JOURNAL OF TROPICAL LIFE SCIENCE
2024, Vol. 14, No. 2, 331 – 348
http://dx.doi.org/10.11594/jtls.14.02.13

Research Article

Three-Dimensional Structure of Human Epididymis Protein 4 (HE4): A Protein Modelling of an Ovarian Cancer Biomarker Through *In Silico* Approach

Nur Nadiah Abdul Rashid [1, 2, 3], Mohd Hamzah Mohd Nasir [2, 4], Nurasyikin Hamzah [2, 5], Che Muhammad Khairul Hisyam Ismail [2, 4], Siti Aishah Sufira Nor Hishamuddin [2, 4], Izzat Fahimuddin Mohamed Suffian [1], Azzmer Azzar Abdul Hamid [2, 4]*

[1] Department of Pharmaceutical Technology, Kulliyyah of Pharmacy, International Islamic University Malaysia, Bandar Indera Mahkota, 25200, Kuantan, Pahang, Malaysia
[2] Research Unit for Bioinformatics and Computational Biology (RUBIC), Kulliyyah of Science, International Islamic University Malaysia, Bandar Indera Mahkota, 25200, Kuantan, Pahang, Malaysia
[3] Department of Chemistry, Centre for Foundation Studies, International Islamic University Malaysia, 26300, Gambang, Pahang, Malaysia
[4] Department of Biotechnology, Kulliyyah of Science, International Islamic University Malaysia, Bandar Indera Mahkota, 25200, Kuantan, Pahang, Malaysia
[5] Department of Chemistry, Kulliyyah of Science, International Islamic University Malaysia, Bandar Indera Mahkota, 25200, Kuantan, Pahang, Malaysia

Article history:
Submission December 2023
Revised March 2024
Accepted March 2024

**Corresponding author:*
E-mail: azzmer@iium.edu.my

ABSTRACT

The Human Epididymis Protein 4 (HE4) biomarker has been extensively investigated for its potential in diagnosing ovarian cancer (OC). For the application of diagnostic techniques and drug delivery, it is crucial to understand the protein tertiary structure. However, the Protein Data Bank (PDB) does not currently contain the three-dimensional (3D) structure of HE4. Therefore, an *in silico* analysis was conducted to model the HE4 protein using AlphaFold, I-TASSER, and Robetta servers, with the sequence retrieved from UniProt (ID: Q14508). These three servers employed deep learning algorithms, threading templates, and *de novo* methods, respectively. Subsequently, Molecular Dynamics (MD) simulation using the GROMACS software package improved each 3D structure model, resulting in optimised and refined structures: RF1, RF2, and RF3. PROCHECK and ERRAT programmes were employed to assess the structure quality. The Ramachandran plots from PROCHECK indicated that 100% of residues were within the allowed regions for all servers except for I-TASSER. For the refined structures, RF1 and RF3, all residues were concentrated within the allowed regions. According to the ERRAT programme, the RF1 model exhibited the highest overall quality factor of 97.701, followed by RF3 and AlphaFold models with scores of 94.643 and 93.750, respectively. After these validations, RF1 emerged as the most accurately predicted 3D structure of HE4 and has one tunnel identified by CAVER 3.0 tool that facilitates the transportation of small particles to the active site, supported by FTsite and PrankWeb binding site predictions. This model holds potential for various computational studies, including the development of OC diagnostic kits. It will enhance our comprehension of the interactions between the protein and other biomolecules.

Keywords: AlphaFold, De novo, Human Epididymis Protein 4 (HE4), Ramachandran plot, Threading template

Introduction

In 2022, ovarian cancer (OC) was the fourth most common cancer among women in Malaysia, with 1,838 incidences and 1,167 deaths [1]. In the same year, 324,603 new cases were detected

How to cite:
Rashid NNA, Nasir MHM, Hamzah N, et al. (2024) Three-dimensional structure of Human Epididymis Protein 4 (HE4): A Protein Modelling of an Ovarian Cancer Biomarker Through *In Silico* Approach. Journal of Tropical Life Science 14 (2): 331 – 348. doi: 10.11594/jtls.14.02.13.

globally, with an estimation of 206,956 deaths, making OC one of the leading causes of mortality among gynaecological malignancies [2]. Even with advanced therapies and treatments, the overall 5-year survival rate among OC patients is still poor, reportedly less than 40% [3]. It was disclosed that the 5-year survival rates for patients diagnosed at early stages (I and II) and advanced stages (III and IV) are approximately 90% and 20%, respectively [4]. This shows early detection of the disease improves the long-term survival of the patients [5].

Human epididymis protein 4 (HE4), also recognised as WAP four disulphide core domain protein 2 (WFDC2), is a promising OC biomarker [6]. It is a secretory protein that belongs to the family of whey acidic protein domains and was discovered to be overexpressed in ovarian carcinomas in 1999 [7]. For decades, cancer antigen 125 (CA125) has been widely utilised in the screening of ovarian cancer, but new findings suggested that the combination of both biomarkers improves the diagnostic efficiency [5, 8–10]. HE4 is a glycoprotein and one of the challenges in obtaining the crystal structure of glycoproteins is the difficulty in purifying the protein, resulting from its nature of complex structure [11]. Due to this, the crystal structure of the HE4 protein is not available in the Protein Data Bank (PDB) database and elsewhere. Therefore, to develop a screening technique or other application that uses the *in silico* approach, it is crucial to determine the protein's three-dimensional (3D) structure.

Recently, research related to drug discovery, drug delivery and development of diagnostic techniques has seen significant utilisation of computational or *in silico* methods, which are more cost- and time-effective [12]. Furthermore, determining the tertiary structure of proteins is essential for many aspects of biomedical and biotechnology applications, as the structure ensures the protein's stability and functionality [13]. Most of the protein structures are precisely determined by using X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, but the 3D protein structures can also be predicted by the utilisation of different tools and software, such as AlphaFold, Iterative Threading Assembly Refinement (I-TASSER), and Robetta web servers [14, 15]. The development of the *in silico* approach to infer 3D protein structures from the primary protein sequences broadens the knowledge of the protein's physical interactions, stability, and potential [16].

The AlphaFold program is a revolutionary machine-learning technique that uses a deep learning algorithm enhanced by physical and biological knowledge about protein structures [17]. On the other hand, the precision of the I-TASSER server predictions is determined by the C-score, which is a scoring function based on the consensus significance score and the relative clustering structure density of various threading templates that are already established in the PDB database [18]. Another protein structure prediction server, Robetta, is a fully automated program that implements the combination of template-based and *de novo* methods and it covers every residue of the sequence provided by the user [19]. In the 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP14), the structures predicted by AlphaFold were tremendously more accurate than other competing teams. It was recorded that the median backbone accuracy of the AlphaFold structures was 0.96 Å RMSD (root-mean-square-deviation), while the closest competitor method resulted in 2.8 Å RMSD [17]. Table 1 shows the differences between the AlphaFold, I-TASSER and Robetta servers.

In 2021, the human proteome 3D models were successfully predicted by AlphaFold with high accuracy, which displayed satisfactory backbone prediction and the side chains were precisely oriented [23]. The protein structure validation should take place after the structure prediction, as it is a crucial step in evaluating the reliability and accuracy of the structures. The accuracy of the tertiary structures depends on a few factors, including the usage of different prediction tools. Consequently, the rapid development of protein structure validation tools can be seen in the recent decade, such as MolProbity, PROCHECK, ERRAT, and Verify3D [24–26]. PROCHECK and ERRAT are two widely used programs that evaluate the quality of protein 3D structures, and they are available on the SAVeS web server (https://saves.mbi.ucla.edu/). A two-dimensional (2D) plot called the Ramachandran plot, obtained from the PROCHECK programme, reveals the φ (phi) and ψ (psi) torsional angles of the amino acids in a protein sequence, determining the folding of a protein structure. It establishes the allowed and disallowed regions of conformational space [27].

The ERRAT program assesses the quality of

Table 1. The differences between the AlphaFold, I-TASSER and Robetta prediction servers

	Developer	Method	No. of amino acid allowed	Reference
AlphaFold	DeepMind Technologies (Alphabet Inc.)	https://github.com/deepmind/alphafold Deep learning algorithm: Utilises neural network topologies based on geometric and physical limitations of the protein structures	1000	[17]
I-TASSER	Zhang Lab (University of Michigan)	https://zhanggroup.org/I-TASSER/ 1. Retrieve protein templates with alike folds from the PDB database 2. *Ab-initio* modelling, rebuild the fragments into full-length 3D models	1500	[18, 20]
Robetta	Baker Lab (University of Washington)	https://robetta.bakerlab.org/ Combination of comparative modelling (based on PDB100 template database and *de novo* method)	1201	[21, 22]

the protein 3D structures by calculating overall quality factor scores, which analyses the interactions of non-bonded atoms in the amino acids of the protein [28]. A score of more than 50 is considered a good quality structure, but a higher score indicates a higher quality structure [29]. It is important to identify a high-quality predicted model of the protein to provide a complete understanding of its physical, chemical and biological interactions with various molecules. Hence, this study aims to apply the *in silico* technique to determine the best predicted 3D structure of HE4 protein for numerous applications in the future, including the development of the OC screening technique.

Material and Methods

Protein sequence search

The primary sequence of HE4 is required before its tertiary structure modelling, which was retrieved from UniProt (https://www.uniprot.org/) using 'HE4' as a keyword. The full amino acid sequence for the human type of HE4 protein was selected and retrieved. Based on the protein sequence, the ProtParam tool (https://web.expasy.org/protparam/) was used to determine many parameters, including the molecular weight, amino acid compositions and atomic components [30].

Protein tertiary structure modelling by protein structure prediction servers

The primary sequence of HE4 protein containing 124 amino acids was applied in three different protein structure prediction web servers: AlphaFold (https://alphafold.ebi.ac.uk/), I-TASSER (https://zhanggroup.org/I-TASSER/), and Robetta (https://robetta.bakerlab.org/). The FASTA format of the protein sequence was downloaded from UniProt and used as the sequence input for each program. The default settings and parameter were retained, and the process was initiated by the sequence submission. The confidence score (C-score), predicted TM (pTM), and predicted local distance difference test (pLDDT) of the model were recorded. Good quality protein structure has a higher than 0.5 C-score, high pTM score (-5 to 2), and high pLDDT score (0 to 100) [31–33]. Each predicted structure was visualised using PyMOL (version 2.4.1) molecular viewer [34].

Molecular dynamic simulations of HE4 tertiary models

Molecular dynamic (MD) is the most in-demand computational method in analysing the equilibrium structures and dynamic interactions of biological systems. Prior to the MD simulations, the best 3D model predicted by AlphaFold, I-TASSER, and Robetta was downloaded in pdb files. The MD simulation of each model was conducted using GROMACS 5.1 software package, separately, with triplicate for each system [35]. The proteins were simulated within a virtual cubic box with 1.0 nm distance between the protein and

the box faces, at constant pressure and temperature of 1 atm and 310 K, respectively. The water molecules were added in the solvation step using simple point charge (SPC216), followed by the neutralisation of the protein by using sodium ions, Na^+ and chloride ions, Cl^-. The simulations were conducted using the OPLS forcefield for 100 ns. All MD simulations were analysed for their root-mean-square deviation (RMSD) and radius of gyration (Rg). The most stable HE4 conformation, which was the middle structure of the top cluster after the clustering step of each simulation, was extracted as pdb files: RF1, RF2 and RF3, from the MD simulations of AlphaFold, I-TASSER and Robetta, respectively. The 3D structures of RF1, RF2 and RF3 were viewed on PyMOL (version 2.4.1) molecular viewer and their secondary structure elements were analysed by PDBsum (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/) [27].

Structure assessment of HE4 protein structure

Two protein structure validation programs, PROCHECK and ERRAT (https://saves.mbi.ucla.edu/) were used to evaluate the quality [28] of the best-predicted models by AlphaFold, I-TASSER, Robetta, and the middle structures extracted from the top cluster after the MD simulations of AlphaFold, I-TASSER and Robetta systems: RF1, RF2 and RF3. The pdb files were submitted to the validation programme server and evaluated. Ramachandran plots, ERRAT plots, and overall-quality-factor values presented information on the backbone conformation and residue interactions. The best HE4 model was scanned with CAVER 3.0 tool (https://loschmidt.chemi.muni.cz/caverweb/) to identify and characterise the transport pathways or the tunnels of static protein structures, where the tunnels aid the transportation of small molecules to the active site of the protein. The binding site residues of CAVER 3.0 were compared to the residues predicted by other binding site prediction servers: FTsite (https://ftsite.bu.edu/) and PrankWeb (https://-prankweb.cz/).

Results and Discussion

HE4 primary sequence

The search on the UniProt website resulted in the protein sequence with UniProt ID Q14508, with a length of 124 amino acids and a mass of 12993 Da. The molecular formula is $C_{541}H_{874}N_{154}O_{178}S_{19}$, consisting of 1766 atoms with a primary sequence of:

MPACRLGPLAAALLLSLLLFGFTLVSGTGA-EKTGVCPELQADQNCTQECVSDSECADNL-KCCSAGCATFCSLPNDKEGSCPQVNINFPQ-LGLCRDQCQVDSQCPGQMKCCRNGCGKV-SCVTPNF

Based on the ProtParam analysis, the most common amino acid present in the sequence is cysteine (13.7%), followed by leucine (11.3%). 15.3% of the full sequence are amino acids with electrically charged side chains: arginine, lysine, aspartic acid, and glutamic acid.

HE4 3D models by prediction servers

The tertiary structure predictions were successfully performed by AlphaFold, I-TASSER, and Robetta (Table 2), and each resulted in five predicted 3D models. AlphaFold utilised a template-free prediction approach while I-TASSER generated the model from various threading alignments based on the PDB library. The I-TASSER model was developed from the 1zlgA, 1udkA, 7mn5B, 1udk, and 7fdeP templates in the PDB. The full-length structure was then modelled based on the combination of the fragments from the available templates. The accuracy of the overall topology is determined by the predicted TM (pTM) score, where the models with a value higher than 0.5 are considered to have a highly similar fold with related proteins [31]. The pTM-score of the best model of I-TASSER was 0.35 ± 0.12, with a C-score of -3.31. The confidence score or C-score, with a range of -5 to 2, evaluates the structure quality based on the alignments of the threading templates [32]. A higher C-score value defines a higher confidence model. Predicted local distance difference test (pLDDT) scores range from 0 to 100, with 100 as the highest confidence of a predicted model as it resembles the true structure of a protein [33]. AlphaFold predicted five models, and the best model was model 3, with pLDDT and pTM scores of 82.1 and 0.635, respectively. For Robetta, the protein folding prediction of implemented the combination of template-based and *de novo*, template-free approach. The confidence score for the best structure modelled by Robetta was 0.72, considered a good model. The best-predicted models were viewed using the PyMOL molecular visualisation viewer, where the alpha helix and beta sheets can be observed clearly (Table 2).

Table 2. Predicted models retrieved from AlphaFold, I-TASSER, and Robetta, and their scores according to the scoring system

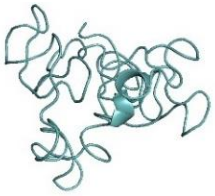
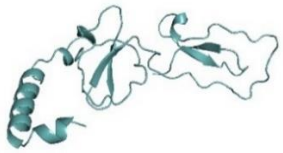
Prediction server	Scoring system	Predicted model	Scores	Best model
AlphaFold	a) pLDDT b) pTM-score	Model 1	pLDDT: 80.7 pTM-score: 0.614	 Model 3
		Model 2	pLDDT: 80.7 pTM-score: 0.611	
		Model 3	pLDDT: 82.1 pTM-score: 0.635	
		Model 4	pLDDT: 80.3 pTM-score: 0.567	
		Model 5	pLDDT: 73.0 pTM-score: 0.450	
I-TASSER	a) C-score b) pTM-score (provided only for model 1)	Model 1	C-score: -3.31 pTM-score: 0.35±0.12	 Model 1
		Model 2	C-score: -3.62	
		Model 3	C-score: -3.87	
		Model 4	C-score: -4.07	
		Model 5	C-score: -3.62	
Robetta	C-score	Model 1	0.72	 Model 1

Table 3. Average RMSD and standard deviation (in Å) of HE4 predicted model after 15 ns.

MD simulation	AlphaFold		I-TASSER		Robetta	
	Average RMSD	Standard deviation	Average RMSD	Standard deviation	Average RMSD	Standard deviation
1	12.427	0.2427	5.8880	0.2034	5.6747	0.3618
2	14.346	0.5646	6.5305	0.2967	6.7649	1.1630
3	14.103	0.4789	6.4497	0.2759	8.4413	0.7639

MD simulations of HE4 predicted models

The refinement of the predicted tertiary structures was carried out by MD simulations to analyse the motions of the molecules and atoms over a specific time. This approach aims to adjust and adapt the moderately accurate protein models predicted by the prediction servers closer to their native state. The conformational differences between the protein backbones from their initial structural conformation until the end of the MD simulation are measured using the root mean square deviation (RMSD) [36]. Table 3 presents the average RMSD and standard deviation values of three MD simulations of each predicted model, in which the MD simulation 1 of each system was selected and further analysed. This is based on the lower standard deviation value compared to MD simulations 2 and 3, indicating less fluctuations and higher similarity of conformations, resulting in greater stability [37]. A study on MD simulation of hydrogenated amorphous carbon has shown greater stability and reduced fluctuation, with lower standard deviation [38].

The RMSD value for the AlphaFold protein fluctuated significantly at the beginning of the simulation course and equilibrated at 12 Å after 8 ns (Figure 1d). The RMSD for I-TASSER and Robetta reached equilibrium at approximately 15 ns with a lower RMSD value (6 Å) than AlphaFold. Relevantly, low RMSD values, ideally ~2 Å, during the simulation course indicate higher stability of conformation [39]. However, for certain molecules, higher and more significant fluctuation of RMSD values is expected, describing that

the degree of deviation of the trajectory is higher, where the conformations differ significantly from its initial conformation. In MD simulations, equilibration is an important phase which involves the protein reaching a stable and consistent state. During this phase, the system adjusts to the simulation conditions, settles into its minimum energy, and achieves a balanced distribution of velocities and positions for its components, such as atoms [40]. Due to these, large fluctuations are expected at the beginning of the MD simulations for most molecules until the protein transitions to a more stable conformation.

Even though the RMSD fluctuations occurred at the beginning of the simulations for all three systems, the AlphaFold system shows a higher degree of deviation of the trajectory where the RMSD value fluctuated extensively between 5 ns to 8 ns (Figure 1d). This is affected by notable structural changes between its initial conformation and its stable conformations after equilibrium has been reached. The equilibration phase was achieved after 8 ns until the end of the simulation, when the flattening of the RMSD curve deduced that stable conformations were achieved. Table 3 shows that the average RMSD of AlphaFold is significantly higher compared to I-TASSER and Robetta due to its cylindrical-shaped initial

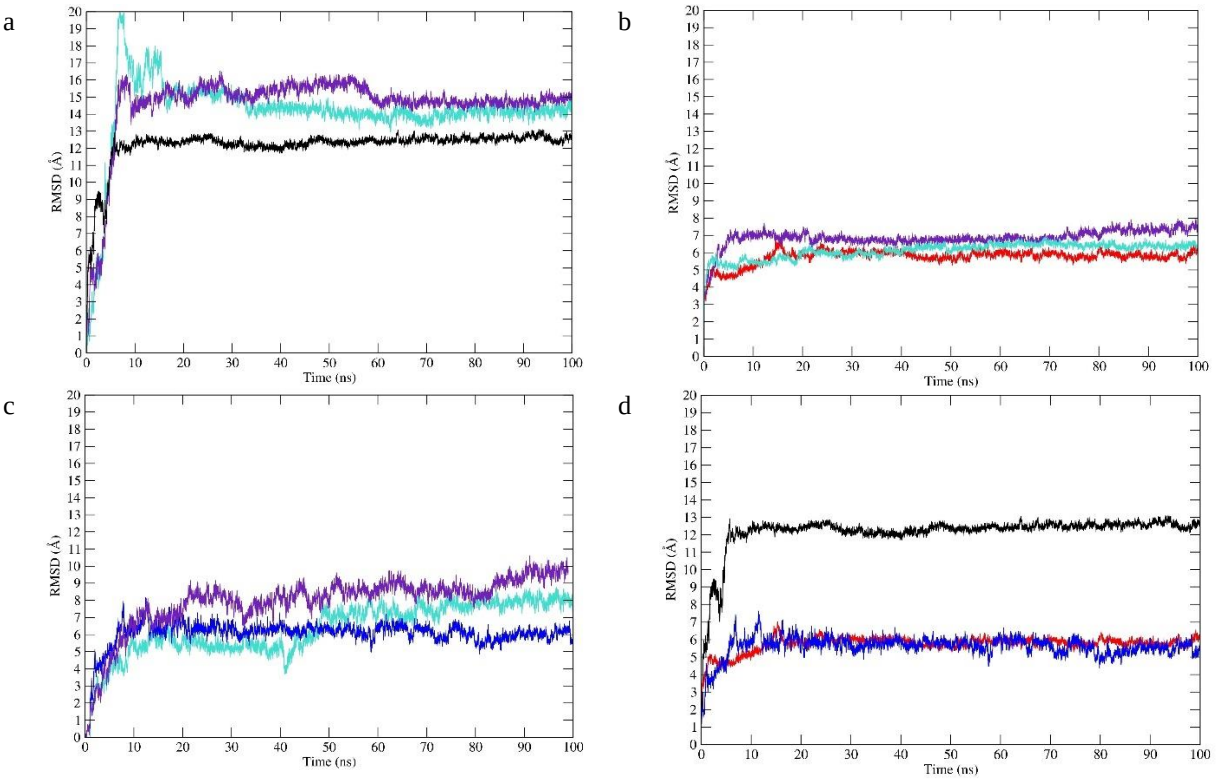


Figure 1. The RMSD of (a) AlphaFold, (b) I-TASSER, and (c) Robetta triplicates during 100 ns of MD simulations using OPLS force field. Turquoise and purple lines represent the second and third MD simulations, respectively. (d) illustrates the RMSD of the selected replicate, which is MD simulation 1, for AlphaFold (black), I-TASSER (red), and Robetta (blue).

Table 4. Clustering analysis after MD simulations of the AlphaFold, I-TASSER and Robetta predicted HE4 protein models.

Clustering analysis	AlphaFold	I-TASSER	Robetta
Total number of structures	9001	9001	9001
Number of clusters	353	150	542
Number of structures in the top cluster	429	1062	150
Middle structure of top cluster	RF1 At 83770 (83.77 ns)	RF2 At 84040 (84.04 ns)	RF3 At 48930 (48.93 ns)

conformation (Table 2), requiring greater conformational changes for the protein to fold into a compact protein structure. Each MD simulation was followed by a clustering step, which serves the purpose of identifying similar structures and reducing the complexity of the post-simulation analysis [38,39]. This clustering step determines the most dominant conformations within the simulation ensemble while highlighting the essential structural motifs. It clusters similar conformations together and simplifies analysis.

Based on the clustering analysis (Table 4), which applied Gromos clustering method with RMSD cut-off of 0.15 nm, 429 structures were found to be highly similar in conformations throughout the MD simulation of AlphaFold predicted model. Approximately 11.80 % (1062) of the 9001 I-TASSER conformations display a high degree of similarity. The refined HE4 model for the AlphaFold, I-TASSER and Robetta systems was derived from the middle structure of the top cluster and was denoted as RF1, RF2 and RF3, respectively (Figure 2). It is worth noting that the refined model represents the most dominant conformation.

For the structural conformations, the RF2 and RF3 models show slight changes compared to the I-TASSER and Robetta predicted models, respectively, while RF1 model has distinguished changes compared to its initial conformation which is the AlphaFold predicted model. These notable changes between AlphaFold and RF1 models (Figure 3) matched the RMSD fluctuations that occurred during the simulations, where the AlphaFold system encountered a high degree of deviation. The alterations in the structures of RF2 and RF3 from their initial conformations are less noticeable, but the arrangement of the structures still exists, which corresponds to the minor fluctuations of the RMSD curve.

The significant structural changes between the AlphaFold and RF1 models are obviously noticed at the helix motif from glycine-7 (GLY7) to threonine-23 (THR23) (Figure 4). As stable conformations were achieved, this region folded inwards, making more atom-atom interactions with neighbouring residues, which leads to an overall stable 3D structure. These interactions may occur through hydrogen bonds, disulphide bonds, Van der Waals forces, ionic and hydrophobic interactions [43–45]. These various means of interaction play crucial part in stabilising the protein's secondary and tertiary structures. The folding of protein structure is vital to ensure good functionality

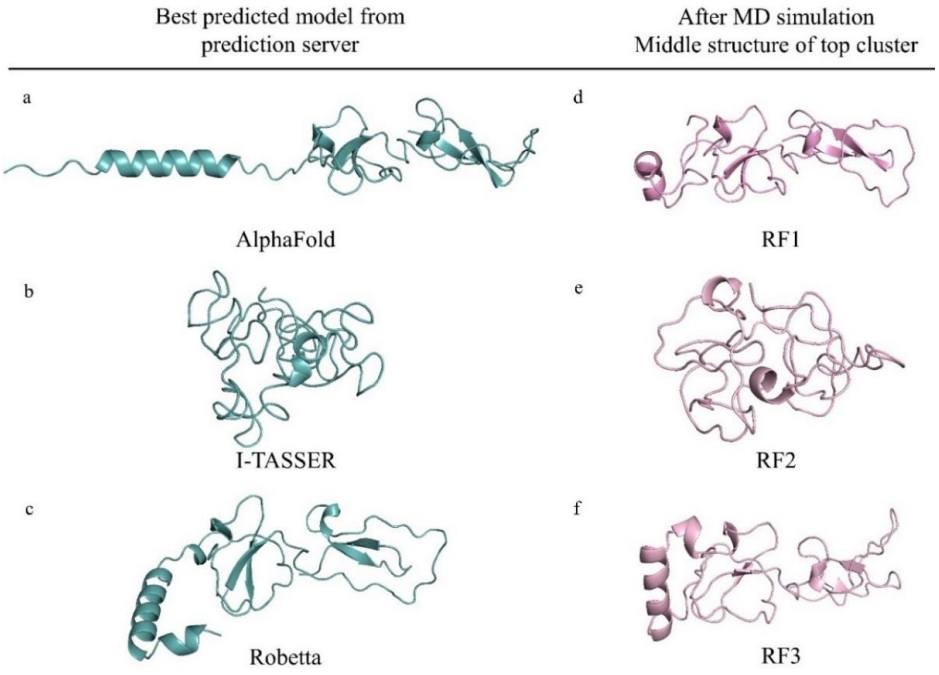


Figure 2. The 3D models of HE4 protein, including the best model predicted by (a) AlphaFold, (b) I-TASSER, and (c) Robetta), and the middle structures of the top cluster after the MD simulation of each predicted model: (d) RF1, (e) RF2, and (f) RF3. The models were viewed and extracted from PyMOL molecular viewer software.

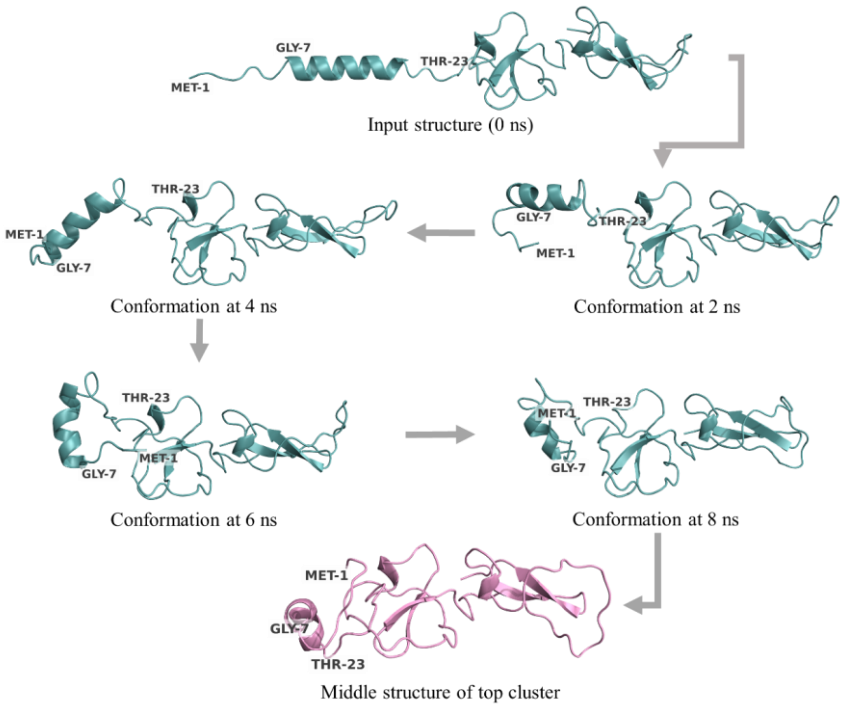


Figure 3. The progress of HE4 protein structure over MD simulation time for the AlphaFold system until the middle structure of the top cluster (RF1) was obtained. The RMSD values fluctuated significantly from 0 ns to 8 ns, showing significant structural changes, until it reaches equilibrium with flattened RMSD curve. The protein undergoes a conformational change towards a state of greater stability, forming dominant conformations.

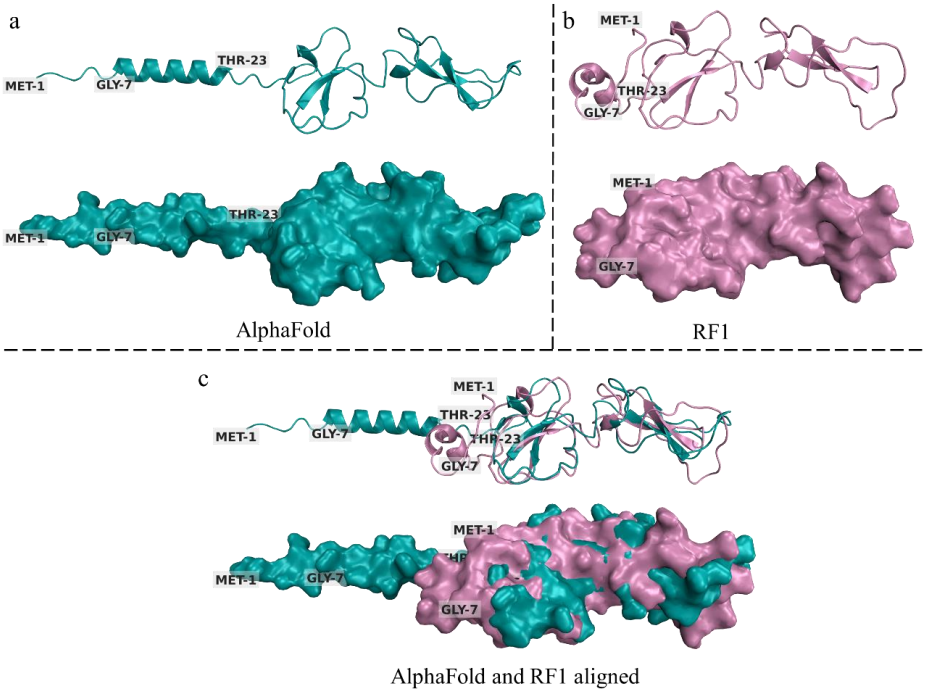


Figure 4. The tertiary structures of (a) HE4 modelled by AlphaFold that was used as the input for the MD simulation, (b) RF1, the middle structure of the top cluster after the MD simulation of the AlphaFold model using OPLS force field for 100 ns, and (c) the alignment of (a) and (b), showing the structural differences between the two HE4 models. The significant changes observed from (a) to (b) corresponded to the notable fluctuations in the RMSD graph (Figure 1) during the MD simulation. The 3D structures are shown in cartoon and surface viewing modes.

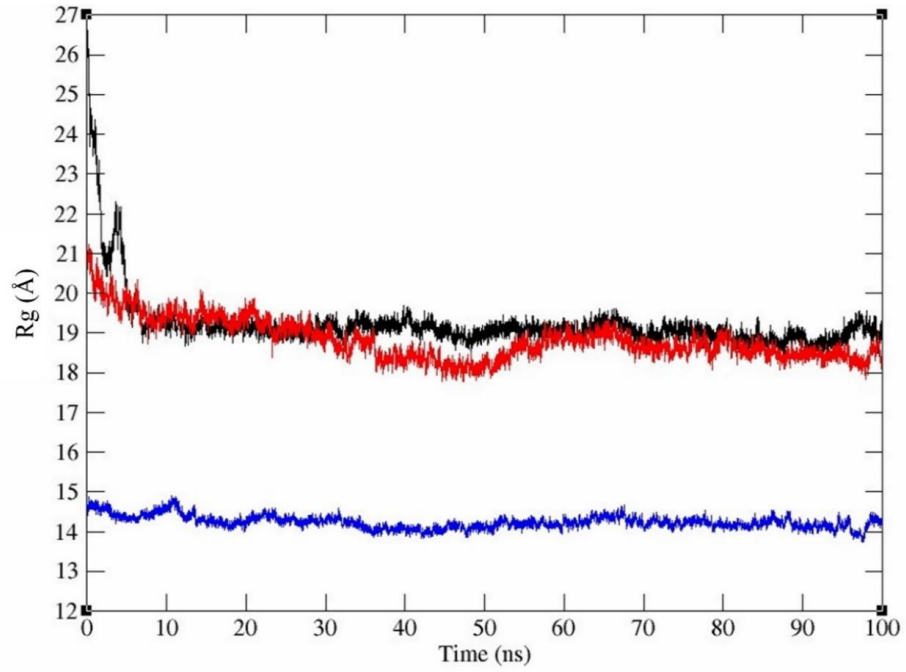


Figure 5. The radius of gyration of AlphaFold (black), I-TASSER (red), and Robetta (blue) 3D models during 100 ns of MD simulations using OPLS force field.

and stability, as properly folded protein structure is less prone to denaturation.

The stability and compactness of a protein structure can be determined by the radius of gyration (Rg), apart from the RMSD curve. Rg describes a high rigidity and compactness of the protein structure by achieving steady values during the simulation [46]. As depicted in Figure 5, the Rg graph of the AlphaFold system fluctuated significantly before stabilising after 8 ns at approximately 19 Å. Similar to the RMSD, this is due to the notable structural differences between its initial and stable conformation, RF1, where the initial conformation was elongated and cylindrical-like before it was folded compactly. For the I-TASSER and Robetta systems, no prominent Rg fluctuations occurred. They maintained Rg values of approximately 19Å and 14.5 Å, respectively. From the Rg graph, it can be deduced that all three systems maintained relatively steady values after 8 ns, which means that they were compact and spent most of the simulation time as stable folded structures.

The RF1, RF2 and RF3 models were further analysed using PBDsum, where the wiring diagrams of the secondary structure were obtained (Figure 6). This diagram is a helpful tool for understanding the interaction between the various components of the protein and it visualises the overall protein structure by showing the helices, strands, and different motifs such as the beta-turn and beta-hairpin [27]. The RF1, RF2 and RF3 models presented common structural motifs of native proteins such as alpha-helices and beta-strands. The RF1 protein model reveals three helices (H1, H2, and H3) and multiple strands from different beta sheets (A, B, and C), while the RF3 model displays more helices between amino acids 1 to 124 with a lesser number of strands. All three wiring diagrams contained disulphide bonds, which are represented by the yellow-linking bars. As for RF2, the protein structure lacked helices and strands, making it a less preferred HE4 model compared to RF1 and RF3.

The alpha-helices are essential structural components in proteins as their main role is to maintain the protein's stability and shape [47]. They allow for efficient packing of the polypeptide chain, enabling the protein to achieve a more compact structure while maximising interactions between the amino acids. The beta-strand serves the same purpose of maintaining the protein's structural stability by forming hydrogen bonds with adjacent strands, resulting in the formation of beta sheets. Based on the presence of multiple helices and strands, the protein structures of RF1 and RF3 highly resembled the common secondary structure elements of native proteins.

HE4 structure validations

The most dominant HE4 conformation for the AlphaFold, I-TASSER and Robetta systems: RF1, RF2 and RF3, were verified using PROCHECK and ERRAT, on the SAVeS server. The PROCHECK programme generated a Ramachandran

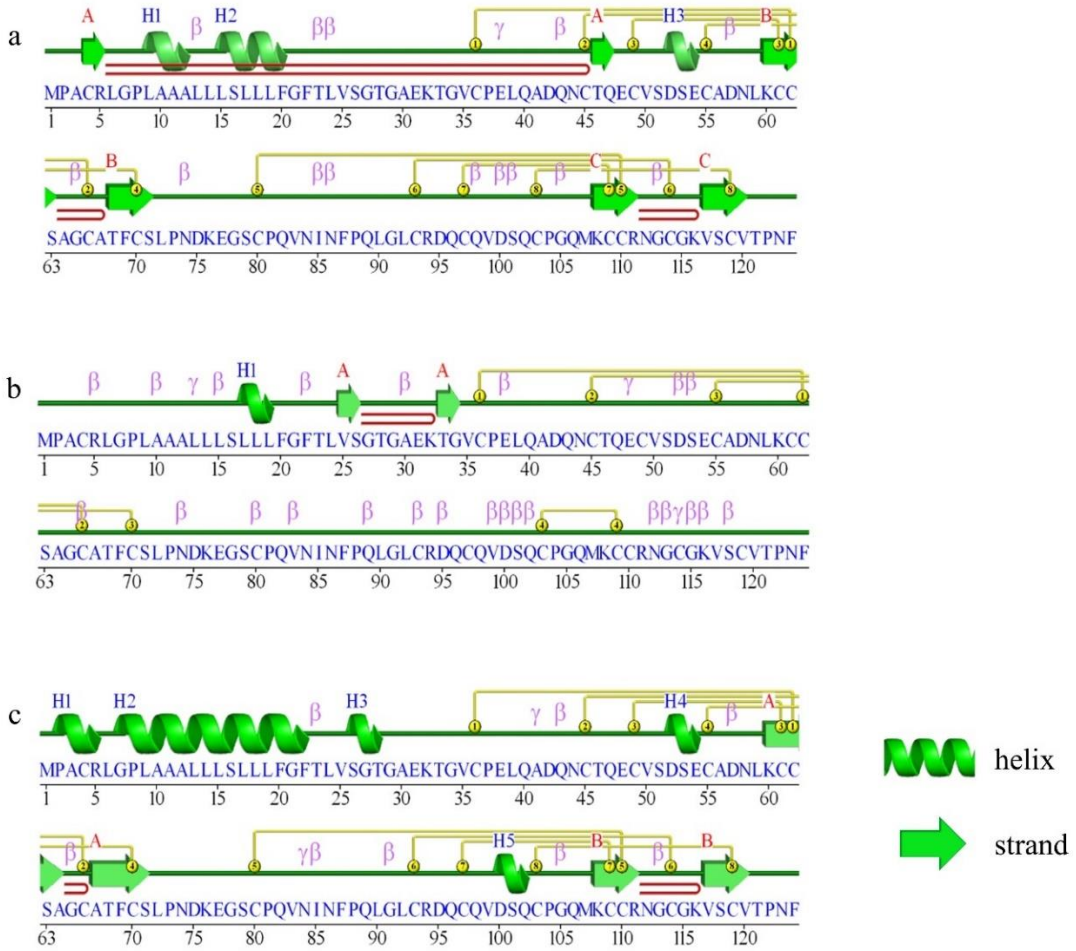


Figure 6. The schematic wiring diagram of the HE4 representing the secondary structure elements: alpha-helices and beta-sheets, obtained from the PDBsum analysis; (a) RF1; (b) RF2; (c) RF3. The green springs represent the helices, while the green arrows are the strands. The helices are marked as H1, H2, H3, and H4, and the strands are labelled according to which beta sheet they belong to (A, B, C). These diagrams were retrieved from PDBsum.

Table 5. The percentage of HE4 residues in different regions of the Ramachandran plot

Region	Region colour	AlphaFold	RF1	I-TASSER	RF2	Robetta	RF3
Residues in most favoured regions (A, B, L)	Red	85.4%	87.4%	35.0%	65.0%	80.6%	82.5%
Residues in additional allowed regions (a, b, l, p)	Yellow	13.6%	12.6%	40.8%	28.2%	17.5%	17.5%
Residues in generously allowed regions (~a, ~b, ~l, ~p)	Pale yellow	1.0%	-	11.7%	1.9%	1.9%	-
Residues in disallowed regions	White	-	-	12.6%	4.9 %	-	-

plot that represents the φ (phi) and ψ (psi) torsional angles of the polypeptide backbone of the amino acids. Excluding glycine and proline, 100% of the residues of the AlphaFold and Robetta models, including their refined models, RF1 and RF3 were in the allowed regions, while for I-TASSER model and its refined model (RF2), there were 13 and five residues (glutamic acid-38, valine-35, cysteine-66, valine-99, and aspartic acid-100) found at the disallowed region, respectively (Figure 7 and Table 5). The disallowed regions are defined as where significant steric hindrance between the torsions occurs [27]. This region represents the combinations of φ and ψ angles that are less commonly

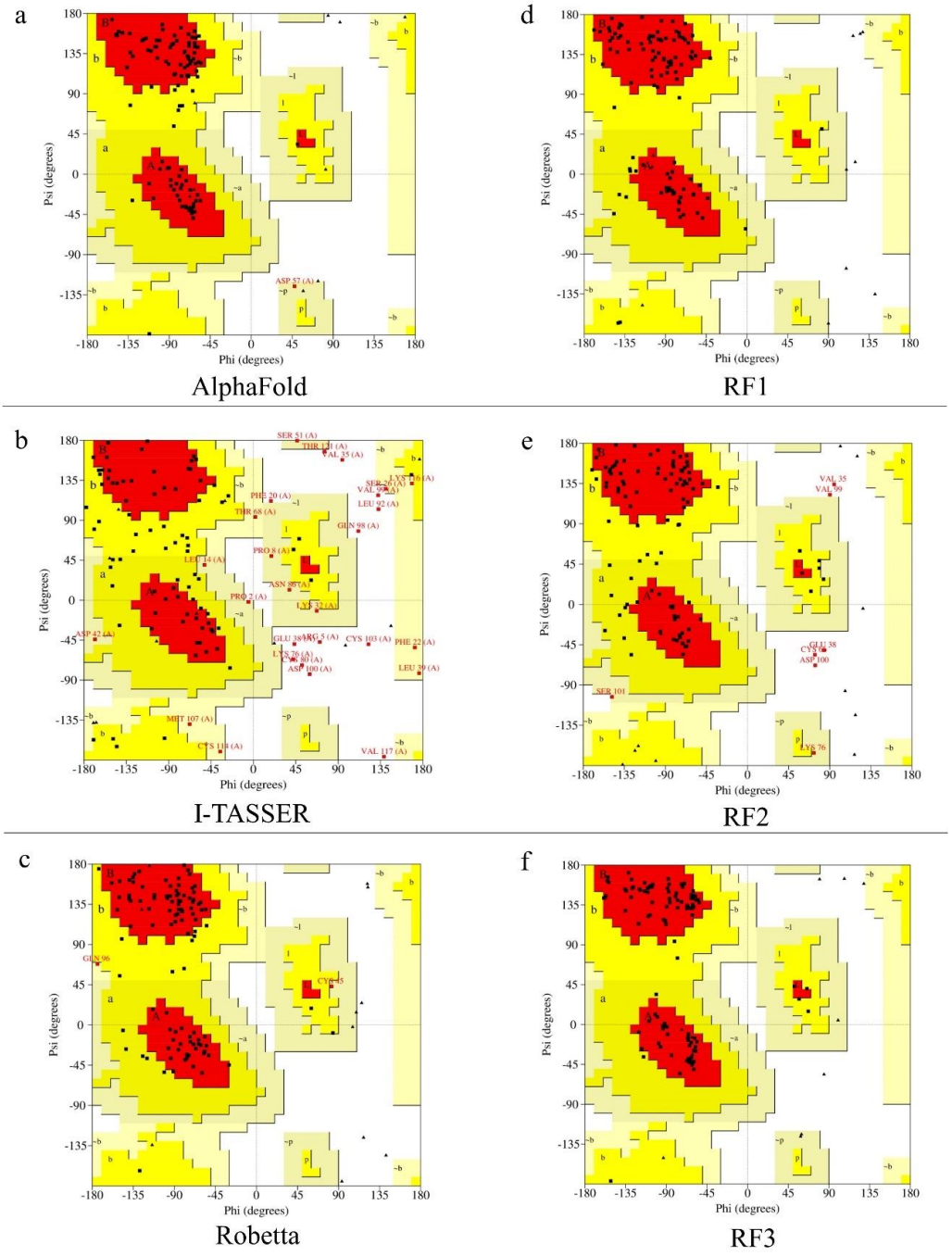


Figure 7. The Ramachandran plots of the HE4 protein models: (a) AlphaFold, (b) I-TASSER, (c) Robetta, (d) RF1, (e) RF2, and (f) RF3. The residues were plotted in different regions: most favoured (red), additional allowed (yellow), generously allowed (pale yellow), and disallowed region (white). The plots were obtained from the SAVeS web server (https://saves.mbi.ucla.edu/).

presented in well-folded native protein structures due to the strain they would generate within the protein's backbone. Glycine does not contain any side chain; hence, it is permissible to adopt the torsional angles in any of the quadrants of the Ramachandran plot. Ho and Brasseur clarified that the first quadrant (top-left) represents the beta-sheet region, the bottom-left quadrant is the right-handed alpha-helix, while the top-right quadrant is the left-handed alpha-helix region [48, 49].

The most dominant conformations (RF1, RF2 and RF3) of the MD simulations of the three systems, AlphaFold, I-TASSER and Robetta, were noticed to be more stable and accurate based on the distribution of the φ and ψ dihedral angles of the amino acids. This is shown in Table 3 where the percentages of the amino acids located in the most favoured regions were improved. Amino acid residues within the most favoured regions have stable backbone geometries without steric strains, while the residues in the additional and generously allowed regions are less favourable but are still allowed where steric clashes are possible due to specific interactions or local structural constraints. From the Ramachandran plot analysis, the RF1 protein model has the highest quality, with 87.4% of the non-glycine and non-proline residues located within the most favoured regions, and none of the residues were located at the generously allowed and disallowed regions.

Apart from the Ramachandran plot validation, ERRAT programme is utilised to verify the

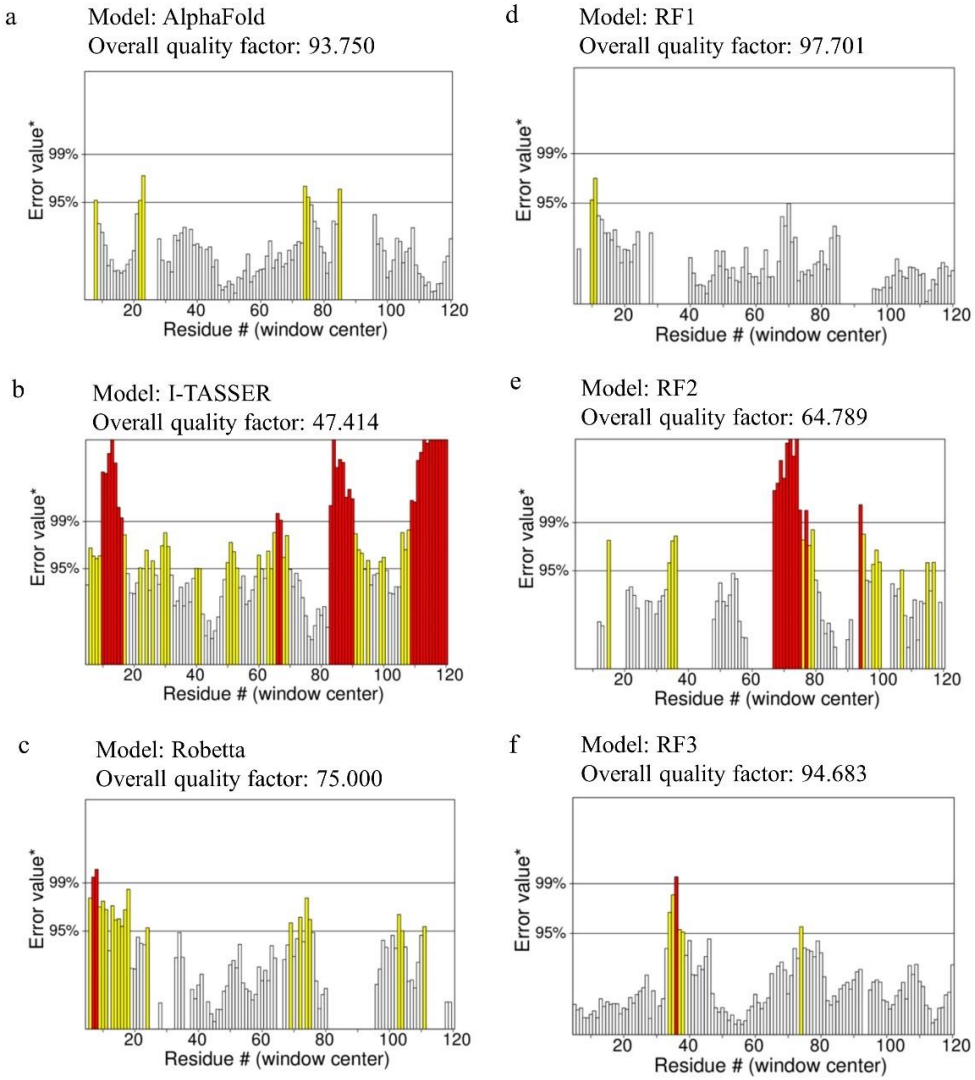


Figure 8. The ERRAT plots of the HE4 protein models; (a) AlphaFold, (b) I-TASSER, (c) Robetta, (d) RF1, (e) RF2, and (f) RF3. The red bars represent outlier residues where the region significantly deviate from typical, well-folded, native protein structures. The plots were obtained from the SAVeS web server.

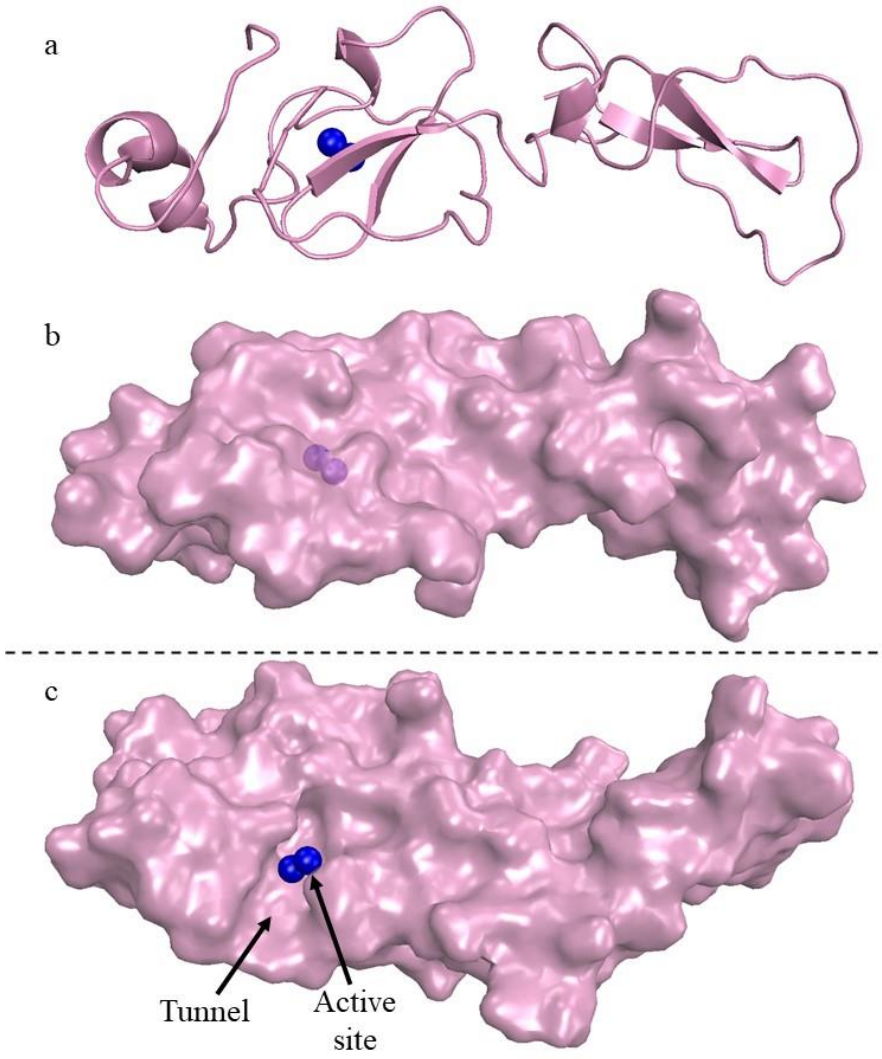


Figure 9. The RF1 structure, representing the best HE4 tertiary model obtained from this study, in (a) cartoon and (b) surface viewing modes. (c) Rotation of structure (b), showing the tunnel that facilitates the transportation of small molecules to the active site, identified by CAVER 3.0 tool. The region with blue dots indicates the predicted active site of the protein.

protein structures by generating overall-quality-factor scores, based on the interactions of the non-bonded atoms between the residues [28, 50]. It investigates the atomic interaction distribution inside the protein and compares it to statistical data acquired from high-resolution crystallographic structures. ERRAT is denoted by a value between 0 to 100, with 100 being the highest quality model. The evaluation by ERRAT showed that RF1 represents the highest quality of HE4 protein structure, with an overall quality factor score of 97.701, an improvement from the AlphaFold model with a score of 93.750. The I-TASSER model has the poorest quality according to ERRAT validation, with only a 47.414 overall quality factor score. An error value exceeding 99% represents a poorly modelled region [51], in which 29 residues of the I-TASSER HE4 model are plotted at this region (Figure 8b), resulting in the lowest score. Even after the MD simulation of the I-TASSER system, the overall quality factor scores of the most dominant conformation, RF2, were still poor (64.789) with 11 outliers (Figure 8e). These outliers, highlighted in red bars, indicate potential inaccuracies or regions of lower quality within the protein model that could impact its functionality. Based on the Ramachandran and ERRAT plots, both RF1 and RF3 models are of high-quality HE4 tertiary structures, where RF1 achieved a slightly better result with no outlier in both plots.

No outlier was recorded for either the AlphaFold or RF1 models, which showed high accuracies in the structural conformations. The AlphaFold protein prediction server was developed according to a hybrid of physical and bioinformatics methodologies, where the developer designed its components to imitate the available PDB data. This was done with the minimum imposition of handcrafted characteristics, using a physical and geometric inductive bias [17]. As a result, the network adapts more effectively from the minimal data in the PDB, while being able to manage the complexity of the structural data, and AlphaFold was reported to be the best protein prediction server available today. Even so, a more stable and dominant HE4 conformation, RF1 (Figure 9) was obtained from the MD simulation of the best predicted model by AlphaFold, with satisfactory validations by the Ramachandran and ERRAT plots, which was better than the valida-

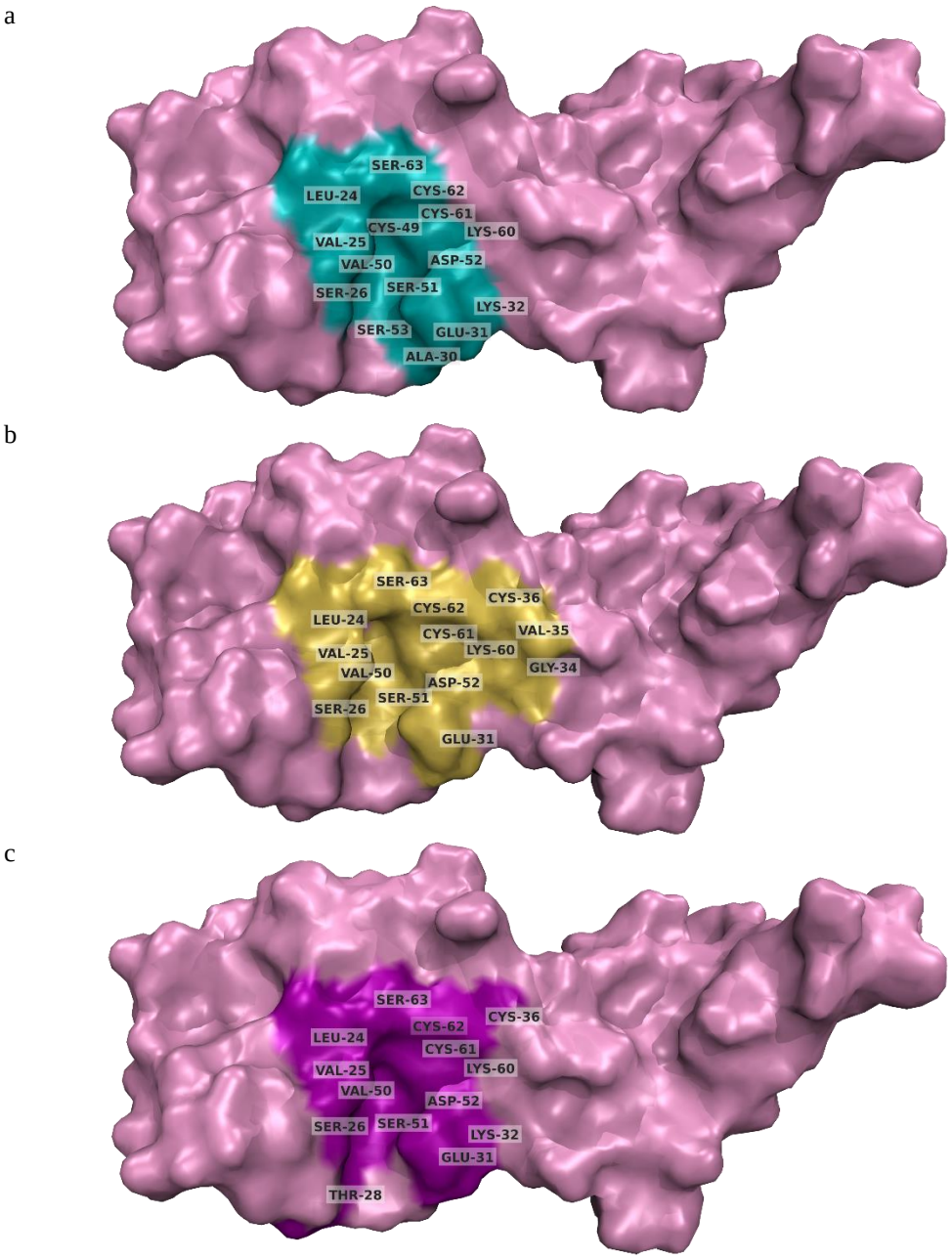


Figure 10. The binding site and its residues, predicted by (a) CAVER 3.0, (b) PrankWeb, and (c) FTsite servers. They show high similarity as 11 identical residues were identified in estimating the interactions with other molecules.

Table 6. The binding site residues predicted by CAVER 3.0, PrankWeb, and FTsite servers.

Binding site residue	CAVER	PrankWeb	FTsite
LEU24	●	●	●
VAL25	●	●	●
SER26	●	●	●
THR28			●
ALA30	●		
GLU31	●	●	●
LYS32	●		●
GLY34		●	
VAL35		●	
CYS36		●	●
CYS49	●		
VAL50	●	●	●
SER51	●	●	●
ASP52	●	●	●
SER53	●		
LYS60	●	●	●
CYS61	●	●	●
CYS62	●	●	●
SER63	●	●	●
Total residues	**15**	**14**	**14**

tions of the AlphaFold predicted model.

In a separate analysis using CAVER 3.0 tool, one tunnel was identified for the RF1 protein structure, with a bottleneck radius of 1.8 Å, length of 1.0 Å, and distance-to-surface of 1.0 Å. A total of 15 residues contributed to the formation of the bottleneck and predicted active site: LEU24, VAL25, SER26, ALA30, GLU31, LYS32, CYS49, VAL50, SER51, ASP52, SER53, LYS60, CYS61, CYS62, and SER63. CAVER 3.0 is a tool designed for the analysis of protein structures, focusing on the identification of tunnels and binding site within the protein structures [52]. This tunnel aids the movement of small molecules such as water, ions and substrates, in and out of the proteins, guiding specific binding of ligand to its active site. This finding improves the understanding of the protein interactions with other biomolecules, elevating the potential to be used in a variety of computational investigations, including facilitating the development of ovarian cancer detection kits.

Two binding site prediction servers, FTsite and PrankWeb, were compared to the prediction made by CAVER 3.0, revealing high degree of similarity, with 11 identical residues (LEU24, VAL25, SER26, GLU31, VAL50, SER51, ASP52, LYS60, CYS61, CYS62, and SER63) contributed in the formation of the binding regions. PrankWeb applies a machine learning approach and predicts the ligand binding site using a template-free method. It assigns random forests to assess the ligand ability of points on the protein's accessible surface, which the points indicate the potential ligand contact atom locations [53]. Figure 10 and Table 6 display the RF1 binding site, highlighting the residues involved in the potential interactions with other ligands. The identification of binding site of proteins enhances the understanding of protein function, estimating interactions with ligands and other proteins, and assisting the process of drug discovery and design.

Conclusion

The tertiary structure of human epididymis protein four was successfully predicted with satisfactory validations from various computational tools. The MD simulations revealed that the stability of the conformation was achieved after 8 ns until the end of the simulation course for all systems. The wiring diagram that shows the secondary structure elements has proven that the most dominant model by the AlphaFold system, RF1, imitates the actual elements of most protein structures with the presence of multiple helices and strands. The RF1 model was deduced as the highest quality HE4 protein tertiary structure based on the

structure evaluation programmes PROCHECK and ERRAT. 100% of the amino acids were found in the favoured regions of the Ramachandran plot, with ERRAT's overall score of 97.701. The CAVER 3.0, PrankWeb, and FTsite tools detected the presence of a tunnel that may facilitate the passage of tiny molecules towards the binding site. Certainly, the outcome of this *in silico* study can be widely utilised for future research such as the development of diagnostic techniques and drug delivery. The best modelled HE4 tertiary structure (RF1) obtained from this research can be further used in computational studies such as molecular docking and molecular dynamic (MD) simulations with various ligands.

Acknowledgement

We would like to express our gratitude to the Research Unit for Bioinformatics & Computational Biology (RUBIC) laboratory of the International Islamic University Malaysia for providing the computational facilities to complete this work, under the RMCSDG grant (RMCG20-002-0002).

References

1. Ferlay J, Ervik M, Lam F et al. (2024) Malaysia Fact Sheets. France: International Agency for Research on Cancer 1–2.
2. Ferlay J, Ervik M, Lam F et al. (2024) World Fact Sheets. France: International Agency for Research on Cancer 1–2.
3. Reid BM, Permuth JB, Sellers TA (2017) Epidemiology of ovarian cancer: a review. Cancer biology & medicine 14 (1): 9–32. doi: 10.20892/j.issn.2095-3941.2016.0084.
4. Charkhchi P, Cybulski C, Gronwald J et al. (2020) CA125 and Ovarian Cancer: A Comprehensive Review. Cancers 12 (12): 3730. doi: 10.3390/cancers12123730.
5. Barr CE, Funston G, Jeevan D et al. (2022) The Performance of HE4 Alone and in Combination with CA125 for the Detection of Ovarian Cancer in an Enriched Primary Care Population. Cancers 14 (9): 1–18. doi: 10.3390/cancers14092124.
6. James NE, Chichester C, Ribeiro JR (2018) Beyond the Biomarker: Understanding the Diverse Roles of Human Epididymis Protein 4 in the Pathogenesis of Epithelial Ovarian Cancer. Frontiers in Oncology 8: 124. doi: 10.3389/fonc.2018.00124
7. Schummer M, Ng W V, Bumgarner RE et al. (1999) Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 238 (2): 375–385. doi: 10.1016/s0378-1119(99)00342-x.
8. Dochez V, Caillon H, Vaucel E et al. (2019) Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review. Journal of Ovarian Research 12 (1): 1–9. doi: 10.1186/s13048-019-0503-7.
9. Ren X, Zhang H, Cong H et al. (2018) Diagnostic Model of Serum miR-193a-5p, HE4 and CA125 Improves the Diagnostic Efficacy of Epithelium Ovarian Cancer. Pathology & Oncology Research 24 (4): 739–744. doi: 10.1007/s12253-018-0392-x.
10. Yao S, Xiao W, Chen H et al. (2019) The combined detection of ovarian cancer biomarkers HE4 and CA125 by a fluorescence and quantum dot dual-signal immunoassay. Analytical Methods 11 (37): 4814–4821. doi: 10.1039/c9ay01454c.
11. Woods RJ (2018) Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. Chemical Reviews 118 (17): 8005–8024. doi: 10.1021/acs.chemrev.8b00032.
12. Jisna VA, Jayaraj PB (2021) Protein Structure Prediction: Conventional and Deep Learning Perspectives. The Protein Journal 40 (4): 522–544. doi: 10.1007/s10930-021-10003-y.
13. Haim A, Neubacher S, Grossmann TN (2021) Protein Macrocyclization for Tertiary Structure Stabilization. ChemBioChem 22 (17): 2672–2679. doi: 10.1002/cbic.202100111.
14. Yang J, Zhang Y (2016) Protein Structure and Function Prediction Using I-TASSER. Current Protocols in Bioinformatics 52 5.8.1-5.815. doi: 10.1002/0471250953.bi0508s52.
15. Agnihotry S, Pathak RK, Singh DB et al. (2022) Chapter 11 - Protein structure prediction. In: Singh DB, Pathak RKBT-B eds. Academic Press. 177–188.
16. Gupta Y, Savytskyi O V, Coban M et al. (2022) Protein structure-based in-silico approaches to drug discovery: Guide to COVID-19 therapeutics. Molecular Aspects of Medicine 101151. doi: 10.1016/j.mam.2022.101151.
17. Jumper J, Evans R, Pritzel A et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873): 583–589. doi: 10.1038/s41586-021-03819-2.
18. Zhang C, Mortuza SM, He B et al. (2018) Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins: Structure, Function, and Bioinformatics 86 (S1): 136–151. doi: 10.1002/prot.25414.
19. Park H, Kim DE, Ovchinnikov S et al. (2018) Automatic structure prediction of oligomeric assemblies using Robetta in CASP12. Proteins: Structure, Function, and Bioinformatics 86 (S1): 283–291. doi: 10.1002/prot.25387.
20. Yang J, Yan R, Roy A et al. (2014) The I-TASSER suite: Protein structure and function prediction. Nature Methods 12 (1): 7–8. doi: 10.1038/nmeth.3213.
21. Baek M, DiMaio F, Anishchenko I et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science (New York, NY) 373 (6557): 871–876. doi: 10.1126/science.abj8754.
22. Baker Lab (2023) Robetta. https://robetta.bakerlab.org/. (2023) Accessed date: February 2023.
23. Tunyasuvunakool K, Adler J, Wu Z et al. (2021) Highly accurate protein structure prediction for the human proteome. Nature 596 (7873): 590–596. doi: 10.1038/s41586-021-03828-1.
24. Lengths M, Angles M (2018) Limitations of structure evaluation tools errat. Quick Guideline Comput Drug Des 16: 75.
25. Williams CJ, Headd JJ, Moriarty NW et al. (2018) MolProbity: More and better reference data for improved all-atom structure validation. Protein Science 27 (1): 293–315. doi: 10.1002/pro.3330.

26. Rashmi D (2018) In Silico Homology Modeling and Validation of a-Glucosidase Enzyme. Journal of Drug Delivery & Therapeutics 8 (6): 124–8.
27. Laskowski RA, Jabłońska J, Pravda L et al. (2018) PDBsum: Structural summaries of PDB entries. Protein Science 27 (1): 129–134. doi: 10.1002/pro.3289.
28. Pradeepkiran JA, Sainath SB, Balne PK, Bhaskar M (2021) Chapter 3 - Computational modeling and evaluation of best potential drug targets through comparative modeling. In: Pradeepkiran JA, Sainath SBBT-BM eds. Academic Press. 39–78.
29. L. S, Vasu P (2017) In silico designing of therapeutic protein enriched with branched-chain amino acids for the dietary treatment of chronic liver disease. Journal of Molecular Graphics and Modelling 76 192–204. doi: 10.1016/j.jmgm.2017.06.015.
30. Saikat ASM, Islam R, Mahmud S et al. (2020) Structural and Functional Annotation of Uncharacterized Protein NCGM946K2_146 of Mycobacterium Tuberculosis: An In-Silico Approach. Proceedings. doi: 10.3390/proceedings2020066013
31. Yin R, Feng BY, Varshney A, Pierce BG (2022) Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Science 31 (8): e4379. doi: 10.1002/pro.4379.
32. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9 (1): 40. doi: 10.1186/1471-2105-9-40.
33. Guo H-B, Perminov A, Bekele S et al. (2022) AlphaFold2 models indicate that protein sequence determines both structure and dynamics. Scientific reports 12 (1): 10696. doi: 10.1038/s41598-022-14382-9.
34. Barber RD (2021) Software to Visualize Proteins and Perform Structural Alignments. Current Protocols 1 (11): e292. doi: 10.1002/cpz1.292.
35. Bauer P, Hess B, Lindahl E (2022) GROMACS 2022 Source code. doi: 10.5281/zenodo.6103835
36. Aier I, Varadwaj PK, Raj U (2016) Structural insights into conformational stability of both wild-type and mutant EZH2 receptor. Scientific Reports 6 (1): 34984. doi: 10.1038/srep34984.
37. Pitera JW (2014) Expected Distributions of Root-Mean-Square Positional Deviations in Proteins. The Journal of Physical Chemistry B 118 (24): 6526–6530. doi: 10.1021/jp412776d.
38. Mangolini F, Hilbert J, McClimon JB et al. (2018) Thermally Induced Structural Evolution of Silicon- and Oxygen-Containing Hydrogenated Amorphous Carbon: A Combined Spectroscopic and Molecular Dynamics Simulation Investigation. Langmuir 34 (9): 2989–2995. doi: 10.1021/acs.langmuir.7b04266.
39. Bahaman AH, Wahab RA, Abdul Hamid AA et al. (2021) Molecular docking and molecular dynamics simulations studies on β-glucosidase and xylanase *Trichoderma asperellum* to predict degradation order of cellulosic components in oil palm leaves for nanocellulose preparation. Journal of biomolecular structure & dynamics 39 (7): 2628–2641. doi: 10.1080/07391102.2020.1751713.
40. Frenkel D, Smit B (1996) Understanding molecular simulation : from algorithms to applications. 2nd ed. Physics Today. doi: 10.1063/1.881812
41. Peng J, Wang W, Yu Y et al. (2018) Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems†. Chinese Journal of Chemical Physics 31 (4): 404–420. doi: 10.1063/1674-0068/31/cjcp1806147.
42. Liu Y, Amzel LM (2018) Conformation Clustering of Long MD Protein Dynamics with an Adversarial Autoencoder. https://www.academia.edu/95826900/Conformation_Clustering_of_Long_MD_Protein_Dynamics_with_an_Adversarial_Autoencoder?uc-sb-sw=87388001. Accessed date: November 2023
43. Narayan M (2021) Securing Native Disulfide Bonds in Disulfide-Coupled Protein Folding Reactions: The Role of Intrinsic and Extrinsic Elements vis-à-vis Protein Aggregation and Neurodegeneration. ACS Omega 6 (47): 31404–31410. doi: 10.1021/acsomega.1c05269.
44. Raschle T, Rios Flores P, Opitz C et al. (2016) Monitoring Backbone Hydrogen-Bond Formation in β-Barrel Membrane Protein Folding. Angewandte Chemie International Edition 55 (20): 5952–5955. doi: 10.1002/anie.201509910.
45. Li J, Wang Y, An L et al. (2018) Direct Observation of CH/CH van der Waals Interactions in Proteins by NMR. Journal of the American Chemical Society 140 (9): 3194–3197. doi: 10.1021/jacs.7b13345.
46. Kumar DD, Pandian L et al. (2017) A Molecular Docking and Dynamics Approach to Screen Potent Inhibitors Against Fosfomycin Resistant Enzyme in Clinical Klebsiella pneumoniae. Journal of Cellular Biochemistry 118. doi: 10.1002/jcb.26064
47. Aizawa H (2018) Allosteric effect by synchronized resonance of amide bonds through alpha-helix. Trends in Research 1 (2): 1–2. doi: 10.15761/tr.1000109.
48. Tam B, Sinha S, Wang SM (2020) Combining Ramachandran plot and molecular dynamics simulation for structural-based variant classification: Using TP53 variants as model. Computational and Structural Biotechnology Journal 18: 4033–4039. doi: 10.1016/j.csbj.2020.11.041.
49. Spencer RK, Butterfoss GL, Edison JR et al. (2019) Stereochemistry of polypeptoid chain configurations. Biopolymers 110 (6): e23266. doi: 10.1002/bip.23266.
50. Mohammed S, Sa'idu H, Manzo JO et al. (2022) Prediction and Validation of 3-Dimensional Structure of Rice OsTHIC Abiotic Stress Responsive Gene. Asian Journal of Plant Biology 4 (1): 1–4. doi: 10.54987/ajpb.v4i1.671.
51. Al-Khayyat MZS, Al-Dabbagh AGA (2016) In silico Prediction and Docking of Tertiary Structure of LuxI, an Inducer Synthase of Vibrio fischeri. Reports of biochemistry & molecular biology 4 (2): 66–75.
52. Stourac J, Vavra O, Kokkonen P et al. (2019) Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. Nucleic Acids Research 47 (W1): W414–W422. doi: 10.1093/nar/gkz378.
53. Jendele L, Krivak R, Skoda P et al. (2019) PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic Acids Research 47 (W1): W345–W349. doi: 10.1093/nar/gkz424.

This page is intentionally left blank.