molecules

MDPI

Article

Mechanism of Action of Non-Synonymous Single Nucleotide Variations Associated with α-Carbonic Anhydrase II Deficiency

Taremekedzwa Allan Sanyanga [1], Bilal Nizami [1,2] and Özlem Tastan Bishop [1,*]

[1] Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Grahamstown 6140, South Africa; asanyanga@gmail.com (T.A.S.); nizamibilal1064@gmail.com (B.N.)

[2] Institute of Materials and Environmental Chemistry, Research Centre for Natural Sciences of the Hungarian Academy of Sciences, Magyar tudósok körútja 2, 1117 Budapest, Hungary

* Correspondence: O.TastanBishop@ru.ac.za; Tel.: +27-46-603-8072; Fax: +27-46-603-7576

Received: 31 August 2019; Accepted: 24 October 2019; Published: 4 November 2019

Abstract: Human carbonic anhydrase II (CA-II) is a Zinc (Zn^{2+}) metalloenzyme responsible for maintenance of acid-base balance within the body through the reversible hydration of CO_2 to produce protons (H^+) and bicarbonate (BCT). Due to its importance, alterations to the amino acid sequence of the protein as a result of single nucleotide variations (nsSNVs) have detrimental effects on homeostasis. Six pathogenic CA-II nsSNVs, K18E, K18Q, H107Y, P236H, P236R and N252D were identified, and variant protein models calculated using homology modeling. The effect of each nsSNV was analyzed using motif analysis, molecular dynamics (MD) simulations, principal component (PCA) and dynamic residue network (DRN) analysis. Motif analysis identified 11 functionally important motifs in CA-II. RMSD data indicated subtle SNV effects, while PCA analysis revealed that the presence of BCT results in greater conformational sampling and free energy in proteins. DRN analysis showed variant allosteric effects, and the average *betweenness centrality (BC)* calculations identified Glu117 as the most important residue for communication in CA-II. The presence of BCT was associated with a reduction to Glu117 usage in all variants, suggesting implications for Zn^{2+} dissociation from the CA-II active site. In addition, reductions to Glu117 usage are associated with increases in the usage of the primary and secondary Zn^{2+} ligands; His94, His96, His119 and Asn243 highlighting potential compensatory mechanisms to maintain Zn^{2+} within the active site. Compared to traditional MD simulation investigation, DRN analysis provided greater insights into SNV mechanism of action, indicating its importance for the study of missense mutation effects in proteins and, in broader terms, precision medicine related research.

Keywords: precision medicine, carbonic anhydrase II; single nucleotide variation; allosteric effect; dynamic residue network analysis; MD-TASK

1. Introduction

Carbonic anhydrases (CAs) are metalloenzymes responsible for the catalysis of the reversible interconversion of carbon dioxide (CO_2) and water (H_2O) to bicarbonate (HCO_3^- or BCT) and protons (H^+), and the reaction is illustrated in Equation (1) [1,2]. At least six distinct CA families have been identified to date, namely, α (alpha), β (beta), γ (gamma), δ (delta), η (eta) and ζ (zeta) [3–6]. Vertebrates contain only the α family. The β-CAs are found in all organisms except for chordates [3,4,6]; γ-CAs are found widely in bacteria and plants in addition to methanogenic archaea [3,4]; δ-CAs are found in diatoms and in some marine algae [7]; and ζ-CAs are not found in some bacteria but only in diatoms [8]. It should be noted that the α family can also be found in bacteria. Although CAs are

found in a variety of organisms, these enzyme families do not contain significant amino acid sequence similarity and are regarded as an example of convergent evolution [9,10].

$$CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+ \quad (1)$$

The α-CA proteins in humans are divided into four subgroups depending on intracellular location, cytosolic, mitochondrial, secreted and membrane associated, with each consisting of several isoforms. The isoforms across all subgroups total 15 enzymes. CA-I, II, III, VII and XIII belong to the cytosolic subgroup, while CA-IV, IX, XII and XIV are members of the membrane associated subgroup. CA-VA and VB are mitochondrial enzymes, and CA-VI is part of the secreted subgroup [1,10,11]. CA-VIII, X and XI are regarded as the carbonic anhydrase related proteins (CARP) and do not possess CO_2 hydration activity. From the α-CAs, CA-II has the highest rate of reaction and assists in the maintenance of homoeostasis within the body [12]. Reaction substrates and products (CO_2, HCO_3^- and H^+) are essential for the regulation of biological processes such as, but not limited to, respiration, cerebrospinal fluid (CSF) formation and bone resorption within cells [13,14]. CA-II consists of 260 residues and a Zinc ion (Zn^{2+}) within the active site that is in tetrahedral coordination geometry with three histidine residues (His94, His96, His119) and a water molecule [1,6,10,11,15,16]. These coordinating residues are also known as coordination ligands. The active site also comprises of hydrogen bonded waters that form part of the proton transfer network with His64 acting as the main proton shuttle, transporting protons to and away from the active site [17]. His64 facilitates its proton shuttling role by alternating between two conformations ("in" and "out") during catalysis [16–19] or through tautomerisation of the histidine ring [20].

Three CO_2 binding pockets have been identified in CA-II and are located approximately 3–4, 5–7 and 10–12 Å away from the Zn^{2+} [21–24]. For the purposes of this study, these pockets have been designated as the primary, secondary and tertiary binding pockets, respectively. The primary binding pocket is made up of hydrophobic residues Val121, Val142, Leu197 and Trp208, while the secondary CO_2 binding site is made up of aromatic residues Phe66, Phe95, Trp97 and Phe225. The tertiary binding pocket comprises the residues Trp7, His64, Thr199, Pro200 and Asn243 and is located along a tunnel leading to the primary pocket. Of the three binding pockets, the secondary pocket is the only non-catalytic binding pocket, and its role in CA-II is yet to be fully investigated [1,21–24].

To assist with protein stability, CA-II contains two groups of aromatic residues known as the primary and secondary aromatic clusters. The primary aromatic cluster consists of the residues Trp5, Tyr7, Trp16 and Phe20, and joins the N terminal to the rest of the protein [1]. The secondary aromatic cluster is larger and is comprised of the residues Phe66, Phe70, Phe93, Phe95, Trp97, Phe175, Phe178 and Phe225 [1,25,26].

In the absence of CA-II, CO_2 is hydrated at a rate constant between 0.030 and 0.15 s^{-1}, compared to 1×10^6 s^{-1} in the enzyme mediated reaction [15,27–30]. The large difference in reaction rates coupled with the importance of CA-II to other biological processes indicates that any impairment to the function of CA-II could have detrimental effects to cells and the body. In humans, poor CA-II function causes CA deficiencies resulting in the phenotypes' osteopetrosis with renal tubular acidosis and cerebral calcification [31]. Advances in genomic research identified non-synonymous single nucleotide variations (nsSNVs) occurring in CA-II to be the main cause of these diseases [32,33]. Numerous studies have been done associating CA-II SNVs with CA deficiencies. For instance, research in 2004 by Shah et al. [34] identified 11 novel CA-II mutations, such as G144R, in individuals suffering from CA deficiencies leading to osteopetrosis with renal tubular acidosis and cerebral calcification. The changes to the amino acid sequence of CA might influence residue interactions and communication within the protein resulting in poor enzyme function and stability causing protein deficiencies. As variations might lead to dysfunctional proteins and cause the indicated diseases, it is important to understand the mechanism of these SNVs to identify "activator" compounds reversing the effect of variations and rescuing the protein function.

To date, most of the research into CA has focused on inhibition for the management of conditions such as, but not limited to, glaucoma and altitude sickness, which are related to the overexpression of CA-II [35–38]. CA inhibitors have also found use as diuretics [38,39]. However, prolonged use of CA inhibitors is not without consequence; for example, prolonged use of acetazolamide is associated with a reduction in osteoclast function and bone resorption [40] that could potentially lead to osteopetrosis. Factoring in the potential presence of SNVs and their effect on CA-II function, specific inhibitors would have varying efficacies across different individuals depending on the SNV that is present within the CA-II proteins, highlighting a research gap for precision medicine related studies for CA inhibitors.

The aim of the current study is to characterize the structural and functional effects of six validated nsSNVs (K18E, K18Q, H107Y, P236H, P236R and N252D) on CA-II protein structure as proposed previously [41,42], by combining homology modelling, molecular dynamics (MD) simulation, principal component analysis (PCA) and dynamic residue network (DRN) analysis [41,42] to identify underlying mechanisms responsible for CA-II deficiencies. Previous studies have focused only on MD simulations to analyze the effect of SNVs in CA-II. Within this research, not only have we used MD to analyze the variant effects but we have also employed DRN to analyze SNV effects on residue and protein network communication. DRN analysis demonstrated differences to the variant mechanisms of action, and revealed all six SNVs to be associated with allosteric effects in variant proteins. DRN, further, showed that Glu117 is the most important residue within the protein. We were also able to predict that H107Y is the most deleterious variant due to its direct reduction to Glu117 communication in all three protein states (apo, BCT and CO_2 bound). In general, this research enhances the importance of the previously proposed approach in variant analysis [41,42], in that MD alone is insufficient to understand the effects of missense mutations to protein structure, and shows the importance of coupling MD with DRN analysis. This method is particularly important for precision medicine related studies, which aim to analyze the cause of disease at the molecular level and then to utilize targeted treatment for patients.

2. Results and Discussion

In this study, six validated and pathogenic SNVs of CA-II were investigated using a combination of motif analysis, MD, PCA and DRN analysis to determine potential variant effects on the protein structure and function at the molecular level.

2.1. Six Disease Related SNVs and Their Spatial Positions Are Identified

Data retrieval from the Ensembl [43] and HUMA [44] databases identified six validated SNVs with pathogenic effects (Table 1). Even though there were several more SNVs, we limited our CA-II dataset to those validated within dbSNP via frequency, or via cluster, and containing a phenotype annotation. Table 1 also presents the predicted variant effects on protein structure using the Variant Analysis Portal (VAPOR) [44]. VAPOR is a SNP/SNV analysis tool that combines PhD-SNP [45], PROVEAN [46], FATHMM [47] and PolyPhen-2 [48] to predict SNV presence as either damaging or tolerated. VAPOR also incorporates I-Mutant 2.0 [49] and MUpro [50] to predict SNV effects on stability [44]. MUpro analysis showed that all SNVs are expected to cause reductions to protein stability. Data in Table 1 also presents the global minimum allele frequency (MAF) of the SNVs obtained from gnomAD [51]. Data indicates that N252D occurs at the highest frequency.

Table 1. α-CA-II single SNVs rs IDs, associated residue-variant substitutions, model z-DOPE score, MAF and predicted variant consequences. Global MAF obtained from gnomAD.

rs ID	Variation	z-DOPE Score	MAF	VAPOR Analysis			
				I-Mutant		MUpro	
				ΔΔG	Stability	ΔΔG	Stability
rs118203931	K18E	−2.204	<0.01	−1.30	Decrease	−0.499	Decrease
	K18Q	−2.203	<0.01	−1.24	Decrease	−0.655	Decrease
rs118203933	H107Y	−2.204	<0.01	0.26	Increase	−0.867	Decrease
rs118203932	P236H	−2.159	<0.01	−1.46	Decrease	−0.586	Decrease
	P236R	−2.186	<0.01	−0.60	Decrease	−0.311	Decrease
rs2228063	N252D	−2.197	0.007	0.06	Increase	−0.474	Decrease

Homology modeling of variant proteins was performed by Schrödinger Prime [52,53]. Z-DOPE (normalized discrete protein energy) score [54] and Ramachandran plot were used to assess protein model quality and select the best structure for each case. The z-DOPE statistical method shows that all structures have scores less than −2.00 (Table 1). This indicates that models resemble native protein structures. Within the scope of the studied literature, no evidence as to the possibility of variant linkage disequilibrium in CA-II was observed. As a result, all modeled proteins contained one SNV only, and no structures with combinations of SNVs were generated and analyzed.

Next, 3-dimensional (3D) spatial analysis of variant location in relation to the active site and CO_2 binding pockets was done. Figure 1 presents the SNV positions within the CA-II protein. K18E and K18Q are located within a helix whereas the other variants are present in loop regions. H107Y is located within the protein interior and closest to the active site and primary CO_2 binding pocket, whereas the other variants are positioned towards the exterior of CA-II. The primary and secondary binding sites of CO_2 are also presented in Figure 1A,B. As the secondary CO_2 binding pocket is acatalytic, during catalysis (see Equation (1)) the substrates BCT and CO_2 both bind to the primary pocket.

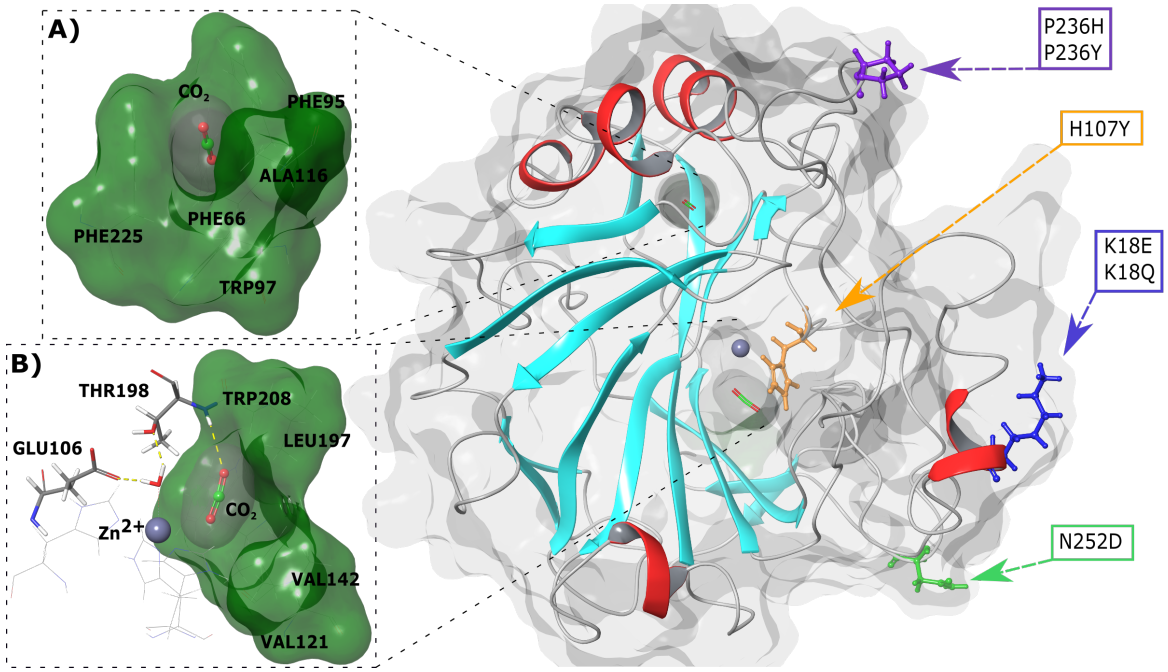

Figure 1. 3D structure of human CA-II WT protein and SNV locations. (**A**) Secondary CO_2 binding pocket. (**B**) Primary CO_2 binding pocket. Cyan and red secondary structure represents beta sheets and helices, respectively. The grey sphere represents the Zn^{2+}. Representations were generated using Schrödinger Maestro and Inkscape.

2.2. Identified SNVs May Have an Indirect Effect on Protein Structure and Function

To date, numerous researches have been conducted on CA-II structure. These studies allowed the identification of key residues essential to CA-II structure and function as presented in Table S1. Interestingly, it is evident that none of the selected variants are located at important residues, suggesting that as opposed to having a direct effect on the protein structure, variant action may occur via indirect or secondary mechanisms. This hypothesis has been uncovered in the rest of the article.

2.3. Identified SNVs Are Located within or around the Highly Conserved Motifs

Important functional residue groups in protein families are generally highly conserved. Hence, as a next step, motif analysis was performed across all selected α-CA proteins (Table S2) to identify conserved motifs that might contain new functional residue groups in addition to those in Table S1. The default length of the motifs that can be searched in the Multiple Expectation Maximization for Motif Elicitation (MEME) program [55] is up to 50 residues. Short linear motifs to that function in protein–protein interactions are known to be between 3 and 11 residues [56,57]. To our knowledge, there is no defined length to the functional motifs within a protein; hence, we set MEME parameters to lengths of 3–20 amino acids to include possible motifs marginally longer than average short linear motifs as applied in our previous study [56]. MEME calculations initially gave 100 motifs across all α-CA sequences. Motif pairwise correlation data analysis reduced the final dataset to 77 as explained in the Materials and Methods (Section 3.2).

A heat map of the remaining 77 motif dataset is presented in Figure S1, with motif conservation represented as the number of sites per total number of protein sequences. Motifs are numbered according to the MEME output. From the heat map data, it is evident that only motifs 1–11 are conserved in CA-II and have significant E-values (Table 2). As a result, motifs 1–11 were selected for further analysis. Data in Figure 2A presents a combination of the motif logo and conservation results. The motif logo is a representation of the amino acids present in each motif (M1–M11), and shows the conservation of each protein residue in that motif sequence. The heat map highlights motif conservation within the selected proteins (Table S2). A value of 1 represents 100% motif conservation in all sequences, whereas a value of 0 suggests that the motif does not exist in any of the proteins within the selected set. Since the motif analysis included the non-catalytic CA isoforms CA-VIII, CA-X and CA-XI, it should be noted that the most significantly discovered motifs are most probably important for the maintenance of structure and stability.

Heat map results in Figure 2A demonstrate that motifs 1–6 and 8 are conserved across all α-CA proteins, indicating that residues within these motifs have most likely important function and/or structural (i.e., stability) roles within the α-CAs. Motif 7 is conserved in all except CA-XV, and motif 9 is conserved in all except CA-XI. Motif 10 is conserved in all cytosolic CAs (I, II, III, VII and XII) and is also conserved in two membrane CA proteins (XII and XIV). CA-XII and XIV are single pass type I membrane proteins with the N-terminus existing extracellular and C-terminus towards the cytosol [11]. Existence of these proteins within the cytosol suggests that this motif could play a role in protein function, solubility and/or stability within the cell. Motif 11 is conserved in all α-CAs located within the cytoplasm, and includes acatalytic isoform CA-VIII. Conservation within CA-VIII suggests that this motif most likely is important for protein stability within the intracellular environment, as opposed to catalytic function. These motifs are mapped to the 3D structure of CA-II (Figure 2B).

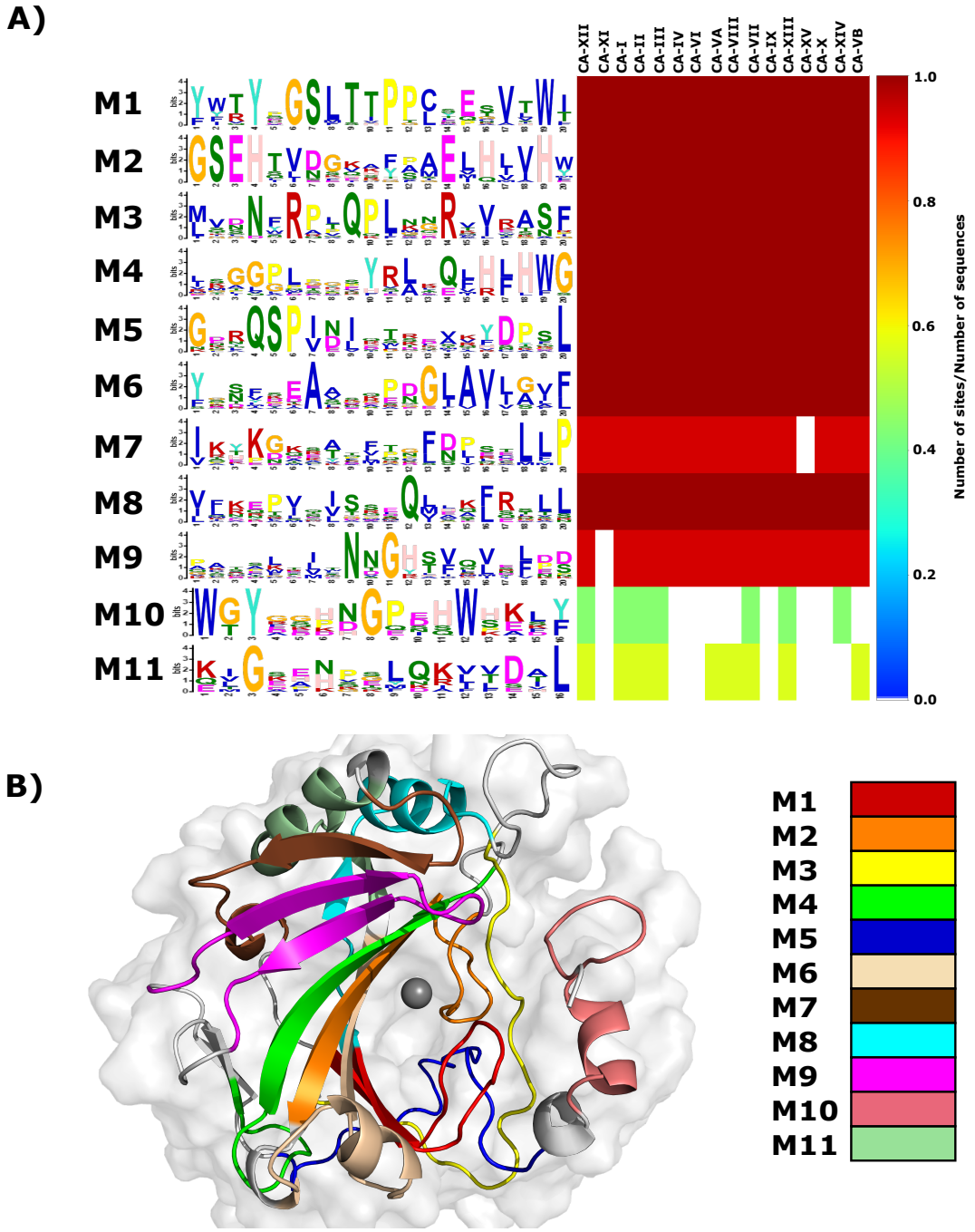

Figure 2. Motifs present in the α-CA family. (**A**) Web logo presenting residue conservation within a motif in each sequence and a heat map showing motif conservation as the number of motif sites per total number of protein sequences. A value of zero is for motifs that are not conserved in any sequence, whereas a value of 1 shows that motif exists in all selected sequences. The prefix 'M' represents the word "motif". (**B**) Motif location in 3D space within CA-II. The grey sphere represents the Zn^{2+}. Representations were generated using Schrödinger Maestro and Inkscape.

As a next step, we checked to see if motifs contain Table S1 residues as well as variant data. Where applicable, the mapping of residues in Table S1 onto motifs allowed for the assignment of motif function. Since proteins are networks of residues communicating with one another [41,42], the function of specific protein residues is affected by other surrounding residues. Assignment of function via residue mapping was conducted as follows, using motif 1 as an example. Data in Table 2 shows motif

residue range, sequence and associated E-value, and from the table, it is noted that motif 1 contains the residues 190 to 209. Within this group of residues exists, Thr198 and Thr199 that are essential in the catalytic pathway via the catalytic orientation of the Zn^{2+} water ligand molecule and active site water network coordination (see Table S1). In addition to Thr198 and Thr199, motif 1 also contains Leu197, Pro200 and Trp208 that are essential for primary and tertiary CO_2 binding pocket formation. Therefore, a combination of the functions Leu197, Thr198, Thr199, Pro200 and Trp208 in CA-II suggests that motif 1 is involved with catalysis and CO_2 binding pocket formation. A similar process was used to assign functions of the remaining 10 CA-II motifs. Assigned motif functions are presented in Table 2.

Results in Table 2 further indicate the respective SNV positions in each motif. K18E and K18Q are present on motif 10 while H107Y is located on motif 2. The variants P236H and P236R are not located on any conserved motifs, whereas N252D is present on motif 3. Analysis of SNV locations with respect to their motifs suggests that H107Y could have the greatest effect on protein structure and function. P236H and P236R are located between the two highly conserved motif 3 and motif 8. Variant presence could indirectly influence these motifs. In addition, as none of the variants are located on motifs responsible for catalytic mechanism, SNVs could still influence catalysis due to stability decreases in or near active site residues. Potential effects of variants on enzyme mechanism were also investigated later within this article.

Table 2. Starting and ending positions of conserved motifs in CA-II, associated E-values and motif contribution to function. Residues from Table S1 are underlined and highlighted in bold. SNV positions (K18E, K18Q, H107Y and N252D) are underlined, italicized and presented in bold red.

Motif	Residue Range	Residues	E-Value	Residue Count	Contribution to Functions
1	190–209	YWTYPGS**LTTP**PLLECVT**W**I	2.2×10^{-164}	20	Catalytic mechanism and Primary CO_2 binding pocket formation
2	104–123	GS**E*H***TVDKKKYAA**E**L**H**L**V**HW	4.4×10^{-140}	20	Active site and/or Zn^{2+} stability
3	240–259	MVD**N**W**R**PAQPLK***N***RQIKASF	1.2×10^{-108}	20	Tertiary CO_2 binding pocket formation and enzyme stability
4	79–98	LKGGPLDGTYRLIQ**FHFHW**G	3.0×10^{-89}	20	Enzyme stability and/or Zn^{2+} coordination
5	25–44	GERQ**S**PVDIDTHTAKYDPSL	6.5×10^{-86}	20	Enzyme stability
6	127–146	YGDFGKAVQQPDGLA**V**LGIF	3.8×10^{-73}	20	Primary CO_2 binding pocket formation
7	166–185	IKTKGKSAD**F**TN**F**DPRGLLP	6.0×10^{-54}	20	Participated in secondary aromatic cluster
8	210–229	VLKEPISVSSEQVLK**F**RKLN	6.4×10^{-40}	20	Enzyme stability and secondary CO_2 binding pocket formation
9	53–72	QATSLRILN**N**G**H**A**FN**VE**F**DD	1.9×10^{-41}	20	Enzyme stability and/or catalytic mechanism
10	5–20	**W**G**Y**GKHNGPEH**W**H***K***D**F**	7.7×10^{-25}	16	Enzyme stability
11	148–163	KVGSAKPGLQKVVDVL	2.3×10^{-6}	16	Enzyme stability

2.4. Zn^{2+} Parametrization Helps the Ion to Maintain Its Position over MD Simulations

As traditional MD force fields are designed for protein amino acids, these built-in force fields are incapable of handling metals and potential interactions with amino acids. A consequence of this would be metal escape from the active site of the protein during the MD simulations and diminishing their accuracy. Within this study Zn^{2+} force field parametrization was essential to govern how Zn^{2+} interacts with its coordinating atoms and prevent it from escaping the active site. Parametrization was performed using AmberTools17 [58]. Gaussian 09 [59] quantum mechanical (QM) optimizations showed no evidence of bond breakage between Zn^{2+} and coordinating atoms, giving a strong indication that the derived force field parameters would hold the Zn^{2+} in place during MD. In addition, bond length

measurements in between the Zn^{2+} and coordinating atoms were compared to those previously derived by Harding [60], and by Bernadat et al. [61]. Results indicated that bond lengths calculated by the Metal Centre Parameter Builder (MCPB) [62] fell within previously reported ranges of His $<$ 2.03 Å and H_2O $<$ 2.18 Å [60,63]. Bond angles calculated within this study were also similar to those derived in previous research [61]. Parametrization of the active site further determined the Zn^{2+}, His94 (ND), His96 (ND) and His119 (NE) to have atom charges of 0.592, −0.089, −0.033 and −0.160, respectively. Results agree with those previously calculated by Bernadat et al. [61] that show Zn^{2+} to have a charge less than one, and the coordinating His ND and NE atoms having lower negative charge compared to their standard charges [61]. The calculated final Zn^{2+} parameters are presented in Table S3.

2.5. Global Level Analyses Hint at the Functional/Structural Effect of Certain Variants

21 protein structures (WT and six variant proteins for each state; apo, BCT and CO_2 bound) were taken into 200 ns MD simulations. The trajectories were analyzed using root mean square deviation (RMSD), PCA and radius of gyration (Rg) metrics to observe potential variant structural effects, in the absence and presence of BCT and CO_2 molecules at the global level.

2.5.1. Proteins with Variations Occupy Different Conformational Spaces to the WT

RMSD Analysis

As VAPOR results (Table 1) predicted a decrease in stability within the variants, RMSD differences between WT and variant proteins were expected. RMSDs of WT and variants were calculated for each frame over the MD simulation for each protein system and compared (Figure S2). Overall RMSD values did not show any drastic changes in the presence of variations, probably an indication of more subtle effect. To be able to understand the variation effect and to observe potential discrete conformational changes in proteins, RMSD distribution histograms of the WT and variant proteins were generated, as previously applied in Penkler and Tastan Bishop 2019 [64] (Figure 3). WT and variant histogram width is the number of conformations sampled by proteins over the MD simulation. Frequency represents number of times a specific conformation was sampled during the MD simulation, and the histogram peaks are indicative of the most commonly occupied conformation. Variant apo, BCT and CO_2 proteins were plotted against the WT apo, BCT and CO_2 proteins, respectively. The variant apo structures were compared to the WT apo protein, whereas the variant BCT proteins were compared to the BCT bound WT protein, and variant CO_2 proteins were compared to the WT CO_2 bound protein.

Figure 3 presents the average RMSD differences between the WT and variant proteins. Though the differences are small, previous research in 2004 by Almstedt [32] proved that residue movements and rotations as small as 0.3–0.4 Å are enough to destabilize CA-II. To check the significance of our results, statistical calculations were performed (Table S4). Using a 5% level of significance, statistical data indicated that the changes and differences to structural conformations are statistically significant between the respective WT and variant proteins.

Apo protein structure analysis (Figure 3) shows that the H107Y has the closest average structure to that of the WT, whereas K18E has the largest structural differences from the WT. In addition, with exception to K18E and P236R, it is observed that the average RMSD coincides with the histogram peaks. This finding can be explained by the plot shape of K18E and P236R. The wide histogram base suggests that these variants sampled the most conformations during MD. Statistical results however present an interesting finding with regards to the WT CO_2 bound protein and H107Y apo variant. These two proteins are not statistically different (p-value: 0.5124) indicating some conformational similarity. In the presence of substrate BCT, the P236H has the closest average RMSD to that of the WT, while K18Q has the largest average RMSD difference to that of the WT.

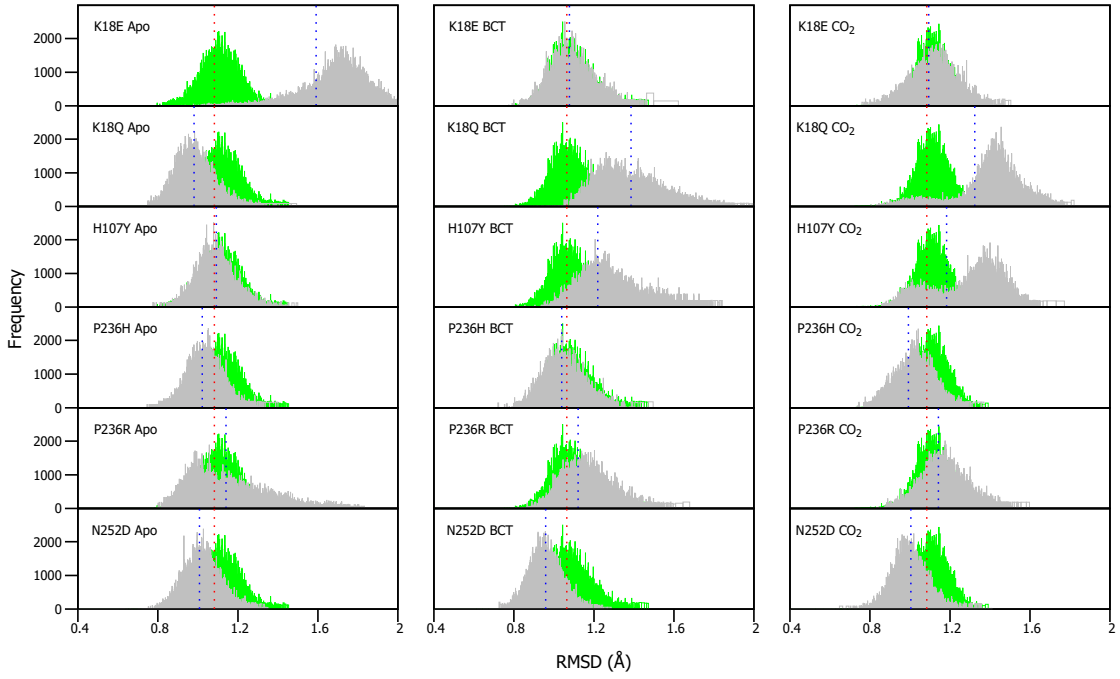

Figure 3. α-carbon RMSD distribution of the WT and variant proteins over the 200 ns MD simulation. The green and grey histograms indicate the RMSD distribution of the WT and variant protein, respectively. The red and blue dashed lines represent the mean RMSD of the WT and variant proteins, respectively. Variant apo, BCT and CO_2 proteins are each plotted against the WT apo, BCT and CO_2 bound proteins, respectively. The x-axis represents RMSD of sampled conformations during MD simulation, whereas the *y*-axis (Frequency) represents number of times a specific conformation was sampled.

When the substrate CO_2 is bound, P236R has the closest average RMSD to that of the WT, whereas, K18Q has the largest difference in average RMSD from the WT protein. As was observed when BCT is bound, variants K18Q and H107Y are associated with the greatest conformational changes. K18Q and H107Y histograms also demonstrate two RMSD peaks each, indicating that during MD, K18Q and H107Y occupied two conformations (one is more dominant than the other). Using the WT as reference, the conformations which are represented with the peaks closest to the WT, at RMSD of 1.1 Å would most likely be the ones capable of catalysis by comparison of structures sampled during MD simulation for WT and variant proteins.

PCA Analysis

RMSD data represents a 2D interpretation of the conformations sampled by the proteins during MD. PCA analysis was performed to expand on the RMSD data and to observe the 3D conformational sampling and internal dynamics of the WT and variant proteins. Figure S3 shows the 3D PCA plots of the WT and variant proteins. The eigenvalue fraction of each PC is presented in Table S5 and indicates that most of the conformational sampling space is covered by PC1 and PC2. PC1 represents the largest possible variance while PC2 is representative of the second largest variance within the structures. Conformations associated with low free energy are expected to be more stable.

In the apo form, the WT samples two distinct protein conformations along PC1 and one conformation along PC2. Though two conformations are sampled along PC1, RMSD results only show evidence of one major conformation in Figure 3. This can be explained by size of the low energy well in Figure S3. The larger well suggests that the majority of the structures are at that conformation, and therefore these conformations represent the peak observed in the RMSD results. When BCT is bound, the protein samples a larger conformation space than that of the apo and CO_2 bound structures.

The BCT bound conformations also have a higher free energy. When CO_2 is bound, the WT forms two structures along PC1; however, only one structural cluster has low energy, suggesting that this is the conformation observed within RMSD results.

K18E and K18Q data indicate that although the variants occur at the same position, the mechanism of action could be different. When BCT is bound, K18E occupies a larger conformational sampling space than that of K18Q. Free energy data suggests that the BCT bound structures in K18E have higher energy than those of K18Q. When CO_2 is bound to K18Q, two conformational clusters are formed along PC1, and these conformations are also observed within the RMSD results. When BCT is bound to H107Y, compared to the WT, the variant has higher free energy and samples more conformations. This indicates evidence of variant associated greater conformational changes and supports findings in RMSD. When substrate CO_2 is bound to H107Y, the results indicate two distinct low energy conformational clusters that are also observed within the RMSD results. P236H and N252D conformations exhibit the highest free energy when BCT is bound, compared to the WT and other variant proteins. The apo protein of P236R demonstrates evidence of potential instability. The variant forms two conformational clusters along both PC1 and PC2. Occupation of multiple conformations is indicative of variant associated conformational changes, and this finding agrees with that observed within the RMSD results.

Rg Analysis

The RMSD and PCA analyses of the WT and variant proteins allowed for the determination of SNV effects on protein; conformation, stability and possible functional consequences in global level. To expand on these results, the Rg of all protein systems were calculated. Calculation of Rg allows for identification of potential relationships between protein compactness, conformation and stability within the variant proteins [65]. Considering the effects of protein compactness on enzymatic function, decreases would result in key enzyme residues moving away from each other, whereas increases in protein compactness would result in residues moving closer to each other. Previous CA-II studies have suggested that residue-residue distance has an effect on enzyme kinetics, and protein residues can be neither too close nor too far from each other for optimal enzyme activity [16,27,66–69].

WT and variant Rg were calculated for each frame over the MD simulation for each protein system and compared (Figure S4). As observed in RMSD, variant presence shows minimal effect on protein compactness. Figure 4 presents a histogram of the Rg distribution between the WT and variant proteins. As with the RMSD distribution, variant apo structures are compared to the WT apo protein, whereas the variant BCT proteins are compared to the BCT bound WT protein, and variant CO_2 proteins are compared to the WT CO_2 bound protein.

The Rg distribution in Figure 4 indicates that H107Y is associated with a reduction to compactness for all three protein states. When the H107Y has substrates BCT and CO_2 bound, histograms have a wider base compared to the WT. This observation can be explained by the previous RMSD findings whereby the proteins exhibited greater conformational sampling. When BCT is bound to K18Q, an interesting result is observed. The wide base of the Rg distribution histogram demonstrates evidence of greater Rg sampling compared to the WT. In addition, the histogram exhibits two peaks, indicating major changes to compactness during MD. Comparison of RMSD and Rg distribution peaks for K18Q suggests that the conformations sampled during MD (RMSD) could cause a shift to the protein centre of mass and cause the two observed Rg peaks. When CO_2 is bound to K18Q, an increase in protein compactness is observed.

Within this section we analyzed potential global functional and/or structural effects associated with SNV presence and noted that the variants have some subtle effects where the significance is supported by statistical calculations on CA-II compactness and conformation (see Figure S4). At the 5% level of significance, statistical data shows that WT and variant protein compactness is significantly different for most of the protein systems. The WT apo and the CO_2 bound K18E variant do however share similar compactness (p-value: 0.06527). The CO_2 containing N252D and BCT bound P236R variants indicate no significant differences to protein compactness when compared to the CO_2 bound

WT protein (*p*-values: 0.1479 and 0.3288, respectively). In the following section we take this research forward to identify residues that are responsible for the global effects observed.

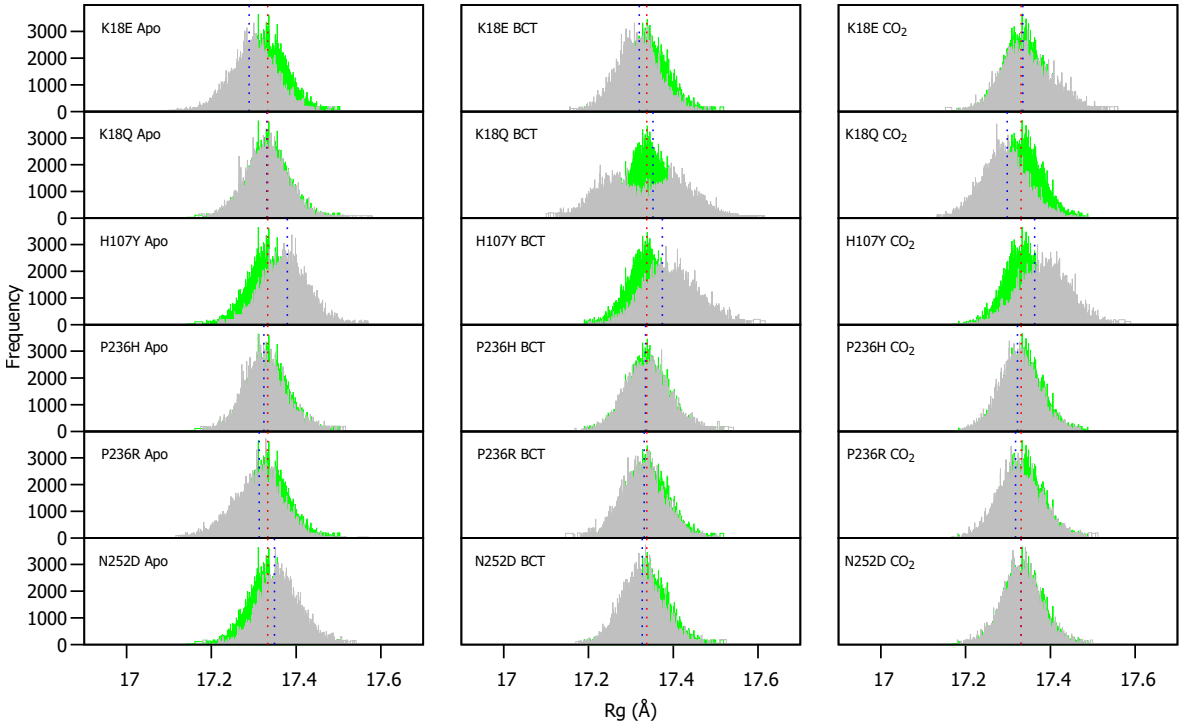

Figure 4. Rg distribution of the WT and variant proteins over the 200 ns MD simulation. The green and grey histograms indicate the Rg distribution of the WT and variant protein, respectively. The red and blue dashed lines represent the mean Rg of the WT and variant proteins, respectively. Variant apo, BCT and CO_2 proteins are each plotted against the WT apo, BCT and CO_2 bound proteins, respectively. The x-axis represents Rg of sampled conformations during MD simulation, whereas the *y*-axis (Frequency) represents number of times a specific conformation was sampled.

2.6. Residue Level Analysis Reveals Further Differences between WT and Variant Protein Systems

In the first part of residue level analysis, we looked at the root mean square fluctuation (RMSF) values. WT and variant protein system RMSF values are presented in Figure S5. To improve data resolution, ΔRMSF between the WT and variant proteins was calculated for apo, BCT and CO_2 protein states by subtracting WT RMSF from the variant RMSF data. Thus, a positive ΔRMSF is indicative of higher flexibility in variant residues, and a negative ΔRMSF shows a reduction in RMSF within the variant residues.

Analysis of Figure 5 ΔRMSF data demonstrates that for the apo proteins, SNV presence only influences the first 20 N-terminus residues of K18E and P236R. This group of residues are members of motif 10 (Table 2), and contain residues (Trp5, Tyr7, Trp16 and Phe20) within the initial aromatic cluster that are involved in maintaining CA-II stability. Increases to flexibility of residues in this region might have detrimental effects on protein stability. The previously observed RMSD results for K18E and P236R could be due to the increases in the fluctuations in motif 10 residues. Comparing the RMSF values between the two variant proteins, data indicate that the first 20 residues of K18E are more flexible than those of P236R.

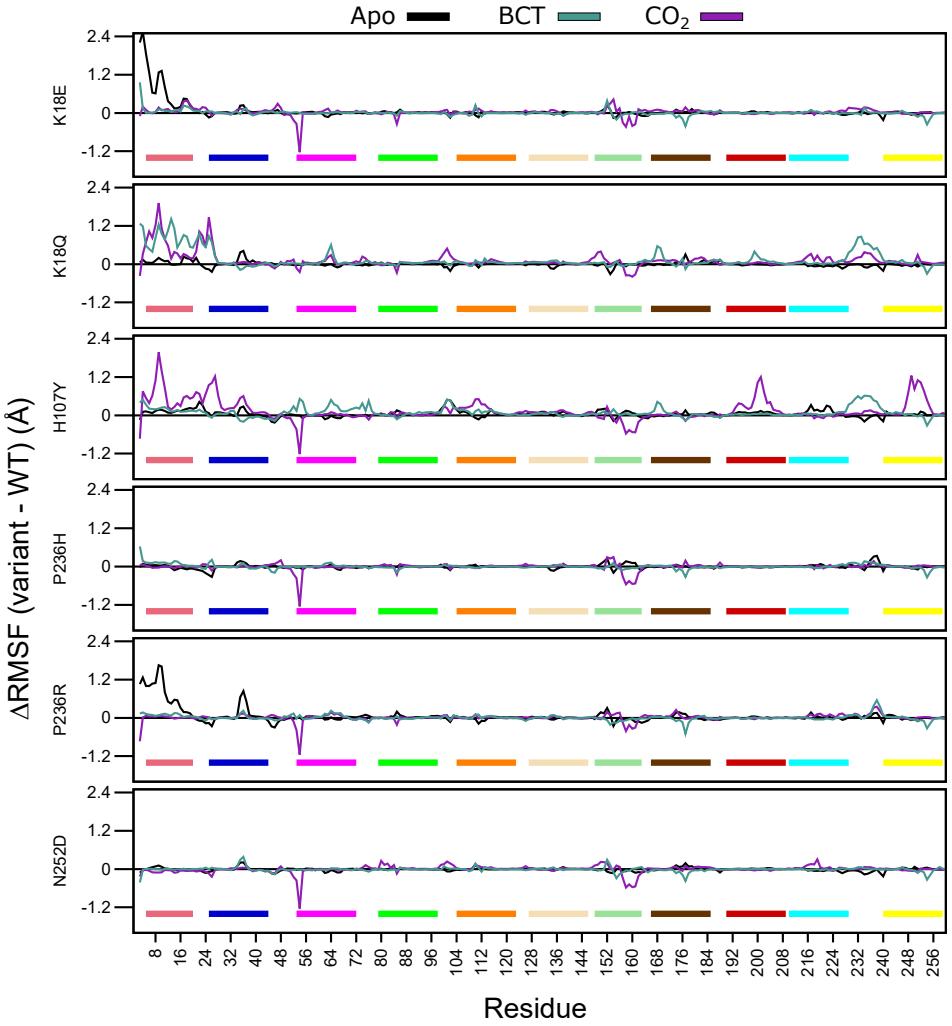

Figure 5. Δ α-carbon RMSF comparison of the WT and variant proteins (variant minus WT). Respective motifs are indicated as bars at the bottom of each plot. Motif colors correspond to those in Figure 2B.

When BCT is bound to the variant proteins, SNV effects are observed to the greatest extent in K18Q and H107Y. Both proteins are associated with increases to RMSF between residues 230 and 240. These groups of residues are located between motif 3 and motif 8; hence, increases to the flexibility of these residues could have an impact on stability. K18Q shows increases to the flexibility of motif 10 residues containing Trp5, Tyr7, Trp16 and Phe20 of the initial aromatic cluster. This observed RMSF increase could explain the potential instability and greater conformational sampling noted within the RMSD results. The substitution of Lys with Gln at position 18 is also associated with increases to RMSF. With respect to H107Y, RMSF increases are observed in some residues located in motif 2 (104–124), motif 7 (166–186) and motif 9 (53–73). RMSF increases to these motifs highlight potential implications for H107Y function and stability. Motif 2, motif 7 and motif 9 are all involved in enzyme stability (Table 2). The increase in the flexibility of these residue groups could explain the potential instability and greater conformational sampling observed within the RMSD when BCT is bound. In addition, flexibility increases in these motifs also suggest potential implications on enzyme function, as both motifs include active site residues and could have implications on enzyme function. This could also explain the poor activity associated with this variant.

When CO_2 is bound, all variants with the exception to K18Q exhibit large RMSF decreases to residues 53–54 and 157–162. Residues 53–54 are located on motif 9. However, as the decrease to RMSF is not spanning multiple residues, variant presence could have minimal effect on the function of motif 9. Comparison of RMSD and RMSF results of K18Q and H107Y with the other variants when CO_2 is bound highlights the potential role of motif 11. Both K18Q and H107Y are the only variants exhibiting RMSF increases in motif 10 residues Trp5 and Tyr7 of the initial aromatic cluster; however, all variants show RMSF decreases between residues 157 and 162. Noting that only K18Q and H107Y exhibit potential instability and greater conformational sampling in RMSD, this could be associated only with motif 10. This finding suggests that motif 11 may not play a role in CA-II stability.

H107Y shows most changes to RMSF when CO_2 is bound compared to the other variants. In addition to the flexibility increases in motif 10 residues, RMSF increases are also noted in residues 192–208 (motif 1) and 246–254 (motif 3). Overall flexibility increases demonstrate that SNV effects are centered around the active site of the protein, and RMSF increases to Thr199 and Pro200 show potential effect on the binding of CO_2 to the tertiary pocket. This result suggests that, compared to the other variants, H107Y could have the greatest effect to enzyme function.

Throughout all variants, the presence of substrates BCT and CO_2 has the greatest effect on RMSF. Overall data further supports the initial finding that the SNV presence has an allosteric effect on CA-II structure and function. This is evidenced by RMSF changes away from SNV positions. In the next sections, we explore DRN to analyze the SNVs effect at residue level and identify the potential allosteric effects of variants.

2.7. Short Range Effects of Each Variation Are Deciphered Using Weighted Contact Map Analysis

Since RMSD and Rg analysis showed subtle changes associated with SNV presence, network analysis was performed to obtain a more thorough and robust understanding of variant effects on the CA-II protein. Through previous research, network analyses has proved useful in the identification of key regions and residues essential for communication, function and stability in proteins [70–72]. In 2017, Brown et al. [41] utilized DRN to successfully investigate SNV effects on protein dynamics. Here, contact map analysis was performed to identify SNV associated short-range changes to the protein network and potential variant mechanisms. Contact maps show all weighted/frequency of interactions occurring between a network of residues within 6.7 Å over the MD simulation. These weighted contacts include all interactions such as, but not limited to, van der Waals, hydrogen bonds and/or electrostatic interactions [42].

Results in Figure 6 show a heat map of the weighted contacts (frequency of interaction) occurring between all variant protein SNVs and their corresponding neighboring residues. Weighted contacts with the value 0 indicate that there is no contact occurring between the two residues, whereas a value of 1 shows that respective amino acids are constantly interacting over the MD simulation.

Analysis of K18E contact maps in Figure 6 indicates no changes to interactions occurring between either Lys18 or Glu18 and the other protein residues; however K18Q demonstrates that when BCT or CO_2 are bound, Gln18 forms new interactions with His3. Since His3 is not conserved nor located on any CA-II motif group (Table 2) formation of these interactions may be of minimal importance in maintaining enzyme structure and function. On the other hand, comparison of His15 interaction analysis with Glu18 (K18E) and Gln18 (K18Q), with the corresponding RMSD results, shows an interesting finding. Decreases in interactions between His15 and Glu18/Gln18 are associated with potential instability and greater conformational sampling (Figure 3). Glu18 in the K18E apo protein shows a decrease to interactions with His15 (Figure 6), and the resulting RMSD histogram for K18E apo in Figure 3 exhibits potential instability and greater conformational sampling. Gln18 in K18Q indicates decreases to interactions with His15 only when substrates are bound, and the resulting RMSD distribution of K18Q indicates potential instability and greater conformational sampling when BCT and CO_2 are bound. Noting that His15 is located on motif 10 and factoring in its functional importance to CA-II, data suggests that the potential instability and greater conformational sampling observed in K18E and K18Q could be as a result of interaction loses with His15. The potential instability and greater

conformational sampling associated with His15, may however have minimal effect on the catalytic function of K18E and K18Q. Previous enzyme kinetic research performed in 1988 by Eriksson [25] showed that even after removal of the first 23 N-terminal residues, CA-II maintained function and hydrated CO_2 at rate constant of 1.5×10^5 s^{-1} [25].

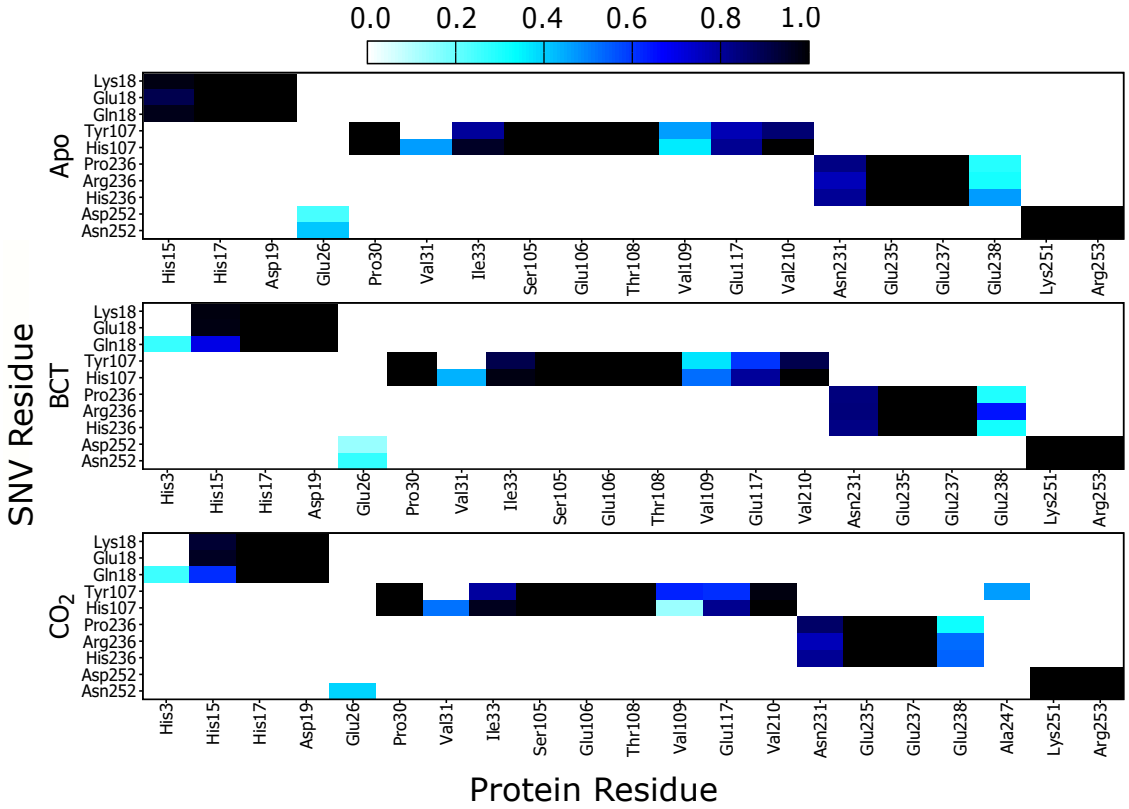

Figure 6. Heat map presenting the contact map weighted interactions between the respective SNV residues in; K18E, K18Q, H107Y, P236H, P236R and N252D with neighboring CA-II residues.

Analyzing the contacts of H107Y in Figure 6, results indicate that Tyr107 loses all interactions with Val31. Tyr107 also shows a decrease in interactions with Glu117, and this decline in weighted contacts is greater when the substrates BCT and CO_2 are bound to the protein. Active site Zn^{2+} is coordinated by the direct ligand His119 and the indirect (secondary) ligand Glu117. Residue Glu117 stabilizes the Zn^{2+} through interactions with the direct ligand His119 [1]. It is observed that Tyr107 reduction in weighted contacts with Glu117 is not as great as that observed with Val31. Due to the importance of Glu117 in Zn^{2+} affinity, additional interactions could have formed between Tyr107 and Glu117 as a compensatory measure to maintain enzyme structure, function and catalytic efficiency. Potential compensatory mechanisms occurring in H107Y were investigated later in this article using DRN analysis. When CO_2 is bound to H107Y, Tyr107 creates new interactions with residue Ala247, as Ala247 is located on motif 3 (Table 2) and formation of new interactions could assist with stability.

As contact maps do not discriminate between bond/interaction types, hydrogen bond analysis was performed to analyze changes occurring to interactions between residue 107 and neighboring residues. Hydrogen bonds are one of the stronger interactions occurring in proteins, and loss would have significant effects to enzyme stability and structure. Results in Figure 7 compare hydrogen bond fractions (proportion of total MD frames bonds were present) between residue 107 and neighboring atoms in the WT and variant protein. Results show the loss of hydrogen bonds between Tyr107 and the neighboring residues; Val31, Glu117 and His119. One hydrogen bond is lost between Tyr107 and Val31, whereas two hydrogen bonds each are lost between Tyr107 and the residues Glu117 and His119. From data, it is evident that H107Y loses half of the hydrogen bonds present in the WT. Residue 107 in the WT makes 10 hydrogen bonds within neighboring atoms over MD, whereas in the variant, only

5 hydrogen bonds are made with neighboring atoms. Previous studies into H107Y have also proved that at least two hydrogen bonds are lost between Tyr107 and Glu117 [32,73–75]. The loss of these hydrogen bonds was identified as being responsible for active site distortion and protein misfolding. In addition, studies in 1991 by Venta et al. [75] demonstrated hydrogen bond loss between Tyr107 and Tyr193 in H107Y. However, within this study, no hydrogen bond loss between these two residues was observed.

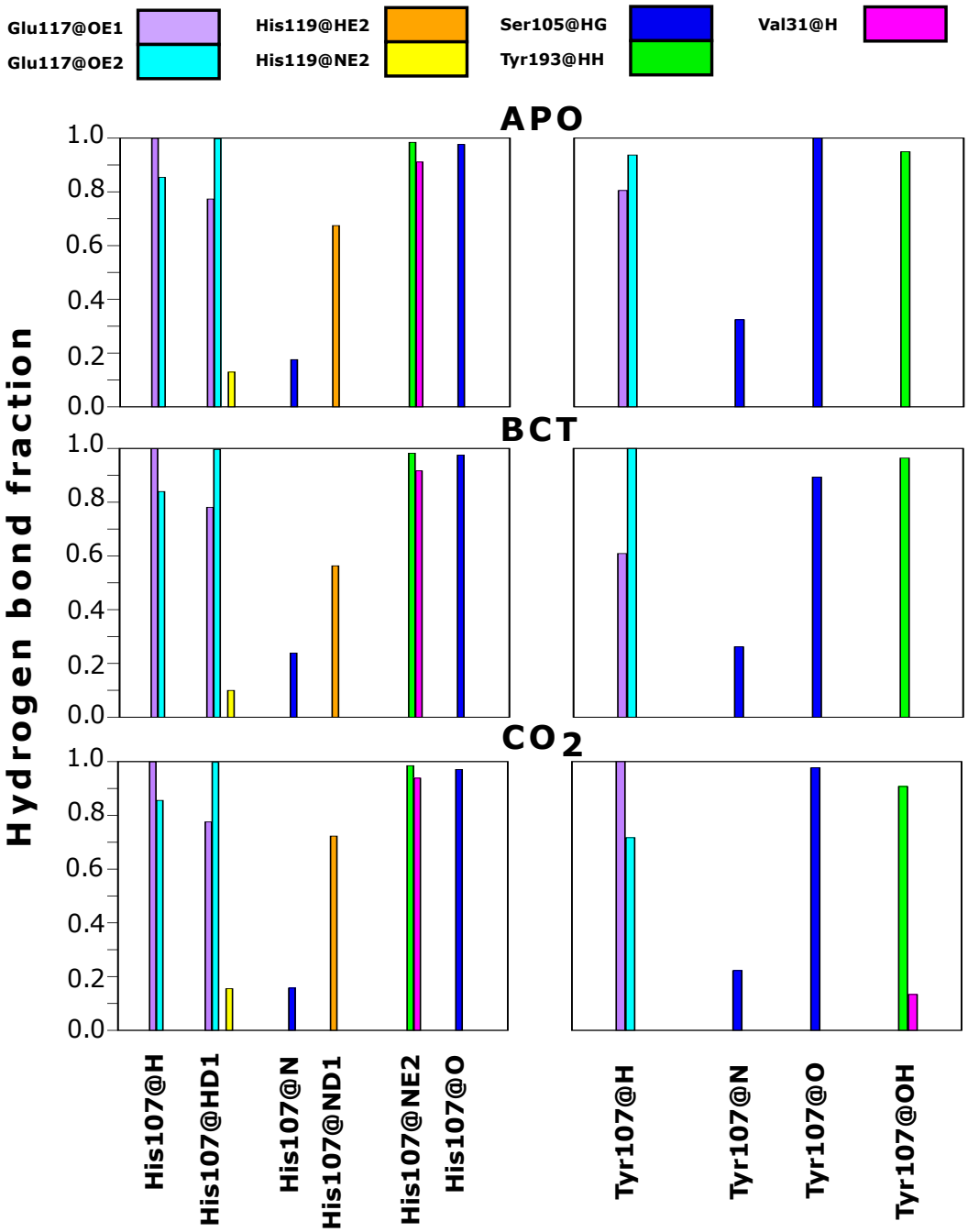

Figure 7. Hydrogen bond fraction between residue 107 and neighboring atoms in the WT and variant proteins. Fraction represents proportion of the total MD simulation frames the hydrogen bond existed between the specific atoms. **Left:** WT. **Right:** H107Y. Suffix @ refers to residue atoms; **H:** hydrogen; **HD:** delta hydrogen; **HE:** epsilon hydrogen; **HG:** gamma hydrogen; **HH:** eta hydrogen; **N:** nitrogen; **ND:** delta nitrogen; **NE:** epsilon nitrogen; **O:** oxygen; **OE:** epsilon oxygen; **OH:** hydroxyl.

The coupling of data in Tables 2 and S2 and Figure 7 allows for the identification of H107Y mechanism of action. Hydrogen bond loss with both Glu117 and His119 could destabilize the active site Zn^{2+} and possibly increase Zn^{2+} dissociation from the active site and cause instability. Glu117 has already been shown to have influence on Zn^{2+} rate of dissociation and enzyme activity [1,26,76].

Analysis of P236H contact maps indicates that there are increases to interactions between His236 and Glu238 in the apo and BCT bound protein; however, P236R shows increases to interactions between Arg236 and Glu238 when substrates BCT and CO_2 are bound. Differences to interaction loss between P236H and P236R suggest that although these two variants are located at the same position, the mechanism of action differs. Glu238 is not located in any of the identified CA-II motifs and lies between motif 3 and motif 8. Changes to Glu238 interactions could indirectly affect the motif function.

N252D contact map analysis indicates decreases in weighted interactions between Asp252 and Glu26 for the apo and BCT bound protein. When CO_2 is bound, interactions with Glu26 are completely lost. The interaction losses with Glu26 appear to have minimal effect on N252D stability as evidenced by its RMSD distribution.

2.8. Dynamic Residue Networks Show Changes in Residue Accessibility and Communication within CA-II

We have identified the direct effects to residue interactions occurring as a result of SNV presence in CA-II and possible consequences to enzyme structure and function. However, results have also suggested potential indirect variant mechanisms of action (i.e., allosteric effects) evidenced by motif analysis (Table 2) and RMSF (Figure 5). To fully understand SNV effects, investigation into indirect and/or compensatory variant mechanisms of action was performed using the average shortest path (L) and *betweenness centrality* (BC).

2.8.1. Average Shortest Path (L)

L measures the accessibility of a residue within a protein, while average L refers to the mean protein residue accessibility across all MD frames. Normalized average ΔL (variant minus WT) values were calculated for each WT and variant protein system across all MD frames (Figure 8). A decrease to ΔL indicates that variant residues are moving closer to each other with respect to the WT and becoming more accessible, whereas, an increase to ΔL indicates a decrease in residue accessibility within the variant in comparison to the WT. In general, ΔL calculations identified subtle differences. Table S6 presents residues with an average ΔL greater or less than 2 standard deviations from the mean ΔL for each protein system indicating significant changes to accessibility.

Overall, in almost all proteins, an increase to residue accessibility was observed for the region of residues ≈150–180 (motif 7 and motif 11), and this trend was also observed within the RMSF results. Previous studies involving L and RMSF have suggested that a linear correlation exists between the two metrics [65]. Figure 8 also presents changes to accessibility of Ala54 in all variants. K18Q, H107Y and P236R show decreases to Ala54 accessibility while the other variants do not.

The K18E apo protein shows an increase to residue accessibility in the motif 10 residues His10 and Trp16. With regards to the BCT and CO_2 bound K18Q proteins, results show increases to residue accessibility between residues 3 and 11. As motif 10 contains initial aromatic cluster residues (Trp5, Trp7 and Trp16), increases to motif 10 accessibility could explain the greater conformational sampling observed in the RMSD results for K18E apo and BCT and CO_2 bound K18Q (Figure 3). The apo protein for K18E, P236H and P236R shows increases to residue accessibility between residues 230 and Glu238. ΔL decreases in P236H could be as a result of an increase in contacts between residue 236 and Glu238 observed within the contact maps.

Results in Table S6 indicate that the majority of changes to residue accessibility occur away from residues previously identified as important for CA-II structure and function (Table S1). It should however be noted that although changes to ΔL do not occur at these residues, accessibility changes to neighboring residues could still affect important residues as proteins function as a network.

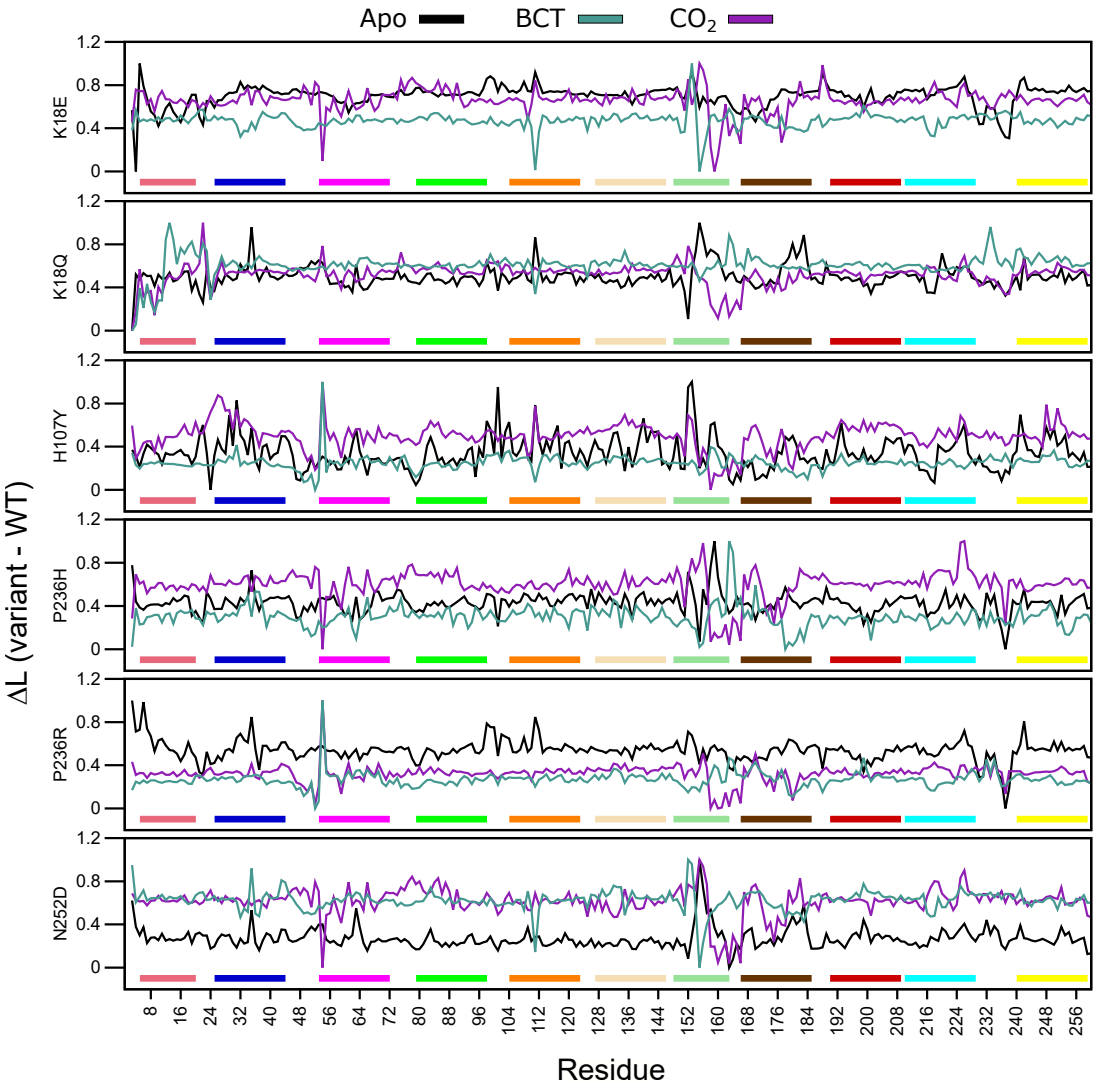

Figure 8. Change in average shortest path *L* between the WT and variant protein (variant minus WT) over the MD simulation. Respective motifs are indicated as bars at the bottom of each plot. Motif colors correspond to those in Figure 2B.

2.8.2. *Betweenness Centrality* (*BC*)

Average *L* allowed for the determination of residues accessibility within the WT and variant proteins. The effect of residue accessibility on protein communication was then measured using *BC*. During MD simulation, protein residues are constantly communicating with one another, and *BC* allows for identification of the most important residues for protein structure and function. Residues with the highest *BC* are most important for protein communication. When SNVs are present, residue communication could change, and analysis of these changes would allow for the identification of compensatory mechanisms employed by variants in an attempt to maintain protein function and stability [41,77].

Average *BC* analysis (data not shown) revealed that Glu117 has the highest BC in all WT and variant proteins. This suggests that Glu117 is the most important residue for communication in CA-II. To observe the *BC* differences between the WT and variant proteins, normalized Δ*BC* (variant minus WT) was calculated. A decrease to Δ*BC* indicates a decrease in residue usage within the variant, whereas an increase to Δ*BC* demonstrates increased residue usage. Figure 9 presents the average normalized Δ*BC* of the WT and variant proteins, and Table S7 shows residues with an average Δ*BC*

greater or less than 2 standard deviations from the mean Δ*BC* for each protein system indicating significant changes to residue communication.

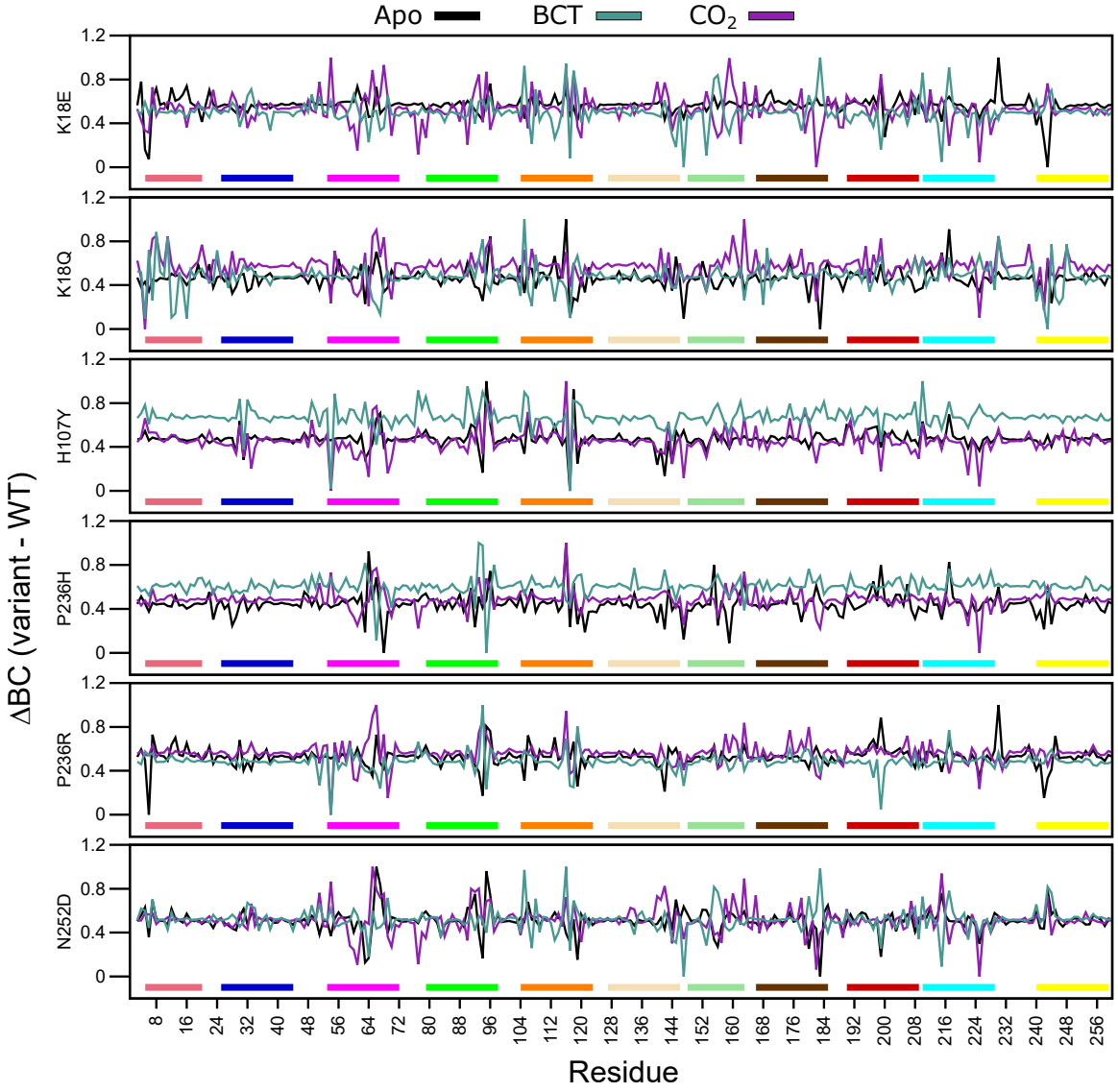

Figure 9. Change in *betweenness centrality* (*BC*) between the WT and variant protein (variant minus WT) over the MD simulation. Respective motifs are indicated as bars at the bottom of each plot. Motif colors correspond to those in Figure 2B.

K18E apo and K18Q with BCT and CO_2 bound demonstrate decreases to Trp5 Δ*BC*. As Trp5 is part of the initial aromatic cluster, the decreased residue usage could explain the greater conformational sampling observed within the RMSD. Analysis of H107Y Δ*BC* results indicates a reduction to the usage of Glu117 for all three protein states. The reduction in usage of Glu117 can be explained by loss of weighted contacts and hydrogen bonds observed. The results further suggest that H107Y could influence Zn^{2+} dissociation through changes to interactions with Glu117. When CO_2 was bound to H107Y, Zn^{2+} ligand His96 showed an increase in Δ*BC* highlighting a potential compensatory measure to maintain Zn^{2+} affinity and stability within the active site. More Δ*BC* increases to secondary aromatic cluster residues were observed in the apo and CO_2 bound proteins, suggesting potential greater variant effects. In all three H107Y protein systems (apo, BCT and CO_2), increases to Δ*BC* of Phe95 were noted. Phe95 is part of motif 4 and is a member of the secondary aromatic cluster. Increases to Phe95 Δ*BC* could be a compensatory measure by H107Y to maintain structure and function.

Within the other variant proteins, the presence of BCT is associated with a reduction in Glu117 usage, highlighting at potential effects on Zn^{2+} affinity for the active site. In all protein systems Glu117 Δ*BC* decreases are associated with Δ*BC* increases to either of the Zn^{2+} primary coordinating residues His94, His96 and/or His119 and the Zn^{2+} secondary ligand Asn243 evidenced in Table S7, highlighting potential compensatory mechanisms in maintaining Zn^{2+} within the active site. An interesting result is also observed where a decrease in Δ*BC* of one or more Zn^{2+} coordinating residues is associated with an increase in Δ*BC* of another coordinating residue. For example, Δ*BC* decreases in His94 and His119 in the P236R and N252D apo proteins are associated with Δ*BC* increases to His96. With regards to P236R apo protein the increase to His96 communication could be a result of the decrease in the usage of Asn243. Asn423 is a secondary Zn^{2+} coordinating ligand that maintains Zn^{2+} stability through direct interactions with His96 [1].

Δ*BC* results indicate that apart from H107Y, variant effects occur away from the SNV site, and variant presence has allosteric effect on residues important for the structure and function of CA-II as highlighted in Table S7 data. Although SNV effects are more evident at active site residues, some indirect effects are observed at aromatic cluster residues as observed in N252D. This suggests potential implications for precision medicine related studies in the treatment of CA-II deficiencies. Treatment options would have to be targeted to either rescue the primary and secondary aromatic cluster residues or the active site residues.

2.9. Variant Presence Shows Remote Effects on Proton Shuttle Residue

As DRN analysis demonstrated evidence that variant effects occur away from the SNV site, the proton shuttle residue His64 was investigated for potential variant effects occurring during protein dynamics that may not have been observed through MD and/or DRN analysis. His64 behavior is also governed by the pKa of surrounding residues [78–80] therefore changes to His64 conformations may not be visible using RMSF analysis, and the resulting conformations may not change residue accessibility or usage.

During MD simulation His64 was observed to rotate between two main conformations ("in" and "out"). Figure 10A shows an example of these conformations in the WT apo (green) protein. The presence of the two conformations agrees with previous literature findings [16–19]. Data in Table 3 shows the average distances of the imidazole ring of His64 from the Zn^{2+} for the "in" and "out" conformations of the majority representative structural clusters. It is evident that the apo proteins are associated with greater His64 conformations.

From the K18E apo results, it is observable that His64 does not occupy an "out" conformation. The variant, however, occupies what we have termed a "faux in" conformation (Figure 10A), which is being observed for the first time in this study, and has not been noted in previously studied literature [1,6,10,16,81]. The "faux in" conformation was observed for a fraction of 0.972 in all MD frames, and interestingly compared to the other protein systems, this conformation brings the imidazole ring closest to the Zn^{2+}. As no "out" conformation for His64 was observed, data suggests that His64 in K18E may not be able to adopt an "out" conformation that could significantly impact proton shuttling as a result of a larger water network being required to shuttle protons out of the active site [82–84]. His64 in the "faux in" conformation may also not be able to assist with the stabilization of active site water network to the same extent as the "in" conformation, which could have an effect on reaction rates [83].

K18Q results when BCT is bound (Table 3) show that His64 did not occupy the "out" conformation and remained in the "in" conformation for the duration of the MD simulation. Data in Table 3 also shows an unusual result with regards to P236R apo results. His64 can occupy all three conformations. The "in", "out" and "faux in" conformations were present for a fraction of 0.049, 0.123 and 0.827 of all MD frames, respectively, suggesting a strong preference for "faux in". Analysis of the trajectory (results not shown) and the order of emergence of conformations in the MD frames suggests that His64 when in the "faux i" conformation may not be able to rotate directly to the "out" conformation, without transitioning through the "in" conformation initially. With respect to K18E and P236R, the "faux in"

conformation was not observed when substrate was bound, suggesting that substrate presence could influence the behavior of His64.

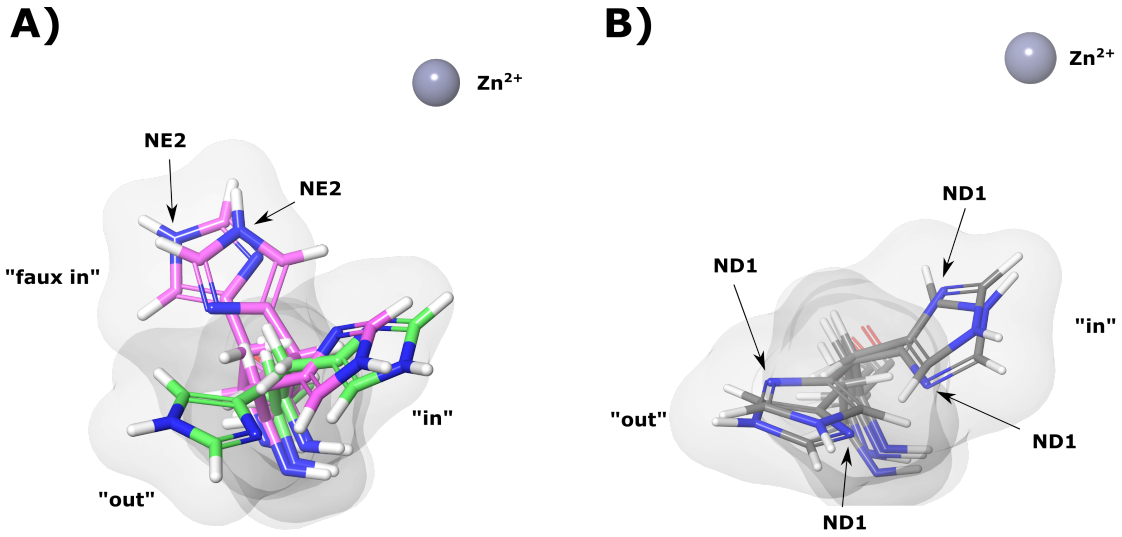

Figure 10. CA-II apo protein His64 "in", "out" and "faux in" conformations. (**A**) WT and K18E proteins. Green and magenta colors represent WT and K18E, respectively. (**B**) N252D variant. The grey sphere represents the Zn^{2+}, white, blue and red structures represent hydrogen, nitrogen and oxygen atoms, respectively.

Table 3. Distance of His64 imidazole group from Zn^{2+} for the "in" and "out" conformations. All distances are measured from the His64 imidazole ring centroid to the Zn^{2+}. Faux refers to other conformations observed excluding traditional "in" and "out" occupied by His64.

Variant	Imidazole-Zn^{2+} Distance (Å)								
	apo			BCT			CO_2		
	In	Out	Faux in	In	Out	Faux in	In	Out	Faux in
K18E	8.65	*	7.30	8.08	11.09	*	6.96	11.20	*
K18Q	8.11	11.01	*	8.22	*	*	8.96	10.42	*
H107Y	8.50	10.63	*	8.24	10.86	*	7.98	10.67	*
P236H	8.57	11.26	*	8.70	12.02	*	7.12	11.71	*
P236R	8.26	11.02	7.36	7.99	10.84	*	7.50	10.43	*
N252D	8.11	11.33	*	8.65	10.93	*	8.63	11.20	*
WT	8.24	11.21	*	8.57	11.07	*	8.45	11.31	*

* Conformation not observed.

Additional analysis of the "in" and "out" conformations of His64 and other structural clusters also showed another interesting finding in which rotation of the His64 CB-CG (beta carbon atom and gamma carbon atom) bond was observed. An example of this bond rotation is shown in Figure 10A,B evidenced by the NE2 and ND1 atoms of His64 occupying different spatial orientations in K18E and N252D. Rotation of the His64 CB-CG in addition to being observed in the variants was also noted in the WT proteins. This result could suggest potential implications for the mechanism of action of His64 in proton shuttling. Previous research has suggested that either His64 "in" and "out" rotation and/or imidazole tautomerization to be responsible for proton shuttling [16–20]; however, this finding could suggest that in addition to tautomerization of His64 in the "out" conformation to shuttle protons, the His64 CB-CG bond rotation could facilitate imidazole ring rotation to shuttle a proton to and from the active site during catalysis. Additional research is however required to confirm this.

3. Materials and Methods

Figure 11 summarizes the overall approach, methods and software used to conduct research in this article.

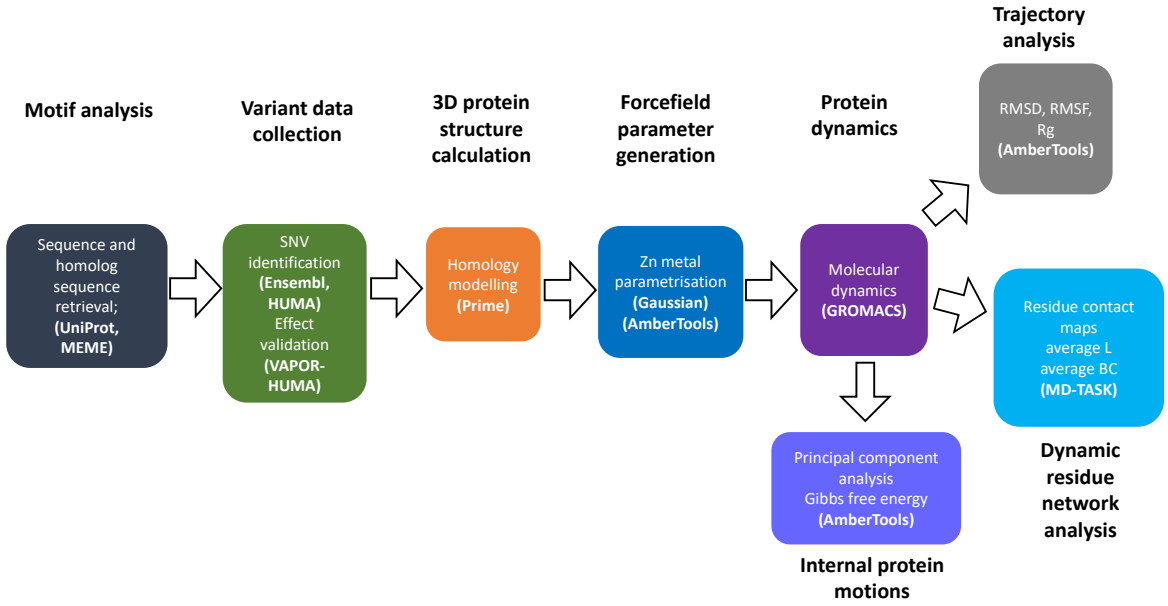

Figure 11. Flow chart of methodology used in analysis of the effects SNVs on CA-II structure and function.

3.1. Data Retrieval

The amino acid sequence of CA-II was retrieved from the Universal Protein Resource (UniProt) [11]. UniProt BLAST was then used to identify other CA sequences within the α-CA family using the CA-II sequence as the query. The BLASTp program was used to search for homologous sequences against the UniProtKB target database using the BLOSUM-62 matrix [85] and the UniProt E threshold search parameter value of 100. From the BLAST results, all human CA proteins CA-I to XIV, including mouse CA-XV, were selected to create the final dataset (Table S2).

3.2. Motif Analysis

Discovery and analysis of motifs was performed according to methodology by Ross et al. and Nyamai and Tastan Bishop [56,86] with slight modification using the retrieved sequences as a query. Briefly, motif discovery was performed using online MEME SUITE version 5.05 [55], using the 0-order model of sequences. A minimum and maximum motif width of 5–20 residues was set. A total of 100 motifs for each human α-CA protein were set up for discovery. Discovered MEME motifs were then validated using MAST and E-value for inclusion or exclusion. Only motifs with an E-value less than 0.001 were retained. In addition, motifs with pairwise correlations greater than 0.6 were removed using MAST to create the final dataset. If the pairwise correlation is greater than 0.6, this may cause some p-values and E-values to be underestimated and diminishes accuracy [55]. The existence of each final dataset motif was then checked for in each CA sequence, and matplotlib [87] was used to construct a heat map representing motif conversation as the number of sites per total number of protein sequences. Motif logos showing the conservation of each amino acid residue in each motif sequence were also downloaded from the MEME server and visualized using Inkscape [88].

3.3. Homology Modeling

3.3.1. Wild-Type

There are 652 entries in the Protein Data Bank (PDB) [89] for the 3D structure of CA-II protein (UniProt accession number: P00918; accessed 20/06/2018); however, the majority of them are missing residue number 126 [90], hindering accurate CA-II SNV modeling, parametrization and MD. Residue 126 exists only in CA-I [91]; therefore, template numbering proceeds from 125 and continues to 127 even though the FASTA sequence matches the ATOM sequence 100%. Some of the structures also lack Zn^{2+} in the correct coordination geometry, which would have implications on metal site parametrization.

Potential templates from the 652 structures were filtered using AmberTools17 [58] by observing all atoms bonded to the Zn^{2+} within a radii of 2.5 Å in a tetrahedral geometry. The 2.5 Å is the maximum Zn-ligand bond distance [60,63]. PDBs containing an HHHX (3 Histidines and 1 H_2O) coordinating formation with Zn^{2+} were considered, and the best template was selected based on 3D structure resolution and PDB validation. Crystal structure 2VVA with resolution of 1.56 Å, and a sequence similarity of 99% to the UniProt protein (covering residues 3–260) was chosen as the template. 2VVA also contains CO_2 co-crystallized to the primary pocket of CA-II. Due to errors encountered when using *pdb4amber* with 2VVA, renumbering was insufficient to correct missing residue 126, and therefore, homology modeling was necessary.

Clustal Ω (Omega) [92,93] was utilized to align the target CA-II sequence (UniProt accession: P00918) and template sequence (2VVA). Schrödinger Prime [52,53] was used to generate five homology models that included the ligands Zn and CO_2. Wild-type (WT) model quality was validated using the z-DOPE score and Ramachandran plot. If the z-DOPE score is less than −1.0, homology models are regarded as being like native structures. A Ramachandran plot allows for the identification of energetically favorable amino acid backbone conformations by comparison of φ (phi) and ψ (psi) torsion angles.

3.3.2. Variants

SNV analysis was carried out according to proposed protocol by Brown and Tastan Bishop, 2017 [94]. Briefly, to the downloaded SNVs from HUMA [44] and Ensembl [43], datasets were filtered to include only CA-II nsSNVs. Datasets were further filtered to include only SNVs validated within dbSNP [95]. SNVs were then cross referenced with the Clinvar [96] and OMIM [97] databases to isolate pathogenic variants. VAPOR in HUMA [44] was used to predict the expected effects of these SNVs on CA-II structure and function.

Homology modeling was performed to introduce the identified nsSNVs into the CA-II structure, with unique structures generated for each nsSNV. The CA-II amino acid sequence was first modified to contain the required SNVs. Homology modeling was then performed according to the WT methodology. The WT protein model was used as the template, and five models including all HETATMs (hetero-atoms) were generated per variant protein. The z-DOPE score and Ramachandran plots were used to validate the quality of the models.

WT and variant apo protein models were generated by removing CO_2 from the modeled proteins.

3.3.3. Bicarbonate Bound Structure

Besides the WT and variant apo and CO_2 bound protein models, bicarbonate (BCT) bound WT and variant complexes were also calculated using the crystal structure (PDB ID: 2VVB). In this case, the apo WT and variant protein models were superimposed with 2VVB using PyMOL [98]. The coordinates of BCT were then added to the apo proteins to generate the final BCT containing structures. Superposition to incorporate BCT into the protein structure was preferred as opposed to modeling. The superposition allows for the maintenance of one reference structure (CO_2 bound model) for apo, CO_2 and BCT bound

structures comparison. Modeling could have introduced some structural changes to the proteins and inhibited direct comparison of all three protein models for WT and variant proteins.

In total, 21 models (7 apo, 7 with CO_2 and 7 proteins with BCT) were calculated and taken into the further calculations as explained below.

3.4. Zn^{2+} Parametrization

Due to the presence of Zn^{2+} in the CA-II active site and its importance in protein function, AmberTools17 was used to perform metal site parametrization for the 21 WT and variant CA-II protein models, utilizing the bonded model approach and the MCPB [62]. Zn^{2+} parametrization was necessary to develop atom forcefields for interactions and to ensure the metal remained within the active site during MD simulations.

CA-II protein structures where protonated using the H++ server [99] at pH 7.0 [100,101], with an internal and external dielectric of 10 and 80, respectively, and a system salinity of 0.15 M. AMBER topology and coordinate files (*top* and *crd*) were then downloaded from the server, and *ambpdb* used to generate a protonated PDB file. To the generated PDB file, protonation states of all structures were then validated by ensuring that the Zn^{2+} coordinating ligands; His94, His96 and His119 were in the correct protonation states (HID, HID, and HIE, respectively).

MCPB was then used to generate Gaussian 09 input files, prior to QM calculations. QM calculations were performed according to methodology by Li and Merz 2017 [62] using the B3LYP/6-31G basis set in Gaussian 09 [59], using 192 CPU cores at the Center for High Performance Computing (CHPC) cluster, Cape Town South Africa. To the Gaussian 09 output files, MCPB was then used to calculate bond lengths, angles and dihedrals between Zn^{2+} and coordinating atoms, to derive the final Zn^{2+} forcefield parameters.

3.5. Molecular Dynamics

For MD simulations, CA-II protein structures were protonated by the Schrödinger Maestro [102] Protein Preparation Wizard in conjunction with PROPKA [103] at pH 7.0 [100,101]. The previously generated forcefield parameters were used in conjunction with Leap modeling [104] to generate AMBER topologies, utilizing the AMBER ff14SB forcefield [105] and a cubic box of cut-off distance 10 Å (distance between protein molecule and box). To the box, water molecules adhering to the TIP3P water model were added as solvent, and a concentration of 0.15 M NaCl was used to neutralize the system. Generated AMBER topology files were then converted to GROMACS [106] topology using ACPYPE [107] to generate *gro* and *top* files, prior to energy minimization. Generated topology files were then manually inspected to ensure accurate neutralization by comparing the total protein charge (*qtot*) to the quantity of counter-ion added. AMBER to GROMACS topology conversion, using ACPYPE, maintains all previously set Leap parameters, such as cubic box cut-off distance, across the different programs, and generated topologies can be directly minimized.

MD simulations were conducted for all 21 protein structures using GROMACS 2018.2 [106]. Energy minimization was setup for 50,000 steps using the steepest descent algorithm and terminated when a maximum force (F_{max}) of no greater than 1000 kj mol^{-1} nm^{-1} was attained, and the system converged. Temperature and pressure equilibration (*NVT* and *NPT* ensemble, respectively) were performed after energy minimization, using the modified Brenson thermostat and the Particle Mesh Ewald (PME) coulomb type for long-range electrostatics. All bonds were constrained under the LINCS algorithm. The *NVT* ensemble was performed for 100 ps at 300 K, followed by the NPT ensemble, until the system stabilized at a pressure of 1 bar. MD simulations were performed at the CHPC cluster, on one Nvidia Tesla v100 GPU in conjunction with 10 CPU cores over a period of 200 ns, with a time integration step of 2 fs. Coordinates were written to file every 10 ps.

3.6. Molecular Dynamics Trajectory Analysis

To ensure efficient and accurate protein analysis, the trajectories resulting from the MD simulations were stripped of all periodic boundary conditions and centered within the simulation box using *cpptraj* [108]. From the corrected trajectory, new PDB files for the WT and protein variants were also generated for MD analysis. The Visual Molecular Dynamics program (VMD) [109] was used to visualize the trajectory and to evaluate whether the system adhered to the parameters set and whether the Zn^{2+} remained in place.

The AmberTools17 package *cpptraj* was used to calculate RMSD, RMSF and the Rg of the protein α-carbons.

3.7. Statistical Analysis

Statistical analysis of the RMSD and the Rg data was performed using the z-test for parametric data and the Mann–Whitney U test (MWU) for non-parametric data. All statistical analysis and data generation was performed using the RStudio v1.1.456 [110] integrated development environment, in combination with R v3.5.1 [111].

3.8. Proton Shuttle Analysis

To observe changes to proton shuttle residue behavior during MD simulation, structural clustering based on His64 conformations was performed for all protein systems using the hierarchical agglomerative (bottom up) algorithm and the average-linkage method [112–114], in conjunction with *cpptraj* and the *cluster* command. Cluster command was set to perform clustering every four frames to generate a total of four protein clusters to account for His64 "in" and "out" conformations, flips in the imidazole ring and other potential orientations.

Conformational clusters were generated using the following three criteria: angle between His64 CB, CG atoms and the Zn^{2+}; dihedral angles between His64 N, CA, CB and CG atoms (chi1) and CA, CB, CG and ND1 (chi2); and distance between His64 ND1 atom and the Zn^{2+}. Protein structures representative of each conformational cluster were also generated.

3.9. Dynamic Residue Network Analysis

DRN analysis for each protein was done using MD-TASK [42]. Residue interaction was predicted through the evaluation of pairwise distances between all Cβ (Cα for glycine) atoms, across all complexes present in each frame during the MD trajectory. Within the DRN, each protein residue is a node within the network [42].

3.9.1. Weighted Contact Map Analysis

Contact maps show the frequency of interaction between two residues over an MD trajectory. Residue–residue interactions in the WT and protein variants over the 200 ns simulations were calculated using the *contact_map.py* script in MD-TASK, using a cut-off distance of 6.7 Å previously reported for Cα-Cα node interaction [115]. The weighted contacts of the WT and SNV residues were compared to observe the effect of SNVs on short-range residue–residue interaction.

3.9.2. Average Shortest Path (L)

L measures the accessibility of a protein node (specific residue) by computing the total number of shortest paths to that node and dividing by the total number of nodes minus one [42]. Equation (2) is used to calculate *L*.

$$\alpha = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \quad (2)$$

From Equation (2), V is the set of nodes in the network. $d(s,t)$ is the shortest path from s to t, while n represents the total number of nodes within the network [116]. An increase in L shows a decrease in residue accessibility, while decreases in L signify an increase in residue accessibility. L was calculated across all MD frames for a threshold of 6.7 Å using the *calc_network.py* script in MD-TASK. Calculated L was then normalized on a scale of 0 to 1 using unity-based normalization to generate normalized L. The average normalized L was determined by averaging normalized residue L across all frames of trajectory for each residue in each protein. The ΔL for each residue in the DRN was then calculated by subtracting the average normalized L of the WT and variant proteins (variant minus WT).

3.9.3. Average *Betweenness Centrality (BC)*

BC governs the importance of a residue for protein communication. The *BC* of a node equates to the number of shortest paths passing through that node, interlinking one specific node to numerous others. *BC* is determined according to Equation (3).

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t))} \tag{3}$$

Within Equation (3), V represents network nodes, whereas $\sigma(s,t)$ is the number of shortest (s,t) paths. $\sigma(s,t|v)$ is the number of those paths passing through node v other than s,t [116]. The more frequent the communication of a residue within a protein, the higher the *BC*, indicating the importance of that residue within the protein. *BC* was determined following methodology to that of L (Section 3.9.2). Briefly, *BC* for each protein residue was calculated using the *calc_network.py* script for a threshold of 6.7 Å. Unity based normalization was applied to generate normalized *BC*. The average normalized *BC* for each protein residue was then determined by averaging the normalized *BC*. WT and variant average normalized *BC* were then subtracted (variant minus WT) to calculate ΔBC.

3.10. Principal Component Analysis (PCA)

PCA was performed using *cpptraj* according to methodology in 2014 by Roe et al. [117]. Global rotational/translational motion was removed from the trajectories by applying an RMS best-fit to an average structure. The coordinate covariance matrix was then calculated for all heavy atoms (excluding hydrogen) for the WT protein and variants, followed by diagonalizing to obtain the eigenvectors and eigenvalues. Variant coordinates were then projected along each eigenvector to obtain separate projections for each trajectory set. The first and second projections where then normalized and plotted against each other to obtain a graph of PC1 against PC2 using the *hist* analysis command of *cpptraj*. The free energy associated with PC1 and PC2 at 300 K was also calculated using the *hist* command.

4. Conclusions

In conclusion, motif analysis, MD, PCA and DRN techniques were used to investigate the potential effects of the six validated SNVs (K18E, K18Q, H107Y, P236H, P236R and N252D) on the structure, stability and function of human α-CA-II in both the presence and absence of substrates BCT and CO_2. The study was divided into two parts comprising global and local level protein analysis. Motif studies identified 11 motifs potentially important to CA-II structure, function and stability. Only the variants K18E, K18Q, H107Y and N252D are, however, located on these motifs. Global level analysis indicated subtle effects of variant presence on CA-II structure. The presence of substrates BCT and CO_2 is associated with greater SNV effects, and in general, when substrate BCT is bound, PCA analysis showed an increase to conformational sampling and free energy in all protein structures. Local analysis using DRN demonstrated that the effects of the SNVs on CA-II structure and function occur away from the SNV location, indicating an allosteric/indirect SNV effect. In all cases, Glu117 was identified as the most important residue for communication within CA-II. Variant H107Y results showed a reduction to interactions between Tyr107 and Glu117. A loss of at least two hydrogen

bonds was also noted for these two residues. Average *BC* analyses indicated a reduction to the usage of residue Glu117 in H107Y for all three protein states as a result of the interaction loss, while the other variants showed a reduction to Glu117 usage mainly when BCT was bound. The reduction to Glu117 usage suggests potential implications for metal ion affinity for the CA-II active site and was associated with increases in the usage of the Zn^{2+} coordinating residues His94, His96 and His119, indicating compensatory mechanisms to maintain Zn^{2+} within the active site. Proton shuttle analysis also highlighted the formation of a novel His64 conformation "faux in" in K18E and P236R apo proteins. The imidazole ring of the "faux in" conformation is located closer to the Zn^{2+} compared to the "in" and "out" conformations.

In addition to the findings summarized above, we also emphasize that MD simulations coupled with DRN analysis [41,42] would provide detailed insights into the understanding of SNV mechanism of action at the molecular level and would help with precision medicine related studies. In light of our findings, this study proposes taking steps towards the treatment of CA-II deficiencies, which would either be by rescuing the primary and secondary aromatic cluster residues or the active site residues. Possible future work could include quantum mechanical calculations as to the effect of SNVs on the proton transfer pathways in CA-II.

Supplementary Materials: The following are available online, Table S1: Identified CA-II residues important for structure, function and stability. Table S2: Table of UniProt CA accession numbers for sequences used in motif discovery. Figure S1: Heat map of all conserved motifs within human α-CA family and associated UniProt accession Motif conservation is represented as number of motif sites per total protein sequences. A value of zero shows that motifs are not conserved in any sequence, whereas a value of 1 shows that motif is conserved in all sequences. The prefix 'M' represents the word motif. Table S3: Zn^{2+} non-bonded, bonds, angles and dihedral parameters derived within this study $\mathbf{K}_b$**:** bond force constant; $\mathbf{K}_\theta$**:** angle force constant; $\mathbf{R}_{min}$**:** vdW radius; ϵ**:** LJ potential well energy. Figure S2: RMSD comparison between the WT and variant proteins during MD. Table S4: Associated *p*-values for the RMSD and Rg distribution data. Figure S4: 3-dimensional (3D) plot of PC1 vs PC2 of the WT and variants proteins as a function of free energy. Free energy is represented in kcal/mol. Table S5: Eigenvalue fraction of each principal component for the active site residues within the wild-type and variant proteins in the absence (apo) and presence (non apo) of CO_2. Figure S4: Rg comparison between the WT and variant proteins during MD. Figure S5: Residue RMSF comparison between the WT and variant proteins. Table S6: Variant residues showing decreases and increases to ΔL during MD simulation. Residues from Table S1 are underlined and highlighted in bold. SNV positions are underlined, italicized and highlighted in bold red. Table S7: Variant residues showing decreases and increases to ΔBC during MD simulation. Residues from Table S1 are underlined and highlighted in bold. SNV positions are underlined, italicized and highlighted in bold red.

Author Contributions: Conceptualization, Ö.T.B.; methodology, T.A.S., B.N. and Ö.T.B.; validation, T.A.S. and Ö.T.B.; formal analysis, T.A.S. and Ö.T.B.; investigation, T.A.S.; resources, Ö.T.B.; writing—original draft preparation, T.A.S.; writing—review and editing, Ö.T.B.; visualization, T.A.S.; supervision, Ö.T.B.; project administration, Ö.T.B.; funding acquisition, Ö.T.B.

Funding: This research was funded by National Research Foundation (NRF) South Africa grant numbers 105267 and 111212. The authors are responsible for the content of this publication and it is not a representation of the funder's official views.

Acknowledgments: The authors would like to thank the Center for High Performance Computing (CHPC) Cape Town, South Africa, for providing the cluster to perform molecular dynamics and quantum mechanical calculations. Authors thank Prof Reza Zolfaghari Emameh, National Institute of Genetic Engineering and Biotechnology, Iran, for the introduction to the carbonic anhydrase family of enzymes.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BC	*Betweenness centrality*
BCT	Bicarbonate
BLAST	Basic local alignment search tool
CA	Carbonic anhydrase
DRN	Dynamic residue network

L	Average shortest path
LJ	Lennard–Jones
MCPB	Metal Centre Parameter Builder
MD	Molecular dynamics
PCA	Principal component analysis
QM	Quantum mechanical
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
Rg	Radius of gyration
SNV	Single nucleotide variation
WT	Wild type

References

1. Lindskog, S. Structure and mechanism of carbonic anhydrase. *Pharmacol. Ther.* **1997**, *74*, 1–20. [CrossRef]
2. Silverman, D.N.; Lindskog, S. The catalytic mechanism of carbonic anhydrase: implications of a rate-limiting protolysis of water. *Accounts Chem. Res.* **1988**, *21*, 30–36. [CrossRef]
3. Tripp, B.C.; Smith, K.; Ferry, J.G. Carbonic anhydrase: New insights for an ancient enzyme. *J. Biol. Chem.* **2001**, *276*, 48615–48618. [CrossRef] [PubMed]
4. Di Fiore, A.; Alterio, V.; Monti, S.M.; De Simone, G.; D'Ambrosio, K. Thermostable carbonic anhydrases in biotechnological applications. *Int. J. Mol. Sci.* **2015**, *16*, 15456–15480. [CrossRef] [PubMed]
5. Somalinga, V.; Buhrman, G.; Arun, A.; Rose, R.B.; Grunden, A.M. A High-Resolution Crystal Structure of a Psychrohalophilic α–Carbonic Anhydrase from Photobacterium profundum Reveals a Unique Dimer Interface. *PLoS ONE* **2016**, *11*, e0168022. [CrossRef]
6. Supuran, C.T. Structure and function of carbonic anhydrases. *Biochem. J.* **2016**, *473*, 2023–2032. [CrossRef]
7. Soto, A.R.; Zheng, H.; Shoemaker, D.; Rodriguez, J.; Read, B.A.; Wahlund, T.M. Identification and preliminary characterization of two cDNAs encoding unique carbonic anhydrases from the marine alga Emiliania huxleyi. *Appl. Environ. Microbiol.* **2006**, *72*, 5500–5511. [CrossRef]
8. Lane, T.W.; Saito, M.A.; George, G.N.; Pickering, I.J.; Prince, R.C.; Morel, F.M. Biochemistry: A cadmium enzyme from a marine diatom. *Nature* **2005**, *435*, 42. [CrossRef]
9. Hewett-Emmett, D.; Tashian, R.E. Functional diversity, conservation, and convergence in the evolution of the α-, β-, and γ-carbonic anhydrase gene families. *Mol. Phylogenetics Evol.* **1996**, *5*, 50–77. [CrossRef]
10. McKenna, R.; Frost, S.C. Overview of the carbonic anhydrase family. In *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications*; Springer: Berlin, Germany, 2014; pp. 3–5.
11. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
12. Seifter, J.L.; Chang, H.Y. Disorders of acid-base balance: new perspectives. *Kidney Dis.* **2016**, *2*, 170–186. [CrossRef]
13. Orešković, D.; Klarica, M. The formation of cerebrospinal fluid: Nearly a hundred years of interpretations and misinterpretations. *Brain Res. Rev.* **2010**, *64*, 241–262. [CrossRef] [PubMed]
14. Blair, H.C.; Teitelbaum, S.L.; Ghiselli, R.; Gluck, S. Osteoclastic bone resorption by a polarized vacuolar proton pump. *Science* **1989**, *245*, 855–857. [CrossRef] [PubMed]
15. Berg, J.M.; Tymoczko, J. Chapter 9, Making a Fast Reaction Faster: Carbonic Anhydrase. In *Biochemistry*, 5th ed.; WH Freeman and Company: New York, NY, USA, 2002.
16. Silverman, D.N.; McKenna, R. Solvent-mediated proton transfer in catalysis by carbonic anhydrase. *Accounts Chem. Res.* **2007**, *40*, 669–675. [CrossRef] [PubMed]
17. Tu, C.; Silverman, D.N.; Forsman, C.; Jonsson, B.H.; Lindskog, S. Role of histidine 64 in the catalytic mechanism of human carbonic anhydrase II studied with a site-specific mutant. *Biochemistry* **1989**, *28*, 7913–7918. [CrossRef]
18. Nair, S.K.; Christianson, D.W. Unexpected pH-dependent conformation of His-64, the proton shuttle of carbonic anhydrase II. *J. Am. Chem. Soc.* **1991**, *113*, 9455–9458. [CrossRef]

19. Boone, C.D.; Gill, S.; Tu, C.; Silverman, D.N.; McKenna, R. Structural, catalytic and stabilizing consequences of aromatic cluster variants in human carbonic anhydrase II. *Arch. Biochem. Biophys.* **2013**, *539*, 31–37. [CrossRef]
20. Shimahara, H.; Yoshida, T.; Shibata, Y.; Shimizu, M.; Kyogoku, Y.; Sakiyama, F.; Nakazawa, T.; Tate, S.I.; Ohki, S.Y.; Kato, T.; et al. Tautomerism of histidine 64 associated with proton transfer in catalysis of carbonic anhydrase. *J. Biol. Chem.* **2007**, *282*, 9646–9656. [CrossRef]
21. Merz, K.M., Jr. Carbon dioxide binding to human carbonic anhydrase II. *J. Am. Chem. Soc.* **1991**, *113*, 406–411. [CrossRef]
22. Liang, J.Y.; Lipscomb, W.N. Binding of substrate CO2 to the active site of human carbonic anhydrase II: A molecular dynamics study. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 3675–3679. [CrossRef]
23. Domsic, J.F.; Avvaru, B.S.; Kim, C.U.; Gruner, S.M.; Agbandje-McKenna, M.; Silverman, D.N.; McKenna, R. Entrapment of carbon dioxide in the active site of carbonic anhydrase II. *J. Biol. Chem.* **2008**, *283*, 30766–30771. [CrossRef]
24. Alexander, R.S.; Nair, S.K.; Christianson, D.W. Engineering the hydrophobic pocket of carbonic anhydrase II. *Biochemistry* **1991**, *30*, 11064–11072. [CrossRef] [PubMed]
25. Eriksson, A.E.; Jones, T.A.; Liljas, A. Refined structure of human carbonic anhydrase II at 2.0 Å resolution. *Proteins Struct. Funct. Bioinform.* **1988**, *4*, 274–282. [CrossRef] [PubMed]
26. Hunt, J.A.; Fierke, C.A. Selection of carbonic anhydrase variants displayed on phage aromatic residues in zinc binding site enhance metal affinity and equilibration kinetics. *J. Biol. Chem.* **1997**, *272*, 20364–20372. [CrossRef] [PubMed]
27. Fisher, Z.; Hernandez Prada, J.A.; Tu, C.; Duda, D.; Yoshioka, C.; An, H.; Govindasamy, L.; Silverman, D.N.; McKenna, R. Structural and kinetic characterization of active-site histidine as a proton shuttle in catalysis by human carbonic anhydrase II. *Biochemistry* **2005**, *44*, 1097–1105. [CrossRef] [PubMed]
28. Khalifah, R.G. The carbon dioxide hydration activity of carbonic anhydrase I. Stop-flow kinetic studies on the native human isoenzymes B and C. *J. Biol. Chem.* **1971**, *246*, 2561–2573.
29. Ho, C.; Sturtevant, J.M. The kinetics of the hydration of carbon dioxide at 25. *J. Biol. Chem.* **1963**, *238*, 3499–3501.
30. Steiner, H.; Jonsson, B.H.; Lindskog, S. The Catalytic Mechanism of Carbonic Anhydrase: Hydrogen-Isotope Effects on the Kinetic Parameters of the Human C Isoenzyme. *Eur. J. Biochem.* **1975**, *59*, 253–259. [CrossRef]
31. Jakubowski, M.; Szahidewicz-Krupska, E.; Doroszko, A. The Human Carbonic Anhydrase II in Platelets: An Underestimated Field of Its Activity. *BioMed Res. Int.* **2018**, *2018*, 4548353. [CrossRef]
32. Almstedt, K.; Lundqvist, M.; Carlsson, J.; Karlsson, M.; Persson, B.; Jonsson, B.H.; Carlsson, U.; Hammarström, P. Unfolding a folding disease: folding, misfolding and aggregation of the marble brain syndrome-associated mutant H107Y of human carbonic anhydrase II. *J. Mol. Biol.* **2004**, *342*, 619–633. [CrossRef]
33. Roth, D.E.; Venta, P.J.; Tashian, R.E.; Sly, W.S. Molecular basis of human carbonic anhydrase II deficiency. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 1804–1808. [CrossRef]
34. Shah, G.N.; Bonapace, G.; Hu, P.Y.; Strisciuglio, P.; Sly, W.S. Carbonic anhydrase II deficiency syndrome (osteopetrosis with renal tubular acidosis and brain calcification): Novel mutations in CA2 identified by direct sequencing expand the opportunity for genotype-phenotype correlation. *Hum. Mutat.* **2004**, *24*, 272. [CrossRef]
35. Scozzafava, A.; Supuran, C.T. Glaucoma and the applications of carbonic anhydrase inhibitors. In *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications*; Springer: Berlin, Germany, 2014; pp. 349–359.
36. Swenson, E.R. Carbonic anhydrase inhibitors and high altitude illnesses. In *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications*; Springer: Berlin, Germany, 2014; pp. 361–386.
37. Supuran, C.T.; Scozzafava, A.; Casini, A. Carbonic anhydrase inhibitors. *Med. Res. Rev.* **2003**, *23*, 146–189. [CrossRef] [PubMed]
38. Supuran, C.T. Carbonic anhydrases: Novel therapeutic applications for inhibitors and activators. *Nat. Rev. Drug Discov.* **2008**, *7*, 168. [CrossRef] [PubMed]
39. Puscas, I.; Coltau, M.; Baican, M.; Pasca, R.; Domuta, G. The inhibitory effect of diuretics on carbonic anhydrases. *Res. Commun. Mol. Pathol. Pharmacol.* **1999**, *105*, 213–236. [PubMed]

40. Shinohara, C.; Yamashita, K.; Matsuo, T.; Kitamura, S.; Kawano, F. Effects of carbonic anhydrase inhibitor acetazolamide (AZ) on osteoclasts and bone structure. *J. Hard Tissue Biol.* **2007**, *16*, 115–123. [CrossRef]
41. Brown, D.K.; Amamuddy, O.S.; Tastan Bishop, Ö. Structure-based analysis of single nucleotide variants in the renin-angiotensinogen complex. *Glob. Heart* **2017**, *12*, 121–132. [CrossRef]
42. Brown, D.K.; Penkler, D.L.; Sheik Amamuddy, O.; Ross, C.; Atilgan, A.R.; Atilgan, C.; Tastan Bishop, Ö. MD-TASK: A software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **2017**, *33*, 2768–2771. [CrossRef]
43. Zerbino, D.R.; Achuthan, P.; Akanni, W.; Amode, M.R.; Barrell, D.; Bhai, J.; Billis, K.; Cummins, C.; Gall, A.; Girón, C.G.; et al. Ensembl 2018. *Nucleic Acids Res.* **2017**, *46*, D754–D761. [CrossRef]
44. Brown, D.K.; Tastan Bishop, Ö. HUMA: A platform for the analysis of genetic variation in humans. *Hum. Mutat.* **2018**, *39*, 40–51. [CrossRef]
45. Capriotti, E.; Calabrese, R.; Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **2006**, *22*, 2729–2734. [CrossRef]
46. Choi, Y.; Sims, G.E.; Murphy, S.; Miller, J.R.; Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **2012**, *7*, e46688. [CrossRef] [PubMed]
47. Shihab, H.A.; Gough, J.; Cooper, D.N.; Stenson, P.D.; Barker, G.L.; Edwards, K.J.; Day, I.N.; Gaunt, T.R. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **2013**, *34*, 57–65. [CrossRef] [PubMed]
48. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nat. Methods* **2010**, *7*, 248. [CrossRef] [PubMed]
49. Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2. 0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **2005**, *33*, W306–W310. [CrossRef]
50. Cheng, J.; Randall, A.; Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Bioinform.* **2006**, *62*, 1125–1132. [CrossRef]
51. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alföldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* **2019**, 531210, doi:10.1101/531210. [CrossRef]
52. Jacobson, M.P.; Friesner, R.A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002**, *320*, 597–608. [CrossRef]
53. Jacobson, M.P.; Pincus, D.L.; Rapp, C.S.; Day, T.J.; Honig, B.; Shaw, D.E.; Friesner, R.A. A hierarchical approach to all-atom protein loop prediction. *Proteins Struct. Funct. Bioinform.* **2004**, *55*, 351–367. [CrossRef]
54. Šali, A.; Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815. [CrossRef]
55. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef]
56. Ross, C.; Knox, C.; Tastan Bishop, Ö. Interacting motif networks located in hotspots associated with RNA release are conserved in Enterovirus capsids. *FEBS Lett.* **2017**, *591*, 1687–1701. [CrossRef] [PubMed]
57. Davey, N.E.; Van Roey, K.; Weatheritt, R.J.; Toedt, G.; Uyar, B.; Altenberg, B.; Budd, A.; Diella, F.; Dinkel, H.; Gibson, T.J. Attributes of short linear motifs. *Mol. Biosyst.* **2012**, *8*, 268–281. [CrossRef] [PubMed]
58. Case, D.; Cerutti, D.; Cheatham, T.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Greene, D.; Homeyer, N.; et al. *Amber 2017*; University of California: San Francisco, USA, 2017; pp. 1–950.
59. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; et al. *Gaussian 09 Revision E.01*; Gaussian Inc.: Wallingford, CT, USA, 2009.
60. Harding, M.M. Small revisions to predicted distances around metal sites in proteins. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2006**, *62*, 678–682. [CrossRef] [PubMed]
61. Bernadat, G.; Supuran, C.T.; Iorga, B.I. Carbonic anhydrase binding site parameterization in OPLS-AA force field. *Bioorgan. Med. Chem.* **2013**, *21*, 1427–1430. [CrossRef]
62. Li, P.; Merz, K.M. MCPB.py: A python based metal center parameter builder. *J. Chem. Inf. Model.* **2016**, *56*, 599–604.10.1021/acs.jcim.5b00674. [CrossRef]

63. Peters, M.B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M.N.; Merz, K.M., Jr. Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *J. Chem. Theory Comput.* **2010**, *6*, 2935–2947. [CrossRef]
64. Penkler, D.L.; Bishop, Ö.T. Modulation of human Hsp90α conformational dynamics by allosteric ligand interaction at the c-terminal domain. *Sci. Rep.* **2019**, *9*, 1600. [CrossRef]
65. Penkler, D.; Atilgan, C.; Tastan Bishop, Ö. Allosteric Modulation of Human Hsp90α Conformational Dynamics. *J. Chem. Inf. Model.* **2018**, *58*, 383–404. [CrossRef]
66. Elder, I.; Tu, C.; Ming, L.J.; McKenna, R.; Silverman, D.N. Proton transfer from exogenous donors in catalysis by human carbonic anhydrase II. *Arch. Biochem. Biophys.* **2005**, *437*, 106–114. [CrossRef]
67. Bhatt, D.; Tu, C.; Fisher, S.Z.; Hernandez Prada, J.A.; McKenna, R.; Silverman, D.N. Proton transfer in a Thr200His mutant of human carbonic anhydrase II. *Proteins Struct. Funct. Bioinform.* **2005**, *61*, 239–245. [CrossRef]
68. Bhatt, D.; Fisher, S.Z.; Tu, C.; McKenna, R.; Silverman, D.N. Location of binding sites in small molecule rescue of human carbonic anhydrase II. *Biophys. J.* **2007**, *92*, 562–570. [CrossRef] [PubMed]
69. An, H.; Tu, C.; Duda, D.; Montanez-Clemente, I.; Math, K.; Laipis, P.J.; McKenna, R.; Silverman, D.N. Chemical rescue in catalysis by human carbonic anhydrases II and III. *Biochemistry* **2002**, *41*, 3235–3242. [CrossRef] [PubMed]
70. Liang, Z.; Verkhivker, G.M.; Hu, G. Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications. *Briefings Bioinform.* **2019**, doi:10.1093/bib/bbz029. [CrossRef] [PubMed]
71. Smith, I.N.; Thacker, S.; Seyfi, M.; Cheng, F.; Eng, C. Conformational Dynamics and Allosteric Regulation Landscapes of Germline PTEN Mutations Associated with Autism Compared to Those Associated with Cancer. *Am. J. Hum. Genet.* **2019**, *104*, 861–878. [CrossRef] [PubMed]
72. Hu, G.; Di Paola, L.; Liang, Z.; Giuliani, A. Comparative study of elastic network model and protein contact network for protein complexes: The hemoglobin case. *BioMed Res. Int.* **2017**, *2017*, 2483264. [CrossRef]
73. Tu, C.; Couton, J.; Van Heeke, G.; Richards, N.; Silverman, D. Kinetic analysis of a mutant (His107–> Tyr) responsible for human carbonic anhydrase II deficiency syndrome. *J. Biol. Chem.* **1993**, *268*, 4775–4779.
74. Almstedt, K.; Mårtensson, L.G.; Carlsson, U.; Hammarström, P. Thermodynamic interrogation of a folding disease. Mutant mapping of position 107 in human carbonic anhydrase II linked to marble brain disease. *Biochemistry* **2008**, *47*, 1288–1298. [CrossRef]
75. Venta, P.; Welty, R.; Johnson, T.; Sly, W.; Tashian, R. Carbonic anhydrase II deficiency syndrome in a Belgian family is caused by a point mutation at an invariant histidine residue (107 His—-Tyr): Complete structure of the normal human CA II gene. *Am. J. Hum. Genet.* **1991**, *49*, 1082.
76. Hurst, T.K.; Wang, D.; Thompson, R.B.; Fierke, C.A. Carbonic anhydrase II-based metal ion sensing: Advances and new perspectives. *Biochim. Biophys. Acta (BBA) Proteins Proteom.* **2010**, *1804*, 393–403. [CrossRef]
77. Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* **2004**, *344*, 1135–1146. [CrossRef]
78. Fisher, S.Z.; Tu, C.; Bhatt, D.; Govindasamy, L.; Agbandje-McKenna, M.; McKenna, R.; Silverman, D.N. Speeding up proton transfer in a fast enzyme: Kinetic and crystallographic studies on the effect of hydrophobic amino acid substitutions in the active site of human carbonic anhydrase II. *Biochemistry* **2007**, *46*, 3803–3813. [CrossRef] [PubMed]
79. Zheng, J.; Avvaru, B.S.; Tu, C.; McKenna, R.; Silverman, D.N. Role of hydrophilic residues in proton transfer during catalysis by human carbonic anhydrase II. *Biochemistry* **2008**, *47*, 12028–12036. [CrossRef] [PubMed]
80. Buonanno, M.; Di Fiore, A.; Langella, E.; D'Ambrosio, K.; Supuran, C.; Monti, S.; De Simone, G. The crystal structure of a hCA VII variant provides insights into the molecular determinants responsible for its catalytic behavior. *Int. J. Mol. Sci.* **2018**, *19*, 1571. [CrossRef] [PubMed]
81. Mikulski, R.L.; Silverman, D.N. Proton transfer in catalysis and the role of proton shuttles in carbonic anhydrase. *Biochim. Biophys. Acta (BBA) Proteins Proteom.* **2010**, *1804*, 422–426. [CrossRef] [PubMed]
82. Jackman, J.E.; Merz, K.M.; Fierke, C.A. Disruption of the active site solvent network in carbonic anhydrase II decreases the efficiency of proton transfer. *Biochemistry* **1996**, *35*, 16421–16428. [CrossRef]
83. Maupin, C.M.; McKenna, R.; Silverman, D.N.; Voth, G.A. Elucidation of the proton transport mechanism in human carbonic anhydrase II. *J. Am. Chem. Soc.* **2009**, *131*, 7598–7608. [CrossRef]

84. Cui, Q.; Karplus, M. Is a "proton wire" concerted or stepwise? A model study of proton transfer in carbonic anhydrase. *J. Phys. Chem. B* **2003**, *107*, 1071–1078. [CrossRef]
85. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919. [CrossRef]
86. Nyamai, D.W.; Tastan Bishop, Ö. Aminoacyl tRNA synthetases as malarial drug targets: a comparative bioinformatics study. *Malar. J.* **2019**, *18*, 34. [CrossRef]
87. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90. [CrossRef]
88. Inkscape. Inkscape: A Vector Drawing Tool. 2004. Available online: http://www.inkscape.org (accessed on 2 June 2019).
89. Rose, P.W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A.R.; Christie, C.H.; Costanzo, L.D.; Duarte, J.M.; Dutta, S.; Feng, Z.; et al. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **2016**, *45*, D271–D281. [PubMed]
90. Behnke, C.A.; Le Trong, I.; Godden, J.W.; Merritt, E.A.; Teller, D.C.; Bajorath, J.; Stenkamp, R.E. Atomic resolution studies of carbonic anhydrase II. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66*, 616–627. [CrossRef] [PubMed]
91. Henderson, L.E.; Henriksson, D.; Nyman, P.O. Primary structure of human carbonic anhydrase C. *J. Biol. Chem.* **1976**, *251*, 5457–5463. [PubMed]
92. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef]
93. Sievers, F.; Higgins, D.G. Clustal Omega, accurate alignment of very large numbers of sequences. In *Multiple Sequence Alignment Methods*; Springer: Berlin, Germany, 2014; pp. 105–116.
94. Brown, D.K.; Tastan Bishop, Ö. Role of structural bioinformatics in drug discovery by computational SNP analysis: Analyzing variation at the protein level. *Glob. Heart* **2017**, *12*, 151–161. [CrossRef]
95. Sherry, S.T.; Ward, M.H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef]
96. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **2017**, *46*, D1062–D1067. [CrossRef]
97. OMIM. Online Mendelian Inheritance in Man. 2018. Available online: https://www.omim.org/entry/611492 (accessed on 23 August 2018).
98. Schrodinger, L. *The PyMOL Molecular Graphics System, Version 1.8*; Schrodinger LLC: New York, NY, USA, 2015.
99. Gordon, J.; B Myers, J.; Folta, T.; Shoja, V.; S Heath, L.; Onufriev, A. H++: A server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33*, W368–W371. [CrossRef]
100. Demir, Y.; Demir, N.; Nadaroglu, H.; Bakan, E. Purification and Characterization of Carbonic Anhydrase from Bovine Erythrocyte Plasma Membrane. *Prep. Biochem. Biotechnol.* **2000**, *30*, 49–59. [CrossRef]
101. Demir, N.; Demir, Y.; Coşkun, F. Purification and characterization of carbonic anhydrase from human erythrocyte plasma membrane. *Turk. J. Med Sci.* **2001**, *31*, 477–482.
102. Schrödinger. *Schrödinger Release 2018-3: Maestro*; Schrödinger LLC: New York, NY, USA, 2018.
103. Olsson, M.H.; Søndergaard, C.R.; Rostkowski, M.; Jensen, J.H. PROPKA3: Consistent treatment of internal and surface residues in empirical pka predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537. [CrossRef] [PubMed]
104. Schafmeister, C.; Ross, W.; Romanovski, V. *LEaP*; University of California: San Francisco, CA, USA, 1995.
105. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713. [CrossRef] [PubMed]
106. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Softwarex* **2015**, *1*, 19–25. [CrossRef]
107. da Silva, A.W.S.; Vranken, W.F. ACPYPE-Antechamber python parser interface. *BMC Res. Notes* **2012**, *5*, 367. [CrossRef]

108. Roe, D.R.; Cheatham, T.E., III. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095. [CrossRef]
109. Humphrey, W.; Dalke, A.; Schulten, K. VMD—Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [CrossRef]
110. RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio, Inc.: Boston, MA, USA, 2015.
111. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
112. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv. (CSUR)* **1999**, *31*, 264–323. [CrossRef]
113. Shao, J.; Tanner, S.W.; Thompson, N.; Cheatham, T.E. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334. [CrossRef]
114. Bouguettaya, A.; Yu, Q.; Liu, X.; Zhou, X.; Song, A. Efficient agglomerative hierarchical clustering. *Expert Syst. Appl.* **2015**, *42*, 2785–2797. [CrossRef]
115. Chakrabarty, B.; Parekh, N. NAPS: Network analysis of protein structures. *Nucleic Acids Res.* **2016**, *44*, W375–W382. [CrossRef]
116. Hagberg, A.; Swart, P.; Chult, D.S. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Technical report; Los Alamos National Lab. (LANL): Los Alamos, NM, USA, 2008.
117. Roe, D.R.; Bergonzo, C.; Cheatham, T.E., III. Evaluation of enhanced sampling provided by accelerated molecular dynamics with Hamiltonian replica exchange methods. *J. Phys. Chem. B* **2014**, *118*, 3543–3552. [CrossRef] [PubMed]

Sample Availability: Not available.