UNIVERSITÀ DEGLI STUDI DI PADOVA

Master Degree in Physics of Data

Biological Datasets for Computational Physics

Final Project

Sara's disease: an exploration of MMACHC and its implications in CblC

Professor

Prof. Monika Fuxreiter

Candidate

Luca Brocco

Academic Year 2024/2025

Contents

1 Introduction

Sara is a cousin of mine, second child of her parents and fourth nephew of my grandparents. We were happy that our family got larger; at first glance everything was fine. But when Sara started growing, things looked weird. She had a visible development delay, with severe vision problems. Her parents investigated the problems with the best doctors. The diagnosis was not easy, but in the end they came to the following conclusion: Sara was hit by **Methylmalonic acidemia with homocystinuria of type C (CblC)**. CblC is an inborn error of vitamin B12 metabolism that usually manifests in hit individuals during neonatal age. It involves the inability to convert Cobalamin into active forms, with several serious consequences such as neurological deterioration, anemia and cardiomyopathy. CblC is an Inherited Metabolic Disease and is transmitted in an autosomal recessive manner: both parents must be carriers and then they have a 1/4 chance for every child to have the disease.
Molecular analysis of **MMACHC** gene is a screening technique for identifying the carrier status of parents. Several other proteins are involved in vitamin B12's metabolism. This work is going to primarily focus on exploration of MMACHC structure and its possible mutations.

2 Methods and Objectives

2.1 MMACHC structure

MMACHC is a widely expressed protein that catalyzes the reductive decyanation of cyanocobalamin to yield Cobalamin and cyanide using FAD or FMN as cofactors and NADPH as cosubstrate [1], [2]. Cyanocobalamin constitutes the inactive form of vitamin B12 introduced from the diet, and it is converted into the active cofactors methylcobalamin and 5-deoxyadenosylcobalamin [3].
Human MMACHC is composed by 282 amino acids arranged in a specific sequence available on UniProt database [4]:

MEPKVAELKQKIEDTLCPFGFEVYPFQVAWYNELLPPAFHLPLPGPTLAFLVLSTPAMFD
RALKPFLQSCHLRMLTDPVDQCVAYHLGRVRESLPELQIEIIADYEVHPNRRPKILAQTAAHV
AGAAYYYQRQDVEADPWGNQRISGVCIHPRFGGWFAIRGVVLLPGIEVPDLPPRKPHDCVPT
RADRIALLEGFNFHWRDWTYRDAVTPQERYSEEQKAYFSTPPAQRLALLGLAQPSEKPSSPS
PDLPFTTPAPKKPGNPSRARSWLSPRVSPPASPGP

MMACHC has a N-terminal region which is highly disordered, and a Rossmann-like folded core with a $\alpha/\beta/\alpha$ sandwich motif, typical of cofactor-binding proteins. Within the core there is also a B12-binding pocket for anchoring and orienting the vitamins and an active site flexible loop. The sequence is ended by a C-terminal domain for structural stability [5]. The 3D structure of the protein can be obtained from the AlphaFold Protein Structure Database [6]. The structure is ranked according to

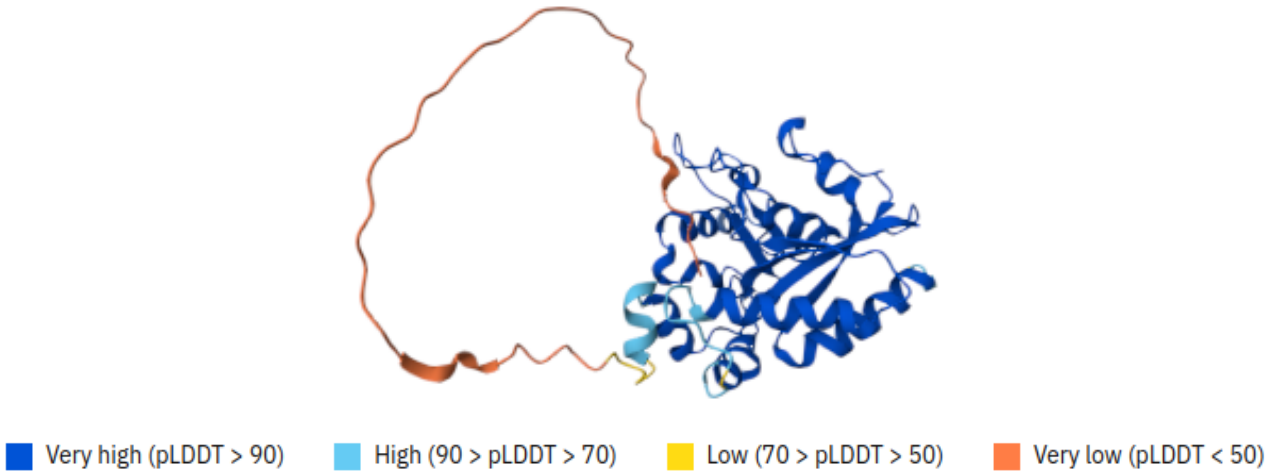

Figure 2.1: Structure of MMACHC according to AlphaFold Database with Confidence Levels per residue

pLLDT confidence score and is mostly defined with very high confidence level.

2.2 Methods

MMACHC has a small but relevant disordered part in its structure. Using machine learning methods, we want to inspect the protein's structure, as well as its possible mutations. Results will be compared with various available sources, such as ESpritz tool [7], DisProt Database [8] and MobiDB [9], as well as ClinVar [10] and UniProt for known disease-associated mutations. We will train a Random Forest for predicting the disordered part of the protein, then we will use a Convolutional Neural Network for predicting the secondary structure of the protein. Later on, we will exploit a pre-trained Variational Auto Encoder for predicting single amino acid mutations in MMACHC structure and make use of AlphaFold [6] and FoldX [11] to perform an analysis.

2.3 Objectives

In this work we want to explore MMACHC's structure using ML algorithms and test their accuracy on the target protein. We also want to investigate possible mutations of the primary structure of the protein, identify their nature as benign or possible disease-related and explore their interaction with other proteins involved in vitamin B12's methabolism, such as MMADHC, MTRR and MTR [12]. Possibly, we want the model to correctly predict already known illness-related mutations, and predict

the outcome of mutations which at present have not been defined as pathogenic or benign.

3 Models and predictions

3.1 Identification of disordered residues

As seen before, MMACHC has a well defined primary structure. The secondary structure is for the most part well defined, with a small region being disordered. First of all, we want to be able to detect automatically this region and identify any possible disordered residues left in the structure. For this goal we are going to train a Random Forest on DisProt data. We are feeding the forest with other proteins structures, and we are using it to classify each residue of MMACHC as disordered or not. This step ensures that our methods are powerful enough for this work's main goals. Results are shown in Table 3.1.

Table 3.1: Summary of MMACHC disorder

Method	% disorder	dis. regions > 30 res.	tot. dis. segments
Random Forest (trained on DisProt)	18.4	1	1
DisProt	18.4	1	1
ESpritz (X-ray)	35.1	1	3
ESpritz (DisProt)	31.6	1	2
ESpritz (NMR)	46.8	1	7

We can see that the Random Forest trained on DisProt (which provides easily accessible APIs) obtains the same results as DisProt database itself. This is to be expected if the method works well on the dataset (composed by other proteins data from DisProt) and features that characterize disordered residues are consistent across different structures. We conclude that ML algorithms can be used for protein structural analysis and let us explore their features.

3.2 Secondary structure predictions

In this part we are going to train a Convoluted Neural Network with the goal to predict the secondary structure of a protein, with a particular focus on MMACHC.
As training dataset we initially tried to use Cullpdb dataset [13]. The dataset contains information about 5365 polypeptides. Each polypeptide has there encoded its primary and secondary structure, as well as some important features obtained by PSSM (position-specific scoring matrices) calculated by PSI-BLAST (Position Specific Iterative-Basic Local Alignment Search Tool) [14]. Both primary and secondary structure are one-hot encoded. 3 classes are going to be defined for this classification: H (helix), E (strand) and C (coil). This simplifies the purpose of the dataset, that features 8 classes, in fashion of DSSP classification [15], but allows us to have a better performance on a single polypeptide, which is the goal of our machine. Input data is going to be fed with a 'sliding window' approach, allowing the CNN to get neighborhood knowledge for each residue. Since this approach is computationally expensive, we are going to feed only a limited number of proteins, chosen by terms of a similarity metric with MMACHC. Results are shown in 3.2.

After training the model we realized that the performance on MMACHC sequence was consistently worse than on a random sequence extracted from the dataset. This is probably due to the fact that

Table 3.2: Summary of first DSSP prediction model

Method	Total proteins	f1 validation score	Accuracy on MMACHC
CNN	1000	0.899	35.5%

MMACHC's structure is much different from the one provided in Cullpdb.
We then decided to build our own custom dataset. The idea is to select specific proteins that share purposes or features with MMACHC. We are using BIOGrid [16] to get code references for about 100 proteins. We are then training again the CNN on the new data. Results are shown in Table 3.3.

Table 3.3: Summary of second DSSP prediction model

Method	Total proteins	f1 validation score	Accuracy on MMACHC
CNN	86	0.63	77.7%

Even though the model was trained on a significantly lower number of polypeptides, the results on MMACHC are significantly better and in-line with model's overall performance. Best results were obtained with windows covering 11 residues. We show model's DSSP prediction in Figure 3.1.

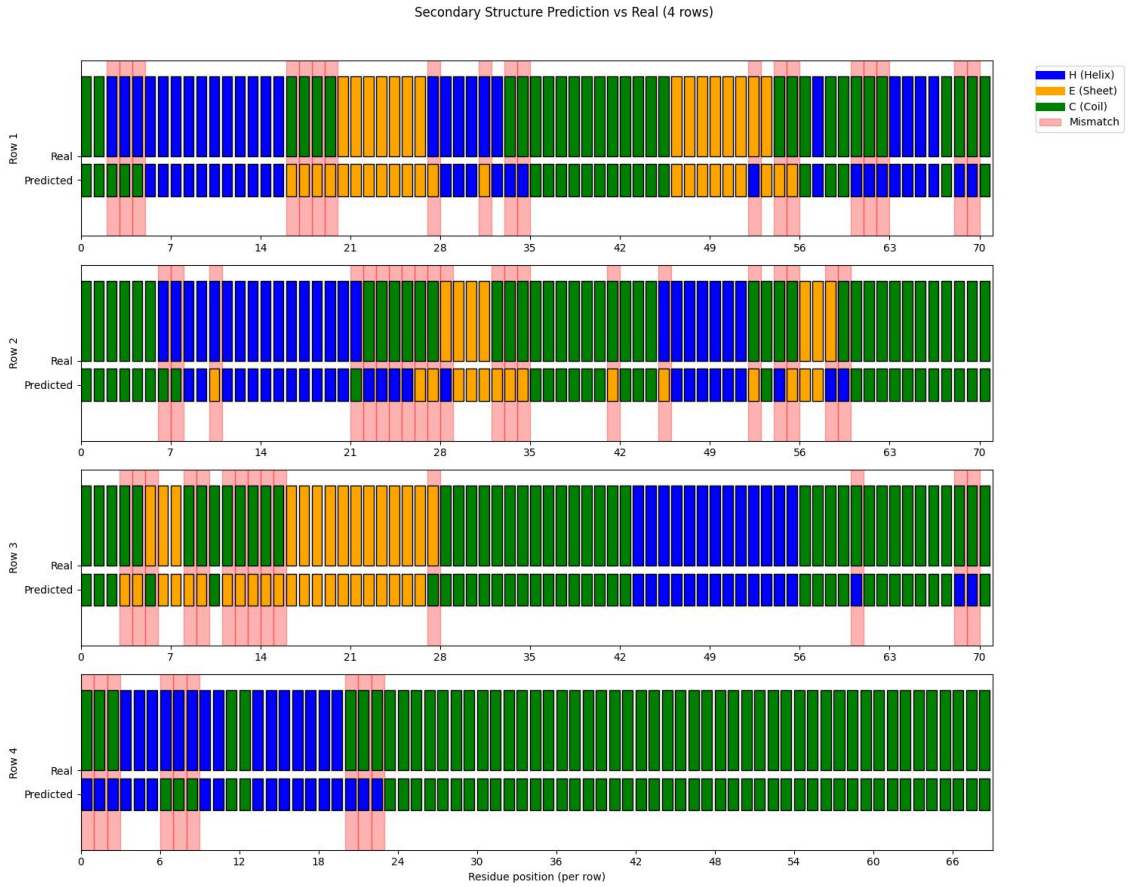

Figure 3.1: DSSP of MMACHC according to our newly trained CNN. Residues whose predictions are labeled wrong are highlighted in red.

We observe that there is no clear overlap between the errors of this section and the residues identified

as disordered in the previous one.

3.3 Mutations prediction

In this section we want to predict possible mutations and their nature (benign or pathogenic) on our target protein using ML techniques. We tried implementing a *n-gram* algorithm from scratch [17], but failed in obtaining acceptable performances due to lack of data and computational resources. N-gram approaches, in fact, scale badly with the size n of the gram (20^n). In addition to that, they do not take count of large-scale dependencies in protein structures. Other models are available in literature and are known to work, but their training is computationally expensive and time demanding. For these reasons we are relying on an already trained and publicly available model for this task: EVE [18]. EVE is a model for the prediction of clinical significance of human variants based on sequences of diverse organisms across evolution. It uses fully unsupervised deep learning trained on amino acid sequences of over 140K species. EVE allows for predictions of single-amino acid mutations, and classifies each mutation in a 0-1 scale, where 0 represents a most benign mutation, and 1 a most pathogenic one.
MMACHC's mutation spectrum is almost complete (residues not computable by EVE are shown in white) and results are shown in Figure 3.3.

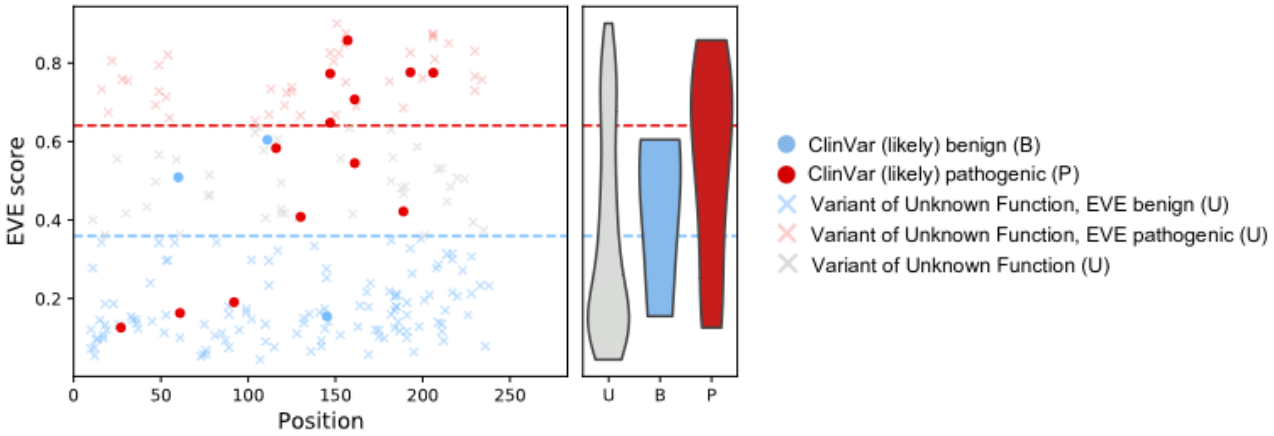

Figure 3.2: Statistics summary for variants seen in humans to date (dots). Dashlines represent EVE's classification boundaries for the 75% most confident class assignments.

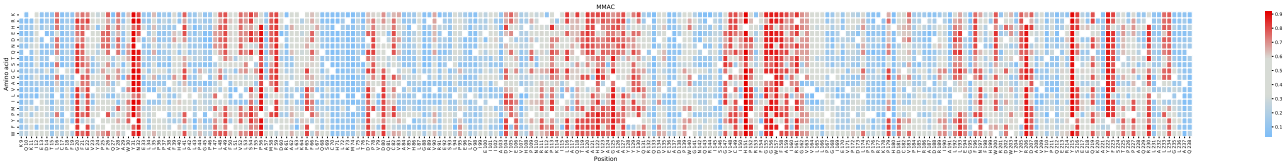

Figure 3.3: MMACHC's mutation spectrum computed by EVE. Cell colors range from red (pathogenic) to light blue (benign)

All MMACHC studied mutations (to date) are involved with CblC. EVE predictions agree upon 87.5% of mutations present in literature. Disagreements come for mutations Q27R [19], R61P and E92D. However, for the last 2 there is no sufficient literature evidence for their classification [20]. We conclude that EVE predictions for single residue mutations in MMACHC are accurate enough in terms of damage done to the protein, and thus must be explored further.

3.4 Mutations exploration

As previously pointed out, EVE predictions agree in large part with literature and known mutations relative to CblC disease. EVE's most selling point, however, is the capability of predicting pathogenicity

of single residue mutations for all residues of a sequence, regardless of the existing specific documentation, thanks to advanced ML and exploiting huge and comprehensive datasets. We now want to explore possible mutations that are scored most pathogenic according to EVE and test their consequences on MMACHC and its co-expression partners' structure and functionality.
Since MMACHC and its diseases are mainly related to vitamin B12 metabolism, targets of this section will be MMADHC [21], being MMACHC's direct binding partner, MTR, responsible for cytosolic B12 enzyme downstream of MMACHC and MTRR [22], which maintains MTR activity and is co-complex with MMACHC in cytosol.
A possible way to study structural instability due to mutations is to feed AlphaFold with the modified sequence of the protein [23]. Another useful tool to get instability metrics is FoldX [11]. We compute EVE's predictions for most pathogenic and most benign mutations, with most interest towards MMACHC functional sites, and check their effects on MMACHC's stability. We measured metrics such as Effective Strain (ES) and Total Energy (ΔE), which are reported as differences values from one of the wild-type. Results are averaged on all the simulations outputted by AlphaFold, and are shown in Table 3.4. Errors are computed statistically over all AlphaFold simulations. R206Q mutation is used as reference, as it is a known pathogenic mutation [24].

Table 3.4: Summary of structural analysis of mutated MMACHC structures.

Position	Residue	Alteration	EVE Score	ΔpLLDT	ES	ΔE
90	V	Q	0.671	0.8 $\pm$ 0.1	19.5 $\pm$ 1.6	26.8 $\pm$ 5.7
95	P	F	0.689	1.1 $\pm$ 0.1	25.9 $\pm$ 2.9	20.6 $\pm$ 8.6
104	D	F	0.830	1.36 $\pm$ 0.07	16.8 $\pm$ 1.3	15.9 $\pm$ 5.5
161	R	S	0.843	2.65 $\pm$ 0.09	18.2 $\pm$ 1.3	0.4 $\pm$ 2.9
196	F	H	0.811	0.7 $\pm$ 0.2	22.8 $\pm$ 2.4	19.3 $\pm$ 3.0
197	N	I	0.826	0.9 $\pm$ 0.2	24.9 $\pm$ 1.9	20.2 $\pm$ 4.7
200	W	E	0.837	0.8 $\pm$ 0.3	25.1 $\pm$ 2.3	9.6 $\pm$ 3.7
202	D	C	0.770	1.1 $\pm$ 0.2	15.3 $\pm$ 4.9	14.8 $\pm$ 4.2
203	W	P	0.705	1.3 $\pm$ 0.1	19.1 $\pm$ 3.4	8.3 $\pm$ 5.2
233	L	P	0.777	0.6 $\pm$ 0.1	23.1 $\pm$ 3.5	9.4 $\pm$ 3.9
234	L	C	0.778	1.1 $\pm$ 0.1	26.9 $\pm$ 1.7	13.5 $\pm$ 3.8
206	R	Q	0.775	1.0 $\pm$ 0.3	26.9 $\pm$ 1.3	22.8 $\pm$ 9.2

We observe that for all mutations ΔpLLDT is small and positive. This means that AlphaFold predicts the structure of all mutants with a confidence at least equal to the one associated with wild-type MMACHC. While this does not ensure correctness of predictions, it still measures the capability of the software in predicting unseen structures. We also observe that the energy increase of the mutants with respect to the wild-type is high (> 6 $kcal/mol$) for specific mutations (V90Q, P95F, D104F, F196H, N197I, D202C, L234C), and the same happens for the energy increase of R206Q, which is a known pathogenic mutation. This suggests that the analyzed mutations can lead to structural destabilization and even be pathogenic, with consequences in Cobalamin processing and metabolic activity. Even Effective Strain is high for all mutations [25]. This is another metric useful in determining the loss of stability from the wild-type.

In addition to mutation consequences on MMACHC, we are able to retrieve information about the whole MMACHC, MMADHC, MTR and MTRR complex and how single mutations in MMACHC impact the structure and stability of the whole system. To do that, we are using the same AlphaFold and FoldX pipeline as before, but instead of the "Stability" command we are using "AnalyzeComplex" from FoldX, which allows for studying several metrics relative to polypeptide complexes. Unfortunately, AlphaFold was not able to generate the structure and predict the interactions for some of the analyzed

mutations (P95F, D104F). These have to be considered as strongly damaging in terms of the interaction between the proteins, hence the inability of the software on providing a stable state for the complex. We were still able to analyze the remaining mutations with the pipeline and obtained several metrics for all mutations. Final results are shown in Table 3.5 and will be discussed in later chapters.

Table 3.5: Summary of structural analysis of mutated MMACHC structures.

Position	Residue	Alteration	ΔpLLDT	ΔE	ΔE Interaction
90	V	Q	-0.1 ± 0.2	-1.2 ± 3.8	17 ± 31
196	F	H	-0.5 ± 0.2	13.5 ± 7.4	59 ± 40
197	N	I	-0.13 ± 0.05	-0.2 ± 2.6	8.2 ± 6.5
202	D	C	0.05 ± 0.09	-1.3 ± 2.0	3.7 ± 11.9
234	L	C	0.1 ± 0.2	4.3 ± 2.1	-9 ± 29
206	R	Q	-0.4 ± 0.1	1.4 ± 1.9	-34 ± 8

4 Discussion

In this work several ML techniques were used to perform predictions in terms of structural composition and consequences of single mutations in proteins. In 3.1 we trained a Random Forest on DisProt data. Our goal was to predict if a MMACHC residue is disordered or not starting from the primary structure and hydrophobicity of amino acids. This task was relatively simple, being a binary classification well distinguished over the feature space. We were able to replicate DisProt's best performance, showing the capability of ML algorithms over predicting protein structures and features.

In 3.2 we trained a CNN with the goal of predicting the secondary structure of MMACHC. We faced several problems running this task. First off, Cullpdb dataset [13], even though being large and comprehensive, was inadequate for our objective. Proteins in this dataset had very low similarity scores and probably did not share enough features with MMACHC. This led our CNN to perform very well on the Validation set, with results comparable with best performances in literature [26], but poorly on MMACHC specifically. Since our goal was to predict this specific polypeptide, not to get a universal predictor, we decided to create a custom training dataset. This way we were able to train the CNN on a dataset in which we ensured to include proteins that shared structural features or functionality with MMACHC. BIOgrid tool [16] helped us identifying which proteins to include in the dataset, considering the human metabolism of vitamin B12 [22]. With this *caviat* we were able to get an accuracy on MMACHC of 77.7%, slightly off the best accuracy performances obtainable from Deep Learning approaches in implementing DSSPs [27]. This is probably due to the fact that our dataset was far smaller than the ones used by the best models (which train on the whole UniProt, in range of TB of space allocation), and our computational resources were far more limited. Keeping a small amount of proteins in the dataset, while artificially engineering it, showed promising results in terms of a specific protein prediction. With more resources available, more comprehensive and wide-use models can be trained. However, no errors were made in the region identified as disordered in 3.1. This is probably due to the fact that the available classification for this region is more simplistic than the real MMACHC structure. However, we can only train models on available experimental data. Until no data is found, our model cannot output better predictions.

After having shown the capabilities of ML in terms of structural prediction, we also wanted to inspect possible mutations in MMACHC primary structure and check whether we were able to predict new pathogenic mutations 3.3. As pointed out in [28], several ML techniques are available for this task: many of them have now a long history of implementations and improvements. However, they all need a very large dataset to train on, including a wide range of proteins and their clinical ensured mutations, and many numeric features for helping out the training process. We tried a simple n-gram statistical model [17], but were not able to obtain significant results. This approach scales exponentially (as 20^n) with the size n of gram considered, so it needs a training dataset of a comparable size, which we could not afford. In addition to that, this method does not consider large distance dependencies, which many proteins have already shown to have [29], including MMACHC.
Seen the limitations in training a model *ex novo*, we decided to take a pre-trained model and inspect its predictions on our target protein. EVE seemed to be the best candidate for this purpose [18], being a model that exploits Variational Auto Encoders to train on a large amount of data and retrieve from it only the components essential for the regression it is trained on. EVE provides a score for each analyzed mutation, from 0 (most benign) to 1 (most pathogenic). We took EVE predictions for MMACHC and cross-checked with available clinic documentation to see whether the predictions were accurate or not. Results showed that EVE agreed with literature on 87.5% of mutation outcomes, and disagreed only on those of which there is not sufficient evidence of benignity/pathogenicity at time of this work. We concluded that EVE predictions on single amino acid mutations were accurate and needed inspection.
EVE predicts the outcome of every possible mutation in every possible residue of the chain of MMACHC, highlighting a score for each. We thought that the most reasonable analysis had to be done over MMACHC residues mutations known for being located in binding sites or functional cores and with high EVE scores. Mutations in Cobalamin binding sites (residues 103-115, 196-203) can alter the folding capability of the protein [30], reducing its binding affinity with Cbl, while those localized in Glutathione Binding Pocket (161, 206, 230, all Arg) may influence Gsh ability to bind to MMACHC [31]. The so called PNRRP loop (90-94) functions as lid for Cobalamin site, and its malfunctioning leads to improper dealkylation activity [32]. Residues from 150-180 and 250-260 are instead hypotized to be the interaction site for MMADHC [33]. To avoid inaccurate predictions on mutations, we took in analysis mutations in residues out of the disordered regions.
After selecting the mutations, we attempted to use AlphaFold [6] to predict the structure of the mutants. This has already been shown to be possible and useful to retrieve stability information about the mutant structures [23]. For all mutations we computed ΔpLLDT, which quantifies the confidence of AlphaFold in predicting the structure, Effective Strain ES and difference in energy from the wild-type ΔE, which are both metrics that quantify the stability of the mutant. It is notable that ES was computed directly on AlphaFold output, while ΔE was obtained using FoldX [11] on AlphaFold output.
$\Delta pLLDT$ scores were all slightly higher than 0. We interpreted this as AlphaFold capability of representing mutant structures with sufficient confidence, at least equal to the one of wild-type. General trends in terms of ES and ΔE were observed: while not strictly correlated, almost all mutants showed significant increase from the wild-type in terms of these two metrics. We selected the subset of mutations that had a significantly very high increase of energy (> 6 $kcal/mol$) for the second part of the analysis, since they were the most likely to be harmful for the complex in charge for vitamin B12 metabolism. Our results from this section, being related to not clinical-confirmed pathogenic mutations, cannot be compared with literature-proven results. However, our mutations find confirmation in databases such as ClinVar Miner [34], where we were able to find information about possibly harmful mutations that are not confirmed as pathogenic.

Finally, in 3.4 we analyzed the whole vitamin B12 complex, made of MMACHC, MMADHC, MTR and MTRR, and tried to retrieve information about the structural and functional impact of a single amino acid mutation in MMACHC structure on the system. We used the same AlphaFold

and FoldX pipeline as 3.3, but using the "AnalyzeComplex" FoldX command instead of "Stability". AlphaFold was not able to generate the structures of some mutations (P95F, D104F). We assume that these mutations were the most destabilizing in terms of the structure of the whole complex, seen the inability of obtaining a structural prediction. However, we were still able to retrieve information about several mutations and compute $\Delta pLLDT$, showing how AlphaFold is still highly confident of the outputted predictions, and ΔE, highlighting one of its specific components, the one related to protein-protein interaction $\Delta E\ Interaction$. It is notable that errors obtained in table 3.5 are far bigger in the ones presented in table 3.4. This is probably due to the fact that we are predicting systems with a higher degree of complexity, and each run of the software produces results that are slightly more distributed. However, in light of the obtained values, we affirm that the most harmful mutation on the complex, out of the analyzed ones, seems to be F196H, being the only one with a high value of ΔE with respect to the wild-type complex. Other mutations seem to have impacted less the complex than MMACHC structure itself. However, being high the range of error, is not impossible that other amino acid alterations taking place in the same positions have significantly worse impacts.

5 Conclusions

In this work we ran a comprehensive study of MMACHC structure and the impact of single amino acid mutations on it. The whole process was based on the hypotesis that structural mutations can damage the protein in terms of stability and functionality. While not inspecting directly the functionality, we were able to predict whether a mutation can be harmful in terms of protein stability or not. This work has to be taken as a pathway driver for direct clinical analysis, ensuring that experimentalist waste no time in inspecting mutations that have no impact in human illnesses.
We need to point out that the mutation part analysis was performed with predictions from AlphaFold and FoldX, being the state-of-the-art in terms for software for sequence structural predictions. However, this software was not directly designed for predicting the effects of mutations in known structures. The results then suffer strongly from dataset bias and are correlated to existing structures [35]. Thus, they must be handled carefully and *cum grano salis*: mutations identified as mostly pathogenic and destabilizing must be explored further. We also point out that we analyzed only the highest EVE score mutation for each targeted position. However, for most position there were multiple amino acid substitutions that had similar EVE scores. Investigating the outcomes of these other mutations may provide information about specificity of mutations. Given that for each residue there are 19 possible mutations it is safe to say that work like this one can really reshape the way we think about protein bioengineering. Time and resources can be saved by exploiting ML capabilities in this field, and this translates in human lives and health gain. Chances are that with help of ML even rare illnesses, such as CblC, can be tackled in time, and allow suffering people, such as Sara, to live a better, brighter life.

Bibliography

[1] J. Kim et al. "Decyanation of vitamin B12 by a trafficking chaperone". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.38 (2008), pp. 14551–14554. DOI: 10.1073/pnas.0805989105.

[2] D. S. Froese et al. "Mechanism of vitamin B12-responsiveness in cblC methylmalonic aciduria with homocystinuria". In: *Molecular Genetics and Metabolism* 98.4 (2009), pp. 338–343. ISSN: 1096-7192. DOI: 10.1016/j.ymgme.2009.07.014. URL: https://www.sciencedirect.com/science/article/pii/S1096719209002364.

[3] J. Kim et al. "A human vitamin B12 trafficking protein uses glutathione transferase activity for processing alkylcobalamins". In: *The Journal of Biological Chemistry* 284.48 (2009), pp. 33418–33424. DOI: 10.1074/jbc.M109.057877.

[4] *UniProt*. https://www.uniprot.org/uniprotkb/Q9Y4U1/entry.

[5] Markos Koutmos et al. "Structural Basis of Multifunctionality in a Vitamin B12-processing Enzyme". In: *The Journal of Biological Chemistry* 286.34 (2011), pp. 29780–29787.

[6] *AlphaFold Protein Structure Database*. https://alphafold.ebi.ac.uk/entry/Q9Y4U1.

[7] *ESpritz*. http://old.protein.bio.unipd.it/espritz/.

[8] *DisProt*. https://disprot.org/DP03867.

[9] *MobiDB*. https://mobidb.org/Q9Y4U1.

[10] *ClinVar*. https://www.ncbi.nlm.nih.gov/clinvar/.

[11] Buß Oliver et al. "FoldX as Protein Engineering Tool: Better Than Random Based Approaches?" In: *Computational and Structural Biotechnology Journal* 16 (2018), pp. 25–33. DOI: 10.1016/j.csbj.2018.01.002.

[12] Maria Plesa et al. "Interaction between MMACHC and MMADHC, two human proteins participating in intracellular vitamin B12 metabolism". In: *Molecular Genetics and Metabolism* 102.2 (2011), pp. 139–148. DOI: 10.1016/j.ymgme.2010.10.011.

[13] *CullPdb Dataset*. http://dunbrack.fccc.edu/lab/.

[14] Stephen F. Altschul et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Research* 25.17 (Sept. 1997), pp. 3389–3402. ISSN: 0305-1048. DOI: 10.1093/nar/25.17.3389. URL: https://doi.org/10.1093/nar/25.17.3389.

[15] Jan Zacharias and Ernst-Walter Knapp. "Protein Secondary Structure Classification Revisited: Processing DSSP Information with PSSC". In: *Journal of Chemical Information and Modeling* 54.7 (2014). PMID: 24866861, pp. 2166–2179. DOI: 10.1021/ci5000856. URL: https://doi.org/10.1021/ci5000856.

[16] *BioGRID*. https://wiki.thebiogrid.org/doku.php/aboutus.

[17] Djoerd Hiemstra. "N-Gram Models". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 1910–1910. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_935. URL: https://doi.org/10.1007/978-0-387-39940-9_935.

[18] J. Frazer, P. Notin, M. Dias, et al. "Disease variant prediction with deep generative models of evolutionary data". In: *Nature* 599 (2021), pp. 91–95. DOI: 10.1038/s41586-021-04043-8.

[19] Lerner-Ellis J et al. "Identification of the gene responsible for methylmalonic aciduria and homocystinuria, cblC type". In: *Nat Genet* 38 (2006), pp. 93–100. DOI: 10.1038/ng1683.

[20] *LOVD Global Variome shared: MMACHC entries.* https://databases.lovd.nl/shared/variants/MMACHC.

[21] Maria Plesa et al. "Interaction between MMACHC and MMADHC, two human proteins participating in intracellular vitamin B12 metabolism". In: *Molecular Genetics and Metabolism* 102.2 (2011), pp. 139–148. ISSN: 1096-7192. DOI: https://doi.org/10.1016/j.ymgme.2010.10.011. URL: https://www.sciencedirect.com/science/article/pii/S1096719210003707.

[22] Mucha P et al. "Vitamin B12 Metabolism: A Network of Multi-Protein Mediated Processes". In: *Int. J. Mol. Sci.* 25.15 (May 2024), p. 8021. ISSN: 1477-4054. DOI: 10.3390/ijms25158021. URL: https://doi.org/10.1093/bib/bbae178.

[23] John M. McBride et al. "AlphaFold2 Can Predict Single-Mutation Effects". In: *Phys. Rev. Lett.* 131 (21 Nov. 2023), p. 218401. DOI: 10.1103/PhysRevLett.131.218401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.131.218401.

[24] P Anthony et al. "Interaction of Glutathione with MMACHC Arginine-Rich Pocket Variants Associated with Cobalamin C Disease: Insights from Molecular Modeling". In: *Biomedicines* 11.12 (2023), p. 3217. DOI: 10.3390/biomedicines11123217. URL: https://doi.org/10.3390/biomedicines11123217.

[25] John M. McBride and Tsvi Tlusty. "AI-Predicted Protein Deformation Encodes Energy Landscape Perturbation". In: *Phys. Rev. Lett.* 133 (9 Aug. 2024), p. 098401. DOI: 10.1103/PhysRevLett.133.098401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.133.098401.

[26] S Wang et al. "Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields". In: *Sci Rep* 6 (2016), p. 18962. DOI: 10.1038/srep18962. URL: https://doi.org/10.1038/srep18962.

[27] Yu C H et al. "End-to-End Deep Learning Model to Predict and Design Secondary Structure Content of Structural Proteins". In: *ACS biomaterials science and engineering* 8.3 (2022), pp. 1156–1165. DOI: 10.1021/acsbiomaterials.1c01343. URL: https://doi.org/10.1021/acsbiomaterials.1c01343.

[28] Diaz D J et al. "Using machine learning to predict the effects and consequences of mutations in proteins". In: *Current opinion in structural biology* 78 (2023), p. 102518. DOI: 10.1016/j.sbi.2022.102518. URL: https://doi.org/10.1016/j.sbi.2022.102518.

[29] Maksimenko O and Georgiev P. "Mechanisms and proteins involved in long-distance interactions". In: *Front. Genet.* 5.28 (2014). DOI: 10.3389/fgene.2014.00028. URL: http://dx.doi.org/10.3389/fgene.2014.00028.

[30] Esser A J et al. "Versatile enzymology and heterogeneous phenotypes in cobalamin complementation type C disease". In: *iScience* 25.9 (2022), p. 104981. DOI: 10.1016/j.isci.2022.104981. URL: https://doi.org/10.1016/j.isci.2022.104981.

[31] Lisa Longo et al. "Missense mutations in MMACHC protein from cblC disease affect its conformational stability and vitamin B12-binding activity: The example of R161Q mutation". In: *Molecular Genetics and Metabolism* 145.3 (2025), p. 109150. ISSN: 1096-7192. DOI: https://doi.org/10.1016/j.ymgme.2025.109150. URL: https://www.sciencedirect.com/science/article/pii/S1096719225001416.

[32] D Sean Froese et al. "Structure of MMACHC Reveals an Arginine-Rich Pocket and a Domain-Swapped Dimer for Its B12 Processing Function". In: *Biochemistry* 51.25 (2012). PMID: 22642810, pp. 5083–5090. DOI: 10.1021/bi300150y. eprint: https://doi.org/10.1021/bi300150y. URL: https://doi.org/10.1021/bi300150y.

[33] D Sean Froese et al. "Structural Insights into the MMACHC-MMADHC Protein Complex Involved in Vitamin B12 Trafficking". In: *Journal of Biological Chemistry* 49 (2015), pp. 29167–77. DOI: 10.1074/jbc.M115.683268.

[34] *ClinVar Miner*. https://clinvarminer.genetics.utah.edu/variants-by-gene/MMACHC/significance/any.

[35] Marina A. Pak et al. "Using AlphaFold to predict the impact of single mutations on protein stability and function". In: *PLOS ONE* 18.3 (Mar. 2023), pp. 1–9. DOI: 10.1371/journal.pone.0282689. URL: https://doi.org/10.1371/journal.pone.0282689.