

Mechanism of Action of Non-Synonymous Single Nucleotide Variations Associated with α -Carbonic Anhydrases II, IV and VIII

A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN BIOINFORMATICS
OF
RHODES UNIVERSITY, SOUTH AFRICA

IN THE DEPARTMENT OF
BIOCHEMISTRY AND MICROBIOLOGY
FACULTY OF SCIENCE

BY
T. ALLAN SANYANGA
ORCID: 0000-0001-8744-1957

APRIL 2020

© COPYRIGHT BY T. ALLAN SANYANGA, 2020. ALL RIGHTS RESERVED.

ABSTRACT

The carbonic anhydrase (CA) group of enzymes are Zinc (Zn^{2+}) metalloproteins responsible for the reversible hydration of CO_2 to bicarbonate (HCO_3^-) and protons (H^+) for the facilitation of acid-base balance and homeostasis within the body. Across all organisms, a minimum of six CA families exist, including, α (alpha), β (beta), γ (gamma), δ (delta), η (eta) and ζ (zeta). Some organisms can have more than one family, with exception to humans that contain the α family solely. The α -CA family comprises of 16 isoforms (CA-I to CA-XV) including the CA-VIII, CA-X and CA-XI acatalytic isoforms. Of the catalytic isoforms, CA-II and CA-IV possess one of the fastest rates of reaction, and any disturbances to the function of these enzymes results in CA deficiencies and undesirable phenotypes. CA-II deficiencies result in osteopetrosis with renal tubular acidosis and cerebral calcification, whereas CA-IV deficiencies result in retinitis pigmentosa 17 (RP17). Phenotypic effects generally manifest as a result of poor protein folding and function due to the presence of non-synonymous single nucleotide variations (nsSNVs). Even within the acatalytic isoforms such as CA-VIII that allosterically regulates the affinity of inositol triphosphate (IP_3) for the IP_3 receptor type 1 (ITPR1) and regulates calcium (Ca^{2+}) signalling, the presence of SNVs also causes phenotypes cerebellar ataxia, mental retardation, and dysequilibrium syndrome 3 (CAMRQ3). Currently the majority of research into the CAs is focused on the inhibition of these proteins to achieve therapeutic effects in patients via the control of HCO_3^- production or reabsorption as observed in glaucoma and diuretic medications. Little research has therefore been devoted into the identification of stabilising or activating compound that could rescue protein function in the case of deficiencies.

The main aim of this research was to identify and characterise the effects of nsSNVs on the structure and function of CA-II, CA-IV and CA-VIII to set a foundation for rare disease studies into the CA group of proteins. Combined bioinformatics approaches divided into four main objectives were implemented. These included variant identification, sequence analysis and protein characterisation, force field (FF) parameter generation, molecular dynamics (MD) simulation and dynamic residue network analysis (DRN).

Six variants for each of the CA-II, CA-IV and CA-VIII proteins with pathogenic annotations were identified from the HUMA and Ensembl databases. These included the pathogenic variants K18E, K18Q, H107Y, P236H, P236R and N252D for CA-II. CA-IV included the pathogenic R69H, R219C and R219S, and benign N86K, N177K and V234I variants. CA-VIII included pathogenic S100A, S100P, G162R and R237Q, and benign S100L and E109D variants. CA-II has been more extensively studied than CA-IV and CA-VIII, therefore residues essential to its function and stability are known. To discover important residues and regions within the CA-IV and CA-VIII proteins sequence and motif analysis was performed across the α -CA family, using CA-II as a reference. Sequence analysis identified multiple conserved residues between the two acatalytic CA-II and CA-IV, and the acatalytic CA-VIII isoforms that were proposed to be essential for protein stability. With exception to the benign N86K CA-IV variant, none of the other pathogenic or benign CA-II, CA-IV and CA-VIII SNVs were located at functionally or structurally important residues. Motif analysis identified 11 conserved and important motifs within the α -CA family. Several of the identified variants were located on these motifs including K18E, K18Q, H107Y and N252D (CA-II); N86K, R219C, R219S and V234I (CA-IV); and E109D, G162R and R237Q (CA-VIII). As there were no x-ray crystal structures of the variant proteins, homology modelling was performed to calculate the

protein structures for characterisation. In CA-VIII, the substitution of Ser for Pro at position 100 (variant S100P) resulted in destruction of the β -sheet that the SNV was located on. Little is known about the mechanism of interaction between CA-VIII and ITPR1, and residues involved. SiteMap and CPORT were used to identify binding site amino for CA-VIII and results identified 38 potential residues.

Traditional FFs are incapable of performing MD simulations of metalloproteins. The AMBER ff14SB FF was extended and Zn^{2+} FF parameters calculated to add support for metalloprotein MD simulations. In the protein, Zn^{2+} was noted to have a charge less than +1. Variant effects on protein structure were then investigated using MD simulations. Root mean square deviation (RMSD) and radius of gyration (Rg) results indicated subtle SNV effects to the variant global structure in CA-II and CA-IV. However, with regards to CA-VIII RMSD analysis highlighted that variant presence was associated with increases to the structural rigidity of the protein. Principal component analysis (PCA) in conjunction with free energy analysis was performed to observe variant effects on protein conformational sampling in 3D space. The binding of BCT to CA-II induced greater protein conformational sampling and was associated with higher free energy. In CA-IV and CA-VIII PCA analysis revealed key differences in the mechanism of action of pathogenic and benign SNVs. In CA-IV, wild-type (WT) and benign variant protein structures clustered into single low energy well hinting at the presence of more stable structures. Pathogenic variants were associated with higher free energy and proteins sampled more conformations without settling into a low energy well. PCA analysis of CA-VIII indicated the opposite to CA-IV. Pathogenic variants were clustered into low energy wells, while the WT and benign variants showed greater conformational sampling. Dynamic cross correlation (DCC) analysis was performed using the MD-TASK suite to determine variant effects on residue movement. CA-II WT protein revealed that BCT and CO_2 were associated with anti-correlated and correlated residue movement, highlighting at opposite mechanisms. In CA-IV and CA-VIII variant presence resulted in a change to residue correlation compared to the WT proteins.

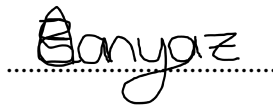
DRN analysis was performed to investigate SNV effects of residue accessibility and communication. Results demonstrated that SNVs are associated with allosteric effects on the CA protein structures, and effects are located on the stability assisting residues of the aromatic clusters and the active site of the proteins. CA-II studies discovered that Glu117 is the most important residue for communication, and variant presence results in a decrease to the usage of the residue. This effect was greatest in the CA-II H107Y SNV, and suggests that variants could have an effect on Zn^{2+} dissociation from the active site. Decreases to the usage of Zn^{2+} coordinating residues were also noted. Where this occurred, compensatory increases to the usage of other primary and secondary coordination residues were observed, that could possibly assist with the maintenance of Zn^{2+} within the active site. The CA-IV variants R69H and R219C highlighted potentially similar pathogenic mechanisms, whereas N86K and N177K hinted at potentially similar benign mechanisms. Within CA-VIII, variant presence was associated with changes to the accessibility of the N-terminal binding site residues. The benign CA-VIII variants highlighted possible compensatory mechanisms, whereby as one group of N-terminal residues loses accessibility, there was an increase to the accessibility of other binding site residues to possibly balance the effect. Catalytically, the proton shuttle residue His64 in CA-II was found to occupy a novel conformation named the “faux in” that brought the imidazole group even closer to the Zn^{2+} compared to the “in” conformation.

Overall, compared to traditional MD simulations the incorporation of DRN allowed more detailed investigations into the variant mechanisms of action. This highlights the importance of network analysis in the study of the effects of missense mutations on the structure and function of proteins. Investigations of diseases at the molecular level is essential in the identification of disease pathogenesis and assists with the development of specifically tailored and better treatment options especially in the cases of genetically associated rare diseases.

Declaration

I TAREMEKEDZWA ALLAN SANYANGA, declare that this is my own unaided work, except where duly acknowledged. It is being submitted for the degree of Doctor of Philosophy in Bioinformatics for the Faculty of Science at Rhodes University. It has not been submitted before for any degree for examination in any other university.

TAREMEKEDZWA ALLAN SANYANGA

A handwritten signature in black ink that reads "Sanyanga". The signature is written in a cursive style and is positioned above a horizontal dotted line.

Date:.....30/04/2020.....

Acknowledgments

SUPERVISOR

Professor Özlem Tastan Bishop for her tireless work and long nights to provide academic and financial support, and for her role as my academic mother and mentor. Thank you for all your faith and patience, without you this PhD would not have been possible.

RESEARCH

Computational resources to perform this research were provided by the Center for High Performance Computing (CHPC), South Africa.

FUNDING

This research and success would not have been possible without the generosity of the National Research Foundation (NRF) of South Africa grant number 105267. The content expressed within this thesis are not a representation of the funders conclusions and opinions.

PERSONAL

“Alice Sanyanga”, for her constant motivation and encouragement to keep pursuing my ambitions dreams. “Dr A and A Hungwe”, and “M Ussi” for their roles as my academic and life mentors. My close friends “Rethabile Mofokeng”, “Fadzayi Hare”, “Cleopas Watama”, “Tendai Chonzi”, and “Tanaka Nyakonda” for the encouragement and inspiration to drive forward, and all those late nights working.

SUPPORT

I would like to thank my family and fellow RUBi research group members, for their advice and assistance in the face of adversity.

Dedication

“EDWIN SANYANGA”, ROLE MODEL, INSPIRATION AND THE MAN WHO HAS ALWAYS LED FROM THE FRONT.

“SHYLET SANYANGA”, UNCONDITIONAL LOVE AND SUPPORT.

Contents

ABSTRACT	iii
DECLARATION	v
ACKNOWLEDGMENTS	vi
DEDICATION	vii
TABLE OF CONTENTS	viii
LISTING OF EQUATIONS	xii
LIST OF FIGURES	xiii
LIST OF TABLES	xvi
LISTING OF SUPPLEMENTARY FIGURES	xviii
LISTING OF SUPPLEMENTARY TABLES	xix
LISTING OF CODE	xx
LISTING OF ABBREVIATIONS	xxi
LISTING OF AMINO ACIDS	xxii
LISTING OF WEB SERVERS	xxiii
RESEARCH OUTPUTS	xxiv
THESIS OVERVIEW	xxv
I LITERATURE REVIEW	1
1.1 Background	1
1.2 Carbonic anhydrases	3
1.2.1 Carbonic anhydrase II (CA-II)	5
1.2.1.1 Deficiencies	7
1.2.1.1.1 Osteopetrosis	7
1.2.1.1.2 Renal tubular acidosis (RTA)	9
1.2.1.1.3 Cerebral calcification	11
1.2.1.2 Clinical implications of CA-II inhibition	11
1.2.1.3 Active site of CA-II	12
1.2.1.4 Mechanism of action of CA-II	14
1.2.2 Carbonic anhydrase IV (CA-IV)	15
1.2.2.1 Deficiencies	16

1.2.2.1.1	Retinitis pigmentosa	16
1.2.2.2	Active site of CA-IV	18
1.2.2.3	Mechanism of action of CA-IV	19
1.2.3	Carbonic anhydrase VIII (CA-VIII)	19
1.2.3.1	Deficiencies	22
1.3	Knowledge Gap	22
1.4	Research Aim	23
2	CHARACTERISATION OF CA-II, CA-IV AND CA-VII, AND PROTEIN VARIANTS	25
2.1	Introduction	26
2.1.1	Protein sequence analysis	27
2.1.1.1	Multiple sequence alignment (MSA)	27
2.1.1.1.1	ClustalΩ	29
2.1.1.1.2	MAFFT	29
2.1.1.2	Motif sequence analysis	30
2.1.1.2.1	Motif database query	30
2.1.1.2.2	Motif identification from sequences	30
2.1.2	Three dimensional structural analysis	32
2.1.3	Protein-protein interaction analysis	32
2.1.3.1	Domain and motif pairs	33
2.1.3.2	Genomic methods	33
2.1.4	Variant identification and characterisation	34
2.1.4.1	ClinVar	35
2.1.4.2	OMIM	35
2.1.4.3	VAPOR	36
2.1.5	Homology modelling	36
2.1.5.1	Template identification	37
2.1.5.2	Multiple sequence alignment	37
2.1.5.3	Model building	37
2.1.5.3.1	MODELLER	37
2.1.5.3.2	Prime	38
2.1.5.4	Validation	38
2.1.6	Binding site identification	39
2.2	Methodology	41
2.2.1	Data Retrieval	41
2.2.1.1	Protein Sequences	41
2.2.1.2	3D protein Structures	41
2.2.1.2.1	CA-II and CA-IV	41
2.2.1.2.2	CA-VIII	42
2.2.2	Protein-protein interaction prediction	43
2.2.3	Motif analysis	43
2.2.4	Homology modelling	44
2.2.4.1	Wild-type	44
2.2.4.2	Variants	45
2.2.4.3	Bicarbonate bound structure	46
2.2.5	Identification of CA-VIII binding site residues	46
2.3	Results and Discussion	48
2.3.1	CA characterisation reveals potential association with other proteins	48

2.3.2	Data retrieval identifies pathogenic and benign CA SNVs	50
2.3.3	Variant 3D spatial location may disrupt protein function and integrity . .	52
2.3.4	Association of CA-II and CA-IV with membrane carriers	55
2.3.5	Potential CA-VIII and ITPR1 association residues identified	55
2.3.6	Structurally important CA-IV and CA-VIII residues identified via sequence analysis	58
2.3.7	SNVs are located around or within conserved motifs	61
2.4	Conclusion	65
3	Zn²⁺ FORCE FIELD PARAMETER GENERATION	67
3.1	Introduction	68
3.1.1	Force fields	68
3.1.2	Protein force fields	68
3.1.2.1	AMBER force field	69
3.1.3	General AMBER force field	71
3.1.4	Extending the AMBER force fields	72
3.1.4.1	Metal Center Parameter Builder	73
3.2	Methodology	78
3.2.1	Protein preparation	78
3.2.2	Zn ²⁺ parametrisation	79
3.3	Results and Discussion	80
3.4	Conclusion	83
4	EFFECTS OF VARIANTS ON THE STRUCTURE AND FUNCTION OF CA-II, CA-IV AND CA-VIII	84
4.1	Introduction to Molecular dynamics	85
4.1.1	Principal component analysis	86
4.1.2	Dynamic cross correlation	87
4.1.3	Dynamic residue networks (DRN)	88
4.1.3.1	Weighted contact maps	89
4.1.3.2	Average shortest path	89
4.1.3.3	<i>Betweenness centrality</i> (<i>BC</i>)	89
4.2	Methodology	91
4.2.1	Protein Preparation	91
4.2.2	Molecular dynamics	93
4.2.3	Molecular dynamics trajectory analysis	93
4.2.3.1	Proton shuttle analysis	94
4.2.4	Dynamic cross correlation (DCC)	94
4.2.5	Dynamic residue network analysis	94
4.2.5.1	Weighted contact map analysis	94
4.2.5.2	Average shortest path (<i>L</i>)	95
4.2.5.3	<i>Betweenness centrality</i> (<i>BC</i>)	95
4.2.6	Principal component analysis (PCA)	95
4.3	Results and Discussion	97
4.3.1	Variant presence is associated with conformational changes to the global structure of CAs	97
4.3.1.1	RMSD analysis	98
4.3.1.1.1	CA-II	98

4.3.1.1.2	CA-IV	100
4.3.1.1.3	CA-VIII	101
4.3.1.2	PCA analysis	102
4.3.1.2.1	CA-II	103
4.3.1.2.2	CA-IV	105
4.3.1.2.3	CA-VIII	107
4.3.1.3	Rg analysis	108
4.3.1.3.1	CA-II	108
4.3.1.3.2	CA-IV	109
4.3.1.3.3	CA-VIII	110
4.3.2	Local residue analysis hints at variant effects to protein structure	111
4.3.2.1	DCC analysis	111
4.3.2.1.1	CA-II	112
4.3.2.1.2	CA-IV	114
4.3.2.1.3	CA-VIII	115
4.3.2.2	RMSF analysis	118
4.3.2.2.1	CA-II	118
4.3.2.2.2	CA-IV	120
4.3.2.2.3	CA-VIII	122
4.3.3	Short range residue interactions are affected by variant presence	123
4.3.3.1	CA-II	124
4.3.3.2	CA-IV	128
4.3.3.3	CA-VIII	130
4.3.4	SNVs are associated with changes to residue accessibility and communication	131
4.3.4.1	Average shortest path	131
4.3.4.1.1	CA-II	132
4.3.4.1.2	CA-IV	135
4.3.4.1.3	CA-VIII	137
4.3.4.2	<i>Betweenness centrality</i> (BC)	138
4.3.4.2.1	CA-II	139
4.3.4.2.2	CA-IV	143
4.3.4.2.3	CA-VIII	145
4.3.5	Variant effects on protein shuttle behaviour	148
4.3.6	Considerations of CA-II, CA-IV and CA-VIII SNVs towards drug discovery	153
4.4	Conclusion	154
5	CONCLUSION	155
	REFERENCES	160
	SUPPLEMENTARY MATERIAL	200

Listing of Equations

1.1	Reversible hydration of CO_2 catalysed by carbonic anhydrase (CA).	3
1.2	CA-II half reaction mechanisms (ping pong) occurring during proton (H^+) shuttling and $\text{CO}_2/\text{HCO}_3^-$ interconversion.	14
2.1	Determination of the correlation between two amino acid sequences. Symbols are, $c(k)$: correlation; $c_v(k)$: correlation of volume component; $c_p(k)$: polarity component.	29
2.2	Calculation of position information within alignment during motif identification.	31
3.1	Relationship between total potential energy, and bonded and nonbonded interactions.	69
3.2	Basic Hamiltonian for potential energy calculations using the basic AMBER force field.	70
3.3	δx displacement of a molecular systems N atoms giving rise to force δF	75
3.4	Cartesian Hessian matrix calculation.	75
3.5	Eigen analysis of k to determine force constants, eigenvalues λ_i and eigenvectors \hat{v}_i	75
3.6	Potential energy function resulting from the relation of force field to internal coordinates.	76
4.1	Newtons second law of motion.	85
4.2	PCA covariance matrix calculation.	87
4.3	Determination of residue dynamic cross correlation.	88
4.4	Calculation of average shortest path.	89
4.5	Determination of the betweenness centrality of a specific residue.	89

List of Figures

1.1	CO ₂ binding pockets of CA-II. A) Secondary pocket; B) Primary pocket; C) Tertiary pocket. Green and blue colours represent hydrophobic and hydrophilic residues. Modified from Sanyanga <i>et al.</i> 2019 [66].	6
1.2	Osteoclast function and role of CA-II in proton (H ⁺) generation during bone resorption. Adapted from Tolar <i>et al.</i> 2004 [75].	8
1.3	Role of carbonic anhydrases in the blood carbonic acid (H ₂ CO ₃) · bicarbonate (HCO ₃ ⁻) buffer system. Modified from Pereira <i>et al.</i> 2009 [88].	10
1.4	Active site of CA-II showing primary and secondary coordination spheres. Dashed linkages between O and H indicate hydrogen bonds. Adapted from Lindskog 1997 [19].	13
1.5	CA-II mechanism of action. Modified from Berg <i>et al.</i> 2002 [103].	15
1.6	CA-IV protein structure and Zn ²⁺ primary coordination sphere.	16
1.7	CA-IV Zn ²⁺ primary and secondary coordination spheres.	19
1.8	CA-VIII protein structure.	20
1.9	Mouse ITPR1 ion channel protein binding domains and corresponding residue number. Adapted from Bosanac <i>et al.</i> 2002 [132].	21
2.1	STRING protein association interaction network of CA-II (CA2), CA-IV (CA4) and CA-VIII (CA8) proteins. Line colour represents the method of interaction prediction.	49
2.2	Illustration of the respective CA-II SNVs, and their proximity to the primary and secondary CO ₂ binding pockets. A) Secondary CO ₂ binding pocket. B) Primary CO ₂ binding pocket. The Zn ²⁺ is represented by the grey sphere. Adapted from Sanyanga <i>et al.</i> 2019 [66].	53
2.3	Illustration of the respective CA-IV SNVs, and their spatial location from the protein active site. The Zn ²⁺ is represented by the grey sphere.	54
2.4	3-dimensional image of CA-VIII showing the SNV location, predicted binding site residues and location of each motif. Orange: S100A, S100L and S100P; Green: E109D; Purple: G162R; Magenta: R237Q. Adapted from Sanyanga and Tastan Bishop 2020 [269].	57
2.5	Motif logo demonstrating amino acid conservation in each of the valid motif sequences, and corresponding motif conservation. Adapted from Sanyanga <i>et al.</i> 2019 [66].	62
2.6	Motif mapping onto the 3D structure of the CA-II, CA-IV and CA-VIII proteins.	63
3.1	Illustration of the relationship between terms in the AMBER Hamiltonian. Different coloured spheres indicate various atoms. Solid and dashed lines represent bonded and nonbonded interactions respectively.	71
3.2	Summary of MCPB workflow of Zn ²⁺ parametrisation and metal ion modelling.	74
3.3	Geometry optimisation of the first Zn ²⁺ coordination sphere during metal ion parametrisation Gaussian QM calculations.	80

3.4	MK-RESP charges calculated for Zn^{2+} and coordinating atoms. M1: Zn; Y1: His94 NE2 (epsilon nitrogen); Y2: His96 NE2 (epsilon nitrogen) Y3: His199 ND1 (delta hydrogen); Y4: O (H_2O); CC: CG (gamma carbon); CR: CE1 (epsilon carbon); CV: CD2 (delta carbon).	82
4.1	Parameter validation presenting bond distances during MD between Zn^{2+} and coordinating atoms, and angles of: M1–Y1–CR (His94); M1–Y2–CR (His96); M1–Y3–CC (His119) and M1–Y4–HW (H_2O) during MD simulation.	97
4.2	RMSD distributions of the WT and variant CA-II protein systems. Average RMSD for each plot is presented as a dashed line on each plot of the corresponding colour.	99
4.3	RMSD distributions of the WT and variant CA-IV and CA-VIII protein systems. Average RMSD for each plot is presented as a dashed line on each plot of the corresponding colour. CA-VIII plot adapted from Sanyanga and Tastan Bishop 2020 [269].	101
4.4	PCA analysis of the WT and variant CA-II proteins and associated free energy for each conformational cluster. The x -axis and y -axis represent the 2D PCA plot. Adapted from Sanyanga <i>et al.</i> 2019 [66].	104
4.5	PCA analysis of the WT and variant CA-IV proteins and associated free energy for each conformational cluster. The x -axis and y -axis represent the 2D PCA plot.	106
4.6	PCA analysis of the WT and variant CA-IV proteins and associated free energy for each conformational cluster. The x -axis and y -axis represent the 2D PCA plot.	107
4.7	Rg distributions of the WT and variant CA-II protein systems. Average Rg for each plot is presented as a dashed line on each plot of the corresponding colour.	109
4.8	Rg distributions of the WT and variant CA-IV and CA-VIII protein systems. Average Rg for each plot is presented as a dashed line on each plot of the corresponding colour. CA-VIII adapted from Sanyanga and Tastan Bishop 2020 [269].	110
4.9	DCC analysis showing residue movement in CA-II. The x -axis and y -axis represent protein residues.	113
4.10	DCC analysis showing residue movement in CA-IV. The x -axis and y -axis represent protein residues.	115
4.11	DCC analysis showing residue movement in CA-VIII. The x -axis and y -axis represent protein residues. Adapted from Sanyanga and Tastan Bishop 2020 [269].	116
4.12	Δ RMSF comparison of the α -carbon atoms of the CA-II WT and variant protein systems (WT – variant). Colour coded bars at the bottom of each plot represent the similar coloured motifs in Figure 2.6.	119
4.13	Δ RMSF comparison of the α -carbon atoms of the CA-IV and CA-VIII WT and variant protein systems (WT – variant). Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6. CA-VIII adapted from Sanyanga and Tastan Bishop 2020 [269].	122
4.14	Contact map weighted interactions of the CA-II WT and SNV residues. Adapted from Sanyanga <i>et al.</i> 2019 [66].	124
4.15	Proportion of the total MD simulation frames the hydrogen bond existed (hydrogen bond fraction) between residue 107 and neighbouring atoms within the WT and variant proteins. Left: WT. Right: H107Y. Adapted from Sanyanga <i>et al.</i> 2019 [66].	126
4.16	Heat map presenting the contact map weighted interactions between the SNV residues in CA-IV and CA-VIII, and respective neighbouring amino acids.	130

4.17	ΔL (WT – variant) comparison of the CA-II protein systems. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6.	132
4.18	ΔL (WT – variant) of the CA-IV and CA-VIII respectively during MD simulation. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6. CA-VIII adapted from Sanyanga and Tastan Bishop 2020 [269].	136
4.19	ΔBC (WT – variant) comparison of the CA-II protein systems. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6.	140
4.20	ΔBC (WT – variant) of the CA-IV and CA-VIII respectively during MD simulation. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6.	144
4.21	Illustration of the His64 proton shuttle “in”, “out” and “faux in” conformations during MD in the CA-II _{apo} protein. A: WT and K18E proteins. Green and magenta colours represent WT and K18E respectively. B: N252D variant. Adapted from Sanyanga <i>et al.</i> 2019 [66].	149
4.22	Contact maps of His64 in the WT and respective CA-II variant apo proteins. Tyr7 has been circled in red.	152

List of Tables

1.1	The α carbonic anhydrase (CA) isoforms, and intracellular location and/or properties (subgroup).	4
2.1	CA-II, CA-IV and CA-VIII identified SNVs and potential variant consequences. . .	52
2.2	Mapping of CA-II residues onto the CA-IV and CA-VIII proteins via MSA (Figure S1) and identified key residue function. Orange represents conserved residues. Adapted from Sanyanga <i>et al.</i> 2019 [66].	59
2.3	Motif residue ranges within the CA-II, CA-IV and CA-VIII proteins, and amino acid sequence. Residues from Table 2.2 are underlined and highlighted in bold. SNVs for each protein are underlined, italicised and presented in bold red. Adapted from Sanyanga <i>et al.</i> [66].	66
3.1	Zn ²⁺ non-bonded, bonds, angles and dihedral parameters derived within this study. K_b : bond force constant; K_θ : angle force constant; R_{min} : vdW radius; ϵ : LJ potential well energy. Adapted from Sanyanga <i>et al.</i> 2019.	81
4.1	Residue renaming for parameter export in the CA-II and CA-IV protein coordination residues. MCPB refers to the metal center parameter builder.	91
4.2	CA-II residues showing significant changes to accessibility during MD simulation. Important CA-II residues from Table 2.2 are highlighted in bold and underlined. SNV positions are underlined, italicised and highlighted in bold red. Adapted from Sanyanga <i>et al.</i> 2019.	134
4.3	CA-IV residues showing changes to ΔL greater or less than two standard deviations.	137
4.4	CA-VIII residues showing changes to ΔL greater or less than two standard deviations. SNV positions are underlined, italicised and highlighted in bold red. Residues located within the CA-VIII binding site are underlined and highlighted in bold blue. Important CA-VIII residues from Table 2.2 are highlighted in bold and underlined. Overlapping potential PPIs and important structural residues are underlined and highlighted in bold green. Adapted from Sanyanga and Tastan Bishop 2020 [269]. . .	138
4.5	CA-II residues showing significant changes to communication/usage during MD simulation. Important CA-II residues from Table 2.2 are highlighted in bold and underlined. SNV positions are underlined, italicised and highlighted in bold red. . .	141
4.6	CA-IV residues showing significant changes to communication/usage during MD simulation. Important CA-IV residues from Table 2.2 are highlighted in bold and underlined.	145
4.7	CA-VIII residues showing significant changes to communication/usage during MD simulation. SNV positions are underlined, italicised and highlighted in bold red. CA-VIII binding site residues are underlined and highlighted in bold blue. Important CA-VIII residues from Table 2.2 are underlined and highlighted in bold. PPI and important structural residues are underlined and highlighted in bold green. Adapted from Sanyanga and Tastan Bishop 2020 [269].	147

4.8 Distance of His64 imidazole group from Zn^{2+} for the “in” and “out” conformations within CA-II. 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. Adapted from Sanyanga *et al.* 2019 [66]. 150

Listing of Supplementary Figures

S1	Multiple sequence alignment of CA-II, CA-IV and CA-VIII. CA-II in bold has been selected as the reference sequence.	203
S2	Conserved motifs within the human α -CA family, and associated proteins and UniProt accessions. Motif conservation is expressed as a heat map representing the number of motif sites per total protein sequences.	204
S3	RMSD of the apo, BCT and CO ₂ bound CA-II protein over the 200 ns MD simulation.	209
S4	RMSD of the CA-IV and CA-VIII proteins over the 200 ns MD simulation.	210
S5	Rg of the apo, BCT and CO ₂ bound CA-II protein over the 200 ns MD simulation.	214
S6	Rg of the CA-IV and CA-VIII proteins over the 200 ns MD simulation.	215
S7	RMSF of the apo, BCT and CO ₂ bound CA-II protein residues over the 200 ns MD simulation.	216
S8	RMSF of the CA-IV and CA-VIII protein residues over the 200 ns MD simulation.	217
S9	Average L for the CA-II protein residues in the apo, BCT and CO ₂ bound states.	218
S10	Average L for the CA-IV and CA-VIII over the MD simulation.	219
S11	Average BC for the CA-II protein residues in the apo, BCT and CO ₂ bound states.	220
S12	Average BC for the CA-IV and CA-VIII over the MD simulation.	221

Listing of Supplementary Tables

S1	Table of UniProt CA accession numbers for final CA family dataset.	200
S2	CA STRING predicted functional partners and confidence (approximate probability) values representing strength of data support.	201
S3	CA-VIII potential protein-protein binding sites and residues. SNV positions are italicised, underlined and highlighted in bold red.	202
S4	Identified human α -CA motifs and associated E-values.	205
S5	Eigenvalue fraction of the various CA-II WT and variant protein states.	211
S6	Eigenvalue fraction of the various CA-IV WT and variant proteins.	212
S7	Eigenvalue fraction of the various CA-VIII WT and variant proteins.	213

Listing of Code

S1	Generated CA-II <i>frmod</i> parameters for molecular dynamics parameter export. . .	206
S2	Example of CA-II LEaP input for MD protein topology preparation.	207
S3	Example of PCA script using to calculate 3D structural differences between the WT and variant proteins.	208

Listing of Abbreviations

Abbreviation	Definition
2D	Two dimensional
3D	Three dimensional
AE1	chloride/bicarbonate exchanger
AMBER	Assisted Model Building with Energy Refinement
<i>BC</i>	<i>Betweenness centrality</i>
BCT	Bicarbonate
BLAST	Basic local alignment search tool
CA	Carbonic anhydrase
CAMRQ3	cerebellar ataxia, mental retardation and disequilibrium syndrome 3
CARP	Carbonic anhydrase related protein
DNA	Deoxyribonucleic acid
DOPE	Discrete optimised protein energy
EM	Expectation Maximization
ER	Endoplasmic reticulum
FF	Force field
GPI	Glycosylphosphatidylinositol
GROMACS	GRONingen MACHine for Chemical Simulations
HUMA	Human mutation analysis platform
IC	Intercalated cell
KDE	Kernel density estimate
<i>L</i>	<i>Shortest path</i>
MD	Molecular dynamics
MM	Molecular mechanics
MSA	Multiple sequence alignment
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
OMIM	Online mendelian inheritance in man
PCA	Principal component analysis
PDF	Probability density function
PPI	Protein-protein interaction
PWM	Position-specific Weight Matrices
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
QM	Quantum mechanics
R _g	Radius of gyration
RP	Retinitis pigmentosa
RTA	Renal tubular acidosis
SNV	Single nucleotide variation
VAPOR	Variant analysis portal
vdW	Van der Waals
WT	Wild-type

Listing of Amino Acids

Single letter code	Three letter code	Amino acid
A	ALA	Alanine
C	CYS	Cystein
D	ASP	Aspartic acid
E	GLU	Glutamic acid
F	PHE	Phenylalanine
G	GLY	Glycine
H	HIS	Histidine
I	ILE	Isoleucine
K	LYS	Lysine
L	LEU	Leucine
M	MET	Methionine
N	ASN	Asparagine
P	PRO	Proline
Q	GLN	Glutamine
R	ARG	Arginine
S	SER	Serine
T	THR	Threonine
V	VAL	Valine
W	TRP	Tryptophan
Y	TYR	Tyrosine

Listing of Web Servers

Webserver	URL
Ensembl	https://www.ensembl.org/index.html
H++	http://biophysics.cs.vt.edu
HUMA	https://huma.rubi.ru.ac.za
MEME	http://meme-suite.org/tools/meme
OMIM	https://omim.org
RCSB	https://www.rcsb.org
VAPOR	https://huma.rubi.ru.ac.za/#vapor

Research Ouputs

RESEARCH ARTICLES

1. Sanyanga, T.A.; Nizami, B.; Tastan Bishop, Ö. Mechanism of Action of Non-Synonymous Single Nucleotide Variations Associated with α -Carbonic Anhydrase II Deficiency. *Molecules* 2019, 24. doi:10.3390/molecules2421398.
2. Sanyanga, T.A.; Tastan Bishop, Ö. Structural Characterization of Carbonic Anhydrase VIII and Effects of Missense Single Nucleotide Variations to Protein Structure and Function. *International Journal of Molecular Sciences* 2020, 21. doi:10.3390/ijms21082764.

CONFERENCE ATTENDANCE

Oral Presentations:

1. Sanyanga, T.A.; Nizami, B.; Tastan Bishop, Ö. Carbonic Anhydrase II: The Effect of Single Nucleotide Polymorphisms on Enzyme Structure and Function. *6th International BAU Drug Design Congress*. 2018. Istanbul, Turkey

Thesis Overview

The main objective of the research was to characterise the effects of non-synonymous single nucleotide variants (nsSNVs) on the structure and function of α -carbonic anhydrases (CAs) II, IV and VIII through the use of a combination of bioinformatics approaches to set the foundation for drug discovery towards rare diseases involving the CA group of proteins. This thesis totals five chapters covering the approaches utilised to achieve the main objective, excluding supplementary information.

CHAPTER 1

This chapter provides an introduction into the CA group of enzymes and describes the features of CA-II, CA-IV and CA-VIII. Included is also a description of the mechanism of action of the catalytic isoforms, and a description of the phenotypes associated with the respective CA deficiencies.

CHAPTER 2

This chapter focuses on the sequence analysis of CA-II, CA-IV and CA-VIII to identify conserved residue regions that are important to the structure and function of the proteins. Identification and characterisation of variants associated with phenotypes in these proteins is also included in this chapter, and the homology modelling of variant proteins.

CHAPTER 3

Chapter 3 describes the generation of Zn^{2+} parameters for the AMBER force field which are necessary to conduct molecular dynamics (MD) simulations on metalloproteins. Force field parametrisation was necessary to prevent metal ion escape during MD.

CHAPTER 4

Expands on chapter 3 and involves implementation of the generated force field parameters to conduct MD simulations using GROMACS, in order to investigate the SNV effects on the global structure of the protein. Traditional MD simulation techniques were also expanded on by the addition of dynamic residue network (DRN) analysis to investigate the effects of SNVs to local protein structure, and to note changes that may occur in the protein network as a result of variant presence.

CHAPTER 5

Summarises the findings from chapters 2–4 and their significance. This chapter also details possible future work.

1

Literature Review

1.1 BACKGROUND

Rare diseases are described as phenotypes affecting a small proportion of the population [1–4]. These conditions are also regarded as orphan diseases due to the limited research conducted on them and the poor treatment options available in their management [5, 6]. The implications resulting from the lack of research into these rare diseases pose a public health issue for respective patients in multiple ways. Firstly, since public health institutions focus to a greater extent on the more common diseases, this suggests that healthcare facilities may not be equipped to handle the orphan diseases [7]. Secondly, as the phenotypes are poorly studied this also results in limited knowledge into the mechanism of disease which in-turn results in poor treatment and management strategies for individuals [7, 8]. Lastly, patients with these diseases live for many years and experience a diminished quality of life, especially

if the disease onset occurs during early childhood [3, 5, 9, 10]. This often results in a greater burden on healthcare institutions and individuals required to provide care for the patients. This also includes the generally high cost of treatment for the management of these conditions [7].

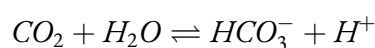
Rare diseases can originate from numerous circumstances. These include exposure to toxins, pathogens, abnormalities within the genome of an individual or disruptions to function of the immune system. The majority of rare diseases are due to genetic changes occurring within individuals, as a result of mutations that either affect individual or multiple genes. These mutations can be inherited from parents to offspring [1, 3, 5, 6, 11]. Multiple inheritance mechanisms for rare diseases exist such as including but not limited to; autosomal (dominant and/or recessive), X-linked and mitochondrial. Autosomal inheritance however forms the greatest distributions [3, 12]. Autosomal dominant inheritance refers to when the disease is inherited from a parent with the phenotype, whereas, autosomal recessive refers to when the disease is inherited from parents that carry the abnormal gene [13].

In addition to mutations causing genetic rare diseases through changes to the genes of individuals, rare diseases can also be the result of genetic polymorphisms occurring within the DNA of an individual. These polymorphisms are divided into two broad categories depending on the frequency within the population. Genetic polymorphisms occurring in greater than 1% of the population are regarded as single nucleotide polymorphisms (SNPs) whereas at a frequency of less than 1% are regarded as single nucleotide variations (SNVs) [14]. Non-synonymous (ns) SNVs are single point substitutions naturally occurring within the nucleotide sequence of a protein [15]. These substitutions alter respective amino acid codons resulting in a different protein sequence, and could potentially change the 3D (3-dimensional) structure of the protein. Structural alterations could then have an effect on protein expression, function, residue-residue interactions and stability, resulting in disease or specific phenotypes in individuals [15–18]. The disease's extent is dependant upon the

function of the gene the nsSNV is located on, therefore, nsSNVs occurring on genes responsible for the maintenance of homeostasis and acid base balance within the body such as in the carbonic anhydrases (CA) could result in the dysregulation of key physiological and biological processes within the body via a domino like effect, and cause CA deficiencies.

1.2 CARBONIC ANHYDRASES

CAs (EC 4.2.1.1) are Zinc (Zn^{2+}) metalloenzymes responsible for the catalysis of the reversible hydration of carbon dioxide (CO_2) and water (H_2O) to bicarbonate (HCO_3^- or BCT) and protons (H^+). The reaction mechanism is illustrated in Equation 1.1 [19, 20]. This interconversion is essential for the maintenance of acid-base balance and homeostasis within the body [19, 21]. In addition, reaction substrates and products are also required by cells to conduct numerous biological processes such as but not limited to; cerebrospinal fluid formation, signal transduction, calcification, oncogenesis and respiration [22–26]. In the absence of CA, Equation 1.1 (forward direction) proceeds slowly at a rate constant ranging from 0.03 – $0.15\ s^{-1}$.



Equation 1.1. Reversible hydration of CO_2 catalysed by carbonic anhydrase (CA).

At least six distinct CA families have been identified to date. These include α (alpha), β (beta), γ (gamma), δ (delta), η (eta) and ζ (zeta). Of these families, vertebrates possess mainly the α family [27]. The α -CAs can however also be found in bacteria [28]. The β -CAs are located in vascular plants and numerous organisms with the exception to vertebrates and chordates [29–31]; γ -CAs are common to methanogenic archaea and plants, and are also present in numerous bacteria [30, 31]; δ -CAs are located in diatoms and some marine algae [32]; The η -CAs are members of *Plasmodium* [33]; and

ζ -CAs are found in diatoms and in some bacteria [34]. Although CAs are ubiquitous, these enzyme families do not contain significant amino acid sequence similarity, and therefore represent convergent evolution [27, 35]. Although the CAs contain Zn^{2+} the γ and ζ families are capable of using other metal ions as cofactors. The γ -CAs can use either cobalt (Co^{2+}) or iron (Fe^{2+}) as cofactors, whereas the ζ -CAs are capable of using cadmium (Cd^{2+}) as a cofactor [29, 34, 36, 37]. These CA cofactors are maintained within the active site of the protein through three coordinating His residues.

The α -CA family can be further subdivided into five subgroups (depending on intracellular location and protein properties) comprising of several isoforms [19, 27, 38, 39]. These subgroups include cytosolic, glycosylphosphatidylinositol (GPI) anchored, mitochondrial, transmembrane associated and secreted. All subgroups total 16 isoforms and these are presented in Table 1.1. The three cytosolic isoforms presented in Table 1.1 namely; CA-VIII, X and XI are acatalytic, and are therefore regarded as the carbonic anhydrase related proteins (CARP) [27, 39]. Their lack of activity is as a result of at least one missing His residue necessary to maintain the Zn^{2+} cofactor within the active site [40].

Table 1.1. The α carbonic anhydrase (CA) isoforms, and intracellular location and/or properties (subgroup).

Isoforms	Subgroup	Source
CA-I; CA-II; CA-III; CA-VII; CA-VIII; CA-X; CA-XI; CA-XIII	Cytosolic	[19, 27, 29, 39]
CA-VA; CA-VB	Mitochondria	[41, 42]
CA-IV; CA-XV*	GPI anchored	[40, 43, 44]
CA-VI	Secreted	[40, 45, 46]
CA-IX; CA-XII; CA-XIV	Transmembrane associated	[27, 47]

*Not expressed in humans.

Though the CAs are expressed in multiple organisms, this thesis will focus on the three human CA isoforms CA-II, CA-IV and CA-VIII which are explained in further detail within the following sections.

1.2.1 Carbonic Anhydrase II (CA-II)

CA-II (previously known as carbonic anhydrase C and carbonic anhydrase B) is a 260 amino acid cytosolic protein that possess the greatest rate of reaction compared to the other α -CAs. The enzyme is capable of hydrating CO_2 at a rate constant of $1 \times 10^6 \text{ s}^{-1}$ [48]. This is a large difference from the reaction rate of the uncatalysed reaction indicating the importance of this enzyme to the body. The CA-II active site Zn^{2+} exists in tetrahedral coordination geometry with three His residues; His94, His96 and His119, and a water molecule [19, 27, 29, 40, 49, 50]. These coordinating residues are also known as coordination ligands. During catalysis His64 transports H^+ to and from the active site by alternating between two conformations (“in” and “out”) [50–53] or through His ring tautomerisation [54]. This action is necessary in order to reduce the distance to the active site exit and allow efficient H^+ shuttling. Water molecules are present within the active site that also assist with the H^+ shuttling by forming a network of water molecules to transport the H^+ [51].

To further maintain pH homeostasis, CA-II has been shown to form metabolon complexes with the, sodium/hydrogen (Na^+/H^+) exchanger (NHE1) [55, 56], chloride/bicarbonate ($\text{Cl}^-/\text{HCO}_3^-$) exchanger (AE1/SLC4A1) [57–59] and sodium bicarbonate ($\text{Na}^+/\text{HCO}_3^-$) cotransporter (NBC1) [60, 61]. The metabolon complexes increase the flux of ions across membranes. Unusually, however, the CA-II and AE1 metabolon complex has been reported to have minimal impact on ion flux [58].

The CA-II enzyme contains three CO_2 binding pockets termed the primary, secondary and tertiary that are located approximately 3–4 (primary pocket), 5–7 (tertiary pocket) and 10–12 Å (secondary pocket) away from the Zn^{2+} [62–65]. These sites are presented in Figure 1.1. Protein residues that form each pocket include; Val121, Val142, Leu197 and Trp208 for the primary pocket; Phe66, Phe95, Trp97, and Phe225 for the secondary pocket; and Trp7, His64, Thr199, Pro200 and Asn243 for the tertiary pocket. The primary and tertiary pockets are both catalytic, however, the secondary pocket’s function is still yet to be fully understood [19, 62–65]. The tertiary pocket lies along a tunnel

terminating at the primary pocket, and is believed to function as a potential CO₂ reservoir for the primary pocket [62].

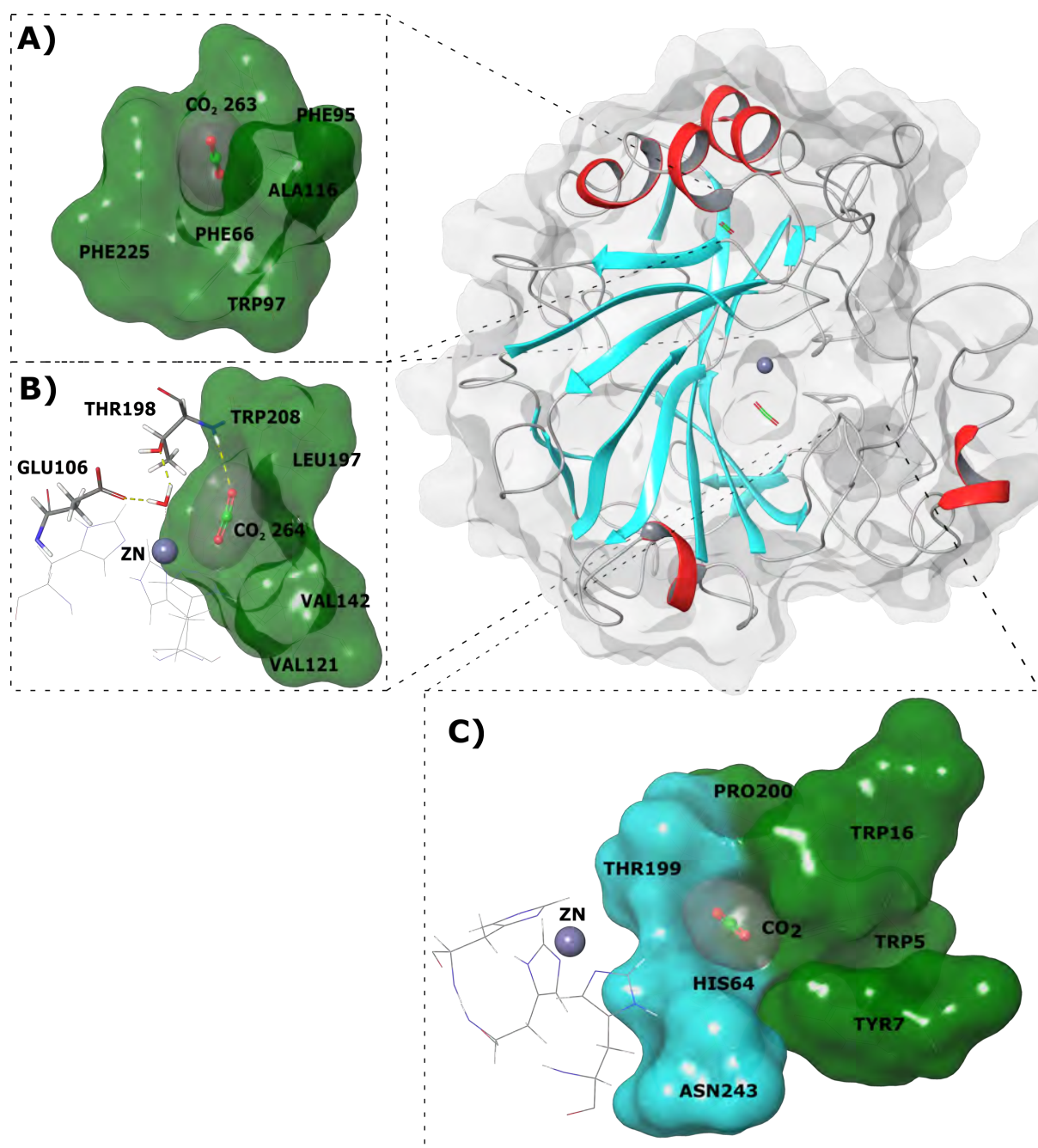


Figure 1.1. CO₂ binding pockets of CA-II. A) Secondary pocket; B) Primary pocket; C) Tertiary pocket. Green and blue colours represent hydrophobic and hydrophilic residues. Modified from Sanyanga *et al.* 2019 [66].

The CA-II structure also comprises of two aromatic clusters also termed as primary and secondary [19]. These two aromatic clusters are responsible for the maintenance of protein stability. The primary aromatic cluster merges the CA-II N-terminal to the rest of protein and comprises of the

residues; Trp5, Tyr7, Trp16 and Phe20 [19]. The secondary aromatic cluster comprises of the residues; Phe66, Phe70, Phe93, Phe95, Trp97, Phe175, Phe178 and Phe225 and is larger than the primary aromatic cluster [19, 67, 68]. It is notable that some of the primary aromatic cluster residues also form the tertiary CO₂ binding pocket, and some members of the secondary aromatic cluster also assist with formation of the secondary binding pocket (Figure 1.1).

Noting the role and function of CA-II in biological systems, the following section describes conditions associated with CA-II deficiencies and their impact on these systems.

1.2.1.1 Deficiencies

1.2.1.1.1 *Osteopetrosis*

CA-II is expressed in numerous cells and tissues within the body. The protein is expressed at high levels during bone resorption that is carried out by osteoclasts [69]. A CA-II deficiency within this process results in the development of the rare conditions osteopetrosis with renal tubular acidosis (RTA) and cerebral calcification (MIM No: 259730) [70].

Osteopetrosis (“marble bone disease”) is a genetic disorder that results in an increase to bone density and mass caused by failures to bone resorption [71]. Bone density is controlled by a balance between the function of osteoblasts that synthesise bone, and osteoclasts that break down bone. Poor osteoclast function disturbs this balance, leading to unregulated bone synthesis thereby increasing bone density. The increase to density makes the bones more brittle as opposed to harder, resulting in osteomyelitis and bone fractures. In addition to bone defects, patients also present with optic nerve compression resulting in vision impairment, mental retardation, short stature and dental malocclusion, and could also present with RTA [72].

During bone resorption H⁺ generated by CA-II is transported into the osteoclast resorption cavity by a vacuolar H⁺–ATPase pump resulting in the lowering of the extracellular space’s (adjacent

to bone) pH to within a range of 4–5 [73]. When the extracellular environment is acidified the hydroxyapatite ($\text{Ca}_5(\text{PO}_4)_3(\text{OH})$) mineral component of bone is degraded to inorganic phosphate (HPO_4^{2-}), H_2O , calcium (Ca^{2+}) [74, 75]. The intracellular function of CA-II, and mechanism of bone resorption by osteoclasts is summarised in Figure 1.2. The $\alpha_v\beta_3$ integrin contained in podosomes facilitates the attachment of the osteoclast to the bone. CA-II hydrates CO_2 to produce H^+ that is necessary to acidify the resorption cavity, and initiate bone demineralisation. The vacuolar H^+ –ATPase transports H^+ across the membrane into the resorption cavity. Cathepsin K is then responsible for the removal of the bone organic matrix.

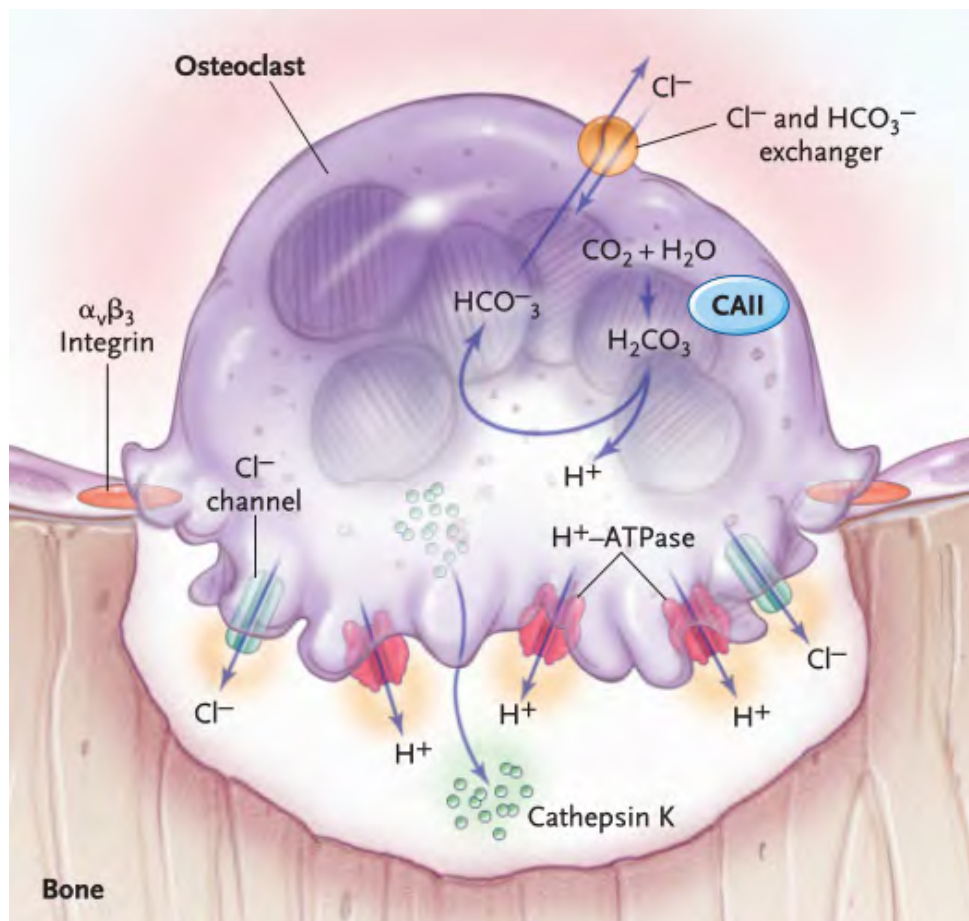


Figure 1.2. Osteoclast function and role of CA-II in proton (H^+) generation during bone resorption. Adapted from Tolar *et al.* 2004 [75].

Osteopetrosis can be inherited as either an autosomal recessive or autosomal dominant disorder. Autosomal recessive osteopetrosis is the more severe of the two, and affects infants a few months after

birth. Currently the only cure for this condition involves hematopoietic stem cell transplantation (HSCT), however this does not work for all cases [76]. Gene therapy is however being investigated as an alternative to HSCT [76]. Generally without treatment, autosomal recessive osteopetrosis results in fatalities. In infants, a diminished life expectancy can also result. Autosomal dominant osteopetrosis is the more mild condition manifesting in adults [72, 77–79]. This form of osteopetrosis is also the most common of the two [78]. Although both conditions are rare, autosomal recessive and dominant osteopetrosis occur at a rate of 1 in 250 000 and 1 in 20 000 of the population respectively [71, 76, 80]. Due to the malignancy of autosomal recessive osteopetrosis, its incidence could be higher as a result of numerous deaths within populations.

CA-II deficiencies are associated with autosomal recessive osteopetrosis as a result of a homozygous or compound heterozygous mutations within the CA-II gene on chromosome 8q21 [72] in the form of nsSNVs. One such mutation that has been linked to osteopetrosis with RTA and cerebral calcification is H107Y, though the exact mechanism of pathogenesis is not known [81–85]. Mutations/variations to CA-II could inhibit the enzyme's ability to generate the H^+ required to acidify the osteoclast resorption cavity. This in-turn would inhibit bone demineralisation leading to osteopetrosis. Variation effects causing the deficiency could either be stability and/or function altering.

1.2.1.1.2 Renal Tubular Acidosis (RTA)

The kidney intercalated cells (ICs); α and β , express high levels of CA-II. However the collecting duct principal cells, proximal tubules and loop of Henle show low expression levels [27, 86, 87]. RTA results from decreases to blood pH occurring at a greater extent than the body can buffer. The acidosis can be caused by different physiological effects within the body, and currently three main types of RTA have been discovered; distal (type 1), proximal (type 2) and a combination of both distal and proximal (type 3) RTA. Distal RTA occurs as a result of the failure by ICs to secrete H^+ into the lumen resulting

in ion build-up and acidosis. Proximal RTA (RTA type 2) is caused by the inability of cells to reabsorb HCO_3^- into the blood which in turn disturbs the HCO_3^- buffer system and results in acidosis [21, 88, 89].

To maintain the carbonic acid (H_2CO_3) · HCO_3^- buffer system in blood, H_2CO_3 within the kidney lumen is converted to CO_2 and H_2O through the action of CA-IV. CO_2 then freely diffuses across the membrane into the cell where CA-II hydrates it to form HCO_3^- and H^+ . The H^+ is then transported out of the cell by the H^+ – ATPase pump, whereas HCO_3^- is transported to the basolateral space then the blood by AE1 [88]. This process is illustrated in Figure 1.3. The ratio of HCO_3^- to H_2CO_3 necessary to maintain the pH of blood is 20:1 (HCO_3^- : H_2CO_3).

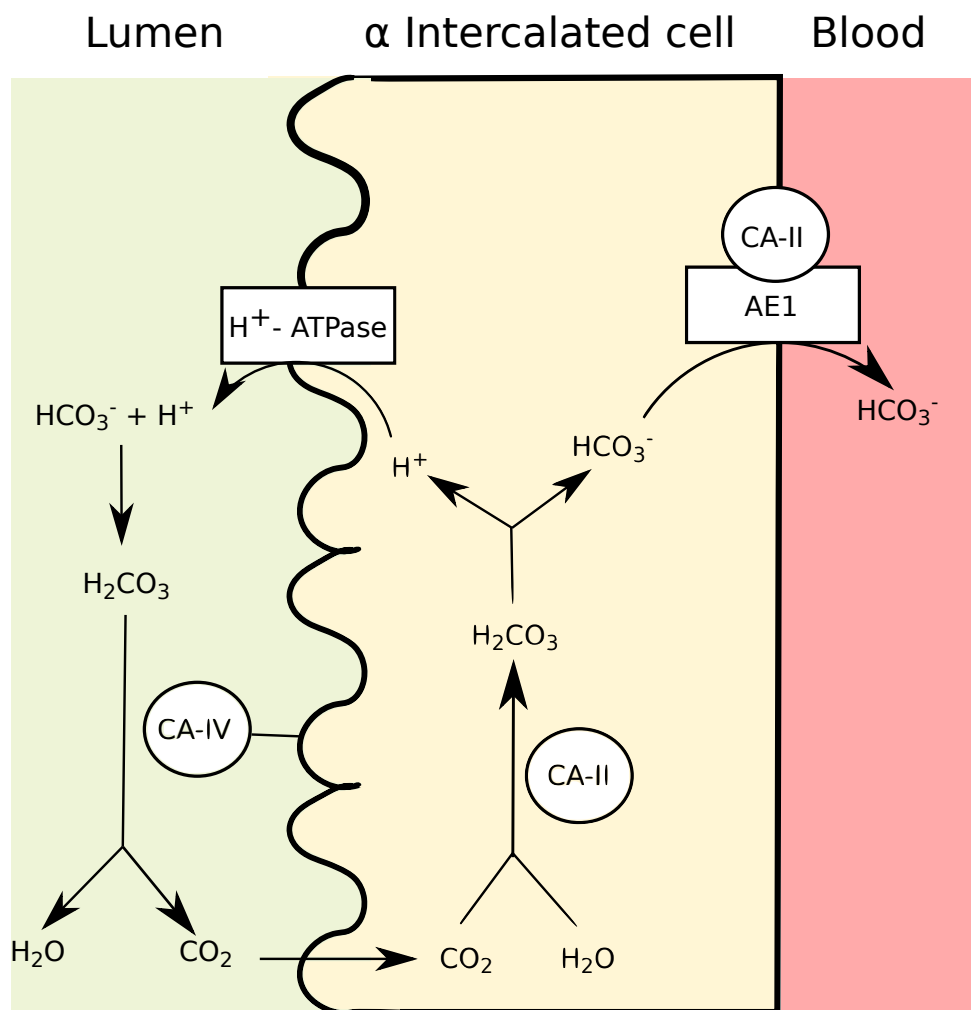


Figure 1.3. Role of carbonic anhydrases in the blood carbonic acid (H_2CO_3) · bicarbonate (HCO_3^-) buffer system. Modified from Pereira *et al.* 2009 [88].

CA-II associated RTA is inherited as an autosomal recessive phenotype, and presents as RTA type 3. The phenotype generally manifests along with osteopetrosis and cerebral calcification [72, 79, 88, 90, 91]. Though the exact mechanism is not known, poor CA-II function could result in the inability to generate H^+ and/or HCO_3^- that can either be absorbed or secreted. These ions are both products of forward reaction of Equation 1.1. CA-II associated RTA usually presents with osteopetrosis therefore SNV H107Y again has been linked to this disease.

1.2.1.1.3 Cerebral Calcification

Cerebral calcification is defined by anomalous Ca^{2+} deposits in the blood vessels of the cerebral cortex and the basal ganglia of the brain [92, 93]. As a result of the Ca^{2+} deposits, the condition can be characterised by mental retardation, dementia, slurred speech and seizures [92]. The ion deposition within blood vessels could be as a result of RTA whereby a lower blood pH could result in Ca^{2+} precipitation.

1.2.1.2 Clinical Implications of CA-II Inhibition

Most CA-II disorders such as glaucoma and altitude sickness are associated with enzyme overexpression [94–97]. The inhibition of CA-II has also found use in diuretics [97, 98]. As a result, CA-II drug discovery has been focused on the inhibition of enzymatic activity presenting a research gap in CA-II studies, whereby there is insufficient research being conducted on CA-II structural and functional rescue. In addition, the prolonged used of CA inhibitors such as acetazolamide has been found inhibit osteoclast function [99, 100] and could potentially lead to osteopetrosis. It has also not been well documented as to what effect CA inhibitors could have on the function of other CA proteins (i.e. CA-IV or CA-VIII) within the body. As a result of potential mutations having varying effects on cellular functions, inhibitors could have non-uniform and varying efficacies and cause undesirable effects. This presents a knowledge gap for CA drug discovery studies. No effective

osteopetrosis treatment currently exists and in some instances symptoms can only be managed. This is usually achieved through nutrient supplementation, transfusions or transplantations [76].

1.2.1.3 Active Site of CA-II

The Zn^{2+} cofactor within the active site of CA-II exists in a tetrahedral coordination geometry. This geometry is illustrated in Figure 1.4 and shows both the primary and secondary coordination spheres of CA-II. The primary coordination residues are directly involved with the maintenance of Zn^{2+} within the active site of the protein. The secondary (indirect) ligands exert an influence of the primary ligands that also assists in the stabilisation of Zn^{2+} within the active site [19]. Previous research has shown that Gln92 and Glu117 have an impact on the rate of dissociation of Zn^{2+} from the active site [68, 101, 102]. The Thr198 (Thr199 using historic numbering) residue assists with stabilisation of the Zn^{2+} bound hydroxide (OH^-). It should be noted that majority of the CA-II structures that have been crystallised all skip residue 126. Within this thesis these structures have been renumbered to include residue 126. Therefore residue numbers 127 and higher have been reduced by one. Thus as opposed to the residue numbering ending at 261, they end at 260.

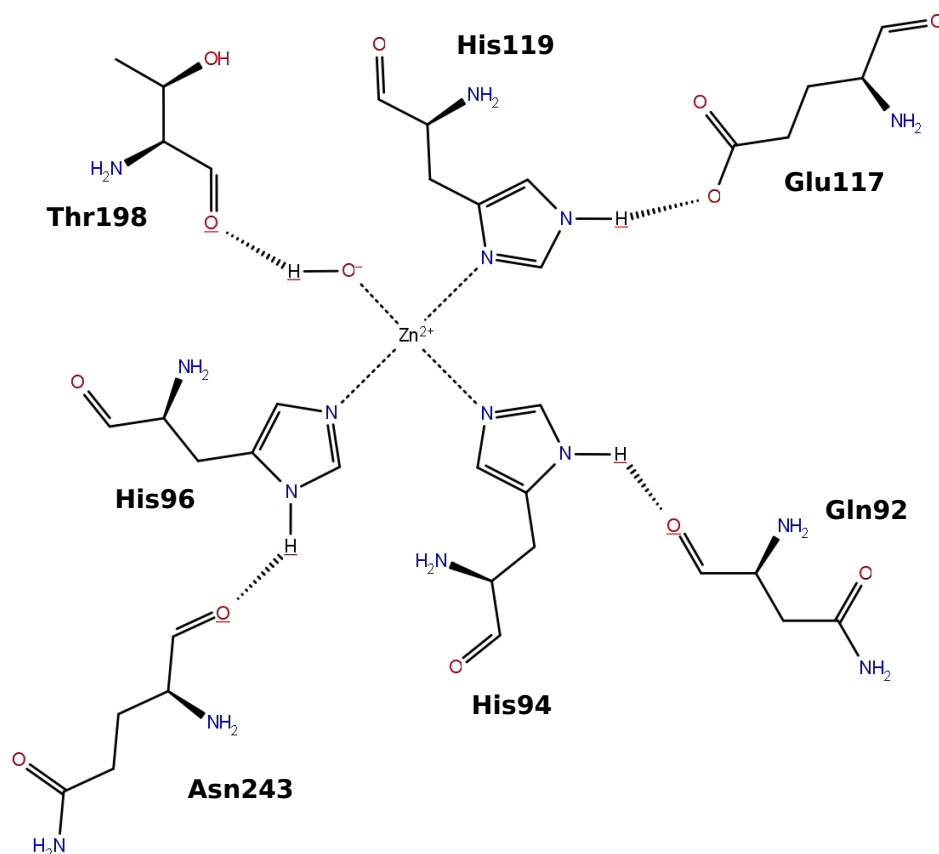


Figure 1.4. Active site of CA-II showing primary and secondary coordination spheres. Dashed linkages between O and H indicate hydrogen bonds. Adapted from Lindskog 1997 [19].

From Figure 1.4 it is noted that although three His coordinate Zn^{2+} , the coordinating atoms and manner of coordination are different. His94 and His96 stabilise the cofactor through interactions with their NE2 (epsilon nitrogen) atoms, whereas His119 stabilises the Zn^{2+} through interactions with the ND1 (delta nitrogen) atom. This orientation has implications for protein coordination. His residues are usually capable of occupying three protonation states:

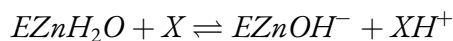
1. HID: His delta nitrogen is protonated.
2. HIE: His epsilon nitrogen is protonated.
3. HIP: Both the delta and epsilon His are protonated.

Inspection of the coordinating atoms (Figure 1.4) and the listed protonation states indicates that to maintain Zn^{2+} stability within the active site, the NE2 atoms of His94 and His96, and the ND1 atom of His119 should never be protonated to avoid repulsion effects with the cofactor. This therefore

disqualifies a HIP protonation state which is positively charged, and indicates that His94 and His96 occupy HID protonation states, whereas His119 occupies a HIE protonation state.

1.2.1.4 Mechanism of Action of CA-II

The interconversion of CO_2 and H_2O to HCO_3^- to H^+ and HCO_3^- across all catalytic CA isoforms proceeds by similar reaction mechanism as that for CA-II which is illustrated in Equation 1.2 and Figure 1.5 [19, 38]. The reaction is facilitated by the function of two fundamental ionisable groups within the enzyme, the Zn^{2+} bound H_2O that can be deprotonated and ionised to a hydroxide (OH^-). The metal hydroxide is the catalytically active form [29].



Equation 1.2. CA-II half reaction mechanisms (ping pong) occurring during proton (H^+) shuttling and $\text{CO}_2/\text{HCO}_3^-$ interconversion.

The four stages in Figure 1.5 occur as a result of a two step ping pong reaction mechanism [27]. The ping pong mechanism indicates that H^+ shuttling and the interconversion of $\text{CO}_2/\text{HCO}_3^-$ occur as two separate half reactions as illustrated in Equation 1.2. The first equation shows the interconversion of $\text{CO}_2/\text{HCO}_3^-$, whereas the second equation represents the H^+ shuttling.

Although the reaction in Figure 1.5 focuses on Zn^{2+} and the bound $\text{H}_2\text{O}/\text{OH}^-$, numerous other CA-II residues are important for this process. Initially CO_2 enters the active site for catalysis followed by the nucleophilic attack on the substrate by the Zn^{2+} bound OH^- . This then results in the formation of an intermediate bond with Zn^{2+} . HCO_3^- then dissociates from the active and is replaced by H_2O . The Zn^{2+} bound OH^- is then regenerated and the H^+ shuttled out of the active site by the

proton shuttle (X). These stages correspond to Equation 1.2 above.

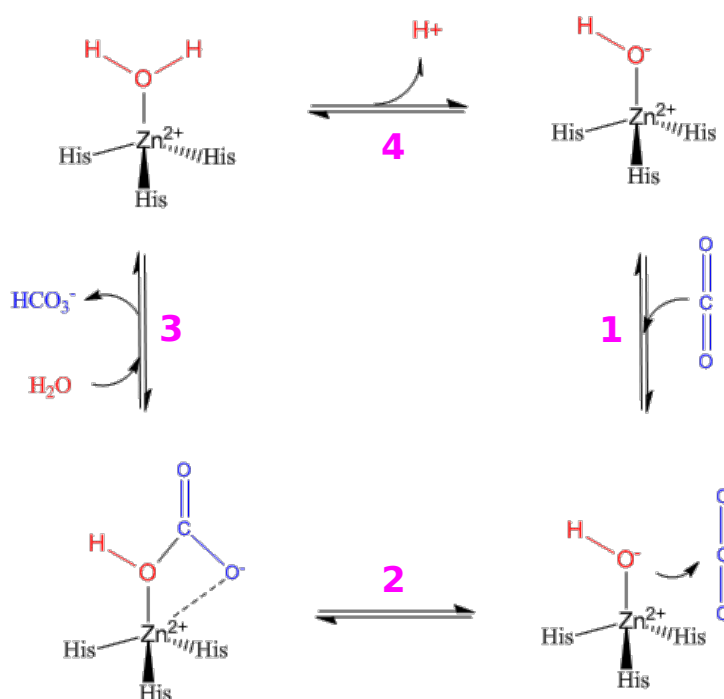


Figure 1.5. CA-II mechanism of action. Modified from Berg *et al.* 2002 [103].

1.2.2 Carbonic Anhydrase IV (CA-IV)

CA-IV is a membrane bound and catalytic enzyme that is localised to the plasma membrane. The enzyme is a 312 amino acid protein that is synthesised within the endoplasmic reticulum (ER) of cells [40]. The Zn^{2+} is coordinated by three His residues; His115, His117 and His140 [40], and one water molecule creating a tetrahedral coordination geometry. The CA-IV enzyme structure is presented in Figure 1.6. This enzyme also shows similar catalytic regions to CA-II [27]. A GPI anchor is responsible for CA-IV membrane attachment with Ser284 serving as the omega-site [19, 104]. Signal peptidases also cleave off the first 20 N-terminal amino acids within the ER prior to the protein reaching its final destination. Although CA-II is regarded as the fastest CA with regards to the hydration of CO_2 , when dehydrating HCO_3^- (Equation 1.1 reverse reaction) CA-IV exhibits faster catalysis [27, 105].

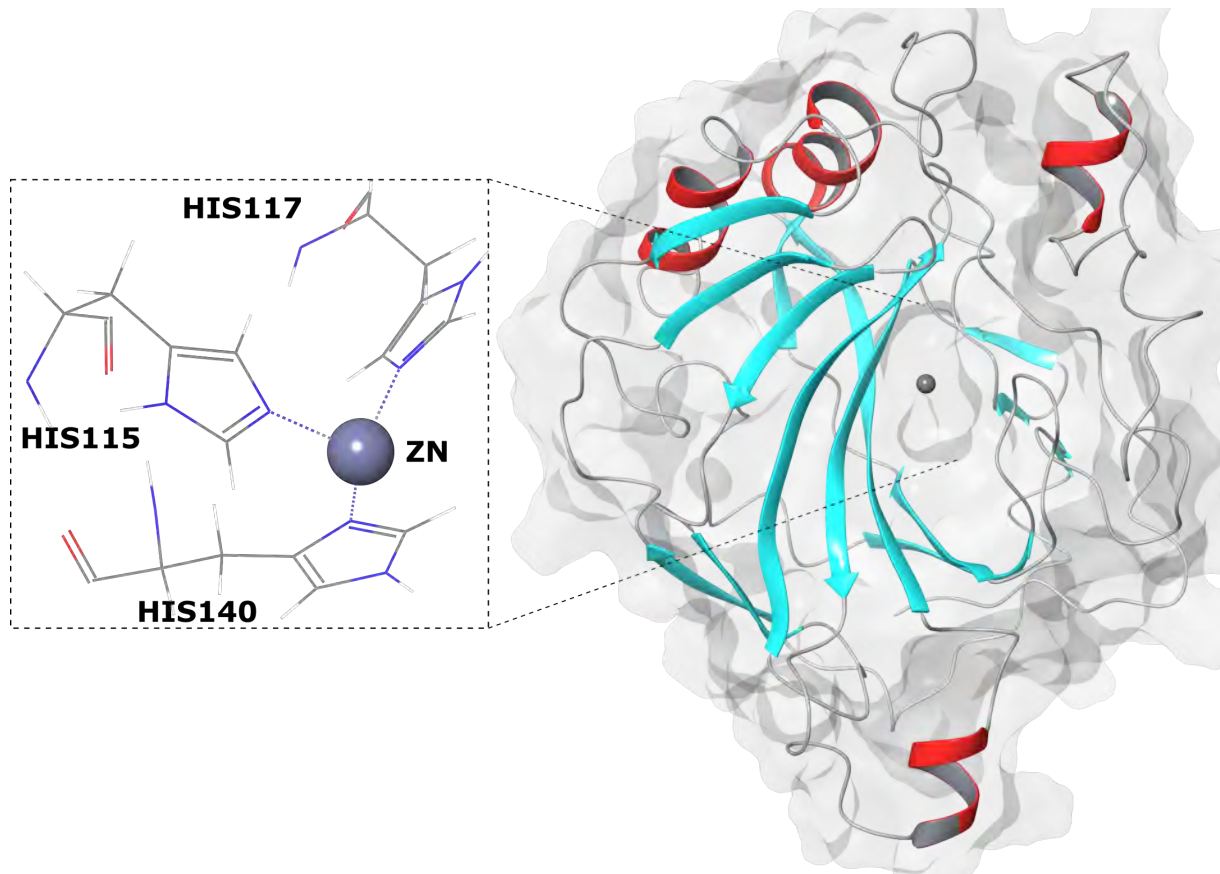


Figure 1.6. CA-IV protein structure and Zn^{2+} primary coordination sphere.

CA-IV is located in various organs within the body, including but not limited to kidneys, heart, brain and lungs, and is critically important for normal ocular function within individuals evidenced by phenotypic eye disorders observed during CA-IV deficiencies [19, 27, 104]. As observed in CA-II, CA-VI is also capable of forming transport metabolons with other membrane proteins to enhance the flux of ions. These membrane carriers include; chloride/bicarbonate (Cl^-/HCO_3^-) exchanger (AE1) and the sodium bicarbonate cotransporter (NBC1) [106–108].

1.2.2.1 Deficiencies

1.2.2.1.1 *Retinitis Pigmentosa*

Within the eye, CA-IV is important for the removal of acid from the retina, retinal epithelium and choroid choriocapillaris [107], and failure to do so results in the hereditary phenotype retinitis pigmentosa (RP) (MIM No: 600852). RP is a rare degenerative ocular disorder that results in the

irreversible worsening of vision within individuals, and is capable of affecting both eyes [109, 110]. The disease is characterised by advancing damage to the retinal neuroepithelium, and can be classified as either neuro-retinal or chorioretinal degradation. In addition to the neuroepithelium, the retinal pigment epithelium layer is also damaged, and is the initial site of retinal degeneration [109]. When suffering from RP, individuals present with a constricted field of view, reduction to peripheral vision and/or night blindness as a result of photoreceptor death [109, 111]. Although RP is normally localised to the eye, non-ocular associations have been identified in numerous other syndromes [112].

RP is inherited from parents as either an autosomal recessive (50–60% of cases), autosomal dominant (30–40% of cases) or X-linked (5–15% of cases) disorder [112–117], and is associated with more than 45 gene loci including CA-IV [112]. RP has a prevalence of 1 in 4000 within the population [112]. Autosomal-recessive RP can be subdivided into two main stages of disease onset, early and late. The early onset RP occurs during individuals late teens and contains largest number of cases. It involves accelerated changes to retinal function followed by changes to retinal morphology. The observable signs include cataract formation (39–72% of patients) and/or macular degeneration [109, 111]. The late onset RP affects retinal function to a lower extent than early onset. RP progresses slower within patients and results in minor vision loss. These cases are however much rarer compared to the early onset cases. The autosomal-dominant RP progresses at slow rate but however results in the severe damage to the retina [109, 111]. The condition manifests during an individuals early to late teens. The X-linked form of RP is related to the sex of an individual and in terms of severity is between autosomal recessive and dominant [109, 111]. Males tend to manifest greater changes to the retina compared to females due to the single X chromosome since there is no additional X chromosome to offset the phenotype.

The autosomal dominant form of RP is associated with CA-IV and is described as RP17 (retinitis pigmentosa 17). RP17 is inherited as a result of heterozygous mutations on chromosome 17q23

which contains the CA-IV gene [118]. Previous association studies between RP and CA-IV identified nsSNVs as being responsible for causing RP17, and examples of these variants include; R14W, R69H and R219S [119]. In a study, R14W was found to induce apoptosis within the choriocapillaris of endothelial cells. In addition, surface cell trafficking of CA-IV from the ER was reported to be inhibited by R69H and R219S causing anomalous intracellular retention, and resulting in cell death via apoptosis [108, 119]. This imparts that, although CA-IV assists in pH maintenance, RP17 could be due to apoptosis induced by ER stress as a result of a toxic gain of function caused by protein misfolding [119]. This hypothesis is further supported by the of lack of retinal abnormalities observed in CA-IV knockout mice [120].

Though numerous researches have been conducted on the disease there is no cure to the condition, however CA-IV inhibitors have been found to slow progression. Nutrient supplementation using vitamin A also assists in slowing disease progression and visual aid help manage vision deterioration [121]. The lack of a cure or additional treatment methods highlights a potential research gap within the study of RP17.

1.2.2.2 Active Site of CA-IV

Structurally the active site of CA-IV is similar to that of CA-II [19, 27]. The Zn^{2+} cofactor exhibits tetrahedral coordination geometry and this is illustrated in Figure 1.7. Comparing to CA-II, the primary and secondary coordination sphere residues are completely conserved in both proteins and share similar protonation states. As observed in CA-II the Zn^{2+} secondary ligands have an effect on cofactor dissociation within the active site [19, 27, 68, 101].

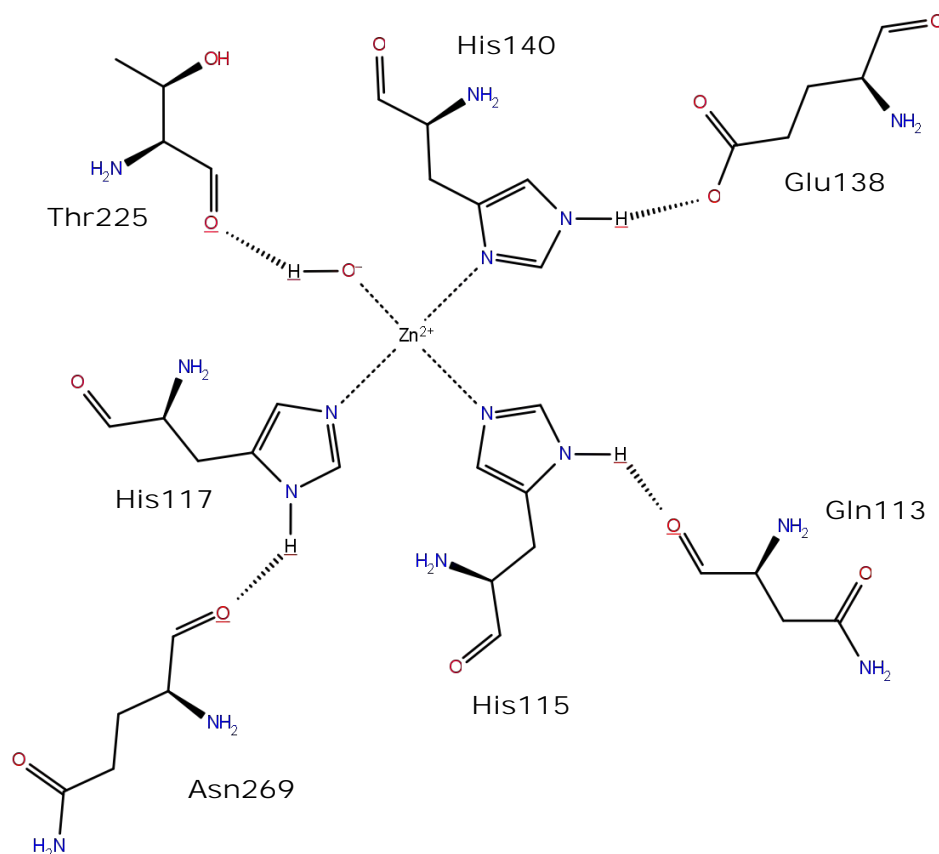


Figure 1.7. CA-IV Zn^{2+} primary and secondary coordination spheres.

1.2.2.3 Mechanism of Action of CA-IV

During catalysis the mechanism of action of CA-IV proceeds in a similar manner as to that of CA-II as illustrated in Equation 1.2 and Figure 1.5, whereby CO_2 is initially hydrated followed by H^+ shuttling from the active site by its respective proton shuttle His88 [19, 27, 39].

1.2.3 Carbonic Anhydrase VIII (CA-VIII)

CA-VIII is an acatalytic member of the cytosolic CA subgroup that comprises of 290 amino acids. The protein has an Arg substitution as opposed to a His at position 116 and therefore lacks CO_2 hydration activity, as Zn^{2+} cannot be coordinated and maintained within the active site [40]. Across all catalytic α -CAs, His116 (His94 in CA-II) is the first Zn^{2+} coordination residue and is completely conserved. CA-VIII structure is presented in Figure 1.8 and has been found to be most similar to CA-II and CA-XIII of the cytosolic subgroup [122]. From the protein structure it is evident that the

Zn²⁺ cofactor is not present.

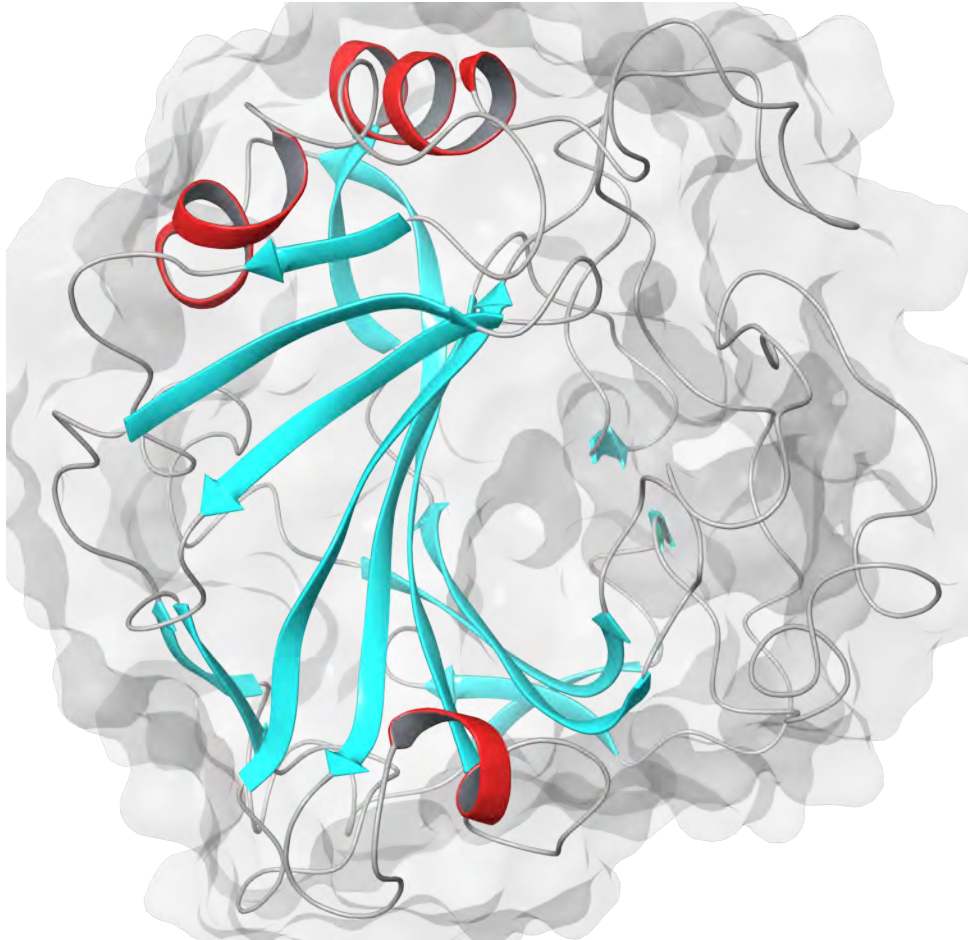


Figure 1.8. CA-VIII protein structure.

The majority of research into the CA-VIII mechanism of action have been performed in mouse and zebrafish models. The mouse and zebrafish CA-VIII proteins share 98% and 84% sequence identity with human CA-VIII [122, 123]. Protein localisation studies have identified that CA-VIII is expressed highly in the cerebellum [124], and is necessary for motor coordination [123, 125] and central nervous system development [126, 127]. In the cerebellum, CA-VIII associates with the inositol 1,4,5-triphosphate receptor type 1 (IP₃R1 or ITPR1) in the Purkinje cells, where it allosterically inhibits the binding of inositol triphosphate (IP₃). This is achieved by lowering IP₃ affinity for the ion channel protein without altering the maximum number of ligand binding sites [127]. Ca²⁺ release is regulated by the binding of IP₃ to ITPR1, and this regulation is necessary for the facilitation of motor learning and synaptic plasticity [128–131].

Investigations into the structure of ITPR1 in 2003 by Hirota *et al.* [127] defined three domains within the protein; ligand binding, modulatory and channel domains. However, research in 2002 and 2005 by Bosanac *et al.* [132, 133] identified the existence of two additional domains; the suppressor and coupling domains. The suppressor domain is located before the ligand binding domain, and also regulates IP₃ affinity for ITPR1 [134]. The coupling domain is located after the channel domain. All five domains of ITPR1 are illustrated in Figure 1.9.

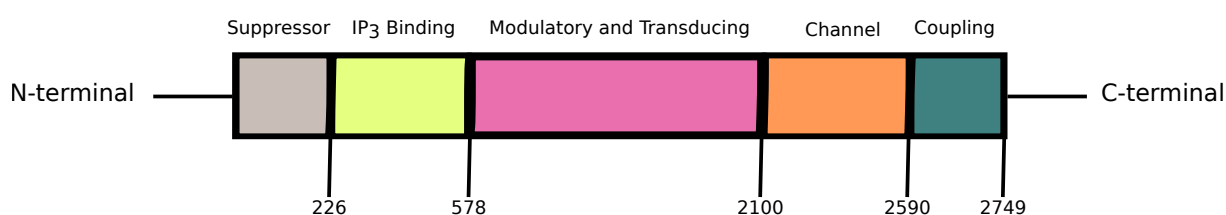


Figure 1.9. Mouse ITPR1 ion channel protein binding domains and corresponding residue number. Adapted from Bosanac *et al.* 2002 [132].

Previous literature into CA-VIII and ITPR1 association studies in mouse models discovered that the CA-VIII residues 45–291_{mouse} are all members of the CA domain (residues 28–290_{mouse}) and form the minimum binding site for association. Identified ITPR1 residues associating with CA-VIII however present a research gap. Hirota *et al.* [127] identified that CA-VIII interacted with residues 1387–1647 within the modulatory domain. The modulatory domain has also been identified as being responsible for binding other cellular proteins such as calmodulin (CAM) [135]. CAM like CA-VIII assists in the regulation of Ca²⁺ within the body, and has been confirmed to bind the ITPR1 residues 1564–1585, which are located within the experimentally determined binding region of CA-VIII [127]. In 2002 a separate study by Sienaert *et al.* [134] noted that the suppressor domain was also capable of binding CAM. As CAM and CA-VIII both regulate IP₃ affinity for its receptor and Ca²⁺ release, these two proteins could bind to similar regions on ITPR1. Within the scope of studied literature no association studies between CA-VIII and the suppressor domain have been performed. Association studies have only been investigated for the modulatory domain [127]. With respect to the modulatory domain, the exact residues essential for the interaction with the ion channel are still yet

to be identified.

1.2.3.1 Deficiencies

Improperly functioning CA-VIII leads to Ca^{2+} dysregulation and results in the rare phenotypes cerebellar ataxia, mental retardation and disequilibrium syndrome 3 (CAMRQ3) (MIM No: 613227). Initially mutations to ITPR1 were identified to be associated with CAMRQ3 due to disturbances to Ca^{2+} signalling by the ion channel protein [131, 136–140]. Later research found out that heterogeneous mutations to the CA-VIII gene at chromosome 8q12 also cause CAMRQ3. Genetic studies have revealed that the mutant protein S100P causes CAMRQ3 [82, 131, 141–143]. Patients with the mutation were observed to have mild mental retardation and ataxia. Additionally the G162R mutation has also been identified as being associated with cerebellar ataxia [144].

Currently the mechanism of association between CA-VIII and ITPR1 is not well understood presenting an obstacle to the treatment of CA-VIII disorders and drug discovery [38, 143, 145].

1.3 KNOWLEDGE GAP

Most research into CA drug discovery has focused on the inhibition of proteins to achieve therapeutic effect. Some of these inhibitors also tend to be non-specific and inhibit multiple CA isoforms. This has left a research gap in drug discovery with respect to the identification of “activator” compounds to assist with poorly functioning proteins, and/or allosteric compounds that may rescue the function and structure of poorly folded proteins. This coupled with inadequate studies into the rare phenotypes leaves individuals with underwhelming treatment options. When the diseases are degenerative as with CA-II, CA-IV and CA-VIII this also results in a diminished quality of life.

Understanding the effects of SNVs on CA structure and function would advance drug discovery and precision medicine in two main ways. Identification of disease pathogenesis and role of SNVs would allow for the determination of an individual's chances of developing disease, and possible

drug responses, thereby aiding precision medicine. A comparison of benign and pathogenic variant mechanisms in addition to identifying disease pathogenesis, and undesirable changes to protein structure, may also assist in the identification of protective structural changes or variant mechanisms of action. This would have biotechnological implications for CA associated technology where robust variant proteins may need to be genetically engineered to further research into artificial lungs, blood substitutes or CO₂ sequestration [97, 146].

For simplicity, the terms nsSNVs, SNVs and variants will be used interchangeably to refer to missense variations unless otherwise stated. Due to literature differences on the definitions of mutations and variants, the terms mutations and variations have both been used interchangeably to describe nsSNVs throughout this thesis, and all variants are non-synonymous in nature.

I.4 RESEARCH AIM

The aim of the current research was to use a combination of molecular dynamics (MD) and dynamic residue network (DRN) analysis approaches to study and analyse the effect of nsSNVs on the structure and function of CA-II, CA-IV and CA-VIII. This research was broken down into five main goals to achieve this:

1. Protein characterisation: Sequence analysis methods including motif identification using CA-II as a reference to characterise the structures of CA-IV and CA-VIII (Chapter 2).
2. SNV identification: Discovery of validated SNVs associated with phenotypes in CA-II, CA-IV and CA-VIII through the use of online resources and literature. Including homology modelling to note structural differences between wild-type (WT) and variant proteins (Chapter 2).
3. Cofactor parametrisation: Quantum mechanical (QM) and molecular mechanical (MM) calculations using Gaussian09 and the AmberTools packages to develop Zn²⁺ force field (FF) parameters in order to allow metal ion inclusion in MD simulation (Chapter 3).
4. MD simulation: Simulations of the variant effects on CA-II, CA-V and CA-VIII structures to investigate changes occurring to the global protein structure using the GROningen MACHine for Chemical Simulations (GROMACS) package (Chapter 4).

5. DRN analysis: Analysis of the variant effects on the local protein structure of CA-II, CA-IV and CA-VIII to observe changes in residue-residue interactions, accessibility and residue usage between wild-type (WT) and variant proteins through the use of the MD-TASK suite tools (Chapter 4).

*There is no passion to be found playing small, in settling for
a life that is less than the one you are capable of living.*

Nelson Mandela

2

Characterisation of CA-II, CA-IV and CA-VII, and Protein Variants

CHAPTER OVERVIEW

The study into the pathogenesis of rare diseases presents a complex task due to inadequate information as to the function and mechanism of the causative proteins. In addition, even when the function of the protein is understood, variant effects to function and structure within these proteins remains relatively unknown presenting an additional hurdle [15]. Where variants have been identified in proteins, predictions as to whether these variants are benign or pose health risks also presents a challenging task [15]. This chapter is divided into two key aspects; *in-silico* characterisation of CA-II, CA-IV and CA-IV; and the identification of pathogenic SNVs, and homology modelling of variant

proteins. As the phenotypes associated with CA-II, CA-IV and CA-VIII are rare, characterisation allows for the complete analysis of the proteins from primary to tertiary structure. This will include; motif discovery, multiple sequence alignment (MSA) techniques and binding site identification. Functionally important residues are conserved to preserve protein function, and can be used to deduce unknown residue or motif functions. Additionally, potentially novel protein-protein interactions (PPIs) will be investigated to discover CA protein associations not observed within the scope of studied literature or yet to be reported. Identification of pathogenic SNVs and homology modelling would set the foundation for investigations into the pathogenesis of CA-II, CA-IV and CA-IV deficiencies through assessment of the effects of variant presence on protein secondary structure and potential consequences to 3D structure.

2.1 INTRODUCTION

Due to the ubiquitous nature of CA-II and its high rate of catalysis this protein has been extensively studied and used as a standard model to understand the mechanism of action of the other catalytic CAs [19, 27, 50]. In addition, mutagenesis experiments have also been employed to investigate individual residue importance within the protein [147, 148]. The investigations into CA-II, though beneficial, have also resulted in far less research being conducted on other CA isoforms such as CA-IV. The acatalytic isoform CA-VIII is also poorly understood. Though CA-II, CA-IV and CA-VIII exhibit low sequence similarity to each other they share some conserved residues suggesting that CA-II could be used to characterise the less extensively studied CA-IV and CA-VIII proteins. The identification and characterisation of the important residues or motifs within these isoforms would set the foundation for investigations into the pathogenesis of the CA-II, CA-IV and CA-III deficiencies; osteopetrosis with RTA and cerebral calcification, RP17, and CAMRQ3 respectively, which are poorly understood.

There are multiple techniques and tools available for the characterisation of protein functions in bioinformatics. These include and are not limited to; protein sequence analysis, 3D structural analysis and network interaction analysis.

2.1.1 Protein Sequence Analysis

Protein sequence analysis centres around the identification of conserved residues or residue groups within proteins, in that, catalytically and functionally important residue groups within proteins are seldom mutated or substituted, and remain conserved to preserve protein function [149–153]. Protein structure is also preserved if substituted amino acids have similar physico-chemical properties [150, 154]. Examples of techniques utilised in protein sequence analysis include; MSAs and motif identification.

2.1.1.1 Multiple Sequence Alignment (MSA)

MSA involves the comparison of homologous sequences for protein function prediction, structural and phylogenetic analysis [155, 156]. These comparisons can also be used to infer homology and relationships between protein sequences. This technique is however highly dependent on accuracy of the alignment. Two types of alignment exist; local and global. Local alignment involves aligning specific segments of the sequences, whereas in global alignment the whole sequences are aligned [157–159].

As alignment forms one of the corner stones of bioinformatics, multiple alignment algorithms exist that attempt to balance biological accuracy and computational complexity [155]. Biological accuracy refers to the quality of the alignment, whereby deletions, gaps and insertions should be in their correct positions. In addition, the MSA should be close to the true alignment of the sequences. To determine MSA quality, alignment scores are calculated. These scores are a summation of amino acid substitution scores between pairs of sequences, with higher values indicating better quality alignments.

Amino acid substitution matrices contain all possible residue alignment scores [160, 161].

The speed and accuracy of the MSA is governed by computational complexity. Complexity is defined as $O(L^n)$ [162]. O defines the complexity, L is the sequence length and n the number of sequences. Complexity is directly proportional to n . Two sequence (pairwise) alignments are calculated using dynamic programming algorithms, however as n increases the speed of alignment gets slower presenting a complex problem [163, 164].

To solve the complexity problem, a number of algorithms employing the “progressive alignment” heuristic [165] were developed, whereby protein sequences are progressively aligned using either the Smith-Waterman [166], Needleman-Wunsch [163], k-tuple [167] or k-mer algorithms [168]. Afterwards the relationship between sequences is determined through clustering using the k-means and/or mBed methods [169]. Guide trees are then calculated from distance scores (derived from normally converted similarity scores), and used as a reference to add sequences to the MSA. The Neighbour-Joining (NJ) [170] and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [171] methods are used to calculate guide trees. Similar sequences are initially added to the MSA followed by more distant sequences [155]. The drawbacks to progressive alignment include, the entire MSA is dependant on the initial alignment, therefore any mistakes occurring initially cannot be rectified in the later steps. The “iterative alignment” heuristic was developed as an improvement to the traditional progressive. The difference is that in iterative alignment the sequences are realigned and progressively added to improve alignment quality. This allows for the correction of mistakes that may occur initially [143, 155].

The algorithms; ClustalΩ (Omega) [172, 173], ClustalW [174], MAFFT (Multiple Alignment using Fast Fourier Transform) [175] and MUSCLE (MUltiple Sequence Comparison by Log-Expectation) [168] all employ progressive alignment. From this list ClustalΩ, MAFFT and MUSCLE are capable of iterative alignment. The former two alignment algorithms are

summarised below.

2.1.1.1.1 *ClustalΩ*

The ClustalΩ algorithm shares similar quality with other algorithms, but tends to be more efficient for the alignment for a larger number of sequences. The k-tuple method is used to generate a pairwise, followed by sequence clustering via the a modified mBed method, and finally clustering using the k-means [155, 172, 173]. The modified mBed method has a complexity of $O(N \log N)$ increasing efficiency from $O(N^2)$ or $O(N^3)$. Each sequence is embedded in a space of n dimensions whereby $n \propto \log N$ [172]. An n element vector that can be rapidly clustered by k-means is then used to replace each sequence. The UPGMA method is utilised to build a guide tree, and the MSA constructed using HHalign which aligns a two profile hidden Markov model (HMM) [176–178].

2.1.1.1.2 *MAFFT*

The MAFFT algorithm utilises fast Fourier transform (FFT) to rapidly identify homologous regions within sequences [175]. The progressive alignment methods (FFT-NS-1, FFT-NS-2) and iterative methods (FFT-NS-i) can be applied to protein sequences during MSA.

$$c(k) = c_v(k) + c_p(k)$$

Equation 2.1. Determination of the correlation between two amino acid sequences. Symbols are, $c(k)$: correlation; $c_v(k)$: correlation of volume component; $c_p(k)$: polarity component.

The correlation (level of similarity) between two protein sequences is determined according to Equation 2.1. The $c(k)$, $c_v(k)$ and $c_p(k)$ refer to; correlation, correlation of volume component and polarity component respectively [175]. Homologous regions between two sequences are indicated by a high value of $c(k)$, that also correspond to these regions. During pairwise alignment when two sequences are of approximately equal length $c_v(k)$ they have a complexity of $O(N^2)$. FFT is capable of improving the efficiency of the alignment to $O(N \log N)$ and thus reduce CPU time.

Homologous region positions are not determined by the FFT analysis, but by the positional lag k . Sliding window analysis using a window size of 30 sites is utilised to determine the homologous regions, whereby the highest 20 peaks of correlation $c(k)$ are used to calculate the degree of local homologies [175]. When a minimum of two consecutive segments are homologous, they are merged into a single segment. Homologous segments are then arranged consistently in both sequences, to obtain an alignment.

2.1.1.2 Motif Sequence Analysis

Motifs are defined as conserved patterns in a sequence that identify important functional or structural regions occurring within in a set of proteins [179]. Motif analysis provides an additional dimension to protein sequence analysis. MSAs identify conserved residues whereas motif analysis identifies conserved patterns. Highly conserved patterns across numerous species/proteins highlight essential residue groups. Unknown and known motifs can be compared, and/or functional residues within the motifs assessed to characterise the respective protein. Motifs can either be queried within a database or manually identified from a set of homologous sequences.

2.1.1.2.1 Motif Database Query

Databases such as Pfam [180] and PROSITE [181] allow protein characterisation through querying a sequence to identify known protein domains. These assist with potential functional identification, and each domain can contain specific motifs that responsible for specific functions within the protein such as; stability maintenance, intracellular localisation adaptation and/or functional roles [182].

2.1.1.2.2 Motif Identification From Sequences

During identification from homologous sequences, motifs can be represented in two ways, namely; consensus string and Position-specific Weight Matrices (PWM) [183].

The consensus string representation shows the conserved and variable residues within a motif sequence. The PWM representation employs a position frequency matrix (PFM) where the residue frequency at each position is calculated via division of the residue sum, by the number of sequences. Within the PWM, each row presents an amino acid under the IUPAC (International Union of Pure and Applied Chemistry) naming system and columns represent a particular pattern position [183–185]. Information of each element within the PWM is calculated according to Equation 2.2 [186]. I_i is the position information at i within the alignment; A is the set of residues inclusive of gaps; p_j is the background frequency; t_{ij} represents site-specific frequency of residue j at position i . The motif logo that shows amino acid conservation in each motif sequence is generated by the graphical plotting of I_i [186, 187].

$$I_i = \sum_{j \in A} t_{ij} \log_2 \frac{t_{ij}}{p_j}$$

Equation 2.2. Calculation of position information within alignment during motif identification.

The two main tools involved with the identification of motifs include; Multiple Expectation Maximization for Motif Elicitation (MEME) [188–190] and Motif Alignment and Search Tool (MAST) [191]. MEME utilises a probabilistic approach in the identification of motifs. The Expectation Maximization (EM) algorithm is used to simultaneously optimise PWMs [183, 188–190]. During motif identification, an initial motif discovered by MEME is improved by EM until the number of iterations is acquired and until the PWM values do not change. An E-value is then calculated by merging the best sequence motifs to generate an overall match between the motif and sequence. Motifs with E-values less than 0.001 are regarded as being accurate [190]. MAST is however sequence orientated and therefore provides a single sequence score making it well suited for protein analysis. MAST determines the best sequence match to each motif based on the MEME output [190], and can be used to validate MEME output. Motifs with MAST pairwise

correlations greater than 0.6 are usually omitted from analysis as associated E-values and p-values may be underestimations resulting in diminished accuracy [190]. The MEME web-server is capable of automatically conducting MAST analysis on its output.

2.1.2 Three Dimensional Structural Analysis

The 3D structure of a protein governs its function and is more well conserved than its sequence. This indicates that structural comparison between homologous structures can be used to characterise proteins [182, 192]. Homologous structures can be identified by querying the Protein Data Bank (PDB) [193] with the sequence. Sequences containing at least 30–40% identity are regarded as homologous, but this is dependant on the length of the sequences. Incorporation of E-values eliminates the dependency on sequence length with sequences of E-values less than 10^{-10} being regarded as homologous [194]. This threshold holds true even when sequences share less than 30% identity which can be an underestimation of homolog detected via sequence similarity. Homologous proteins usually share similar function [194]. Previous research found that structures sharing a root mean square deviation (RMSD) average ranging from 0–1.2 Å within the PDB usually represents identical proteins, with the differences being as a result of crystallography resolution or protein flexibility [195].

2.1.3 Protein-Protein Interaction Analysis

In addition to the above techniques that are dependent on some form of comparison and functional inference, proteins can be characterised through the use of PPI networks [196]. PPIs define how different proteins interact with each other, and are of particular interest in protein characterisation as some biological functions are dependent on PPIs. The disruption of PPIs could therefore result in disease [197].

Network interaction predictions can either be determined experimentally or computationally.

Experimental determination involves high-throughput laboratory analysis of potential PPIs. Computational predictions of PPIs uses various tools and is divided into numerous categories depending on the prediction data-type [197]. Some examples of these include but are not limited to domain and motif pairs and genomic methods.

2.1.3.1 Domain and Motif Pairs

As domains each perform a specific task within a protein, interaction partners between similar domains/proteins are conserved to an extent. Interactions between close homologs occur in a similar manner [198]. Interaction similarity would also allow for the identification of PPI sites which would be invaluable in the study of drug design and disease [197].

PPI interaction diagrams though showing protein associations do not indicate interacting domains. Various techniques have been developed to identify domains. The majority however include domain identification using databases such as Pfam, and training models to identify domain-domain interactions using known PPIs.

2.1.3.2 Genomic Methods

PPI prediction using genomic methods involves the analysis of gene fusion, interacting protein sequence pair co-evolution and gene order conservation [199]. Gene fusion utilises domain/protein homologs to determine potential associations. The principle is that, other genomes contain homologs of interacting domains/proteins fused into a single protein chain [199]. Sequence co-evolution analysis involves similarity assessment between a pair of distance matrices or phylogenetic trees [198, 199]. Gene neighbourhood is based upon the principle that genes encoding potentially interacting proteins within closely related functions are co-regulated in eukaryotes as operons [198]. Distantly related organisms tend to have a shuffled gene order as a result of neutral evolution, however, operons in co-regulated genes are highly conserved. Functional linkage between genes and proteins

can therefore be predicted using gene neighbourhood. Previous research has found that physical interaction occurs in 63–75% of co-regulated genes [200, 201].

The STRING database [199, 202] is a common tool that uses protein network interactions to characterise proteins and determine PPIs. A combination of high-throughput experimental analysis, computational predictions and data mining from publicly available information resources are used to determine potential protein associations in STRING. Post prediction, STRING outputs an interaction diagram where the nodes indicate respective proteins and the edges show interactions of associations between the proteins [197, 199, 202]. Edge thickness correlates to confidence scores that indicate the approximate probability of the existence of an interaction between two proteins. Low, medium, high confidence are represented by scores of 0.15, 0.4, 0.7 and 0.9 respectively [202].

2.1.4 Variant Identification and Characterisation

Variant characterisation involves investigations of potential SNV effects to protein structure, and the determination of potential associations to phenotypes.

To identify potential linkages or non-linkages between variants, and specific phenotypes and traits, association studies are conducted. These involve the genetic sequencing of populations and genomes comparison between healthy and unhealthy individuals to confirm associations [15, 203]. Correlations between phenotypes and variations do not however prove causality. As these studies generally involve populations, this has allowed for the determination of frequencies to study variant prevalence. It should however be noted that in addition to the population, frequency is also affected by how deleterious the variant is. Variants associated with high mortality in populations would occur at a lower frequency due to the number of deaths.

Projects currently involved with investigations into population genetics include; 1000 Genomes [204], Genome Aggregation Database (gnomAD) exomes and genomes [205], Trans-Omics for Precision Medicine (TOPMed) [206] and Exome Aggregation Consortium (ExAC) [207]. Although

the information within these project databases is useful not all populations are represented in each, and since the databases are independent of one another they could report varying frequencies for different population groups. These databases also complement scientific research that has been conducted linking variations to phenotypes, and serve as an information validation platform confirming that the variation occurring is indeed a true variation and not the result of sequencing errors.

Variant identification involves the filtering of these large databases/datasets to identify variations associated with a particular trait. To assist with this numerous resources exist, including ClinVar [82], Online Mendelian Inheritance in Man (OMIM) [208] and Variant Analysis Portal (VAPOR) [209].

2.1.4.1 ClinVar

ClinVar is a public resource involved with the processing of variations found in patient samples and their reported clinical significance to identify relationships between genetic variations and disease [82]. The reported variant clinical significances can be divided into three categories; benign, pathogenic and protective [210]. Benign refers to variants present within the protein that do not result in a particular phenotype, whereas pathogenic refers to variants resulting in a specific phenotype or that could make an individual more susceptible to it. Protective variants make individuals more resilient to a particular phenotype.

2.1.4.2 OMIM

The OMIM database is a freely available and comprehensive resource of human genes and genetic phenotypes. The database prioritises and focuses on linkages between genotypes and phenotypes [208]. The databases also links other genetic resources to validate its findings.

2.1.4.3 VAPOR

VAPOR is a consensus of bioinformatics tools developed in 2018 that allows filtering of SNV datasets to obtain predictions of clinical significance and potential variant effects to stability. VAPOR merges FATHMM [211], PhD-SNP [212], PolyPhen-2 [213] and PROVEAN [214] to determine potential clinical significance as either damaging or tolerated. I-Mutant 2.0 [215] and MUpro [216] are also employed to determine potential SNV effects on stability [209].

The comparison between benign and pathogenic variants for CA-II, CA-IV and CA-VIII allows for the identification of key effects resulting in the development of pathogenic phenotypes.

2.1.5 Homology Modelling

Homology modelling is a technique that is used to investigate the 3D structures of proteins. This technique is normally implemented when crystal structure of the respective proteins are not available for investigations. Homology modelling allows for the mapping of specific motifs onto the protein structure to observe their location within 3D space. In addition, differences between wild-type (WT) and mutant proteins via the determination of changes to protein secondary structure can also be investigated [18, 217].

Homology modelling is conducted via four consecutive steps listed below:

1. Template identification
2. Multiple sequence alignment
3. Model building
4. Model validation

In the case of low quality models, structural refinement can be performed prior to model validation to improve quality.

2.1.5.1 Template Identification

Homology modelling of unknown proteins is performed using the atomic coordinates of a template (usually X-ray crystallography structure or homology model of the WT protein for SNVs), and is thus dependent on the template quality for accuracy. A combination of structure resolution, R-factor and Rfree values determine the quality of a template [218–220]. This suggests that template identification is therefore the first and most important step of homology modelling. Templates can be identified and have quality inspected through the RCSB database [193].

2.1.5.2 Multiple Sequence Alignment

MSA of the query and template sequences is the second most important step, whereby amino acid conservation and secondary structure conservation has to be thoroughly analysed. During this step it is important to ensure that there are minimal alignment gaps occurring within the protein secondary structure (α -helices and β -sheets) to have high quality structures. Where the positions of α -helices and β -sheets are unknown, the PSIPRED program [221] can be used to predict protein secondary structure from a respective sequence.

2.1.5.3 Model Building

After alignment homology models can then be calculated using programs like MODELLER [222] and/or Prime [223, 224]. These programs will generate multiple protein structures which can then all be compared to obtain the best representative protein structure.

2.1.5.3.1 *MODELLER*

MODELLER generates models and predicts conformations through the use of experimentally generated protein structures. This is achieved through the obtainment of spatial restraints from the MSA, and optimally satisfying them. The models are initially expressed as probability density

functions (PDFs) of the restrained features [222]. The PDFs restraining separate spatial features are then combined to generate the molecular PDFs that are then optimised to generate the final 3D models. The PDF optimisation is performed to ensure minimal deviation between the input restraints and the model, by applying an energy minimisation using the conjugate gradient algorithm [222, 225].

2.1.5.3.2 *Prime*

Prime is a homology modelling tool developed by Schrödinger. Atomic positions obtained from the MSA of proteins are used to calculate structures, and the OPLS2000 all-atom force field (FF) used to determine model energy score [223, 224, 226]. The orientation of side chains including potential electrostatic and van der Waals (vdW) interactions are predicted using a Generalized Born model of solvation. In the event whereby the template and query sequence do not align, a solvation incorporating *ab initio* procedure is utilised to generate unaligned sections of the query sequence [224].

2.1.5.4 Validation

A number of tools and servers are available to validate homology models and confirm accuracy. Some of these include, the discrete protein energy (DOPE) score [222], PROSA [227], Ramachandran plot and Verify 3D [228].

DOPE is a pairwise atomic distance-dependent statistical potential derived using probability theory from a sample of native protein structures [229], is automatically implemented by MODELLER and optimised during model building [222, 229]. The energy of atom pairs within the protein is summed to generate the model energy, which can be utilised to distinguish model quality. As DOPE is protein specific it cannot be used to compare different proteins. To facilitate this, normalised DOPE (z-DOPE) score is used as it is not dependent on adjustable parameters. Negative values are indicative better protein structures [222]. Models with a z-DOPE score less than -1.00 are regarded as having

native-like structures.

PROSA uses a combination of z-score and residue scores to validate protein models. The z-score is defined as the energy difference between a protein's native-fold and an ensemble of protein misfolds in units of the ensemble standard deviation [230]. The calculated z-score is then compared to the scores of other proteins within the PDB to observe whether the score lies within the range of native proteins of similar size [227]. Residue score is a plot of local energies as a function of amino acid sequence position [227]. Poor regions of the protein structure are represented by positive energy values, and thus allows the determination of protein model quality. Brobdingnagian energy fluctuations are typically observed when plotting single residues. Average energy over each 40 residue range is therefore calculated to smooth the plot.

In Ramachandran plot, φ (phi) and ψ (psi) torsion angles are used to identify energetically favourable amino acid backbone conformations [231, 232]. Verify 3D validates structures by comparison of respective 3D atomic models to their amino acid sequence through the use of a 3D profile, and comparing to good structures [233].

2.1.6 Binding Site Identification

With homology models constructed, binding site residues of the less commonly studied CAs can be identified, and variants assessed for potential effects to binding site residues. This would be of great interest especially in the analysis of PPIs for the study of diseases or drug design. Numerous programs exist for the identification of PPI binding site residues, however, the two main ones utilised within this research include; SiteMap [234, 235] and CPORT (Consensus Prediction Of interface Residues in Transient complexes) [236].

SiteMap selects site points based on energetic and geometric properties through the use of an algorithm similar to Goodford's GRID algorithm [237]. To each grid point, hydrophilic and hydrophobic properties are calculated and contour maps prepared [235]. Determined binding site

residues are ranked according to a SiteScore metric which represents a weighted average of enclosure and hydrophobic scores, and number of sites which all aggregate to a total of 1.0. A SiteScore of 0.80 can accurately determine binding and non-binding PPI sites on proteins. Greater SiteScores signal greater binding site confidence.

CPORT uses a combination of cons-PPISP (consensus Protein-Protein Interaction Site Predictor) [238], PIER (Protein IntErface Recognition) [239], PINUP (Protein Interface residUe Prediction) [151], ProMate [240] and SPPIDER (Solvent accessibility based Protein-Protein Interface iDentification and Recognition) [241] to predict PPI residues.

2.2 METHODOLOGY

2.2.1 Data Retrieval

2.2.1.1 Protein Sequences

The protein sequence of CA-II (UniProt accession: P00918; accessed 20/06/2018) was downloaded from the Universal Protein Resource (UniProt)[40]. A UniProt BLAST using the CA-II as the query sequence was then performed to identify other α -CA isoforms. Homologous sequences were identified using the BLASTp algorithm in conjunction with BLOSUM-62 matrix [242] within the UniProtKB target database and an E-threshold search parameter value of 1000. All human α -CA homologs (CA-I to CA-XV) were then selected from the BLAST results to create the final CA family dataset. This dataset consisted of 16 sequences in total, including CA-IV and CA-VIII (Uniprot accessions: P22748 and P35219 respectively), and are presented in Table S1.

The protein sequences of CA-IV and CA-VIII were also separately downloaded from UniProt. As these proteins were part of the initial CA-II homolog dataset, the sequences were used in conjunction with BLOSUM-62 matrix to perform a reverse BLAST within the UniProtKB database using an E-threshold value of 1000 to include potentially weak similarities. This was to observe whether homologs within the initial dataset were identified.

2.2.1.2 3D Protein Structures

2.2.1.2.1 CA-II and CA-IV

The PDB contains 652 crystal structures for the CA-II protein (UniProt accession number: P00918) [40]. The majority of these structures contain the common CA-II feature whereby residue 126 is missing [243]. This feature results in inaccurate SNV modelling, and MD simulation. Only CA-I was regarded as containing residue 126 [244], therefore crystal structure numbering

proceeds from 125 and continues to 127 even though the ATOM and FASTA sequences have a 100% match. Within some of the structures, Zn^{2+} is not in the correct coordination geometry (tetrahedral: $3 \times \text{His}$ and $1 \times \text{H}_2\text{O}$) hindering accurate metal ion parametrisation.

The crystal structure of CA-IV (UniProt accession number: P22748) has been solved in 10 structures. From the crystal structures, there are no zymogen (containing the signal peptide region and the GPI-anchored region) templates, therefore only residues 20–284 could be modelled using template based modelling. No crystal structures containing CA-IV with co-crystallised CO_2 were also observed.

To discover potential templates for the homology modelling of CA-II and CA-IV, the numerous templates for each protein were downloaded and filtered using the *PdbSearcher.py* script bundled with AmberTools17 [245]. To lists containing the templates of CA-II and CA-IV each, *PdbSearcher.py* was set to discover all bonds existing between Zn^{2+} and ligating atoms within a radius (cut-off) of 2.5 Å. The maximum Zn-ligand bond distance is 2.5 Å [246, 247]. The resulting environment and summary output files were then analysed to identify PDB templates containing Zn^{2+} in a tetrahedral coordination geometry. These templates contained Zn^{2+} in an HHHX ($3 \times \text{His}$ and $1 \times \text{H}_2\text{O}$) coordination formation. From the identified templates, the best crystal structures for the homology modelling of CA-II and CA-IV were then selected using a combination of PDB validation and resolution. As CAs reversibly hydrate CO_2 , templates containing either CO_2 or BCT were prioritised.

2.2.1.2.2 CA-VIII

The PDB contains one structure for CA-VIII (PDB ID: 2W2J) that has been solved via X-ray crystallography. The first 23 N-terminal residues are however not crystallised, and the structure is missing; CG, CD, OE1 and OE2 atoms from the residues Glu23, Glu24, Glu28 and Glu264. The atoms CG, CD, OE1 and NE2 are missing from Gln187; and CG1 and CG2 atoms are missing from Val263.

2.2.2 Protein-Protein Interaction Prediction

CA-II and CA-IV have previously been found to interact and form transport metabolons with membrane carriers, whereas CA-VIII has been found to associate with ITPR1. To identify other proteins CA-II, CA-IV and CA-VIII potentially associate with, protein-protein association networks were predicted using the STRING server by querying each sequence (UniProt accession numbers; P00918, P22748 and P35219 respectively). Post-prediction, the resultant protein-protein network diagram was viewed using the data acquisition type, and results download as scaled vector graphics (*.svg).

2.2.3 Motif Analysis

Motif discovery and analysis of the final CA family (Table S1) dataset was performed according to Ross *et al.*, and Nyamai and Tastan Bishop [248, 249] with minor modification. To the individual datasets, the online MEME SUITE version 5.05 [190] was used to conduct motif discovery using a minimum and maximum motif width of 5–20 residues and the 0-order model of sequences. A maximum of 100 motifs were set to be discovered from each dataset. After discovery, the E-value and MAST were then used to validate the motifs. Motifs with an E-value less than 0.001 and a MAST pairwise correlation less than 0.6 were retained for further analysis.

Validated motifs from each dataset were then mapped onto their respective protein structures (CA-II, CA-IV and CA-VIII) using PyMOL version 2.4.0 [250]. This mapping also served to verify the existence of each motif in the respective CA protein sequence. After verification, a heat map illustrating motif conservation as the number of sites per total number of protein sequences was then constructed using Matplotlib [251]. In addition, the motif logos were also downloaded from the MEME server and visualised using Inkscape [252].

2.2.4 Homology Modelling

2.2.4.1 Wild-Type

With respect to CA-II, crystal structure 2VVA of 1.56 Å resolution and a sequence similarity of 99% to the UniProt protein (covering residues 3–260) was selected as the main template. CO₂ also was co-crystallised to the primary pocket of this template. An additional template 2VVB containing a co-crystallised BCT was also identified and set aside as a secondary template. The renumbering of 2VVA using *pdb4amber* was insufficient to correct the missing residue 126 and resulted in errors. As a result homology modelling was necessary.

For CA-IV no crystal structures containing co-crystallised CO₂ or BCT were identified, therefore X-ray crystal structure 5KU6 of resolution 1.80 Å was selected as the main template. This template covered residues 21–284.

Homology modelling of CA-II and CA-IV was performed using Schrödinger Maestro [253] and Prime [223, 224]. For each protein, the templates 2VVA and 5KU6 were loaded into Prime and aligned to their respective target sequences using ClustalΩ. Prime was then set to calculate five homology models for both CA-II and CA-IV that included the Zn²⁺ ligand, and the coordinating H₂O where possible. For CA-II, the CO₂ HETATM in 2VVA was also included for homology modelling. The quality of the homology models was then validated using the z-DOPE score and Ramachandran plot.

To introduce the missing atoms to the CA-VIII structure, homology modelling of CA-VIII (UniProt accession number: P35219) was performed using 2W2J as a template. MAFFT using the E-INS-i alignment strategy in conjunction with the BLOSUM62 matrix was used to align the target and template protein sequences. After alignment the unaligned first 23 N-terminal residues were then trimmed off as the template does not cover these residues. A total of 100 CA-VIII homology models were then calculated using MODELLER v9.19 [222]. The z-DOPE scores of each model was then

determined, and models ranked according to the best score (lowest values). The top three models with the best scores were then selected and subjected to further validation by PROSA [227] and Verify 3D [228]. The validation results were then collated and the best CA-VIII protein model selected.

For comparative purposes the calculated CA-II (without CO₂), CA-IV and CA-VIII homology models were all superimposed and analysed using PyMOL [250].

2.2.4.2 Variants

SNV analysis was carried out according to proposed protocol by Brown and Tastan Bishop, 2017 [254] and Sanyanga *et al.*, 2019 [66]. The HUMA [209] and Ensembl [210] databases were utilised to identify and download CA-II, CA-IV and CA-VIII nsSNVs. To the downloaded variants, the dbSNP was used to filter and select SNVs that have been validated by frequency within either 1000 Genomes, gnomAD, TOPMed or ExAC. The ClinVar [82] and OMIM [208] databases were then used to cross reference the validated SNVs to identify and isolate all variants associated with a phenotype annotation (pathogenic, benign or protective). Variant effects to protein structure and function were further predicted using VAPOR for additional characterisation.

The identified SNVs were then introduced into their respective structures (CA-II, CA-IV and CA-VIII), with unique structures generated for each nsSNV and protein. The amino acid FASTA sequences of CA-II, CA-IV and CA-VIII were each initially modified to introduce the required SNVs. Homology modelling of each CA protein was then performed according to the respective wild-type (WT) methodology. For all protein SNVs the corresponding WT protein was used as the main template and all HETATMs included in the modelling. Prime was used to calculate homology models for the CA-II and CA-IV variants, whereas MODELLER was utilised to calculate models for the CA-VIII variants. The z-DOPE score was then used to validate the quality of the variant proteins. The two different programs Prime and MODELLER were used to preserve consistency with the respective WT proteins.

With respect to CA-II, CO₂ was removed from the models to generate WT and variant apo proteins.

2.2.4.3 Bicarbonate Bound Structure

The previously downloaded X-ray crystal structure of 2VVB was used to generate BCT bound complexes for CA-II (WT and variants). PyMOL was used to superimpose 2VVB with the apo WT and variant protein models generated previously. To the apo proteins, the coordinates of BCT were added, and the complexes saved to generate the final BCT bound structures. Through superposition, one reference structure (CO₂ bound model) could be maintained for all three CA-II protein systems, apo, BCT and CO₂ bound (both WT and variants) for easier comparison. Structural changes could have been introduced via homology modelling thereby inhibiting direct comparison of the three CA-II protein states.

Though the active sites of CA-II and CA-IV are structurally similar, the CO₂ and BCT bound CA-II protein structures could have been superimposed to generate substrate containing CA-IV structures however this was not performed. RP17 has been attributed to a toxic gain of function due to poor protein folding within the ER inducing stress, and not poor hydration and dehydration of CO₂ and BCT respectively [119].

Total generated protein structures included; 21 models for CA-II (7 apo, 7 BCT and 7 CO₂, 1 WT and 6 SNVs each); 7 apo models for CA-IV (1 WT and 6 SNVs); and 7 models for CA-VIII (1 WT and 6 SNVs).

2.2.5 Identification of CA-VIII Binding Site Residues

Due to the limited literature on the mechanism of CA-VIII and residues essential for enzyme function, potential PPI residues were identified using the Schrödinger SiteMap tool v4.8.012 [234, 235] and the CPORT server from HADDOCK [255, 256]. The CA-VIII WT protein (homology model) was

prepared using the Schrödinger Maestro Protein Preparation Wizard [253] and PROPKA [257] by protonating at pH 7.0. The top five potential binding sites were then identified using SiteMap site recognition. Sites containing a minimum of 15 site points per reported site were retained. A fine grid size of 0.35 Å and restrictive hydrophobicity were set as the search parameters. From the reported binding sites, the top five sites with a SiteScore >0.80 were selected for further analysis. The previously prepared protein was also submitted to the CPORT PPI prediction server.

Results from both SiteMap and CPORT were then collated to build a consensus of potential PPI residues for CA-VIII.

2.3 RESULTS AND DISCUSSION

The main focus of the current chapter was to characterise the CA-II, CA-IV and CA-VIII proteins using a combination of motif and PPI analysis and to identify variants associated the phenotypes; osteopetrosis with RTA and cerebral calcification, RP17 and CAMRQ3.

2.3.1 CA Characterisation Reveals Potential Association With Other Proteins

PPI analysis was performed for all three CA proteins using STRING server [202] and results are presented in Figure 2.1.

The colour of the lines represents the method of interaction prediction. Data in Table S2 represents the associated confidence scores for the results in Figure 2.1. Associated proteins do not have to physically interact with each other, but could jointly contribute to a shared function. Though the STRING server identifies potential interacting proteins, it does not identify regions or residues potentially interacting within the PPIs.

Results in Figure 2.1 demonstrates that the CA-II protein has been predicted to interact with multiple proteins, however experimental evidence has only been conducted for association with SLC9A1 (NHE1) and SLC4A4 (NBC). These STRING results are in agreement with the previous literature findings that support the formation of transport metabolons with these proteins. Numerous other interactions have also been predicted from curation databases with high confidence scores. CA-II is a pathway neighbour of these proteins so even though association linkages were predicted, no physical interaction exists between the proteins.

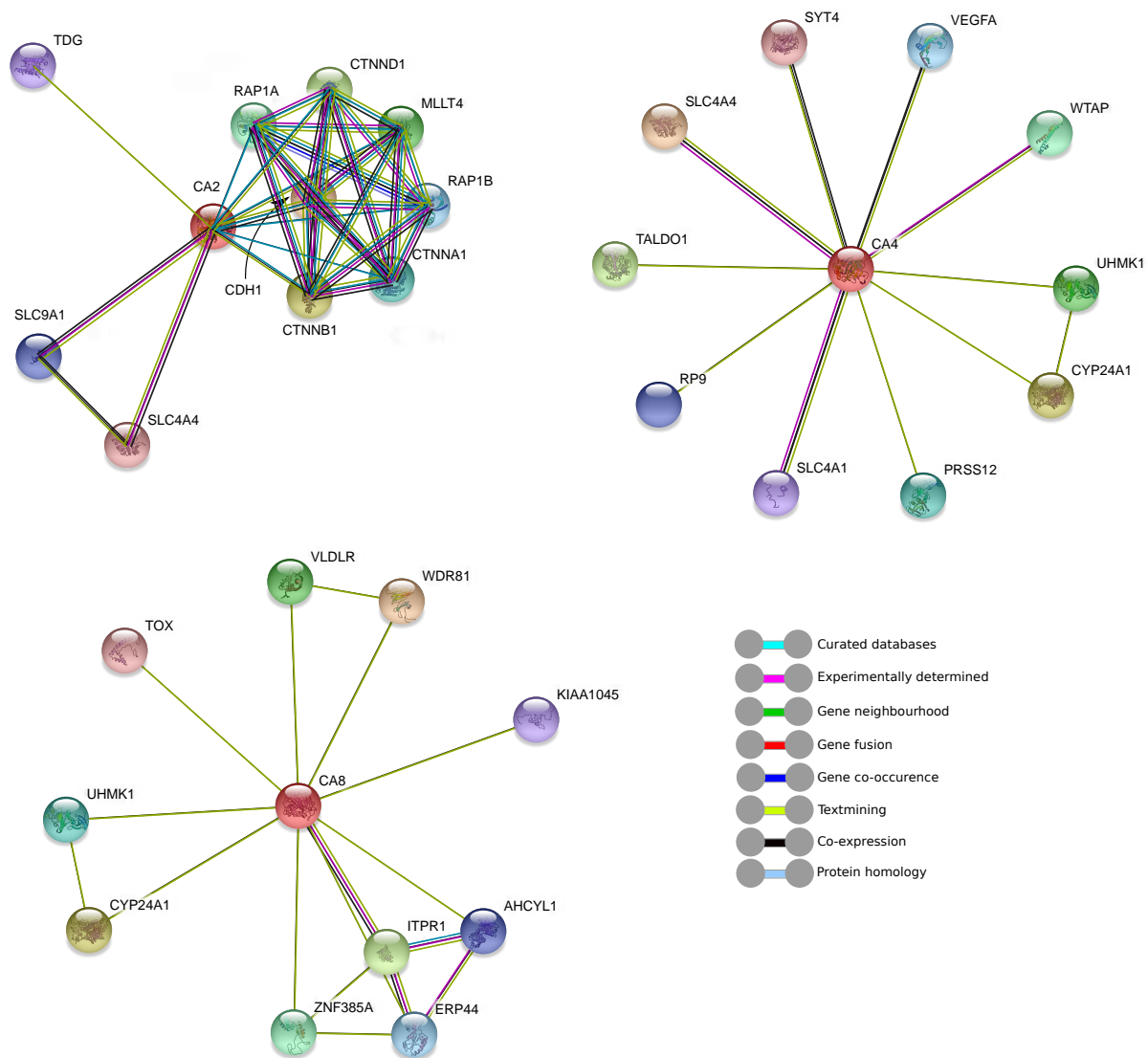


Figure 2.1. STRING protein association interaction network of CA-II (CA2), CA-IV (CA4) and CA-VIII (CA8) proteins. Line colour represents the method of interaction prediction.

With respect to CA-IV, experimental evidence has only been conducted with regards to three proteins; SLC4A1 (AE1), SLC4A4 and WTAP (Wilms' tumour 1-associating protein). The former two also interact with CA-II. Unexpectedly is the association of CA-IV with WTAP. The WTAP protein is highly expressed in several cancers [258–261], and CA-IV was found to be a novel tumour suppressor that acts by binding to WTAP and inducing protein degradation through polyubiquitination [259]. This finding could have implications for the treatment of RP17. CA inhibitors are used to slow progression of RP17 in patients, and should CA-IV become an anti-cancer

target CA inhibitors could interfere with its role suggesting the need for alternative medicines.

Analysis of CA-VIII the data in Figure 2.1 shows that CA-VIII associates with ITPR1 which was expected as previous experimental analyses have verified this interaction [127]. Within the scope of the studied literature and results in Figure 2.1, experimental analysis has been conducted for interactions between CA-VIII and ITPR1, whereas the other associations have been predicted through text-mining. It should also be noted that interactions via text-mining have also been predicted between CA-VIII and ITPR1.

The confidence scores (text-mining) in Table S2 indicate that STRING predicted that CA-VIII was more likely to associate with WDR81 (WD repeat-containing protein 81) and CYP24A1 (1,25-dihydroxyvitamin D(3) 24-hydroxylase) as opposed to ITPR1. WDR81 is a transmembrane protein containing two domains namely the BEACH (Beige and Chediak-Higashi) and WD40 (WD or beta-transducin repeats). The BEACH domain is responsible for lysosomal trafficking, and mutations to the gene have been associated with cerebellar ataxia and disequilibrium syndrome [82, 217]. The WD40 domain is essential to signal transduction, apoptosis and cell cycle control [262]. WD motifs also act as binding sites for protein-protein interactions, suggesting that CA-VIII could bind to the WD motif. CYP24A1 is a cytochrome (CYP) P450 and like CA-VIII, it is important for Ca^{2+} homeostasis within the body [40, 263]. As cytochromes have previously been found to bind other proteins [264] CA-VIII could directly interact with it. Although these associations have been predicted via text-mining, additional research is however required to validate the potential associations between CA-VIII with these other proteins. It still remains unknown whether these proteins could jointly contribute to a specific phenotype such as ataxia [265].

2.3.2 Data Retrieval Identifies Pathogenic and Benign CA SNVs

The Ensembl and HUMA databases identified multiple CA-II, CA-IV and CA-VIII nsSNVs. From the initial screening the CA proteins contain numerous variants, however, to set the foundation for

rare variant effect prediction and rare disease pathogenesis, only validated SNVs with a phenotype annotation were selected. From the screened variants, not all were associated with the phenotypes; osteopetrosis with RTA and cerebral calcification, RP17 and CAMRQ3. To this effect variants annotated as pathogenic and benign were also included for analysis to identify differences behind variant mechanisms. These variants and associated phenotype annotations are presented in Table 2.1. The CA-VIII variant G162R contained no phenotypic annotation within the Ensembl or HUMA databases but had been regarded as pathogenic within scientific literature [144]. This SNV was included within the final dataset to investigate it. Table 2.1 data also presents VAPOR [209] variant effect predictions to protein structure. From the results it is observable that variant presence is expected to cause stability reductions within the proteins. The global minimum allele frequency (MAF) of each variant is also presented in Table 2.1. With respect to CA-II, variant N252D occurs at the highest frequency, whereas in CA-IV N177K and V234I occur at the highest frequency. E109D has the highest MAF in CA-VIII. From the data in Table 2.1 the benign variants generally occur at higher frequencies than pathogenic ones. From the table N252D contains both a pathogenic and benign annotation.

The z-DOPE scores in Table 2.1 present values less than -1.00 for all variants indicating that structures resemble native proteins and are of high quality. No experimental evidence as to the possibility of variant linkage disequilibrium in either of CA-II, CA-IV and CA-VIII was noted within the scope of the studied literature. Therefore no models containing a combination or multiple SNVs were generated. Each homology model contained one SNV only.

Table 2.1. CA-II, CA-IV and CA-VIII identified SNVs and potential variant consequences.

rs ID	Variation	z-DOPE score	MAF	I-Mutant		MUpro		Clinical significance
				$\Delta\Delta G$	Stability	$\Delta\Delta G$	Stability	
CA-II								
rs118203931	K18E	-2.204	<0.01	-1.30	Decrease	-0.499	Decrease	Pathogenic
	K18Q	-2.203	<0.01	-1.24	Decrease	-0.655	Decrease	Pathogenic
rs118203933 ^a	H107Y	-2.204	<0.01	0.26	Increase	-0.867	Decrease	Pathogenic
rs118203932	P236H	-2.159	<0.01	-1.46	Decrease	-0.586	Decrease	Pathogenic
	P236R	-2.186	<0.01	-0.60	Decrease	-0.311	Decrease	Pathogenic
rs2228063	N252D	-2.197	0.007	0.06	Increase	-0.474	Decrease	Pathogenic / Benign
CA-IV								
rs267606695 ^b	R69H	-1.726	<0.01	-0.91	Decrease	-1.32	Decrease	Pathogenic
rs149391728	N86K	-1.737	<0.01	-0.35	Decrease	-1.40	Decrease	Benign
rs267606695	N177K	-1.730	0.003	-0.45	Decrease	-0.78	Decrease	Benign
rs118203931 ^b	R219C	-1.760	<0.01	-0.71	Decrease	-1.39	Decrease	Pathogenic
	R219S	-1.744	<0.01	-1.32	Decrease	-1.57	Decrease	Pathogenic
rs387906598	V234I	-1.758	0.003	0.17	Increase	-0.39	Decrease	Benign
CA-VIII								
rs267606695	S100A	-1.374	<0.01	-0.66	Decrease	-1.23	Decrease	Pathogenic
rs149391728	S100L	-1.381	<0.01	-0.31	Decrease	-0.17	Decrease	Benign
rs267606695 ^c	S100P	-1.410	<0.01	-0.27	Decrease	-1.39	Decrease	Pathogenic
rs149391728	E109D	-1.436	0.50	-0.16	Decrease	-0.41	Decrease	Benign
rs149391728 ^c	G162R	-1.357	<0.01	-0.84	Decrease	-0.54	Decrease	Pathogenic
rs387906598 ^c	R237Q	-1.393	<0.01	-0.58	Decrease	-1.05	Decrease	Pathogenic

^a Associated with osteopetrosis with RTA and cerebral calcification, ^b Associated with RP17, ^c Associated with CAMRQ3.

2.3.3 Variant 3D Spatial Location May Disrupt Protein Function and Integrity

Knowing the respective binding sites for each catalytic CA, 3D spatial analysis of CA-II and CA-IV, and the respective variants was then performed. Data in Figure 2.2 demonstrates variant spatial 3D location within CA-II, and the primary and secondary CO₂ binding pockets (Figure 2.2A,B respectively) [66]. With the exception to K18E and K18Q that are located on an α -helix, all other CA-II variants are located within loop secondary structures. In addition, structural inspection reveals that H107Y is located closest to the active site and primary CO₂ binding pocket towards the interior of the protein. The remaining variants are located closer to the protein exterior surface. Both CA-II substrates; BCT and CO₂ (see Equation 1.1) bind the primary pocket since the secondary pocket is acatalytic.

Analysis of the amino acid physiochemical properties could explain the VAPOR results and the variant clinical significance. In K18E, Lys and Glu have different charges (positive and negative respectively). Replacement of Lys with a negatively charged amino acid could affect salt bridge formation. The pK_a of Lys is 10.5 whereas that of Glu is 4.2. The large differences to the pK_a could also affect neighbouring residues and have an impact on stability. This effect would also apply to K18Q, H107Y, P236H and P236R. In H107Y a hydrophilic amino acid is replaced with an aromatic one, whereas in P236H and P236R aliphatic residues are replaced with hydrophilic ones. These substitutions could also have an effect on protein solubility and key residue-residue interactions.

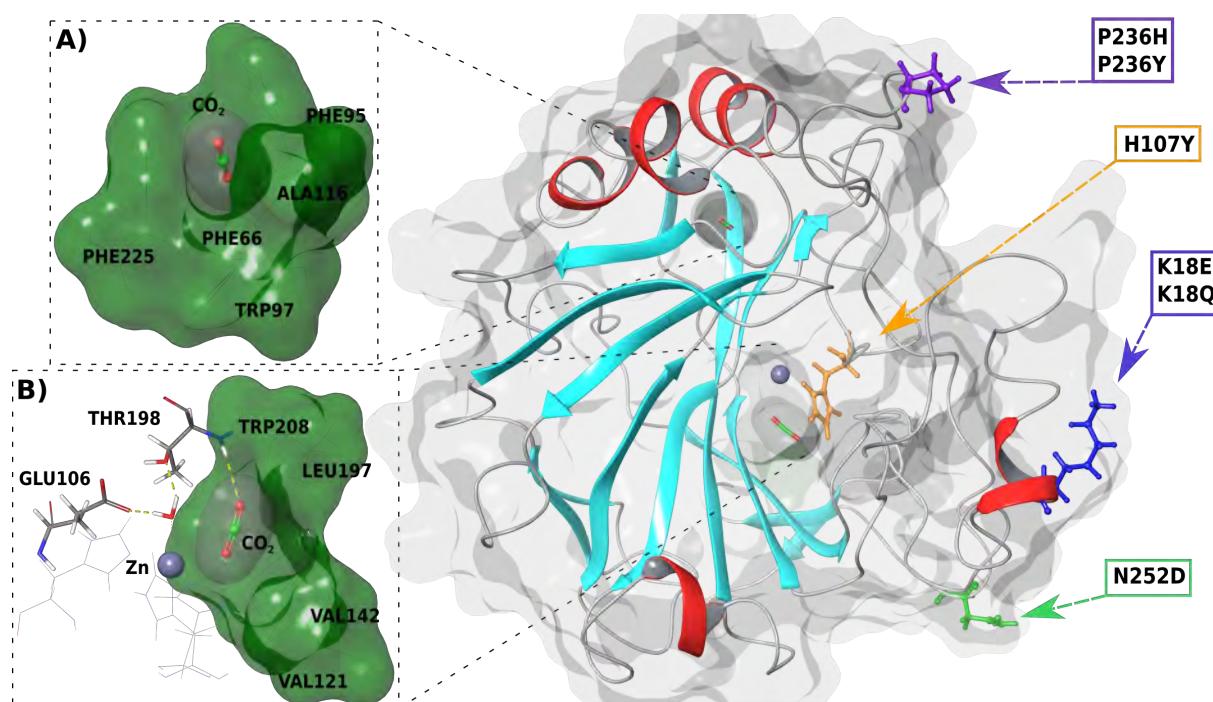


Figure 2.2. Illustration of the respective CA-II SNVs, and their proximity to the primary and secondary CO_2 binding pockets. A) Secondary CO_2 binding pocket. B) Primary CO_2 binding pocket. The Zn^{2+} is represented by the grey sphere. Adapted from Sanyanga *et al.* 2019 [66].

The CA-IV variants and their spatial location in relation to the active site are presented in Figure 2.3. Data indicates that none of the variants are located within the active site, but the majority are located close to the protein surface. Variants R219C, R219S and V234I are located on β -sheets whereas the other variants are located on loops. R69H is situated next to a β -sheet. Since R219C, R219S and V234I are located on β -sheets, but V234I is benign, this highlights at significant variant differences to

the mechanism of actions of pathogenic and benign SNVs. On first inspection, variant contribution towards pathogenesis could be due to the type of amino acid substitution. Arg is a positively charged polar amino acid. The substitutions Cys219 and Ser219 are polar and uncharged. Replacement of Arg at position 219 could have an effect on the salt bridges that the residue forms with its side chains. The loss of salt bridges could have detrimental effects to protein stability [266, 267]. Arg is also larger than Cys and Ser, therefore replacement with a smaller molecule would also mean that less electronegative atoms are available to form hydrogen bonds with the adjacent β -sheets which could affect protein stability.

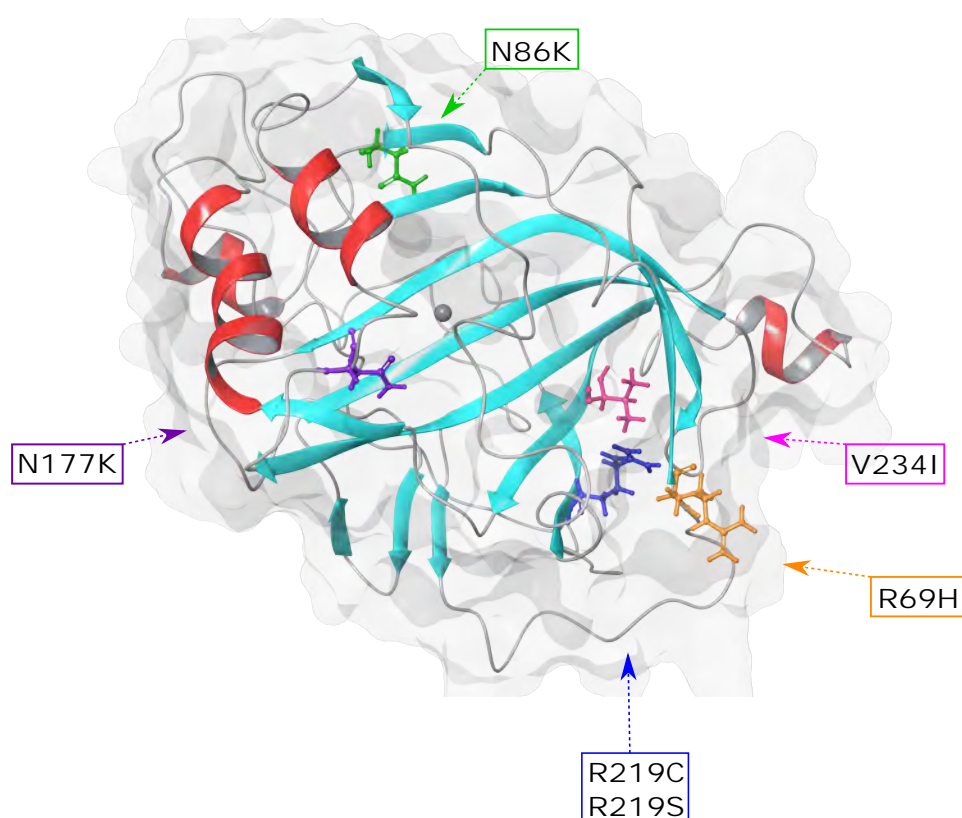


Figure 2.3. Illustration of the respective CA-IV SNVs, and their spatial location from the protein active site. The Zn^{2+} is represented by the grey sphere.

In R69H, a charged amino acid replaces another charged residue therefore potential salt bridge formation may not be affected. The pK_a s of Arg and His are 12.5 and 6.0 respectively. The difference to pK_a could have an effect on the behaviour of neighbouring residues and affect stability. N86K and N177K could be benign for two main reasons. Firstly, the substitution is located on loop secondary

structures as opposed to α -helices or β -sheets and could therefore have minimal effect. Secondly, as Lys residues are charged and replace an uncharged Asp, the Lys side chain could form salt bridges in CA-IV and assist with protein stability.

2.3.4 Association of CA-II and CA-IV With Membrane Carriers

Previous research highlighted the formation of transport metabolons between CA-II and CA-IV, and respective membrane carriers [55–61, 107, 108]. It should be noted that even without associating with the transport metabolons, CA-II and CA-IV are fully capable of hydrating CO₂ and dehydrating HCO₃⁻, and association with membrane carriers only has an effect on ion flux across the membrane and not catalysis. With respect to CA-II, this effect is not uniform as association with AE1 does not enhance ion flux [58]. With regards to CA-IV, contradicting research with regards to protein association with NBC1 and the pathogenesis of RP17 have been published [107, 119].

Due to the conflicting literature the association of CA-II and CA-IV with membrane carriers was not investigated. CA-VIII is the only protein being studied that requires interaction with a receptor to function, the effects of CA-VIII SNVs and binding site residue identification are discussed in the next section.

2.3.5 Potential CA-VIII and ITPR1 Association Residues Identified

With high quality protein models calculated, binding site investigations of the less commonly studied CA-VIII protein were then performed. The step was necessary to enhance the understanding of variant effects on CA-VIII. SiteMap [234, 235] and CPORT [236] analysis was performed to expand on the previous research by Hirota *et al.* [127] and identify the exact residues within the minimum binding site of CA-VIII (residues 44–290) that could associate with ITPR1.

SiteMap and CPORT analysis results are presented in Table S3. A total of five binding sites were identified by SiteMap, however only four binding sites has SiteScores >0.80 (Table S3). These four

binding sites each comprise of at least 30 amino acids located on the exterior surface of the protein. Binding site 1 comprised of the most residues. Mapping of SNVs to respective binding sites shows that SNVs are localised to binding sites 2 (R237Q) and 4 (S100A, S100L, E109D and S100P), with the exception to G162R that is not located within any binding site. CPORT analysis only discovered one binding site comprising of more than 30 residues as with SiteMap. From the CPORT results R237Q is the only variant located within the binding site.

As limited research has been performed towards the identification of the exact CA-VIII residues interacting with ITPR1, SiteMap and CPORT results were merged to obtain a consensus of all binding site residues (Table S3). The merging was preferred due to the large number of residues identified by SiteMap, and binding site 1 alone could not be used as it spans the entire protein. The consensus of the 38 identified binding site residues and their 3D spatial location is presented in Figure 2.4, and data indicates that R237Q is the only variant located within the binding site residues. The majority of the binding site amino acids are located between residues 44–290 which agrees with previous literature research [127]. This analysis has expanded on past studies whereby key amino acids between the residues 26–40 have also been identified as important binding site residues demonstrating the importance of the N-terminal residues. Cleavage of the first 43 N-terminal residues was noted to result in a 16-fold decrease to CA-VIII activity [127]. It is also observed that the N-terminal (green) and C-terminal (red) residues are situated within close proximity to each other. The CA-VIII region covering residues 150–157 was previously suggested to contain essential ITPR1 binding site residues in 2013 by Aspatwar *et al.* [268]. This research has confirmed their findings in that Gly151 and Ile153 are two binding site residues located within this region.

Though experimental studies have indicated that CA-VIII interacts with ITPR1, it is not known as to whether CA-VIII interacts with other cellular proteins therefore binding site residues could interact with other proteins. All residues have however been assumed to interact with ITPR1 for

the purposes of this study. The spheres in Figure 2.4 represent the binding site residues identified by SiteMap and CPORT. Residues include; Green: Gly26, Val27, Glu28, Trp29, Gly30, Tyr31, Glu32, Glu33, Gly34, Val35, Glu36, Leu39, Val40, Ala44; Blue: Leu93, Lys94, Glu111, Tyr113, Arg116, Ser147, Gly151, Ile153, Asp214, Ile224, Arg237, Tyr238; and Red: Thr255, His256, Leu262, Val263, Glu264, Gly265, Ile269, Phe274, Pro276, Gln278, Phe289, Gln290.

Through observations of the variant positions in relation to the binding site residues of CA-VIII potential effects of variants on protein structure were investigated. As R237Q is the only variant located within the binding site residues a direct variant effect to CA-VIII binding to ITPR1 is expected. This however does not rule out potential indirect effects to binding site residues that may be caused by the other variants. This hypothesis has been investigated within the rest of this thesis.

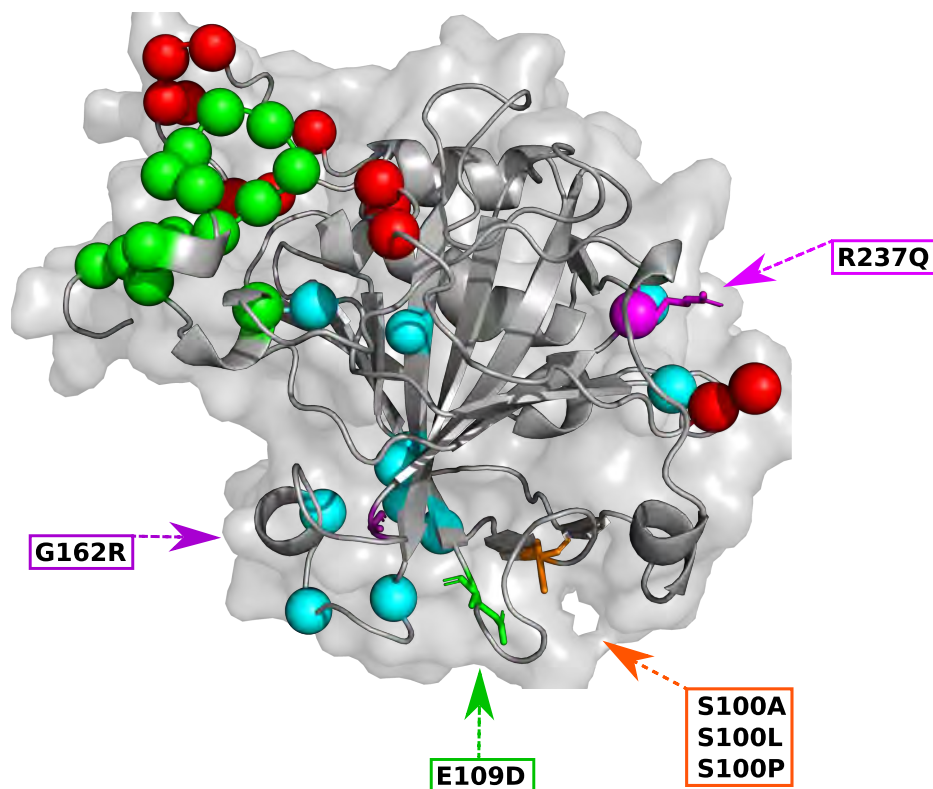


Figure 2.4. 3-dimensional image of CA-VIII showing the SNV location, predicted binding site residues and location of each motif. Orange: S100A, S100L and S100P; Green: E109D; Purple: G162R; Magenta: R237Q. Adapted from Sanyanga and Tastan Bishop 2020 [269].

Data in Figure 2.4 also presents the 3D spatial location of the SNVs on the CA-VIII protein. Variants S100A, S100L and S100P are located at the end of a β -sheet, whereas the remaining variants

are located on loops. Replacement of serine with proline at position 100 results in the complete destruction of the shorter adjacent (residues 71–73) and the respective β -sheets highlighting variant effects to protein secondary structure. Disruption at the secondary structure level would result in the loss of key hydrogen bonding between the β -sheets necessary to maintain protein stability. Previous research had suggested that loop residues 147–162 could be affected by the presence of S100P [268], however changes to secondary structure for this group of residues were not observed. This effect was further investigated using dynamic residue network (DRN) analysis in Chapter 4. Analysis of the E109D SNVs possibly explains the benign nature of this variant. Glu and Asp are both negatively charged therefore their substitution may not affect salt bridge formation. In addition, the pK_a s of Glu and Asp are 4.2 and 3.9 respectively making them almost similar. This suggests a minimal effect on neighbouring residues as the amino acids share similar physiochemical properties.

2.3.6 Structurally Important CA-IV and CA-VIII Residues Identified Via Sequence Analysis

As CA-II is well studied, an MSA of CA-II, CA-IV and CA-VIII (Figure S1) was performed to identify functionally important CA-IV and CA-VIII residues. Essential amino acids are expected to be conserved across the homologous proteins. Using CA-II as the reference sequence would allow for the identification of these residues. The mechanism of catalytic CAs and active site structure is similar making comparison of CA-II to CA-IV relatively simple. With regards to CA-VIII, the protein structure was aligned with CA-II prior to residue mapping to observe similarity. Superposition of CA-II and CA-VIII revealed an RMSD difference of 1.302 Å, showing structural similarity. Although the respective protein sequences share 40% sequence identity, structural similarity indicates that CA-II can be used to identify functionally and structurally important residues. The CA-II, CA-IV and CA-VIII functional and structurally important residues are presented in Table 2.2. The orange highlight shows amino acids conserved between CA-II and either CA-IV or CA-VIII. Due to the roles of the aromatic clusters in CA-II, aromatic amino acid substitutions were also regarded as conserved.

Table 2.2. Mapping of CA-II residues onto the CA-IV and CA-VIII proteins via MSA (Figure S1) and identified key residue function. Orange represents conserved residues. Adapted from Sanyanga *et al.* 2019 [66].

CA-II	Residue		Function in CA-II	Reference
	CA-IV	CA-VIII		
Trp5	Trp23	Trp29	Aromatic cluster residue (primary), His64 stabilisation for the "out" conformation and tertiary CO ₂ binding site formation	[19, 67, 68]
Tyr7	Tyr25	Tyr31	Aromatic cluster residue, active site water network coordination	[19, 48, 50, 62–65, 67, 68]
Trp16	Tyr34	Trp37	Aromatic cluster residue (primary)	[19, 67, 68]
Phe20	Val38	Phe41	Aromatic cluster residue (primary)	[19, 67, 68]
Ser29	Ser52	Ser50	Stability of enzyme	[270, 271]
Asn62	Asn86	Asp85	Active site water network coordination	[50]
His64	His88	His87	Proton shuttling residue	[19, 50, 62–65]
Phe66	Val90	Ile89	Aromatic cluster residue (secondary), secondary CO ₂ binding pocket formation	[19, 62–65]
Asn67	Met91	Gln90	Active site water network coordination	[20]
Phe70	Leu94	Leu93	Aromatic cluster residue (secondary)	[19, 67, 68]
Gln92	Gln113	Glu114	Secondary Zn ²⁺ ligand	[19, 50]
Phe93	Leu114	Val115	Aromatic cluster residue (secondary)	[19, 67, 68]
His94	His115	Arg116	Zn ²⁺ coordination residue	[19, 50]
Phe95	Leu116	Phe117	Aromatic cluster residue, secondary CO ₂ binding pocket formation	[19, 62–65]
His96	His117	His118	Zn ²⁺ coordination residue	[19, 50]
Trp97	Trp118	Trp119	Aromatic cluster residue (secondary), secondary CO ₂ binding pocket formation	[19, 62–65]
Glu106	Glu127	Glu128	Catalytic orientation of Zn ²⁺ water ligand molecule	[19]
Glu117	Glu138	Glu139	Zn ²⁺ affinity and catalytic efficiency, secondary Zn ²⁺ ligand	[272]
His119	His140	His141	Zn ²⁺ coordination residue	[19, 50]
Val121	Val142	Ile143	Primary CO ₂ binding pocket formation	[19, 62–65]
Val142	Val165	Ile165	Primary CO ₂ binding pocket formation	[19, 62–65]
Phe175	Met199	Ile198	Aromatic cluster residue (secondary)	[19, 67, 68]
Phe178	Ser202	Phe201	Aromatic cluster residue (secondary)	[19, 67, 68]
Leu197	Leu224	Leu222	Primary CO ₂ binding pocket formation	[19, 62–65]
Thr198	Thr225	Thr223	Orientation of Zn ²⁺ water ligand for catalysis, and deep water molecule stabilisation	[50]
Thr199	Thr226	Ile224	Active site water coordination, CO ₂ binding pocket formation (tertiary)	[50]
Pro200	Pro227	Pro225	Tertiary CO ₂ binding pocket formation	[19, 62–65]
Trp208	Trp235	Trp233	Primary CO ₂ binding pocket formation	[19, 62–65]
Phe225	Phe252	Phe250	Aromatic cluster residue (secondary), secondary CO ₂ binding pocket formation	[19, 62–65]
Asn243	Asn269	Asn273	Tertiary CO ₂ binding pocket formation, secondary Zn ²⁺ ligand	[19, 62–65]
Arg245	Arg271	Arg275	Enzyme stability	[273]

Analysis of results in Table 2.2 and the SNV locations shows that none of the selected variants occur at functionally or structurally important residues with exception to CA-IV_{N86K}. This suggests that variant action may occur via indirect or secondary mechanisms as opposed to direct effects for the rest of the protein variants. Results in Table 2.2 also shows that residues are divided into two broad categories; (A) those essential for maintaining stability such as aromatic cluster residues, and (B)

those essential to catalytic function such as the Zn^{2+} coordinating residues, active site water network residues and CO_2 binding site residues. Comparison of CA-II and CA-IV residue mapping highlights that not all primary and secondary aromatic cluster residues are conserved possibly indicating an adaptation of the protein residues to assist with function and cellular location stability. CA-IV residues identical to CA-II were therefore regarded as important. Due to the acatalytic nature of CA-VIII, the aromatic residues; Trp29, Tyr31, Trp37, Phe41, Phe117, Trp119, Phe201 and Phe250 were all regarded as essential for the maintenance of stability within the protein. These residues could perform a similar role as that of the primary and secondary aromatic cluster residues in CA-II. In addition, though not aromatic; Ser50, Leu93, Val115, Ile198 and Arg275 could assist with protein stability. The remaining CA-VIII residues in Table 2.2 could have an acatalytic adaptive function, however due to the limited research into acatalytic CAs these functions cannot be deduced from the MSA.

The CA-VIII Arg for His substitution at position 116 that makes the protein unable to coordinate Zn^{2+} and acatalytic in nature, is also observed in Table 2.2. Interestingly, mutagenesis studies have shown that replacing Arg116 with His116 assisted with the restoration of CO_2 hydration activity [274]. This suggests that Arg116 could have an adaptive role within the acatalytic CA-VIII, and is further supported by identification of Arg116 as a potential ITPR1 binding site residue in Figure 2.4. It is currently not known whether the other catalytically substituted residues in CA-VIII could have an adaptive role for acatalytic function. Further research is however required. From the study by Sjöblom *et al.* [274] it was also unclear as to whether the catalytic CA-VIII_{His116} mutant could be capable of associating with ITPR1, which would assist with the identification of adaptive residues changes that CA-VIII might possess to facilitate acatalytic function.

2.3.7 SNVs Are Located Around Or Within Conserved Motifs

Phylogenetic analysis has previously been performed for each CA-II, CA-IV and CA-VIII respectively [19, 145, 275]. The research however did not identify conserved residue regions that may be important to the structure and function of the CAs. Noting this research gap, motif analysis was performed to identify highly conserved residue regions within the proteins. As proteins function as networks of amino acids, motif discovery would expand on the residues in Table 2.2 and allow further identification of functionally important residue segments.

Key short linear motifs involved with PPIs are on average 3–11 residues long [248, 276]. Within the scope studied literature the CA proteins do not have a defined length for functional motifs, and as a result MEME parameters were set to identify motifs 3–20 residues long in order to include motifs that could be longer than the short linear average. The default MEME motif width is however 50 residues. From the 100 motifs set to be identified by MEME, pairwise correlation analysis using MAST reduced the dataset to 77 valid motifs. Results presented in Figure S2 represent a heat map of motif conservation in the human α -CA family expressed as the number of motif sites per total number of protein sequences. These motifs are numbered according to the MEME output. A value of 0 within the heat map indicates that the motif does not exist in any of the protein sequences, whereas a value of 1 is indicative of 100% motif conservation in all sequences.

The E-values of motifs in Figure S2 are demonstrated in Table S4. From the data in the table it is observed that motifs 1–11 have E-values less than 0.001, therefore this subset was selected for further analysis. As the acatalytic CA isoforms; CA-VIII, CA-X and CA-XI were included in the motif analysis, discovered motifs are most probably essential to the maintenance of protein stability and structure, as opposed to biological function. The motif logos showing the amino acid conservation within the 11 valid CA motif sequences are shown in Figure 2.5.

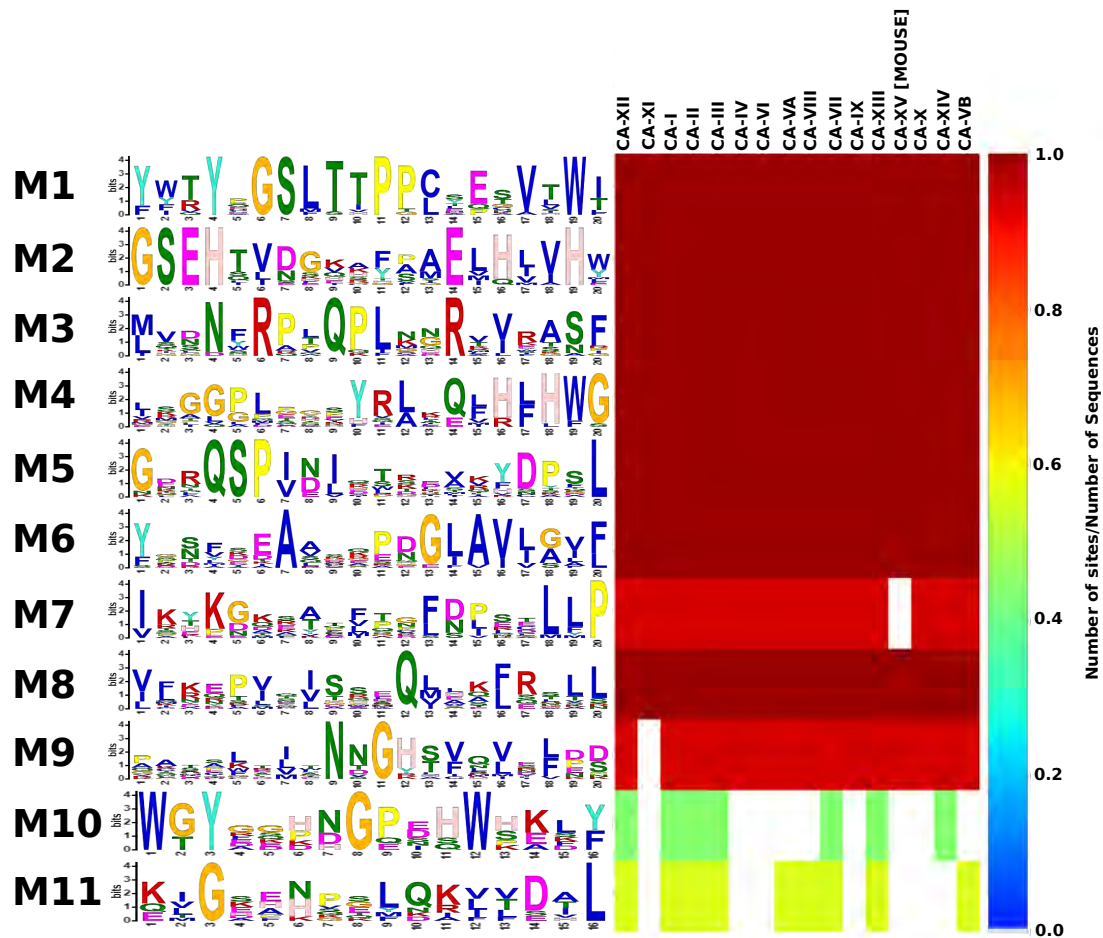


Figure 2.5. Motif logo demonstrating amino acid conservation in each of the valid motif sequences, and corresponding motif conservation. Adapted from Sanyanga *et al.* 2019 [66].

Data in Figure 2.5 demonstrates that across all α -CA proteins, motifs 1–6 and motif 8 are conserved. Conservation across all human α -CAs suggests that motif residues could be of significant importance to protein structure and function. Motif 7 and motif 9 are conserved in all CAs except for CA-XV and CA-XI respectively. Motif 10 is conserved in the catalytic cytosolic CA isoforms (CA-I, CA-II, CA-III, CA-VII and CA-XIII) and the membrane associated CA isoforms (CA-XII and CA-XIV). CA-XII and CA-XIV exist as single pass type I membrane proteins comprising of an extracellular N-terminus and a cytosolic C-terminus [40]. Existence of motif 10 in the cytosolic proteins could suggest that this motif is involved with protein solubility and/or stability within cells. In addition, it is noted that the acatalytic CA isoforms (CA-VIII, CA-X and CA-XI) do not contain this motif. This finding suggests that this motif could also have some role in catalytic function. Motif 11 is conserved in

all cytosolic catalytic α -CA isoforms and in the acatalytic CA-VIII isoform. Conservation in CA-VIII indicates that as opposed to this motif being essential to catalytic function, it could have a role in protein stability within the cellular environment.

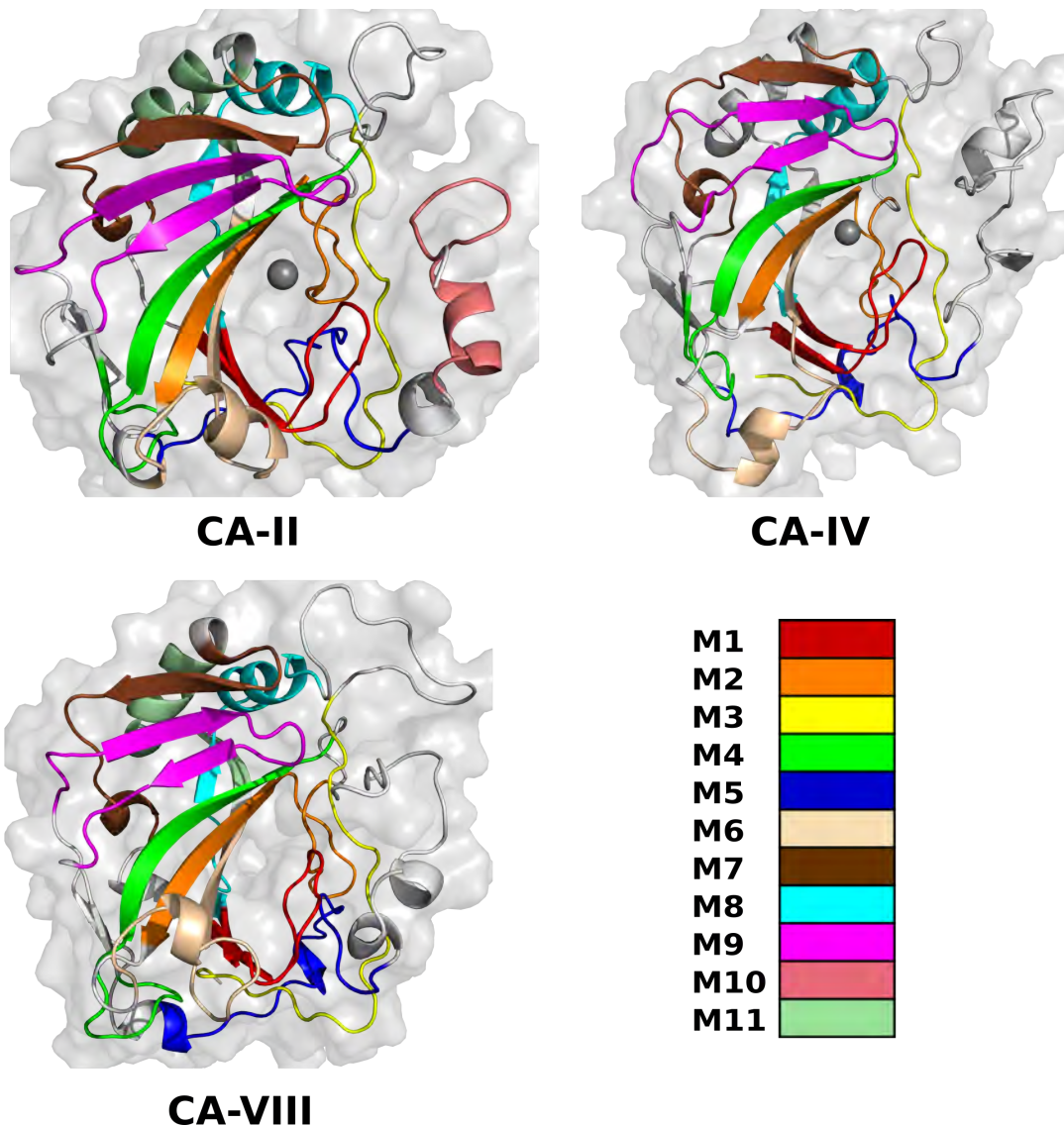


Figure 2.6. Motif mapping onto the 3D structure of the CA-II, CA-IV and CA-VIII proteins.

Motifs in Figure 2.5 were then mapped onto the 3D protein structures of CA-II, CA-IV and CA-VIII, and results are presented in Figure 2.6. The individual motifs are colour coded onto the respective protein. Variant presence on each motif was then analysed to determine the motifs that the SNVs are located on. Of interest is CA-VIII whereby the green binding site residues (Figure 2.4) are not located on any motif whereas the C-terminus binding site residues (red colour) are located on

motif 3. This result is however expected, as CA-VIII has been compared to other catalytic CAs. The terminal region is potentially important for binding to ITPR1 and not stability. Additionally, this could suggest that some of the red coloured binding site residues could be essential for CA-VIII stability, and this is supported by the residue mapping in Table 2.2. Motif function for each protein was assigned using the residue mapping data in Table 2.2 since proteins are dynamic in nature the amino acid residues are in constant communication with one another [18, 277]. The behaviour of specific residues is affected by the neighbouring residues. Residue mapping to assign motif function was conducted as follows, using motif 4 that covers CA-II residues 79–98 as an example. Motif 4 contains the Zn^{2+} coordination residues His94 and His96, and the secondary aromatic cluster residues Phe93, Phe95 and Trp97. Combining the function of these two residue groups suggests that motif 4 assists with enzymatic stability and Zn^{2+} coordination. Data in Table 2.3 presents the assigned motif functions of the 11 valid motifs. From the results it is noted that motif 3 in CA-IV covers the omega site for the GPI-anchor therefore this motif could also assist with its attachment.

SNV locations on each motif are also highlighted in Table 2.3. With regards to CA-II, K18E and K18Q are located on motif 10, H107Y is located on motif 2, and N252D is located on motif 3. The P236H and P236R variants are however located between motif 3 and motif 8 that are highly conserved. P236H has been found to have an effect on protein folding [278], and this could be due to an effect exerted on either of these two motifs. In CA-IV, N86K is located on motif 9 while R219C, R219S and V234I are located on motif 1. With respect to CA-VIII, E109D is located on motif 4, G162R is located on motif 6 and R237Q is located on motif 11. It should be noted that although the remaining CA-II, CA-IV and CA-VIII variants are not located on any motifs, SNVs could have an indirect effect on the motifs in Table 2.3. Of the catalytic CAs none of the variants are located on motifs essential to catalytic function with exception to N86K. Results suggest that variants could have a greater influence on CA stability as opposed to catalytic function. This suggestion is supported by the VAPOR result and

previous research [119, 125, 131, 271, 278, 279]. SNV effects on stability and catalytic mechanism were further investigated within this thesis.

2.4 CONCLUSION

This chapter presents sequence and structural analysis of the CA-II, CA-IV and CA-VIII proteins which were separated into two main parts, variant identification and CA protein characterisation. The Ensembl and HUMA databases identified a total of 18 variants (6 CA-II, 6 CA-IV and 6 CA-VIII) associated with phenotype annotations, pathogenic and benign for comparative purposes. Next CA proteins were characterised. Essential residues to CA-II have been identified through previous literature. This allowed for the identification of potentially important residues within the less extensively studied CA-IV and CA-VIII. This is however not without some limitations. Each CA has adapted to function within its cellular environment therefore some amino acids essential to this may not be conserved. This obscures potential variant effects as variants are not located on any residues essential to catalysis or stability within either protein, potentially highlighting at allosteric mechanisms of action. Motif analysis was employed to identify conserved sequence patterns within the 16 α -CA family proteins, and investigate whether SNVs were located on these motifs. A total of 11 motifs were identified, and results further highlighted at allosteric SNV mechanisms of action as not all variants are located on conserved motifs. The lack of essential residue knowledge also greatly limits the analyses that can be performed with regards to each protein.

In the following chapter we explore Zn^{2+} cofactor parameter generation to allow for metalloprotein molecular dynamics (MD) simulations of proteins to investigate variant mechanisms of action.

Table 2.3. Motif residue ranges within the CA-II, CA-IV and CA-VIII proteins, and amino acid sequence. Residues from Table 2.2 are underlined and highlighted in bold. SNVs for each protein are underlined, italicised and presented in bold red. Adapted from Sanyanga *et al.* [66].

Motif	CA-II		CA-IV		CA-VIII		Contribution to function according to CA-II
	Residue range	Residues	Residue range	Residues	Residue range	Residues	
1	190-209	<u>YW</u> <u>TY</u> <u>PG</u> <u>SL</u> <u>TP</u> <u>PL</u> <u>LE</u> <u>CV</u> <u>T</u> <u>WI</u>	217-236	<u>Y</u> <u>R</u> <u>Y</u> <u>L</u> <u>G</u> <u>S</u> <u>L</u> <u>T</u> <u>P</u> <u>TC</u> <u>DE</u> <u>KV</u> <u>V</u> <u>WT</u>	215-234	<u>YW</u> <u>VY</u> <u>EG</u> <u>SL</u> <u>TP</u> <u>PC</u> <u>SE</u> <u>GV</u> <u>T</u> <u>WI</u>	CO ₂ binding pocket formation (primary) and catalysis
2	104-123	<u>GSE</u> <u>H</u> <u>TV</u> <u>D</u> <u>KK</u> <u>KY</u> <u>AA</u> <u>EL</u> <u>HL</u> <u>V</u> <u>HW</u>	125-144	<u>GSE</u> <u>H</u> <u>SL</u> <u>D</u> <u>GE</u> <u>H</u> <u>F</u> <u>AM</u> <u>EM</u> <u>H</u> <u>I</u> <u>V</u> <u>HE</u>	126-145	<u>GSE</u> <u>HT</u> <u>VN</u> <u>FK</u> <u>AF</u> <u>PM</u> <u>EL</u> <u>HL</u> <u>LI</u> <u>HW</u>	Active site and/or Zn ²⁺ stability
3	240-259	<u>MVD</u> <u>N</u> <u>WR</u> <u>PA</u> <u>QP</u> <u>L</u> <u>K</u> <u>NR</u> <u>Q</u> <u>IK</u> <u>AS</u> <u>F</u>	266-285	<u>MKD</u> <u>N</u> <u>VR</u> <u>PL</u> <u>QQ</u> <u>L</u> <u>GQR</u> <u>T</u> <u>VI</u> <u>K</u> <u>S</u> <u>G</u>	270-289	<u>LGD</u> <u>N</u> <u>FR</u> <u>PT</u> <u>QP</u> <u>L</u> <u>SDR</u> <u>V</u> <u>IR</u> <u>AA</u> <u>F</u>	CO ₂ binding pocket formation (tertiary) and stability
4	79-98	<u>LKG</u> <u>GL</u> <u>D</u> <u>G</u> <u>TYR</u> <u>LI</u> <u>Q</u> <u>F</u> <u>H</u> <u>F</u> <u>H</u> <u>W</u> <u>G</u>	100-119	<u>ISG</u> <u>GL</u> <u>P</u> <u>AP</u> <u>Y</u> <u>Q</u> <u>AK</u> <u>Q</u> <u>L</u> <u>H</u> <u>L</u> <u>H</u> <u>W</u> <u>S</u>	101-120	<u>GG</u> <u>PL</u> <u>P</u> <u>Q</u> <u>GH</u> <u>FE</u> <u>LY</u> <u>EV</u> <u>RF</u> <u>H</u> <u>W</u> <u>G</u>	Zn ²⁺ coordination and stability
5	25-44	<u>GER</u> <u>Q</u> <u>SP</u> <u>VD</u> <u>ID</u> <u>I</u> <u>T</u> <u>HT</u> <u>AK</u> <u>Y</u> <u>D</u> <u>P</u> <u>S</u> <u>L</u>	48-67	<u>KDR</u> <u>Q</u> <u>SP</u> <u>IN</u> <u>I</u> <u>V</u> <u>T</u> <u>TK</u> <u>AK</u> <u>V</u> <u>D</u> <u>K</u> <u>KL</u>	46-65	<u>GE</u> <u>Y</u> <u>Q</u> <u>SP</u> <u>IN</u> <u>LN</u> <u>S</u> <u>RE</u> <u>AR</u> <u>Y</u> <u>D</u> <u>P</u> <u>S</u> <u>L</u>	Stability
6	127-146	<u>YGD</u> <u>FG</u> <u>K</u> <u>AV</u> <u>QQ</u> <u>P</u> <u>DGL</u> <u>AV</u> <u>L</u> <u>G</u> <u>IF</u>	150-169	<u>SR</u> <u>NV</u> <u>KE</u> <u>AQ</u> <u>D</u> <u>PE</u> <u>DE</u> <u>IA</u> <u>V</u> <u>L</u> <u>A</u> <u>FL</u>	150-169	<u>FG</u> <u>SIDE</u> <u>AV</u> <u>G</u> <u>K</u> <u>P</u> <u>H</u> <u>G</u> <u>IA</u> <u>IA</u> <u>I</u> <u>AL</u> <u>F</u>	CO ₂ binding pocket (primary) formation
7	166-185	<u>IK</u> <u>T</u> <u>K</u> <u>G</u> <u>K</u> <u>S</u> <u>AD</u> <u>F</u> <u>T</u> <u>N</u> <u>F</u> <u>D</u> <u>P</u> <u>R</u> <u>G</u> <u>L</u> <u>L</u> <u>P</u>	190-209	<u>IP</u> <u>K</u> <u>PE</u> <u>M</u> <u>S</u> <u>T</u> <u>T</u> <u>M</u> <u>A</u> <u>ES</u> <u>LL</u> <u>D</u> <u>L</u> <u>L</u> <u>P</u>	189-208	<u>I</u> <u>Q</u> <u>Y</u> <u>K</u> <u>G</u> <u>K</u> <u>S</u> <u>K</u> <u>T</u> <u>I</u> <u>P</u> <u>CF</u> <u>NP</u> <u>N</u> <u>T</u> <u>I</u> <u>L</u> <u>L</u> <u>P</u>	Participated in secondary aromatic cluster
8	210-229	<u>VL</u> <u>KE</u> <u>P</u> <u>IS</u> <u>V</u> <u>S</u> <u>SE</u> <u>Q</u> <u>V</u> <u>L</u> <u>K</u> <u>F</u> <u>R</u> <u>K</u> <u>L</u> <u>N</u>	237-256	<u>V</u> <u>F</u> <u>R</u> <u>E</u> <u>P</u> <u>I</u> <u>Q</u> <u>L</u> <u>H</u> <u>R</u> <u>E</u> <u>Q</u> <u>I</u> <u>L</u> <u>A</u> <u>F</u> <u>S</u> <u>Q</u> <u>K</u> <u>L</u>	235-254	<u>L</u> <u>F</u> <u>R</u> <u>Y</u> <u>P</u> <u>L</u> <u>T</u> <u>I</u> <u>S</u> <u>Q</u> <u>L</u> <u>Q</u> <u>J</u> <u>E</u> <u>F</u> <u>F</u> <u>R</u> <u>R</u> <u>L</u> <u>R</u>	CO ₂ binding pocket formation (secondary) and stability
9	53-72	<u>Q</u> <u>A</u> <u>T</u> <u>S</u> <u>L</u> <u>R</u> <u>I</u> <u>L</u> <u>N</u> <u>N</u> <u>G</u> <u>H</u> <u>A</u> <u>F</u> <u>N</u> <u>V</u> <u>E</u> <u>F</u> <u>E</u> <u>D</u> <u>D</u>	77-96	<u>KK</u> <u>Q</u> <u>T</u> <u>W</u> <u>T</u> <u>V</u> <u>Q</u> <u>N</u> <u>V</u> <u>G</u> <u>H</u> <u>S</u> <u>V</u> <u>M</u> <u>M</u> <u>L</u> <u>L</u> <u>E</u> <u>N</u>	76-95	<u>V</u> <u>C</u> <u>R</u> <u>D</u> <u>C</u> <u>E</u> <u>V</u> <u>T</u> <u>N</u> <u>D</u> <u>G</u> <u>H</u> <u>T</u> <u>I</u> <u>Q</u> <u>V</u> <u>I</u> <u>L</u> <u>K</u> <u>S</u>	Stability and catalysis
10	5-20	<u>W</u> <u>G</u> <u>Y</u> <u>G</u> <u>K</u> <u>H</u> <u>N</u> <u>G</u> <u>P</u> <u>E</u> <u>H</u> <u>W</u> <u>H</u> <u>K</u> <u>D</u> <u>F</u>	-	-	-	-	Stability
11	148-163	<u>K</u> <u>V</u> <u>G</u> <u>S</u> <u>A</u> <u>K</u> <u>P</u> <u>G</u> <u>L</u> <u>Q</u> <u>K</u> <u>V</u> <u>V</u> <u>D</u> <u>V</u> <u>L</u>	-	-	171-186	<u>Q</u> <u>I</u> <u>G</u> <u>K</u> <u>E</u> <u>H</u> <u>V</u> <u>G</u> <u>L</u> <u>K</u> <u>A</u> <u>V</u> <u>T</u> <u>E</u> <u>I</u> <u>L</u>	Stability

Only those who attempt the absurd can achieve the impossible.

Albert Einstein

3

Zn²⁺ Force Field Parameter Generation

CHAPTER OVERVIEW

The vast majority of biological processes occurring within cells are facilitated by proteins of which approximately 30–40% require a metal cofactor for function [280, 281]. The roles of metal ions in biological functions include; cross linking agents due to the ability to bind numerous ligands, redox reactions through multiple oxidation states, structural roles and substrate activation [282]. As a result of their importance to catalytic function, metals have to be considered and correctly modelled prior to MD simulations as their accuracy is dependent on this. This chapter introduces Zn²⁺ ion modelling and force field (FF) parameter derivation for CA-II that will be utilised for MD simulations within the next chapter. The derived FF parameters will govern the manner in which Zn²⁺ interacts with other amino acid residues within the CA proteins, and prevent active site escape during MD simulation.

3.1 INTRODUCTION

CA-II and CA-IV are the catalytic CAs that require a Zn^{2+} cofactor to function, unlike CA-VIII. The Zn^{2+} is critical to the hydration of CO_2 and dehydration of HCO_3^- . It is therefore imperative to consider metal ions when preparing metalloproteins for MD simulations.

3.1.1 Force Fields

In MM and MD simulations a FF is defined as the Hamiltonian and set of parameters used to calculate potential energy [245, 283–285]. These parameter sets include bond lengths, bond angles and dihedrals between specific groups of molecules, and govern their intra and inter-molecular interactions [245, 286]. Numerous FF types have been developed, and depend on the respective molecules including; protein, DNA (deoxyribonucleic acid), RNA (ribonucleic acid), carbohydrates, lipids and a general FF (for organic molecules) [245]. Although the FFs cover a great number of biological molecules, even when combined, there are limitations to supported atoms, making FF parameter derivation necessary for metal cofactors.

3.1.2 Protein Force Fields

Numerous protein FFs exist for the analysis of MM and MD simulations involving these types of atoms [286]. From these, the most common FFs are:

1. AMBER (Assisted Model Building with Energy Refinement) [287]
2. CHARMM (Chemistry at Harvard Macromolecular Mechanics) [288]
3. OPLS (Optimized Potentials for Liquid Simulations) [289]
4. GROMOS (GRONingen MOlecular Simulation) [290]

It should be noted that the FFs above are not ranked, and come bundled with their respective MD software therefore it is essential to consider compatibility prior to simulations. These protein FFs are

also not capable of handling metal ion cofactors such as Zn^{2+} hence custom parameter generation is necessary. Therefore in addition to compatibility, the ease of custom FF parameter imports should also be considered when selecting the FF and MD simulation program. The respective FF and corresponding MD simulation programs have been briefly described below.

The GROMACS simulation package [291] is capable of implementing the listed protein FFs and is optimised for performance, however, offers limited flexibility for custom FFs. Schrödinger Desmond [292] implements OPLS, and although metalloprotein simulations are possible [293, 294], both Desmond and Epik [295] for Het state generation require a license. AMBER [245] and CHARMM [296] FFs are each implemented by their respective programs. Both these programs offer support for custom FFs, but are not optimised as well GROMACS. Through ACPYPE (AnteChamber PYthon Parser interfacE) [297] AMBER topologies can be converted to GROMACS format to take advantage of the optimisation. Of the previously mentioned programs, Desmond is the only one that comes with inbuilt support of the CA Zn^{2+} cofactor. Comparison of the above suggests AMBER to be most ideal for metal ion modelling and FF parameter generation due to its free license and ability to port to GROMACS offering excellent customisability [66].

3.1.2.1 AMBER Force Field

In this FF the addition of bonded and nonbonded interaction energy gives the total potential energy of the molecule [245, 287]. This relationship is presented in Equation 3.1.

$$V_{total} = V_{bonded} + V_{nonbonded}$$

Equation 3.1. Relationship between total potential energy, and bonded and nonbonded interactions.

Bonded interactions refer to covalent linkages of atoms including; bonds, angles and dihedrals,

whereas nonbonded interactions refer to non-covalent linkages such as long range electrostatic and van der Waals (vdW) forces. The Hamiltonian of the AMBER FF takes on the basic form illustrated in Equation 3.2. The first three terms refer to the bonded interactions (bond, angle and dihedral respectively). The fourth and fifth terms denote the electrostatic and vdW interactions respectively. Molecule forces and velocities are described by the integration of this Hamilton during MD.

$$V_{AMBER} = \sum_i^{n_{bonds}} k_r (r_i - r_{i,eq})^2 + \sum_i^{n_{angles}} k_\theta (\theta_i - \theta_{i,eq})^2 + \sum_i^{n_{dihedrals}} \sum_n^{n_{i,max}} \frac{V_{i,n}}{2} [1 + \cos(n\phi_i - \gamma_{i,n})] \\ + \sum_{i<j}^{n_{atoms}} \frac{q_i q_j}{4\pi\epsilon_0 r_{i,j}} + \sum_{i<j}^{n_{atoms}} \left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^6} \right)$$

Equation 3.2. Basic Hamiltonian for potential energy calculations using the basic AMBER force field.

For the bonded interactions, the i and j represent pairs of atoms. The k_r , r_i and $r_{i,eq}$ represent the bond stretching force constant, bond distance and equilibrium bond distance, whereas k_θ , θ_i and $\theta_{i,eq}$ denote the angle bending force constant, bond angle and equilibrium bond angle respectively. The $V_{i,n}$, n , ϕ and $\gamma_{i,n}$ represent the barrier to free rotation, rotation periodicity, dihedral angle and phase shift. With respect to the nonbonded terms, $r_{i,j}$ denotes the interatomic distance between atoms i and j . The q_i and q_j represent atom point charges, whereas ϵ_0 denotes the dielectric constant. In the final term representing vdW interactions, $A_{i,j}$ and $B_{i,j}$ control potential energy well depth and position for an atom pair. The relationships between the Hamiltonian terms are illustrated in Figure 3.1.

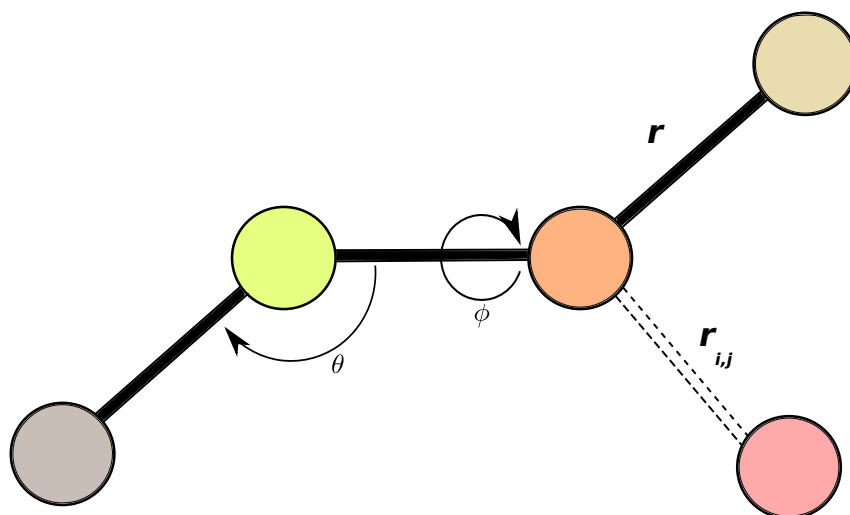


Figure 3.1. Illustration of the relationship between terms in the AMBER Hamiltonian. Different coloured spheres indicate various atoms. Solid and dashed lines represent bonded and nonbonded interactions respectively.

Currently the latest and most improved AMBER protein FF is ff14SB which expands on the older ff99SB [298] and its modified version (ff12SB) [299, 300]. The ff94 [287] and ff99 [301] FF lacked a good energy balance between peptide and protein backbones, and helical regions. Glycine backbone parameters were also incorrectly treated, and helical systems were over-stabilised by ff94 FF variants [302]. The ff99SB was introduced to correct these shortcomings and again later improved in ff12SB. These FFs however still contained limitations to side chain dihedral and backbone parameters [303], which were subsequently improved in the ff14SB forcefield thereby increasing accuracy and reproducibility.

3.1.3 General AMBER Force Field

The general AMBER FF (GAFF) includes a set of parameters for organic molecules such as ligands. The AMBER protein FFs cannot accommodate ligands therefore an additional set of FF parameters is required. The two main FF that were developed to handle most pharmaceutical molecules are GAFF and GAFF2 [304]. GAFF contains parameter set for at least 50 special atom types [245, 304]. Carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), sulphur (S), fluorine (F), chlorine

(Cl), bromine (Br) and iodine (I) make up the organic chemical space. GAFF2 was created as an improvement for GAFF and reduces RMS errors by approximately 50% improving reproducibility.

The combination of the ff14SB and GAFF FFs allows AMBER to cover complex protein systems and the FFs do not cause any conflicts.

3.1.4 Extending the AMBER Force Fields

Although combination of the ff14SB and GAFF offers great support for complex protein systems, it lacks parameters for the metallic cofactors. To facilitate MD of metalloproteins, the AMBER FFs require extending to add support for Zn^{2+} . Parameters developed for a specific coordination geometry such as tetrahedral can be inferred onto other metalloproteins of similar coordination geometry thus extending the FF.

As parameter sets include bonded and nonbonded terms (Equation 3.1), there are three main models of extending the AMBER FF. [247, 305, 306]. These are described below:

1. Bonded model: The bonded model involves the definition of explicit bonds between the metal ion and ligands and treating them as covalent bonds. Nonbonded potentials via the determination of formal or quantum mechanically calculated charges are usually added to this model. Coordination number and ligand changes cannot be simulated using this model though.
2. Nonbonded Model: In this model all interactions between the metal and coordinating ligands are treated as nonbonded. Oxidation state is used to refer to metal ion charge. These interactions are described by the 12-6 and 12-6-4 Lennard-Jones (LJ) parameters. Both parameters offer excellent transferability from monovalent atoms to tetravalent ions [305, 307, 308]. Multiple binding modes can be accommodated by the nonbonded model giving it an advantage over the bonded module.
3. Cationic dummy atom model: Within this model, covalent bonds between the metal and coordinating atoms are mimicked through the placement of charges or dummy atoms between the metal ion and ligands. The numerous empirical parameters involved with this model indicates that it requires high intricacies during parametrisation.

Upon comparison of the models it is worth noting that nonbonded terms are associated with long range electrostatics and vdW forces therefore the metal ion may not be held firmly in place during

MD and could escape. Within the bonded model the metal ion is less likely to escape. As no catalysis or changes to Zn^{2+} coordination occur within the nanosecond (ns) time scale usually simulated, the bonded model is the most ideal metal ion modelling for AMBER FF extension.

3.1.4.1 Metal Center Parameter Builder

Metalloprotein molecular models can be efficiently constructed, prototyped and validated using the Metal Center Parameter Builder (MCPB) program [247, 309]. MCPB is capable of building over 80 metalloprotein models through implementation of the bonded model [309]. In addition to AMBER, parameter sets generated by MCPB also fit the FF functional form of CHARMM. The parametrisation process uses *ab initio* calculations to determine AMBER-like FF parameters for the first metal coordination sphere in two broad main parts summarised in Figure 3.2. Bonded term (r and θ) parameters are initially determined. Metal-ligand torsion barriers are however lower than kT (where k is the Boltzmann constant and T the temperature), therefore dihedrals are omitted [247, 309]. Secondly, the electrostatic term point charges are obtained. Since most metals are buried within proteins, LJ parameters are also not parametrised. Electrostatic interactions tend to be more important than the vdW type.

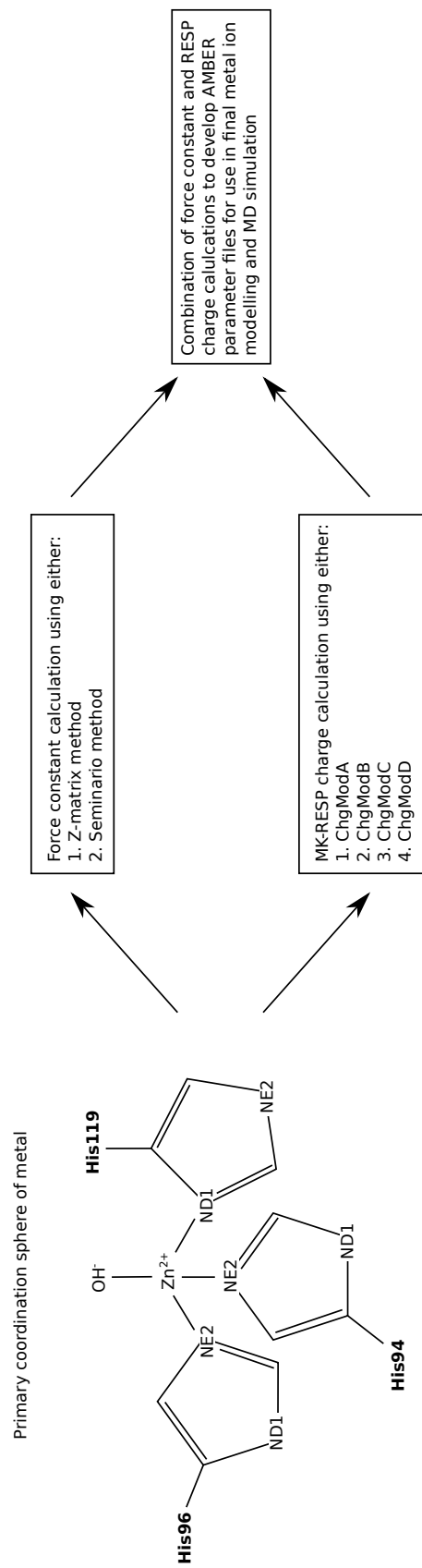


Figure 3.2. Summary of MCPB workflow of Zn^{2+} parametrisation and metal ion modelling.

Force constant determination k_r and k_θ can be performed using either one of two methods:

- Z-Matrix method
- Seminario method

Both of the above methods make use of a Cartesian Hessian matrix to attain a balance of accuracy and speed during parametrisation. Within the Z-matrix method k_r and k_θ are calculated from the Cartesian Hessian matrix, whereas the Seminario method [310] calculates the FF force constants using submatrices of the Cartesian Hessian Matrix [247, 310]. Both methods take on the basic equations indicated below. The Hessian is represented by $[k]$ which is a tensor of dimensions $3N \times 3N$ and rank 2, and represents the intramolecular FF to second order of small molecule displacements (δx) as illustrated in Equation 3.3. The $[k]$ can be obtained by *ab initio* programs such as Gaussian 09 (G09) [311]. Force constant calculation is performed through the use of second derivative of energy with respect to coordinates in the form of a Cartesian Hessian matrix (Equation 3.4).

$$\delta F = -[k]\delta x$$

Equation 3.3. δx displacement of a molecular systems N atoms giving rise to force δF .

$$[k] = k_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j}$$

Equation 3.4. Cartesian Hessian matrix calculation.

$$F_i = -[k]\hat{v}_i \delta d = -\lambda_i \hat{v}_i \delta r$$

Equation 3.5. Eigen analysis of k to determine force constants, eigenvalues λ_i and eigenvectors \hat{v}_i .

Force constants, eigenvalue (λ_i) and the eigenvectors (\hat{v}_i) are provided by the eigen analysis of k (Equation 3.5). The \hat{v}_i also represents the normal modes. If the \hat{v}_i and δr (magnitude of displacement) share the same direction, the reaction force (F_i) will always act anti-parallel or parallel. Due to the large

number of individual elements in $[k]$, its direct use during MD simulation may be unmanageable. The number of terms can be reduced by relating the FF to the internal coordinates, giving rise to the Hamiltonian in Equation 3.6. The r , θ , φ and ω represent bonds, angles, dihedral angles and improper dihedrals (out of plane).

$$V = \sum_{bonds} \frac{1}{2} k_r (r - r_{eq})^2 + \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{1}{2} k_\varphi (\varphi - \varphi_{eq})^2 + \sum_{impropers} \frac{1}{2} k_\omega (\omega - \omega_{eq})^2$$

Equation 3.6. Potential energy function resulting from the relation of force field to internal coordinates.

The Z-matrix or Seminario derived parameters are then imported and analysed by MCPB and the respective bonds and angles assigned their values [309]. The Merz-Singh-Kollman (MK) [312] and restrained electrostatic potential (RESP) [313–315] methods are used to calculate the partial charges of the metal centre and the surrounding atoms. The vdW atomic radii are however obtained from literature. The MK-RESP method is preferred as it allows for the adjustment of charges in four methods [247] namely:

1. ChgModA: All ligating residue charges can be changed
2. ChgModB: Backbone heavy atom (C, CA, N, O) charges are restrained to those in the AMBER ff94 FF
3. ChgModC: Backbone atom (C, CA, H, HA, HN, N, O) charges are restrained to those in the AMBER ff94 FF
4. ChgModD: The CB atoms are also restrained in addition to all backbone atoms

Compared to the Z-matrix method, the Seminario method does not define internal coordinates thereby offering it an advantage over the Z-matrix method. The combinations of the Seminario and either the ChgModB or ChgModD gives the best results in that order [247]. As a result the Seminario ChgModB approach was preferred for this thesis.

After parametrisation minimisation techniques are used to verify FF stability. It should be noted that custom residue names are given to the parametrised residues to avoid conflicts with the built in FFs.

3.2 METHODOLOGY

This section explains the Zn^{2+} metal ion modelling of CA-II to extend the AMBER ff14SB FF parameter set to include the Zn^{2+} cofactor. The preparation was performed using scripts bundled in together with AmberTools17 [245]. The parametrisation, MM and QM calculations were performed using G09 [311]. This was necessary as Zn^{2+} is important for the catalytic functions of CA and needs to be maintained within the active site.

3.2.1 Protein Preparation

Uploading of crystal structure 2VVA to the H++ server [316] for protonation produced errors and crashes on the server that could not be solved via renumbering or remodelling using MODELLER. As a result, this protein could not be selected as the template for metal parametrisation. An additional template 4WL4 (HHHX coordination geometry) of resolution 1.1 Å was therefore selected from the previously generated environment and summary files (see subsection 2.2.4.1), and downloaded.

To the downloaded 4WL4 crystal structure, all HETATMS excluding the coordinating H_2O and Zn^{2+} were removed. The remaining ATOM records, coordinating H_2O and Zn^{2+} HETATMs were then separated into three new PDB files using *awk* scripting. The extracted ATOM records were renumbered using *pdb4amber* to correct the inconsistent numbering caused by the missing residue 126. The renumbered PDB file was then uploaded to the H++ server for protonation. The protonation was set at pH 7.0 with a system salinity of 0.15 M. The external and internal dielectric were set to 80 and 10 respectively. After protonation, AMBER topology and coordinate files (**.top* and **.crd*) were then downloaded from the server. The H++ server will generate an error if the ATOM sequence contains missing residues, or if the protein structure is already protonated. All HETATMs within the PDB file are also removed during protonation.

The downloaded AMBER topology files were then processed using *ambpdb* to generate a

protonated PDB file of CA-II. The separate Zn^{2+} containing PDB file was then concatenated together with the protonated CA-II PDB and the complex saved separately. The protonation state of the Zn^{2+} ligands was then visually inspected using Schrödinger Maestro [253] to ensure that the residues; His94, His96 and His119 were in the correct protonation states (HID, HID, and HIE respectively). If not, the erroneous atoms were deleted and resultant structure saved.

3.2.2 Zn^{2+} Parametrisation

Metal ion parametrisation was performed using the Seminario and ChgModB methods. To the correctly protonated structure, MCPB [309] was used to detect all atoms within 2.5 Å of the Zn^{2+} and generate the G09 input files (**opt.com*, **fc.com* and **large.com*). The **opt*, **fc* and **large.com* files are important for the optimisation, force constant and MK-RESP charge calculation respectively during QM. GaussView 5 (GView 5) [317] was used to visualise the generated input files to ensure that all coordinating residues were identified, and within the correct orientation. The input files were then opened using a text editor and the “%MEM” and “%NProcShared” directives manually changed to 10 000 MB and 72 respectively. QM calculations using the G09 B3LYP/6-31G basis set on 192 CPU cores was then performed at the CHPC (Center for High Performance Computing) cluster, Cape Town South Africa.

After optimisation, the resulting **.log* files and intermediate geometries were visualised using GView 5 to ensure that no bond breakage had occurred and that optimisation was complete. The *formchk* command was then used to generate the final **.fchk* files using the optimised G09 check point file (**opt.chk*). MCPB was utilised to calculate the bond lengths, angles and dihedrals between Zn^{2+} and the coordinating atoms, and to derive the final FF parameters for Zn^{2+} which were written out to a parameter modification file (**.frcmod*). MCPB additionally generates **.mol2* files containing the atoms and associated RESP charges for each of the coordinating atoms to complement the parameters (His94: HD1.mol2; His96: HD2.mol2; His119: HE1.mol2).

3.3 RESULTS AND DISCUSSION

Due to the importance of Zn^{2+} in CAs, metal ion parametrisation was performed to extend the built-in FFs and add metal ion support. Parametrisation would also be invaluable to DRN analysis as variant presence could have an effect on the metal ion and coordinating residues which was investigated in Chapter 4.

QM optimisation of the first Zn^{2+} coordination sphere completed in 87 steps. Results from the optimisations are presented in Figure 3.3, and data demonstrates that the system converges and no changes to total energy occur within the last 10 steps. The optimised structures at each step during geometry optimisation were then manually inspected across all 87 steps for errors. No evidence of bond breakage between Zn^{2+} and the coordinating residues was observed and in addition, the metal ion remained coordinated in the correct orientation over all steps suggesting that the derived FF parameters were accurate. Within Chapter 4, MD was used to further investigate whether derived parameters would be able to maintain the Zn^{2+} within the active site.

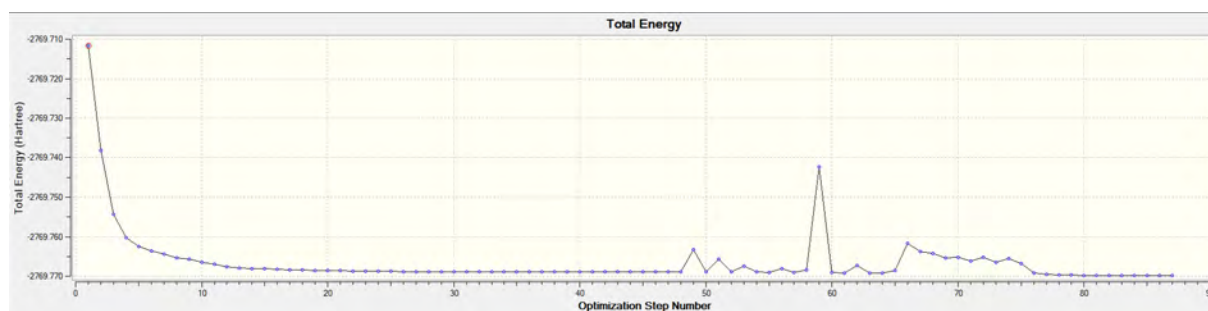


Figure 3.3. Geometry optimisation of the first Zn^{2+} coordination sphere during metal ion parametrisation Gaussian QM calculations.

After ensuring accurate optimisation, bond length, angle and dihedrals were then calculated using MCPB and results are presented in Table 3.1. Generated MCPB parameters were compared to those calculated in 2006 by Harding [246], and 2013 by Bernadat et al. [318]. Bond length measurements between Zn^{2+} and coordinating residues were within the reported threshold of His $< 2.03 \text{ \AA}$ and $\text{H}_2\text{O} < 2.18 \text{ \AA}$ [246, 247]. Calculated bond angles were also similar to previous findings [318].

Table 3.1. Zn^{2+} non-bonded, bonds, angles and dihedral parameters derived within this study. K_b : bond force constant; K_θ : angle force constant; R_{min} : vdW radius; ϵ : LJ potential well energy. Adapted from Sanyanga *et al.* 2019.

Non-bonded:

Atom	R_{min} (Å)	ϵ (kcal/mol)
M1	1.40	0.02
Y1	1.82	0.17
Y2	1.82	0.17
Y3	1.82	0.17
Y4	1.77	0.15

Bonds:

Bond type	K_r (kcal/mol/Å ²)	Bond length (Å)
M1-Y4	41.20	2.12
Y1-M1	91.70	1.98
Y2-M1	94.30	1.98
Y3-M1	93.00	1.98

Angles:

Angle type	K_θ (kcal/mol/radian ²)	Equilibrium angle degrees (θ)
CC-Y3-M1	56.44	127.30
CR-Y1-M1	38.97	125.70
CR-Y2-M1	53.67	127.46
M1-Y1-CV	39.62	128.05
M1-Y2-CV	54.89	126.22
M1-Y3-CR	54.18	125.48
M1-Y4-HW	44.55	122.59
Y1-M1-Y2	39.89	116.07
Y1-M1-Y3	37.44	115.90
Y1-M1-Y4	31.02	101.10
Y2-M1-Y3	36.29	114.63
Y2-M1-Y4	36.83	105.42
Y3-M1-Y4	30.04	100.62

Dihedral:

Definition	Divider	Barrier (kcal/mol)	Phase degrees (θ)	Periodicity
X-CC-Y3-X	2	4.80	180.0	2.0
X-CR-Y1-X	2	10.00	180.0	2.0
X-CR-Y2-X	2	10.00	180.0	2.0
X-CV-Y1-X	2	4.80	180.0	2.0
X-CV-Y2-X	2	4.80	180.0	2.0
X-Y3-CR-X	2	10.00	180.0	2.0
CX-CT-CC-Y3	1	0.05	180.0	-4.0
CX-CT-CC-Y3	1	0.74	0.0	-3.0
CX-CT-CC-Y3	1	0.20	0.0	-2.0
CX-CT-CC-Y3	1	0.69	0.0	1.0

M1: Zn; Y1: His94 NE2 (epsilon nitrogen); Y2: His96 NE2 ((epsilon nitrogen)); Y3: His199 ND1 (delta hydrogen); Y4: O (H₂O); CC: CG (gamma carbon); CR: CE1 (epsilon carbon); CV: CD2 (delta carbon)

The 3D spatial locations of the parameters and atoms reported in Table 3.1 are illustrated in Figure 3.4, and show the relationship between Zn^{2+} and coordinating atoms within 3D space. The charges of each atom within the coordination sphere were then evaluated for accuracy by comparison to literature. Figure 3.4 also shows the calculated atom MK-RESP charges, rounded up to two decimal places. Results in Figure 3.4 demonstrate that Zn^{2+} has a charge less than +1, whereas His94 (NE2/Y1), His96 (NE2/Y2) and His119 (ND1/Y3) to have atomic charges of -0.09 , -0.03 and -0.16 respectively. Calculated charges show agreement with those previously determined in 2013 by Bernadat *et al.* [318] whereby Zn^{2+} was found to have a charge smaller than +1, while the coordination N atoms on His94, His96 and His119 contained lower negative charges in comparison to their standard charges [318]. The generated CA-II *.frmod file to be used to export parameters is also presented in Listing S1.

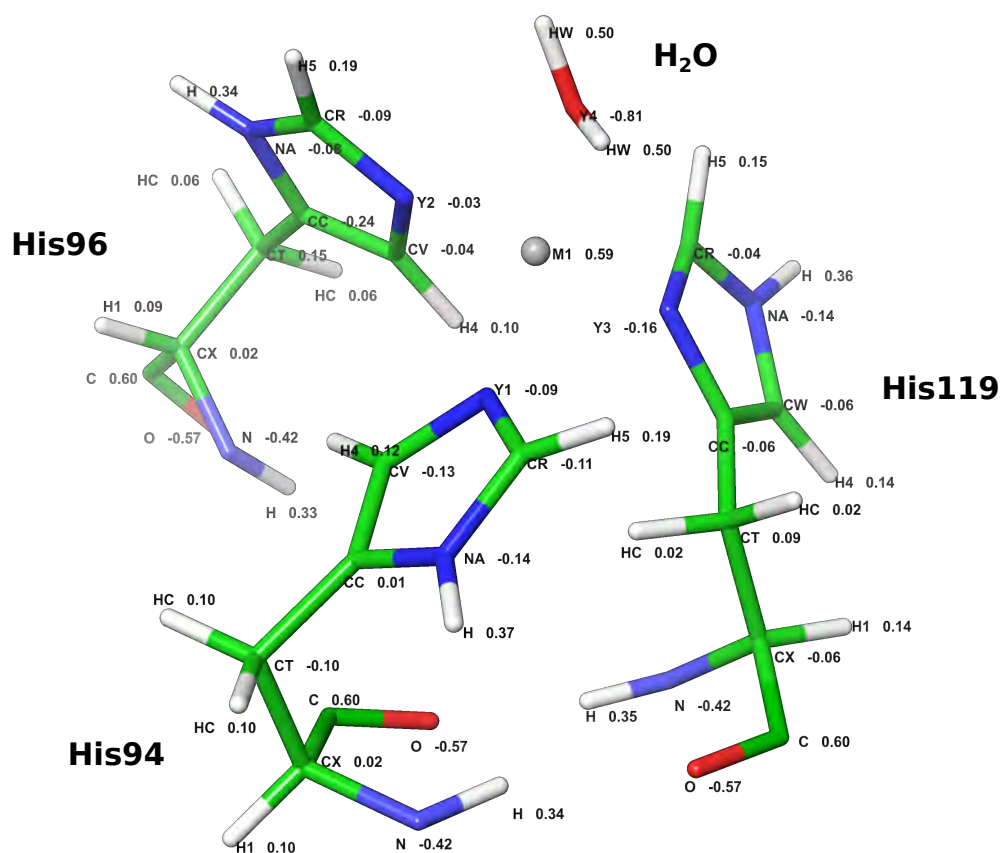


Figure 3.4. MK-RESP charges calculated for Zn^{2+} and coordinating atoms. M1: Zn; Y1: His94 NE2 (epsilon nitrogen); Y2: His96 NE2 (epsilon nitrogen) Y3: His119 ND1 (delta hydrogen); Y4: O (H_2O); CC: CG (gamma carbon); CR: CE1 (epsilon carbon); CV: CD2 (delta carbon).

3.4 CONCLUSION

Traditional MD FF are not capable of handling simulations involving metalloproteins, and as a result cofactors escape the proteins. The AMBER ff14SB FF was extended by the generation of Zn^{2+} FF parameters for MD simulations. Additionally these parameters are required in an easily exportable and customisable format. Generated parameters held the Zn^{2+} cofactor in place during optimisation and could potentially allow for the investigations of potential effects SNVs could have on Zn^{2+} and coordinating atoms within the CA-II and CA-IV proteins.

In the next chapter MD simulations were used to validate the generated FF parameters and investigate whether they are capable of maintaining the cofactor within the active site. Simulations are also used to investigate variant effects on the respective CA proteins.

*A man who views the world the same at fifty as he did at
twenty has wasted thirty years of his life.*

Muhammad Ali

4

Effects of Variants on the Structure and Function of CA-II, CA-IV and CA-VIII

CHAPTER OVERVIEW

Bioinformatic approaches such as MD offer a cost effective and accurate means to understand the structure and functions of proteins compared to experimental analysis. Improvements to FFs defining how atoms interact with one another have improved the efficacy of MD simulations and optimised them to closely resemble experimental outputs. Wet laboratory analysis is still however required to confirm and validate MD simulation findings. Unusually, even where experimental analysis is available, the mechanism or changes to protein structure or function associated with a specific phenotype may not easily be decipherable highlighting the need for *in-silico* analysis. Within this

chapter, MD simulation was used to validate the previously generated FF parameters. Additionally, MD simulation in-conjunction with principal component analysis (PCA) and dynamic residue network analysis (DRN) was also used to identify variant associated changes to the structure and function of CAs and investigate the pathogenesis of CA deficiencies.

4.1 INTRODUCTION TO MOLECULAR DYNAMICS

MD is a technique involved with analysis of the physical motions of atoms in order to simulate (mimic) the transformations undergoing a protein as if it were in a natural biological system. The macromolecular structure governs molecular interactions which are essential for biological function therefore an initial protein model is required for the process [319]. The Hamiltonian in Equation 3.2 determines the forces that atoms exert on each other, with Newton's laws determining their effect on atomic motion over discrete time steps of less than a few femtoseconds, during the simulation. Newtonian laws also allow for the determination of atom positions and velocities during simulation [320]. During simulation long range electrostatic interactions are handled by the particle-mesh Ewald (PME) method [245, 321, 322], whereas a continuum model is used to estimate vdW interactions. In biological cells, protein atoms are always in constant motion as with MD.

The Newtonian second law of motion factored into protein dynamics is presented in Equation 4.1. The F , m and a represent the force, mass and the acceleration of an atom. From Equation 4.1 acceleration is the first derivative of velocity, whereas velocity is the derivative of position. Simulations can either be all atomistic or coarse grained depending on computational resources available [320, 323, 324]. All atom simulations do however offer higher accuracy and reproducibility [325, 326].

$$F = ma$$

Equation 4.1. Newtons second law of motion.

For accurate MD analysis the protein environment (solvation representation) is of significant

importance to accuracy and reproducibility. The solvent system will have an effect on how various atoms interact with each other during the simulation, thereby having an effect on the potential energy (Equation 3.6). The three main types of solvation include:

1. *In-vacuo*: MD is performed in vacuum
2. Implicit solvation: Solvent molecule effects are estimated by the FF and mathematical equations
3. Explicit solvation: Solvent molecules are represented explicitly from the protein molecule. Periodic boundary conditions (PBCs) within a simulation box are used to main the protein and solvent. To facilitate continuous simulations, a molecule exiting the box on the left side re-emerges on the right

From the solvent representations listed, it should be noted that the selected FF should complement the desired environment. Explicit solvation offers the best reproducibility but is however computationally expensive. In the case of insignificant computational resources, the other representations can be selected for MD. Through MD, relationships between protein structure and function can be found at the residue level offering an additional dimension to experimental analysis [320]. This makes MD incredibly useful for investigations into variant associated changes to protein structure in order to determine disease pathogenesis.

4.1.1 Principal Component Analysis

As MD computes large complex datasets of the changes a protein undergoes during simulation, this makes analysis of all variables and their relationships to each other difficult to interpret. Principal component analysis (PCA) is a statistical procedure used to reduce the dimensions of large complex (multidimensional) datasets to enhance interpretability [327]. The dimensionality reduction is performed with minimal loss to information.

PCA is used to extract the protein's most dominant modes of motion from the simulation trajectory, and is performed on the molecule's mass-weighted Cartesian coordinates [328–330]. Prior to internal motion analysis, the overall global and translational motion have to be removed from the

trajectory which is mainly achieved through alignment of the trajectory to a reference structure and applying a least squared fit [245, 330, 331]. A covariance matrix of $3N \times 3N$ elements, describes the correlated movement of a protein with N atoms (Equation 4.2). The $\langle \rangle$ represent the mean across all sampled conformations, and r describes the mass weighted Cartesian coordinates. The σ_{ij} defines the covariance between the i and j coordinates.

$$\sigma_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle$$

Equation 4.2. PCA covariance matrix calculation.

When the covariance matrix is diagonalised, this results in $3N$ eigenvectors and eigenvalues [328]. These account for the modes and amplitudes of the motions. Eigenvectors are the data projections of the principal components and indicate the direction of motion, whereas eigenvalues describe the variance and energies of each eigenvector. Time-dependent motions of each component are then obtained by projecting coordinates along each eigenvector [332]. The first principal component (PC1) shows the largest variation, whereas while PC2 shows the second-most, and PC3 the least variation.

4.1.2 Dynamic Cross Correlation

During simulations, MD analysis is involved with the movement of the respective protein atoms during simulation. Dynamic cross correlation (DCC) measures the degree to which these protein atoms move together (the correlation of their movement) [277]. DCC is calculated according to Equation 4.3.

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle} \cdot \sqrt{\langle \Delta r_j^2 \rangle}}$$

Equation 4.3. Determination of residue dynamic cross correlation.

The Δr_i , $\langle \rangle$ represent the displacement of atom i from its average position and average time over the entire trajectory respectively. DCC values of -1 , 0 and 1 indicate; anti-correlation (residues moving in different directions), no-correlation (the movement of a residue as no effect on the movement of another residue) and correlated (residues are moving in the same direction) residue movement respectively.

4.1.3 Dynamic Residue Networks (DRN)

Since proteins are networks of interacting residues working in unison to maintain protein structure and function, DRN analysis is involved with investigations into the changes of these residue interactions over the MD simulation. It is calculated using the MD-TASK suite [277]. DRN analysis is divided into three main areas of analysis:

1. Weighted contact maps
2. Average shortest path (L)
3. *Betweenness centrality* (BC)

Analysis of the above is especially useful in the study of the effects of SNVs on protein structure and function as variant presence may disrupt key interactions essential for the maintenance of stability and function, thereby resulting in a particular phenotype. Proteins could also change residue interactions to compensate for the presence of variants. To predict residue interactions, pairwise distances between all C_β (C_α for glycine) atoms are evaluated. Each protein residue represents a node within the network (DRN) [277].

4.1.3.1 Weighted Contact Maps

Weighted contact maps are calculated using MD-TASK and show how frequently two residues within a protein interact over an MD trajectory [277]. Variant presence may have effects on the interaction frequency/weighted interactions between two residues making it invaluable for variant associated analyses [18, 66].

4.1.3.2 Average Shortest Path

The accessibility of a specific residue (protein node) is defined by L . It is computed by the division of the total number of shortest paths to that specific node by the total number of nodes minus one [277]. Its calculation is represented in Equation 4.4.

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

Equation 4.4. Calculation of average shortest path.

The V , $d(s,t)$ and n represent the set of nodes in the network, the shortest path from s to t and the total number of nodes within the network respectively [333]. Increases to L denote decreases to residue accessibility, whereas decreases to L are indicative of increases to residue accessibility.

4.1.3.3 Betweenness Centrality (BC)

The importance of a residue to protein communication is defined by the BC . This metric is computed by measurement of the number of shortest paths passing through a specific node interlinking numerous nodes to others. Equation 4.5 illustrates how to calculate BC .

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

Equation 4.5. Determination of the betweenness centrality of a specific residue.

The V , $\sigma(s,t)$ and $\sigma(s,t|v)$ denote the network nodes, number of shortest (s,t) paths and the number of those paths passing through node v other than s,t [333] respectively. Higher values of BC are indicative of more frequent communication of a specific residue within the DRN. More frequently communicating residues are regarded as being associated with higher usage in a protein and are more important for function and/or stability within the protein.

4.2 METHODOLOGY

4.2.1 Protein Preparation

Prior to MD simulation, the Schrödinger Maestro [253] Protein Preparation Wizard implementing PROPKA [257] was used to protonate the CA-II, CA-IV and CA-VIII proteins at a pH of 7.0 [334, 335]. To the CA-II and CA-IV proteins, the protonation states of the Zn^{2+} coordinating residues (see Table 4.1) were then verified to ensure that the first two coordination His ligands in both proteins were in the HID protonation state, with the final His coordination residue in the HIE protonation state. A combination of LEaP modelling [336] and manual file modification was then used to export the previously generated FF parameters (see Chapter 3) to add Zn^{2+} support to the AMBER ff14SB FF as follows. To the respective PDB files for each WT and variant CA-II and CA-IV protein, the primary Zn^{2+} coordination residue names were modified to that of the respective MCPB (metal center parameter builder) identifier as illustrated in Table 4.1. MCPB was used during metal ion parametrisation (Chapter 3). His₉₄^{CA-II} and His₁₁₅^{CA-IV} were edited to HD1, His₉₆^{CA-II} and His₁₁₇^{CA-IV} modified to HD2, and His₁₁₉^{CA-II} and His₁₄₀^{CA-IV} edited to HE1 each. The ZN residue name was also modified to ZN1, and the coordinating water renamed to WT1.

Table 4.1. Residue renaming for parameter export in the CA-II and CA-IV protein coordination residues. MCPB refers to the metal center parameter builder.

CA-II		CA-IV	
Residue	MCPB identifier	Residue	MCPB identifier
His94	HD1	His115	HD1
His96	HD2	His117	HD2
His119	HE1	His140	HE1
ZN	ZN1	ZN	ZN1
HOH	WT1	-	-

After the residue name modification, protein topologies were generated for MD simulation using LEaP separately. Listing S1 and Listing S2 show an example of the custom FFs and LEaP input used

together. Residues were set to use the AMBER ff14SB FF [303], whereas non standard residues were set to use the gaff2 FF [304]. BCT and CO₂ are non-standard residues therefore support for these molecules was also added. This was achieved by calculating the AM1-BCC charges for each molecule using *antechamber* [337] for the gaff2 atom type through the use of the AmberTools17 built-in quantum chemistry program *sgm*. Generated **.mol2* files containing the charge for each atom were then processed using *parmchk2* to generate the respective **.frcmod* files containing parameters for each molecule. The respective **.mol2* files for BCT and CO₂ were then added to the LEaP program along with the generated **.frcmod* files to extend the FFs to adequately support BCT and CO₂.

As the coordinating residue names for the primary Zn²⁺ coordination spheres in CA-II and CA-IV were modified, and their respective **.mol2* files generated by MCPB had custom atom names (M1; Y1; Y2; Y3 and Y4, see Table 3.1 and Figure 3.4), these atom types were manually added to LEaP. This was then followed by the addition of their respective **.mol2* files identifying the HD1, HD2, HE1 and ZN1 PDB residue names. Since HD1, HD2, HE1 and ZN1 constitute non-standard residues not recognised by LEaP, bonds to the ZN1 by the coordinating atoms were manually specified to let LEaP know that these atoms were connected.

AMBER topologies were then solvated using the TIP3P water model as a solvent in a cubic box of 10 Å cut-off distance (distance between protein molecule and box). The system was then neutralised using Na⁺ and Cl⁻ counter-ions. ACPYPE was used to convert the AMBER topology files to GROMACS [291] topologies (**.gro* and **.top* files). Conversion of AMBER topologies using ACPYPE allows the porting of all previously generated parameters including MCPB, LEaP, solvation and cubic box dimensions to GROMACS. To the generated **.top* file, total protein charge (*qtot*) was manually inspected to ensure that the net charge was 0.00. This was confirmed through comparison of *qtot* with the quantity of counter-ion added to ensure correct neutralisation.

4.2.2 Molecular Dynamics

The previously generated protein topologies were now ready for minimisation. MD simulations were set up for all 35 protein systems (21 CA-II, 7 CA-IV and 7 CA-VIII). Protein energy minimisation was performed using the steepest descent algorithm, and set to terminate when the system had converged and an F_{max} (maximum force) of $1\,000\text{ kJ mol}^{-1}\text{ nm}^{-1}$ had been attained. Minimised structures were subjected to temperature (*NVT*) and pressure (*NPT*) equilibration. *NVT* and *NPT* were performed by constraining all bonds through the use of the LINCS algorithm. The Particle Mesh Ewald (PME) coulomb type was set for long-range electrostatics in-conjunction with the modified Brenson thermostat. The *NVT* ensemble was performed at a temperature of 300 K for a period of 100 ps. Once equilibrated the *NPT* ensemble was performed using the Parinello-Rahman barostat algorithm [338] until the system stabilised at a pressure of 1 bar. The CHPC Cluster in Cape Town was used to perform MD simulations using a combination of 10 CPU cores and one Nvidia Tesla v100 GPU for a 200 ns duration. A 2 fs time integration step was also set. Coordinates were written to file over 10 ps intervals.

4.2.3 Molecular Dynamics Trajectory Analysis

After MD simulation, the proteins were centred within the simulation box and periodic boundary conditions (PBCs) removed. All water molecules were also stripped using *cpptraj* [339]. The resulting trajectory was visualised using VMD [340] to ensure that the Zn^{2+} ion was maintained within the active site and the system adhered to the set parameters. PDB files at the start and end of the simulation were also generated for analysis. Root mean square deviation (RMSD), root mean square fluctuation (RMSF) and the radius of gyration (Rg) of the protein α -carbons were then calculated using *cpptraj*.

4.2.3.1 Proton Shuttle Analysis

As the CA-II and CA-IV proton shuttles are essential to catalysis, the proton shuttle behaviour was evaluated during MD simulation. Catalytic CA proton shuttles were analysed through structural clustering using the average-linkage method [341–343] and the hierarchical agglomerative (bottom up) algorithm implemented by the *cluster* command of *cpptraj*. A total of four conformational clusters were set to be generated. These would allow for the inclusion of the “in” and “out” His64 conformations, and other unexpected conformations including imidazole ring flips. The clustering was performed for every four frames. Conformations were set to be clustered using the following three criteria; distance between His64 ND1 atom and the Zn²⁺, angle between His64 CB, CG atoms and the Zn²⁺, and dihedral angles between His64 N, CA, CB and CG atoms (chi1), and CA, CB, CG and ND1 (chi2). Representative structures from each cluster were also set to be generated.

4.2.4 Dynamic Cross Correlation (DCC)

The extent to which the WT and variant protein residues move together was calculated using the *dcc.py* script of the MD-TASK suite [277]. The correlated residue motions of the C_α atoms of the proteins over the MD simulation are presented as a heat map showing residue correlation.

4.2.5 Dynamic Residue Network Analysis

DRN analysis was performed to analyse changes occurring to the CA-II, CA-IV and CA-VIII protein networks due to SNV presence using MD-TASK [277]. Calculations were performed every 100 MD simulation frames using a cut-off of 6.7 Å for C_α-C_α node interaction [344].

4.2.5.1 Weighted Contact Map Analysis

The MD-TASK script *contact_map.py* was used to determine changes to residue-residue interactions between the SNVs and neighbouring residues over the 200 ns simulation. The interaction changes

at each SNV location were then compared to corresponding WT residue to observe short-range interaction differences. Gnuplot [345] was used to generate a heat map showing the degree of weighted interaction changes for each residue during MD.

4.2.5.2 Average Shortest Path (L)

The *calc_network.py* MD-TASK script was used to calculate L for the WT and variant proteins every 100 MD frames. The calculated L for each protein was then averaged across all selected frames to obtain average L . Unity-based normalisation on a scale of 0 to 1, was then performed for each WT and variant CA protein group (21 CA-II; 7 CA-IV and 7 CA-VIII) separately, to normalise all respective WT and variant proteins onto the same scale and generate average normalised L . The ΔL showing accessibility changes between the WT and variant proteins was then calculated by subtracting the average normalised L of the WT and variant proteins (WT – variant).

4.2.5.3 Betweenness Centrality (BC)

Similar to L , residue BC of the WT and variant proteins was determined using the *calc_network.py* script of MD-TASK. Residue BC was averaged across all selected MD frames to obtain average BC . Each WT and variant CA protein group BC was then normalised separately on a scale of 0 to 1 using unity-based normalisation to get respective data onto the same scale and calculate average normalised BC . WT and variant average normalised BC was then subtracted (WT – variant) to determine ΔBC to compare changes between WT and variant proteins.

4.2.6 Principal Component Analysis (PCA)

PCA was performed to analyse the 3D structural changes occurring to the CA proteins as a result of variant presence [331]. An RMS best-fit to the first structure was applied to each CA protein to remove global rotational/translational motion. All heavy atoms (excluding hydrogen) were used

to calculate the coordinate covariance matrix of the WT and variant protein structures. Respective eigenvectors and eigenvalues were then obtained by diagonalising the matrix. Along each eigenvector, variant coordinates were projected to obtain each trajectory sets separate projections. Normalisation was applied to the first two projections, and data plotted to obtain a graphical representation of PC1 vs PC2 using the *cpptraj hist* command. The *hist* command was also set to also calculate the Gibbs Free Energy at 300 K associated with PC1 and PC2. An example of the PCA script utilised is presented in Listing S3.

4.3 RESULTS AND DISCUSSION

4.3.1 Variant Presence Is Associated With Conformational Changes To the Global Structure of CAs

MD simulations were performed for 35 protein systems; 21 CA-II structures (7 apo, 7 BCT and 7 CO₂ bound), 7 CA-IV and 7 CA-VIII structures for 200 ns each (77.60 CPU hours). Variant-associated changes to the global structure of the proteins were then analysed using the RMSD, PCA and Rg.

For CA-II and CA-IV proteins, the Zn²⁺ remained in place during the MD simulation indicating successful parametrisation. Results in Figure 4.1 demonstrate the average distance between Zn²⁺ and coordinating residues during MD. Data illustrates that the Zn²⁺ maintained the same distance from its ligands, and bond angles were close to those presented in Table 3.1. This strongly suggests that parameters were valid and accurate.

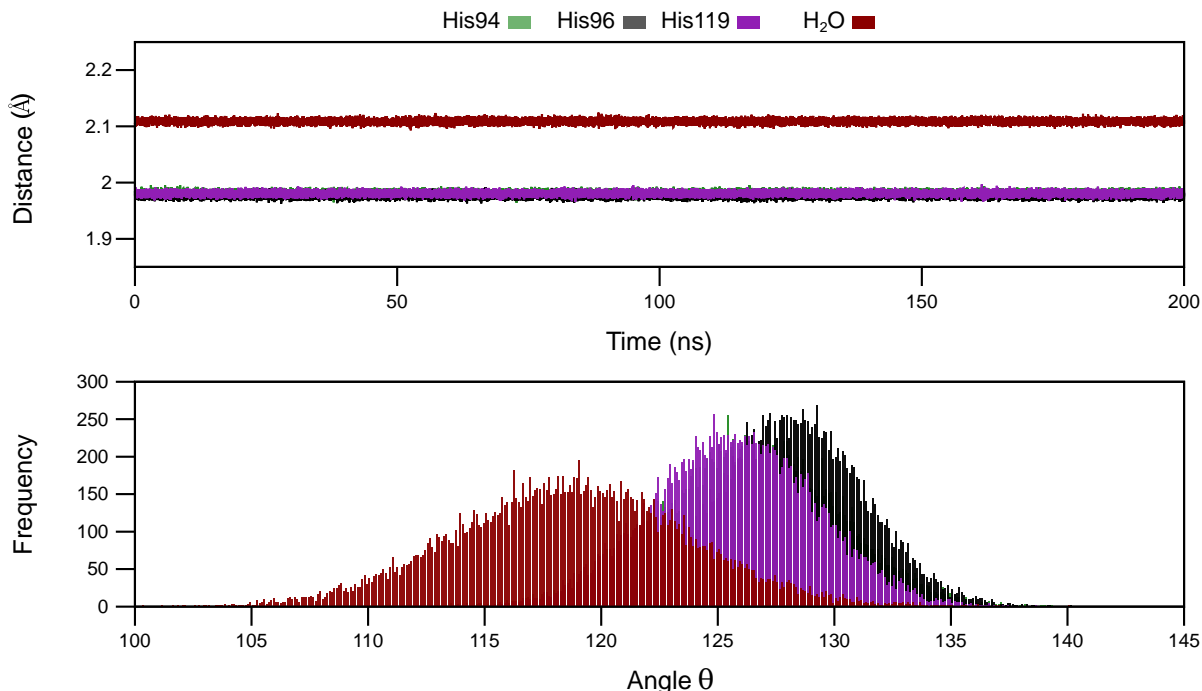


Figure 4.1. Parameter validation presenting bond distances during MD between Zn²⁺ and coordinating atoms, and angles of: M1-Y1-CR (His94); M1-Y2-CR (His96); M1-Y3-CC (His119) and M1-Y4-HW (H₂O) during MD simulation.

4.3.1.1 RMSD Analysis

In Chapter 2, VAPOR analysis (Table 2.1) predicted variant-associated stability reductions to protein structure. As a result, RMSD differences were expected between proteins. Across all MD frames, WT and variant protein system RMSDs were calculated using *cpptraj* and results are presented in Figure S3. Data in Figure S3 suggests that the SNVs presence in CA-II could have a subtle effect to protein structure as evidenced by lack of drastic changes to RMSD. To observe the discrete changes to RMSD occurring during MD simulation, RMSD distributions demonstrating sampled conformations of the WT and variant proteins as the Kernel density estimate (KDE) were calculated, and results are shown in Figure 4.2.

The KDE is a statistical procedure that is non-parametric in nature and utilised to calculate the probability density function (PDF) of a variable. KDEs share a close relationship with histograms but however, have the added advantage whereby, there is no informational loss occurring during calculation as a result of binning which is a feature observed in histograms [346, 347]. The distribution shape and spread of the data is also easier to understand due to the data smoothing implemented in KDEs [346, 347].

4.3.1.1.1 CA-II

From the data in Figure 4.2, the width of the distribution represents the number of sampled conformations during MD, whereas peaks describe the most sampled protein conformation. The *y*-axis can be thought of as the frequency of conformational sampling during MD. Distributions sampling numerous conformations in relation to the WT protein could be indicative of potential instability within the variant. Results in Figure 4.2 separately compare each CA-II protein system (apo, BCT and CO₂). Even though RMSD distributions between the WT and variant proteins show minor differences, in 2004 research by Almstedt [271] discovered that residue displacements as small

as 0.30–0.40 Å are capable of destabilising the structure of CA-II.

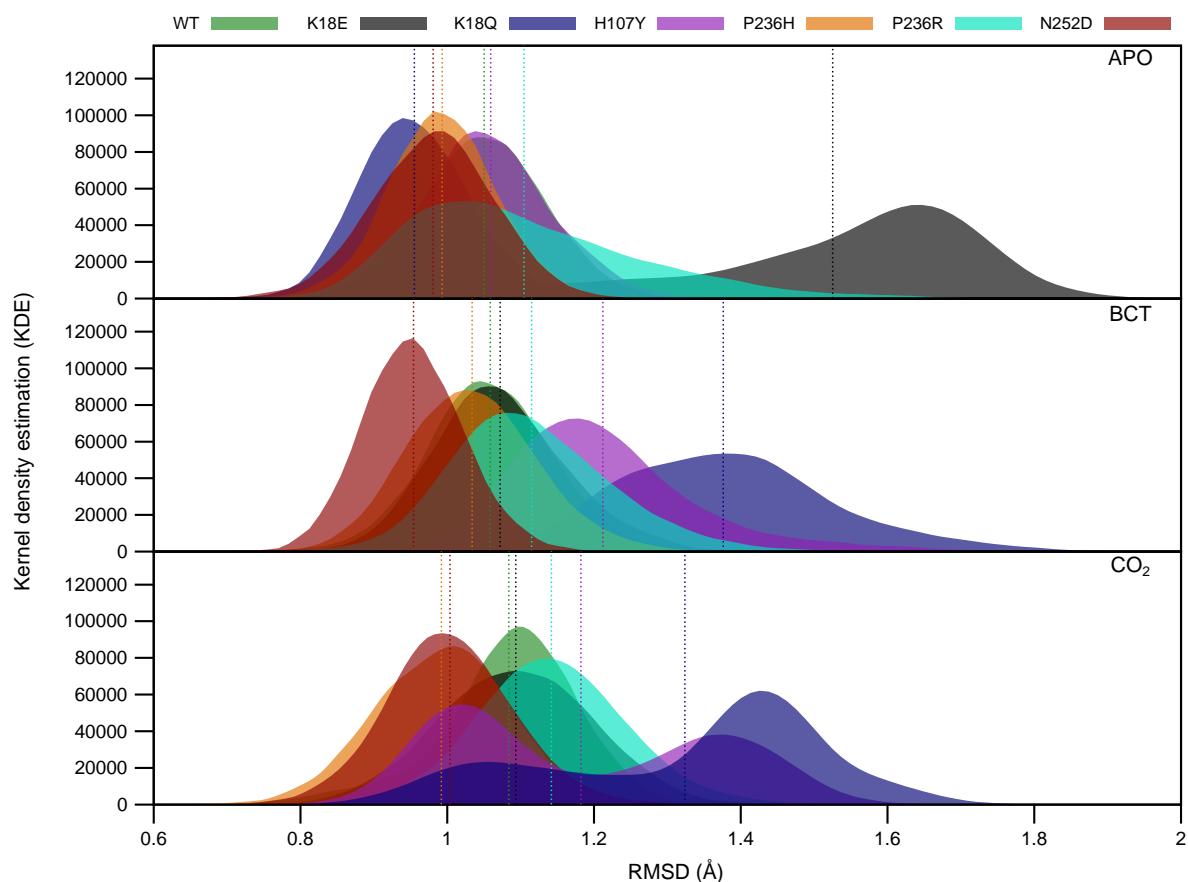


Figure 4.2. RMSD distributions of the WT and variant CA-II protein systems. Average RMSD for each plot is presented as a dashed line on each plot of the corresponding colour.

The average RMSDs of the CA-II proteins are presented in Figure 4.2 as dashed lines of corresponding colour on each plot. Data highlights that not all average RMSDs coincide with the peaks of each plot, therefore direct comparison of structures using average RMSDs was not performed. The distribution size and shapes were compared instead.

Analysis of the apo proteins in Figure 4.2 shows that H107Y has the greatest structural overlap with the WT protein, while K18E shows the greatest structural differences. Variants K18E and P236R sampled the greatest number of structural conformations during MD evidenced by the wider distribution bases. When BCT is bound to CA-II, P236H shares the greatest structural overlap with the WT protein, whereas K18Q exhibits the largest differences. In the presence of CO₂, P236R has the greatest structural overlap with the WT, while K18Q demonstrates the largest differences. In

the presence of both BCT and CO₂ variants K18Q and H107Y exhibit the greatest conformational sampling. In addition, two peaks are observed in the RMSD distributions of each indicating that during MD both variants formed two distinct conformational clusters. The greater peaks are indicative of the more preferred conformation. Assessment of the observed peaks in relation to the WT suggests that, the conformational clusters sharing the greatest structural overlap with WT proteins could represent the more stable and catalytically viable conformations.

4.3.1.1.2 *CA-IV*

The RMSD of CA-IV over the MD simulation is presented in Figure S4 and the respective distribution is presented in Figure 4.3. CA-IV data in Figure S4 highlights that the individual protein RMSDs remain constant during MD, with the exception to the WT protein that presents a change to RMSD at approximately 50 ns. Comparison of the pathogenic and benign variant demonstrates that the pathogenic variants; R69H, R219C and R219S generally have lower RMSDs compared to their benign counterparts. This could suggest increases to the rigidity of these structures. Analysis of the RMSD distribution (Figure 4.3) indicates that all CA-IV proteins are associated with one major conformational cluster during MD. N86K samples the fewest protein conformations during MD, whereas the WT protein samples the greatest number for structural conformations. The low conformational sampling of N86K could indicate increases to structural rigidity as a result of variant presence. The benign variants share greater structural overlap with the WT protein compared to the pathogenic variants. The pathogenic variants also each share great conformational overlap with each other. Further inspection of the RMSD distribution indicates that variants N86K and N177K, and R69H and R219C show evidence of large structural overlap between each individual pair of proteins respectively. This finding could suggest that for each pair of the aforementioned benign and pathogenic variants, SNV mechanism of action could be similar. This has been investigated throughout the rest of this chapter.

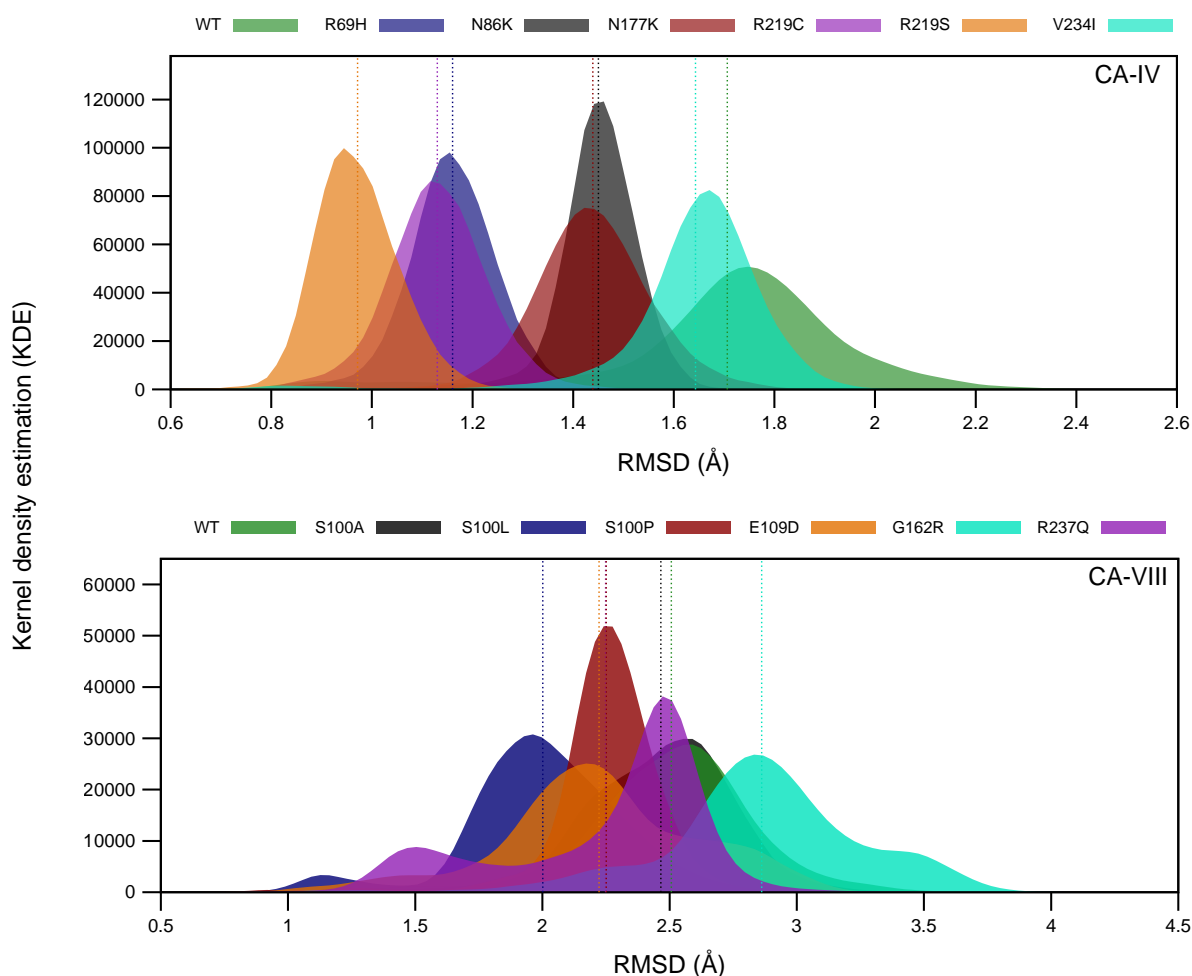


Figure 4.3. RMSD distributions of the WT and variant CA-IV and CA-VIII protein systems. Average RMSD for each plot is presented as a dashed line on each plot of the corresponding colour. CA-VIII plot adapted from Sanyanga and Tastan Bishop 2020 [269].

4.3.1.1.3 CA-VIII

CA-VIII RMSD results in Figure S4 indicate that over the duration of the MD simulation, G16R was associated with the greatest RMSD changes highlighting at potential variant instability. Assessment of the RMSD distribution in Figure 4.3 indicates that S100A exhibits the greatest structural overlap with the WT, and S100L shows the largest RMSD difference from the WT. Given that S100L and E109D are benign and both variants share some structural overlap with the pathogenic S100P, results suggest that the pathogenic effects of the variants may have an effect on the manner in which amino acids interact with each other (local changes) as opposed to global changes. This is further supported by the structural overlap observed between the WT protein, and the variants S100L and S100P. The variants

S100L and R237Q also form two distinct conformational clusters during MD simulation. The 237Q conformational cluster at approximately 2.45 Å samples similar conformations to that of the WT. In both S100L and R237Q conformations at 1.1 Å and 1.5 Å are sampled to a lesser extent than the other cluster. G162R samples three potential conformational clusters during MD simulation with peaks occurring at 2.3, 2.8 and 3.5 Å. The presence of three conformational clusters could indicate variant associated instability within the protein. The clusters at 2.3 and 2.8 Å sample similar conformations to the WT protein. The clusters at 2.3 Å and 3.5 Å are not sampled as frequently as the 2.8 Å structural cluster.

Further comparison of RMSD distributions shows that S100P samples the least conformations of all the proteins, suggesting increases to structural rigidity. Previous S100P research in 2009 by Turkmen *et al.* in 2009 [131] suggested that the variant could be associated with a reduction to protein stability. Additional research by Aspatwar *et al.* in 2010 [145] highlighted that the when Pro substitutes Ser at position 100 this would result in a shorter and more constrained β -sheet and loops due to poor protein folding, which could be a direct effect of the β -sheet destruction observed previously in Chapter 2. This finding could explain the smaller conformational sampling observed. Increases to CA-VIII rigidity by S100P could have an effect on the ability of the protein to allosterically regulate ITPR1. A structure that is too constrained may not be able to induce conformational changes within the receptor.

4.3.1.2 PCA Analysis

PCA analysis was performed to analyse the 3D conformational sampling, internal dynamics, and associated of the CA protein conformations. As RMSD only shows conformational sampling in 2D space, PCA would facilitate multidimensional analysis.

4.3.1.2.1 CA-II

The 3D PCA analysis of the WT and variant CA-II proteins is presented in Figure 4.4. To ensure that the majority of the conformational sampling was covered by PC1 and PC2, the eigenvalue fraction of each PC was calculated and results are presented in Table S5. Data shows that the majority of conformational sampling is covered by PC1 and PC2 which represent the largest and second largest possible variances of the structures respectively. More stable protein conformations are expected to be associated with lower free energy.

Analysis of the WT protein shows that, when neither BCT nor CO₂ are bound, two protein conformations are sampled along PC1 whereas only one conformation is sampled along PC2. The PC1 result is in disagreement with the RMSD results presented in Figure 4.2 that show only one major conformational cluster. The difference in results can be explained through the analysis of the free energy landscapes associated with each cluster. Results indicate one free energy well is larger than the other. This suggests that the majority of conformations sampled during MD are located within the larger well explaining the single peak observed in the RMSD results. When BCT is bound, the WT samples a larger conformational space compared to the apo and CO₂ bound structures. The BCT bound structures are associated with higher free energy. As observed in the apo, when CO₂ is bound, two conformational clusters are observed along PC1. The low energy structural cluster could be the one observed within the RMSD results.

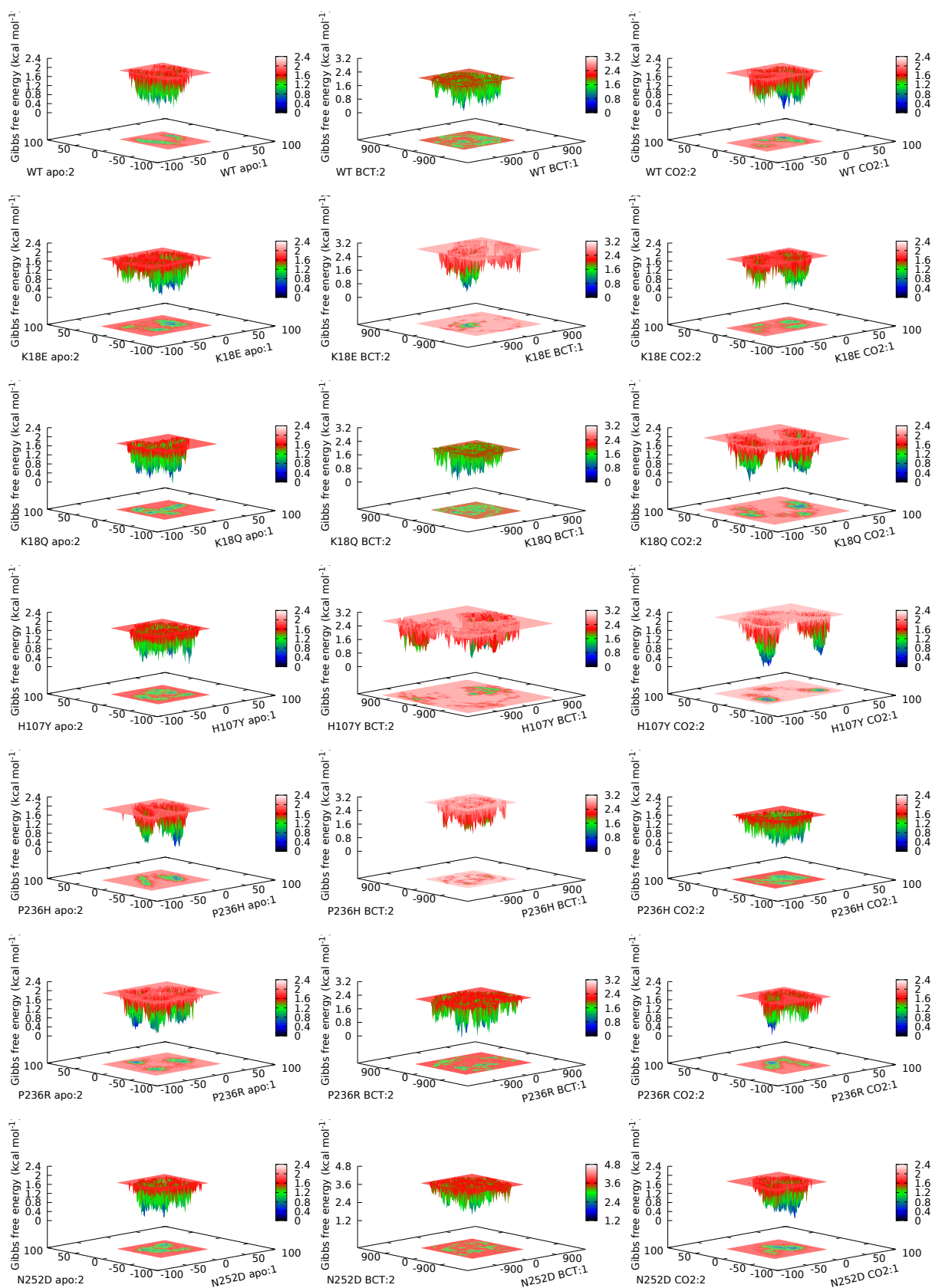


Figure 4.4. PCA analysis of the WT and variant CA-II proteins and associated free energy for each conformational cluster. The x -axis and y -axis represent the 2D PCA plot. Adapted from Sanyanga *et al.* 2019 [66].

Data in Figure 4.4 also suggests that the mechanism of K18E and K18Q could be different even though the variations occur at the same position. This difference is highlighted by the free energy data that shows the K18E structures have higher free energy than the K18Q ones. The PCA and RMSD results of K18Q with CO₂ show agreement. The two structural clusters observed within the RMSD results are also evident along PC1. The BCT bound H107Y structure samples more conformations than the WT and also has higher free energy. The result is complemented by the RMSD findings that also greater conformational sampling in the variant compared to the WT. In addition, H107Y forms two conformational clusters of low energy when BCT is bound. These clusters are were also observed within the RMSD results. The P236H and N252D structures demonstrate the highest free energies of all the proteins. The P236H result shows some agreement with previous literature findings indicated that the SNV affected CA-II folding [278]. The high free energy could be as a result of erroneous protein folding. Results in Figure 4.4 also present evidence of the potential instability of P236R when in the apo state. Two distinct conformational clusters are formed by this variant along both PC1 and PC2. The multiple conformations occupied by this variant during MD are in agreement with the respective RMSD findings.

4.3.1.2.2 CA-IV

Data in Figure 4.5 presents the PCA analysis results of CA-IV WT and variant proteins, while results in Table S6 show the eigenvalue fraction of each PC. Initial inspection of data reveals the differences between the conformational sampling of the pathogenic and benign variants. The WT and benign variants; N86K, N177K and V234I are all associated with one distinct low energy structural well compared to the pathogenic variants. The conformations for the WT protein are associated with lower free energy compared to the variants. The PCA results also follow a similar trend as to that observed within the RMSD results whereby, one low energy structural well is observed for each protein which is in agreement with the single peaks observed within the RMSD distribution. Comparison of benign

and pathogenic SNVs also shows that the disease causing proteins are generally associated with higher free energy highlighting potential decreases to protein stability.

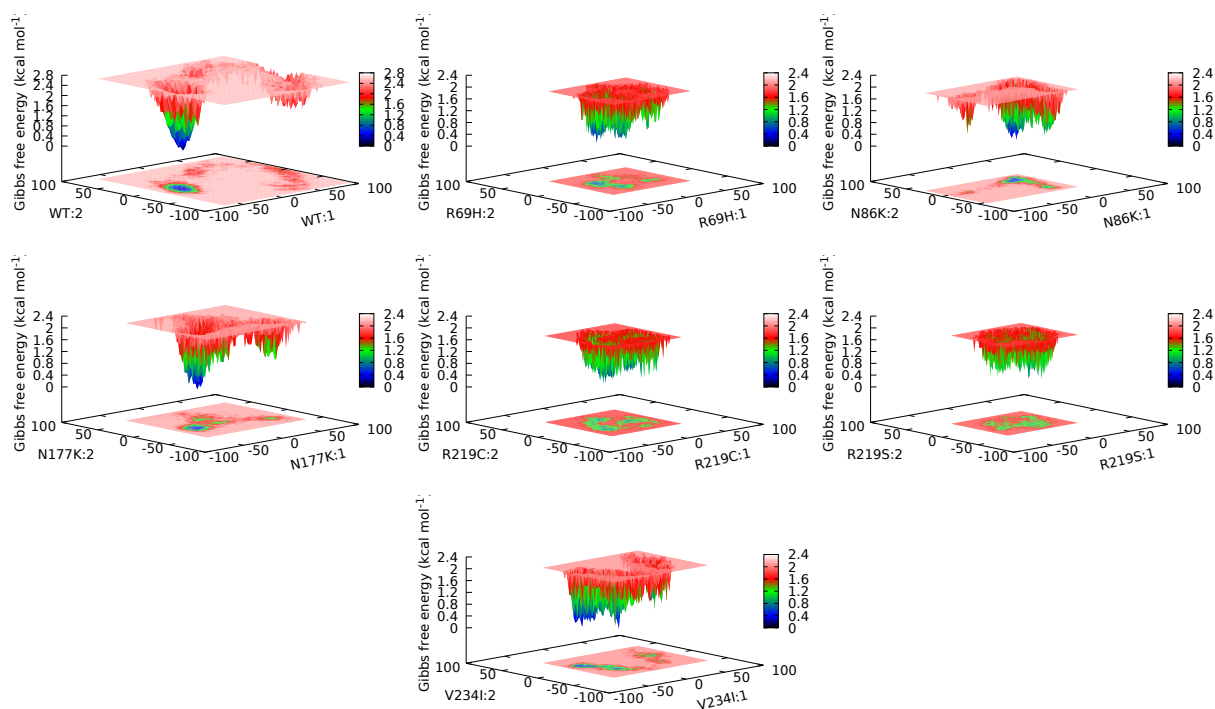


Figure 4.5. PCA analysis of the WT and variant CA-IV proteins and associated free energy for each conformational cluster. The x -axis and y -axis represent the 2D PCA plot.

The R69H variant conformations have the lowest free energy of the pathogenic SNVs. This indicates that during MD simulation the variant occupied more stable conformations compared to R219C and R219S. The occupation of these stable conformations could explain previous experimental results that found functional loss of R69H to not be as severe as that of R219S [119]. This could be as a result of the ability of R69H to occupy a low energy conformations which could assist with catalysis. PCA results also support the hypothesis by Datta *et al.*, [119] that suggests that RP17 may not be caused by the inability of CA-IV to remove excess acid from the retina, but could be due to a toxic gain of function. Results also evidence that along PC1 and PC2, the SNVs R219C and R219S occupy greater conformational sampling in 3D space.

4.3.1.2.3 CA-VIII

The CA-VIII PCA results for the WT and variant proteins are presented in Figure 4.6. Eigenvalue fractions are presented in Table S7 and demonstrate that the majority of the conformational sampling during MD simulation is covered by PC1 and PC2.

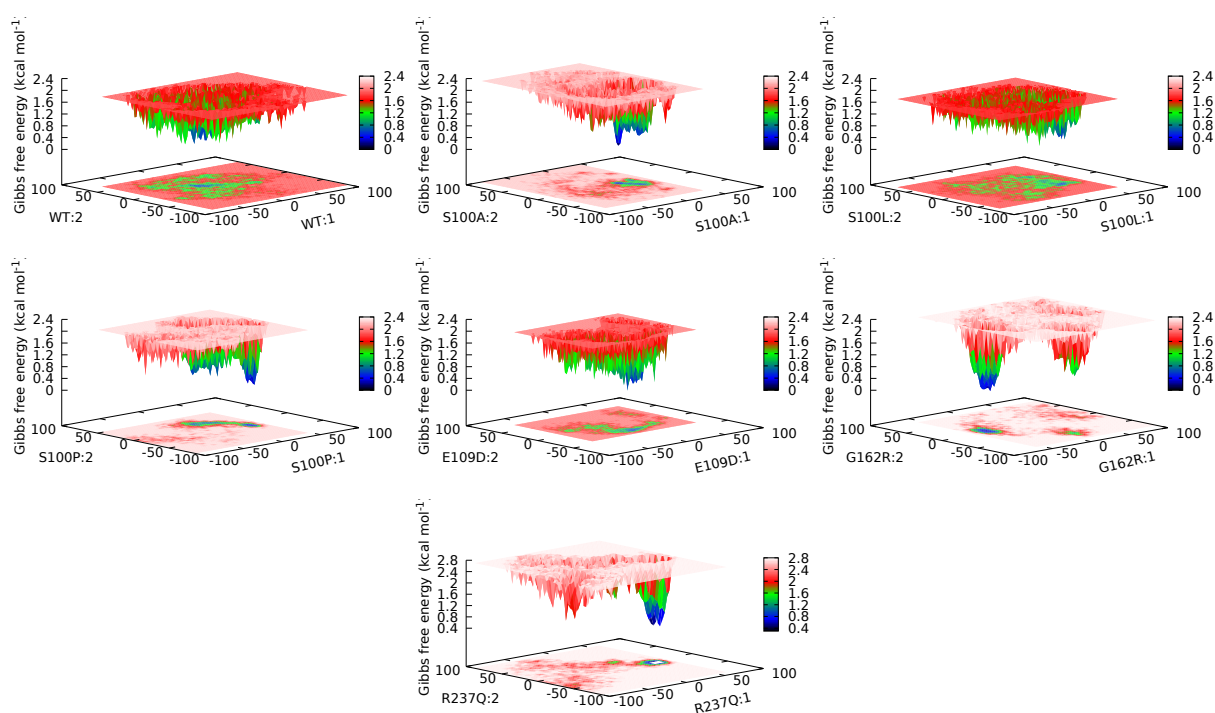


Figure 4.6. PCA analysis of the WT and variant CA-IV proteins and associated free energy for each conformational cluster. The *x*-axis and *y*-axis represent the 2D PCA plot.

Preliminary analysis of Figure 4.6 illustrates differences between the conformational sampling and free energy of the benign and pathogenic variants. Opposite to that observed in CA-IV, the pathogenic variations of CA-VIII are associated with lower free energy compared to the WT and benign variants evidenced by the existence of low energy structural wells. This indicates potential increases to structural stability of the structures within these wells and could explain the structural rigidity increases observed within the RMSD results. The non-pathogenic proteins also show greater conformational sampling along both PC1 and PC2. This greater conformational sampling in 3D space could be necessary to allow CA-VIII to facilitate its allosteric effect on ITPR1. G162R analysis shows the presence of two low energy structural wells and a third less defined one. The result is in

agreement with the RMSD findings that highlighted three potential conformational clusters.

4.3.1.3 Rg Analysis

In the previous sections variant effects to CA structure and conformation were investigated using a combination of RMSD and PCA analysis. In this section, variant effects on the Rg of the protein was investigated. Analysis of Rg would allow for the determination of potential relationships between protein compactness and the structural differences noted in the previous section [348]. Potential increases to protein compactness would result key residues moving closer to each other whereas, decreases to compactness could result in protein residues moving further apart from each other. CA residue-residue distance through previous studies has been shown to have an effect on enzyme kinetics. Specific residue-residue distances are necessary for optimal enzymatic activity [50, 349–353].

4.3.1.3.1 CA-II

Data in Figure S5 compares the CA-II WT and variant protein Rg over the 200 ns MD simulation. Results show subtle variant effects on the compactness of the proteins. To better understand the spread of the data, Rg distributions were calculated and the shape analysed. Figure 4.7 illustrates a distribution of the various Rgs sampled by the proteins during MD. Alike protein systems were compared to each other. The apo, BCT and CO₂ protein systems are each compared individually.

Figure 4.7 highlights that a reduction to the compactness of H107Y occurs for all three protein states. Compared to the WT protein H107Y also shows a wider distribution base suggesting greater Rg sampling. RMSD findings do support this result as H107Y had greater conformational sampling. The K18Q_{BCT} protein shows an unexpected result in the presence of substrate, though the Rg distribution shows a wide base, two distinct Rg peaks are also observed. This could suggest that during MD, K18Q underwent conformational changes that could have resulted in a shift to the protein's centre of mass resulting in the clusters. K18Q is more compact in the presence of CO₂.

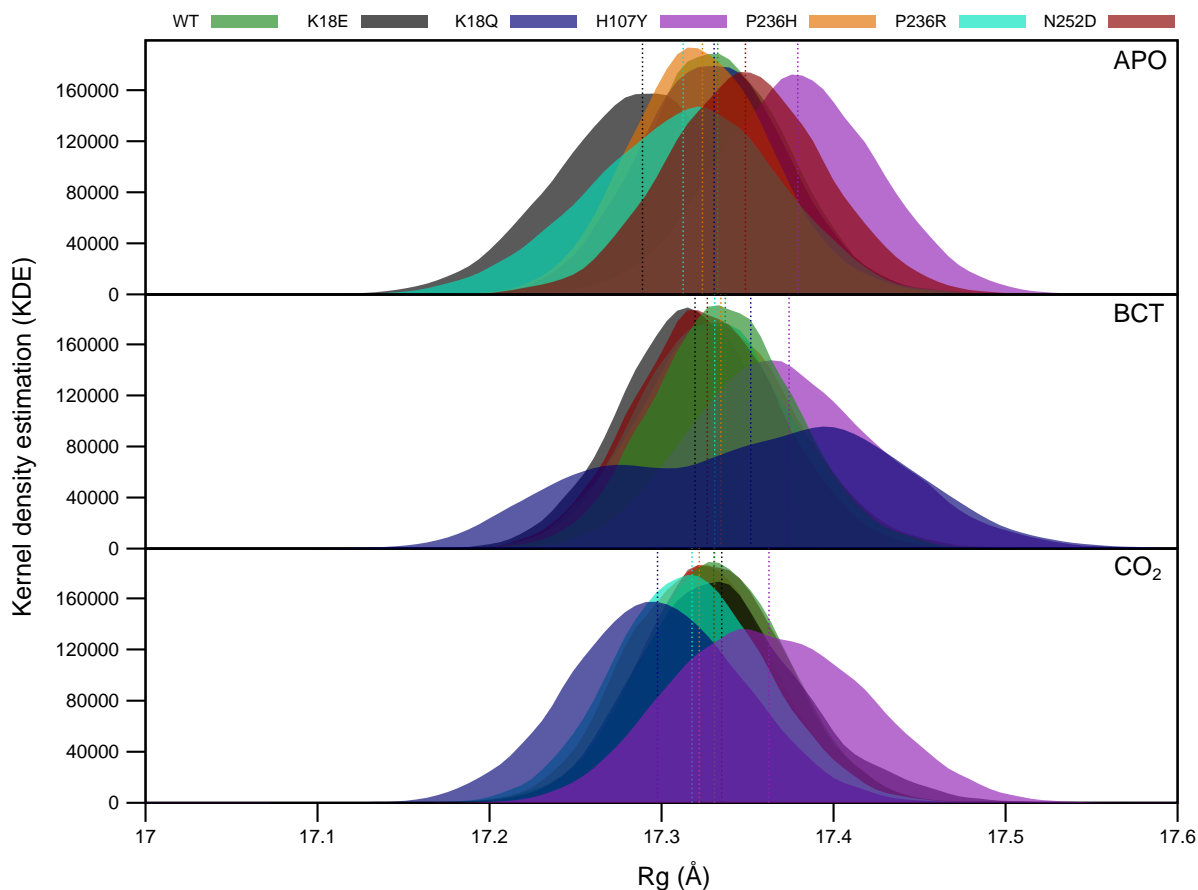


Figure 4.7. Rg distributions of the WT and variant CA-II protein systems. Average Rg for each plot is presented as a dashed line on each plot of the corresponding colour.

4.3.1.3.2 CA-IV

The Rg of CA-IV during over the 200 ns MD are presented in Figure S6. With regards to CA-IV, over the duration of the MD simulation there are no major differences to the Rgs of the WT and variant proteins. To observe subtle differences in the Rgs of the WT and variant proteins, the Rg distribution of the CA-IV protein systems was calculated and results are presented in Figure 4.8. Data shows that the variant proteins are slightly more compact than the WT. Each distribution presents only one peak indicating that CA-IV structures only sampled one Rg conformation during MD and no changes to the protein centre of mass occurred. Unexpectedly, as RMSD results showed differences in conformations between benign and pathogenic proteins, data suggests that variant presence has minimal effect on the compactness of CA-IV. Although the WT and variant proteins have different

compactness, all the CA-IV proteins sample approximately the same number of Rg conformations (distribution width).

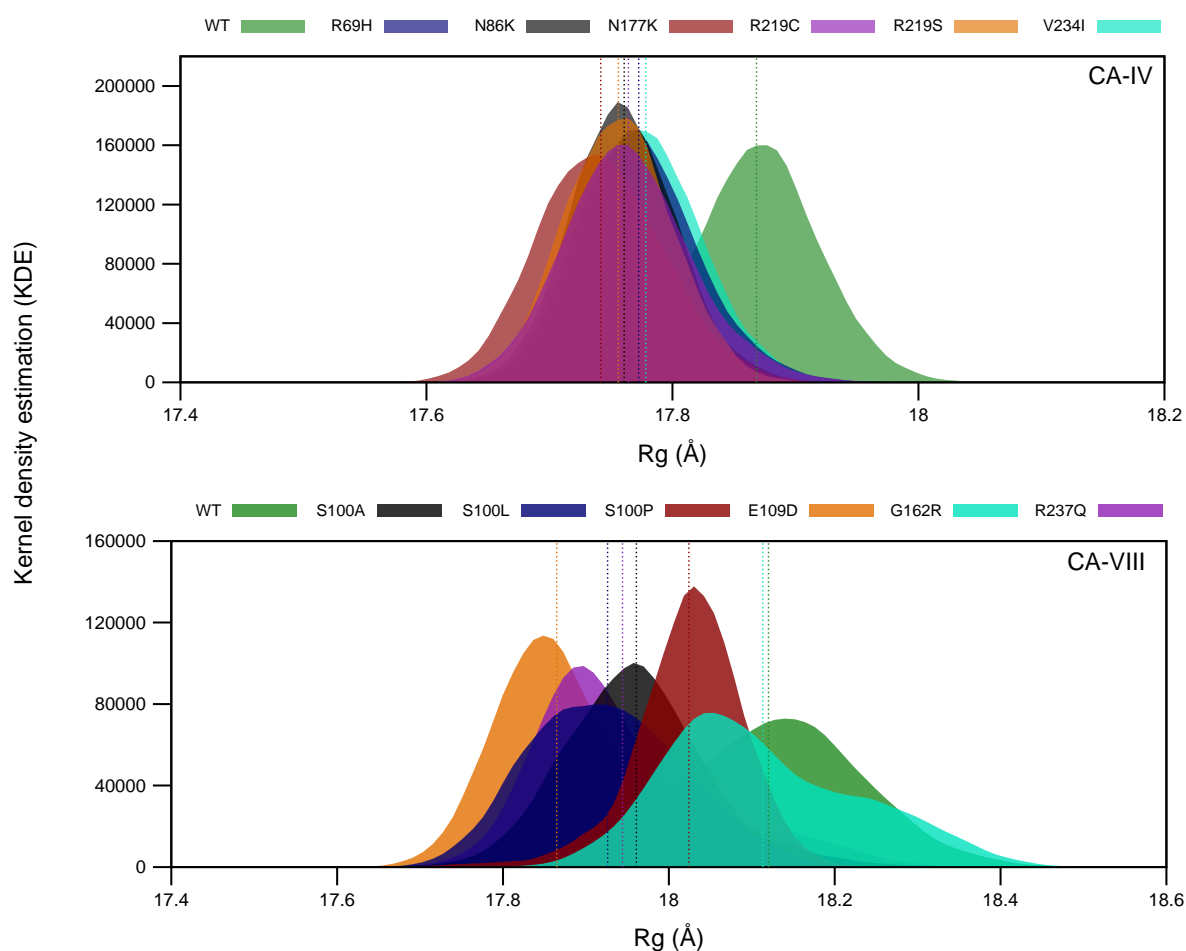


Figure 4.8. Rg distributions of the WT and variant CA-IV and CA-VIII protein systems. Average Rg for each plot is presented as a dashed line on each plot of the corresponding colour. CA-VIII adapted from Sanyanga and Tastan Bishop 2020 [269].

4.3.1.3.3 CA-VIII

Figure S6 highlights that the CA-VIII WT protein is less compact than variant proteins after 100 ns, then for the rest of the simulation WT and G162R proteins maintain similar Rg. This finding is further supported by data in Figure 4.8 that highlights the WT protein is less compact than the other variants. E109D is however the most compact CA-VIII protein. The S100P distribution has the smallest base indicating that this variant sampled the fewest Rg conformations of all proteins. As the RMSD results indicated variant increases to rigidity, the low number of Rg conformations

sampled could be due to a more constrained S100P. Interestingly the WT protein samples the most Rg conformations. This effect can be explained by the PCA results that showed greater conformational sampling along both PC1 and PC2. The binding of CA-VIII to ITPR1 could be affected by increases in compactness of the variants, as essential binding site residues could have their accessibilities altered thereby inhibiting Ca²⁺ homeostasis and regulation. This effect is most likely due to local changes in residue behaviour as opposed to global Rg changes as a benign variant is the most compact. Potential instability of G162R is also observed in Figure 4.8 whereby data shows the potential formation of another conformational cluster at 18.25 Å. Formation of the cluster could be as a result of the change to G162R Rg after 100 ns. The Rg results complement the RMSD ones that indicate potential stability changes to G162R as a result of variant presence.

4.3.2 Local Residue Analysis Hints At Variant Effects To Protein Structure

In the previous sections, subtle variant effects to protein structure were observed for RMSD, PCA and Rg analysis. In addition results hinted at potential SNV effects to local protein structure (residues), DCC and RMSF analysis was performed to assess the impact of SNVs to CA residues.

4.3.2.1 DCC Analysis

Since the physical motions of atoms are computed during MD, DCC analysis was performed to analyse the extent to which protein residues move together [277]. DCC analysis of the CA-II WT and variant proteins is presented in Figure 4.9. Globally results indicate that along with the variant type, the presence of substrate has an effect on the correlation of residue motion. All seven protein sets show differing residue correlations and behaviour, as the proteins change from the apo to BCT and CO₂ bound states.

4.3.2.1.1 CA-II

Analysis of CA-II_{WT} illustrates that the apo and CO₂ bound states show greater correlated residue movement, whereas the BCT bound protein shows anti-correlated movement.

Factoring in Equation 1.1 the WT DCC results are expected. From Equation 1.1 the hydration of CO₂ is reversible, and BCT and CO₂ can both act as substrates for this reaction. Opposite residue behaviours are therefore expected for the BCT and CO₂ bound structures, that is, anti-correlated motion may be required to dehydrate BCT, whereas correlated motion maybe necessary to hydrate CO₂. Substrate CO₂ binds to the apo structure of the WT by possibly displacing the deep water molecule [50, 354] and is part of the forward reaction. Similarity to residue correlation is therefore expected between the apo and CO₂ bound proteins. Noting this P236H_{apo} may struggle to bind CO₂ during catalysis as this protein shows the greatest extent of residue anti-correlation.

Comparison of the WT and variant protein data when BCT is bound shows that all variants exhibit some degree of anti-correlation with exception to N252D. However, with minimal experimental research performed on BCT binding to the variants, it is hard to determine whether a possible relationship between residue anti-correlation and the rate of BCT dehydration exists. With regards to CO₂, K18E, H107Y, P236R and 252D show similar residue correlation as to that observed in the presence of BCT. Similarity in residue correlation between H107Y_{BCT} and H107Y_{CO₂} could be indicative of poor CO₂ hydration activity and explain the lower CO₂ hydration (64% of the WT) activity associated with the variant in previous studies [84, 271].

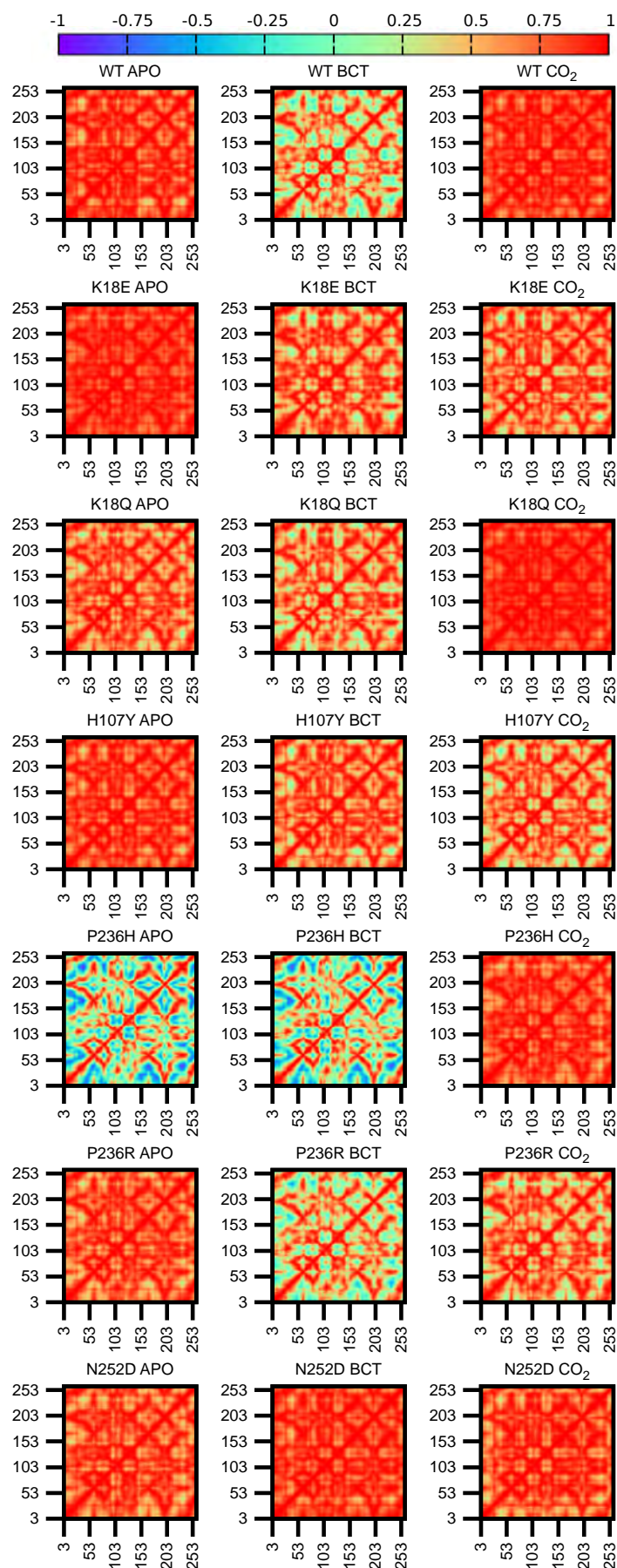


Figure 4.9. DCC analysis showing residue movement in CA-II. The x -axis and y -axis represent protein residues.

4.3.2.1.2 CA-IV

Residue correlation of the WT and variant CA-IV proteins is presented in Figure 4.10. Data highlights differences to the residue motions of the WT and variant proteins. In the WT, majority of the residues demonstrate no correlation to minor anti-correlated residue movements. Variants R69H, N177K and R219C show more correlated residue movement compared to the other proteins, while N86K, R219S and V234I show no correlation for a larger number of amino acids.

Comparison of residue motion between pathogenic and benign variants does not give a clear indication as to the differences in mechanism of SNV action. However, assessment of the CA-IV RMSD and DCC results hints at which variants could have a similar mechanism of action. From data in Figure 4.3, R69H and R219C showed similar structural sampling and a great degree of conformational overlap. DCC results for these variants also shows similar variant effects to residue correlation. These findings suggest that R69H and R219C may have a similar mode of pathogenesis in CA-IV. Conversely, with N86K and N117K though also showing similar structural sampling and some overlap, the DCC results show differences to residue movement. This could suggest differences to variant mechanisms. As CA-II and CA-IV form transport metabolons with other proteins, changes to residue correlation could also have an impact on PPIs however more research is required to confirm this. Impaired interactions with SLC4A4 for R69H and R219S has been observed in previous studies [107, 355].

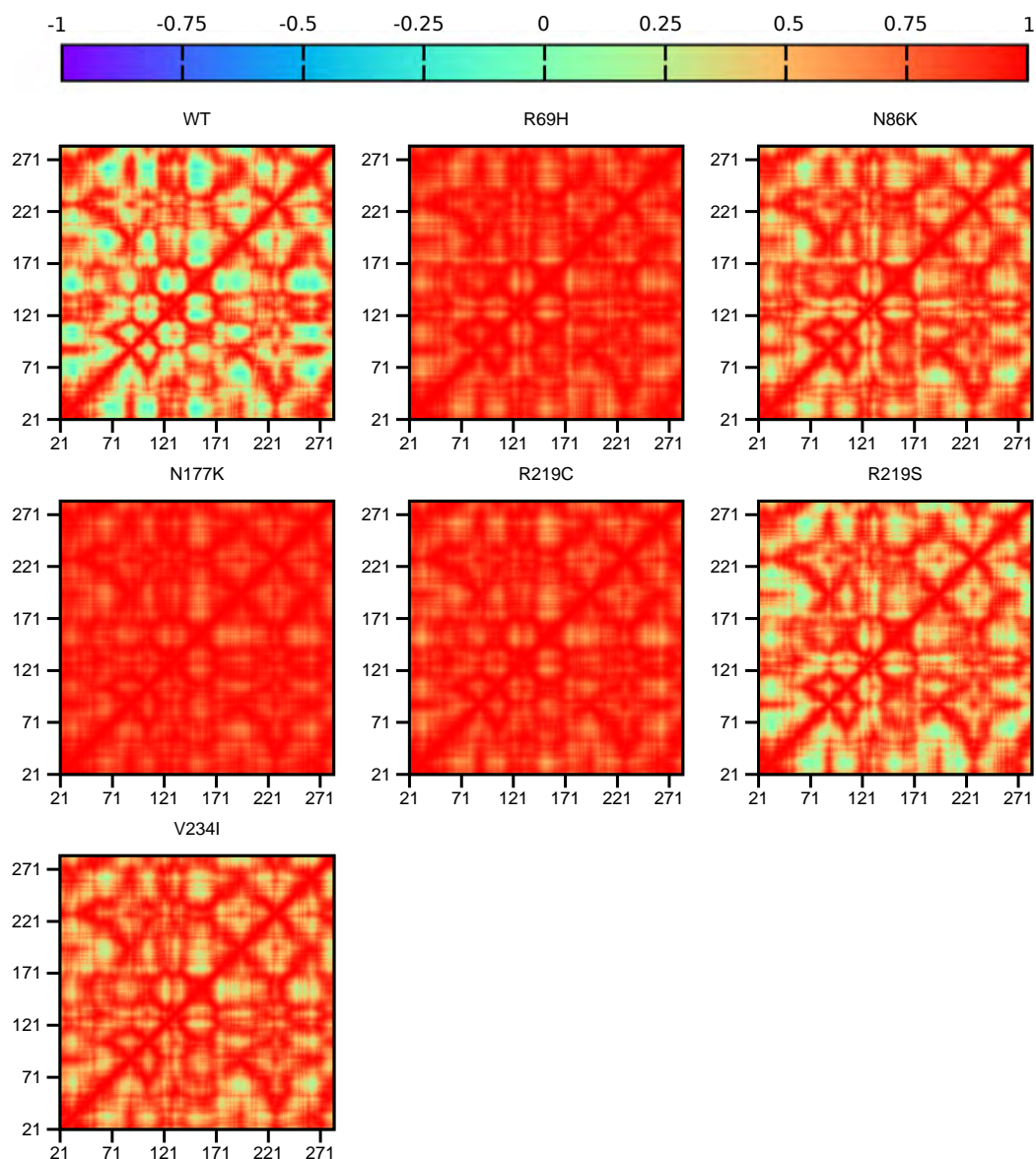


Figure 4.10. DCC analysis showing residue movement in CA-IV. The x -axis and y -axis represent protein residues.

4.3.2.1.3 CA-VIII

The correlations of CA-VIII residues during MD simulation are presented in Figure 4.11. Results demonstrate a clear difference to residue correlation of the WT compared to the variant proteins. The majority of amino acids within the WT protein exhibit anti-correlated movement whereas the variant residues exhibit no correlation to greater correlation to residue movement. Analysis of RMSD and DCC results suggests that this anti-correlated motion within the WT could be the cause the greater

conformational sampling observed (Figure 4.3). The anti-correlated movement observed in the WT results means that these residues could either moving towards each other or moving away from each other. Likewise, the lack of correlation to greater correlation of the variant proteins could also explain the increases to structural rigidity as observed in S100P and the smaller conformational sampling.

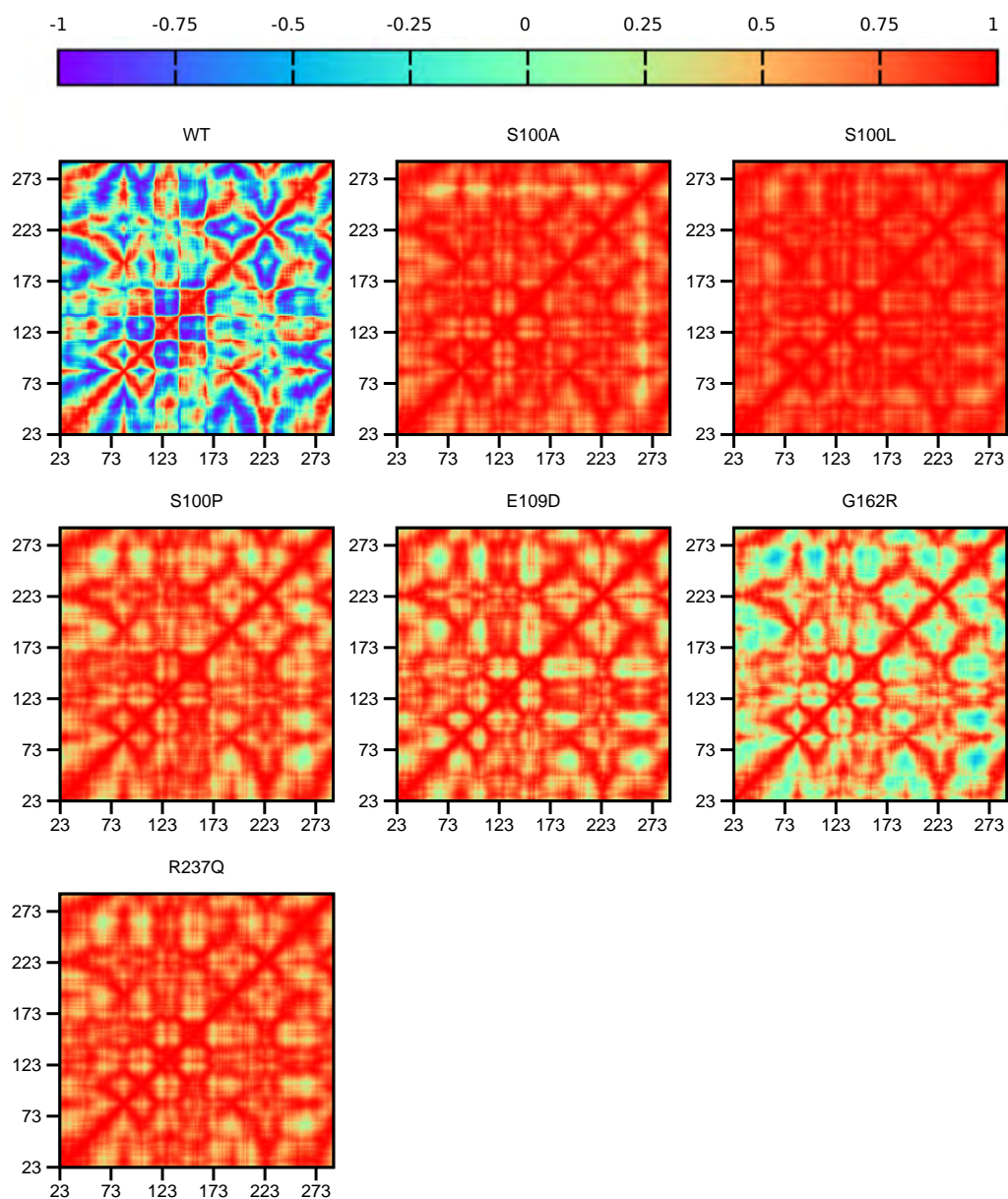


Figure 4.11. DCC analysis showing residue movement in CA-VIII. The x -axis and y -axis represent protein residues. Adapted from Sanyanga and Tastan Bishop 2020 [269].

Variant analysis in Figure 4.11 reveals that S100L has the most correlated residue movement compared to the other SNVs. No correlation to multiple of residues in S100A, S100P, E109D, G162R and R237Q was also observed. This lack of correlation is greatest in the benign variant E109D.

Combination of the CA-VIII binding site identification with the DCC results potentially highlights at variant mechanisms. The anti-correlated movement in the WT protein may be necessary to facilitate the allosteric changes in ITPR1 to regulate IP₃ affinity for the receptor.

The lack of correlation within S100A, S100L, E109D, G162R, S100P and R237Q potentially highlights disturbances to the variant protein networks. No correlation within these variants suggests that upon binding to ITPR1, inconsistent (random) movements and changes to the CA-VIII residues may result thereby affecting the allosteric function of the protein. Interestingly G162R shows minor anti-correlation to residue movement, and is the only variant showing this effect. Even though the differences between the WT and variant proteins are evident in Figure 4.11, the differences between benign and pathogenic variants cannot be determined using DCC alone.

To address the lack of correlation and possible benign mechanisms, the PCA results were also incorporated into DCC interpretation. It is noted that S100L is benign and shows more correlated residue movement. This suggests that upon binding to ITPR1 the variant protein may be able to achieve a constant and non-random motion. With regards to E109D, the lack of correlation suggests that E109D binding to ITPR1 could induce random non-consistent residue movements. The lack of a stable low energy conformation (Figure 4.6) however, also suggests that potential structural changes to the receptor or its flexibility may be able to induce changes in the conformational sampling of E109D to that of one that facilitates the allosteric regulation of IP₃. The flexibility of a receptor has been shown to have an effect on ligand docking in addition to ligand flexibility [356–358]. The pathogenic variants are more rigid and occupy lower energy structures, therefore changes to protein conformation may not occur readily thereby hindering the regulation of Ca²⁺ release.

In the next section the effects of correlation on residue flexibility were investigated to further understand variant effects on the CA proteins.

4.3.2.2 RMSF Analysis

The RMSF of the WT and variant CA-II protein residues for each system are presented in Figure S7. Results show that each individual variant has a different effect on the flexibility of the residues within each protein system. However from the data it is difficult to compare the effect of each SNV to the WT. To solve this problem, Δ RMSF (WT – variant) was calculated for each protein system and results are presented in Figure 4.12. A positive Δ RMSF is indicative of decrease to the flexibility of variant residues, whereas a negative Δ RMSF symbolises an increase to variant residue flexibility.

4.3.2.2.1 CA-II

Figure 4.12 apo results demonstrate that with respect to K18E and P236R, variant presence has an effect on the first 20 N-terminus residues, that cover motif 10 and include the initial aromatic cluster residues Trp5, Tyr7, Trp16 and Phe20 involved with stability maintenance within CA-II. Protein stability could therefore be affected by increases to the stability of these residues resulting in potential instability, as was predicted by the VAPOR results. Comparison of these two variants also illustrate that the K18E residues show larger changes to Δ RMSF which could be as a result of the variant being located within this group of residues. This effect is also noted in the RMSD of the apo K18E whereby structures show larger conformational sampling compared to P236R, which could be as a result of flexibility changes to motif 10 residues.

Analysis of the BCT bound proteins in Figure 4.12 highlights that K18Q and H107Y exhibit the greatest SNV effects. Decreases to Δ RMSF are observed between residues 230–240 that cover motif 3 and motif 8 amino acids. Increases to the flexibility of these residues could have an impact on protein stability. Increases to residue flexibility within the primary aromatic cluster (motif 10) is also noted in K18Q. Disturbances to the residues within the primary aromatic cluster could explain the greater conformational sampling observed within the RMSD results. Replacement of Lys with Gln

at position 18 also increases the flexibility of the residue, possibly through disruptions to interactions with neighbouring residues. This has however been investigated later on. Globally numerous residues in H107Y show a negative Δ RMSF indicating potential increases to flexibility of the majority of the protein structure. Flexibility increases are observed for; in motif 2 (104–124), motif 7 (166–186) and motif 9 (53–73). These motifs are all involved with the maintenance of CA-II stability, and Δ RMSF changes could also explain the greater conformational sampling observed within the RMSD. Motif 2 and motif 9 are also involved with catalysis therefore changes to residue flexibility could have an effect on protein function, and explain the poor activity associated with H107Y.

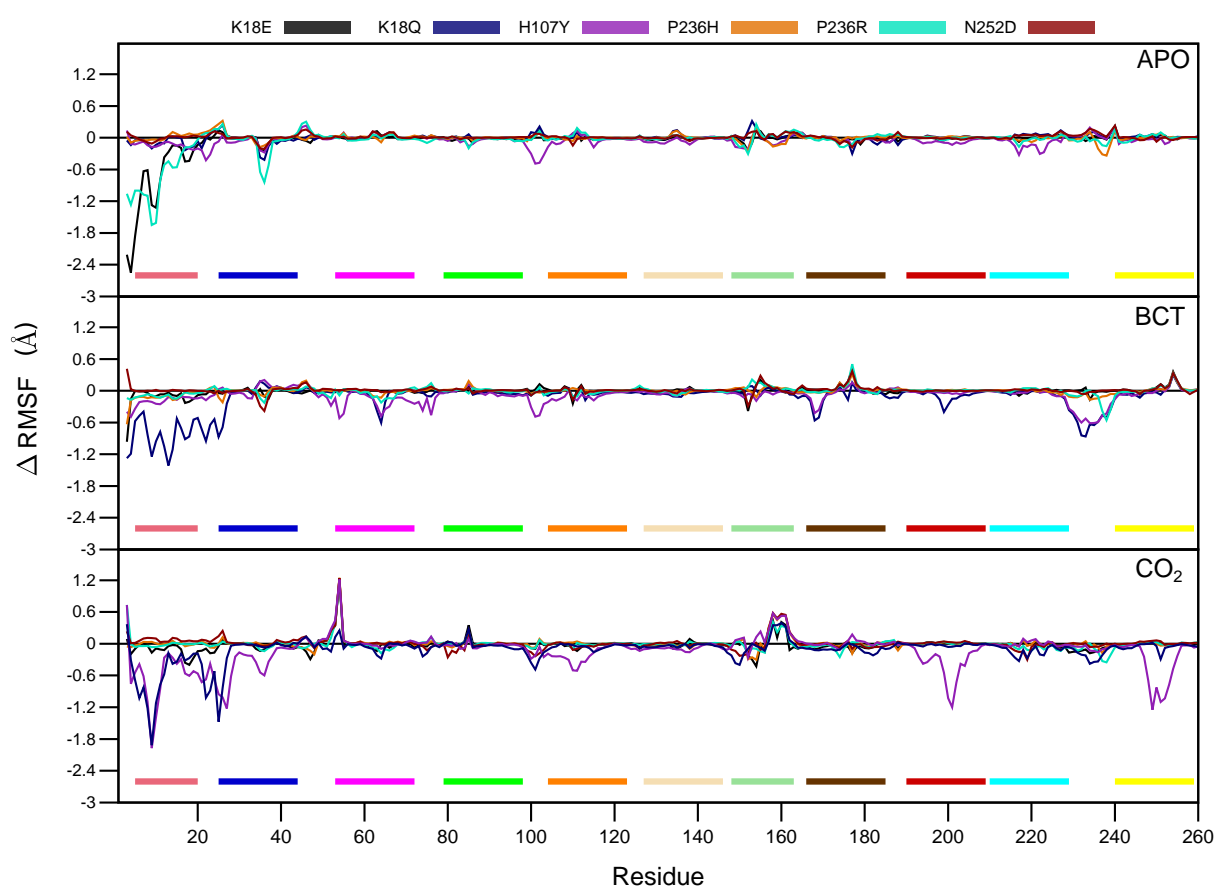


Figure 4.12. Δ RMSF comparison of the α -carbon atoms of the CA-II WT and variant protein systems (WT – variant). Colour coded bars at the bottom of each plot represent the similar coloured motifs in Figure 2.6.

Figure 4.12 results indicate that in the presence of CO_2 , decreases to flexibility of residues 53–54 (motif 9) are noted for all variants with the exception to K18Q. The flexibility decrease does not span

multiple residues therefore this change to Δ RMSF could affect the function of motif 9 with minimal impact. Inspection of K18Q_{CO₂} and H107Y_{CO₂} RMSF results in-conjunction with RMSD results hints at the role of motif 11. Within these variants increases to Trp5 and Tyr7 residue flexibility is observed. In addition, all variants display Δ RMSF increases to residues 157–162. Noting the large conformational sampling and structural clusters observed in Figure 4.2, this effect could be as a result of motif 10. The observation suggests that motif 11 may not assist with stability in CA-II. H107Y_{CO₂} exhibits the most changes to Δ RMSF compared to the other variants. Not only is an increase to the flexibility of motif 10 residues noted, flexibility increases are also evident in residues 192–208 (motif 3) and 246–254 (motif 3). This highlights that in the presence of CO₂ flexibility increases to H107Y are centred around the active site of the protein. Increases to Thr199 and Pro200 could affect the binding of CO₂ to the tertiary pocket of the protein, and potentially suggest that H107Y could have the greatest effect on the structure and function of CA-II.

Comparison of the three protein systems, apo, BCT and CO₂ indicates that substrate presence has the greatest effect on residue flexibility. Data also suggests that variant presence is associated with allosteric effects in CA-II that could affect structure and/or function. This is supported by SNV effects occurring further away from the variant location.

4.3.2.2.2 CA-IV

The residue RMSF of the CA-IV is presented in Figure S8. CA-IV results indicate similar residue flexibility between the WT and variant proteins except for residues 144–164. These amino acids are part of a loop secondary structure containing a small α -helix. Reduction to flexibility of these loop residues could also explain the rigid RMSDs observed in Figure S4 for the variant proteins. As with CA-II, Δ RMSF (WT – variant) for CA-IV was calculated to further resolve the RMSF differences, and results are presented in Figure 4.13. Data highlights residues 41–43 are associated with a slight reduction to flexibility in all variants except V234I. R219C and R219S are associated with a slight

increase to residue flexibility for residue 104 on motif 4. This motif may assist with stability (Table 2.3), however this increase to flexibility is unlikely to have an effect on protein function or stability, as His115, His117 and Trp118 (also located on motif 4) show no major changes to residue flexibility. For residues 144–164 variant proteins show flexibility reductions ranging from 0.6–4.4 Å. This result hints at significant variant effects on motif 6 (wheat). Flexibility reductions of these residues could be as a result of the poor protein folding associated with the CA-IV SNVs [119] and since motif 6 is located between residues that assist with CO₂ binding pocket formation (Val142 and Val165), reductions to its flexibility may affect interactions with the substrate and cause the reported variant associated toxic gain of function or poor enzyme activity [119]. What is unexpected however is that both benign and pathogenic variants show flexibility reductions to these residues. This indicates that in addition to protein flexibility reductions, differences to pathogenic and benign SNV mechanisms may be more complex and could involve disturbances to interactions with other key residues which are not visible within the RMSF results. Although motif 6 was assigned as CO₂ binding pocket formation, its conservation in CA-VIII could suggest that a secondary role of motif 6 that is yet to be discovered.

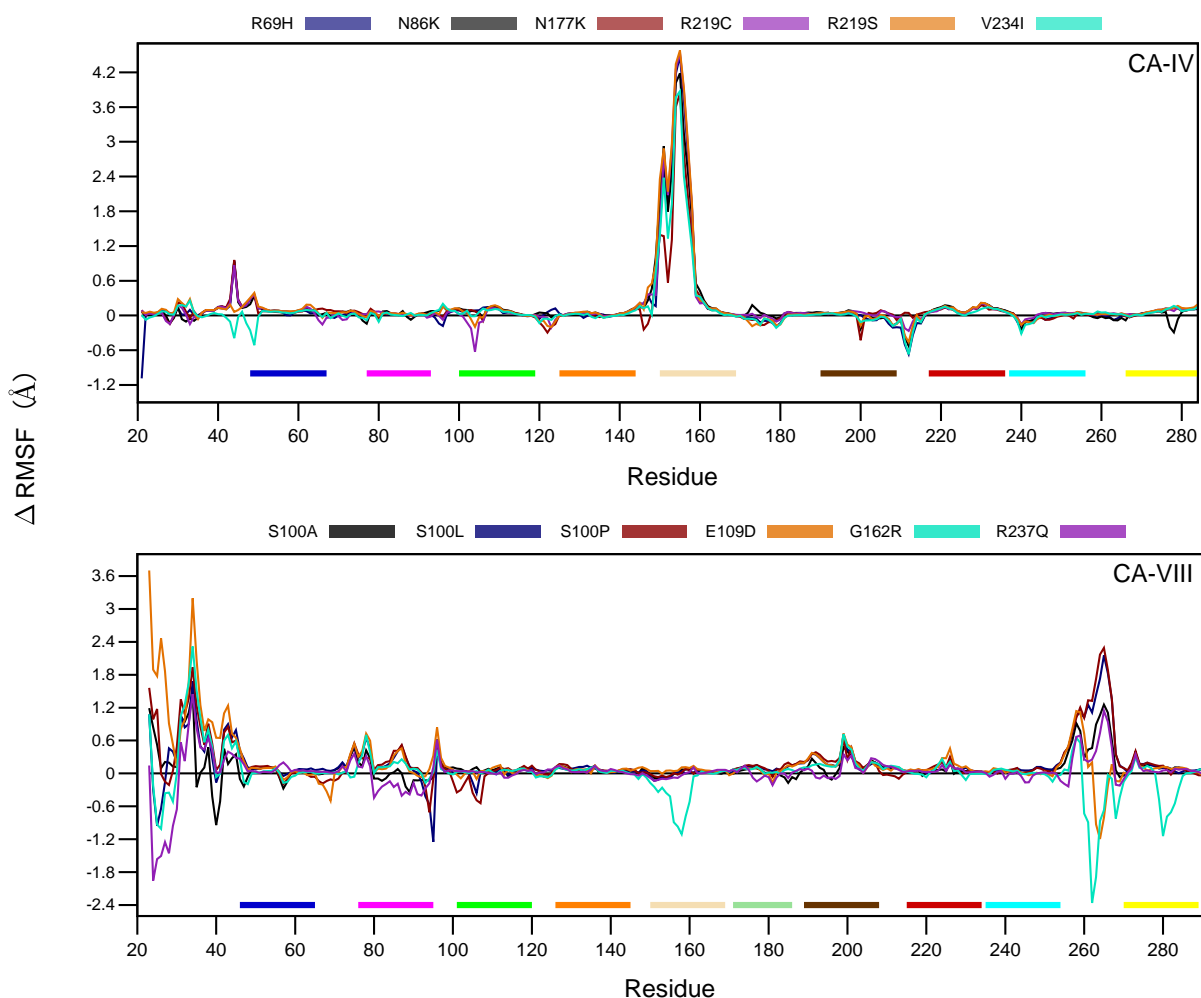


Figure 4.13. Δ RMSF comparison of the α -carbon atoms of the CA-IV and CA-VIII WT and variant protein systems (WT – variant). Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6. CA-VIII adapted from Sanyanga and Tastan Bishop 2020 [269].

4.3.2.2.3 CA-VIII

The RMSF of CA-VIII residues over the MD simulation is presented in Figure S8, and as with CA-IV the Δ RMSF (WT – variant) was also calculated. Analysis of CA-VIII results in Figure 4.13 demonstrates large changes to Δ RMSF for the residues 23–46 and 253–275. Increases to flexibility of these regions is expected as these residues are located with loop secondary structures. It should however be noted that within these regions, the CA-VIII binding site residues (Figure 2.4) are also contained. Flexibility reductions in all variants are observed for; Trp29, Gly30, Tyr31, Glu32, Glu33, Gly34, Leu39 and Ala44 which are all members of the N-terminus binding site residues. Reductions to the flexibility of Trp37 are also noted. Trp29, Tyr31 and Trp37 potentially assist with stability of

CA-VIII therefore reductions to residue flexibility could have an effect on stability. In addition, the binding site residues, Thr255, His256, Leu262, Val263, Glu264, Gly265, Ile269 and Phe274, and the stability assisting Arg275 also show a reduction to residue flexibility at the C-terminus of the protein, with exception to E109D and G162R that show increases to residue flexibility.

Amino acids showing consistent increases to Δ RMSF to the global protein structure include 35–48, 73–79 and 181–214. More variant residues have a Δ RMSF that is greater than zero indicating a wide spread reduction to protein flexibility and potentially resulting in more constrained and rigid protein structures [145]. This could explain the lower conformational sampling as was observed within the RMSD and greater compactness in the R_g results. Analysis of G162R data illustrates further increases to flexibility for the residues 150–160. Gly151 and Gly153 have been predicted as potential binding site residues therefore, increases to the flexibility of these residues could have an effect on protein interaction with ITPR1. Decreases to residue flexibility in the variant proteins could also explain the increase to correlation observed in DCC.

Overall CA-IV and CA-VIII RMSF results are in agreement with those of CA-II. Variants do not cause RMSF changes at the variant location but elsewhere within the protein. This further suggests that the CA variants could have allosteric effects on protein function and structure. In the next sections we explore DRN to analyse the SNVs effect at residue level and identify the potential allosteric effects of variants.

4.3.3 Short Range Residue Interactions Are Affected By Variant Presence

Data from previous sections highlighted that the variants were associated with subtle effects to the global structure of the CA proteins, and potentially had an allosteric impact on the structure and function of the CA proteins. Analysis of the changes to interactions occurring as a result of SNV presence and their effects on the protein network is useful in the identification of SNV mechanism of action, and effects on residues important for communication, function and stability [18, 66,

359–361]. Contact map analysis was performed to identify and analyse short-range changes to residue interactions occurring as a result of SNV presence. These interactions were calculated and set up to include all residues within a threshold of 6.7 Å participating in; vdW, hydrogen bonds and/or electrostatic interactions with the SNV residue over MD simulation [277].

4.3.3.1 CA-II

A heat map of the frequency interactions occurring in CA-II during MD is presented in Figure 4.14.

A value of 1 indicates constant interaction of protein residues during MD, whereas a value of 0 shows that there was no communication between residues during MD simulation.

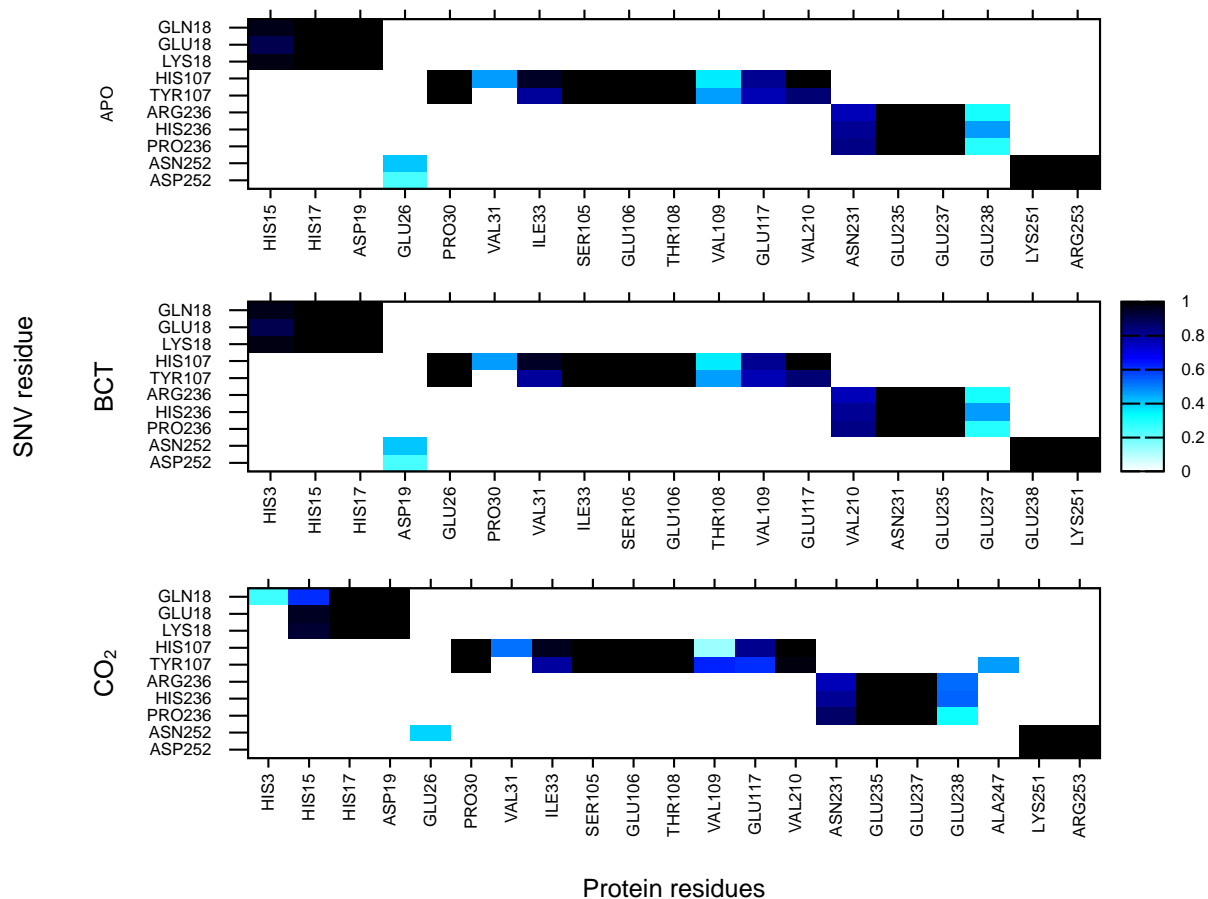


Figure 4.14. Contact map weighted interactions of the CA-II WT and SNV residues. Adapted from Sanyanga *et al.* 2019 [66].

Data from Figure 4.14 demonstrates that K18E (Glu18) is not associated with any changes in interactions with residues. Gln18 in K18Q however forms new interactions with His3 when substrate

CO₂ is bound. Formation of these interactions with His3 may not have an effect on protein structure as this residue is neither conserved nor located on any motif (Table 2.3). Interestingly when the RMSD analysis of K18E and K18Q are combined with the contact map analysis, this assists with the identification of variant mechanisms of action. Decreases in interactions between Glu18/Gln18 and His15 are associated with greater conformational sampling (Figure 4.2). K18E_{apo} shows a decrease in interactions between His15 and Glu18, and the corresponding RMSD distribution shows an increase to conformational sampling. With respect to K18Q_{BCT} and K18Q_{CO₂}, Gln18 also exhibits decreases to interactions with His15 and the corresponding RMSD distribution shows greater structural sampling indicating potential instability. Motif 10 is involved with protein stability therefore interaction losses could affect protein stability. Previous studies in 1988 by Eriksson [67] showed that CA-II maintained function and hydrated CO₂ at rate constant of $1.5 \times 10^5 \text{ s}^{-1}$ in the absence of the first 23 N-terminal residues. This suggests that interaction loss with His15 could have limited effect on CA-II catalytic function.

Assessment of H107Y results in Figure 4.14 shows a complete loss of interactions between Tyr107 and Val31. Decreases to interactions are also observed with residue Glu117 for all three protein states. The largest decreases to the interactions are observed when H107Y is in the presence of substrates BCT and CO₂. Glu117 is a secondary Zn²⁺ ligand that stabilises the metal ions through direct interaction with the primary ligand His119 (see Figure 1.4) [19, 68]. Comparison the loss of interaction between Tyr107 · Val31, and Tyr107 · Glu117 highlights that the interaction loss is greater for Val31. This observation can be explained by the role of Glu117 in Zn²⁺ affinity. Since Glu117 assists with stabilising Zn²⁺, additional compensatory interactions could have been formed between Tyr107 and Glu117, and possibly neighbouring residues to maintain the integrity of Glu117 within the protein. DRN analysis was investigated later in this chapter to identify whether compensatory measures occur within H107Y to maintain protein structure and function. Tyr107 in H107Y_{CO₂}

forms new interactions with Ala247 that is located on motif 3. These interactions could assist with stability maintenance within the protein.

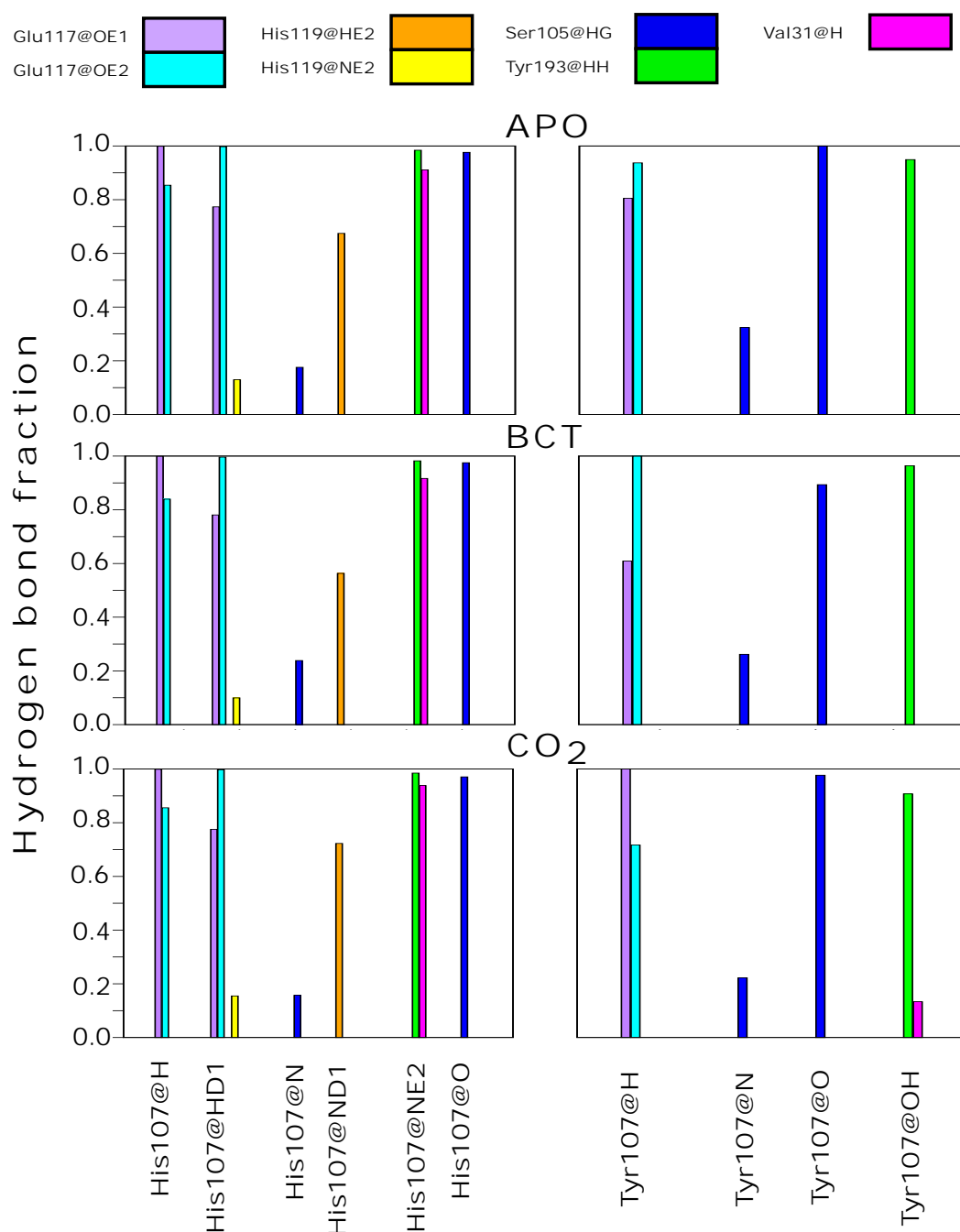


Figure 4.15. Proportion of the total MD simulation frames the hydrogen bond existed (hydrogen bond fraction) between residue 107 and neighbouring atoms within the WT and variant proteins. Left: WT. Right: H107Y. Adapted from Sanyanga *et al.* 2019 [66].

Since H107Y showed decreases to interactions with the important residue Glu117, hydrogen bond analysis was performed to further investigate the variant effects. Contact maps do not differentiate between interaction types therefore additional analysis was required. Within proteins, hydrogen

bonds are one of the stronger interaction types, therefore any changes or disruptions to hydrogen bonding could have significant impact on the structure and function of the protein.

Data in Figure 4.15 shows hydrogen bond analysis of Tyr107 with its neighbouring residues, and compares the hydrogen bond loss for all three protein systems. Data is presented as hydrogen bond fractions which represents the duration of the MD simulation hydrogen bonds existed. Only residues with a hydrogen bond proportion greater than 0.1 were regarded as having formed stable and not intermittent hydrogen bonds with Tyr107. Results in Figure 4.15 show hydrogen bond loss occurs between Tyr107 and the residues; Val31, Glu117 and His119. The Tyr107 · Val31 interaction loses one hydrogen bond whereas the Tyr107 · Glu117 and Tyr107 · His119 interactions lose two hydrogen bonds each. This loss is in agreement with previous studies that noted the loss of at least two hydrogen bonds between Tyr107 and Glu117 in H107Y [81, 83, 271, 279]. This study has however identified a novel finding whereby, Tyr107 also loses hydrogen bonds with the primary Zn²⁺ coordination ligand His119. Disruptions to interactions with Zn²⁺ coordinating residues would have an impact on metal ion stability, and could result in the protein misfolding and active site distortion. The mechanism of H107Y could thus involve Zn²⁺ destabilisation and/or an increased tendency for Zn²⁺ to dissociate from the active site. Glu117 has previously been found to have an effect on the dissociation of Zn²⁺ from the CA-II active site [19, 68, 101].

Overall His107 forms 10 hydrogen bonds with neighbouring residues, whereas Tyr107 only forms 5. The result indicates that the presence of Tyr107 in CA-II results in a loss of half the hydrogen bonds. Previous studies in 1991 by Venta *et al.* [83] suggested that Tyr107 also lost a hydrogen bond with Tyr193. Within this study no evidence of hydrogen bond formation between these two residues was noted.

Contact map results of P236H and P236R suggests differences in the mechanism of the variants. Increases to interactions between His236 and Glu238 were observed for the P236H_{apo} and P236H_{CO₂}

proteins, while in the P236R_{BCT} and P236R_{CO₂} proteins, Arg236 also shows reduction to interactions with Glu238. Glu238 is situated between motif 3 and motif 8, and interaction changes could have an effect on the neighbouring motifs. The N252D_{apo} and N252D_{BCT} proteins demonstrate decreases to interactions between Asp252 and Glu26. Interestingly interactions with Glu26 are completely lost when CO₂ is bound to N252D. Comparison of the N252D RMSD and contact map results suggest that interactions with Glu26 may have minimal importance on the stability of CA-II as a large conformational sampling was not observed. This however does not rule out the variant effects on catalysis. Results however do further support the findings that the presence of substrate could affect the mechanism of the variants.

4.3.3.2 CA-IV

Contact map interactions between the CA-IV SNVs and neighbouring residues are presented in Figure 4.16. Data demonstrates that in R69H shows a reduction to weighted interactions with Phe71 while there is a complete loss of interactions with Gly102. Assessment of the data in Table 2.3 reveals that neither of these residues have an assigned function. Phe71 is however a ringed amino acid and is next to Phe70. These Phe residues could have a stabilising effect on the structure of CA-IV in a similar mechanism as that observed within the primary and secondary aromatic clusters of CA-II. Previous studies into R69H had found that there are no losses in interaction between the His69 and Gly103 [355]. This is in agreement with our findings as Figure 4.16 data does not show any loss of interactions with Gly103.

N86K forms new contacts with Val90. The increases to interactions with this residue could help stabilise the CA-IV protein. Using the data in Table 2.3, Val90 aligns with Phe66 of the secondary aromatic cluster of CA-II. Though not aromatic, this could suggest that Val90 could have an adaptive stabilising effect for the GPI anchored CA-IV. Formation of these new interactions could also explain the benign nature of the variant. N177K demonstrates a slight decrease to interactions with Gln181

and loses all contacts with His245. This variant however gains new contacts with Ile245.

Comparison of the WT, R219C and R219S proteins illustrates a decrease in contacts with Leu67 within the variants. A complete loss of interactions with Lys283 and Ser284 is also observed. As Ser284 functions as the omega-site for the GPI anchor in CA-IV, losses to interactions with these proteins could have an effect on the attachment of the GPI anchor and could explain the impaired protein trafficking to the cell surface observed in previous studies [119]. Lys283 and Ser284 are also located on motif 3 which assists with stability. These interaction losses could also have implications for protein stability. R219C and R219S forms new contacts with Val234. Residue 234 is also the SNV location of the benign variant V234I. Increases with Val234 could be a compensatory measure to help the protein maintain stability.

R219C also forms an additional contact with Gly104. Increases in contacts with Trp235 are observed within both variants as well. Increases in interactions with this residue could have implications for CO₂ binding to CA-IV. This could also explain the extremely poor activity (loss of at least 90% of hydration activity) associated with the R219S variant [119]. This could also explain the difference in activities with R69H which only loses 14% of CO₂ hydration activity [119]. The mild loss observed in R69H could be as a result of no interference to Trp235 of the CO₂ binding pocket (see Table 2.2). Previous experimental analysis on the impact of interference of CO₂ binding site residues on CA-II activity have been performed [362]. Results showed decreased catalytic function, when Val142 was substituted to Tyr142 in addition to disrupted metabolon function. This effect could also apply to CA-IV. V234I analysis indicates a slight decrease in weighted contacts with Leu105.

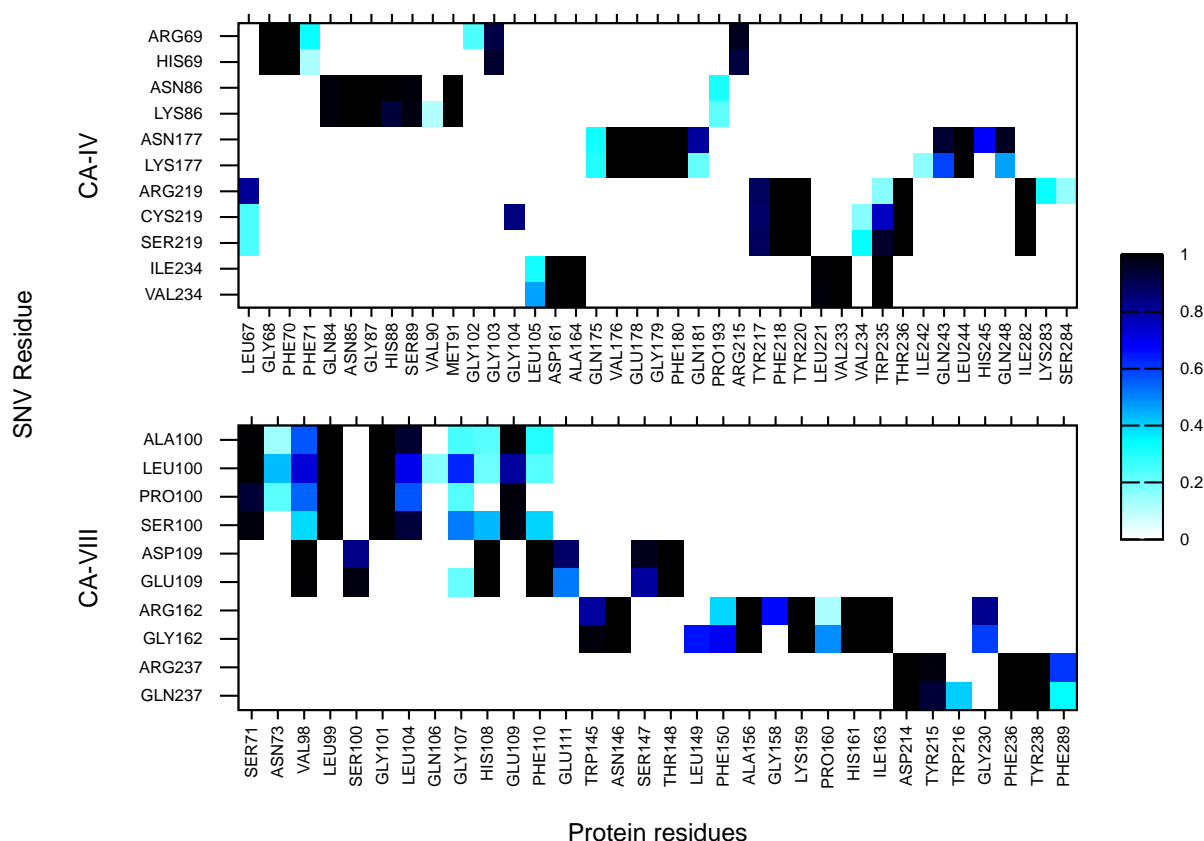


Figure 4.16. Heat map presenting the contact map weighted interactions between the SNV residues in CA-IV and CA-VIII, and respective neighbouring amino acids.

4.3.3.3 CA-VIII

Data in Figure 4.16 also presents the changes to weighted interactions between the WT and variant CA-VIII proteins. Comparison of CA-VIII_{WT}, S100A, S100L and S100P shows an increase to interactions with Asn73 and Val98 for the variant proteins. The effect is greater in S100L. In addition, S100L forms new contacts with Gln106 and shows a minor reduction to weighted contacts with Glu109. Comparison of S100P to the WT indicates that the variant protein loses interactions with His108 and Phe110. Interestingly, pathogenic variants S100A and S100P show a significant reduction in interactions with Gly107 compared to the WT and benign S100L. E109D data also highlights at the formation of a new contact with Gly107. Results suggest that interactions with Gly107 could have an effect on variant stability and CAMRQ3 pathogenesis. This is further supported by the presence of Gly107 on motif 4 (Table 2.3 and Figure 2.6) which could potentially assist

with CA-VIII stability. G162R results demonstrate the formation of interactions with Leu149 and increases to interactions with Phe150. R237Q forms new contacts with Trp216 and shows a reduction in interactions with Phe289. Overall results show changes to contacts with some residues not listed in Table 2.2, and of unknown functional significance. Results could suggest an acatalytic adaptation of these residues to assist with key function and stability roles within CA-VIII.

4.3.4 SNVs Are Associated With Changes To Residue Accessibility and Communication

As data analysis from the previous sections hinted at possible allosteric (indirect) SNV effects to CA structure and function, and potential disturbances to the protein network. In this section the L and BC analysis was used to investigate SNV associated effects to the protein network and to identify allosteric mechanisms.

4.3.4.1 Average Shortest Path

The accessibility of residues within a protein is determined using L . The L for each residue was determined across every 100 MD frames and averaged to determine the mean accessibility of the CA residues during MD.

Data in Figure S9 compares the average L of the WT and variant CA-II proteins during MD simulation. From the results it is difficult to differentiate the changes to residue accessibility as a result of SNV presence, therefore ΔL (WT – variant) was calculated, and results are presented in Figure 4.17. Positive ΔL are due to a decreased average L value with respect to the WT for a specific residue. Hence increased accessibility for that specific residue within the variant protein, whereas negative values are indicative of decreases to residue accessibility within the variant proteins.

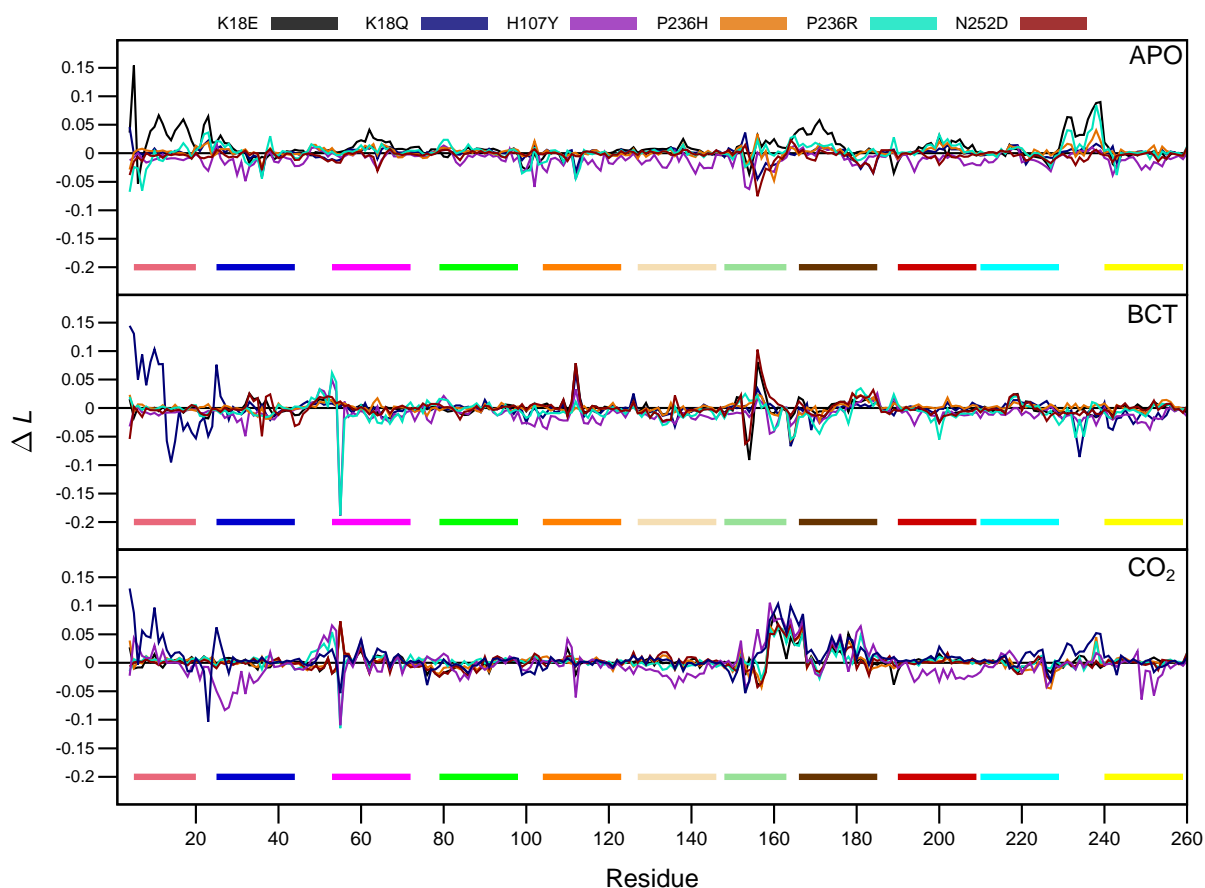


Figure 4.17. ΔL (WT – variant) comparison of the CA-II protein systems. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6.

4.3.4.1.1 CA-II

Globally, comparison of ΔL for all three protein systems indicates that SNV effects are more pronounced when substrates are bound to the protein. Motif 9 (magenta region) residues within all variants in the presence of BCT and CO_2 demonstrate differing behaviours possibly suggesting different variant mechanisms. Accessibility decreases to this motif are observed for K18Q, H107Y and P236R. Motif 7 and motif 11 residues ($\approx 150-180$) show increases to accessibility in the presence of CO_2 . This finding complements the RMSF results whereby increases to ΔRMSF were also noted for this group of residues. The RMSF and L metrics have previously been found to share a linear correlation [348].

Residues with a ΔL less or greater than two standard deviations from the mean were then identified

and are presented in Table 4.2. . These residues are regarded as showing significant changes to residue accessibility. K18E_{apo} with respect to the motif 10 residues His10 and Trp16 shows an increase to accessibility. With regards to K18Q, when substrates BCT and CO₂ are bound the N-terminal residues display similar behaviour of showing an increase to residue accessibility. This region covers residues 3–11 and contains Trp5 and Tyr7 of the initial aromatic cluster. If residues within the initial aromatic cluster become more accessible to each other, this could have detrimental effects to the rest of the structure as they could become less accessible to other residues, thus reducing structural stability elsewhere. This is further supported by the decrease to the accessibility of Trp5 in the K18E apo protein. The RMSD distributions for K18E_{apo}, K18Q_{BCT} and K18Q_{CO₂} show greater conformational sampling and this could be an effect of this. K18E and P236R apo proteins display residue accessibility increases for Phe230 and Glu238. This finding could be due to the increase to interactions between Arg236 and Glu238 observed within the contact maps. H107Y_{apo} results shows decreases to the accessibility of Ser29 which assists with stability. Decreases to the accessibility of Val31 are also observed to the protein when in the apo and BCT bound state. This reduction to residue accessibility could be due to the loss of interactions observed with the contact map analysis.

ΔL results further hint at the allosteric effects of the variants as most of the changes to residue accessibility occur away from the SNV location with exception to P236H and P236R. The majority of the residues however have a ΔL close to zero this showing that the SNV effects do not affect the global protein structure to a great extent, and could have an effect on the local residues and the network.

Table 4.2. CA-II residues showing significant changes to accessibility during MD simulation. Important CA-II residues from Table 2.2 are highlighted in bold and underlined. SNV positions are underlined, italicised and highlighted in bold red. Adapted from Sanyanga *et al* 2019.

Variant	Apo	BCT	CO ₂
Positive ΔL (Residue accessibility increase)			
K18E	His4 His10 Trp16 Ile22 Gly170 Phe230 <u>Asn231</u> Glu235 Pro236 Glu237 Glu238	Lys111 Gly155 Leu156	Ala54 Lys158 Val159 Val160 Asp161 Leu163 Asp164 Ser165 Ile166 Asn177 Phe178
K18Q	His3 Ile22 Ala152	His3 His4 Trp5 Gly6 Gly8 Lys9 His10 <u>Asn11</u> Lys24 Lys111	His3 His4 Lys9 Lys24 Lys158 Val159 Val160 Asp161 Leu163 Asp164 Ser165 Ile166
H107Y	Lys24	Val49 Ser50 Asp52 Gln53 Lys111 Leu184	Asp52 Gln53 Gly155 Lys158 Val159 Val160 Asp161 Val162 Leu163 Ser165 Ile166 Asp179 Pro180
P236H	Ile22 Asp101 Gly155 Thr199 His236 Glu237 Glu238	His3 Gly155 Leu156 Phe178 Asp179 Pro180 Arg181 Leu183 Pro201	His3 Ala54 Lys158 Val159 Val160 Asp161 Val162 Leu163 Asp164 Ser165 Ile166 Ala173 Phe175 Asn177 Glu237
P236R	Pro21 Ile22 Phe230 Asn231 Glu235 Arg236 Glu237 Glu238	Val49 Asp52 Gln53 Pro180	Asp52 Ile59 Lys158 Val159 Val160 Asp161 Val162 Leu163 Ile166 Pro180 Glu237
N252D	Leu163 Asp164	Lys111 Gly155 Leu156 Gln157 Leu183	Ala54 Lys158 Val159 Val160 Asp161 Val162 Leu163 Asp164 Ser165 Ile166 Phe175
Negative ΔL (Residue accessibility decrease)			
K18E	Trp5 Lys111 Lys153 Leu188	Ala152 Lys153	Gly155 Leu156 Leu188
K18Q	Thr35 Lys111 Pro154 Gly155 Leu156 Lys158 Val159 Pro180 Gly182 Leu183 Glu220	Gly12 Pro13 Glu14 Asp19 Leu163 Glu233	Ile22 Ala54 Ala152
H107Y	Ser29 Val31 Asp101 Lys111 Leu140 Ala152 Lys153 Val241	Val31 Ala54 Lys158	Gly25 Glu26 Arg27 Ala54 Lys111 Gln248 Lys251
P236H	His3 Thr35 Ala152 Lys158 Val159 Val160	His36 Thr37 Leu163 Asp164 Gly170	Leu156 Phe225 Arg226
P236R	His3 Gly6 Tyr7 Thr35 Gly98 Ser99 Leu100 Lys111 Asp242	Ala54 Leu163 Asp164 Gly170 Thr199 Gly232 Gly234	Ala54 Leu156 Gly170
N252D	His3 Thr35 Gly63 Pro154 Gly155 Leu156 Gln157 Lys158 Pro180 Leu183	His3 Thr35 Ser43 Ala152 Lys153	Gly155 Leu156 Arg226

4.3.4.1.2 CA-IV

The average L of the CA-IV WT and variant proteins during MD are shown in Figure S10. Greater average L values indicate decreases to accessibility. CA-IV data illustrates that the least accessible residues within the proteins are located at approximately positions 32–35, 65, 148–157 and 260–263. To resolve the average L differences, ΔL of the CA-IV WT and variant proteins was calculated and results are presented in Figure 4.18. Data illustrates that most of the ΔL values of the proteins are located close to 0 indicating minimal changes to residue accessibility across the protein. Results however demonstrate increases to the accessibility of residues 148–157 (motif 6). Increases to the accessibility of these residues could be due to the reduction to residue flexibility of the same amino acids within the RMSF results (Figure 4.13). Changes to the accessibility of these residues could have implications for metabolon formation and stability. Residues 148–157 are located on a loop. Functionally and structurally, loops are involved with facilitating protein-protein interactions, and linking secondary structure elements together [363, 364]. The CA-IV residues involved with metabolon complex formation have yet to be identified, however within CA-VIII motif 6 contains residues essential to the binding of the protein to ITPR1 [268]. This therefore suggests that in CA-IV residues 148–157 could contain amino acids essential to metabolon formation.

N86K results show an interesting finding, data demonstrates a decrease to the accessibility of residues 276–281. As this group of residues is located in close proximity to the omega site of CA-IV (Ser284), decreases to accessibility could have implications for the attachment of the GPI anchor. Given the benign nature of the variant, data suggests that this effect could be a compensatory mechanism that assists with stability of the omega site. It is still however yet to be investigated as to the potential effects of N86K on the trafficking of CA-IV to the cell surface. It should be noted that as proteins function as a network, neighbouring residues also have an effect on residue accessibility [18] as a result of their movements during protein dynamics. Therefore changes to the accessibility of

a specific residue does not mean that the specific residue itself has moved. R219C shows accessibility increases approximately to residues 95–110. This effect could be due to the increase in weighted interactions with Gly104 previously observed. CA-IV WT and variant protein residues showing changes to ΔL greater or less than two standard deviations are presented in Table 4.3.

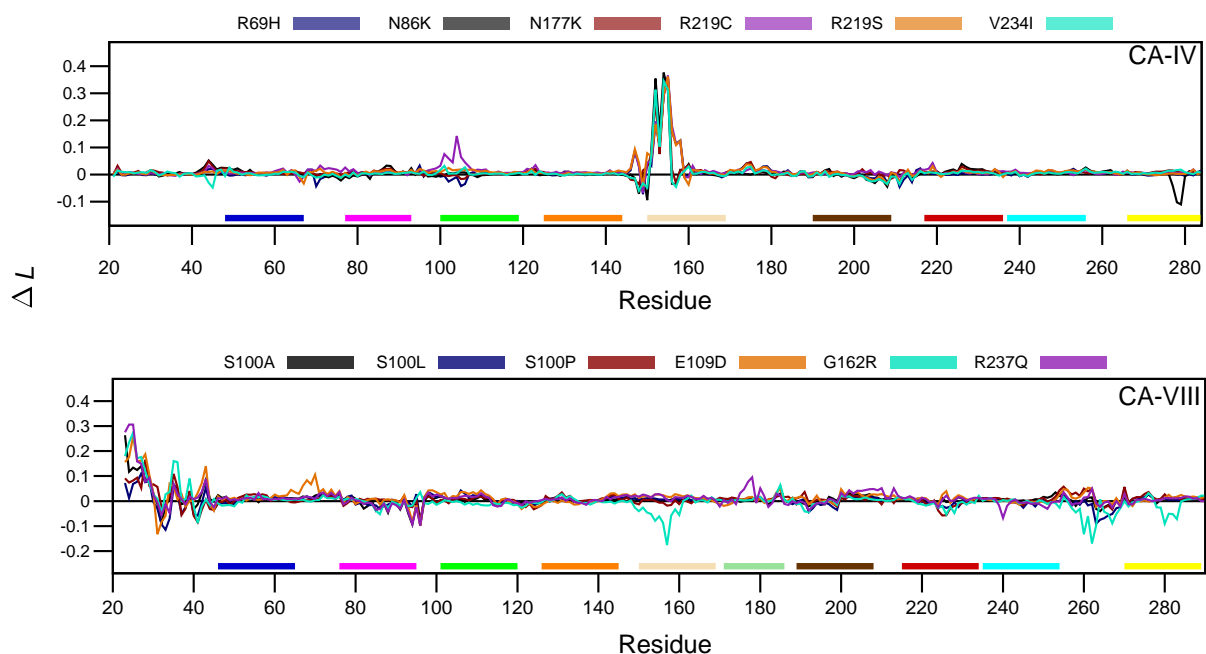


Figure 4.18. ΔL (WT – variant) of the CA-IV and CA-VIII respectively during MD simulation. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6. CA-VIII adapted from Sanyanga and Tastan Bishop 2020 [269].

Data in Table 4.3 illustrates that N177K, V234I and R219S do not contain any residues showing significant accessibility decreases. R219C data indicates that the accessibility change between residues 95–110 is significant. Global inspection of the data in Table 4.3 indicates a potential difference between the accessibility of residues within the benign and pathogenic SNVs. Though all variants show accessibility increases between residues 148–157 the difference is that while all benign variants show increases to residues 152–155, in the pathogenic variants additional increases to the accessibility of residues Ala156, Gln157 and Asp158 are observed. The functional significance of these residues to CA-IV is yet to be fully determined.

Table 4.3. CA-IV residues showing changes to ΔL greater or less than two standard deviations.

Variant	Positive ΔL (Residue accessibility increase)
R69H	Lys147 Asn152 Val153 Lys154 Glu155 Ala156 Gln157 Asp158
N86K	Asn152 Val153 Lys154 Glu155
N177K	Asn152 Lys154 Glu155
R219C	Gly104 Asn152 Val153 Lys154 Glu155 Ala156 Gln157 Asp158
R219S	Lys147 Asn152 Val153 Lys154 Glu155 Ala156 Gln157 Asp158
V234I	Asn152 Val153 Lys154 Glu155
Negative ΔL (Residue accessibility decrease)	
R69H	Thr149
N86K	Ser150 Gln278 Arg279
N177K	
R219C	Thr149
R219S	
V234I	

4.3.4.1.3 CA-VIII

The average L analysis of CA-VIII reveals that the N-terminal and residues 259–266 are associated with the least accessibility (Figure S10). The ΔL results are presented in Figure 4.18 and initial inspection highlights that WT and variant ΔL values remain close to 0 indicating minor changes to global residue accessibility. Residues showing ΔL changes greater or less than two standard deviations were then extracted from Figure 4.18, and data is presented in Table 4.4. Results show that most of the amino acids showing changes to ΔL comprise Glu rich N-terminal residues (residues 21–36) within all variants. S100L N-terminal residues 26–29 and 35 demonstrate an increase to accessibility, while residues 32–34 indicate a decrease. Additionally, residues 263–265 also show accessibility decreases in S100L. E109D results show similarity with those of S100L. Binding site residues 26–29 show accessibility increases, whereas 31–33 are associated with decreases to accessibility.

The increases and decreases in accessibility to the N-terminal binding residues of S100L and E109D

could indicate a compensatory mechanism whereby, as one set of residues moves further apart another set of residues move closer together to balance out the effect. Since the green and red binding site regions in Figure 2.4 are in close proximity, the accessibility changes could assist with the maintenance of binding site integrity. This mechanism could be essential for the maintenance of binding site integrity, and could explain the variant clinical significance of benign.

Table 4.4. CA-VIII residues showing changes to ΔL greater or less than two standard deviations. SNV positions are underlined, italicised and highlighted in bold red. Residues located within the CA-VIII binding site are underlined and highlighted in bold blue. Important CA-VIII residues from Table 2.2 are highlighted in bold and underlined. Overlapping potential PPIs and important structural residues are underlined and highlighted in bold green. Adapted from Sanyanga and Tastan Bishop 2020 [269].

Variant	Positive ΔL (Residue accessibility increase)
S100A	Glu23 Glu24 Glu25 <u>Gly26 Val27 Glu28 Trp29 Val35</u> Asp43
S100L	Glu23 Glu25 <u>Gly26 Val27 Glu28 Trp29 Val35</u> Asp43
S100P	Glu23 Glu24 Glu25 <u>Gly26 Glu28 Trp29 Gly30 Val35 Thr255</u>
E109D	Glu23 Glu24 Glu25 <u>Gly26 Val27 Glu28 Trp29</u> Asp43 Leu70
G162R	Glu23 Glu24 Glu25 <u>Gly26 Val27 Glu28 Val35 Glu36 Leu39</u>
R237Q	Glu23 Glu24 Glu25 <u>Gly26 Val27 Trp29</u> Gly178
Negative ΔL (Residue accessibility decrease)	
S100A	<u>Glu32 Lys94</u> Lys96
S100L	<u>Glu32 Glu33 Gly34 Trp37</u> Val40 <u>Phe41</u> Lys96 <u>Val263 Glu264 Gly265</u> Cys266 Asp267
S100P	<u>Trp37 Val40 Phe41</u> Lys96 Pro225 Pro226
E109D	<u>Tyr31 Glu32 Glu33 Lys94</u> Lys96
G162R	Val157 Ala260 Leu262 <u>Val263</u> Leu280
R237Q	<u>Lys94</u> Lys96

4.3.4.2 Betweenness Centrality (BC)

In the previous section results demonstrated that variant presence has an effect on residue accessibility. In this section the effects of changes to residue accessibility on protein residue communication was investigated using *BC*. Residues showing the greatest communication (large average *BC*) are regarded as being the most important for protein structure and function, as proteins are dynamic in nature and

residues are always in constant communication [18, 277]. Changes within the communication/usage of residues during MD would allow for identification of compensatory mechanisms in variant proteins as a measure to maintain function and stability [18, 365].

4.3.4.2.1 CA-II

Results in Figure S11 indicate that Glu117 is the most important residue for communication in CA-II as evidenced by the high BC . Data also shows other important residues essential to communication within CA-II namely; His64, Ala65, Phe66, Asn67, Val142, Gly144 Val206 and Asn243. These residues include the proton shuttle, secondary aromatic cluster residues, active site water coordination and CO₂ binding pocket formation residues (Table 2.2), highlighting the importance of these active site residues to CA-II. As with L , ΔBC (WT – variant) was calculated to allow for the easier identifications of residues showing significant changes to average BC , and results are presented in Figure 4.19.

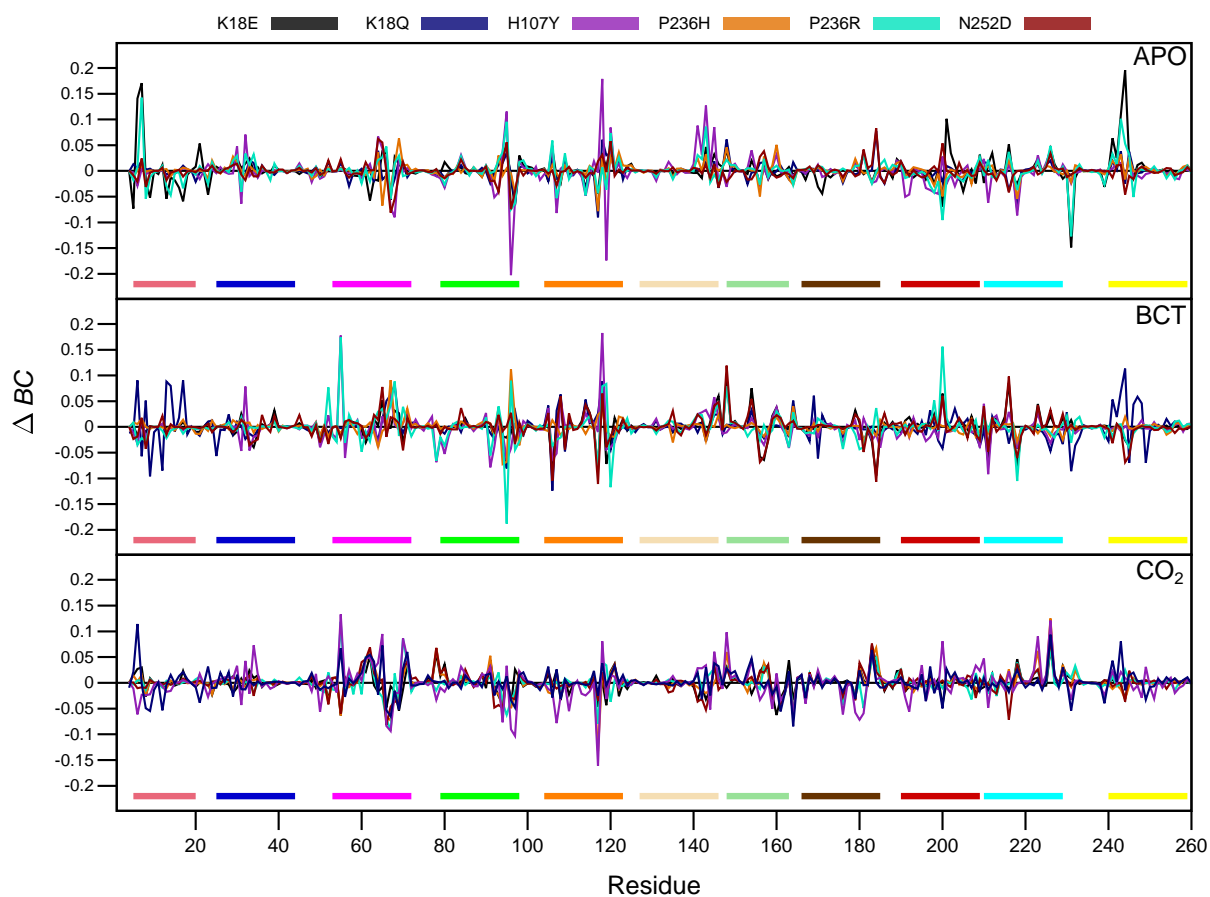


Figure 4.19. ΔBC (WT – variant) comparison of the CA-II protein systems. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6.

Positive ΔBC values indicate a decrease to specific residue usage within variant proteins during MD simulation. This would also indicate that the specific residue is more important within the WT protein compared to the variant. A negative ΔBC demonstrates an increase to specific variant residue communication. Results in Figure 4.19 show that depending on the substrate, the SNVs have different effects on the CA-II protein. Residues with an average ΔBC greater or less than two standard-deviations were calculated and results are presented in Table 4.5.

Table 4.5. CA-II residues showing significant changes to communication/usage during MD simulation. Important CA-II residues from Table 2.2 are highlighted in bold and underlined. SNV positions are underlined, italicised and highlighted in bold red.

Variant	Apo	BCT	CO ₂
Positive ΔBC (reduction in residue communication)			
K18E	<u>Trp5</u> Gly6 Pro200 Met240 <u>Asp242</u> <u>Asn243</u>	<u>His64</u> His107 Lys113 <u>Glu117</u> Ile145 Leu147 Lys153 <u>Thr199</u> Ile215 Val222	Asn61 Ala77 Leu90 Val162 Gly182 Leu183 Val217 <u>Phe225</u>
K18Q	Gly63 <u>His94</u> <u>Glu117</u> <u>His119</u> Leu147 Pro180 Leu183 Asp242 Trp244	<u>Trp5</u> Gly12 Pro13 <u>Trp16</u> <u>Phe66</u> <u>Asn67</u> His107 <u>Glu117</u> <u>Thr168</u> Asp242 <u>Asn243</u> Pro246	<u>Trp5</u> Ala54 Leu60 Asn61 <u>His64</u> Phe70 Gly182 <u>Phe225</u> Asp242
H107Y	Val31 Gly63 <u>His94</u> <u>Glu117</u> <u>His119</u> Leu140 <u>Val142</u> Gly144	Val31 Ala54 <u>Phe66</u> <u>Tyr107</u> <u>Glu117</u> Gly144 <u>Thr199</u>	Ile33 Ala54 <u>His64</u> Glu69 <u>Glu117</u> Leu147 <u>Thr199</u> Val222 <u>Phe225</u>
P236H	Gly63 Val68 <u>Gln92</u> Val121 <u>Val142</u> Leu147 Val159 Asp242	<u>Phe66</u> <u>Phe70</u> <u>Gln92</u> <u>Phe95</u> <u>Glu117</u> Leu163	Asn61 Ala77 Leu90 Leu147 Gly182 Leu183 Val222 <u>Phe225</u>
P236R	Gly6 <u>His94</u> Ser105 <u>His119</u> <u>Val142</u> Asp242 <u>Asn243</u>	Tyr51 Ala54 <u>Asn67</u> <u>Phe95</u> <u>Glu117</u> Leu118 Leu147 <u>Thr199</u>	Ala54 Leu60 Asn61 Glu69 Gly182 Leu183 <u>Phe225</u>
N252D	Gly63 <u>His64</u> <u>His94</u> <u>His119</u> Leu156 Pro180 Leu183 <u>Thr199</u> <u>Phe225</u>	<u>His64</u> His107 <u>Glu117</u> Leu147 Lys153 <u>Thr199</u> Ile215	Leu60 Asn61 <u>Phe70</u> Ala77 Gly182 <u>Phe225</u>
Negative ΔBC (residue communication increase)			
K18E	His4 <u>Trp16</u> <u>His96</u> <u>Thr199</u> Phe230	Ser105 Ala116 Leu118 Gly155 Leu156 Leu183 Val210 Val217	Ala54 Ala65 Val68 <u>Phe93</u> <u>Phe95</u> Ala116 Leu118 Val159 <u>Thr199</u>
K18Q	<u>Phe66</u> <u>His96</u> Lys113 Ala116 Ile145 Ile215 Val217	Gly8 Asn11 <u>His94</u> Ser105 Lys169 Phe230 Trp244 Gln248	<u>Tyr7</u> Gly8 Asn11 Ala65 <u>Phe66</u> Val68 <u>His96</u> Val160 Leu163 <u>Thr199</u> Phe230
H107Y	<u>Phe66</u> <u>Asn67</u> <u>Phe95</u> Glu106 Leu118 Val217	Thr55 Ala77 Leu90 <u>His94</u> Ser105 Val210	Ala65 <u>Phe66</u> <u>Phe93</u> <u>Phe95</u> <u>His96</u> Glu106 Ala116 Asp179
P236H	<u>His64</u> <u>Phe66</u> <u>His96</u> Ala116 Gly155 <u>Thr199</u> Val217	Gly63 <u>Phe93</u> <u>His94</u> <u>Trp97</u> Val134	Ala54 Ala65 <u>Phe66</u> <u>Phe93</u> <u>Phe95</u> Ala116 Leu163
P236R	<u>Tyr7</u> <u>Phe66</u> <u>Phe95</u> <u>His96</u> Glu106 <u>Thr199</u> Phe230 <u>Arg245</u>	Ala77 <u>His94</u> Ala116 <u>His119</u> Val217	Ala65 <u>Phe66</u> <u>His94</u> Ala116 Leu163 Phe175 Pro180
N252D	<u>Phe66</u> <u>Asn67</u> <u>Gln92</u> <u>Phe95</u> <u>His96</u> Ile215 <u>Asn243</u>	Ser105 Ala116 Gly155 Leu156 Leu183 Val217 <u>Asn243</u> Trp244	Tyr51 Ala54 Ala65 <u>Phe66</u> Val68 Ile91 <u>Gln92</u> <u>Phe93</u> <u>Val142</u> Val160 Leu163 Ile215

Decreases in the usage of Trp5 were observed for K18E_{apo}, K18Q_{BCT} and K18Q_{CO₂} proteins. As

Trp5 assists with stability in CA-II, the lower residue usage could explain the large conformational

sampling demonstrated in the RMSD results. This decrease in the usage of Trp5 in K18E_{apo} could be a direct result of the decrease to residue accessibility observed. With regards to H107Y, ΔBC results are in agreement with the contact map findings. For all three protein states H107Y shows a reduction in the usage of Glu117. This could be consequence of the decreases to residue interactions and hydrogen bonds between Tyr107 and Glu117. This further supports that H107Y could have an effect on the dissociation of Zn²⁺ from the CA-II active site through disruption of interactions with Glu117. H107Y_{CO₂} analysis highlights a possible compensatory measure to prevent Zn²⁺ dissociation from the active site. Results demonstrate an increase in the usage of primary ligand His96 which could compensate for the interaction losses with Glu117 and His119 to maintain Zn²⁺ within the active site. Increases to usage of residues within the secondary coordination sphere as also observed. In addition, all H107Y systems showed increases in the usage of Phe95 which assists with stability in motif 4. These increases might assist the variant with maintaining stability and indicate compensatory measures.

Reductions in the usage of Glu117 is also observed within the other variants. This further suggests that variants have an allosteric effect on CA-II that affects the active site of the protein. These decreases in Glu117 in-turn are associated with increases in the usage and communication of primary ligands His94, His96 and/or His119, and Asn243 which is a secondary Zn²⁺ ligand. This highlights at compensatory mechanisms occurring with the CA-II active to maintain active site integrity and Zn²⁺ coordination. This mechanism is observed for P236R_{apo} and N252D_{apo} whereby decreases in communication by His94 and His119, is complemented by an increase to the usage of His96. In addition, as P236R_{apo} showed decreases to Asn243 which directly interacts with His96. The increases to usage of His96 could also be to compensate for the interaction losses with Asn243.

Overall from the results in Figure 4.19 data highlights variant effects occur away from the SNV location indicating that the CA-II variants have an allosteric effect on the protein structure. SNV effects are observed at active site residues, and at the aromatic cluster residues. Towards precision

medicine related studies of CA-II and associated variants, treatment strategies would have to be designed to assist with active site stability, and/or the rescue of primary and secondary aromatic cluster residues.

4.3.4.2.2 CA-IV

The average *BC* of the WT and variant CA-IV proteins is presented in Figure S12. Data indicates that residues Glu138, Val165 and Ala167 are the most important residues for communication in CA-IV. The Glu138 high usage is expected as it aligns with Glu117 in CA-II, and is well conserved throughout all CAs [19, 66]. With minimal research performed on CA-IV, we hypothesise that Glu138 could also have a effect of Zn^{2+} affinity and dissociation from the CA-IV active site as observed in CA-II. Val165 and Ala167 are members of motif 6. Val165 could assist with formation of the CA-IV CO_2 binding pocket (Table 2.2) explaining its high usage. Unusually however is Ala167 which aligns with Gly144 in CA-II. It should be noted that both Val165 and Ala167 are located in close proximity to the region showing the greatest reduction to residue flexibility and increases to accessibility. Assessment of the results evidences that compared to CA-II, CA-IV has fewer residues generally showing high usage. Interestingly Val90, Gln113, His115, Glu127, Gly128, Trp235 and Asn269 also show high average *BC*. With the exception to Gly128 these residues are listed in Table 2.2 supporting the accuracy of the residue function assignment, and illustrating their importance to CA-IV. Although Val90 aligns with Phe66 and is not aromatic, this result suggests that Val90 could have an adaptive function in CA-IV.

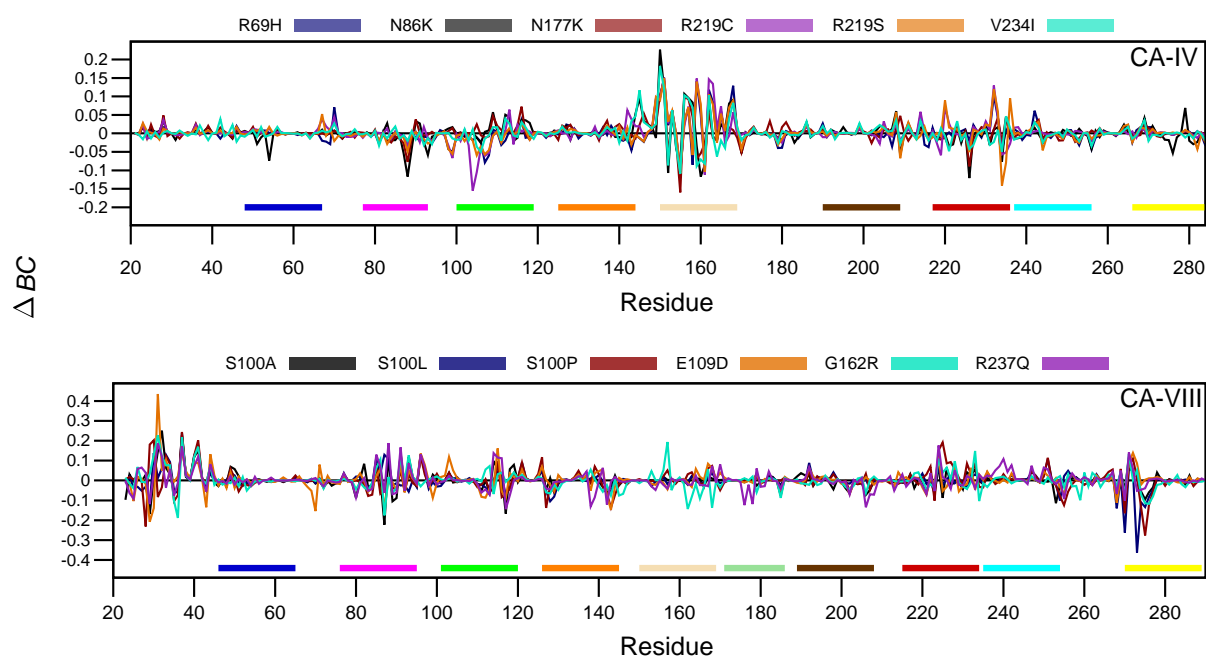


Figure 4.20. ΔBC (WT – variant) of the CA-IV and CA-VIII respectively during MD simulation. Colour coded bars at the bottom of the plot represent the similar coloured motifs in Figure 2.6.

Results in Figure 4.20 present the ΔBC of the CA-IV proteins. Initial inspection of results shows that motif 1 and motif 6 (red and wheat regions) show the greatest changes to ΔBC . Numerous residues within the variant proteins show ΔBC values greater or less than 0, indicating that variant presence has an effect on residue communication throughout the protein. Residues showing ΔBC changes greater or less than two standard deviations are presented in Table 4.6. Figure 4.20 data reveals that there are no changes to residue usage occurring at the positions of the SNVs but changes are located elsewhere within the protein, and centred around the motif 6 residues (residues 150–169). R69H, N177K and R219S show decreases to the usage of Val165 that may assist with the CO₂ binding pocket, with N177K showing an additional decrease to the usage of Leu116. R219S highlights an unusual result, although this variant is associated with poor catalytic activity there are no decreases to the usage of the assigned catalytic residues.

Benign variants N86K and N177K both show increased usage in the proton shuttle residue (His88) and Thr226 that may assist with the CO₂ binding pocket formation. Increases to the communication of these residues could indicate a compensatory measure by these variants in maintaining CA-IV

stability. In addition, both SNVs show increases in the usage of Asn152, Glu155, Glu160 which could also highlight a compensatory mechanism for benign variants. This result complements the RMSD distributions (see Figure 4.2) that hinted at potentially similar mechanisms of N86K and N177K. These residues also show usage increases in V234I. Increases to Asn152 and Glu160 are evident solely to the benign SNVs hinting at potential importance in benign mechanisms. The R219S increases to weighted interactions for Trp235 (Figure 4.16) are associated with increases to residue communication as evidenced in Table 4.6. Overall from the table, pathogenic variants show communication decreases to greatest number of residues.

Table 4.6. CA-IV residues showing significant changes to communication/usage during MD simulation. Important CA-IV residues from Table 2.2 are highlighted in bold and underlined.

Variant	Positive ΔBC (Residue usage decrease)
R69H	Phe70 Thr149 Ser150 Arg151 Val153 Gln157 Pro159 Glu162 Ile163 <u>Val165</u> Ala167 Phe168 Lys232 Ile242
N86K	Lys145 Ser150 Arg151 Ala156 Gln157 Asp158 Glu162 Ala167 Phe168 Arg279
N177K	<u>Leu116</u> Glu146 Ser150 Arg151 Val153 Ala156 Gln157 Glu162 <u>Val165</u>
R219C	Thr149 Ser150 Arg151 Gln157 Pro159 Glu160 Glu162 Ile163 <u>Val165</u> Tyr220 Lys232
R219S	Thr149 Ser150 Arg151 Gln157 Pro159 Glu160 Glu162 Ile163 Ala167 Phe168 Tyr220 Lys232 Thr236
V234I	Lys145 Ser150 Arg151 Val153 Ala156 Gln157 Asp158 Glu162 Phe168
	Negative ΔBC (Residue usage increase)
R69H	Ala107 Pro108 Glu155 Asp158 Asp161
N86K	Ile54 <u>His88</u> Asn152 Glu155 Glu160 Asp161 <u>Thr226</u> Val234
N177K	<u>His88</u> Asn152 Glu155 Pro159 Glu160 Val170 <u>Thr226</u>
R219C	Ser99 Gly104 Leu105 Glu155 Asp161
R219S	Glu155 Asp161 Pro209 Val234 <u>Trp235</u>
V234I	Asn152 Glu155 Pro159 Glu160 Asp161 Ala164

4.3.4.2.3 CA-VIII

The average BC of CA-VIII was also calculated to understand the effect of L on the BC of key residues with the protein, and results are presented in Figure S12. Data shows that residues; Glu139, Ile165, Ala167, Val231, Trp233 and Asn273 are the most important residues for communication in CA-VIII evidenced by the higher average BC values. Alignment of these residues with CA-II (Figure S1)

demonstrates that these amino acids map onto residues; Glu117, Val142, Gly144, Val206, Trp208 and Asn243 of CA-II (Table 2.2), and essential for communication [66]. Residues in Figure S12 that are non-aromatic and associated with high average BC values may be indicative of an acatalytic adaptation in the maintenance of structure, stability and function of CA-VIII.

Results in Figure S12 indicate that residue Asn273 in S100L has the highest average BC compared to the other proteins. High average BC is also observed for the E109D residues Trp29 and Gly30 compared to the WT and other variants. These increases to BC could be indicative of variant compensatory measures in order to main structural stability and binding site stability through Trp29 and Gly30 respectively.

The ΔBC of the CA-VIII WT and variant residue is presented in Figure 4.20, whereas the residues showing ΔBC value greater of less than two standard deviations are presented in Table 4.7. From the results in Figure 4.20, data shows that multiple residues have ΔBC values greater or less than 0 indicating that SNV presence has an effect on residue communication within CA-VIII. These effects are more pronounced than the N-terminus and C-terminus of the protein, which contain essential ITPR1 binding site residues.

Results in Table 4.7 highlight that apart from G162R there are no communication changes occurring at the SNV positions. This further hints at an allosteric SNV mechanism of action, since positions 100 (S100A, S100P and S100L), 109 (E109D) and 237 (R237Q) do not show direct changes to residue communication. Of the variants, G162 contains the most binding site residues showing decreases to communication, whereas R237Q in general has the greatest number of residues showing decreases to residue communication.

Table 4.7. CA-VIII residues showing significant changes to communication/usage during MD simulation. SNV positions are underlined, italicised and highlighted in bold red. CA-VIII binding site residues are underlined and highlighted in bold blue. Important CA-VIII residues from Table 2.2 are underlined and highlighted in bold. PPI and important structural residues are underlined and highlighted in bold green. Adapted from Sanyanga and Tastan Bishop 2020 [269].

Variant	Positive ΔBC (Residue usage decrease)
S100A	<u>Tyr31</u> <u>Glu32</u> <u>Trp37</u> <u>Val40</u> <u>Phe41</u> Thr88 Val91 <u>Leu93</u> Lys96 Gly271
S100L	<u>Gly30</u> <u>Glu33</u> <u>Trp37</u> <u>Val40</u> <u>Phe41</u> Asp85 His87 Thr88 Lys96 Arg251
S100P	<u>Trp29</u> <u>Gly30</u> <u>Glu33</u> <u>Trp37</u> <u>Val40</u> <u>Phe41</u> Lys96 Gly126 <u>Ile224</u> Pro225
E109D	<u>Tyr31</u> <u>Trp37</u> Ala44 Lys96 Pro103 Val115 <u>Ile224</u> Asp272
G162R	<u>Gly30</u> <u>Tyr31</u> <u>Glu32</u> <u>Trp37</u> <u>Val40</u> <u>Phe41</u> Gly86 Val91 Leu93 Val157 <u>Ile224</u> <u>Trp233</u> Gly271
R237Q	<u>Tyr31</u> <u>Glu32</u> <u>Trp37</u> <u>Phe41</u> <u>Asp85</u> Thr88 Val91 Lys96 <u>Glu114</u> <u>Val115</u> <u>Arg116</u> <u>Ile224</u> Leu240 Gly271
Negative ΔBC (Residue usage increase)	
S100A	Glu23 <u>Trp29</u> His87 <u>Ile89</u> <u>Phe117</u> Leu253 <u>Arg275</u>
S100L	<u>Trp29</u> Glu128 Arg254 Gly268 Leu270 Asn273 Phe274 <u>Arg275</u>
S100P	<u>Glu28</u> Ser127 <u>Ile143</u> <u>Thr255</u> Leu270 <u>Phe274</u> <u>Arg275</u> Pro276
E109D	Glu25 <u>Trp29</u> <u>Gly30</u> Asp43 Leu70 His87 Ile143 Gly268
G162R	<u>Val35</u> <u>Glu36</u> His87 <u>Arg162</u> Leu168 Ile234 <u>Arg275</u> Pro276
R237Q	<u>Trp29</u> <u>Phe117</u> Met138 Leu142 His176 Gly178 Leu206 Leu270

Assessment of the ΔBC increases and associated residues may highlight the possible variant mechanisms. In all variants, at least two aromatic cluster residues show decreases to communication. Trp37 is the only amino acid showing a consistent decrease in all variants, possibly suggesting its importance to CA-VIII. Reductions to the usage of Trp29, Trp37 and Phe41 at the N-terminal could explain the development of CAMRQ3 and reported poor CA-VIII stability associated with the variants [131]. The lack of residue correlation within the DCC results could also be as a result of usage changes to these residues during MD. Additionally, from Table 4.7 data, it observed that excluding G162R, all other variants how decreases in the usage of Lys96. Overall the decline in the usage of binding site residues in the CA-VIII variants could have an effect on association with ITPR1, therefore resulting in Ca^{2+} dysregulation and homeostasis disruptions. Pathogenic variants S100P and G162R, present increases to usage with the stability assisting residue Trp29. This could highlight

a potential compensatory mechanism by the variants to maintain protein structure and function.

Benign variants S100L and E109D however show usage reductions in the fewest residues potentially essential to stability (bold black and green, and underlined) compared to the pathogenic variants. This finding could hint at a lower extent of CA-VIII destabilisation within the benign variants compared to the pathogenic ones.

4.3.5 Variant Effects on Protein Shuttle Behaviour

DRN analysis showed evidence of variant allosteric effects on the CA-II protein structure. In this section variant associated effects on the behaviour of the protein shuttle residue His64 were investigated. This was performed to investigate as to whether changes to this residue had occurred that might not have been detected using traditional MD approaches or DRN analysis. It should be noted that the pK_a of surrounding residues also governs the behaviour of His64 [366–368]. In addition, changes occurring to the proton shuttle may not change residue flexibility, accessibility or usage but may still have an effect on residue function.

His64 was observed to rotate between the “in” and “out” conformations during MD simulations which was expected as this has been observed in previous studies and literature [50–53]. Examples of these conformations are presented in Figure 4.21A with respect to the WT_{apo} protein (green).

In previous literature in 2007 by Silverman and McKenna [50] it was suggested that the distance between the proton shuttle His64 and Zn^{2+} has an effect on the rate of CO_2 hydration in CA-II. Thus for the WT and variant proteins, the average distances between Zn^{2+} and the His64 imidazole ring was measured and results are presented in Table 4.8.

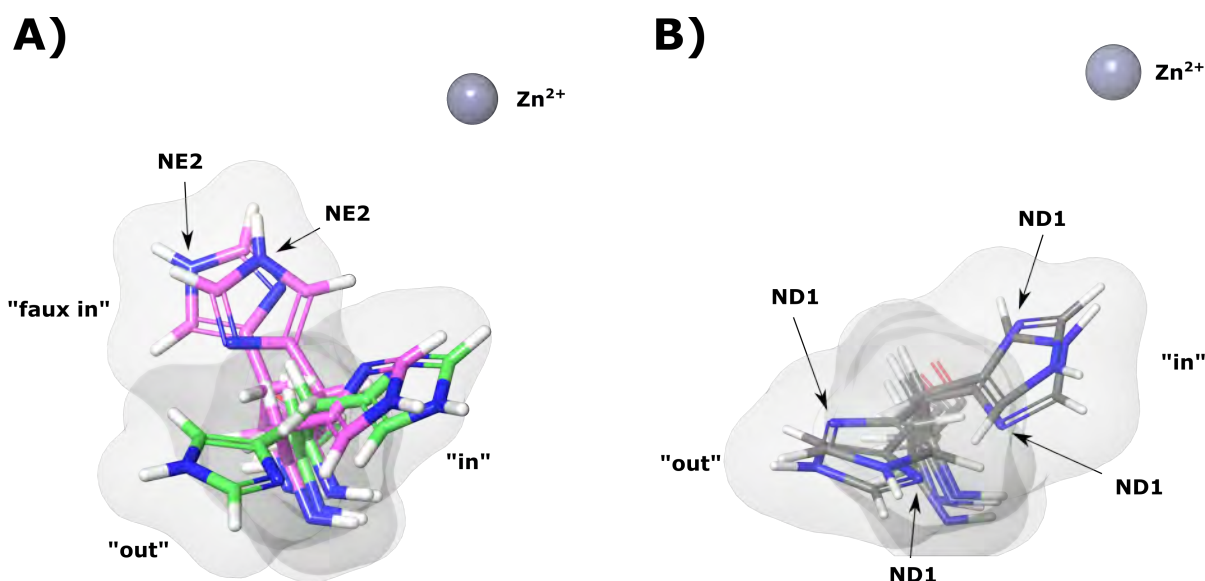


Figure 4.21. Illustration of the His64 proton shuttle “in”, “out” and “faux in” conformations during MD in the CA-II_{apo} protein. A: WT and K18E proteins. Green and magenta colours represent WT and K18E respectively. B: N252D variant. Adapted from Sanyanga *et al.* 2019 [66].

Analysis of the data in Table 4.8 shows an unexpected finding whereby His64 does not occupy an “out” conformation in K18E_{apo}. Results showed occupation of a His64 orientation that has not yet been reported in previous literature [19, 27, 29, 48, 50], and has been observed for the first time within this study. This orientation has been termed as the “faux in” conformation, and is illustrated in Figure 4.21A. To ensure that this novel conformation was not a random occurrence during MD, the prevalence of the conformation over the 200 ns MD simulation was determined, and results showed that this conformation existed for a fraction of 0.972 of all MD frames. This indicates that the proton shuttle remained in this position for the majority of the MD simulation. Furthermore, the “faux in” conformation brings the imidazole ring of His64 closest to the Zn²⁺ of all conformations. Since His64 did not adopt the “out” conformation during MD, this could suggest implications for proton shuttling during catalysis. Firstly this “faux in” conformation may not be able to stabilise the active site water network like the “in” conformation [369]. Secondly, without an “out” conformation protons would have to travel a greater distance to exit the active site during CO₂ hydration suggesting that a large water network would be needed to shuttle protons [369–371]. This would negatively impact on

proton shuttling and rate of catalysis.

In addition to K18E_{apo}, K18Q_{BCT} did not occupy an “out” conformation during MD simulation. The proton shuttle maintained the “in” conformation throughout the entire MD simulation. P236_{apo} results showed existence of this novel “faux in” conformation but however unlike K18E_{apo}, His64 in P236R was also able to adopt the “out” conformation indicating that the proton shuttle was capable of occupying all three conformations. Analysis of orientation prevalence revealed an unusual finding. Results showed that the “in”, “out” and “faux in” conformations existed for a fraction of 0.049, 0.123 and 0.827 of all MD simulation frames, indicating that the “faux in” conformation was adopted for the majority of the simulation, and was preferred to the “in” and “out” conformation.

Table 4.8. Distance of His64 imidazole group from Zn²⁺ for the “in” and “out” conformations within CA-II. All distances are measured from the His64 imidazole ring centroid to the Zn²⁺. Faux refers to other conformations observed excluding traditional “in” and “out” occupied by His64. Adapted from Sanyanga *et al.* 2019 [66].

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.

Emergence analysis of the conformations revealed that the “faux in” was incapable of directly rotating to the out conformation without transitioning through the “in” orientation first. Analysis of K18E and P236R in the presence of substrate further suggests that the presence of BCT and CO₂ have an effect on SNV behaviour, as the “faux in” conformation was not observed in their presence. Assessment of these results with *BC* indicates the the appearance of the “faux in” conformation is

associated with the reduction of the usage of Asn243, whereas in K18Q_{BCT} a the reduction in Asn243 usage is associated with the lack of an “out” conformation. Asn243 constantly interacts with His64 as shown in Figure 4.22.

Further analysis of Figure 4.21 demonstrates that the imidazole ring of His64 is capable of rotation at the CB-CG (beta carbon atom and gamma carbon atom) bond. This is evidenced by the change of the position and orientation of either the ND1 or NE2 atom in relation to Zn²⁺ (Figure 4.21A,B). This rotation is not limited to the variant proteins only, and includes the WT. Previous studies into the proton shuttling by His64 suggested that protons are shuttled to and from the active site through imidazole ring tautomerisation and/or His64 “in” and “out” rotation [50–54]. Discovery of the His64 rotation along the CB-CG bond could suggest that imidazole ring rotation could also assist with proton shuttling. Additional research is however necessary to validate this.

Results also confirm the previous findings by [50], whereby the side chain of His64 of the “in” conformation forms no interactions with other active site residues, however analysis of the “out” conformations revealed the formation of a π -stack with the indole ring of Trp5 (sandwich π -stack) [50]. Additionally, formation of a π -stack with Phe230 was also observed for His64 in the “out” conformation of the WT. Visual inspection of His64 in the “faux in” conformation suggests that the proton shuttle could form π -stacks or additional interactions with either Trp5 or Tyr7. To investigate this contact maps of the proton shuttle were calculated and results are presented in Figure 4.22.

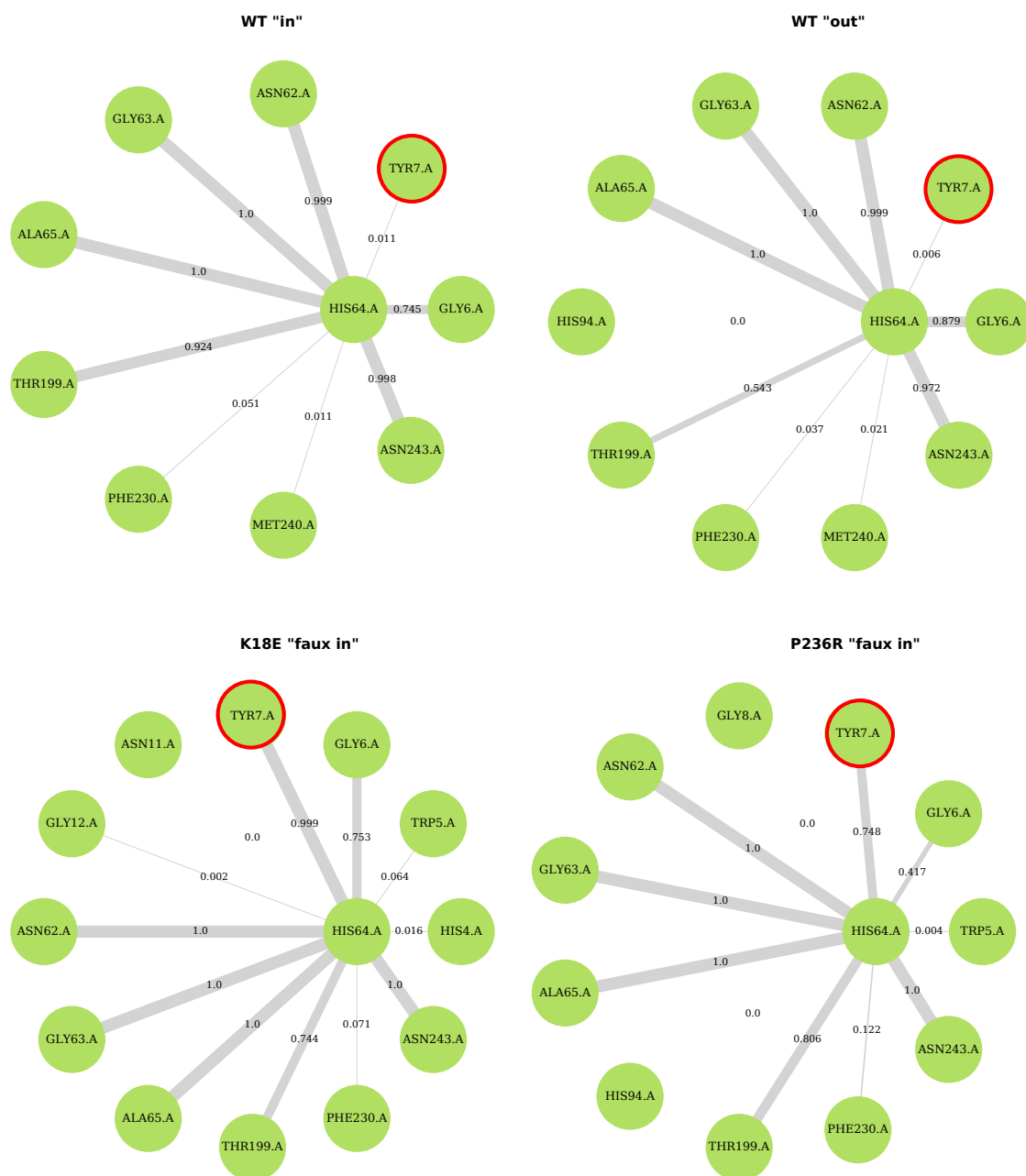


Figure 4.22. Contact maps of His64 in the WT and respective CA-II variant apo proteins. Tyr7 has been circled in red.

Results show an increase to weighted interactions between Tyr7 and His64 within the K18E_{apo} and P236R_{apo} proteins for the “faux in” conformation. Interestingly, interactions with the aromatic cluster residue are not observed for the WT proteins when occupying either the “in” or “out” conformations. This shows that Tyr7 may have an effect on the occurrence of the “faux in” conformation. The P236R_{apo} His64 (“faux in”) and Tyr5 residues also formed a T-shaped π -stack,

however this interaction was not observed within the K18E representative structure.

Investigations into the proton shuttle residue His88 of CA-IV revealed that none of the variants adopted the “faux in” conformation. This finding highlights minimal variant impact, suggesting that the poor catalysis associated with R219S is not due to distortions of the proton shuttle, and the cause lies elsewhere.

4.3.6 Considerations of CA-II, CA-IV and CA-VIII SNVs Towards Drug Discovery

The previous chapters have detailed information on the mechanisms of actions of the various SNVs within the CA-II, CA-IV and CA-VIII proteins. Variant effects on CA-II are mainly located within the active site, whereas effects on CA-IV are located away from it. Despite the different SNV effect locations, CA-II and CA-IV PCA results demonstrated that WT and benign variants are associated with lower free energies. Drug discovery with regards to the pathogenic catalytic isoforms should be focused on allosteric compounds or molecular chaperones to stabilise the respective protein structures. Pathogenic CA-IV SNVs have been partially stabilised by CA inhibitor dorzolamide. Since research demonstrated RP17 was attributed to a toxic gain of function [119] in CA-IV, this molecular basis could have rendered the inhibitor effective. Given the non-specific nature of CA inhibitors, the effects of dorzolamide could significantly be reduced *in vivo* and could inhibit other CAs non-specifically [372]. Thus CA-IV isoform specific inhibitors are required. As pathogenic CA-II SNV data evidences potentially poor catalytic function, inhibitors may not be effective in the treatment of osteopetrosis with RTA and cerebral calcification. CA-II would require compounds capable of stabilising the Zn²⁺ within the active site, and DRN analysis employed to validate compound effects. With regards to CA-VIII, as the mechanism of interaction with ITPR1 is not well understood it is difficult to ascertain potential considerations for drug discovery into CAMRQ3. Data however suggests that CA-VIII may need to be destabilised by compounds to achieve therapeutic action as evidenced in RMSD, PCA and Rg results.

4.4 CONCLUSION

MD simulations and analysis are essential in the investigation of SNV effects on protein structure analysis. This approach is however limited with regards to investigations of mutant effects therefore DRN analysis was also incorporated. Previously generated parameters were capable of maintaining the Zn^{2+} cofactor within the active site. Throughout the CA-II, CA-IV and CA-VIII proteins, variant presence has a subtle effect on the protein structure. Interestingly with regards to CA-VIII, the Rg results were in agreement with DCC and PCA, and demonstrated that the WT protein was less stable than the variants. This was in contradiction to *in silico* predictions using I-Mutant and MUpro that predicted stability reductions. This demonstrates that although *in silico* prediction methods assist with filtering of large datasets they are not always accurate with regards to variant effects on stability. Correlations between stability predictions and clinical significance of variants for these tools also require investigations. PCA analysis revealed differing behaviour of conformational sampling between catalytic and acatalytic CA isoforms. Catalytic WT and benign variants of the CA-II and CA-IV proteins are associated with lower conformational sampling and energy within 3D space, whereas the WT and benign variants in CA-VIII are associated with higher free energy and greater conformational sampling.

Greater insights into variant mechanism of action was provided through DRN as opposed to traditional MD approaches. This suggests DRN analysis is necessary for the study of mutation effects as it also investigates changes within the protein network as opposed to traditional MD analysis that factors in global structural changes and residue flexibility.

5

Conclusion

The phenotypes associated with the CA group of enzymes are as a result of poor protein function or folding. To date most of the research into CAs has focused on the inhibition of these proteins to achieve therapeutic effect in individuals, leaving a large research gap with regards to the functional rescue of these proteins and the identification of activator and stabilising compounds. The main objective of this research was to characterise the effects of pathogenic and benign validated SNVs on the structure and function of CA-II, CA-IV and CA-VIII. The understanding of disease pathogenesis associated with the SNVs would set the foundation for precision medicine related studies within the CA group of enzymes in two main ways. Firstly, numerous CA variants have been identified within online databases without any associated phenotype. This research would allow rapid characterisation of CA variants to determine which are more likely to cause disease. Secondly, understanding

the mechanism of action associated with the SNVs would greatly assist with drug discovery into conditions associated with CA deficiencies, and offer individuals suffering from osteopetrosis with RTA and cerebral calcification, RP17 and CAMRQ3 with better treatment options and a higher quality of life.

CHAPTER 2

Within this chapter the sequence and structures of the CA-II, CA-IV and CA-VIII proteins were characterised using sequence and motif analysis, and homology modelling. Prior to these analyses, the STRING server was used to identify potential PPIs associated with each protein. The catalytic CAs CA-II and CA-IV do not require a membrane to function, however CA-VIII functions through interactions with ITPR1. The exact associating residues are not known. STRING analysis revealed no new potential PPI interactions. Unusually however was the discovery of associations between CA-IV and the WTAP protein suggesting that CA-IV could have an anti-cancer role. This suggested that the long term use of CA inhibitors in slowing RP17 progression may not be ideal should CA-IV ever become an anti-cancer target. SiteMap and CPORT were then utilised to identify potential binding site residues in CA-VIII. Binding site residue investigations into CA-VIII discovered 38 potential amino acids that could associate with ITPR1.

As CA-IV and CA-VIII are less well investigated, sequence analysis using CA-II as the reference sequence was performed to identify conserved residues. The three proteins showed conservation of multiple important amino acids to CA-II, and motif analysis of the entire α -CA family identified a maximum of 11 valid motifs. CA-II contained all 11 motifs, whereas CA-IV and CA-VIII comprised of 9 and 10 conserved motifs respectively. SNV identification identified six variants for each of the CA proteins associated with a phenotype annotation in within the HUMA and Ensembl databases. These included, CA-II: K18E, K18Q, H107Y, P236H, P236R and N252D; CA-IV: R69H, N86K, N177K,

R219C, R219S and V234I; CA-VIII: S100A, S100L, S100P, E109D, G162R and R237Q. Many more variants were available, however the only those with associated with a phenotype annotation were selected to observe similarities and differences between the known phenotypes, and to study trends. Each of the variants was then modelled. CA-VIII_{S100P} was found to destroy the β -sheet secondary structure it was located on. CA-VIII_{R237Q} is located on a potential ITPR1 binding site residue. Across all three CAs none of the SNVs occur at residues important for function with the exception to CA-IV_{N86K}. The CA-II variants K18E, K18Q, H107Y and N252D; CA-IV variants N86K, R219C, R219S and V234I; and CA-VIII variants E109D, G162R and R237Q are however located on conserved motifs.

CHAPTER 3

The AMBER ff14sb FF was extended to add support for the Zn^{2+} cofactor that is essential to catalysis in CAs. QM calculations showed that within CA-II, the Zn^{2+} cofactor has a charge less than +1, and the respective coordinating histidine nitrogen atoms have smaller negative charges compared to their formal charges. These parameters would be necessary to conduct accurate MD simulations.

CHAPTER 4

This chapter focused on the use of MD simulations and DRN to investigate the effects of SNVs on the structure and function of CA-II, CA-IV and CA-VIII. RMSD and Rg analysis revealed subtle variant effects on the global structure of the CA proteins.

CA-II

PCA analysis revealed that the presence of either substrate BCT or CO_2 is associated with differences to conformational sampling. In the presence of BCT, protein structures present highest free energies and the greatest 3D conformational sampling. Variant effects occurred away from the SNV location

and were centred around the active site of the enzyme and residues of the aromatic clusters necessary for the maintenance of stability. *BC* analysis revealed that Glu117 is the most important residue for communication within CA-II. Variants were associated with decreases to the usage of Glu117 in the presence of BCT highlighting at potential effects on Zn^{2+} dissociation from the active site of the enzyme. These effects were greatest in H107Y which demonstrated residue communication reduction for the apo, BCT and CO_2 bound proteins due to the loss of hydrogen bonds and weighted interactions between Tyr107 and Glu117. Additional hydrogen bonds were also lost between Tyr107 and the coordinating His119. In all variants a reduction in the usage of a Zn^{2+} primary coordination ligand was associated with increases to the usage of other primary and secondary coordination illustrating variant compensatory mechanisms. Within CA-II, a novel proton shuttle conformation name the "faux in" conformation was observed for the K18E_{apo} and P236R_{apo} proteins. This conformation was occupied to the greatest extent in during MD and was associated with increases to weighted contacts between Tyr7 and His64. Residue correlation analysis indicated that the presence of BCT results in anti-correlated residue movement in the WT protein, whereas CO_2 results in increases to residue correlation.

CA-IV

The analysis of the pathogenic and benign variants of CA-IV using PCA revealed that the WT and benign variants are associated with lower conformational sampling in 3D space and lower free energy compared to the pathogenic variants, indicating greater stability. Greatest changes to the flexibility and accessibility of CA-IV residues was observed for motif 6 which includes the residues 150–169. These residues were also associated with increases to accessibility. Drastic *BC* changes were also associated with this motif suggesting physiological importance of these residues. *BC* identified Glu138, Val165 and Ala167 as the most important residues for communication in CA-IV. N86K and N177K both showed increases to the usage of the proton shuttle residue His88. Generally results suggested that

R69H and R219C could have a similar mechanism of pathogenesis, whereas N86K and N177K could employ similar mechanisms to tolerate variant presence.

CA-VIII

Variant presence in CA-VIII was associated with increases to rigidity of the proteins. The differences between pathogenic and benign variants were observed through PCA analysis. Pathogenic SNVs were associated with lower conformational sampling with structures clustering into a single well of low free energy. The WT and benign variants showed greater conformational sampling and were associated with higher free energy. WT residues showed greater anti-correlated movement whereas the variant DCC results demonstrated a lack of correlation to greater correlation to residue movement. *L* revealed possible compensatory mechanisms employed by benign variants. S100L and E109D revealed both increases and decreases to N-terminal residue accessibility which could assist with the maintenance of binding site integrity. Within all variants, Trp37 which could possibly assist with enzyme stability was associated with a reduction to usage.

FUTURE STUDIES

This research covered associations between the CA-II, CA-IV and CA-VIII enzymes, and the mechanism of action of SNVs. Future studies includes the alanine scanning of the respective proteins to identify which residues are essential to stability especially for the CA-IV and CA-VIII proteins. Additionally since SNVs may have an effect on the pK_a of the protein, *in-silico* pK_a calculations could also be performed to observe these changes to the protein. The pK_a of key residues could have an effect on the association of the respective CA proteins with membranes. Expanding on this, constant pH MD could also be investigated to observe changes to protonation states as a result of pH to further determine variant effect. Lastly to investigate effects on proton shuttling by the SNVs, QM simulations could also be performed to simulate the proton transfer process within the catalytic CAs.

References

1. Orphanet. *About Rare Diseases* [Online; accessed 06 March 2020]. 2012. https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN.
2. Song, P., Gao, J., Inagaki, Y., Kokudo, N. & Tang, W. Rare diseases, orphan drugs, and their regulation in Asia: Current status and future perspectives. *Intractable & rare diseases research* 1, 3–9 (2012).
3. Wakap, S. N. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics* 28, 165–173 (2020).
4. Gong, S. & Jin, S. Current progress in the management of rare diseases and orphan drugs in China. *Intractable & rare diseases research* 1, 45–52 (2012).
5. De Vrueth, R., Baekelandt, E. & De Haan, J. Priority medicines for Europe and the world:” a public health approach to innovation.”. *WHO Background Paper* 6 (2013).
6. Van Weely, S. & Leufkens, H. Priority medicines for Europe and the world-A public health approach to innovation. *Orphan diseases. Geneva (Switzerland): World Health Organization*, 95–100 (2004).
7. Stoller, J. K. The challenge of rare diseases. *Chest* 153, 1309–1314 (2018).
8. Mueller, T., Jerrentrup, A., Bauer, M. J., Fritsch, H. W. & Schaefer, J. R. Characteristics of patients contacting a center for undiagnosed and rare diseases. *Orphanet journal of rare diseases* 11, 81 (2016).

9. WHO. *Priority diseases and reasons for inclusion* [Online; accessed 06 March 2020]. 2020. https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf.
10. Stoller, J. K., Smith, P., Yang, P., Spray, J., *et al.* Physical and social impact of alpha 1 antitrypsin deficiency: results of a survey. *Cleveland Clinic journal of medicine* **61**, 461–467 (1994).
11. Boat, T. F., Field, M. J., *et al.* *Rare diseases and orphan products: Accelerating research and development* (National Academies Press, 2011).
12. Jackson, M., Marks, L., May, G. H. & Wilson, J. B. The genetic basis of disease. *Essays in biochemistry* **62**, 643–723 (2018).
13. Mensink, K. A. & Hand, J. L. Autosomal recessive inheritance: An updated review. *Pediatric dermatology* **23**, 404–409 (2006).
14. Kassam, S., Meyer, P., Corfield, A., Mikuz, G. & Sergi, C. Single nucleotide polymorphisms (SNPs): history, biotechnological outlook and practical applications. *Current Pharmacogenomics* **3**, 237–245 (2005).
15. Bhattacharya, R., Rose, P. W., Burley, S. K. & Prlić, A. Impact of genetic variation on three dimensional structure and function of proteins. *PloS one* **12**, e0171355 (2017).
16. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *science* **322**, 881–888 (2008).
17. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics* **33**, 228 (2003).
18. Brown, D. K., Amamuddy, O. S. & Tastan Bishop, Ö. Structure-based analysis of single nucleotide variants in the renin-angiotensinogen complex. *Global Heart* **12**, 121–132 (2017).
19. Lindskog, S. Structure and mechanism of carbonic anhydrase. *Pharmacology & Therapeutics* **74**, 1–20 (1997).

20. Silverman, D. N. & Lindskog, S. The catalytic mechanism of carbonic anhydrase: implications of a rate-limiting protolysis of water. *Accounts of Chemical Research* **21**, 30–36 (1988).
21. Seifter, J. L. & Chang, H.-Y. Disorders of acid-base balance: new perspectives. *Kidney Diseases* **2**, 170–186 (2016).
22. Orešković, D. & Klarica, M. The formation of cerebrospinal fluid: nearly a hundred years of interpretations and misinterpretations. *Brain Research Reviews* **64**, 241–262 (2010).
23. Blair, H. C., Teitelbaum, S. L., Ghiselli, R. & Gluck, S. Osteoclastic bone resorption by a polarized vacuolar proton pump. *Science* **245**, 855–857 (1989).
24. Chegwiddden, W. R., Dodgson, S. J. & Spencer, I. M. in *The Carbonic Anhydrases* 343–363 (Springer, 2000).
25. Henry, R. P. & Swenson, E. R. The distribution and physiological significance of carbonic anhydrase in vertebrate gas exchange organs. *Respiration physiology* **121**, 1–12 (2000).
26. Chaput, C. D. *et al.* A proteomic study of protein variation between osteopenic and age-matched control bone tissue. *Experimental Biology and Medicine* **237**, 491–498 (2012).
27. McKenna, R. & Frost, S. C. in *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications* 3–5 (Springer, 2014).
28. Angeli, A. *et al.* Inhibition of bacterial α -, β - and γ -class carbonic anhydrases with selenazoles incorporating benzenesulfonamide moieties. *Journal of enzyme inhibition and medicinal chemistry* **34**, 244–249 (2019).
29. Supuran, C. T. Structure and function of carbonic anhydrases. *Biochemical Journal* **473**, 2023–2032 (2016).
30. Tripp, B. C., Smith, K. & Ferry, J. G. Carbonic anhydrase: new insights for an ancient enzyme. *Journal of Biological Chemistry* **276**, 48615–48618 (2001).

31. Di Fiore, A., Alterio, V., Monti, S. M., De Simone, G. & D'Ambrosio, K. Thermostable carbonic anhydrases in biotechnological applications. *International Journal of Molecular Sciences* **16**, 15456–15480 (2015).
32. Soto, A. R. *et al.* Identification and preliminary characterization of two cDNAs encoding unique carbonic anhydrases from the marine alga *Emiliana huxleyi*. *Applied Environmental Microbiology* **72**, 5500–5511 (2006).
33. Del Prete, S. *et al.* Discovery of a new family of carbonic anhydrases in the malaria pathogen *Plasmodium falciparum*—The η -carbonic anhydrases. *Bioorganic & medicinal chemistry letters* **24**, 4389–4396 (2014).
34. Lane, T. W. *et al.* Biochemistry: A cadmium enzyme from a marine diatom. *Nature* **435**, 42 (2005).
35. Hewett-Emmett, D. & Tashian, R. E. Functional diversity, conservation, and convergence in the evolution of the α -, β -, and γ -carbonic anhydrase gene families. *Molecular Phylogenetics and Evolution* **5**, 50–77 (1996).
36. Xu, Y., Feng, L., Jeffrey, P. D., Shi, Y. & Morel, F. M. Structure and metal exchange in the cadmium carbonic anhydrase of marine diatoms. *Nature* **452**, 56 (2008).
37. Ferreira-Martins, D. *et al.* A cytosolic carbonic anhydrase molecular switch occurs in the gills of metamorphic sea lamprey. *Scientific Reports* **6**, 33954 (2016).
38. Frost, S. C. & McKenna, R. *Carbonic anhydrase: mechanism, regulation, links to disease, and industrial applications* (Springer Science & Business Media, 2013).
39. Supuran, C. T. Carbonic anhydrases-an overview. *Current pharmaceutical design* **14**, 603–614 (2008).

40. Consortium, U. UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **45**, D158–D169 (2017).
41. Boriack-Sjodin, P. A., Heck, R. W., Laipis, P. J., Silverman, D. N. & Christianson, D. W. Structure determination of murine mitochondrial carbonic anhydrase V at 2.45-Å resolution: implications for catalytic proton transfer and inhibitor design. *Proceedings of the National Academy of Sciences* **92**, 10949–10953 (1995).
42. Shah, G. N. *et al.* Targeted mutagenesis of mitochondrial carbonic anhydrases VA and VB implicates both enzymes in ammonia detoxification and glucose metabolism. *Proceedings of the National Academy of Sciences* **110**, 7423–7428 (2013).
43. Schneider, H.-P. *et al.* GPI-anchored carbonic anhydrase IV displays both intra- and extracellular activity in cRNA-injected oocytes and in mouse neurons. *Proceedings of the National Academy of Sciences* **110**, 1494–1499 (2013).
44. Hilvo, M. *et al.* Characterization of CA XV, a new GPI-anchored form of carbonic anhydrase. *Biochemical journal* **392**, 83–92 (2005).
45. Murakami, H. & Sly, W. Purification and characterization of human salivary carbonic anhydrase. *Journal of Biological Chemistry* **262**, 1382–1388 (1987).
46. Occhipinti, R. & Boron, W. F. Role of Carbonic Anhydrases and Inhibitors in Acid–Base Physiology: Insights from Mathematical Modeling. *International journal of molecular sciences* **20**, 3841 (2019).
47. Karhumaa, P. *et al.* Expression of the transmembrane carbonic anhydrases, CA IX and CA XII, in the human male excurrent ducts. *Molecular human reproduction* **7**, 611–616 (2001).

48. Mikulski, R. L. & Silverman, D. N. Proton transfer in catalysis and the role of proton shuttles in carbonic anhydrase. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1804**, 422–426 (2011).
49. Berg, J. M. & Tymoczko, J. *Biochemistry* 5th ed. Chap. 9, Making a Fast Reaction Faster: Carbonic Anhydrase (WH Freeman and Company, New York, 2002).
50. Silverman, D. N. & McKenna, R. Solvent-mediated proton transfer in catalysis by carbonic anhydrase. *Accounts of Chemical Research* **40**, 669–675 (2007).
51. 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* **28**, 7913–7918 (1989).
52. Nair, S. K. & Christianson, D. W. Unexpected pH-dependent conformation of His-64, the proton shuttle of carbonic anhydrase II. *Journal of The American Chemical Society* **113**, 9455–9458 (1991).
53. 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. *Archives of Biochemistry and Biophysics* **539**, 31–37 (2013).
54. Shimahara, H. *et al.* Tautomerism of histidine 64 associated with proton transfer in catalysis of carbonic anhydrase. *Journal of Biological Chemistry* **282**, 9646–9656 (2007).
55. Krishnan, D. *et al.* Carbonic anhydrase II binds to and increases the activity of the epithelial sodium-proton exchanger, NHE3. *American Journal of Physiology-Renal Physiology* **309**, F383–F392 (2015).

56. Li, X., Alvarez, B., Casey, J. R., Reithmeier, R. A. & Fliegel, L. Carbonic anhydrase II binds to and enhances activity of the Na⁺/H⁺ exchanger. *Journal of Biological Chemistry* **277**, 36085–36091 (2002).
57. Vince, J. W. & Reithmeier, R. A. Carbonic anhydrase II binds to the carboxyl terminus of human band 3, the erythrocyte Cl⁻/HCO₃⁻ exchanger. *Journal of Biological Chemistry* **273**, 28430–28437 (1998).
58. Al-Samir, S. *et al.* Activity and distribution of intracellular carbonic anhydrase II and their effects on the transport activity of anion exchanger AE1/SLC4A1. *The Journal of physiology* **591**, 4963–4982 (2013).
59. Sterling, D., Reithmeier, R. A. & Casey, J. R. A transport metabolon Functional interaction of carbonic anhydrase II and chloride/bicarbonate exchangers. *Journal of Biological Chemistry* **276**, 47886–47894 (2001).
60. Pushkin, A. *et al.* Molecular mechanism of kNBC1–carbonic anhydrase II interaction in proximal tubule cells. *The Journal of physiology* **559**, 55–65 (2004).
61. Becker, H. M. & Deitmer, J. W. Carbonic anhydrase II increases the activity of the human electrogenic Na⁺/cotransporter. *Journal of Biological Chemistry* **282**, 13508–13521 (2007).
62. Merz Jr, K. M. Carbon dioxide binding to human carbonic anhydrase II. *Journal of The American Chemical Society* **113**, 406–411 (1991).
63. Liang, J.-Y. & Lipscomb, W. N. Binding of substrate CO₂ to the active site of human carbonic anhydrase II: a molecular dynamics study. *Proceedings of The National Academy of Sciences* **87**, 3675–3679 (1990).
64. Domsic, J. F. *et al.* Entrapment of carbon dioxide in the active site of carbonic anhydrase II. *Journal of Biological Chemistry* **283**, 30766–30771 (2008).

65. Alexander, R. S., Nair, S. K. & Christianson, D. W. Engineering the hydrophobic pocket of carbonic anhydrase II. *Biochemistry* 30, 11064–11072 (1991).
66. Sanyanga, T. A., Nizami, B. & Tastan Bishop, Ö. Mechanism of Action of Non-Synonymous Single Nucleotide Variations Associated with α -Carbonic Anhydrase II Deficiency. *Molecules* 24 (2019).
67. Eriksson, A. E., Jones, T. A. & Liljas, A. Refined structure of human carbonic anhydrase II at 2.0 Å resolution. *Proteins: Structure, Function, and Bioinformatics* 4, 274–282 (1988).
68. 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. *Journal of Biological Chemistry* 272, 20364–20372 (1997).
69. Laitala, T. & Väänänen, H. Inhibition of bone resorption in vitro by antisense RNA and DNA molecules targeted against carbonic anhydrase II or two subunits of vacuolar H (+)-ATPase. *The Journal of clinical investigation* 93, 2311–2318 (1994).
70. 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. *Human Mutation* 24, 272–272 (2004).
71. Stark, Z., Savarirayan, R. & Orphanet, O. Osteopetrosis. *Orphanet Journal of Rare Diseases* 4 (2009).
72. OMIM. *Online Mendelian Inheritance in Man* [MIM Number: 259730] [Last edited: 08/03/2016]. <https://www.omim.org/entry/259730>.

73. Silver, I., Murrills, R. & Etherington, D. Microelectrode studies on the acid microenvironment beneath adherent macrophages and osteoclasts. *Experimental cell research* 175, 266–276 (1988).
74. Blair, H. C. How the osteoclast degrades bone. *Bioessays* 20, 837–846 (1998).
75. Tolar, J., Teitelbaum, S. L. & Orchard, P. J. Mechanisms of Disease: Osteopetrosis. *The New England Journal of Medicine* 351, 2839–2849 (2004).
76. NORD. *National Organization for Rare Disorders: Rare Disease Database* [accessed: 01/12/2019]. 2019. <https://rarediseases.org/rare-diseases/osteopetrosis/>.
77. Villa, A., Guerrini, M. M., Cassani, B., Pangrazio, A. & Sobacchi, C. Infantile malignant, autosomal recessive osteopetrosis: the rich and the poor. *Calcified tissue international* 84, 1 (2009).
78. Del Fattore, A., Cappariello, A. & Teti, A. Genetics, pathogenesis and complications of osteopetrosis. *Bone* 42, 19–29 (2008).
79. Sly, W. S., Hewett-Emmett, D., Whyte, M. P., Yu, Y.-S. & Tashian, R. E. Carbonic anhydrase II deficiency identified as the primary defect in the autosomal recessive syndrome of osteopetrosis with renal tubular acidosis and cerebral calcification. *Proceedings of the National Academy of Sciences* 80, 2752–2756 (1983).
80. Mazzolari, E. *et al.* A single-center experience in 20 patients with infantile malignant osteopetrosis. *American journal of hematology* 84, 473–479 (2009).
81. 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. *Journal of Biological Chemistry* 268, 4775–4779 (1993).

82. Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, D1062–D1067 (2017).
83. 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. *American Journal of Human Genetics* **49**, 1082 (1991).
84. Roth, D. E., Venta, P. J., Tashian, R. E. & Sly, W. S. Molecular basis of human carbonic anhydrase II deficiency. *Proceedings of The National Academy of Sciences* **89**, 1804–1808 (1992).
85. Soda, H. *et al.* A point mutation in exon 3 (His 107→ Tyr) in two unrelated Japanese patients with carbonic anhydrase II deficiency with central nervous system involvement. *Human genetics* **97**, 435–437 (1996).
86. LöNnerholm, G., Wistrand, P. J. & Bárány, E. Carbonic anhydrase isoenzymes in the rat kidney. Effects of chronic acetazolamide treatment. *Acta physiologica scandinavica* **126**, 51–60 (1986).
87. Brown, D., Kumpulainen, T., Roth, J. & Orci, L. Immunohistochemical localization of carbonic anhydrase in postnatal and adult rat kidney. *American Journal of Physiology-Renal Physiology* **245**, F110–F118 (1983).
88. Pereira, P., Miranda, D., Oliveira, E. & Simões e Silva, A. Molecular pathophysiology of renal tubular acidosis. *Current genomics* **10**, 51–59 (2009).
89. Laing, C. M., Toye, A. M., Capasso, G. & Unwin, R. J. Renal tubular acidosis: developments in our understanding of the molecular basis. *The international journal of biochemistry & cell biology* **37**, 1151–1161 (2005).

90. Strisciuglio, P., Sartorio, R., Pecoraro, C., Lotito, F. & Sly, W. Variable clinical presentation of carbonic anhydrase deficiency: evidence for heterogeneity? *European journal of pediatrics* **149**, 337–340 (1990).
91. Aramaki, S. *et al.* Carbonic anhydrase II deficiency in three unrelated Japanese patients. *Journal of inherited metabolic disease* **16**, 982–990 (1993).
92. Ramos, E., Oliveira, J., Sobrido, M. & Coppola, G. in *GeneReviews [Internet]* [updated: 24/08/2017] (University of Washington, 2004). <https://www.ncbi.nlm.nih.gov/books/NBK1421/>.
93. Bosley, T. M. *et al.* The neurology of carbonic anhydrase type II deficiency syndrome. *Brain* **134**, 3502–3515 (2011).
94. Scozzafava, A. & Supuran, C. T. in *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications* 349–359 (Springer, 2014).
95. Swenson, E. R. in *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications* 361–386 (Springer, 2014).
96. Supuran, C. T., Scozzafava, A. & Casini, A. Carbonic anhydrase inhibitors. *Medicinal Research Reviews* **23**, 146–189 (2003).
97. Supuran, C. T. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nature Reviews Drug Discovery* **7**, 168 (2008).
98. Puscas, I., Coltau, M., Baican, M., Pasca, R. & Domuta, G. The inhibitory effect of diuretics on carbonic anhydrases. *Research Communications In Molecular Pathology and Pharmacology* **105**, 213–236 (1999).

99. Shinohara, C., Yamashita, K., Matsuo, T., Kitamura, S. & Kawano, F. Effects of carbonic anhydrase inhibitor acetazolamide (AZ) on osteoclasts and bone structure. *Journal of Hard Tissue Biology* **16**, 115–123 (2007).
100. Lehenkari, P., Hentunen, T. A., Laitala-Leinonen, T., Tuukkanen, J. & Väänänen, H. K. Carbonic anhydrase II plays a major role in osteoclast differentiation and bone resorption by effecting the steady state intracellular pH and Ca²⁺. *Experimental cell research* **242**, 128–137 (1998).
101. Hurst, T. K., Wang, D., Thompson, R. B. & Fierke, C. A. Carbonic anhydrase II-based metal ion sensing: Advances and new perspectives. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1804**, 393–403 (2010).
102. Lesburg, C. A. & Christianson, D. W. X-ray crystallographic studies of engineered hydrogen bond networks in a protein-zinc binding site. *Journal of the American Chemical Society* **117**, 6838–6844 (1995).
103. Berg, J., Tymoczko, J. & Stryer, L. in *Biochemistry* 5th ed. (WH Freeman, 2002).
104. Sly, W. S. & Hu, P. Y. Human carbonic anhydrases and carbonic anhydrase deficiencies. *Annual review of biochemistry* **64**, 375–401 (1995).
105. Baird, T. T., Waheed, A., Okuyama, T., Sly, W. S. & Fierke, C. A. Catalysis and inhibition of human carbonic anhydrase IV. *Biochemistry* **36**, 2669–2678 (1997).
106. Sterling, D., Alvarez, B. V. & Casey, J. R. The extracellular component of a transport metabolon extracellular loop 4 of the human AE1 Cl⁻/HCO⁻ exchanger binds carbonic anhydrase IV. *Journal of Biological Chemistry* **277**, 25239–25246 (2002).
107. Yang, Z. *et al.* Mutant carbonic anhydrase 4 impairs pH regulation and causes retinal photoreceptor degeneration. *Human molecular genetics* **14**, 255–265 (2005).

108. Rebello, G. *et al.* Apoptosis-inducing signal sequence mutation in carbonic anhydrase IV identified in patients with the RP17 form of retinitis pigmentosa. *Proceedings of the National Academy of Sciences* **101**, 6617–6622 (2004).
109. Dewar, A. & Reading, H. The biochemical aspects of retinitis pigmentosa. *International Journal of Biochemistry* **6**, 615–641 (1975).
110. Veltel, S., Gasper, R., Eisenacher, E. & Wittinghofer, A. The retinitis pigmentosa 2 gene product is a GTPase-activating protein for Arf-like 3. *Nature structural & molecular biology* **15**, 373 (2008).
111. OMIM. *Online Mendelian Inheritance in Man* [MIM Number: 268000] [Last edited: 10/03/2019]. <https://www.omim.org/entry/268000>.
112. Hartong, D. T., Berson, E. L. & Dryja, T. P. Retinitis pigmentosa. *The Lancet* **368**, 1795–1809 (2006).
113. Novak-Lauš, K., Kukulj, S., Zoric-Geber, M. & Bastaic, O. Primary tapetoretinal dystrophies as the cause of blindness and impaired vision in the republic of Croatia. *Acta Clin Croat* **41**, 23–27 (2002).
114. Grøndahl, J. Estimation of prognosis and prevalence of retinitis pigmentosa and Usher syndrome in Norway. *Clinical genetics* **31**, 255–264 (1987).
115. Rozet, J. *et al.* Dominant X linked retinitis pigmentosa is frequently accounted for by truncating mutations in exon ORF15 of the RPGR gene. *Journal of medical genetics* **39**, 284–285 (2002).
116. Souied, E. *et al.* Severe manifestations in carrier females in X linked retinitis pigmentosa. *Journal of medical genetics* **34**, 793–797 (1997).

117. Bunker, C. H., Berson, E. L., Bromley, W. C., Hayes, R. P. & Roderick, T. H. Prevalence of retinitis pigmentosa in Maine. *American journal of ophthalmology* 97, 357–365 (1984).
118. OMIM. *Online Mendelian Inheritance in Man* [MIM Number: 600852] [Last edited: 11/30/2018]. <https://www.omim.org/entry/600852>.
119. Datta, R., Waheed, A., Bonapace, G., Shah, G. N. & Sly, W. S. Pathogenesis of retinitis pigmentosa associated with apoptosis-inducing mutations in carbonic anhydrase IV. *Proceedings of the National Academy of Sciences* 106, 3437–3442 (2009).
120. Ogilvie, J. M. *et al.* Carbonic anhydrase XIV deficiency produces a functional defect in the retinal light response. *Proceedings of the National Academy of Sciences* 104, 8514–8519 (2007).
121. Zhao, Y. *et al.* Vitamins and Mineral Supplements for Retinitis Pigmentosa. *Journal of ophthalmology* 2019 (2019).
122. Picaud, S. S. *et al.* Crystal structure of human carbonic anhydrase-related protein VIII reveals the basis for catalytic silencing. *Proteins: Structure, Function, and Bioinformatics* 76, 507–511 (2009).
123. Huang, M.-S. *et al.* Roles of carbonic anhydrase 8 in neuronal cells and zebrafish. *Biochimica et Biophysica Acta (BBA)-General Subjects* 1840, 2829–2842 (2014).
124. Kato, K. Sequence of a novel carbonic anhydrase-related polypeptide and its exclusive presence in purkinje cells. *FEBS letters* 271, 137–140 (1990).
125. Aspatwar, A., Tolvanen, M. E., Ortutay, C. & Parkkila, S. in *Carbonic Anhydrase: Mechanism, Regulation, Links to Disease, and Industrial Applications* 135–156 (Springer, 2014).
126. Taniuchi, K. *et al.* Developmental expression of carbonic anhydrase-related proteins VIII, X, and XI in the human brain. *Neuroscience* 112, 93–99 (2002).

127. Hirota, J., Hideaki, A., Hamada, K. & Mikoshiba, K. Carbonic anhydrase-related protein is a novel binding protein for inositol 1, 4, 5-trisphosphate receptor type 1. *Biochemical Journal* 372, 435–441 (2003).
128. Hur, E.-M. *et al.* Junctional membrane inositol 1, 4, 5-trisphosphate receptor complex coordinates sensitization of the silent EGF-induced Ca²⁺ signaling. *The Journal of cell biology* 169, 657–667 (2005).
129. Lamont, M. G. & Weber, J. T. The role of calcium in synaptic plasticity and motor learning in the cerebellar cortex. *Neuroscience & Biobehavioral Reviews* 36, 1153–1162 (2012).
130. Gruol, D., Manto, M. & Haines, D. Ca²⁺ Signaling in Cerebellar Purkinje Neurons. *The Cerebellum* 11, 605–608 (2012).
131. Türkmen, S. *et al.* CA8 mutations cause a novel syndrome characterized by ataxia and mild mental retardation with predisposition to quadrupedal gait. *PLoS genetics* 5, e1000487 (2009).
132. Bosanac, I. *et al.* Structure of the inositol 1, 4, 5-trisphosphate receptor binding core in complex with its ligand. *Nature* 420, 696 (2002).
133. Bosanac, I. *et al.* Crystal structure of the ligand binding suppressor domain of type 1 inositol 1, 4, 5-trisphosphate receptor. *Molecular cell* 17, 193–203 (2005).
134. Sienaert, I. *et al.* Localization and function of a calmodulin-apocalmodulin-binding domain in the N-terminal part of the type 1 inositol 1, 4, 5-trisphosphate receptor. *Biochemical Journal* 365, 269 (2002).
135. Yamada, M. *et al.* The calmodulin-binding domain in the mouse type 1 inositol 1, 4, 5-trisphosphate receptor. *Biochemical Journal* 308, 83–88 (1995).
136. Hsiao, C.-T. *et al.* Mutational analysis of ITPR1 in a Taiwanese cohort with cerebellar ataxias. *PloS one* 12, e0187503 (2017).

137. Parolin Schnekenberg, R. *et al.* De novo point mutations in patients diagnosed with ataxic cerebral palsy. *Brain* **138**, 1817–1832 (2015).
138. Van Dijk, T. *et al.* A de novo missense mutation in the inositol 1, 4, 5-triphosphate receptor type 1 gene causing severe pontine and cerebellar hypoplasia: Expanding the phenotype of ITPR1-related spinocerebellar ataxia's. *American journal of medical genetics Part A* **173**, 207–212 (2017).
139. Chen, X. *et al.* Deranged calcium signaling and neurodegeneration in spinocerebellar ataxia type 3. *Journal of Neuroscience* **28**, 12713–12724 (2008).
140. Liu, J. *et al.* Deranged calcium signaling and neurodegeneration in spinocerebellar ataxia type 2. *Journal of Neuroscience* **29**, 9148–9162 (2009).
141. OMIM. *Online Mendelian Inheritance in Man* [MIM Number: 613227] [Last edited: 06/11/2013]. <https://www.omim.org/entry/613227>.
142. Hirasawa, M. *et al.* Carbonic anhydrase related protein 8 mutation results in aberrant synaptic morphology and excitatory synaptic function in the cerebellum. *Molecular and Cellular Neuroscience* **35**, 161–170 (2007).
143. Mori, S. *et al.* Nucleotide variations in genes encoding carbonic anhydrase 8 and 10 associated with femoral bone mineral density in Japanese female with osteoporosis. *Journal of bone and mineral metabolism* **27**, 213–216 (2009).
144. Kaya, N. *et al.* Phenotypical spectrum of cerebellar ataxia associated with a novel mutation in the CA8 gene, encoding carbonic anhydrase (CA) VIII. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **156**, 826–834 (2011).

145. Aspatwar, A., EE Tolvanen, M., Ortutay, C. & Parkkila, S. Carbonic anhydrase related protein VIII and its role in neurodegeneration and cancer. *Current pharmaceutical design* **16**, 3264–3276 (2010).
146. Boone, C. D., Habibzadegan, A., Gill, S. & McKenna, R. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Biomolecules* **3**, 553–562 (2013).
147. Maupin, C. M. *et al.* Effect of active-site mutation at Asn67 on the proton transfer mechanism of human carbonic anhydrase II. *Biochemistry* **48**, 7996–8005 (2009).
148. Turkoglu, S. *et al.* Mutation of active site residues Asn67 to Ile, Gln92 to Val and Leu204 to Ser in human carbonic anhydrase II: Influences on the catalytic activity and affinity for inhibitors. *Bioorganic & medicinal chemistry* **20**, 2208–2213 (2012).
149. Rehm, B. Bioinformatic tools for DNA/protein sequence analysis, functional assignment of genes and protein classification. *Applied microbiology and biotechnology* **57**, 579–592 (2001).
150. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).
151. Liang, S., Zhang, C., Liu, S. & Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic acids research* **34**, 3698–3707 (2006).
152. Kalinina, O. V., Mironov, A. A., Gelfand, M. S. & Rakhmaninova, A. B. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Science* **13**, 443–456 (2004).
153. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology* **257**, 342–358 (1996).

154. Duan, Y., Reddy, B. V. & Kaznessis, Y. N. Physicochemical and residue conservation calculations to improve the ranking of protein–protein docking solutions. *Protein science* **14**, 316–328 (2005).
155. Daugeilaite, J., O’Driscoll, A. & Sleator, R. D. An overview of multiple sequence alignments and cloud computing in bioinformatics. *ISRN Biomathematics 2013* (2013).
156. Kemena, C. & Notredame, C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**, 2455–2465 (2009).
157. Barton, C., Flouri, T., Iliopoulos, C. S. & Pissis, S. P. Global and local sequence alignment with a bounded number of gaps. *Theoretical Computer Science* **582**, 1–16 (2015).
158. Huang, M., Shah, N. D. & Yao, L. Evaluating global and local sequence alignment methods for comparing patient medical records. *BMC Medical Informatics and Decision Making* **19**, 263 (2019).
159. Zhou, Z.-m. & Chen, Z.-w. Dynamic programming for protein sequence alignment. *Int. J. Bio-Sci. Bio-Technol.* **5**, 141–150 (2013).
160. Altschul, S. F. Amino acid substitution matrices from an information theoretic perspective. *Journal of molecular biology* **219**, 555–565 (1991).
161. Pearson, W. R. Selecting the right similarity-scoring matrix. *Current protocols in bioinformatics* **43**, 3–5 (2013).
162. Waterman, M. S., Smith, T. F. & Beyer, W. A. Some biological sequence metrics. *Advances in Mathematics* **20**, 367–387 (1976).
163. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**, 443–453 (1970).

164. Wallace, I. M., Blackshields, G. & Higgins, D. G. Multiple sequence alignments. *Current opinion in structural biology* **15**, 261–266 (2005).
165. Feng, D.-F. & Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of molecular evolution* **25**, 351–360 (1987).
166. Smith, T. F., Waterman, M. S., *et al.* Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–197 (1981).
167. Wilbur, W.J. & Lipman, D.J. Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences* **80**, 726–730 (1983).
168. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797 (2004).
169. Blackshields, G., Sievers, F., Shi, W., Wilm, A. & Higgins, D. G. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology* **5**, 21 (2010).
170. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**, 406–425 (1987).
171. Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* **38**, 1409–1438 (1958).
172. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011).
173. Sievers, F. & Higgins, D. G. in *Multiple Sequence Alignment Methods* 105–116 (Springer, 2014).

174. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673–4680 (1994).
175. Katoh, K., Misawa, K., Kuma, K.-i. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).
176. Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. Fast and accurate automatic structure prediction with HHpred. *Proteins: Structure, Function, and Bioinformatics* **77**, 128–132 (2009).
177. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960 (2004).
178. Eddy, S. R. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**, 755–763 (1998).
179. Mohamed, S. A. E. H., Elloumi, M. & Thompson, J. D. in *Pattern Recognition-Analysis and Applications* (IntechOpen, 2016).
180. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic acids research* **47**, D427–D432 (2018).
181. Sigrist, C. J. *et al.* New and continuing developments at PROSITE. *Nucleic acids research* **41**, D344–D347 (2012).
182. Sleator, R. D. & Walsh, P. An overview of in silico protein function prediction. *Archives of microbiology* **192**, 151–155 (2010).
183. Hashim, F. A., Mabrouk, M. S. & Al-Atabany, W. Review of Different Sequence Motif Finding Algorithms. *Avicenna journal of medical biotechnology* **11**, 130 (2019).

184. Zhang, X. in *Encyclopedia of Systems Biology* (eds Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 1721–1722 (Springer New York, New York, NY, 2013). ISBN: 978-1-4419-9863-7. https://doi.org/10.1007/978-1-4419-9863-7_439.
185. Xia, X. Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica* 2012 (2012).
186. Gorodkin, J., Heyer, L. J., Brunak, S. & Storomo, G. Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics* 13, 583–586 (1997).
187. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* 18, 6097–6100 (1990).
188. Bailey, T. L., Elkan, C., *et al.* Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994).
189. Bailey, T. L. & Elkan, C. *The value of prior knowledge in discovering motifs with MEME.* in *Ismb* 3 (1995), 21–29.
190. Bailey, T. L. *et al.* MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Research* 37, W202–W208 (2009).
191. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics (Oxford, England)* 14, 48–54 (1998).
192. Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics* 36, 307–340 (2003).
193. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, gkw1000 (2016).
194. Pearson, W. R. An introduction to sequence similarity (“homology”) searching. *Current protocols in bioinformatics* 42, 3–1 (2013).

195. Kufareva, I. & Abagyan, R. in *Homology Modeling* 231–257 (Springer, 2011).
196. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Molecular systems biology* 3 (2007).
197. Qi, Y. & Noble, W. S. in *Handbook of Statistical Bioinformatics* 427–459 (Springer, 2011).
198. Shoemaker, B. A. & Panchenko, A. R. Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLOS Computational Biology* 3, 1–7 (2007).
199. Von Mering, C. *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research* 33, D433–D437 (2005).
200. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences* 23, 324–328 (1998).
201. Huynen, M., Snel, B., Lathe, W. & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome research* 10, 1204–1210 (2000).
202. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* 47, D607–D613. <https://string-db.org/> (2018).
203. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517 (1996).
204. Consortium, 1. G. P. *et al.* A global reference for human genetic variation. *Nature* 526, 68 (2015).
205. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, 531210 (2019).

206. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*, 563866 (2019).
207. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic acids research* **45**, D840–D845 (2016).
208. OMIM. *Online Mendelian Inheritance in Man* [Online; accessed 15 May 2019]. 2018. <https://www.omim.org/entry/114815>.
209. Brown, D. K. & Tastan Bishop, Ö. HUMA: A platform for the analysis of genetic variation in humans. *Human Mutation* **39**, 40–51 (2018).
210. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (2017).
211. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutation* **34**, 57–65 (2013).
212. 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* **22**, 2729–2734 (2006).
213. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248 (2010).
214. 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* **7**, e46688 (2012).
215. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* **33**, W306–W310 (2005).

216. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* 62, 1125–1132 (2006).
217. Türkmen, S. *et al.* Cerebellar hypoplasia and quadrupedal locomotion in humans as a recessive trait mapping to chromosome 17p. *Journal of medical genetics* 43, 461–464 (2006).
218. Arendall, W. B. *et al.* A test of enhancing model accuracy in high-throughput crystallography. *Journal of structural and functional genomics* 6, 1–11 (2005).
219. Fiser, A. in *Computational biology* 73–94 (Springer, 2010).
220. Read, R. J. *et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* 19, 1395–1412 (2011).
221. Buchan, D. W. & Jones, D. T. The PSIPRED protein analysis workbench: 20 years on. *Nucleic acids research* 47, W402–W407 (2019).
222. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* 234, 779–815 (1993).
223. Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *Journal of Molecular Biology* 320, 597–608 (2002).
224. Jacobson, M. P. *et al.* A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics* 55, 351–367 (2004).
225. Papadrakakis, M. & Ghionis, P. Conjugate gradient algorithms in nonlinear structural analysis problems. *Computer methods in applied mechanics and engineering* 59, 11–27 (1986).

226. Nayeem, A., Sitkoff, D. & Krystek Jr, S. A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. *Protein Science* 15, 808–824 (2006).
227. Wiederstein, M. & Sippl, M. J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 35, W407–W410 (2007).
228. Eisenberg, D., Lüthy, R. & Bowie, J. U. [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods In Enzymology* 277, 396–404 (1997).
229. Shen, M.-y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein science* 15, 2507–2524 (2006).
230. Zhang, L. & Skolnick, J. What should the Z-score of native protein structures be? *Protein science* 7, 1201–1207 (1998).
231. Ramachandran, G. N. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7, 95–99 (1963).
232. Lovell, S. C. *et al.* Structure validation by $C\alpha$ geometry: ϕ, ψ and $C\beta$ deviation. *Proteins: Structure, Function, and Bioinformatics* 50, 437–450 (2003).
233. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* 356, 83 (1992).
234. Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chemical biology & drug design* 69, 146–148 (2007).
235. Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling* 49, 377–389 (2009).
236. De Vries, S. J. & Bonvin, A. M. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS one* 6, e17695 (2011).

237. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry* **28**, 849–857 (1985).
238. Chen, H. & Zhou, H.-X. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins: Structure, Function, and Bioinformatics* **61**, 21–35 (2005).
239. Kufareva, I., Budagyan, L., Raush, E., Totrov, M. & Abagyan, R. PIER: protein interface recognition for structural proteomics. *Proteins: Structure, Function, and Bioinformatics* **67**, 400–417 (2007).
240. Neuvirth, H., Raz, R. & Schreiber, G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *Journal of molecular biology* **338**, 181–199 (2004).
241. Porollo, A. & Meller, J. Prediction-based fingerprints of protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics* **66**, 630–645 (2007).
242. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of The National Academy of Sciences* **89**, 10915–10919 (1992).
243. Behnke, C. A. *et al.* Atomic resolution studies of carbonic anhydrase II. *Acta Crystallographica Section D: Biological Crystallography* **66**, 616–627 (2010).
244. Henderson, L. E., Henriksson, D. & Nyman, P. O. Primary structure of human carbonic anhydrase C. *Journal of Biological Chemistry* **251**, 5457–5463 (1976).
245. Case, D. *et al.* Amber 2017, 1–950 (2017).
246. Harding, M. M. Small revisions to predicted distances around metal sites in proteins. *Acta Crystallographica Section D: Biological Crystallography* **62**, 678–682 (2006).

247. Peters, M. B. *et al.* Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *Journal of Chemical Theory and Computation* **6**, 2935–2947 (2010).
248. Ross, C., Knox, C. & Tastan Bishop, Ö. Interacting motif networks located in hotspots associated with RNA release are conserved in Enterovirus capsids. *Febs Letters* **591**, 1687–1701 (2017).
249. Nyamai, D. W. & Tastan Bishop, Ö. Aminoacyl tRNA synthetases as malarial drug targets: a comparative bioinformatics study. *Malaria Journal* **18**, 34 (2019).
250. DeLano, W. L. *et al.* Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography* **40**, 82–92 (2002).
251. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **9**, 90 (2007).
252. Team, I. Inkscape: A vector drawing tool. <http://www.inkscape.org> (2004).
253. Schrödinger. Schrödinger Release 2018-3: Maestro, Schrödinger, LLC, New York, NY, 2017. *Received: February* (2018).
254. Brown, D. K. & Tastan Bishop, Ö. Role of structural bioinformatics in drug discovery by computational SNP analysis: analyzing variation at the protein level. *Global Heart* **12**, 151–161 (2017).
255. Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein- protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731–1737 (2003).
256. Van Zundert, G. *et al.* The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology* **428**, 720–725 (2016).

257. Olsson, M. H., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent treatment of internal and surface residues in empirical pka predictions. *Journal of Chemical Theory and Computation* 7, 525–537 (2011).
258. Chen, Y. *et al.* WTAP facilitates progression of hepatocellular carcinoma via m6A-HuR-dependent epigenetic silencing of ETS1. *Molecular cancer* 18, 127 (2019).
259. Zhang, J. *et al.* Carbonic anhydrase IV inhibits colon cancer development by inhibiting the Wnt signalling pathway through targeting the WTAP–WT1–TBL1 axis. *Gut* 65, 1482–1493 (2016).
260. Yu, H.-L. *et al.* WTAP is a prognostic marker of high-grade serous ovarian cancer and regulates the progression of ovarian cancer cells. *Oncotargets and therapy* 12, 6191 (2019).
261. Chen, L. & Wang, X. Relationship between the genetic expression of WTAP and bladder cancer and patient prognosis. *Oncology letters* 16, 6966–6970 (2018).
262. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic acids research* 47, D351–D360 (2018).
263. Colussi, G. *et al.* Chronic hypercalcaemia from inactivating mutations of vitamin D 24-hydroxylase (CYP24A1): implications for mineral metabolism changes in chronic renal failure. *Nephrology Dialysis Transplantation* 29, 636–643 (2013).
264. Kandel, S. E. & Lampe, J. N. Role of protein–protein interactions in cytochrome P450-mediated drug metabolism and toxicity. *Chemical research in toxicology* 27, 1474–1486 (2014).
265. Alsahli, S., Alrifai, M. T., Al Tala, S., Mutairi, F. A. & Alfadhel, M. Further delineation of the clinical phenotype of cerebellar ataxia, mental retardation, and disequilibrium syndrome type 4. *Journal of central nervous system disease* 10, 1179573518759682 (2018).

266. Chan, C.-H., Yu, T.-H. & Wong, K.-B. Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding. *PLoS One* 6, e21624 (2011).
267. Bandyopadhyay, A. K., Islam, R. N. U., Mitra, D., Banerjee, S. & Goswami, A. Stability of buried and networked salt-bridges (BNSB) in thermophilic proteins. *Bioinformatics* 15, 61–67 (2019).
268. Aspatwar, A. *et al.* Abnormal cerebellar development and ataxia in CARP VIII morphant zebrafish. *Human molecular genetics* 22, 417–432 (2013).
269. Sanyanga, T. A. & Tastan Bishop, Ö. Structural Characterization of Carbonic Anhydrase VIII and Effects of Missense Single Nucleotide Variations to Protein Structure and Function. *International Journal of Molecular Sciences* 21, 2764 (2020).
270. Mårtensson, L. G. *et al.* Role of an evolutionarily invariant serine for the stability of human carbonic anhydrase II. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* 1118, 179–186 (1992).
271. Almstedt, K. *et al.* Unfolding a folding disease: folding, misfolding and aggregation of the marble brain syndrome-associated mutant H107Y of human carbonic anhydrase II. *Journal of Molecular Biology* 342, 619–633 (2004).
272. Kiefer, L. L., Paterno, S. A. & Fierke, C. A. Hydrogen bond network in the metal binding site of carbonic anhydrase enhances zinc affinity and catalytic efficiency. *Journal of The American Chemical Society* 117, 6831–6837 (1995).
273. Tashian, R. E. in *Advances in genetics* 321–356 (Elsevier, 1992).
274. Sjöblom, B., Elleby, B., Wallgren, K., Jonsson, B.-H. & Lindskog, S. Two point mutations convert a catalytically inactive carbonic anhydrase-related protein (CARP) to an active enzyme. *Febs Letters* 398, 322–325 (1996).

275. Aspatwar, A., Tolvanen, M. E. & Parkkila, S. Phylogeny and expression of carbonic anhydrase-related proteins. *BMC molecular biology* **11**, 25 (2010).
276. Davey, N. E. *et al.* Attributes of short linear motifs. *Molecular BioSystems* **8**, 268–281 (2012).
277. Brown, D. K. *et al.* MD-TASK: A software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **33**, 2768–2771 (2017).
278. Wu, M.-J., Jiang, Y. & Yan, Y.-B. Impact of the 237th residue on the folding of human carbonic anhydrase II. *International journal of molecular sciences* **12**, 2797–2807 (2011).
279. 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* **47**, 1288–1298 (2008).
280. Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L. & Thornton, J. M. Metal ions in biological catalysis: from enzyme databases to general principles. *JBIC Journal of Biological Inorganic Chemistry* **13**, 1205–1218 (2008).
281. Andreini, C., Bertini, I. & Rosato, A. Metalloproteomes: a bioinformatic approach. *Accounts of chemical research* **42**, 1471–1479 (2009).
282. Berg, J. *Metal ions in proteins: structural and functional roles in Cold Spring Harbor symposia on quantitative biology* **52** (1987), 579–585.
283. González, M. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique* **12**, 169–200 (2011).
284. Beauchamp, K. A., Lin, Y.-S., Das, R. & Pande, V. S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *Journal of chemical theory and computation* **8**, 1409–1414 (2012).

285. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods* **14**, 71–73 (2017).
286. Ponder, J. W. & Case, D. A. in *Advances in protein chemistry* 27–85 (Elsevier, 2003).
287. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **117**, 5179–5197 (1995).
288. MacKerell Jr, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B* **102**, 3586–3616 (1998).
289. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **118**, 11225–11236 (1996).
290. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry* **25**, 1656–1676 (2004).
291. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Softwarex* **1**, 19–25 (2015).
292. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters in SC'06: *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing* (2006), 43–43.
293. Ferreira, G. M. *et al.* Inhibition of Porcine Aminopeptidase M (pAMP) by the Pentapeptide Microginins. *Molecules* **24**, 4369 (2019).
294. Shelley, J. C. *et al.* Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *Journal of computer-aided molecular design* **21**, 681–691 (2007).

295. Greenwood, J. R., Calkins, D., Sullivan, A. P. & Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of computer-aided molecular design* **24**, 591–604 (2010).
296. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **30**, 1545–1614 (2009).
297. Da Silva, A. W. S. & Vranken, W. F. ACPYPE-Antechamber python parser interface. *BMC Research Notes* **5**, 367 (2012).
298. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).
299. Zgarbová, M. *et al.* Refinement of the Cornell *et al.* nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation* **7**, 2886–2902 (2011).
300. Pérez, A. *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of [alpha]/[gamma] conformers *Biophys. J* **92**, 3817–3829 (2007).
301. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry* **21**, 1049–1074 (2000).
302. Wickstrom, L., Okur, A. & Simmerling, C. Evaluating the performance of the ff99SB force field based on NMR scalar coupling data. *Biophysical journal* **97**, 853–856 (2009).
303. Maier, J. A. *et al.* ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation* **11**, 3696–3713 (2015).

304. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25**, 1157–1174 (2004).
305. Li, P., Roberts, B. P., Chakravorty, D. K. & Merz Jr, K. M. Rational design of particle mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *Journal of chemical theory and computation* **9**, 2733–2748 (2013).
306. Pang, Y.-P. Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method. *Proteins: Structure, Function, and Bioinformatics* **45**, 183–189 (2001).
307. Li, P., Song, L. F. & Merz Jr, K. M. Systematic parameterization of monovalent ions employing the nonbonded model. *Journal of chemical theory and computation* **11**, 1645–1657 (2015).
308. Li, P., Song, L. F. & Merz Jr, K. M. Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water. *The Journal of Physical Chemistry B* **119**, 883–895 (2014).
309. Li, P. & Merz, K. M. MCPB.py: A python based metal center parameter builder. *Journal of Chemical Information and Modeling* **56**. PMID: 26913476, 599–604. eprint: <http://dx.doi.org/10.1021/acs.jcim.5b00674>. <http://dx.doi.org/10.1021/acs.jcim.5b00674> (2016).
310. Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *International journal of quantum chemistry* **60**, 1271–1277 (1996).
311. Frisch, M. J. *et al.* *Gaussian 09 Revision E.01* Gaussian Inc. Wallingford CT 2009.
312. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **5**, 129–145 (1984).

313. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).
314. Besler, B. H., Merz Jr, K. M. & Kollman, P. A. Atomic charges derived from semiempirical methods. *Journal of Computational Chemistry* **11**, 431–439 (1990).
315. Cieplak, P., Cornell, W. D., Bayly, C. & Kollman, P. A. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *Journal of Computational Chemistry* **16**, 1357–1377 (1995).
316. Gordon, J. *et al.* H⁺⁺: A server for estimating pK_as and adding missing hydrogens to macromolecules. *Nucleic Acids Research* **33**, W368–71 (Aug. 2005).
317. Dennington, R., Keith, T., Millam, J., *et al.* GaussView, version 5. *Semichem Inc.: Shawnee Mission, KS* (2009).
318. Bernadat, G., Supuran, C. T. & Iorga, B. I. Carbonic anhydrase binding site parameterization in OPLS-AA force field. *Bioorganic & Medicinal Chemistry* **21**, 1427–1430 (2013).
319. Karplus, M. & McCammon, J. Molecular dynamics simulations of biomolecules. *Nature structural and molecular biology* **9**, 646–652 (2002).
320. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC* **8**, 37 (2015).
321. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N · log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **98**, 10089–10092 (1993).
322. Hockney, R. & Eastwood, J. Computer Simulation Using Particles. *Computer Simulation Using Particles, New York: McGraw-Hill, 1981* (1981).

323. Orozco, M. *et al.* in *Advances in protein chemistry and structural biology* 183–215 (Elsevier, 2011).
324. Garg, A. & Pal, D. Exploring the use of molecular dynamics in assessing protein variants for phenotypic alterations. *Human mutation* (2019).
325. Kapla, J., Stevansson, B. & Maliniak, A. Coarse-Grained Molecular Dynamics Simulations of Membrane–Trehalose Interactions. *The Journal of Physical Chemistry B* **120**, 9621–9631 (2016).
326. Shimizu, M. & Takada, S. Reconstruction of Atomistic Structures from Coarse-Grained Models for Protein–DNA Complexes. *Journal of chemical theory and computation* **14**, 1682–1694 (2018).
327. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202 (2016).
328. Sittel, F., Jain, A. & Stock, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *The Journal of chemical physics* **141**, 07B605_1 (2014).
329. Amadei, A., Linssen, A. B. & Berendsen, H. J. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **17**, 412–425 (1993).
330. Amadei, A., Linssen, A., De Groot, B., Van Aalten, D. & Berendsen, H. An efficient method for sampling the essential subspace of proteins. *Journal of Biomolecular Structure and Dynamics* **13**, 615–625 (1996).
331. Roe, D. R., Bergonzo, C. & Cheatham III, T. E. Evaluation of enhanced sampling provided by accelerated molecular dynamics with Hamiltonian replica exchange methods. *The Journal of Physical Chemistry B* **118**, 3543–3552 (2014).

332. Haider, S., Parkinson, G. N. & Neidle, S. Molecular dynamics and principal components analysis of human telomeric quadruplex multimers. *Biophysical journal* **95**, 296–311 (2008).
333. Hagberg, A., Swart, P. & Chult, D. *Exploring network structure, dynamics, and function using NetworkX* tech. rep. (Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008).
334. Demir, Y., Demir, N., Nadaroglu, H. & Bakan, E. Purification and characterization of carbonic anhydrase from bovine erythrocyte plasma membrane (2000).
335. Demir, N., Demir, Y. & Coşkun, F. Purification and characterization of carbonic anhydrase from human erythrocyte plasma membrane. *Turkish Journal of Medical Sciences* **31**, 477–482 (2001).
336. Schafmeister, C., Ross, W. & Romanovski, V. LEaP. *University of California, San Francisco* (1995).
337. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* **25**, 247–260 (2006).
338. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics* **52**, 7182–7190 (1981).
339. Roe, D. R. & Cheatham III, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation* **9**, 3084–3095 (2013).
340. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

341. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: A review. *ACM Computing Surveys (CSUR)* **31**, 264–323 (1999).
342. Shao, J., Tanner, S. W., Thompson, N. & Cheatham, T. E. Clustering molecular dynamics trajectories: 1. Characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation* **3**, 2312–2334 (2007).
343. Bouguettaya, A., Yu, Q., Liu, X., Zhou, X. & Song, A. Efficient agglomerative hierarchical clustering. *Expert Systems With Applications* **42**, 2785–2797 (2015).
344. Chakrabarty, B. & Parekh, N. NAPS: Network analysis of protein structures. *Nucleic Acids Research* **44**, W375–W382 (2016).
345. Williams, T., Kelley, C. & many others. *Gnuplot 5.2: An interactive plotting program* 2018. <http://www.gnuplot.info>.
346. Silverman, B. W. *Density estimation for statistics and data analysis* (CRC press, 1986).
347. Scott, D. W., Tapia, R. A. & Thompson, J. R. Kernel density estimation revisited. *Nonlinear Analysis: Theory, Methods & Applications* **1**, 339–372 (1977).
348. Penkler, D., Atilgan, C. & Tastan Bishop, Ö. Allosteric Modulation of Human Hsp90 α Conformational Dynamics. *Journal of Chemical Information and Modeling* (2018).
349. Elder, I., Tu, C., Ming, L.-J., McKenna, R. & Silverman, D. N. Proton transfer from exogenous donors in catalysis by human carbonic anhydrase II. *Archives of Biochemistry and Biophysics* **437**, 106–114 (2005).
350. Bhatt, D. *et al.* Proton transfer in a Thr200His mutant of human carbonic anhydrase II. *PROTEINS: Structure, Function, and Bioinformatics* **61**, 239–245 (2005).
351. Fisher, Z. *et al.* Structural and kinetic characterization of active-site histidine as a proton shuttle in catalysis by human carbonic anhydrase II. *Biochemistry* **44**, 1097–1105 (2005).

352. 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. *Biophysical Journal* **92**, 562–570 (2007).
353. An, H. *et al.* Chemical rescue in catalysis by human carbonic anhydrases II and III. *Biochemistry* **41**, 3235–3242 (2002).
354. Håkansson, K., Carlsson, M., Svensson, L. A. & Liljas, A. Structure of native and apo carbonic anhydrase II and structure of some of its anion-ligand complexes. *Journal of molecular biology* **227**, 1192–1204 (1992).
355. Alvarez, B. V. *et al.* Identification and characterization of a novel mutation in the carbonic anhydrase IV gene that causes retinitis pigmentosa. *Investigative ophthalmology & visual science* **48**, 3459–3468 (2007).
356. Kokh, D. B., Wade, R. C. & Wenzel, W. Receptor flexibility in small-molecule docking calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 298–314 (2011).
357. Antunes, D. A., Devaurs, D. & Kaviraki, L. E. Understanding the challenges of protein flexibility in drug design. *Expert opinion on drug discovery* **10**, 1301–1313 (2015).
358. Lexa, K. W. & Carlson, H. A. Protein flexibility in docking and surface mapping. *Quarterly reviews of biophysics* **45**, 301–343 (2012).
359. 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 In Bioinformatics* (2019).
360. 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. *The American Journal of Human Genetics* **104**, 861–878 (2019).
361. 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 Research International* **2017** (2017).
362. Alvarez, B. V., Loisel, F. B., Supuran, C. T., Schwartz, G. J. & Casey, J. R. Direct extracellular interaction between carbonic anhydrase IV and the human NBC1 sodium/bicarbonate co-transporter. *Biochemistry* **42**, 12321–12329 (2003).
363. Balasco, N., Esposito, L., Simone, A. D. & Vitagliano, L. Role of loops connecting secondary structure elements in the stabilization of proteins isolated from thermophilic organisms. *Protein Science* **22**, 1016–1023 (2013).
364. Papaleo, E. *et al.* The role of protein loops and linkers in conformational dynamics and allostery. *Chemical reviews* **116**, 6391–6423 (2016).
365. Amitai, G. *et al.* Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology* **344**, 1135–1146 (2004).
366. Fisher, S. Z. *et al.* 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* **46**, 3803–3813 (2007).
367. 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* **47**, 12028–12036 (2008).

368. Buonanno, M. *et al.* The crystal structure of a hCA VII variant provides insights into the molecular determinants responsible for its catalytic behavior. *International Journal of Molecular Sciences* **19**, 1571 (2018).
369. Maupin, C. M., McKenna, R., Silverman, D. N. & Voth, G. A. Elucidation of the proton transport mechanism in human carbonic anhydrase II. *Journal of The American Chemical Society* **131**, 7598–7608 (2009).
370. 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* **35**, 16421–16428 (1996).
371. Cui, Q. & Karplus, M. Is a “proton wire” concerted or stepwise? A model study of proton transfer in carbonic anhydrase. *The Journal of Physical Chemistry B* **107**, 1071–1078 (2003).
372. Lomelino, C. L., Supuran, C. T. & McKenna, R. Non-classical inhibition of carbonic anhydrase. *International journal of molecular sciences* **17**, 1150 (2016).

Supplementary Material

Table S1. Table of UniProt CA accession numbers for final CA family dataset.

Sequence	UniProt accession
CA-I	P00915
CA-II	P00918
CA-III	P07451
CA-IV	P22748
CA-VA	P35218
CA-VB	Q9Y2D0
CA-VI	P23280
CA-VII	P43166
CA-VIII	P35219
CA-IX	Q16790
CA-X	Q9NS85
CA-XI	O75493
CA-XII	043570
CA-XIII	Q8N1Q1
CA-XIV	Q9ULX7
CA-XV	Q99N23

Table S2. CA STRING predicted functional partners and confidence (approximate probability) values representing strength of data support.

CA Protein	Associated protein/gene name	Confidence Score
CA-II	CCDH1	0.925
	CTNNB1	0.916
	CTNND1	0.905
	MLLT4	0.9
	RAP1A	0.9
	RAP1B	0.9
	CTNNA1	0.9
	SLC9A1	0.878
	TDG	0.874
	SLC4A4	0.8
CA-IV	SLC4A4	0.824
	CYP24A1	0.739
	TALDO1	0.643
	UHMK1	0.641
	WTAP	0.633
	VEGFA	0.632
	PRSS12	0.632
	RP9	0.57
	SLC4A1	0.566
	SYT4	0.565
CA-VIII	WDR81	0.903
	CYP24A1	0.796
	ITPR1	0.761
	VLDLR	0.725
	ZNF385A	0.713
	ERP44	0.670
	UHMK1	0.670
	AHCYL1	0.649
	KIAA1045	0.627
	TOX	0.622

Table S3. CA-VIII potential protein-protein binding sites and residues. SNV positions are italicised, underlined and highlighted in bold red.

SiteMap		
Binding site	SiteScore	Binding site residues
1	0.885	23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 43 44 45 46 47 49 51 52 53 54 55 58 59 60 61 62 76 77 78 79 80 81 83 84 85 86 87 88 90 92 93 94 95 96 111 113 114 116 118 119 121 122 123 124 125 126 127 128 130 133 134 135 136 137 141 143 145 152 153 154 155 157 158 159 163 165 171 172 177 178 180 181 183 184 185 187 188 190 191 192 193 194 196 198 199 200 201 202 205 222 223 224 225 226 228 229 233 245 246 248 249 251 252 254 255 256 257 258 260 261 262 263 264 266 269 270 271 272 274 275 276 277 278 279 280 281 282 283 284 285 286
2	0.908	56 57 59 60 61 65 66 69 70 72 102 131 132 134 136 169 171 210 211 212 213 214 215 216 236 <u>237</u> 238 239 287 288 289 290
3	0.897	43 46 47 48 64 65 67 68 102 103 104 105 106 110 158 160 161 163 164 217 218 219 220 228 229 230 231 232 282 283 284 285 286
4	0.881	68 69 71 72 73 74 75 96 97 98 <u>100</u> 101 106 107 <u>109</u> 174 175 176 177 178 204 205 207 208 209 210 212 213 238 239 240 241
5	0.792	190 191 252 254 255 256 257 258 259 260 262 263 266 267 268 269
CPORT		
1	NA*	26 27 28 29 30 31 32 33 34 35 36 39 40 44 93 94 111 113 116 147 151 153 214 224 <u>237</u> 238 255 256 262 263 264 265 269 274 276 278 289 290
Consensus of SiteMap and CPORT		
1	NA*	26 27 28 29 30 31 32 33 34 35 36 39 40 44 93 94 111 113 116 147 151 153 214 224 <u>237</u> 238 255 256 262 263 264 265 269 274 276 278 289 290

*NA refers to not applicable.

```

      4-          4-          5W          15H          25G          30P          40Y          48S          58R
CA-II/1-260  1 MSH-----HWGYGKHNGPEHWHKDFPIAKG-----EROSPVDIDTHTAKYDPSL--KPLSVSYDQATSLRILNNGHAFN 67
CA-IV/1-312  1 MRMLLALLALSARPSASAES-----HWCYEVAESSNYPCLVYVKGWGNQKDRQSPINIVITKAKVDKKL-GRFFSGYDKKQWTWVONNGHSVM 91
CA-VIII/1-290 1 MADL-SFIEDTVAFPEKEEDEEEEEEGVEWGY--EEGVE-WGLVFDANG-----EYQSPINLNSREARYDPSLDVRLSPNYVVCRCDCVMDGHTIQ 90

      68V          78V          86G          96H          106E          116A          125T          135Q          144G          153K
CA-II/1-260  68 VEFDDSDKAVLKGGPL--DGTYRLIQFHFWGSLDGGSEHTVDKKEYAAELHLVHWN-ISKYEDFGKAVQOP-DGLAVLGIFLKVSA-KPGLQKVVVD 161
CA-IV/1-312  92 ML---ENKASISGGGLP--APYQAKQLHLHWSDLPYKGEHSLDGEHFAMEMHIVHEKEKGTSRNVKEAQQPEDEI AVLAFLEAGTQVNEGFQPLVE 185
CA-VIII/1-290 91 VIL---KSKSVLSGGPLPQGHFEFLYEVRFHWGRENRQSGSEHTVNFKAEPMELHLIHWNSTLFGS IDEAVGKH-HGIAIIALFVQIGKE-HVGLKAVTE 184

      163L          173A          183L          189D          199T          209I          219S          229N          237-          244W
CA-II/1-260  162 VLDSTIKTKGKSADFTNFDPRGLLP--ESL--DYWTYPGSLTTPPLLECVTWIVLKEPI SVSSEQVLFKRLNFNGEGEP-----EELMVDNWRFAOPLK 251
CA-IV/1-312  186 ALSNIPKPEMSTTMAESSLLDLLPKEEKL-RHYFRYLGSLTTPTCDEKVVWTVFREP IQLHREQILAFSQKLYYDKEQTVS-----MKDNVRPLQOLG 277
CA-VIII/1-290 185 ILQDIQYKGSKTIIPCFNPNNTLLP--DPLLDYVWVYEGSLTI PPCSEGVTVWILFRYPLTISQLQIEEFRRRLRTHVKGAE LVEGCDGILGDNFRRTQPLS 281

      254Q          256-          256-          259F
CA-II/1-260  252 NROII-----KASFK- 260
CA-IV/1-312  278 QRTVIKSGAPGRPLPWALPALLGPMLACLLAGFLR 312
CA-VIII/1-290 282 DRVLI-----RAAQ- 290

```

Figure S1. Multiple sequence alignment of CA-II, CA-IV and CA-VIII. CA-II in bold has been selected as the reference sequence.

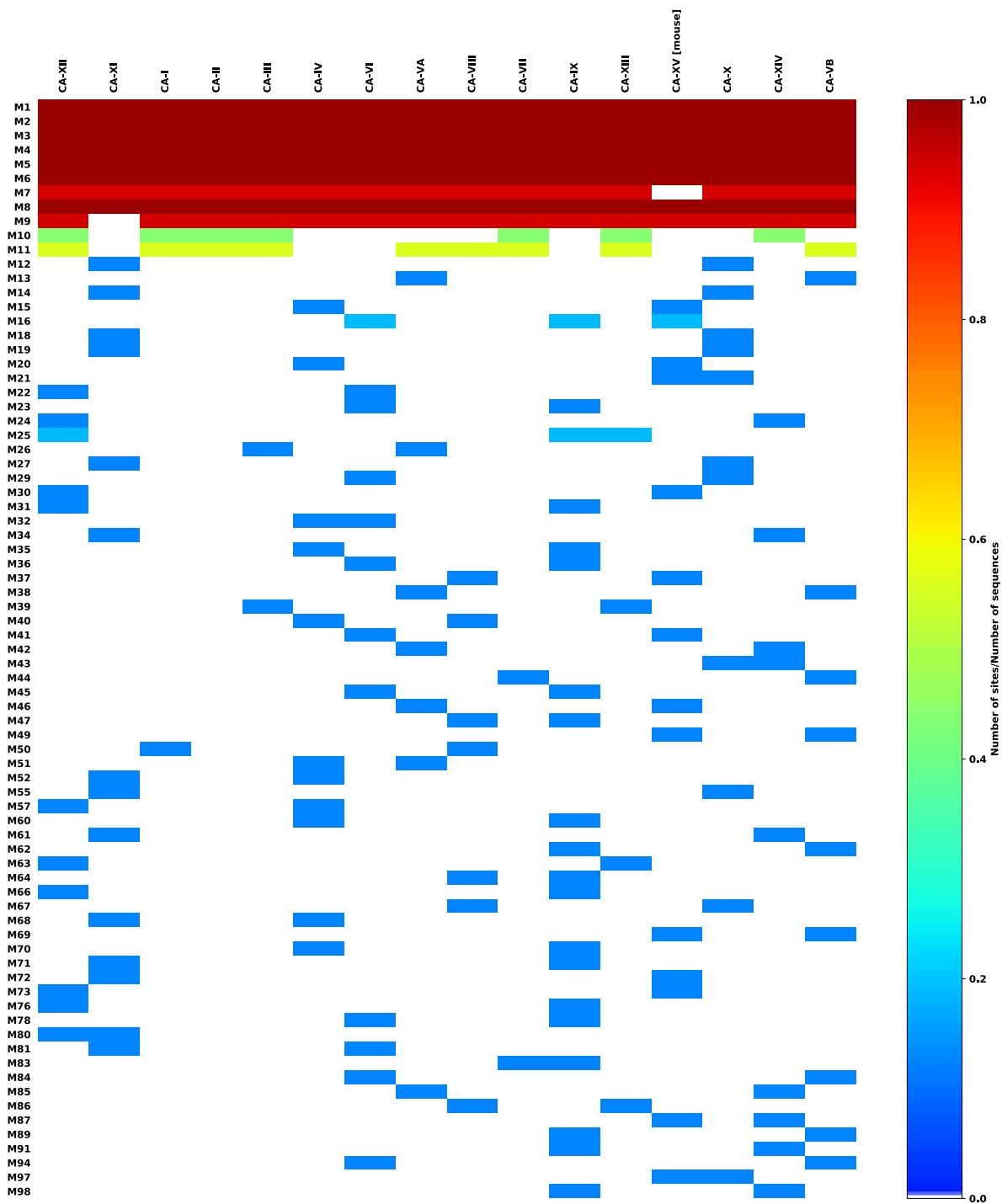


Figure S2. Conserved motifs within the human α -CA family, and associated proteins and UniProt accessions. Motif conservation is expressed as a heat map representing the number of motif sites per total protein sequences.

Table S4. Identified human α -CA motifs and associated E-values.

Motif	E-value	Motif	E-value
M1	2.2×10^{-164}	M43	5.6×10^{004}
M2	4.4×10^{-140}	M44	6.2×10^{004}
M3	1.2×10^{-108}	M45	8.8×10^{004}
M4	3.0×10^{-089}	M46	1.1×10^{005}
M5	6.5×10^{-086}	M47	2.3×10^{005}
M6	3.8×10^{-073}	M49	6.5×10^{004}
M7	6.0×10^{-054}	M50	1.1×10^{005}
M8	6.4×10^{-040}	M51	1.3×10^{005}
M9	1.9×10^{-041}	M52	1.7×10^{005}
M10	7.7×10^{-025}	M55	3.3×10^{005}
M11	2.3×10^{-006}	M57	7.3×10^{004}
M12	2.6×10^{-002}	M60	1.2×10^{005}
M13	6.9×10^{-001}	M61	1.6×10^{005}
M14	8.1×10^{-001}	M62	1.6×10^{005}
M15	8.5×10^{000}	M63	5.9×10^{004}
M16	2.8×10^{001}	M64	2.1×10^{005}
M18	1.3×10^{002}	M66	2.0×10^{005}
M19	2.2×10^{002}	M67	2.5×10^{005}
M20	2.5×10^{002}	M68	2.5×10^{005}
M21	1.0×10^{003}	M69	4.8×10^{005}
M22	1.6×10^{003}	M70	4.8×10^{005}
M23	2.5×10^{003}	M71	5.5×10^{005}
M24	2.8×10^{003}	M72	1.0×10^{006}
M25	3.0×10^{003}	M73	7.6×10^{005}
M26	4.8×10^{003}	M76	4.2×10^{005}
M27	5.7×10^{003}	M78	1.5×10^{006}
M29	1.2×10^{003}	M80	9.3×10^{005}
M30	5.4×10^{003}	M81	3.4×10^{006}
M31	6.0×10^{003}	M83	7.7×10^{005}
M32	6.8×10^{003}	M84	5.2×10^{005}
M34	1.4×10^{004}	M85	6.2×10^{005}
M35	1.8×10^{004}	M86	6.0×10^{005}
M36	2.0×10^{004}	M87	7.9×10^{005}
M37	2.0×10^{004}	M89	2.9×10^{006}
M38	1.1×10^{004}	M91	2.4×10^{006}
M39	2.1×10^{004}	M94	8.7×10^{005}
M40	3.8×10^{004}	M97	2.6×10^{006}
M41	4.0×10^{004}	M98	2.7×10^{006}
M42	4.5×10^{004}		

```

REMARK GOES HERE, THIS FILE IS GENERATED BY MCPB.PY
MASS
M1 65.4 Zn ion
Y1 14.01 0.530 sp2 N in 5 memb.ring w/LP (HIS,ADE,GUA)
Y2 14.01 0.530 sp2 N in 5 memb.ring w/LP (HIS,ADE,GUA)
Y3 14.01 0.530 sp2 N in 5 memb.ring w/LP (HIS,ADE,GUA)
Y4 16.000 0.465 same as ow

BOND
M1-Y4 41.2 2.1088 Created by Seminario method using MCPB.py
Y1-M1 91.7 1.9812 Created by Seminario method using MCPB.py
Y2-M1 94.3 1.9764 Created by Seminario method using MCPB.py
Y3-M1 93.0 1.9806 Created by Seminario method using MCPB.py
CC-Y3 410.0 1.394 JCC,7,(1986),230; HIS
CR-Y1 488.0 1.335 JCC,7,(1986),230; HIS
CR-Y2 488.0 1.335 JCC,7,(1986),230; HIS
Y1-CV 410.0 1.394 JCC,7,(1986),230; HIS
Y2-CV 410.0 1.394 JCC,7,(1986),230; HIS
Y3-CR 488.0 1.335 JCC,7,(1986),230; HIS
Y4-HW 553.0 0.9572 ! TIP3P water

ANGL
CC-Y3-M1 56.44 127.30 Created by Seminario method using MCPB.py
CR-Y1-M1 38.97 125.70 Created by Seminario method using MCPB.py
CR-Y2-M1 53.67 127.46 Created by Seminario method using MCPB.py
M1-Y1-CV 39.62 128.05 Created by Seminario method using MCPB.py
M1-Y2-CV 54.89 126.22 Created by Seminario method using MCPB.py
M1-Y3-CR 54.18 125.48 Created by Seminario method using MCPB.py
M1-Y4-HW 44.55 122.59 Created by Seminario method using MCPB.py
Y1-M1-Y2 39.89 116.07 Created by Seminario method using MCPB.py
Y1-M1-Y3 37.44 115.90 Created by Seminario method using MCPB.py
Y1-M1-Y4 31.02 101.10 Created by Seminario method using MCPB.py
Y2-M1-Y3 36.29 114.63 Created by Seminario method using MCPB.py
Y2-M1-Y4 36.83 105.42 Created by Seminario method using MCPB.py
Y3-M1-Y4 30.04 100.62 Created by Seminario method using MCPB.py
CC-CV-Y1 70.0 120.00 AA his
CC-CV-Y2 70.0 120.00 AA his
CC-Y3-CR 70.0 117.00 AA his
CR-Y1-CV 70.0 117.00 AA his
CR-Y2-CV 70.0 117.00 AA his
CT-CC-Y3 70.0 120.00 AA his
CW-CC-Y3 70.0 120.00 AA his
HW-Y4-HW 100.000 104.520 same as hw-ow-hw, penalty score= 0.0
NA-CR-Y1 70.0 120.00 AA his
NA-CR-Y2 70.0 120.00 AA his
Y1-CR-H5 50.0 120.00 AA his
Y1-CV-H4 50.0 120.00 AA his
Y2-CR-H5 50.0 120.00 AA his
Y2-CV-H4 50.0 120.00 AA his
Y3-CR-H5 50.0 120.00 AA his
Y3-CR-NA 70.0 120.00 AA his

DIHE
X -CC-Y3-X 2 4.8 180.0 2.0 JCC,7,(1986),230
X -CR-Y1-X 2 10.0 180.0 2.0 JCC,7,(1986),230
X -CR-Y2-X 2 10.0 180.0 2.0 JCC,7,(1986),230
X -CV-Y1-X 2 4.8 180.0 2.0 JCC,7,(1986),230
X -CV-Y2-X 2 4.8 180.0 2.0 JCC,7,(1986),230
X -Y3-CR-X 2 10.0 180.0 2.0 JCC,7,(1986),230
CC-CV-Y1-M1 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CC-CV-Y2-M1 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CC-Y3-M1-Y4 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CR-Y1-M1-Y2 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CR-Y1-M1-Y3 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CR-Y1-M1-Y4 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CR-Y2-M1-Y3 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CR-Y2-M1-Y4 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CT-CC-Y3-M1 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CW-CC-Y3-M1 3 0.00 0.00 3.0 Treat as zero by MCPB.py
CX-CT-CC-Y3 1 0.047 180.0 -4.0
CX-CT-CC-Y3 1 0.74 0.0 -3.0
CX-CT-CC-Y3 1 0.204 0.0 -2.0
CX-CT-CC-Y3 1 0.69 0.0 1.0
M1-Y1-CR-H5 3 0.00 0.00 3.0 Treat as zero by MCPB.py
M1-Y1-CV-H4 3 0.00 0.00 3.0 Treat as zero by MCPB.py
M1-Y2-CR-H5 3 0.00 0.00 3.0 Treat as zero by MCPB.py
M1-Y2-CV-H4 3 0.00 0.00 3.0 Treat as zero by MCPB.py
M1-Y3-CR-H5 3 0.00 0.00 3.0 Treat as zero by MCPB.py
M1-Y3-CR-NA 3 0.00 0.00 3.0 Treat as zero by MCPB.py
NA-CR-Y1-M1 3 0.00 0.00 3.0 Treat as zero by MCPB.py
NA-CR-Y2-M1 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y1-M1-Y2-CR 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y1-M1-Y2-CV 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y1-M1-Y3-CC 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y1-M1-Y3-CR 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y1-M1-Y4-HW 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y2-M1-Y1-CV 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y2-M1-Y3-CC 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y2-M1-Y3-CR 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y2-M1-Y4-HW 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y3-M1-Y1-CV 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y3-M1-Y2-CV 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y3-M1-Y4-HW 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y4-M1-Y1-CV 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y4-M1-Y2-CV 3 0.00 0.00 3.0 Treat as zero by MCPB.py
Y4-M1-Y3-CR 3 0.00 0.00 3.0 Treat as zero by MCPB.py

IMPR
CT-CW-CC-Y3 1.1 180. 2.

NONB
M1 1.3950 0.0149170000 IOD set for Zn2+ ion from Li et al. JCTC, 2013, 9, 2733
Y1 1.8240 0.1700 OPLS
Y2 1.8240 0.1700 OPLS
Y3 1.8240 0.1700 OPLS
Y4 1.7683 0.1520 same as ow

```

Listing S1: Generated CA-II *frmod* parameters for molecular dynamics parameter export.

```

source leaprc.protein.ff14SB
source leaprc.water.tip3p
source leaprc.gaff2
addAtomTypes {
    { "M1" "Zn" "sp3" }
    { "Y1" "N" "sp3" }
    { "Y2" "N" "sp3" }
    { "Y3" "N" "sp3" }
    { "Y4" "O" "sp3" }
}
HD1 = loadmol2 HD1.mol2
HD2 = loadmol2 HD2.mol2
HE1 = loadmol2 HE1.mol2
ZN1 = loadmol2 ZN1.mol2
WT1 = loadmol2 WT1.mol2
BCT = loadmol2 BCT.mol2
CO2 = loadmol2 CO2.mol2
loadamberparams BCT.frcmod
loadamberparams CO2.frcmod
loadamberparams WAT.frcmod
loadamberparams frcmod.ions1lm_126_tip3p
loadamberparams ca2_mcpbpy.frcmod
mol = loadpdb ca2_final_zaff.pdb
bond mol.94.NE2 mol.261.ZN
bond mol.96.NE2 mol.261.ZN
bond mol.119.ND1 mol.261.ZN
bond mol.261.ZN mol.262.O
bond mol.93.C mol.94.N
bond mol.94.C mol.95.N
bond mol.95.C mol.96.N
bond mol.96.C mol.97.N
bond mol.118.C mol.119.N
bond mol.119.C mol.120.N
savepdb mol ca2_dry.pdb
saveamberparm mol ca2_dry.prmtop ca2_dry.inpcrd
solvatebox mol TIP3PBOX 10.0
addions mol Na+ 0
addions mol Cl- 0
savepdb mol ca2_solv.pdb
saveamberparm mol ca2_solv.prmtop ca2_solv.inpcrd
quit

```

Listing S2: Example of CA-II LEaP input for MD protein topology preparation.


```

parm ca2_dry_initial_noWAT.pdb
trajin md_0_1_noWAT_PBC.xtc
rms first :1-258 out rmsd_CA_WT.agr mass time 0.01
average crdset WT_CA_average
createcrd WT_CA_trajectories
run
crdaction WT_CA_trajectories rms ref WT_CA_average :1-258
crdaction WT_CA_trajectories matrix covar name WT_CA_covar :1-258
runanalysis diagmatrix WT_CA_covar out WT_CA_evecs.dat vecs 3 \
    name myEvecs nmwiz nmwizvecs 3 nmwizfile prot_WT.nmd nmwizmask :1-258
runanalysis modes eigenval name myEvecs out WT_CA_evalfrac.dat
crdaction WT_CA_trajectories projection WT modes \
    myEvecs beg 1 end 3 :1-258 crdframes 1,20001
hist WT:1 bins 200 out WT_CA_hist_1.agr norm name WT_CA_1
hist WT:2 bins 200 out WT_CA_hist_2.agr norm name WT_CA_2
hist WT:3 bins 200 out WT_CA_hist_3.agr norm name WT_CA_3
hist WT:1 WT:2 free 300 bins 200 out WT_CA_hist_1_2_fes.gnu name WT_CA_1_2
hist WT:2 WT:3 free 300 bins 200 out WT_CA_hist_2_3_fes.gnu name WT_CA_2_3
hist WT:1 WT:3 free 300 bins 200 out WT_CA_hist_1_3_fes.gnu name WT_CA_1_3
run
quit

```

Listing S3: Example of PCA script using to calculate 3D structural differences between the WT and variant proteins.

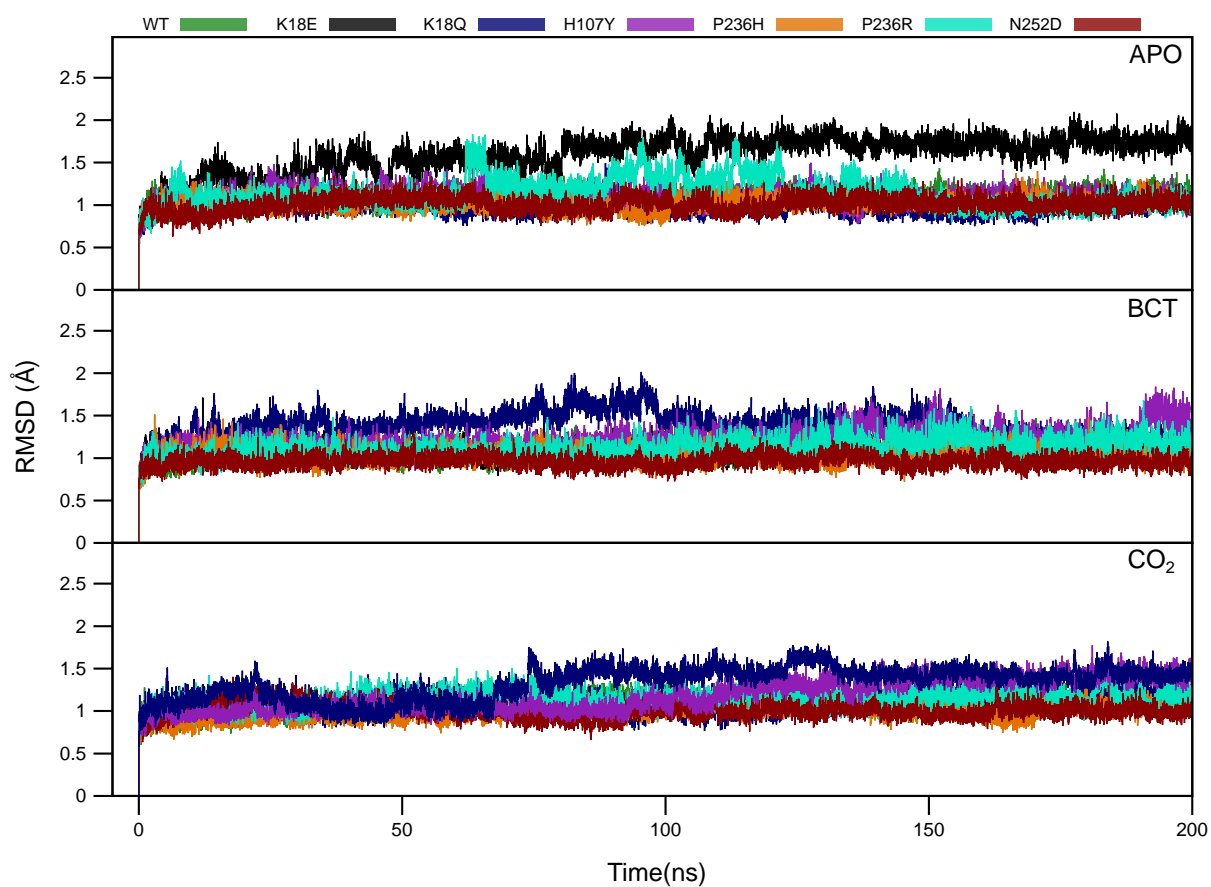


Figure S3. RMSD of the apo, BCT and CO₂ bound CA-II protein over the 200 ns MD simulation.

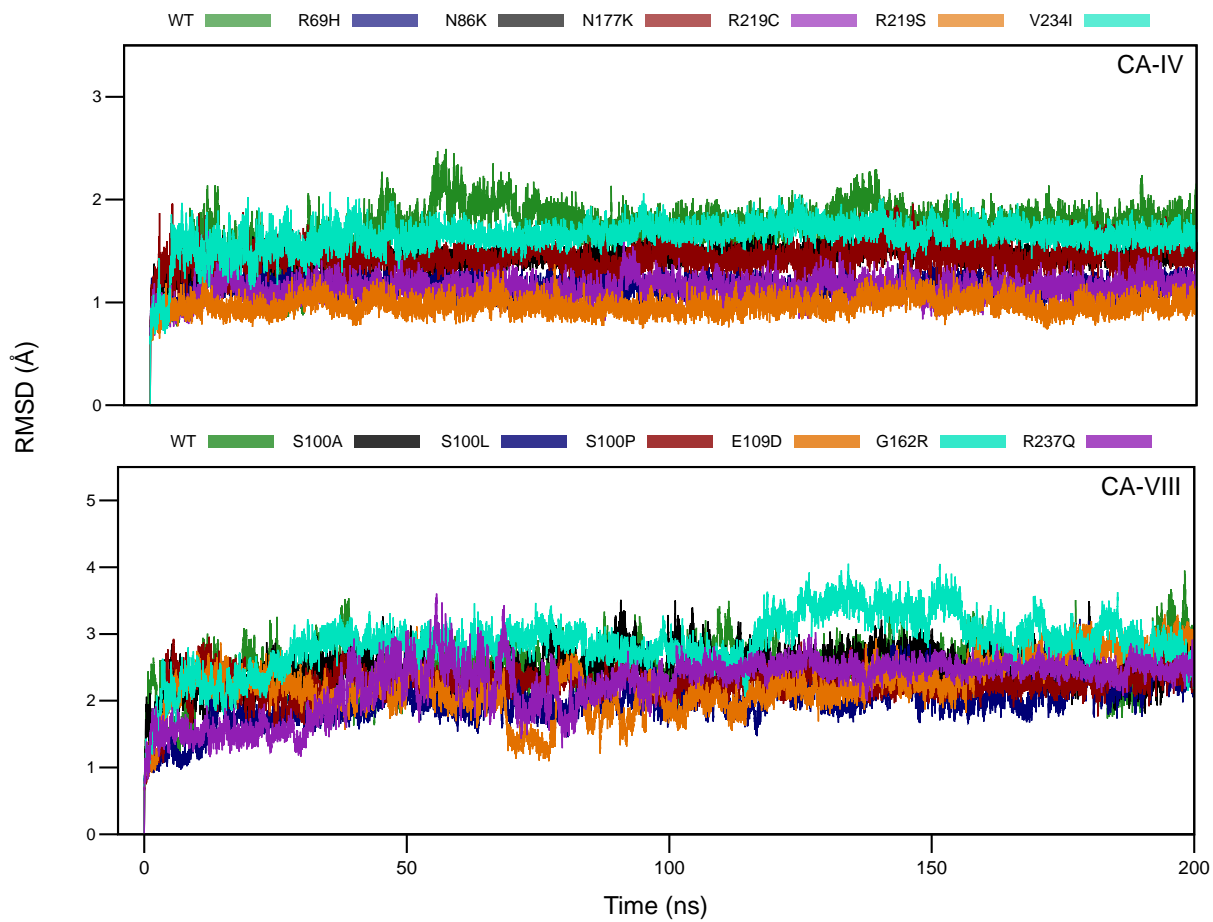


Figure S4. RMSD of the CA-IV and CA-VIII proteins over the 200 ns MD simulation.

Table S5. Eigenvalue fraction of the various CA-II WT and variant protein states.

Protein	Substrate	PC1	PC2	PC3
WT	Apo	0.4673	0.3253	0.2074
	BCT	0.4256	0.3550	0.2194
	CO ₂	0.5399	0.2589	0.2012
K18E	Apo	0.5503	0.2500	0.1997
	BCT	0.4836	0.3165	0.1999
	CO ₂	0.5772	0.2218	0.2010
K18Q	Apo	0.5260	0.2732	0.2009
	BCT	0.4497	0.3598	0.1905
	CO ₂	0.6091	0.2647	0.1261
H107Y	Apo	0.4439	0.3322	0.2239
	BCT	0.5688	0.2764	0.1548
	CO ₂	0.7201	0.1670	0.1129
P236H	Apo	0.5535	0.2628	0.1837
	BCT	0.5195	0.2741	0.2064
	CO ₂	0.4846	0.2977	0.2177
P236R	Apo	0.4693	0.3107	0.2200
	BCT	0.4772	0.3415	0.1813
	CO ₂	0.4218	0.3208	0.2574
N252D	Apo	0.4415	0.3358	0.2227
	BCT	0.4635	0.3607	0.1758
	CO ₂	0.4348	0.3235	0.2417

Table S6. Eigenvalue fraction of the various CA-IV WT and variant proteins.

Protein	PC1	PC2	PC3
WT	0.7306	0.1883	0.0811
R69H	0.4246	0.3094	0.2661
N86K	0.4290	0.3208	0.2502
N117K	0.6188	0.2277	0.1534
R219C	0.4190	0.3348	0.2462
R219S	0.4495	0.2981	0.2523
V234I	0.6594	0.1989	0.1417

Table S7. Eigenvalue fraction of the various CA-VIII WT and variant proteins.

Protein	PC1	PC2	PC3
WT	0.4272	0.3532	0.2196
S100A	0.4239	0.3479	0.2281
S100L	0.4816	0.3540	0.1644
S100P	0.6399	0.2364	0.1237
E109D	0.4889	0.2854	0.2257
G162R	0.5675	0.3060	0.1264
R237Q	0.7264	0.1903	0.0833

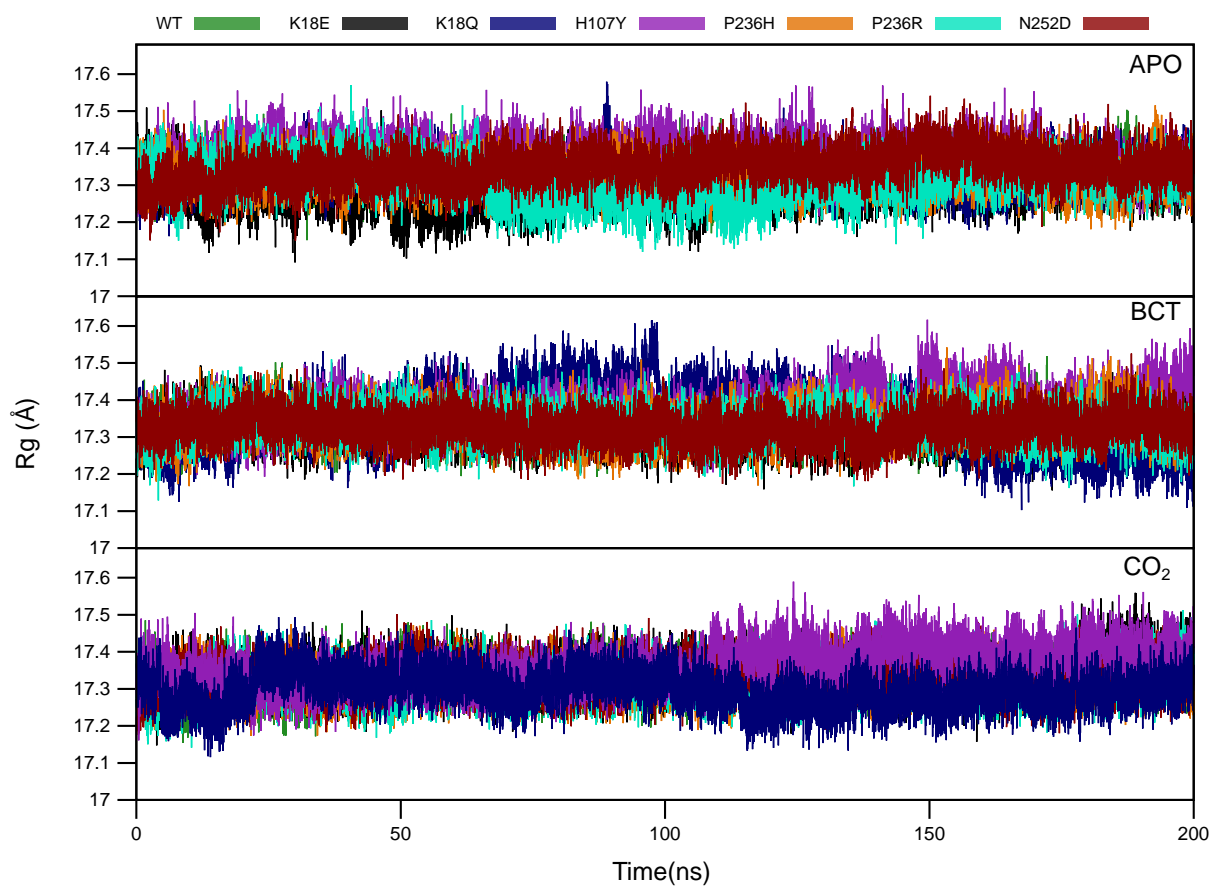


Figure S5. Rg of the apo, BCT and CO₂ bound CA-II protein over the 200 ns MD simulation.

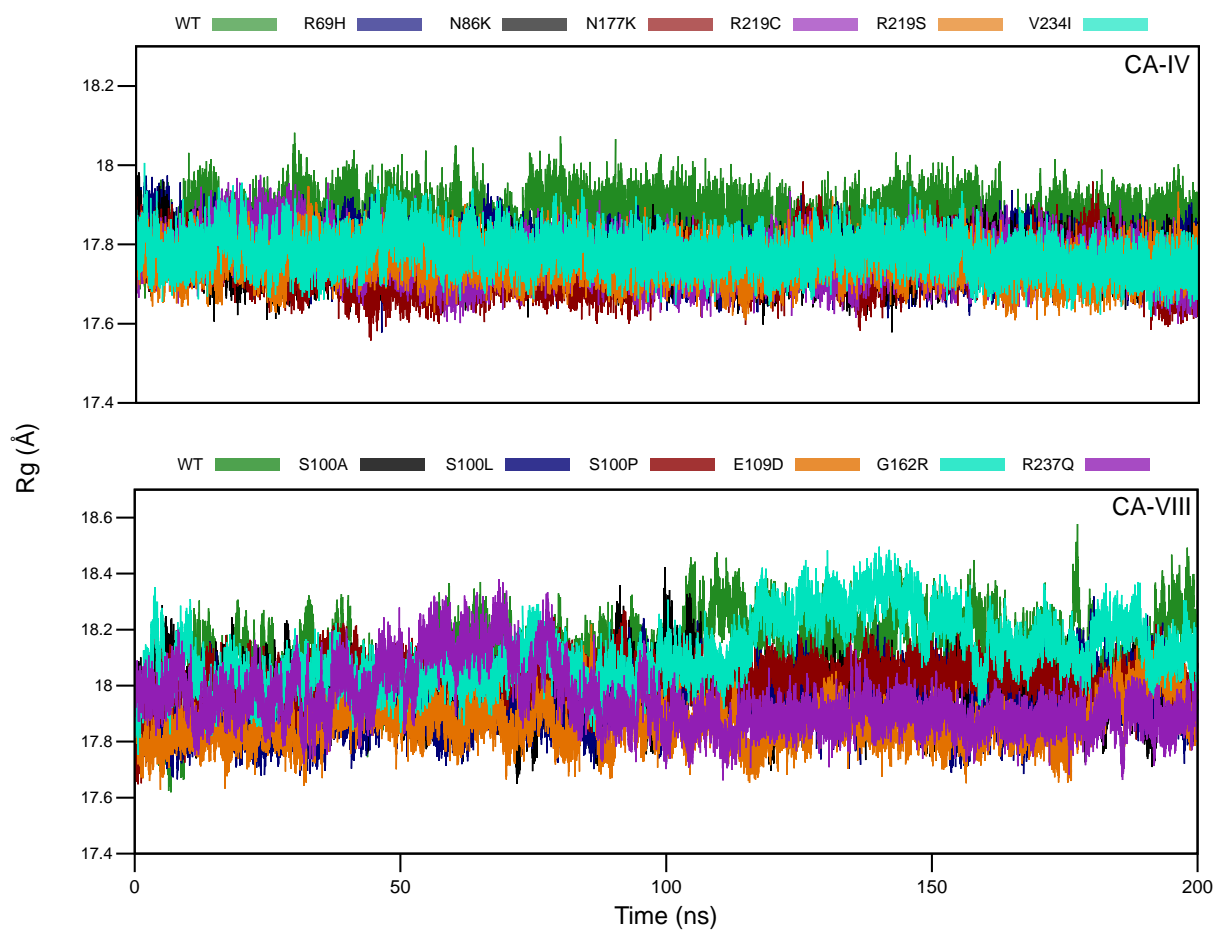


Figure S6. Rg of the CA-IV and CA-VIII proteins over the 200 ns MD simulation.

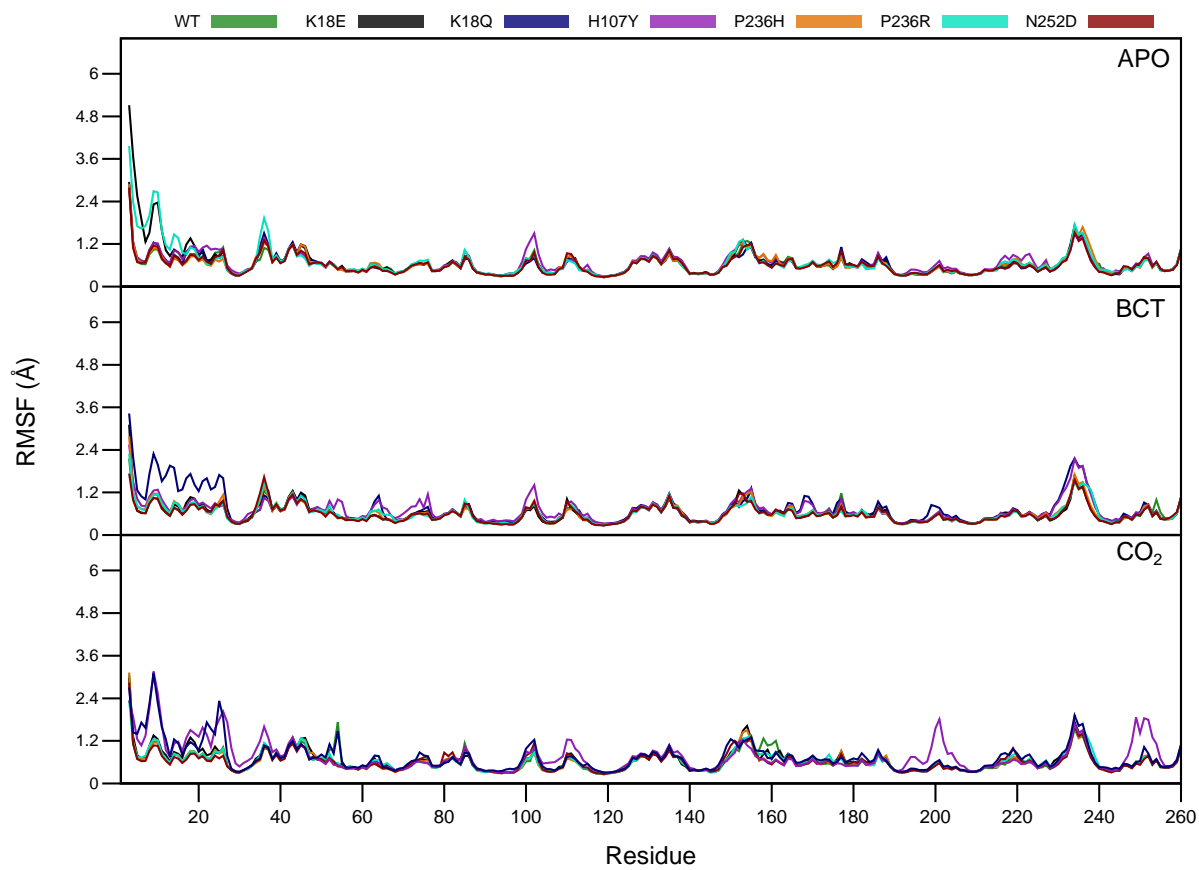


Figure S7. RMSF of the apo, BCT and CO₂ bound CA-II protein residues over the 200 ns MD simulation.

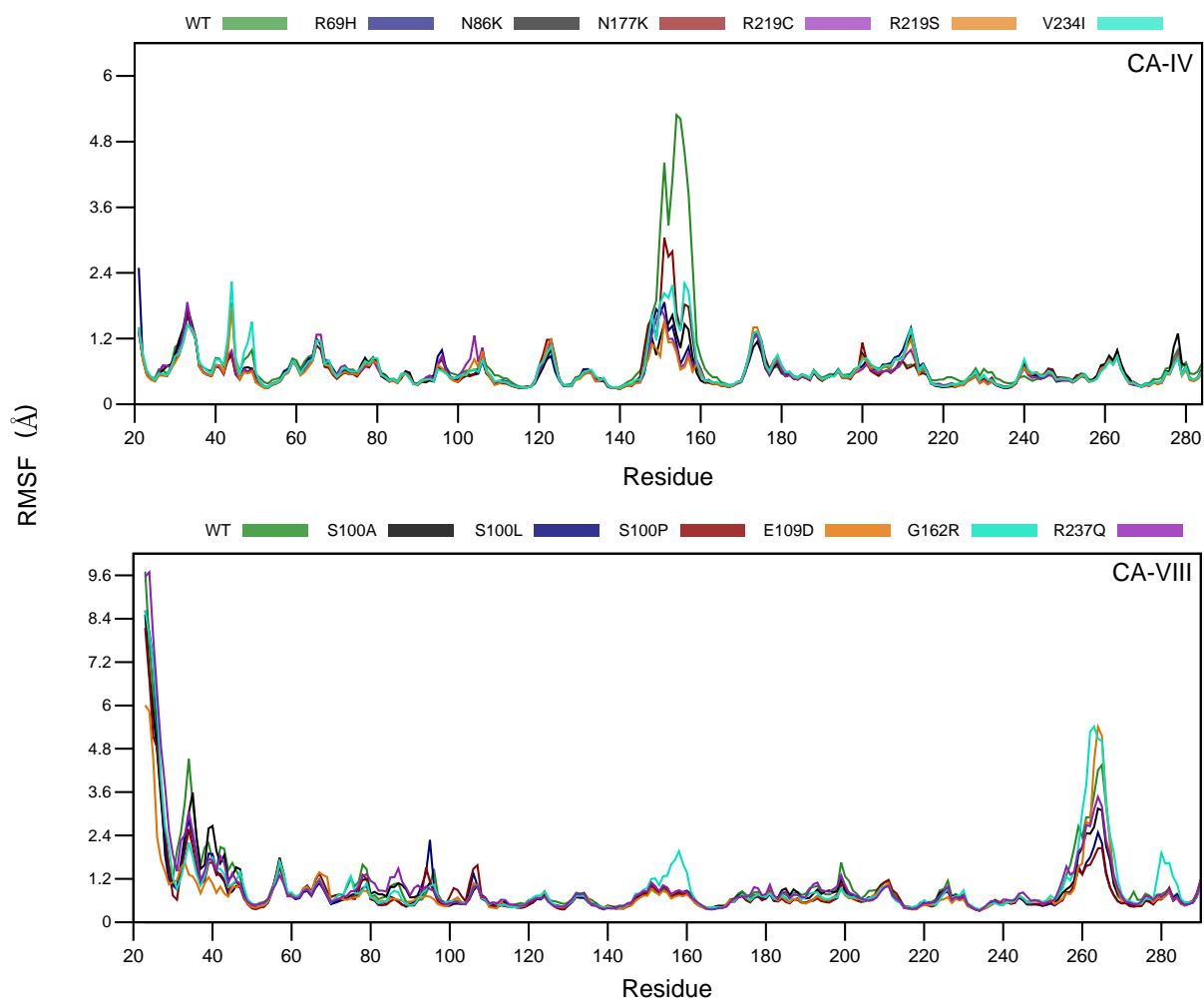


Figure S8. RMSF of the CA-IV and CA-VIII protein residues over the 200 ns MD simulation.

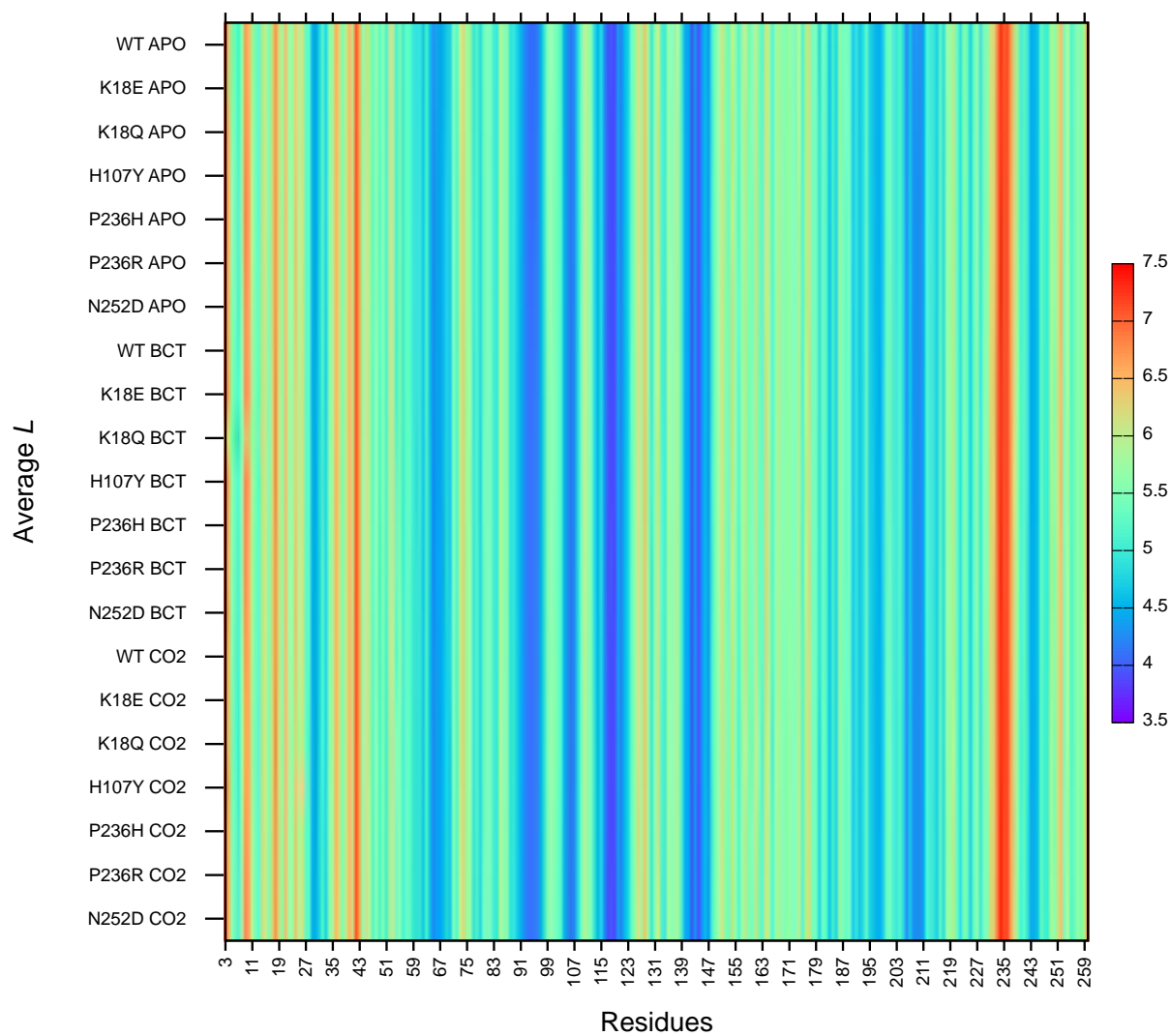


Figure S9. Average L for the CA-II protein residues in the apo, BCT and CO₂ bound states.

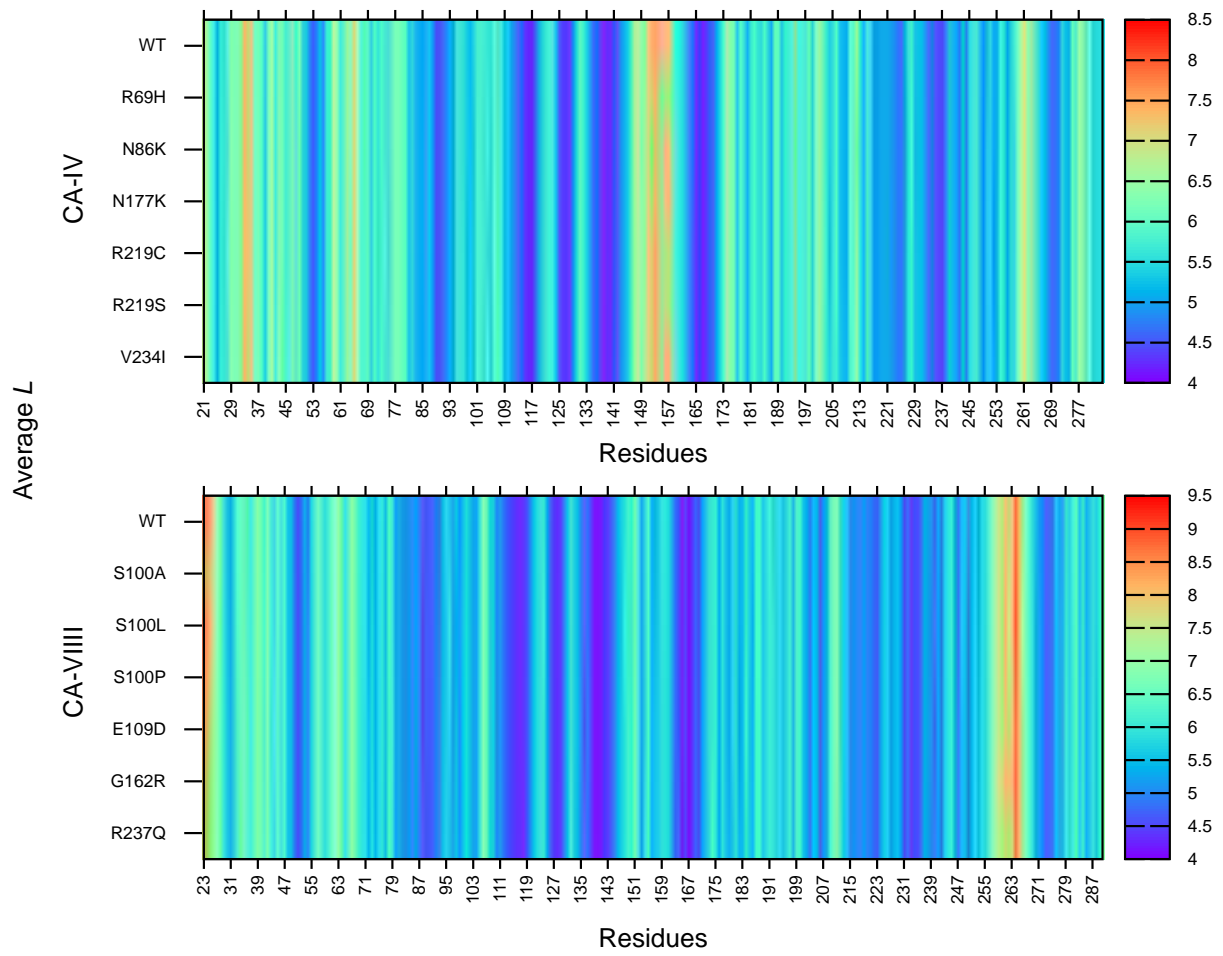


Figure S10. Average L for the CA-IV and CA-VIII over the MD simulation.

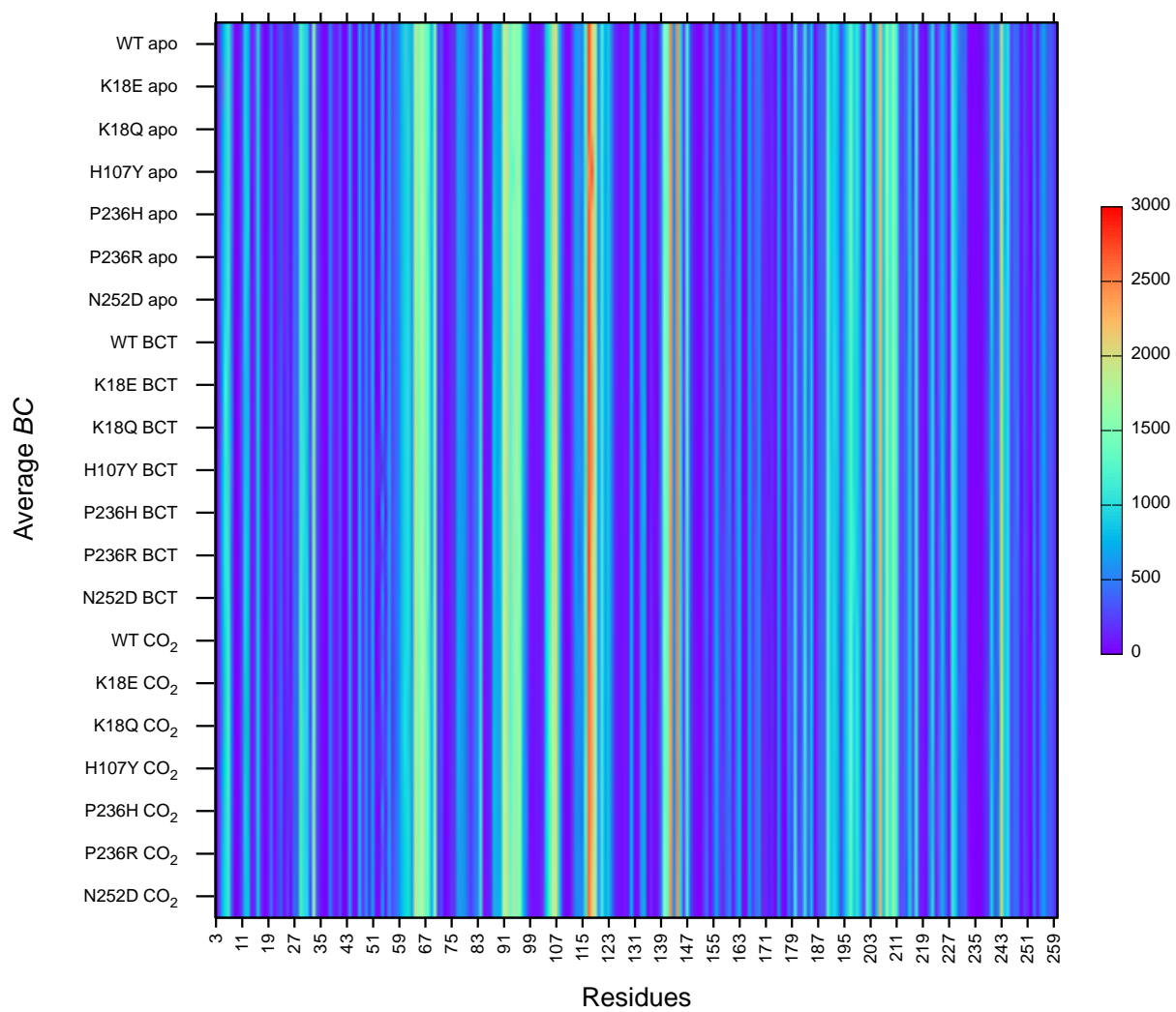


Figure S11. Average *BC* for the CA-II protein residues in the apo, BCT and CO₂ bound states.

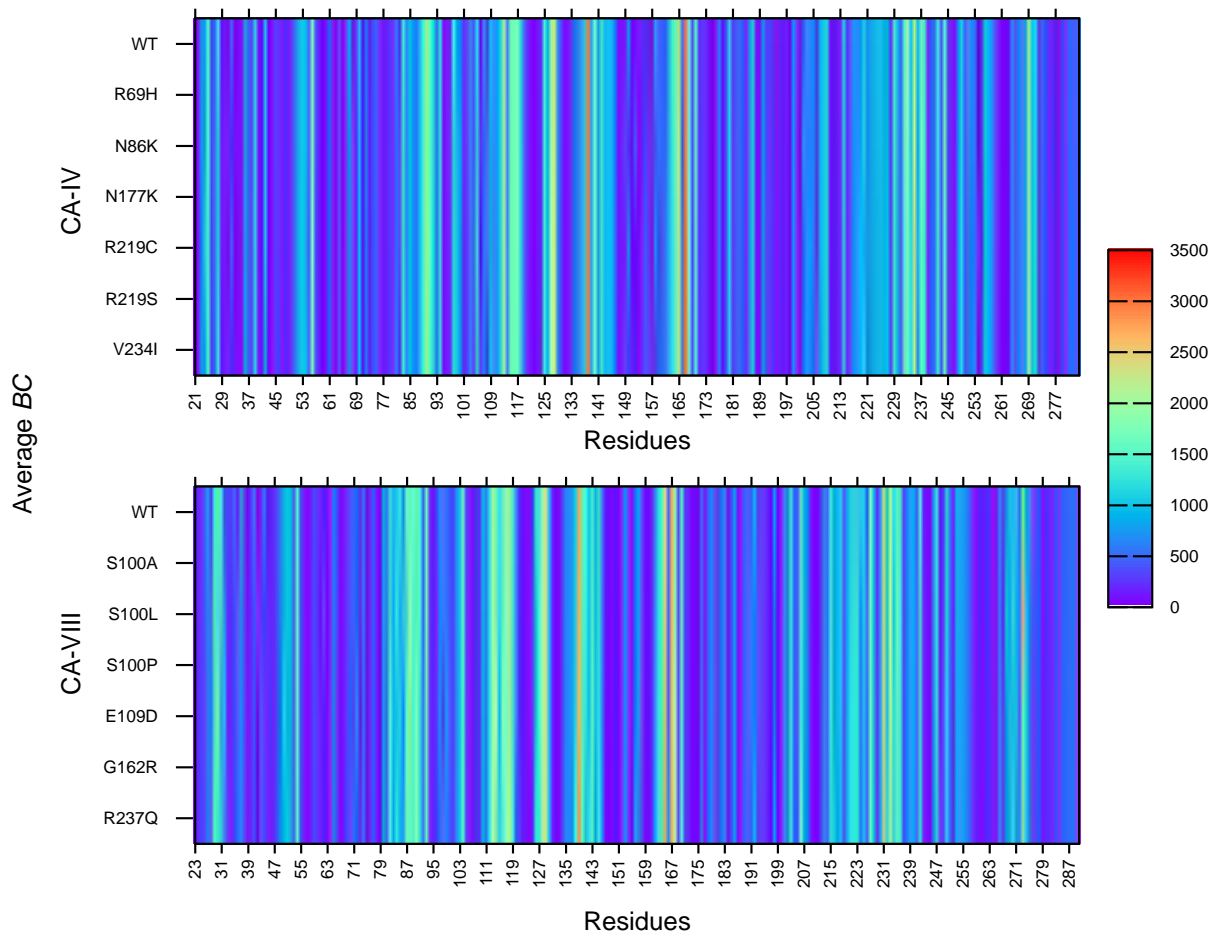


Figure S12. Average *BC* for the CA-IV and CA-VIII over the MD simulation.