Quarterly Reviews of Biophysics

www.cambridge.org/qrb

Graph theory approaches for molecular dynamics simulations

Amun C. Patel[1], Souvik Sinha[1] and Giulia Palermo[1,2]

[1]Department of Bioengineering, University of California Riverside, 900 University Avenue, 92521, Riverside, CA, United States and [2]Department of Chemistry, University of California Riverside, 900 University Avenue, 92521, Riverside, CA 52512, United States

Review

Cite this article: Patel AC, Sinha S and Palermo G (2024). Graph theory approaches for molecular dynamics simulations. *Quarterly Reviews of Biophysics*, **57**, e15, 1–12
https://doi.org/10.1017/S0033583524000143

Received: 25 August 2024
Revised: 27 September 2024
Accepted: 01 October 2024

Keywords:
Allostery; Proteins; RNA; Nucleic Acids; Network Theory

Corresponding author:
Giulia Palermo;
Email: giulia.palermo@ucr.edu

Abstract

Graph theory, a branch of mathematics that focuses on the study of graphs (networks of nodes and edges), provides a robust framework for analysing the structural and functional properties of biomolecules. By leveraging molecular dynamics (MD) simulations, atoms or groups of atoms can be represented as nodes, while their dynamic interactions are depicted as edges. This network-based approach facilitates the characterization of properties such as connectivity, centrality, and modularity, which are essential for understanding the behaviour of molecular systems. This review details the application and development of graph theory-based models in studying biomolecular systems. We introduce key concepts in graph theory and demonstrate their practical applications, illustrating how innovative graph theory approaches can be employed to design biomolecular systems with enhanced functionality. Specifically, we explore the integration of graph theoretical methods with MD simulations to gain deeper insights into complex biological phenomena, such as allosteric regulation, conformational dynamics, and catalytic functions. Ultimately, graph theory has proven to be a powerful tool in the field of molecular dynamics, offering valuable insights into the structural properties, dynamics, and interactions of molecular systems. This review establishes a foundation for using graph theory in molecular design and engineering, highlighting its potential to transform the field and drive advancements in the understanding and manipulation of biomolecular systems.

Table of contents

Introduction

A network or a graph provides a structured representation of the relationships among entities within a complex system. Graphs are composed of a set of nodes/vertices or components and edges, which refers to direct interaction between nodes. Graph theory, a branch of mathematics, is then employed to analyse and study these networks, revealing global properties as well as deciphering the rules governing the local interactions (Barabási and Pósfai, 2016). For several decades, graph theory has been extensively used to study diverse systems, including social networks, transportation networks, electrical circuits, communication systems, chemical systems, and biomolecular systems. Molecular dynamics (MD) studies, which involve the simulation of the physical movements of atoms and molecules, benefit significantly from graph theoretical methods to understand complex molecular systems and predict interactions by analysing large datasets. This review explores the integration of graph theoretical models with MD simulations to enhance the understanding of complex biological phenomena, such as allosteric regulation, conformational dynamics, and catalytic functions.

The dynamic motions obtained through MD simulations can be described as graphs, representing atoms or groups of atoms as nodes, and their interactions are depicted as edges (Barabási and Pósfai, 2016). This network-based approach enables the characterization of various properties of graphs, including connectivity, centrality, and modularity, which are crucial for understanding the behaviour of molecular systems. Network models derived from graph theory have been

instrumental in elucidating the mechanisms of allosteric regulation in large biomolecular complexes, allowing for the identification of key residues or regions that play significant roles in the propagation of allosteric signals (Cui and Karplus, 2008; Dokholyan, 2016; Guo and Zhou, 2016; Liu and Nussinov, 2016; Wagner et al., 2016; Bowerman and Wereszczynski, 2016a; Wodak et al., 2019; Arantes et al., 2022). In recent years, network models of biomolecular complexes have been improved through the integration of enhanced sampling simulation methods, obtaining enhanced network models to evaluate the effect of long-timescale dynamics on the biomolecular network (East et al., 2020a). This approach allows describing how the allosteric signalling is transmitted along slow dynamical motions, which are typical of the allosteric response (Kern and Zuiderweg, 2003) and can be compared with relaxation data from solution NMR experiments (Kern and Zuiderweg, 2003; Nierzwicki et al., 2021; Skeens et al., 2024).

The application of graph theory in MD simulations has led to significant advancements in understanding the allosteric mechanisms in large protein-nucleic acid complexes. Early studies by Luthey-Schulten and colleagues implemented the combination of graph theory and MD to decipher the allosteric network in a tRNA synthase (Sethi et al., 2009). More recent applications include studies of the transcription proteins, obtaining graphs that define the signal transduction in the initiation complex (Yan et al., 2019), the nucleosome core particle, revealing how allosteric signals propagate through histone proteins to influence DNA packaging and accessibility and drug – drug synergy (Bowerman and Wereszczynski, 2016b; Adhireksan et al., 2017; Bowerman et al., 2019), and the spliceosome, with insights into the dynamic assembly of small ribonucleoproteins and RNA (Casalino et al., 2018; Saltalamacchia et al., 2020). Graph theoretical analyses integrated with MD simulations have also identified critical residues and conformational changes that govern catalysis and selectivity in the CRISPR-Cas9 system (Saha et al., 2022), a transformative tool for gene editing that relies on precise allosteric regulation for its function (Nierzwicki et al. 2020; Zuo and Liu, 2020). In this review article, we detail the application and development of graph theory-based models using the CRISPR-Cas9 genome editing system and its Cas12a and Cas13a colleagues as case studies (Saha et al., 2022). CRISPR-associated proteins (Cas) are RNA-guided enzymes that use a guide RNA to recognize and cleave any matching DNA sequence, enabling the editing of DNA and RNA (Chen and Doudna, 2017). These studies allow us to introduce the key concepts of graph theory and demonstrate their practical applications on large protein/nucleic acid complexes.

Overall, graph theory has proven to be a powerful tool for analysing molecular dynamics simulations, providing insights into structural properties, dynamics, and interactions of molecular systems. This review delves into the application of graph theoretical techniques and concepts in the analysis of molecular structures, interactions, and dynamics. We discuss the fundamental theory and report innovative approaches that are transforming the field. We showcase applications to characterize allosteric mechanisms in biomolecules, and we highlight how innovative graph-theory approaches can be used to design biomolecular systems with improved functionality. This establishes the foundations to use graph theory for molecular design and engineering.

Network models derived from graph theory

Graph theory emerged from Leonard Euler's work in the 18th century, inspired by his exploration of the seven bridges of Königsberg (Barabási and Pósfai, 2016). This mathematical discipline began with Euler's solution to whether a path could traverse each bridge exactly once, marking the inception of graph theory. Since then, the field has flourished, finding applications across various domains such as communication science (e.g., social media networks), economics, geology, and physics. Graph theory's impact extends notably to systems biology, where it models complex interactions within biological systems.

Early representations of proteins as graphs were performed by Vishveswara and co-workers reporting topological networks of biomolecules (Brinda and Vishveshwara, 2005) and by the group of Luthey-Schulten, who proposed a protocol to study the dynamic allosteric communication in biomolecules based on their studies on a tRNA synthase (Sethi et al., 2009). The authors used correlation analysis to construct graphs that represent long-range communication (Figure 1). In these dynamical network models, the biomolecular system is represented as a graph where nodes correspond to amino acids (Cα atoms) and nucleotides (P atoms, N1 in purines and N9 in pyrimidines) (Melo et al., 2020), and edges denote their connections. The length of edges reflects the degree of correlations, positioning strongly correlated nodes closer together (resulting in shorter edges). This approach fundamentally builds upon

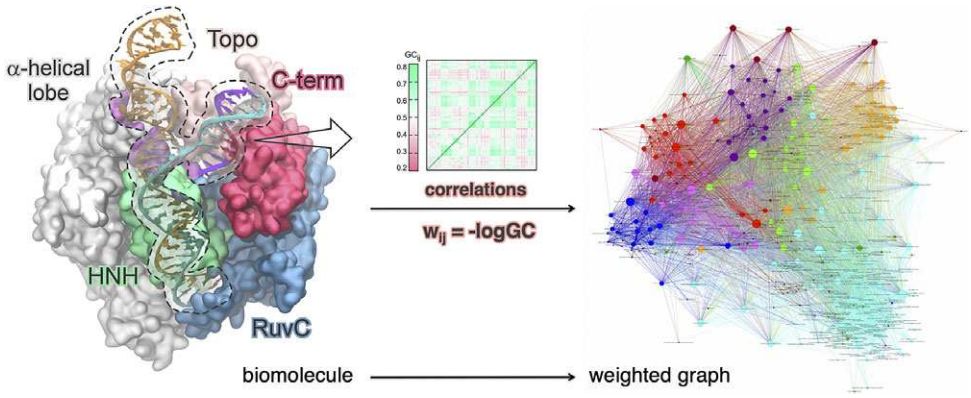

Figure 1. Biomolecular dynamic network. Overview of a biomolecular complex (a) and its representation as a network of nodes and edges (b) through correlation analysis. In panel a, the CRISPR-Cas9 system (PDB 4UN3) represents a typical protein/nucleic acid complex. Adapted with permission from Palermo et al.(Palermo et al., 2017) Copyright 2017 American Chemical Society. In panel b, a network map is shown from Pacific RISA Core Network Map Eigenvector FA2 Region 10 K (https://www.flickr.com/photos/pacificrisa/11345330443 (CC BY-NC-ND 2.0)).

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

correlation analysis for identifying dynamic correlations between spatially distant sites.

One common method is cross-correlation analysis, which involves computing Pearson's correlations between the fluctuations of Cα atoms relative to their average positions (Freddolino et al., 2013). The cross-correlation coefficients, CC_{ij}, are computed over the course of the simulation using equation (1), where Δ_{ri} and Δ_{rj} are the fluctuation vectors of the atoms i and j, respectively. The angle bracket represents an average over the sampled period. The value of CC_{ij} ranges from -1 to 1. Positive CC_{ij} values represented a correlated motion between atoms i and j, while negative CC_{ij}values describe anticorrelated motions.

$$CC_{ij} = \frac{\langle \Delta \overrightarrow{r_i}(t) \cdot \Delta \overrightarrow{r_j}(t) \rangle}{\left(\langle \Delta \overrightarrow{r_i}(t)^2 \rangle \langle \Delta \overrightarrow{r_j}(t)^2 \rangle \right)^{\frac{1}{2}}} \quad (1)$$

This approach helps reveal patterns of coordinated motion within biomolecular systems but overlooks correlated motions that occur out of phase (i.e., that are not collinear), prompting for the use of alternative correlation analysis approaches. The Generalized Correlation (GC) method (Lange and Grubmüller, 2006) assesses the correlation between residues by considering mutual information, which captures non-linear correlations. Through this method, two variables (x_i, x_j) can be considered correlated when their joint probability distribution $p(x_i,x_j)$, is larger than the product of their marginal distributions $p(x_i) \cdot p(x_j)$. The mutual information (MI) is a measure of the degree of correlation between x_i and x_j defined as a function of, $p(x_i,x_j)$ $p(x_i)$, $p(x_j)$ according to:

$$MI[x_i,x_j] = \iint p(x_i,x_j) \, ln \frac{p(x_i,x_j)}{p(x_i) \cdot p(x_j)} dx_i dx_j \quad (2)$$

Notably, MI is closely related to the definition of the Shannon entropy, $H[x]$, i.e., the expectation value of the information of a random variable x, with probability distribution $p(x_i)$:

$$H[x] = -\int p(x) \ln p(x) dx \quad (3)$$

and thus it can be computed as:

$$MI[x_i,x_j] = H[x_i] + H[x_j] - H[x_i,x_j] \quad (4)$$

Where $H[x_i]$ and $H[x_j]$ are the marginal Shannon entropies, and $H[x_i,x_j]$ is the joint entropy. Since MI varies from $0 to +\infty$, normalized generalized correlation coefficients (GC_{ij}), ranging from 0 (independent variables) to 1 (fully correlated variables), are defined as:

$$GC_{ij}[x_i,x_j] = \left\{ 1 - e^{-\frac{2MI[x_i,x_j]}{d}} \right\}^{-\frac{1}{2}} \quad (5)$$

where $d = 3$ is the dimensionality of x_i and x_j. In this approach, the marginal and joint Shannon entropies for atomic vector displacements are computed as ensemble averages over multiple trajectories. Since the correlated motions are derived from mutual information, this method effectively identifies any form of dependence in atomic motions, irrespective of their directional relationships. Building on this correlation analysis approach, one can create detailed graph models of proteins and nucleic acids that reveal insights into their structural and functional dynamics (Rivalta et al., 2012; Gasper et al. 2012; Miao et al. 2013). Correlation analysis is also extremely helpful *per se*, to identify the coupled dynamics of spatially distant sites, which is at the basis of allostery in biomolecules (Guo and Zhou, 2016). As an example, it was used to describe how DNA binding induces coupled protein dynamics and triggers activation of the Cas12a genome editing system (Figure 2a). Cas12a is an extraordinarily rapid enzyme that enables the swift and ultrasensitive detection of nucleic acids (Chen et al., 2018) through the DETECTR technology. This technology has been instrumental in detecting the SARS-CoV-2 coronavirus during the COVID-19 pandemic (Broughton et al., 2020). As part of the tremendous effort to combat COVID-19 through MD simulations (Arantes et al., 2020), extensive MD simulations of Cas12a were carried out (Rossetti et al., 2022; Saha et al., 2024; Strohkendl et al., 2024). Correlation analysis on MD trajectories of Cas12a was performed before and after DNA binding (i.e., by comparing the RNA- and DNA-bound forms).

The cross-correlations matrix (i.e., a two-by-two plot of the Cα CC coefficients) of Cas12a shows a conserved pattern of correlated/anticorrelated motions in both RNA- and DNA- bound states (Figure 2b). Notably, the recognition lobe (REC) of the enzyme preserves anticorrelated motions (i.e., $CC < 0$) with the nuclease lobe (NUC).

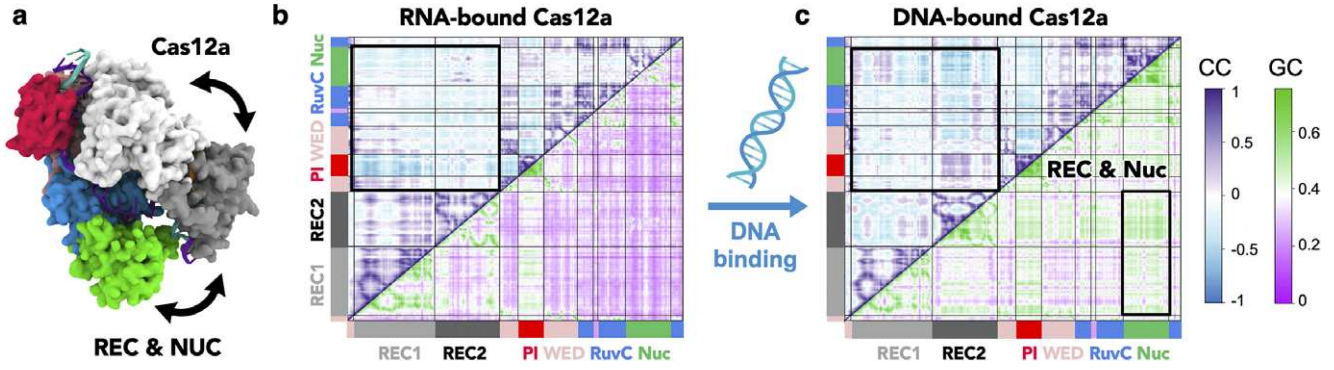

Figure 2. Correlation analysis of Cas12a. a. Overview of the CRISPR-Cas12a complex. The Cas12a protein is shown as molecular surface, highlighting the individual domains using different colours (REC1: light grey, REC2: dark grey, PAM-interacting, PI: red, RuvC: blue, Nuc: green). Nucleic acids are shown as ribbons. b-c. Cross-correlation (CC, upper triangles) and generalized correlations (GC, lower triangles) matrices were computed for Cas12a in both the RNA-bound state (b) and upon DNA binding (c). The strength of the CC and GC coefficients is represented according to the scales on the right. The protein sequence is also displayed. Boxes highlight anticorrelated motions (CC ≤ 0) and highly coupled GC between REC and NUC, which are also illustrated in the cartoon of Cas12a (a). Adapted with permission from Saha et al.(Saha et al. 2020) Copyright 2020 American Chemical Society.

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

This indicates a tendency for REC to move in the opposite direction relative to NUC, facilitating the 'open-to-close' conformational transition characteristic of Cas proteins (Palermo et al., 2016), which is essential for nucleic acid binding. The generalized correlation matrix, which goes beyond the reach of Pearson-like *CC* analysis, shows that while in the RNA-bound state of Cas12a coupled motions are mainly detected among the REC lobe, upon DNA binding, correlated motions of the REC and NUC lobes become more prominent (Figure 2c). This revealed that DNA binding induces a switch in the conformational dynamics of Cas12a, activating the distal REC and NUC lobes to enable nucleic acid cleavage. Notably, the highly coupled dynamics of the REC2 and NUC regions suggest that REC2 could regulate the nuclease function, similar to the CRISPR-associated nuclease Cas9 (Dagdas et al., 2017; Palermo et al., 2018). These mutual domain dynamics may be critical for the nonspecific binding of DNA and the mechanistic functioning of DETECTR technology (Broughton et al., 2020). Since REC is a key determinant of the system's specificity, these findings also provided a rational basis for engineering improved genome editing and viral detection tools.

Graph construction and community network analysis

As introduced above, networks are constructed by 'weighting' the edges through correlations obtained through dynamics, where the weight (w_{ij}) of the edge between nodes i and j is computed as:

$$w_{ij} = -\log[GC_{ij}] \tag{6}$$

This 'weighted graph' represents the system as a dynamic network that highlights critical nodes crucial for communication within the complex (Figure 1). Two nodes are commonly considered connected if any heavy atom of the two residues is within 5 Å of each other (i.e., distance cutoff) for at least the 75% of the simulation time (i.e., frame cutoff). These cutoffs are selected according to extensive convergence studies following community network analysis (CNA). In this analysis, a set of 'communities' can be identified as groups of nodes with dense internal connections but sparse connections between groups. These local substructures can be detected using the Girvan-Newman algorithm, a divisive method that partitions the network based on 'edge betweenness' (EB). The EB, $g(v)$, is defined as:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7}$$

where $\sigma_{st}(v)$ is the total number of shortest paths from node s to node t that crosses the edge v, whereas σ_{st} is the total number of shortest pathways existing between nodes s and t. Edges with the highest betweenness connect multiple pairs of nodes through the shortest paths, serving as critical links between different communities. Hence, pairs of nodes associated with edges of high betweenness are crucial for communication flow within the weighted network. It quantifies the 'traffic' flowing through edges, measuring how often an edge serves as a bridge in the communication flow between nodes. Using the EB parameter, the Girvan-Newman algorithm (Newman and Girvan, 2004) establishes community structure through an iterative process. Here, the edge with the highest betweenness is removed from the network, and the betweenness of the remaining edges is recalculated. As the process continues, it progressively isolates communities until the network is divided into the desired number of communities, or each node represents its own community. The optimal division of the network should ensure that each community contains nodes that are highly interconnected internally, while different communities are poorly interconnected. Toward this end, the 'modularity' parameter, denoted as Q, measures the strength or quality of the community structure and is used to determine the optimal division of the network. Q represents the difference in probability between intra-community and inter-community connections for a given network division and is defined as follows:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{8}$$

where e_{ii} is the fraction of edges that link nodes in community i to other nodes in community i, while $a_i = \sum_j e_{ij}$ is the fraction of edges that connect to at least one node in the community i. The modularity value falls in the range of 0 to 1, with larger values indicating higher community structure quality.

The convergence of the community repartitioning is important to estimate the appropriate distance and frame cutoff for CNA. The Community Repartition Difference (CRD) is defined as:

$$CRD(c_1, c_2) = 1 - \frac{\sum_{n_i,n_j} z(n_i, n_j, c_1)\, z(n_i, n_j, c_2)}{\sum_{n_i,n_j} z(n_i, n_j, c_1)} \tag{9}$$

where $z(n_i, n_j, c_i)$ is defined as 1 if nodes n_i and n_j belong to the same community in a given partition c_i (i.e., the community structure) and 0 otherwise. CRD provides a normalized count of pairs that are grouped together in two community structures, offering a reliable estimate of the similarities between different network partitions. By computing the CRD for different frame and distance cutoffs, one can evaluate the convergence of the community structure and evaluate the appropriate cutoff values.

Overall, in a typical community network visualization, communities are linked by bonds whose thickness corresponds to the total EB corresponding to the intercommunity edges, indicating the strength of communication between communities. Indeed, the total EB between pairs of communities serves as a significant indicator of their communication strength. This metric helps in understanding the extent to which communities within a network are interconnected and how strongly they influence each other's dynamics and functions.

Community network analysis (CNA) was applied to the CRISPR-Cas9 system to characterize the allosteric role of the Protospacer Adjacent Motif (PAM) which is a short DNA sequence that facilitates targeting of the desired DNA sequence across the genome by the Cas9 enzyme (Figure 3a) (Jinek et al., 2012). Biochemical studies indicated that in CRISPR-Cas9, PAM binding activates the concerted function of the two catalytic domains, HNH and RuvC, which are located distally from PAM (Sternberg et al., 2014, 2015). Upon PAM binding, the striking conformational plasticity of HNH would activate its nuclease function while simultaneously activating the other RuvC nuclease, leading to the cleavage of both of the DNA strands through a mechanism that involves metal ions (Casalino et al., 2020; Nierzwicki et al., 2022). However, the allosteric mechanism by which PAM binding communicates with the HNH and RuvC domains remained unknown. CNA identified communities of closely correlated residues and quantified the strength of correlations between them, represented by a set of nodes and edges weighted according to GCs (*vide supra*) (Palermo et al., 2017). When comparing the structural communities in Cas9 with and without PAM (wPAM and w/oPAM, respectively), CNA revealed that PAM

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

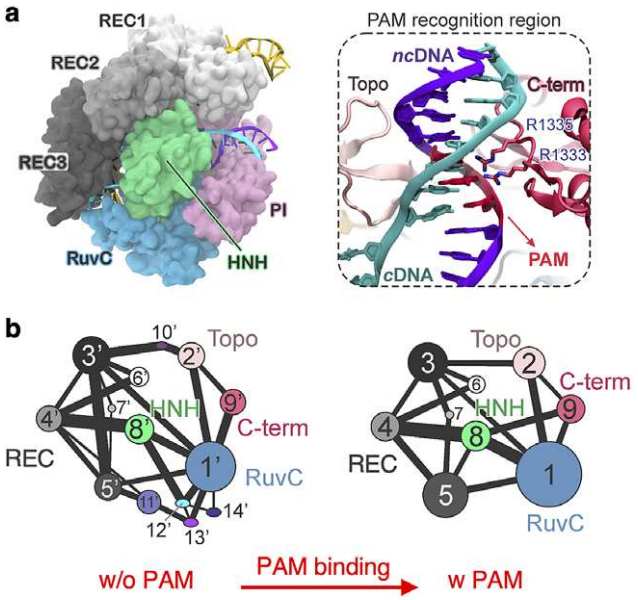

Figure 3. Community Network Analysis (CNA) of CRISPR-Cas9. a. Overview of the CRISPR-Cas9 system, highlighting the individual domains using different colours (α-helical lobe: light grey, PAM-interacting C-terminal: red, RuvC: blue, HNH: green). Nucleic acids are shown as ribbons. A close-up view shows the PAM recognition region, highlighting the PAM sequence in red). b. CNA of CRISPR-Cas9 in the absence of PAM (without PAM, w/oPAM, left) and upon PAM binding (with PAM, wPAM, right), shown in a 2D representation of the community network. Bonds connect communities and measure their intercommunication strength. Adapted with permission from Ricci et al.(Ricci et al., 2019) Copyright 2019 American Chemical Society; and from Palermo et al.(Palermo et al., 2017) Copyright 2017 American Chemical Society.

binding reduces the number of communities (Figure 3b), thereby enhancing allosteric signalling.

In the absence of PAM, the communities are significantly more fragmented, weakening the essential correlations for effective allosteric signalling. Additionally, PAM strengthens the correlation between communities #1 and #8, which comprise the RuvC and HNH domains, respectively. This is depicted in Figure 3b by a thicker bond between these communities. The connection between communities #1 and #8 is notably weaker in the absence of PAM due to the increased fragmentation. Thus, PAM clearly induces a stronger communication channel between the HNH and RuvC domains, activating their catalytic activity and offering a rationale for previous biochemical studies (Sternberg et al., 2014, 2015).

Shortest path calculations

Network analysis provides valuable insights by identifying 'shortest pathways' between distally located node pairs using algorithms like Floyd-Warshall (Floyd, 1962). The shortest path calculation involves finding the path between a pair of nodes such that the sum of weights of constituent edges is minimized. These pathways often represent efficient communication routes among allosteric sites and help identify 'signal transducers' in biomolecular systems. The Floyd-Warshall algorithm sums the edge lengths (w_{ij}) across different paths to identify the shortest path. It is a well-established tool for shortest-path calculations, which finds the optimal route for all node pairings. Another useful algorithm for shortest path calculations is the Dijkstra algorithm (Dijkstra, 1959), commonly used in cartography to find the shortest routes to destinations. As opposed to Floyd-Warshall, which is best for finding the shortest path between all pairs of nodes, especially in a dense network, Dijkstra is best when the shortest path between a single node and all other nodes is required, especially for sparse graphs. The algorithm's shortest path search begins with defined starting and destination points and aims to minimize the total distance (maximize correlation) between nodes connected by (w_{ij}) inter-node connections (Figure 4a). Together, these shortest path algorithms are remarkable for uncovering the potential pathways through which molecular signals propagate. It is important to note, however, that in large biological systems, the number of possible pathways between distant nodes grows with the interconnectedness of nodes. Moreover, signalling transfer within biomolecular systems does not always follow a single optimal path; instead, it can involve a variety of alternative sub-optimal routes. For this reason, shortest path analysis commonly integrates the top sub-optimal routes alongside the shortest 'optimal' pathway in the evaluation of the signal transduction.

Shortest path analysis through the Dijkstra algorithm was used to find the shortest pathways that connect the sites of DNA binding (i.e., three recognition domains, REC1–3) with the nuclease domains (i.e., HNH and RuvC) in CRISPR-Cas9 (Figure 4b). In this study, shortest path analyses were performed on MD trajectories obtained through enhanced sampling, using a Gaussian accelerated MD (GaMD) approach (Miao et al., 2015; Wang et al., 2021), to improve the configurational sampling and access long-timescale dynamics, which is critical for protein allostery (Kern and Zuiderweg, 2003). This enhanced network model was used to evaluate the effect of long-timescale dynamics on the biomolecular network and was compared with the slow dynamical motions measured through solution NMR in the isolated Cas9 domains (East et al., 2020a). The computed pathways consisted of residue-to-residue steps that optimize the overall correlation between amino acids 789/794 and 841/858, which belong to the HNH domain but are adjacent to RuvC and REC2, respectively. This calculation provided an estimation of the principal channels of communication between RuvC and REC2. Notably, the top-ranked shortest pathway that maximizes the dynamic transmission between RuvC and REC2 through HNH shows a ~ 70% overlap with the pathway experimentally identified in an isolated construct of the HNH domain using NMR CPMG relaxation dispersion. We note that, while the experimental approach is limited to the isolated domain, the computation of the optimal pathways considers the full-length complex, offering an interpretation of HNH allosterism within the context of the Cas9 protein bound to RNA and DNA (East et al., 2020b).

Building on the agreement between NMR and the theoretical approach, the pathway connecting REC-HNH-RuvC was used as a reference to measure alterations in the allosteric communication in a *Geobacillus stearothermophilus* Cas9 species that is functional in human plasma and opens new avenues for applications in the genome editing field (Belato et al., 2022a; Belato et al., 2022b). Alterations in the allosteric pathway connecting REC and HNH were also observed in the presence of mutations in the REC3 domain that increase the specificity of Cas9 against off-target effects (Ricci et al., 2019; Mitchell et al., 2020; Skeens et al., 2024).

These findings indicated that CRISPR-Cas9 achieves altered selectivity by modulating its allosteric communication, a concept that paved the way for successful engineering toward improved gene editing (Chen et al., 2017; Schmid-Burgk et al., 2020). Furthermore, the agreement between shortest-path calculations and CPMG relaxation dispersion underscores the effectiveness of using shortest-path calculations to analyse allosteric pathways (East et al., 2020b).

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

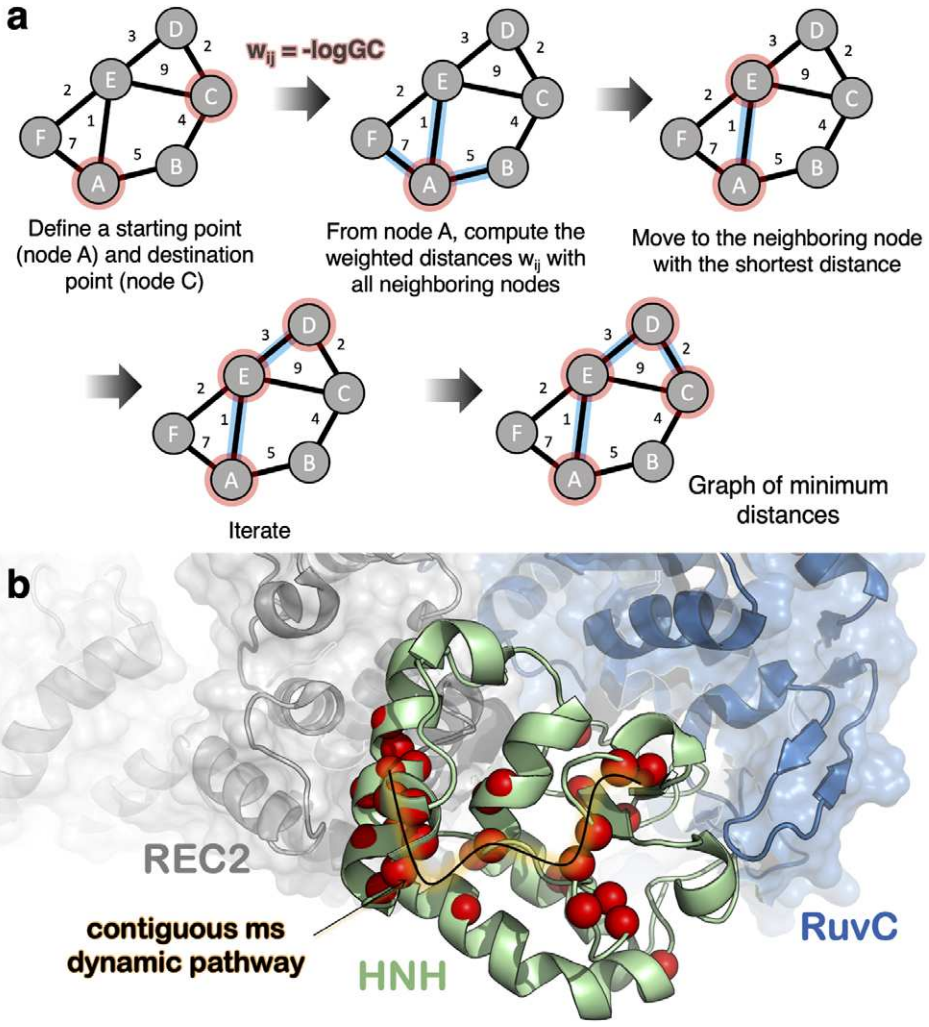

Figure 4. Shortest path calculation. a. The Dijkstra algorithm is used for shortest path calculation by defining a starting point and a destination (i.e., nodes A and C) and iteratively optimizing the path from the former to the latter. In each iteration, the algorithm designates the closest unvisited node as the current node, updating the distances to the remaining unvisited nodes until the destination is reached. For biomolecular allostery, the algorithm employs correlation coefficients as metrics to identify the closest nodes (i.e., $w_{ij} = -logCG$) thereby maximizing the correlation between the starting and destination nodes. b. In the context of the HNH domain of CRISPR-Cas9, the Dijkstra algorithm identifies an allosteric pathway connecting the DNA recognition region (REC2) to the RuvC cleavage site. The signaling route identified through this algorithm (illustrated by the pink line) overlaps with the slow dynamic residues found through solution NMR (represented by purple spheres). Adapted with permission from East et al.(East et al., 2020a) Copyright 2020 American Chemical Society.

Networks of communication gain and loss

As mentioned earlier, edge betweenness (EB) is a crucial metric for assessing the 'traffic' through network edges. It has been further employed to create circular networks depicting mutation-induced allosteric gain or loss. In a detailed study investigating the improved specificity in CRISPR-Cas9 obtained through three lysine-to-alanine mutations (K810A, K855A, K848A, Figure 5a) (Slaymaker et al., 2016), the mutation-induced change in EB (ΔEB) was calculated as the difference between the EB of the mutant and wild-type (WT) systems (Nierzwicki et al., 2021). The normalized ΔEB values were visualized using circular networks, where the HNH communities computed through Community Network Analysis (*vide supra*) are arranged in a circle and connected by links with thickness proportional to the ΔEB (Figure 5b).

Communities hosting the residues that form the allosteric pathways, i.e., the 'allosteric communities' (A), were defined based on the agreement between CPMG relaxation dispersion and shortest path analysis (Figure 4) and distinguished from non-allosteric communities (NA). Negative ΔEB values (ΔEB < 0, blue) indicate a loss of communication, while positive values (ΔEB > 0, red) signify a communication gain due to the mutation. The study revealed a significant loss of communication between the allosteric routes that connect the functional sites (i.e., the A1–A3 communities Figure 5b). Conversely, non-allosteric sites (NA1–NA4) showed increased communication, suggesting that mutations enhancing Cas9 specificity also disrupt its allosteric signalling. To experimentally validate this observation, an 'NMR-derived network analysis' that fully integrates NMR relaxation dispersion with MD and graph theory was introduced (Figure 5c). Here, the computational communities were used as a reference, while the dynamic exchange among them was based on NMR, confirming a loss in crosstalk between allosteric communities. This approach showed that the K-to-A mutations dramatically reduce the dynamic exchange between allosteric communities in HNH, corroborating the theoretical outcomes. In summary, these networks effectively highlight allosteric communication changes induced by mutations in biomolecular systems.

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

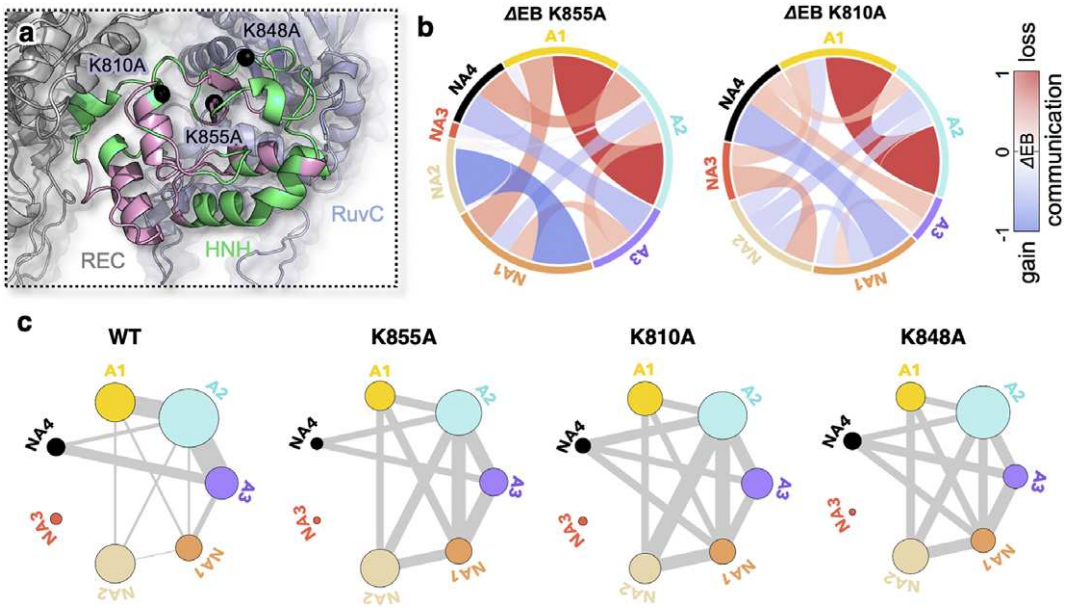

Figure 5. Circular Networks. a. Allosteric pathway of information transfer (pink) spanning HNH (green) from the DNA recognition lobe (REC) to the RuvC core. The K810A, K855a, and K848A enhancing specificity mutations are indicted. b. Circular networks of mutation-induced edge betweenness change (ΔEB) noting gain (blue) or loss (red) in allosteric crosstalk between MD-derived communities (shown for the K855A and K810A mutants). HNH communities are plotted on a circle, connected through links the thickness of which is proportional to ΔEB. **c**. Networks integrating the MD-derived communities (circles), with the experimental dynamic exchange among them (bonds with thickness proportional to CPMG relaxation dispersion NMR upon normalizing the number of flexible residues in each community). Adapted with permission from Nierzwicki et al. (Nierzwicki et al., 2021) 2021 eLIFE.

Signal-to-noise ratio (SNR) of communication efficiency

Traditional shortest-path measurements obtained from the dynamic network serve a valuable purpose in identifying the most probable communication pathways between predefined sites (Sethi et al., 2009). However, they fall short of providing insight into how these pathways compare within the broader communication network, as assessing the favourability is crucial for identifying dominant allosteric communication pathways. To address this gap, we recently introduced a Signal-to-Noise Ratio (SNR) metric (Sinha et al., 2024), which quantifies the preference for communication between pathways of similar length in the network (the "noise") connecting distant sites (the "signal") (Figure 6). In this approach, the 'noise' is computed as the distribution of the sum of edge betweennesses, EB, $g(v)$, (i.e., the cumulative betweennesses, measuring the total traffic passing through the edges) among all residues and nucleobases. The 'signal' is derived from the distribution of cumulative edge betweennesses of pathways connecting the regions of interest. In detail, the cumulative betweennesses of each pathway (S_k) is calculated as the sum of the betweennesses of all the constituent edges in that specific pathway:

$$S_k = \sum_{i=1}^{n-1} g_k(i) \tag{10}$$

where $g_k(i)$ is the edge betweenness of the edge in the $\boldsymbol{k^{th}}$ between node i and $i+1$, and n is the number of edges in the k^{th} pathway. Then, SNR is determined as:

$$SNR = \frac{\frac{E[S]}{Var(S)}}{\frac{E[N]}{Var(N)}} \tag{11}$$

where $E\left[\frac{S}{N}\right]$ and $Var\left(\frac{S}{N}\right)$are the expectation and variance of the signal/noise distribution, respectively. High SNR values indicate the preference of the network to communicate through the signal over other noisy routes. We recently introduced the SNR to identify the allosteric signalling responsible for the activation of Cas13a (Figure 6a), an RNA-targeting protein that shows promise for RNA detection and imaging (O'Connell, 2019). Cas13a uses a CRISPR RNA (crRNA) to target RNA sequences and trigger the catalytic action of a composite active site, which is distally located with respect to the sites of RNA binding and is formed by two 'Higher Eukaryotes and Prokaryotes Nucleotide' (HEPN) domains, cleaving any solvent-exposed RNA.

To establish how the allosteric communication controls the RNA cleavage activity in Cas13a, multi-microsecond MD simulations, reaching almost ~170 μs of sampling, were performed. SNR analysis was performed over the obtained trajectories to elucidate the signalling efficiency between the guide RNA and catalytic sites. This analysis showed that, in the inactive protein (i.e., the binary complex formed by the protein bound to the crRNA), the signal overlaps with the noise (Figure 6b) resulting in lower SNR values. Upon target RNA binding, the signal stands out over the noise, indicating remarkable communication efficiency, and identifies the region of the guide RNA sourcing most allosteric communications (i.e., the switch region, displaying high SNR values).

This explains previous experimental observations reporting that the binding of the target RNA to the switch region of the guide RNA results in the activation of the protein toward RNA cleavage. This approach also pinpointed the critical activation hotspots in the protein (R377, N378, and R973, Figure 6c). We have also shown that alanine mutation of these residues increases sensitivities to single-nucleotide mismatches. These variants have also been tested for detecting singlenucleotide polymorphisms in SARS-CoV-2 variants, demonstrating their potential for disease diagnostics. This has paved new avenues for the development of highly selective RNA-based cleavage and detection tools. Building on these findings, the SNR emerges as a useful tool to distinguish allosteric signals from non-allosteric inter-residue communications in

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

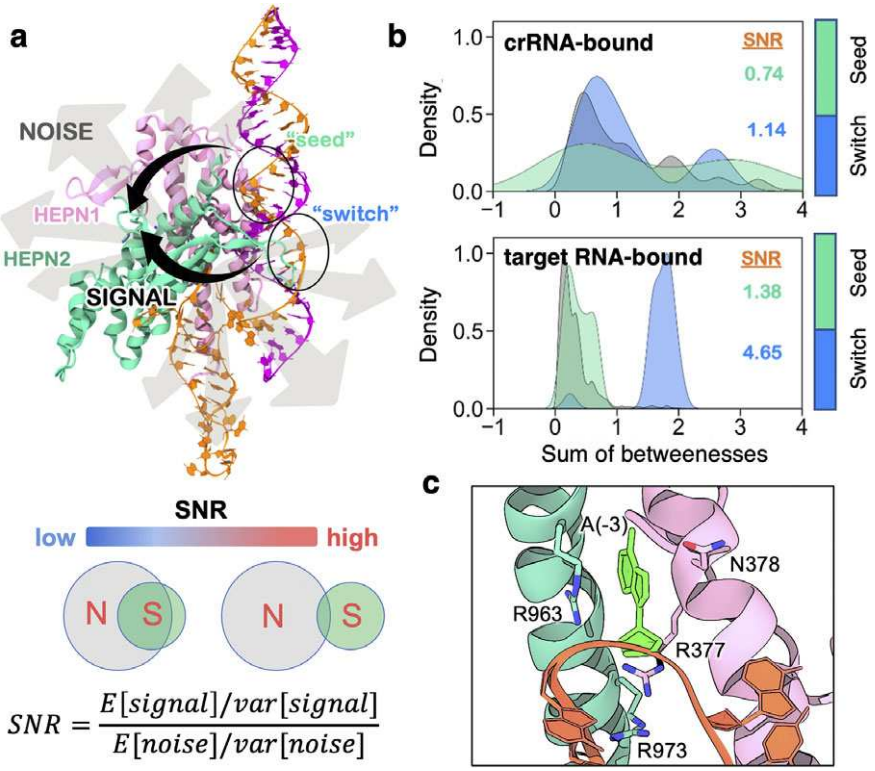

Figure 6. Signal-to-Noise Ratio of communication efficiency. a. Schematic of the Signal-to-Noise Ratio (SNR) of communication efficiency on the 3D structure of Cas13a. Two black arrows indicate the signal standing up over the noise, shown using grey arrows. High SNR indicates that the signal stands out over the noise. b. Distribution of the signals from the crRNA "seed" (green) and "switch" (blue) regions to the catalytic core residues, plotted on the background of noise (grey) in the crRNA- (top) and target RNA-bound (bottom) Cas13a. **c** Sites of increased specificity in Cas13a identified through computational analysis and tested through mutagenesis and DNA cleavage experiments. Adapted with permission from Sinha et al.(Sinha et al., 2024) Copyright 2023, Published by Oxford University Press on behalf of Nucleic Acids Research.

biomolecular complexes and identify critical hotspots for mutational analysis aimed at improving biomolecular function and specificity.

Centrality analysis

Centrality is a fundamental concept in network theory, illustrating the relative influence of a node or cluster of nodes within a network (Barabási and Pósfai, 2016). Its application in graph theory, particularly in social media networks, underscores its importance in information flow. In social networks, nodes with numerous connections act as hubs where information centralizes and transfers efficiently. Similarly, in biomolecular systems, residues that serve as hubs govern the system's behaviour. Three primary measures define centrality in a network. Degree Centrality (DC) is computed as the number of edges connected to a node, serving as a local centrality measure. Betweenness Centrality (BC) quantifies the number of shortest paths passing through a node, indicating how often a node acts as a bridge between others. Eigenvector Centrality (EC) measures a node's influence based on its connections and the influence of those connections. The EC is derived from the first eigenvector of the adjacency matrix A, which is a square matrix used to describe connections between vertices in a graph. The EC of a node, c_I, is the sum of the centralities of all nodes that are connected to it by an edge:

$$c_i = \frac{1}{\lambda}\sum_{j=1}^{n} A_{ij} c_j \qquad (12)$$

where the edges A_{ij} are elements of the adjacency matrix A and λ is the eigenvalue associated with the eigenvector composed by c_I elements. EC quantifies the connectivity of each amino acid or nucleobase within the system, identifying elements that significantly influence the network. It provides a normalized measure of connectivity, facilitating comparisons across different conditions such as mutations or effector binding (CFA et al., 2018).

The EC was used to provide a comparable measure of the allosteric signalling among the three variants of the Cas9 nuclease. These variants, namely VQR, VRER, and EQR, have introduced the ability to recognize alternative PAM sequences (Figure 7a) (Kleinstiver et al., 2015) as opposed to wild-type Cas9 (WT Cas9) that recognizes only the canonical *5'*-NGG-*3'* PAM sequence (where N can be any base) significantly constraining its application towards recognizing a wide-array of genome sequences. These variants recognize *5'*-NGA-*3'*, *5'*-NGAG-*3'*, and 5'-NGCG-3' PAM sequences through mutations of the PAM-interacting (PI) domain, remarkably expanding the DNA targeting capability of Cas9. The EC distribution analysis plotted on the 3D structure of the Cas9 variants (Figure 7b) shows that the most relevant domains in terms of correlation with the overall motion of the system are REC1 – 3 HNH, and RuvC. Higher EC values are associated with HNH and REC2. This is consistent with Community Network Analysis (Figure 7c), showing that the interconnection between REC2 and HNH is the strongest in all variants (i.e., thicker bonds connect communities #8 and #4). This evidence also agrees well with the EC distribution for the WT Cas9 (Figure 7d) and with single-molecule FRET experiments, indicating that in the WT Cas9 the motions of REC2 regulate HNH

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

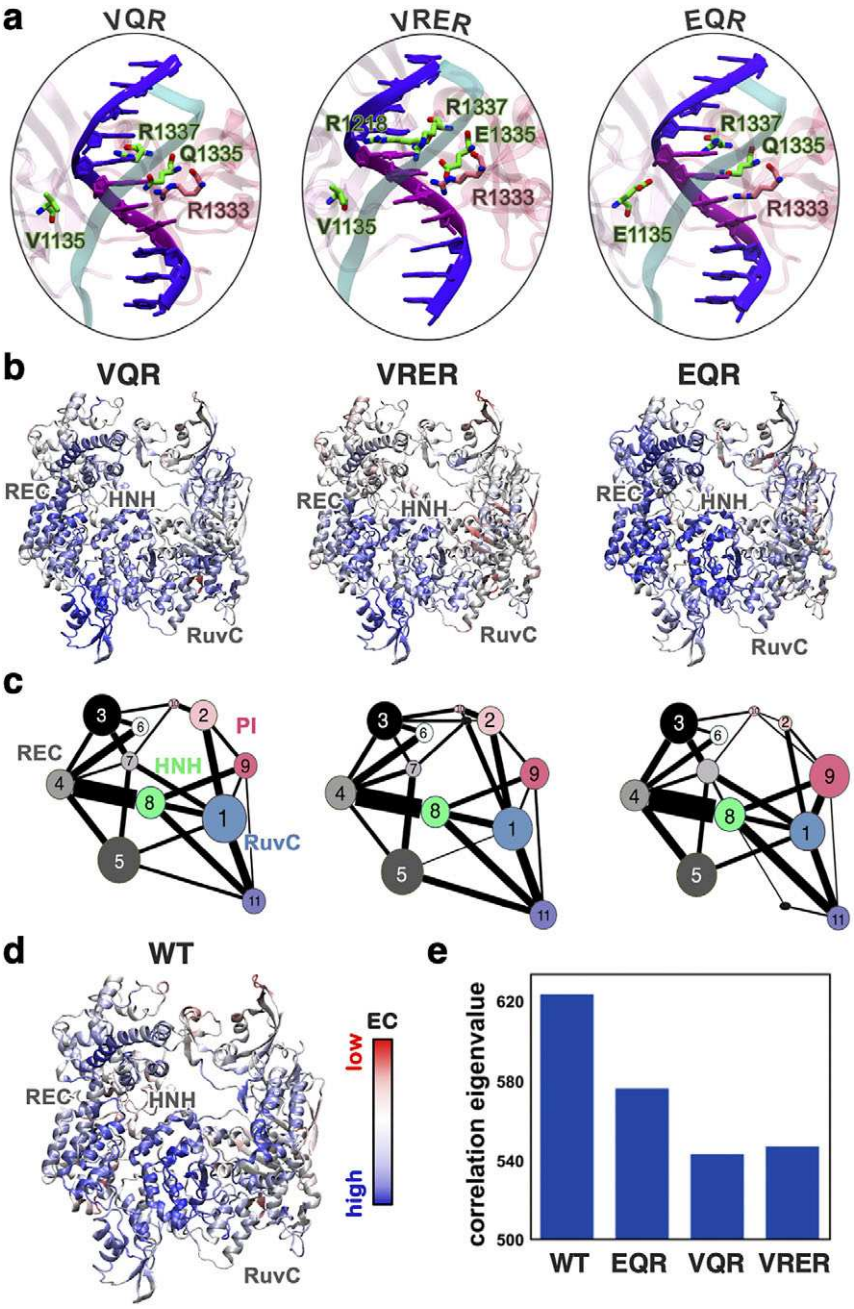

Figure 7. Eigenvector centrality analysis of Cas9 variants. a. Close-up view of the PAM binding domain in three variants of Cas9 (*viz.*, VQR, VRER and EQR) with mutations in the PAM binding region(Kleinstiver et al., 2015). The DNA target is shown as ribbons, highlighting the PAM sequence in magenta. The Cas9 residues mutated to alter the protein's selectivity (green) and the residues preserved (pink) are shown as sticks. b. Eigenvector centrality distribution plotted on the 3D structure of the Cas9 variants, color-coded from red (lowest EC) to blue (highest EC). c. Community structures for the EQR, VQR and VRER Cas9 variants. d. EC distribution shown for the WT Cas9, coloured according to the colour scale on the right. e. Highest eigenvalues of the generalized correlation matrix for the WT Cas9 and its variants.

(Chen et al., 2017; Dagdas et al., 2017; Sung et al., 2018). Noteworthy, the eigenvalues associated with the EC distribution (Figure 7e) follow the same activity trend experimentally measured for these mutants. Indeed, Kleinstiver et al. have shown that the activity of these variants in human cells for the *5'*-NGG-*3'* PAM seems to be WT > EQR > VQR ~ VRER (Kleinstiver et al., 2015), as also confirmed by independent studies (Hirano et al., 2016). The eigenvalues associated with the EC distribution are a quantitative measure of the overall correlation of the system, indicating the efficiency of the communication.

In the Cas9 variants, a decreased EC distribution strongly indicates a lower degree of communication, reflecting the decreased experimental activity of these systems. Hence, single point mutations in the PI domain affect the system's communication, strengthening the notion that PAM acts as an allosteric effector of the Cas9 dynamics. Overall, these findings show that the EC analysis allows reliable comparisons of the system connectivity in comparison with mutated systems. Moreover, EC helps identify the main mode of collective correlations in the network, making it a valuable tool for understanding the connectivity in biomolecular systems.

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

Perspective applications

Here, we propose a series of potential applications aimed at advancing the use of graph theory to elucidate biophysical mechanisms and inform drug design.

Graph theory helps pinpoint critical residues or domains that serve as hubs for functional interactions. This can be applied to map interaction pathways between key players in respiratory chains, predicting how perturbations or mutations might disrupt these processes. Complex I of the mitochondrial respiratory chain is essential for electron transport and proton pumping. Notably, the ubiquinone reduction site – crucial for Complex I function – extends nearly 200 Å from the proton channels, raising questions about the molecular origin of this long-range coupling (Sharma et al., 2015). Graph representations can be employed to identify structural motifs or key residues that mediate this coupling. By analysing the connectivity within the complex, critical pathways for energy transfer and signal propagation during the coupling event can be elucidated.

Applying graph theory could also be instrumental in identifying efficient pathways for electron transfer in biomolecules (Gray and Winkler, 1996). For instance, it could provide insights into the plastocyanin-cytochrome f interaction, where subtle conformational changes in one protein facilitate efficient electron transfer within the photosynthetic electron transport chain (Cruz-Gallardo et al. 2012). Additionally, graph theory can be used to model and compare the effects of mutations or environmental changes (such as pH or ionic strength) on the interaction network, offering a deeper understanding of how these factors impact electron transfer efficiency. Theoretical studies of electron transfer necessitate a quantum mechanical description, and the integration of this framework with graph theory offers promising insights (Arantes et al., 2022) to elucidate, for instance, how long-range effects influence the evolution of chemical reactions (Brunk et al., 2011). By merging graph theory with quantum mechanical descriptions, we can gain valuable insights into the mechanisms of electron transfer and their broader implications for cellular processes.

Allosteric regulation in large RNAs, such as group II introns and the spliceosome, involves long-range interactions where changes in one region of the RNA can influence distant sites (Casalino et al. 2016, Casalino et al., 2018; Saltalamacchia et al., 2020). In these systems, RNA splicing entails several conformational changes, driven by both RNA – RNA and RNA-protein interactions. Graph-based analyses can aid in deciphering RNA folding pathways and elucidating how communication between distinct regions affects catalysis, intron cleavage, and exon ligation. This approach could enhance our comprehensive understanding of the structure – function relationships within these complex RNA systems.

Combining graph-theoretical approaches with MD simulations enables the identification of critical nodes (key amino acids or domains) that may regulate function or serve as allosteric drug targets (Bernetti et al., 2024). This approach is particularly valuable for discovering new drug targets, especially in complex systems characterized by allosteric responses and feedback mechanisms. Notable examples include imidazole glycerol phosphate synthase (Rivalta et al., 2012; Calvó-Tusell et al., 2022), which is a target for the development of antifungal, antibacterial, and herbicidal agents, and the signal-transducing GTPase K-Ras, a quintessential example of a small yet allosterically complex protein that is highly relevant in oncology (Castelli et al., 2024).

In summary, graph theory provides an in-depth analysis of connectivity and interaction networks, helping to unravel the mechanistic details of various biological functions. These include electron transfer in respiratory and photosynthetic chains, the conformational dynamics of complex RNA – RNA and RNA-protein interactions, and the allosteric mechanisms of drug targets. These insights have practical applications in fields like bioenergetics, drug discovery, and the design of biomolecular machines.

Outlook and challenges

Graph theory has significantly deepened our understanding of complex molecular systems, providing valuable insights into information transfer and the structure of biomolecular interaction networks. Challenges in graph theory are particularly pronounced when modeling the intricate relationships between cause and effect, especially in systems where multiple variables interact. While graph theory is invaluable for representing these relationships, inferring causation demands careful consideration of underlying assumptions, such as the presence or absence of key effectors. One approach to constructing causal graphs involves using causal inference methods, such as Directed Acyclic Graphs (DAGs) (Barabási and Pósfai, 2016). In a DAG, edges have a specified direction, and the graph is acyclic, meaning it is impossible to start at one node and follow a sequence of directed edges that leads back to the same node. In causal models, DAGs effectively represent causal relationships between variables, with each directed edge indicating a causal influence. In probabilistic models, DAGs can be employed to represent dependencies between random variables, where nodes signify variables and edges denote conditional dependencies.

Applying DAGs to decipher causation in biomolecular dynamics is both highly challenging and innovative, as it has yet to be fully realized in complex biomolecular systems. Key challenges include the ability to manage large datasets with numerous variables and the complexity of high-dimensional data, typical in MD simulations. This complexity can obscure causal discovery due to the vast number of potential relationships. Moreover, creating clear, interpretable visualizations and explanations of causal relationships that accurately reflect the underlying mechanisms adds another layer of difficulty. These challenges underscore the need for interdisciplinary approaches that bridge graph theory, statistics, and causal inference to develop effective solutions.

Conclusions

In conclusion, the integration of graph theory with molecular dynamics simulations has significantly advanced our understanding of complex molecular systems, particularly in the context of allosteric regulation, conformational dynamics, and catalytic functions. By providing a structured framework to represent and analyse the interactions and dynamics within biomolecular networks, graph theory has enabled the identification of key residues and pathways critical to these processes. The application of these methods to systems like the CRISPR-Cas9 genome editing tool underscores the potential of graph theory not only to elucidate biological mechanisms but also to guide the design and engineering of biomolecular systems with enhanced functionality. As this review highlights, ongoing developments in graph theoretical approaches promise to further transform the field, offering new avenues for the exploration and manipulation of complex biological phenomena.

Acknowledgments. This material is based upon work supported by the National Institutes of Health (Grant No. R01GM141329 to G.P.) and the National Science Foundation (Grant No. CHE- 2144823 to G.P.). GP

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press

acknowledges support by the Sloan Foundation (Grant No. FG-2023-20431) and the Camille and Henry Dreyfus Foundation (Grant No. TC-24-063). The computational studies reviewed here were carried out using Expanse at the San Diego Supercomputing Center through allocation MCB160059 and Bridges2 at the Pittsburgh Supercomputer Center through allocation BIO230007 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation supports grants #2138259, #2138286, #2138307, #2137603, and #2138296. Computer time was also provided by the National Energy Research Scientific Computing Center (NERSC) under Grant No M3807.

Competing interest. The authors declare no competing financial interest.

Additional information. Analysis codes and script files for the analyses presented in this article can be downloaded from Github: https://github.com/palermolab.

References

Adhireksan Z, **Palermo G**, **Riedel T**, ***et al.*** (2017) Allosteric cross-talk in chromatin can mediate drug-drug synergy. *Nature Communications* **8**, 14860.

Arantes PR, **Patel AC**, **Palermo G** (2022) Emerging Methods and Applications to Decrypt Allostery in Proteins and Nucleic Acids. *Journal of Molecular Biology* **434**, 167518.

Arantes PR, **Saha A**, **Palermo G** (2020) Fighting covid-19 using molecular dynamics simulations. *ACS Central Science* **6**, 1654–1656.

Barabási A-L, **Pósfai M** (2016) *Network science*. Cambridge: Cambridge University Press.

Belato HB, **D'Ordine AM**, **Nierzwicki L**, ***et al*** (2022a) Structural and dynamic insights into the HNH nuclease of divergent Cas9 species. *Journal of Structural Biology* **214**, 107814.

Belato HB, **Norbrun C**, **Luo J**, ***et al*** (2022b) Disruption of electrostatic contacts in the HNH nuclease from a thermophilic Cas9 rewires allosteric motions and enhances high-temperature DNA cleavage. *Journal of Chemical Physics* **157**, 225103.

Bernetti M, **Bosio S**, **Bresciani V**, ***et al*** (2024) Probing allosteric communication with combined molecular dynamics simulations and network analysis. *Current Opinion in Structural Biology* **86**, 102820.

Bowerman S, **Hickok RJ**, **Wereszczynski J** (2019) Unique Dynamics in Asymmetric macroH2A–H2A Hybrid Nucleosomes Result in Increased Complex Stability. *Journal of Physical Chemistry B* **123**:419–427.

Bowerman S, **Wereszczynski J** (2016a) Detecting Allosteric Networks Using Molecular Dynamics Simulation. *Methods in Enzymology* **578**, 429–447.

Bowerman S, **Wereszczynski J** (2016b) Effects of MacroH2A and H2A.Z on Nucleosome Dynamics as Elucidated by Molecular Dynamics Simulations. *Biophysical Journal* **110**, 327–337.

Brinda KV, **Vishveshwara S** (2005) A Network Representation of Protein Structures: Implications for Protein Stability. *Biophysical Journal* **89**, 4159–4170.

Broughton JP, **Deng X**, **Yu G**, ***et al*** (2020) CRISPR–Cas12-based detection of SARS-CoV-2. *Nature Biotechnologies* **38**, 870–874.

Brunk E, **Ashari N**, **Athri P**, ***et al.*** (2011) Pushing frontiers of first-principles based computer simulations of chemical and biological systems. *Chimia (Aarau)* **65**, 667–671.

Calvó-Tusell C, **Maria-Solano MA**, **Osuna S**, **Feixas F** (2022) Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. *Journal of the American Chemical Society* **144**, 7146–7159.

Casalino L, **Nierzwicki Ł**, **Jinek M**, **Palermo G** (2020) Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by Ab Initio Molecular Dynamics. *ACS Catalysis* **10**, 13596–13605.

Casalino L, **Palermo G**, **Rothlisberger U and Magistrato A** (2016) Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group IIIntrons. *J. Am. Chem. Soc.* **138**, 10374–10377.

Casalino L, **Palermo G**, **Spinello A**, ***et al.*** (2018) All-atom simulations disentangle the functional dynamics underlying gene maturation in the intron lariat spliceosome. *Proceedings of the National Academy of Sciences U S A* **115**, 6584–6589.

Castelli M, **Marchetti F**, **Osuna S**, ***et al.*** (2024) Decrypting Allostery in Membrane-Bound K-Ras4B Using Complementary *In Silico* Approaches Based on Unbiased Molecular Dynamics Simulations. *Journal of the American Chemical Society* **146**, 901–919.

Negre CFA, **Morzan UN**, **Hendrickson HP**, ***et al.*** (2018) Eigenvector Centrality Distribution for Characterization of Protein Allosteric Pathways. *Proceedings of the National Academy of Science U S A* **115**, 12201–12208.

Chen JS, **Dagdas YS**, **Kleinstiver BP**, ***et al.*** (2017) Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410.

Chen JS, **Doudna JA** (2017) The Chemistry of Cas9 and its CRISPR Colleagues. *Nature Reviews Chemistry* **1**, 78.

Chen JS, **Ma E**, **Harrington LB**, ***et al.*** (2018) CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436–439.

Cruz-Gallardo I, **Díaz-Moreno I**, **Díaz-Quintana A**, **De la Rosa MA** (2012) The cytochrome f–plastocyanin complex as a model to study transient interactions between redox proteins. *FEBS Letters*, **586**, 646–652.

Cruz-Gallardo I, **Diaz-Moreno I**, **Diaz-Quintana A**, **De la Rosa MA** (2012) The cytochrome f–plastocyanin complex as a model to study transient interactions between redox proteins. *FEBS Letters*, **586**, 646–652.

Cui Q, **Karplus M** (2008) Allostery and cooperativity revisited. *Protein Science* **17**, 1295–1307.

Dagdas YS, **Chen JS**, **Sternberg SH**, **Doudna JA** (2017) A Conformational Checkpoint between DNA Binding and Cleavage by CRISPR-Cas9. *Science Advances* **3**, eaao002

Dijkstra EW (1959) A Note on Two Problems in Connection with Graphs. *Numerical Mathematics (Heidelberg)* **1**, 269–271.

Dokholyan N V (2016) Controlling Allosteric Networks in Proteins. *Chemical Reviews* **116**, 6463–6487.

East KW, **Newton JC**, **Morzan UN**, ***et al.*** (2020a) Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *Journal of the American Chemical Society* **142**, 1348–1358.

East KW, **Skeens E**, **Cui JY**, ***et al.*** (2020b) NMR and computational methods for molecular resolution of allosteric pathways in enzyme complexes. *Biophysical Reviews* **12**, 155–174.

Floyd RW (1962) Algorithm 97: Shortest path. *Communications of the ACM* **5**, 345.

Freddolino PL, **Gardner KH**, **Schulten K** (2013) Signaling mechanisms of LOV domains: new insights from molecular dynamics studies. *Photochemical & Photobiological Sciences* **12**, 1158.

Gasper PM, **Fuglestad B**, **Komives EA**, **Markwick PR**, & **McCammon JA** (2012) Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(52):21216–21222. PMC3535651.

Gray HB, **Winkler HR** (1996) Electron transfer in proteins. *Annual Review of Biochemistry* **65**, 537–561.

Guo J, **Zhou HX** (2016) Protein Allostery and Conformational Dynamics. *Chemical Reviews* **116**, 6503–6515.

Hirano S, **Nishimasu H**, **Ishitani R**, **Nureki O** (2016) Structural Basis for the Altered PAM Specificities of Engineered CRISPR-Cas9. *Molecular Cell* **61**, 886–894.

Jinek M, **Chylinski K**, **Fonfara I**, ***et al.*** (2012) A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821

Kern D, **Zuiderweg ER** (2003) The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology* **13**, 748–757

Kleinstiver BP, **Prew MS**, **Tsai SQ**, ***et al.*** (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485.

Lange OF, **Grubmüller H** (2006) Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function and Genetics* **62**, 1053–1061.

Liu J, **Nussinov R** (2016) Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Computational Biology* **12**, e1004966.

Melo MCR, **Bernardi RC**, **de la Fuente-Nunez C**, **Luthey-Schulten Z** (2020) Generalized correlation-based dynamical network analysis: a new high-performance approach for identifying allosteric communications in molecular dynamics trajectories. *Journal of Chemical Physics* **153**, 134104.

Miao Y, **Feher VA**, **McCammon JA** (2015) Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation* **11**:3584–3595

Miao Y, **Nichols SE**, **Gasper PM**, **Metzger VT**, & **McCammon JA** (2013) Activation and dynamic network of the M2 muscarinic receptor. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(27): 10982–10987. PMCID: PMC3703993.

Mitchell BP, **Hsu R V.**, **Medrano MA**, ***et al.*** (2020) Spontaneous Embedding of DNA Mismatches Within the RNA:DNA Hybrid of CRISPR-Cas9. *Frontiers in Molecular Biosciences* **7**, 39.

Newman MEJ, **Girvan M** (2004) Finding and evaluating community structure in networks. *Physical Review E* **69**, 026113.

Nierzwicki Ł, **Arantes PR**, **Saha A**, **Palermo G** (2021) Establishing the Allosteric Mecanism in CRISPR-Cas9. *WIREs Computational Molecular Sciences* **11**, e1503.

Nierzwicki Ł, **Arantes PR**, **Saha A and Palermo G** (2020).† Establishing the Allosteric Mechanism in CRISPR-Cas9. *WIREs Comp. Mol. Biosci.* **2020**, e1503.

Nierzwicki L, **East KW**, **Binz MB**, ***et al.*** (2022) Principles of Target DNA Cleavage and Role of Mg2+ in the Catalysis of CRISPR-Cas9. *Nature Catalysis* **10**, 912–922.

Nierzwicki L, **East KW**, **Morzan UN**, ***et al.*** (2021) Enhanced specificity mutations perturb allosteric signaling in CRISPR-Cas9. *Elife* **10**, e73601.

O'Connell MR (2019) Molecular Mechanisms of RNA Targeting by Cas13-containing Type VI CRISPR–Cas Systems. *Journal of Molecular Biology* **431**: 66–87

Palermo G, **Chen JS**, **Ricci CG**, ***et al.*** (2018) Key role of the REC lobe during CRISPR–Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Quarterly Reviews of Biophysics* **51**, e9.

Palermo G, **Miao Y**, **Walker RC**, ***et al.*** (2016) Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Central Science* **2**, 756–763.

Palermo G, **Ricci CG**, **Fernando A**, ***et al.*** (2017) Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *Journal of the American Chemical Society* **139**, 16028–16031.

Ricci CG, **Chen JS**, **Miao Y**, ***et al.*** (2019) Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. *ACS Central Science* **5**, 651–662.

Rivalta I, **Sultan MM**, **Lee NS**, ***et al.*** (2012) Allosteric pathways in imidazole glycerol phosphate synthase. *Proceedings of the National Academy of Sciences U S A* **109**, 1428–1436.

Rossetti M, **Merlo R**, **Bagheri N**, ***et al.*** (2022) Enhancement of CRISPR/Cas12a *trans*-cleavage activity using hairpin DNA reporters. *Nucleic Acids Research* **50**, 8377–8391.

Saha A, **Ahsan M**, **Arantes PR**, ***et al.*** (2024) An alpha-helical lid guides the target DNA toward catalysis in CRISPR-Cas12a. *Nature Communications* **15**, 1473.

Saha A, **Arantes P**, **Hsu R**, ***et al.*** (2020) Molecular Dynamics Reveals a DNA-Induced Dynamic Switch Triggering Activation of CRISPR-Cas12a. *Journal of Chemical Information and Modeling* **60**, 6427–6437.

Saha A, **Arantes PR**, **Palermo G** (2022) Dynamics and mechanisms of CRISPR-Cas9 through the lens of computational methods. *Current Opinion in Structural Biology* **75**, 102400.

Saltalamacchia A, **Casalino L**, **Borišek J**, ***et al.*** (2020) Decrypting the Information Exchange Pathways across the Spliceosome Machinery. *Journal of the American Chemical Society* **142**, 8403–8411.

Schmid-Burgk JL, **Gao L**, **Li D**, ***et al.*** (2020) Highly Parallel Profiling of Cas9 Variant Specificity. *Molecular Cell.* **78**, 794–800.

Sethi A, **Eargle J**, **Black AA**, **Luthey-Schulten Z** (2009) Dynamical networks in tRNA: protein complexes. *Proceedings of the National Academy of Sciences U S A* **106**, 6620–6625.

Sharma V, **Belevich G**, **Gamiz-Hernandez AP**, ***et al.*** (2015) Redox-induced activation of the proton pump in the respiratory complex I. *Proceedings of the National Academy of Sciences U S A* **112**, 11571–11576.

Sinha S, **Molina Vargas AM**, **Arantes PR**, ***et al.*** (2024) Unveiling the RNA-mediated allosteric activation discloses functional hotspots in CRISPR-Cas13a. *Nucleic Acids Research* **52**, 906–920.

Skeens E, **Sinha S**, **Mohd A**, ***et al.*** (2024) High-Fidelity, Hyper-Accurate, and Evolved Mutants Rewire Atomic Level Communication in CRISPR-Cas9. *Science Advances* **10**, eadl1045.

Slaymaker IM, **Gao L**, **Zetsche B**, ***et al.*** (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88.

Sternberg SH, **LaFrance B**, **Kaplan M**, **Doudna JA** (2015) Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature* **527**, 110–113.

Sternberg SH, **Redding S**, **Jinek M**, ***et al.*** (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–68.

Strohkendl I, **Saha A**, **Moy C**, ***et al.*** (2024) Cas12a domain flexibility guides R-loop formation and forces RuvC resetting. *Molecular Cell.* **84**, 2717–2731.

Sung K, **Park J**, **Kim J**, ***et al.*** (2018) Target Specificity of Cas9 Nuclease via DNA Rearrangement Regulated by the REC2 Domain. *Journal of the American Chemical Society* **140**, 7778–7781

Wagner JR, **Lee CT**, **Durrant JD**, ***et al.*** (2016) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chemical Reviews* **116**, 6370–6390.

Wang J, **Arantes PR**, **Bhattarai A**, ***et al.*** (2021) Gaussian accelerated molecular dynamics: Principles and applications. *WIREs Computational Molecular Sciences* **11**, e1521.

Wodak SJ, **Paci E**, **Dokholyan N V.**, ***et al.*** (2019) Allostery in Its Many Disguises: From Theory to Applications. *Structure* **27**, 566–578.

Yan C, **Dodd T**, **He Y**, ***et al.*** (2019) Transcription preinitiation complex structure and dynamics provide insight into genetic diseases. *Nature Structure Molecular Biology* **26**, 397–406.

Zuo Z, **Liu J** (2020) Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Current Opinion in Structural Biology* **62**, 166–174.

https://doi.org/10.1017/S0033583524000143 Published online by Cambridge University Press