Investigating hydrogen production enzymes using gaming GPUs

Weining Wang *

University School, OH 44022, US

* Corresponding Author Email: wweining06@gmail.com

Abstract. Hydrogen production is crucial to use hydrogen as an energy source. Hydrogenase, an enzyme found in green algae, plays a key role in this process. The function of an enzyme is determined by its structure, and the structure of hydrogenase suggests that certain residues in the proton transfer pathway (PTP) are critical for its function. Simulation results from the initial study support the hypothesis that mutations in these residues can alter enzymatic activity by changing the dynamics of amino acids in the proton transfer pathway. The goal of the current study was to understand which structural and geometric features of the PTP in hydrogenase affect its efficiency in hydrogen production, and to predict mutations that could improve this efficiency. To investigate the relationship between protein structure/dynamics and function in this enzyme, molecular dynamics simulations and various analysis techniques, including GPU cards and cloud computing, have been used to study eight variations of hydrogenase. Based on the data, mutations in the PTP of the protein can affect its dynamics and might be linked to changes in its efficiency. The C299A has similar structure and behavior to the normal protein, but the loss of a key sulfur atom in its structure significantly reduces its effectiveness. The charged D variant of C299 had higher activity than the neutral S variant while both had different behavior than the Wildtype. It paves way for future expansion to other variants, lengthening of simulation time, establishment of complete structure and functional relationships, and prediction of activity enhanceable amino acid in silico mutagenesis.

Keywords: Hydrogen; energy source; Hydrogenase; Experiment.

1. Introduction

Aextracted from the Earth's crust (Richard, 2017) Fossil fuels have had significant impacts on the planet, including global warming, sea level rise, changes in agriculture, an increase in severe weather, and the spread of tropical diseases (Melissa, 2016). The only way to reduce the frequency of these disasters is to transition away from dependence on fossil fuels. However, many barriers to this transition remain, including erratic production, weather and climate constraints, cost issues, immature mining technologies, and potential hazards (BBC bitesize, 2021). Hydrogen energy has the potential to replace conventional energy sources as the most efficient, environmentally friendly and promising new energy source. It can also contribute to achieving a carbon-free economy and support future scientific breakthroughs (Fairley, 2021). One of the main advantages of hydrogen energy is that it is clean and environmentally friendly. It is non-toxic, non-polluting, and the only product of combustion is water, making it a potential solution for reducing carbon emissions and mitigating the greenhouse effect (TWI, 2021).

Although the potential of hydrogen energy is widely recognized, there is still a question of how to produce it efficiently. Currently, the primary method of producing hydrogen energy is through electrolysis of water (Office of Hydrogen and Fuel Cell Technologies, 2020). However, when it comes to producing hydrogen on a large scale, it may be more efficient to use strains of algae like Cladosporium rhizogenes that produce hydrogen directly from solar energy. Creating an anaerobic environment is key to inducing hydrogen production through the use of an enzyme called hydrogenase. Hydrogenase has an efficient proton transfer pathway that is critical to its function. This pathway shuttles protons from the bulk water to the catalytic center as a substrate or releases products from the catalytic center into the bulk water. According to Senger et al (2018), the proton transfer pathway consists of five residues: Cysteine 169 (C169), Glutamic acid 141 (E141), Serine 189 (S189),

Glutamic acid 144 (E144), and Arginine 148 (R148). Mutation of these residues to other amino acids reduces hydrogen production activity.

In order to study the effect of residue mutations on hydrogen-producing activity, it is necessary to first determine the structural and geometric characteristics of these mutated residues. Chemical reactions occur rapidly and it is difficult to experimentally map each step in the chemical process (Karplus et al., 2013). Molecular dynamics simulations (MD), use molecular mechanics from classical physics to analyze the physical motion of atoms and molecules over a fixed period of time to understand the "evolution" of the system (Hollingsworth et al., 2018). Graphics processing units (GPUs) are able to run the data needed to simulate the structure and action of different mutations in protein dynamics (Rich, 2020).

In a previous study, the goal was to identify the structural dynamics in the protein sequence that determine hydrogen production function. Senger et al (2018) examined a series of mutations along the proton transfer pathway and demonstrated that they had different effects on hydrogen production activity. To test the hypothesis that mutations along the proton transfer pathway affect enzyme activity by altering the dynamic properties of amino acids, the first phase of the experiment used gaming GPU cards to perform molecular dynamics simulations on four hydrogenase site-directed mutagenesis variants (Table 1) to assess the structure and dynamic role of residues along the proton transfer pathway on hydrogen production efficiency. The structure and dynamics differences between the wild-type protein and mutants were identified, and these structural features were associated with the functional activity of hydrogenase.

To do this, Charmm-Gui was used to prepare input files for molecular dynamics (MD) simulations for each hydrogenase variant (E141A, E141Q, E144A, and C377H). Colab Notebooks were used for MD simulation, following the instructions to set up and perform 5-nanosecond MD simulations for each variant. After the MD simulations, Colab Notebooks directly computed and plotted various measures, including the root-mean-square deviation (RMSD), radius of gyration (Rg), root-mean-square fluctuation (RMSF), eigenvectors of Principal Component Analysis (PCA), and Pearson's Cross-Correlation (CC). The data showed that global structures, such as the radius of gyration for the whole protein, were similar during the 5 nanosecond MD simulations. However, 2D RMSD and PCA showed that the variants had different time-dependent dynamics. Using RMSF and Pearson's cross-correlation, the different dynamic features along the proton transfer pathway were identified.

Specifically, the mean objective of the current study is to determine the structural/geometric features on hydrogenase proton transfer pathway that affect the efficiency of hydrogen production, and predict the mutations that can improve the efficiency of hydrogen production. To do so, the MD simulations have been expanded to include all 11 variants (Table 2) to validate that in silico mutagenesis yields the same dynamic information compared to previous simulations starting from X-ray crystallographic structures, and to establish structure/dynamics and function relationship of proton transfer pathway in determining the efficiency of hydrogen production. The identification of dynamics in the wild type preserved in the wild type structure have provided insights into how to reverse engineer the protein to enhance hydrogen production. The activity assay results in Senger, M. et al (2018) are the key data for associating the observed dynamics difference with different functional outcomes observed in that study. Ultimately, the hydrogen production has been optimized using molecular dynamics simulation results from gaming GPU cards.

Table 1. Variants of hydrogenase used in the initial study (Hofmann et al, 2018).

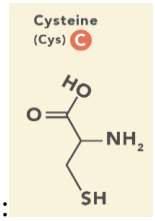
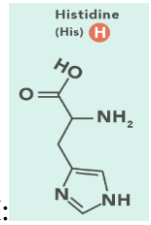
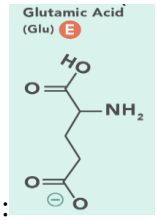
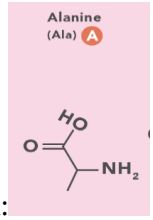
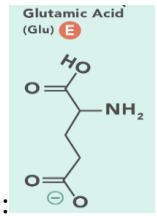
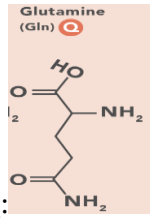
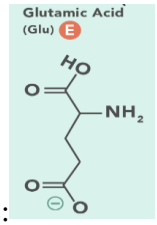
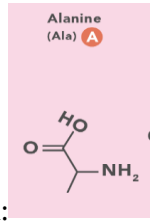
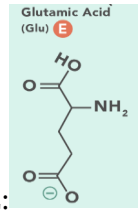
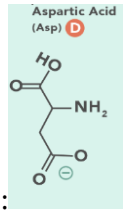
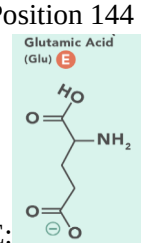
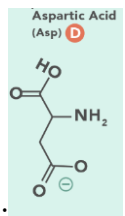
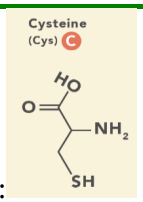
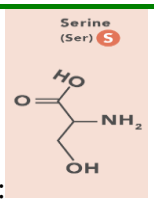
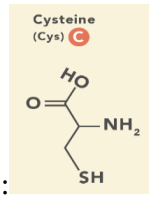
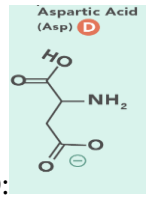
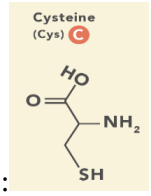
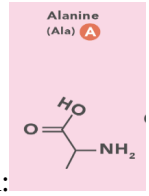
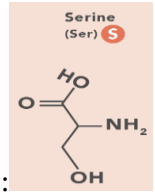
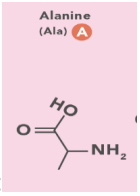
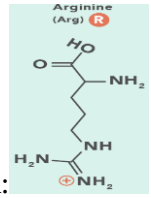
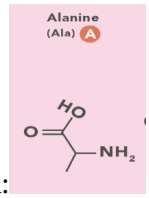
Wild Type	Mutation	Position	Experimental method	Resolution	Simulation length (ns)
	C377H	Position 377 C:  to H: 	X-ray diffraction	1.80 Å	5
	E141A	Position 141 E:  to A: 	X-ray diffraction	1.45 Å	5
4XDD	E141Q	Position 141 E:  to Q: 	X-ray diffraction	1.61 Å	5
	E144A	Position 144 E:  to A: 	X-ray diffraction	2.14 Å	5

Table 2. Variants of hydrogenase that will be use in the current study.

Wild Type	Mutation	Position	Experimental method	Simulation length (ns)
	E141D	Position 141 E:  to D: 	In silico mutagenesis	5
	E144D	Position 144 E:  to D: 	In silico mutagenesis	5
	C169S	Position 169	In silico mutagenesis	5

		C:  to S: 		
4XDD	C169D	Position 169	In silico mutagenesis	5
		C:  to D: 		
	C169A	Position 169	In silico mutagenesis	5
		C:  to A: 		
	S189A	Position 189	In silico mutagenesis	5
		S:  to A: 		
	R148A	Position 148	In silico mutagenesis	5
		R:  to A: 		

2. Method

2.1. MD simulation setup

Charmm-Gui (Kim et al, 2008) was used to prepare input files for MD simulations for each protein (Table 2). Specific Colab Notebooks were used for MD simulation (Arantes et al, 2021), the main procedure for this experiment was to follow the instructions on Colab Notebooks to set up and perform molecular dynamics: The first step was to set the environment for MD calculation, which included installing dependencies, import Google Drive and check if the GPU nodes were correctly allocated. The second step was to load the necessary input files, which included providing the PRMTOP filename, CRD filename, PDB filename and Google Drive Path, then clicked the button to see the 3D structure of that specific protein. The third step was to equilibrate the simulation box, the number of parameters for MD Equilibration protocol was set, which included setting the minimization steps to 1000, simulation time to 2 nanoseconds, integration time 2 femtoseconds, temperature to 298 Kelvin, pressure to 1 bar, position restraints force constant to 800 kJ/mol, frequency to write the trajectory file to 10 picoseconds, and frequency to write the log file to 10 picoseconds. Then an Equilibration MD simulation (NPT ensemble) was run at the end of step 3. The main job for step 4, which was the final step, was to run a Production MD simulation. But before that the input file names for equilibrated PDB and state files needed to be provided and the parameters for MD production

protocol needed to be set: set the simulation time to 1 nanosecond, number of strides (integers) to 5, integration timestep to 2 femtoseconds, temperature to 298 Kelvin, pressure to 1 bar, frequency to write the trajectory file to 10 picoseconds, and frequency to write the log file to 10 picoseconds. A total of 5 nanosecond simulation was performed for each variant.

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

2.2. Analysis of MD trajectories

After the MD simulation part, Colab Notebooks directly computed and plotted root-mean-square deviation (RMSD) as a distribution, computed and plotted radius of gyration (Rg) as a distribution, computed root-mean-square fluctuation (RMSF) and showed 2D RMSD, calculated eigenvectors of Principal Component Analysis (PCA) as distributions, and plotted Pearson's Cross-Correlation (CC).

The following equation was used to calculate the radius of gyration (Rg):

$$R_g = \left(\frac{\sum_i |r_i|^2 m_i}{\sum_i m_i}\right)^{1/2} \tag{1}$$

Where m_i is the mass of atom *i* and $\boldsymbol{r_i}$ the position of atom *i* with respect to the center of mass of the molecule. The calculation was implemented in the Colab notebook.

The following equation was used to calculate the root-mean-square deviation (RMSD):

$$\mathrm{RMSD}(t) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i \left|\mathrm{r}_i(t) - \mathrm{r}_i^{ref}\right|^2\right]^{1/2} \tag{2}$$

Where $M = \Sigma_i m_i$ and $\boldsymbol{r_i}(t)$ is the position of atom *i* at time *t* after least square fitting the structure to the reference structure. The calculation was implemented in the Colab notebook.

Principle component analysis (PCA) was used to reduce the dimensionality of MD simulation datasets. Each snapshot from MD simulation trajectories was projected onto the first and second principal components (PC1 and PC2) to illustrate the maximal variance, which represents the concerted motion of the protein.

The following equation was used to calculate the root-mean-square fluctuation (RMSF):

$$\mathrm{RMSD}_i = \left[\frac{1}{T}\sum_{t_j=1}^{T} \left|\mathrm{r}_i(t_j) - \mathrm{r}_i^{ref}\right|^2\right]^{1/2} \tag{3}$$

Where *T* is the time over which one wants to average and $\mathrm{r}_i^{\mathrm{ref}}$ is the reference position of particle *i*. This reference position will be the time-averaged position of the same particle *i*. The calculation was implemented in the Colab notebook.

The following equation was used to calculate the Pearson's Cross-Correlation (CC):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{4}$$

Where *r* is the correlation coefficient, $\mathrm{x_i}$ are values of the x-variable in a sample, $\underline{\mathrm{x}}$ is the mean of the values of the x-variable, $\mathrm{y_i}$ are values of the y-variable in a sample, and $\underline{\mathrm{y}}$ is the mean of the values of the y-variable. The calculation was implemented in the Colab notebook

All the graphs and data from the MD simulations will automatically be collected to the specific Google Drive folder that was filled in the settings.

3. Results

The radius of gyration (Rg), root-mean-square deviation (RMSD), Principal Component Analysis (PCA), root-mean-square fluctuation (RMSF), and Pearson's Cross-Correlation (CC) of protein's CA atoms for each variant of hydrogenase (Table 2) in sampling time of 5 nanoseconds have been calculated and graphed after the data collection.

The radius of gyration (Rg) tells how much the mass of an object is spread out around its center. Figure 1 shows the Rg values for different protein variants with mutations in their Cα atoms. These results show that the size of the proteins is not significantly changed by the mutations, with differences of less than 0.4 angstroms compared to the original, "wild type" protein. This means that the mutations do not significantly alter the overall structure of the protein. However, it's worth noting that the E279D variant had a different type of Rg distribution compared to the other proteins, which may indicate that it behaves differently in some way.

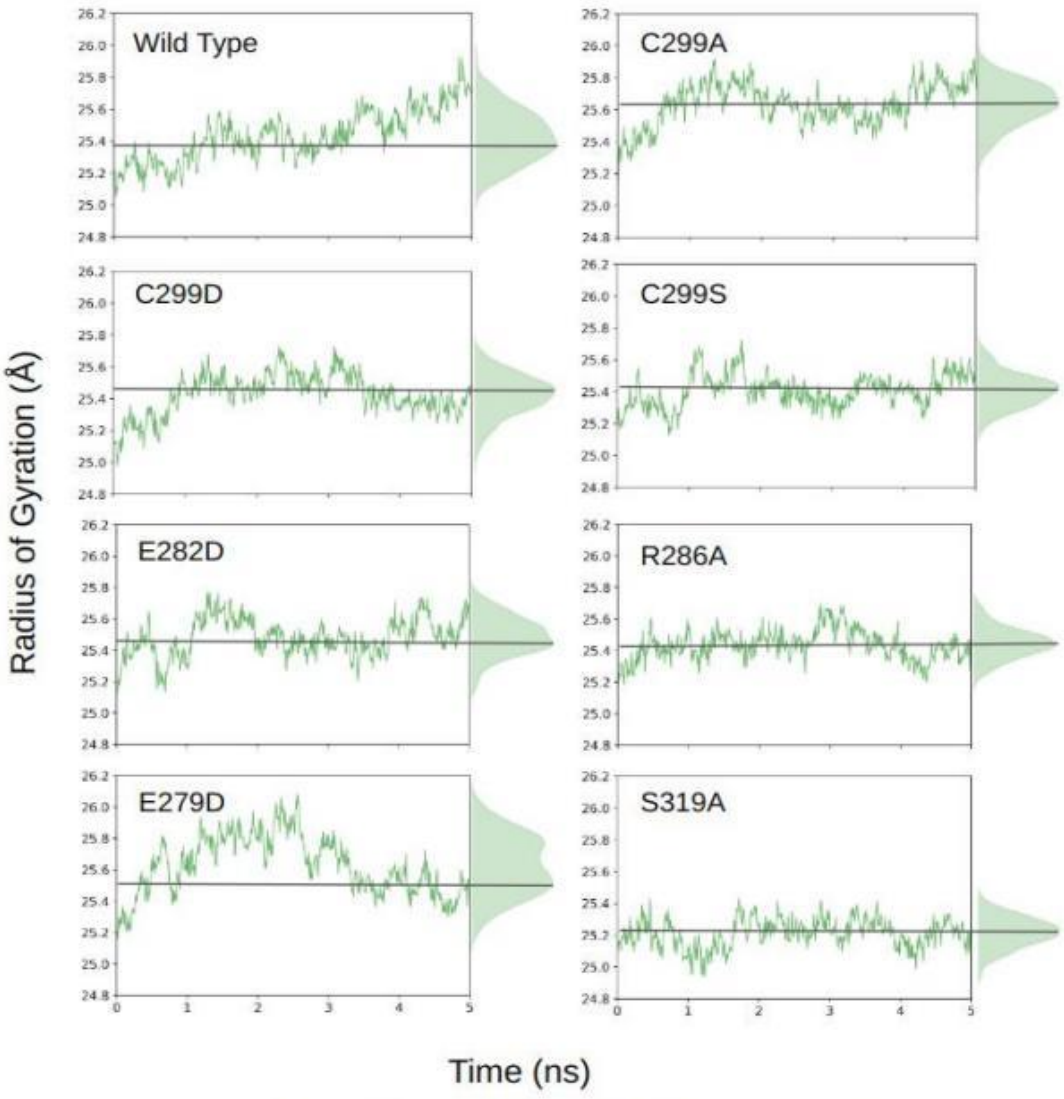


Fig 1. Radius of gyration (Rg) of protein's Cα atoms over 5 ns simulation.

The stability of different proteins has been measured by looking at the root mean square deviation (RMSD) of their Cα atoms. The RMSD is a way to compare the 3D structure of a protein at different times and see how much it has changed. In this case, it has been found that all of the proteins stayed relatively stable, with the RMSD remaining within 3 angstroms of the starting structure. This suggests that the proteins did not change much in shape over the course of the 5 nanosecond simulations. However, the graph did notice some differences in the RMSD between the different protein constructs, which may indicate that they have different levels of stability.

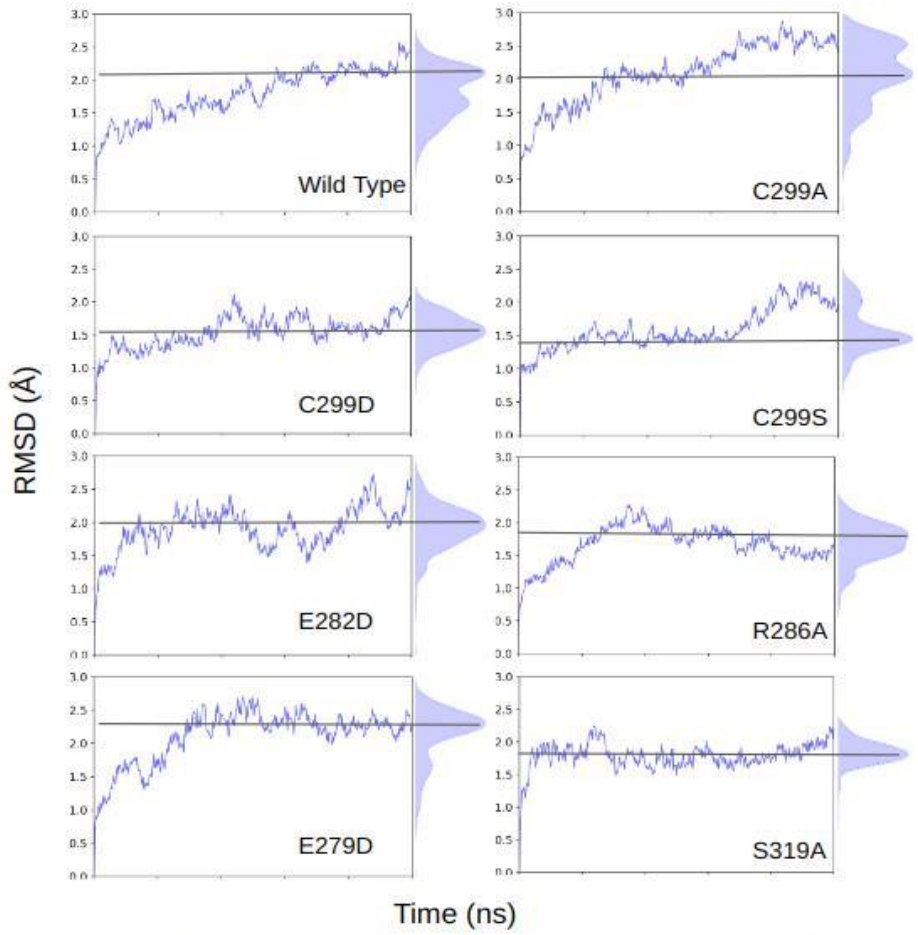


Fig 2. Time evolution of root mean square deviation (RMSD) of protein's Ca atoms with reference to t=0

In Figure 3, the changes in proteins' structures over time have been observed by looking at the 2D root mean square deviation (RMSD). This is calculated by comparing the RMSD of the protein at one time to the RMSD at another time over the course of 5 nanoseconds. The results show that the C299A protein has a structure that is more similar to the original, "wild type" protein, with the largest difference in RMSD between any two time points being 4 angstroms. On the other hand, the E282D and S319A proteins had the least amount of structural variation over the 5 nanoseconds.

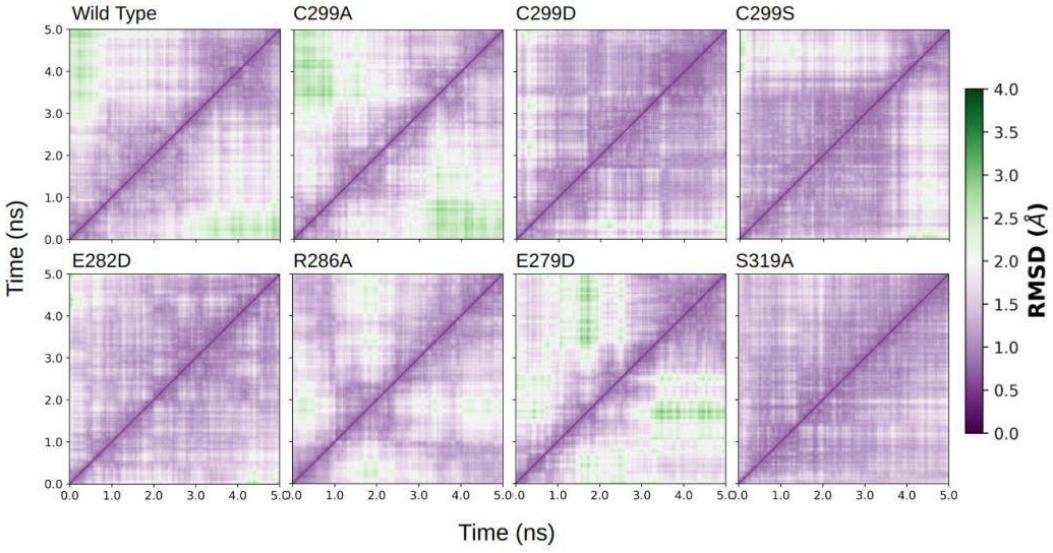


Fig 3. 2D RMSD of protein at time x compare to protein at time y over 5 ns simulation. The color bar of RMSD values was shown on the right.

Principal component analysis (PCA) is a way to study how proteins move and change over time. It involves finding the most important directions in which the protein's movement varies the most, and then creating a simplified representation of the data in those directions. This can be helpful for understanding patterns and trends in the data, as well as for identifying key ways that the protein moves. In the context of these proteins, PCA can be used to understand how the different mutants behave and how they are affected by different conditions. Figure 4 shows the PCA for eight proteins, with the first two components accounting for the majority of the protein dynamics. The C299A protein appears to have the most similar behavior to the original, "wild type" protein, while the other six mutants show more distinct patterns of movement.

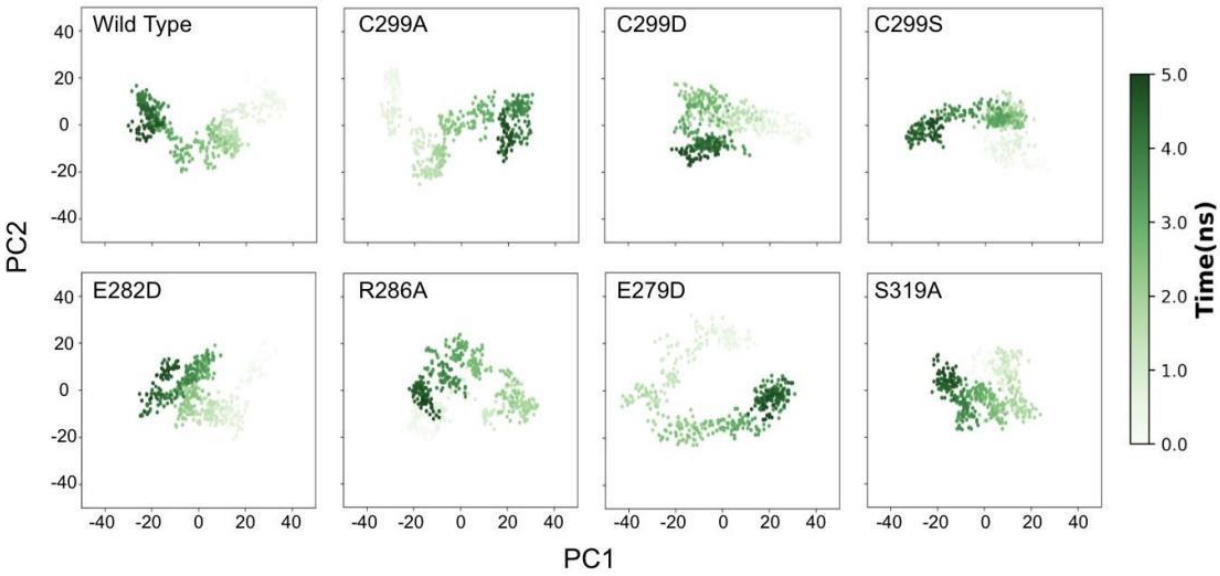


Fig 4. Principal Component Analysis (PCA) showing the projection of individual snapshots from MD trajectories onto the first two principal components.

The graph in Figure 5 shows how much a certain protein's atoms move around, on average, compared to a reference structure. This is called the root-mean-square fluctuation (RMSF). The RMSF for eight different mutations, or changes, to the protein was measured and compared. The results show that these eight mutations do not cause the protein to move significantly differently than the reference structure. This suggests that the changes in the protein's function seen in these mutations are probably not caused by the protein's movement.

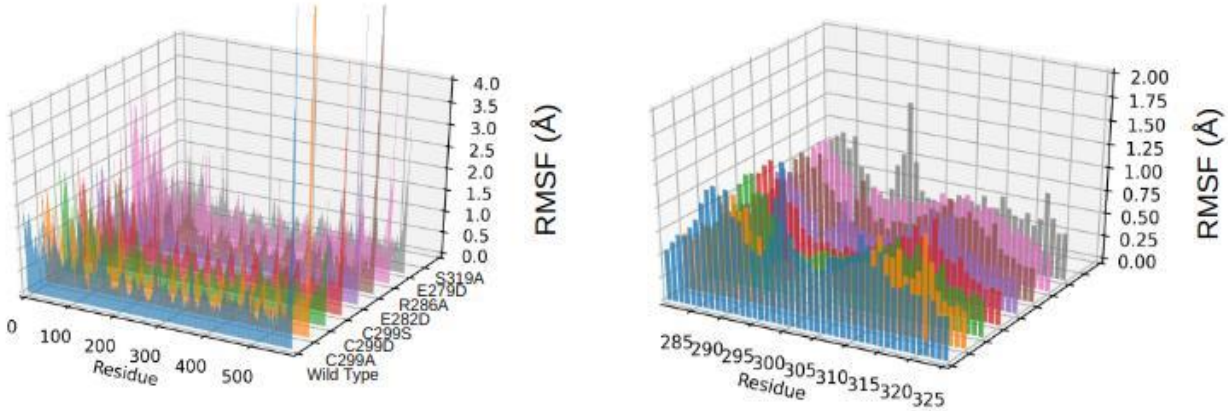


Fig 5. Root-mean-square fluctuation (RMSF) of protein's Ca atoms with reference to the average structure from 5 ns simulation (left) and zoom-in view of the RMSF for the region where mutations occur.

The graph in Figure 6 shows how much two parts of a protein move together or in opposite directions during a 5-nanosecond simulation. This is called Pearson's Cross-Correlation (CC). A positive CC value means that the two parts are moving in the same direction, while a negative value means that they are moving in opposite directions. The results show that there are differences in the movement of different parts of the protein along the proton transfer pathway and some other areas.

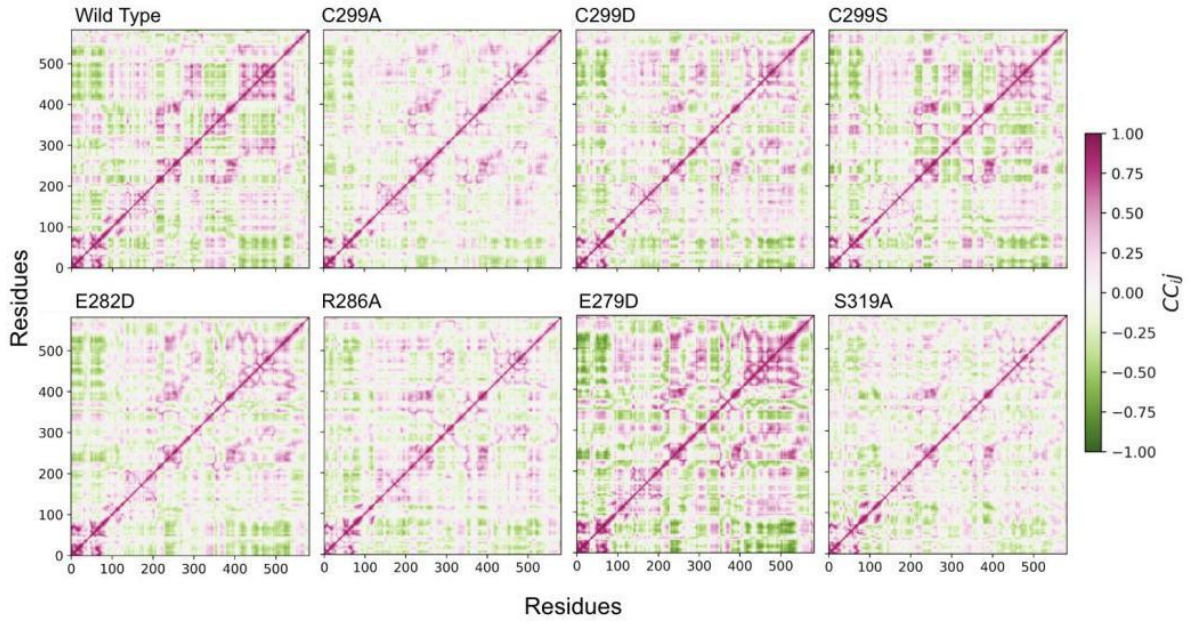


Fig 6. Pearson's cross correlation (CC) between each two residues during 5 ns simulation. The color bar corresponding to the CC values was shown on the right.

4. Discussion

In this study, computer simulations have been used to investigate the behavior of a normal hydrogenase protein and seven variations of it. Previous research has shown that the normal protein is the most effective, followed by R286A, E282D, C299D, E279D, S319A, C299S, and C299A (reference 11). Current simulation results showed that the C299A variant has similar behavior to the normal protein (Figure 4-6), but the loss of a key sulfur atom in its structure significantly reduces its effectiveness. Comparing C299D and C299S, it showed that the charged D variant had higher activity than the neutral S variant. However, both the C299D and C299S variants had a different behavior than the normal protein (Figure 4-6). Therefore, it is important to consider the role of protein behavior (dynamics) in its function in future studies. More detailed analysis of other variations, similar to C299, will be needed to fully understand the relationship between dynamics and function. Currently, these mutations do not have experimental structures available for simulation. In future studies, in silico mutagenesis will be used to test if similar results can be achieved with this approach. Additionally, it is not yet clear if simulating the protein's behavior for 5 nanoseconds is sufficient to accurately capture its key functions. The simulations will be for a longer period in future research. Once the simulations can explain decreased activity in Senger's study (2018), design in silico mutagenesis to predict which amino acid could enhance the activity. The in silico mutagenesis will be performed using ChimeraX (Goddard TD et al, 2021).

5. Conclusion

The simulation results suggest that changes (mutations) in the proton transfer pathway of the protein can affect its behavior (dynamics) and might be linked to changes in its effectiveness. To fully understand the relationship between a protein's behavior and its function, it is needed to study all

available mutations and compare them to experimental data. This will show a complete picture of how dynamics and function are connected.

References

[1] Hollingsworth, S., & Dror, R. (2018, September 19). Molecular dynamics simulation for all. Retrieved January 27, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6209097/

[2] Martin, K., Levitt, M., & Warshel, A. (2013, October 9). The Nobel Prize in chemistry 2013. Retrieved January 27, 2022, from https://www.nobelprize.org/prizes/chemistry/2013/press-release/

[3] Pell, R. (2020, March 18). Nvidia to gamers: Use your GPUs to fight COVID-19. Retrieved January 27, 2022, from https://www.smart2zero.com/news/nvidia-gamers-use-your-gpus-fight-covid-19

[4] Gray, R. (2017, March 13). The biggest energy challenges facing humanity. Retrieved January 27, 2022, from https://www.bbc.com/future/article/20170313-the-biggest-energy-challenges-facing-humanity

[5] Denchak, M. (2021, April 21). Are the effects of global warming really that bad? Retrieved January 27, 2022, from https://www.nrdc.org/stories/are-effects-global-warming-really-bad

[6] BBC News. (2021). Pros and cons of renewable energy resources - generation of electricity - n ational 4 physics revision - BBC bitesize. Retrieved January 27, 2022, from https://www.bbc.co. uk/bitesize/guides/zbsdmp3/revision/4

[7] Fairley, Peter. (2021, February 24). Cheap renewables could make green hydrogen a practical replacement for fossil fuels. Retrieved March 10, 2021, from

[8] https://www.technologyreview.com/2021/02/24/1017822/green-hydrogen-clean-energy/

[9] TWI. (2020). what are the pros and cons of hydrogen fuel cells? Retrieved January 27, 2022, from https://www.twi-global.com/technical-knowledge/faqs/what-are-the-pros-and-cons-of-hydrogen-fuel-cells

[10] Hydrogen and Fuel Cell Technologies Office. (2020). Hydrogen production: Electrolysis. Retrieved January 27, 2022, from https://www.energy.gov/eere/fuelcells/hydrogen-production-electrolysis

[11] Duan, J., Senger, M., Esselborn, J., Engelbrecht, V., Wittkamp, F., Apfel, U. . . . Winkler, M. (2018). Crystallographic and spectroscopic assignment of the Proton Transfer Pathway in [fefe]-hydrogenases. Nature Communications, 9(1). doi:10.1038/s41467-018-07140-x

[12] Kertess, L., Happe, T., & Hofmann, E. (2018). Apo [fefe]-hydrogenase HYDA1 from Chlamydomonas reinhardtii, variant c377h. doi:10.2210/pdb6gl6/pdb

[13] Duan, J., Engelbrecht, V., Esselborn, J., Hofmann, E., Winkler, M., & Happe, T. (2018). [fefe]-hydrogenase HYDA1 from Chlamydomonas reinhardtii,variant e141a. doi:10.2210/pdb6gm5/pdb

[14] Duan, J., Engelbrecht, V., Esselborn, J., Hofmann, E., Winkler, M., & Happe, T. (2018). [fefe]-hydrogenase HYDA1 from Chlamydomonas reinhardtii,variant E141Q. doi:10.2210/pdb6gm6/pdb

[15] Esselborn, J., Hofmann, E., Kurisu, G., & Happe, T. (2015). Apo [fefe]-hydrogenase CPI. https://doi.org/10.2210/pdb4xdd/pdb

[16] Arantes, P. R., Polêto, M. D., Pedebos, C., & Ligabue-Braun, R. (2021). Making it rain: Cloud-based molecular simulations for everyone. Journal of Chemical Information and Modeling, 61(10), 4852-4856. doi:10.1021/acs.jcim.1c00998

[17] Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry, 29(11), 1859-1865. doi:10.1002/jcc.20945

[18] UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. Protein Sci. 2021 Jan; 30(1):70-82.