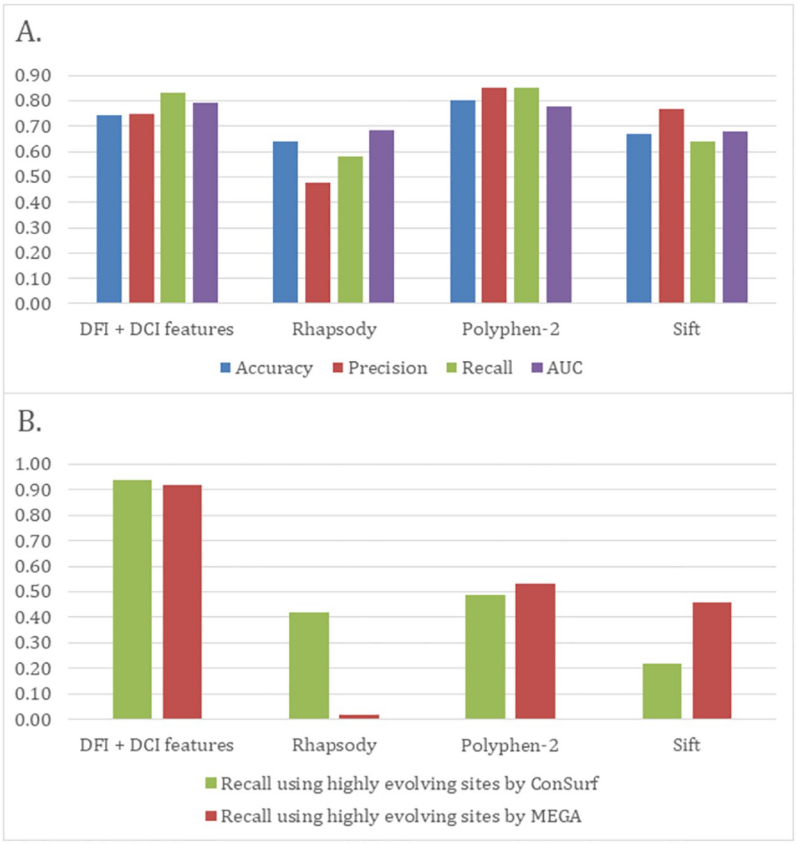

Fig 6. Accuracy of prediction tools tested against 144 enzyme ensemble data. (A) Bar plot showing accuracy, precision, recall, and area-under-the-curve (auc) values for four different methods including our DFI + DCI features. Without using any evolutionary data, our performance matches and may even exceed evolution-based metrics. (B) Recall values for those same metrics tested on fast evolving sites in our data. Higher false negative rates lead to lower recall values for sequence based metrics, but not for DFI + DCI features.

https://doi.org/10.1371/journal.pcbi.1010006.g006

evolution. An example is the emergence of red color from a green ancestor in a close relative of Green Fluorescence Protein (GFP). This protein needs a minimum of 12 mutations and one deletion to convert from green to red color with high efficiency. A majority of the mutations are far from the chromophore. While the flexibility of the mutational sites does not change, in allosteric response to these mutations, both rigidification and increased flexibility occur for regions of the fold widely separated in the 3-D structure of the proteins, accommodating required flexibility for red photoconversion. These synergistic effects allow catalysis to proceed as desired and function without mutations of catalytic residue positions while maintaining fold stability and quaternary structure [19].

Here we also observe that disease-associated (i.e., function altering) mutations follow the same pattern. As neutral variants and DAVs provide the best opportunity to explore the molecular principles of how genetic variations shape phenotypic changes, we observed the same principle of dynamic allostery such that functions become altered through distal mutations while conserving the amino-acid sequence of catalytic residues. We have found that the disruption of the allosteric dynamics with functionally-important sites in a protein is a mechanistic explanation for many missense variants associated with diseases and other biological phenotypes. The patterns of dynamic coupling with the active sites are different for disease and neutral phenotypes for missense mutations that occur at spatially-distant positions to

functional (active) sites. Specific analysis of GCase proteins also provides evidence of the same mechanism observed in resurrected studies. These distal mutations allosterically modify flexibility profiles of different sites, leading to a change in function.

This finding also suggests that rather than affecting only protein stability, the disruption of ligand binding, or both, the allosteric dynamic coupling and stability explain how a large fraction of disease-associated variants impair protein-ligand function or enzymatic activity [6,7,12]. A high-throughput functional assay of over 2,000 variants also show that only a minority of mutations led to a decrease in protein stability [4]. Thus, our findings align with the neutral theory of molecular evolution, as mutations on functionally important catalytic sites must have been eliminated by negative selection due to critical functional loss. On the other hand, the distal mutations remotely fine-tune the native state ensemble to modify function without interfering with folding/folding stability.

We are in the era of rapid development of next-generation methods for whole-genome, whole-exome, and targeted sequencing that has produced an unprecedented amount of data. Among all the genetic variation data, the most commonly observed variants are missense, and identifying the missense variants with pathogenic effects that contribute to disease or drug sensitivities is the primary goal of 21st-century genomic analysis and phylomedicine. As stated in a review of allostery by Liu and Nussinov [52], uniting the genetic code, which constitutes "the first secret of life," and allostery, "the second secret of life," could reveal a generalized disease mechanism and allow for the discovery of novel drugs, as well as blueprints for innovative personalized treatment methods.

Methods

Dataset

A total of 144 individual monomeric protein structures from the Protein Data Bank (PDB) [41] were collected from a BLAST search of sequences with requirements of $\geq$80% sequence identity and $\geq$80% query coverage to ensure only structures that could be accurately mapped to human variation data were included. Human genetic variations were obtained from the HumVar, and HumDiv databases [38] with 1024 amino acid variants, where 433 were neutral and 591 were deleterious.

Determining catalytic sites

The catalytic sites were gathered from the Catalytic Site Atlas (CSA) database [58], which identifies the residues directly involved in catalyzing the reactions of enzymes. Since these residues are critical for protein function, they were used as input into our dynamic coupling index (DCI) metric. The entries in the CSA were either "original entries" derived from the literature itself or "homology entries" based on sequence comparison with the literature-based original entries. In either case, the catalytic sites purported by the CSA should accurately represent functional sites on the protein. Our dataset contained 144 enzymatic proteins that mapped to entries in the CSA database.

Calculating functional-dynamics profiles

Dynamic flexibility index (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 scores (DFIs) do not show 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. Hinges are parts of the protein which are generally rigid. At the same time, they do not exhibit a high fluctuation response to perturbations but transfer these perturbations to the rest of the protein. Like hinges on a door, they stand still, providing an anchor point for other parts to move around.

The method for obtaining the dynamic flexibility index (DFI) is based on the perturbation response scanning (PRS) method [59], in which the C-alpha atom of each residue in the protein is modeled as a node in an elastic network model (ENM). The interaction between each node is modeled by a harmonic potential with a distance-dependent spring constant [59,60]. A small perturbation in the form of an external random force (i.e., Brownian kick) is sequentially applied on each node in the network, and the perturbation response of all nodes is recorded according to linear response theory as

$$[\Delta \mathbf{R}]_{3N\times 1} = [\mathbf{H}]^{-1}_{3N\times 3N}[\mathbf{F}]_{3N\times 1} \tag{1}$$

where **F** is the external random force, $\mathbf{H}^{-1}$ is the inverse Hessian, and **ΔR** is the positional displacement of all *N* nodes in three dimensions.

However, ENM is a coarse-grained model. To improve the accuracy of this model and allow sensitivity to mutations, the hessian inverse can be replaced with the covariance matrices obtained from molecular dynamics simulations.

$$[\Delta \mathbf{R}]_{3N\times 1} = [\mathbf{G}]_{3N\times 3N}[\mathbf{F}]_{3N\times 1} \tag{2}$$

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.

Each perturbation is performed in ten different directions to ensure an isotropic response. The perturbation is repeated for every node in the network, and the positional displacements **ΔR** of each node are stored in a perturbation matrix **A** given by

$$[\mathbf{A}]_{NxN} = \begin{bmatrix} \Delta|R^1|_1 & \Delta|R^2|_1 & \cdots & \Delta|R^N|_1 \\ \Delta|R^1|_2 & \Delta|R^2|_2 & \cdots & \Delta|R^N|_2 \\ \vdots & \vdots & \ddots & \vdots \\ \Delta|R^1|_{N-1} & \Delta|R^2|_{N-1} & \cdots & \Delta|R^N|_{N-1} \\ \Delta|R^1|_N & \Delta|R^2|_N & \cdots & \Delta|R^N|_N \end{bmatrix} \tag{3}$$

where $|\Delta R^j|_i = \sqrt{\langle (\Delta R)^2 \rangle}$ is the magnitude of the positional displacement of each residue *i* in response to a perturbation at residue *j*. The DFI score of residue *i* is defined as the sum of the total displacement of residue *i* induced by a perturbation on all residues, which is computed by taking the sum of the *i*-th row of the perturbation matrix **A**,

$$\mathrm{DFI}_i = \frac{\sum_{j=1}^{N} |\Delta R^j|_i}{\sum_{i=1}^{N}\sum_{j=1}^{N} |\Delta R^j|_i} \tag{4}$$

where the denominator is the total displacement of all residues, used as a normalizing factor. Therefore, the greater the DFI score at position *i*, the more flexible that site will be and the lower the score, the more rigid that site will be, meaning it has less of a response to perturbations throughout the protein. Oftentimes it can be useful to examine the flexibility of certain

residues relative to the flexibility range of that single protein. To do this, DFI values can be ranked on a percentage scale as shown below:

$$\%DFI_i = \frac{n_{\leq i}}{N} \tag{5}$$

where $n_{\leq i}$ is the number of positions having $DFI \leq DFI_i$.

Recently, we have extended this method to identify allosteric links or dynamic coupling between any given residue and functionally important residues by introducing a new metric, the *dynamic coupling index* (DCI) [36]. The DCI metric can identify DARC sites, which are distal to functional sites but control them through dynamic allosteric coupling. This type of allosteric coupling is important; sites with strong dynamic allosteric coupling to functionally critical residues (DARC sites), regardless of separation distance, likely contribute to the function. Thus, a mutation at such a site can disrupt the allosteric dynamic coupling or regulation, leading to functional degradation. As defined, DCI is the ratio of the sum of the mean square fluctuation response of the residue *i* upon functional site j perturbations (i.e., catalytic residues) to the response of residue *i* upon perturbations on all residues. DCI enables us to identify DARC site residues, which are more sensitive to perturbations exerted on residues critical for function. This index can be utilized to determine residues involved in allosteric regulation. It is expressed as

$$\mathrm{DCI}_{ij} = \frac{\sum_{j=1}^{\mathrm{N}_{functional}} |\Delta \mathrm{R}^j|_i / \mathrm{N}_{functional}}{\sum_{j=1}^{\mathrm{N}} |\Delta \mathrm{R}^j|_i / \mathrm{N}} \tag{6}$$

where $|\Delta R^j|_i$ 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 $N_{\mathrm{functional}}$. The denominator is the average mean square fluctuation response over all residues. Just as with DFI, DCI may also be ranked on a percentage scale:

$$\%DCI_{ij} = \frac{m_{\leq i}}{N} \tag{7}$$

where $m_{\leq i}$ is the number of positions having $DCI \leq DCI_{ij}$.

We further investigated the change in dynamics upon mutation compared to the wild type structure using ΔDFI and ΔDCI. The delta-DFI (ΔDFI) profile was calculated as

$$\Delta\mathrm{DFI}_i = \frac{\mathrm{DFI}_{disease} - \mathrm{DFI}_{wt}}{\mathrm{DFI}_{wt}} \tag{8}$$

Where $\mathrm{DFI}_{disease}$ is the dynamics profile for the mutated protein structure and DFI_{wt} is the dynamics profile for the wild-type structure. Similarly, the delta-DCI (ΔDCI) profile was calculated as

$$\Delta\mathrm{DCI}_i = \frac{\mathrm{DCI}_{disease} - \mathrm{DCI}_{wt}}{\mathrm{DCI}_{wt}} \tag{9}$$

One additional tool we use is DCI asymmetry, which measures preferential information transfer through asymmetric dynamic coupling. Simply put, the coupling asymmetry between

positions i and j can be calculated as

$$DCI_{asym} = DCI_{ij} - DCI_{ji} \quad (10)$$

$$\%DCI_{asym} = \%DCI_{ij} - \%DCI_{ji} \quad (11)$$

Where DCI_{ij} represents the relative response of residue i to a perturbation at residue j and DCI_{ji} represents the relative response of residue j to a perturbation at residue i.

It should be once again made clear that all dynamic analysis of the GCase protein was conducted using data from MD simulations only, while analysis of the 144 enzyme ensemble was performed using data from ENM simulations only.

Molecular dynamics simulations

To compute the DFI and DCI profiles of each missense variant of GCase, we first performed MD simulations to obtain the native ensemble of each variant and then applied our analysis. The starting structure for GCase was taken from the Protein Data Bank (accession number 1ogs [34]). The mutagenesis tool was used in Pymol [61] to create variant structures. Next, we loaded structures into TLEAP using the ff14SB force field [62]. We then added protein hydrogens were and a 14.0 Å cubic box of TIP3P surrounding water atoms, followed by Na^+ and Cl^- atoms for neutralization [63]. Then all systems were energy-minimized using the SANDER module of AMBER 14 [64,65]. First, the protein was kept fixed with harmonic restraints to allow surrounding water molecules and ions to relax, followed by a second minimization step in which the restraints were removed and the protein-solution was further minimized. Both minimization steps employ the method of steepest descent followed by conjugate gradient.

We then ran heating, density equilibration and production using the GPU-accelerated PMEMD module of AMBER 14 [65]. Periodic boundary conditions were used in all simulations, and the bond lengths of all covalent hydrogen bonds were constrained using SHAKE [64]. Direct-sum, non-bonded interactions were cut off at distances of 9.0 Å or greater, and long-range electrostatic interactions were calculated using the particle mesh Ewald method [66,67]. During the heating cycle, we heated systems from 0K to 300K over a duration of 250 ps. The density of the system was then allowed to equilibrate over 5 ns at constant temperature and pressure. A Langevin thermostat was used to control the temperature at 300 K and a Berendsen barostat to adjust the pressure at 1 bar. We used a timestep of 2 fs and saved structural conformations every 10 ps. All simulations were allowed to progress to 1 μs of total simulation time, deemed the minimal required simulation time for convergence based on earlier studies [24,68].

In order to calculate DFI and DCI, we calculated covariance matrices using 50 ns moving windows that overlap by 25 ns over the last 500 ns of the trajectory of each simulation. In order to ensure ergodicity where the DFI and DCI profiles present the equilibrium dynamics, there are two of the basic conditions that need to be met: (i) All conformations must be sampled from the same distribution. (ii) The time windows and subsequent covariance matrices obtained ought to be independent of the initial atomic coordinates in order to eliminate global motions and accurately capture equilibrium coordinate information. Because of this, the final average DFI profiles will be independent of the window size; meaning that the averaging of DFI profiles from different time window sizes (i.e. 50 ns vs 75 ns) will give similar results and the calculated covariance matrices extracted from different times of trajectories should also result in similar DFI profiles, such as seen in S2 Fig.

Clustering the DFI values of DARC sites

We clustered the DFI profiles of DARC sites for various mutated GCase proteins by comparing their percentile rankings. To compare the flexibility profiles, the proteins are concatenated into a data matrix $\boldsymbol{X}$. The statistical procedure Singular value decomposition (SVD) is used to factorize the data into the orthonormal basis, which is a representation of the vector space containing data. It is similar to principal component analysis which may be used to assist in understanding the structure of data or to increase the signal-to-noise ratio in data by eliminating the redundant dimensions and mapping it on a lower-dimensional space. Clustering by SVD acts as an effective noise filter by isolating the highest variances among data points in the top principal vectors. Consequently, the remaining insignificant singular vectors can be omitted from the reconstruction.

The DFI profiles of all proteins are merged into a matrix $\boldsymbol{X}$, of dimensions ($\boldsymbol{m} \times \boldsymbol{n}$). Here $\boldsymbol{m}$ is the number of datasets (protein variants) we are clustering together, each having $\boldsymbol{n}$ number of attributes (n = number of DARC sites, thus each element in a given column presents the DFI value of specific DARC site of a given variant). On performing SVD, $\boldsymbol{X}$ is decomposed as follows:

$$[\boldsymbol{X}]_{m\times n} = [\mathbf{U}]_{m\times m}[\zeta]_{m\times n}[\boldsymbol{V}]_{n\times n} \tag{12}$$

Here, $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices with orthonormal columns and are called left singular vectors and right singular vectors, respectively, and ζ is a diagonal matrix with diagonal elements known as the singular values of $\boldsymbol{X}$.

The singular values of $\boldsymbol{X}$, by convention, are arranged in a decreasing order of their magnitude; $\sigma = \{\sigma_i\}$ represent the variances in the corresponding left and right singular vectors. The set of highest singular values representing the largest variance in the orthonormal singular vectors can be interpreted to show the characteristics in the data $\boldsymbol{X}$ and the right singular vectors create the orthonormal basis which spans the vector space representing the data. The left singular vectors contain weights indicating the significance of each attribute in the dataset as $w_i = \sum_{k=1}^{r} \sigma_k |u_{ik}|$. Using these features of the decomposed singular vectors, we can create another matrix, $\boldsymbol{X}^*$ using only the highest 'r' singular values which can mimic the basic characteristics of the original dataset. Thus, $\boldsymbol{X}^*$ can be represented as

$$[\boldsymbol{X}^*]_{m\times r} = [\mathbf{V}^*]_{m\times r}[\zeta^*]_{r\times r} \tag{13}$$

Here, ζ^* contains only largest $\boldsymbol{r}$ singular values and $\mathbf{V}^*$ contains the corresponding right singular vectors. The data are now clustered hierarchically based on the pairwise distance between different protein variants in the reconstructed DFI data with reduced dimensions.

For a pair of datasets (or between flexibility profiles of any two proteins) j_1 and j_2, the distance between them in the original set of data was given by

$$d_{12} = \sqrt{\sum_{i=1}^{n} (\boldsymbol{X}_i^{j1} - \boldsymbol{X}_i^{j2})^2} \tag{14}$$

which in reduced dimensions can be calculated as

$$d_{12} = \sqrt{\sum_{i=1}^{r} (\boldsymbol{X}^{*j1}_i - \boldsymbol{X}^{*j2}_i)^2} \tag{15}$$

These pairwise distances are used as the parameters for clustering the flexibility profiles of GCase. The DFI values of DARC sites are aligned and clubbed into a dataset matrix $\boldsymbol{X}$. The three largest singular values are used for reconstruction of data and clustering. The pairwise distance between each protein using the equation above is used for clustering them hierarchically.

A bottom-up approach is used for the hierarchical clustering, where initially each protein variant is assigned its own cluster and then, in successive iteration, closest clusters are merged together into a common cluster. In this approach, the distance between clusters is defined by the average pairwise distance between their components (average linkage clustering [69]). In the end, the clusters are represented hierarchically using a dendrogram, where the vertical axis denotes the Euclidean distance between various clusters and among their sub-clusters.

Neural network

In an attempt to enhance our prediction accuracy based on protein dynamics we integrated a Neural Network based training and prediction algorithm. With an increased number of dimensions in data space, regular regression methods fall behind machine learning strategies and artificial Neural Networks. Our data contains multiple dynamics driven metrics emerging from per position specific DFI of the position with observed variant and also DFI of the neighborhood positions as well as DCI. These metrics by themselves display strong correlation (Fig 5) [49], but proteins are dynamic systems meaning per residue dynamics cannot grant all the relative information about the global dynamics. Therefore, with the inclusion of several distinct metrics that represent different dynamical features of the proteins, we exploited an Artificial Neural Net based prediction approach.

The feed forward Neural Network architecture deployed in this paper utilizes a single input layer with multiple features and a binary classification model. The features include: DFI_i, $\%DCI_{ji}$, DCI_{asym}, and the average DFI of residues within 7Å. We use residues within 7Å because they have direct interactions with the variant site. These four features along with corresponding sites and ground truth values may be found in the Supplementary Materials as S1 Data. The network includes two hidden layers with 80 nodes each between the input and the output layer. The hidden layers are connected with a 50% dropout scheme to eliminate overfitting. The initial node weights and biases of the network are sampled from a uniform distribution with Rectified Linear Units as the activation function to reach better convergence compared to a sigmoid function. The output layer has initial uniform node weights and biases sampled from Xavier uniform initializer and a sigmoid activation function with binary label output. The optimization algorithm utilizes a stochastic gradient descent with built-in momentum to minimize the cross-entropy loss function. The built-in momentum helps to escape saddle points and reach a global minimum loss. The learning rate for the optimizer is set as 0.001 with 1000 epochs in total for the Neural Network to converge. The Neural Network is trained with 90% of randomly selected data points and tested by the remaining 10%. This process is repeated 10 times to gather improved statistics and eliminate any bias coming from the data itself. Employing a 10-fold training/testing algorithm provides a distribution of accuracies instead a single accuracy.

The evaluation metrics AUC, accuracy, precision and recall are used to evaluate the predictive power of the classification model by comparing prediction values with the ground truth values. The four possible outcomes from the binary classifier are: True positive (TP), true negative (TN), false positive (FP), and false negative (FN). Accuracy, precision, and recall equations for calculation are denoted below:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

AUC is obtained by calculating the area under the Receiver Operating Characteristic (ROC) curve, which is generated using true positive rate and false positive rates.

In order to determine highly evolving sites, we utilized two different methods: (i) the ConSurf database to evaluate conservation of each site, [70,71], and (ii) Molecular Evolutionary Genetics Analysis (MEGA) software [72] to calculate evolutionary rates as described by Kumar et al. [11].

Supporting information

S1 Fig. Stick diagrams of loop 1 which are colored corresponding to their DFI for 19 different disease variants. DFI here is a color code within a spectrum of red-white-blue where red shows the highest, and blue shows the lowest flexible sites.
(TIF)

S2 Fig. %DFI profiles averaged over different time scales demonstrate convergence. Black: average %DFI values calculated using covariance matrix data over 400ns to 600ns of the wild type GCase simulation. Blue: average %DFI values calculated using covariance matrix data over 600ns to 800ns of the wild type GCase simulation. Red: average %DFI values calculated using covariance matrix data over 800ns to 1ms of the wild type GCase simulation. All profiles use 50 ns moving windows that overlap by 25 ns when calculating average %DFI.
(TIF)

S3 Fig. Accuracy of prediction tools tested against highly evolving sites in 144 enzyme ensemble data. Bar plots showing accuracy, precision, recall, and area-under-the-curve (auc) values for four different methods including our DFI + DCI features. (A) The prediction methods were evaluated using only fast evolving sits according to ConSurf. (B) The prediction methods were evaluated using only fast evolving sits according to MEGA. Note that for our highly evolving subset, rhapsody returned 0 true positive and 0 false negative values, causing AUC, precision, and recall to be either zero or incalculable. Using either set of highly evolving sites, we are slightly better in AUC and comparable in precision. However, the dynamics based classifier have slightly lower values for accuracy owing to higher false positive rates.
(TIF)

S1 Table. Enzyme ensemble information. For each missense variant used in our analysis, shows the PDB identification code, mutation site, pathogenicity (disease or neutral), and active sites. Sites are aligned to the associated PDB file.
(CSV)

S2 Table. Shows the mutation site of each GCase DAV used in our analysis. Asterisk denotes sites with multiple DAVs reported. Sites are aligned to PDB ID: 1ogs [34].
(CSV)

S1 Data. Input data for our neural network (see Methods). *dfi_i*, *%dci_ji*, *dci asymmetry*, and *average dfi within 7Å* columns contain input layer features and the *disease(1) or neutral(0)* column contains ground truth values for pathogenicity. Columns *pdb id* and *pdb residue index* exist for identification purposes.
(CSV)

S1 Files. Contains input files for MD simulations of GCase variants.
(RAR)

Author Contributions

Conceptualization: Nicholas J. Ose, Brandon M. Butler, Avishek Kumar, Sudhir Kumar, S. Banu Ozkan.

Data curation: Nicholas J. Ose, Brandon M. Butler, Avishek Kumar, I. Can Kazan, Maxwell Sanderford, Sudhir Kumar.

Formal analysis: Nicholas J. Ose, I. Can Kazan, Maxwell Sanderford, Sudhir Kumar, S. Banu Ozkan.

Funding acquisition: Sudhir Kumar, S. Banu Ozkan.

Investigation: Nicholas J. Ose, Brandon M. Butler, Avishek Kumar, Maxwell Sanderford, Sudhir Kumar.

Methodology: Nicholas J. Ose, Brandon M. Butler, Avishek Kumar, I. Can Kazan, Sudhir Kumar, S. Banu Ozkan.

Project administration: S. Banu Ozkan.

Resources: S. Banu Ozkan.

Software: Nicholas J. Ose, Avishek Kumar, I. Can Kazan.

Supervision: Sudhir Kumar, S. Banu Ozkan.

Visualization: Brandon M. Butler, Avishek Kumar, Sudhir Kumar, S. Banu Ozkan.

Writing – original draft: Nicholas J. Ose, Brandon M. Butler, Avishek Kumar, Sudhir Kumar, S. Banu Ozkan.

Writing – review & editing: Nicholas J. Ose, Brandon M. Butler, Avishek Kumar, I. Can Kazan, Sudhir Kumar, S. Banu Ozkan.

References

1. Alber T. Mutational effects on protein stability. Annu Rev Biochem. 1989; 58: 765–798. https://doi.org/10.1146/annurev.bi.58.070189.004001 PMID: 2673021
2. Guerois R., Nielsen J.E., Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. J Mol Biol. 2002; 320: 369–387. https://doi.org/10.1016/S0022-2836(02)00442-4 PMID: 12079393
3. Yue P., Li Z., Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005; 353: 459–473. https://doi.org/10.1016/j.jmb.2005.08.020 PMID: 16169011
4. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, et al. Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. Cell. 2015; 161: 647–660. https://doi.org/10.1016/j.cell.2015.04.013 PMID: 25910212
5. Kumar S., Dudley J., Filipski A., Liu L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends Genet. 2011; 27. https://doi.org/10.1016/j.tig.2011.06.004 PMID: 21764165
6. Butler B.M., Gerek Z.N., Kumar S., Ozkan S.B. Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association. Proteins. 2015; 83: 428–435. https://doi.org/10.1002/prot.24748 PMID: 25546381
7. Wang X., Wei X., Thijssen B., Das J., Lipkin S.M., Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012; 30: 159–164. https://doi.org/10.1038/nbt.2106 PMID: 22252508

8. Krishnaswamy S, Kanteti R, Duke-Cohan JS, Loganathan S, Liu W, Ma PC, et al. Ethnic Differences and Functional Analysis of MET Mutations in Lung Cancer. Clin Cancer Res. 2009; 15: 5714–5723. https://doi.org/10.1158/1078-0432.CCR-09-0070 PMID: 19723643
9. Campitelli P, Swint-Kruse L, Ozkan SB. Substitutions at Nonconserved Rheostat Positions Modulate Function by Rewiring Long-Range, Dynamic Interactions. Wilke C, editor. Mol Biol Evol. 2020b; 38: 201–214. https://doi.org/10.1093/molbev/msaa202 PMID: 32780837
10. Glembo T., Thorpe M., Farrell D., Gerek Z., Ozkan S. Collective Dynamics Differentiates Functional Divergence in Protein Evolution. PLoS Comput Biol. 2012; 8. https://doi.org/10.1371/journal.pcbi.1002428 PMID: 22479170
11. Kumar S., Suleski M.P., Markov G.J., Lawrence S., Marco A., Filipski A.J. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome Res. 2009; 19: 1562–9. https://doi.org/10.1101/gr.091991.109 PMID: 19546171
12. Kumar A., Butler B.M., Kumar S., Ozkan S.B. Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine. Curr Opin Struct Biol. 2015a; 35: 135–142. https://doi.org/http%3A//dx.doi.org/10.1016/j.sbi.2015.11.002 PMID: 26684487
13. Nussinov R., Tsai C.-J. Allostery in disease and in drug discovery. Cell. 2013; 153: 293–305. https://doi.org/10.1016/j.cell.2013.03.034 PMID: 23582321
14. Swint-Kruse L. Using Evolution to Guide Protein Engineering: The Devil IS in the Details. Biophys J. 2016; 111: 10–18. https://doi.org/10.1016/j.bpj.2016.05.030 PMID: 27410729
15. Kumar S, Sanderford M, Gray VE, Ye J, Liu L. Evolutionary diagnosis method for variants in personal exomes. Nat Methods. 2012; 9: 855–856. https://doi.org/10.1038/nmeth.2147 PMID: 22936163
16. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update: HGMD 2003 UPDATE. Hum Mutat. 2003; 21: 577–581. https://doi.org/10.1002/humu.10212 PMID: 12754702
17. Bhabha G., Ekiert D.C., Jennewein M., Zmasek C.M., Tuttle L.M., Kroon G., Dyson, et al. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat Struct Mol Biol. 2013; 20: 1243–1249. https://doi.org/10.1038/nsmb.2676 PMID: 24077226
18. Campbell E., Kaltenbach M., Correy G., Carr P., Porebski B.T., Livingstone E., et al. The role of protein dynamics in the evolution of new enzyme function. Nat Chem Biol. 2016. https://doi.org/10.1038/nchembio.2175 PMID: 27618189
19. Kim H., Zou T., Modi C., Dörner K., Grunkemeyer T.J., Chen L., et al. A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in GFP-like proteins. Structure. 2015; 23: 34–43. https://doi.org/10.1016/j.str.2014.11.011 PMID: 25565105
20. Zou T., Risso V.A., Gavira J.A., Sanchez-Ruiz J.M., Ozkan S.B. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. Mol Biol Evol. 2015; 32: 132–143. https://doi.org/10.1093/molbev/msu281 PMID: 25312912
21. Modi T, Huihui J, Ghosh K, Ozkan SB. Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. Phil Trans R Soc B. 2018; 373: 10. https://doi.org/10.1098/rstb.2017.0184 PMID: 29735738
22. Villy Isaksen G., Åqvist J., Brandsdal B.O. Enzyme surface rigidity tunes the temperature dependence of catalytic rates. Proc Natl Acad Sci. 2016; 113. https://doi.org/10.1073/pnas.1605237113 PMID: 27354533
23. Romero R, Ramanathan A, Yuen T, Bhowmik D, Mathew M, Munshi LB, et al. Mechanism of glucocerebrosidase activation and dysfunction in Gaucher disease unraveled by molecular dynamics and deep learning. Proc Natl Acad Sci. 2019; 116: 5086–5095. https://doi.org/10.1073/pnas.1818411116 PMID: 30808805
24. Modi T, Risso VA, Martinez-Rodriguez S, Gavira JA, Mebrat MD, Van Horn WD, et al. Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity. Nat Commun. 2021; 12: 1852. https://doi.org/10.1038/s41467-021-22089-0 PMID: 33767175
25. Campitelli P, Modi T, Kumar S, Ozkan SB. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. Annu Rev Biophys. 2020a; 49: 267–288. https://doi.org/10.1146/annurev-biophys-052118-115517 PMID: 32075411
26. Lieberman R.L. A Guided Tour of the Structural Biology of Gaucher Disease: Acid- β -Glucosidase and Saposin C. Enzyme Res. 2011; 2011: 1–15. https://doi.org/10.4061/2011/973231 PMID: 22145077
27. Aharon-Peretz J, Rosenbaum H, Gershoni-Baruch R. Mutations in the Glucocerebrosidase Gene and Parkinson's Disease in Ashkenazi Jews. N Engl J Med. 2004; 351: 1972–1977. https://doi.org/10.1056/NEJMoa033277 PMID: 15525722

28. Do J, McKinney C, Sharma P, Sidransky E. Glucocerebrosidase and its relevance to Parkinson disease. Mol Neurodegener. 2019; 14: 36. https://doi.org/10.1186/s13024-019-0336-2 PMID: 31464647

29. Clark LN, Ross BM, Wang Y, Mejia-Santana H, Harris J, Louis ED, et al. Mutations in the glucocerebrosidase gene are associated with early-onset Parkinson disease. Neurology. 2007; 69: 1270–1277. https://doi.org/10.1212/01.wnl.0000276989.17578.02 PMID: 17875915

30. Velayati A, Yu WH, Sidransky E. The Role of Glucocerebrosidase Mutations in Parkinson Disease and Lewy Body Disorders. Curr Neurol Neurosci Rep. 2010; 10: 190–198. https://doi.org/10.1007/s11910-010-0102-x PMID: 20425034

31. Clark LN, Kartsaklis LA, Wolf Gilbert R, Dorado B, Ross BM, Kisselev S, et al. Association of Glucocerebrosidase Mutations With Dementia With Lewy Bodies. Arch Neurol. 2009; 66. https://doi.org/10.1001/archneurol.2009.54 PMID: 19433657

32. Lieberman R.L., Wustman B.A., Huertas P., Powe A.C., Pine C.W., Khanna R., et al. Structure of acid β-glucosidase with pharmacological chaperone provides insight into Gaucher disease. Nat Chem Biol. 2007; 3: 101–107. https://doi.org/10.1038/nchembio850 PMID: 17187079

33. Hruska K.S., LaMarca M.E., Scott C.R., Sidransky E. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Hum Mutat. 2008; 29: 567–583. https://doi.org/10.1002/humu.20676 PMID: 18338393

34. Dvir H, Harel M, McCarthy AA, Toker L, Silman I, Futerman AH, et al. X-ray structure of human acid-β-glucosidase, the defective enzyme in Gaucher disease. EMBO Rep. 2003; 4: 704–709. https://doi.org/10.1038/sj.embor.embor873 PMID: 12792654

35. Liou B, Kazimierczuk A, Zhang M, Scott CR, Hegde RS, Grabowski GA. Analyses of Variant Acid β-Glucosidases. J Biol Chem. 2006; 281: 4242–4253.

36. Kumar A., Glembo T.J., Ozkan S.B. The Role of Conformational Dynamics and Allostery in the Disease Development of Human Ferritin. Biophys J. 2015b; 109: 1273–1281. https://doi.org/10.1016/j.bpj.2015.06.060 PMID: 26255589

37. Li Z., Bolia A., Maxwell J.D., Bobkov A.A., Ghirlanda G., Ozkan S.B., et al. A Rigid Hinge Region Is Necessary for High-Affinity Binding of Dimannose to Cyanovirin and Associated Constructs. Biochemistry. 2015; 54: 6951–6960. https://doi.org/10.1021/acs.biochem.5b00635 PMID: 26507789

38. Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7: 248–249. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512

39. Butler BM, Kazan IC, Kumar A, Ozkan SB. Coevolving residues inform protein dynamics profiles and disease susceptibility of nSNVs. Jernigan RL, editor. PLOS Comput Biol. 2018; 14: e1006626. https://doi.org/10.1371/journal.pcbi.1006626 PMID: 30496278

40. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006; 22: 2729–2734. https://doi.org/10.1093/bioinformatics/btl423 PMID: 16895930

41. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28: 235–242. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235

42. Eswar Narayanan, Webb Ben, Marti-Renom Marc A., Madhusudhan M.S., Eramian David, Shen Minyi, et al. Comparative Protein Structure Modeling Using Modeller. Curr Protoc Bioinforma. 2014; 5.6.1–5.6.30. https://doi.org/https%3A//doi.org/10.1002/0471250953.bi0506s15 PMID: 18428767

43. Campitelli P, Ozkan SB. Allostery and Epistasis: Emergent Properties of Anisotropic Networks. Entropy. 2020; 22: 667. https://doi.org/10.3390/e22060667 PMID: 33286439

44. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet. 2013; 76. https://doi.org/10.1002/0471142905.hg0720s76 PMID: 23315928

45. Espinosa O, Mitsopoulos K, Hakas J, Pearl F, Zvelebil M. Deriving a Mutation Index of Carcinogenicity Using Protein Structure and Protein Interfaces. Tramontano A, editor. PLoS ONE. 2014; 9: e84598. https://doi.org/10.1371/journal.pone.0084598 PMID: 24454733

46. Wei Q, Xu Q, Dunbrack RL. Prediction of phenotypes of missense mutations in human proteins from biological assemblies: Missense Mutations and Biological Assemblies. Proteins Struct Funct Bioinforma. 2013; 81: 199–213. https://doi.org/10.1002/prot.24176 PMID: 22965855

47. Ye Z-Q, Zhao S-Q, Gao G, Liu X-Q, Langlois RE, Lu H, et al. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics. 2007; 23: 1444–1450. https://doi.org/10.1093/bioinformatics/btm119 PMID: 17384424

48. Cheng TMK, Lu Y-E, Vendruscolo M, Lio' P, Blundell TL. Prediction by Graph Theoretic Measures of Structural Effects in Proteins Arising from Non-Synonymous Single Nucleotide Polymorphisms.

Nussinov R, editor. PLoS Comput Biol. 2008; 4: e1000135. https://doi.org/10.1371/journal.pcbi.1000135 PMID: 18654622

49. Nevin Gerek Z, Kumar S, Banu Ozkan S. Structural dynamics flexibility informs function and evolution at a proteome scale. Evol Appl. 2013; 6: 423–433. https://doi.org/10.1111/eva.12052 PMID: 23745135

50. Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci. 2018; 115: 4164–4169. https://doi.org/10.1073/pnas.1715896115 PMID: 29610305

51. Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Ponty Y, editor. Bioinformatics. 2020; 36: 3084–3092. https://doi.org/10.1093/bioinformatics/btaa127 PMID: 32101277

52. Liu J., Nussinov R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. PLOS Comput Biol. 2016; 12. https://doi.org/10.1371/journal.pcbi.1004966 PMID: 27253437

53. Chen K, Arnold FH. Engineering new catalytic activities in enzymes. Nat Catal. 2020; 3: 203–213. https://doi.org/10.1038/s41929-019-0385-5

54. Wilding M, Hong N, Spence M, Buckle AM, Jackson CJ. Protein engineering: the potential of remote mutations. Biochem Soc Trans. 2019; 47: 701–711. https://doi.org/10.1042/BST20180614 PMID: 30902926

55. Jiménez-Osés G, Osuna S, Gao X, Sawaya MR, Gilson L, Collier SJ, et al. The role of distant mutations and allosteric regulation on LovD active site dynamics. Nat Chem Biol. 2014; 10: 431–436. https://doi.org/10.1038/nchembio.1503 PMID: 24727900

56. Saavedra HG, Wrabl JO, Anderson JA, Li J, Hilser VJ. Dynamic allostery can drive cold adaptation in enzymes. Nature. 2018; 558: 324–328. https://doi.org/10.1038/s41586-018-0183-2 PMID: 29875414

57. Modi T, Ozkan S. Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 β-lactamase. Int J Mol Sci. 2018; 19: 3808. https://doi.org/10.3390/ijms19123808 PMID: 30501088

58. Furnham N., Holliday G.L., de Beer T.A.P., Jacobsen J.O.B., Pearson W.R., Thornton J.M. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res. 2014; 42: D485–D489. https://doi.org/10.1093/nar/gkt1243 PMID: 24319146

59. Atilgan C., Atilgan A.R. Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. PLoS Comput Biol. 2009; 5. https://doi.org/10.1371/journal.pcbi.1000544 PMID: 19851447

60. Atilgan A.R., Durell S.R., Jernigan R.L., Demirel M.C., Keskin O., Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 2001; 80: 505–515. https://doi.org/10.1016/S0006-3495(01)76033-X PMID: 11159421

61. Schrodinger. The PyMOL Molecular Graphics System, Version 2.0.4. 2015.

62. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015; 11: 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255 PMID: 26574453

63. Sun Y., Kollman P.A. Hydrophobic solvation of methane and nonbond parameters of the TIP3P water model. J Comput Chem. 1995; 16: 1164–1169. https://doi.org/10.1002/jcc.540160910

64. Pearlman D.A., Case D.A., Caldwell J.W., Ross W.S., Cheatham T.E. III, DeBolt S., et al. 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. 1995; 91: 1–41.

65. Salomon-Ferrer R., Götz A.W., Poole D., Le Grand S., Walker R.C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput. 2013; 9: 3878–3888. https://doi.org/10.1021/ct400314y PMID: 26592383

66. Darden T, York D, Pedersen L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. J Chem Phys. 1993; 98: 10089–10092. https://doi.org/10.1063/1.464397

67. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys. 1995; 103: 8577–8593. https://doi.org/10.1063/1.470117

68. Sawle L, Ghosh K. Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma. J Chem Theory Comput. 2016; 12: 861–869. https://doi.org/10.1021/acs.jctc.5b00999 PMID: 26765584

69. Day WHE, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. J Classif. 1984; 1: 7–24. https://doi.org/10.1007/BF01890115

70. Ben Chorin A, Masrati G, Kessel A, Narunsky A, Sprinzak J, Lahav S, et al. ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. Protein Sci. 2020; 29: 258–267. https://doi.org/10.1002/pro.3779 PMID: 31702846

71. Goldenberg O, Erez E, Nimrod G, Ben-Tal N. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res. 2009; 37: D323–D327. https://doi.org/10.1093/nar/gkn822 PMID: 18971256

72. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Battistuzzi FU, editor. Mol Biol Evol. 2018; 35: 1547–1549. https://doi.org/10.1093/molbev/msy096 PMID: 29722887