***In silico* structural analysis of secretory clusterin to assess pathogenicity of mutations identified in the evolutionarily conserved regions**

Lipi Dasa, d, Shashank Shekharb, Pratik Chandrania, c, d, ¥, Ashok K Varmaa, d\*

a Advanced Centre for Treatment, Research and Education in Cancer, Kharghar, Navi Mumbai, Maharashtra- 410210, INDIA.

b Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi - 110016, INDIA.

c Medical Oncology Molecular Lab, Tata Memorial Hospital, Parel, Mumbai, Maharashtra- 400012, INIDA.

d Homi Bhabha National Institute, Training School Complex, Anushaktinagar, Mumbai, Maharashtra- 400094, INDIA.

**Keywords:** Clusterin, *ab initio* modelling, mutational pathogenicity, protein stability

**Corresponding Author Information:**

\*ASHOK K. VARMA, Ph.D.,

Principal Investigator & Scientific Officer 'G',

Tata Memorial Centre, Advanced Centre for Treatment, Research and Education in Cancer,

Kharghar, Navi Mumbai, Maharastra- 410210, INDIA.

Email: avarma@actrec.gov.in

Phone: +91-22-2740 5112

Fax: +91-22-2740 5085

¥Email: pchandrani@actrec.gov.in

**ABSTRACT:**

Clusterin (CLU) is a secreted glycoprotein, heterodimeric in nature, and is expressed in a wide variety of tissues and body fluids such as serum and plasma. CLU has also been known to be a promising biomarker for cell death, malignancy, cancer progression, and resistance development. However, the lack of a CLU crystal structure obstructs understanding the possible role of reported mutations on the structure, and the subsequent effects on downstream signalling pathways and cancer progression. Considering the importance of crystal structure, a model structure of the pre-secretory isoform of CLU was built to predict the effect of mutations at the molecular level.

*Ab initio* model was built using RaptorX, and loop refinement and energy minimization were carried out with ModLoop, ModRefiner, and GalaxyWeb servers. The cancer associated mutational spectra of CLU was retrieved from the cBioPortal server and 117 unique missense mutations were identified. Evolutionarily conserved regions and pathogenicity of mutations identified in CLU were analysed using ConSurf and Rhapsody, respectively. Furthermore, sequence and structure-based mutational analysis were carried out with iSTABLE, DynaMut and PremPS servers. Molecular dynamics simulations were carried out with GROMACS for 50ns to determine the stability of the wild type and mutant protein structures.

A dynamically stable model structure of pre-secretory CLU (psCLU) which has high concurrence with the sequence based secondary structure predictions has been explored. Changes in the intra-atomic interactions and folding pattern between wild type and mutant structures were observed. To our conclusion, eleven mutations with the highest structural and functional significance have been predicted to have pathogenic and deleterious effects.

**Keywords:** Clusterin, *ab initio* modelling, mutational pathogenicity, protein stability

**Abbreviations:** Clusterin-CLU; endoplasmic reticulum-ER; ER associated protein degradation-ERAD; glucose regulated protein-GRP; nuclear clusterin-nCLU; Protein Data Bank-PDB; pre-secretory clusterin-psCLU; secretory clusterin-sCLU

1. **Introduction:**

Clusterin (CLU) is a heterodimeric glycoprotein expressed in a wide variety of tissues and is secreted in body fluids such as serum and plasma. It is coded by a highly conserved gene and was first identified in ram’s rete testis fluid ([Fritz, Burdzy, Setchell & Blaschuk 1983](#_ENREF_15)). CLU has three isoforms that arise due to alternative splicing— the secretory isoform (sCLU), the nuclear isoform (nCLU), and the cytoplasmic isoform (cCLU or isoform 11036) ([Koltai 2014](#_ENREF_28)). Interestingly, two of these isoforms have antagonistic roles in apoptosis— sCLU prevents apoptosis whereas nCLU induces cellular apoptosis ([Zoubeidi, Chi & Gleave 2010](#_ENREF_55); [Koltai 2014](#_ENREF_28)). sCLU is synthesized as a 449 amino acid precursor, with an ER-targeting signal peptide ([Li, Zoubeidi, Beraldi & Gleave 2013](#_ENREF_32); [Foster, Dangla-Valls, Lovestone, Ribe & Buckley 2019](#_ENREF_14)). The resulting protein is N-glycosylated at six asparagine residues, with the formation of 4-5 disulfide bonds ([Kapron et al. 1997](#_ENREF_25); [Rohne, Prochnow & Koch-Brandt 2016](#_ENREF_43)). This ER-associated pre-secretory CLU (psCLU) has a molecular weight of 53kDa that is translocated to the Golgi apparatus. psCLU is further glycosylated with complex carbohydrates and then cleaved between Arg227-Ser228 by a furine-type protein to the final heterodimeric form of the protein ([Li et al. 2013](#_ENREF_32); [Rodriguez-Rivera, Garcia, Molina-Alvarez, Gonzalez-Martin & Goicoechea 2021](#_ENREF_42)). Mature sCLU consists of α and β subunits linked by disulfide bonds, structurally assembled in an anti-parallel manner and functions as a molecular chaperone ([Zoubeidi et al. 2010](#_ENREF_55); [Li et al. 2013](#_ENREF_32); [Koltai 2014](#_ENREF_28)). CLU comprises amphipathic and coiled-coil helices with large stretches of intrinsically disordered regions ([Bailey, Dunker, Brown, Garner & Griswold 2001](#_ENREF_4); [Zoubeidi et al. 2010](#_ENREF_55)).

In addition to its function as a molecular chaperone ([Rohne, Prochnow, Wolf, Renner & Koch-Brandt 2014](#_ENREF_44)), differential expression of CLU has been observed in normal versus malignant tissues undergoing apoptosis, and may be considered as a possible biomarker for cell death, malignancy and progression ([Lyu, Wang, Wang, Zhang & Kong 2018](#_ENREF_34)).

The use of taxanes, such as paclitaxel, to treat cancers is known to induce ER stress which leads to the emergence of treatment-resistant phenotypes ([Lamoureux et al. 2011](#_ENREF_30); [Li et al. 2013](#_ENREF_32)). Resistance arises when molecular chaperones help cells cope with stress-induced misfolded and aggregated proteins. In normal conditions, ER stress causes molecular chaperones to activate the ER-associated protein degradation (ERAD) pathway, thereby eliminating the misfolded proteins ([Wu & Kaufman 2006](#_ENREF_50); [Ron & Walter 2007](#_ENREF_45)). Under treatment-induced ER stress, glycoproteins like psCLU, retrotranslocate from the Golgi network to the ER ([Afshar, Black & Paschal 2005](#_ENREF_2); [Nizard et al. 2007](#_ENREF_36); [Li et al. 2013](#_ENREF_32)). This retrotranslocated psCLU binds to the chaperone glucose-regulated protein (GRP78) to avoid the ERAD pathway and translocates to the mitochondrial membrane. At the mitochondrial membrane, psCLU sequesters activated Bax and disrupts the mitochondrial apoptotic pathway, which in turn leads to the development of a treatment-resistant phenotype ([Zhang et al. 2005](#_ENREF_54); [Nizard et al. 2007](#_ENREF_36); [Trougakos et al. 2009](#_ENREF_48)). Therefore, even in its pre-secretory form, CLU plays an important role in building resistance to anti-cancer therapies, by promoting survival of the tumor cells and possibly allowing them to develop an aggressive phenotype ([Redondo et al. 2006](#_ENREF_39); [Liu et al. 2018](#_ENREF_33); [Rodriguez-Rivera et al. 2021](#_ENREF_42)).

The crystal structure of CLU was not reported in the Protein Data Bank (PDB), though a few groups have characterised its secondary structure ([Bailey et al. 2001](#_ENREF_4); [Wyatt, Yerbury & Wilson 2009](#_ENREF_51); [Matukumalli, Tangirala & Rao 2017](#_ENREF_35)). Therefore, *ab initio* model of CLU was built and refined to predict the folding pattern. Furthermore, *in silico* studies were carried out to determine the effect of reported mutations on CLU structure and their possible correlation to pathogenicity.

1. **Materials and methods:**
	1. *Secondary structure analysis:*

The amino acid sequence of CLU was retrieved in FASTA format from UniProt (P10909) and submitted to PSI-blast based secondary structure PREDiction, PSIPRED, (<http://bioinf.cs.ucl.ac.uk/psipred/>). PSIPRED is an online server that investigates protein structure using artificial neural network machine learning algorithms ([Buchan & Jones 2019](#_ENREF_8)). Secondary structure analysis and disorder prediction were carried out using PSIPRED 4.0 and DISOPRED3, respectively ([Jones 1999](#_ENREF_21); [Jones & Cozzetto 2015](#_ENREF_22)).

* 1. *Three Dimensional structure prediction and loop-refinement:*

Due to the low homology of CLU with available crystal structures in PDB, *ab initio* structure prediction methods were chosen. The FASTA sequence was submitted to I-TASSER ([Yang et al. 2015](#_ENREF_53)), Robetta ([Kim, Chivian & Baker 2004](#_ENREF_26)), and RaptorX ([Kallberg et al. 2012](#_ENREF_24)) servers for three-dimensional model prediction. The coordinates of the highest-ranked model structure were downloaded and validated for stereochemistry. Ramachandran plots were generated for each model by submitting the coordinate file to the SAVES server (<https://saves.mbi.ucla.edu/>) ([Laskowski, MacArthur, Moss & Thornton 1993](#_ENREF_31)). The model having the maximum number of residues in the favoured region was selected for further loop refinement. Multiple rounds of loop refinement were carried out using the ModLoop server (<https://modbase.compbio.ucsf.edu/modloop/>) ([Fiser & Sali 2003](#_ENREF_13)) and the GalaxyWeb Loop Refinement server (<http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=LOOP>) ([Ko, Park, Heo & Seok 2012](#_ENREF_27)). Ramachandran plot was analysed after each round of refinement to assess the quality of the model.

To improve the quality of the model further, two tertiary structure refinements tools were implemented sequentially. The initial model was first refined with the standalone version of ModRefiner (<https://zhanglab.ccmb.med.umich.edu/ModRefiner/>) ([Xu & Zhang 2011](#_ENREF_52)) and the resulting structure was further refined using GalaxyRefine (<https://bio.tools/galaxyrefine>) ([Heo, Park & Seok 2013](#_ENREF_18)) to generate five refined model structures. These five model structures were ranked using ProTSAV (<http://www.scfbio-iitd.res.in/software/proteomics/protsav.jsp>) ([Singh, Kaushik, Mishra, Shanker & Jayaram 2016](#_ENREF_47)) and the best DOPE scored model structure was selected for further refinement. A final round of refinement and energy minimization was carried out with the YASARA (<http://www.yasara.org/minimizationserver.htm>) ([Krieger et al. 2009](#_ENREF_29)) and GalaxyRefine servers (<https://bio.tools/galaxyrefine>) ([Heo et al. 2013](#_ENREF_18)). The refinement was terminated when the Ramachandran plot showed no residues in the disallowed region and this final model was considered as the structure of psCLU. Secondary structure prediction of the final model was performed by the STRIDE webserver ([Heinig & Frishman 2004](#_ENREF_17)). Further comparison of the modelled psCLU was carried out with the recently submitted PDB structure of clusterin in the AlphaFold-EBI PSD database ([Jumper et al. 2021](#_ENREF_23)).

* 1. *Mutational analysis:*

All the reported mutations for CLU were retrieved from cBioPortal (<https://www.cbioportal.org/>) ([Cerami et al. 2012](#_ENREF_9); [Gao et al. 2013](#_ENREF_16)) to study their effects in the folding pattern of psCLU. cBioPortal is an open source for cancer genomics data analysis. Structural and functionally conserved regions of psCLU were identified using the ConSurf server (<https://consurf.tau.ac.il/>) ([Berezin et al. 2004](#_ENREF_7); [Ashkenazy et al. 2016](#_ENREF_3)). The pathogenic probability of a mutation was predicted using the Rhapsody (<http://rhapsody.csb.pitt.edu/index.php>) ([Ponzoni, Peñaherrera, Oltvai & Bahar 2020](#_ENREF_37)) and the destabilizing effect of mutations on the structure of the protein was predicted using the iSTABLE server (<http://predictor.nchu.edu.tw/iStable/about.php>) ([Chen, Lin & Chu 2013](#_ENREF_10)). The structure-based prediction software used were DynaMut (<http://biosig.unimelb.edu.au/dynamut/>) ([Rodrigues, Pires & Ascher 2018](#_ENREF_41)) and PremPS (<https://lilab.jysw.suda.edu.cn/research/PremPS/>) ([Chen et al. 2020](#_ENREF_11)).

* 1. *Molecular dynamics simulation:*

The dynamic properties of the wild type and mutant psCLU were studied by molecular dynamics (MD) simulation using GROMACS version 2021 ([Berendsen, Spoel & Drunen 1995](#_ENREF_6); [Abraham et al. 2015](#_ENREF_1)). The protein structure was solvated with the Extended Simple Point Charge (SPC/E) model of water. Solvation was carried out within a cubic box with 1nm space from the edge of the protein with periodic boundary conditions ([Berendsen, Grigera & Straatsma 1987](#_ENREF_5)). An electrically neutral system was built by the addition of Na+ and Cl- counter ions. Energy minimization was carried out using the steepest descent algorithm for 50,000 steps, with a target energy of 1000kJ/mol/nm ([Robertson, Tirado-Rives & Jorgensen 2015](#_ENREF_40)). The system was coupled to an external bath using Berendsen’s method ([Hess, Bekker, Berendsen & Fraaije 1997](#_ENREF_19)) and MD simulation for 25,000,000 steps of 2fs each using the OPLS-AA force field, for a total of 50ns were performed. Trajectories were analysed using GROMACS built-in tools to calculate Root Mean Square Deviation (RMSD). RMSD plots were generated using Grace (<https://plasma-gate.weizmann.ac.il/Grace/>). Furthermore, the PDB coordinates of the simulated psCLU and its mutants were extracted from the MD trajectories at every 10ns. The weak interatomic interactions were visualized using LigPlot ([Wallace et al. 1995](#_ENREF_49)) and Visual Molecular Dynamics (VMD) ([Humphrey, Dalke & Schulten 1996](#_ENREF_20)). The MD trajectory analysis module of VMD was used to visualise the occupancy of the mutated residues across all MD trajectories.

1. **Results and Discussion:**
	1. *Secondary structure analysis:*

*In silico* secondary structure analysis of CLU using PSIPRED reveals that the protein exhibits 61.02% helicity, 33.41% random coil, and 5.57% strand-like nature **(Figure 1a)**. Despite its largely helical content, multiple stretches of the sequence were also predicted to be part of intrinsically disordered regions **(Figure 1b)**. This also explains the difficulty in getting the crystal structure of CLU protein, as these regions would contribute to instability and poor solubility of protein during purification and crystallization process ([Bailey et al. 2001](#_ENREF_4)). Furthermore, the active form of secretory CLU is glycosylated at different asparagine residues, which in turn stabilizes the protein and conserves its chaperone activity.

* 1. *Three dimensional, ab initio model structure:*

Homology modelling of psCLU did not yield well folded three-dimensional structures due to non-availability of homologous structure in the Protein Data Bank (PDB). Hence *ab initio* modelling was performed, and the model obtained from the RaptorX server showed the best folding pattern with favourable Ramachandran plot statistics. This initial structure had 93.0% residues in the most favoured regions, 6.7% residues in additionally allowed regions, and 0.7% residues in the generously allowed regions. In the first round of structure optimization, the residues in the generously allowed regions were further refined by the ModLoop server ([Fiser & Sali 2003](#_ENREF_13)). Furthermore, multiple cycles of loop refinement were carried out with the GalaxyLoop server ([Ko et al. 2012](#_ENREF_27)). To improve the quality assessment score, the model was refined up to P-Score = 0.5978 using ModRefiner ([Xu & Zhang 2011](#_ENREF_52)). The resulting structure from ModRefiner was further refined using GalaxyRefine to rebuild and repack the side-chains, followed by molecular dynamics-based overall structural relaxation which helped in improving the global and local quality of the model structures ([Heo et al. 2013](#_ENREF_18)). The resulting five model structures were assessed based on the ProTSAV P-scores; P-Score for Model-1 = 0.5786; Model-2= 0.5798; Model-3 = 0.5825; Model-4= 0.5775; Model-5= 0.5787. The best P-scored Model-4 was selected for a final round of refinement and energy minimization with the YASARA server ([Krieger et al. 2009](#_ENREF_29)) and GalaxyRefine server ([Heo et al. 2013](#_ENREF_18)).

This modelled structure of psCLU correlates well with the pattern of secondary structure predictions obtained from PSIPRED **(Supplementary figure 1)**. Ramachandran plot for the final refined structure indicates 98.5% residues in the most favoured region and 1.5% residues in the additionally favoured region **(Figure 2)**. The *ab initio* modelled structure of psCLU was further compared with the structure of CLU submitted in the AlphaFold-EBI PSD database. STRIDE analysis of the two PDB structures indicates a close correlation in the high confidence regions predicted by AlphaFold and the modelled psCLU. However, the major difference between the two structures were observed in the low confidence region predicted by AlphaFold **(Supplementary figure 2).** The region from Asp193 to Glu245, which was predicted as loops by the AlphaFold algorithm, was successfully modelled and refined as α-helices in this study.

Furthermore, the accuracy of the model was also confirmed by noting the distances between the pairs of cysteine residues known to form disulphide bonds in the active form of CLU. Therefore, this model was used for structure-based mutational analysis.

* 1. *Mutational analysis:*

Several CLU mutations have been reported in multiple cancers. However, the absence of a crystal structure hampers understanding of the structural effect of these mutations, and subsequently their role in cancer progression. The cBioPortal for Cancer Genomics (<https://www.cbioportal.org/>) is an open-access, open-source resource for multidimensional cancer genomics data that provides access to molecular profiles and clinical attributes from large-scale cancer genomics projects ([Cerami et al. 2012](#_ENREF_9)). The 187 reported mutations for CLU were retrieved from the cBioPortal database, of which 147 were reported to be missense mutations. Some of the mutations were reported from different cancer types, bringing the number of unique missense mutations to 117. The *in silico* pathogenicity and the effect on protein folding and structure were studied for these 117 missense mutations using the modelled structure of psCLU.

Evolutionary conservation was checked by the ConSurf server ([Berezin et al. 2004](#_ENREF_7); [Ashkenazy et al. 2016](#_ENREF_3)) and the regions having high structural and functional conservation were identified **(Figure 3)**. Evaluation of pathogenicity of mutations identified in CLU was carried out using sequence-based and structure-based prediction software.

Sequence-based prediction software such as Rhapsody ([Ponzoni et al. 2020](#_ENREF_37)) and integrated predictor for protein stability change upon single mutation (iSTABLE) ([Chen et al. 2013](#_ENREF_10)) predict the pathogenicity and changes in protein stability due to mutations. Rhapsody is a machine learning method that enables the assessment of pathogenicity by incorporating sequence coevolution data, structure and dynamics-based features. Rhapsody predicts the effect of a mutation on protein activity and also the potential pathogenicity of the mutation ([Ponzoni et al. 2020](#_ENREF_37)). A saturation mutagenesis analysis of CLU was carried out by performing a complete *in silico* scan of all possible 19 amino acid substitutions at each site on the CLU protein sequence **(Figure 4)**.

An integrated predictor, iSTABLE, was used to study the effects of mutations on the protein structure. iSTABLE combines results from different predictors such as iMutant, Mupro and uses the power of meta predictions to determine the effect of the mutation on the protein stability ([Chen et al. 2013](#_ENREF_10)). The difference in folding free energy change (ΔΔG in kcal/mol) between wild type and mutant protein, an impact factor of protein stability change, was computed.

DynaMut webserver ([Rodrigues et al. 2018](#_ENREF_41)) was used to study the effect of the mutations on the modelled structure of psCLU. DynaMut implements a normal mode approach and integrates information regarding the environment characteristics and graph-based signatures of wild type residues to analyse and visualize protein dynamics. This allows for a consensus on results while assessing the impact of mutations on stability ([Rodrigues et al. 2018](#_ENREF_41)). Mutational analysis by DynaMut predicts changes in protein stability (ΔΔG in kcal/mol), variation in entropy energy, change in flexibility, and allows the visualisation of non-covalent molecular interactions ([Rodrigues et al. 2018](#_ENREF_41)). All of these predictions were carried out for 117 missense mutations **(Supplementary Table 1)**.

Structure-based mutational analysis was carried out using the PremPS server ([Chen et al. 2020](#_ENREF_11)). Different existing computational methods use datasets that are dominated by mutations reducing the protein stability. In contrast, the PremPS method is composed of only ten evolutionary- and structure-based features trained on a balanced dataset with an equal number of stabilizing and destabilizing mutations ([Chen et al. 2020](#_ENREF_11)). The structural coordinates of psCLU were submitted as the input structure along with the list of the missense mutations. The PremPS server predicts change in protein stability as ΔΔG in kcal/mol and the location of the mutation either in the hydrophobic core or protein surface. The non-covalent interactions between wild type and mutant residue were also visualised to determine alterations in the protein structure **(Supplementary Table 1)**. Upon comparing with ConSurf results, eleven mutations located in the structurally or functionally conserved regions were identified. These mutations showed a net destabilizing effect from sequence-based (iSTABLE) and structure-based analysis (DynaMut and PremPS), while concurring with the pathogenicity prediction of Rhapsody **(Table 1)**.

* 1. *RMSD analysis to determine dynamic properties of wild type and mutant psCLU:*

MD simulation using GROMACS version 2021 was used to evaluate the dynamic properties of the wild type and mutant psCLU structures ([Berendsen et al. 1995](#_ENREF_6); [Abraham et al. 2015](#_ENREF_1)). The wild type and mutant structures of psCLU were simulated for 50ns and the coordinates of the protein structure were captured at every 10ns. The RMSD profile of wild type psCLU indicated that the structure stabilizes in the first 10ns of simulation before exhibiting a slight increase after 20ns of simulation. The structure, however, attains stability after 30ns of simulation. A similar profile is observed for the Glu45Lys mutation. The other ten mutants exhibit similar RMSD profiles and attain stability within 10-20ns of simulation. The maximum RMSD observed is 1.8nm, indicating the stability of wild type and mutant structures. The differences in the RMSD between wild type and mutant structures indicate the changes and differences in protein stability **(Figure 6)**. Maximum deviation is observed between the wild type and Lys123Thr mutant.

*3.5 Changes in the weak intramolecular interactions indicate structural alterations:*

The eleven mutations identified at the evolutionarily conserved regions exhibit significant changes at the structural level **(Table 1)**. Structure-based mutational analysis by DynaMut and PremPS shows that Arg227His mutation causes a net destabilizing effect on the structure of CLU and concurrent increase in flexibility. In the psCLU wild type structure, Arg at 227 position forms weak hydrogen bonds with Arg214 and Ser223. Hydrophobic interactions were observed with Leu207, Leu211, Phe218, Arg224, Ile225, Leu229, and Met230 . The Arg227His mutation led to the loss of the hydrophobic interactions with Leu207, Leu211, and Phe218. The loss of a hydrogen bond with Arg214 was also observed **(Figure 5, Supplementary figure 5)**. While the hydrogen bonds remain constant as the simulation progresses, the protein structures extracted from the MD trajectories indicates additional hydrogen bonds with Arg224, His217, and Glu65. It has been reported that psCLU is cleaved between Arg227-Ser228 by a furine-type protein to give rise to the heterodimeric form of CLU ([Li et al. 2013](#_ENREF_32); [Rodriguez-Rivera et al. 2021](#_ENREF_42)). A mutation at this protease site may prevent further processing of psCLU, leading to its accumulation in the ER along with activated Bax, thus disrupting the apoptotic pathway ([Zhang et al. 2005](#_ENREF_54)).

Six mutations were identified in the functionally conserved regions. The Glu45Lys mutation leads to a loss in hydrogen bonds with Asn48 and Arg443, along with a loss of hydrophobic interactions with Glu436 and Gln440. The Lys123Thr mutation forms an additional hydrogen bond with Thr120, and loss in hydrophobic interactions with Gly294 and Arg297. A loss of hydrogen bonds with Arg282 and Arg286 were observed in the Glu287Lys mutant, with no significant changes in the hydrophobic interactions. Furthermore, no significant changes are observed in the Arg324Gln mutation. A loss of hydrophobic interaction with Met34 was observed in the Pro428His mutation, with no changes in the hydrogen bonds. The Arg443His mutation led to the loss in hydrogen bonds with Glu45 and His446 with the formation of additional hydrophobic interactions with Leu384 and His446 **(Supplementary figures 3 and 5)**. All these six mutations led to the disruption of different weak intramolecular interactions and a decrease in protein stability.

The remaining four mutations were identified in the structurally conserved regions of psCLU. The Trp110Cys mutation leads to the loss of a hydrogen bond with Thr88 and a significant loss in hydrophobic interactions due to the absence of the aromatic ring of tryptophan. The loss of hydrophobic interactions with Leu331 was observed in the Ile288Val mutant, with no changes in the hydrogen bonds. The Phe430Cys mutation shows a significant loss of hydrophobic interactions due to the absence of the aromatic ring of phenylalanine. However, no changes in the hydrogen bonds were observed. The formation of multiple hydrophobic interactions were observed for the Ala438Thr mutation **(Supplementary figures 4 and 5)**. It is also interesting to note that the Glu287Lys and Ile288Val mutations are adjacent to each other. Both are structurally conserved and show a destabilizing effect on the protein structure.

The extensive post-translational glycosylation of CLU and its ability to form higher-order oligomers in solution has made it difficult to purify as a recombinant protein and subsequent elucidation of its structure ([Poon et al. 2002](#_ENREF_38); [Dabbs & Wilson 2014](#_ENREF_12)). Thus, the specific amino acid residues of CLU that mediate its interactions with other chaperone proteins such as GRP78, cell receptors, and other biological ligands remain unexplored. A recent report elucidated a rapid and efficient method to express and purify large quantities of native recombinant CLU for the first time ([Satapathy, Dabbs & Wilson 2020](#_ENREF_46)). The high-yield system also allowed the possibility of expressing and characterizing panels of CLU mutants, thus, enabling the identification of the exact residues involved in protein-protein interactions. With the development of such techniques, the study of the structure-function relationships of CLU can be actively pursued by researchers. This could also be a major leap in cancer research, as the structural and functional significance of the reported mutations can be studied in much more detail. Eleven mutations, having the highest structural and functional significance, reported in this study provide a novel starting point for further exploration using *in vitro* and *in vivo* approaches, as all of them were predicted to have pathogenic and deleterious effects.

1. **Conclusion:**

Though *in silico* modelling software and prediction tools are not without their limitations, they provide important leads for further studies. In this study, we have explored a reasonably accurate model of psCLU whose structural predictions are in concurrence with sequence-based predictions. We were also able to identify eleven missense mutations having pathogenic and deleterious effects across multiple prediction software, thus emphasizing the importance of further investigations on the functional effect of these mutations. It is also interesting to note that there are at least two stretches of 20-25 amino acids, between 108-133 and 284-308, which are predicted to be functionally and structurally conserved. Among the eleven mutations, Trp110Cys, Lys123Thr, Glu287Lys, and Ile288Val are located in these two conserved stretches and have significant effects on the protein structure. Therefore, evaluating the effect of other mutations in these two stretches with *in silico* and *in vitro* tools, may provide further insights into their ability to act as promising genetic markers and also act as targets for drug-discovery studies.

**Acknowledgement:**

The authors would like to thank the BTIS facility at ACTREC (DBT grant BT/PR40181/BTIS/137/15/2021) and Dr. Jayaram’s group at SCFBio, IIT Delhi for their help with model refinement and energy minimization. Authors would also like to thank the Indian Institute of Technology, Varanasi for PARAM Shivay- National Supercomputing Mission access provided to PC.

**Author contributions:**

LD, PC, AKV designed the experiments; LD, SS, PC performed the experiments; LD and AKV wrote the manuscript.

**Conflict of interest:**

The authors declare no conflict of interest.

**Word count (Abstract to Conclusion):** 3647

**References:**

Abraham, M., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* *1*: 19-25.

Afshar, N., Black, B. E. & Paschal, B. M. (2005). Retrotranslocation of the chaperone calreticulin from the endoplasmic reticulum lumen to the cytosol. *Mol Cell Biol* *25*: 8844-8853.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* *44*: W344-350.

Bailey, R. W., Dunker, A. K., Brown, C. J., Garner, E. C. & Griswold, M. D. (2001). Clusterin, a binding protein with a molten globule-like region. *Biochemistry* *40*: 11828-11840.

Berendsen, H., Grigera, J. R. & Straatsma, T. (1987). THE MISSING TERM IN EFFECTIVE PAIR POTENTIALS. *The Journal of Physical Chemistry* *91*: 6269-6271.

Berendsen, H., Spoel, D. & Drunen, R. V. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* *91*: 43-56.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R. & Ben-Tal, N. (2004). ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* *20*: 1322-1324.

Buchan, D. W. A. & Jones, D. T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res* *47*: W402-W407.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C. & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* *2*: 401-404.

Chen, C. W., Lin, J. & Chu, Y. W. (2013). iStable: off-the-shelf predictor integration for predicting protein stability changes. *BMC Bioinformatics* *14 Suppl 2*: S5.

Chen, Y., Lu, H., Zhang, N., Zhu, Z., Wang, S. & Li, M. (2020). PremPS: Predicting the impact of missense mutations on protein stability. *PLoS Comput Biol* *16*: e1008543.

Dabbs, R. A. & Wilson, M. R. (2014). Expression and purification of chaperone-active recombinant clusterin. *PLoS One* *9*: e86989.

Fiser, A. & Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. *Bioinformatics* *19*: 2500-2501.

Foster, E. M., Dangla-Valls, A., Lovestone, S., Ribe, E. M. & Buckley, N. J. (2019). Clusterin in Alzheimer's Disease: Mechanisms, Genetics, and Lessons From Other Pathologies. *Front Neurosci* *13*: 164.

Fritz, I. B., Burdzy, K., Setchell, B. & Blaschuk, O. (1983). Ram rete testis fluid contains a protein (clusterin) which influences cell-cell interactions in vitro. *Biol Reprod* *28*: 1173-1188.

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. & Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* *6*: pl1.

Heinig, M. & Frishman, D. (2004). STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* *32*: W500-502.

Heo, L., Park, H. & Seok, C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res* *41*: W384-388.

Hess, B., Bekker, H., Berendsen, H. & Fraaije, J. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* *18*: 1463-1472.

Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *J Mol Graph* *14*: 33-38, 27-38.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* *292*: 195-202.

Jones, D. T. & Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* *31*: 857-863.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* *596*: 583-589.

Kallberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H. & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* *7*: 1511-1522.

Kapron, J. T., Hilliard, G. M., Lakins, J. N., Tenniswood, M. P., West, K. A., Carr, S. A. & Crabb, J. W. (1997). Identification and characterization of glycosylation sites in human serum clusterin. *Protein Sci* *6*: 2120-2133.

Kim, D. E., Chivian, D. & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* *32*: W526-531.

Ko, J., Park, H., Heo, L. & Seok, C. (2012). GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res* *40*: W294-297.

Koltai, T. (2014). Clusterin: a key player in cancer chemoresistance and its inhibition. *Onco Targets Ther* *7*: 447-456.

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D. & Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* *77 Suppl 9*: 114-122.

Lamoureux, F., Thomas, C., Yin, M. J., Kuruma, H., Beraldi, E., Fazli, L., Zoubeidi, A. & Gleave, M. E. (2011). Clusterin inhibition using OGX-011 synergistically enhances Hsp90 inhibitor activity by suppressing the heat shock response in castrate-resistant prostate cancer. *Cancer Res* *71*: 5838-5849.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* *26*: 283-291.

Li, N., Zoubeidi, A., Beraldi, E. & Gleave, M. E. (2013). GRP78 regulates clusterin stability, retrotranslocation and mitochondrial localization under ER stress in prostate cancer. *Oncogene* *32*: 1933-1942.

Liu, Y., Men, C., Xu, Y., Zhao, K., Luo, L., Dong, D. & Yu, Q. (2018). Clusterin promotes growth and invasion of clear cell renal carcinoma cell by upregulation of S100A4 expression. *Cancer Biomark* *21*: 915-923.

Lyu, N., Wang, Y., Wang, J., Zhang, Z. & Kong, W. (2018). Study on early diagnosis of epithelial ovarian cancer by analysis of plasma septin-9 and clusterin level. *J Cancer Res Ther* *14*: S444-S449.

Matukumalli, S. R., Tangirala, R. & Rao, C. M. (2017). Clusterin: full-length protein and one of its chains show opposing effects on cellular lipid accumulation. *Sci Rep* *7*: 41235.

Nizard, P., Tetley, S., Le Drean, Y., Watrin, T., Le Goff, P., Wilson, M. R. & Michel, D. (2007). Stress-induced retrotranslocation of clusterin/ApoJ into the cytosol. *Traffic* *8*: 554-565.

Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N. & Bahar, I. (2020). Rhapsody: predicting the pathogenicity of human missense variants. *Bioinformatics* *36*: 3084-3092.

Poon, S., Rybchyn, M. S., Easterbrook-Smith, S. B., Carver, J. A., Pankhurst, G. J. & Wilson, M. R. (2002). Mildly acidic pH activates the extracellular molecular chaperone clusterin. *J Biol Chem* *277*: 39532-39540.

Redondo, M., Esteban, F., Gonzalez-Moles, M. A., Delgado-Rodriguez, M., Nevado, M., Torres-Munoz, J. E., Tellez, T., Villar, E., Morell, M. & Petito, C. K. (2006). Expression of the antiapoptotic proteins clusterin and bcl-2 in laryngeal squamous cell carcinomas. *Tumour Biol* *27*: 195-200.

Robertson, M. J., Tirado-Rives, J. & Jorgensen, W. L. (2015). Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *Journal of Chemical Theory and Computation* *11*: 3499 - 3509.

Rodrigues, C. H., Pires, D. E. & Ascher, D. B. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* *46*: W350-W355.

Rodriguez-Rivera, C., Garcia, M. M., Molina-Alvarez, M., Gonzalez-Martin, C. & Goicoechea, C. (2021). Clusterin: Always protecting. Synthesis, function and potential issues. *Biomed Pharmacother* *134*: 111174.

Rohne, P., Prochnow, H. & Koch-Brandt, C. (2016). The CLU-files: disentanglement of a mystery. *Biomol Concepts* *7*: 1-15.

Rohne, P., Prochnow, H., Wolf, S., Renner, B. & Koch-Brandt, C. (2014). The chaperone activity of clusterin is dependent on glycosylation and redox environment. *Cell Physiol Biochem* *34*: 1626-1639.

Ron, D. & Walter, P. (2007). Signal integration in the endoplasmic reticulum unfolded protein response. *Nat Rev Mol Cell Biol* *8*: 519-529.

Satapathy, S., Dabbs, R. A. & Wilson, M. R. (2020). Rapid high-yield expression and purification of fully post-translationally modified recombinant clusterin and mutants. *Sci Rep* *10*: 14243.

Singh, A., Kaushik, R., Mishra, A., Shanker, A. & Jayaram, B. (2016). ProTSAV: A protein tertiary structure analysis and validation server. *Biochim Biophys Acta* *1864*: 11-19.

Trougakos, I. P., Lourda, M., Antonelou, M. H., Kletsas, D., Gorgoulis, V. G., Papassideri, I. S., Zou, Y., Margaritis, L. H., Boothman, D. A. & Gonos, E. S. (2009). Intracellular clusterin inhibits mitochondrial apoptosis by suppressing p53-activating stress signals and stabilizing the cytosolic Ku70-Bax protein complex. *Clin Cancer Res* *15*: 48-59.

Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* *8*: 127-134.

Wu, J. & Kaufman, R. J. (2006). From acute ER stress to physiological roles of the Unfolded Protein Response. *Cell Death Differ* *13*: 374-384.

Wyatt, A. R., Yerbury, J. J. & Wilson, M. R. (2009). Structural characterization of clusterin-chaperone client protein complexes. *J Biol Chem* *284*: 21920-21927.

Xu, D. & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* *101*: 2525-2534.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat Methods* *12*: 7-8.

Zhang, H., Kim, J. K., Edwards, C. A., Xu, Z., Taichman, R. & Wang, C. Y. (2005). Clusterin inhibits apoptosis by interacting with activated Bax. *Nat Cell Biol* *7*: 909-915.

Zoubeidi, A., Chi, K. & Gleave, M. (2010). Targeting the cytoprotective chaperone, clusterin, for treatment of advanced cancer. *Clin Cancer Res* *16*: 1088-1093.

**Figure legends:**

**Figure 1: (a)** Secondary structure analysis of CLU using PSIPRED and **(b)** Disorder prediction by DISOPRED.

**Figure 2: (a)** *Ab initio*modelled structure of psCLU. **(b)** Ramachandran plot for modelled structure of psCLU.

**Figure 3:** Evolutionary and structurally conserved regions of CLU predicted by the ConSurf server.

**Figure 4:** Pathogenicity prediction of reported mutations using Rhapsody server.

**Figure 5:** Weak intramolecular interactions observed in psCLU Arg227His mutation. Protein structures have been captured at every 10ns of MD simulation. Panel **(a)** depicts the interactions observed in the wild type psCLU structure and panel **(b)** depicts the interactions in the mutant structure. Hydrogen bonds are depicted in green colored dotted lines. Hydrophobic interactions are denoted by red half circles.

**Figure 6:** RMSD plots of wild type psCLU and mutants obtained after 50ns of molecular dynamics (MD) simulation.

**Table legend:**

**Table 1**: Mutation induced changes in protein stability and pathogenicity as predicted by DynaMut, iSTABLE, PremPS, and Rhapsody servers.

**Supplementary legends:**

**Supplementary Table 1**: Mutation induced changes in protein stability and results predicted by DynaMut, iSTABLE, and PremPS servers.

**Supplementary Figure 1:** Correlation between **(a)** secondary structure prediction by PSIPRED and **(b)** secondary structure prediction of *ab initio* model structure by STRIDE. The evolutionarily conserved residues are highlighted by green boxes. **(c)** Percentage of α-helix, random coil, and β-sheet calculated by PSIPRED and STRIDE. Most parts of the secondary structure predicted by PSIPRED correlate well with the final model of psCLU. Coiled coil regions are usually unstable in nature and hence the degree of confidence in predicting them is lower, as indicated by the shorter blue bars in the PSIPRED prediction result. However, this gets resolved during model refinement, which ultimately generates a more stable structure. Therefore, some differences are observed in the coiled coil regions of the PSIPRED prediction and the final model.

**Supplementary Figure 2:** Correlation between the low confidence region (Asp193 to Glu245) in the modelled psCLU and AlphaFold CLU structures. The modelled psCLU structure shows refinement of the previously predicted loop region as α-helices. STRIDE secondary structure predictions are indicated in tabular form.

**Supplementary Figure 3:** Weak intramolecular interactions observed in psCLU mutations located in the functionally conserved regions. Protein structures have been captured at every 10ns of MD simulation. Panel **(a)** depicts the interactions observed in the wild type psCLU structure and panel **(b)** depicts the interactions in the mutant structure. Hydrogen bonds are depicted in green colored dotted lines. Hydrophobic interactions are denoted by red half circles.

**Supplementary figure 4:** Weak intramolecular interactions observed in psCLU mutations located in the structurally conserved regions. Protein structures have been captured at every 10ns of MD simulation. Panel **(a)** depicts the interactions observed in the wild type psCLU structure and panel **(b)** depicts the interactions in the mutant structure. Hydrogen bonds are depicted in green colored dotted lines. Hydrophobic interactions are denoted by red half circles.

**Supplementary Figure 5:** Occupancy analysis of mutated residues across all trajectories as visualized in VMD. The differently colored spheres indicate the volume occupied by the atom during the simulation. Blue- maximum volume; Green-medium volume; Red-minimum volume.