Encoding Protein Conformational Distributions Through Allosteric Traffic Graphs

Lucas Lee
CPSC 483
lucas.c.lee@yale.edu

Abstract

Understanding the dynamic behavior of KRAS, a key protein in cellular signaling, is crucial due to its role in regulating cell growth and its implications in cancer research. This study introduces a novel approach for modeling KRAS conformational distributions using allosteric traffic graphs and graph neural networks (GNNs). The proposed methodology combines hierarchical pooling and attention mechanisms within a GNN architecture, termed KRAS-Attention-Pooling (KAP) [1]. By encoding graph-based representations of residue interactions, KAP efficiently predicts conformational shifts and provides insights into the free energy landscape (FEL) of KRAS. Extensive experimentation demonstrates the superiority of KAP over existing methods, both in predictive accuracy and in its ability to identify key allosteric clusters. This work advances the understanding of KRAS dynamics, offering a robust framework for studying conformational changes in complex biological systems.

1 Introduction

GTPases are pivotal in cellular signaling due to their role in transferring phosphate groups; therefore, understanding their conformational behavior is essential for optimizing their functionality and control mechanisms. KRAS in particular is a critical player in cell signaling pathways that regulate cell growth, differentiation, and survival [1]. Though long considered a difficult target to drug, mutants of KRAS are often oncogenic, making the protein a key target in cancer research [1]. Despite advances in structural biology, there remains a particular challenge in correlating such mutations to changes in protein dynamics and behavior. A major hurdle is that constructing a free energy landscape (FEL) that can describe a protein's fully dynamic range is difficult. This FEL is particularly important in that it determines which conformations are thermodynamically favorable and will be most occupied. However, it is often necessary to know descriptive collective variables to map important states and the transitions between them to the FEL. A more efficient method of producing representations of a particular protein's underlying FEL would be therefore enable understanding of how particular mutations affect conformational distributions.

Thus the central goal of this project is therefore to uncover the effects of KRAS mutations on protein dynamics, focusing specifically on how their corresponding traffic graphs can be used to predict differences in conformational distributions between the KRAS active and inactive states (Figure 1). In order to efficiently turn these traffic graphs into meaningful embeddings that can be used to make such predictions, we propose and implement a novel graph neural network (GNN) that uses DiffPool as well as attention-based methods to accurately predict conformational population distributions with respect to active and inactive reference states while also providing interpretability in terms of meaningful node assignment in the pooling process.

[1] All code available at `https://github.com/Lucas-Lee11/KAP`

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

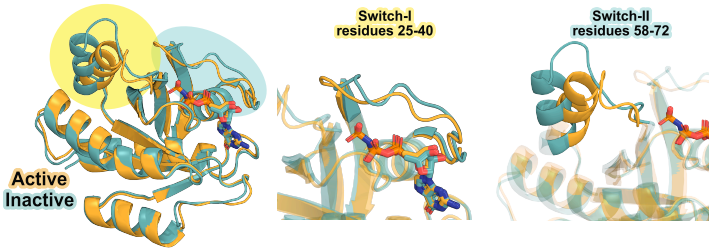

Figure 1: Crystal structures of the inactive (PDB-ID: 4OBE) and active (PDB-ID: 6VJJ) conformations of KRAS bound to GDP and GTP respectively. Key conformational differences in the switch I and switch II loops highlighted.

This model, which we call KRAS-Attention-Pooling (KAP) is a system specific GNN which utilizes the weighted traffic matrix to predict distance to inactive and active population in an embedding space based off the switch I and switch II loops which facilitate the main conformational transition between the two states. In this proposed approach, we create graph embeddings in a hybrid model that incorporates both hierarchical pooling [2] to create interpretable understandings of allosterically relevant clusters and domains of the protein, as well as graph attention [3] to allow for asymmetric weighting of messages in the later pooled part of the network. A distinct first layer without attention that uses the weighted graph directly is also utilized.

This hybrid methodology utilizing both a typical message passing layer and attention achieves significantly better predictive power when compared to straighforward GNN architectures such as pure GCN, GraphSAGE, and GAT and in comparison to DiffPool-based methods that utilize those architectures for their embedding and assignment. In addition, the hierarchical pooling also some relationship between known important structures within the KRAS protein, indicating the network is able to recover such information. Such results both imply that the underlying graph representation is able to encode both allosteric and energetic information about protein dynamics and may be able to be applied to other systems to predict energetic shifts.

2 Related Work

Graph-based methods in particular have emerged as promising tools to encode the underlying dynamic states of proteins. Protein networks have been frequently constructed using metrics such as correlation values and contact maps, enabling for network analysis techniques including eigenvector centrality and community partitioning [4]. Particularly in the context of FEL perturbation, the use of graphs to model proteins provides a way to link individual residue interactions with system-level informational changes [5].

For instance the model of graphs developed by Tyler *et al.* [6], connects protein folding states with their graphical representations. This representation was constructed from the atomic coordinates of the protein during an MD simulation where vertices represent the positions of Cα atoms while edges connect such carbons that are considered contacts [6]. The model was applied to small proteins, and the topological descriptors derived from these graph representation, such as the Kirchhoff index and average shortest path, were able to distinguish between folded, unfolded, and misfolded states [6]. With this method, Tyler *et al.* [6] were able to estimate free energy of the system over time, providing some sense of how an MD simulation is evolving over a particular FEL, this methodology generally involves costly simulation time which could be used to directly FEL without a secondary model. In a different approach, Rana *et al.* [7] utilized a colored subgraphs to predict binding free energy changes. In their methodology, subgraphs of mutated and active sites are created with edges constructed based on distance and atom type. Using aggregated metrics of these subgraphs as predictive features, the trained GGL-PPI model performed effective prediction of binding $\Delta\Delta G$ on mutation [7]. A limitation with this method however, is only examining local changes at the site of mutation and catalysis, whereas in reality such binding and mutation changes can have much longer-range allosteric effects that affect protein conformation. A full understanding of allosteric communication in conjunction with the underlying FEL may require more global information that just local structure can provide.

In terms of GNN usage in the context of MD simulation-based protein characterization, Chiang *et al.* [5] demonstrated that by using protein graphs with a mix of correlation and distance-based edges, GNNs can be utilized to encode aggregate information. Using GCN and GraphSAGE as benchmarks, their ProDAR network as able to accurately classify proteins to their enzymatic function and highlight specific residues of significance in the network [5]. These same distance and correlation metrics are both serve as inputs into our approach to path generation, however the traffic graphs are ultimately constructed from these pathway ensembles rather than only correlation, and are thus able to more directly capture information flow. Yet this method also does not directly address the issue of understanding the protein's underlying FEL and involves the overhead of running a MD simulation to obtain the necessary input data.

3 Methodology

Taking inspiration from Chiang *et al.* [5], we use a protein graph structure with residues as nodes, using one-hot feature vector of length 20 to represent the amino acid identity. The edges between residues are constructed from the path analysis output, summing the number of paths that traverse over each connection, creating a weighted undirected graph. In order to create an expressive embedding of each graph correlating to its active-inactive distribution, we utilize a modified version of the DiffPool architecture to take advantage of its ability to encode graph hierarchical information in its pooling process [2]. Proteins themselves are inherently hierarchical in the formation secondary structures and larger domains that group together residues. Thus we constructed a 3 layer DiffPool-based method, where at each layer the number of nodes is reduced based on learned clusters up to a set limit number of nodes. This use of pre-defined number of clusters inherently limits the transferability of this network to other systems, but will allow for the learning of allosterically meaningful domains within KRAS.

However, as noted above experimental testing has indicated that certain residues as well as larger domains can act as key allosteric sites, significantly altering the flow of information through the protein. Though this behavior is captured through the path analysis that informs the graph creation, such weights would only be relevant for the first layer of pooling, since the DiffPool architecture learns reduced representations for the subsequent layers of pooling. Thus we propose the addition of attention mechanisms for later layers of this network to facilitate the weighted message massing. More specifically we define the embedding $\mathbf{h}_v^{(\ell)}$ of node v within a GNN at pooling level ℓ based on

$$\mathbf{h}_v^{(\ell)} = \begin{cases} \sigma\left(\sum_{u \in N(v)} w_{uv} \mathbf{W}^{(\ell)} \mathbf{x}_u\right), & \ell = 1 \\ \sigma\left(\sum_{u \in N(v)} \alpha_{uv}^{(\ell)} \mathbf{W}^{(\ell)} \mathbf{h}_u^{(\ell-1)}\right), & \ell > 1 \end{cases} \tag{1}$$

where $\mathbf{x}_u$ is the initial one-hot amino-acid representation of the base graph and w_{uv} is the edge weight between vertices v and u. In subsequent pooling layers, this edge weight is replaced with α_{vu}, representing the trained attention mechanism based that weight the relative importance of connections between the pooled clusters. Each pooling layer itself in the DiffPool architecture is parameterized by two GNNs, one for predicting embeddings at that layer and another for predicting assignments for the next level of pooling. Thus with this hybrid paradigm, GNNs in the first level of pooling utilizes the provided weights of the allosteric input graph and the later ones are attention-baed.

In terms of the size of each of the ℓ levels of pooling, we used decreasing maximum cluster size in a ratio of $1/4$ of starting with $168 \to 42 \to 11 \to 3$. We also optionally include skip connections from each layer's embedding embedding with the final hidden layer output $\mathbf{h}^{(k)}$ using global mean to act as the embedding for the entire graph. This skip connection embedding can serve as the input to an multi-layer perceptron (MLP) to convert to the desired dimension of y to create the prediction $\hat{y} \in \mathbb{R}^2$.

As a baseline, we compared this method to both GNN architectures without that pooling behavior as well as DiffPool based methods that do not utilize this hybrid edge weight and attention based message passing.

4 Experiments

4.1 Dataset

Though allosteric traffic graphs can be constructed from full MD simulations, the computational cost of running accurate simulations on a comprehensive dataset of KRAS mutants is significant. However, recent studies have suggested that AlphaFold2 (AF2) is capable of predicting multiple conformations of the same protein, and can be potentially used to elucidate the conformational plasticity of biological systems [8]. Thus using AF2-generated conformational ensembles, we have constructed a dataset of traffic graphs for a set of KRAS point mutations, constructed through a pipeline starting with residue-level mutual information between the individual residues in each ensemble based on φ and ψ dihedral angles. This was then incorporated with residue distance to create an ensemble of allosteric paths between all residue pairs. Through summing number of paths traversing each inter-residue edge, we construct a undirected weighted adjacency matrix $A \in \mathbb{R}_{+}^{n \times n}$ where n is the number of residues; $n = 168$ for the KRAS system. The current dataset consists of a combination of mutational scan and masked data, resulting in 1008 different KRAS traffic graphs with differing underlying mutations and FELs.

Overall, the graphs display a mean unweighted vertex degree of ≈ 7.08 and a standard deviation of ≈ 2.91, indicating relative uniformity in number of connections. This is expected since most residues have some connections with their neighbors within the cutoff distance. However, the graphs display greater variability in their weighted degree distribution, with mean weighted degree ≈ 1324.05 with standard deviation ≈ 939.986. We also note correlation between nodes of with higher mean weight degree also have increased variability in that degree, indicating that mutations can modulate the traffic flow around these key allosteric sites (Appendix, Figure 3).

In order to capture conformationally important information as labels for these graphs, we propose using distance to known active and inactive crystal structures in a reduced PCA space derived from inter-residue distances within the switch I and switch II regions of KRAS. Previous investigations of KRAS dynamics have indicated mutations of the protein generate significant population shifts in these regions [9], [10] and the active-inactive crystal structures (Figure 1) already show large conformational differences in these domains. Thus we create a two-feature label for each mutant $y = (y_0, y_1) \in \mathbb{R}^2$ where y_0 represents distance to inactive, and y_1 represents distance to active state. This dataset was divided randomly in a 80-10-10 train-validation-test split for the training and evaluation process.

4.2 Training Details

During the hyperparameter optimization process, we utilized only the validation set. Based on this experimentation, we used as a baseline a 3-layer GNNs with a hidden dimension of 64 to parameterize the DiffPool embedding and assignment predictions. For the edge weight based first level, we utilized the GCN architecture due to its simplicity, while the later layers are GAT based to accomplish this attention mechanism.

The model was optimized for MSE loss added to the sum of the link prediction and entropy regularization losses that DiffPool uses to optimize its cluster assignments. This weighting between the three loss values can be represented as

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{MSE}} + \lambda_b \sum_{i=1}^{\ell} \mathcal{L}_{\text{LP}_i} + \lambda_c \sum_{i=1}^{\ell} \mathcal{L}_{\text{ER}_i} \tag{2}$$

Though empirical testing, we set $\lambda_a = 50$ and $\lambda_b = \lambda_c = 1$. This was optimized via the Adam optimizer, with a learning rate of 1e-4 and weight decay of 1e-2 is able to provide the best results, converging in around 400-700 epochs. Final results were obtained after hyperparameter optimization on the test dataset.

4.3 Results

After training our model and its baseline comparisons, we evaluated each of the architectures on the testing dataset both with the DiffPool pooling technique as well as without it, using MSE as a quality

Table 1: MSE Results

	Architecture				
Pooling	MLP	GCN	GraphSAGE	GAT	GCN-GAT
None	0.946 ± 0.006	0.530 ± 0.015	0.504 ± 0.023	0.509 ± 0.031	$\mathbf{0.448 \pm 0.029}$
DiffPool	0.906 ± 0.012	0.501 ± 0.055	0.459 ± 0.051	0.948 ± 0.905	$\mathbf{0.402 \pm 0.067}$

measure. We trained each model for 500 epochs 5 separate times to calculate a mean and standard deviation (Table 1). Based on this data, we found that in both pooling scenarios, the hybrid edge weight and attention based method outperforms all other architectures, including both GCN and GAT alone. Using the differentiable pooling technique also achieves a lower MSE compared to without for all of the architectures.

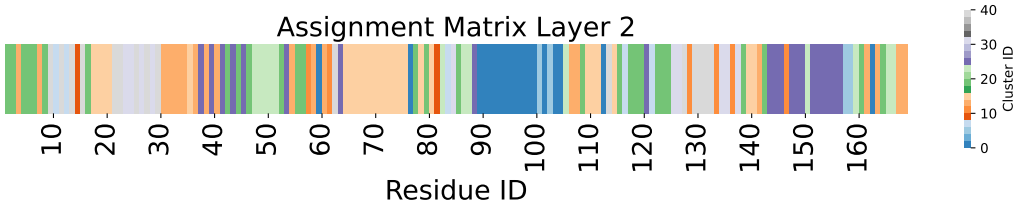

Figure 2: Visualization of a the nodewise cluster assignment at the second pooling layer for an example input graph in the test dataset. Each color represents a particular cluster assignment for the node.

One of the main reasons to utilize this pooling is to enable learning of meaningful clusters within the graph as a means of interpretability. In this case, we sought to see if the model could learn allosterically or functionally relevant domains of the KRAS protein system based only on the traffic graph provided. Examining an example of the cluster assignments of the model at the second layer of pooling for which there are a maximum of 42 different clusters to assign nodes to, we visualized how each node in the original graph is assigned (Figure 2). In particular, we found that the switch II loop (residues 60-75) is generally preserved as a single cluster as well as many important GTP interface residues in the range of 10-30.

Thus we found that our KAP model, which incorporates both the differentiable pooling technique and the hybrid edge weight and attention architecture is best able to predict underlying population distribution characteristics and thus encode information about its FEL. It also further reinforces that such allosteric graph models are able to encode information pertaining to population distributions with respect to active and inactive states.

5 Conclusion

This study presents a significant advancement in the computational modeling of protein conformational dynamics, particularly for KRAS. The proposed KRAS-Attention-Pooling (KAP) framework effectively combines graph-based residue interaction data with hierarchical pooling and attention mechanisms to accurately predict conformational distributions. Results demonstrate that KAP outperforms traditional GNN models, providing both superior predictive accuracy and deeper interpretability regarding allosteric interactions. By identifying key structural domains and enabling efficient exploration of the KRAS free energy landscape, this approach has implications for understanding protein behavior and informing therapeutic strategies. Future applications of this methodology could extend to other proteins, enabling broader insights into molecular dynamics and biological function.

6 Appendix

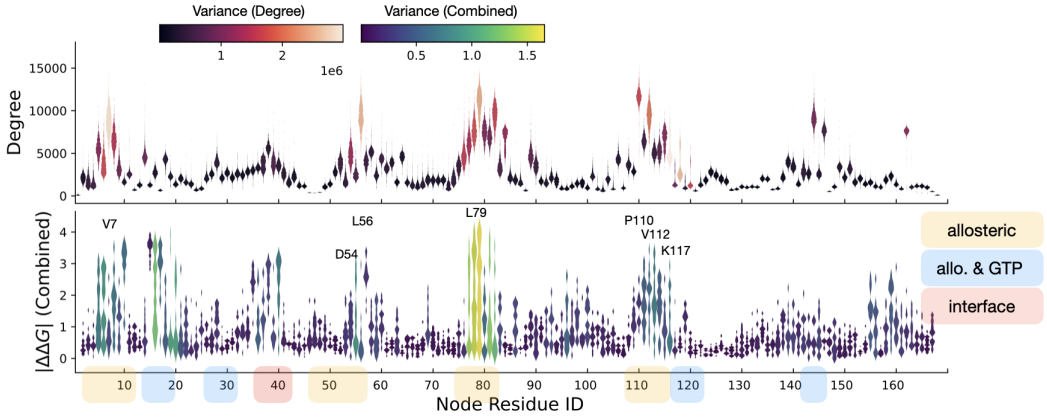

Figure 3: Violinplots of weighted node degree (above) and combined folding and binding $\Delta\Delta G$ experimental data taken from Weng *et al.* [1]. Respective mutated node residue ID labeled below, with key allosteric and binding interface sites highlighted.

References

[1] C. Weng, A. J. Faure, A. Escobedo, and B. Lehner, "The energetic and allosteric landscape for kras inhibition," *Nature*, vol. 626, no. 7999, pp. 643–652, 2024, ISSN: 0028-0836. DOI: 10.1038/s41586-023-06954-0.

[2] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," *arXiv*, 2018. DOI: 10.48550/arxiv.1806.08804.

[3] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv*, 2017. DOI: 10.48550/arxiv.1710.10903.

[4] F. Maschietto, B. Allen, G. W. Kyro, and V. S. Batista, "Mdigest: A python package for describing allostery from molecular dynamics simulations," *The Journal of Chemical Physics*, vol. 158, no. 21, p. 215 103, 2023, ISSN: 0021-9606. DOI: 10.1063/5.0140453.

[5] Y. Chiang, W.-H. Hui, and S.-W. Chang, "Encoding protein dynamic information in graph representation for functional residue identification," *Cell Reports Physical Science*, vol. 3, no. 7, p. 100 975, 2022, ISSN: 2666-3864. DOI: 10.1016/j.xcrp.2022.100975.

[6] S. Tyler, C. Laforge, A. Guzzo, A. Nicolaï, G. G. Maisuradze, and P. Senet, "Einstein model of a graph to characterize protein folded/unfolded states," *Molecules*, vol. 28, no. 18, p. 6659, 2023. DOI: 10.3390/molecules28186659.

[7] M. M. Rana and D. D. Nguyen, "Geometric graph learning to predict changes in binding free energy and protein thermodynamic stability upon mutation.," *The journal of physical chemistry letters*, vol. 14, no. 49, pp. 10 870–10 879, 2023, ISSN: 1948-7185. DOI: 10.1021/acs.jpclett.3c02679.

[8] H. K. Wayment-Steele, A. Ojoawo, R. Otten, *et al.*, "Predicting multiple conformations via sequence clustering and alphafold2," *Nature*, vol. 625, no. 7996, pp. 832–839, 2024, ISSN: 0028-0836. DOI: 10.1038/s41586-023-06832-9.

[9] P. Grudzien, H. Jang, N. Leschinsky, R. Nussinov, and V. Gaponenko, "Conformational dynamics allows sampling of an "active-like" state by oncogenic k-ras-gdp," *Journal of Molecular Biology*, vol. 434, no. 17, p. 167 695, 2022, ISSN: 0022-2836. DOI: 10.1016/j.jmb.2022.167695.

[10] T. Pantsar, "The current understanding of kras protein structure and dynamics," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 189–198, 2020, ISSN: 2001-0370. DOI: 10.1016/j.csbj.2019.12.004.