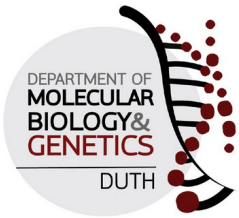
Structural modeling of rRM6
a reversed sequence protein

Μοντελοποίηση πρωτεΐνης με αντίστροφη αλληλουχία



Μαγδαληνή Χατζοπούλου, ΑΕΜ 1938

Τμήμα Μοριακής Βιολογίας και Γενετικής, Δημοκρίτειο Πανεπιστήμιο Θράκης

Φεβρουάριος 2022

Επιβλέπων καθηγητής Νικόλαος Γλυκός

Εργαστήριο Δομικής και Υπολογιστικής Βιολογίας

~~~
~~~

Acknowledgments

First, I would like to thank my supervisor for his guidance and support throughout this project and for equipping me with all the necessary tools to continue my journey. In addition I would like to thank my NMG groupie and friend Panagiotis for his constant support and company over the course of my undergraduate years.

Next, I would like to express my gratitude to my partner Angelos for his inspiring words, love and monumental support all these years and assistance during my final year of studies. I am beyond grateful also for my friends Efstathia, Kostas, Lito and Elisavet who have helped me not only during the process of my thesis but also in every step of my academic years.

Table of Contents

Abstract

The RM6 protein (PDB ID:1QX8) represents a regular and canonical helical bundle. Recent experiments studying its retro-isomer (rRM6) have suggested that it also forms an α-helical structure of high stability. Molecular replacement calculations using the RM6 protein as a model on crystallographic data from rRM6 failed to determine its structure. Hence, it was necessary to generate rRM6 models of sufficient quality in order to assist this process. The ISAMBARD program was utilized which uses geometric modeling to construct coiled-coil backbones. This tool managed to determine the structure of RM6 by producing models which exhibited RMSD values of only ~1.16 Å for 196 residues of the bundle. Therefore, it was also used in the case of rRM6 where it was shown that the left anti-parallel topology exhibited the lowest energy value after refinement (-12514 kcal/mol) while also sharing the same hydrophobic core with RM6. However, since the results did not exhibit a pronounced solution, multiple models with different topologies could be of use to facilitate the molecular replacement calculations. The determination of the rRM6 structure could contribute to our current knowledge regarding the foldability of retro-proteins and thus decipher the role of backbone directionality on protein folding.

Abbreviations

rRM6: retro-RM6
CMA -ES: Covariance Matrix Adaptation Evolution Strategy
PSO: Particle Swarm Optimization
DE: Differential Evolution
GA: Genetic Algorithm
HB: Helical Bundle

Περίληψη

Η πρωτεΐνη RM6 (PDB ID: 1QX8) αποτελεί ένα μετάλλαγμα της πρωτεΐνης ROP (PDB ID: 1ROP) και αναδιπλώνεται ως ένα σταθερό και κλασσικό α-ελικοειδές δεμάτιο. Αρχικές προσπάθειες για τον προσδιορισμό της ρέτρο-δομής της με τη μέθοδο της μοριακής αντικατάστασης και χρησιμοποιώντας την RM6 ως μοντέλο δεν οδήγησαν σε επίλυση της δομής. Προκειμένου να δημιουργηθούν μοντέλα της ρέτρο-RM6 χρησιμοποιήθηκε η μέθοδος της γεωμετρικής μοντελοποίησης η οποία αξιοποιεί παραμέτρους για να κατασκευάσει κύριες αλυσίδες σπειρωμένων σπειραμάτων. Συγκεκριμένα, χρησιμοποιήθηκε το πρόγραμμα ISAMBARD, αρχικά για την κατασκευή της RM6 με στόχο την αξιολόγηση της ικανότητας αυτου του εργαλείου. Τα αποτελέσματα έδειξαν πως το ISAMBARD κατασκεύασε μοντέλα με τιμές RMSD κοντά στα 1.16 Å για τα 196 κατάλοιπα της δομής. Τα συγκεκριμένα θεωρήθηκαν αρκετά ικανοποιητικά ακόμα και για υπολογισμούς μοριακής αντικατάστασης, εφαρμόζοντας έτσι την ίδια διαδικασία και για τη ρέτρο-RM6. Το θεωρητικό μοντέλο του αριστερού και αντιπαράλληλου α-ελικοειδές δεματίου παρουσίασε τη χαμηλότερη ενέργεια ύστερα από βελτιστοποίηση (-12514 kcal/mol). Ωστόσο εφόσον οι ενεργειακές διαφορές με τις υπόλοιπες δομές δεν ήταν σημαντικές, παραπάνω από ένα μοντέλα, με διαφορετικές τοπολογίες θα μπορούσαν να χρησιμοποιηθούν για την επίλυση της ρέτρο-RM6 με τη μέθοδο της μοριακής αντικατάστασης. Ο προσδιορισμός της δομής της ρέτρο-RM6 θα μπορούσε να απαντήσει στο ερώτημα για το αν η κατεύθυνση της κύριας αλυσίδας συμβάλλει στην αναδίπωση μιας πρωτεΐνης και ακόμα, για το πως μία πρωτεΐνη με αντεστραμμένη αλυσίδα αναδιπλώνεται.

Chapter 1 Introduction

1.1 Coiled-Coils

Coiled-coils represent an abundant motif and is present in proteins which are involved in signal transduction, molecule recognition and refolding events [1,2]. As described by Crick (1952) [3,4] two or more α-helices forming a coiled-coil interact in a “knobs-into-holes” manner meaning that one side chain (knob) of a helix is placed in a space between four side chains in the facing helix (hole) [5]. They constitute homo- or hetero- structures with parallel or anti-parallel orientation and left or right helical twist [5,6]. This packing mode entails strict regular contacts between the interacting side chains in every seven residues which are arranged over two helical turns, altering the residues per turn from 3.63 to 3.5 (7/2) [5,6]. This creates a heptad repeat -*abcdefg* (hpphppp, h: hydrophobic residue, p: polar residue), with *a* and *d* residues mostly being hydrophobic and the positioning of equivalent residues next to each other in the amphipathic α-helices [5,6]. Due to the latter this model is also referred to as in -register with some exceptions to this rule [5]. The residues occupying *a* and *d* positions form a hydrophobic stripe along the helical axis which constitutes the bundle's driving force for oligomerization [5,7]. On the contrary, the residues occupying *e* and *g* positions are mostly charged amino acids and contribute to the ionic interactions between the chains [8].

A different mode of packing for the coiled-coil structures (referred to as out-of register) is the “ridges-into-grooves” model where the interacting residue packs above or beneath its equivalent [5,9]. The main geometric difference between these packing modes is the crossing angle (the angle of a helix relative to the superhelical axis calculated in degrees) where in the former model is approximately $+20^{o}$ while in the latter $+23^{o}$ [3,4,9]. Paradigms of coiled-coil structures and their hydrophobic cores are shown in Figure 1, for the Repressor of Primer protein (ROP, PDB ID: 1ROP) and the the yeast transcriptional activator GCN4 (PDB ID: 2ZTA). The former represents a well studied four α-helical anti-parallel bundle while both of them follow the proposed “knobs-into-holes” model [10,11]. This regular nature of the α-helical coiled-coils has led to their parameterization, thus there are established equations describing key geometrical features of these structures [6,12,13]. These parameters describe each helix and its orientations in the superhelix bundle [5] and can be used to investigate backbone conformations which are not present in nature [14]. This way the unexplored “protein-fold” space can be studied through *de novo* modeling without the use of an experimentally determined structure or any sequence homology [12,14,15].

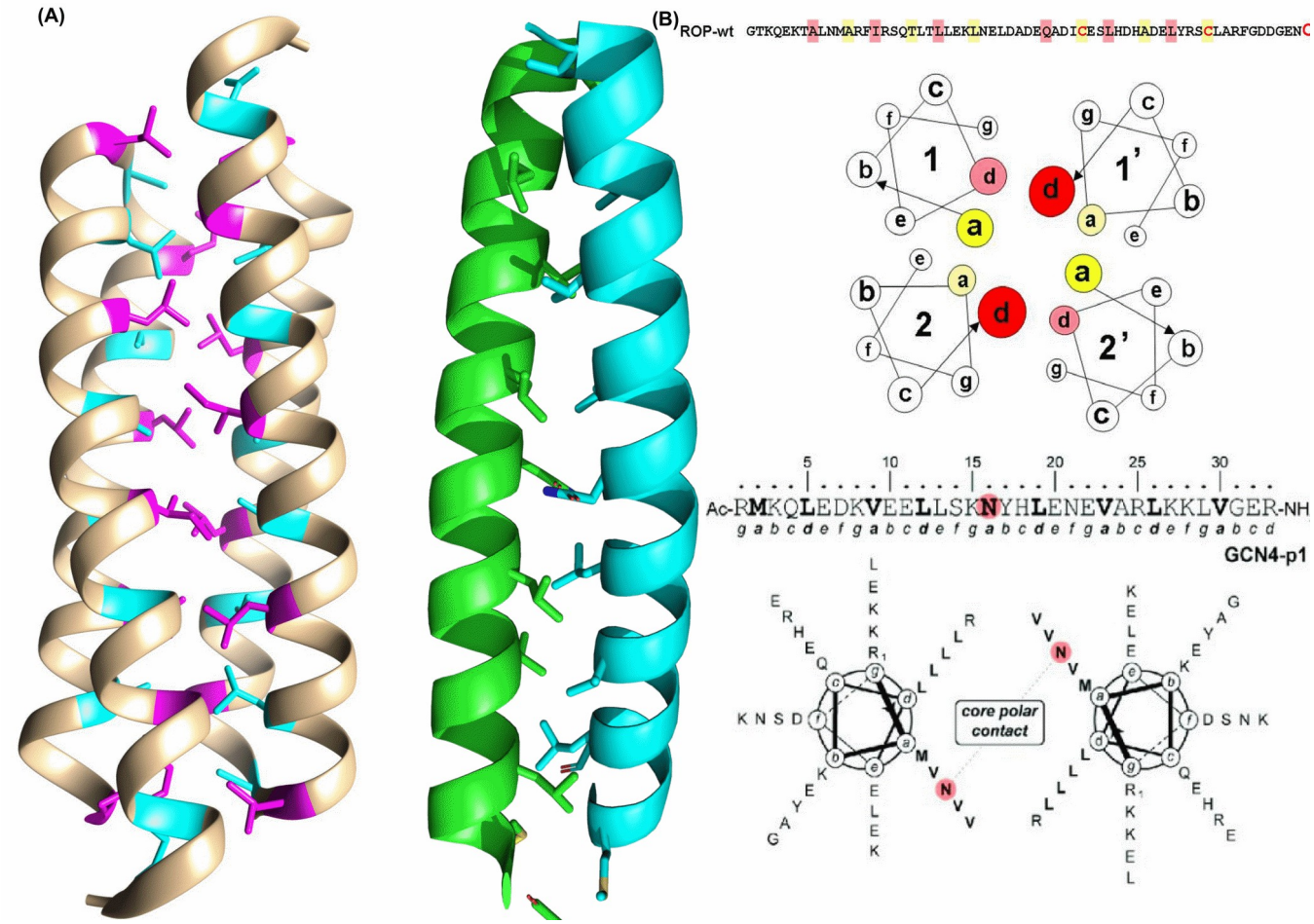


Figure 1| The ROP (PDB ID: 1ROP) and GCN4 (PDB ID: 2ZTA) proteins. (A) On the left the four α-helix Repressor of Primer protein and on the right the GCN4 protein are presented as cartoons. In both cases, the side chains that participate in the formation of the hydrophobic core are depicted as sticks. (B) The heptad motif and the helical wheel representations of the ROP (up) and GCN4 (down) proteins. In the case of ROP the residues that occupy the *a* position are colored yellow while the ones occupying the *d* position are colored red. In the case of GCN4 the heptad motif starts with the Met_2 residue occupying the *a* position. Both of these structures follow the "knobs-into-holes" packing mode as seen in the organization of their hydrophobic cores (*adad*). The images of the protein structures were captured with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311 and the PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC . The helical wheel representations and the sequences were adapted from [90,91].

1.2 Coiled-Coil Parameters and Geometric modeling

As previously mentioned, coiled-coils' backbone features can be described geometrically as parameters (Figure 2) [3,4,12,14]. These include the superhelical pitch, which characterizes the distance (Å) required for the superhelix to complete a full turn (Formula 1) in addition to the superhelical radius, which refers to the vector connecting the center of each individual helix and the center of the superhe-

lix (°) (Figure 2) [3–6]. An additional significant geometric element is the Crick's angle (or positional orientation angle) which describes the angle (°) between the α-helix and superhelix radius vectors and denotes the position of a residue with respect to the superhelix (Figure 2) [3–6]. Last but not least, the axial translation of the individual helices is calculated via the z -shift (Å) parameter or Δzoff (Formula 2) [12]. Other general geometric parameters include the superhelical frequency (ω_0), the helical frequency (ω_1), chain superhelical phase offset ($\Delta\phi_0$), and starting helical phase ($\Delta\phi_1$) (Figure 2) [5].

$$P = \pm \sqrt{\frac{d^2}{\left(N/n - \frac{1}{m}\right)^2 - 4\pi^2 \times r0^2}} \quad , \tag{1}$$

where d is the residue translation in an α-helix, N is the integer nearest to the number of, turns n residues make in a straight α-helix, r0 is the superhelical radius and m the residues per turn in a straight α-helix.

$$\Delta Z = \left(\frac{t2 - t1}{a2 - a1}\right) * \left(sina2 - sina1\right) \quad , \tag{2}$$

where t is the distance along the major helix and α the tilt angle.

The aforementioned parameters have been employed for the modeling of coiled-coils [5,6,12,14]. This type of protein structure prediction method is termed as *geometric modeling*, which utilizes our prior knowledge on coiled-coil geometry in order to *de novo* build protein Cα backbones [12,14]. There are multiple programs that use coiled-coil parameters either for protein structure prediction [12,14,15], motif identification and analysis [16,17], or the quantification of the properties of experimental structures [18]. With regard to geometric modeling, it proves highly useful when modeling sequences without an available reference structure or any known homology [14]. One important tool that applies geometric modeling on coiled-coil structures is ISAMBARD (Figure 3) [14]. This modular program is able to model any "parameterizable" protein fold and therefore offers the S*pecifications* module that contains geometric parameters [14]. All molecules (proteins and nucleic acids) are represented as AMPAL (Atom, Monomer, Polymer, Assembly, Ligand) objects in order to easier browse through the different organization levels of the structure [14]. Through the use of python's inheritance, the user is able to construct structures starting from either the *Polymer* or *Assembly* organization level [14]. The former describes how to arrange *Monomers* into a chain while the latter each *Polymer* with respect to each other [14]. In order to construct the protein models, ISAMBARD searches the parameter space of coiled-coils either exhaustively via grid scanning or by following a metaheuristic approach [14]. Once

the parameters have been generated the protein backbone models are constructed in addition to the modeling of the side chains with an external program [14,19]. Finally, the protein structures are assessed by the BUFF (Bristol University Docking Engine Force Field) which is an implementation of the BUDE (Bristol University Docking Engine) program [20].

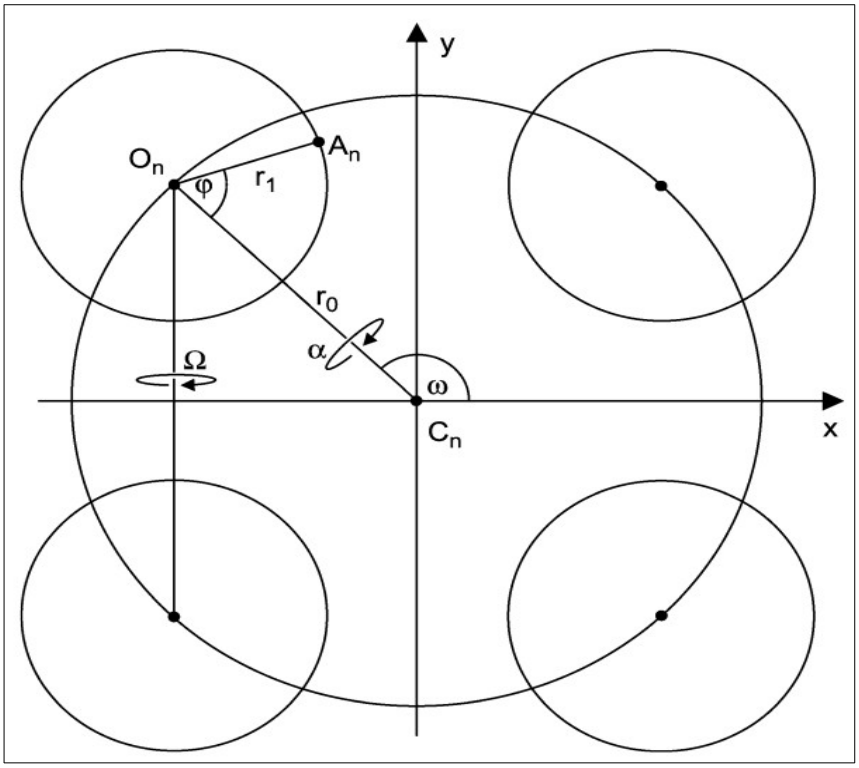


Figure 3| A schematic representation of the coiled-coil parameters at a tetrameric coiled coil. Points C_n and O_n represent the centers of the superhelix and an individual helix respectively while A_n a Cα carbon atom. Each vector represents a different parameter. Namely, vector r_0 represents the superhelical radius and r_1 the α-helix radius. The Crick's angle (φ) corresponds to the angle between the r_0 and r_1 vectors for the same residue. The pitch angle (α) represents the relative angle of a helix relative to the superhelix. The Ω parameter corresponds to the angle between two neighboring helices. Δφ is given by the angle between $r_{1,2}$ which are the radii vectors for two consecutive residues termed as the phase shift of the α-helix. The same parameter for the superhelix (for $r_{01,2}$) is termed as Δω.

The top model is selected and the process is continued until a specified number of iterations of this workflow has been reached [14]. The BUDE tool describes the atomic properties of all 20 standard amino acids and calculates the ligand binding energy (in BUFF between component chains) via the equation [14,15,20]:

$$Ecomplex = Esteric + Eelectrostatic + Edesolvation \quad , \tag{3}$$

calculated in kcal/mol. Where Esteric is the repulsion cause by overlapping atoms, Eelectrostatic is the resulting energy by charged-charged interactions and finally Edesolvation for each amino acid is an experimentally derived solvation energy value [20]. Ecomplex follows the thermodynamic rule where the lesser the value the more stable the structure.

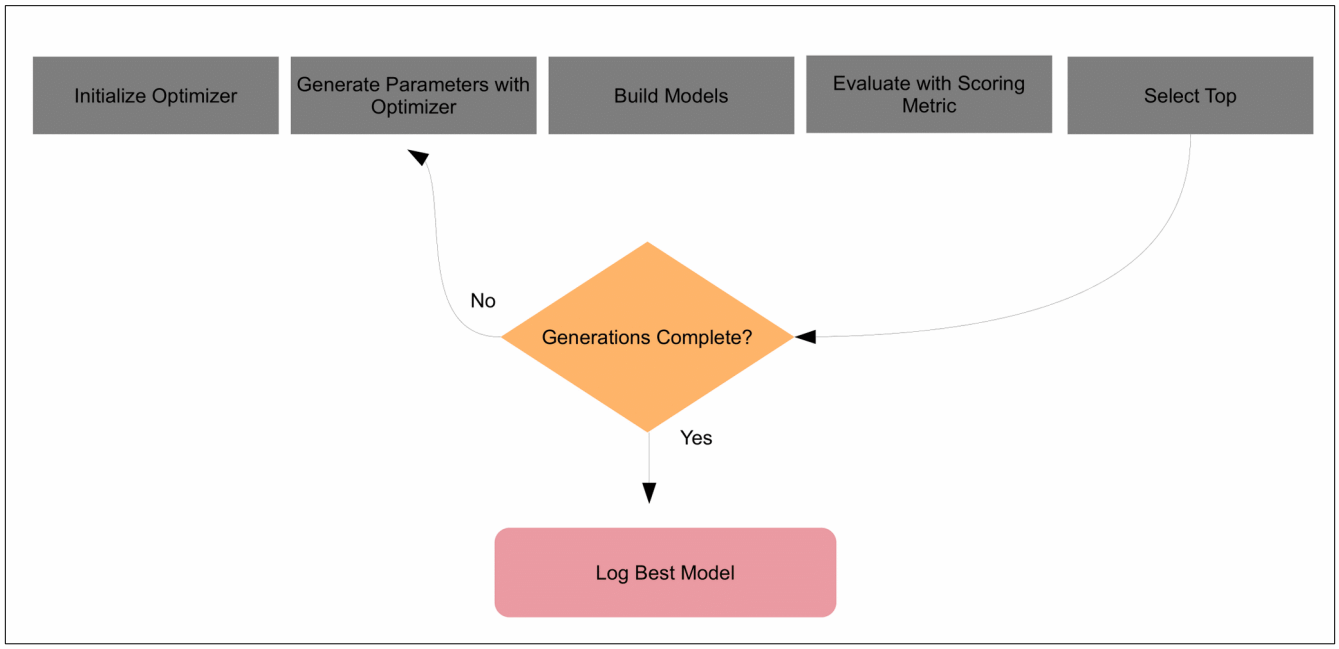


Figure 4| The general workflow of ISAMBARD. It starts with the initiation of the modeling process with coiled-coil specifications (parameters) offered by this module and a protein sequence for a specific oligomeric state. These parameters are passed on to an optimizer which is an algorithm responsible for the parameter optimization. This tool uses the GA (Genetic Algorithm), PSO (Particle Swarm Optimization), CMA -ES (Covariance Matrix Adaptation Evolution Strategy) and DE (Differential Evolution) algorithms. Next is the model building step. The Cα backbone is generated by the application of the coiled-coil parameters while the side chains are modeled using the SCWRL4 program. Then the models are assessed by using the BUFF (Bristol University Docking Engine Force Field) program which is integrated to ISAMBARD. The process is terminated when a specified number of generations has been achieved.

The construction of protein backbone conformations by the application of the aforementioned parameters [3,4,14,15] facilitates the exploration of the “Dark matter” of protein folding. This besides contributing to our future understanding of the conformational space in general [21], it could also find applications on synthetic biology and protein design [14,22].

1.3 Exploring the parameter space

In order to construct *de novo* coiled-coil models via geometric modeling, it is important to find the optimal values of the aforementioned parameters. *Optimization* techniques are divided into two categories; the *function* and *parameter* optimization [23]. In the former the optimum form of a function that describes an object is searched, while in the the latter the optimal values of a set of variables is explored [23]. This category includes methods such as *Metaheuristics* or *Grid scan* which search the parameter space differently [24,25]. Metaheuristic approaches will not necessarily guarantee to find the optimal solution and a grid scan searches the parameter space exhaustively even though only particular regions could be promising for the optimization problem [23–25].

1.3.1 Metaheuristics

Metaheuristics (similarly to the ones in the ISAMBARD tool) are algorithmic frameworks commonly derived from nature which aim to solve complicated optimization problems [24]. The DE (Differential Evolution) and GA (Genetic Algorithm) are two algorithmic variants which are inspired by Darwin's Evolution [24,26–28]. In general, the evolution-inspired optimization algorithms consist of *individuals* which are candidate solutions to the problem and a set of those comprise a *population [24,26,27]*. The process starts from a set of randomly produced individuals which are iteratively altered (*recombined* or *mutated*) and then assessed based on a *fitness function* (Table 1) [24,29]. The fittest individuals are more likely to be selected to breed the next generation [24,29,30]. The procedure is terminated when a specific termination condition has been met, most commonly when a specified number of generations has been reached [24,29,30]. These strategies are of high importance due to their easy parallelization where the calculation of the evaluation function for each individual is assigned to a different processor [31].

Besides the two frameworks mentioned above, ISAMBARD also offers the CMA -ES (Covariance Matrix Adaptation Evolution Strategy) and PSO (Particle Swarm Optimization) algorithms for parameter optimization [14]. The latter is a nature-inspired optimization algorithm that searches the parameter space by simulating the flight of birds in flocks (Table 2) [24]. In this case, instead of individuals there are *particles* that are placed in the search space of interest and move around based on the results of the objective function (in this case Formula 3) for their current location (where a location represents a specific set of parameters) [24]. Subsequently, each particle decides to move to a different location on the basis of their own results and of their fellow birds (particles) [24]. The iteration starts again when all the particles have moved [24]. This method mimics the collective attempt of a bird flock to find food.

In the case of parameter optimization the particle swarm searches the parameter space for the minima of the objective function (Formula 3) [24]. On the other hand, the CMA -ES algorithm is the result of the de-randomization of the Evolution Strategy (ES) with a *covariance matrix* (Table 3) [31,32] which gives the covariance between every pair of random elements.

Table 1| *The* structure of an Evolutionary Computation algorithm. Adapted from [24] and ISAMBARD docs.

Evolutionary Computation (EC) -Algorithm

P = Generate_Initial_Population()
while termination condition not met **do**
P' = Vary (P)
Evaluate (P')
P = Select (P'UP)
end while

Table 2| *The structure of the Particle Swarm Optimization algorithm. Adapted from [33] and ISAMBARD docs.*

Particle Swarm Optimization (PSO) -Algorithm
for each particle Initialize_particle **end** **Do** **for** each particle calculate_fitness_value **if** the fitness_value is better than the best_value (pBest) in the history (of the swarm) **set best value as pBest** **end** choose the particle with the best_value of all (gBest) **for** each particle calculate velocity update particle position **end** **while maximum iterations or minimum error criteria are not attained**

In this case individuals (elements) are adapted in each iteration in order to produce a new population [30,31,34] which directs the creation of new individuals towards the optimal solution [31,34]. There have been experiments suggesting the advantageous convergence properties of the aforementioned algorithm when compared to other ESs [31]. For instance, the particular algorithm has been shown to prevent the population from converging prematurely [35]. It is important to mention that when using a CMA -ES algorithm there is no need to apply large population sizes to avoid degeneration of the population (negative definite covariance matrix) [35]. The population sizes (number of individuals per iterations) are freely assigned, with small values mostly leading to faster convergence and large ones preventing reaching local optima [35].

Table 3| The structure of the Covariance Matrix Adaptation Evolution Strategy algorithm. Adapted from [34,35] and ISAMBARD docs.

Covariance Matrix Adaptation Evolution Strategy (CMA -ES)-Algorithm
set population_size (λ)
initialize state_parameters (m, σ, p_σ=0, p_c=0)*
while not terminate **do** #iterate
for i in {1...λ} **do** #sample and evaluate λ new solutions
χ_i = sample_multivariate_normal(mean =m, covariance_matrix = σ^2C)
f_i = fitness(χ_i)
$\chi i...\lambda \leftarrow \chi_{s(1)...s(\lambda)}$ with s(i)=argsort(f_1...λ,i) #sort solutions
update m, σ, p_σ, p_c, C
return m or χ_i

*where m, σ, p_σ, p_c, C are the distribution mean and best solution of the problem, step-size, two evolution paths and a positive definite covariance matrix respectively.

1.3.2 Grid scan

Instead of metaheuristics the search of the optimal parameters for a particular coiled-coil sequence can be performed via the *grid scan* approach [25,36]. This method entails the exhaustive search for a parameter set within specified value ranges [25]. In the case of ISAMBARD the optimal values of the superhelical radius and pitch, interface angle, register and z-shift of the provided coiled-coils sequences represent the parameter set under investigation (Table 4). Similarly, for model evaluation the BUFF tool is utilized. In a greater parameter space this exhaustive search might be inefficient [36] thus either the parallelization of the grid scan could be applied [36] or the use of one of the aforementioned metaheuristic approaches.

Table 4| The structure of a Grid scan. Adapted from ISAMBARD tutorials.

Grid scanning
Initialize parameters #assign value ranges
for all parameters
model = build_model(parameters)
energy = evaluate_model()
return model, energy

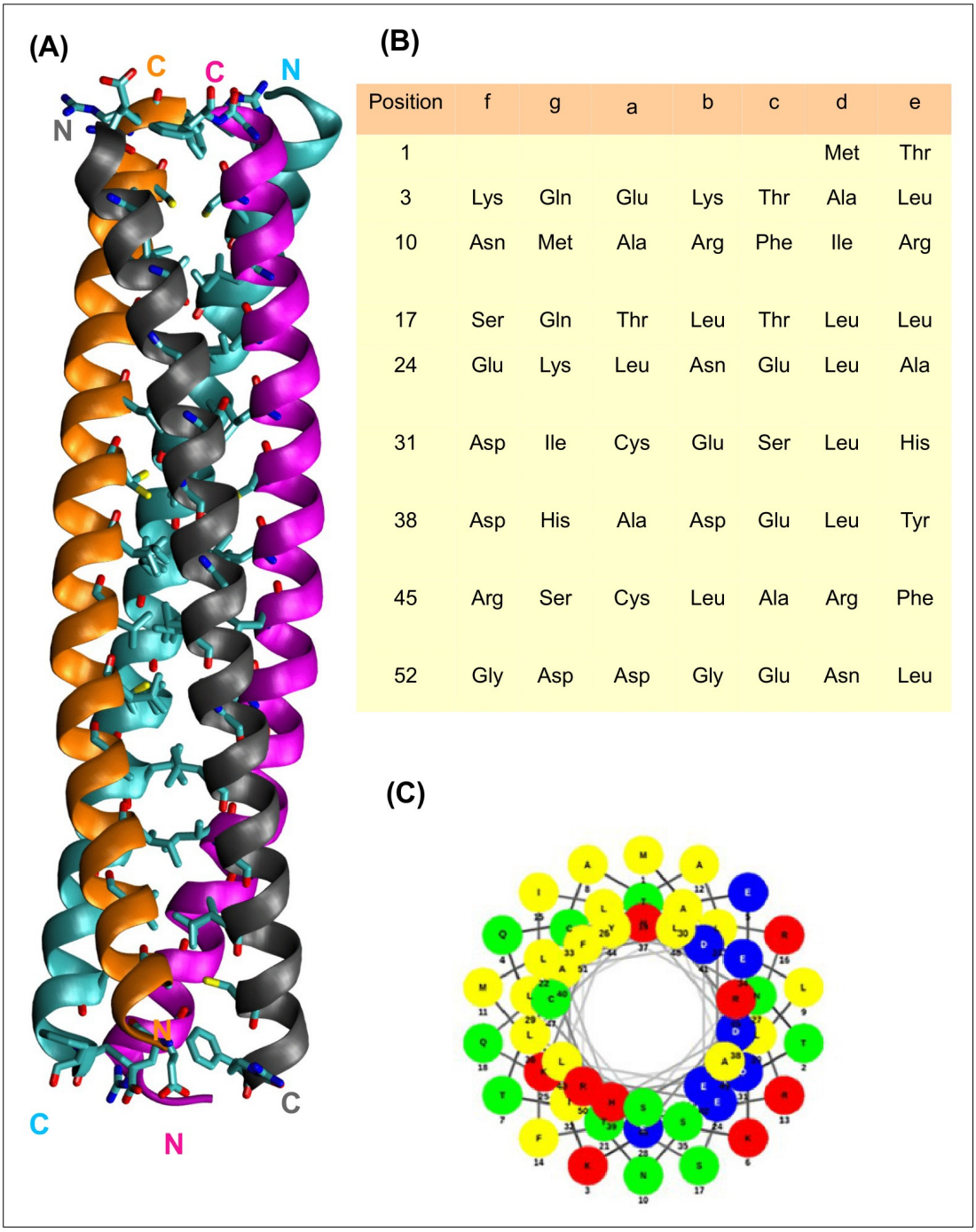
Chapter 2 The study of ROP and its variants

2.1 The Repressor of Primer protein model system

Helical Bundles in general represent a simple tertiary motif and thus a plethora of folding studies have been conducted to study them [37]. As previously mentioned, the dimeric ROP protein (PDB ID: 1ROP) constitutes a well studied representative of an anti-parallel four α-helix bundle with each monomer consisting of two anti-parallel α-helices connected by a short loop [10]. This protein is a highly regular and simple 4 α-Helix Bundle that exhibits a heptad motif only disrupted in the loop re-

gion [10,38]. Its role is to repress the rate of replication in *Escherichia coli* via increasing the association of RNA polymerase II to the complementary RNAI sequence of the primer precursor [39]. Regarding its scientific importance, this particular protein represents a model system for HBs due to its small size and solubility [10,40]. Additionally, the absence of a pro-peptide sequence, disulfide bonds, Proline residues and co-factors also contribute to its simplicity and thus easier study [40]. Notably, it also exhibits a well-defined hydrophobic core of eight hydrophobic layers comprised by the residues (in *adad* order): Ala_{45}-Leu_{41}-Thr_{19}-Ile_{15}, Ala_{12}-Leu_{22}-Cys_{38} -Leu_{48} , Cys_{52}-Gln_{34}-Leu_{26}-Ala_{8}, and Glu_{5}-Leu_{29}-Ala_{31}-Phe_{56} [10,41].

In order to investigate the sequence-structure relationship and folding processes of 4HBs, six ROP variants have been constructed [37]. In most cases, the effect of the introduced mutation on the loop region and the hydrophobic core was investigated and consequently the overall changes in the structure organization and stability [37,41,42]. One of the ROP variants is the $\Delta_{30\text{-}34}$ deletion mutant termed as RM6 (PDB ID: 1QX8) (Figure 5) [41,42]. In contrast to ROP, RM6 is a homo-tetrameric left anti-parallel -α -helix bundle with a five residue deletion (DADEQ) in the turn region which restores the heptad motif (Figure 5) [41]. This variant each monomer of the wild-type ROP is transformed into a continuous α-helix (absence of loop) and results in a highly thermostable protein [41]. This variant exhibits a reorganized hydrophobic core with completely different layer composition [41]. This protein contains seven symmetric layers in contrast to the asymmetric ones of ROP and consist of the residues: Leu_{26} -Leu_{29}-Leu_{26}-$Leu_{29,}$ Cys_{33}-Leu_{22}-Cys_{33}-Leu_{22}, Thr_{19}-Leu_{36}-Thr_{19}-Leu_{36}, Ala_{40}-Ile_{15}-Ala_{40}-Ile_{15}, Ala_{12}-Leu_{43} -Ala_{12}-Leu_{43}, Cys_{47}-Ala_{8}-Cys_{47}-Ala_{8} and Glu_{5}-Phe_{51}-Arg_{50}-Arg_{50} [41]. This symmetry is disrupted however in the last layer which is also accessible to water molecules, thus exhibiting high residue mobility [41]. An additional ROP variant with an uninterrupted heptad repeat has been constructed and it was the result of a two residue insertion [43]. However, in this case the mutant protein does not exhibit drastic changes in its structural properties [37,43]. An additional interesting result supporting the “flexible” nature of ROP is of the Ala31Pro variant with one substitution in the loop region affecting the overall organization of the protein [44]. This variant is right-handed, mixed (parallel and anti-parallel) 4HB with a distinguishable U -like topology [44]. These results support the extreme plasticity of the ROP protein since this particular sequence has folded into a diverse group of hydrophobic cores among its variants [37].



Position	f	g	a	b	c	d	e
1						Met	Thr
3	Lys	Gln	Glu	Lys	Thr	Ala	Leu
10	Asn	Met	Ala	Arg	Phe	Ile	Arg
17	Ser	Gln	Thr	Leu	Thr	Leu	Leu
24	Glu	Lys	Leu	Asn	Glu	Leu	Ala
31	Asp	Ile	Cys	Glu	Ser	Leu	His
38	Asp	His	Ala	Asp	Glu	Leu	Tyr
45	Arg	Ser	Cys	Leu	Ala	Arg	Phe
52	Gly	Asp	Asp	Gly	Glu	Asn	Leu

Figure 5| The structure of the RM6 variant. (A) The cartoon representation of the RM6 variant. The residues that participate in the formation of the hydrophobic core are displayed as sticks and face the interior of the protein. (B) The sequence of the RM6 variant. In this table the position and the corresponding position in the heptad repeat are depicted. (C) The helical wheel representation of the RM6 variant. The structure was displayed with the VMD tool and the helical wheel was produced by NETWHEELS.

2.2 Structural Determinants of Folding and Stability in HBs

Such α-helical bundles demonstrate characteristic determinants in their folding process besides the ones observed in all protein types [45]. In general, both the amino acid sequence of a protein and the path that it follows in order to adopt a native fold encode its structure [46,47]. A term that describes this ability of a protein, to acquire its native and unique three-dimensional structure, is termed as conformational specificity [48]. In the case of the ROP model system (and HBs generally) it can be divided into four levels [48]. First the correct oligomeric state contributes to a right protein fold followed by the correct specification of the HB's topology [48]. Another contributing factor is the relative spatial orientation of the individual secondary structure elements and last but not least the correct and favorable packing of the hydrophobic core [48]. All these describe the structural characteristics of the intrinsic tendency of proteins to adopt their native conformation. In addition, there is a thermodynamic aspect of the conformational specificity which refers to the necessity for a large energy gap between the native structure and the molten globule state (non-native folded state) [48]. On that note, homodimeric 4HBs such as ROP display six possible topologies [48]. Four out of six exhibit clockwise or counterclockwise -turning loops which connect adjacent helices on the same or opposite side of the bundle [48]. The rest topological options are of the bisecting U motif mentioned above [44,48]. Thus, the design of HBs requires the stabilization of only one of these possible conformations and the destabilization of the rest [48].

There are particular driving forces of great significance during protein folding processes which are non-covalent interactions i.e. the hydrophobic effect, hydrogen bonds, Coulombic and van der Waals forces [47,49]. There have been proposed 4HB-specific contributors to protein folding which include the interhelical turns and the helical dipoles that are present in these structures [50]. The former in the case of ROP has been extensively investigated via mutagenesis and structural studies [37,41,42,44]. The latter refers to the impact of the two distinct ends of protein structures on protein folding [50] which conventionally are a free amino (N -terminal end) and carboxyl group (C -terminal end). These helical dipoles are formed due to the intrinsic directionality of the peptide bond, that is also expressed in the polypeptide chain [50]. The side chains of α-helical structures always point towards the N -terminal end and potential divergence of this rule could affect local hydrophobic core geometry [51–53]. Therefore the peptide bonds are characterized as *non-palindromic* and potential alterations, namely with a *retro modification* (sequence reversal) could affect the connectivity between chiral residues and thus the packing of the hydrophobic core [51,54]. Nevertheless, it is still unclear how the backbone direction affects the folding process [54,55]. Computational experiments of retro-modification on three

sequences prone to form α-helical structures were conducted and suggested that the formation of α-helices was not affected by altering the backbone directionality [51].

2.3 Retro-Proteins

In order to investigate the importance of backbone directionality in protein folding, experiments with *retro-proteins*, which are molecules with reversed amino acid sequences compared to their parents, have been conducted [52,54–59]. The inversion of a protein sequence produces a new polypeptide chain that does not exhibit any homology with its parent and thus its foldability is unknown [51–57]. However, there are multiple factors to consider when hypothesizing how these proteins fold.

First, the retro-polypeptide chain is not equivalent to a randomly generated one since it shares some characteristics with its folded native parent [51]. Those include the physicochemical properties of the parent sequence (polar/non-polar patterns) and its hydrophobicity profile in addition to potentially being prone to form similar secondary structure elements [52,55,58]. Nevertheless, there have been inconsistent results regarding the conservation of the secondary structures in the retro-isomers [56]. These have shown that sequence inversion affects secondary structure propensities [56] and dipole interactions which lead to overall structure instability [52,60]. Furthermore, geometrically restrictive elements present in the native protein could contribute to the foldability of their retro-isomers [61]. This was examined in experiments conducted using retro-isomers from the alpha domain of human metallothionein-2 (MT) [61]. It was concluded that the presence of the metal-tetrathiolate nucleus was potentially positively affecting the foldability of the retro-protein [61]. Even though the retro-protein was foldable, the inversion of the backbone direction had an impact on structure formation [61]. Since, hydrophobic collapse is a significant folding factor [46,47,62,63] the extend in which the hydrophobic core is affected upon sequence-reversal also potentially determines the fold of the retro-isomer [51]. The results of the latter study also concluded that the foldability of retro-proteins depends on the size and structural characteristics of its parent and the flexibility of its hydrophobic core [51]. Lastly, theoretically if the protein sequence is regarded as a “string of beads” then upon sequence-reversal the protein fold should remain unaffected [58]. In support of this side-chains only notion, primary results with lattice model computer simulations indicated that the retro-beta-domain of staphylococcal protein A could formed a stable and foldable structure [64]. Contradictory subsequent experimental results from Circular Dichroism (CD) and NMR spectroscopes however, supported that this protein was disordered [58]. Retro-studies on the SH3 domain of a-spectrin, the B1 domain from staphylococcal G protein, and rubredoxin resulted in unfoldable proteins as well [58,65].

On the other hand, the inversion of the protein sequence alters the chirality of the amino acids thus creating a D-protein [59,61]. Commonly, the D-isomer of a protein is expected to adopt a mirror image of its L counterpart [59,61,66]. However, the mirror retro-all-L structure would only be a topological equivalent of the native (non-inversed)-all-D protein since the pairs C=O and N-H would be interchanged [67]. The first part of the term retro-all-L refers to the inverted direction of the polypeptide chain while the second half to the chirality of the amino acids. However, primary computational studies opposed this notion [58,68].

In general, experimental results from retro-studies have been contradictory. They range from retro-proteins with similar but less stable folds compared to their parents [64,69], unfolded [55,58] to retro-proteins with stable structures of high similarity to the native [57]. In the latter case, a characteristic example is the retro-GCN4 protein. This isomer adopts a stable three-dimensional structure with an RMSD value of 0.37 Å when compared to the native [57]. However, retro-GCN4 forms a very stable, parallel, four-helix bundle similarly to the GCN4-pLI mutant, in contrast to the two-stranded native GCN4 [57]. Notably, this structure exhibits a heptad repeat in accordance to the geometric principles of coiled-coils and represents the only resolved retro-α-helical bundle [57]. Another interesting case of a retro-study is the GroES co-chaperonin of *Escherichia coli* [52]. This protein, even though similar to its parent, does not maintain the characteristic interactions with the native sub-unit; GroEL [52]. It also acquires the ability of being unaffected by heat unlike its parent [52]. However, it did not exhibit the same oligomeric state as the native protein, which forms heptamers whereas the retro-GroEs formed either trimeric or pentameric structures [52].

Regarding the ROP paradigm and its variants, there are no resolved retro-structures available to better investigate the effect of sequence reversal on HBs. However, there have been recent experiments on the rROP and rRM6 retro-polypeptides in comparison to their parent proteins that have elucidated some of their structural properties [53]. This study suggested that rRM6 exhibits high similarity to its parent (RM6, PDB ID: 1QX8) on a secondary structure and oligomerization level, in addition to being slightly less compact to RM6 [53]. On the other hand, the rROP protein displayed an unclear oligomerization state and a disordered, molten-globule state [53]. The necessity for further computational studies on these retro-isomers has been expressed and in this case *de novo* modeling represents a highly useful tool [48,70,71] particularly on the aforementioned isomers [53].

2.4 Structural studies on α-Helical Bundles (HBs) and their retro-isomers

The *de novo* protein design of the HB structures could elucidate multiple still unanswered questions regarding the protein folding mechanism [48,72]. Structural and mutational analysis of HB proteins has contributed to our understanding of the determinants that govern the formation of native folds [37,41,44,48,72]. The prediction of a protein's tertiary structure starts with an amino acid sequence which is used to generate a model [71]. There are two categories of protein structure prediction methods; *ab initio* and template based [71]. The latter techniques employ an already available structure as a template which is chosen after sequence alignment [73]. The *ab initio* or *de novo* protein structure prediction techniques are applied when there is no available template for the protein of interest and only the primary sequence is used to build a model from scratch [70,74]. In the case of retro-proteins where there is no available homologous sequence [53] *de novo* methods are highly useful. The model system of ROP and its (retro-) variants represent regular and representative HBs. Thus, the use of geometric modeling in the study of backbone directionality could provide useful insights for the folding of HBs in general. More specifically, via the CCBuilder and ISAMBARD tools, the retro-studies on α-helical structures are feasible [14,15]. These programs have been used in the past for the modeling of coiled-coils [22,75,76]. In more detail, ISAMBARD has been utilized in order to model homotetrameric all-parallel coiled-coils for the investigation of their core geometry [75] and the generation of coiled-coil dimers used for model refinement [76]. In addition, it was utilized for the rational design of α-Helical Barrels via mutagenesis of the *g* position in the heptad repeat [22]. Whereas CCBuilder, has been employed in multiple occasions for the modeling and design of different proteins [77–80]. Notably, this software has been proposed as a potential tool for modeling coiled-coils for Molecular Replacement (MR) calculations [81]. MR calculations use an available three-dimensional model (search model) which is placed into the crystal lattice in order to determine the crystal structure of a protein [81,82]. On that note, MR attempts on crystallographic data obtained from the rRM6 crystals have failed to allow a complete structure determination, when using the RM6 as a search model [83]. This implies potential differences between the retro-isomer and its parent [83]. Thus, the application of geometric modeling on the rRM6 protein could also assist future determination of the crystal structure of an additional retro-isomer [81,83].

2.5 Scope of the study

So far, the effect of backbone directionality on protein folding, especially in the case of HBs, remains unknown. Thus, *de novo* protein structure prediction of the rRM6 could contribute to our current knowl-

edge on this matter [48,52,54,56,67,67–71]. This could assist the future determination of the rRM6 structure [81,83] which in addition to the already available retro-GCN4, may contribute to our understanding of how retro-proteins fold [57]. Taking all of the above into account, the aims of this study were to:

1. Assess ISAMBARD as a tool for geometric modeling.
2. Generate and assess models of the rRM6 protein using the aforementioned software.

The RM6 protein exhibits a highly regular and simple topology even compared to ROP (no loops or turns) and its modeling could potentially contribute to our understanding of HB folding in general. Here, we hypothesize that the modeled retro-isomer adopts a stable protein structure of high similarity to its parent. We believe that the presence of the geometric restrictive element of the heptad repeat is a contributing factor to the conservation of the oligomeric state, shape, stability and secondary structure propensities of the retro-isomer. It is expected that the model with the lowest energy generated by ISAMBARD would be of left-anti-parallel topology similarly to RM6. This would suggest that the retro-variant of the RM6 mutant remains unaffected upon sequence reversal. Results from this study, could further support that geometric modeling is a significant tool for the construction of coiled-coil structures and could be employed in special cases such as the modeling of a retro-HB.

Chapter 3 Methods

3.1 Heptad repeat and sequences used for RM6

The native amino acid sequence of RM6 was obtained from the Protein Data Bank (PDB) (Table 5). The retro-sequence was reversed by using a simple python script that inverted the sequences of interest (Script 1). At the beginning of the rRM6 sequence a Met was added in order to match the sequence that was used for the retro-experiments on RM6 [53,83]. The heptad motifs (*a/d* residues) of both the native and the retro-protein were found via the DeepCoil2 Bioinformatics toolkit through accessing its web server [84]. This tool uses neural networks to identify coiled-coil domains in protein sequences [84].

Table 5| The native and retro sequences used for the modeling of RM6, ROP and their retro-variants.

Native sequence	Retro-sequence
MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHD-HADELYRSCLARFGDDGENL	MLNEGDDGFRALCSRYLEDAHDHLSECIDALEN-LKELLTLTQSRIFRAMNLATKEQKTM

3.2 Mutagenesis using Pymol and graphics

The initial computational investigation of the sequence reversal on RM6 entailed the mutation (reversal) of the side chains on the backbone of the native protein. This was achieved via the *Mutagenesis* tool of the Pymol program [85]. In addition, the helical wheel representation of the proteins were generated via the NetWheels and EMBOSS:pepewheel web servers [86,87]. The structures were displayed in Pymol, Chimera or VMD [85,88,89]. The graphs were generated via the Xmgrace plotting software or via Python scripts [90].

3.3 Geometric modeling using ISAMBARD

The ISAMBARD program described in detail was used for the geometric modeling of the proteins (see Introduction) [14]. Both the Grid scan and Metaheuristic methods were assessed based on their ability to model the RM6 protein. The metaheuristic algorithms used were the GA, DE, PSO and CMA -ES. Again their performance on modeling the RM6 HB was assessed. The ISAMBARD modeling scripts are provided at the *Scripts* section at the end.

3.3.1 Grid scan

Script 2 displays an example of a grid scan script used in this particular study. The python modules Numpy and Itertools were imported in addition to the modeling and specifications packages which are used for the construction and description of the coiled-coil proteins. Python's inheritance is applied in order to initialize the coiled coil parameters superhelix radius and pitch, Crick's angle (phiCα angle) and the z-shift in addition to the oligomeric state of the bundle and the helix length (number of amino acids). Also the superhelix orientation was added to the parameters investigated. Furthermore, a python dictionary with ideal phiCα values was provided in order to assist the search of optimal values for this parameter. Then via numpy's *arange* method, value ranges for the parameters of interest were assigned in addition to the preferred searching step. A nested for loop was applied in order to obtain all possible parameter combinations which were later used to construct the protein backbone. The SCWRL4 program is integrated to ISAMBARD which predicts side chain conformations [19]. In order to minimize the running time and since each model in the grid is independently calculated, a parallelization approach was applied. The ranges for each parameter are shown in Table 6. Twelve scripts were generated in twelve different directories and each contained a different range value for the radius parameter. This way all possible combinations were generated and assessed with the BUFF program [14,15,20]. All possible models, each having a different starting register (*abcdefg*) were generated by using an individual function for model building.

Table 6: The overall parameters and ranges tested using the grid scan approach.

Parameter	Min	Max	Step
Radius	6.0	8.21	0.2
Pitch	50	350.1	10
Interface angle	-30	30.1	2
z-shift	-10	10	1

3.3.2 Metaheuristics

All the available metaheuristic frameworks offered by ISAMBARD [14,24,30] were utilized and assessed in this study. Different number of generations and individuals were applied in order to evaluate their performance on modeling the RM6 protein. The type of parameters used in the evolutionary optimizers are either *STATIC* or *DYNAMIC*. The latter requires a provided mean value and a value range in order assign and search the parameter space while the former only the static value. For instance the oligomeric state of the RM6 (or rRM6) represents a static value, while radius, pitch, phiCα and z-shift parameters dynamic (Table 7). The ideal phiCα angles were calculated after the equation:

$$Delta = n * delta - \left(\frac{delta}{2}\right) \quad , \tag{4}$$

where delta = 360/7 and n in the range (1,8)

Table 7| The overall parameters and ranges tested using the metaheuristics approach.

Parameter	Mean value	Value range
Radius	7.0	2.0
Pitch	200.0	150.0
PhiCα angle	ideal_phica	27
Z-shift	0.0	20.0

In addition to the parameters mentioned in Table 7, when using the optimizers, models with all possible super helical twists and orientations were constructed. Thus, the generated models were left all-parallel, left anti-parallel, right-all parallel and right anti-parallel. In order to identify the optimal sequence to be used for the modeling of the proteins and test the sensitivity of geometric modeling using ISAMBARD, different variations of the length of the resolved RM6 structure were used (Table 8-9). Furthermore, different z-shift values (Table 10) were applied on the whole rRM6 sequence in order to recreate all possible hydrophobic cores for rRM6 obtained by DeepCoil2.

Table 8| All the sequences used for the modeling of RM6 to test the sensitivity of the method.

Helix length	Sequence
58	MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLAR-FGDDGENL
55	MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLAR-FGDDG
54	MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLAR-FGDD
53	MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLAR-FGD
52	MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLARFG

Table 9| All the sequences used for the modeling of rRM6 to find the optimal one.

Helix length	Sequence
59	MLNEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKE-QKTM
58	LNEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
57	NEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
56	EGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
55	GDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
54	DDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
53	DGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
52	GFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM
51	FRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM

Table 10| The z-shift values for rRM6.

Z-shift	Orientation-Twist	Value range
-18	left/right-anti-parallel	0
3	left/right-anti-parallel	0
-5	left/right-anti-parallel	0
8	left/right-all-parallel	0

3.4 Model refinement and evaluation

3.4.1 Galaxy

The Galaxy web server and especially the RefineComplex tool was used to refine the inter-helical contacts in the HBs and the orientation of the helices [91]. The identification of the interface residues (8Å Cα-Cα distance) marks the initial step of the web server's workflow in addition to the characterization of the complex's symmetry only for homo-structures [91]. Next, cycles of energy minimization and re-

laxation are carried out via Molecular Dynamics (MD) simulations after calculating equations of physics-, knowledge- and restraint-based terms [91]. The refined models are generated after a distance restraints protocol (protocol 1) and a both distance and position restraints one (protocol 2) [91]. Models 1-5 are constructed after protocol 1 while models 6-10 after protocol 2 [91].

3.4.2 MM-align

The MM-align algorithm was utilized to calculate the RMSD values of the generated models [92]. Its web server performs sequence independent alignment of protein structures using a modified version of the common Needlman-Wunsch dynamic programming algorithm [92]. Notably, it applies a weighting factor on the interacting side chains when aligned [92].

Chapter 4 Results

As mentioned before, geometric modeling and specifically the ISAMBARD program was assessed by modeling the RM6 protein structure. These models were refined using GALAXY and compared to the crystallographic structure (PDB ID: 1QX8) using MM-align [91,92]. At first, the grid scan and then all the available metaheuristic methods were examined. The sensitivity and accuracy of the latter method was investigated by intentionally utilizing wrong parameters and sequences for the modeling of RM6. After that the modeling of the retro-isomer was performed. In order to accomplish that, different workflows were performed which entailed the deletion of residues from the retro-sequence and the z-shift manipulation.

4.1 Hydrophobic core of RM6

In this step, the DeepCoil2 web server was used in order to identify and compare its results with the hydrophobic core of the resolved structure [41,84]. In Figure 6, the probabilities to occupy an *a/d* position of each residue of the native RM6 are displayed. It was found that the coiled-coil region of the input sequence was between residues 10 to 45. These results excluded the Cys_{47}-Ala_{8}-Cys_{47}-Ala_{8} and Glu_{5}-Phe_{51}-Arg_{50}-Arg_{50} layers which are part of the hydrophobic core. In addition, results indicated that three residues (Ala_{30}, His_{37} and Tyr_{44}) which occupy *e* positions in the native sequence exhibited high *a*-position probabilities (0.325, 0.297 and 0.240) (Figure 6). This step was carried out in order to assess the accuracy of DeepCoil2 on identifying the hydrophobic layers of the RM6 structure and thus utilizing it on the retro-RM6.

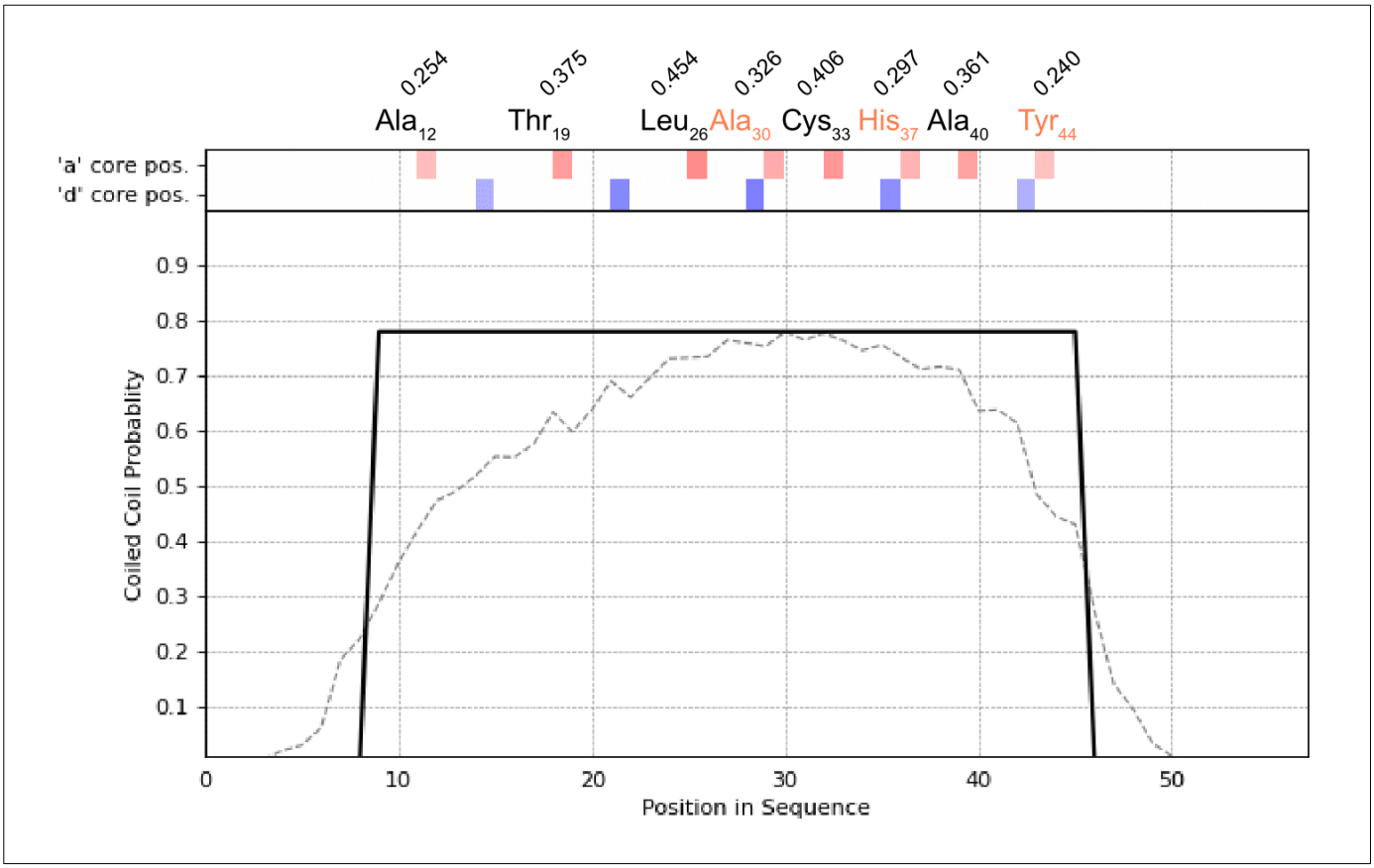


Figure 6| The a/d residues of the RM6 protein sequence, retrieved from the DeepCoil2 web server. The residues occupying *a* positions are colored red while the ones occupying *d* blue. The probabilities for the *a* residues are marked above the three letter code of each residue. Residues Ala_{30}, His_{37} and Tyr_{44} in the protein sequence occupy an *e* position, however, they exhibited relatively high probabilities for an *a* position in the heptad repeat.

4.2 Grid scan

The initial modeling steps included the calibration and assessment of the grid scan method on the RM6 protein sequence. The models generated were only of left handedness and both all-parallel and anti-parallel orientation. The parameters investigated were the superhelix radius, superhelix pitch, the interface angle, register and z-shift. The results included 33,264 models with unique parameter values and starting register. Due to the long running time (several days) and the comparably high energy values (data not shown) the grid scan method was excluded from the study and thus the metaheuristic frameworks were calibrated and assessed.

4.3 Metaheuristics

Initial trial runs using all the optimizers (GA, DE, PSO, CMA -ES) were carried out where 10 number of generations and 200 number of individuals were applied (Supplementary 1). In this step, the whole sequence of the RM6 protein was used and the parameters investigated were the helix length, superhelix radius and pitch, phiCα angle and z-shift. Only left anti-parallel models were generated aiming at identifying the most suitable optimizer to use when modeling this protein. The results included energy values and the corresponding parameters used for model generation for each heptad position. Results from the optimizers also contain the standard deviations calculated for the models in each generation, the running time and the number of models generated (Supplementary 1). An additional step included the calibration of the optimal number of generations and number of individuals. The population sizes investigated were 200, 500 and 1000 while the number of generations were 10 and 15.

4.3.1 Optimizer selection via modeling the RM6 protein

Results from all the optimizers indicated that all the algorithms constructed models that reached a minimal energy value when starting with a residue occupying the *d* position in the heptad repeat (Figure 7). Interestingly the CMA -ES's *d* model exhibited the lowest minimum energy (-4068.658 kcal/mol) compared to all the models generated by all the metaheuristic frameworks. This particular algorithm generated and assessed 2200 models for each heptad position and the parameter values of the best model are shown in Table 11 (Supplementary 1). In the cases of the DE and GA algorithms, even though the *d* models exhibited the lowest energy values, the energies of all the *abcdefg* models were approximate (Supplementary 1). Notably, the *b* and *d* models of the latter algorithm demonstrated minimum values of -3900 kcal/mol and -3917 kcal/mol respectively (Supplementary 1). This remark was supported by the standard deviations of the optimizers as well which indicated that the final generations consisted of energetically similar individuals (DE and GA algorithms, Supplementary 1). In the case of PSO, the *d* model reached the minimum energy value of -3972.967 kcal/mol, however, this optimizer did not search and evaluate the same number of individuals as the CMA -ES algorithm. The former optimizer searched 1797, 1892, 1704, 1903, 1752, 1576 and 1624 models for the *a-b-c-d-e-f-g* starting registers respectively (Supplementary 1). Both PSO and CMA -ES algorithms generated the *d* left anti-parallel models in total time of 23 minutes. This suggested that the CMA -ES algorithm for the same amount of time, searched more extensively the parameter space and found the optimal solution to the optimization problem. Thus, for the next modeling steps, only the CMA -ES framework was used to model the RM6 and rRM6 proteins.

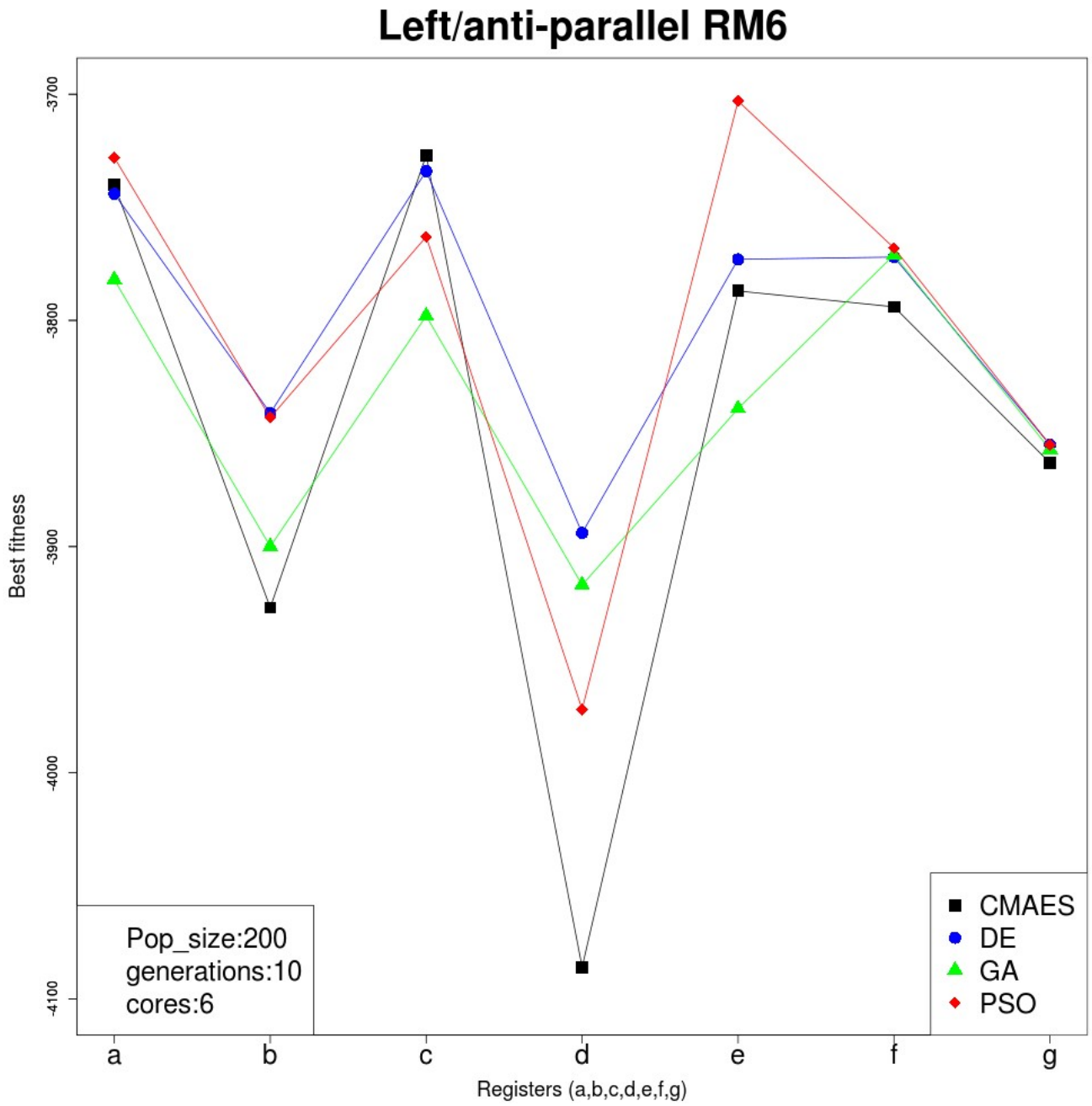


Figure 7| The results from the initial runs of the optimizers. The labeling of each optimizer is displayed on the figure. The models generated were of left helical twist and anti-parallel orientation. The number of generations applied were 10 and the population size (number of individuals) was 200. All the algorithms produced models which reached minimal energy when starting with the *d* residue. The *d* CMA -ES model exhibited the lowest energy value of all the models.

Table 11: The best parameters fitted for the RM6 protein when using the CMA -ES algorithm.*

Parameter	Value
Radius	6.58003914300648
Pitch	201.74795126373593
PhiCα angle	322.8023176842898
Z-shift	1.933370217717626

*(number of generations 10, number of individuals)

4.3.2 Selecting population size and number of generations via RM6 modeling

In addition to selecting the ideal optimizer, the suitable population sizes and the number of generations for the CMA -ES algorithm were investigated and identified. In this set of experiments models with right handedness and parallel orientation were also introduced. This was performed in order to assess the sensitivity of the process in its ability to identify the correct orientation and superhelix twist. The fitness of the individuals was again calculated by the BUDE force field and resulted in four potential topologies (left/all-parallel, left/anti-parallel, right/parallel, right/anti-parallel) for each register (a, b, c, d, e, f, g) (Figure 8). Results from the increase of the number of generations while keeping the population size at 200, suggested that the *d* was the starting register exhibiting the lowest energy value in the left/all-parallel, left/anti-parallel and right/anti-parallel models. On the other hand, the *g* right/all-parallel model displayed the lowest fitness score for the same number of generations. However, the *d* left/anti-parallel model demonstrated the lowest energy value (-4086 kcal/mol) compared to all (Figure 8.B). The alteration of the number of generations did not improve the energy of the *d* left anti-parallel structure (Figure 8). As for the models with the rest starting registers, they either did not exhibit a drastic change on their energy values (Figure 8.A) or were not affected at all (*e,f*) (Figure 8.A). Hence, overall the increase in the number of generation for the same population size, did not improve the optimization process thus for the rest of the modeling experiments it was kept constant at 10. The results from the alteration of the number of generations when using the CMA -ES algorithm on the whole RM6 sequence are displayed on Figure 9. The running times for the values 200, 500 and 1000 of the population size were 23 minutes, 59 minutes and ~2 hours respectively. In all cases, the *d* left anti-parallel models exhibited the lowest energy values which were -4068 kcal/mol, -4069 kcal/mol and -4074 kcal/mol respectively. However, the increase in the population size to 500 and 1000 resulted in models of lower energy values for the *abcefg* registers (Figure 9). A bigger population size leads to the generation and assessment of more individuals per iteration. The selection was based on the fact that the general geometric parameters of the retro-RM6 are unknown, thus a rigorous and efficient search of the parameter space is required. Taking into account that large population sizes prevent reaching local

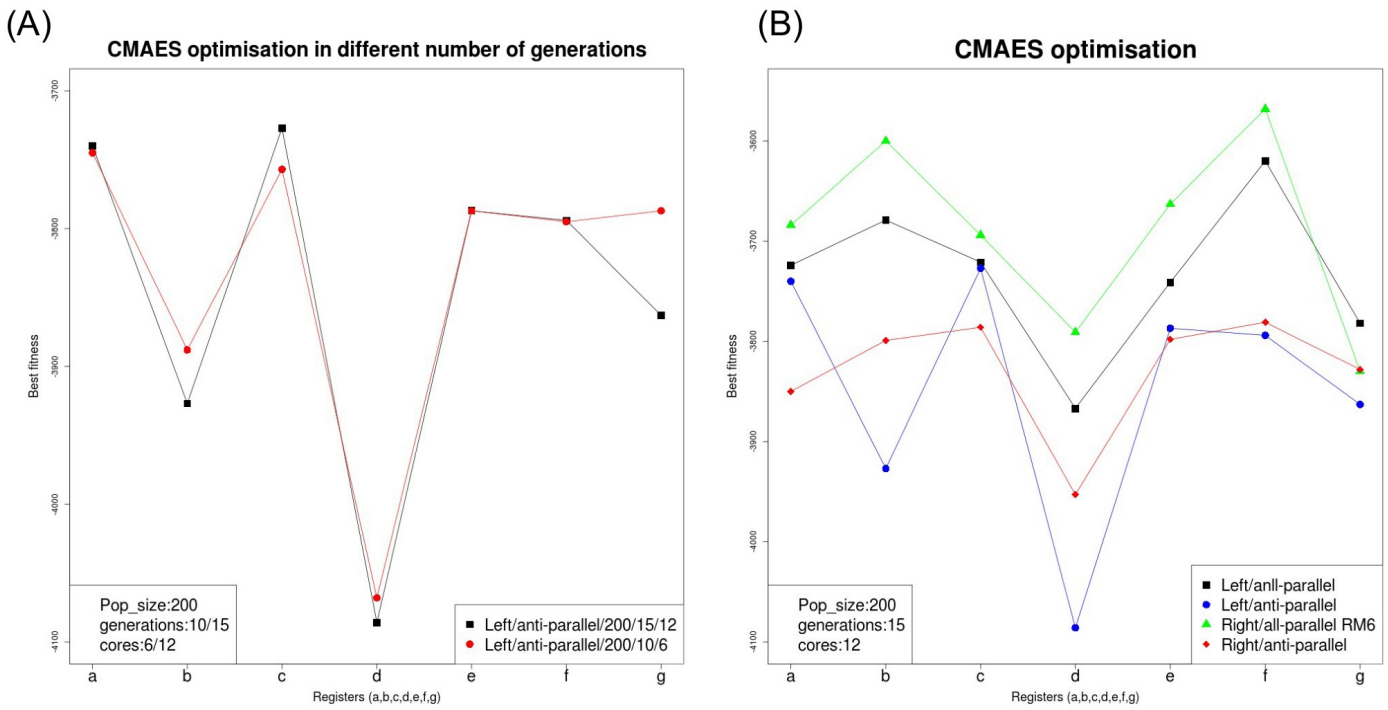


Figure 8| The calibration of the number of generation using the CMA -ES algorithm. The labeling of the two runs is displayed on the figure. The cores utilized for the experiments were also increased to 12. This was performed in order to reduce the running time. (A) The left/anti-parallel models generated by CMA -ES. The population size was kept constant at 200. The results suggested that the increase of the aforementioned variable did not drastically affect the left/anti-parallel generated models of the RM6 protein. (B) The left/all-/anti- parallel and right/all-/anti-parallel models produced by CMA -ES with population size kept constant at 200 and the number of generations at 15. The results indicated that the model with the lowest energy at all cases was the one with a *d* starting register, besides the right/all-parallel model which was *g*.

optima [34] and the aforementioned results, the 500 value was selected. The CMA -ES optimizer managed to search the parameter space sufficiently, in a relatively short amount of time and produced models of low energy values in addition to finding the optimal parameters (Met_1 -*d* left/anti-parallel). Hence, for the rest of the modeling procedures the population size was assigned to 500. The models generated for a population size equaled to 500 and number of iterations equaled to 10 are depicted on Figure 10. The models left all-parallel, left anti-parallel and right anti-parallel exhibited minimum energy values -3855 kcal/mol, -4069 kcal/mol and -3948 kcal/mol for *d* starting register residue whereas the right all-parallel model for *g* starting register residue -3815 (Figure 10). The best model (left/anti-parallel) was constructed utilizing the parameter values in Table 12.

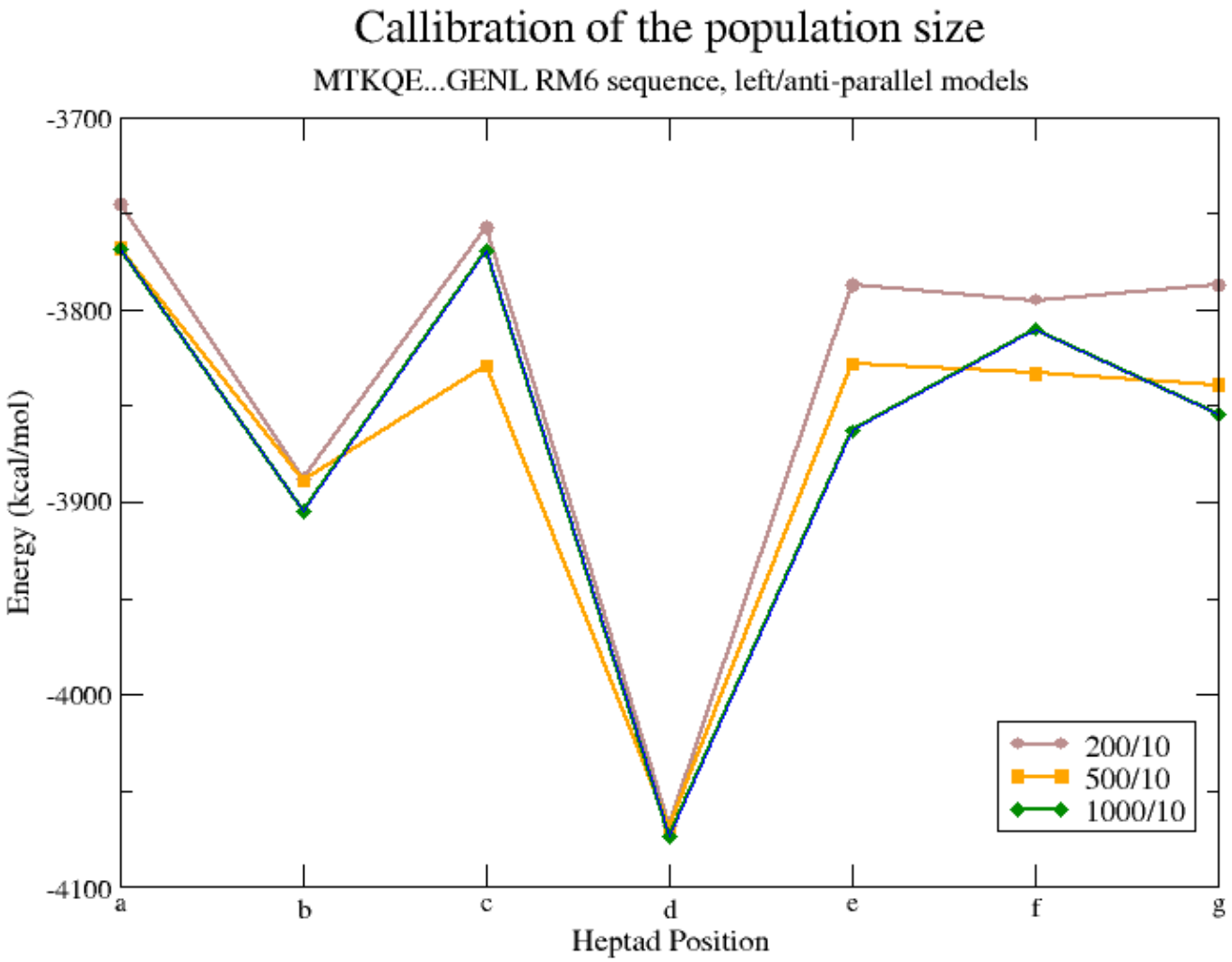


Figure 9| The calibration of the population size using the CMA -ES algorithm. The labeling of the different runs are displayed on the graph. The population sizes applied were 200, 500 and 1000. Only the left/anti-parallel models are included in this plot. The whole sequence of the RM6 protein was used on this modeling step. All the runs produces *d* models exhibiting the lowest energy value compared to the rest.

Table 12: The best parameters fitted for the RM6 protein (left/anti-parallel model) when using the CMA -ES algorithm

*(number of generations 10, number of individuals 500)

Parameter	Value
Radius	6.57324992797141
Pitch	199.22124109702992
PhiCα angle	323.1378318669839
Z-shift	1.865982531023413

Figure 10| The models of RM6 generated using the CMA -ES algorithm by applying population size 500 and number of generations (iterations) equal to 10. The parameters investigated were the superhelix radius, superhelix pitch, phiCα angle and z-shift. In this case models with all possible helix twist and orientation were generated resulting in left/all-parallel, left/anti-parallel, right/all-parallel and right/anti-parallel models. Their labeling is depicted on the graph. The whole sequence of RM6 was used. The model exhibiting the lowest energy was the *d* left/anti-parallel which corresponds to the native structure of the RM6. Thus, the algorithm efficiently generated and distinguished the native fold by identifying the appropriate geometric parameters.

At this point, the hydrophobic layers of the modeled left anti-parallel RM6 were identified and compared to the ones of the resolved protein structure (Figure 11, Table 13). This was achieved by inspecting the generated model in the graphics (see Methods). The results indicated that the hydrophobic layers of the modeled RM6 differed from the ones observed in the resolved structure. More specifically, its hydrophobic core was out-of-register (Figure 11, Table13). Even though they shared the same amino acid content, the hydrophobic layers in the modeled RM6 were translocated which resulted in the formation of the faulty core. The RMSD value of the superposition of the modeled RM6

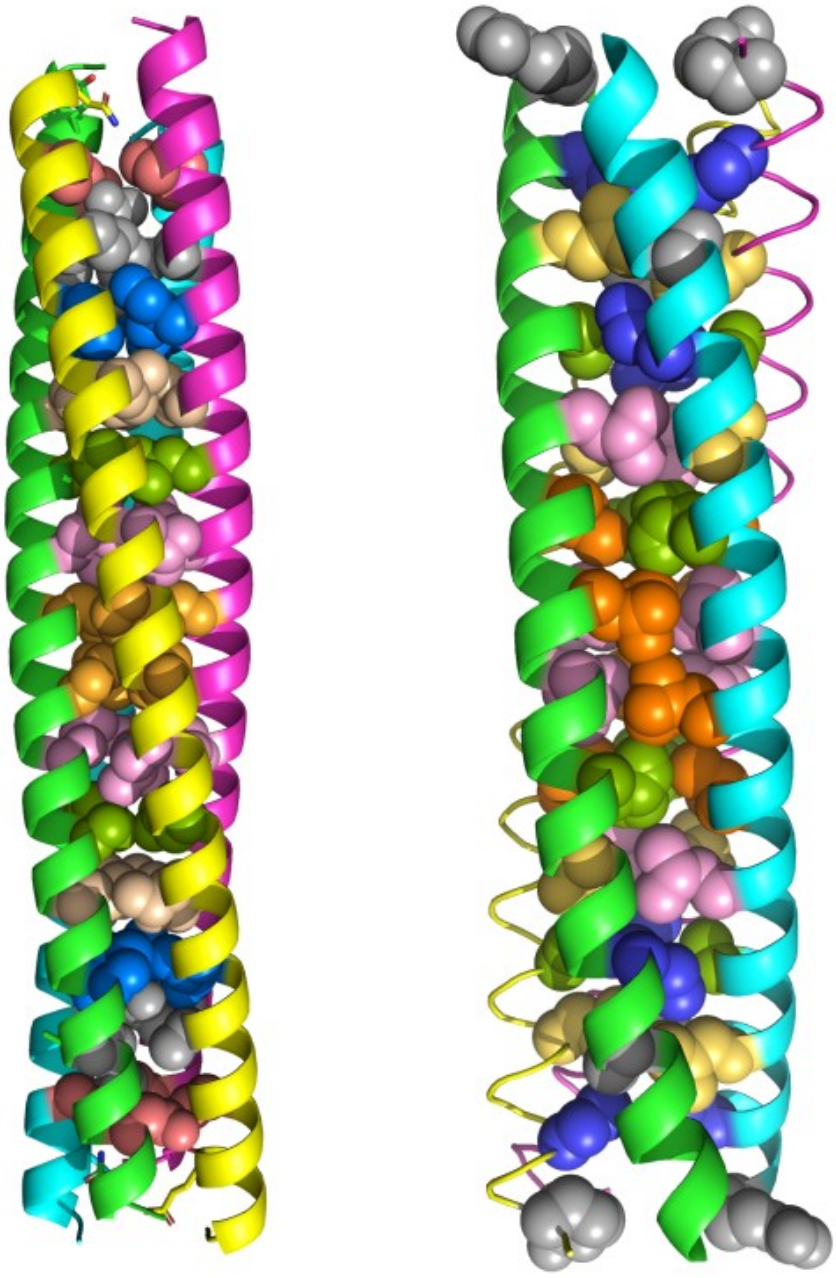


Figure 11| On the left, the modeled RM6 with the residues forming the hydrophobic layers displayed as spheres. Starting from the inside of the structure the Cys33 -Leu29 -Cys33 -Leu29 layer is colored orange, the Leu26 -Leu36 -Leu26 -Leu36 pink, the Ala40 -Leu22 -Ala40 -Leu22 green, the Thr19 -Leu43 -Thr19 -Leu43 light yellow, the Cys47 -Ile15 -Cys47 -Ile15 blue, the Ala12 -Arg50 -Ala12 -Arg50 gray and the Asp54-Ala8-Asp54-Ala8 light red. Chains A, B, C and D are colored green, cyan, magenta and yellow respectively. On the right, the corresponding residual positions are presented in the native RM6 crystal structure, using the same color coding for both the residues and the chains.

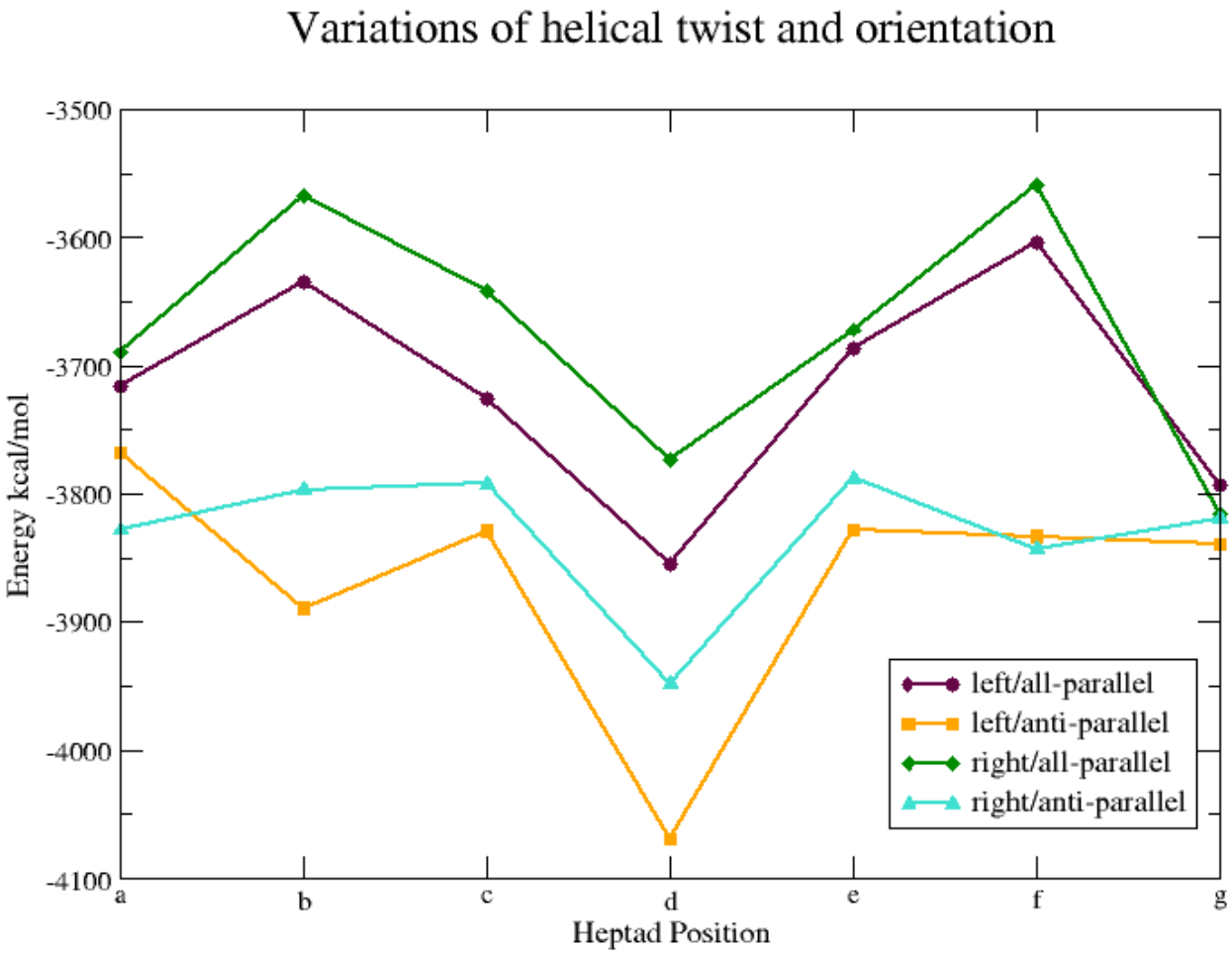
(*d* left/anti-parallel, 500/10, MT.GENL) compared to the resolved RM6 (PDB ID: 1QX8) was 1.71 Å. However, this value refers only to Cα atoms while also exhibiting the wrong alignment. Hence, a potentially correct superposition would result in much higher RMSD values.In addition to identifying the appropriate population size and number of iterations, the ability of ISAMBARD to distinguish between right and wrong coiled-coil parameters was also examined. On that note, models of right superhelix twist and all-parallel orientation were constructed using the population sizes and number of iterations mentioned above via modeling the whole RM6 sequence (MT...GENL) (Figure 12). The results indicated that the model with the lowest energy value was the *d* left anti-parallel. All the models achieved minimum energy when starting with a residue occupying a *d* position besides the *g* right all-parallel one.

Table 13: The hydrophobic layers of the modeled (left) and resolved (right) RM6

Modeled RM6	Resolved RM6
Cys_{33} -Leu_{29} -Cys_{33} -Leu_{29}	Leu_{26} -Leu_{29} -Leu_{26} -Leu_{29}
Leu_{26} -Leu_{36} -Leu_{26} -Leu_{36}	Cys_{33} -Leu_{22} -Cys_{33} -Leu_{22}
Ala_{40} -Leu_{22} -Ala_{40} -Leu_{22}	Thr_{19} -Leu_{36} -Thr_{19} -Leu_{36}
Thr_{19} -Leu_{43} -Thr_{19} -Leu_{43}	Ala_{40} -Ile_{15} -Ala_{40} -Ile_{15}
Cys_{47} -Ile_{15} -Cys_{47} -Ile_{15}	Ala_{12} -Leu_{43} -Ala_{12} -Leu_{43}
Ala_{12} -Arg_{50} -Ala_{12} -Arg_{50}	Cys_{47} -Ala_{8} -Cys_{47} -Ala_{8}
Asp_{54}-Ala_{8}-Asp_{54}-Ala_{8}	Glu_{5} -Phe_{51} -Arg_{50} -Arg_{50}

4.3.3 Sequence derivatives and hydrophobic cores of the RM6

Besides the input parameters, the sequence used to perform geometric modeling (using ISAMBARD) plays a significant role. On that note, various sequence derivatives from the RM6 sequence were used to model the protein (see Methods) aiming at assessing the sensitivity of this method and generating the correct in-register hydrophobic core. The appropriate sequence used to for RM6 was the MT...GDD whereas the one used for the all-parallel was the MT...GENL (see Methods). The selection was based on the heptad repeat of the RM6 in order to construct the *a-d-a-d* core geometry in the left anti-parallel structure and the *a-a-a-a/d-d-d-d* in the left all-parallel. For the modeling of the anti-parallel structures the sequences examined are shown in Table 8. Results from this set of experiments indicated that the best, energy-wise, model was the one constructed using the MT...GENL sequence (Figure 13). In addition to assessing their energy values, the hydrophobic cores of the generated structures were examined using graphics (Table 13,14). The energy values of the *d* left anti-parallel models were -4069 kcal/mol, -4048 kcal/mol, -3964 kcal/mol, -3858 kcal/mol and -3783 kcal/mol for the



MT...GENL, MT...GDDG, MT...GDD, MT...GD and MT...G sequences and their hydrophobic cores are shown

Figure 12| Models with different helical twists and orientations. In order to validate the accuracy of ISAMBARD, models with faulty input parameter were constructed and the ability of the aforementioned program to distinguish the right ones (known from the resolved structure) was investigated. For this purpose the energy values of the left/all-parallel, left/anti-parallel, right/all-parallel and right/anti-parallel models were compared. The model exhibiting the lowest energy value was the left/anti-parallel one which is in accordance with the structural characteristics of the resolved RM6. The labeling of the different models is displayed on the figure. The graph was generated using the XmGrace plotting software.

below (Table 14). At this step their RMSD values were calculated before and after refinement (Table 15) and their sequence alignment was assessed (Figure 14). The resulting models were refined using the Galaxy web server [91] and the RMSD values of the best, minimized models were calculated again with MM-align [92]. All the models generated exhibited low RMSD values after refinement (<1.3 Å). It was observed that the longer the sequence used for modeling, the lower the energy value before and after refinement. This was also the case for the RMSD values after refinement.

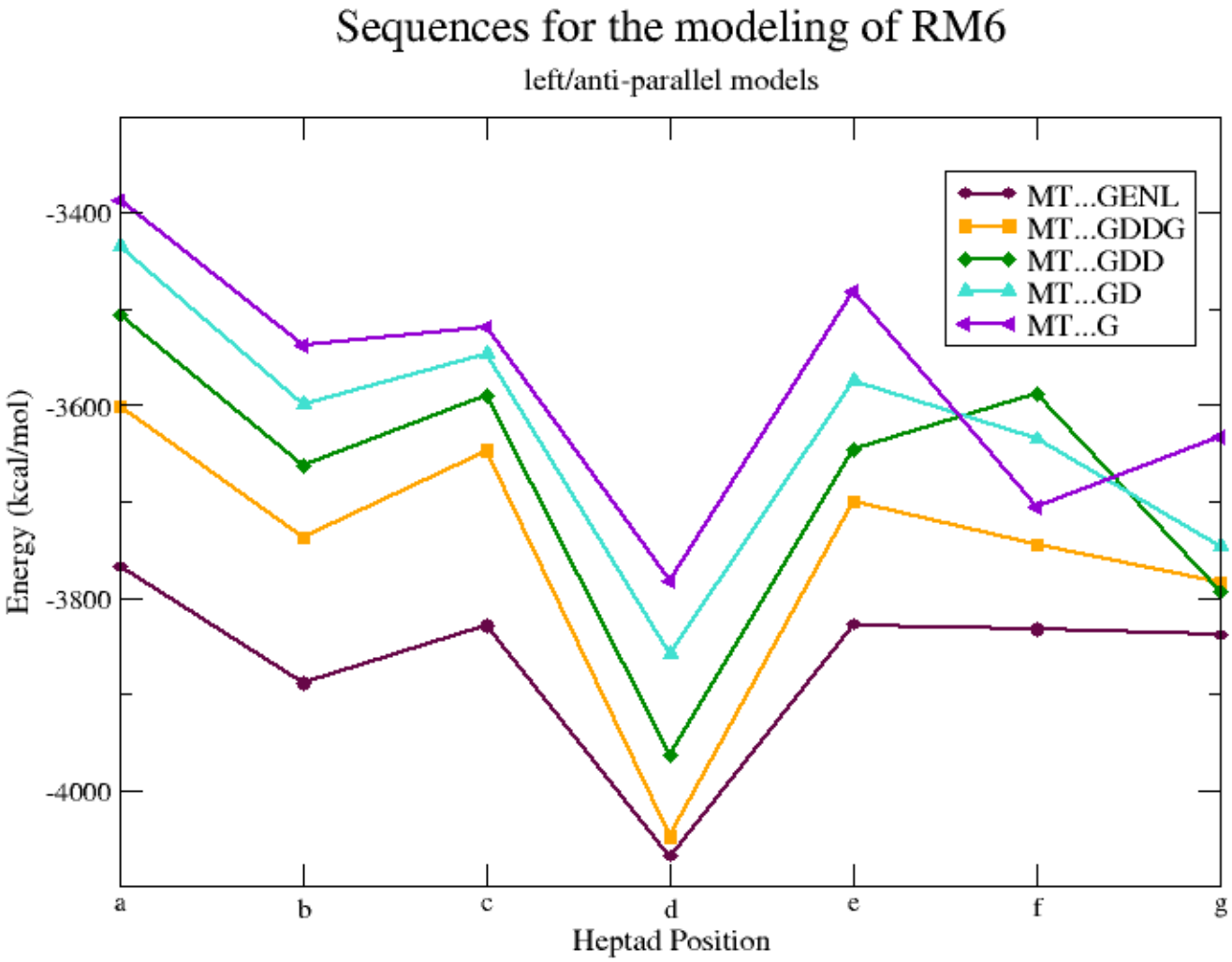


Figure 13| The sequences used for the modeling of the left/anti-parallel RM6.The labeling of the different sequences is displayed on the Figure. In all cases the *d* models were exhibiting the lowest negative value. The energetically best model (-4048 kcal/mol) was the one constructed using the MT...GENL sequence (Table 8). The *d* model that was created the sequence recreating the hydrophobic core of RM6 (MT...GDD) exhibited an energy value of -3964 kcal/mol.

Table 14: The hydrophobic layers of the left/ anti-parallel models generated with the various RM6 sequence derivatives

MT...GENL	MT...GDDG	MT...GDD	MT...GD	MT...G
Asp_{54}-Ala_{8} -Asp_{54}-Ala_{8}	Leu_{26} -Leu_{29} -Leu_{26} -Leu_{29}	Leu_{26} -Leu_{29} -Leu_{26} -Leu_{29}	Leu_{26} -Leu_{29} -Leu_{26} -Leu_{29}	Leu_{26} -Leu_{29} -Leu_{26} -Leu_{29}
Ala_{12} -Arg_{50} -Ala_{12} -Arg_{50}	Cys_{33} -Leu_{22} -Cys_{33} -Leu_{22}	Cys_{33} -Leu_{22} -Cys_{33} -Leu_{22}	Cys_{33} -Leu_{22} -Cys_{33} -Leu_{22}	Cys_{33} -Leu_{22} -Cys_{33} -Leu_{22}
Cys_{47} -Ile_{15} -Cys_{47} -Ile_{15}	Thr_{19} -Leu_{36} -Thr_{19} -Leu_{36}	Thr_{19} -Leu_{36} -Thr_{19} -Leu_{36}	Thr_{19} -Leu_{36} -Thr_{19} -Leu_{36}	Thr_{19} -Leu_{36} -Thr_{19} -Leu_{36}
Thr_{19} -Leu_{43} -Thr_{19} -Leu_{43}	Ala_{40} -Ile_{15} -Ala_{40} -Ile_{15}	Ala_{40} -Ile_{15} -Ala_{40} -Ile_{15}	Ala_{40} -Ile_{15} -Ala_{40} -Ile_{15}	Ala_{40} -Ile_{15} -Ala_{40} -Ile_{15}

Ala_{40} -Leu_{22} -Cys_{47} -Ile_{15}	Ala_{12} -Leu_{43} -Ala_{12} -Leu_{43}	Ala_{12} -Leu_{43} -Ala_{12} -Leu_{43}	Ala_{12} -Leu_{43} -Ala_{12} -Leu_{43}	Ala_{12} -Leu_{43} -Ala_{12} -Leu_{43}
Leu_{26} -Leu_{36} -Leu_{26} -Leu_{36}	Cys_{47} -Ala_{8} -Cys_{47} -Ala_{8}	Cys_{47} -Ala_{8} -Cys_{47} -Ala_{8}	Cys_{47} -Ala_{8} -Cys_{47} -Ala_{8}	Cys_{47} -Ala_{8} -Cys_{47} -Ala_{8}
Cys_{33} -Leu_{29} -Cys_{33} -Leu_{29}	Glu_{5} -Phe_{51} -Arg_{50} -Arg_{50}	Glu_{5} -Phe_{51} -Arg_{50} -Arg_{50}	Glu_{5} -Phe_{51} -Arg_{50} -Arg_{50}	Glu_{5} -Phe_{51} -Arg_{50} -Arg_{50}

Table 15: RMSD and energy values before and after refinement of d left/anti-parallel models for the various sequences

Sequence (*d* left/anti parallel models)	Energy (kcal/mol)	Correct hydrophobic core	RMSD (Å) /196 residues	Galaxy Energy (kcal/mol)	Model with lowest RMSD value	RMSD after refinement (Å)/ 196 residues
MT...GENL	-4069	N	1.71	-12,192	3	1.27
MT…GDDG	-4048	Y	1.89	-11,544	2	1.17
MT….GDD	-3964	Y	1.76	-11,427	6	1.16
MT...GD	-3858	Y	1.84	-11,177	5	1.13
MT...G	-3783	Y	1.53	-11,018	10	1.17

Regarding the all-parallel structures the hydrophobic core (*aaaa/dddd*) was constructed using the MT...GENL sequence and the hydrophobic layers of the left all-parallel models were (in *aaaa/dddd* order, starting from the N-terminal to the C-terminal): $Glue_{5}$ -Glu_{5} -Glu_{5} -Glu_{5}, Ala_{8} -Ala_{8} -Ala_{8} -Ala_{8}, Ala_{12} -Ala_{12} -Ala_{12} -Ala_{12}, Ile_{15} -Ile_{15} -Ile_{15} -Ile_{15}, Thr_{19} -Thr_{19} -Thr_{19} -Thr_{19}, Leu_{22} -Leu_{22} -Leu_{22} -Leu_{22}, Leu_{26} -Leu_{26} -Leu_{26} -Leu_{26}, Leu_{29} -Leu_{29} -Leu_{29} -Leu_{29}, Cys_{33} -Cys_{33} -Cys_{33} -Cys_{33}, Leu_{36} -Leu_{36} -Leu_{36} -Leu_{36}, Ala_{40} Ala_{40} -Ala_{40} -Ala_{40}, Leu_{43} -Leu_{43} -Leu_{43} -Leu_{43}, Cys_{47} -Cys_{47} -Cys_{47} -Cys_{47}, Arg_{50} -Arg_{50} -Arg_{50} -Arg_{50}, Asp_{54} -Asp_{54} -Asp_{54} -Asp_{54} and Asn_{57} -Asn_{57} -Asn_{57} -Asn_{57}.

4.3.4 Modeling the retro-RM6

After modeling the RM6 protein in order to troubleshoot the ISAMBARD program, models of its retro-isomer were generated following a similar workflow. At first, the a/d residues of the retro-polypeptide chain were identified, using the DeepCoil2 program (Figure 15). These results were utilized in order to narrow down the potential hydrophobic cores of the retro-isomer and assist the modeling process (Table 16). The Ile_{28} and Lys_{35} residues exhibited a propensity to occupy an *a* position at the heptad repeat, however, their probability values were lower compared to the rest and hence not considered.

Since the heptad motif of the retro-RM6 was predicted by the DeepCoil2, potential hydrophobic cores of the retro-isomer were constructed either by deleting residues from the sequence (similarly to the

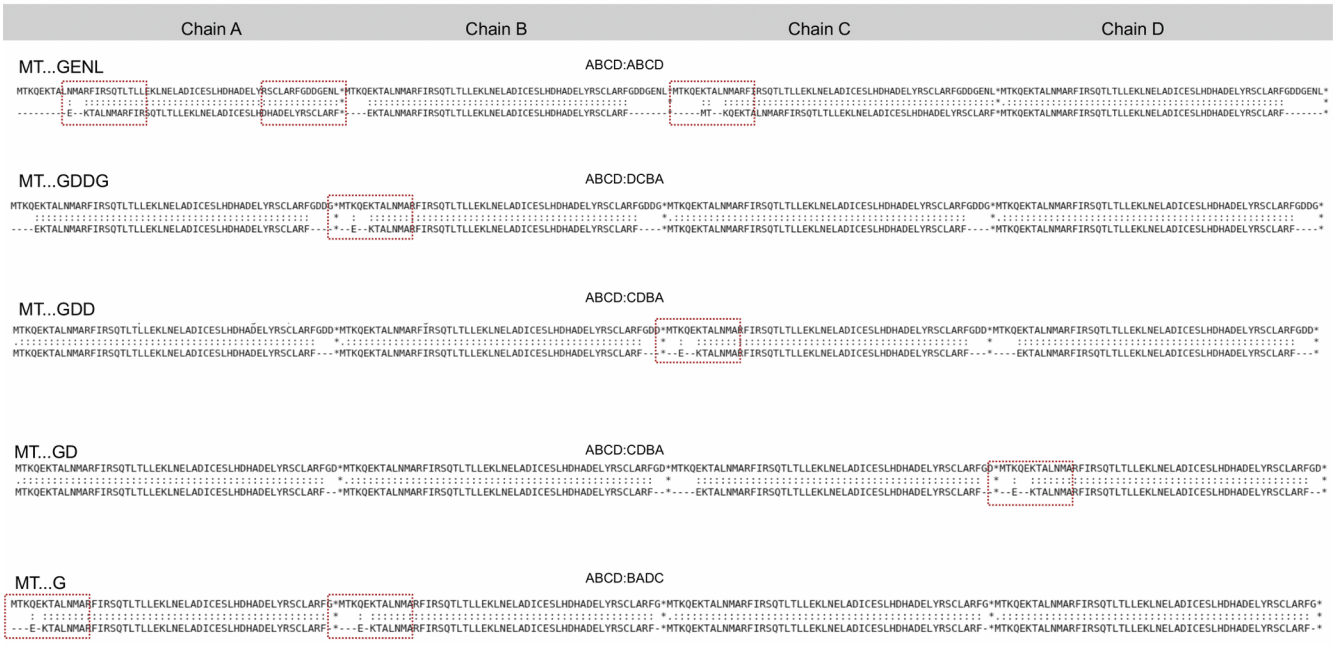


Figure 14| The sequence alignment between the best models (generated by the sequence derivatives) and RM6. The joining of the helices is displayed above each sequence alignment. The first part (ABCD) which is the same in all alignments corresponds to the RM6 protein. The seconds part corresponds to the superimposed chains of each model. Potentially interesting parts of the alignment or mismatches are highlighted with the red boxes.

RM6, see Table 9) or by keeping the register of the starting residue fixed to *f* (phiCα = 180.0°). In the latter workflow, the appropriate z-shift value was searched in order to manipulate the translocation of the helices and thus recreate the probable hydrophobic cores. No more than 5-10 residues were excluded from each helix when searching for the z-shift values to construct the potential hydrophobic cores.

Table 16: The heptad motif of the retro-RM6 as defined by DeepCoil2

	a	b	c	d	e	f	g
						Met	Leu
3	Asn	Glu	Gly	Asp	Asp	Gly	Phe
10	Arg	Ala	Leu	Cys	Srer	Arg	Tyr
17	Leu	Glu	Asp	Ala	His	Asp	His
24	Leu	Ser	Glu	Cys	Ile	Asp	Ala
31	Leu	Glu	Asn	Leu	Lys	Glu	Leu
38	Leu	Thr	Leu	Thr	Gln	Ser	Arg
45	Ile	Phe	Arg	Ala	Met	Asn	Leu
52	Ala	Thr	Lys	Glu	Gln	Lys	Thr
59	Met						

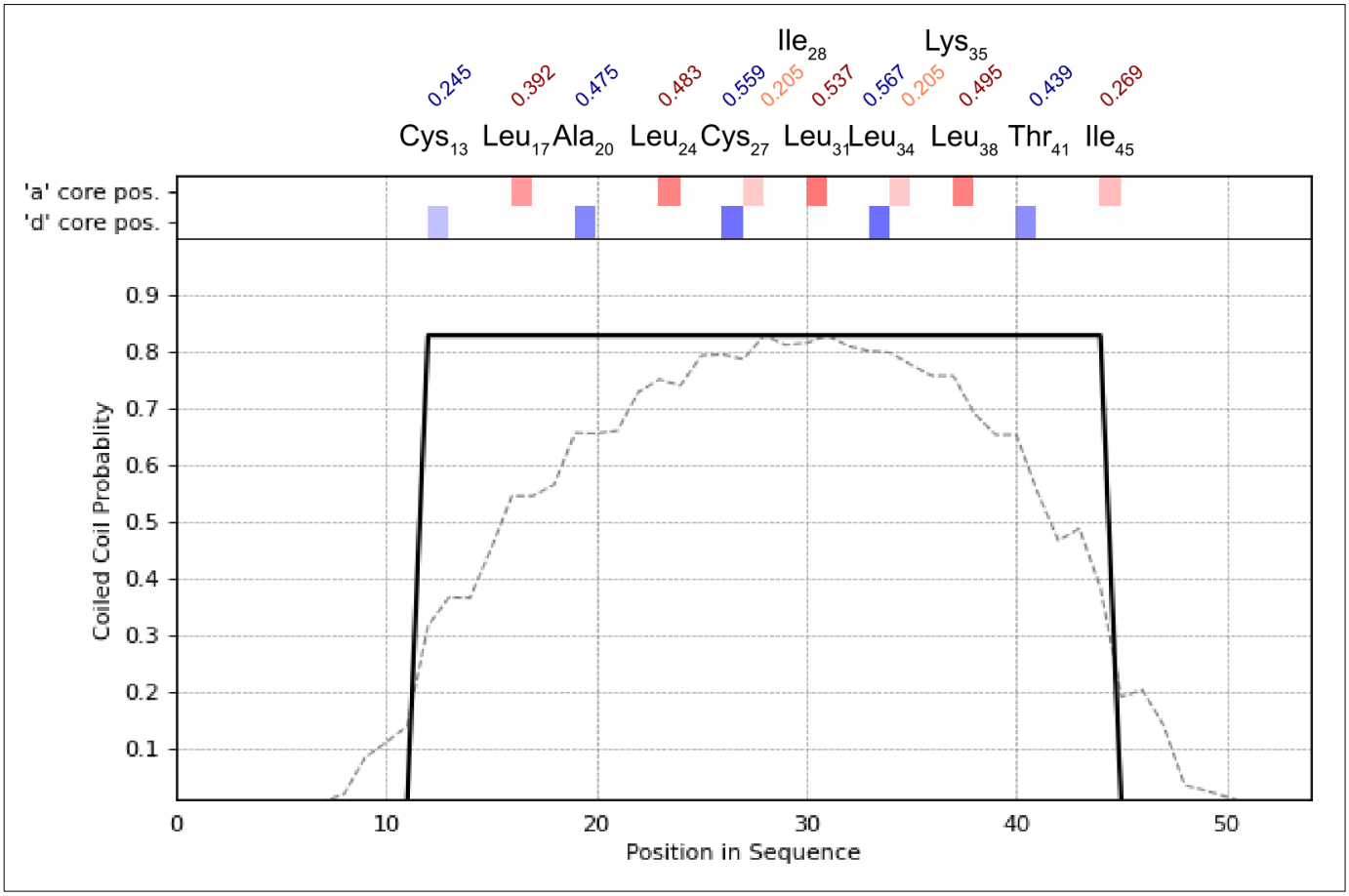


Figure 15| The *a/d* positions of the retro-RM6 as predicted by the DeepCoil2 program. The *a* positions of the heptad repeat are colored dark red, the *d* blue and the potential *a* of low probability scores light red. The heptad motif starts at residue Cys_{13} with a probability score of 0.245 and terminates at the Ile_{45} residue with a probability score of 0.249. The Ile_{28} and Lys_{35} residues exhibited a probability score which indicated that they may occupy an *a* position at the heptad repeat.

This was performed in order to recreate the hydrophobic layers of rRM6 without deleting an extended region of the helices, thus modeling potentially more stable HBs. Results from the sequence derivatives consisted of energy values for every starting register in each sequence (Figure 16) and the models generated were refined using again the GalaxyRefineComplex webserver (Table 17) [91].

The left anti-parallel and right all-/anti- parallel models generated by the MLN...TM sequence exhibited the lowest energy values for their respective topologies (Table 17). The starting register of these models agreed with the results of the DeepCoil2 for that particular sequence. Regarding the left all-parallel

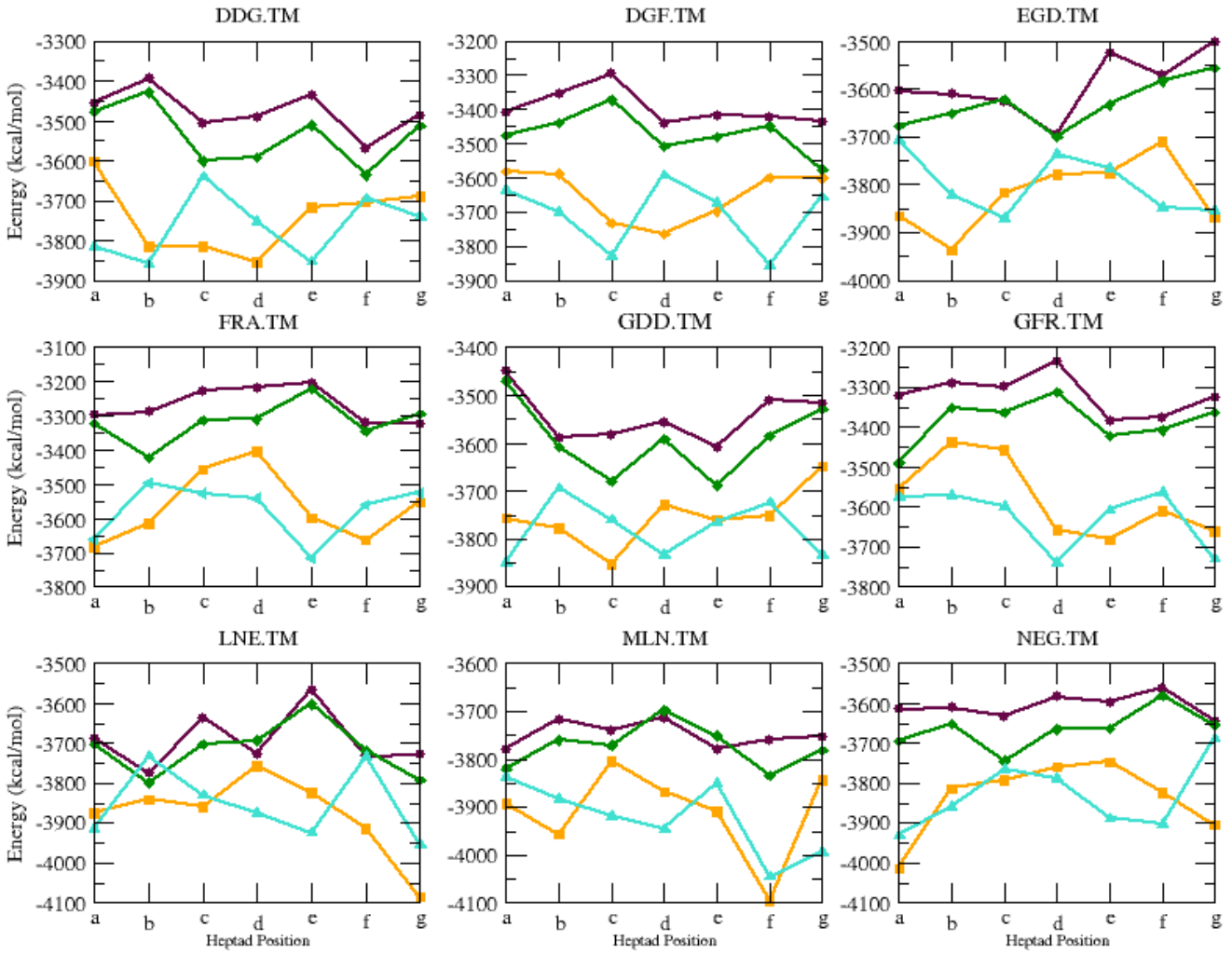


Figure 16| All the derivatives from the rRM6 sequence used to model the retro-isomer. The graphs display the energy values in every sequence for each heptad position. The left all-parallel models are colored maroon (circles), the left anti-parallel orange (squares), the right all-parallel green (diamonds) and the right anti-parallel turquoise (triangle up). In general, the all-parallel models exhibited higher energy values compared to the anti-parallel ones. The CMA -ES algorithm was used to generate the models with a number of generations equaled to 10 and population size equaled to 500.

topology, the model with lowest energy value was again the one constructed with the MLN...TM sequence (Table 17). It did not however display the same starting register as the rest models generated by the sequence. The only left all-parallel model starting with a register agreeing with the results of DeepCoil2 was the one created by the FRA...TM sequence (Table 18). The hydrophobic cores of the aforementioned structures were:

- left all-parallel retro-RM6 (*a* starting register, MLN...TM):

Glu_4 -Asp_7 -Glu_4 -Asp_7, Ala_{11} -Asp_7 -Ala_{11} -Asp_7, Ala_{11} -Arg_{15} -Ala_{11} -Arg_{15}, Arg_{15} -Glu_{18} -Arg_{15} -Glu_{18}, Glu_{18} -Asp_{22} -Glu_{18} -Asp_{22}, Asp_{22} -Ser_{25} -Asp_{22} -Ser_{25}, Ser_{25} -Asp_{29} -Ser_{25} -Asp_{29}, Asp_{29} -Glu_{32} -Asp_{29} -Glu_{32}, Glu_{32} -Glu_{36} -Glu_{32} -Glu_{36}, Glu_{36} -Thr_{39} -Glu_{36} -Thr_{39}, Thr_{39} -Ser_{43} -Thr_{39} -Ser_{43}, Ser_{43} -Phe_{46} -Ser_{43} -Phe_{46}, Phe_{46} -Asn_{50} -Phe_{46} -Asn_{50}, Asn_{50} -Thr_{52} -Asn_{50} -Thr_{52}.

- left all-parallel retro-RM6 (*e* starting register, MLN..TM):

Glu_4 -Ala_{11} -Glu_4 -Ala_{11}, Asp_7 -Ser_{14} -Asp_7 -Ser_{14}, Glu_{18} -Ala_{11} -Glu_{18} -Ala_{11}, Ser_{14} -His_{21} -Ser_{14} -His_{21}, Glu_{18} -Ser_{25} -Glu_{18} -Ser_{25}, His_{21} -Ile_{28} -His_{21} -Ile_{28}, Ser_{25} -Glu_{32} -Ser_{25} -Glu_{32}, Ile_{28} -Lys_{35} -Ile_{28} -Lys_{35}, Glu_{32} -Thr_{39} -Glu_{32} -Thr_{39}, Lys_{35} -Gln_{42} -Lys_{35} -Gln_{42}, Thr_{39} -Phe_{46} -Thr_{39} -Phe_{46}, Gln_{42} -Met_{49} -Gln_{42} -Met_{49}, Phe_{46} -Thr_{53} -Phe_{46} -Thr_{53}, Met_{49} -Gln_{56} -Met_{49} -Gln_{56}.

- left all-parallel retro-RM6 (*g* starting register, FRA...TM):

Cys_5 -Cys_5 -Cys_5 -Cys_5, Leu_9 -Leu_9 -Leu_9 -Leu_9, Ala_{12} -Ala_{12} -Ala_{12} -Ala_{12}, Leu_{16} -Leu_{16} -Leu_{16} -Leu_{16}, Cys_{19} -Cys_{19} -Cys_{19} -Cys_{19}, Leu_{23} -Leu_{23} -Leu_{23} -Leu_{23}, Leu_{26} -Leu_{26} -Leu_{26} -Leu_{26}, Leu_{30} -Leu_{30} -Leu_{30} -Leu_{30}, Thr_{33} -Thr_{33} -Thr_{33} -Thr_{33}, Ile_{37} -Ile_{37} -Ile_{37} -Ile_{37}, Ala_{40} -Ala_{40} -Ala_{40} -Ala_{40}, Ala_{44} -Ala_{44} -Ala_{44} -Ala_{44}, Glu_{47} -Glu_{47} -Glu_{47} -Glu_{47}.

- left anti-parallel retro-RM6 (*f* starting register, MLN...TM):

Met_{59} -Asp_6 -Met_{59} -Asp_6, Arg_{10} -Glu_{55} -Arg_{10} -Glu_{55}, Ala_{48} -Leu_{17} -Ala_{48} -Leu_{17}, Ile_{45} -Ala_{20} -Ile_{45} -Ala_{20}, Leu_{24} -Thr_{41} -Leu_{24} -Thr_{41},- Cys_{27} -Leu_{38} -Cys_{27} -Leu_{38}, Leu_{31} -Leu_{34} -Leu_{31} -Leu_{34}.

Regarding the models with right superhelix twist the hydrophobic cores of their best models were:

- right all-parallel retro-RM6 (*f* starting register, MLN...TM):

Asn_3 -Arg_{10} -Asn_3 -Arg_{10}, Asp_6 -Cys_{13} -Asp_6 -Cys_{13}, Arg_{10} -Leu_{17} -Arg_{10} -Leu_{17}, Cys_{13} -Ala_{20} -Cys_{13} -Ala_{20}, Leu_{17} -Leu_{24} -Leu_{17} -Leu_{24}, Cys_{27} -Ala_{20} -Cys_{27} -Ala_{20}, Leu_{31} -Leu_{24} -Leu_{31} -Leu_{24}, Cys_{27} -Leu_{34} -Cys_{27} -Leu_{34}, Leu_{38} -Leu_{31} -Leu_{38} -Leu_{31}, Leu_{34} -Thr_{41} -Leu_{34} -Thr_{41}, Leu_{38} -Ile_{45} -Leu_{38} -Ile_{45}, Ala_{48} -Thr_{41} -Ala_{48} -Thr_{41}, Ile_{45} -Ala_{52} -Ile_{45} -Ala_{52}, Ala_{48} -Glu_{55} -Ala_{48} -Glu_{55}.

- right anti-parallel retro-RM6 (*f* starting register, MLN...TM):

Met_{59} -Asp_6 -Met_{59} -Asp_6, Arg_{10} -Glu_{55} -Arg_{10} -Glu_{55}, Ala_{48} -Leu_{17} -Ala_{48} -Leu_{17}, Ile_{45} -Ala_{20} -Ile_{45} -Ala_{20}, Leu_{24} -Thr_{41} -Leu_{24} -Thr_{41},- Cys_{27} -Leu_{38} -Cys_{27} -Leu_{38}, Leu_{31} -Leu_{34} -Leu_{31} -Leu_{34}.

The cores of the anti-parallel structures were comprised by seven, symmetric hydrophobic layers. The next best solutions for the left anti-parallel models, considering their energy values and display of a correct staring register (agreeing with DeepCoil2 results), were (from lowest to highest energy): LNE...TM (-4085 kcal/mol), NEG...TM (-4014 kcal/mol), EGD...TM (-3937 kcal/mol), DDG...TM (-3854

kcal/mol), DDG...TM (-3854 kcal/mol) and GDD...TM (-3852 kcal/mol) (Table 17). The hydrophobic layers of all the aforementioned structures were: Asp6 -Met59 -Asp6 -Met59, Glu55 -Arg10 -Glu55 -Arg10, Cys13 -Ala52 -Cys13 -Ala52, Ala48 -Leu17 -Ala48 -Leu17, Ala20 -Ile45 -Ala20 -Ile45, Thr41 -Leu24 -Thr41 -Leu24, Cys27 -Leu38 -Cys27 -Leu38 and Leu34 -Leu31 -Leu34 -Leu31. As for the right anti-parallel models, the ones exhibiting low energy values and correct starting register (besides the MLN...TM model) were the LNE...TM (-3953 kcal/mol) and NEG..TM (-3931 kcal/mol) (Table 17). The hydrophobic cores of these structures were the same as the one mentioned above for the LNE, NEG, EGD, DDG and GDD left anti-parallel models.

There are three potential hydrophobic layers that emerge from the heptad motif of the retro-RM6. For the anti-parallel models and z-shift value equaled to -5, (in *adad* order from the N-terminal to the C-terminal) the hydrophobic core consisted of the residues: Asn_3 -Glu_{55} -Asn_3 -Glu_{55}, Arg_{10} -Ala_{48} -Arg_{10} -Arg_{48}, Leu_{17} -Tyr_{41} -Leu_{17} -Tyr_{41}, Leu_{24} -Leu_{34} -Leu_{24} -Leu_{34}, Leu_{31} -Cys_{27} -Leu_{31} -Cys_{27}, Leu_{38} -Ala_{20} -Leu_{38} -Ala_{20}, Ile_{45} -Cys_{13} -Ile_{45} -Cys_{13}, Ala_{52} -Asp_6 -Ala_{52} -Asp_6, for a z-shift value of -18 they were: Asn_3 -Ala_{48} -Asn_3 -Ala_{48}, Arg_{10} -Tyr_{41} -Arg_{10} -Tyr_{41}, Leu_{17} -Leu_{34} -Leu_{17} -Leu_{34}, Leu_{24} -Cys_{27} -Leu_{24} -Cys_{27}, Leu_{31} -Ala_{20} -Leu_{31} -Ala_{20}, Leu_{38} -Cys_{13} -Leu_{38} -Cys_{13} and Ile_{45} -Asn_6 -Ile_{45} -Asn_6, while for a z-shift value of 3 the hydrophobic layers were: Arg_{10} -Glu_{55} -Arg_{10} -Glu_{55}, Leu_{17} -Ala_{48} -Leu_{17} -Ala_{48}, Leu_{24} -Tyr_{41} -Leu_{24} -Tyr_{41}, Leu_{31} -Leu_{34} -Leu_{31} -Leu_{34}, Leu_{38} -Cys_{27} -Leu_{38} -Cys_{27}, Ile_{45} -Ala_{20} -Ile_{45} -Ala_{20}, Ala_{52} -Cys_{13} -Ala_{52} -Cys_{13} and Met_{59} -Asp_6 -Met_{59} -Asp_6.

The selection of the appropriate translocation was performed after trials of different z-shift values by applying a population size of 100 and number of generations equaled to 2. Regarding the all-parallel structures the same workflows were carried out. In the case of the z-shift manipulation, the two values used to recreate the hydrophobic cores of the all-parallel structures were 10 and 0. The hydrophobic layers of the all-parallel structures for z-shift equaled to 0 and its mean value equaled to 0 were (starting from the N -terminal to the C -terminal in *aaaa/dddd* order): Asn_3 -Asn_3 -Asn_3 -Asn_3, Asp_6 -Asp_6 -Asp_6 -Asp_6, Arg_{10} -Arg_{10} -Arg_{10} -Arg_{10}, Cys_{13} -Cys_{13} -Cys_{13} -Cys_{13}, Leu_{17} -Leu_{17} -Leu_{17} –Leu_{17}, Ala_{20} -Ala_{20} -Ala_{20} -Ala_{20}, Leu_{24} -Leu_{24} -Leu_{24} -Leu_{24}, Cys_{27} -Cys_{27} -Cys_{27} -Cys_{27}, Leu_{31} -Leu_{31} -Leu_{31} -Leu_{31}, Leu_{34} -Leu_{34} -Leu_{34} -Leu_{34}, Leu_{38} -Leu_{38} -Leu_{38} -Leu_{38}, Thr_{41} -Thr_{41} -Thr_{41} -Thr_{41}, Ile_{45} -Ile_{45} -Ile_{45} -Ile_{45}, Ala_{48} -Ala_{48} -Ala_{48} -Ala_{48}, Ala_{52} - Ala_{52} - Ala_{52} – Ala_{52}, Glu_{55} -Glu_{55} -Glu_{55} -Glu_{55} and Met_{59} -Met_{59} -Met_{59} -Met_{59}. While for a z-shift value equaled to 10 (mean value equaled to 0): Arg_{10} -Asn_3 -Arg_{10} -Asn_3, Asp_6 -Cys_{13} -Asp_6 -Cys_{13}, Arg_{10} -Leu_{17} -Arg_{10} -Leu_{17}, Cys_{13} -Ala_{20} -Cys_{13} -Ala_{20}, Leu_{17} -Leu_{24} -Leu_{17} -Leu_{24}, Ala_{20} -Cys_{27} -Ala_{20}-Cys_{27}, Leu_{24} -Leu_{31} - Leu_{24} -Leu_{31}, Cys_{27} -Leu_{34} -Cys_{27} -Leu_{34}, Leu_{31} -Leu_{38} -Leu_{31} -Leu_{38}, Leu_{34} -Thr_{41}-Leu_{34} -Thr_{41}, Leu_{38} -Ile_{45} -Leu_{38} -Ile_{45}, Thr_{41} -Ala_{48} -Thr_{41} -Ala_{48}, Ile_{45} -Ala_{52} -Ile_{45} -Ala_{52}, Ala_{48} -Glu_{55} -

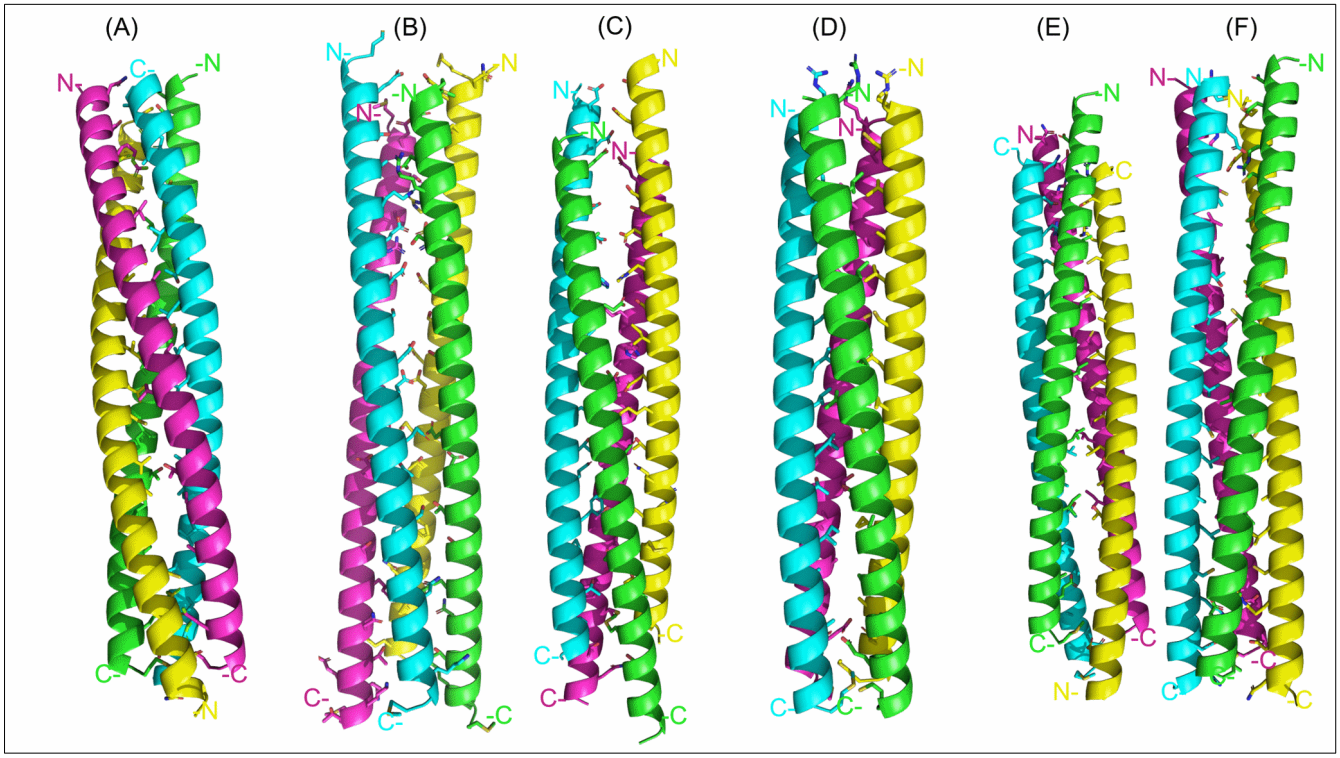


Figure 15| The best models generated by the residue deletion workflow. They were selected on the basis of their energy and agreement to the DeepCoil2 results for the retro-RM6 sequence. (A) The left anti-parallel retro-RM6 generated by the MLN...TM sequence. This model exhibited the lowest energy (Table 17) for this particular topology in addition to exhibiting the correct starting register (compared to the DeepCoil2 results). (B) The MLN...TM *a* left all-parallel model that displayed the lowest energy value (Table 17).(C) The MLN...TM *e* left all-parallel model that exhibited the same low energy value as the *a* left all-parallel model. (D) The FRA...TM *g* left all-parallel model with the same heptad motif as the one identified by the DeepCoil2 program. The right all-parallel (E) and right anti-parallel (F) models that exhibited the lowest energy value and agreeing starting register for that topology. The residues that comprised the hydrophobic cores of those structures are displayed as sticks. Chain A is colored green, chain B cyan, chain C magenta and chain D yellow. The coloring scheme is the same for all the structures.

Ala_{48} -Glu_{55} and Ala_{52} -Met_{59} -Ala_{52} -Met_{59}. Similarly, the generated models (Figure 16) were refined using the aforementioned tool (Table 18). Results from the anti-parallel structures indicated that there was an energy preference to the models (right and left) produced by the 3 z-shift value (Table 18). These structures exhibited the lowest energy values compared to the rest (anti – and all- parallel). As for the all-parallel structures of rRM6, the left- and right- all parallel models with a z-shift value equaled to 0 exhibited the lowest energy values.

Table 17: Energies from the best models (model1) of the left/right all-/anti- parallel retro-rRM6 structures.

Sequence	Starting register/ Model	Energy before refinement (kcal/mol)	Energy after refinement (kcal/mol)	Same hydrophobic core as the DeepCoil2 results
DDG.TM				
	f left all-parallel	-3568	-11,000	N
	d left anti-parallel	-3854	-11,634	Y
	f right all-parallel	-3636	-10,819	N
	b right anti-parallel	-3857	-10,730	N
DGF.TM				
	d left all-parallel	-3441	-10,869	N
	d left anti-parallel	-3764	-10,842	N
	g right all-parallel	-3575	-10,625	N
	f right anti-parallel	-3855	-10,505	N
EGD.TM				
	d left all-parallel	-3696	-11, 341	N
	b left anti-parallel	-3937	-11,958	Y
	d right all-parallel	-3701	-11,155	N
	c right anti-parallel	-3871	-11,163	N
FRA.TM				
	g left all-parallel	-3324	-10,964	Y
	a left anti-parallel	-3684	-10,305	N
	b right all-parallel	-3422	-10,248	N
	e right anti-parallel	-3716	-10,136	N
GDD.TM				
	e left all-parallel	-3609	-11,120	N
	c left anti-parallel	-3852	-11,761	Y
	e right all-parallel	-3688	-11,016	N
	a right anti-parallel	-3850	-10,889	N
GFR.TM				
	e left all-parallel	-3383	-10,629	N
	e left anti-parallel	-3682	-10,575	N
	a right all-parallel	-3490	-10,462	N
	d right anti-parallel	-3738		N
LNE.TM				
	b left all-parallel	-3775	-11,648	N
	g left anti-parallel	-4085	-12,310	Y
	b right all-parallel	-3799	-11,475	N
	g right anti-parallel	-3953	-12,025	Y
MLN.TM				
	a/e left all-parallel	-3780	-11,558, -11,639	N
	f left anti-parallel	-4095	-11,645	Y
	f right all-parallel	-3834	-11,580	Y

	f right anti-parallel	-4045	-11,504	Y
NEG.TM				
	g left all-parallel	-3643	-11,637	N
	a left anti-parallel	-4014	-12,110	Y
	c right all-parallel	-3744	-11,401	N
	a right anti-parallel	-3931	-11,826	Y

Table 18: Energies from the different models generated by keeping the z-shift and phiCα values constant

z-shift	Model Starting register *f*	Energy before refinement (kcal/mol)	Energy after refinement (kcal/mol)
3			
	left anti-parallel	-3987	-12514
	right anti-parallel	-4034	-12111
-5			
	left anti-parallel	-3836	-12416
	right anti-parallel	-3821	-12085
-18			
	left anti-parallel	-3680	-12180
	right anti-parallel	-3785	-11918
0			
	left all-parallel	-3802	-12244
	right anti-parallel	-3806	-12087
10			
	left all-parallel	-3634	-12289
	right anti-parallel	-3638	-12039

Chapter 5 Discussion

As previously stated, the structural regularity of the coiled-coil proteins is evident in the presence of a heptad repeat [3–5]. This motif ensures the adoption of particular backbone conformations that can be described and recreated by a set of parameters [3,4,12,14,15]. The parameterization of coiled-coil geometry paves the way for a new modeling method termed as geometric modeling [12,14,15]. The latter approach could enable the investigation of the uncharted parameter space of coiled-coil folding [14,21]. This “Dark matter” of conformations could entail protein structures that are not commonly encountered in nature, such as retro-proteins [56,57,65,68,69,69]. These molecules are constructed by the reversal of amino acid sequences which are derived from native or native-like proteins [58].

The study of the aforementioned isomers could answer prominent questions regarding the process of protein folding [49,51,53,58,67]. In more detail, retro-studies could enrich our current understanding on what extend does backbone directionality contribute to protein folding, specifically in the case of HBs [49,51,53,57,58,68]. Early hypotheses supported that the retro-proteins are a mirror image of their

parents, while subsequent studies opposed this concept [58,68]. There have been additional studies which proposed that retro-proteins could adopt similar folds to their parents [55,57,69]. However, on those occasions, the stability of the retro-isomer was questionable [68,69]. A significant paradigm of a retro-structure is the one of GCN4, which represents the only retro-isomer of a HB that is stable and experimentally determined [57].

Subsequent attempts to determine the structure of a different retro-isomer, the one of the RM6 protein (PDB ID: 1QX8) failed [83]. The latter protein represents a ROP (PDB ID: 1ROP) variant which forms a highly stable and canonical HB [41]. Molecular replacement attempts on crystallographic data obtained from the rRM6 crystals [83] failed to allow a complete structure determination, implying potential differences between the retro-isomer and its parent. This signifies the need to construct possibly useful rRM6 models for molecular replacement calculations. In order to accomplish that geometric modeling was employed which does not require any prior knowledge about the protein, besides that it forms a coiled-coil structure [14,15]. Taking all of the above into account, the primary aim of present study was to evaluate the ability of ISAMBARD to generate RM6 (PDB ID: 1QX8) [41] models of quality and accuracy on an atomic level. On that notion, it was necessary to identify an appropriate workflow, as well, in order for ISAMBARD to identify energetically acceptable parameter combinations. Based on those results, the final goal was to produce a plethora of retro-RM6 models in order to assess and select a group that could be utilized in crystallographic molecular replacement calculations.

5.1 Constructing the RM6 models; metaheuristics

In order to generate the RM6 models and assess the efficiency and accuracy of the metaheuristic algorithms results from the DeepCoil2 program were used (Figure 6) [84]. This program appeared to be accurate at predicting the *a/d* positions of the RM6 structure, while excluding hydrophobic layers that were mobile, asymmetric and penetrated by water molecules [41]. The method of grid scanning was excluded at early steps of the study due to long running times and higher energy values compared to results of the metaheuristic algorithms (data not shown).

The modeling process started with the troubleshooting of the optimizer selection (Figure 7) and then the calibration of the population size and number of iterations (Figures 8,9). These steps are crucial when working with these types of algorithms since the values of these parameters highly influence their performance [35]. Regarding the selection of CMA -ES, this particular algorithm represents a “state-of-the-art” framework in evolutionary computing and identifies the optimal solution of the optimization problem in few generations [32]. The application of a small number of generations (iterations)

proves highly useful when encountering a complex optimization problem and reduces the time complexity [32,34]. Thus, the selection of a small number of iterations (10) for the problem of interest agreed with previous results on that matter [31,32,34]. As for the suitable population size when using this algorithm, it has been suggested that a reasonable parameter value is preferable since adaptation time is more or less independent of it [31,34]. Notably, values below 10 are not preferred since they affect the robustness of the method, and an increasing number of λ (population size) decreases the performance of the algorithm in both simple and complex optimization problems [34]. When testing the different population sizes -200,500 and 1000, the resulting energy values did not change drastically for the (best) *d* left anti-parallel models (-4068 kcal/mol, -4069 kcal/mol and -4074 kcal/mol). However, the running times significantly increased (23', 1 hour and ~3 hours). In addition, the energies of the rest of the models were comparably lower for the 500/10 run (Figure 8). As a result, the selected values of 10 for the number of generations and 500 for the population size (λ) were considered suitable for identification of the RM6's (and future rRM6) coiled-coil parameters.

After the identification of the appropriate workflow and input parameters when working with a metaheuristics framework, additional steps to assess the ability of ISAMBARD to generate and identify the correct model for the RM6 protein were carried out. For that purpose, models with wrong parameters were constructed. More specifically, models with wrong superhelical twist (right) and orientation (parallel). When comparing the energies obtained from the left all- (Figure 12, maroon) and left anti- (Figure 12, orange) parallel arrangements of the RM6 models, the correct left anti-parallel model exhibited the lowest energy value. Hence, ISAMBARD identified the native orientation of that particular sequence (Figure 12). On top of that, these results exhibited a pronounced energy minimum for the *d* register which is only present when the (correct) anti-parallel helical twist is applied (Figure 12). In addition, this program correctly differentiated between solutions for the helices' major handedness (Figure 12). Both right all-parallel (Figure 12, green) and right anti-parallel (Figure 12, turquoise) hypothetical models of RM6 exhibited higher energies that the correct (normal) left anti-parallel structure (Figure 12, orange). The energy results from the geometric modeling of RM6 suggest that the provided sequence folds into a left anti-parallel HB starting with a residue occupying the *d* position in the heptad motif, which agree with the experimental structure [41]. It was necessary to evaluate the similarity of the structures (modeled and resolved RM6) on an atomic level as well. This was accomplished via calculating the RMSD value of the *d* left anti-parallel model to the RM6 structure. Impressively, the RMSD value was only 1.7 Å for 196 Cα atoms before refinement. After refinement this value was dropped down to 1.21 Å for the same number of atoms (Table 15, MT...GENL sequence). These results are

clearly impressive since the modeled RM6 was not constructed by a knowledge-based method (i.e. homology modeling) but solely by using geometric backbone parameters [14]. The robustness and efficiency of this method was also evident when testing other parameters as well. This program is powerful enough to be able to find convincing solutions even when initiated from wrong sequences. For instance, instead of the MT…GENL, the appropriate sequence to use when modeling this protein based on its heptad repeat it is MT...GDD (Table 14) [41]. This is also evident from the fact that the former sequence does not recreate the hydrophobic core of RM6 (Table 13), while the MT...GDDG, MT...GDD, MT...GD and MT...G do (Table 14).

Despite the fact that the last four sequences resulted in the same hydrophobic core as RM6 (Table 14), the results suggested that the left anti-parallel structure modeled with the MT...GDD sequence represented the best solution (Figures 13,14) (Tables 14,15). Even though the MT...GENL-left anti-parallel model displayed the lowest energy value (Table 15) it did not exhibit the lowest RMSD (Table 15). The energy values of -4069 kcal/mol before and -12192 after refinement could be explained by the fact that longer helical sequences commonly produce more stable structures [93]. Nevertheless, the energy values from ISAMBARD & GALAXY are not comparable. They are based on different force fields with different assumptions. This is, however, the case for many synthetic sequences and it has been proposed that by increasing the helical sequence the amount of hydrophobic contacts surges [93]. Energy results from the sequence derivatives of the RM6 protein (Table 15) agree with this concept [93]. The super-positioning of the MT...GD left anti-parallel and MT..G left anti-parallel models with the RM6 also exhibited low RMSD values comparable or lower than those of the best model (MT...GDD) (Table 15). In order to further evaluate the alignment of each model, the joining of the chains was assessed (Figure 14). For instance, the MT...G left anti-parallel model even though exhibited an RMSD value near the best model (1.17 Å), the joining of the models' chains with the resolved structure was wrong (ABCD:BADC) (Figure 14). On top of that, the deletion of just two residues per helix increased the energy value before (~180 kcal/mol) and after (~410 kcal/mol) refinement (Table 15). This was also the case for the MT...GD left anti-parallel model which even though exhibited the lowest RMSD value (1.13 Å) and an acceptable chain joining (ABCD:CDAB, RM6 is a dimer of dimers) [41], the deletion of just one residue per helix increased the energies of the model again, before (106 kcal/mol) and after (263 kcal/mol) refinement (Table 15).

5.2 Constructing the retro-RM6 models

These results suggested that the geometric modeling of an α-helical bundle, like RM6, using ISAMBARD, gave models of quality and accuracy sufficient enough even for demanding calculations such

as crystallographic molecular replacement. For the modeling of rRM6, which does not exhibit an available crystallographic structure, a similar workflow was carried out. Only *priori* of this set of experiments was that the retro-sequence folds into an α-helical bundle. Results from the structural study of this protein suggested that it is a stable and foldable α-helical structure [53]. Also, the oligomeric state of the retro-isomer was assumed to be tetrameric based on the results of the same study [53]. Further results indicated that like its parent, the rRM6 protein formed a coiled-coil structure [53]. In addition, SAXS (Small-angle X-ray scattering) experiments suggested that the rRM6 and RM6 proteins shared the same shape with the former molecule being potentially larger [53]. Since the heptad repeat of this protein is unknown and unclear whether it shares the same one with RM6, the DeepCoil2 program was utilized. This tool provided satisfactory results for the parent sequence, thus it was also employed in the case of rRM6.

Results from the DeepCoil2 server suggested that the register of the starting residue was *f*. Again there were two residues (Ile_{28} and Lys_{35}) that diverged from this pattern, but their probabilities were lower than the rest and hence not considered (Figure 15). There are four potential topologies for the rRM6 protein; left anti-parallel, left all-parallel, right anti-parallel and right all-parallel. This in addition to the fact that the retro-structure might exhibit potential differences compared to the parent protein, were the main criteria for the final model selection. The models and their energies generated by both the z-shift manipulation and the sequence derivatives were compared. This was performed in order to identify the hydrophobic core exhibiting the lowest energy value and hence the potentially most representative rRM6 model.

In general, the energy results from the retro-RM6 sequence derivatives suggested that the anti-parallel models exhibited the lowest energy values in all starting registers (Figure 16). Starting with the left anti-parallel topology, results from residue deletion in the N-terminal end of the retro-sequence suggested that the MLN...TM left anti-parallel model exhibited the lowest energy value (Table 17). In addition, both the left and right anti-parallel MLN...TM models displayed the same hydrophobic cores as the anti-parallel models generated by a z-shift value equaled to 3 (Table 18). However, the low energy values of the MLN...TM models could also be contributed to the length of the sequence [93]. Regarding the all-parallel structures generated by residue deletion, only one left all-parallel model exhibited the correct starting register (Table 17) (FRA...TM), whereas no right all-parallel model exhibited a starting register agreeing with the results of DeepCoil2 (Figure 15). These results imply, that the particular sequence could potentially favor the formation of an anti-parallel structure instead of a parallel one. However, the energy values from the sequence derivatives cannot be confidently compared with each

other, since the deletion of residues is a drastic intervention to a structure which affects its stability [93]. In addition, by following this method, residues which could be potentially useful are deleted from the structure.

Nevertheless, the preference for an anti-parallel structure was displayed by the experiments with manipulation of the z-shift value (Table 18). In this set of experiments the most energetically preferred hydrophobic core was the one constructed with z-shift value equaled to 3 with the left anti-parallel model displaying the lowest energy after refinement (Table 18). Regarding the rest of the left anti-parallel structures, as the chains C and D were translocated towards the N-terminal (z-shift values -5 and -18) the energy values after refinement were becoming greater by ~100 kcal/mol and ~334 kcal/mol respectively (Table 18). This was also the case for the right anti-parallel models with their energies being increased by 26 kcal/mol and 193 kcal/mol respectively (Table 18). This model also exhibited the lowest energy when displaying the hydrophobic core generated by a z-shift value equaled to 3. By changing only the orientation and keeping the major handedness left, the energy values are raised by 270 kcal/mol and 225 kcal/mol for z-shifts values equaled to 0 and 10 respectively, which also the case for the right handed topology (Table 18).

5.3 Conclusions

Results from the present study suggest that the best models for the retro-RM6 are potentially of left major handedness and anti-parallel orientation. These could be useful for molecular replacement calculations in order to determine the rRM6 structure. All the anti-parallel models which displayed the lowest energy either generated by residue deletion or by a z-shift value equaled to 3, exhibited the same hydrophobic core as the one observed in the crystallographic structure of RM6. Remarkably, even upon sequence reversal, the hydrophobic core present in the native structure is energetically preferred in the theoretical models of the retro-RM6 generated by ISAMBARD. These results, further support the notion that the preserved physicochemical properties of the amino acids in the retro-RM6 contribute to the conservation of the native structural characteristics [52,53,57]. On top of that, these results agree with previous studies which suggest that structural changes in the retro-structure are in accordance with alterations in the native hydrophobic core [51,53]. These findings further add to the already available information about the retro-RM6 [53] by proposing that this isomer potentially maintains the major handedness and orientation of the native structure.

Since the results did not exhibit a pronounced energy minima which could suggest that a particular topology is most preferred, more than one models could be utilized fore molecular replacement calcu-

lations. For instance, the potential structural differences between the rRM6 and RM6 structures, which hindered the determination of the former, could be attributed to their different major handedness. Therefore, the right anti-parallel model constructed by ISAMBARD might represent an additional candidate for future experiments on rRM6 determination besides the left anti-parallel one. Different mutational studies on the ROP left-all-parallel structure have resulted in the transformation into a right-handed mixed parallel and anti-parallel protein [94]. Similarly to the retro-GCN4, the retro-RM6 could have retained the secondary structure elements of its parent protein but form a distinct protein structure [57]. This was also supported by computational experiments on retro-peptides, which exhibited the same secondary structure propensities as their counterparts, but not necessarily the same fold [51]. As for the all-parallel structures, the left- and right- models generated by a z-shift value equaled to 0 and 10 could be also considered.

5.4 Limitations and Future work

The present data supported the notion that the ISAMBARD program created RM6 models sufficient for molecular replacement calculations and hence this could also be the case for the left anti-parallel model of retro-RM6. However, these results must be interpreted with caution and a number of limitations should be borne in mind. Notably, even though the aforementioned studies suggested that the retro-RM6 might fold into an α-helical structure [53], it is not certain whether it folds into a canonical helical bundle like its parent. In case it does not, this could nullify the whole procedure from the beginning since it was assumed that it does in order to calculate the geometrical parameters. Another limitation of the study is that geometric modeling via ISAMBARD is powerful enough to construct energetically acceptable models of switching registry and z-shift translocation. This further complicated the model selection since the results did not display a significant and pronounced solution. An already mentioned solution would be to use multiple retro-RM6 models (see 5.3) which would facilitate the structure determination of the isomer. In addition, further evaluation steps could be implemented in order to better assess the produced structures. For instance, the ISAMBARD program offers python libraries which are used to evaluate the quality of the models. More specifically, they are able to calculate the contact order of a molecule normalized by its sequence length [14,95,96] (see Evaluation package). On top of that, the packing quality of a protein's hydrophobic core can be assessed via calculating the hydrophobic fitness [14,97] (see Evaluation package). This scoring method can potentially recognize the native fold among the potential models of retro-RM6 and suggest the pronounced topology of the structure [97]. A different scoring function could be employed and adapted for assessing inter-helical interactions, similarly to the BUDE force field. Last but not least, *ab initio* modeling protocols

could be integrated in the study as an alternative method for generating potential structures of retro-RM6 for molecular replacement calculations [98].

Supplementary 1| The left-anti-parallel models of RM6 in the 200/10 run

CMAES

Running register a

gen evals avg std min max

0 200 -2263.7 1858.4 -3689.51 9179.81

1 200 -2751.83 1474.57 -3663.62 9849.3

2 200 -3080.63 790.912 -3710.55 730.916

3 200 -3281.66 411.752 -3707.16 -1107.59

4 200 -3389.34 261.58 -3697.98 -2388.53

5 200 -3387.2 292.347 -3677.34 -2071

6 200 -3486.96 178.925 -3740.93 -2241.51

7 200 -3507.8 174.8 -3739.5 -2432.41

8 200 -3543.93 88.3934 -3745.83 -3257.89

9 200 -3551.94 109.387 -3710.51 -2764.88

Evaluated 2200 models in total in 0:32:56.170391

Best fitness is (-3745.834317242302,)

Best parameters are [58, 7.203610817182088, 318.9881099641674, 18.67469695040517, -1.6353546160084265]

Running register b

gen evals avg std min max

0 200 -2322.48 1542.59 -3722.87 4001.68

1 200 -2503.59 2410.53 -3771.82 22156.2

2 200 -3152.65 552.932 -3643.19 129.334

3 200 -3401.15 307.82 -3767.58 -1611.58

4 200 -3455.59 367.505 -3774.88 1038.9

5 200 -3518.38 149.566 -3771.51 -2161.61

6 200 -3552.19 108.426 -3821.95 -3238.79

7 200 -3604.05 108.318 -3824.09 -3254.4

8 200 -3610.06 121.198 -3839.1 -3040.54

9 200 -3626.16 199.435 -3888.87 -1284.93

Evaluated 2200 models in total in 0:32:59.164348

Best fitness is (-3888.8681338049532,)

Best parameters are [58, 7.209947443385686, 197.28012769504835, 121.2169975956489, -2.3308680360565424]

Running register c

gen evals avg std min max

0 200 -1527.34 2089.89 -3613.26 5685.92

1 200 -1858.05 2669.99 -3676.38 11766.2

2 200 -2370.95 2781.77 -3575.7 20383.6

3 200 -3195.62 530.13 -3639.79 540.012

4 200 -3023.74 1837.3 -3652.61 14412.4

5 200 -3235.96 397.114 -3664.32 -1406.2

6 200 -3311.63 348.188 -3757.17 -732.922

7 200 -3210.15 480.745 -3699.56 -5.29113

8 200 -3388 264.215 -3687.41 -2216.36

9 200 -3500.32 166.877 -3720.05 -2612.38

Evaluated 2200 models in total in 0:37:24.126057

Best fitness is (-3757.1703844209574,)

Best parameters are [58, 7.575446608160528, 323.419172446283, 258.42857142857144, -6.018977590641277]

Warning! Parameter 3 is at or near maximum allowed value

Running register d

gen evals avg std min max

0 200 -2550.54 1823.72 -3853.09 12400.5

1 200 -2988.23 1370.77 -3875.32 8534.98

2 200 -3336.43 660.537 -3911.28 1368.48

3 200 -3657.97 291.712 -4015.35 -1526.9

4 200 -3614.51 469.609 -3994.74 391.875

5 200 -3674.27 396.218 -4053.93 -278.535

6 200 -3702.62 364.587 -4049.38 -1012.88

7 200 -3852.64 126.532 -4067.3 -3032.48

8 200 -3907.8 73.4105 -4057.27 -3538.1

9 200 -3939.27 68.0728 -4068.66 -3640.43

Evaluated 2200 models in total in 0:23:08.224060

Best fitness is (-4068.6582206340418,)

Best parameters are [58, 6.58003914300648, 201.74795126373593, 322.8023176842898, 1.933370217717626]

Running register e

gen evals avg std min max

0 200 -2304.88 1463.81 -3651.66 4680.54

1 200 -2000.84 2802.99 -3624.24 18721.6

2 200 -2995.33 1047.08 -3761.44 5678.42

3 200 -3217.21 687.483 -3770.16 5024.47

4 200 -3305.9 266.844 -3678.18 -1394.48

5 200 -3366.35 266.193 -3716.45 -1284.6

6 200 -3343.36 257.665 -3755.56 -2346.17

7 200 -3362.46 284.256 -3787.54 -1339.65

8 200 -3382.1 290.522 -3782.57 -651.623

9 200 -3355.67 540.003 -3770.96 3096.86

Evaluated 2200 models in total in 0:29:17.599519

Best fitness is (-3787.5434849989256,)

Best parameters are [58, 7.161478839962543, 298.53431981029775, 97.42727603906239, 0.8353219948130775]

Running register f

gen evals avg std min max

0 200 -2160.58 1721.03 -3604.15 5048.9

1 200 -2662.49 1417.33 -3676.34 7944.47

2 200 -2622.26 2338.46 -3582.27 17821.7

3 200 -3162.46 666.857 -3701.66 1448.64

4 200 -3317.44 252.159 -3701.02 -1351.99

5 200 -3290.94 306.84 -3795.59 -1424.15

6 200 -3305.97 303.537 -3693.91 -1853.03

7 200 -3327.83 273.154 -3665.8 -1607.47

8 200 -3305.15 306.19 -3678 -1645.07

9 200 -3257.42 381.735 -3613.67 -1153.31

Evaluated 2200 models in total in 0:33:43.178407

Best fitness is (-3795.588241130886,)

Best parameters are [58, 7.2339928490270236, 249.4277558805806, 154.9201168686517, -2.2340935967569644]

Running register g

gen evals avg std min max

0 200 -1939.72 1947.98 -3740.67 13351.9

1 200 -2420.43 2152.62 -3692.7 14667.6

2 200 -3100.12 856.681 -3726.69 5506.28

3 200 -3352.54 450.199 -3732.26 503.446

4 200 -3406.67 341.732 -3754.22 -1675.77

5 200 -3467.42 290.028 -3744.23 -1420.16

6 200 -3580.92 121.047 -3760.63 -2812.56

7 200 -3634.13 100.003 -3763.68 -2889.38

8 200 -3643.94 124.244 -3784.3 -3029.14

9 200 -3657.57 204.122 -3787.07 -1140.03

Evaluated 2200 models in total in 0:21:49.690280

Best fitness is (-3787.069736773192,)

Best parameters are [58, 7.434176313256784, 227.00234109707864, 289.2790058430281, 1.2338891732129433]

PSO

Running register a

gen evals avg std min max

0 200 -1712.27 3377.26 -3679.08 27204.6

1 179 -2290.47 3438.83 -3635.02 27204.6

2 152 -2126.65 2303.33 -3683.24 20017.8

3 165 -1694.06 2044.48 -3728.63 9128.21

4 190 864.142 7744.62 -3644.88 42278.7

5 171 -561.5 7349.52 -3716.05 42278.7

6 182 -1257.24 3557.19 -3644.14 24061

7 187 -1808.37 3347.19 -3727.23 26888.5

8 186 -2259.98 2432.93 -3698.7 20051.6

9 185 -2735.01 2102.33 -3728.48 15129.4

Evaluated 1797 models in total in 0:27:28.286429

Best fitness is (-3728.6329211457437,)

Best parameters are [58, 7.6731291124090735, 311.8147103496983, 38.69926659776473, 8.209928934769327]

Running register b

gen evals avg std min max

0 200 -1916.23 2785.72 -3843.19 16566.1

1 188 -1931.48 2890 -3769.04 17686.4

2 194 -2136.33 2163.87 -3822.18 11082.2

3 196 20.2222 3585.52 -3647.6 16382

4 192 -2120.99 2781.88 -3713.52 16382

5 186 -2731.79 1422.57 -3756.94 8197.36

6 183 -1515.56 2632.39 -3749.4 12099

7 186 930.806 8405.27 -3767.77 58327.6

8 181 -0.220355 8368.11 -3484.73 58327.6

9 186 -2891.26 1220.85 -3673.84 6278.79

Evaluated 1892 models in total in 0:29:30.299556
Best fitness is (-3843.191723225236,)
Best parameters are [58, 7.215313576558515, 173.98478612498883, 131.28635524280716, -3.205906066304327]
Running register c

gen evals avg std min max

0 200 -1830.77 2911.61 -3650.35 18264.5

1 169 -2234.04 2066.01 -3763.81 17081.1

2 185 -2562.1 1518.25 -3761.41 9205.98

3 186 1095.77 4453.65 -3622.57 25244

4 185 -79.3148 4164.39 -3620.24 25244

5 177 -2194.3 2567.28 -3675.67 16209.8

6 174 -2485.97 2526.24 -3680.64 15960.3

7 167 -1772.57 2391.41 -3680.64 15960.3

8 132 -843.781 2915.83 -3680.64 15960.3

9 129 -842.021 3296.29 -3680.64 15374.7

Evaluated 1704 models in total in 0:28:38.075763
Best fitness is (-3763.813468006548,)
Best parameters are [58, 7.534038359054897, 195.151651231609, 221.69721220416403, -5.778720752652821]
Running register d

gen evals avg std min max

0 200 101.279 5755.57 -3862.64 27341.5

1 183 2889.21 8745.4 -3922.64 37242

2 181 1817.59 14438.1 -3866.27 170953

3 191 765.467 13714.3 -3897.08 170953

4 193 -887.41 4879.89 -3924.47 25447

5 189 -1101.03 4679.42 -3842.37 25618.7

6 189 -1060.86 4803.09 -3939.33 25618.7

7 192 -1296.15 4263.48 -3917.69 21441.2

8 194 -841.692 5030.07 -3896.4 26712.4

9 191 -1376.97 4580.23 -3972.97 26712.4

Evaluated 1903 models in total in 0:23:23.413153

Best fitness is (-3972.96780212029,)
Best parameters are [58, 6.7296580671526804, 108.37372670592491, 328.36862230707527, 0.3311373507659721]
Running register e

gen evals avg std min max

0 200 -1371.77 4004.71 -3680.09 20174.2

1 187 -1632.28 3772.97 -3702.89 27310.5

2 187 -2777.37 2408.89 -3688.97 27310.5

3 197 -3107.07 814.085 -3703.63 2683.41

4 184 -2091.95 2299.65 -3611 13953.4

5 175 -2600.53 1896.71 -3611 11125.7

6 177 -3118.99 741.34 -3627.14 2305.96

7 126 -1999.01 2413 -3644.75 8566.74

8 164 237.245 5438.61 -3674.15 22370

9 155 -1044.74 4377.04 -3559.29 28953

Evaluated 1752 models in total in 0:24:28.128215
Best fitness is (-3703.6321288102927,)
Best parameters are [58, 7.250096732423234, 349.45870640268697, 68.59077107717573, -13.501662405705622]
Warning! Parameter 2 is at or near maximum allowed value
Running register f

gen evals avg std min max

0 200 -1089.03 15324.4 -3652.5 212675

1 159 -1702.63 15319.9 -3705.14 212675

2 103 -3063.41 1074.56 -3723.9 8968.04

3 162 -959.731 6155.74 -3612.83 68677.5

4 176 -1688.38 5579.17 -3616.57 68677.5

5 177 -2503.15 2342.45 -3685.65 20890.3

6 113 -2227.4 3208.32 -3619.22 25453.2

7 152 -2653.91 2547.93 -3632.87 25453.2

8 173 -2594.54 1975.19 -3619.22 15838.4

9 161 -2625.78 1341.48 -3768.68 3405.64

Evaluated 1576 models in total in 0:27:56.591209
Best fitness is (-3768.6832003137083,)
Best parameters are [58, 7.311483553894328, 292.67358940633846, 153.0506997679393, -1.9879321171086128]
Warning! Parameter 3 is at or near minimum allowed value
Running register g

gen evals avg std min max

0 200 -1003.28 5490.84 -3703.36 38765.1

1 166 -1820.05 4689.46 -3718.2 38765.1

2 171 -2466.47 2475.24 -3786.36 13916.6

3 151 -2767.39 1413.96 -3806.89 12145.7

4 153 -2352.7 1540.85 -3760.43 5480.88

5 168 -1768.35 2431.09 -3722.66 9286.02

6 172 -1958.38 2717.56 -3855.32 18378.6

7 182 -1953.91 2916.76 -3855.32 13524.5

8 130 -1776.08 4362.45 -3855.32 30255

9 131 -1994.36 3633.82 -3775 30255

Evaluated 1624 models in total in 0:18:24.129491
Best fitness is (-3855.317159632557,)
Best parameters are [58, 7.624638740763023, 154.3211213938332, 316.8997237966096, -7.474548294442053]
Warning! Parameter 3 is at or near maximum allowed value

GA

Running register a

gen evals avg std min max

0 123 -2816.66 907.138 -3655.05 151.313

1 146 -3402.65 121.993 -3676.69 -3181.11

2128 -3485.05 89.8886 -3676.69 -3339.2

3 161 -3572.59 60.1891 -3710.9 -3471.2

4 137 -3618.27 43.6413 -3728.05 -3546.65

5 153 -3647.25 33.0902 -3733.87 -3589.51

6 150 -3666.66 27.0574 -3767.63 -3625.48

7 148 -3683.66 22.9919 -3782.66 -3649.17

8 135 -3694.73 19.3884 -3782.66 -3665.16

9 146 -3704.01 16.4338 -3782.66 -3678.82

Evaluated 1627 models in total in 0:27:26.203894
Best fitness is (-3782.6577053157885,)
Best parameters are [58, 7.43843892362984, 165.1960570348028, 29.331743887072978, -1.6620870819393938]
Running register b

gen evals avg std min max

0 153 -3105.58 488.762 -3742.04 -1575.56

1 155 -3432.76 110.928 -3742.04 -3248.69

2 150 -3518.13 86.7715 -3787.83 -3399.29

3 128 -3567.53 80.8665 -3817.71 -3461.67

4 134 -3618.16 74.6219 -3817.71 -3501.87

5 138 -3667.39 68.9648 -3837.77 -3567.23

6 150 -3723.95 58.591 -3849.71 -3627.34

7 114 -3751.28 50.69 -3870.71 -3666.36

8 143 -3788.55 30.3672 -3900.34 -3726.72

9 135 -3799.3 24.903 -3900.34 -3759.08

Evaluated 1600 models in total in 0:23:41.537673
Best fitness is (-3900.3446673012127,)
Best parameters are [58, 7.102697584971639, 175.2853863939617, 119.35820862776255, -2.5695363780145524]
Running register c

gen evals avg std min max

0 140 -3050.34 507.923 -3688.22 -1038.88

1 135 -3368.35 109.699 -3688.22 -3185.81

2 148 -3455.27 91.5412 -3755.8 -3319.55

3 135 -3502.76 81.4525 -3755.8 -3378.69

4 153 -3563.43 67.781 -3763.12 -3463.11

5 133 -3602.94 59.5575 -3770.56 -3517.39

6 147 -3647.64 52.5706 -3786.68 -3568.07

7 136 -3680.58 46.5367 -3786.68 -3606.8

8 125 -3712.95 39.5501 -3790.28 -3645.56

9 124 -3738.9 29.6493 -3798.65 -3682.53

Evaluated 1576 models in total in 0:26:05.622464
Best fitness is (-3798.653959002797,)
Best parameters are [58, 7.484645156417415, 268.2359381812384, 245.08887941102196, -5.3570287613318825]
Running register d

gen evals avg std min max

0 162 -3264.02 356.77 -3882.84 -2117.09

1 160 -3527.18 145.559 -3882.84 -3321.47

2 127 -3616.33 122.038 -3882.99 -3429.67

3 150 -3706.95 90.5299 -3882.99 -3552.69

4 160 -3763.03 66.5525 -3902.79 -3639.08

5 145 -3798.46 55.3168 -3910.85 -3706.24

6 127 -3822.52 47.575 -3910.85 -3740.88

7 136 -3848.94 38.1135 -3912.41 -3774.75

8 134 -3868.12 29.0952 -3917.79 -3808.39

9 153 -3888.18 15.4235 -3917.79 -3853.01

Evaluated 1654 models in total in 0:17:24.889920
Best fitness is (-3917.7887333047615,)
Best parameters are [58, 7.241131619736361, 204.2423421908887, 321.57594042861035, -7.9439555584872945]
Running register e

gen evals avg std min max

0 138 -2930.39 661.521 -3652.72 -812.987

1 136 -3378.97 97.4546 -3713.46 -3200.46

2 129 -3441.55 78.4761 -3713.46 -3331.56

3 154 -3493.92 74.1087 -3713.46 -3394.82

4 139 -3537.02 70.9353 -3753.5 -3442.85

5 137 -3572.74 65.3842 -3753.5 -3470.38

6 138 -3609.49 52.5023 -3753.5 -3523.2

7 144 -3633.69 47.5068 -3788.34 -3560.56

8 143 -3659.28 47.203 -3839.29 -3592.48

9 142 -3682.83 47.0983 -3839.29 -3618.6

Evaluated 1600 models in total in 0:19:58.305442
Best fitness is (-3839.2855530606503,)
Best parameters are [58, 7.310261424743253, 172.03377045441763, 79.64950003346735, 7.346747568497819]
Running register f

gen evals avg std min max

0 153 -3106.61 427.179 -3620.12 -1642.96

1 139 -3365.91 87.0832 -3643.52 -3219.84

2 142 -3425.76 73.5373 -3673.59 -3321.01

3 127 -3452.75 67.4847 -3733 -3366.43

4 146 -3484.36 64.7304 -3733 -3401.99

5 149 -3509.17 61.9016 -3733 -3426.87

6 154 -3544.25 69.7285 -3733 -3458.11

7 140 -3571.65 74.3971 -3755.54 -3476.62

8 159 -3620.38 75.9548 -3761.99 -3513.03

9 143 -3688.46 54.3722 -3771.27 -3559.48

Evaluated 1652 models in total in 0:22:59.706454
Best fitness is (-3771.269584046744,)
Best parameters are [58, 7.338510921788755, 342.22583920047936, 172.37091140673826, -1.1762254072447065]
Running register g

gen evals avg std min max

0 139 -2898.91 756.109 -3747.07 -358.958

1 146 -3397.72 135.15 -3747.07 -3156.02

2 136 -3502.59 99.6004 -3747.39 -3349.8

3 148 -3569.18 74.6606 -3765.92 -3451.68

4 135 -3604.39 63.6616 -3798.25 -3498.27

5 139 -3633.64 55.3028 -3798.25 -3544.98

6 143 -3668.35 54.5296 -3811.79 -3583.74

7 123 -3697.71 49.4541 -3842.53 -3628.41

8 142 -3732.98 45.051 -3857.88 -3665.31

9 153 -3755.36 38.6372 -3857.88 -3694.01

Evaluated 1604 models in total in 0:24:44.346691

Best fitness is (-3857.8766174149996,)

Best parameters are [58, 7.473104462727869, 179.74898607307537, 265.14280405248894, -7.351178560022423]

DE

Running register a

gen evals avg std min max

0 200 -1885.89 2542.68 -3603.87 11561

1 200 -2941.35 1030.44 -3628.95 2982.71

2 200 -3300.15 340.842 -3744.61 -1129.58

3 200 -3427.91 141.953 -3744.61 -2853.57

4 200 -3480.27 113.093 -3744.61 -2977.79

5 200 -3511.07 97.855 -3744.61 -3232.59

6 200 -3537.53 84.6602 -3744.61 -3289.43

7 200 -3564.91 75.4809 -3744.61 -3300.96

8 200 -3583.42 68.7065 -3744.61 -3300.96

9 200 -3604.51 53.2817 -3744.61 -3405.18

Evaluated 2000 models in total in 0:28:57.406535

Best fitness is (-3744.6144996875178,)

Best parameters are [58, 7.156509628795848, 337.7761165894732, 17.078860138679552, -2.5967621705941077]

Running register b

gen evals avg std min max

0 200 -1534.61 2797.94 -3667.23 12806.5

1 200 -2799.84 1163.15 -3676.85 4368.91

2 200 -3256.62 375.347 -3796.95 -906.408

3 200 -3380.77 143.621 -3796.95 -2438.33

4 200 -3432.45 115.165 -3796.95 -3088.43

5 200 -3469.09 106.619 -3796.95 -3214.09

6 200 -3502.76 99.5192 -3796.95 -3286.54

7 200 -3528.55 88.8683 -3796.95 -3296.46

8 200 -3549.67 84.4618 -3796.95 -3296.46

9 200 -3575.34 80.6463 -3841.22 -3353.27

Evaluated 2000 models in total in 0:28:26.706560

Best fitness is (-3841.223787636084,)

Best parameters are [58, 7.245054672982067, 156.381639227941, 117.75828266624534, -2.3117302044474424]

Running register c

gen evals avg std min max

0 200 -1489.97 3900.93 -3642.86 22683.8

1 200 -2919.36 1073.96 -3677.07 5533.4

2 200 -3297.47 297.337 -3677.07 -1492.38

3 200 -3378.89 210.354 -3677.07 -1492.38

4 200 -3439.78 123.526 -3688.71 -2710.56

5 200 -3472.92 99.839 -3702.38 -2795.13

6 200 -3502.71 84.5305 -3734.4 -3255.35

7 200 -3521.92 80.1384 -3734.4 -3255.35

8 200 -3539.71 74.9963 -3734.4 -3255.35

9 200 -3561.2 67.8441 -3734.4 -3264.34

Evaluated 2000 models in total in 0:32:34.964323
Best fitness is (-3734.401095566741,)
Best parameters are [58, 7.452799152965777, 349.16455894113085, 235.69451122579076, -4.661327703505561]
Warning! Parameter 2 is at or near maximum allowed value
Running register d

gen evals avg std min max

0 200 -1593.89 3003.01 -3755.39 12565.2

1 200 -2904.16 967.207 -3803.9 1011.51

2 200 -3304.56 466.186 -3803.9 -408.398

3 200 -3468.36 221.12 -3854.39 -2264.66

4 200 -3545.04 145.995 -3854.39 -2588.84

5 200 -3586.14 135.052 -3854.39 -2588.84

6 200 -3626.24 108.095 -3876.13 -3134.68

7 200 -3665.49 98.3419 -3894.68 -3347.47

8 200 -3688.75 91.79 -3894.68 -3428.55

9 200 -3710.94 85.6719 -3894.68 -3428.55

Evaluated 2000 models in total in 0:23:39.819206
Best fitness is (-3894.68080739762,)
Best parameters are [58, 6.945549866786462, 308.70325120635215, 324.40714106106117, -9.017463394116593]
Running register e

gen evals avg std min max

0 200 -1297.76 3286.32 -3614.15 17147.8

1 200 -2663.23 1582.75 -3700.41 10506.6

2 200 -3138.9 662.315 -3700.41 901.558

3 200 -3345.52 327.534 -3734.83 -1167.83

4 200 -3443.06 120.633 -3742.65 -2909.19

5 200 -3478.42 109.53 -3742.65 -2909.19

6 200 -3507.69 94.6422 -3762.17 -3214

7 200 -3527.87 86.8956 -3762.17 -3240.37

8 200 -3547.37 87.234 -3773.9 -3240.37

9 200 -3572.19 84.1563 -3773.9 -3284.52

Evaluated 2000 models in total in 0:29:03.093330
Best fitness is (-3773.9034999339106,)
Best parameters are [58, 7.480161522054298, 315.3719826142643, 102.7364772897914, -0.8452874512086186]
Running register f

gen evals avg std min max

0 200 -1704.64 2801.83 -3631.5 17541.5

1 200 -2853.92 1295.15 -3631.5 5616.34

2 200 -3283.45 304.757 -3631.5 -1533.04

3 200 -3385.32 146.716 -3711.45 -2713.5

4 200 -3428.53 112.15 -3711.45 -2787.02

5 200 -3454.32 88.052 -3711.45 -3151.09

6 200 -3474.37 81.3599 -3711.45 -3151.09

7 200 -3494.97 73.7719 -3711.45 -3254.65

8 200 -3511.98 69.4625 -3772.26 -3284.93

9 200 -3527.16 66.3544 -3772.26 -3342.06

Evaluated 2000 models in total in 0:31:00.346490
Best fitness is (-3772.264125423857,)
Best parameters are [58, 7.303757093051794, 266.57331722308044, 157.0297825730657, -2.3683466964730537]
Running register g

gen evals avg std min max

0 200 -1538.45 3029.55 -3732.86 20613.2

1 200 -2774.67 1228.08 -3745.4 3983.18

2 200 -3202.44 617.378 -3745.4 216.339

3 200 -3434.4 183.058 -3745.4 -2763.06

4 200 -3492.48 149.006 -3755.92 -2763.06

5 200 -3529.63 122.905 -3755.92 -2912.87

6 200 -3561.16 109.834 -3755.92 -2912.87

7 200 -3592.26 80.8393 -3855.91 -3364.52

8 200 -3609.23 74.7516 -3855.91 -3364.52

9 200 -3631.85 69.9729 -3855.91 -3364.52

Evaluated 2000 models in total in 0:22:35.580805
Best fitness is (-3855.907402613946,)
Best parameters are [58, 7.681792862768369, 282.87138746869, 295.21774257585025, -10.84957084116523]

Scripts

Script 1

The inversion of the RM6 sequence

```
#!/usr/bin/env python3

sequence_rRM6 = "MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLARFGDDGENL"[::-1]
print(sequence_rRM6)
```

Script 2

Grid scan -tetramer example

```
#!/usr/bin/env python3
import numpy as np
import budeff
import isambard.specifications as specifications
import isambard.modelling as modelling
import itertools

class APHomoTetramer(specifications.CoiledCoil):
        oligomeric_state = 4
        def __init__(self, helix_length, radius, pitch, ideal_phica, zshift):
                super().__init__(self.oligomeric_state, auto_build = False)
                self.aas = [helix_length, helix_length, helix_length, helix_length]
                self.major_radii = [radius, radius, radius, radius]
                self.major_pitches = [pitch, pitch, pitch, pitch]
                self.z_shifts = [0, zshift, zshift, zshift]
                self.phi_c_alphas = [ideal_phica, ideal_phica, ideal_phica, ideal_phica]
                # self.major_handedness = ['r', 'r', 'r', 'r'] only for right all- and anri- parallel structures
                self.orientations = [1, -1, 1, -1]
                self.build()

ideal_phica_for_register = {
   "a": 25.714285714285715,
   "b": 128.57142857142856,
   "c": 231.42857142857144,
   "d": 334.2857142857143,
   "e": 77.14285714285714,
   "f": 180.0,
   "g": 282.85714285714283,
}

# CoiledCoil.from_parameters

radii = np.arange(8.4, 8.61, 0.2) # min max step
interface_angles = np.arange(-30, 30.1, 2) # min max step
major_pitches = np.arange(50, 350.1, 10) # min max step
z_shifts = np.arange(-10, 10.1, 1)


def build_tetramer(radius, interface_angle, major_pitch, zshift):
   sequences = [
   "MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLARFGDDGENL",
```

```
   "MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLARFGDDGENL",
   "MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLARFGDDGENL",
   "MTKQEKTALNMARFIRSQTLTLLEKLNELADICESLHDHADELYRSCLARFGDDGENL"]

   gs_tetramer = specifications.CoiledCoil.from_parameters(4, 58, radius, major_pitch, ideal_phica_for_register['d']
+interface_angle)
   gs_tetramer = modelling.pack_side_chains_scwrl(gs_tetramer, sequences)
   return gs_tetramer

#iterating
for register, ideal_phica in ideal_phica_for_register.items():
        for i, radius in enumerate(radii):
                for j, interface_angle in enumerate(interface_angles):
                        for m, major_pitch in enumerate(major_pitches):
                                for z, zshift in enumerate(z_shifts):
                                        tetramer_model= build_tetramer(radius, interface_angle, major_pitch, zshift)
                                        results = budeff.get_internal_energy(tetramer_model).total_energy
                                        print(register, radius, interface_angle, major_pitch, zshift, results )
```

Script 3

Example of the CMA -ES script used for the modeling of a tetrameric HB

```
#!/usr/bin/env python3

import sys
import warnings
warnings.simplefilter("ignore")

import isambard.specifications as specifications
import isambard.modelling as modelling
import isambard.optimisation
import budeff
import isambard.optimisation.evo_optimizers as ev_opts
from isambard.optimisation.evo_optimizers import Parameter

def get_buff_total_energy(ampal_object):
   return budeff.get_internal_energy(ampal_object).total_energy


class APSwitchRegistry(specifications.CoiledCoil):
   """ Specification for creating anti-parallel coiled coils with switching registry"""
   oligomeric_state = 4
   def __init__(self, helix_length, radius, pitch, phica, zshift):
      super().__init__(self.oligomeric_state, auto_build=False)
      self.aas = [helix_length, helix_length, helix_length, helix_length]
      self.major_radii = [radius, radius, radius, radius]
      self.major_pitches = [pitch, pitch, pitch, pitch]
      self.major_handedness = ['r', 'r', 'r', 'r']
      self.phi_c_alphas = [phica, phica, phica, phica]
      self.z_shifts = [0, zshift, 0, zshift]
      self.orientations = [1, -1 , 1, -1]
      self.build()

sequences = [
```

```
    'MLNEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM',
    'MLNEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM',
    'MLNEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM',
    'MLNEGDDGFRALCSRYLEDAHDHLSECIDALENLKELLTLTQSRIFRAMNLATKEQKTM'
]

parameters = [
        Parameter.static('Helix Length', 59),
        Parameter.dynamic('Radius', 7.0, 2.0),
        Parameter.dynamic('Pitch', 200, 150),
        Parameter.dynamic('Phi_CA', 180.0, 27),
        Parameter.dynamic('z-shift', -18.0, 0.0)
    ]
opt_cmaes = ev_opts.CMAES(APSwitchRegistry, sequences, parameters, get_buff_total_energy)
opt_cmaes.run_opt(500, 10, cores=12)
optimized_model = opt_cmaes.best_model
with open('18_right_anti-parallel_rRM6.pdb', 'w') as f:
            print(optimized_model.pdb, file=f)
```

References

[1] Hartmann MD. Functional and Structural Roles of Coiled Coils. Subcell Biochem 2017;82:63–93. https://doi.org/10.1007/978-3-319-49674-0_3.

[2] Burkhard P, Stetefeld J, Strelkov SV. Coiled coils: a highly versatile protein folding motif. Trends Cell Biol 2001;11:82–8. https://doi.org/10.1016/S0962-8924(00)01898-5.

[3] Crick FHC. The Fourier transform of a coiled-coil. Acta Crystallogr 1953;6:685–9. https://doi.org/10.1107/S0365110X53001952.

[4] Crick FHC. The packing of α-helices: simple coiled-coils. Acta Crystallogr 1953;6:689–97. https://doi.org/10.1107/S0365110X53001964.

[5] Lupas AN, Bassler J, Dunin-Horkawicz S. The Structure and Topology of α-Helical Coiled Coils. In: Parry DAD, Squire JM, editors. Fibrous Proteins Struct. Mech., vol. 82, Cham: Springer International Publishing; 2017, p. 95–129. https://doi.org/10.1007/978-3-319-49674-0_4.

[6] Lupas AN, Gruber M. THE STRUCTURE OF a-HELICAL COILED COILS n.d.:42.

[7] Pratap JV, Luisi BF, Calladine CR. Geometric principles in the assembly of α-helical bundles. Philos Trans R Soc Math Phys Eng Sci 2013;371:20120369. https://doi.org/10.1098/rsta.2012.0369.

[8] Parry DAD, Fraser RDB, Squire JM. Fifty years of coiled-coils and α-helical bundles: A close relationship between sequence and structure. J Struct Biol 2008;163:258–69. https://doi.org/10.1016/j.jsb.2008.01.016.

[9] Chothia C, LEVITTt M, Richardson D. Structure of proteins: Packing of a-helices and pleated sheets n.d.:5.

[10] Banner DW, Kokkinidis M, Tsernoglou D. Structure of the ColE1 rop protein at 1.7 A resolution. J Mol Biol 1987;196:657–75. https://doi.org/10.1016/0022-2836(87)90039-8.

[11]X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/1948029/ (accessed December 23, 2021).
[12] Grigoryan G, DeGrado WF. Probing Designability via a Generalized Model of Helical Bundle Geometry. J Mol Biol 2011;405:1079–100. https://doi.org/10.1016/j.jmb.2010.08.058.
[13] Walshaw J, Woolfson DN. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. J Struct Biol 2003;144:349–61. https://doi.org/10.1016/j.jsb.2003.10.014.
[14] Wood CW, Heal JW, Thomson AR, Bartlett GJ, Ibarra AÁ, Brady RL, et al. ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. Bioinformatics 2017;33:3043–50. https://doi.org/10.1093/bioinformatics/btx352.
[15] Wood CW, Bruning M, Ibarra AA, Bartlett GJ, Thomson AR, Sessions RB, et al. CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. Bioinformatics 2014;30:3029–35. https://doi.org/10.1093/bioinformatics/btu502.
[16] Walshaw J, Woolfson DN. SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures11Edited by J. Thornton. J Mol Biol 2001;307:1427–50. https://doi.org/10.1006/jmbi.2001.4545.
[17] Strelkov SV, Burkhard P. Analysis of alpha-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation. J Struct Biol 2002;137:54–64. https://doi.org/10.1006/jsbi.2002.4454.
[18] Dunin-Horkawicz S, Lupas AN. Measuring the conformational space of square four-helical bundles with the program samCC. J Struct Biol 2010;170:226–35. https://doi.org/10.1016/j.jsb.2010.01.023.
[19] Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77:778–95. https://doi.org/10.1002/prot.22488.
[20] McIntosh-Smith S, Wilson T, Ibarra AA, Crisp J, Sessions RB. Benchmarking Energy Efficiency, Power Costs and Carbon Emissions on Heterogeneous Systems. Comput J 2012;55:192–205. https://doi.org/10.1093/comjnl/bxr091.
[21] Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I. Probing the “Dark Matter” of Protein Fold Space. Structure 2009;17:1244–52. https://doi.org/10.1016/j.str.2009.07.012.
[22] M. Dawson W, O. Martin FJ, G. Rhys G, L. Shelley K, Leo Brady R, N. Woolfson D. Coiled coils 9-to-5: rational de novo design of α-helical barrels with tunable oligomeric states. Chem Sci 2021;12:6923–8. https://doi.org/10.1039/D1SC00460C.
[23] Sahab MG, Toropov VV, Gandomi AH. 2 - A Review on Traditional and Modern Structural Optimization: Problems and Techniques. In: Gandomi AH, Yang X-S, Talatahari S, Alavi AH, editors. Metaheuristic Appl. Struct. Infrastruct., Oxford: Elsevier; 2013, p. 25–47. https://doi.org/10.1016/B978-0-12-398364-0.00002-4.
[24] Bianchi L, Dorigo M, Gambardella LM, Gutjahr WJ. A survey on metaheuristics for stochastic combinatorial optimization. Nat Comput 2009;8:239–87. https://doi.org/10.1007/s11047-008-9098-4.

[25] Gulzat T, Lyazat N, Siládi V, Sembina G, Maksatbek S. Research on predictive model based on classification with parameters of optimization. Neural Netw World 2020;30:295–308. https://doi.org/10.14311/NNW.2020.30.020.
[26] Storn R. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. Differ Evol n.d.:19.
[27] Hertz A, Kobler D. A framework for the description of evolutionary algorithms. Eur J Oper Res 2000;126:1–12. https://doi.org/10.1016/S0377-2217(99)00435-X.
[28] Roeva O, Fidanova S, Paprzycki M. Influence of the Population Size on the Genetic Algorithm Performance in Case of Cultivation Process Modelling 2013:6.
[29] Whitley D. A genetic algorithm tutorial. Stat Comput 1994;4. https://doi.org/10.1007/BF00175354.
[30] Fortin F-A. DEAP: Evolutionary Algorithms Made Easy n.d.:5.
[31] Hansen N, Müller SD, Koumoutsakos P. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). Evol Comput 2003;11:1–18. https://doi.org/10.1162/106365603321828970.
[32] Gagganapalli SR. Implementation and Evaluation of CMA-ES Algorithm 2015.
[33] Hudaib A, Al Hwaitat A. Movement Particle Swarm Optimization Algorithm. Mod Appl Sci 2017;12:148. https://doi.org/10.5539/mas.v12n1p148.
[34] Hansen N, Ostermeier A. Completely Derandomized Self-Adaptation in Evolution Strategies. Evol Comput 2001;9:159–95. https://doi.org/10.1162/106365601750190398.
[35] Hansen N. The CMA Evolution Strategy: A Tutorial. ArXiv160400772 Cs Stat 2016.
[36] Dufour J-M, Neves J. Chapter 1 - Finite-sample inference and nonstandard asymptotics with Monte Carlo tests and R. In: Vinod HD, Rao CR, editors. Handb. Stat., vol. 41, Elsevier; 2019, p. 3–31. https://doi.org/10.1016/bs.host.2019.05.001.
[37] Amprazi M, Kotsifaki D, Providaki M, Kapetaniou EG, Fellas G, Kyriazidis I, et al. Structural plasticity of 4- -helical bundles exemplified by the puzzle-like molecular assembly of the Rop protein. Proc Natl Acad Sci 2014;111:11049–54. https://doi.org/10.1073/pnas.1322065111.
[38] Castagnoli L, Scarpa M, Kokkinidis M, Banner DW, Tsernoglou D, Cesareni G. Genetic and structural analysis of the ColE1 Rop (Rom) protein. EMBO J 1989;8:621–9.
[39] Som T, Tomizawa J. Regulatory regions of ColE1 that are involved in determination of plasmid copy number. Proc Natl Acad Sci U S A 1983;80:3232–6. https://doi.org/10.1073/pnas.80.11.3232.
[40] Willis MA, Bishop B, Regan L, Brunger AT. Dramatic Structural and Thermodynamic Consequences of Repacking a Protein's Hydrophobic Core. Structure 2000;8:1319–28. https://doi.org/10.1016/S0969-2126(00)00544-X.
[41] Glykos NM, Papanikolau Y, Vlassi M, Kotsifaki D, Cesareni G, Kokkinidis M. Loopless Rop: Structure and Dynamics of an Engineered Homotetrameric Variant of the Repressor of Primer Protein. Biochemistry 2006;45:10905–19. https://doi.org/10.1021/bi060833n.
[42] Kokkinidis M, Vlassi M, Papanikolaou Y, Kotsifaki D, Kingswell A, Tsernoglou D, et al. Correlation between protein stability and crystal properties of designed ROP variants. Proteins 1993;16:214–6. https://doi.org/10.1002/prot.340160208.
[43] Vlassi M, Steif C, Weber P, Tsernoglou D, Wilson KS, Hinz HJ, et al. Restored heptad pattern continuity does not alter the folding of a four-alpha-helix bundle. Nat Struct Biol 1994;1:706–16. https://doi.org/10.1038/nsb1094-706.

[44] Glykos NM, Cesareni G, Kokkinidis M. Protein plasticity to the extreme: changing the topology of a 4-alpha-helical bundle with a single amino acid substitution. Struct Lond Engl 1993 1999;7:597–603. https://doi.org/10.1016/s0969-2126(99)80081-1.
[45] Kamtekar S, Hecht MH. Protein Motifs. 7. The four-helix bundle: what determines a fold? FASEB J Off Publ Fed Am Soc Exp Biol 1995;9:1013–22. https://doi.org/10.1096/fasebj.9.11.7649401.
[46] Travaglini-Allocatelli C, Ivarsson Y, Jemth P, Gianni S. Folding and stability of globular proteins and implications for function. Curr Opin Struct Biol 2009;19:3–7. https://doi.org/10.1016/j.sbi.2008.12.001.
[47] Finkelstein AV. 50+ Years of Protein Folding. Biochem Mosc 2018;83:S3–18. https://doi.org/10.1134/S000629791814002X.
[48] Hill R.B., Raleigh D.P., Lombardi A., Degrado W.F. De Novo Design of Helical Bundles as Models for Understanding Protein Folding and Function. Acc Chem Res 2000;33:745–54.
[49] Newberry RW, Raines RT. Secondary Forces in Protein Folding. ACS Chem Biol 2019;14:1677–86. https://doi.org/10.1021/acschembio.9b00339.
[50] Kamtekar S, Hecht MH. The four-lielix bundle: what determines a fold? FASEB J 1995;9:1013–22. https://doi.org/10.1096/fasebj.9.11.7649401.
[51] Zhang Y, Weber JK, Zhou R. Folding and Stabilization of Native-Sequence-Reversed Proteins. Sci Rep 2016;6. https://doi.org/10.1038/srep25138.
[52] Ahmed S, Shukla A, Guptasarma P. Folding behavior of a backbone-reversed protein: Reversible polyproline type II to β-sheet thermal transitions in retro-GroES multimers with GroES-like features. Biochim Biophys Acta BBA - Proteins Proteomics 2008;1784:916–23. https://doi.org/10.1016/j.bbapap.2008.02.009.
[53] Kefala A, Amprazi M, Mylonas E, Kotsifaki D, Providaki M, Pozidis C, et al. Probing Protein Folding with Sequence-Reversed α-Helical Bundles. Int J Mol Sci 2021;22:1955. https://doi.org/10.3390/ijms22041955.
[54] Zerze GH, Stillinger FH, Debenedetti PG. Computational investigation of retro-isomer equilibrium structures: Intrinsically disordered, foldable, and cyclic peptides. FEBS Lett 2020;594:104–13. https://doi.org/10.1002/1873-3468.13558.
[55] Lorenzen S, Gille C, Preissner R, Frömmel C. Inverse sequence similarity of proteins does not imply structural similarity. FEBS Lett 2003;545:105–9. https://doi.org/10.1016/S0014-5793(03)00450-2.
[56] Kutyshenko VP, Prokhorov DA, Molochkov NV, Sharapov MG, Kolesnikov I, Uversky VN. Dancing retro: solution structure and micelle interactions of the retro-SH3-domain, retro-SHH-‘Bergerac.’ J Biomol Struct Dyn 2014;32:257–72. https://doi.org/10.1080/07391102.2012.762724.
[57] Mittl PRE, Deillon C, Sargent D, Liu N, Klauser S, Thomas RM, et al. The retro-GCN4 leucine zipper sequence forms a stable three-dimensional structure. Proc Natl Acad Sci 2000;97:2562–6. https://doi.org/10.1073/pnas.97.6.2562.
[58] Lacroix E, Viguera AR, Serrano L. Reading protein sequences backwards. Fold Des 1998;3:79–85. https://doi.org/10.1016/S1359-0278(98)00013-3.
[59] Sabaté R, Espargaró A, de Groot NS, Valle-Delgado JJ, Fernàndez-Busquets X, Ventura S. The Role of Protein Sequence and Amino Acid Composition in Amyloid Formation:

Scrambling and Backward Reading of IAPP Amyloid Fibrils. J Mol Biol 2010;404:337–52. https://doi.org/10.1016/j.jmb.2010.09.052.
[60] Uversky VN. What does it mean to be natively unfolded? Eur J Biochem 2002;269:2–12. https://doi.org/10.1046/j.0014-2956.2001.02649.x.
[61] Pan PK, Zheng ZF, Lyu PC, Huang PC. Why reversing the sequence of the alpha domain of human metallothionein-2 does not change its metal-binding and folding characteristics. Eur J Biochem 1999;266:33–9. https://doi.org/10.1046/j.1432-1327.1999.00811.x.
[62] Chandler D. Interfaces and the driving force of hydrophobic assembly. Nature 2005;437:640–7. https://doi.org/10.1038/nature04162.
[63] Dill KA, MacCallum JL. The Protein-Folding Problem, 50 Years On. Science 2012.
[64] Does a backwardly read protein sequence have a unique native state? - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/9053902/ (accessed January 6, 2022).
[65] Folding behaviour of retro-rubredoxin n.d. https://www.researchgate.net/publication/285907653_Folding_behaviour_of_retro-rubredoxin (accessed January 10, 2022).
[66] Milton RC, Milton SC, Kent SB. Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity [corrected]. Science 1992;256:1445–8. https://doi.org/10.1126/science.1604320.
[67] Guptasarma P. Reversal of peptide backbone direction may result in the mirroring of protein structure. FEBS Lett 1992;310:205–10. https://doi.org/10.1016/0014-5793(92)81333-H.
[68] Olszewski KA, Kolinski A, Skolnick J. Does a backwardly read protein sequence have a unique native state? Protein Eng 1996;9:5–14. https://doi.org/10.1093/protein/9.1.5.
[69] Witte K, Skolnick J, Wong C-H. A Synthetic Retrotransition (Backward Reading) Sequence of the Right-Handed Three-Helix Bundle Domain (10-53) of Protein A Shows Similarity in Confomation as Predicted by Computation. J Am Chem Soc - J AM CHEM SOC 1998;120. https://doi.org/10.1021/ja982203h.
[70] Baker D. What has de novo protein design taught us about protein folding and biophysics? Protein Sci 2019;28:678–83. https://doi.org/10.1002/pro.3588.
[71] Khoury GA, Smadbeck J, Kieslich CA, Floudas CA. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol 2014;32:99–109. https://doi.org/10.1016/j.tibtech.2013.10.008.
[72] Boyle AL, Woolfson DN. De novo designed peptides for biological applications. Chem Soc Rev 2011;40:4295–306. https://doi.org/10.1039/C0CS00152J.
[73] Fiser A. Template-Based Protein Structure Modeling. Methods Mol Biol Clifton NJ 2010;673:73–94. https://doi.org/10.1007/978-1-60761-842-3_6.
[74] Kondabala R, Kumar V. Computational Intelligence Tools for Protein Modeling. In: Yadav N, Yadav A, Bansal JC, Deep K, Kim JH, editors. Harmony Search Nat. Inspired Optim. Algorithms, Singapore: Springer; 2019, p. 949–56. https://doi.org/10.1007/978-981-13-0761-4_89.
[75] Szczepaniak K, Ludwiczak J, Winski A, Dunin-Horkawicz S. Variability of the core geometry in parallel coiled-coil bundles. J Struct Biol 2018;204:117–24. https://doi.org/10.1016/j.jsb.2018.07.002.
[76] Lapenta F, Aupič J, Vezzoli M, Strmšek Ž, Da Vela S, Svergun DI, et al. Self-assembly and regulation of protein cages from pre-organised coiled-coil modules. Nat Commun 2021;12:939. https://doi.org/10.1038/s41467-021-21184-6.

[77] Burgess NC, Sharp TH, Thomas F, Wood CW, Thomson AR, Zaccai NR, et al. Modular Design of Self-Assembling Peptide-Based Nanotubes. J Am Chem Soc 2015;137:10554–62. https://doi.org/10.1021/jacs.5b03973.
[78] Turgay Y, Eibauer M, Goldman AE, Shimi T, Khayat M, Ben-Harush K, et al. The molecular architecture of lamins in somatic cells. Nature 2017;543:261–4. https://doi.org/10.1038/nature21382.
[79] Burton AJ, Thomson AR, Dawson WM, Brady RL, Woolfson DN. Installing hydrolytic activity into a completely de novo protein framework. Nat Chem 2016;8:837–44. https://doi.org/10.1038/nchem.2555.
[80] Bhamidimarri SP, Zahn M, Prajapati JD, Schleberger C, Söderholm S, Hoover J, et al. A Multidisciplinary Approach toward Identification of Antibiotic Scaffolds for Acinetobacter baumannii. Structure 2019;27:268-280.e6. https://doi.org/10.1016/j.str.2018.10.021.
[81] Guzenko D, Chernyatina AA, Strelkov SV. Crystallographic Studies of Intermediate Filament Proteins. In: Parry DAD, Squire JM, editors. Fibrous Proteins Struct. Mech., Cham: Springer International Publishing; 2017, p. 151–70. https://doi.org/10.1007/978-3-319-49674-0_6.
[82] Rossmann MG. The molecular replacement method. Acta Crystallogr A 1990;46:73–82. https://doi.org/10.1107/S0108767389009815.
[83] Kefala A, Kotsifaki D, Providaki M, Amprazi M, Kokkinidis M. Expression, purification and crystallization of a protein resulting from the inversion of the amino-acid sequence of a helical bundle. Acta Crystallogr Sect F Struct Biol Commun 2017;73:51–3. https://doi.org/10.1107/S2053230X16020173.
[84] Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinforma Oxf Engl 2019;35:2790–5. https://doi.org/10.1093/bioinformatics/bty1062.
[85] Schrödinger, LLC. The AxPyMOL Molecular Graphics Plugin for Microsoft PowerPoint, Version 1.8 2015.
[86] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet TIG 2000;16:276–7. https://doi.org/10.1016/s0168-9525(00)02024-2.
[87] NetWheels: A web application to create high quality peptide helical wheel and net projections | bioRxiv n.d. https://www.biorxiv.org/content/10.1101/416347v1 (accessed January 16, 2022).
[88] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–12. https://doi.org/10.1002/jcc.20084.
[89] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14:33–8, 27–8. https://doi.org/10.1016/0263-7855(96)00018-5.
[90] Turner: XMGRACE, Version 5.1. 19 - Μελετητής Google n.d. https://scholar.google.com/scholar_lookup?title=XMGRACE, +Version+5.1.19&author=PJ+Turner&publication_year=2005& (accessed January 12, 2022).
[91] Heo L, Lee H, Seok C. GalaxyRefineComplex: Refinement of protein-protein complex model structures driven by interface repacking. Sci Rep 2016;6:32153. https://doi.org/10.1038/srep32153.

[92] Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res 2009;37:e83. https://doi.org/10.1093/nar/gkp318.
[93] Effect of Chain Length on the Formation and Stability of Synthetic .alpha.-Helical Coiled Coils | Biochemistry n.d. https://pubs.acs.org/doi/10.1021/bi00255a032 (accessed February 2, 2022).
[94] Levy Y, Cho SS, Shen T, Onuchic JN, Wolynes PG. Symmetry and frustration in protein energy landscapes: A near degeneracy resolves the Rop dimer-folding mystery. Proc Natl Acad Sci 2005;102:2373–8. https://doi.org/10.1073/pnas.0409572102.
[95] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–94. https://doi.org/10.1006/jmbi.1998.1645.
[96] Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci 2000;97:1525–9. https://doi.org/10.1073/pnas.97.4.1525.
[97] Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol 1995;252:709–20. https://doi.org/10.1006/jmbi.1995.0529.
[98] Lee J, Freddolino PL, Zhang Y. Ab Initio Protein Structure Prediction. In: J. Rigden D, editor. Protein Struct. Funct. Bioinforma., Dordrecht: Springer Netherlands; 2017, p. 3–35. https://doi.org/10.1007/978-94-024-1069-3_1