Exhaustive In-silico Simulation of Single Amino Acid Insertion and Deletion Mutations

Alistair Turcan
Dept. of Computer Science
Western Washington University
Bellingham, WA, USA
turcana@wwu.edu

Grant Chou
Department of Computer Science
Western Washington University
Bellingham, WA, USA
choug@wwu.edu

Lilu Martin
Department of Computer Science
Western Washington University
Bellingham, WA, USA
marti400@wwu.edu

Theo Miller
Department of Computer Science
Western Washington University
Bellingham, WA, USA
mille550@wwu.edu

Dylan Thompson
Department of Computer Science
Western Washington University
Bellingham, WA, USA
thomp289@wwu.edu

Filip Jagodzinski
Department of Computer Science
Western Washington University
Bellingham, WA, USA
jagodzf@wwu.edu

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995279

***Abstract*—The effects of insertion and deletion mutations (InDels) on protein structures are understudied because performing such experiments in vitro is prohibitive. Consequently, little real world data exists for these types of mutations, despite many diseases being caused by InDels. Computational modeling can support researchers in their efforts to understand the impacts of InDels. In this work we present an *in silico* approach to generate all exhaustive InDel mutants for a protein. We analyze the effects of the InDels via heatmaps and other visualizations using a variety of metrics, including rigidity analysis and structural data about the mutants relative to the wild type.**

***Index Terms*—computational structural biology, protein indel mutations**

I. Introduction

Insertion and deletion mutations (InDels) are a common type of protein mutation. Amino acid substitution mutations have been explored by significantly more researchers, leaving the effects of InDels much less understood and in need of further study [1, 7]. The high cost of performing wet-lab experiments when doing sequence-level insertions and deletions, preceding transcription and translation, and ultimately obtaining a new protein structure [19] is one of the primary limiting factors in the study of InDels [9, 18].

Structurally, InDels occur when non-frameshift insertions or deletions in the DNA sequence cause at least one amino acid to be inserted or deleted from the protein's sequence, resulting in an InDel mutant protein. The causes of InDel mutations include replication errors and genome duplication [3, 15].

InDels can theoretically be at any location in a protein, but most frequently they occur in loop regions [14]. InDels are also known to occur in the secondary structures of a proteins, which results in larger effect than when they occur in other regions [8]. In particular, multiple deletions in an alpha helix should typically prevent the protein's expression entirely. InDel mutations have recently been found to be highly correlated with functional changes in proteins, more than substitutions [2, 12], indicating that they could be an important factor in the structural evolution of proteins.

Many diseases are caused by InDel mutations. Cystic fibrosis is one such example [20]. The F508del mutation in nucleotide-binding domain-1 (NBD1) of the cystic fibrosis transmembrane conductance regulator (CFTR) is the predominant cause of cystic fibrosis [4]. Also, several types of cancer have been associated with InDel mutations [10, 15]. In addition, multiple severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants are caused by InDels [6]. It has been shown that InDels at the 1/S2 subunits of this virus results in mutants with greater resistance to vaccines [16].

Also, the number of InDels has been shown to correlate with the extent of structural change that the mutation has on a protein. InDels of length two or more have the largest impact on protein structure, whereas single InDels are found to have smaller effects [22]. Due to the limited research into InDels, these findings have not been thoroughly investigated.

Furthermore, due to the lack of real world InDel data, and the time and cost burden of obtaining that data via experiments on physical proteins, a sudden increase in publicly available wet-lab datasets for InDel mutants is not likely to occur soon. As a result, computational modeling of these mutations will play a key part in the future of InDel studies. Such modeling might complement and even inform wet-lab approaches, providing hints about where a mutation experiment in the physical protein should be performed so to realize an expected outcome. In this paper, we present our compute pipeline for generating and analyzing protein InDels. We rely on existing mutation modeling and our own protein rigidity analysis software. In this work specifically, we generate and analyze every possible single insertion and deletion mutation for a set of proteins. Our aim is to better understand the effects of different InDels.

Our prior work describes methods of *in silico* generating individual InDel mutations [11] and exhaustive pairwise

Funded by NSF award 2031283 of Jagodzinski

978-1-6654-6819-0/22/$31.00 ©2022 IEEE

substitution mutations for a given protein [17]. This paper describes the first attempt at *in silico* generating and analyzing exhaustive single InDel mutations for a protein.

II. Methods

A. Generating Mutants in silico

To computationally generate InDel mutants, we use Rosetta [13]. We rely on a compute pipeline developed in our recent work, which takes a PDB file and mutation parameters as input and performs *in-silico* structural modeling of a single insertion or deletion mutation using an inverse kinematic robotics approach [11]. The pipeline outputs a PDB file with accompanying energetics and structural data about the mutant.

We generate an exhaustive set of single InDels for each protein. For a protein consisting of n residues, there are n single residue deletion sites (one for each residue) and $n+1$ single residue insertion sites (at the start, end, and between each residue) where any of 20 amino acids may be inserted for a total of $n + 20(n+1)$ total single InDels per protein. We generated exhaustive single InDel mutants for 21 proteins (Table I), which were chosen primarily due to their size (50-207 residues), to make the computational run-time tractable.

TABLE I
PDB files used (number of residues)

1APC (106)	1BJ8 (109)	1BW6 (56)	1C05 (159)
1GXG (85)	1HCE (118)	1HHP (99)	1IET (98)
1IP0 (50)	1JA6 (129)	1M8L (96)	1N3H (207)
2B0G (83)	2MYO (188)	7RAT (124)	1CTX (71)
1SRL (64)	1IFJ (50)	1NM4 (102)	1IFD (50)
1CRN (46)			

B. Protein Rigidity Analysis

We perform protein rigidity analysis of the wild type and InDel mutants using KINARI-lib [5]. Rigidity analysis creates a Body-Bar-Hinge (BBH) mechanical model of the protein, which is modeled as an associated graph where nodes represent the Body and Bar components of the BBH and edges correspond to constraints among the bodies. A pebble game algorithm is run on the associated graph to identify rigid and flexible components, from which rigid clusters of atoms in the protein can be inferred. We apply the same rigidity analysis to the wild type protein to establish a baseline.

C. Distributed Compute Pipeline

The compute pipeline (Figure 1) described above takes 1 to 10 minutes to model a single mutation for a given protein, depending on the number of residues. To scale this approach to the task of generating tens of thousands of mutants, we distributed the workload across a pool of 150 general purpose university lab computers during their unused hours, each with 16GB RAM and 8 to 16 4GHz CPU cores.

Distributing the workload sped up computation by approximately 150x over the sequential approach, depending on the availability of machines. However, even with this distributed approach, the pipeline could still take months to run to completion for larger proteins. With this constraint in mind, we chose to examine proteins in the range of 50 to 207 residues in length.

D. Focus on Analysis

There were two main areas of analysis of interest to us:

1) Understand better the structural effects of InDel mutations via modeling
2) Assess the destructiveness of InDel mutations.

Studying the structural effects of InDel Mutations via modeling involves generating mutants *in silico* via our distributed compute pipeline, followed by scoring the change between each mutant and the wild type. We used our Two Largest Clusters Comparison Score (TLCCS) [21] as our scoring metric. This metric we display in a heatmap to show which InDel mutations have the most impact on structure.

Predicting the destructiveness of InDel mutations involves reasoning about the likelihood that a given InDel will exist in nature. During the modeling process, Rosetta outputs a numerical energy score that can be interpreted as such a probability - i.e. high energy mutants are less energetically favored, and thus less likely to exist in nature. We evaluate the validity of this approach by modeling multiple deletions inside of an alpha helix, which is known to be a very destructive mutation, giving us a baseline for comparison for generated InDels.

III. Results and Discussion

A. Structural Effects of InDels

Figures 2, 3, 4 show the range of effects that result from inserting single residues into various proteins at all insertion points. Similarly, 5, 6, and 7 show the same for single deletion mutations at all deletion points. A few notable things can be seen. First, for insertions, the average effect doesn't vary much between different residues on the same protein, but it does vary between different proteins. For deletions, however, the average effect varies more between different residues on the same protein. Second, some residues have a more widely varying effect than others. For example, inserting Phenylalynine (F) has a more predictable effect on the structure of 1SRL than inserting Proline (P). That is not surprising, considering that Phenylalanine along with Tryptophan and Tyrosine is among the largest amino acids, so inserting it into the sequence of amino acids is more than likely to create a steric clash.

B. Analysis across multiple proteins

We also performed an analysis across proteins to better understand the impacts due to InDel mutations that are not specific to any one protein. Figure 8 shows the average effect caused by inserting each residue into every position across 21 different proteins. Such an meta analysis might be useful in predicting which residues tend to result in the greatest number of InDel mutant outliers, and also the direction (above or below the mean) those outliers tend to be in.

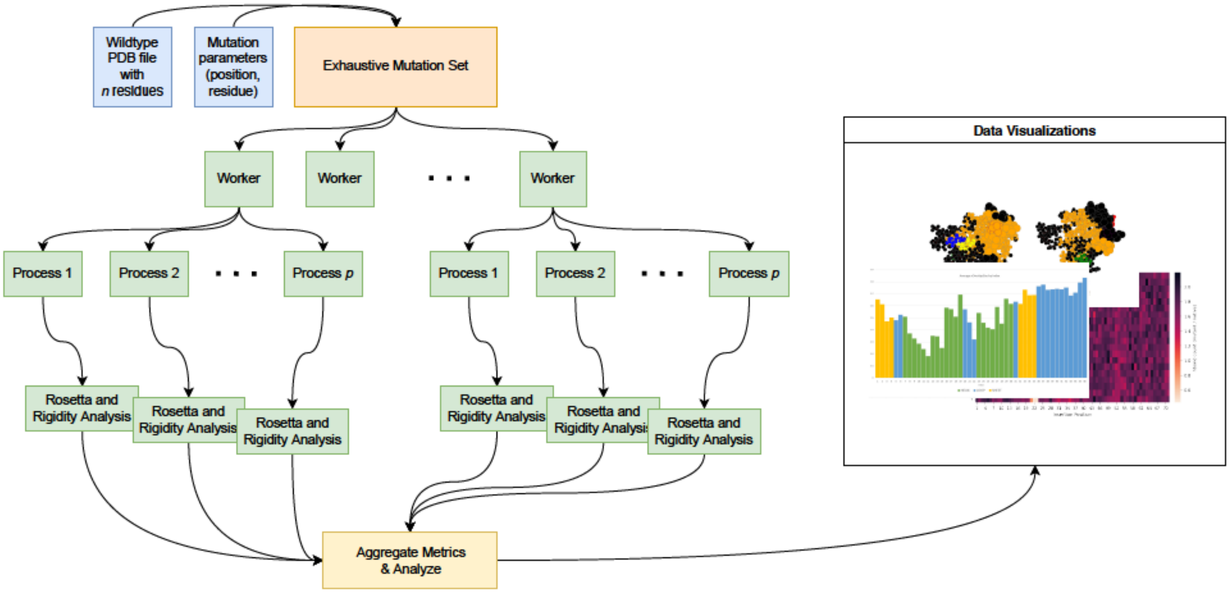

Wildtype PDB file with *n* residues
Mutation parameters (position, residue)
Exhaustive Mutation Set
Worker
Worker
. . .
Worker
Process 1
Process 2
. . .
Process *p*
Process 1
Process 2
. . .
Process *p*
Rosetta and Rigidity Analysis
Rosetta and Rigidity Analysis
Rosetta and Rigidity Analysis
Rosetta and Rigidity Analysis
Rosetta and Rigidity Analysis
Rosetta and Rigidity Analysis
Aggregate Metrics & Analyze
Data Visualizations

Fig. 1. Distributed Compute Pipeline: All single InDel mutations are generated, rigidity analysis is performed, rigidity metrics are collected, and results are aggregated to create data visualizations.

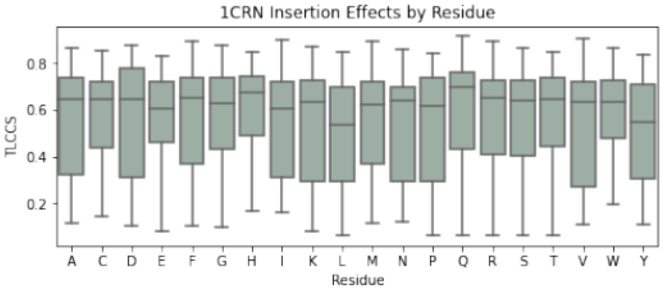


Fig. 2. *TLCCS* for inserting each residue into 1CRN. Each box and whisker plot displays the range of effects that result from inserting a particular residue at all possible insertion points.

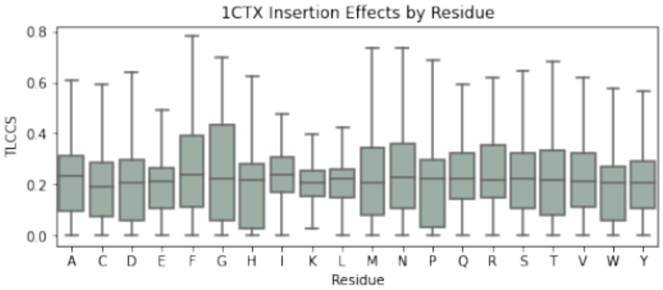


Fig. 3. *TLCCS* for inserting each residue into 1CTX. Each box and whisker plot displays the range of effects that result from inserting a particular residue at all possible insertion points.

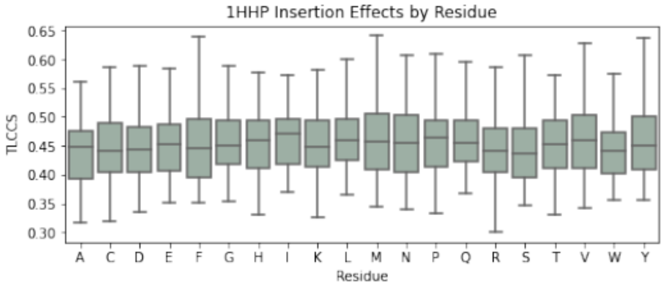


Fig. 4. *TLCCS* for inserting each residue residue into 1HHP. Each box and whisker plot displays the range of effects that result from inserting a particular residue at all possible insertion points.

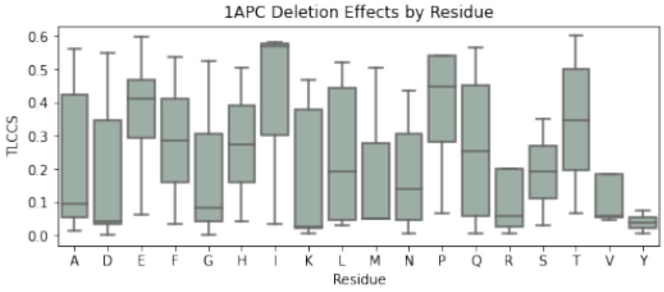


Fig. 5. *TLCCS* for deleting each residue in 1APC. Each box and whisker plot displays the range of effects that result from deleting that residue at all possible deletion points.

C. Heatmap Visualizations

We visualize the effects of InDel locations and residue by using heatmaps, with the colorbar represeent the metric that we chose to analyze. In this study, two metrics are particularly important: (1) change in protein structure (TLCCS) and (2) ratio of hydrogen bonds (hbonds_mut_native_ratio). For instance, as can be seen from Figures 9 and 10, in 1HHP,

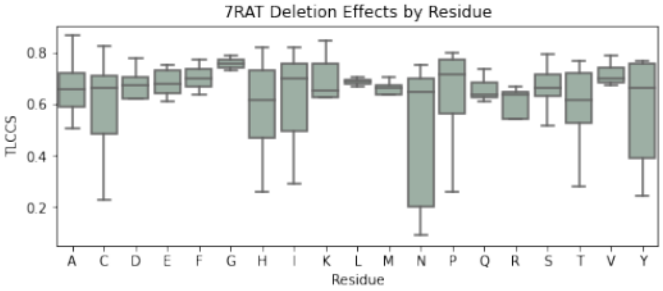


Fig. 6. *TLCCS* for deleting each residue in 7RAT. Each box and whisker plot displays the range of effects that result from deleting that residue at all possible deletion points.

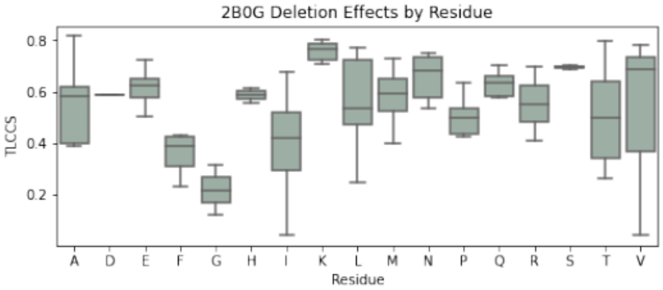


Fig. 7. *TLCCS* for deleting each residue in 2B0G. Each box and whisker plot displays the range of effects that result from deleting that residue at all possible deletion points.

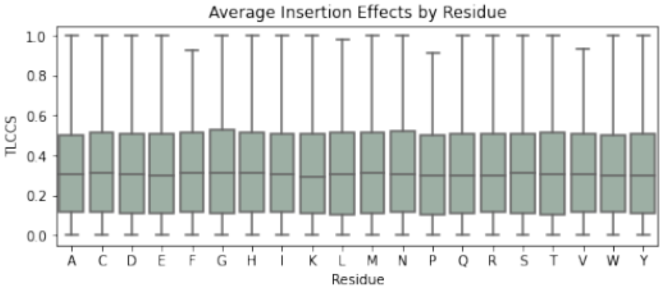


Fig. 8. *TLCCS* for each residue averaged across all of the proteins simulated. Each box and whisker plot displays the range of effects that result from inserting a particular residue at all possible insertion locations.

insertion of most residues at indices 84 or 85 result in a large structural change and a decrease in the number of hydrogen bonds in the mutant. Although the heatmaps for the analyses using the hydrogen bonds and TLCCS metrics look similar, there are differences (residues 22 and 32, for example), hinting that the utility afforded by both metrics are not identical. The extent that one metric is useful over the other, or whether using them in combination provides the most complete assessment of the effects of InDels, is something that we leave for future work. Similarly, assessing to what extent an InDel affects the count and distribution of hydrogen bonds, including those near and far removed from the mutation site, we leave for future work.

To identify other potentially noteworthy metrics which yielded significant signals, we employed principal component analysis. We found that in many proteins, the change in hydrophobic interactions often makes up a large portion of a non-leading principal component. This led us to create the same type of heatmap as described above with the color map corresponding to changes in hydrophobic interactions based on the insertion index and the residue being inserted. This visualization (Figure 11) revealed hot spots with a rather large increase such as the insertion of Valine (V) at index 16.

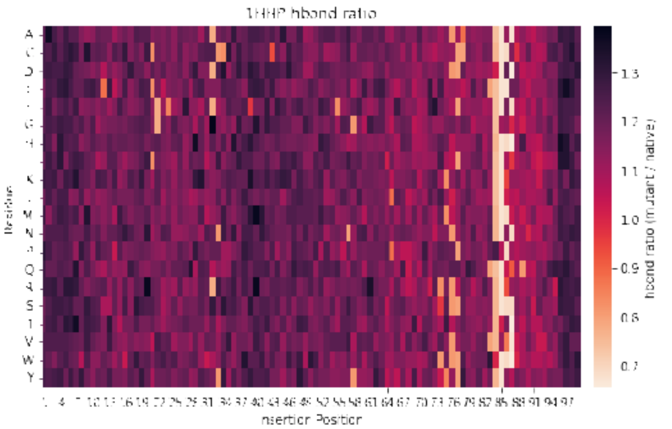


Fig. 9. Effects on number of hydrogen bonds overall of each possible insertion into 1HHP.

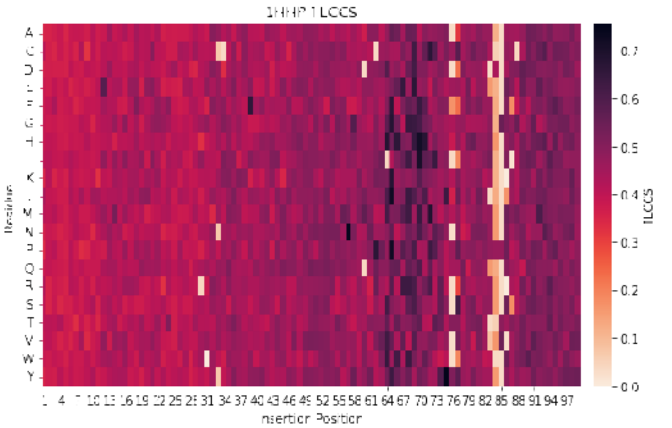


Fig. 10. Effects on the TLCCS of each possible insertion into 1HHP.

D. Rigid Cluster Visualization

Figure 12 shows a visualization of the rigid clusters of a protein's wild type and an InDel mutation. As can be seen from this comparison, the protein's distribution of large rigid clusters changes significantly in response to the InDel. This type of analysis reveals that the effects on a protein's rigid clusters caused by an InDel mutation can have effects beyond the vicinity of the InDel.

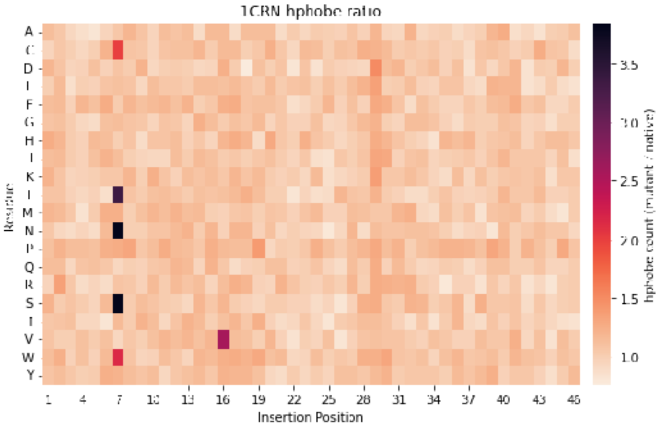


Fig. 11. Effects on number of hydrophobic interactions of each possible insertion into 1CRN.

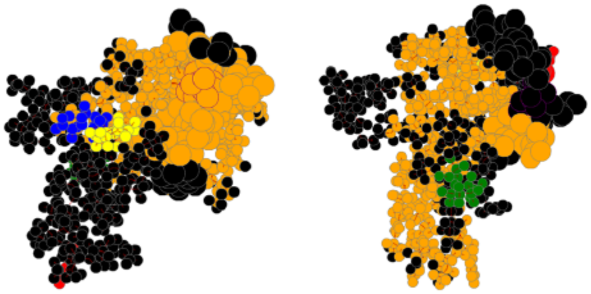
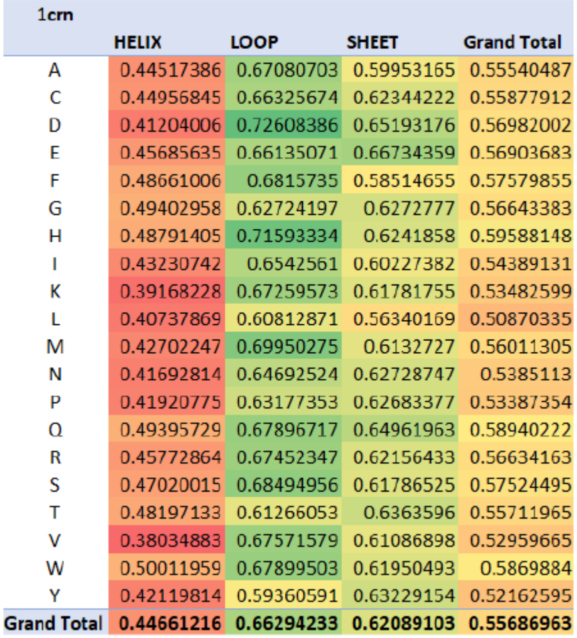


Fig. 12. Visualization of 5 largest rigid clusters of 1CTX wild-type (left) and InDel mutation after insertion of Cysteine (C) at index 62 (right). Each circle represents one atom with colors corresponding to each of the five largest rigid clusters.

E. Residue and Secondary Structure Comparison

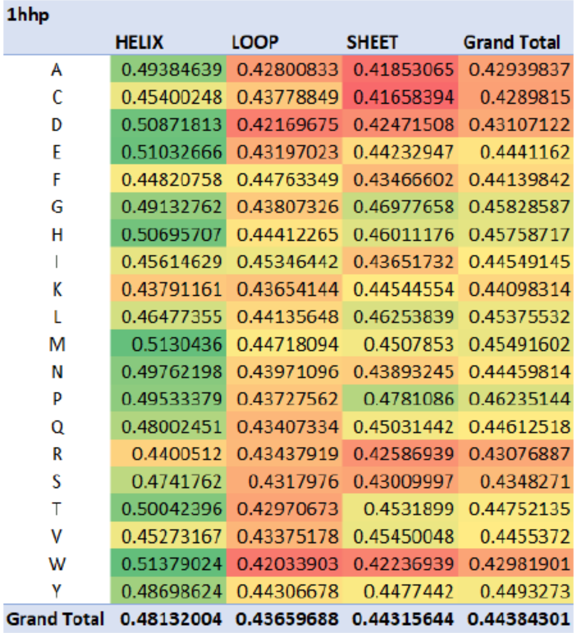
Figures 13 and 14 show two examples of the effect of various insertions on a protein's rigid clusters, based on the residue that is inserted, as well as the secondary structure that the residue is inserted into. This visualization allows us to see the distribution of InDel effects, including patterns within those. For instance, insertions of Aspartic Acid (D) and Tryptophan (W) into the β-sheets of 1CTX seems to have a very pronounced effect on the protein's rigid clusters when compared to the rest of the possible insertions. Interesting to note is that Aspartic Acid is acidic, while Trypotphan is aromatic, so their effect upon insertion to the structural properties of a protein in this case seems to be due to different mechanisms.

The analysis based on secondary structured yielded additional interesting results. For single insertions in individual proteins, hot spots appear on the heatmap visualizations, indicating a larger effect at specific locations where InDels were performed. Typically, these were found in the secondary structures, primarily alpha helices. However, surprisingly, some InDels not in secondary structures also had significant effects.



1crn	HELIX	LOOP	SHEET	Grand Total
A	0.44517386	0.67080703	0.59953165	0.55540487
C	0.44956845	0.66325674	0.62344222	0.55877912
D	0.41204006	0.72608386	0.65193176	0.56982002
E	0.45685635	0.66135071	0.66734359	0.56903683
F	0.48661006	0.6815735	0.58514655	0.57579855
G	0.49402958	0.62724197	0.6272777	0.56643383
H	0.48791405	0.71593334	0.6241858	0.59588148
I	0.43230742	0.6542561	0.60227382	0.54389131
K	0.39168228	0.67259573	0.61781755	0.53482599
L	0.40737869	0.60812871	0.56340169	0.50870335
M	0.42702247	0.69950275	0.6132727	0.56011305
N	0.41692814	0.64692524	0.62728747	0.5385113
P	0.41920775	0.63177353	0.62683377	0.53387354
Q	0.49395729	0.67896717	0.64961963	0.58940222
R	0.45772864	0.67452347	0.62156433	0.56634163
S	0.47020015	0.68494956	0.61786525	0.57524495
T	0.48197133	0.61266053	0.6363596	0.55711965
V	0.38034883	0.67571579	0.61086898	0.52959665
W	0.50011959	0.67899503	0.61950493	0.5869884
Y	0.42119814	0.59360591	0.63229154	0.52162595
Grand Total	**0.44661216**	**0.66294233**	**0.62089103**	**0.55686963**

Fig. 13. Average effect of inserting each residue in each type of secondary structure in 1CRN. A lower value means a larger effect.



1hhp	HELIX	LOOP	SHEET	Grand Total
A	0.49384639	0.42800833	0.41853065	0.42939837
C	0.45400248	0.43778849	0.41658394	0.4289815
D	0.50871813	0.42169675	0.42471508	0.43107122
E	0.51032666	0.43197023	0.44232947	0.4441162
F	0.44820758	0.44763349	0.43466602	0.44139842
G	0.49132762	0.43807326	0.46977658	0.45828587
H	0.50695707	0.44412265	0.46011176	0.45758717
I	0.45614629	0.45346442	0.43651732	0.44549145
K	0.43791161	0.43654144	0.44544554	0.44098314
L	0.46477355	0.44135648	0.46253839	0.45375532
M	0.5130436	0.44718094	0.4507853	0.45491602
N	0.49762198	0.43971096	0.43893245	0.44459814
P	0.49533379	0.43727562	0.4781086	0.46235144
Q	0.48002451	0.43407334	0.45031442	0.44612518
R	0.4400512	0.43437919	0.42586939	0.43076887
S	0.4741762	0.4317976	0.43009997	0.4348271
T	0.50042396	0.42970673	0.4531899	0.44752135
V	0.45273167	0.43375178	0.45450048	0.4455372
W	0.51379024	0.42033903	0.42236939	0.42981901
Y	0.48698624	0.44306678	0.4477442	0.4493273
Grand Total	**0.48132004**	**0.43659688**	**0.44315644**	**0.44384301**

Fig. 14. Average effect of inserting each residue in each type of secondary structure in 1HHP. A lower value means a larger effect.

IV. Conclusions

Despite being less common than substitutions, InDels are responsible for important structural and functional changes in proteins. However, they are understudied due to a lack of efficient experimental techniques for conducting exhaustive mutation experiments in physical proteins. In this work we present our proof-of-concept approach to generate exhaustive InDel mutants and analyze their effects.

Using existing open-source software, including Rosetta and rigidity analysis algorithms, we computationally generated protein structures for every possible single InDel mutation for a set of proteins. We presented a variety of analysis approaches to help infer how InDels affect the structural properties of proteins.

We focused primarily on structural analysis, using the change in rigid clusters upon an InDel as a metric of the effect of the mutation. This preliminary study hopes to provide useful information for future projects wishing to study InDel mutations which are limited by the lack of real-world data, demonstrating the applications of computationally modeling InDels at scale. In future work, we will develop a model to take into account a mix of rigidity and structural-based metrics to make predictions about the effects of InDels, both on the scale of an individual protein as well as across protein families.

References

[1] Monica Berrondo and Jeffrey J Gray. Computed structures of point deletion mutants and their enzymatic activities. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2844–2860, 2011.

[2] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, 2012.

[3] Nadia A Chuzhanova, Emmanuel J Anassis, Edward V Ball, Michael Krawczak, and David N Cooper. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local dna sequence complexity. *Human Mutation*, 21(1):28–44, 2003.

[4] Garry R Cutting. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics*, 16(1):45–56, 2015.

[5] Naomi Fox, Filip Jagodzinski, and Ileana Streinu. Kinarilib: A c++ library for mechanical modeling and pebble game rigidity analysis. *Minisymposium on Publicly Available Geometric/-Topological Software*, pages 29–32, 2012.

[6] Robert F Garry and William R Gallaher. Naturally occurring indels in multiple coronavirus spikes. *Virological*, 2020.

[7] Courtney E Gonzalez, Paul Roberts, and Marc Ostermeier. Fitness effects of single amino acid insertions and deletions in tem-1 β-lactamase. *Journal of Molecular Biology*, 431(12):2320–2330, 2019.

[8] Dirk W Heinz, Walter A Baase, Frederick W Dahlquist, and Brian W Matthews. How amino-acid insertions are allowed in an α-helix of t4 lysozyme. *Nature*, 361(6412):561–564, 1993.

[9] Fereydoun Hormozdiari, Raheleh Salari, Michael Hsing, Alexander Schönhuth, Simon K Chan, S Cenk Sahinalp, and Artem Cherkasov. The effect of insertions and deletions on wirings in protein-protein interaction networks: a large-scale study. *Journal of Computational Biology*, 16(2):159–167, 2009.

[10] Prathima Iengar. An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic acids research*, 40(14):6401–6413, 2012.

[11] Muneeba Jilani, Alistair Turcan, Nurit Haspel, and Filip Jagodzinski. Assessing the effects of amino acid insertion and deletion mutations. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2511–2518. IEEE, 2021.

[12] RyangGuk Kim and Jun-tao Guo. Systematic analysis of short internal indels and their impact on protein folding. *BMC Structural Biology*, 10(1):24, 2010.

[13] Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian W Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487:545–574, 2011.

[14] Sara Light, Rauan Sagit, Oxana Sachenkova, Diana Ekman, and Arne Elofsson. Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions. *Molecular Biology and Evolution*, 30(12):2645–2653, 09 2013.

[15] Maoxuan Lin, Sarah Whitmire, Jing Chen, Alvin Farrel, Xinghua Shi, and Jun-tao Guo. Effects of short indels on protein structure and function in human genomes. *Scientific Reports*, 7(1):9313, 2017.

[16] Zhe Liu, Huanying Zheng, Huifang Lin, Mingyue Li, Runyu Yuan, Jinju Peng, Qianling Xiong, Jiufeng Sun, Baisheng Li, Jie Wu, et al. Identification of common deletions in the spike protein of severe acute respiratory syndrome coronavirus 2. *Journal of Virology*, 94(17):e00790–20, 2020.

[17] Nicholas Majeske and Filip Jagodzinski. Elucidating which pairwise mutations affect protein stability: An exhaustive big data approach. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 01, pages 508–515, 2018.

[18] Stefano Pascarella and Patrick Argos. Analysis of insertions/deletions in protein structures. *Journal of molecular biology*, 224(2):461–471, 1992.

[19] Liat Rockah-Shmuel, Ágnes Tóth-Petróczy, Asaf Sela, Omri Wurtzel, Rotem Sorek, and Dan S Tawfik. Correlated occurrence and bypass of frame-shifting insertion-deletions (indels) to give functional proteins. *PLoS genetics*, 9(10):e1003882, 2013.

[20] Lap-Chee Tsui and Ruslan Dorfman. The cystic fibrosis gene: a molecular genetic perspective. *Cold Spring Harbor Perspectives in Medicine*, 3(2):a009472, 2013.

[21] Alistair Turcan, Anna Zivkovic, Dylan Thompson, Lorraine Wong, Lauren Johnson, and Filip Jagodzinski. Cgrap: A web server for coarse-grained rigidity analysis of proteins. *Symmetry*, 13(12), 2021.

[22] Zheng Zhang, Jie Huang, Zengfang Wang, Lushan Wang, and Peiji Gao. Impact of Indels on the Flanking Regions in Structural Domains. *Molecular Biology and Evolution*, 28(1):291–301, 07 2010.