Neural relational inference to learn allosteric long-range interactions in proteins from molecular dynamics simulations

Jingxuan Zhu
Jilin University
Juexin Wang
University of Missouri https://orcid.org/0000-0002-2260-4310
Weiwei Han
Jilin University
Dong Xu (xudong@missouri.edu)
University of Missouri - Columbia https://orcid.org/0000-0002-4809-0514

Article

Keywords: Protein Allostery, Long-range Intra-protein Communication, Graph Neural Network, Encoder-decoder Architecture, Latent Interactions

Posted Date: February 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-219933/v1

Version of Record: A version of this preprint was published at Nature Communications on March 29th, 2022. See the published version at https://doi.org/10.1038/s41467-022-29331-3.

Neural relational inference to learn allosteric long-range interactions in proteins from molecular dynamics simulations

Jingxuan Zhu[1,2], Juexin Wang[2], Weiwei Han[1*], Dong Xu[2*]

1-Key Laboratory for Molecular Enzymology and Engineering of Ministry of Education, School of Life Science, Jilin University, Changchun, China

2-Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO, United States

* Corresponding Authors

Abstract

Protein allostery is a biological process facilitated by spatially long-range intra-protein communication, whereby ligand binding or amino acid mutation at a distant site affects the active site remotely. Molecular dynamics (MD) simulation provides a powerful computational approach to probe the allosteric effect. However, current MD simulations cannot reach the time scales of whole allosteric processes. The advent of deep learning made it possible to evaluate both spatially short and long-range communications for understanding allostery. For this purpose, we applied a neural relational inference (NRI) model based on a graph neural network (GNN), which adopts an encoder-decoder architecture to simultaneously infer latent interactions to probe protein allosteric processes as dynamic networks of interacting residues. From the MD trajectories, this model successfully learned the long-range interactions and pathways that can mediate the allosteric communications between the two distant sites in the Pin1, SOD1, and MEK1 systems.

Introduction

Many protein functions are regulated by specific dynamic biomolecular processes, such as allostery, protein folding/unfolding, and protein activation. A biomolecular motion can be considered a dynamic system driven by atomic/residue interactions. Molecular dynamics (MD) simulations can directly probe the biomolecular motion but may fail to capture meaningful functional information due to the limited time scale of simulation, as well as the high dimensionality and complexity of 3D trajectory data. To extract the biological information from the massive data, statistical techniques such as principal component analysis (PCA)[1] and cross-correlation analysis[2] have been applied in MD analyses. PCA reduces the data dimension while maintaining essential information, making the biomolecular motion more interpretable and allowing for visualization. Cross-correlation analysis assesses the extent to which the atomic/residue fluctuations are correlated with one another by examining the magnitude of all pairwise cross-correlation coefficients. However, these methods are inherently restricted to the linear relationships between data features[2, 3], and therefore miss the nonlinear correlation in dynamics closely related to long-range communication in protein[4]. As a result, many challenging MD analysis problems lack good methods to probe long-range communications and nonlinear effects. For example, allosteric communication is commonly understood in proteins, but understanding how signals are transmitted over long distances within a protein or across different protein molecules has been a challenging problem for a long time.

To model protein allosteric communication, many graph models have been developed. In general, a protein can be mapped to a graph, in which each node represents a residue, and each weighted edge represents an interaction between two nodes. The shortest paths between the allosteric site and active sites in a protein may be important for propagating signals in the allosteric communication. Earlier graph models used a static crystal structure to calculate the shortest paths between one residue and other residues[5]. Later, information from MD simulations was used to define the interaction graph and shortest path but was not sufficient to characterize allostery[6].

The advent of deep learning has provided new opportunities to explore allosteric effects. The emerging graph neural network (GNN) is designed to model data systems in graphs, and it has facilitated great success in solving many graph-related problems[7, 8]. Recently, the GNN helped fulfill long-term research goals in modeling complex dynamic systems in traffic scenes, dynamic

physical systems, and computer vision tasks by using implicit interaction models with message passing[9, 10] or attention mechanisms[11]. Even more noteworthy is an unsupervised neural relational inference (NRI) model can infer an explicit interaction structure while simultaneously predicting the dynamic model in physical simulations, such as the movement of basketball players on the court[12]. This model trains a form of variational autoencoders using motion capture data to model dynamics of the input system, in which the learned embedding (latent code) translates the underlying interaction into an interpretable graph structure and predicts time-related dynamics. Interestingly, this model successfully distinguishes whether a basketball player favors right-hand focus or left-hand focus by only depending on the state of the movement without knowledge of the underlying interactions[12].

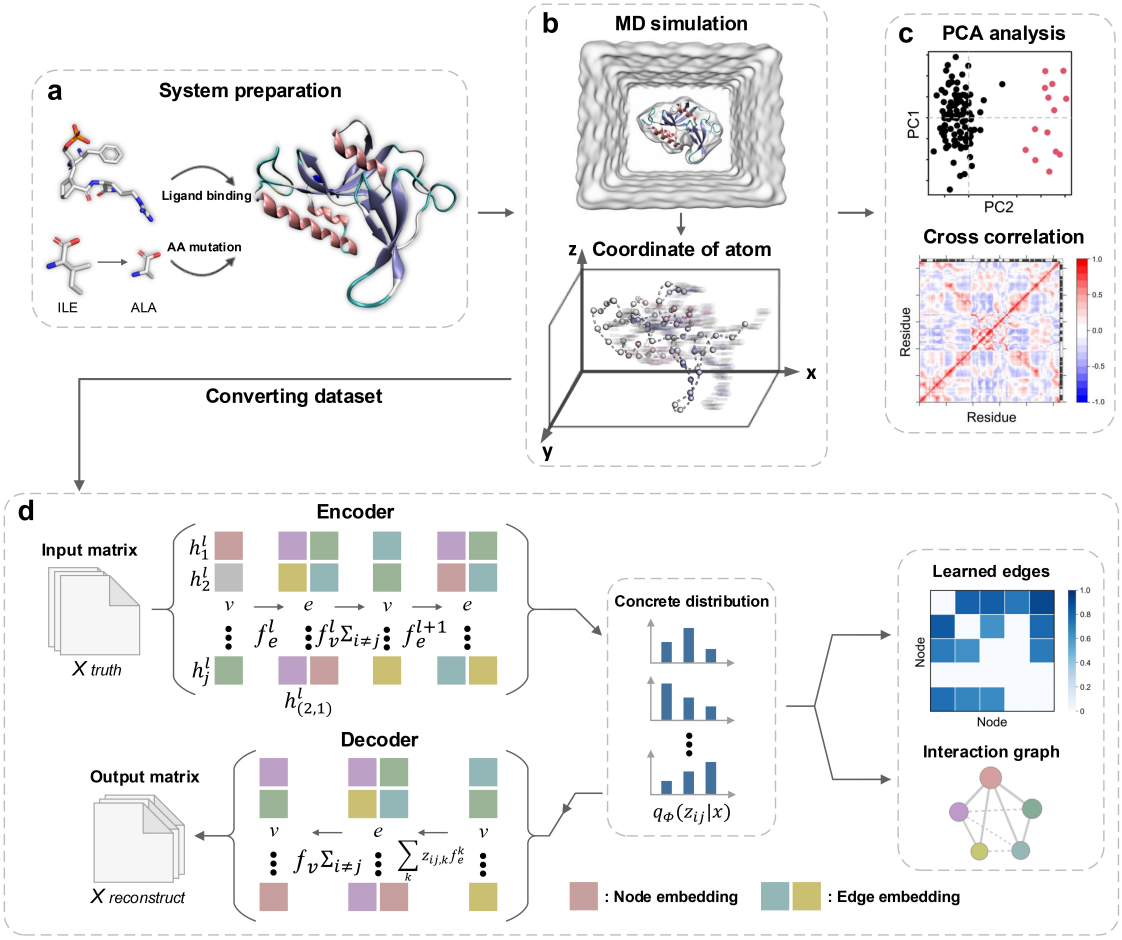

Fig. 1 The process of inferring an interaction graph by reconstructing an MD simulation trajectory. The process including the system preparation of a ligand-binding complex or mutant protein structure with allostery (**a**), the MD simulation of a prepared allosteric system to obtain the trajectory with the dynamic 3D coordinates (**b**), the conventional analysis for the trajectory (**c**), such as PCA or cross-correlation calculation, and the training using the

NRI model with two jointly trained components (**d**). The NRI model consists of an encoder, which infers a factorized distribution $q_{\Phi}(z|x)$ over the latent interactions based on input trajectories and a decoder, which reconstructs several time steps of the dynamic systems given the latent graph learned from the encoder. Based on the MD trajectory, the NRI model formulates the protein allosteric process as a dynamic network of interacting residues. The interaction graph learned from this model is compared with the conventional analysis for a better understanding of the allosteric pathway in the protein.

MD simulation can describe the behavior of biological macromolecules, such as proteins, fats, sugars, etc. These biomolecules are formed by atoms connected by chemical bonds, whose motion rules conform to quantum mechanics, and also approximately to Newtonian mechanics. The NRI model is suitable to learn the simulated motion trajectory of biological macromolecules. In this work, we adapted the NRI model (Fig. 1) to understand how the allosteric pathway mediates remote regulation from ligand binding or mutation site to the active center in a protein. Based on the trajectories running from MD simulations, we formulated the protein allosteric processes as the dynamic networks of interacting residues. This model uses GNN to learn the embedding of the network dynamics by minimizing the reconstruction error between the reconstructed and input trajectories; then, our NRI model infers edges between residues represented by latent variables. The learned embedding inherently abstracts the essential roles of the key residues in the conformational transition, which helps decipher the mechanism of protein allostery. In this study, we performed MD simulations for three allosteric systems, i.e., (*i*) the allosteric regulation of Pin1 induced by ligand binding, (*ii*) the conformational transition of SOD1 by G93A amyotrophic lateral sclerosis-linked mutation, and (*iii*) the activation of MEK1 by oncogenic mutations. Compared to conventional MD analysis, our results show that the GNN-based model can learn interpretable interaction patterns and paths in a nonlinear protein system. To the best of our knowledge, this study is the first attempt to use GNN, particularly NRI to analyze MD simulations in biological systems.

Results

Pathways mediate inter-domain allosteric communication in Pin1

Pin1 utilizes allostery to alter functional activities by changing the local effective modulus of protein without conformational changes[13]. Pin1 as an attractive therapeutic target, contains an inactive N-terminal Trp-Trp (WW) domain (residues 1-39) and an enzymatically active C-terminal peptidyl-prolyl isomerase (PPIase) domain (residues 50-163) connected by a linker (residues 40-

49)[14]. The PPIase domain is composed of a PPIase core (α4-helix and β4-β7 sheets), α1-α3 helices, and a semi-disordered catalytic loop (Fig. 2a). While both domains bind phospho-Ser/Thr-Pro containing substrate motifs, only the PPIase domain can isomerize the peptidyl-prolyl bond through the catalytic site[15]. Moreover, the isolated PPIase has a binding affinity that is typically 100 times weaker than the PPIase in the full-length Pin1, suggesting that the non-catalytic WW domain has the potential to remotely modulate the catalytic activity of the PPIase domain[16]. We performed two simulations of Pin1 in the apo and FFpSPR-bound forms to evaluate the long-range effect of substrate binding to the WW domain on the flexibility of the protein backbone. The root-mean-square deviation (RMSD) and root-mean-square fluctuation (RMSF) values for the simulations (Fig. 2a and Fig. S1a) show that the apo form exhibits high flexibility in the WW domain (β1-β2), catalytic loop, α2-helix, and the PPIase core (β5/α4). In contrast, the flexibilities of these domains are significantly quenched when the FFpSPR binds to the WW domain, indicating that the ligand binding not only stabilizes the conformation of the WW domain but also significantly reduces the dynamic flexibility of the PPIase domain.

To explore the pathways mediating the allosteric communication by the WW domain in Pin1, we introduced a dynamic model of allostery learned by our NRI model. We utilized the MD trajectories to train the NRI model with an encoder and decoder (See Methods for details). We selected 50 uniform timesteps at 4 ns intervals as the ground truth to the model and reconstructed these timesteps. The ground truth and reconstructed trajectories (Supplementary Video 1), the RMSF values (Fig. S2), and the MSE values (Table S1) show that the model can correctly reconstruct the trajectory of motion with small errors while learning the interaction graph.

According to the distribution of learned edges between residues (Fig. 2b), we integrated adjacent residues as blocks for a more straightforward observation of the interactions (Fig. 2c and d). Clearly, the learned edges often occur between the WW domain and other domains, suggesting that the WW domain is the key element in protein movement. Also, we calculated the shortest pathways from the residues in the WW domain to the residues in the catalytic loop based on the learned edges (Table S2). Notably, when the FFpSPR binds to the WW domain, the correlation between the WW domain and the PPIase core is reinforced to launch the first two types of pathways, i.e., from the WW domain to Q131 or P133 in the PPIase core; then, the direct coupling

between the PPIase core and the catalytic loop completes the allosteric communication from the WW domain to the catalytic loop via the WW-PPIase core link (Fig. 3a, *left and middle*).

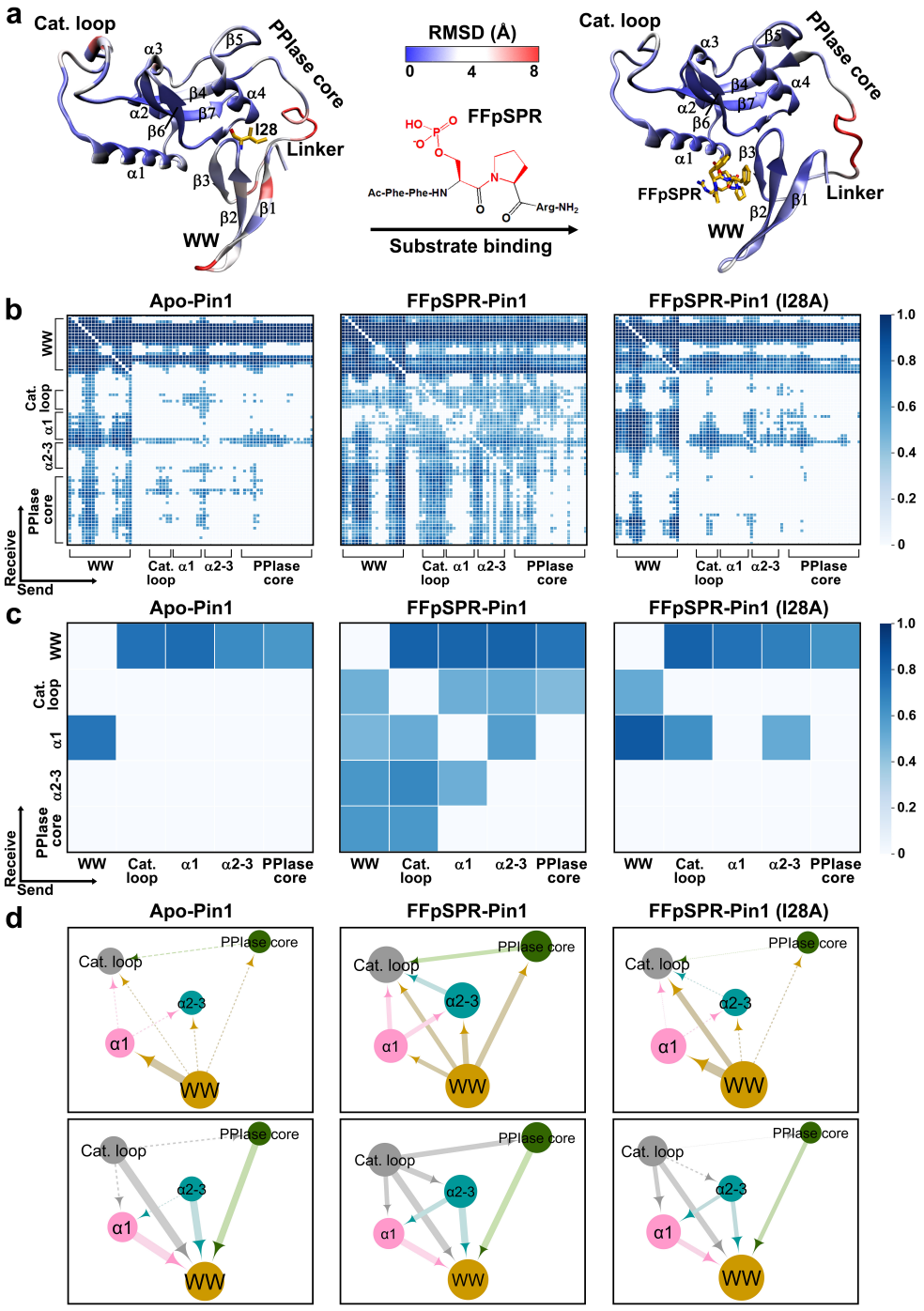

Fig. 2 Changes in protein flexibility and interacting patterns upon ligand binding or mutation in Pin1. a shows the protein flexibility of apo-Pin1 (*left*) and FFpSPR-Pin1 (*right*), where the color scale represents the backbone RMSD. **b** presents the distribution of learned edges between residues in the MD simulations of apo-Pin1 (*left*), FFpSPR-Pin1 (*middle*), and I28A FFpSPR-Pin1 (*right*). **c** presents the distribution of learned edges between

domains/blocks in the MD simulations of apo-Pin1 (*left*), FFpSPR-Pin1 (*middle*), and I28A FFpSPR-Pin1 (*right*). **d** indicates the interacted domains/blocks of apo-Pin1 (*left*), FFpSPR-Pin1 (*middle*), and I28A FFpSPR-Pin1 (*right*), mapped from the learned edges. The size of a node represents the number of edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. The arrows point to the directionality of a learned edge, i.e., the influence from the one starting domain to the ending domain.

Besides, the FFpSPR binding strengthens another communication from the WW domain via K97 in the α1-helix and S105/C113 in the α2-3 helices to the catalytic loop (Fig. 3a, *right*). The frequency of each residue on the paths may demonstrate the relative importance of each residue in enhancing the global connectivity and mediating capabilities that can strengthen the allosteric communication upon the substrate binding (Fig. S3). In particular, both T29 in the interdomain interface and C113 near the catalytic site appear on the allosteric pathways (Fig. 3a, *left and right*). Interestingly, the I28/T29 in the interdomain interface and C113 have been noted as vital mutation sites for impacting the activity of Pin1[17, 18, 19].

However, in the absence of ligand binding, no pathway is found from the WW domain to the catalytic loop. Although the WW domain can interact with the α1-helix, the communication cannot pass from the α1-helix to the catalytic loop (Fig. 3b and Table S2). Thus, the ligand binding makes the WW domain and the PPIase domain more coordinated and compact to strengthen the interdomain communication in Pin1. Moreover, through the learning of different time scales' trajectories, it became evident that this coordination is achieved step by step (Fig. S4). No signal could be transmitted to the catalytic loop at first, but as the simulation time increased, the α-helices and PPIase core promoted the formation of allosteric pathways. Finally, a network of allosteric interactions connecting the WW domain with the catalytic loop was formed to regulate the activity of the PPIase domain.

An NMR study[18] reported that the I28A mutation weakens interdomain interactions between the WW domain and the PPIase domain to reduce the binding affinity of the catalytic site. We simulated the I28A Pin1 of the FFpSPR-bound form, and the trajectory's RMSF value shows that the I28A mutation increases the mobility of the whole protein structure, especially in the WW domain, the catalytic loop, and the α1-α3 helices (Fig. S1b). The learned interaction graph between key domains in Fig. 2c and d (*right*) shows that the I28A mutation dramatically weakens the interactions between the WW domain and PPIase core/α2-α3 helices, which indicates that the fluctuation of the WW domain blocks the propagation of the allosteric signals from the WW to the PPIase core and α2-α3 helices. Although the WW domain is still partially connected to the α1-

helix, the α1-helix cannot bridge to the catalytic loop, resulting in the breakdown of the pathway from the WW domain to the catalytic loop via the α1-helix (Fig. 2d, *right* and Fig. 3c).

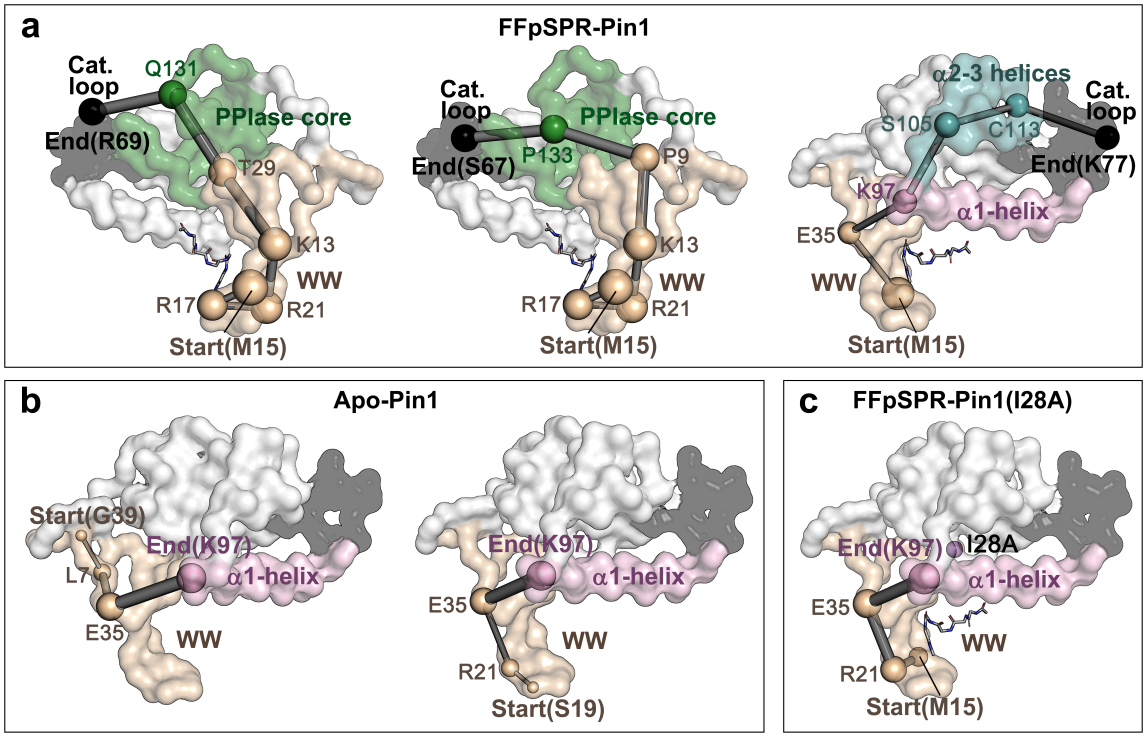

Fig. 3 Pathways mediating inter-domain allosteric communications in Pin1, obtained from shortest pathway calculation. In **a**, the allosteric pathways mediate remote communication from the WW-domain (WW) to the catalytic loop in FFpSPR-Pin1. On the *left*, the allosteric pathway starts from the WW and continues through Q131 in the PPIase core to R69 in the catalytic loop. In the *middle*, the allosteric pathway starts from the WW and continues through P133 in the PPIase core to S67 in the catalytic loop. On the *right,* the allosteric pathway opens communication from the WW and continues through K97 in the α1-helix and the S105/C113 in the α2-3 helices to K77 in the catalytic loop. We used the residues in the WW domain as the starting point and the residues in the catalytic loop as the ending points to present the shortest pathways (additional pathways are shown in Table S2). In **b**, the two pathways in apo-Pin1 are illustrated on the *left*, starting from G39, extending through L7/E35 and ending in K97 in the α1-helix. The *right* pathway begins at S19, then extends onward through R21/E35 and ends at K97 in the α1-helix; **c** represents the pathway in I28A FFpSPR-Pin1 starts in M15 and extends through R21/E35, ending at K97 in the α1-helix. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction.

Allosteric effect of the G93A amyotrophic lateral sclerosis-linked mutation in SOD1

Copper-zinc superoxide dismutase-1 (SOD1) is an oxidoreductase responsible for decomposing toxic superoxide radicals into molecular oxygen and hydrogen peroxide in two rapid steps by alternately reducing and oxidizing active-site copper[20]. The overall structure is composed of eight antiparallel β-strands, plus two loops forming an active site (Fig. 4a). The long active loop

(residues 49-83) can be divided into a dimerization loop (DL), disulfide loop (DiL), and a zinc-binding loop (ZL). The small active loop is an electrostatic loop (EL) with residues 122-142 near the metal site[21]. A study of the SOD1-linked neurodegenerative disorder amyotrophic lateral sclerosis (ALS) shows that the G93A mutation forces the EL to move away from ZL, decreasing the Zn (II) affinity of the protein[22], which affects the pathogenic process of the SOD1-linked ALS[23]. Since the G93A mutation occurs away from the metal site (Fig. 4b), this process is allosteric.

We performed MD simulations for wild type (WT) and G93A SOD1 to generate trajectories for learning the interactions in SOD1 (see Methods for details). The RMSF values (Fig. S5) show that the EL of the G93A SOD1 becomes more flexible than that of the WT SOD1. Correspondingly, the motion mode reveals that the G93A mutation induces the EL far away from the metal site, while the EL of the WT SOD1 can be stabilized in the proximity of the metal site (Fig. 4b). In addition, we found that the G93A mutation makes the A93(O)-L38(N) distance increase, resulting in a decrease in hydrogen bond interaction (Fig. S6a and Table S3). And many hydrogen-bond interactions between the β-barrel and active loops are weakened to make the SOD1 structure looser compared to the WT SOD1 (Figs. S6b-i, S7, and Table S3). Also, the overall dimension of the protein calculated by the radius of gyration (Rg) demonstrates a decrease in protein compactness upon the G93A mutation (Fig. S8). To explore how G93A mutation at the distant site significantly alters the cooperative dynamics near the active loops, we ran the NRI model on the trajectories and compared the performance of motion reconstruction. From the ground truth and reconstructed trajectories (Supplementary Video 2), the RMSF values (Fig. S9), and the MSE values (Table S1), we found that our NRI model can reconstruct the dynamics' many time steps based on the interaction graph learned by the encoder.

The interacting domains mapped from the learned graph show that the long active loop (DL, DiL, and ZL) and the small active loop (EL) interact with each other closely in the WT SOD1, which stabilizes the Zn (II) binding environment (Fig. 4c, d, and e, *left*). A close look at the learned edges graph in Fig. 4c, *left* reveals that the long and small active loops also connect to the residues in the β-barrel, causing a close EL state. Moreover, the pathways in the WT SOD1 further explain the communication pathways, starting from G93 through DL, DiL, and ZL to the EL (Fig. 4f, *left* and Table S4). In contrast, during the EL opening induced by the G93A mutation, the inner connections originally in the long active loop of the WT SOD1 are almost broken, thereby

loosening the network of Zn (II) binding sites (Fig. 4c, d, and e, *right*). Then the allosteric pathways emanating from the A93 no longer propagate through the long active loop, but directly through the residues in the β-barrel to the EL (Fig. 4f, *right,* and Table S4). Overall, the G93A mutation weakens the interaction networks within the active loops, which significantly enlarges the Zn (II) binding pocket and decreases the Zn (II) affinity with the SOD1.

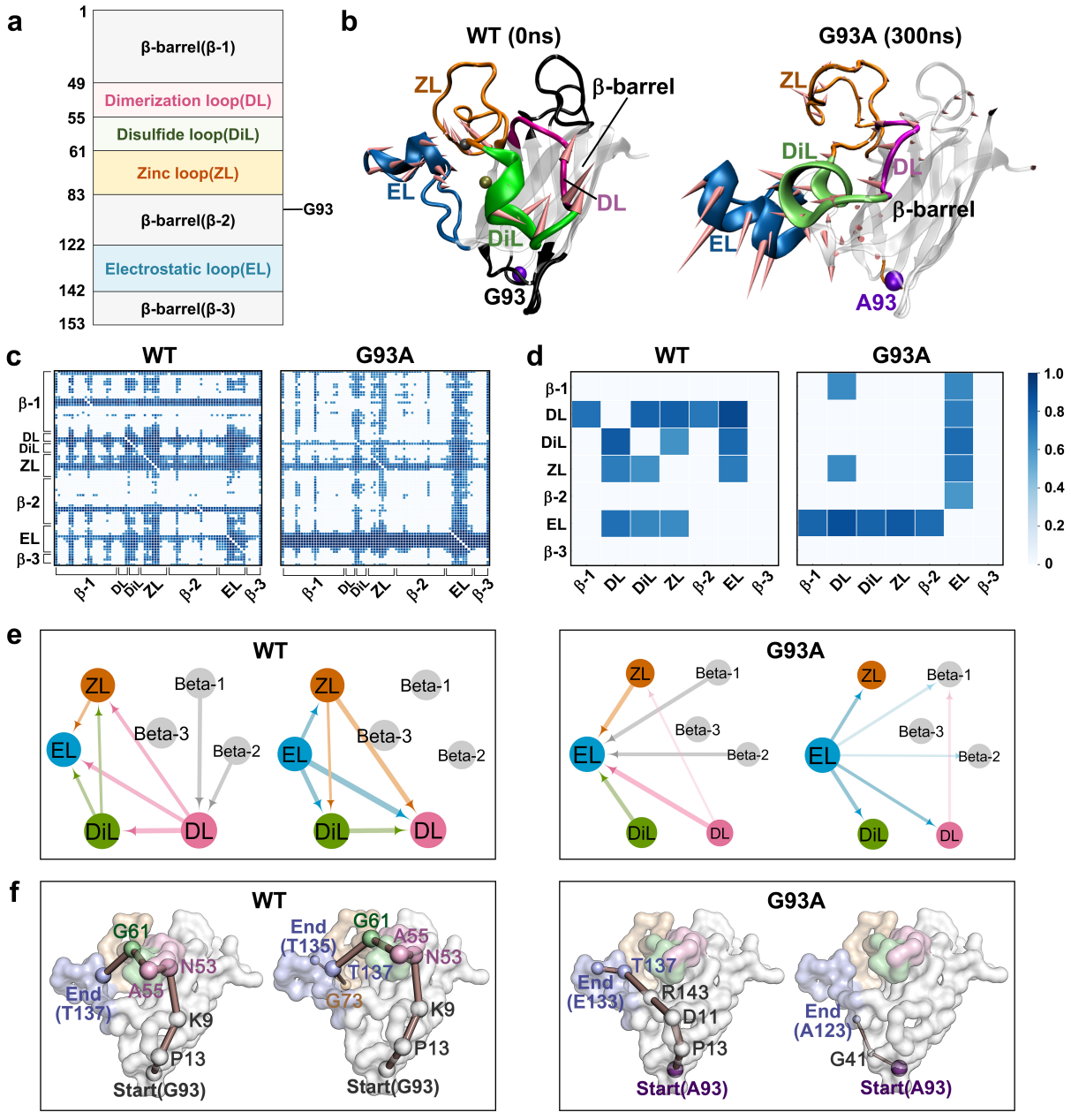

Fig. 4 Change of interactions between residues/domains upon G93A mutation in SOD1. a shows the domain partitions of the SOD1 protein, which includes the position of the G93A mutation. **b** presents the initial structure of the WT SOD1 and G93A SOD1 structure at 300 ns, including a β-barrel (gray), a dimerization loop (DL colored pink), a disulfide loop (DiL colored green), a zinc-binding loop (ZL in orange), and an electrostatic loop (EL in blue). The

directions shown in the graphic denote the motion mode of the protein. **c** presents the distribution of learned edges between residues in the MD simulations of the WT (*left*) and the G93A (*right*) for SOD1. **d** shows a block distribution chart of learned edges between domains in the MD simulations of the WT (*left*) and G93A (*right*) for SOD1. In **e**, the interaction graph is mapped from the learned edges for the WT (*left*) and G93A (*right*) in SOD1. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. The arrows point toward the directionality of the learned edge. In **f**, the pathways from the G93 run through residues in the β-barrel, and residues in the long active loop go to the EL loop for the WT SOD1 (*left*); moreover, the pathways from the A93 go through residues in the β-barrel to the EL loop for the G93A SOD1 (*right*). The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. We used G93/A93 as the starting point, and the residues in the EL as the ending points to present the pathways.

Mechanism of oncogenic mutations activating MEK1

Mitogen-activated protein kinase (MAPKK, also known as MEK) acts as an integration point in the RAS-RAF-MEK-ERK mitogen-activated protein kinase (MAPK) signaling cascade[24]. The activation of MEK requires its phosphorylation by upstream kinases referred to as RAF kinases[25]. The human MEK1 protein consists of a small N-terminal lobe (N-lobe) and a large C-terminal lobe (C-lobe)[26]. As shown in Fig. 5a and b, the small N-lobe is dominated by five antiparallel β-strands (core kinases domain-1) and two conserved αA/αC helices. In these two helices, the αC-helix is critical in the regulation of MEK1 activity[27]. The active site of MEK1 is located at the interface of the N-lobe and C-lobe, binding to the substrate (such as ATP) or the competitive inhibitor known as A-type natriuretic peptide (ANP). The large C-lobe contains three core kinase domains, an activation segment, and a proline-rich loop, where the activation segment and proline-rich loop are crucial in regulating the activation of MEK1 and downstream ERKs in cells[27, 28]. Recent studies reported that the E203K mutation remotely affects the active site of MEK1 to increase the phosphorylation of ERK1/2[29]. Similarly, the phosphorylation of Ser218 and Ser222 is also required for MEK1 activation to promote cell proliferation and transformation, which eventually leads to various human cancers[25].

To explore the allosteric effect of the mutation on MEK1, we performed MD simulations and analyses for two nonactive MEK1s (WT and A52V)[29], two active forms (mutation E203K[29], and a phosphorylated MEK1, where both Ser218 and Ser222 were phosphorylated[25]). The secondary structure changes (as shown in Fig. S10) show that the activation segment experiences a helix-to-loop transition in the active MEK1 (E203K and phosphorylated Ser218/222). In contrast, this segment's helix content in the WT and in the A52V MEK1 increased significantly compared to

the active MEK1. The principal component analysis (Fig. S11) reflects the activation segment's open trend in the active MEK1.

The above analysis only shows the changes in the dynamic motions of MEK1, which may fail to identify the common interaction features in two active MEK1s. Thus, the NRI model was applied to learn the trajectories. The motion reconstruction results show that the trajectory reconstructed by our NRI model almost coincides with the ground-truth trajectory (Fig. S12, Supplementary Video 3, and Table S1), indicating a high degree of confidence based on the learned interaction graph. As shown in the learned interaction graph of nonactive MEK1 (WT and A52V) (Fig. 5c and d), few interactions occur between the domains. In contrast, the αA-helix, core kinase domain-1, activation segment, and the proline-rich loop of phosphorylated MEK1 strongly interact with other domains, which indicates that they drive the slow motion in the activation of phosphorylated MEK1 (Fig. 5c and d).

We mapped the graph of the phosphorylated MEK1 as the interacting domains (Fig. 5e, *left*) and calculated the allosteric pathways (Fig. 5f, *left* and Table S5). Interestingly, four domains (the αA-helix, the core kinase domain-1, the activation segment, and the proline-rich loop) form an interaction pattern, in which the activation segment connects all the way to the core kinase domain-1 and the αA-helix, which may affect the binding affinity of ANP in the active pocket. Meanwhile, the activation segment also connects to the proline-rich loop, which may activate downstream extracellular signal-regulated kinases (ERKs) in cells. Then, we applied the NRI model to learn the inner-domain correlation from the dynamic motion of E203K MEK1. A closer look at the learned graph reveals that similar to the phosphorylated MEK1, the active mutation (E203K) strengthens the interactions between the activation segment/proline-rich loop and the rest of MEK1 (Fig. 5c, d, and e). From the allosteric pathways starting with R201 (Fig. 5f, *right* and Table S5), we found that the activation segment has a great effect on passing messages from R201 (near E203K) to the proline-rich loop. The communication propagates through the αA-helix to the αC-helix due to the effect of the E203K mutation on the αC-helix. Hence, phosphorylated Ser218/222 and E203K mutations have a similar effect on the proline-rich loop, i.e., the activation segment as a “messenger” can interact with the proline-rich loop in their dynamics thereby enhancing communication to the proline-rich loop.

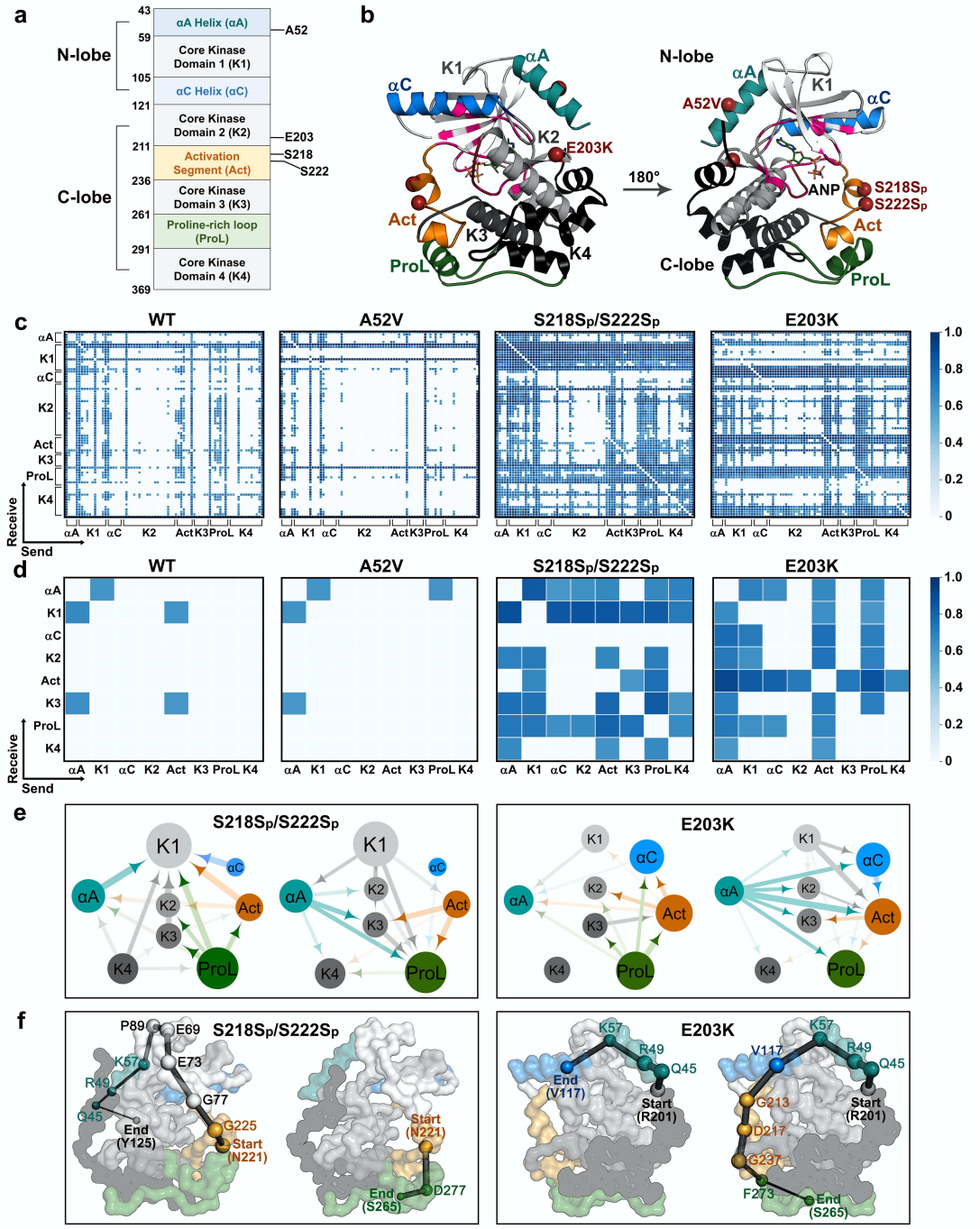

Fig. 5 Changes in domain communications upon active mutations in MEK1. a shows the domain partition of MEK1 protein including the positions of mutations (A52V, S218Sp/S222Sp, and E203K). **b** shows two different views of the MEK1 structure. The N-terminal lobe (N-lobe) contains one core kinase (gray) and two conserved α-helices (blue). The C-terminal lobe (C-lobe) contains three core kinase domains (gray and black), an activation segment (orange), and a proline-rich loop (green). **c** presents the distribution of learned edges between residues in the MD simulations of WT, A52V, S218Sp/S222Sp, and E203K MEK1. **d** presents the distribution of learned edges between domains in the MD simulations of WT, A52V, S218Sp/S222Sp, and E203K MEK1. In **e**, the interaction graph is mapped from the learned edges of active mutant MEK1. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. The arrows denote

the directionality of a learned edge. In **f**, the allosteric pathways start from N221 in the activation segment and lead to the αA-helix and the proline-rich loop in the S218Sp/S222Sp MEK1 (*left*). On the (*right*), the allosteric pathways start from R201 (near E203K) and lead to the αC-helix and the proline-rich loop in the E203K MEK1. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. We used N221 and R201 (near E203K) as the starting points, and the residues in the αA/αC helices and the proline-rich loop as the ending points to present the pathways.

Discussion

This study applied a GNN-based NRI model to analyze latent interactions between residues from reconstructing MD trajectories of proteins. We carried out three case studies to explore the allosteric long-range interactions for the Pin1, SOD1, and MEK1 systems. We have demonstrated that our NRI model can effectively generate the interaction graphs related to protein's slow-motion through embedding reconstructed MD trajectories, and the shortest pathways between the allosteric site and the active site in the interaction graphs can reveal the pathways mediating allosteric communications. Although a number of methods have applied graph theory to model allosteric communication in protein, they are mostly based on the static crystal structures of proteins only. These theory-based methods are ineffective in modeling allostery, which is dynamic in nature. One major advance in our allostery model is the use of information from MD simulations. Notably, we take a major step in recognizing allostery-related long-range interactions by adding an instantaneous velocity feature to each node over the continuous snapshots, which can explain the allosteric communication related to biomolecular slow-motion.

The allosteric pathways derived from the shortest paths provide valuable information when considering protein design. It may be possible to mutate the residues in the allosteric pathways to alter the biological functions and regulatory properties of proteins. One example we demonstrated is residue I28 in Pin1 with a known impact of allostery appearing next to T29. The exemplary residue T29 is a key residue on one pathway identified in Fig. 3a. Furthermore, both residues R49 and K57 in the αA-helix appear on the allosteric pathways of two activated MEK1s (Fig. 5f). Since the αA-helix is the critical interface interacting with the rest of the kinase domain, the mutations of residues R49 and K57 are likely to cause significant alterations in the helical structure thereby inducing ERK phosphorylation[30]. In addition, residues E69, E73, and G77 showed a significantly high frequency on the pathways of the phosphorylated MEK1, which confirms their roles as global mediating sites in allosteric communication (Fig. S13). Mutations in this region indeed lead to

constitutive activation of the MAPK pathway[31]. Thus, the allosteric pathways learned by our NRI model may have the potential to greatly reduce the spectrum of mutation screening. Future research directions may include a longer MD simulation or experimental works to validate the predictions from our NRI model.

With regard to traditional analysis methods applied to MD trajectories, PCA obtains principal components representing the movement patterns determined in MD simulations, and the cross-correlation analysis is a measurement that tracks the movements over time of two variables to determine the degree of linear correlation between them. However, these two classical methods are limited to linear correlations mainly associated with harmonic/isotropic local torsional motion that occur on relatively short-time scales. It is well known that dynamic biomolecules often undergo large-scale structural changes during their biological functions. Due to large energy barriers, the conformational changes of biomolecules usually occur in a nonlinear fashion in milliseconds or longer time scales, which is typically inaccessible to MD. On the other hand, the driving forces leading to the long-term conformational changes and the underlying nonlinear relationships reveal themselves long before the conformational changes are revealed in trajectories. The NRI model is a powerful tool used to model such driving forces and nonlinear relationships through the learned embedding and dynamic interaction graph. The NRI model is not restricted to allosteric regulation. Many other biological and pharmaceutical processes, such as protein folding/unfolding, protein activation, or drug molecule binding targets can also be formulated as a dynamic interaction graph by the NRI model. In particular, the NRI model is appealing when probing non-periodic biomolecular motion. Unlike the periodic physical movement where interactions do not change over time, proteins during performing functions are often accompanied by considerable conformation and interaction changes. Using NRI in those cases will retrieve interactions over time. We believe the NRI model can be developed to recover the interactions between residues at every time interval in the process of performing protein functions. Additional NRI methods, such as dynamic NRI[32] can be applied for this purpose.

Methods

NRI model

The NRI model[12] consists of two co-training parts: an encoder to predict the interaction given the dynamic system's trajectories, and a decoder to predict the trajectories of the dynamic system given in the interaction graph. The NRI model simultaneously learns the edge values and reconstructs the trajectories of the dynamic system in an unsupervised manner based on an unknown graph z (where z_{ij} represents the discrete attribute value of the edge between nodes v_i and v_j). The input consists of N nodes. The feature vector (position and velocity in the dimensions of x, y, and z) of node v_i (input/output dimension of 6 for each node) is denoted as x_i^t at time t. All N nodes' feature set is denoted as $x^t = \{x_1^t, \cdots, x_N^t\}$. The trajectory of node i is denoted as $x_i = \{x_i^1, \cdots, x_i^T\}$, where T is the number of time steps. Finally, all trajectory data are recorded as $x = \{x^1, \cdots, x^T\}$. The structure of the model is presented in Fig. 1.

As shown in Fig. 1d (on the left), the encoder

$$q_\Phi(z_{ij}|x) = softmax\left(f_{\mathrm{enc},\Phi}(x)_{ij,1:K}\right) \qquad (1)$$

infers the discrete categorical variable z_{ij} based on the input trajectories $x_1, \cdots, x_K$, in which $f_{enc,\Phi}(x)$ is a GNN performed on the fully connected networks (without self-connection) to predict the latent graph structure. The encoder runs two rounds of node-to-edge ($v \to e$) and an edge-to-node ($e \to v$) message passing. The node-to-edge operation generates the edge features connecting the node embeddings and the edge-to-node operation aggregates the message of edge embeddings from all incoming edges. Since the graph is fully connected, each node obtains a message from the entire graph. Finally, all messages pass from nodes to edges. In our implementation model, every message passing operation is performed by a 2-layer perceptron[12].

The distribution of z, $q_\Phi(z|x)$, is learned from the encoder. Then the sampling is performed to generate z_{ij} only available in the K edge type. We sampled from a continuous approximation of the discrete distribution and used reparameterization to obtain gradients from this approximation, which were calculated as[33]:

$$z_{i,j} = softmax((h_{(i,j)}^2 + g)/\tau) \qquad (2)$$

where $g \in \mathbb{R}^K$ is an independent and uniformly distributed vector from the Gumbel distribution (0,1), and τ (softmax temperature) represents the smoothness of sampling. The distribution tends to become one-hot encoded samples when $\tau \to 0$.

The decoder expressed as:

$$p_\theta(x|z) = \prod_{t=1}^{T} p_\theta(x^{t+1}|x^t, \cdots, x^1, z) \qquad (3)$$

reconstructs the dynamic systems $p_\theta(x^{t+1}|x^t, \cdots, x^1, z)$ with a GNN given the latent graph structure z. A recurrent decoder with a GRU unit[34] is required to model $p_\theta(x^{t+1}|x^t, \cdots, x^1, z)$. The decoder runs multiple GNNs in parallel to the encoder. In the node-to-edge ($v \rightarrow e$) message passing, the input is the recurrent hidden state from the previous time step. The hidden state of an edge is determined by the hidden state of its connecting nodes, and it allows the message at each time step to pass through the hidden state. Thus, the prediction at $t + 1$ is based not only on the previous time step but also on messages from all the previous time steps. In the edge-to-node ($e \rightarrow v$) message passing, the concatenation of the aggregated messages, the current input, and the previously hidden state, is denoted as the input of a GRU update to generate the hidden state at the next time step. Then, the value observed previously and the hidden state at the current time step are used to predict the state's distribution (position and velocity) in future time steps.

This model is formalized as a variational autoencoder (VAE)[35, 36] that maximizes the evidence lower bound (ELBO):

$$\mathcal{L} = \mathbb{E}_{q_\Phi(\mathrm{z}|\mathrm{x})}[\log p_\theta(\mathrm{x}|\mathrm{z})] - \mathrm{KL}[q_\Phi(\mathrm{z}|\mathrm{x})||p_\theta(\mathrm{z})] \quad (4)$$

The ELBO objective has two terms, namely the reconstruction error $\mathbb{E}_{q_\Phi(\mathrm{z}|\mathrm{x})}[\log p_\theta(\mathrm{x}|\mathrm{z})]$ and KL divergence $\mathrm{KL}[q_\Phi(\mathrm{z}|\mathrm{x})||p_\theta(\mathrm{z})]$, in which the encoder $q_\Phi(\mathrm{z}|\mathrm{x})$ returns a factorized distribution of z_{ij}. Therefore, a one-hot encoding representation of the K interaction types (K=4 presents edges/no edge) is used on z_{ij}. $p_\theta(x|z)$ representing the decoder that reconstructs the dynamic systems given the distribution of z_{ij}. The autoencoder maps the input X to a feature space and then maps from this feature space back to the input space to minimize the reconstruction error. The reconstruction error is calculated through:

$$\mathbb{E}_{q_\Phi(\mathrm{z}|\mathrm{x})}[\log p_\theta(\mathrm{x}|\mathrm{z})] = -\sum_j \sum_{t=2}^{T} \frac{||x_j^t - \mu_j^t||^2}{2\sigma^2} + const \qquad (5)$$

and the KL divergence is the sum of entropies and a constant:

$$\mathrm{KL}[q_\Phi(\mathrm{z}|\mathrm{x})||p_\theta(\mathrm{z})] = \sum_{i \neq j} H(q_\Phi(z_{ij}|x)) + const \qquad (6)$$

The whole training process was carried out as follows: (*i*) We first performed the encoder to calculate $q_{\Phi}(z_{ij}|x)$ given a training MD trajectory X; (*ii*) we then sampled z_{ij} from a continuous approximation of the discrete distribution, and (*iii*) we finally ran the decoder to reconstruct the interacting dynamics $p_{\theta}(x^{t+1}|x^{t},\cdots,x^{1},z)$ for the Pin1, SOD1 and MEK1 systems.

Simulation data

1. Preparation of protein structures

The crystal structures for the three systems (Pin1, SOD1, and MEK1) were retrieved from the Protein Data Bank (www.rcsb.org). The apo Pin1 structure was obtained from PDB 3TDB[14]. To obtain the Pin1-FFpSPR complex, we docked the substrate (FFpSPR) into the WW-domain of apo Pin1 using Autodock 4.2[37]. Missing residues of the protein (residues 39-50) in the inter-domain linker were modeled by SWISS-MODEL[38]. The SOD1 and MEK1 structures were taken directly from PDB 2C9V[21] and 3SLS[26]. The structures of corresponding mutations were also constructed using the SWISS-MODEL.

2. Conventional molecular dynamics (cMD) simulations

The cMD simulations of four MEK1 structures (WT, A52V, E203K, and phosphorylated MEK1) were performed using the GROMACS 5.1.4 package with the 53A6 GROMOS force field[39]. All complexes of MEK1 structures and an A-type natriuretic peptide (ANP), which was used as an inhibitor, were performed in a periodic boundary box with the simple point charge (SPC) water model[40]. Two sodium ions were added to the box to neutralize the WT and A52V MEK1 systems, and four sodium ions were added to the box to neutralize the S218Sp/S222Sp MEK1 system. Note that the E203K MEK1 system does not need to add ions for neutralization. In addition, energy minimization was performed using the steepest descent method to obtain the energy-minimized initial structure for the next simulations. Subsequently, 100 ps of NVT (Berendsen temperature coupled with constant particle number, volume, and temperature)[41] and 100 ps of NPT (Parrinello-Rahman pressure coupled with constant particle number, pressure, and temperature)[41] were performed to maintain the stability of the system (300 K, 1 bar). The coupling constants for temperature and pressure were set at 0.1 and 2.0 ps, respectively. Long-range electrostatic interactions were described using the particle mesh Ewald algorithm with an interpolation order

of 4 and a grid spacing of 1.6 Å[42]. van der Waals interactions were calculated according to the cutoff value of 12 Å. All bond lengths were constrained using the LINear Constraint Solver (LINCS) algorithm[42]. After stabilizing all thermodynamic properties, the molecular systems were simulated for 200 ns with a time interval of 2 fs, whereas the coordinates for all models were stored every 2 ps.

3. Gaussian accelerated molecular dynamics (GaMD) simulation

GaMD simulation is an enhanced sampling technique performed by adding a harmonic boost potential to smoothen the system's potential energy surface[43, 44]. To enhance the conformational sampling of Pin1 and SOD1 structures, GaMD simulations were performed on the apo Pin1, FFpSPR bound Pin1, FFpSPR bound I28A Pin1, WT SOD1, and G93A SOD1 structures. For each of the systems, the graphic processing unit (GPU) version of AMBER18 was applied to perform GaMD simulation[45]. The GaMD simulation has five stages: (*i*) conventional MD preparatory stage for the equilibration of the system, (*ii*) conventional MD stage to collect potential statistics for calculating the GaMD acceleration parameters, (*iii*) GaMD pre-equilibration stage with boost potential, (*iv*) GaMD equilibration stage to update the boost parameters, and (*v*) multiple independent GaMD production runs with randomized initial atomic velocities[44]. The molecular systems were simulated 200 ns for the Pin1 system and 300 ns for SOD1 system, with a time interval of 2 fs. The GaMD simulation trajectories were analyzed using CPPTRAJ[46] and VMD[47] for RMSF calculation, secondary structure analysis, principal component analysis, and hydrogen bond calculation.

NRI model construction details

All NRI trainings were performed using Adam optimizer[48] with a learning rate of 0.0005 and a batch size of 1, decayed by a factor of 0.5 every 200 epochs. The concrete distribution was used with $\tau = 0.5$. During testing, we replaced the concrete distribution with a categorical distribution to obtain discrete latent edge types. All experiments were run for 500 training epochs. The discrete samples were used in the training forward pass. We saved model checkpoints after every epoch whenever the validation set performance improved and loaded the best performing model for the test set evaluation. We used a standard Nvidia GeForce GTX 1080Ti GPU card and a Core solo

CPU to train our models. Each CPU was allocated 48G memory. The training time for one experiment took about 5 hours.

We used the MD trajectories to generate the input data for the next training, validation, and test. The dataset of the Pin1 system has a total size of 2000 frames for 73 Cα atoms each. The dataset of the SOD1 system has 3000 frames for 77 Cα atoms each. The dataset of the MEK1 system has 2500 frames for 76 Cα atoms each. We normalized the position and velocity features to the maximum absolute value of 1. The overall input/output dimension of the model is 6 (3D position and velocity). Training, validation, and test samples each contain 50 frames uniformly extracted from each trajectory. For the training part, the model received a ground truth input in each timestep. The dynamics for our three systems changed considerably over time, the protein conformations in the early stage of the simulation are quite different from that in the end stage. Therefore, in the experimental tests, we fed in the 50 timesteps as ground truth to the encoder and then reconstructed these timesteps. All experiments used the multi-layer perceptron (MLP) encoder and recurrent neural network (RNN) decoder to have a capacity comparable to the full graph model. The first edge type is "hard-coded" as non-edge (no messages are passed along this type). The basic building block of the MLP encoder is a 2-layer MLP with a hidden and output (embedding) dimension of 256, together with batch normalization, dropout, and ELU activations. The RNN decoder adds a GRU style update to the single-step prediction. Given the interaction graph learned from the NRI model, we took the allosteric site as the starting point and the active site as the terminal point to calculate the shortest pathways using Dijkstra's algorithm[49].

Acknowledgements

This work was supported by the China Scholarship Council to JZ, the US National Institutes of Health [R35-GM126985] to DX, and the Overseas Cooperation Project of Jilin Province [20200801069GH] to WH. We also thank Ms. Carla Roberts for thoroughly proofreading this paper.

References

1. Altis A, Nguyen PH, Hegger R, Stock G. Dihedral angle principal component analysis of molecular dynamics simulations. *The Journal of chemical physics* **126**, 244111 (2007).

2. Hünenberger PH, Mark AE, van Gunsteren WF. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *Journal of molecular biology* **252**, 492-503 (1995).

3. Li DW, Meng D, Brüschweiler R. Short-range coherence of internal protein dynamics revealed by high-precision in silico study. *Journal of the American Chemical Society* **131**, 14610-14611 (2009).

4. Lange OF, Grubmüller H. Generalized correlation for biomolecular dynamics. *Proteins* **62**, 1053-1061 (2006).

5. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. *Biophysical journal* **86**, 85-91 (2004).

6. Girvan M, Newman ME. Community structure in social and biological networks. *Proc Natl Acad Sci U S A* **99**, 7821-7826 (2002).

7. Wang J*, et al.* scGNN: a novel graph neural network framework for single-cell RNA-Seq analyses. *bioRxiv*, 2020.2008.2002.233569 (2020).

8. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans Neural Netw Learn Syst* **32**, 4-24 (2021).

9. Li C, Cui Z, Zheng W, Xu C, Ji R, Yang J. Action-Attending Graphic Neural Network. *IEEE Transactions on Image Processing* **27**, 3657-3670 (2018).

10. Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent Models of Visual Attention. In: *Neural Information Processing Systems*) (2014).

11. Steenkiste Sv, Chang M, Greff K, Schmidhuber J. Relational Neural Expectation Maximization: Unsupervised Discovery of Objects and their Interactions. In: *International Conference on Learning Representations*) (2018).

12. Kipf TN, Fetaya E, Wang K-C, Welling M, Zemel RS. Neural Relational Inference for Interacting Systems. In: *International Conference on Machine Learning*) (2018).

13. Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57**, 433-443 (2004).

14. Zhang M*, et al.* Structural and Kinetic Analysis of Prolyl-isomerization/Phosphorylation Cross-Talk in the CTD Code. *ACS Chemical Biology* **7**, 1462-1470 (2012).

15. Yaffe MB*, et al.* Sequence-Specific and Phosphorylation-Dependent Proline Isomerization: A Potential Mitotic Regulatory Mechanism. *Science* **278**, 1957 (1997).

16. Peng JW. Investigating Dynamic Interdomain Allostery in Pin1. *Biophysical reviews* **7**, 239-249 (2015).

17. Namanja AT, Peng T, Zintsmaster JS, Elson AC, Shakour MG, Peng JW. Substrate recognition reduces side-chain flexibility for conserved hydrophobic residues in human Pin1. *Structure* **15**, 313-327 (2007).

18. Wilson KA, Bouchard JJ, Peng JW. Interdomain interactions support interdomain communication in human Pin1. *Biochemistry* **52**, 6968-6981 (2013).

19. Xu N*, et al.* The C113D Mutation in Human Pin1 Causes Allosteric Structural Changes in the Phosphate Binding Pocket of the PPIase Domain through the Tug of War in the Dual-Histidine Motif. *Biochemistry* **53**, 5568-5578 (2014).

20. Hart PJ*, et al.* A structure-based mechanism for copper-zinc seperoxide dismutase. *Biochemistry* **38**, 2167-2178 (1999).

21. Strange RW, Antonyuk SV, Hough MA, Doucette PA, Valentine JS, Hasnain SS. Variable metallation of human superoxide dismutase: Atomic resolution crystal structures of Cu-Zn, Zn-Zn and as-isolated wild-type enzymes. *Journal of molecular biology* **356**, 1152-1162 (2006).

22. Kayatekin C, Zitzewitz JA, Matthews CR. Zinc binding modulates the entire folding free energy surface of human Cu,Zn superoxide dismutase. *Journal of molecular biology* **384**, 540-555 (2008).

23. Smith AP, Lee NM. Role of zinc in ALS. *Amyotrophic Lateral Sclerosis* **8**, 131-143 (2007).

24. Kolch W. Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nature reviews Molecular cell biology* **6**, 827-837 (2005).

25. Shi H, Kong X, Ribas A, Lo RS. Combinatorial treatments that overcome PDGFRβ-driven resistance of melanoma cells to V600EB-RAF inhibition. *Cancer research* **71**, 5067-5074 (2011).

26. Meier C*, et al.* Engineering human MEK-1 for structural studies: A case study of combinatorial domain hunting. *Journal of Structural Biology* **177**, 329-334 (2012).

27. Hanks SK, Hunter T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification1. *The FASEB Journal* **9**, 576-596 (1995).

28. Dang A, Frost JA, Cobb MH. The MEK1 proline-rich insert is required for efficient activation of the mitogen-activated protein kinases ERK1 and ERK2 in mammalian cells. *The Journal of biological chemistry* **273**, 19909-19913 (1998).

29. Gao J*, et al.* 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Medicine* **9**, 4 (2017).

30. Arcila ME*, et al.* MAP2K1 (MEK1) Mutations Define a Distinct Subset of Lung Adenocarcinoma Associated with Smoking. *Clinical Cancer Research* **21**, 1935-1943 (2015).

31. Estep AL, Palmer C, McCormick F, Rauen KA. Mutation Analysis of BRAF, MEK1 and MEK2 in 15 Ovarian Cancer Cell Lines: Implications for Therapy. *PLoS One* **2**, 7 (2007).

32. Graber C, Schwing AG. Dynamic Neural Relational Inference. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*) (2020).

33. Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax. In: *International Conference on Learning Representations*) (2016).

34. Cho K*, et al.* Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: *Empirical Methods in Natural Language Processing*) (2014).

35. Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: *International Conference on Learning Representations*) (2014).

36. Rezende DJ, Mohamed S, Wierstra D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In: *International Conference on Machine Learning*) (2014).

37. Morris GM*, et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **30**, 2785-2791 (2009).

38. Biasini M*, et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research* **42**, W252-258 (2014).

39. Oostenbrink C, Soares TA, van der Vegt NF, van Gunsteren WF. Validation of the 53A6 GROMOS force field. *European biophysics journal : EBJ* **34**, 273-284 (2005).

40. Mark P, Nilsson L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *The Journal of Physical Chemistry A* **105**, 9954-9960 (2001).

41. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **81**, 3684-3690 (1984).

42. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of chemical physics* **98**, 10089-10092 (1993).

43. Miao Y, Feher VA, McCammon JA. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of chemical theory and computation* **11**, 3584-3595 (2015).

44. Miao Y, McCammon JA. Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *Annual reports in computational chemistry* **13**, 231-278 (2017).

45. Lee T-S*, et al.* GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features. *J Chem Inf Model* **58**, 2043-2050 (2018).

46. Roe DR, Cheatham TE, 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of chemical theory and computation* **9**, 3084-3095 (2013).

47. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33-38, 27-38 (1996).

48. Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*) (2015).

49. Dijkstra EW. A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269-271 (1959).

Figures

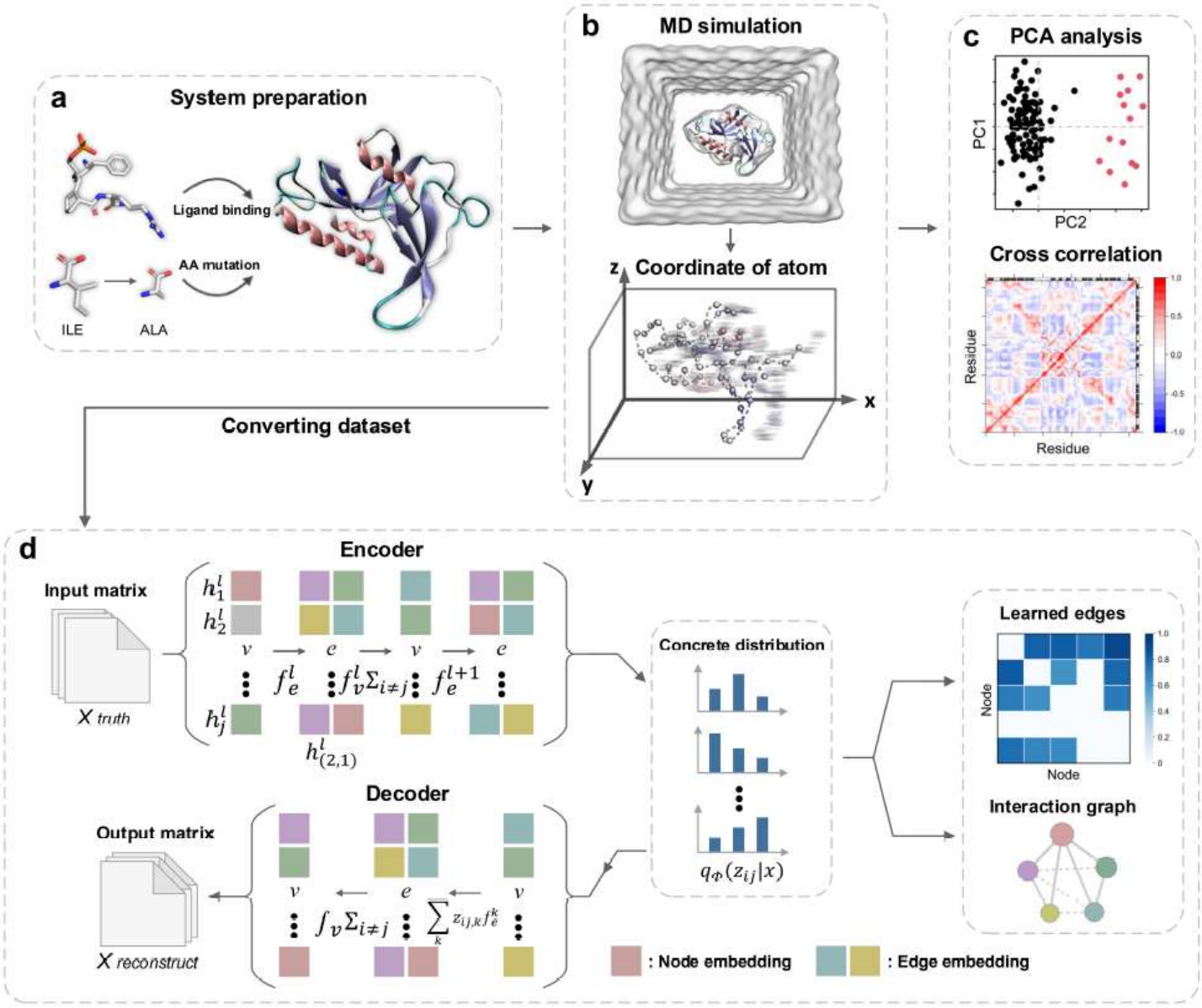

Figure 1

The process of inferring an interaction graph by reconstructing an MD simulation trajectory. The process including the system preparation of a ligand-binding complex or mutant protein structure with allostery (a), the MD simulation of a prepared allosteric system to obtain the trajectory with the dynamic 3D coordinates (b), the conventional analysis for the trajectory (c), such as PCA or cross-correlation calculation, and the training using the NRI model with two jointly trained components (d). The NRI model consists of an encoder, which infers a factorized distribution Φ(z|x) over the latent interactions based on input trajectories and a decoder, which reconstructs several time steps of the dynamic systems given the latent graph learned from the encoder. Based on the MD trajectory, the NRI model formulates the protein allosteric process as a dynamic network of interacting residues. The interaction graph learned from this

model is compared with the conventional analysis for a better understanding of the allosteric pathway in the protein.

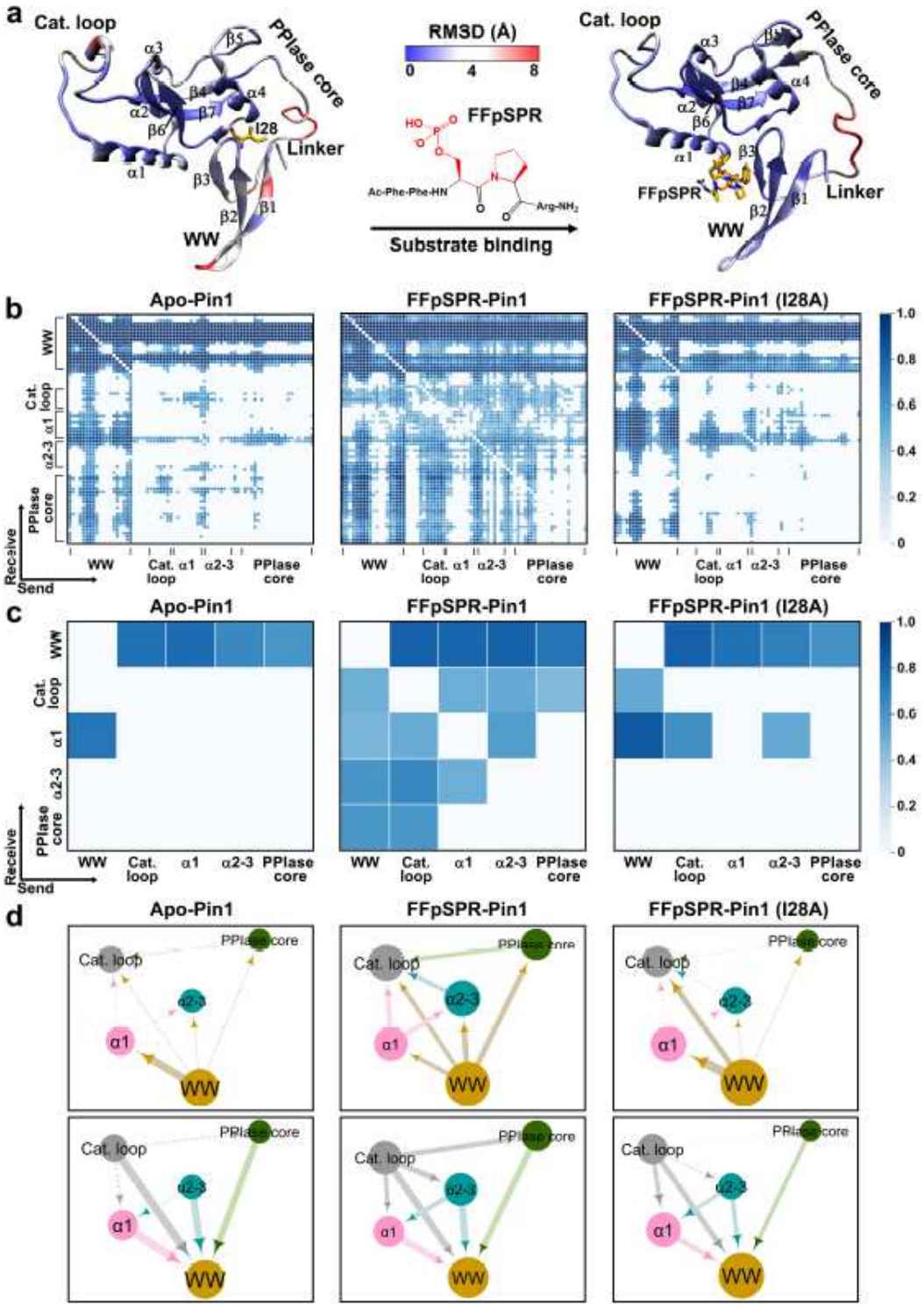

Figure 2

Changes in protein flexibility and interacting patterns upon ligand binding or mutation in Pin1. a shows the protein flexibility of apo-Pin1 (left) and FFpSPR-Pin1 (right), where the color scale represents the backbone RMSD. b presents the distribution of learned edges between residues in the MD simulations of

apo-Pin1 (left), FFpSPR-Pin1 (middle), and I28A FFpSPR-Pin1 (right). c presents the distribution of learned edges between domains/blocks in the MD simulations of apo-Pin1 (left), FFpSPR-Pin1 (middle), and I28A FFpSPR-Pin1 (right). d indicates the interacted domains/blocks of apo-Pin1 (left), FFpSPR-Pin1 (middle), and I28A FFpSPR-Pin1 (right), mapped from the learned edges. The size of a node represents the number of edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. The arrows point to the directionality of a learned edge, i.e., the influence from the one starting domain to the ending domain.

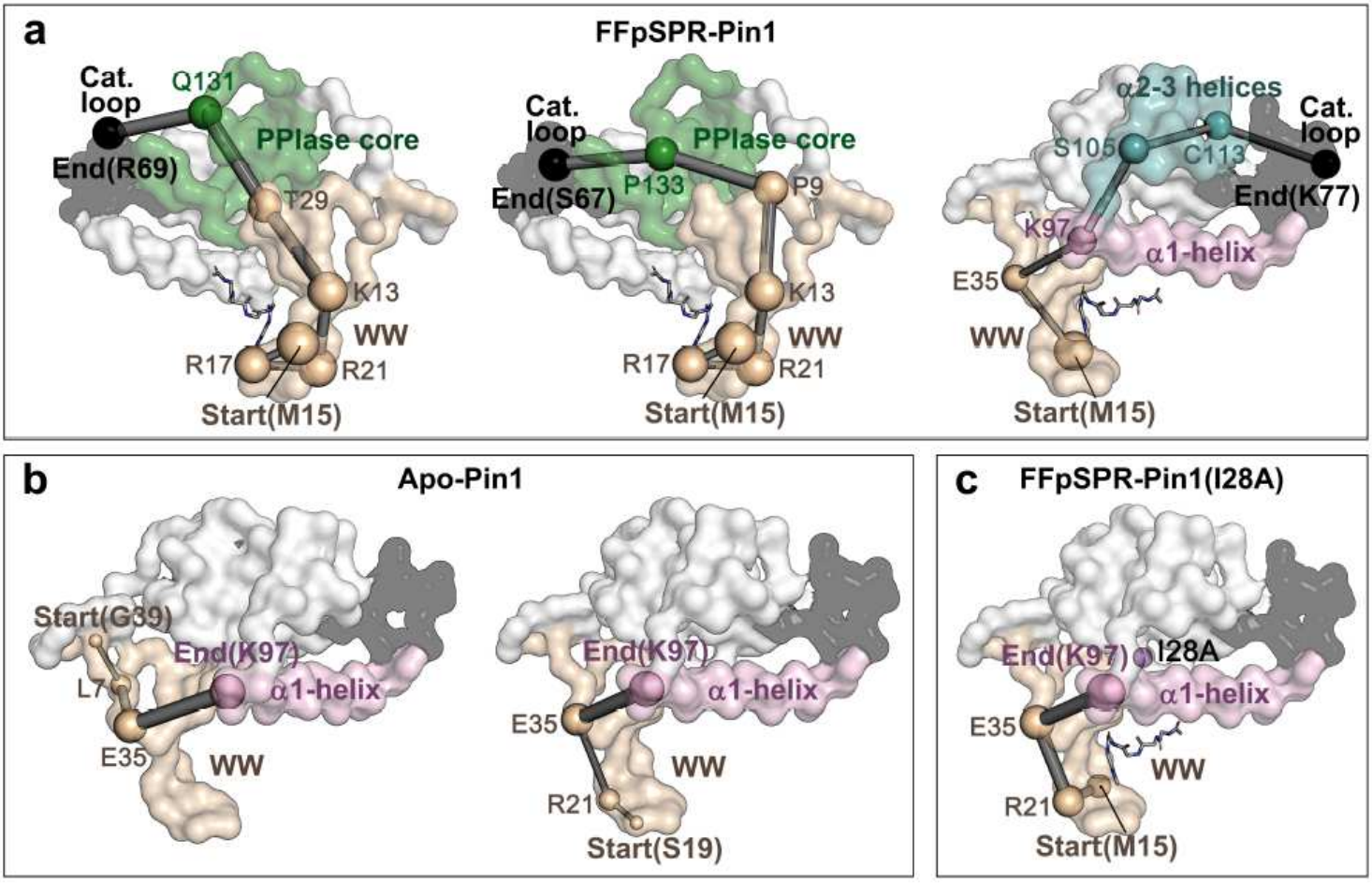

Figure 3

Pathways mediating inter-domain allosteric communications in Pin1, obtained from shortest pathway calculation. In a, the allosteric pathways mediate remote communication from the WW-domain (WW) to the catalytic loop in FFpSPR-Pin1. On the left, the allosteric pathway starts from the WW and continues through Q131 in the PPIase core to R69 in the catalytic loop. In the middle, the allosteric pathway starts from the WW and continues through P133 in the PPIase core to S67 in the catalytic loop. On the right, the allosteric pathway opens communication from the WW and continues through K97 in the α1-helix and the S105/C113 in the α2-3 helices to K77 in the catalytic loop. We used the residues in the WW domain as the starting point and the residues in the catalytic loop as the ending points to present the shortest pathways (additional pathways are shown in Table S2). In b, the two pathways in apo-Pin1 are illustrated on the left, starting from G39, extending through L7/E35 and ending in K97 in the α1-helix. The right pathway

begins at S19, then extends onward through R21/E35 and ends at K97 in the α1-helix; c represents the pathway in I28A FFpSPR-Pin1 starts in M15 and extends through R21/E35, ending at K97 in the α1-helix. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction.

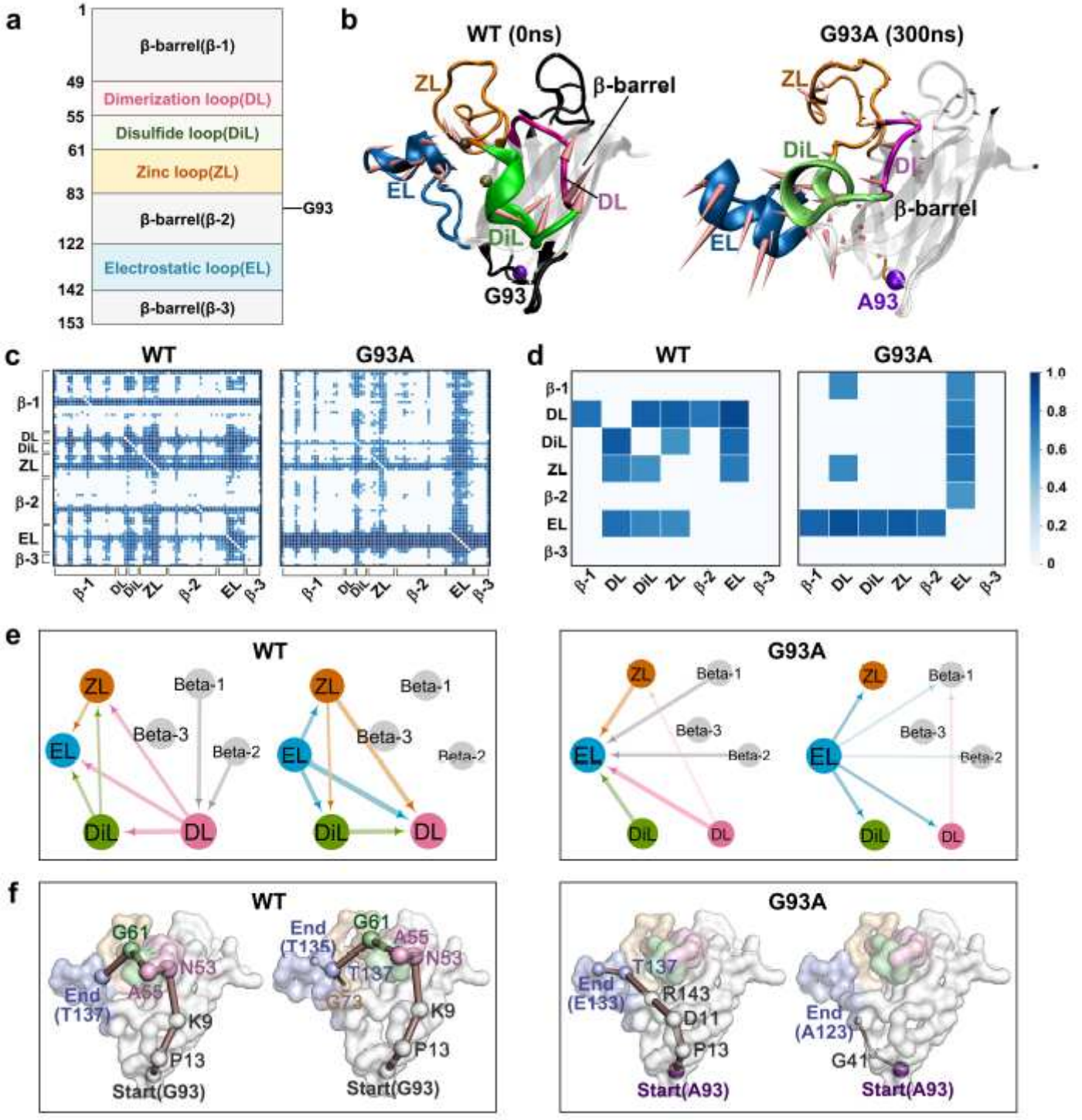

Figure 4

Change of interactions between residues/domains upon G93A mutation in SOD1. a shows the domain partitions of the SOD1 protein, which includes the position of the G93A mutation. b presents the initial

structure of the WT SOD1 and G93A SOD1 structure at 300 ns, including a β-barrel (gray), a dimerization loop (DL colored pink), a disulfide loop (DiL colored green), a zinc-binding loop (ZL in orange), and an electrostatic loop (EL in blue). The directions shown in the graphic denote the motion mode of the protein. c presents the distribution of learned edges between residues in the MD simulations of the WT (left) and the G93A (right) for SOD1. d shows a block distribution chart of learned edges between domains in the MD simulations of the WT (left) and G93A (right) for SOD1. In e, the interaction graph is mapped from the learned edges for the WT (left) and G93A (right) in SOD1. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. The arrows point toward the directionality of the learned edge. In f, the pathways from the G93 run through residues in the β-barrel, and residues in the long active loop go to the EL loop for the WT SOD1 (left); moreover, the pathways from the A93 go through residues in the β-barrel to the EL loop for the G93A SOD1 (right). The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. We used G93/A93 as the starting point, and the residues in the EL as the ending points to present the pathways.

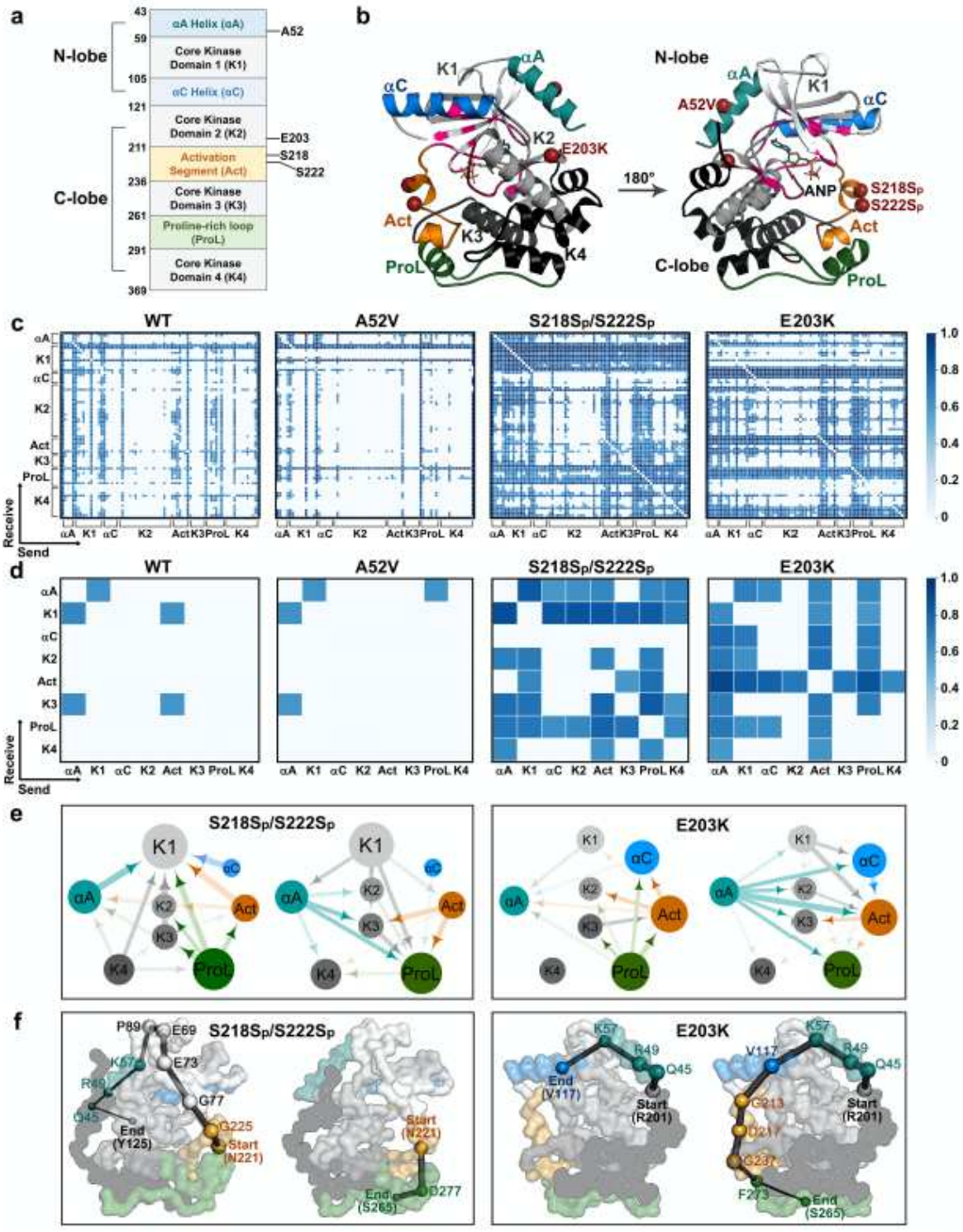

Figure 5

Changes in domain communications upon active mutations in MEK1. a shows the domain partition of MEK1 protein including the positions of mutations (A52V, S218Sp/S222Sp, and E203K). b shows two different views of the MEK1 structure. The N-terminal lobe (N-lobe) contains one core kinase (gray) and two conserved α-helices (blue). The C-terminal lobe (C-lobe) contains three core kinase domains (gray and black), an activation segment (orange), and a proline-rich loop (green). c presents the distribution of

learned edges between residues in the MD simulations of WT, A52V, S218Sp/S222Sp, and E203K MEK1. d presents the distribution of learned edges between domains in the MD simulations of WT, A52V, S218Sp/S222Sp, and E203K MEK1. In e, the interaction graph is mapped from the learned edges of active mutant MEK1. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. The arrows denote the directionality of a learned edge. In f, the allosteric pathways start from N221 in the activation segment and lead to the αA-helix and the proline-rich loop in the S218Sp/S222Sp MEK1 (left). On the (right), the allosteric pathways start from R201 (near E203K) and lead to the αC-helix and the proline-rich loop in the E203K MEK1. The size of a node represents the number of learned edges that directly connect to the node. The thickness of an edge represents the strength of the interaction. We used N221 and R201 (near E203K) as the starting points, and the residues in the αA/αC helices and the proline-rich loop as the ending points to present the pathways.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryvideo.zip
- supplementary210207submit.pdf