Master Thesis

Computational Design of Functional 2D and 3D Assemblies for Photovoltaic Enzymes

Christopher K. Weingarten

M.Sc. Biotechnology

Department of Biosystems Science and Engineering

ETH Zurich

14th February 2024

Co-Examiner:

Dr. Adrian Bunzel

Group Leader and
SNSF Ambizione Fellow

Department of Biosystems
Science and Engineering

ETH Zurich

Examiner:

Prof. Sven Panke

Professor of Bioprocess
Engineering

Department of Biosystems
Science and Engineering

ETH Zurich

ETH zürich

Department of Biosystems
Science and Engineering

Table of Contents

Acknowledgments

First and foremost, I want to thank Dr. Adrian Hans Bunzel for enabling me to work on this project. It was an amazing experience and only deepened my interest in the field of computational chemistry and protein design. I appreciated the balance between learning from Adrian's expertise in this field while also being given enough space to try and figure things out myself. I will always look back on this time with great fondness.

I also want to thank Prof. Dr. Sven Panke for hosting me in his lab. This project would not have been possible without him and the amazing lab he has set up.

A big thanks also goes out to the BioEM facility of the university of Basel and their group head, Dr. Mohamed Chami, who provided the instruments and technical expertise to make the EM analysis possible. I specifically want to thank Carola Alampi for handling the EM grids and taking the images, as well as Dr. Rémi Ruedas and Dr. Carlos Fernandez Rodriguez for doing the analysis on the collected datasets.

Additionally, I want to thank the entire Bioprocess Laboratory for their warm welcome. The good discussions and quick banter in and out of the lab made this project all the more enjoyable. I especially want to thank my colleagues in the Bunzel subgroup, namely Eleftheria Kelefioti Stratidaki, Philipp Elbers, and Jannik Neumann, for showing me the ropes and making me feel at home in the group.

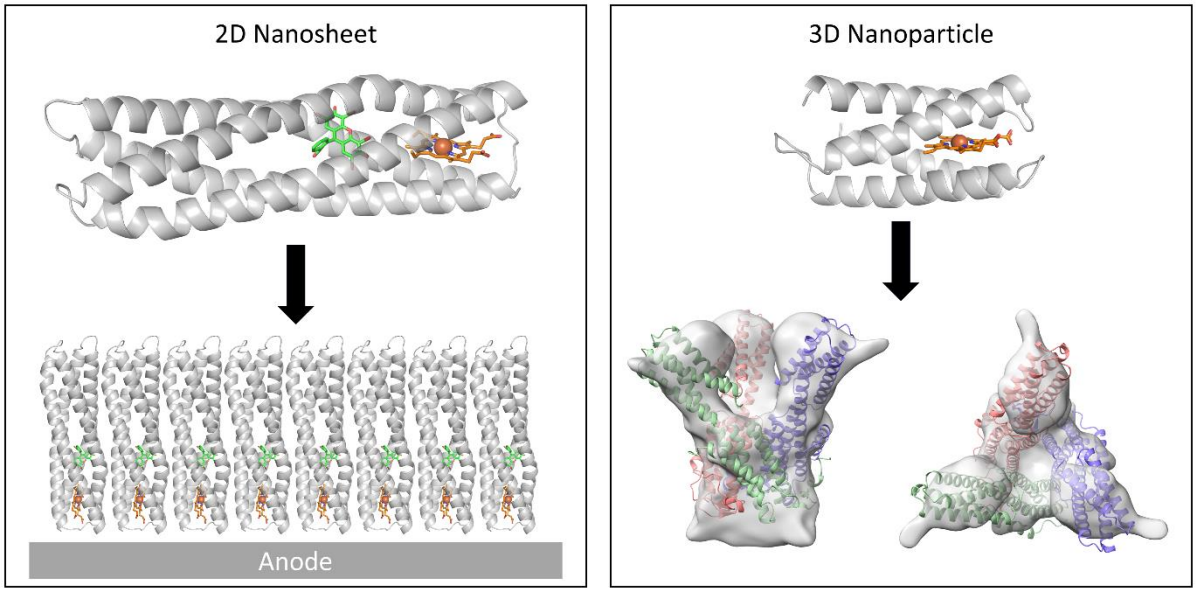
Finally, I want to thank my friends and family who supported me through this project and listened to my ramblings in their free time. I could not have done it without them.

Abstract

Photovoltaics may allow meeting the population's evergrowing energy consumption while reducing greenhouse gas emissions through fossil fuels. To provide a more sustainable avenue towards photovoltaics, the de novo photoenzyme EOY4D2.2 was recently created for application in biohybrid solar cells. EOY4D2.2 was engineered by design and evolution of a binding site for the small-molecule dye eosin Y into a heme protein, which has resulted in improved efficiency and stability in a proof-of-concept solar cell. This work aims to exploit nanoscale engineering to enhance the efficiency of EOY4D2.2 and to provide a robust framework for cryo-EM structure determination.

Solar efficiency would likely benefit from highly structured protein complexes with tight interactions between the protein and photoelectrodes. To this end, a computational pipeline was established to design 2D protein nanosheets and screen them in silico by molecular dynamics simulation. Further engineering of EOY4D2.2 is currently limited by a lack of experimental structural data. To overcome this engineering bottleneck, a 3D nanoparticle design approach was established, where a large protein scaffold serves as a scaffold for structural determination of N- or C-terminally linked photoenzymes. Following this approach, a low-resolution electron density map from a small test dataset was obtained that demonstrates the viability of this approach.

In conclusion, this work showcases the potential of using protein complexes for structural determination of EOY4D2.2 variants as well as increasing the efficiency of these systems. The 3D nanoparticle will help with the design of further variants based on EOY4D2.2, allowing for quick and easy structural determination, aiding in further design efforts. The established pipeline for 2D nanosheet development will aid in the design of more and better-performing protein arrays. Harnessing nanoscale assembly will allow for substantial improvement in photoefficiency, paving the way towards commercially viable biohybrid solar cells.



1 Introduction

1.1 Protein Nanostructure Design

Proteins are the workhorse in biology, performing an impressive array of different functions, including structural support and modification of the cell, recognition of and action upon various molecules, as well as catalysis of most reactions inside the cell. The function of a protein is defined by its structure, which can be broken down into four hierarchical levels. The amino acid sequence of a protein is its primary structure. Based on their sequence, proteins form distinct secondary structure elements, namely α-helices and ß-sheets. The tertiary structure of a protein is defined as the final total fold of a single peptide chain. For monomeric proteins, this is the final structural level. However, certain proteins interact with other proteins to spontaneously create larger, multimeric assemblies, forming quaternary structures.[1] Quaternary structures allow for a whole new avenue of functions in biology, from spatially bringing together polypeptide chains with different functions to create reaction cascades, to assembling large nanostructures to store molecules or form structural scaffolds (Fig. 1).[2]

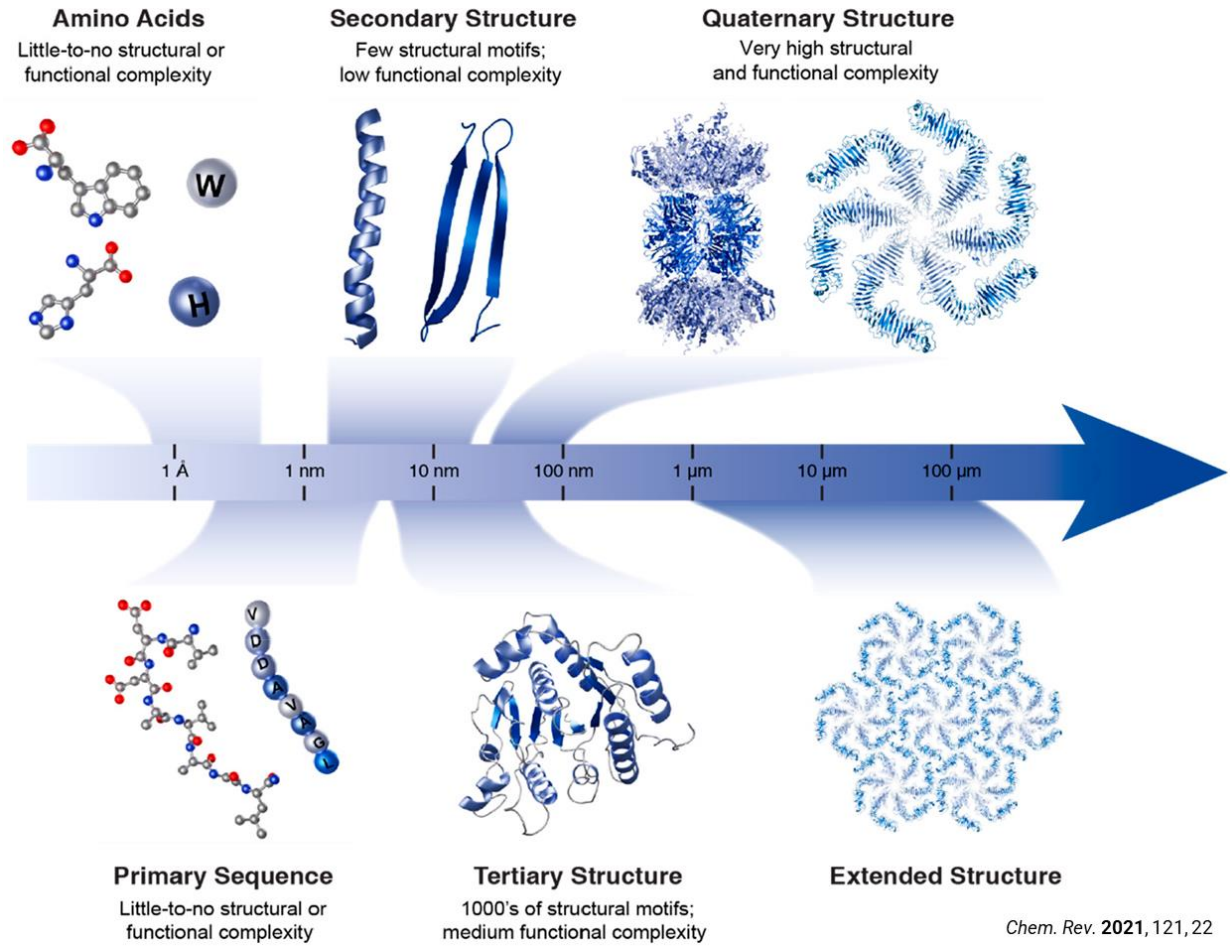


Fig. 1 | Protein structural hierarchy on scale. The four different levels of the hierarchical structure of proteins set on scale, including the primary structure as amino acid sequence, the secondary structure as α-helices and ß-sheets, the tertiary structure as the final fold of a single peptide chain, and finally the quaternary structure, spanning over three orders of magnitude in size when including repeating structures.[3]

Given the functions emerging from tertiary protein assemblies, there has been great interest in controlling and designing self-assembly of proteins. This has, however, proven to be quite challenging. The prediction of the structure of a single polypeptide chain based on its amino acid sequence has long been seen as a central challenge in protein sciences.[4] Recently, the use of machine learning, especially deep learning-based methods, in protein structure prediction has gotten a lot of attention. In 2020, AlphaFold2 highly outperformed other models in CASP14, reaching a median backbone accuracy of 0.96 Å r.m.s.d.$_{95}$ (Cα root-mean-square deviation at 95% residue coverage).[5] Predicting large multimeric assemblies, however, has proven to be more difficult due to the increase of factors influencing self-assembly, such as protein concentration, solvent conditions, as well as higher degrees of freedom due to the absence of any inherent steric constraint as in monomeric proteins, where all residues are connected through a single peptide backbone.[3] Nevertheless, significant progress has been made with machine learning algorithms to predict multimeric structures, such as AlphaFold-Multimer, an AlphaFold2 model trained specifically for multimeric inputs.[6]

On the other hand, protein design works on the inverted problem of protein structure prediction, i.e., instead of finding a structure for the given protein sequence, the aim is to find a sequence that stabilizes a specific structure.[7] One software commonly used for this purpose is Rosetta, a molecular modeling program used for analysis and design of protein structures.[8] To perform design, Rosetta defines a score function which combines several physics-based metrics, such as the Lennard-Jones potential to model van der Waals forces, as well as statistical and knowledge-based metrics, e.g., by including parameters for hydrogen bond potentials to favor angles and distances typically found in nature. Rosetta then stochastically searches for well-scored sequences using a Monte-Carlo based sampling algorithm to find optimal rotamers and amino acids for each position. Constraints can be set by the user to reduce the search space, allowing for a more focused and rational design than without constraints. For instance, during the design of protein oligomers, mutations can be limited to a few residues at the protein-protein interfaces, preventing the redesign of residues with minimal effect.[9] In the past, Rosetta has successfully allowed for the creation of a range of novel protein assemblies, including closed assemblies such as dimers[10] and protein cages[11], or repeating assemblies, including 1D filaments[12] and 2D arrays[13] (Fig. 2).

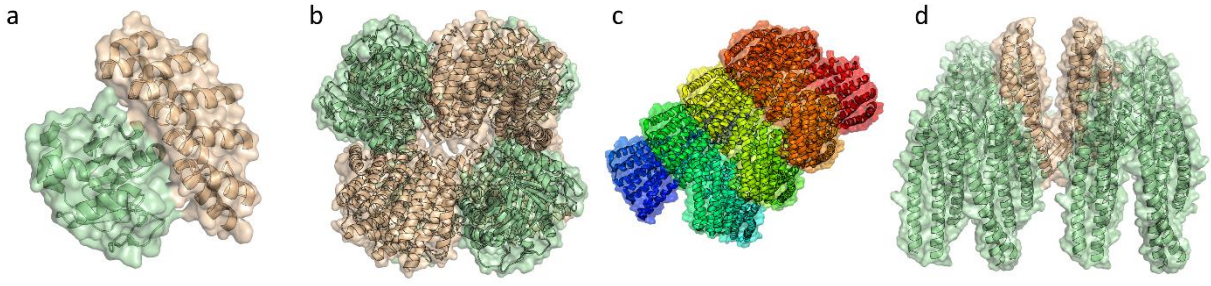


Fig. 2 | Examples of previously designed protein complexes using Rosetta. a) Protein dimer Pdar_PrbC10 (PDB: 3Q9N)[10] **b)** protein cage T33-21 (PDB: 4NWP)[11], **c)** 1D Helical filament DHF38 (PDB: 6E9Y)[12], and **d)** 2D homomeric array p4Z_9[13]

Upon the advent of AlphaFold2 and accurate protein structure prediction, protein engineers quickly realized the potential of these AI-based structure prediction tools for application in the inverse problem of protein design. Just two years later, ProteinMPNN was released, a deep learning-based protein sequence designing method implementing a message-passing neural network architecture. ProteinMPNN consistently outperformed Rosetta in overall native sequence recovery, measured by calculating the percent identity between the native and predicted conformation of the designed structure, at around 50%, while also reducing computational cost. In the case of multimeric protein design, protein-protein interface residues saw similar results in sequence recovery, allowing for complex design.[14] However, ProteinMPNN is currently still unable to design proteins with complex symmetries, requiring protocols combining docking and symmetry methods from other softwares. For instance, it has been recently shown that a hybrid pipeline with Rosetta and ProteinMPNN could be used to facilitate *in vitro* assembly of previously designed nanoparticles.[15]

1.2 Photovoltaics

Photovoltaics aims to harvest light for the solar-driven generation of electricity. In the face of humanity's ever-growing energy demands as well as climate change driven by greenhouse gas emissions from fossil fuel usage, solar power provides a promising avenue to sustainably produce energy while reducing green-house gas emissions. The earth is continuously struck with $>10^5$ TW in the form of solar radiation on average, which is impressive given that the average global electricity demand is 2.9 TW.[16] However, only 4.6% of the global electricity production came from solar in 2022, composing 2.1% of the global primary energy consumption.[17] This demonstrates the potential for higher electricity generation through photovoltaics.

Photovoltaic cell technologies are split into four different generations (Fig. 3a). Solar cells from the first generation, defined by the use of thick crystalline layers composed of silicon, are the most common commercially used cells. Currently, the dominating technology is polycrystalline silicon modules, capturing a market share of over 97%.[18] While these cells have an efficiency of ~20%, optimizing this efficiency (e.g., through formation of monocrystalline silicon cells) can lead to high manufacturing costs, making more efficient cells from the first generation commercially less attractive. Second generation photovoltaic cells, also referred to as thin film photovoltaic cells, introduced alternative materials to silicon, such as cadmium telluride (CdTe) or copper indium gallium selenide (CIGS). However, these materials are either too expensive or too toxic for large scale production, making silicon-based solar cells the preferred technology in most cases.[19]

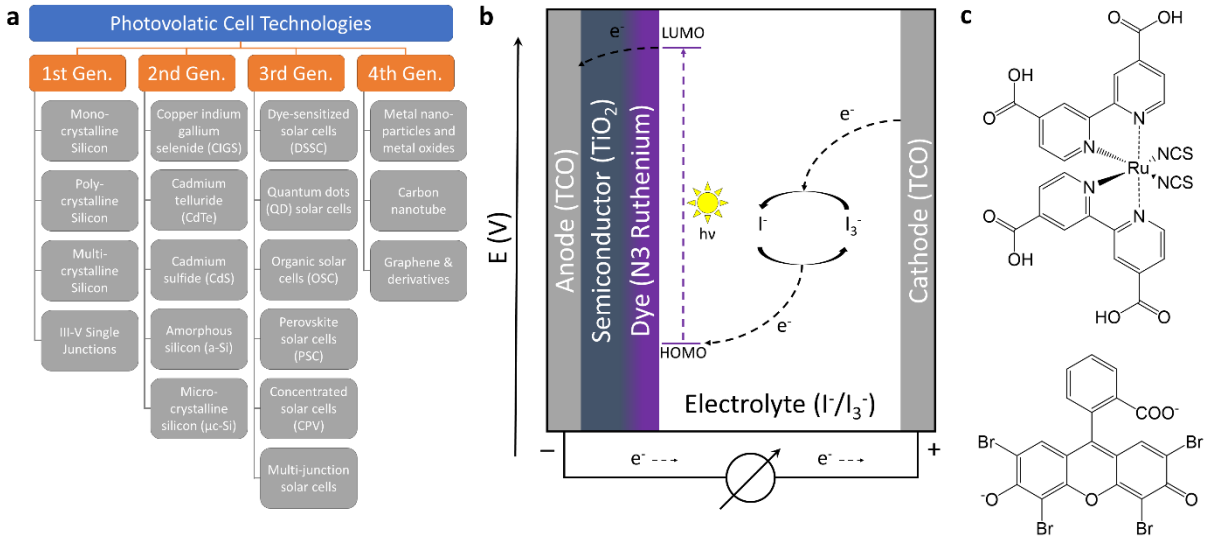


Fig. 3 | Different photovoltaic cell technologies with a focus on dye-sensitized solar cells. a) The four generations of photovoltaic cell technologies with the major technologies listed for each generation.[19] **b)** Schematic of a classical architecture of a dye-sensitized solar cell. **c)** N3 ruthenium dye (top), an example of a dye typically used in DSSCs, and eosin Y (bottom), a simple and abundantly available dye commonly used in photochemistry.

In the third generation, solar cells with a wide range of approaches and materials moving away from the first and second generation emerged, comprising both inexpensive low-efficiency and expensive high-efficiency systems.[19] One technology of note is dye-sensitized solar cells (DSSCs). The first of these cells was manufactured by Michael Grätzel in 1991, with an efficiency of 7%. The main component of the DSSC is the dye or photosensitizer which is adsorbed onto a semiconductor on the working electrode to form the photoanode. Upon absorption of photons by the dye, electrons get excited from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO). The excited electrons are injected into the semiconductor, typically porous TiO_2, leaving the dye in an oxidized state. The electrons travel from the anode to the cathode, where an electrolyte such as iodine is reduced to facilitate the regeneration of the dye to close the redox cycle (Fig. 3b). The choice of dye for this technology is crucial, as the efficiency of DSSCs depends on the dye's absorption spectrum, HOMO-LUMO gap, and excited state lifetime. Additionally, the dye needs to work with other factors in the cell, requiring proper parameters such as the difference in redox potentials compared to the electrolyte and semiconductor, as well as the injection speed for dye stability. Dyes based on transition metal complexes, such as the ruthenium complex N3 (Fig. 3c), have found wide application in DSSCs. The extended absorption spectra of transition metal complexes, their long lifetime of the excited state, as well as their high photostability, have made these complexes ideal candidates for DSSCs. Ruthenium dyes were used in some of the first DSSCs and have been some of the best-performing examples with respect to solar cell efficiency and lifetime. These complexes, however, are typically based on toxic metals with high environmental impact and production costs, hindering large-scale production.[20]

Finding competitive organic dyes has been an ongoing effort in DSSC engineering. Attempts with simple and abundant organic compounds, such as eosin Y[21] (Fig. 3c) or chlorophyll,[22] typically suffer from

limited stability and photo efficiency. Remarkably, examples of DSSCs using co-photosensitization of larger organic dyes have shown efficiencies of up to 15.2%, setting a new record for DSSCs.[23] Nevertheless, the size and complexity of these dyes increase their synthesis cost, which limits the economic viability of these systems.

1.3 EOY4D2.2 – Function & Design

This thesis is based on the previously engineered photoenzyme EOY4D2.2, a four-helix bundle protein binding heme as well as the organic dye eosin Y (EOY).[24] EOY4D2.2 has shown promise for improving the performance of DSSCs by protecting EOY from degradation. This has been previously tested in a model solar cell, in which EOY4D2.2 showed reduced photobleaching compared to EOY alone (Fig. 4a).

The design of EOY4D2.2 was based on another synthetic protein, 4D2, which was originally designed based on the transmembrane D_2-symmetric diheme four-helix bundle in cytochrome bc_1. Mimicking the structure of bc_1, an α-helical peptide was designed that self-assembled into a tetrameric four-helical bundle upon the addition of heme.[25] These four α-helical peptides were then fused together with a flexible linker for *in vivo* expression and to enhance heme affinity by preorganizing the binding sites.[26] The resulting protein was dubbed 4D2, for its 4-helix bundle and *pseudo* D_2 symmetry. In the same work, two different variants of this protein were designed computationally, including m4D2, a monoheme variant of 4D2, and e4D2, an extended variant with four heme binding pockets. An X-ray crystal structure could be achieved for 4D2, however crystallization proved difficult for the other two variants, a problem found to be typical for 4D2-like proteins. While design of 4D2-derived variants has gone forward, the lack of a consistent structural determination method for these proteins makes structural validation of new designs difficult, limiting further engineering efforts.

EOY4D2 was computationally designed from 4D2 by redesigning one of the heme binding sites to a binding pocket for the dye EOY. Design was performed using RosettaMatch and RosettaDesign to design the binding pocket, which was followed by an *in silico* screening step with molecular dynamics (MD) simulations. The best variants were experimentally tested for binding, which resulted in EOY4D2 binding EOY with an affinity of ~10 µM. After two rounds of directed evolution aiming to enhance EOY binding, the binding affinity was reduced to ~500 nM, a variant dubbed EOY4D2.2.[24]

To exploit the modularity of the helical bundle scaffold, an extended m4D2 variant with a terminal heme and three designable modules was subsequently created (Fig. 4b). With this extension, the nomenclature for the variants was changed to match the modularity of the proteins, with one letter corresponding to each binding site. The new variant was called XXXH, with "X" standing for an unfunctionalized module and "H" for a heme binding pocket. This extension would allow for the design of potentially four different binding sites in the protein. Initial functionalization was achieved by transplanting the EOY binding pocket from EOY4D2.2 into the X module adjacent to the heme in XXXH, resulting in the XXE2H, where "E2" stands for the second design iteration for the EOY binding pocket.

Unfortunately, XXE2H showed drastically reduced EOY affinity, most likely because EOY forms tight interactions with the loop in EOY4D2.2, which could not be realized in the new protein. Thus, the E2 binding site was computationally redesigned using Rosetta and screened *in silico* with MD simulations, generating a third iteration of the EOY binding pocket, "E3".

While the binding affinities of the new variants are still to be verified, the extended XXE2H would allow for additional binding sites for small molecules, such as an electron donor to speed up regeneration of the dye, offering a distinct advantage over the EOY4D2.2 variant. Given the large improvements seen in binding affinity in EOY4D2.2 after directed evolution, there is a good chance that similar improvements will be seen after further finetuning in XXE2H.[27] Directed evolution is the first and most obvious option to further improve binding affinity of the variants to EOY. However, by designing highly ordered protein nanostructures, the entire photovoltaic system could be improved beyond just the protecting qualities of the protein.

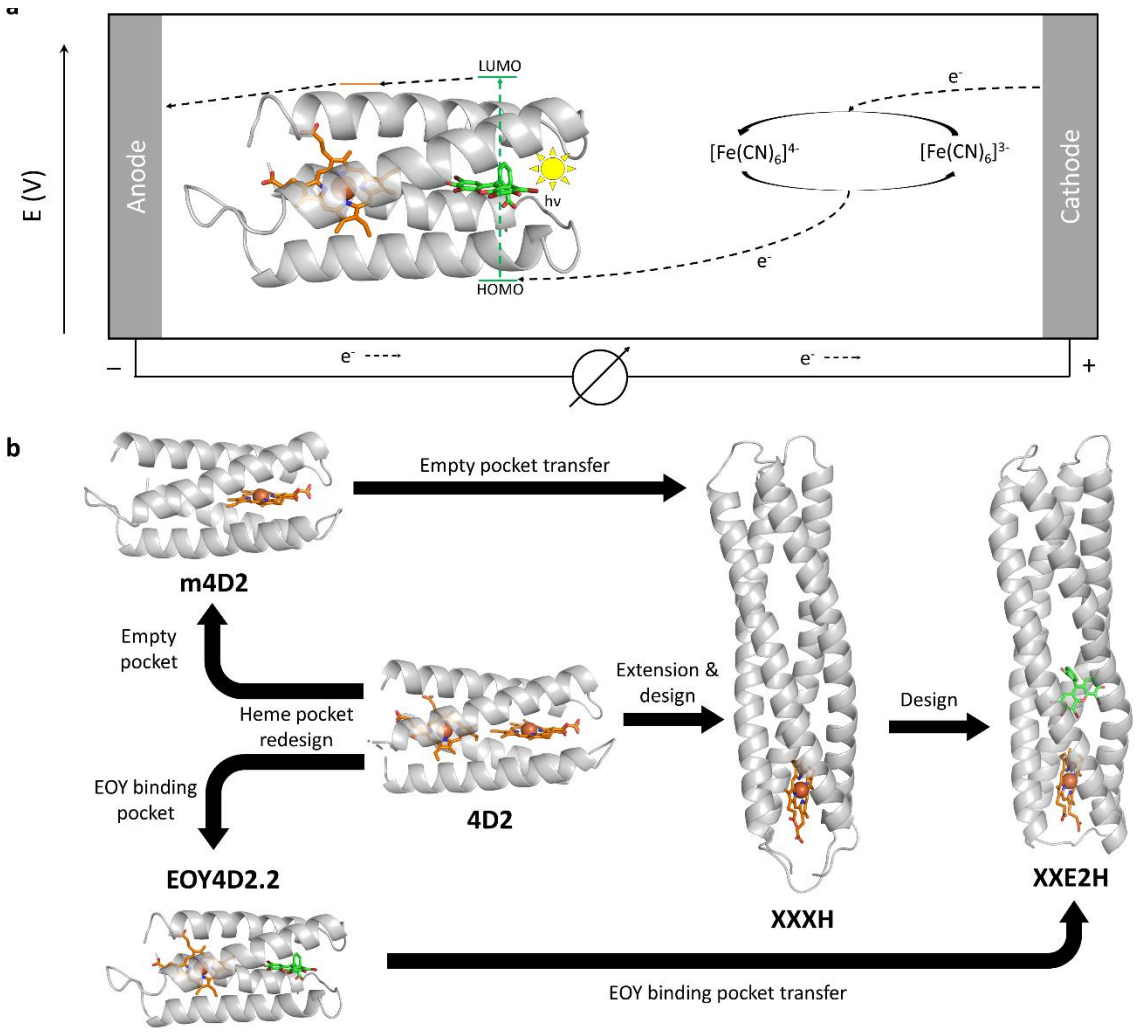


Fig. 4 | EOY4D2.2 working principle and design. a) Schematic overview of the benchtop setup used to test EOY4D2.2 in a photovoltaic system. **b)** Design overview of the 4D2-derived variants relevant to this work, including redesign of binding pockets in 4D2 (EOY4D2.2, m4D2) as well as extended variants with transferred binding pockets (XXXH, XXE2H).

Note: The naming scheme of the extended version was changed during the course of the work. Some original data might point to (X2)E2H or X2E2H, which is identical to XXE2H.

2 Aims

Dye sensitized solar cells (DSSCs) have shown great promise as a sustainable and cheap alternative to silicon-based photovoltaics. Although DSSCs achieve efficiencies of up to 15.2%,[23] they still lag behind other photovoltaics in terms of efficiency and stability. In previous work, Bunzel *et al.* have shown an increase in photostability of the photosensitizer EOY by design and evolution of a binding site for the dye in the 4D2 scaffold. Their photovoltaic protein, dubbed EOY4D2.2, showed lower levels of photobleaching of EOY during testing in a benchtop photovoltaic setup, demonstrating the potential of bioengineering to improve DSSCs.[24]

Building on previously designed proteins, this work will focus on the design of highly ordered assemblies with two main goals in mind: (i) increasing the efficiency of the previously described photovoltaic system by designing a self-assembling 2D-nanosheet of 4D2-derived variants, and (ii) facilitating the structural determination of 4D2-derived variants with cryo-EM by linkage of the protein to a 3D-nanoparticle.

2.1 Increase Photoefficiency with 2D-Nanosheet Design

In the current photovoltaic setup for EOY4D2.2, the protein is dissolved between two electrodes. Electron transport, and thus photoefficiency, would likely benefit from densely adsorbing the protein onto the anode, resulting in a system similar to established DSSCs. To achieve this, the previously designed XXE3H variant will be redesigned in Rosetta to induce spontaneous self-assembly into a 2D-array (Fig. 5a). Promising variants will be screened *in silico* by molecular dynamics simulations and the best variants will be probed for nanosheet formation.

2.2 Facilitate Structural Determination with 3D-Nanoparticle Design

Given some promising cryo-EM structures of the extended e4D2 variant[28], we hypothesize that displaying the protein onto a larger protein scaffold could substantially improve its resolution. To that end, 4D2-derived variants will be linked to a larger, multimeric scaffold mimicking a previously described approach for structure-determination of small proteins.[29] Computational design will be exploited to enhance the rigidity of the variant while also introducing symmetry, both important factors for cryo-EM (Fig. 5b). Using AlphaFold2 and ProteinMPNN, a suitable scaffold will be designed, using the single-heme m4D2 variant as a proof of concept.

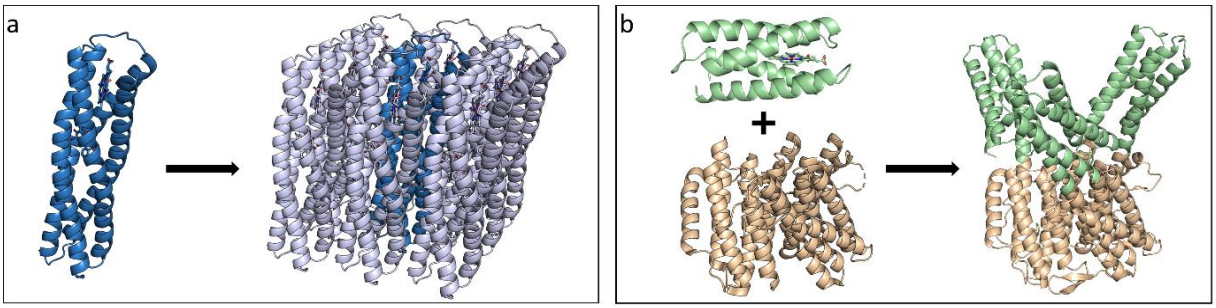


Fig. 5 | Aims of this thesis. This work aims at **(a)** Increasing the efficiency of photovoltaic systems with 2D-nanosheet design and **(b)** facilitating structural determination through cryo-EM with 3D-nanoparticle design.

3 Results & Discussion

3.1 2D-Nanosheet Design

The *de novo* photoenzyme EOY4D2.2 has previously been successfully applied to drive a photocurrent in a model solar cell.[24] This success demonstrates the potential of the designed four-helix bundle protein for application in biohybrid solar cells and enables various avenues for further optimization. One promising strategy, inspired by classical DSSCs, is to have the photoenzyme adsorbed to the anode in a tightly packed 2D array. The localization of the photoenzyme would reduce the distance between redox cofactors and electrodes, thereby enabling faster electron transfer upon excitation in the dye. A tightly packed 2D array would furthermore maximize the protein-electrode interaction, potentially preventing non-productive recombination events. To achieve self-assembly of a 2D array, favorable protein-protein interfaces needed to be designed to drive association.

Designing novel protein-protein interfaces can drive the self-assembly of proteins into order complexes.[3] In the case of 2D array designs, only a few examples have been published by the Baker lab showing successful design of 2D protein arrays using Rosetta that were experimentally validated by transmission electron microscopy (TEM) or atomic force microscopy (AFM)[13]. Here, a similar approach was taken by combining Rosetta design with *in silico* screening through MD simulations. MD was used to narrow down the number of selected variants to proteins showing rigid and defined interfaces. Finally, the most promising variants were expressed recombinantly and tested for 2D-array formation.

3.1.1 Characterizing Parental Variant

Previously, an extended version of EOY4D2.2 dubbed XXE2H was designed in the Bunzel group to allow for more binding sites to be designed in the future. However, EOY forms several interactions with the loops connecting the helices in EOY4D2.2, which could not be realized in XXE2H because the EOY binding site is located at the center of the protein. Thus, XXE2H suffered from low EOY affinity. To enhance EOY binding, the binding site was redesigned using Rosetta with the same strategy used to design the initial binding site.[24] The resulting variants were then screened *in silico* by MD simulations, resulting in three promising variants XXE2H-1, XXE2H-2 and XXE2H-3.[27] Since this work builds on the extended variants, EOY binding had to be validated for the selected variants.

EOY binding was measured using two different methods, circular dichroism (CD) and isothermal calorimetry (ITC). Both methods were previously established to measure the binding affinity of 4D2-derived variants to EOY.[24] CD measurements were conducted by titration of EOY into the designs (Tab. 3) and measuring the CD absorption between 380 nm and 600 nm (Fig. 6a-c). The K_d was then determined by fitting a quadratic binding equation (Eq. 3), with representative wavelengths for binding at 420 nm and 545 nm (Fig. 6d-f). The titration showed two binding signals: one stemming from specific binding to the EOY binding site and the other resulting from weak non-specific interactions with the

protein. The strongest binder was identified to be XXE2H-2, with a K_d of 1.96 µM. Weaker binding was also observed for XXE2H-1 (11.93 uM) and no binding could be detected for XXE2H-3. However, in all cases the intensity of the non-specific signal strongly outweighs the binding signal, and the binding signal is very noisy. This most likely leads to inaccuracies in the fitting and error-prone K_d values.

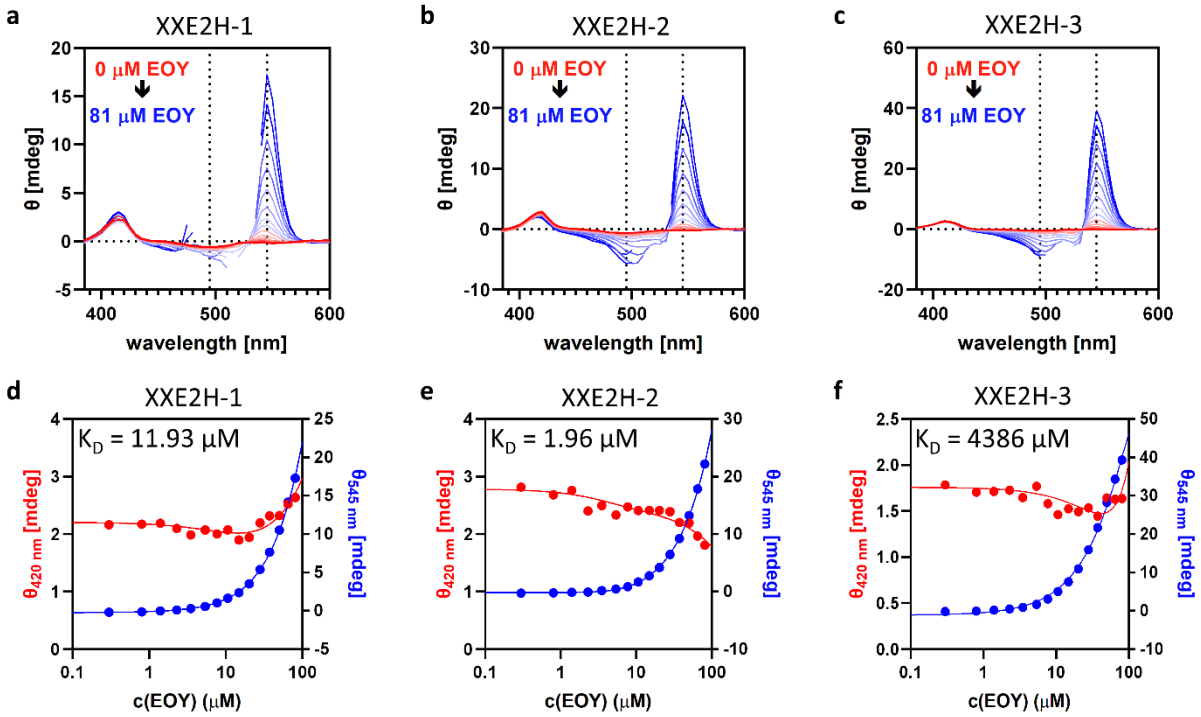


Fig. 6 | CD titration curves of the three XXE2H variants. a-c) CD absorption spectra with the light wavelength in nanometers on the x-axis and the molar ellipticity on the y-axis. Curves go from 0 µM EOY (red) to 81 µM EOY (blue) with constant protein concentrations. **d-f)** The molar ellipticity of representative wavelengths for binding (420 nm, red) and non-specific signal (545 nm, blue) plotted against EOY concentrations. The K_d is calculated from a quadratic binding equation **(Eq. 3)**.

ITC measurements for all three variants were conducted to confirm the CD results and get reliable K_d values. The differential power (DP) peaks indicate strong binding in XXE2H-2, some binding in XXE2H-1, and no binding in XXE2H-3 (Fig. 7). The peaks were integrated to get the heat of injections. Binding curves were fitted using a single binding site model, resulting in a K_d of 14.1 µM for XXE2H-2, somewhat higher than the K_d estimated from CD. Because of the high non-specific signal in the CD spectra, as well as the excellent fit of the ITC data, the K_d determined from ITC is likely more reliable. In any case, both the CD and the ITC data agreed that the XXE2H-2 variant was the strongest binder among the three variants. Thus, this variant, dubbed XXE3H, was used for further design of the 2D-protein array.

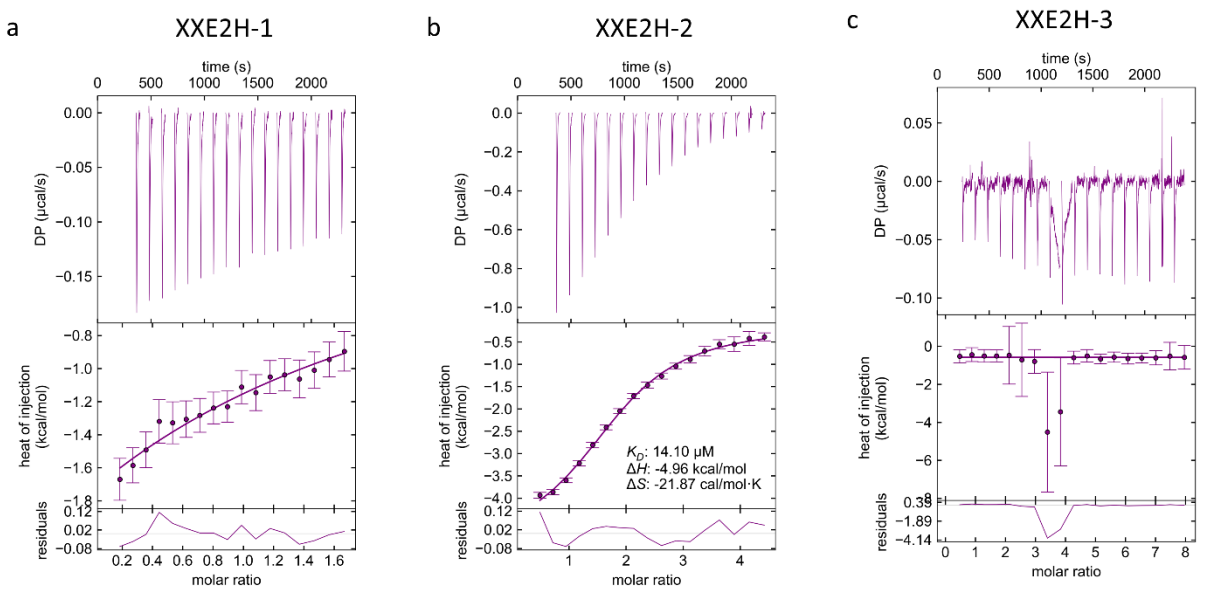


Fig. 7 | ITC data for the three XXE2H variants. ITC data for variants **a)** XXE2H-1, **b)** XXE2H-2, and **c)** XXE2H-3. The plots show the baseline-corrected differential power peaks (top), heat of injection datapoints calculated by integration of each peak (middle), as well as the residuals of each peak to the fit (bottom). For XXE2H-2, for which the fit was most reliable, the K_d, ΔH, and ΔS values determined by the fit are also noted.

3.1.2 Rosetta Docking & Design

Given the successful redesign of XXE2H to increase EOY binding affinity, the resulting variant XXE3H was used for protein nanosheet design. To that end, an ensemble of input structures was taken from previously run MD simulations for *in silico* screening. From this 10 ns simulation, 100 snapshots of the structure were extracted from the second half of the simulation, i.e., in the last 5 ns. With these 100 structures, design was performed similar to a previously published protocol for 2D-protein array design. The protocol included a docking step, where the rigid bodies are brought together, and a design step, optimizing the interface energies to allow for self-assembly.[13]

Docking. Prior to designing the protein-protein interfaces of XXE3H to drive self-assembly into a 2D protein array, the photoenzyme needed to be aligned to copies of itself in an orientation that maximizes the protein-protein interface area while avoiding any steric clashes. This assembly was described by a periodic lattice, with the photoenzyme acting as a subunit in an infinitely repeating plane. To dock the photoenzymes into 2D nanosheets with Rosetta, a symmetry definition file was written, which defines how the subunits are related to each other, the degrees of freedom for the movement between subunits, and how the score of the system is calculated.[13] To represent the unit cell of the 2D array, eight symmetrical subunits were placed around a central subunit and connected through a series of jumps (Fig. 8a). The complete score of the resulting unit cell is determined by five terms: the score of a single subunit, horizontal and vertical interaction scores comprising most of the protein-protein interaction interface, as well as two diagonal interaction scores (see Fig. 8b for a definition of the directional terms). The scores were weighted based on how much they contribute to the unit cell, with subunits and interactions on the edge of the cell being counted ½ and the subunits in the corner ¼. By summing all these terms, the total score E_{tot} could be calculated by Eq. 1,

$$E_{tot} = 4 * E(S0) + 4 * E(S0 \leftrightarrow S1) + 4 * E(S0 \leftrightarrow S2) + 4 * E(S0 \leftrightarrow S5) + 4 * E(S0 \leftrightarrow S6) \quad \text{Eq. 1}$$

where E(S0) is the total score of the central subunit, and the other terms are the interaction scores between the central subunit and the surrounding subunits.

With the symmetry file set up, the docking process was started. The symmetrical docking protocol in Rosetta was used in conjunction with a Monte Carlo mover to sample 20 docked conformations for each input structure.[30] To bring the protein backbones as close as possible to each other, the protein sequence was converted to poly-alanine before the docking step and reverted to the original residues afterwards. The scores of each docked conformation were calculated in a reduced centroid representation, and the best scoring structure among the 20 docked poses was used for further design in a continuous protocol. Next to the centroid scoring, the best docked structure was also assessed by counting the contacts between the poly-alanine subunits. These contact counts ranged from ≈600 to ≈1600, with a peak at ≈1000 (Fig. 8d). Despite the wide distribution, the generally high number of contacts indicated successful docking of the structures.

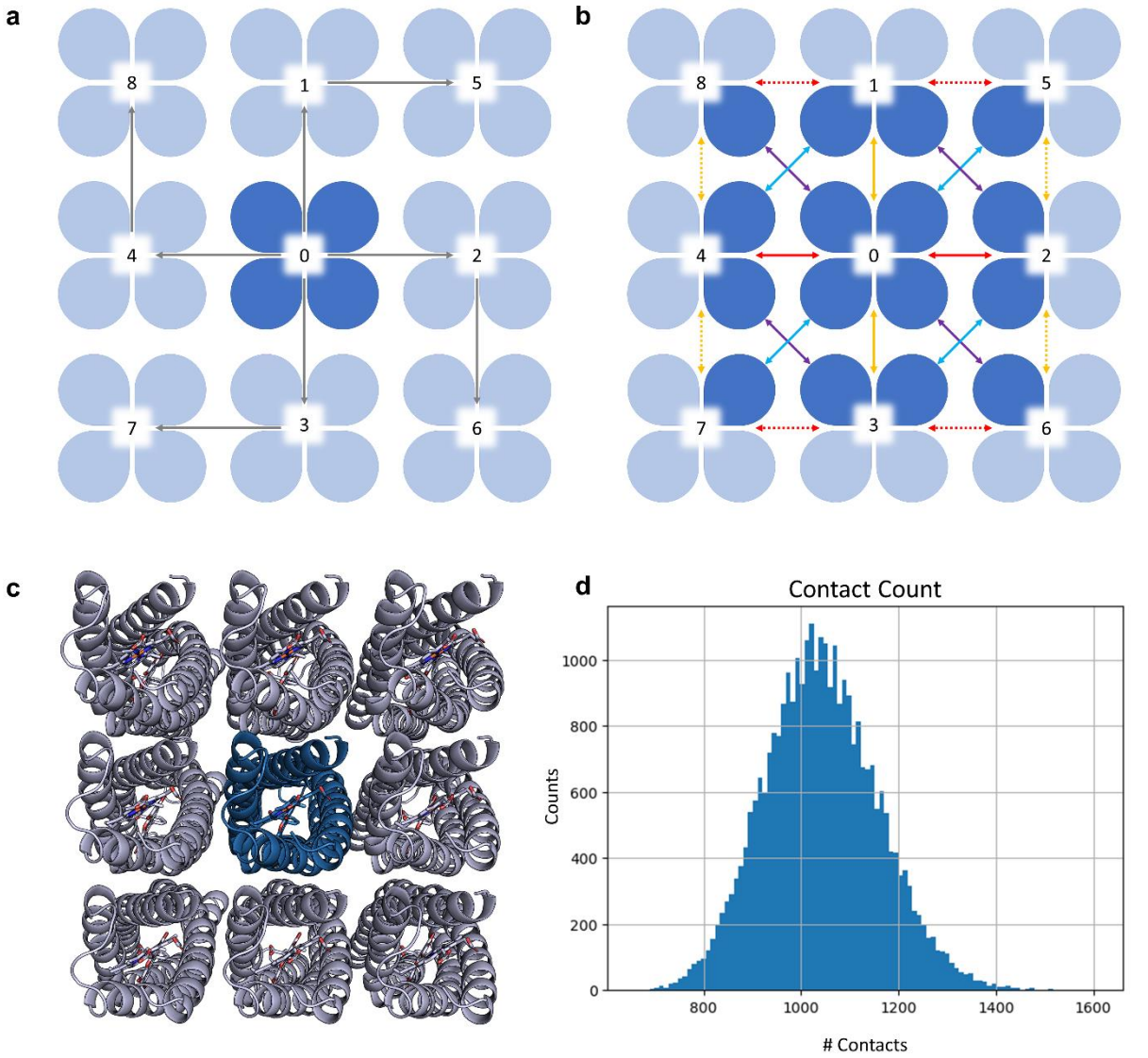


Fig. 8 | Rosetta symmetry definition and docking results. a) The 8 outer subunits (light blue) are connected to the central subunit (dark blue) through a series of jumps, which are grouped into 2 jump groups (vertical & horizontal). **b)** Subunits and interactions considered in the total score (Eq. 1). The subunits considered during scoring are shown in dark blue, summing up to a total of 4 subunits. The arrows show all interaction scores, including horizontal (red), vertical (yellow), and two diagonal (blue & purple) interactions. The color of the arrows matches the color of the corresponding term in Eq. 1. Each interaction term has 4 interactions in the unit cell, with the dotted arrows being considered as half an interaction. **c)** An example of a docked structure colored as in the graphical depiction in (a). **d)** Histogram of contacts counts with the number of contacts between different subunits after docking on the x-axis.

Design. The goal of the design was to optimize the amino acid sequence of the protein given the docked structure without reducing the binding affinity to the dye. To achieve this, 121 residues that could potentially contribute to the interface between the different subunits were chosen to be designed, while the remaining residues were fixed throughout the design process. Additionally, the 9 residues involved in the interaction with EOY were prevented from being redesigned (Fig. 9a,b). To maintain the hydrogen bond network involved in binding EOY, distance constraints were added between atoms forming a hydrogen bond within the network (Fig. 9c). These restrictions were aimed at conserving the already existing properties of the protein, specifically the heme and EOY binding affinities, during the design of the protein-protein interfaces.

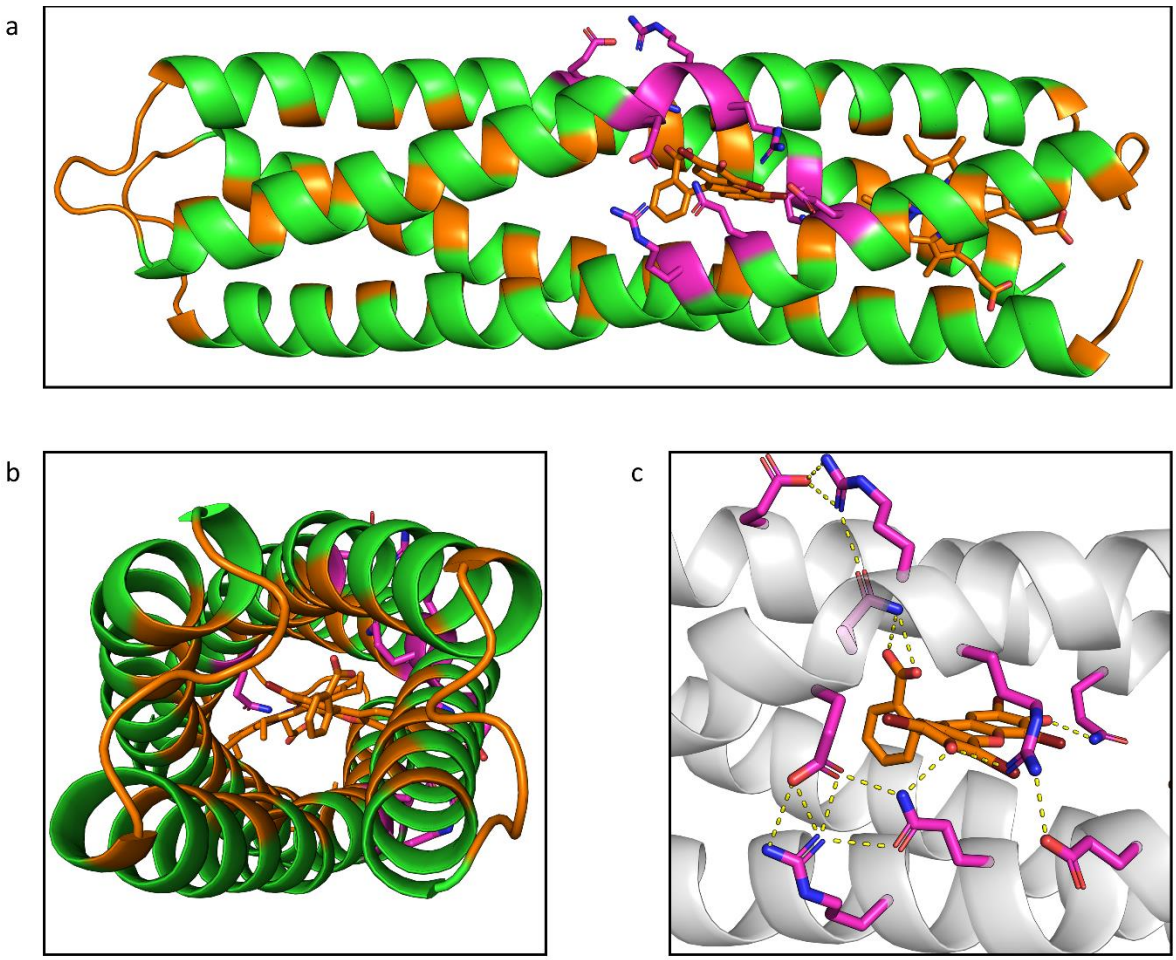


Fig. 9 | Residues selected for interface design. Interface residues (green) in XXE3H were allowed to be redesigned, while the loop and core residues (orange) were restricted solely to moving. A single subunit is shown both in a **(a)** side and **(b)** top view. Residues binding to EOY or participating in the connecting hydrogen bond network (magenta) were also restricted to moving, with additional distance constraints implemented between hydrogen bonding atoms. **c)** Hydrogen bonds involved in the network binding EOY are shown as yellow dotted lines.

The design process itself comprised three iterations of FastDesign[31,32]. In the first iteration, movement of the subunits along their connecting jumps was disabled to enforce the docked confirmation. Additionally, the amino acids available to mutate to were limited to a subset of small amino acids (Ala, Asn, Asp, Glu, Gly, Ile, Leu, Ser, Thr, Val) to encourage close and rigid interactions. In the next two iterations, the FastDesign[31,32] algorithm was given minimal restrictions to maximize the sequence-structure space explored by Rosetta. A total of 29'291 structures were designed this way. Several metrics were analyzed to evaluate the resulting designs. The total score of the system was used to determine the overall stability of the protein assembly. In addition, the interface score of the central subunit to surrounding subunits was calculated by measuring the difference between the total score before and after moving the central subunit out of the assembly[33]. The total score and interface score were taken as a selection metric. The 1'000 highest-scoring variants were taken forward for detailed analysis by MD simulation (Fig. 10a).

Design characterization. Various other metrics were analyzed to further characterize the obtained hits. Firstly, the contact count between subunits was determined. Notably, the interface score

correlated well with the contact count after docking (r^2 = 0.582, Fig. 10b). Given that the interface score was used as a selection criterion, the 1'000 variants selected after design had a higher average contact count (1'214 ± 103) compared to the overall average of all designs (1'042 ± 117). The enrichment for high contact counts indicates that it would be beneficial for the overall design pipeline to put a larger emphasis on selecting good structures after docking prior to the design step. Selecting for better initially docked structures, for instance by running a large number of docking runs and only taking the top docked poses for design, might substantially boost design quality by improving the interface scores and contact counts in the final designs.

The interface score was also compared to the solvent-accessible surface area (SASA) and the shape complementarity (SC) between subunits. The interface score correlated very well with the SASA (r^2 = 0.655), further demonstrating the validity of choosing the interface score as the selection metric. However, the SC did not correlate with the interface score (r^2 = 0.064), indicating that SC might be a valuable additional selection metric.

Hydrogen bonding and buried unsatisfied hydrogen bond scores were used to assess the polarity of the interface, and looking at the hydrogen bond score, we can see that hydrogen bonds generally contributed very little to the interface score (Fig. 10c). Additionally, there is no large increase in the hydrogen bond score for the selected variants, indicating no enrichment for polar interactions with our selection metrics. The number of buried unsatisfied hydrogen bonds thus also stayed roughly the same (Fig. 10d). Given that we were not aiming to specifically design polar interfaces, no selection was made for these criteria. However, these results show additional scores or selection metrics would need to be added to design for a polar interface.

In an attempt to increase the polarity of the interfaces in a different run, a test design run including a score for hydrogen bonding networks and buried unsatisfied hydrogen bonds in one iteration of the three FastDesign steps was performed. Unfortunately, this drastically increased the design time while scoring lower on the interface scores (Fig. S1). Thus, the design presented here was performed entirely without the hydrogen bonding network score.

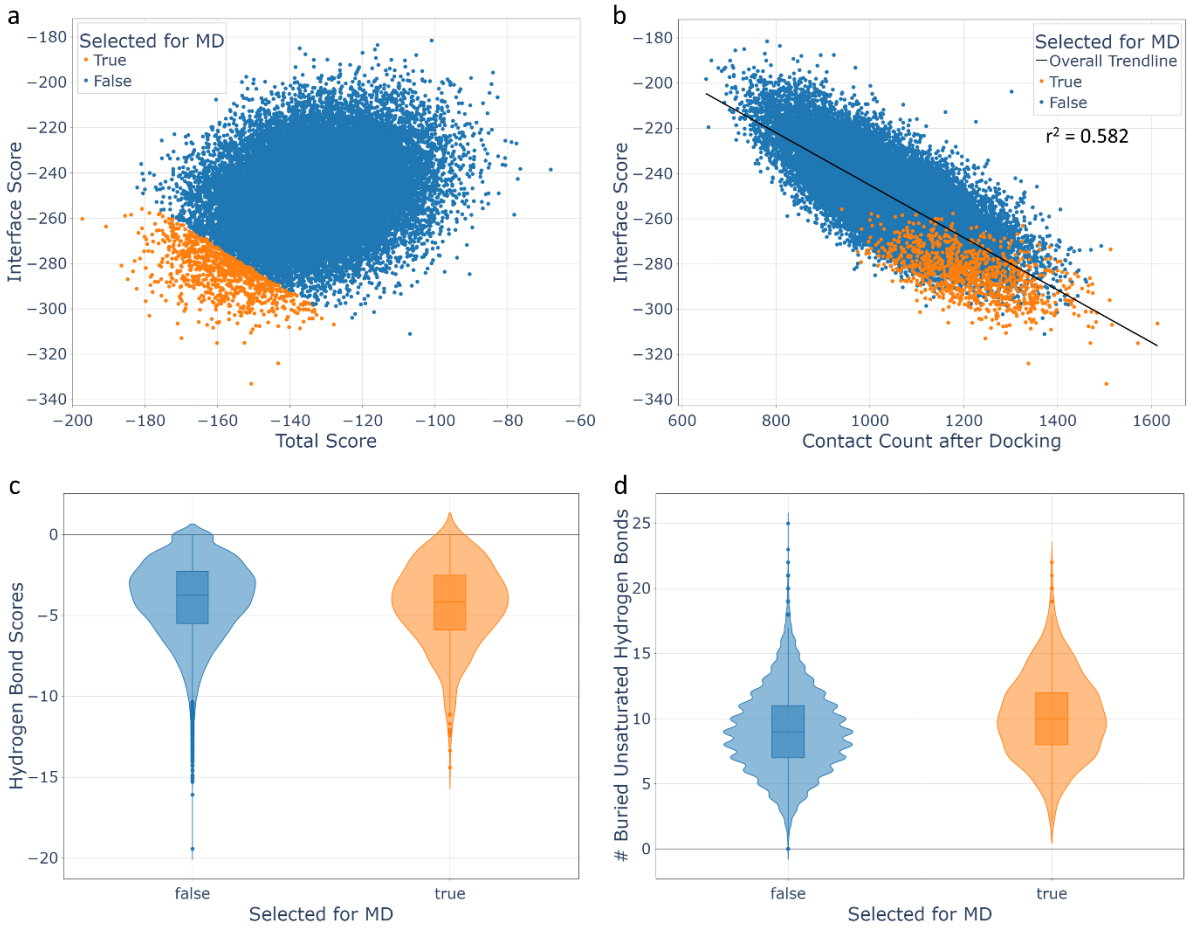


Fig. 10 | Design scoring. From all designs (blue), the top 1000 variants were selected for further analysis (orange). **a)** Scatter plot of the total score against the interface score. Selection of variants was decided by taking the sum of the total and interface score and taking the top 1000 variants. **b)** Scatter plot of the interface contact count after docking against the interface score. The interface score strongly correlates with the contact count (r^2 = 0582, p<<0.05), resulting in an enrichment of structures with high contact counts in the selection. **c)** Violin plots of hydrogen bond scores at the interface and **d)** violin plots of the number of buried unsatisfied hydrogen bonds at the interface for the selected and discarded variants. Both scores show little to no change between the selected and discarded variants.

3.1.3 Molecular Dynamics (MD) Simulations

The 1'000 most promising variants from the Rosetta design were further characterized using molecular dynamics (MD) simulations to narrow down the selection before experimental testing. In these simulations, we wanted to observe (i) if the variants stayed closely bound to each other in a rigid, defined structure, (ii) the composition of the interface for the different variants, and (iii) if the variants still bound EOY in a similar way to the parental variant.

MD setup. For the variants and their interfaces to be analyzed, we needed to simulate a repeating 2D array of the protein. To achieve this, we took advantage of the periodicity in MD simulations. By looking at the symmetric 9-mer generated in the Rosetta design step, the box was defined in a way that it encompassed a single subunit. The subunit was aligned to the y- and z-axes of the box, and the box dimensions were adjusted to provide an exact fit to the subunit's periodic images. Extra room was added in the x-axis to allow for a layer of solvent to prevent interference between the different arrays during the simulation (Fig. 11a). When applying periodicity to the system, the subunit forms the same protein-

protein interfaces as in the 9-mer from Rosetta, simulating an infinite repeating 2D array in the y- and z-axes (Fig. 11b,c).

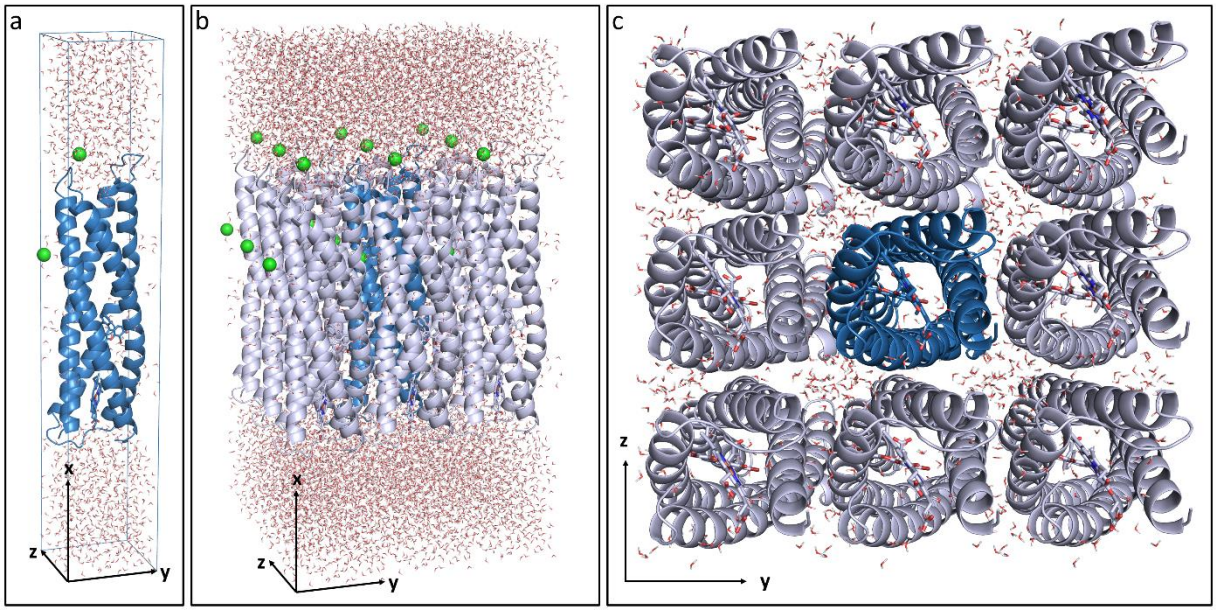


Fig. 11 | MD setup for infinite 2D array simulations. a) A single box of the MD simulation, with the XXE3H variant in blue, chloride ions in green and waters in red. **b)** The main MD box expanded by the surrounding periodic boxes, giving an idea of the infinite 2D array. Proteins not in the main box are shown in white. **c)** Top view of the expanded array with the water stripped above and below the array.

For testing the EOY binding properties of the variants, original hydrogen bonds needed to be conserved during the initial equilibration of the MD simulations (1 ns). To achieve this, distance constraints were implemented between the hydrogen bonding residues and EOY. However, unlike during the Rosetta design, no constraints were applied to the connecting hydrogen bond network. Constraints were removed during the MD production runs. The number of hydrogen bonds to EOY could thus be used as an important benchmark to characterize conservation of EOY binding properties in the final analysis.

***In silico* screening by short MD simulations.** Short 10 ns MD simulations were performed for the 1'000 variants selected from Rosetta design. The simulations were stable for 970 variants. For 30 variants, the MD simulations failed, likely due to the instability of the starting designs. These variants were thus discarded. To select variants for further testing, EOY binding, the protein foldedness, as well as 2D layer formation were analyzed in the final 2 ns of the simulations. As previously established, the average hydrogen bond count between the protein and EOY as well as the average RMSD of EOY were used to identify good EOY binders.[24] Variants with an average of >3 hydrogen bonds as well as an RMSD of EOY <2 Å were selected as good EOY binders (Fig. 12a). To ensure the structural integrity of the scaffold, the number of residues which were in an α-helical conformation of the protein was determined. Only structures with >160 residues in an α-helical were kept for further analysis (Fig. 12b). After applying cutoffs for EOY H-bonds and RMSD, as well as for α-helicity, 204 structures remained.

To identify the best variants, the interfaces and the overall assembly of the 2D array were assessed. The average contact count was used together with the average number of water molecules between the

subunits in the 2D array (Fig. 12c). The overall interface water count remained quite high for all the variants, which can be attributed to the waters to the sides of the protein-protein interfaces, which are not packed tightly and allow for water to pass through the array. Nonetheless, the two metrics correlated well at the lower end of water counts with higher contact counts indicating tighter packing. Interestingly, the correlation between water and contact count breaks down for badly packed structures. The worse structures showed large degrees of unfolding and aggregation, showing high contact counts but also high interface water counts. Thus, it is necessary to analyze both water content and contacts between subunits to reliably identify promising variants. To isolate these variants, cutoffs of >8'000 contacts and <180 interface water molecules were applied, resulting in 135 variants.

To further characterize the composition of the interface, the number of hydrogen bonds spanning the interface between two subunits were also counted. While the counts ranged up to 25 hydrogen bonds, the top variants were composed of unfolded models where backbone-backbone hydrogen bonds began forming. To increase the interface polarity of the selected variants, a cutoff was set at 5 hydrogen bonds (Fig. 12d). Combining all the cutoffs for the metrics, out of the 930 variants, 109 variants were selected for further analysis.

Tab. 1 | Cutoffs for each selection criteria as well as the number of variants after each step.

		10 ns		100 ns	
		Cutoff	Number	Cutoff	Number
Simulation Input		-	1000	-	109
Simulation Output		-	930	-	104
EOY Selection	EOY RMSD	<2 Å	316	<2 Å	51
	EOY Hydrogen Bonds	>3		>3	
α-Helicity Selection		>160	645	>165	45
Contact Selection	Contact Count	>8000	522	>9000	49
	Water Count	<180		<150	
Hydrogen Bond Selection		>5	649	>15	85
Final Selection		-	109	-	11

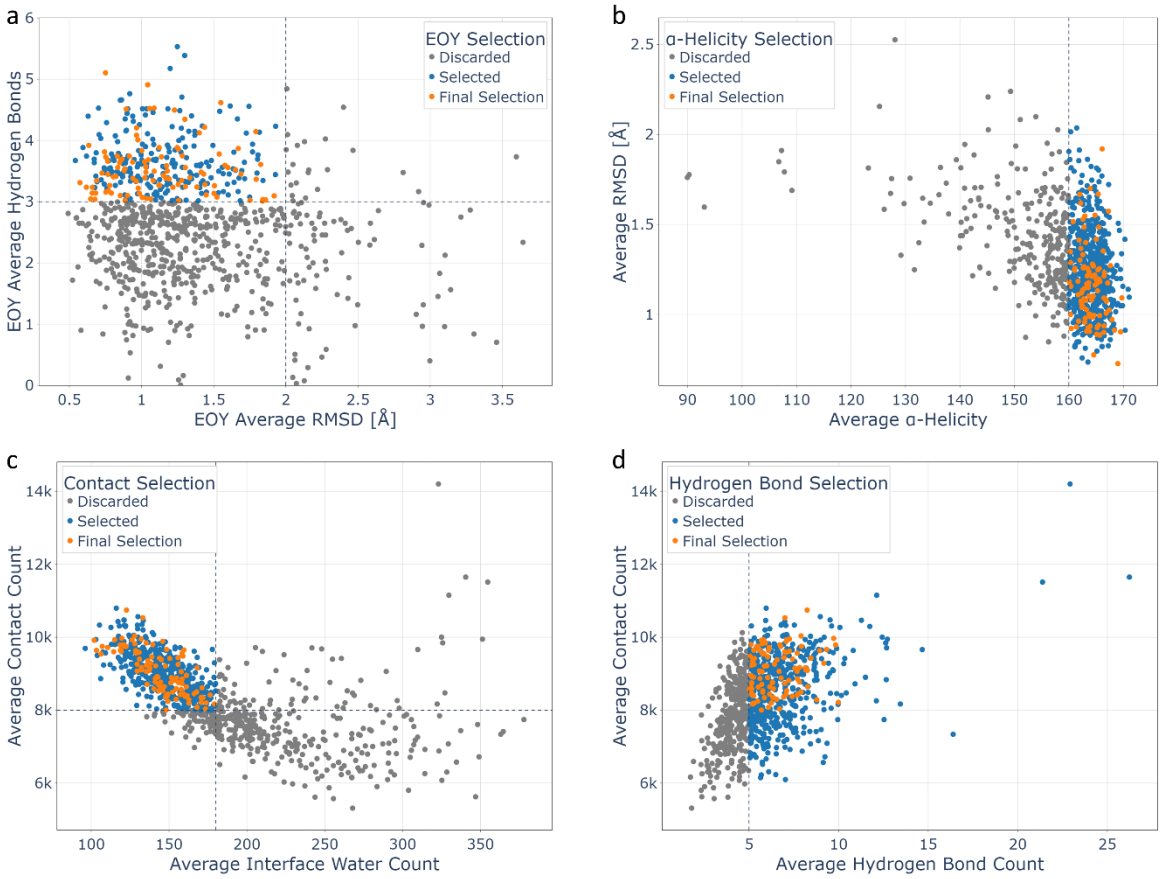


Fig. 12 | Selection of the most promising variants from 10 ns MD simulations. All points represent averages over the last 2 ns of the simulation. Blue and grey points correspond to variants that were either selected or discarded due to the shown cutoff. Orange points indicate variants that survived all cutoffs. **a)** Cutoffs used to select for EOY binding, including the number of hydrogen bonds between the protein and EOY on the y-axis (cutoff: >3) and the RMSD of EOY (cutoff: <2 Å). **b)** Cutoff used to select for α-helicity (cutoff: >160), plotted against the overall protein RMSD **c)** Cutoffs used to select for protein interface, including the contact count between proteins (cutoff: >8k) and the count of water molecules around the interface (cutoff: <180).**d)** Cutoff used to select for number of hydrogen bonds across the interface (cutoff: >5), plotted against the contact counts between proteins.

***In silico* screening by long MD simulations.** While the 10 ns MD simulations help remove the worst designs, we hypothesized that longer MD runs could allow to identify the best hits among the remaining variants. A further 100 ns of MD simulations were thus run for the previously selected 109 variants. During the long simulation, another 5 variants failed due to instabilities in the system, leaving 104 potential candidates. To narrow down the selection for experimental testing, more stringent cutoffs of the same metrics used for the 10 ns results were applied to the 100 ns simulations. The averages for all the metrics were taken over the last 20 ns of the simulations.

The cutoffs for the EOY selection metrics were kept as before at an RMSD of <2 Å and hydrogen bond count of >3 (Fig. 13a), while the cutoff for the α-helicity was increased to >165 residues (Fig. 13b). After selection with these two metrics, 24 variants remained. Cutoffs for contact and water count were also defined more stringently, setting an average contact count of >9'000 and an average interface water count of <150 (Fig. 13c), narrowing the selection down to 12 variants. Finally, the cutoff of average hydrogen bond count remained the same at >5 (Fig. 13d). Altogether, a final selection of 11 variants was defined for possible experimental testing.

The final 11 designs were studied in depth to select the hits for experimental testing. To better visualize the metrics of these last candidates, all scores were normalized from 0 to 1, and the average normalized score was calculated (Fig. 13e). The first variant chosen for experimental testing, dubbed XXE3H-2D1, had the highest average score and performed well on all metrics except the hydrogen bonding. Structural inspection revealed a well-defined binding pocket for EOY with conformations of the residues involved in the binding hydrogen bond network very similar to the parental variant. The protein-protein interfaces were comprised mostly of hydrophobic residues, with small residues where the backbones were close. Large, often aromatic residues were present where the backbones were further apart, filling the gaps and increasing the interface surface. This resulted in a large, hydrophobic interface that allowed for an energetically favorable association. As predicted by the average number of hydrogen bonds at the interface, only one hydrogen bond was stably formed between the subunits. Assembly of the XXE3H-2D1 sheet is thus mostly driven by general hydrophobic aggregation instead of any specific interactions.

Interested in more specific interactions, variants with higher counts of hydrogen bonds across the interface were deemed of higher interest. Unfortunately, due to some mistakes in the analysis, the variants XXE3H-2D7 and XXE3H-2D9 were identified as the two variants with the highest hydrogen bond counts (Fig. S2). Thus, these two variants were ordered instead of other, better performing variants, such as XXE3H-2D2 or XXE3H-2D8. Nonetheless, the structures of XXE3H-2D7 and XXE3H-2D9 were analyzed more in depth.

XXE3H-2D7 forms an extended hydrogen bonding network spanning two subunits, as well as several isolated hydrogen bonds across the interface. However, the hydrogen bonding network spanning the two subunits is not very rigid, dissociating several times over the 100ns simulation. XXE3H-2D9 shows no hydrogen bonding networks, with only a few single residues forming hydrogen bonds across the interface. In general, the interfaces were still mostly hydrophobic, and the packing of both variants seemed much looser than the packing of XXE3H-2D1, which agrees with their lower contact count and higher interface water count.

Looking at the rest of the variants, the structures follow the same patterns, with many large, hydrophobic residues but little specific interactions. Of the 11 different variants, the three variants discussed more in-depth here, XXE3H-2D1, XXE3H-2D7, and XXE3H-2D9, were selected for experimental testing.

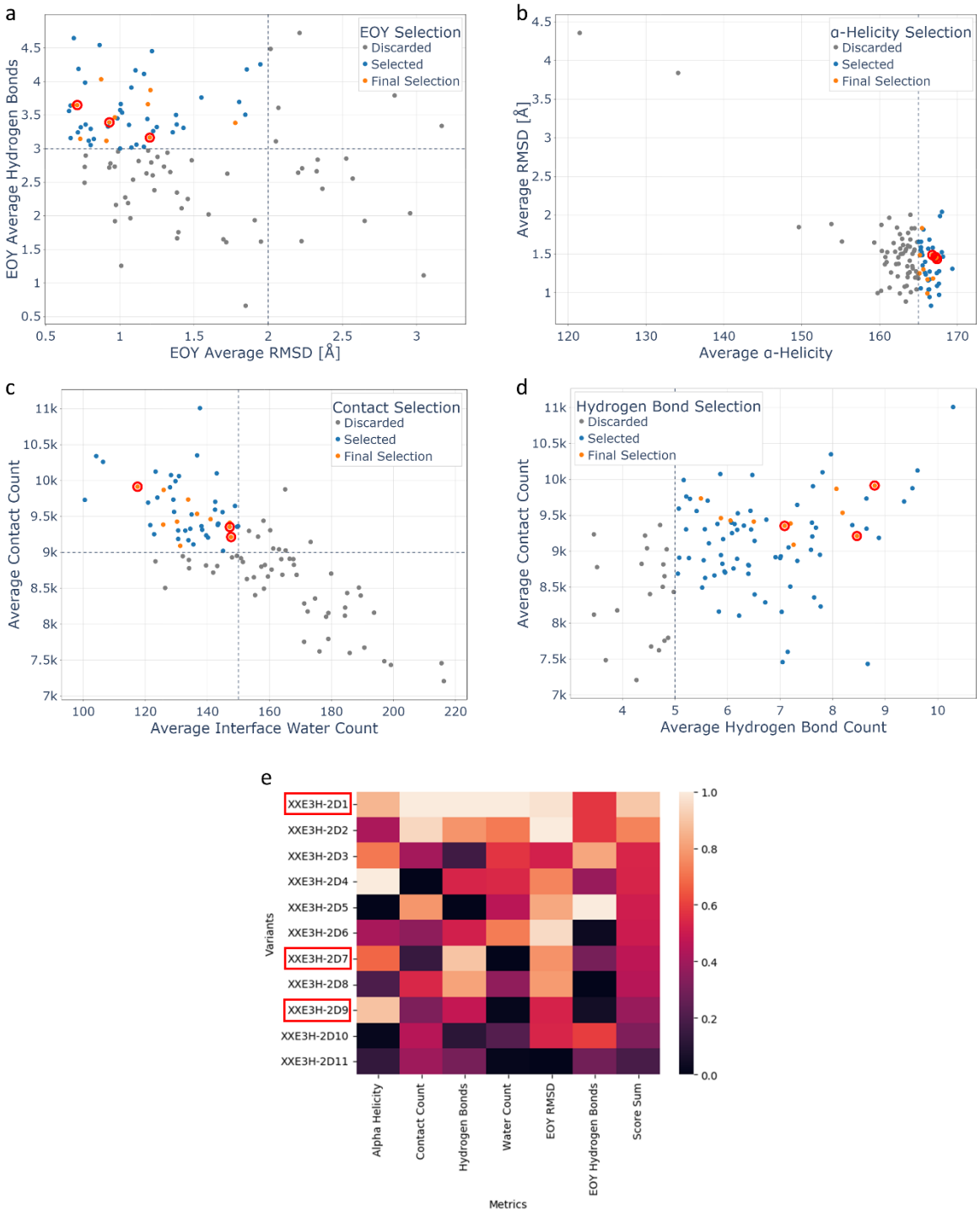


Fig. 13 | MD metrics used in the variant selection process in the 100 ns simulation with marked cutoffs used for selection. All data points represent averages taken over the last 20 ns of the 100 ns simulation. Blue and grey points correspond to variants that were either selected or discarded due to the shown cutoff. Orange points indicate variants that survived all cutoffs. Red circles indicate variants that were chosen for experimental testing. **a)** Cutoffs used to select for EOY binding, including the number of hydrogen bonds between the protein and EOY on the y-axis (cutoff: >3) and the RMSD of EOY (cutoff: <2 Å). **b)** Cutoff used to select for α-helicity (cutoff: >165), plotted against the overall protein RMSD. **c)** Cutoffs used to select for protein interface, including the contact count between proteins (cutoff: >9k) and the count of water molecules around the interface (cutoff: <150). **d)** Cutoff used to select for number of hydrogen bonds across the interface (cutoff: >15), plotted against the contact counts between proteins. **e)** Heatmap with the 7 variants selected by the cutoffs shown above as the rows and the relative scores for the different metrics as the columns. The three variants selected for experimental testing are shown circled in a red box. The data points of these three variants are also circled in the scatter plots above.

3.1.4 Variant purifications & testing

The most promising variants were ordered and commercially synthesized in a pET-21 vector. T7 express *E. coli* were transformed with the plasmids. All variant sequences were confirmed by sanger sequencing after DNA preparation from the transformed cells. Due to the highly hydrophobic properties of all the variants, expression in inclusion bodies was likely, calling for purification under denaturing conditions. SDS-PAGE was performed on both the supernatant and pellet of the lysate, confirming the protein remained in the pellet after lysis. In an attempt to denature and thus resuspend the protein, the pellet was resuspended in buffers containing 8M Urea or 6M Guanidinium Chloride (GdmCl). Both were spun down again, and both the pellet and supernatant tested for the protein with SDS-PAGE **(Fig. 14)**.

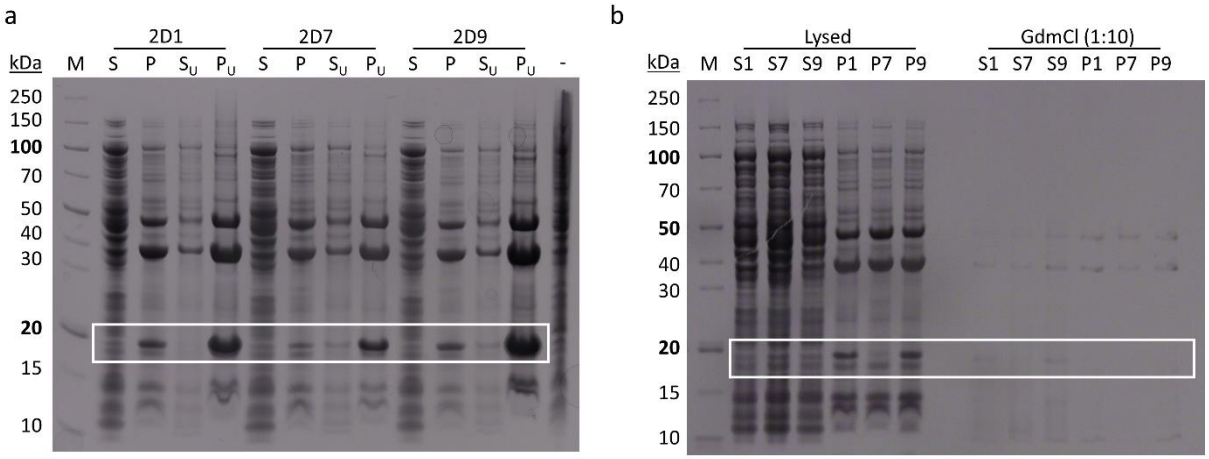


Fig. 14 | SDS-PAGE of samples taken from pellet and supernatant after lysis and attempted resuspension. a) Samples taken after lysis from the supernatant (S) and pellet (P), as well as samples after resuspension in Urea from the supernatant (S_U) and pellet (P_U) clustered by the variants. **b)** Samples taken after lysis and after resuspension in GdmCl, both from the supernatant of XXE3H-2D1 (S1), -2D7 (S7), and -2D9 (S9) and from the pellet, labeled P1, P7, and P9, respectively. The bands from resuspension are weaker due to necessary dilution (1:10) in buffer without GdmCl before gel loading due to incompatibilities of GdmCl and Laemmli buffer.

While the variants could not be resuspended in 8M Urea (Fig. 14a), the protein seemed to dissolve in 6M GdmCl. Despite the bands being weak due to necessary 1:10 dilution before SDS-PAGE to prevent GdmCl from precipitating out in contact with Laemmli buffer, bands are visible in the supernatant of the variants, while none can be seen in the pellet (Fig. 14b). However, an initial attempt to further purify the variants by Ni-NTA chromatography under 6M GdmCl was unsuccessful, with no detectable elution peak in the chromatogram as well as no protein in the eluted fraction. Due to time constraints, no further attempts could be made to purify the protein. Nonetheless, proteins solubilized in 6 M GdmCl, indicating this to be a good strategy for the protein purification. If successful purification can be achieved, it might be possible to subsequently refold the protein, e.g., through dialysis, to induce 2D array formation which could be tested with negative stain EM.

3.2 3D- Nanoparticle Design

Structural determination of protein variants stemming from the 4D2 has proven challenging. Although 4D2, containing 2 heme binding sites, has been crystallized and solved for a structure with a resolution of 1.9 Å[28], attempts at crystallizing other variants such as m4D2 and e4D2, containing 1 or 4 heme binding sites respectively, have not been successful. While the amide 1H-15N HSQC spectrum of m4D2 has been assigned[28], structural determination has not yet concluded. Notably, for the relatively large e4D2 (22.78 kDa instead of 13.89 kDa for 4D2), an 8.4 Å low-resolution cryo-EM structure could also be acquired[28]. This electron density map allowed tracing of the helical bundle backbone as well as the four heme groups, although details such as residue side chains were not apparent.

Given the structural challenges, a method for reliable structural determination would be highly desirable. It has been recently demonstrated that structural determination through cryo-EM of a 17 kDa α-helical protein could be achieved at high resolution up to 3.5 Å by attaching the protein to a larger capsid-forming protein scaffold.[29] Here, we aimed at implementing this approach for proteins of the 4D2 family. To that end, we focused on the single-heme variant m4D2 as a simple representative target structure and initially sought to display it onto the de novo designed nanosphere T33-21 (PDB ID: 4NWP) that has been previously used for structural determination.[11] The nanosphere consists of two subunits, EMA and EMB. Both subunits are homotrimers derived from proteins found in nature. The two subunits have been previously designed to assemble into a larger assembly through a hydrophobic interaction site, with four of each subunit assembling into a spherical shape. The N-terminus of subunit EMA looks outward from the sphere, allowing covalent linkage to another protein, in this case, m4D2. The assembly can then be analyzed under the EM.[29]

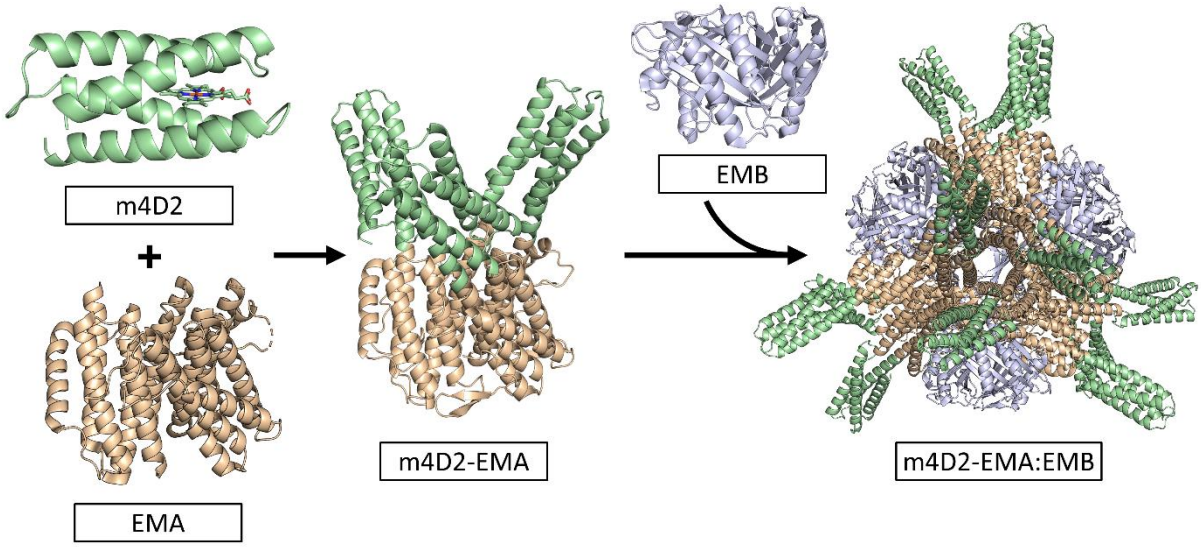


Fig. 15 | Illustration of the nanosphere assembly for structural determination of small proteins. m4D2 (green) is linked to subunit EMA of T33-21 (beige), which together with subunit EMB of T33-21 (purple) forms a spherical assembly with the m4D2 variant covalently linked to EMA. m4D2 is thus displayed on the surface of the assembly.

Note: The scaffold used in Liu, Y. *et al.*, PNAS (2018), corresponds to the T33-21 protein in King, N. *et al.*, Nature (2014).[11,29] Here, the complex scaffold will be referred to as EMA:EMB.

3.2.1 Nanoparticle Design & Assembly

Linker Design. In a first step, the linker between the subunit A (EMA) of the T33-21scaffold and m4D2 had to be designed. To increase the rigidity of the attached variant, we wanted to create a linker where the N-terminal α-helix of EMA is extended into the C-terminal helix of m4D2. For this purpose, prolines at the C-terminal end of the variant were replaced with alanine to prevent α-helix breakage. Four different linker lengths (7-10 residues, Tab. 2) were tested by looking at the predicted AlphaFold2 structures to find the best confirmations of the variant with respect to EMA (Fig. 16). Both the 7- and 8-residue linkers resulted in a continuous helical linkage with no steric interference between the trimer subunits. While the 9-residue linker had a continuous helical linkage, the subunits were close to each other, resulting in some steric clashes. For the 10-residue linker, the connecting α-helix had a small kink, leading us to discard this and all longer linkers.

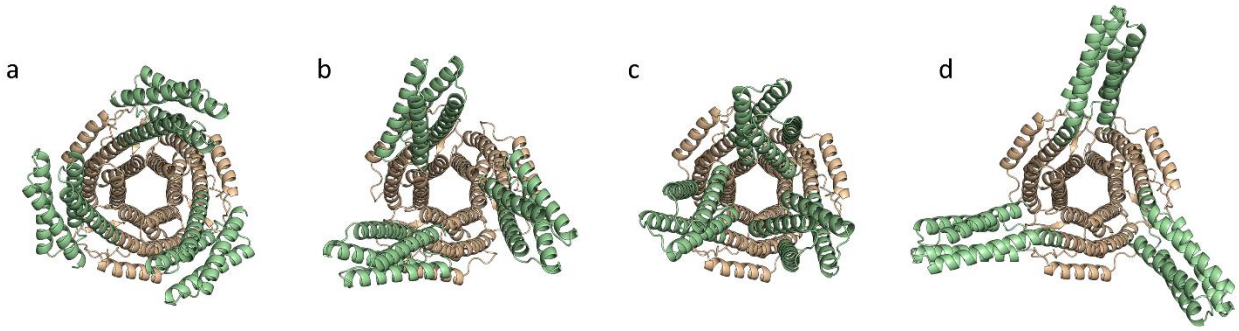


Fig. 16 | EMA-m4D2 trimers with different linker lengths predicted with AlphaFold2. The EMA scaffold is shown in beige and m4D2 in green. The linker length increases from 7 to 10 from **(a)** to **(d)**.

For the two most promising linkers, the 7- and 8-residue lengths, an additional linker was tested. For all the sequences, next to visually assessing the structures, the total per-residue model confidence score (pLDDT) from AlphaFold2 was used as an approximation for the rigidity of the structure.[34] Finally, the linker as well as the interface area between m4D2 and EMA were redesigned with ProteinMPNN in an attempt to find more rigid structures while also removing the Cys and His residues in the scaffold to reduce unwanted interactions with heme (Tab. 2). Interestingly, the MPNN redesign increased the pLDDT of each variant except the previously highest scoring one, of which the pLDDT was substantially reduced. The three most promising linker variants were chosen for experimental testing, choosing either before or after MPNN redesign based on which score was higher. The full sequence of all tested variants can be found in table**Tab. S1**.

Tab. 2 | The linker sequences and lengths of the m4D2-EMA structures chosen for prediction with AlphaFold2.

Linker Length[a]	7	7	8	8	9	10
Linker	VIGSAEL	VIEIGEL	**VIEGSAEL**	VIEIEGEL	VIEIGSAEL	VIEIEGSAEL
pLDDT[b]	73.7	75.4	**78.2**	76.9	75.6	74.1
MPNN Linker[c]	**VIGSAEL**	VIDAGEL	VRDGSAEL	**NRIITGEL**	VIDEKSAEL	-
MPNN pLDDT[b,c]	**76.3**	75.5	72.8	**79.2**	75.6	-
Ordered[d]	**m4D2-EM3**	-	**m4D2-EM2**	**m4D2-EM1**	-	-

[a] Four different linker lengths were tested, with a truncated m4D2 variant for the most promising linker lengths.

[b] The pLDDT from the AlphaFold2 predictions were used as an approximation for rigidity.

[c] For all but the longest linker length, the EMA linker and interface residues between m4D2 and EMA were redesigned in MPNN and the new sequences predicted again with AlphaFold2.

[d] The 3 ordered variants, dubbed m4D2-EMA1, -EMA2, and -EMA3, in order of descending pLDDT.

Trimer Assembly. The genes of the selected variants were ordered and commercially synthesized in a pET-21 vector and expressed in *E. coli.* In a first step, we wanted to test the assembly state of the variant before heme loading or his-tag cleavage. To that end, the cleared lysate was run through a Ni-NTA column before being run on a S200 size exclusion column (Fig. 17a-c) The size exclusion chromatograms of each variant is characterized by a large void peak. This indicates higher levels of aggregation, which could be explained by the association of multiple trimers through unspecific hydrophobic interactions. Next to the void peak, all three variants have a prominent peak at ~75 mL and ~85 mL. m4D2-EMA2 had an additional peak at ~65 mL, agreeing well with the expected elution volume of the trimer. Nonetheless, to confidently assign the peaks to different multimeric states, the mass of the proteins in each peak was measured by mass photometry (Fig. 17d) The mass photometry measurements showed peaks at three different masses, including ~60 kDa, ~120 kDa, and ~240 kDa. The 60 kDa peak is close to the lower limit of mass photometry, which could lead to inaccuracies. It is therefore likely that the assay overestimated its mass, and the peak actually corresponds to the monomer (33 kDa). The trimer of ~100 kDa was assigned to the 120 kDa peak, and thus the ~240 kDa peak likely corresponds to a dimer of trimers.

Formation of higher oligomers, such as trimer dimers is most likely due to aggregation of m4D2, as aggregation of 4D2-derived variants is a common problem seen in previous work. Also, it should be noted that the EMA scaffold displays hydrophobic surface areas designed for assembly of the large EMA:EMB complex. Without the EMB subunit, these exposed hydrophobics might also lead to aggregation.

As the counts for the trimer were highest in the ~75 mL peak from the size exclusion chromatogram of m4D2-EMA2, fractions for this peak were taken for further analysis by EM. Negative stain EM images of m4D2-EMA2 were taken, giving a first impression of the particle shape as well as the heterogeneity of the sample (Fig. 17e). From these images, particles with promising trimeric shapes of the right size were observed, indicating successful assembly of the trimer. However, negative-stain EM also revealed

that the sample was quite heterogeneous, showing large aggregates and smaller particles in agreement with the mass photometry measurements.

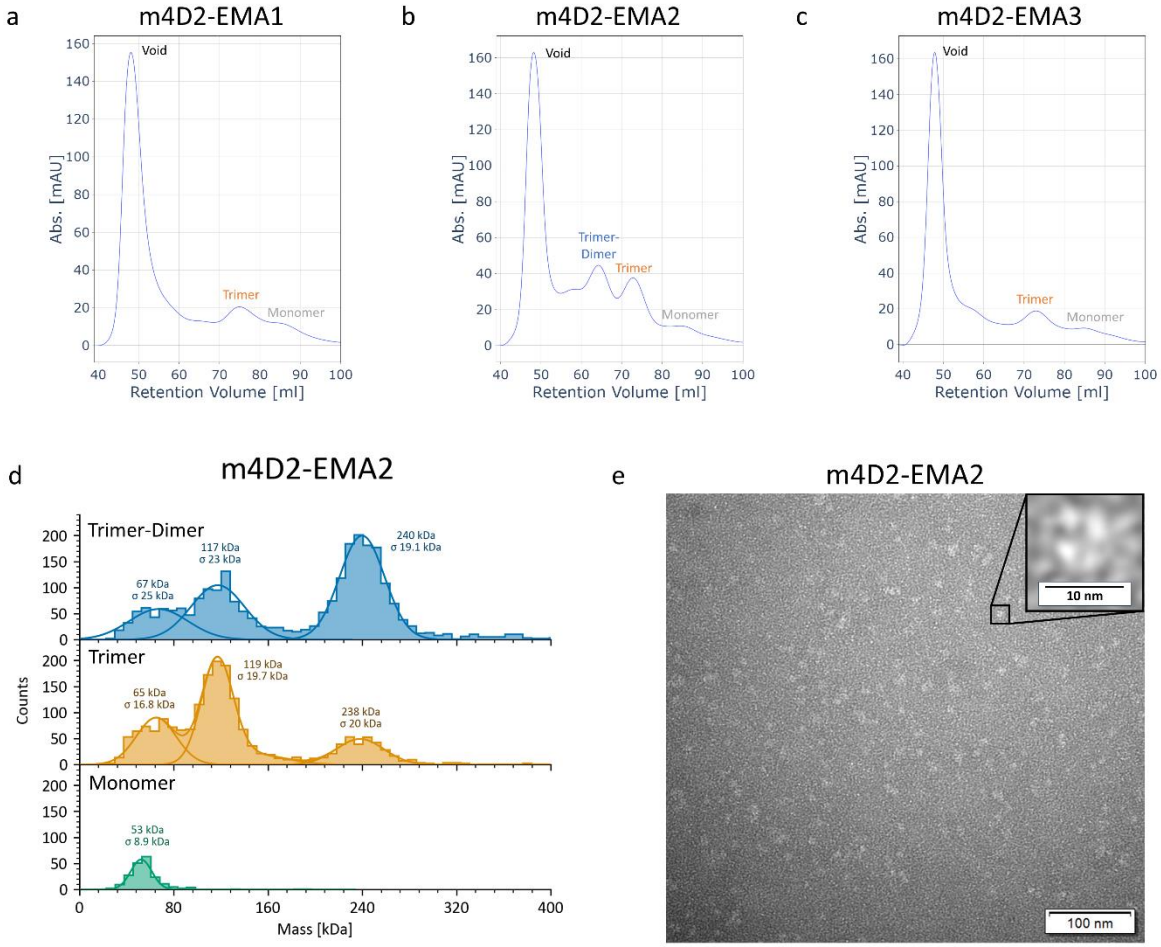


Fig. 17 | Assembly characterization of the different m4D2-EMA variants. The size exclusion chromatograms of **a)** m4D2-EMA1, **b)** m4D2-EMA2, and **c)** m4D2-EMA3. Each variant has a strong void peak (~50 ml) and weaker peaks at ~75 ml and ~85 ml. m4D2-EMA2 has an additional peak at ~65 ml. **d)** The protein mass of the three peaks in m4D2-EMA2 were determined by mass photometry, identifying the 75 ml peak as predominantly the trimer. **e)** Negative stain EM image of the trimer peak of m4D2-EMA2 with a close-up of a single particle potentially representative of the trimer.

Nanosphere assembly. To improve homogeneity and assemble larger particles suitable for cryo-EM analysis, full assembly of the EMA:EMB complex was attempted. Unfortunately, full *in vitro* assembly of EMA:EMB had not been shown to be possible in previous work, showing assembly only in co-expression.[11] Since problems with aggregation of the 4D2-derived variants were already known, co-expression was deemed a poor option due to the high number of displayed variants. To test for conditions in which the full EMA:EMB complex would assemble, one of the engineered variants, m4D2-EMA1, was mixed with the EMB subunit in an equimolar ratio of 20 µM. m4D2-EMA1 was chosen instead of m4D2-EMA2 for the low level of trimer-dimer formation. Assembly was tested at different concentrations of two different lyotropic salts, namely NaCl and Na_2SO_4. NaCl levels varied from 20 mM to 2 M, and Na_2SO_4 concentrations between 0 mM and 1 M, with a baseline of 20mM NaCl and 20 mM NaH_2PO_4 at pH 8.0. The mixtures were incubated overnight and checked for assembly by mass photometry (Fig. 18). For most conditions, we could see a peak at ~100 kDa for the m4D2-EMA1 trimer

and a peak at ~30-60 kDa, attributed to both the 45 kDa EMB trimer as well as the 30 kDa m4D2-EMA1 monomer. Unfortunately, we see no peak with a higher mass than the previously observed trimer-dimer, which is drastically lower than the expected mass of ~580 kDa for the full assembly. For the NaCl titration, we see little to no differences between the different salt concentrations, with all three peaks present in each histogram. Interestingly, for the Na_2SO_4 titration, the peak for the trimer and the trimer-dimer slowly disappears for increasing levels of Na_2SO_4, almost exclusively leaving the monomer and EMB peak. This could be due to EMA crashing out in the sulfate, or perhaps the more polar nature of the interface between the subunits of the trimer, for which the higher salt concentrations are not beneficial. In conclusion, however, attempts at driving the assembly of m4D2-EMA1:EMB *in vitro* were unsuccessful, leaving co-expression as the only option for full assembly.

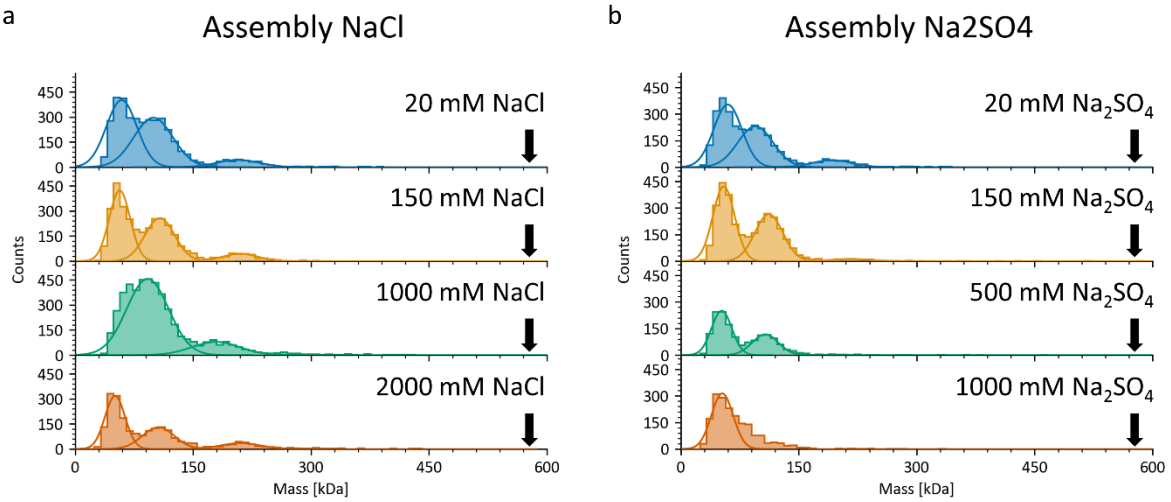


Fig. 18 | Histograms of mass photometry measurements for equimolar mixtures of m4D2-EMA1 and EMB with different buffer conditions. The counts of events are shown on the y-axis and the mass of the events on the x-axis with a bin width of 8.2 kDa. A gaussian curve is fit to each distinguishable peak as line plot overlayed with the histogram. The expected mass of m4D2-EMA1:EMB is indicated by a black arrow in each histogram. The mixtures were incubated overnight in buffers with **a)** different NaCl and **b)** different Na_2SO_4 concentrations. Conditions: 20 µM m4D2-EMA, 20 µM EMB, incubation time at room temperature.

Trimer Redesign Concluding from the promising trimeric shapes identified in the negative stain EM images of the m4D2-EMA2 variant (Fig. 17e) and difficulties in driving full assembly of T33-21 (Fig. 18), we opted for a different approach where we use the trimer for structural analysis. Since the EMA subunit in the designed complex was originally derived from the PH0671 trimer naturally occurring in *P. horikoshii* (PDB ID: 1WY1), we decided to use this WT variant as a basis for redesign using ProteinMPNN. This WT trimer was dubbed EM Trimer or EMT for further purposes.

Based on the previously performed linker design step in the m4D2-EMA variants, a linker length of 7 amino acids was chosen. In a first approach, m4D2-EMT1, ProteinMPNN redesign was performed to optimize the linker and interface area between the m4D2 and the EMT domains as well as replace Cys and His residues in the scaffold. This variant was recombinantly expressed and purified as for the m4D2-EMA variants. Additionally, on first analyses showing promise for trimer assembly, His_6-tag cleavage was performed followed by heme loading with a final size exclusion column (SEC) to achieve high purity.

Compared to the size exclusion chromatograms from the m4D2-EMA variants, the chromatogram for m4D2-EMT1 had a much smaller void peak, with the trimer being the most prominent peak. While a small peak for the trimer-dimer can be seen, the peak is well separated from the trimer peak, and no detectable monomer peak is present (Fig. 19a). Unfortunately, overall protein levels were lower than previous variants. Upon SDS-PAGE analysis of samples at different steps of purification, a strong band was identified in the pellet from centrifugation after lysis, indicating need for a harsher lysis protocol to improve protein yields. The trimer peak was analyzed by negative stain EM (Fig. 19b). These images again showed promising particles with the expected size and general shape of the trimer, while also seeming to be homogeneous and evenly dispersed. Given these promising images, a small dataset of 40 images was collected from this sample and analysis up to 2D classification was performed (Fig. 19c). Although some promising classes could be identified, no clearly recognizable shape could be identified, which is not surprising given the size of the particle.

In a second approach, ProteinMPNN redesign of the entire EMT scaffold domain was performed and the three best performing variants, based off of the ProteinMPNN score and AlphaFold2 pLDDT, were selected, expressed recombinantly, purified and heme loaded before running through the SEC (Fig. 19d-f). Among the three variants, m4D2-EMT4 showed great promise with a single high peak around the expected trimer elution volume. Unfortunately, upon analysis of this variant with negative stain EM, no identifiable particles could be detected. A grid was frozen for cryo-EM analysis to further check for any detectable particles (Fig. 19h). The ensuing cryo-EM images showed small, extended particles which had no resemblance to the trimer. Given the full redesign of the trimer, we hypothesized that, while the trimer still assembled at high concentrations, upon dilution for EM analysis, the trimer dissociated, leading to the small, extended particles we saw in the cryo-EM images. Given these results, we decided to continue with the original m4D2-EMT1 variant.

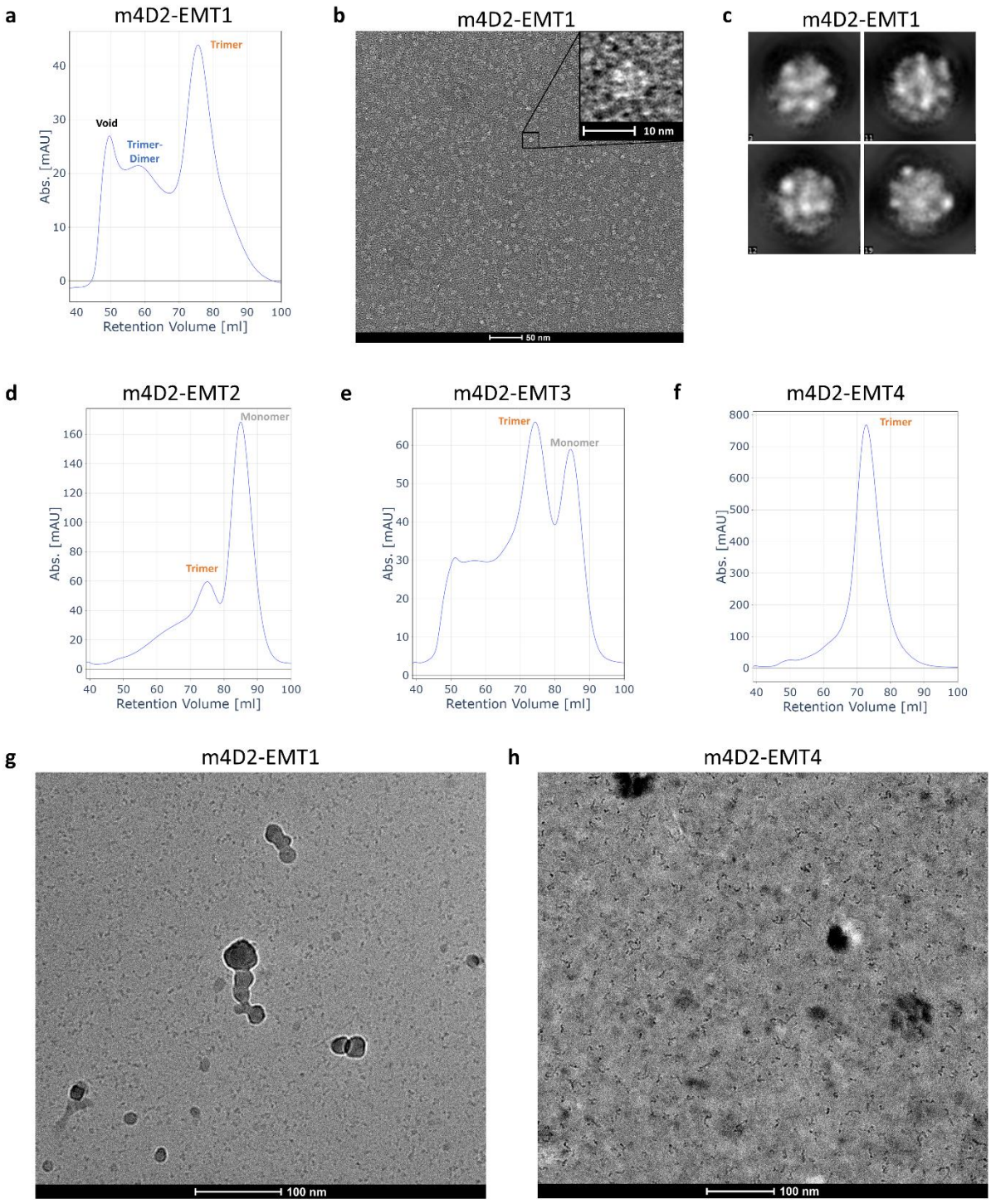


Fig. 19 | Assembly characterization of the different m4D2-EMT variants. a) Size exclusion chromatogram of m4D2-EMT1, with a prominent trimer peak as well as smaller void and trimer-dimer peaks. **b)** Negative stain EM images of m4D2-EMT1, with a close-up of a promising example of a trimer. **c)** 2D-classes determined from a negative stain EM dataset of 40 images potentially representing the m4D2-EMT1 trimer. **d)-f)** Size exclusion chromatograms of m4D2-EMT variants 2-4. All chromatograms are characterized by a small void peak and a small to no trimer-dimer peak. All variants show a prominent trimer peak, with m4D2-EMT4 having much higher protein levels than the other two variants. Monomer peaks are still prominent for m4D2-EMT2 and -3, however no monomer peak can be identified for EMT4.**g)-h)** Raw cryo-EM images of m4D2-EMT1 and m4D2-EMT4.

3.2.2 Structural Validation

Cryo EM. Given the promising negative stain EM images of the m4D2-EMT1 variant, a grid was frozen for cryo-EM analysis. A dataset of 50 micrographs was collected with an average contrast transfer function (CTF) estimation indicating a maximal resolution of ~8 Å (Fig. 19g). Blob particle picking was performed and promising 2D classes were selected for templated particle picking, resulting in 28'086 particles. 2D classes corresponding to the expected shape of the protein were selected, leaving 5'809 particles (Fig. 20a). An *ab initio* 3D model was reconstructed from these particles (Fig. 20b,c), showing good resemblance to the expected structure of m4D2-EMT1. Heterogeneous refinement was performed and, given the trimeric architecture of the protein, a C_3 symmetry expansion could be applied. The resulting electron density map shows characteristics expected of the m4D2-EMT1 variant, with three distinct "arms" reaching out from the more densely packed trimeric scaffold. Fitting an AlphaFold2 predicted structure to the map further emphasizes the similarities between the two (Fig. 20d,e).

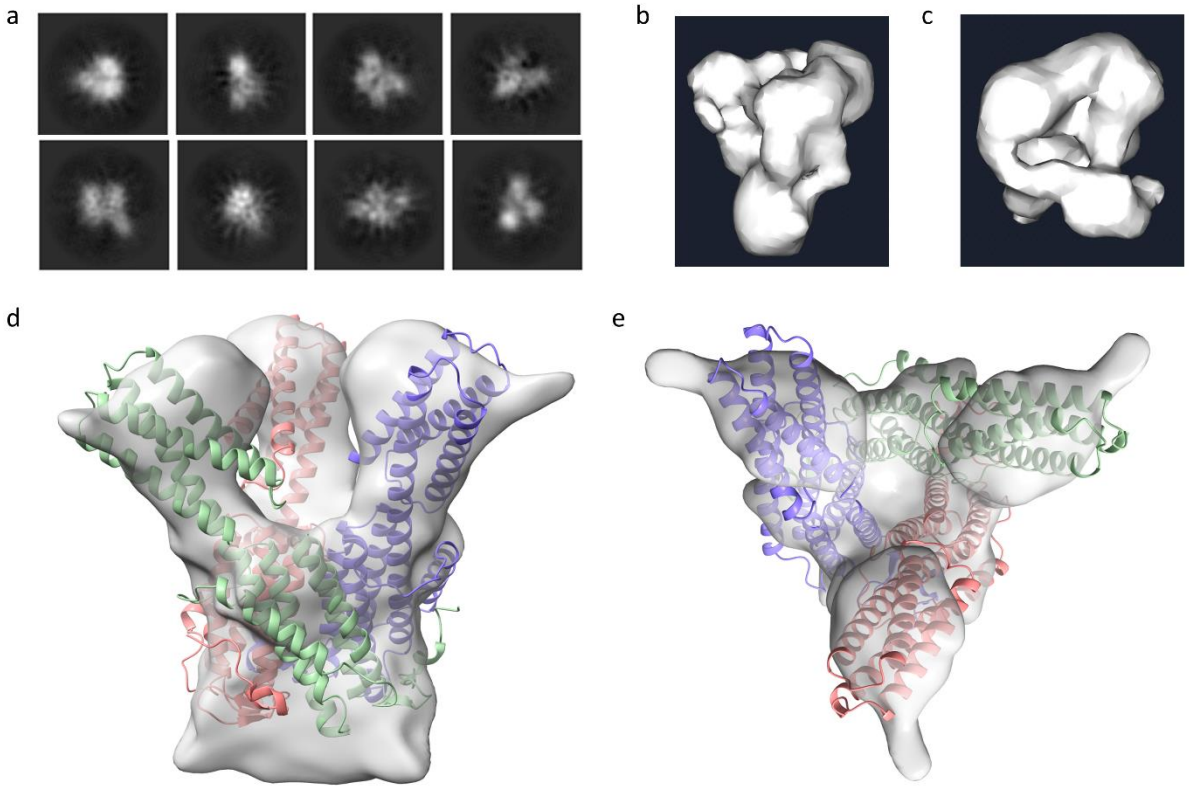


Fig. 20 | Cryo-EM analysis of variant m4D2-EMT1. a) Selected 2D classes from the templated particle picking, with a total of 5809 particles. **b)** Side view and **c)** top view of the *ab initio* reconstructed model. **d)** Side view and **e)** top view of the electron density map after heterogeneous refinement and C_3 symmetry expansion, with an AlphaFold2 model fit to the map.

While these results were promising, the limit was reached for the small dataset recorded. To reach higher resolution, a larger dataset of 3000 micrographs was recorded from the same grid. Analysis of this dataset is currently still underway. In conclusion, a low-resolution electron density map was achieved with a small dataset using the m4D2-EMT1 variant. These promising results indicate this approach could be used for routine structure prediction of 4D2-derived variants.

4 Conclusion & Outlook

4.1 2D-Nanosheet Design

In this thesis, a pipeline to design homomeric 2D-arrays in Rosetta was successfully set up and implemented. Resulting variants were tested *in silico* by MD simulations, exploiting the periodicity of MD to simulate an infinite array. Finally, purification and testing of the most promising variants was attempted in the lab. While the variants expressed in inclusion bodies could successfully be resuspended, purification of these variants was not achieved, preventing any tests on formation of a 2D-nanosheet.

While the pipeline for design seems promising, the main problem was the hydrophobic nature of the designed interfaces, making experimental handling of the variants extremely challenging. Design of more polar interfaces was attempted by adding a HBnet score to one of the steps in the design pipeline. This, however, drastically increased runtime and lowered the overall Rosetta interface score while also adding little to polarize the interfaces, yielding only slightly higher interface hydrogen bond scores.

Continuing in this project, a strategy to effectively increase polarity of the interfaces needs to be found. One option in Rosetta would be to use the HBnetStapleInterface to define a certain number of hydrogen bond networks connecting the subunits and constrain these networks in further design steps. While this would require more finetuning and manual selection of promising networks, this would increase the polarity of the interface. Alternatively, a hybrid design approach using ProteinMPNN in conjunction with Rosetta to redesign the interfaces after docking, a process which has been recently shown to generate more polar interfaces than using just Rosetta.[35]

Additionally, the design process could be performed on a heteromeric 2D-array, allowing for two separate soluble proteins which would assemble into a 2D array upon mixture. A heteromeric array would make loading the individual proteins with their respective ligands prior to mixture and 2D-array formation possible, further facilitating the handling of such samples.

Another major challenge is the lack of inherent symmetry in the protein. To achieve 2D array formation in the current system, at least two new protein-protein interfaces need to be designed, ideally with similar properties and affinities. To make the design process easier, it would be beneficial to introduce a point symmetry before working on the 2D array. One possibility would be to first design a tetramer of the variant. This would introduce a C_4 symmetry, where designing just one interface would suffice for 2D array formation (Fig. 21). This could also be performed for a heteromeric array, where both proteins form a homomeric tetramer.

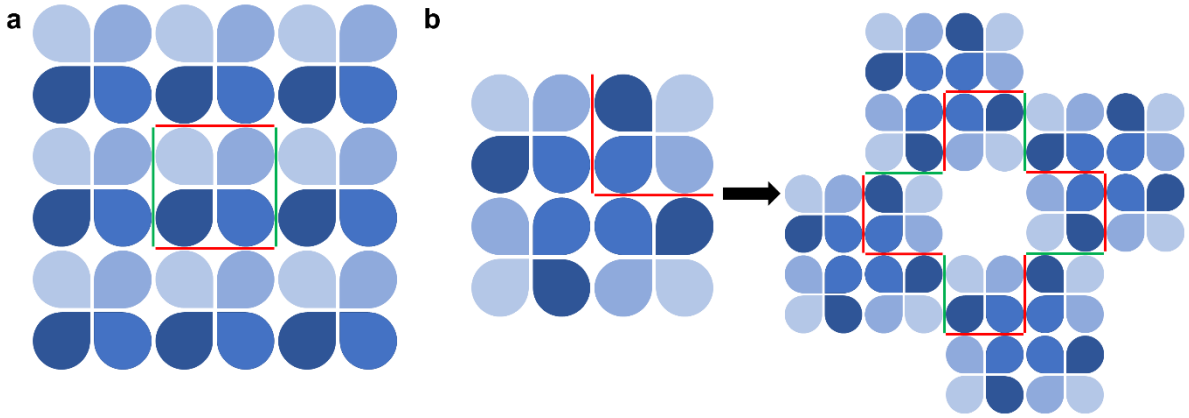


Fig. 21 | Different strategies for 2D nanosheet design with 4D2-derived variants. **a)** The approach used in this thesis for 2D array design. Here, two interfaces (red & green) need to be designed simultaneously. **b)** Alternative approach where a C_4 symmetric tetramer is designed first, focusing on only the red interface. After successful design of the tetramer, only the green interface needs to be designed for 2D array assembly.

In conclusion, the developed pipeline shows great promise for designing 2D nanosheets for proteins. However, more finetuning and possibly additional steps need to be included to increase the ease of handling these sample experimentally to facilitate validation of 2D array formation experimentally. This pipeline will serve as a good starting point for further development.

4.2 3D-Nanoparticle Design

In an attempt to facilitate structural determination of 4D2-derived variants to confirm the structure of newly designed variants and the binding confirmation of the small molecules, m4D2 was displayed on a scaffold protein to increase the particle size and introduce symmetry for cryo-EM analysis. A trimer of 100 kDa displaying m4D2 was analyzed with cryo-EM, resulting in a low-resolution electron density map of the particle, showing great potential for structural determination of the variants. Collection and analysis of a larger dataset is, however, necessary to increase resolution.

Despite this success, particle analysis is still challenging due to the small particle size, especially in particle picking. It would be recommendable to work on increasing the size of the scaffold protein for use in routine structural determination of 4D2-derived variants. This could be attempted by coexpression of the large T33-21 assembly, or identification of another scaffold protein with an outward-facing N-terminal α-helix where *in vitro* assembly is possible. Alternatively, further design could be done on the interface of T33-21, increasing polarity of the interface and facilitating *in vitro* assembly ().

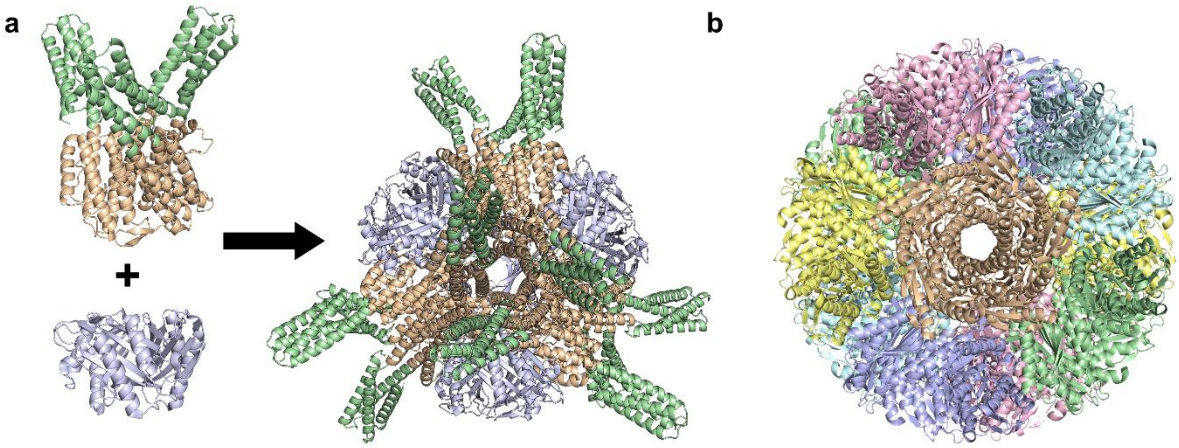


Fig. 22 | Increasing particle size for easier structural determination through cryo-EM with larger scaffolds. a) Full assembly of m4D2-EMA:EMB nanosphere. **b)** Protein cage potentially of interest as an alternative scaffold for displaying 4D2-derived variants (PDB: 1HQK), forming a 12-mer of homopentamers.[36]

Once an electron density map with atomic resolution can be achieved, more variants of interest can be displayed on the same scaffold with little to no design effort, allowing for an easy and modular structural determination strategy for 4D2-derived variants. This is especially important for EOY4D2.2 to confirm the binding confirmation of EOY, since multiple designs have been based on theoretical structures of this variant.

In conclusion, the low-resolution electron density map has demonstrated the validity of the display approach for structural determination of 4D2-derived variants. Upon further optimization of the scaffold, this system can be used for quick and easy structural determination of 4D2-derived variants using cryo-EM with minimal additional design efforts.

5 Materials & Methods

5.1 Computational Methods

5.1.1 Rosetta

In this work, RosettaScripts[30] was used with the Rosetta version "rosetta_src_2021.16.61629_bundle". All XML and bash files were generated in jupyter notebook.

Input files To generate the ensemble from XXE3H, 100 snapshots were taken from the previously performed 10 ns MD simulation[27] by extracting structures every 50 ps in the last 5 ns of the simulation. This was done using CPPTRAJ[37] (Fig. 23). The resulting structures were labeled XXE3H-1 through -100.

```
parm {FOLDER_PAR}/{PARENT}.parm7
change parmindex 0 chainid of * to A
trajin {FOLDER_PAR}/{PARENT}_eq_NPT.nc 505 last 5
autoimage
strip :WAT,Na+,Cl-
strip @H=
rms @CA
outtraj {FOLDER_SNAPS}/{NEW}.pdb multi chain id noter
```

Fig. 23 | CPPTRAJ file for generating the input ensemble for Rosetta design. Specific paths and names are replaced with descriptive terms in curly brackets.

The structures were relaxed in Rosetta with 3 repeats of FastRelax[38,32], using the beta_nov16 weights (Fig. 24). The hydrogen bonding network binding eosin Y (R77, Q80, E84, R121, R122, Q123, E124, E171, Q183) was also constrained to maintain these interactions through the relax process. Individual atoms with hydrogen bonds between each other were constrained by a gaussian distance constraint with a mean distance of 2.5 Å and a standard deviation of 0.5 Å (Fig. 25). The atom pair constraint weight was set to 1 for the score function.

```
<ROSETTASCRIPTS>

    <SCOREFXNS>
        <ScoreFunction name = "b16" weights = "beta_nov16" >
            <Reweight  scoretype = "atom_pair_constraint" weight = "1.0" />
        </ScoreFunction>
    </SCOREFXNS>

    <MOVERS>
        <ConstraintSetMover name             = "mv_cst"
                            add_constraints  = "true"
                            cst_file         = "{FOLDER_PAR}/eosiny.cst" />
        <FastRelax          name             = "mv_relax"
                            scorefxn         = "b16"
                            repeats          = "3" />
    </MOVERS>

    <PROTOCOLS>
        <Add mover_name = "mv_cst"    />
        <Add mover_name = "mv_relax" />
    </PROTOCOLS>

</ROSETTASCRIPTS>
```

Fig. 24 | XML file for Rosetta relax. Specific paths and names are replaced with descriptive terms in curly brackets.

```
AtomPair NE2 80 O3 198 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH1 121 O3 198 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NE2 123 O1 198 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NE2 123 O2 198 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NE2 183 O4 198 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NE2 80 OE1 124 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair OE1 80 NH1 77 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH1 77 OE1 124 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH1 77 OE2 124 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH2 77 OE2 124 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH2 121 OE2 84 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair OE1 123 NH1 122 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH1 122 OE2 171 GAUSSIANFUNC 2.5 0.5 TAG
AtomPair NH2 122 OE2 171 GAUSSIANFUNC 2.5 0.5 TAG
```

Fig. 25 | Rosetta constraint file "eosiny.cst" defining atom pair distances to conserve the hydrogen bonding network around EOY. The atom pair is defined in each line by the residue number and atom name for each atom. The distance constraint is defined by a gaussian function defined by the mean at 2.5 Å and the standard deviation at 0.5 Å.

Symmetry Definition File Before docking and design could be performed, a symmetry file needed to be defined to simulate the protein in a 2D-array, named X2E3H_2D_ARRAY. As generating a symmetry definition file from scratch is quite challenging, the symmetry definition file "3AEIA_P4Z" from previous work on 2D array designs was used as a starting point.[35] The energy term is given as in Eq. 1 to calculate the total score of the pose, with VRTX_1_1 connecting to subunit X. The anchor residue was changed to residue 1 instead of the center of mass due to problems with the docking protocol if the subunit was anchored by eosin Y close to the center of mass of the protein. Coordinates of the virtual residues were defined between "virtual_coordinates_start" and "virtual_coordinates_stop". The coordinates were taken from the symmetry file 3AEIA_P4Z and updated for the added subunits. Underneath the virtual coordinates the jumps are defined, connecting the virtual residues to each other. The jump patterns repeat for each subunit. An example is given here between subunit A and B: VRTA_ctrl is linked to VRTB_outer through JUMPB_to_outer, changing the x- and y-vectors to align with the jump toward the next subunit while keeping the 3D position the same. From VRTB_outer, jump JUMPB_to_redir connects VRTB_redir, updating the 3D position to the position of the next subunit. Finally, VRTB _redir is connected to VRTB_1_1 through JUMPB_1_1, updating the x- and y-vectors to realign the subunit with the array. The subunit is then connected to VRTB_1_1 through JUMPB_1_1_to_subunit.

Finally, three different jump groups were defined, linking movements along these jumps. The first two include the JUMPX_to_redir jumps, the jumps connecting the 3D positions of the different subunits. These jumps are split into two groups based on the directionality of the jump, i.e., as horizontal and vertical, as seen in the arrows in Fig. 8a. Movement along the x-vector was allowed along these two jumps, and were initialized by sliding along this jump by 100 Å. The third jump group comprised all jumps JUMPX_1_1. Several extra degrees of freedom were given here, including rotation with random initialization around and limited movement along the z-axis. Additionally, rotation around the x- and y-axis was allowed, but initial rotation was kept fixed. Finally, the slide order was kept random, setting no fixed movement order for the jump groups.

```
symmetry_name X2E3H_2D_array

E = VRT0_1_1 + (VRT0_1_1:VRT1_1_1) + (VRT0_1_1:VRT2_1_1) + (VRT0_1_1:VRT5_1_1) + (VRT0_1_1:VRT6_1_1)

anchor_residue 1

virtual_coordinates_start
xyz VRT0  1.000000,0.000000,0.000000  0.000000,1.000000,0.000000  34.657441,0.000090,8.516077
xyz VRT0_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  33.657441,0.000090,8.516077
xyz VRT0_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  33.657441,0.000090,8.516077

xyz VRT1_outer  0.152726,-0.988269,-0.000022  0.988269,0.152726,0.000004  33.657441,0.000090,8.516077
xyz VRT1_redir  0.152726,-0.988269,-0.000022  0.988269,0.152726,0.000004  18.384799,98.826941,8.518265
xyz VRT1_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  18.384799,98.826941,8.518265

xyz VRT2_outer  0.988269,0.152726,0.000004  -0.152726,0.988269,0.000022  33.657441,0.000090,8.516077
xyz VRT2_redir  0.988269,0.152726,0.000004  -0.152726,0.988269,0.000022  -65.169409,-15.272552,8.515722
xyz VRT2_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  -65.169409,-15.272552,8.515722

xyz VRT3_outer  -0.152726,0.988269,0.000022  -0.988269,-0.152726,-0.000004  33.657441,0.000090,8.516077
xyz VRT3_redir  -0.152726,0.988269,0.000022  -0.988269,-0.152726,-0.000004  48.930084,-98.826760,8.513889
xyz VRT3_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  48.930084,-98.826760,8.513889

xyz VRT4_outer  -0.988269,-0.152726,-0.000004  0.152726,-0.988269,-0.000022  33.657441,0.000090,8.516077
xyz VRT4_redir  -0.988269,-0.152726,-0.000004  0.152726,-0.988269,-0.000022  132.484292,15.272733,8.516432
xyz VRT4_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  132.484292,15.272733,8.516432

xyz VRT1_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  18.384799,98.826941,8.518265
xyz VRT2_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  -65.169409,-15.272552,8.515722
xyz VRT3_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  48.930084,-98.826760,8.513889
xyz VRT4_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  132.484292,15.272733,8.516432

xyz VRT5_outer  0.988269,0.152726,0.000004  -0.152726,0.988269,0.000022  18.384799,98.826941,8.518265
xyz VRT5_redir  0.988269,0.152726,0.000004  -0.152726,0.988269,0.000022  117.211649,114.099583,8.51862
xyz VRT5_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  117.211649,114.099583,8.51862

xyz VRT6_outer  -0.152726,0.988269,0.000022  -0.988269,-0.152726,-0.000004  -65.169409,-15.272552,8.515722
xyz VRT6_redir  -0.152726,0.988269,0.000022  -0.988269,-0.152726,-0.000004  -80.442052,83.554298,8.51791
xyz VRT6_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  -80.442052,83.554298,8.51791

xyz VRT7_outer  -0.988269,-0.152726,-0.000004  0.152726,-0.988269,-0.000022  48.930084,-98.826760,8.513889
xyz VRT7_redir  -0.988269,-0.152726,-0.000004  0.152726,-0.988269,-0.000022  -49.896767,-
114.099403,8.513534
xyz VRT7_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  -49.896767,-114.099403,8.513534

xyz VRT8_outer  0.152726,-0.988269,-0.000022  0.988269,0.152726,0.000004  132.484292,15.272733,8.516432
xyz VRT8_redir  0.152726,-0.988269,-0.000022  0.988269,0.152726,0.000004  147.756934,-83.554118,8.514244
xyz VRT8_ctrl  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  147.756934,-83.554118,8.514244

xyz VRT5_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  117.211649,114.099583,8.51862
xyz VRT6_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  -80.442052,83.554298,8.51791
xyz VRT7_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  -49.896767,-114.099403,8.513534
xyz VRT8_1_1  0.590817,0.806805,0.000018  0.806805,-0.590817,-0.000013  147.756934,-83.554118,8.514244

virtual_coordinates_stop

connect_virtual JUMP0 VRT0 VRT0_ctrl
connect_virtual JUMP0_1_1 VRT0_ctrl VRT0_1_1
connect_virtual JUMP0_1_1_to_subunit VRT0_1_1 SUBUNIT
connect_virtual JUMP1_to_outer  VRT0_ctrl VRT1_outer
connect_virtual JUMP1_to_redir VRT1_outer VRT1_redir
connect_virtual JUMP1_to_ctrl VRT1_redir VRT1_ctrl
connect_virtual JUMP1_1_1 VRT1_ctrl VRT1_1_1
connect_virtual JUMP1_1_1_to_subunit VRT1_1_1 SUBUNIT
connect_virtual JUMP2_to_outer  VRT0_ctrl VRT2_outer
connect_virtual JUMP2_to_redir VRT2_outer VRT2_redir
connect_virtual JUMP2_to_ctrl VRT2_redir VRT2_ctrl
connect_virtual JUMP2_1_1 VRT2_ctrl VRT2_1_1
connect_virtual JUMP2_1_1_to_subunit VRT2_1_1 SUBUNIT
connect_virtual JUMP3_to_outer  VRT0_ctrl VRT3_outer
connect_virtual JUMP3_to_redir VRT3_outer VRT3_redir
connect_virtual JUMP3_to_ctrl VRT3_redir VRT3_ctrl
connect_virtual JUMP3_1_1 VRT3_ctrl VRT3_1_1
connect_virtual JUMP3_1_1_to_subunit VRT3_1_1 SUBUNIT
connect_virtual JUMP4_to_outer  VRT0_ctrl VRT4_outer
connect_virtual JUMP4_to_redir VRT4_outer VRT4_redir
connect_virtual JUMP4_to_ctrl VRT4_redir VRT4_ctrl
connect_virtual JUMP4_1_1 VRT4_ctrl VRT4_1_1
connect_virtual JUMP4_1_1_to_subunit VRT4_1_1 SUBUNIT
connect_virtual JUMP5_to_outer  VRT1_ctrl VRT5_outer
connect_virtual JUMP5_to_redir VRT5_outer VRT5_redir
connect_virtual JUMP5_to_ctrl VRT5_redir VRT5_ctrl
connect_virtual JUMP5_1_1 VRT5_ctrl VRT5_1_1
connect_virtual JUMP5_1_1_to_subunit VRT5_1_1 SUBUNIT
connect_virtual JUMP6_to_outer  VRT2_ctrl VRT6_outer
connect_virtual JUMP6_to_redir VRT6_outer VRT6_redir
connect_virtual JUMP6_to_ctrl VRT6_redir VRT6_ctrl
connect_virtual JUMP6_1_1 VRT6_ctrl VRT6_1_1
connect_virtual JUMP6_1_1_to_subunit VRT6_1_1 SUBUNIT
connect virtual JUMP7 to outer  VRT3 ctrl VRT7 outer
```

```
connect_virtual JUMP7_to_redir VRT7_outer VRT7_redir
connect_virtual JUMP7_to_ctrl VRT7_redir VRT7_ctrl
connect_virtual JUMP7_1_1 VRT7_ctrl VRT7_1_1
connect_virtual JUMP7_1_1_to_subunit VRT7_1_1 SUBUNIT
connect_virtual JUMP8_to_outer  VRT4_ctrl VRT8_outer
connect_virtual JUMP8_to_redir VRT8_outer VRT8_redir
connect_virtual JUMP8_to_ctrl VRT8_redir VRT8_ctrl
connect_virtual JUMP8_1_1 VRT8_ctrl VRT8_1_1
connect_virtual JUMP8_1_1_to_subunit VRT8_1_1 SUBUNIT

set_dof JUMP1_to_redir x(100)
set_dof JUMP2_to_redir x(100)
set_dof JUMP0_1_1 angle_z(0:360) z[0;-5:5] angle_x angle_y

set_jump_group JUMPGROUP1 JUMP1_to_redir JUMP3_to_redir JUMP6_to_redir JUMP8_to_redir
set_jump_group JUMPGROUP2 JUMP2_to_redir JUMP4_to_redir JUMP5_to_redir JUMP7_to_redir
set_jump_group JUMPGROUP3 JUMP0_1_1 JUMP1_1_1 JUMP2_1_1 JUMP3_1_1 JUMP4_1_1 JUMP5_1_1 JUMP6_1_1 JUMP7_1_1
JUMP8_1_1

slide_type RANDOM
```

Fig. 26 | Rosetta symmetry file used for the docking and design. The file includes 8 sections: i) the symmetry file name, ii) the score function, iii) the residue to which the virtual subunit is linked to the protein subunit, iv) the virtual coordinates, v) the jumps connecting the virtual coordinates, vi) the degrees of freedom for each jump group, vii) the jumps included in each jump group, viii) the slide type.

Docking & Design Docking, design, and filters used variant characterization were all done in one XML file (Fig. 27). For docking, symmetry needed to be set up in a first step. The pose was then saved for the residue sequence and sidechain conformation before mutating the protein to a poly-alanine chain, keeping the proline and glycine residues. The pose was changed to a centroid representation for docking. This was performed by the GenericMonteCarlo[30] mover in conjunction with the SymDockprotocol, setting 20 trials and no drift to perform 20 attempts with each attempt beginning from the original pose. Each pose was scored with the "cen_std_smooth" weights. "recover_low" was set to true to select the best scoring of the 20 docked poses for design. This pose was further characterized with the AtomicContactCount filter, partitioning by chain with a distance of 4.5 Å. The pose was then reverted to a full atom representation, and the side chain identity and conformation were restored from the previously saved pose.

After docking, three iterations of FastDesign[31] were performed to find interfaces that give an energetically favorable conformation. To preserve inward looking and ligand binding amino acids, as well as the loops between the helices, residues were manually selected to restrict to moving (G1, S2, P3, H9, A13, L16, A20, A23, L27, L30, I34, V37, W41, L44, N47, T48, S49, N50, S51, P52, F58, L62, V65, W69, L72, A76, R77, I79, Q80, A83, E84, V86, G90, M93, N96, G97, S98, V99, S100, P101, S102, P103, H109, A113, L116, A120, R121, R122, Q123, E124, L127, F130, L134, V137, W141, L144, N147, T148, S149, N150, S151, P152, L158, I162, V165, W169, E171, L172, A175, A176, L179, Q183, V186, G190, M193), allowing no repacking or redesign. Pro, His, and Cys residues were all restricted to repacking. The hydrogen bond network binding EOY was again constrained with gaussian distance constraints (Fig. 25). In the first design iteration, movement along jumps was forbidden in the MoveMapFactory to enforce the docked conformation and keep the backbones in close proximity. Additionally, only a restricted pool of small amino acids (Ala, Asp, Ile, Leu, Asn, Ser, Thr, Val) was allowed for redesign to promote tight binding with closely packed backbones. In the second and third design iteration, restriction to movement along jumps was removed and the amino acid pool for design was expanded, excluding only design to Cys, Gly, Pro,

and His residues unless native. For the HBnet run, the b16_hbnet score function was used in the second iteration, setting the hbnet and buried_unsatisfied_penalty weights to one together with the b16 weights.

Multiple filters and movers were used for evaluation of the designed structure. Most importantly, the InterfaceScoreCalculator was used to calculate interface score by calculating the difference of the total score before and after removing chain A, i.e., the central subunit. The interfaces were further characterized by the filters ShapeComplementarity, Sasa, and SymUnsatHbonds, all across jump groups 1 and 2, represented by jumps JUMP1_to_redir and JUMP2_to_redir, and a confidence measure of zero. Finally, the number of mutations was counted by the SequenceRecovery filter.

Before the scores could be used for analysis, certain corrections needed to be done. Both the atom pair constraint score as well as the total score of the respective relaxed structures was subtracted from the total score of each pose. The interface atom pair constraint score was also subtracted from the interface score. Additionally, the two interface hydrogen bonding scores, if_A_hbond_bb_sc and if_A_hbond_sc, were added to get the total interface hydrogen bonding score. Since ShapeComplementarity filter cannot take two different jump groups as an input, this value was calculated separately for each interface and averaged to get the total shape complementarity.

```
<ROSETTASCRIPTS>

    <SCOREFXNS>
        <ScoreFunction name = "b16" weights = "beta_nov16">
            <Reweight scoretype = "atom_pair_constraint" weight = "1.0"/>
        </ScoreFunction>
        <ScoreFunction name = "b16_hbnet" weights = "beta_nov16">
            <Reweight scoretype = "hbnet" weight = "1.0"/>
            <Reweight scoretype = "buried_unsatisfied_penalty" weight = "1.0"/>
            <Reweight scoretype = "atom_pair_constraint" weight     = "1.0"/>
        </ScoreFunction>
        <ScoreFunction name = "cen" weights = "cen_std_smooth"/>
    </SCOREFXNS>

    <RESIDUE_SELECTORS>
        <Index name = "sel_nodesign"  resnums = "{res_nodesign}"/>
        <Index name = "sel_eosiny"    resnums = "{res_eosiny}"/>
    </RESIDUE_SELECTORS>

    <TASKOPERATIONS>
        <InitializeFromCommandline name               = "tsk_init"/>
        <IncludeCurrent            name               = "tsk_ic"/>
        <OperateOnResidueSubset    name               = "tsk_nodesign"
                                   selector           = "sel_nodesign">
                                   <PreventRepackingRLT/>
        </OperateOnResidueSubset>
        <RestrictIdentities        name               = "tsk_pghr"
                                   identities         = "PRO,CYS,HIS"/>
        <LimitAromaChi2            name               = "tsk_limitaro"
                                   chi2max            = "110"
                                   chi2min            = "70"/>
        <DisallowIfNonnative       name               = "tsk_disallow_aa"
                                   disallow_aas       = "CGPH"/>
        <DisallowIfNonnative       name               = "tsk_aa_reduced"
                                   disallow_aas       = "CEFGHKMPQRWY"/>
    </TASKOPERATIONS>

    <MOVE_MAP_FACTORIES>
        <MoveMapFactory name = "mm_nojump" jumps = "0"/>
    </MOVE_MAP_FACTORIES>

    <FILTERS>
         <AtomicContactCount     name              = "flt_cc_jump"
                                 partition         = "chain"
                                 normalize_by_sasa = "0"
                                 distance          = "4.5"/>
         <SequenceRecovery       name              = "flt_mutations"
                                 rate_threshold    = "0.0"
                                 mutation_threshold= "1000"
                                 report_mutations  = "1"
                                 verbose           = "1"/>
         <ShapeComplementarity   name              = "flt_sc_1"
                                 sym_dof_name      = "JUMP1_to_redir"
                                 confidence        = "0" />
         <ShapeComplementarity   name              = "flt_sc_2"
                                 sym_dof_name      = "JUMP2_to_redir"
                                 confidence        = "0"/>
         <Sasa                   name              = "flt_sasa"
                                 threshold         = "0"
                                 sym_dof_names     = "JUMP1_to_redir,
                                                      JUMP2_to_redir"
                                 confidence        = "0"/>
         <SymUnsatHbonds         name              = "flt_unsat"
                                 sym_dof_names     = "JUMP1_to_redir,
                                                      JUMP2_to_redir"
                                 cutoff            = "20"
                                 verbose           = "1"
                                 confidence        = "0"/>
    </FILTERS>
```

```
<MOVERS>
    <SetupForSymmetry           name                = "mv_setup_symm"/>
    <SavePoseMover              name                = "mv_savepose"
                                restore_pose        = "0"
                                reference_name      = "ref_pose"/>
    <MakePolyX                  name                = "mv_makepolya"
                                aa                  = "ALA"
                                keep_pro            = "1"
                                keep_gly            = "1"
                                keep_disulfide_cys  = "1"/>
    <SwitchResidueTypeSetMover  name                = "mv_cen"
                                set                 = "centroid"/>
    <SymDockProtocol            name                = "mv_symdock"
                                fullatom            = "0"
                                local_refine        = "0"
                                docking_score_low   = "cen"/>
    <GenericMonteCarlo          name                = "mv_mc_dock"
                                trials              = "20"
                                mover_name          = "mv_symdock"
                                scorefxn_name       = "cen"
                                preapply            = "0"
                                drift               = "0"
                                recover_low         = "1"
                                sample_type         = "low"/>
    <SwitchResidueTypeSetMover  name                = "mv_full"
                                set                 = "fa_standard"/>
    <SaveAndRetrieveSidechains  name                = "mv_sidechains"
                                allsc               = "1"
                                reference_name      = "ref_pose"/>
    <ConstraintSetMover         name                = "mv_cst"
                                add_constraints     = "true"
                                cst_file            = "{FOLDER_PAR}/eosiny.cst"/>

    <FastDesign                 name                = "mv_fdesign_reduced"
                                scorefxn            = "b16"
                                repeats             = "1"
                                movemap_factory     = "mm_nojump"
                                task_operations     = "tsk_init,tsk_ic,
                                tsk_nodesign,tsk_aa_reduced,tsk_pghr,
                                tsk_limitaro,tsk_disallow_aa " />
    <FastDesign                 name                = "mv_fdesign_hbnet"
                                scorefxn            = "b16_hbnet"
                                repeats             = "1"
                                task_operations     = "tsk_init,tsk_ic,
                                tsk_nodesign,tsk_pghr,tsk_limitaro,
                                tsk_disallow_aa" />
    <FastDesign                 name                = "mv_fdesign"
                                scorefxn            = "b16"
                                repeats             = "1"
                                task_operations     = "tsk_init,tsk_ic,
                                tsk_nodesign,tsk_pghr,tsk_limitaro,
                                tsk_disallow_aa" />
    <InterfaceScoreCalculator   name                = "mv_isc"
                                chains              = "A"
                                scorefxn            = "b16" />
</MOVERS>

<PROTOCOLS>
    ### Apply symmetry file ###
    <Add mover_name  = "mv_setup_symm"     />

    ### Dock in centroid mode with poly-Ala chains ###
    <Add mover_name  = "mv_savepose"       />
    <Add mover_name  = "mv_makepolya"      />
    <Add mover_name  = "mv_cen"            />
    <Add mover_name  = "mv_mc_dock"        />
    <Add filter_name = "flt_cc_jump"       />
    <Add mover_name  = "mv_full"           />
    <Add mover_name  = "mv_sidechains"     />
```

```
        ### Design step ###
        <Add mover_name  = "mv_cst"              />
        <Add mover_name  = "mv_fdesign_reduced"/>
        <Add mover_name  = "mv_fdesign"          />  ### either with or without hbnet
        <Add mover_name  = "mv_fdesign"          />

        ### Filters and movers for scoring ###
        <Add mover_name  = "mv_isc"              />
        <Add filter_name = "flt_mutations"       />
        <Add filter_name = "flt_sc_1"            />
        <Add filter_name = "flt_sc_2"            />
        <Add filter_name = "flt_sasa"            />
        <Add filter_name = "flt_unsat"           />
    </PROTOCOLS>

</ROSETTASCRIPTS>
```

Fig. 27 | XML file for Rosetta docking and design.

5.1.2 Molecular Dynamics Simulations

MD Setup In the MD simulations, periodicity was used to simulate an infinitely repeating 2D array. To achieve this, a single subunit needed to be aligned correctly in the simulation box to create an array with the same configuration as in Rosetta. To this end, chain A was extracted from the final Rosetta pose. The monomer was centered and aligned along the principal axes, with the longest protein axis being aligned along the x-axis using "editconf" in GROMACS.[39] The monomer was then rotated around the x-axis based on the angle between the iron heme iron atoms in the Rosetta 9-mer pose. This was repeated in the z-axis to get the final correct alignment to simulate the 2D-array. The box dimensions were then calculated by taking the distance of the heme iron atoms between subunits in both the y- and z-axis. In the x-axis, a fixed size of 140 nm was defined to allow for extra water between the top and bottom of the different arrays across the periodic boundary. The pose was solvated in the defined box using "solvate" in GROMACS.

Finally, the system needed to be parameterized before starting the MD runs. The amber ff19SB force field[40] was used as the protein force field with the OPC water model[41]. GAFF parameters[42] were used for ligands, with additional atom types for heme derived from the b-type heme in cytochrome c oxidase ligated to histidine[43]. The iron of the heme group was bonded to the two ligating histidines. The box was defined with the previously used box sizes and Na^+ or Cl^- ions, depending on the charge of the protein, were added to achieve neutral charge. The tleap[44] input file used to build this system can be found in Fig. 28.

```
source leaprc.protein.ff19SB
source leaprc.gaff
source leaprc.water.opc

### Longhua, 2016; Noodleman, Inorg. Chem 2014; Fee J.Am.Chem.Soc 2008

addAtomTypes {
    { "FE" "Fe" "sp3" }  # Prevents sp0 errors in leap.
    { "NO" "N" "sp2"  }  # Modified by George to define NO and NP atoms as sp2
hybridised.
    { "NP" "N" "sp2"  }
}

### Load ligand parameters:
loadamberprep   """+FOLDER_PAR+"""/EOY_modified.prepi
loadamberparams """+FOLDER_PAR+"""/EOY.frcmod


### Load hemeb and ligand parameters
loadamberparams """+FOLDER_PAR+"""/heme.frcmod
loadoff         """+FOLDER_PAR+"""/hemeb.lib

mol = loadpdb """+FOLDER_PREP+"/"+STRUCTURE+"""_A_princ_rot_solv_corr.pdb

### Bond iron to ligating histidine nitrogens
bond mol."""+RESI[0]+""".NE2 mol."""+RESI[2]+""".FE
bond mol."""+RESI[1]+""".NE2 mol."""+RESI[2]+""".FE


set mol box { """+str(X)+""" """+str(Y)+""" """+str(Z)+""" }
addIonsRand mol Na+ 0
addIonsRand mol Cl- 0

saveamberparm mol """+FOLDER_IN+"/"+STRUCTURE+"/"+STRUCTURE+""".parm7 \
                  """+FOLDER_IN+"/"+STRUCTURE+"/"+STRUCTURE+""".rst7
savepdb mol       """+FOLDER_IN+"/"+STRUCTURE+"/"+STRUCTURE+""".pdb
quit
```

Fig. 28 | Tleap input file. Structure-dependent information is given in all caps and between quotation marks and plus signs.

MD run MD simulations were run in Amber22[45] with PMEMD[46–48] implementation. For initialization from the Rosetta structures, the structure was minimized over 100'000 cycles, switching from the steepest descent to the conjugate gradient method after 10'000 cycles. The nonbonded distance cutoff was set to 8 Å. Periodic boundaries were imposed with constant volume (Fig. 29).

```
minimize
    &cntrl
        imin=1,maxcyc=100000,ncyc=10000,
        cut=8.0,ntb=1,
        ntpr=1000,
        nmropt=1
    /
    DISANG = {FOLDER_IN}/{STRUCTURE}/EOY.rst
    &wt TYPE='END' /
    &end
```

Fig. 29 | Minimization input file. Structure dependent information is give in curly brackets.

The system was heated from 0 K to 300 K over 50 ps in 25'000 2 fs steps, with coordinates from the minimized structure. The nonbonded distance cutoff was set to 8 Å, and the SHAKE[49] algorithm was applied to constrain all bonds involving hydrogen. Periodic boundaries were imposed with constant volume. The Langevin thermostat was used with a collision frequency of 2 ps^{-1} (Fig. 30).

```
0.05 ns heating
    &cntrl
        imin=0,irest=0,ntx=1,
        nstlim=25000,dt=0.002,
        ntc=2,ntf=2,
        cut=8.0, ntb=1,
        ntpr=25000, ntwx=25000,
        ntt=3, gamma_ln=2.0,
        tempi=0.0, temp0=300.0,
        nmropt=1, iwrap=1
    /
    &wt TYPE='TEMP0', istep1=0, istep2=25000,
        value1=0.1, value2=300.0, /
    &wt TYPE='END' /
    DISANG = {FOLDER_IN}/{STRUCTURE}/EOY.rst
```

Fig. 30 | Heating input file. Structure dependent information is given in curly brackets.

During minimization, heating, and the first 1 ns of equilibration, one-sided harmonic restraints were applied for all atoms hydrogen bonding to EOY. An upper distance bound of 2.5 Å and a linear response region after 10 Å was defined with a force constant of 10 kcal/mol*Å. Each structure needed a separate restraint file due to mutations in the variants resulting in different atom numbers. The file was given in the input files as a DISANG file, and nmropt was set to 1 where the restraint file was applied. An example restraint file with CPPTRAJ commands used for generation is shown in Fig. 31.

```
### :EOY@O1 :175@NH2 r1 0 r2 0 r3 2.5 r4 10 rk2 0 rk3 10 ###
### :EOY@O2 :175@NE  r1 0 r2 0 r3 2.5 r4 10 rk2 0 rk3 10 ###
### :EOY@O2 :179@OG1 r1 0 r2 0 r3 2.5 r4 10 rk2 0 rk3 10  ###
### :EOY@O3 :80@NE r1 0 r2 0 r3 2.5 r4 10 rk2 0 rk3 10 ###
### :EOY@O3 :80@NH2 r1 0 r2 0 r3 2.5 r4 10 rk2 0 rk3 10 ###
### :EOY@O4 :183@OG1 r1 0 r2 0 r3 2.5 r4 10 rk2 0 rk3 10 ###

 &rst iat=3516,3053,0
   r1=0.000000, r2=0.000000, r3=2.500000, r4=10.000000, rk2=0.000000,
rk3=10.000000,
   nstep1=0, nstep2=0,
 &end
 &rst iat=3517,3047,0
   r1=0.000000, r2=0.000000, r3=2.500000, r4=10.000000, rk2=0.000000,
rk3=10.000000,
   nstep1=0, nstep2=0,
 &end
 &rst iat=3517,3131,0
   r1=0.000000, r2=0.000000, r3=2.500000, r4=10.000000, rk2=0.000000,
rk3=10.000000,
   nstep1=0, nstep2=0,
 &end
 &rst iat=3497,1449,0
   r1=0.000000, r2=0.000000, r3=2.500000, r4=10.000000, rk2=0.000000,
rk3=10.000000,
   nstep1=0, nstep2=0,
 &end
 &rst iat=3497,1455,0
   r1=0.000000, r2=0.000000, r3=2.500000, r4=10.000000, rk2=0.000000,
rk3=10.000000,
   nstep1=0, nstep2=0,
 &end
 &rst iat=3506,3212,0
   r1=0.000000, r2=0.000000, r3=2.500000, r4=10.000000, rk2=0.000000,
rk3=10.000000,
   nstep1=0, nstep2=0,
 &end
```

Fig. 31 | Restraint file generated in CPPTRAJ. The commands given to CPPTRAJ can be found in comments at the top of the file.

Equilibration was performed using 2 fs time steps, with $5*10^5$ steps for 1 ns, $5*10^6$ steps for 10 ns, and $5*10^7$ for 100 ns. One nanosecond of equilibration was performed after the minimization with EOY hydrogen bond restraints, before removing it for the rest of the simulation. The nonbonded distance cutoff was set to 8 Å, and the SHAKE[49] algorithm was applied to constrain all bonds involving hydrogen. Importantly, to allow for dissociation of the 2D array, the system was changed to an NPT ensemble, keeping the pressure instead of the volume constant with anisotropic pressure scaling. This way, water is allowed to flow in between the protein interfaces if energetically favorable, expanding the y- and z-axes of the box in the process. The Langevin thermostat was used with a collision frequency of 2 ps^{-1} with a constant temperature of 300 K (Fig. 32).

```
1.0 ns equilibration WITH CONSTRAINTS!!! 10 ps per frame
    &cntrl
        imin=0,irest=1,ntx=5,
        ntpr            = 500000,  !save energy every n steps
        ntwx            = 5000,    !save coordinates every n steps
        ntwr            = 500000, !save restrt file every n steps
        nstlim          = 500000, !number steps
        dt=0.002,
        ntc=2,ntf=2,
        cut=8.0,
        ntb=2, ntp=2, taup=1.0,
        ntt=3, gamma_ln=2.0,
        temp0=300.0,
        nmropt=1, iwrap=1
    /
    DISANG = {FOLDER_IN}/{STRUCTURE}/EOY.rst
    &wt TYPE='END' /
10.0 ns equilibration 10 ps per frame
&cntrl
    imin=0,irest=1,ntx=5,
    ntpr            = 500000,  !save energy every n steps
    ntwx            = 5000,    !save coordinates every n steps
    ntwr            = 5000000, !save restrt file every n steps
    nstlim          = 5000000, !number steps
    dt=0.002,
    ntc=2,ntf=2,
    cut=8.0,
    ntb=2, ntp=2, taup=1.0,
    ntt=3, gamma_ln=2.0,
    temp0=300.0, iwrap=1
/
100.0 ns equilibration 10 ps per frame
&cntrl
    imin=0,irest=1,ntx=5,
    ntpr            = 500000,  !save energy every n steps
    ntwx            = 5000,    !save coordinates every n steps
    ntwr            = 5000000, !save restrt file every n steps
    nstlim          = 500000000, !number steps
    dt=0.002,
    ntc=2,ntf=2,
    cut=8.0,
    ntb=2, ntp=2, taup=1.0,
    ntt=3, gamma_ln=2.0,
    temp0=300.0, iwrap=1
/
```

Fig. 32 | MD input files. The first nanosecond of equilibration was performed with the restraints. 10 ns equilibration was run initialized from the first nanosecond, and 100 ns from the 10 ns for selected variants.

Resulting trajectories were analyzed with CPPTRAJ[37]. Evaluation was segmented as much as possible to prevent any unwanted influence of different functions on each other. The RMSD and secondary structure per residue were evaluated in cpptraj_eval_PROT, as well as generating a pdb file for visualization of the structure. Evaluation of EOY RMSD and hydrogen bond number was performed in cpptraj_eval_EOY. Since no alignment was performed for the ligand RMSD, protein RMSD was evaluated before to align the proteins to each other.

To evaluate the protein interfaces across the periodic boundaries, the contact count (cpptraj_eval_CC) and hydrogen bond count (cpptraj_eval_HB) were evaluated before and after removing the periodic box information. The difference could then be calculated between the two generated files for each frame. Additionally, the water count close to the protein-protein interfaces were counted (cpptraj_eval_WAT) by finding all water molecules in a 10 Å distance of residues outside the loops (Fig. 33). The values in the files generated by CPPTRAJ were averaged over the last 2 ns for the 10 ns simulations and over the last 20 ns for the 100 ns simulations.

```
### cpptraj_eval_PROT
parm {FOLDER_EVAL}/{STRUCTURE}.parm7
trajin {FOLDER_EVAL}/{STRUCTURE}_eq_NPT.nc
autoimage
rmsd solv_RMSD @CA out {FOLDER_EVAL}/{STRUCTURE}_prot_RMSD.dat
secstruct HELIX :1-196 out {FOLDER_EVAL}/{STRUCTURE}_secstruct.dat
outtraj {FOLDER_EVAL}/{STRUCTURE}.pdb onlyframes
100,200,300,400,500,600,700,800,900,1000

### cpptraj_eval_EOY
parm {FOLDER_EVAL}/{STRUCTURE}.parm7
trajin {FOLDER_EVAL}/{STRUCTURE}_eq_NPT.nc
autoimage
strip :WAT,Na+,Cl-
rmsd PROT_RMSD @CA out {FOLDER_EVAL}/{STRUCTURE}_prot_RMSD_forEOY.dat
rmsd EOY_RMSD :EOY out {FOLDER_EVAL}/{STRUCTURE}_EOY_RMSD.dat nofit
hbond EOY_HB acceptormask :EOY out {FOLDER_EVAL}/{STRUCTURE}_EOY_HB.dat

### cpptraj_eval_CC
parm {FOLDER_EVAL}/{STRUCTURE}.parm7
trajin {FOLDER_EVAL}/{STRUCTURE}_eq_NPT.nc
autoimage
nativecontacts :1-198 distance 5.0 skipnative out
{FOLDER_EVAL}/{STRUCTURE}_contacts.dat
box nobox
nativecontacts :1-198 distance 5.0 skipnative out
{FOLDER_EVAL}/{STRUCTURE}_contacts_nobox.dat

### cpptraj_eval_WAT
parm {FOLDER_EVAL}/{STRUCTURE}.parm7
trajin {FOLDER_EVAL}/{STRUCTURE}_eq_NPT.nc
autoimage
mask "(:10-40,57-89,110-140,157-189@CA<:10.0)&:WAT" name M out
{FOLDER_EVAL}/{STRUCTURE}_M.dat


### cpptraj_eval_HB
parm {FOLDER_EVAL}/{STRUCTURE}.parm7
trajin {FOLDER_EVAL}/{STRUCTURE}_eq_NPT.nc
autoimage
hbond BOX_HB_test donormask :1-198 acceptormask :1-198 image out
{FOLDER_EVAL}/{STRUCTURE}_HB.dat series uuseries
{FOLDER_EVAL}/{STRUCTURE}_HB_series.dat
box nobox
hbond NOBOX_HB_test donormask :1-198 acceptormask :1-198 image out
{FOLDER_EVAL}/{STRUCTURE}_HB_nobox.dat series uuseries
{FOLDER_EVAL}/{STRUCTURE}_HB_nobox_series.dat
```

Fig. 33 | CPPTRAJ input files. Each comment line defines a new file.

5.1.3 Nanoparticle Linker Design

Initial linker lengths were taken from previous work successfully linking DARPins to the scaffold T33-21.[29] The linker was then extended or shortened by adding more or less residues from the scaffold to the sequences. The integrity of the α-helix was checked by predicting the structure in AlphaFold2, and the pLDDT was taken as an approximation for rigidty[34]. The alphafold2_multimer_v3 model was used with a maximum of 3 recycles (ColabFold v1.5.5: AlphaFold2 using MMseqs2; https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb?pli=1#scrollTo=G4yBrceuFbf3).[50] For well performing linker lengths, an additional sequence was tested with two amino acids removed from the m4D2 variant.

ProteinMPNN was performed on each variant showing a continuous α-helix from the scaffold to m4D2. All residues in the scaffold within 5 Å of m4D2 (selected in PyMol) as well as all His and Cys in the scaffold were selected for redesign. For EMT2, EMT3, and EMT4 the entire scaffold was freed for design, allowing for full redesign of the scaffold. The model v_48_020 was used and the input was given as a homooligomer. A total of 32 sequences were designed at a sampling temperature of 0.1, and a consensus sequence of the best scoring sequences was taken. (ProteinMPNN in Jax!; https://colab.research.google.com/github/sokrypton/ColabDesign/blob/v1.1.1/mpnn/examples/proteinmpnn_in_jax.ipynb#scrollTo=IVmFMidn965N). These sequences were re-evaluated in AlphaFold2 as for the pre-ProteinMPNN variants.

5.2 Experimental Methods

5.2.1 Expression & Purification

Transformation. pET21 plasmids were commercially ordered with the gene for the protein of interest inserted. Plasmids were diluted 10ng/µL with Mili-Q H_2O and stored at -20 °C. An overnight culture of T7 Express Competent *E. coli* was diluted 1:100 in 5 mL Luria-Bertani (LB) medium and grown to an OD_{600} of ~0.8. The culture was cooled on ice and the cells spun down at 10'000 rpm for 20 s at 4 °C (Centrifuge 5424; *Eppendorf*). The pellet was washed three time by adding 1 ml ice-cold sterile water, before finally being resuspended in 50 µL ice-cold sterile water. The cells were transferred to an electroporation tube and 5 µL of previously diluted plasmid was added. Electroporation was performed at a ~1.8 kV pulse for 5 µs (MicroPulser; *BioRad*). 1 mL of LB medium was quickly added, and the cells were allowed to recover at rt for 1 h. 200 µL of culture was then streaked out on a LB agar plate with 0.1 mg/mL carbenicillin.

Expression. All proteins were expressed from a pET-21 vector in T7 Express Competent *E. coli.* Starter overnight cultures, inoculated from one colony of a streak out from transformation or glycerol stock, were grown in LB medium with 0.1 mg/mL Carbenicillin. Expression cultures of 2 L LB medium with 0.1 mg/mL Carbenicillin were inoculated with 2 mL (1:1000) of the overnight cultures. Glycerol stocks (1:1 with 50% Glycerol) and cells for plasmid purification via Miniprep (Plasmid Miniprep - Classic; *ZymoResearch*) for sequencing were taken from the same overnight cultures, if necessary. The expression cultures were incubated at 37 °C and 200 rpm to an OD_{600} of ~0.8 before inducing with 1 mM IPTG. After induction, the cultures were incubated overnight at 18 °C and 200 rpm. Cells were collected through centrifugation (Avanti JXN-26; *Beckman Coulter*; JLA-8.1000 Rotor) at 6'000 rpm for 20 min at 4 °C. The supernatant was discarded, and the pellet resuspended in 2x 40 mL MonoQ A8 buffer (20 mM NaCl, 20 mM NaH_2PO_4, pH 8.0). A small amount of lysozyme and polymyxin b were added before freezing at -20 °C overnight.

Purification. The previously frozen samples were thawed, and all following steps were done on ice or at 5 °C. To further lyse the cells, the samples were sonicated (Sonopuls HD 4100; *Bandelin electronic*)

with a TS 106 Sonotrode for 5 min at 60% amplitude pulses of 1 s on and 1 s off. The debris was removed by centrifugation (Avanti JXN-26; *Beckman Coulter*; JA-20 Rotor) at 18'000 rpm for 40 min at 4 °C. The supernatant was collected, and the His_6-tagged protein of interest was purified using Ni-NTA affinity chromatography (5 mL HisTrap FF; Cytiva). An ÄKTA system (ÄKTA go; *Cytiva*) was used to run all columns. After loading the sample, the column was washed with 20 CV Ni-NTA buffer A (300 mM NaCl, 50 mM NaH_2PO_4, 20 mM Imidazole, pH 8.0) before eluting in 15 mL Ni-NTA buffer B (300 mM NaCl, 50 mM NaH_2PO_4, 300 mM Imidazole, pH 8.0). If no heme loading was performed, the sample was concentrated to 5 mL (Amicon Ultra – 15; *Merck Millipore*) and run through a S75 (HiLoad 16/600, Superdex 75; *Cytiva*) or S200 (HiLoad 16/600, Superdex 200; *Cytiva*) size exclusion column, depending on protein size. Fractions corresponding to the expected mass were collected and used for further analysis or flash frozen in liquid nitrogen and stored in -80 °C.

Denatured Purification. Frozen samples were thawed, with the following steps performed on ice or at 4 °C. Cells were further lysed by sonication (Sonopuls HD 4100; *Bandelin electronic*) with a TS 106 Sonotrode for 5 min at 60% amplitude pulses of 1 s on and 1 s off. The debris was removed by centrifugation (Avanti JXN-26; *Beckman Coulter*; JA-20 Rotor) at 18'000 rpm for 40 min at 4 °C. The supernatant was discarded, and the pellet was resuspended in 40 mL MonoQ buffer A8 (20 mM NaCl, 20 mM NaH_2PO_4, pH 8.0) with an added 8M Urea or 6M Guanidinium Chloride by vortexing and another round of sonication with reduced amplitude of 50%. The sample was stored at 5 °C overnight. Sample was centrifuged again the next day and successful suspension of the protein was checked by SDS-PAGE.

Heme Loading. To prevent unspecific heme-binding through the His_6-tag, a cleavage step with a TEV protease to remove the tag was performed after Ni-NTA affinity chro matography. The protein needed to be changed to TEV buffer for this purpose. This was done by anion exchange chromatography (5 mL HiTrap Q HP; *Cytiva*). The eluted sample in Ni-NTA buffer B (300 mM NaCl, 50 mM NaH_2PO_4, 300 mM Imidazole, pH 8.0) was diluted to 250 mL with TEV buffer A (20mM Na_2HPO_4, 0.5 mM EDTA, pH 9.0) before being loaded onto the column. The column was washed with 20 CV TEV buffer A before eluting in 15 mL TEV buffer B (1 M NaCl, 20 mM Na_2HPO_4, 0.5 mM EDTA, pH 8.0). 1 mM TCEP and ~ 1mg His_6-tagged TEV protease were added and incubated at rt overnight. The next day, another buffer exchange was done to remove EDTA. For this purpose, the protein was diluted to 250 mL with MonoQ buffer A9 (20 mM NaCl, 20 mM NaH_2PO_4, pH 9.0) before being loaded the same anion exchange column. The column was washed with 20 CV MonoQ buffer A9 before eluting in 15 mL MonoQ buffer B (1 M NaCl, 20 mM NaH_2PO_4, pH 8.0). From this sample, to remove the TEV protease, a reverse Ni-NTA column was performed manually, injecting the 15 mL sample and washing the column with Ni-NTA buffer A, collecting the entire 30 mL. The protein concentration of the resulting sample was measured and diluted to 50 mL with MonoQ buffer A8 (20 mM NaCl, 20 mM NaH_2PO_4, pH 8.0) and adding CHES (N-cyclohexyl-2-aminoethanesulfonic acid) to a concentration of 10 mM. Heme was diluted to 1.5 mL in water based on protein concentration, and 1 mL of 500 mM CHES and 2.5 mL DMSO was added. These 5 mL were then

added to the protein sample dropwise over ~1 h with a syringe pump under constant stirring with a magnetic stir bar at ~200 rpm. The sample was then concentrated to 5 mL before being injected into S75 (HiLoad 16/600, Superdex 75; *Cytiva*) or S200 (HiLoad 16/600, Superdex 200; *Cytiva*) SEC, based on the size of the protein, with MonoQ buffer A8 as the running buffer. Elute was collected after 40 mL over 92 1 mL fractions in a 96-well plate, and individual fractions were pooled according to chromatogram peaks and used for further analysis or flash frozen in liquid nitrogen and stored at -80 °C.

Hemochrome Assay Pyridine hemochrome assays were done in transparent 96-well plates (Microplate, 9-well, PS, F-bottom, clear; *Greiner*) to determine protein concentrations after heme loading. 8 wells were filled with 200 µL of protein sample each. 200 µL Pyridine buffer (0.5 M NaOH, 10% Pyridine) was added to each well, adding a small amount Sodium Hydrosulfite to four of the wells to reduce the heme. The pathlength corrected absorbance was measured in a plate reader. (Spark; *Tecan*) at 540 nm, 556 nm, 700 nm, as well as a scan from 300 nm - 700 nm. ΔA was calculated **(Eq. 2)** and the extinction coefficient of pyridine-bound heme of 23970 M^{-1} cm^{-1} was used to calculate the heme concentration. The protein concentration of the sample could thus be deduced under the assumption of a fully heme-loaded sample.

$$\Delta A = (A_{556}^{red} - A_{700}^{red}) - (A_{540}^{ox} - A_{540}^{ox}) \qquad \text{Eq. 2}$$

SDS-PAGE SDS-PAGE was performed using a prebought polyacrylamide gels (Bolt 4 to 12%, Bis-Tris Plus WedgeWell; *Invitrogen*). 20 µL of protein sample was mixed with 5 µL of 5x Laemmli buffer and denatured at 95 °C for 5 min. 10 µL of the sample was then pipetted into a gel well with 1 µL protein ladder (PageRuler Unstained Broad Range; *ThermoFischer*) as a reference in a different well. The gel was run in 300 mL running buffer (Bolt MES SDS; *Invitrogen*) with 160 kV for 30 min. The gel was stained in 25 mL protein stain (QuickBlue; *LubioScience*).

5.2.2 EOY Binding

Circular Dichroism. CD binding titrations were measured on an Applied Photophysics Chirascan Plus CD. All three XXE2H variants were diluted to 10 µM in MonoQ A buffer (20 mM NaCl, 20 mM NaH_2PO_4, pH 8.0) to a final volume of 800 µL in a 1 cm pathlength quartz cuvette. A second 500 µL sample containing10 µM protein and additionally 250 µM EOY was prepared and titrated stepwise to the first sample over 15 injections (Tab. 3). The CD spectra was taken from 280 nm to 650 nm in a 5 nm steps with 2 s per measurement time per step, and three repeats were done for each injection. The buffer background was subtracted from each measurement.

Tab. 3 | Injection steps for CD binding titration.

Injection	V_{added} (µL) [a]	V_{total} (µL) [b]	c_{EOY} (µM) [c]
0	0	0	0
1	1	1	0.3
2	1.5	2.5	0.8
3	2	4.5	1.4
4	3	7.5	2.3
5	4	11.5	3.5
6	6	17.5	5.4
7	8	25.5	7.7
8	10	35.5	10.6
9	15	50.5	14.8
10	20	70.5	20.2
11	30	100.5	27.9
12	40	140.5	37.3
13	60	200.5	50.1
14	80	280.5	64.9
15	100	380.5	80.6

[a] Volume added at each injection step.

[b] Total volume after each infection step.

[c] Total concentration of EOY after each injection step.

The binding affinity of each variant was determined using a quadratic binding equation **(Eq. 3)**, globally fitting the CD absorbance (A) between 385 nm and 435 nm, as well as 535 nm and 560 nm, where [EOY] and [enz] are the concentrations of eosin Y and the enzyme, respectively, and K_d is the dissociation constant. The equation included a non-specific binding term (ns) as well as fitting parameters A_{max} and c.

$$A = A_{max}\left([EOY] + [enz] + K_d - \sqrt{([EOY] + [enz] + K_d)^2 - 4[EOY][enz]}\right) + ns[EOY] + c \qquad \text{Eq. 3}$$

Isothermal Titration Calorimetry. ITC binding titrations were performed on a Malvern Panalytical MicroCal iTC200. EOY was diluted to in MonoQ A buffer (20 mM NaCl, 20 mM NaH_2PO_4, pH 8.0) and added over 17 steps of 2 µL to each XXE2H variant. Protein concentrations were 58.8 µM, 44.4 µM, and 12.3 µM for XXE2H variants 1-3, respectively. EOY concentrations were 1 mM for XXE2H-2 and 0.5 mM for XXE2H-1 and -3. The baseline was determined, and the heat of injection calculated by integrating the DP of each peak over time. The resulting datapoints were fitted with the Levenberg-Marquardt algorithm using a heteroatomic single-site binding model.

5.2.3 Assembly characterization

Mass Photometry. Measurements were performed on a Refeyn OneMP mass photometer. The instrument was calibrated with manufacturer's calibrants. Samples were diluted to 500 nM shortly before measurement. 9 µL MonoQ A buffer (20 mM NaCl, 20 mM NaH_2PO_4, pH 8.0) was applied onto the slide in a gasket chamber, and 1 µL of the sample was added to the buffer droplet and vigorously mixed by pipetting. Events were recorded over 1 min and analyzed with the manufacturer's software.

Negative Stain Electron Microscopy. Copper EM grids were glow discharged and the protein sample was applied and left to adsorb to the grid for 60 s. The grid was washed twice with ddH_2O and before staining in uranyl formate by applying once to the sample quickly and left on the grid for 60 s a second time. Excess liquid was removed by quickly blotting with filter paper between each step. Images were collected on a FEI Tecnai G2 spirit TEM for EMA variants and on a FEI Talos F200C TEM for EMT variants.

Cryo-Electron Microscopy. Protein sample was applied to a glow-discharged copper EM grid, blotted and plunge frozen in liquid ethane (EM GP2; *Leica*). The first dataset of 50 micrographs for m4D2-EMT1 was collected on a Talos 200 microscope at 120kx magnification and an electron dose of 60e/Å^2. All analysis of the dataset was performed in CryoSPARC. The micrographs were motion corrected and CTF estimation was performed. Blob particle picking was performed, selecting 48'448 in a total of 50 2D classes. Two 2D classes with a total of 3'059 particles were selected for template particle picking, selecting 28'086 particles over 50 2D classes. Here, 8 different classes were chosen, resulting in 5'809 particles from which an *ab initio* 3D model was built. Heterogeneous refinement was performed on these particles, and finally C_3 symmetry expansion.

6 References

1. Alberts, B. *et al.* Chapter 3. Proteins. in *Molecular Biology of the Cell* (Garland Science, New York, 2002).

2. Marsh, J. A. & Teichmann, S. A. Structure, dynamics, assembly, and evolution of protein complexes. *Annual Review of Biochemistry* **84**, 551–575 (2015). https://doi.org/10.1146/annurev-biochem-060614-034142

3. Zhu, J. *et al.* Protein Assembly by Design. *Chemical Reviews* **121**, 13701–13796 (2021). https://doi.org/10.1021/acs.chemrev.1c00308

4. Dill, K. A., Ozkan, S. B., Scott Shell, M. & Weikl, T. R. The Protein Folding Problem. *Annual Review of Biophysics* **37**, 289–316 (2008).

5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

6. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2022) doi:10.1101/2021.10.04.463034.

7. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology* **20**, 681–697 (2019). https://doi.org/10.1038/s41580-019-0163-x

8. Leaver-Fay, A. *et al.* ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol* **487**, 545–574 (2011).

9. Kuhlman, B. Designing protein structures and complexes with the molecular modeling program Rosetta. *Journal of Biological Chemistry* **294**, 19436–19443 (2019).

10. Karanicolas, J. *et al.* A De Novo Protein Binding Pair By Computational Design and Directed Evolution. *Molecular Cell* **42**, 250–260 (2011).

11. King, N. P. *et al.* Accurate design of co-assembling multi-component protein nanomaterials. *Nature* **510**, 103–108 (2014).

12. Shen, H. *et al.* De novo design of self-assembling helical protein filaments. *Science* **362**, 705–709 (2018).

13. Gonen, S., DiMaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348**, 1365–1368 (2015).

14. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

15. De Haas, R. J. *et al.* Rapid and automated design of two-component protein nanomaterials using ProteinMPNN. *BioRxiv* (2023). doi:10.1101/2023.08.04.551935.

16. Center for Sustainable Systems University of Michigan. *Photovoltaic Energy Factsheet Pub. No. CSS07-08.* (2023).

17. Ritchie, H., Roser, M. & Rosado, P. Renewable Energy. *Our World in Data* (2020).

18. Bojek, P. Solar PV. *International Energy Agency* (2023).

19. Pastuszak, J. & Węgierek, P. Photovoltaic Cell Generations and Current Research Directions for Their Development. *Materials* **15**, (2022). https://doi.org/10.3390/ma15165542

20. Sharma, K., Sharma, V. & Sharma, S. S. Dye-Sensitized Solar Cells: Fundamentals and Current Status. *Nanoscale Research Letters* **13**, (2018). https://doi.org/10.1186/s11671-018-2760-6

21. Kim, S. S., Yum, J. H. & Sung, Y. E. Improved performance of a dye-sensitized solar cell using a TiO 2/ZnO/Eosin Y electrode. *Solar Energy Materials and Solar Cells* **79**, 495–505 (2003).

22. Syafinar, R., Gomesh, N., Irwanto, M., Fareq, M. & Irwan, Y. M. Chlorophyll Pigments as Nature Based Dye for Dye-Sensitized Solar Cell (DSSC). *Energy Procedia* **79**, 896–902 (2015).

23. Ren, Y. *et al.* Hydroxamic acid pre-adsorption raises the efficiency of cosensitized solar cells. *Nature* **613**, 60–65 (2022).

24. Bunzel, H. A. *et al.* Photovoltaic enzymes by design and evolution. *bioRxiv* (2022). doi:10.1101/2022.12.20.521207.

25. Ghirlanda, G. *et al.* De novo design of a D2-symmetrical protein that reproduces the diheme four-helix bundle in cytochrome bc1. *Journal of the American Chemical Sociesty* **126**, 8141–8147 (2004).

26. Hutchins, G. H. *et al.* An expandable, modular de novo protein platform for precision redox engineering. *PNAS* **120**, (2023).

27. Elbers, P. Extending the light-harvesting spectrum of photovoltaic enzymes by computational design. (RWTH, Aachen, 2023).

28. Hutchins, G. H. *et al.* Precision design of single and multi-heme de novo proteins. *bioRxiv* (2020). doi:10.1101/2020.09.24.311514.

29. Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *PNAS* **115**, 3362–3367 (2018).

30. Fleishman, S. J. *et al.* Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One* **6**, (2011).

31. Bhardwaj, G. *et al.* Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329–335 (2016).

32. Maguire, J. B. *et al.* Perturbing the energy landscape for improved packing during computational protein design. *Proteins: Structure, Function and Bioinformatics* **89**, 436–449 (2021).

33. Lemmon, G. & Meiler, J. Rosetta ligand docking with flexible XML protocols. *Methods in Molecular Biology* **819**, 143–155 (2012).

34. Fowler, N. J. & Williamson, M. P. The accuracy of protein structures in solution determined by AlphaFold and NMR. *Structure* **30**, 925-933 (2022).

35. De Haas, R. J. *et al.* Rapid and automated design of two-component protein nanomaterials using ProteinMPNN. *bioRxiv* (2023). doi:10.1101/2023.08.04.551935.

36. Zhang, X., Meining, W., Fischer, M., Bacher, A. & Ladenstein, R. X-ray structure analysis and crystallographic refinement of lumazine synthase from the hyperthermophile Aquifex aeolicus at 1.6 Å resolution: Determinants of thermostability revealed from structural comparisons. *Journal of Molecular Biology* **306**, 1099–1114 (2001).

37. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

38. Khatib, F. *et al.* Algorithm discovery by protein folding game players. *PNAS* **108**, (2011).

39. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

40. Tian, C. *et al.* Ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation* **16**, 528–552 (2020).

41. Izadi, S., Anandakrishnan, R. & Onufriev, A. V. Building water models: A different approach. *Journal of Physical Chemistry Letters* **5**, 3863–3871 (2014).

42. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *Journal of Computational Chemistry* **25**, 1157–1174 (2004).

43. Yang, L. *et al.* Data for molecular dynamics simulations of B-type cytochrome c oxidase with the Amber force field. *Data Brief* **8**, 1209–1214 (2016).

44. Case, D. A. *et al.* AmberTools. *Journal of Chemical Information and Modeling* **63**, 6183–6191 (2023).

45. Case, D. A., Cheatham, T. E., Simmerling, C. *et al.* Amber2023. *University of California, San Francisco* (2023). https://ambermd.org/contributors.html (2023).

46. Le Grand, S., Götz, A. W. & Walker, R. C. SPFP: Speed without compromise - A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communication* **184**, 374–380 (2013).

47. Götz, A. W. *et al.* Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born. *Journal of Chemical Theory and Computation* **8**, 1542–1555 (2012).

48. Götz, A. W. *et al.* Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* **9**, 3878–3888 (2013).

49. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics* **23**, 321–341 (1977).

50. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022).

7 Supplementary

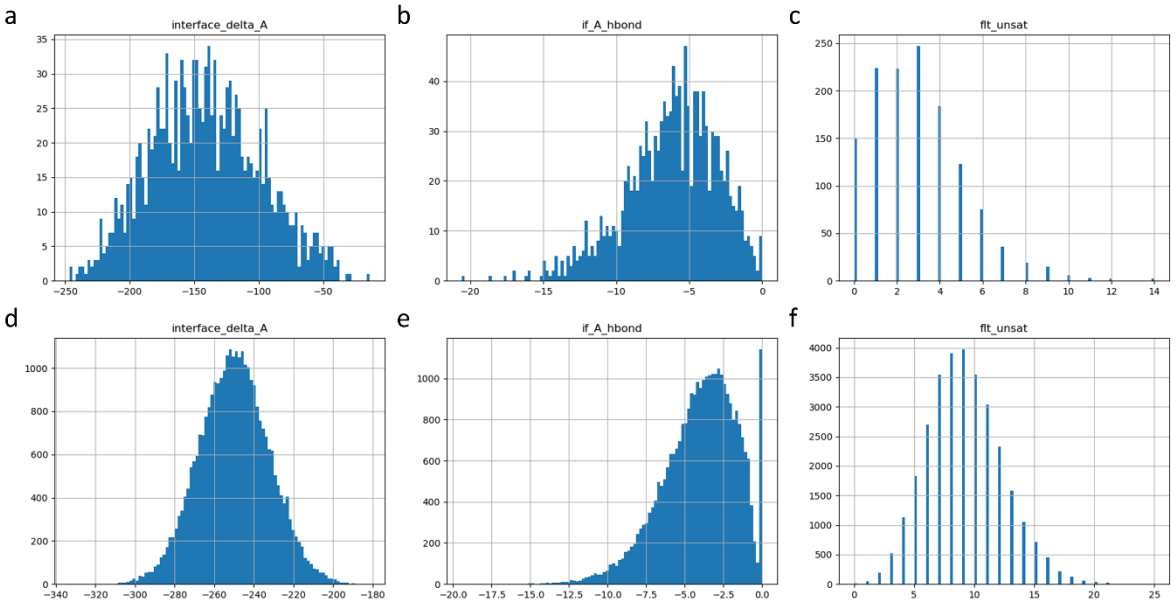


Fig. S1 | Results from Rosetta design run using HBnet compared to design run without HBnet. The interface score from the HBnet score **(a)** with a mean of c.a. -150 was drastically lower than the interface score from the run without HBnet scores **(d)**. Additionally, only slight improvements in the interface hydrogen bond score could be seen **(b,e)**. The largest difference could be seen in the number of unsatisfied buried hydrogen bonds **(c,f)**, most likely due to the inclusion of the "buried_unsat_penalty" score. 1'309 designs were done with HBnet and 37'353 designs without HBnet.

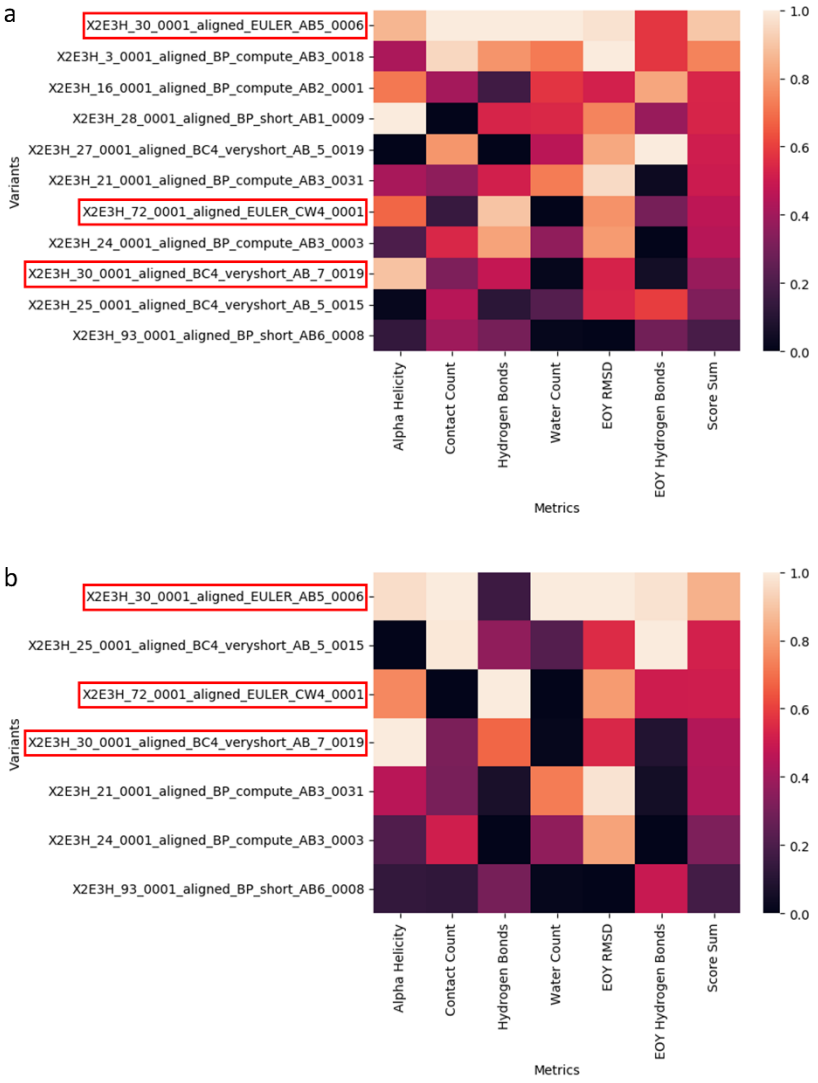


Fig. S2 | Final selection of variants after 100 ns MD simulation with full variant names. Correct analysis led to 11 variants, listed in **(a)**. However, variant ordering was based on a faulty analysis pipeline, leading to the 7 variants seen in **(b)**.

Tab. S1 | Full amino acid sequences of all tested m4D2-EMA variants with the linker sequences in blue.

Linker & Length	Sequence pre-ProteinMPNN	Sequence post-ProteinMPNN
VIGSAEL, 7	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLE SAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIGSAEL**REKLRALIEQVYATGQEMLK NTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPE LREKFRALLEQVYATGQEMLKNTSNSPELREKHRALA EQVIATWQELLKN	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKNYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMKQVDLSGYRLPGYTLA SAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIGSAEL**REKLRALIEQVYATGQEMLK NTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPE LREKFRALLEQVYATGQEMLKNTSNSPELREKHRALA EQVIATWQELLKN
VIEIGEL, 7	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLE SAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIEIGEL**REKLRALIEQVYATGQEMLK NTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPE LREKFRALLEQVYATGQEMLKNTSNSPELREKHRALA EQVIATWQELLKN	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKLYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMKQVGIKERILPGGTLA SAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIDAGEL**REKLRALIEQVYATGQEMLK NTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPE LREKFRALLEQVYATGQEMLKNTSNSPELREKHRALA EQVIATWQELLKN
VIEGSAEL, 8	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLE SAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIEGSAEL**REKLRALIEQVYATGQEML KNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSP ELREKFRALLEQVYATGQEMLKNTSNSPELREKHRAL AEQVIATWQELLKN	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKFYVNPEDKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEKVNLKSFVLPGGTLE SAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VRDGSAEL**REKLRALIEQVYATGQEML KNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSP ELREKFRALLEQVYATGQEMLKNTSNSPELREKHRAL AEQVIATWQELLKN
VIEIEGEL, 8	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLE SAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIEIEGEL**REKLRALIEQVYATGQEML KNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSP ELREKFRALLEQVYATGQEMLKNTSNSPELREKHRAL AEQVIATWQELLKN	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKWYVNPEDKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMERVNLKSFVLPGGNLE SAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**NRIITGEL**REKLRALIEQVYATGQEML KNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSP ELREKFRALLEQVYATGQEMLKNTSNSPELREKHRAL AEQVIATWQELLKN
VIEIGSAEL, 9	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLE SAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIEIGSAEL**REKLRALIEQVYATGQEM LKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPS PELREKFRALLEQVYATGQEMLKNTSNSPELREKHRA LAEQVIATWQELLKN	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKEYVDEEMAGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGQTLA SAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAL**VIDEKSAEL**REKLRALIEQVYATGQEM LKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPS PELREKFRALLEQVYATGQEMLKNTSNSPELREKHRA LAEQVIATWQELLKN
VIEIEGSAEL, 10	MRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDEL TSFIGEAKHYVDEEMKGILEEIQNDIYKIMGEIGSKG KIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLE SAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLA LSDLLFLLAR**VIEIEGSAEL**REKLRALIEQVYATGQE MLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSP SPELREKFRALLEQVYATGQEMLKNTSNSPELREKHR ALAEQVIATWQELLKN	-

7.1 Amino acid sequence of variants

XXE2H-1:
MHHHHHHGKPIPNPLLGLDSTENLYFQGSPELREKHRALAEQVYAIAKKMLELLEKLRALIEQVIATWQELLKNTSNSPELREKFRALLEQVYA
TWQELLEAVKTLRAVLEQVYATGQEMLKNGSVSPSPELREKHRALAEQLIAFAASAEELLEKFRALLEQVYATWQELLKNTSNSPELREKLRAL
IEQVIATWQELLEKAKITWAIAEQVYATGQEMLKN

XXE2H-2 (same as XXE3H):
MHHHHHHGKPIPNPLLGLDSTENLYFQGSPELREKHRALAEQLFALARRAEELLEKLRALIEQVIATWQELLKNTSNSPELREKFRALLEQVYA
TWQELLEVAREIQAVAEQVYATGQEMLKNGSVSPSPELREKHRALAEQLFAYARRQEELLEKFRALLEQVYATWQELLKNTSNSPELREKLRAL
IEQVIATWQELLEAAKRLLAIQEQVYATGQEMLKN

XXE2H-3:
MHHHHHHGKPIPNPLLGLDSTENLYFQGSPELREKHRALAEQLYALMLKMVELLEKLRALIEQVIATWQELLKNTSNSPELREKFRALLEQVYA
TWQELLETAKYARAVAEQVYATGQEMLKNGSVSPSPELREKHRALAEQVYALLKQAAELLEKFRALLEQVYATWQELLKNTSNSPELREKLRAL
IEQVIATWQELLERFLKTLALTEQVYATGQEMLKN

XXE3H-2D1:
MHHHHHHGKPIPNPLLGLDSTENLYFQGSPRLIFVHLVAAAQLWVLALVALMLLLILLVWIAQVLAFWLVLLLNTSNSPWLRWMFLFLLLWVFQ
LWLMLLWVARYIQIIAELVWANGLLMALNGSVSPSPLAWWWHIMWAAALFALARVQEYALVVFIWLLIIVIVIWYALLFNTSNSPLLWVLLTMI
IARVLALWWKLLLAALLLWQIQLLVFAMGVAMLVM

XXE3H-2D7:
MHHHHHHGKPIPNPLLGLDSTENLYFQGSPIQILWHATFAAMLFALAWVAFWALVILTVWIVWVIAVWIALMLNTSNSPLLRNIFLALLAAVLI
LWIVLYFVARLIQYIAEAVYILGLFMIFNGSVSPSPFWLLVHLVLAAMLYAIARRQEFILLIFLLLLAAVWWLWLQLVWNTSNSPIMIWWLVAL
IAQVLVIWQELLIAAAILFELQVRVWILGLAMMWL

XXE3H-2D9:
MHHHHHHGKPIPNPLLGLDSTENLYFQGSPELIWIHMMLAAMLYALAIAAMLVLMLLLIWIAWVWAAWLVLLLNTSNSPWLIWLFIWALAFVWQ
LWLFLLWTARYIQLIAELVYAFGLLMMLNGSVSPSPIFLKWHVMLAAALFALARRQELLLAIFIWLLIVVIVLWFALLFNTSNSPLLWWLLTIA
IARVLAVWYELWLAALLLWAIQILVYAWGVAMVVL

EMB:
MPHLVIEATANLRLETSPGELLEQANKALFASGQFGEADIKSRFVTLEAYRQGTAAVERAYLHACLSILDGRDIATRTLLGASLCAVLAEAVAG
GGEEGVQVSVEVREMERLSYAKRVVARQRLEHHHHHH

m4D2-EMA1:
MHHHHHHGKPIPNPLLGLDSTENLYFQRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKWYVNPEDKGILEEIQNDIYKIMG
EIGSKGKIEGISEERIAWLLKLILRYMERVNLKSFVLPGGNLESAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARNRII
TGELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELREKH
RALAEQVIATWQELLKN

m4D2-EMA2:
MHHHHHHGKPIPNPLLGLDSTENLYFQRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKHYVDEEMKGILEEIQNDIYKIMG
EIGSKGKIEGISEERIAWLLKLILRYMEMVNLKSFVLPGGTLESAKLDVCRTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARVIEG
SAELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELREKH
RALAEQVIATWQELLKN

m4D2-EMA3:
MHHHHHHGKPIPNPLLGLDSTENLYFQRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKNYVDEEMKGILEEIQNDIYKIMG
EIGSKGKIEGISEERIAWLLKLILRYMKQVDLSGYRLPGYTLASAKLDVARTIARRALRKVLTVTREFGIGAEAAAYLLALSDLLFLLARVIGS
AELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELREKHR
ALAEQVIATWQELLKN

m4D2-EMT1:
MHHHHHHGKPIPNPLLGLDSTENLYFQGRITTKVGDKGSTRLFGGEEVWKDSPIIEANGTLDELTSFIGEAKWYVNPELKGILEEIQNDIYKIM
GEIGSKGKIEGISEERIKWLEGLISEYEKLVDLKSFVLPGKTLASAKLDVARTIARRAERKVATVLREFGIGKEALVYLNRLSDLLFLMALVID
GSAELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELREK
HRALAEQVIATWQELLKN

m4D2-EMT2:
MHHHHHHGKPIPNPLLGLDSTENLYFQGKKRVDFERDDGTTKLLNGTVVPLDSPVVTAVNELDALRATLGVAKKYVNPEIAAILDWLQEQLEKA
IAEIASLGAEPGVTEEDIERVLALIKEYEKKVKETEPVLPGATLAAAYLDVAEEQADAAARAVATVLKQYGIGKLTLRYLNNLAVLLDLLALVI
AGSAELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELRE
KHRALAEQVIATWQELLKN

m4D2-EMT3:
MHHHHHHGKPIPNPLLGLDSTENLYFQGKNPISGLEDDGTTLLLNGKRVPLDSPIVQAVTQLDLTIATLGLAKGYVNPELRAILDELQELLEKA
KGEIASEGAIEGVTEEDIEWVKKKVEEYSKLVDLTKKYLPGATLASAYLDIAAANVKLAAQKVAQVLRKYGIGALTLKFLNNLAVLLKLMALVI
DGSAELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELRE
KHRALAEQVIATWQELLKN

m4D2-EMT4:
MHHHHHHGKPIPNPLLGLDSTENLYFQGKEPVTFEEDDGTTKLLTGKKIPLDSPIVEAVNELDLLRAVLGLAKSFVNPETAAILDKIQELIRKA
IGEIASLGAIEGVTEEDIKWVLEQVKKYSKLVKLTSEVLPGQTLASAYLDIASEEADRAARKVATVLKKYGIGKLTLKFLNNLAVLLDLLALVE
AGSAELREKLRALIEQVYATGQEMLKNTSNSPELREKHRALAEQVYATWQELLKNGSVSPSPELREKFRALLEQVYATGQEMLKNTSNSPELRE
KHRALAEQVIATWQELLKN

7.2 DNA sequence of variants

XXE2H-1:

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGTTCTCCTGAACTGCGCGAAAAACACAGAGCATTAGCAGAGCAAGTATACGCCATCGCTAAGAAAATGTTAGAATTATTAG
AAAAATTACGTGCCTTAATTGAACAAGTAATCGCCACATGGCAGGAGCTTCTGAAAAACACCTCCAACAGTCCTGAACTGCGTGAAAAGTTTCG
TGCGTTGTTGGAACAGGTCTACGCAACTTGGCAGGAACTGCTGGAAGCCGTCAAGACGCTGCGTGCGGTATTAGAACAAGTTTATGCTACCGGT
CAGGAGATGTTGAAAAACGGTAGTGTGTCGCCATCGCCAGAACTGCGAGAGAAGCATCGCGCGCTCGCTGAACAGCTTATTGCTTTTGCAGCCT
CAGCCGAGGAACTGCTGGAGAAATTTCGTGCGTTGCTGGAACAAGTCTACGCGACGTGGCAGGAACTGCTGAAAAATACTTCAAATAGCCCGGA
GCTTCGCGAAAAATTACGTGCCCTGATTGAACAGGTTATCGCTACCTGGCAAGAACTGCTGGAGAAAGCGAAAATCACGTGGGCTATAGCCGAG
CAAGTGTATGCCACTGGCCAGGAGATGTTAAAGAATTAATGA

XXE2H-2 (same as XXE3H):

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGAAGTCCTGAATTACGTGAGAAACATCGCGCCTTAGCGGAACAACTGTTTGCACTCGCGCGTCGTGCTGAAGAGCTGCTTG
AAAAACTGCGCGCCCTTATCGAGCAGGTGATCGCTACCTGGCAGGAATTACTGAAAAACACGAGCAATAGCCCGGAACTGCGCGAAAAATTCCG
CGCCCTCCTGGAGCAGGTATATGCAACCTGGCAGGAGTTACTGGAAGTTGCACGCGAAATCCAAGCAGTAGCCGAACAGGTATACGCTACCGGT
CAAGAGATGTTAAAGAACGGCTCAGTTTCCCCGTCTCCCGAGTTACGCGAAAAACATCGTGCATTGGCAGAGCAGTTATTTGCTTATGCGCGTC
GCCAGGAAGAACTCTTAGAGAAGTTTCGTGCACTTCTCGAACAGGTTTACGCGACATGGCAGGAGCTGTTGAAAAACACATCAAACTCACCGGA
ATTACGTGAAAAATTACGTGCCCTCATTGAACAGGTGATCGCCACGTGGCAAGAACTGCTGGAAGCCGCGAAACGTCTGCTGGCTATTCAGGAA
CAAGTATACGCGACAGGACAAGAAATGCTCAAAAACTAATGA

XXE2H-3:

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGCTCTCCAGAACTGCGGGAAAAACACCGCGCATTAGCCGAGCAGCTGTATGCGTTGATGTTGAAAATGGTTGAACTGTTAG
AAAAGCTTCGCGCCCTTATAGAACAAGTGATAGCCACATGGCAAGAGCTGCTGAAAAATACTAGTAACTCTCCGGAACTTCGTGAGAAATTTCG
TGCTCTTCTCGAACAGGTCTACGCTACCTGGCAGGAGTTACTGGAGACGGCAAAATATGCGCGTGCCGTGGCGGAACAAGTGTATGCCACTGGC
CAGGAAATGCTGAAAAATGGGAGTGTATCGCCCAGTCCGGAATTGCGAGAGAAACATCGCGCATTGGCCGAACAGGTATATGCTTTACTCAAAC
AGGCTGCAGAACTTCTGGAAAAATTTCGGGCCTTGCTCGAACAGGTTTATGCTACCTGGCAGGAGCTGTTGAAAAATACGTCGAATTCCCCTGA
ATTGCGAGAAAAGCTCCGAGCGTTAATCGAACAAGTTATCGCGACCTGGCAGGAGTTGCTGGAACGTTTTCTGAAAACGCTTGCCTTAACCGAA
CAGGTTTACGCCACAGGTCAAGAGATGCTTAAAAACTAATGA

XXE3H-2D1:

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGCTCCCCACGTTTGATTTTTGTTCATTTAGTTGCAGCCGCGCAGCTGTGGGTCTTAGCTCTCGTGGCGTTAATGCTCCTTC
TTATCCTCCTGGTTTGGATCGCCCAGGTGCTGGCTTTCTGGTTGGTACTTTTATTGAATACTAGTAATTCTCCCTGGCTGCGCTGGATGTTCCT
GTTTTTGCTGCTTTGGGTTTTCCAGTTGTGGCTGATGTTATTGTGGGTCGCTCGGTATATTCAGATTATTGCGGAATTAGTTTGGGCAAACGGA
CTGTTGATGGCCTTAAACGGTTCTGTGTCCCCGAGTCCCTTAGCCTGGTGGTGGCATATTATGTGGGCAGCCGCCCTGTTTGCCCTTGCACGTG
TCCAGGAATACGCGTTGGTTGTGTTCATTTGGCTGCTCATAATTGTTATTGTTATATGGTACGCATTATTGTTTAACACGTCGAATTCACCACT
GTTGTGGGTCTTATTGACTATGATTATCGCTAGAGTTCTGGCGCTTTGGTGGAAACTCCTGCTGGCAGCCTTACTCTTGTGGCAGATCCAACTC
CTGGTTTTCGCTATGGGCGTTGCAATGCTTGTGATGTAA

XXE3H-2D7:

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGCAGTCCAATCCAGATTCTGTGGCACGCAACCTTCGCTGCCATGCTTTTTGCCTTAGCCTGGGTTGCCTTCTGGGCATTAG
TTATCCTGACAGTCTGGATTGTTTGGGTTATAGCTGTGTGGATCGCACTGATGCTTAATACCAGTAATAGTCCTTTATTACGTAATATATTTCT
GGCATTACTCGCGGCAGTACTGATCCTGTGGATAGTGTTGTATTTCGTGGCACGTCTGATTCAGTATATCGCGGAAGCGGTTTACATCCTTGGG
TTGTTTATGATTTTTAATGGTAGCGTTAGCCCGAGCCCTTTTTGGTTACTGGTGCATCTGGTATTAGCGGCAATGCTGTACGCTATTGCTAGAC
GACAGGAATTCATCCTCCTGATTTTTCTGTTACTGTTAGCAGCCGTTTGGTGGCTTTGGCTGCAGTTGGTCTGGAATACTAGTAATTCCCCGAT
TATGATTTGGTGGCTGGTGGCACTTATAGCGCAGGTGTTGGTTATTTGGCAAGAACTTTTAATTGCAGCCGCAATCCTTTTTGAACTGCAAGTT
CGCGTCTGGATCTTAGGTTTGGCGATGATGTGGTTATAA

XXE3H-2D9:

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGGTCACCGGAACTGATATGGATCCACATGATGCTGGCAGCCATGCTGTACGCGCTTGCGATTGCAGCTATGTTAGTGTTGA
TGCTGCTCCTTATCTGGATTGCATGGGTTTGGGCAGCTTGGCTTGTGCTGCTGCTGAATACCAGTAACTCACCGTGGCTTATTTGGCTTTTTAT
CTGGGCGTTAGCATTCGTTTGGCAGCTTTGGTTATTTCTCCTTTGGACCGCTAGATATATTCAGCTGATAGCAGAATTGGTATATGCATTTGGC
CTGCTGATGATGCTGAATGGCTCCGTCAGCCCCAGCCCAATCTTCCTGAAGTGGCACGTTATGCTGGCTGCGGCTCTTTTTGCTTTAGCACGAC
GGCAGGAACTGCTTCTTGCAATCTTCATCTGGCTGTTAATCGTAGTGATCGTCCTGTGGTTCGCCCTGCTGTTTAATACCTCAAATTCTCCTTT
ACTGTGGTGGCTGCTGACAATCGCAATTGCACGTGTTTTGGCAGTTTGGTATGAGCTGTGGTTAGCAGCACTTCTGCTGTGGGCTATTCAGATT
CTGGTTTATGCATGGGGGGTTGCAATGGTGGTATTATAA

EMB:

TTAACTTTAAGAAGGAGATATACATATGCCGCATTTAGTAATTGAAGCTACGGCAAATCTTAGACTGGAAACCTCACCTGGCGAGTTACTTGAA
CAGGCAAATAAAGCACTGTTCGCATCGGGACAGTTCGGTGAAGCTGATATTAAAAGTCGATTTGTAACACTGGAGGCGTATCGACAAGGGACAG
CAGCAGTTGAACGTGCTTATCTGCATGCATGCTTGAGCATACTTGATGGGCGTGACATCGCAACACGTACACTCTTAGGTGCTTCTTTATGCGC
GGTGCTGGCTGAGGCGGTTGCAGGCGGGGGTGAGGAGGGTGTTCAAGTTTCTGTTGAGGTGCGAGAGATGGAACGTCTGTCCTATGCGAAACGT
GTAGTGGCCCGCCAGCGCCTGGAACATCATCATCATCATCATTGATGA

m4D2-EMA1:
TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGCGGATCACTACTAAAGTAGGGGATAAAGGCAGCACACGACTGTTTGGTGGCGAGGAAGTGTGGAAAGATAGTCCCATAATTG
AGGCGAATGGGACATTAGATGAGCTTACATCATTTATAGGGGAGGCCAAATGGTATGTGAATCCCGAAGACAAAGGTATTCTTGAAGAAATACA
GAACGATATCTATAAAATCATGGGAGAAATCGGATCCAAAGGCAAAATTGAAGGTATATCCGAAGAACGTATTGCCTGGCTGTTAAAACTCATC
CTGAGATATATGGAACGAGTTAATTTAAAGAGCTTTGTACTTCCGGGAGGAAACCTGGAATCAGCGAAACTCGATGTGGCCCGCACCATAGCTC
GTCGGGCTTTGCGTAAGGTACTTACGGTAACCCGTGAATTCGGGATAGGCGCGGAGGCCGCCGCATACCTGCTGGCATTATCCGACTTATTGTT
TCTGCTTGCCCGTAATCGGATTATTACCGGAGAACTTCGCGAAAAACTGCGTGCTTTGATTGAGCAGGTCTATGCCACAGGGCAGGAAATGCTT
AAAAATACATCAAACTCGCCGGAGCTGCGCGAAAAACACCGCGCACTGGCCGAGCAGGTGTATGCAACCTGGCAAGAGCTGCTCAAAAACGGGA
GCGTAAGTCCCTCACCTGAATTGCGTGAAAAATTTCGCGCGTTACTGGAACAGGTGTATGCTACCGGCCAAGAGATGCTGAAAAACACGTCTAA
TTCACCGGAGCTGAGAGAGAAACACCGTGCATTGGCAGAACAAGTAATCGCTACCTGGCAAGAGTTACTTAAAAATTAATGA

m4D2-EMA2:
TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGCGCATCACTACAAAGGTTGGTGATAAAGGCTCCACACGCCTGTTTGGTGGCGAGGAAGTCTGGAAAGACTCTCCGATAATAG
AAGCCAATGGTACACTGGATGAACTCACATCATTTATTGGTGAAGCGAAACATTATGTCGATGAGGAAATGAAGGGTATATTAGAGGAGATACA
GAACGATATTTATAAGATCATGGGTGAAATTGGTTCGAAAGGTAAAATCGAAGGCATTTCGGAAGAACGCATTGCATGGCTTTTGAAGCTTATT
CTGCGCTATATGGAAATGGTCAATCTTAAGTCGTTCGTTCTGCCCGGGGGGACGCTGGAATCCGCCAAACTCGACGTCTGCCGTACCATTGCCC
GTCGAGCACTGCGGAAGGTTTTGACCGTGACGCGTGAATTCGGCATAGGAGCCGAGGCCGCCGCCTATCTGTTGGCACTGTCAGATTTACTGTT
TCTGCTCGCTCGAGTCATTGAGGGATCTGCGGAGTTGCGCGAAAAATTACGCGCTCTTATAGAACAGGTGTACGCAACGGGTCAGGAGATGCTG
AAGAATACGTCAAACTCCCCTGAGCTTCGTGAGAAGCATCGGGCTTTGGCAGAACAAGTATACGCGACTTGGCAAGAATTGCTGAAGAATGGAT
CAGTCTCCCCGTCCCCAGAACTGCGGGAGAAATTTCGAGCATTACTGGAGCAGGTTTATGCGACAGGCCAGGAAATGCTGAAGAACACCTCTAA
TTCTCCAGAACTTAGAGAGAAACACCGAGCCTTAGCCGAGCAAGTTATCGCGACCTGGCAGGAGCTGCTTAAAAACTAATGA

m4D2-EMA3:
TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGAGAATTACCACGAAAGTGGGGGACAAAGGTAGTACTCGCTTATTTGGCGGCGAGGAGGTGTGGAAAGACAGTCCTATTATCG
AAGCGAACGGCACGCTGGACGAGTTAACAAGTTTCATTGGGGAAGCTAAAAATTACGTGGATGAAGAAATGAAAGGTATCCTTGAAGAAATACA
GAATGACATTTACAAAATTATGGGAGAGATTGGCAGTAAAGGGAAGATCGAAGGAATCTCGGAAGAACGCATAGCCTGGTTACTGAAATTAATT
CTTCGGTATATGAAACAGGTGGATCTTTCCGGATATCGCCTCCCTGGTTATACACTGGCAAGCGCGAAACTGGATGTCGCGCGTACAATAGCAC
GTCGGGCTCTTCGTAAGGTTCTTACTGTAACTCGTGAATTTGGCATTGGAGCTGAGGCCGCGGCGTATCTGTTGGCGCTCTCAGATCTTCTGTT
TTTGCTGGCGCGCGTGATTGGAAGCGCAGAGTTGCGGGAAAAACTGCGTGCACTGATTGAACAGGTGTACGCTACAGGACAGGAAATGCTTAAG
AATACCAGTAACTCTCCGGAACTGCGGGAAAAACATCGGGCGCTTGCGGAACAAGTTTACGCTACTTGGCAAGAGTTACTCAAAAATGGGTCTG
TGTCCCCGAGCCCCGAGCTTCGTGAAAAGTTTAGAGCATTGTTGGAACAGGTATACGCAACAGGGCAGGAGATGTTAAAAAATACCTCTAATTC
ACCCGAGTTACGGGAGAAACACCGTGCCCTTGCTGAACAGGTCATCGCCACTTGGCAAGAATTACTGAAAAATTAATGA

m4D2-EMT1:
TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAAT
TTGTACTTCCAGGGCAGAATTACGACCAAAGTAGGTGATAAGGGCAGTACGAGATTATTCGGAGGAGAGGAAGTCTGGAAAGACAGCCCGATCA
TTGAAGCAAACGGTACACTGGACGAACTTACCTCTTTTATTGGCGAAGCCAAATGGTATGTTAACCCGGAGCTTAAGGGGATTCTGGAGGAGAT
TCAGAATGACATATATAAGATAATGGGTGAAATTGGAAGTAAGGGGAAGATTGAAGGGATCTCCGAGGAACGAATCAAATGGCTCGAGGGACTG
ATATCTGAGTACGAAAAACTGGTCGACCTGAAAAGTTTCGTCCTGCCGGGCAAGACATTAGCCTCAGCCAAGTTAGATGTGGCACGTACGATTG
CTCGCCGTGCGGAGCGAAAAGTTGCAACAGTGTTGCGTGAATTCGGGATAGGGAAAGAGGCCCTGGTTTATTTAAATAGACTGAGTGATTTACT
GTTTTTGATGGCGCTGGTCATTGATGGGAGCGCAGAATTGCGGGAAAAACTGCGCGCGCTGATCGAACAAGTATATGCTACCGGCCAGGAAATG
CTTAAAAATACCTCGAATTCACCTGAACTTCGGGAAAAACATCGAGCATTGGCCGAGCAAGTGTACGCCACCTGGCAAGAGCTGCTGAAAAATG
GTAGCGTTAGTCCAAGTCCTGAATTACGAGAAAAATTTAGAGCCTTATTAGAGCAGGTTTATGCAACCGGGCAAGAGATGCTGAAAAACACGAG
CAACTCTCCAGAACTGCGAGAGAAGCATCGCGCACTGGCCGAGCAGGTCATCGCCACATGGCAAGAATTATTAAAAAACTAATGA

m4D2-EMT2:
TTAACTTTAAGAAGGAGATATACATATGCATCACCATCACCATCACGGTAAACCGATTCCTAACCCGCTGTTAGGGCTTGATTCCACTGAGAAC
TTATATTTTCAGGGCAAAAAACGAGTGGATTTTGAAAGAGATGACGGTACTACAAAACTTCTTAACGGCACAGTGGTGCCGCTGGACTCACCGG
TGGTCACCGCCGTGAATGAACTTGACGCACTGCGTGCAACCTTAGGAGTAGCGAAGAAATATGTAAATCCGGAAATTGCAGCCATTCTGGATTG
GCTTCAGGAACAACTTGAGAAAGCCATCGCCGAAATTGCGTCACTGGGCGCCGAGCCAGGCGTCACGGAAGAGGATATAGAGCGCGTGCTGGCA
TTGATTAAGGAATACGAGAAGAAGGTCAAAGAAACAGAACCGGTGCTCCCTGGCGCTACACTGGCGGCTGCATATTTAGATGTCGCTGAAGAGC
AAGCAGACGCAGCCGCGCGCGCCGTCGCGACAGTTTTAAAACAGTATGGAATTGGAAAATTAACACTGCGATATCTGAATAATCTGGCCGTGTT
ATTAGACCTGCTTGCGTTGGTAATTGCTGGCTCTGCTGAGCTGCGGGAAAAACTTCGTGCCCTGATCGAACAGGTCTATGCAACTGGCCAGGAA
ATGCTCAAGAACACCTCCAACTCACCCGAACTCCGCGAGAAACATCGCGCGCTGGCAGAACAGGTTTATGCTACATGGCAGGAGTTACTTAAAA
ACGGGTCTGTATCCCCCTCCCCAGAGCTTCGTGAAAAATTCCGCGCGCTCTTGGAACAGGTGTATGCTACTGGTCAGGAGATGCTGAAAAATAC
ATCCAATTCCCCAGAACTCCGTGAAAAGCATCGTGCTCTGGCGGAGCAGGTTATTGCCACGTGGCAGGAGTTATTGAAAAATTAATGA

m4D2-EMT3:
TTAACTTTAAGAAGGAGATATACATATGCATCATCATCATCACCACGGCAAGCCGATTCCAAACCCACTCCTGGGTCTTGACTCGACAGAAAAT
TTATATTTCCAGGGCAAAAACCCGATTTCTGGTTTGGAAGATGACGGTACAACGCTCTTGCTGAATGGAAAACGCGTTCCTCTGGACTCACCTA
TAGTGCAAGCGGTTACTCAGCTGGATCTGACTATCGCAACTCTGGGCCTGGCAAAAGGGTATGTTAACCCGGAGCTGCGTGCAATCCTTGACGA
GTTACAGGAGTTGTTAGAAAAAGCAAAAGGTGAAATTGCCAGCGAAGGTGCCATCGAAGGAGTTACGGAAGAGGACATCGAATGGGTGAAAAAA
AAAGTTGAAGAATATAGTAAACTGGTTGACTTAACTAAAAAGTACTTACCCGGCGCCACGCTCGCTTCGGCCTACTTAGATATTGCTGCGGCGA
ATGTGAAGCTTGCCGCACAGAAAGTGGCACAAGTGCTGAGAAAGTACGGTATTGGTGCTCTGACCCTGAAATTTCTGAACAACCTGGCAGTCCT
GCTGAAACTTATGGCGTTGGTGATCGACGGGTCGGCAGAGCTTCGAGAAAAACTGCGCGCATTGATTGAGCAGGTGTATGCTACGGGTCAGGAG
ATGCTGAAAAACACGTCAAACTCACCAGAGTTACGTGAGAAACACAGAGCTCTGGCCGAACAGGTCTATGCCACCTGGCAGGAATTGTTAAAGA
ATGGTTCTGTTAGCCCATCTCCCGAGCTGCGTGAAAAATTTAGAGCACTGCTGGAGCAAGTATATGCAACGGGCCAGGAAATGCTGAAAAATAC
GTCGAATTCTCCCGAATTACGAGAAAAACACCGTGCACTCGCGGAACAAGTAATTGCCACTTGGCAGGAACTGCTTAAAAATTAATGA

m4D2-EMT4:

TTAACTTTAAGAAGGAGATATACATATGCACCACCATCATCATCATGGTAAACCTATTCCAAATCCCCTTTTAGGTTTAGATTCCACTGAAAAT
CTTTATTTTCAGGGGAAAGAACCTGTGACTTTCGAGGAGGATGATGGGACCACCAAATTACTGACCGGCAAGAAAATCCCGCTGGACTCTCCCA
TTGTAGAAGCTGTAAATGAACTGGATCTGCTGCGTGCAGTTCTTGGCCTCGCGAAAAGTTTTGTTAATCCCGAGACGGCTGCAATCTTAGACAA
AATCCAGGAACTTATACGTAAGGCCATAGGCGAAATTGCATCTCTGGGTGCGATTGAGGGCGTAACTGAGGAAGACATAAAGTGGGTTCTGGAA
CAGGTCAAAAAATATAGCAAACTGGTAAAGTTGACTTCGGAAGTGCTGCCCGGTCAGACGCTGGCGAGTGCTTACCTGGATATTGCGTCTGAAG
AGGCCGATAGAGCAGCGCGTAAAGTAGCTACAGTACTGAAAAAATATGGCATCGGTAAACTCACGTTAAAATTTCTGAACAACCTGGCGGTTTT
GTTGGATTTGCTGGCTCTCGTTGAAGCAGGTTCCGCCGAACTGCGAGAGAAACTGCGTGCGCTCATTGAACAAGTTTATGCGACAGGTCAAGAA
ATGTTGAAAAATACCAGCAACAGCCCAGAGTTACGAGAGAAGCATCGCGCTCTGGCAGAACAGGTGTATGCAACGTGGCAGGAGTTACTGAAAA
ATGGTTCTGTGTCACCGTCACCCGAATTACGCGAGAAGTTTCGTGCCTTGTTGGAGCAAGTATACGCGACCGGACAGGAGATGCTGAAAAATAC
CAGTAATTCTCCTGAACTCCGTGAAAAACACCGTGCCCTGGCAGAGCAAGTGATCGCCACCTGGCAAGAGCTTTTAAAAAACTAATGA

Full plasmid sequence of XXE3H-2D1 in pet21:

XXE3H-2D1-pET21:

TTAACTTTAAGAAGGAGATATACATATGCACCATCACCACCATCACGGCAAGCCAATTCCGAACCCTCTGCTTGGCCTTGATTCCACCGAAAATTTGTACTTCCAGG
GCTCCCCACGTTTGATTTTTGTTCATTTAGTTGCAGCCGCGCAGCTGTGGGTCTTAGCTCTCGTGGCGTTAATGCTCCTTCTTATCCTCCTGGTTTGGATCGCCCAG
GTGCTGGCTTTCTGGTTGGTACTTTTATTGAATACTAGTAATTCTCCCTGGCTGCGCTGGATGTTCCTGTTTTTGCTGCTTTGGGTTTTCCAGTTGTGGCTGATGTT
ATTGTGGGTCGCTCGGTATATTCAGATTATTGCGGAATTAGTTTGGGCAAACGGACTGTTGATGGCCTTAAACGGTTCTGTGTCCCCGAGTCCCTTAGCCTGGTGGT
GGCATATTATGTGGGCAGCCGCCCTGTTTGCCCTTGCACGTGTCCAGGAATACGCGTTGGTTGTGTTCATTTGGCTGCTCATAATTGTTATTGTTATATGGTACGCA
TTATTGTTTAACACGTCGAATTCACCACTGTTGTGGGTCTTATTGACTATGATTATCGCTAGAGTTCTGGCGCTTTGGTGGAAACTCCTGCTGGCAGCCTTACTCTT
GTGGCAGATCCAACTCCTGGTTTTCGCTATGGGCGTTGCAATGCTTGTGATGTAACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAA
AGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCC
GGATTGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCT
TTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGA
CCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCT
TGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAA
AAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCA
AATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTT
TGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCA
ACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGG
CAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATG
CAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATG
TAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTA
ACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTT
TATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGA
GTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATT
GATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGA
CCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGG
ATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAAC
TCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACC
GGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAA
GCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTT
TATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACG
GTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCC
GCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCAATGGT
GCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCG
CTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAA
CGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAA
TGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGA
TGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAG
AGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCA
GGGCGCTGACTTCCGCGTTTCCAGACTTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTC
GCTCGCGTATCGGTGATTCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATG
CCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGAT
CATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGA
CGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATT
GCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGG
TGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGA
AAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCGTCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGT
AATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGG
CACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCC
GCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGA
GACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGA
GAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACA
ATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGTAATTCAGCTC
CGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACACCGGCATACTCTGCGACAT
CGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCCATTCGATGGTGTCCGGGATC
TCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGC
CCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATA
TAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCT

7.3 Jupyter Notebooks

All jupyter notebooks used as is in the thesis are attached hereafter as PDF printouts. Jupyter notebook files can be found under: https://github.com/ChrisKWeing/NANOPHOTO