Identification of a Third Conformer using ProtTech Experimental Drug

M.T.D.R.Dolan
School of Chemistry, University of Bristol.
(Dated: November 22, 2023)

The objective of this analysis is to test whether or not the novel drug leads to a third conformation of a specific polytropic transmembrane α-helical protein. This is done through a PCA analysis of protein pair distance generated through a classical molecular dynamics simulation. The results find that the drug does indeed form a third protein conformation.

1 Introduction

1.1 Protein Conformation

Protein conformation may be defined as the arrangement in space of its constituent atoms which determine the overall shape of the molecule. The conformation of the protein arises from the bonding arrangements within its structure. [1]

There are 4 distinct levels of this structure, as shown in figure 1. The drug in question will be affecting the tertiary structure of the protein.

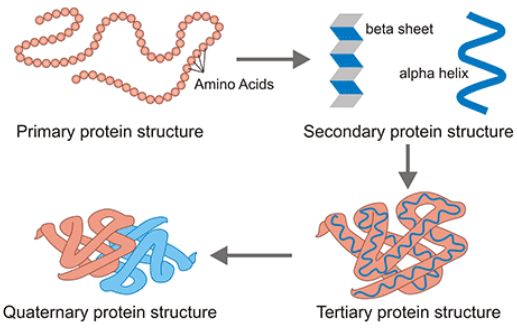


FIG. 1: Diagram showing the four levels of protein structure. [2]

The data represents the 3-dimensional Cartesian coordinates of 8 of the amino acids in the primary protein structure, these form 4 protein pairs: Arg1-Gly1, Ala1-Gly2, His1-Thr1, Pro1-Ala2. The aim of this study is to investigate whether variations in the distances between the amino acids in their pairs during a simulation provides evidence that the drug allows for a third conformer.

It is also known that in the simulation where the drug is present, the first and last time point represent the two known conformers.

1.2 Classical Molecular Dynamics

Molecular dynamics simulates a molecular system by numerically analysing the movement of particles within the system and evolving this over each time period. Specifically classical molecular dynamics means that the particles obey the laws as governed by classical mechanics; quantum effects are ignored.

This is much less labour intensive than an analytic system due to the sheer scale of most systems. It also means that any observable quantity must be able to be expressed as a function of a particle's position and momentum, although in this case it is just these two measurements in which we are interested.

```
program MD          basic MD code
[...]
setlat              function to initialize positions x
initv(temp)         function to initialize velocities vx
t=0
while (t < tmax) do main MD loop
   FandE            function to compute forces and total energy
   Integrate-V      function to integrate equations of motion
   t=t+delt         update time
   sample           function to sample averages
enddo
end program
```

FIG. 2: The typical code structure of a classical molecular dynamics program. [3]

We can use the Newtonian equations of motions to form the following numerical evolutions of each molecules position and velocity [4]:

$$\begin{aligned} x_i(t+\Delta t) &= x_i(t) + v_i(t)\Delta t + \frac{1}{2}a_i(t)\Delta t^2 \\ v_i(t+\Delta t) &= v_i(t) + \frac{1}{2}[a_i(t) + a_i(t+\Delta t)]\Delta t \end{aligned} \tag{1}$$

These two equations are also known as the Velocity-Verlet algorithm and the process of integrating them into the program is indicated by the **Integrate-V** section in figure 2.

These equations are used at each time step to calculate the change in position and momentum of each particle instead of having to solve an integral.

In order to calculate our data, this was done on all the molecules that make up the amino acids within the protein, as well as those in the drug, in the drug simulation. These were then used to calculate the positions of the centres of mass of the amino acids.

2 Analysis and Discussions

2.1 Raw Distances

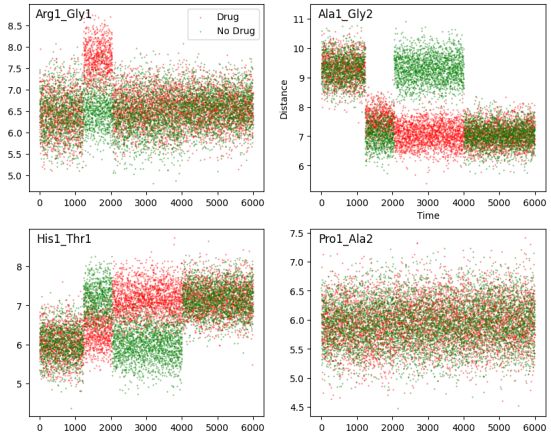


FIG. 3: Graphs showing the distance versus time for all 4 protein pairs.

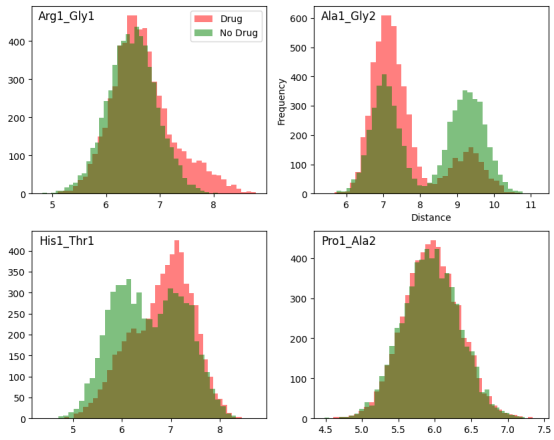


FIG. 4: Histograms plotting the distance frequency in the pairs.

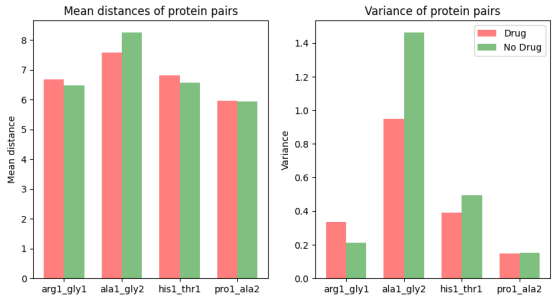


FIG. 5: Bar charts of the means and variances of the pairs.

As shown in figure 3, the addition of the drug clearly produces a change in the pattern of distances, however, it is difficult at this stage to exactly quantify this. It seems that for each pair the distances varies between 2 means (except Pro1-Ala2), this is shown more clearly in the graphs for Ala1-Gly2 and His1-Thr1 in figure 4.

2.1 PCA analysis

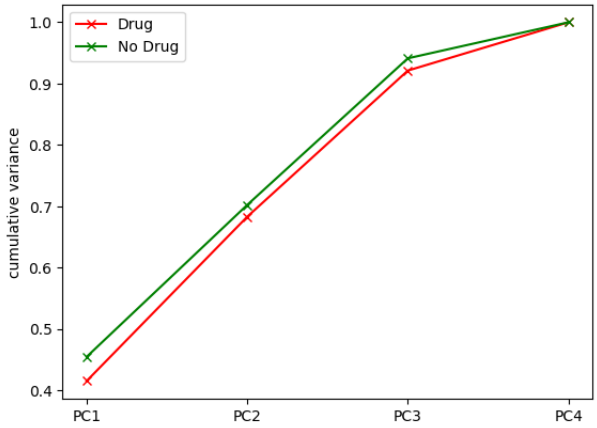


FIG. 6: The cumulative sum of the ratio of the variance described by each principal component.

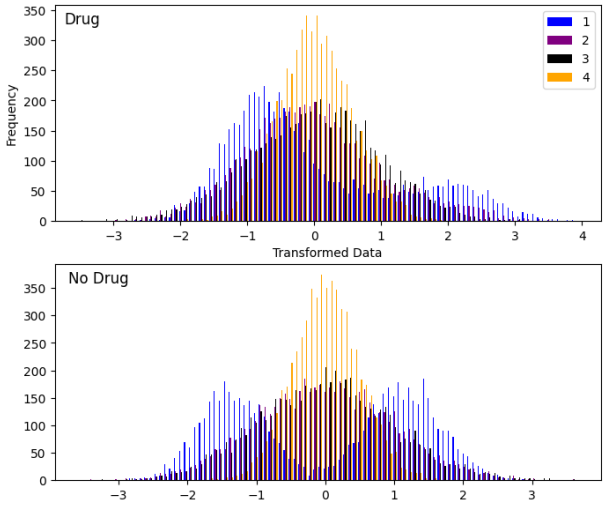


FIG. 7: Histogram of the PCA transformed data.

As seen in figure 6, the variation in the data can be explained almost entirely by the first 3 principal components, and the first 2 explain 70%. This is supported by figure 7 where we can see the bell curve for principal component 4 is the most narrow, indicating low variance. The input data to the PCA model was scaled to ensure greater accuracy.

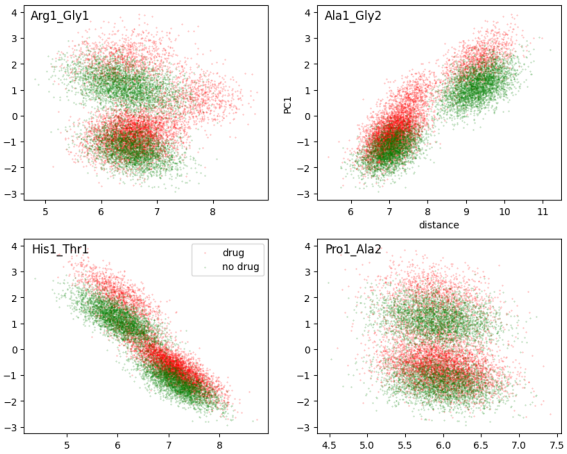


FIG. 8: Graphs of the 1st principal component versus the distances.

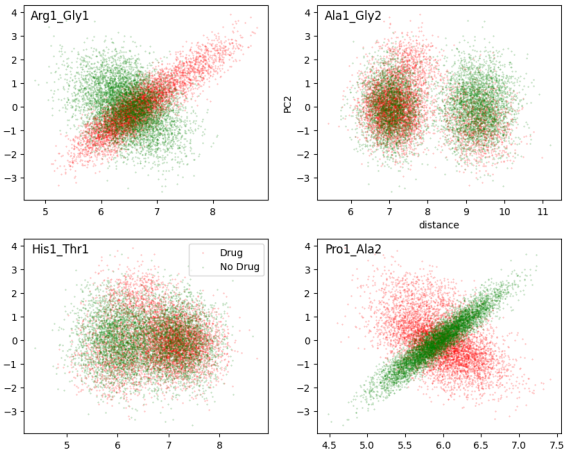


FIG. 9: Graphs of the 2nd principal component versus the distances.

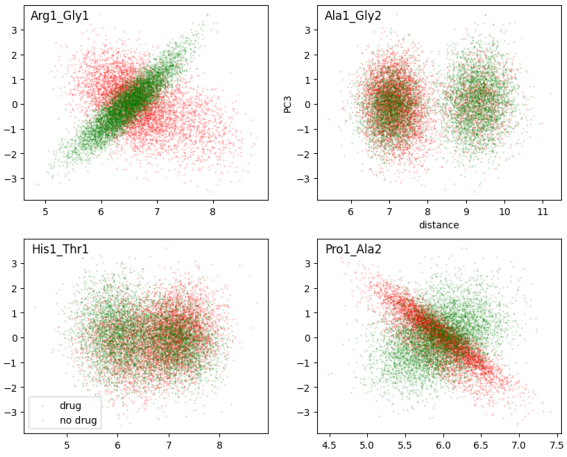
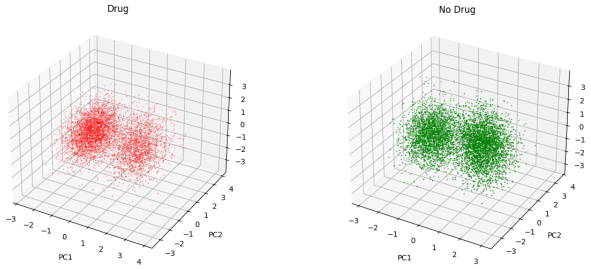


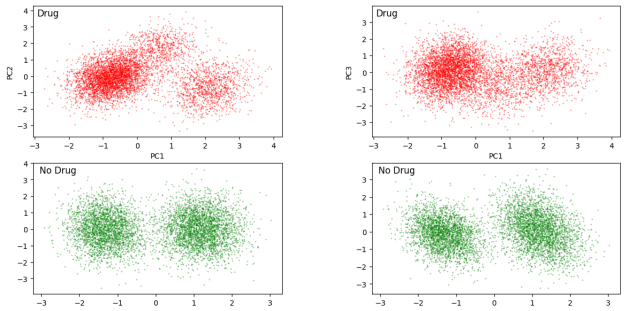
FIG. 10: Graphs of the 3rd principal component versus the distances.



(a) Drug (b) No Drug

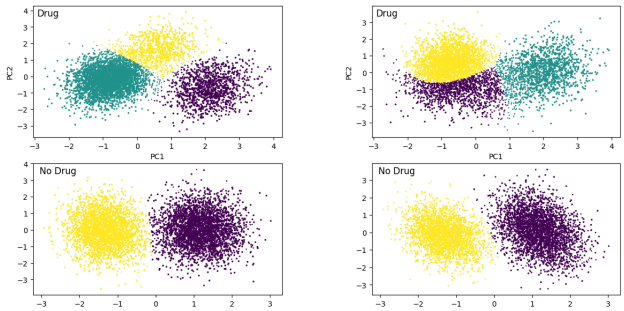
FIG. 11: 3D plots of the the first 3 principal components.

Figures 8, 9, and 10 show the principal components compared to the distances for each protein pair. We can see in figure 8 that we find the best clustering from principal component 1; for components 2 and 3, shapes in the drug data seems to be just the inverse of clustering in the no drug data, which doesn't provide much information on a completely different third conformer. This is supported by the clear clustering of 3 for drugs and 2 for no drugs shown when comparing component 1 to components 2 and 3 in figure 12, whilst when all 3 are compared in figure 11, only 2 clear clusters are seen with and without the drug.



(a) PC1 vs PC2 (b) PC1 vs PC3

FIG. 12: Plots comparing principal component 1 with principal components 2 and 3.



(a) PC1 vs PC2 (b) PC1 vs PC3

FIG. 13: Gaussian clustering of plots in figure 12.

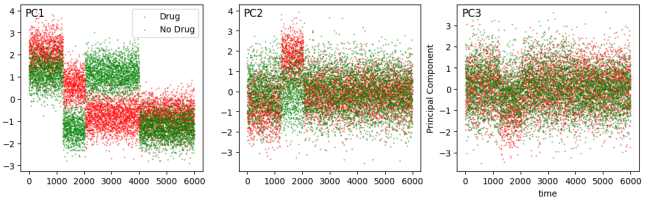


FIG. 14: Graphs showing principal components versus time.

Given that when the when the drug is administered, the first and final timepoints represent the two known conformers, we can see in figure 14 that the protein is in the first conformer for the first 1200 timesteps, and the second conformer for the final 4000. This means that for the period between the 1200th and 2000th timestep, the protein, when administered the drug, can clearly be seen to be in a third conformer. Figure 15 compares the mean distances between pairs during the periods where both models have the protein in the same conformer, and the period that the drug model is in the third conformer.

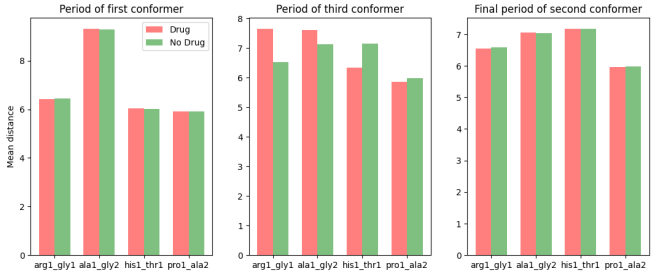


FIG. 15: Bar charts of mean distances during the periods of each conformer

3 Conclusion

It is clear that, as expected, the first principal component was much clearer in showing the three conformers in the drug model, and its comparison to the second principal produced three more distinct clusters than its comparison to the third principal component, but nevertheless they can be seen in both.

The plot of the first principal component versus time in figure 14 clearly shows a third conformer briefly forming during the 1200 - 2000 time period. In the no drug scenario the protein then switches from the first conformer in the 2000 - 4000 timestep period to the second in the 4000 - 6000 timestep period. Hence the reason that in figure 15 it is useful to only consider the time period when both models are in the second conformer, rather than just when the drug model is in the second conformer. But importantly, it shows that when the third conformer is present in the drug model there is a significant difference in mean distances between the raw drug and no drug distance data.

4 References

[1] J.C.Blackstock, *Guide To Biochemistry*, Butterworth-Heinemann, 1989

[2] Cusabio, https://www.cusabio.com/c-20943.html, (accessed November 2023)

[3] D.Frenkel and B.Smit, *Understanding Molecular Simulation*, Academic Press, 1996

[4] A.R.McCluskey, J.Grant, A.R.Symington, T.Snow, J.Doutch, B.J.Morgan, S.C.Parker and K.J.Edler, *J Appl. Crystallogr.*, 2019, **52**, 665-668