Analyzing Conformational Variations in the RSV L Protein with Principal Component Analysis

Nicholas Keung, Chorok Lee, Tim Tan

Abstract

The respiratory syncytial virus L protein plays a crucial role in the virus life cycle. Using six published structures of L in different conformations, we extract the atomic coordinate data and use principal component analysis to identify regions of high variability.

1. Introduction

The respiratory syncytial virus (RSV) belongs to the family Pneumoviridae and infects epithelial cells in the respiratory tract (Collins et al., 2013). Illness often manifests as a minor cold with symptoms such as runny nose, fever, and cough, but they can significantly worsen for patients with comorbidities. Infants, young children, and the elderly have the highest risk of hospitalizations, and illness can progress to pneumonia and bronchiolitis (Collins et al., 2013). In 2019, there were approximately 33 million RSV infections and 3.6 million hospitalizations (Li et al., 2022). Studies on RSV have been extensive, ultimately leading to several FDA-approved vaccines in the last two years.

RSV replicates using its single-stranded RNA genome. The genome encodes for 11 proteins that aid in its viral life cycle. In particular, the L protein plays a significant role as a part of the viral polymerase, responsible for reading genomic RNA to synthesize new genomic RNA (replication) or to synthesize messenger RNA (mRNA) which will produce more viral proteins (transcription). As a result, its central role in producing more RSV virions make it a common target for therapeutics. Aiding in this process are other RSV proteins N, P, and M2-1 (Cao et al., 2020). Furthermore, scientists have identified several domains, or regions, on L that perform specific functions during replication and transcription. Across the various domains that bind to other proteins and RNA, L's function is extremely complicated. As L binds with different molecules, it will adopt new conformations, some of which were identified and published.

Identified protein structures are published in the Protein Data Bank (PDB) where they can be modeled and viewed in software such as PyMOL. The files contain detailed information about the chain of amino acids (residues) which make up the protein, the individual atoms that make up the residue, and the coordinates of the atom in space. For this project, we use atomic coordinate data and principal component analysis to compare multiple conformations of L. We use this analysis to identify highly variable regions which may provide new insight into L's function and can highlight new areas of interest.

The rest of the report will be structured as follows

2. Methodology
    2.1. Dataset
    2.2. Data Extraction and Cleaning
    2.3. Principal Components Analysis
    2.4. Data Visualization
3. Discussion
4. References

Code and data are available at https://github.com/nlkeung/pca-protein-structures.

2. Methodology

2.1 Dataset

For this project, we found six published structures of RSV L. We used two structures of RSV L bound with a tetramer of P, 6UEN (Cao et al., 2020) and 6PZK (Gilman et al., 2019), making up the polymerase complex which we see in the cell.

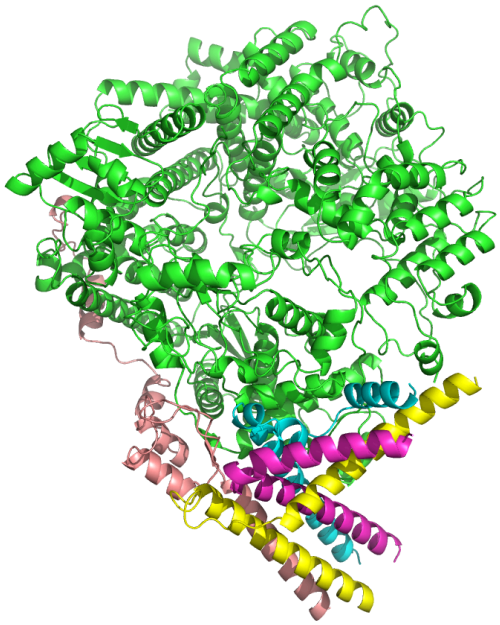


Figure 1. RSV L protein in complex with a tetramer of P. The structure 6UEN is shown with each chain highlighted in different colors. The L protein is shown in green. Meanwhile, the four chains of P are shown in cyan, yellow, magenta, and salmon. This image was obtained using PyMOL.

We also used two structures of L while bound to a fragment of genomic RNA (8SNX) and bound to its antigenome (8SNY), modeling what the structure might look like during transcription and replication respectively (Cao et al., 2024).

Finally, we use an additional two structures of L where it is bound to an inhibitor, serving as potential therapeutics. One structure, 8FPI, is L bound to an inhibitor called MRK-1 which is known to prevent certain conformational changes in L (Kleiner et al., 2023). The other is 8FU3, which is L interacting with another inhibitor called JNJ-8003 (Yu et al., 2023).

2.2 Data Extraction and Cleaning

Before comparing the coordinates of each atom, each structure must be superimposed and aligned to ensure that corresponding atoms match each other in their positions. This was done by uploading each `.pdb` file onto PyMOL and using the `align` command. With 6UEN as the template, the remaining five structures were rotated and translated to minimize the root-mean-square deviation between equivalent atoms.

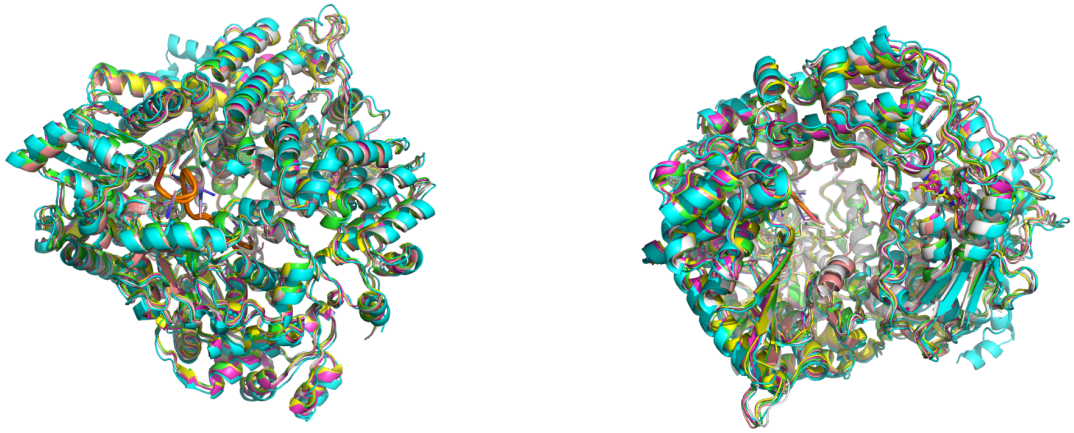


Figure 2. The alignment of all six structures. Each structure was uploaded to PyMOL and aligned to match the same position and orientation as model 6UEN. The chains corresponding to P were removed as well. 6UEN is cyan, 6PZK is green, 8FPI is magenta, 8FU3 is yellow, 8SNX is salmon, 8SNY is white.

After aligning each structure, the files were downloaded and analyzed using the `Bio.PDB` package from Biopython. The `PDBParser` was useful for extracting relevant information by traversing the file in a hierarchical structure the same way proteins are organized. Each `Bio.PDB.Structure` object contains `Model` objects, and each `Model` contains `Chain` objects, corresponding to chains of amino acids (residues). Then, `Chain` is then broken down into `Residue`. In particular, this object is useful because it stores the residue's position in the chain while still accounting for any gaps that could not be resolved experimentally.

Subsequently, there are `Atom` objects for each atom in the residue. The `Atom` object gave us information about its relative position in space through its x, y, and z coordinates.

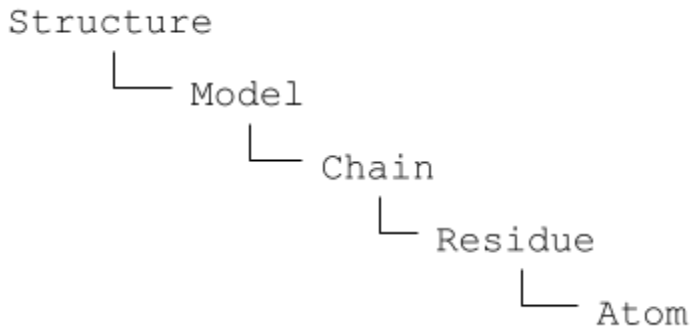


Figure 3. Hierarchy of a PDB structure in Biopython. Using the `Bio.PDB` package, data from each file was organized in a series of hierarchical objects, allowing us to find the relevant information from each residue and atom that we are interested in.

Each of our models contained multiple chains corresponding to L, P, and other molecules, so we extracted each chain and printed out their associated sequence of residues. We were then able to visually identify the relevant chains since L is significantly larger than P or any other chain.

In order to simplify our analysis, we chose to analyze only the alpha carbon from each residue, as these atoms play the most significant role in the protein's shape. This reduced our data from around 11,000 atoms, each with their own x, y, and z values to about 1,400 atoms. Also, in order to account for incomplete data, we iterated across each structure via a set function `remove_missing_atoms()` and kept only the atoms that were common across all six structures, leaving a total of 1,356 atoms. This provides a second-order block matrix with 1356 features (atoms) and 6 observations.

```
6UEN
Length of struct: 1356
CA 11 [ 97.427 125.769 129.548]
CA 12 [ 95.985 126.025 126.049]
CA 13 [ 94.302 123.251 124.039]
CA 14 [ 95.011 123.587 120.337]
CA 15 [ 92.703 121.374 118.283]

6PZK
Length of struct: 1356
CA 11 [ 97.841 126.132 129.73 ]
CA 12 [ 96.72  125.511 126.09 ]
CA 13 [ 95.075 122.765 123.994]
CA 14 [ 95.359 122.902 120.179]
CA 15 [ 92.73  120.798 118.354]

8SNX
Length of struct: 1356
CA 11 [ 98.912 124.662 128.764]
CA 12 [ 96.62  124.967 125.75 ]
CA 13 [ 94.727 122.273 123.85 ]
CA 14 [ 95.201 122.743 120.098]
CA 15 [ 92.766 120.623 118.109]

8SNY
Length of struct: 1356
CA 11 [ 98.784 124.231 129.017]
CA 12 [ 96.755 124.821 125.85 ]
CA 13 [ 94.875 122.237 123.78 ]
CA 14 [ 95.323 122.537 120.011]
CA 15 [ 92.875 120.582 117.851]

8FPI
Length of struct: 1356
CA 11 [ 98.771 125.107 128.848]
CA 12 [ 96.868 125.42  125.568]
CA 13 [ 94.952 122.903 123.475]
CA 14 [ 95.242 123.136 119.699]
CA 15 [ 92.671 120.975 117.954]

8FU3
Length of struct: 1356
CA 11 [ 98.798 125.789 128.502]
CA 12 [ 96.589 125.538 125.442]
CA 13 [ 94.782 122.882 123.487]
CA 14 [ 95.219 123.096 119.755]
CA 15 [ 92.796 120.862 117.917]
```

Figure 4. Final data after cleaning. The alpha carbon atoms were extracted from each structure, where each atom matches its corresponding atom in the other five structures. Information for the first five atoms was printed out, including the atom type (`CA` for alpha carbon), the residue number, and the x, y, and z values.

2.3 Principal Components Analysis

PCA is commonly used for purposes such as 1) dimensionality reduction, 2) principal component extraction, and 3) noise removal. Specifically for our project, our goal is to reduce the dimensionality of each of the 1,356 carbon atoms using PCA and identify which atoms exhibit significant variability across different structures to understand which regions of the structure are changing.

To perform PCA, we needed to flatten this block matrix into a single hierarchy; this was done by using the `reshape` function to turn the three spatial coordinates into additional observations, producing a matrix of 1,356 features and 18 observations. We could then perform singular value decomposition to quickly find the eigenvalues and basic eigenvectors for this matrix.

We then first recentered the data matrix by the mean so all 1,356 features (atoms) could be compared on the same scale, producing a recentered data set B. Given that the covariance matrix can be expressed as the product $\mathbf{S} = (1/\sqrt{(n-1)}\,\mathbf{B}^{\mathbf{T}})^{T}\ (1\sqrt{(n-1)}\,\mathbf{B}^{\mathbf{T}})$, we could then perform singular value decomposition (SVD) on the second factor to extract our transformation matrix $\mathbf{P}^{\mathbf{T}}$, which we can later use to diagonalize the covariance matrix of our original data set. The calculations were performed as shown in the code below.

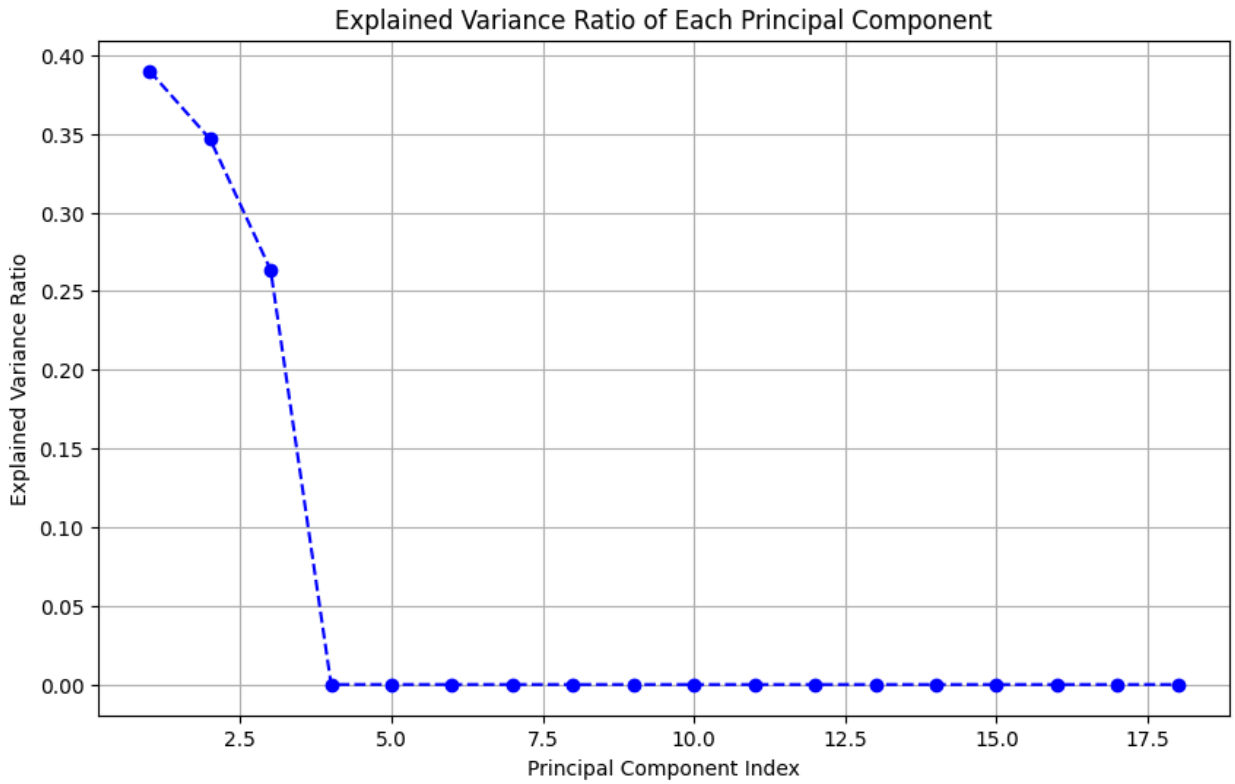
```
# Calculating covariance matrix S (18 x 18)
m, n = matrix.shape
S = np.cov(matrix)

# recentering the data matrix in order to use singular value decomposition
recentered_data = 1/math.sqrt(n-1)*(matrix-matrix.mean(axis=1).reshape(-1,
1))

#performing singular value decomposition to get the PCA transformation
matrix Pt
U , sigma, PT = np.linalg.svd(np.transpose(recentered_data))
P = np.transpose(PT)
```

The transformed data will have new dimensionality (still a total of 1,356) but will compress most of the variance into the first few principal components. In our case we analyzed the variance represented by PCs 1, 2, and 3, the first three column vectors of matrix **P**. The covariance matrix of $\mathbf{Y} = P^{T}\,\mathbf{X}$ where **X** is our original matrix can be represented as a diagonal

matrix where the nonzero entries are the eigenvalues of **S**. The variance represented by PC1, for instance, is the first eigenvalue divided by the trace, or sum of eigenvalues, across the entire diagonal matrix.




```
PC1 accounts for  38.983911990657916 % of the variance
PC2 accounts for  34.65134980447652 % of the variance
PC3 accounts for  26.344872335725956 % of the variance
PC1, PC2, and PC3 account for  99.98013413086038 % of the variance
```

Figure 5. Percentage of variance explained by each principal component. The columns of P were extracted to obtain the principal components, and their associated eigenvalues account for the variance from that principal component. By dividing their eigenvalues by the total variance, we obtained the percentage of variance explained.

2.4 Data Visualization

After determining that the first three principal components account for 99.98% of the variation, we reduced the 18-dimensional alpha carbon atom coordinate data to 3 dimensions. The calculation $\mathbf{Y} = P^T \mathbf{X}$ was performed to project the data onto the principal component axes. Atoms with high variability were then extracted by considering all principal components. We then identified atoms with high variability based on the results. Variability was determined by taking the norm of the corresponding x, y, and z values along each principal component (PC)

axis. A high principal component value indicates that the data point exhibits significant variability along the corresponding principal component direction. This serves as the basis for the calculation described above. The code for these calculations are shown below.

```
# Reducing and transforming data into new basis
top3_P = P[:, :3]
Y = np.dot(top3_P.T, recentered_data)
print(Y.shape)
Y= Y.T

# Calculating Euclidean norm across the columns of Y
# This is done to tell us how much of the variation from the data was due
to each specific atom (the columns of Y)
total_variation = np.sqrt(np.sum(Y**2, axis=1))

# Extracting the top 100 varied atomic indices
top_atoms_indices = np.argsort(total_variation)[-100:]
```

For visualization, we used the `Matplotlib` library to plot the projected data along the principal component axes as seen in Figures 6 and 7.

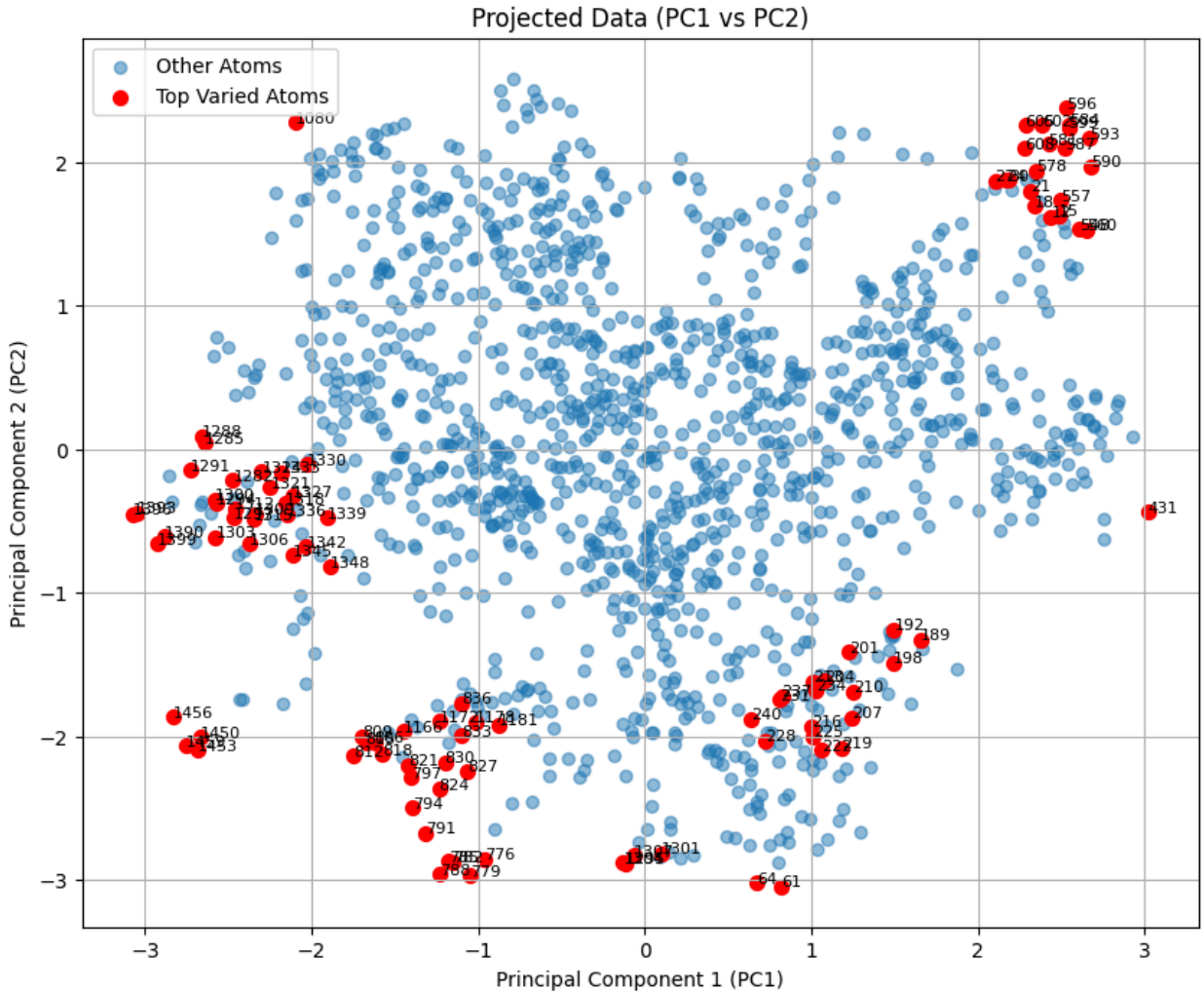


Figure 6. The transformed data plotted along the first two principal components. After PCA, the data was truncated after the first three principal components. The data is then projected using the matrix P^T and plotted along the first two principal components. PC1 and PC2 account for 73.64% of the total variation. The 100 atoms with highest variations are highlighted in red.

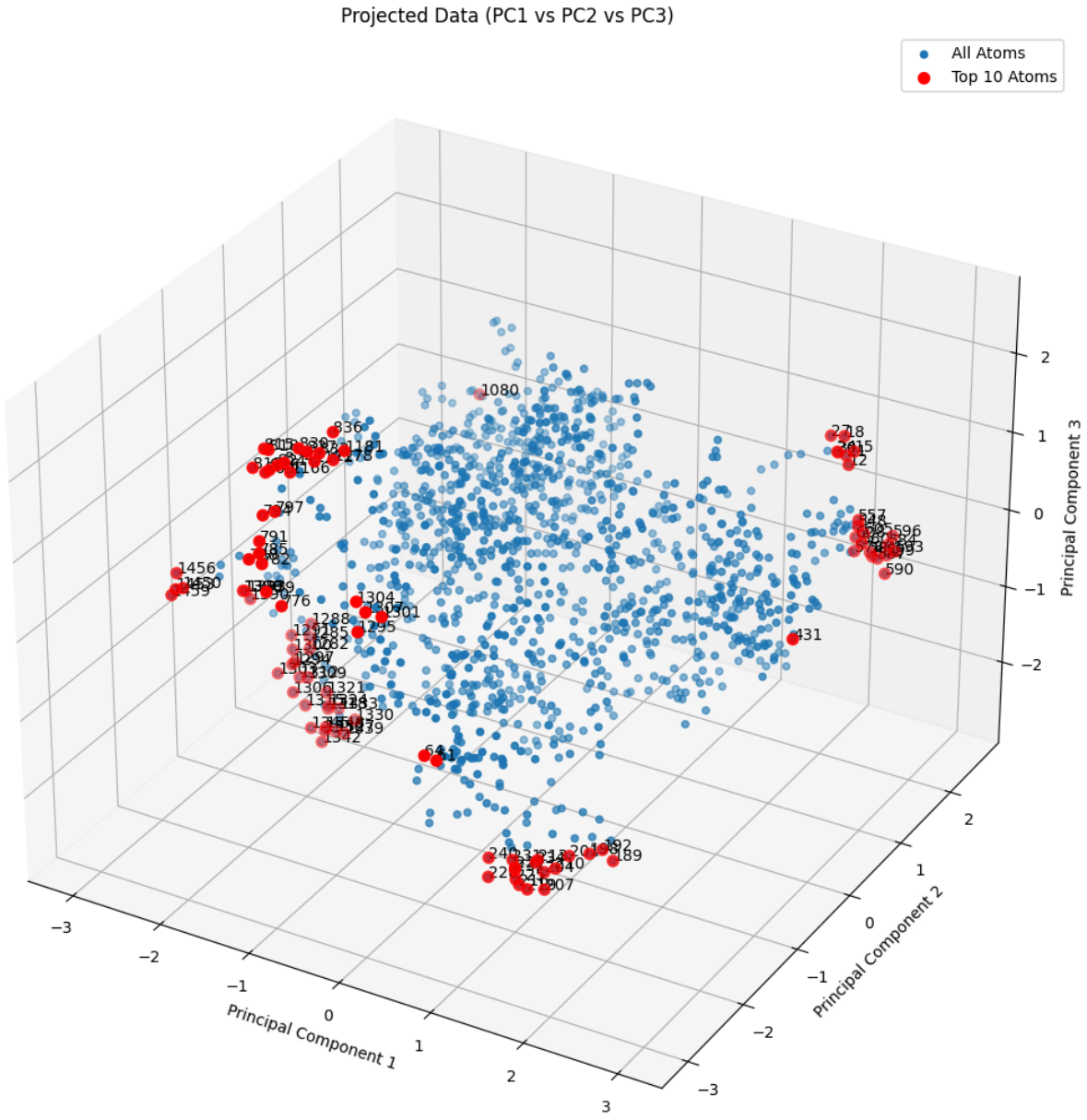


Figure 7. The transformed data plotted along the first three principal components. After projecting the data onto a new basis, the data is plotted on the first three principal components which explain 99.98% of the variance.

Finally, after re-indexing the varied atoms to match their residue numbers, five distinct regions were identified as having the largest variation. These regions include residues 12-30, 189-240, 548-608, 776-836, and 1282-1459. We then used model 6UEN as a reference to identify these regions in PyMOL.

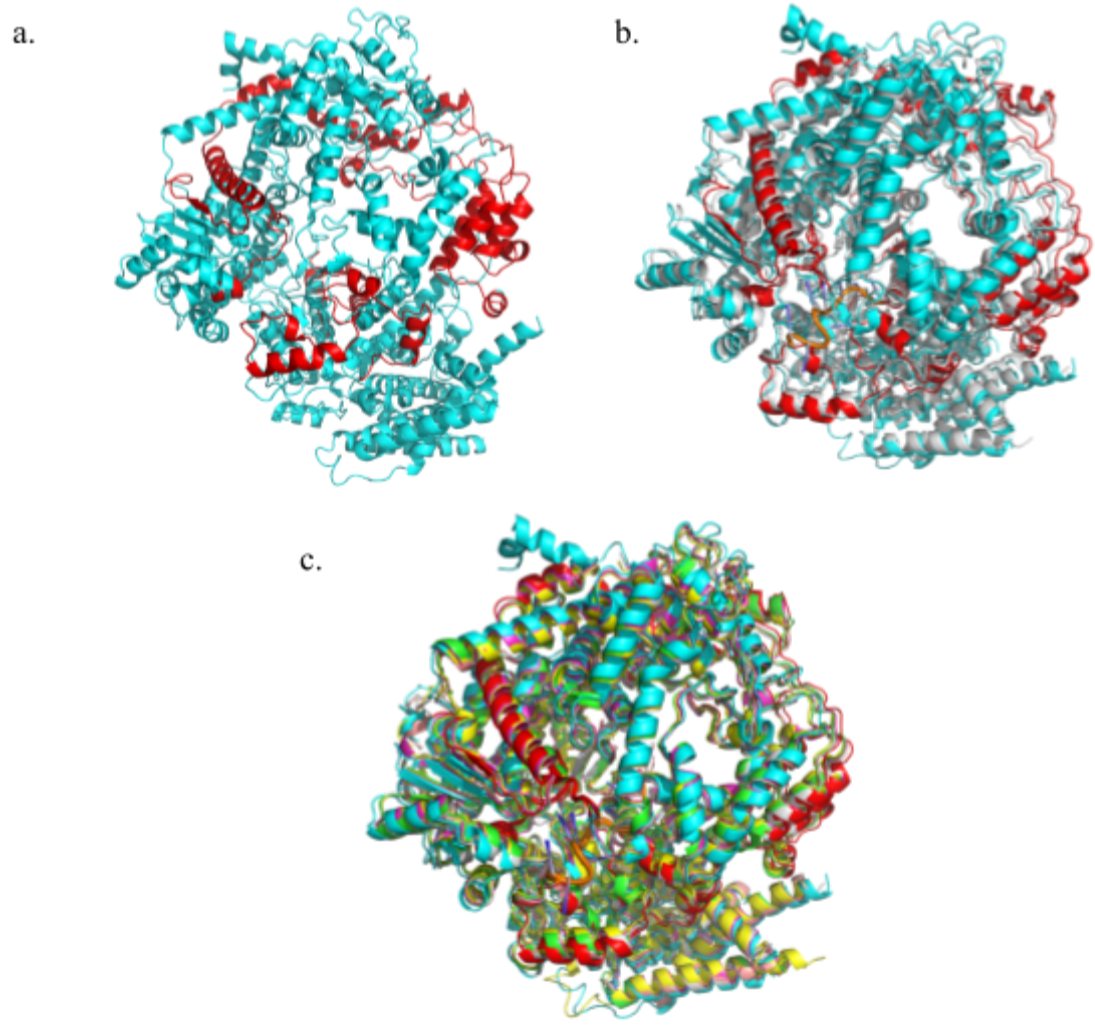


Figure 8. Structures of L with highly varied regions highlighted. a. The regions of highly varied atoms are highlighted in red while using structure 6UEN as a reference. b. Model 8SNX (white) is shown with 6UEN (cyan), along with the highlighted regions in yellow. c. All six models used for this project are shown with the highly varied atoms highlighted.

3. Discussion

Using PCA, we were able to identify specific regions on the L protein that show high variation across different different conformations. We structured our matrix to show 18 features, corresponding to the x, y, and z directions of all six of our structures. Meanwhile, we used the alpha carbons from the protein as our observations. As a result, we were able to identify the variation across these 18 features, and consequently, across the six structures. After finding the matrix P, we were also able to identify the top three principal components. Thus, we were able to

reduce the dimensionality of our data. Finally, by projecting our data using the matrix P^T, we were able to identify which atoms had the highest variation across the 18 features.

From the data, we were able to identify 5 regions on L that varied greatly. This suggests that these regions are more flexible than others which may play a role in its interaction with other molecules. We can see in the models that many of the highly varied regions are the loops connecting one region to another. This makes sense with our understanding of proteins, as these regions tend to be more flexible. In addition, as seen in Figure 8c., residues 548-608 share close proximity to L's binding pocket with RNA in models 8SNX and 8SNY (Cao et al., 2024). The variation in this region makes sense because as L interacts with other genetic material, it must rearrange and shift to accommodate the new substrates. Similarly, the large variation in 1282-1459 also makes sense with published results. Model 8FPI was shown as L interacting with an inhibitor called MRK-1 and 8FU3 shows L interacting with another inhibitor JNJ-8003 (Kleiner et al., 2023; Yu et al., 2023). Both of these inhibitors have binding pockets in the same area which corresponds to the 1282-1459 region. These findings corroborate our results, indicating that the regions identified do correspond to some protein flexibility.

However, there are some limitations to using PCA as a technique for model comparison. One difficulty is finding data to perform this analysis. Because the difficulty in solving structures can vary widely, it may be difficult to find other models to serve as comparisons. In addition, because intrinsically disordered regions do not follow a consistent pattern, it is difficult for scientists to resolve these regions in a structure. As a result, our data will tend to be biased towards repeated and consistent conformations of the protein.

In addition, another consideration is the three-dimensional nature of our data. Our project treats these three dimensions as parts of the features. However, it is possible that splitting apart these triplets of values could separate meaningful information about spatial orientation. This may limit the applications of this variability to the real world. Despite this limitation, we can still use $X=PY$ to return to the original three-dimensional values after PCA.

In the future, considering other methods that minimize information loss while retaining vector information—such as deep learning techniques (e.g. recurrent neural network models)—could yield even better results. These approaches would allow for more sophisticated handling of structural or atomic data with a greater degree of accuracy and contextual understanding.

References

Cao, D., Gao, Y., Chen, Z., Gooneratne, I., Roesler, C., Mera, C., D'Cunha, P., Antonova, A., Katta, D., Romanelli, S., Wang, Q., Rice, S., Lemons, W., Ramanathan, A., & Liang, B. (2024). Structures of the promoter-bound respiratory syncytial virus polymerase. *Nature*, *625*(7995), 611-617. https://doi.org/10.1038/s41586-023-06867-y

Cao, D., Gao, Y., Roesler, C., Rice, S., D'Cunha, P., Zhuang, L., Slack, J., Domke, M., Antonova, A., Romanelli, S., Keating, S., Forero, G., Juneja, P., & Liang, B. (2020). Cryo-EM structure of the respiratory syncytial virus RNA polymerase. *Nat Commun*, *11*(1), 368. https://doi.org/10.1038/s41467-019-14246-3

Collins, P. L., Fearns, R., & Graham, B. S. (2013). Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease. *Curr Top Microbiol Immunol*, *372*, 3-38. https://doi.org/10.1007/978-3-642-38919-1_1

Gilman, M. S. A., Liu, C., Fung, A., Behera, I., Jordan, P., Rigaux, P., Ysebaert, N., Tcherniuk, S., Sourimant, J., Eleouet, J. F., Sutto-Ortiz, P., Decroly, E., Roymans, D., Jin, Z., & McLellan, J. S. (2019). Structure of the Respiratory Syncytial Virus Polymerase Complex. *Cell*, *179*(1), 193-204 e114. https://doi.org/10.1016/j.cell.2019.08.014

Kleiner, V. A., T, O. F., Howe, J. A., Beshore, D. C., Eddins, M. J., Hou, Y., Mayhood, T., Klein, D., Nahas, D. D., Lucas, B. J., Xi, H., Murray, E., Ma, D. Y., Getty, K., & Fearns, R. (2023). Conserved allosteric inhibitory site on the respiratory syncytial virus and human metapneumovirus RNA-dependent RNA polymerases. *Commun Biol*, *6*(1), 649. https://doi.org/10.1038/s42003-023-04990-0

Li, Y., Wang, X., Blau, D. M., Caballero, M. T., Feikin, D. R., Gill, C. J., Madhi, S. A., Omer, S. B., Simoes, E. A. F., Campbell, H., Pariente, A. B., Bardach, D., Bassat, Q., Casalegno, J. S., Chakhunashvili, G., Crawford, N., Danilenko, D., Do, L. A. H., Echavarria, M., . . . investigators, R. (2022). Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. *Lancet*, *399*(10340), 2047-2064. https://doi.org/10.1016/S0140-6736(22)00478-0

Thomas Hamelryck, P. C., Joe Greener, Rob Miller, Lenna X. Peterson, Joao Rodrigues, Kristian Rother, Eric Talevich. (2024, 2024). *Bio.PDB Package Documentation*. Retrieved December 6, 2024 from https://biopython.org/docs/latest/api/Bio.PDB.html

Yu, X., Abeywickrema, P., Bonneux, B., Behera, I., Anson, B., Jacoby, E., Fung, A., Adhikary, S., Bhaumik, A., Carbajo, R. J., De Bruyn, S., Miller, R., Patrick, A., Pham, Q., Piassek, M., Verheyen, N., Shareef, A., Sutto-Ortiz, P., Ysebaert, N., . . . Jin, Z. (2023). Structural and mechanistic insights into the inhibition of respiratory syncytial virus polymerase by a non-nucleoside inhibitor. *Commun Biol*, *6*(1), 1074. https://doi.org/10.1038/s42003-023-05451-4