©2017

Sandeep Vishwanath Belure

ALL RIGHTS RESERVED

CHARACTERIZING COLLAGEN MIMITIC PEPTIDES FOR ORTHOGONAL SELF-ASSEMBLY

By

SANDEEP VISHWANATH BELURE

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Vikas Nanda

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January 2017

ABSTRACT OF THE DISSERTATION

Characterizing Collagen Mimetic Peptides For Orthogonal Self-Assembly

By SANDEEP VISHWANATH BELURE

Dissertation Director:

Vikas Nanda, Ph.D.

A computational design of collagen mimetic peptides (CMPs) that self-assemble orthogonally (mutually exclusively), in the presence of other pre-existing collagen trimer mixtures, *in vitro*, has been proposed. The orthogonality in self-assembly was brought about by orthogonal patterning of ionic salt bridges and residues, along the collagen trimers' axial length. Through the aid of circular dichroism spectroscopy alone, a novel experimental protocol was set-up to rapidly assess the level of cross-talk that may arise in such designed 'heterogeneous monomer to trimer folding' mixture environments. It is shown that the designed collagen mimetic peptides are stable and hetero-specific within their composite 3 chain peptide ecosystem. We experimentally demonstrate the extent to which loss in specificity could possibly occur, upon moving to a higher order 'more than 3 monomers in solution' peptide ensemble. Although the desired level of multi-state orthogonality was not achieved in the current design, the experimental results obtained were used to estimate the stability and specificity barrier threshold that one might run into, if one were to instead design orthogonal systems where-in specificity is incorporated during the computational design stage itself *a priori*. A Pareto frontier plot indicating the

specificity versus stability trade-off is plotted. We conclude that a bottom-up design approach, incorporating design of specificity during the sequence design stage, would be a better way forward for achieving self-assembling orthogonality. In contrast to the complex chaperone assisted protein folding systems existing in nature, our method is a simplistic first step towards the complementary approach of modular synthetic collagen molecule design.

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Dr. Vikas Nanda, for his patient guidance throughout my graduate program. Discussions with him have always been fruitful and his encouragement to try out projects with a riskier component has usually paid off for me in the long run. The freedom to persevere and try out unconventional, creative and highly adaptable approaches under his guidance has shown me how a problem in science may not necessarily be approached in a singular way. This has been probably one of the most valuable lessons that I have learnt under his guidance.

I would also like to extend my gratitude to the other members of my thesis committee Dr. Debashish Bhattacharya, Dr. Joseph Marcotrigiano and Dr. David Shreiber for their valuable time, feedback and discussions. Dr Eddy Arnold has been a valuable guide during my first year of graduate school.

I would like to take this opportunity to also thank my supportive former and current lab members Dr. Avanish Parmar, Dr. Fei Xu, Dr. I. John Khan, Dr. Jim Stapleton, Dr. Daniel Hsieh, Dr. Kenneth McGuinness, Dr. Patrick Nosker, Dr. Stefan Senn, Dr. Hagai Hraanan, Douglas Pike, Jose James, Kaiser Loell, Daniel Grisham and Teresita Silva.

My department Associate Director Gail Ferstandig Arnold has always been available for support in times of need. Finally, I am very thankful for the NIH for its lab project grants that I could avail of, as a graduate assistant at CABM.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Section I. Natural Collagen v/s CMPs:

Naturally occurring collagen is a ubiquitous structural protein that accounts for about 30% of the total human body protein content. As part of the extracellular matrix, collagen aids in the development of tissues and their regeneration. It provides tensile strength to skin, bone, cartilage, tendon and blood vessel walls, and also plays a crucial role in cell adhesion and cell migration during the cell's growth, differentiation and morphogenesis. Re-modeling of collagen has been both directly and indirectly implicated in several pathological conditions, including cancer, osteoporosis, arthritis and fibrosis. Faulty or deleterious mutations in collagen have led to inheritable and debilitating connective tissue disorders such as Osteogenesis Imperfecta and Ehlers-Danlos syndrome. Studying collagen at the molecular level furthers our understanding of these disorders. Comprehending the nature of natural fibrous collagen will also help us in building better collagen biomaterials for artificial livers, blood vessels, skin grafts and corneal tissue [1-3]. Natural, animal derived collagen extracts, have been extensively been used in building biomaterials. Their complex physiochemical properties however, tend to pose problems of immunogenicity and pathogen transmission. As a synthetic alternative to naturally occurring collagen, short chemically prepared ***collagen mimic peptides*** (CMPs) - of the order of 30 amino acids in length for each composite polypeptide chain, are currently being designed for use. One of the issues of structural characterization of human natural collagen is that it contains hydroxyproline and the fibrillar regions of it extend over a thousand amino acids. In this case too, CMPs have proven to be viable substitutes by

extensively serving as toolkits for studying the receptor binding, self-assembling, and folding kinetics of human collagen [4-5].

Study of heterotrimer CMPs, wherein all the 3 chains of the triple helix are of a different sequence, is of considerable interest. We plan to design aggregation resistant heterotrimer CMPs that serve as independent modular pseudo-domains for a futuristic longer chain collagen trimer. Sequence diversity can lead to aggregation resistance [6]. In natural proteins, adjacent domains tend to have relatively lower sequence identities than the ones that are further apart - reducing chances of misfolding [7-8]. Once we have access to well defined synthetic peptide orthogonal domains, one could potentially place them next to each other in tandem and express them in recombinant bacterial systems as long chained collagen molecules (**Fig. s1**). Currently the length that can be successfully expressed in recombinant systems, are the ones up-to around 250 residues in length [9]. Expressing recombinant collagen in bacteria results in higher yields compared to other methods [10]. By co-expressing the genes required for post-translational modifications such as prolyl and lysyl hydroxylation, the recombinant collagen expressed in bacteria could be made to resemble a natural human collagen trimer chain [11]. A couple of other differences between human and recombinant bacterial collagen system are as follows: **a)** The N- and C-terminal of the human triple helical collagen molecule contain domains that are non-triple helical - these domains, particularly the C-terminal one, help in the registration of the 3 chains of the triple helix through cysteine bonds [12]. **b)** There are also molecular chaperones in human cells that prevent misfolding and aggregation of the triple helix [13]. However, given that the triple helix alone can self-assemble *in vitro* to form fibrils [14], a longer chain recombinant bacterial expression system shouldn't be far from reach.

Crystal structure study of **charged** host-guest collagen mimetic peptide sequences have also implied that the sequence differences in the 2 chain types of collagen type 1 for example, may also account for the register specificity in natural collagen, besides the explicit presence of N-C terminal cysteine knots [15].

Section II. Coiled Coil Dimer v/s Collagen Trimer Orthogonality:

Orthogonal self-assembly of proteins and good protein specificity, is known to exist broadly among natural protein-protein interactions [16-18]. It has been studied extensively amongst synthetic coiled coils. For example, in one particular study, from among all the dimerization units possible in a coiled-coil 55-protein interactome toolkit, it was found that 23 synthetic coiled coils and 3 human bZIP coiled-coiled proteins could form 27 pairwise interactions that were mutually orthogonal and heterospecific. These 26 peptides made strong, reciprocal interactions with a partner to form a hetero-dimer. They did not even self-interact to homo-dimerize [19]. Dimer systems of such good orthogonality lead to well separated molecular ecosystems without any undesired crosstalk. We would like to attempt a complementary approach within the synthetic collagen triple helix structure space. By doing so, we would be expanding the synthetic biology playing field. The collagen family is inherently structurally orthogonal to the alpha-helical coiled coil and beta-sheet protein family and all these structural motifs should self-assemble mutually exclusively even when present in close proximity, giving rise to unique modular systems [20]. Although, for developing heterotrimer CMPs, a number of molecular strategies and designs have been considered in the past [21-27], the

cross-reactivity of these heterotrimers and specificity from among a pool of similar structures has seldom been studied before.

When considering and comparing orthogonal ecosystems, it would be interesting to note here that in a 9 peptide ***coiled coil*** toolkit interactome, for a given monomer peptide, the simplest level of orthogonality would involve 1 desired pairwise interaction with the peptide's dimer counterpart and only 8 other undesired competing dimerization interactions (Total no. of pairwise interactions – No. of desired pairwise interactions = No. of undesired pairwise interactions: (1*9) – (1) = 8). This includes the self-association homo-dimer interaction. Contrast this scenario with a 9 peptide collagen toolkit interactome, where-in, for a given monomer peptide, the corresponding lowest complexity orthogonal space is defined by 1 desired collagen trimer association state (with its 2 other collagen monomer counterparts within the target state triple-helix) and at-least 80 (1*9*9 -1 = 80) competing trimer association states, including its very own homo-trimer association state. Thus, achieving specificity/orthogonality in a corresponding 9 peptide ***triple helical*** structure space will prove to be quite challenging. This increase in difficulty level as the level of orthogonality increases is indicated in **Fig. 1**. Nevertheless, our body of work should increase the avenues available for exploring and expanding synthetic biology research, moving us closer to the goal of creating novel nano-molecular engineering tool-kits.

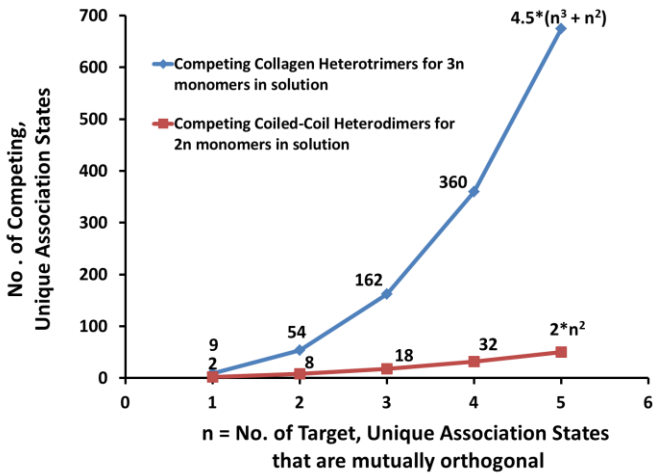


Fig. 1: N-Dimensional orthogonality in coiled coil system versus collagen system. No. of monomers in solutions for the two systems are 2n and 3n respectively. If 'n' is the orthogonality dimension, the Total no. of possible unique chain association states for coiled coil is given by $^{2n}C_2$ with repetition = ((2n+2-1)!)/(2!*(2n-1)!) and for collagen it's given by $^{3n}C_3$ with repetition = ((3n+3-1)!)/(3!*(3n-1)!).

There are other broader impacts of such an undertaking. For instance, wide spectrums of specificities are encoded within the triple-helix scaffold of the natural full-length collagen gene family. Types XI and II collagen for instance, co-assemble as heterotypic fibrils in the cartilage. Human skin and tendon has type III collagen co-assembled with type I collagen. Collagen type I is also known to form heterotypic assemblies in corneal fibrils with collagen V [28]. We hope that through our orthogonal synthetic peptide tool-kit design approach, one might be able to explore how these specificities are brought about within the natural collagen structure scaffold. Also, charge templating of collagen has been experimented with previously to generate useful biomaterials. By modularizing the long chain collagen trimer design through extensive orthogonal patterning of the salt-bridge network, we hope to make way for the study of its folding kinetics and biophysical properties in the future.

Section III. Co-creation v/s Addition to pre-existing Nano-components:

There were two distinct paths to approaching our orthogonal heterotrimers design. The first one was to design these hetero-trimers *ab-initio* (from scratch) and co-design/co-evolve for specificity. Specificity co-evolution is a known phenomenon in the natural world of protein-protein interactions [29-33] and *in vitro* synthetic protein designs [34]. However, we intended to see if the need for specificity co-formation in orthogonal trimers could be circumvented in a purely synthetic design where-in pre-existing peptides were already available and newer peptides needed to be introduced into the environment. This led us to the next alternative approach, whereby we took heterotrimers that were already previously designed in our lab – and put to test, the extent to which the specificity of these timers could be pushed / sustained upon the addition of further mutually exclusively self-assembling heterotrimers into the solution. It's noteworthy to mention here that in a previous work from our lab, it had been proved that one could go up-to 2 orthogonal heterotrimer ensembles, via mimicking natural processes such as circular permutation alone, without the explicit aid of computational design [35]. During the design process, we had to take into consideration the unique challenges that are specific to computational design of collagen. These challenges do not show up when considering traditional approaches adopted for de novo design of other proteins and peptides. For example, unlike globular proteins, collagen being a linear molecule, has no hydro-phobic core that can be computationally modeled as the driving force for the formation of its three dimensional folded structure [36, 37]. Besides this, two-thirds of the residues in CMPs are solvent facing, with little to no quantifiable measure available for estimating the contribution of the resulting hydration network to collagen stability. The hydration

network is an important factor while considering any computation protein design. Last but not the least, collagen is a slow folding molecule exhibiting no clear two state transition between folding and unfolding states [38, 39]. This makes it hard to explicitly assign the contributions of its individual residues to the energy terms of its folded and unfolded state. To overcome these challenges, we choose to extend upon our previous successfully designed and validated sequence based energy scoring function [21] and sought to apply it to our current orthogonal trimer design.

Once the sequences were computationally designed, the peptides would then be synthesized and characterized for stability and specificity. This phase was planned to be carried out in the following manner:

Stability: Stability would be characterized through CD experiment. Hetero-specificity within the respective composite 3 monomer peptide ecosystem, of the newly designed orthogonal trimers – also called here as 'self-orthogonality', would be tested by studying the CD characteristic of the 9 possible competing stoichiometric trimer association states.

Specificity: Specificity of trimer strand association would be checked for, through a series of combinatorial CD experiments – explained in greater detail in the following chapters. Bio-tin tagging the lagging strand in each of the designed target state orthogonal heterotrimers and "fishing" out the monomer chains that associate with these specific lagging strands would be a more effort intensive approach. One would be able to extract only the designed for orthogonal target trimer states and not any other competing trimer association state through this method. The biotin separated out peptide solutions would then be further characterized through HPLC and Mass-spec techniques provided the

peptide peaks in the HPLC retention curves happen to be not well separated. We do not implement biotin tagging and only study the CD characteristics of various ‘monomers to trimer folding mixers’ for our study.

In essence, a computational design for a domain templated collagen toolkit assay that could potentially aid in understanding the independent modular pseudo-domain folding and self-assembly process in synthetic and natural collagen is put forward and the challenges faced at each stage of design are thoroughly documented through a series of incremental and improved design iterations. A sub-hypothesis that a 15 amino acid sequence length spanning 3 chain pseudo-domain could potentially suffice to serve as an independent, modular, hetero-specific collagen domain, folding unit is also rigorously scrutinized and tested.

CHAPTER II

Orthogonal collagen mimetic peptides design

KEYWORDS: Collagen, Peptides, Heterotrimer, Orthogonal Self-assembly, Oligomers, Aggregation, Quantifying Specificity, Sequence search algorithm, Replica exchange, Genetic evolution, Strength Pareto Evolutionary Algorithm 2, Multi-objective optimization, Minimum Set Cover, Denovo design, Energy landscape study.

Section I. Computational Design:

The main goal of our project was to computationally design and test the stability of heterotrimer peptides that self-assembled mutually exclusively in the presence of other pre-existing heterotrimers in solution. An increase in orthogonality by one level would be achieved by the introduction of one new self-assembling heterotrimer into an existing ensemble of peptide heterotrimers. We take a system with 2 pre-existing mutually exclusively self-assembling trimers (orthogonal level 2) and test if this *pre-existing system*'s orthogonality level of 2 can be further increased (level 4 orthogonality) through the addition of 2 more computationally designed mutually exclusively self-assembling heterotrimers into the solution.

Pre-existing, level 1 and 2 orthogonality:

First, let us consider the pre-existing orthogonality levels - level one and level two. Level one consists of a single heterotrimer composed of three distinct 30 amino acid long collagen mimetic peptide monomers. It is known that through the use of attractive

electrostatic interactions between its 3 composite monomer chains named {A, B and C}, the heterotrimer "ABC" has been computationally designed to demonstrate good stability within its '3 monomers in solution only' peptide ecosystem. Specificity for the formation of the specific target state trimer within this 3 composite monomer chain ecosystem, was achieved by lowering the stability of all the other 9 possible competing - 3 chain, stoichiometric association states : "3A", "3B", "3C", "1A2B", "2A1B", "1B2C", "2B1C", "1A2C" and "2A1C". The lowered stability of the competing states was achieved through the use of a relatively much lesser number of attractive electrostatic interactions between the composite trimer chains of the competing states, in comparison to the number of attractive interactions present in the target trimer state. Extensive repulsive interactions were also used to further destabilize these 9 competing states [21]. Level 2 orthogonality moves on to a '6 monomers folding into 2 distinct heterotrimers in solution' peptide ecosystem. Here, an additional set of 3 peptide monomer chains named {D, E and F}, were introduced to the previous {A, B and C} 3 chain peptide group. The target state trimer "DEF" was previously designed in our lab as follows: each half length of the now stably folded heterotrimer "ABC", was treated as an independent modular domain and swapped, emulating the process of 'circular permutation' - a phenomenon that is known to occur naturally in globular proteins. Circular permutation in CMPs retains the salt-bridge network that holds together the three chains of the collagen peptide – with little to no loss in specificity, within a 'composite 3 monomers to a trimer' folding ecosystem. When moving onto a 6 peptide ecosystem, the 2 sets of former three chain peptides {A,B,C} and its newly introduced N-C terminal circular permuted counter-part {D,E,F}, form 2 distinct collagen hetero-timer species { "DEF" and "ABC" }, with a

relatively good amount of specificity within the 6 peptide ecosystem [35]. Although, strong competing cross-talk stoichiometric association states, such as “CEF” and its circular permuted version “BCF” are known to form within a 3 peptide ecosystem, it was shown through streptavidin-bio-tin tagged fishing experiments that when one moved to an all 6 ‘A to F’ peptide solution mixture, only the most stable states: “DEF” and “ABC”, formed out of the possible 56 stoichiometric association states. In this manner, a ‘not computationally designed for’ level 2 orthogonality with 2 independently folding heterotrimers was established. We proposed to test if there was room for further fine tuning of the ‘2 out of 56 states’ specificity attainment, through the means of in-silico computational design of orthogonal positioning of salt-bridges and residues.

Designing a level 3 orthogonality, 9 peptide ecosystem: Our main aim was to introduce one more heterotrimer that self-assembles mutually exclusively in the presence of peptides ‘A to F’. (It is once again noteworthy to mention here that the DEF trimer was not computationally designed for orthogonality). By pushing in more sequence unique, stable target trimer states into the solution ecosystem and destabilizing the rest of the potentially formable competing stoichiometric states, a level 3 orthogonal, non co-derived for specificity design was formulated. We call the specificity or orthogonality being designed for, as non co-derived, since the 2 trimers “ABC” and “DEF” are already pre-exisiting and their sequence cannot be altered. In other words, the newly designed 3 chains, say ‘G, H and I’, should self-assemble into a 3rd new heterotrimer, from among the pool of 9 peptides { A, B, C, D, .. , H, I } present in solution. There should be no crossover of peptide chains from one heterotrimer species {A, B, C}, {D, E, F} or {G, H,

I}, to another, during the simultaneous and independent self-assembly of the 3 “ABC”, “DEF” and “GHI” heterotrimers. For example, 2 possible cross-talk association states are "CEF" and "GFI". Having designed this 9 peptide ensemble computationally, we would be achieving an unprecedented stoichiometric specificity of forming only 3 association states out of ${}^{9}C_{3}$ = 165 possible unique three chain homo-trimer (such as "AAA", "DDD", "GGG" etc.) and heterotrimer (such as "ABG", "DHC", "GHG" etc.) association states.

Designing a level four orthogonality, 12 peptide ecosystem: To further drive home the concept of folding orthogonality, a fourth timer “JKL”, that had a pre-established design constraint of being a circular permutation of “GHI”, was also computationally designed. The orthogonality map between all the independent 4-unique ‘half of full-length’ domains, constituting the 4 heterotrimers “ABC”, “DEF”, “GHI” and “JKL”, is shown in **Fig. 2**. The weak orthogonality resulting through pre-existing circular permutation between the 2 half’s of “ABC” and “DEF”, is shown as a dotted line, and the strong, computationally designed for - orthogonality between the rest of the network, is shown in thick lines. In essence, each of 2 half pseudo-domains of the trimers should be orthogonal to the rest of the pseudo-domains, in-turn resulting in all the 4 full-length heterotrimers being mutually orthogonal.

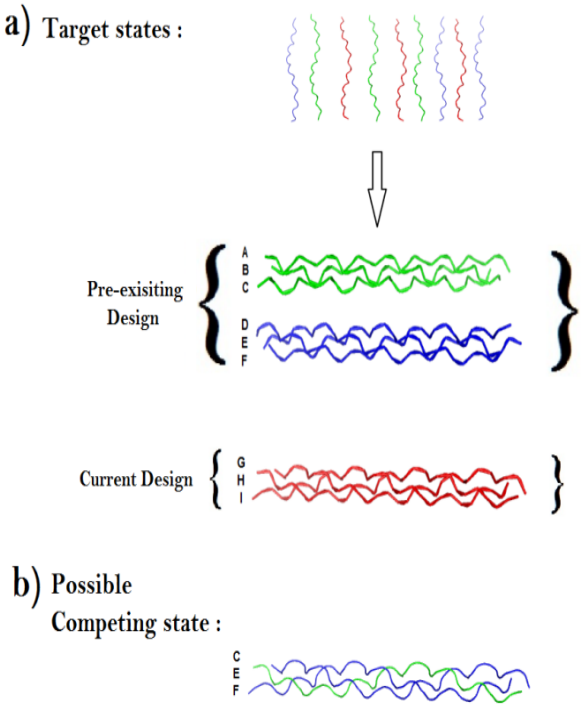


Fig. 1: Level 3 orthogonality: a) Nine peptide chains [A to I] each having a unique amino acid sequence. They self-assemble in-vitro into 3 distinct heterotrimer species. Each species has a different color. GHI is the newly designed trimer species (red) that self assembles orthogonal to the existing ABC (green) and DEF (blue) heterotrimer species. **b)** Example of an undesired cross-talk. Heterotrimer CEF derives its constituent peptide chains from 2 different species (green and blue) as shown here.

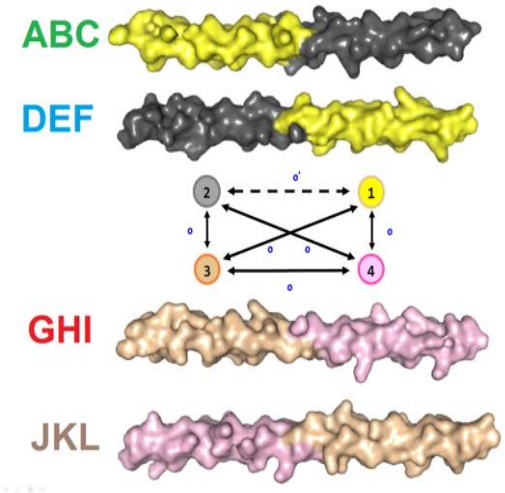


Fig. 2: Level 4 orthogonality: The pre-existing Collagen mimetic heterotrimer peptide "ABC" and its pre-existing circular permuted version "DEF", wherein the N-terminal left

half and the C-terminal right half are interchanged, is shown. The weak orthogonality (o') between the two halves (yellow and grey) is shown as a dotted line. The computationally designed orthogonal trimer pair "GHI" and its circular permuted counterpart "JKL", is shown at the bottom (orange and pink). The colored circles represent the domains and the arrow lines represent the sequence and 3 chain association orthogonality existing between the 4 different colored domains.

Stability and Specificity Score:

The 2 main objectives that had to be optimized for the design of "GHI" and "JKL" were the stability and the specificity score. The protocol followed to achieve this was the same as that for "ABC" design [40, 35]. The protocol in brief: 10 random triplet amino acid sequences are initially selected and laid out in tandem to form a single monomer sequence. Three of these different monomer sequences are used up to define the starting sequence of the target state "GHI" trimer and then circularly permuted to define the starting sequence of the target state "JKL" trimer. The triplet sequence selection pool was restricted to only 5 triplet types –{ 'KOG', 'PKG', 'DOG', 'PDG' and 'POG' }. Stability score of the generated trimers was calculated as follows:

Stability score E = Ionic interaction Energy + Sequence Backbone energy

Specificity score = Target State – Closest Competing State Stability Score

*Backbone energy = -3.8 * (No. of P's or O's in sequence)*

The terms 'energy score' and 'stability score' are used interchangeably; so are the terms 'gap score' and 'specificity score'. A value of '-1' is added to the Ionic interactions energy score sum for every 'K' to 'D' attraction, '+2' for every 'K' to 'K' repulsion and '+3' for every 'D' to 'D' repulsion within the trimer. Only Y-X and Y-X' position

residue interactions were considered for calculation. The Y labeled residue is on one chain and X and X' are present on the adjacent chain of the $[GXYGX'Y']_n$ monomer sequence, 3 chain collagen trimer model [40]. The weight value 3.8, indicating the ratio of energy contribution from the imino acid content to that of attractive charges was chosen based on the correlation between computed energy and the observed melting temperatures (Tms) of synthesized trimers from previous experimental studies. Through a series of repeated single triplet mutations on the initial random sequence, Monte Carlo Simulated Annealing (MCSA) procedure was used next to search through the potential (5^10)^3 sequences search space for minimizing the cost function, with the cost function being defined as :

Cost function ***C = Stability Score + Specificity score***

⇨ $C = E_{GHI} + E_{JKL} + Gap\_{GHI} + Gap\_{JKL}$

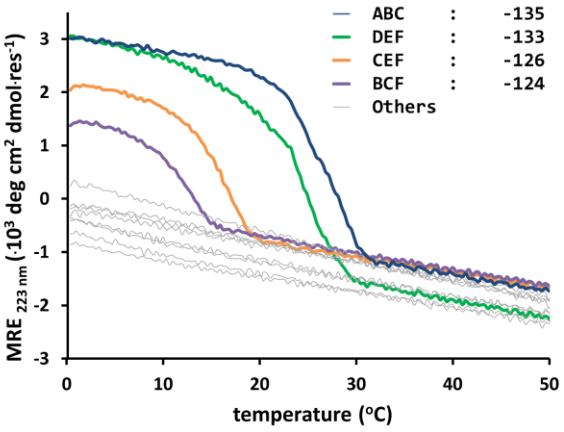
Wherein, the stability or energy score is represented by 'E' and the specificity score is defined as 'Gap score' with Gap score = $E_{\mathbf{Target\_State}} - \min(E_{\mathbf{Competing\_State}})$. The $\min(E_{\mathbf{Competing\_State}})$ term defines the energy score of the most stable competing state or states, from among the peptide ensemble that utilizes 1 or 2 monomers chains of the target trimer association state. For example, 3 chain association state "GFI" may be considered as a competing state to target state "GHI" since the monomers "G" and "I" from {"G", "F", "I"} are being used for the formation of "GHI". All other 5 possible association states of the leading, lagging and middle strands of the competing state "GFI" : {"IGF", "FIG", "GIF", "FGI", "IFG"}, were also taken into consideration, while calculating the $\min(E_{\mathbf{Competing\_State}})$ term value. The stability score and specificity score were normalized as [approx. min (score), approx. max (score)] → [0, 1] and weighted

appropriately for faster convergence to the minimum cost function value, over a predefined number of iterations. To obtain the approximate possible min and max scores for our sequence model, a number of initial short trial simulations were run. Although the 'POG' triplet was included in the initial set of simulation runs – it was later excluded for the following reasons: **a)** Simulation runs including "POG" resulted in lesser specificity (higher specificity/gap score between target and competing state) **b)** "ABC" and "DEF" trimers did not contain any 'POG' triplets, and they had already been validated to give good trimer specificity and stability and **c)** since the contribution of 'POG' to heterotrimers is not yet well characterized.

For the MCSA the cooling temperature schedule used was T = (15000/ i) ºC with the simulation iteration number "i" proceeding from i_initial = 1 to i_final = 15000. Since simulated annealing is known to leave an optimal solution and not find it again, the lowest point visited over the span of the entire simulation run was constantly monitored and saved. Also, multi-start simulated annealing was used in place of the regular simulated annealing. For this, 100 initial starting random sequence conformations were chosen and the best sequence set generated out of the 100 MCSA trials was selected for eventual peptide synthesis. It can be seen from the generated sequence pattern (**Fig. 3**) that the folding orthogonality arises from both the orthogonality in sequence ('K's in place of 'D's and vice-versa) and the orthogonality of presence versus absence of a salt-bridge at a given location along the trimer length and width. It is noteworthy to mention here that "ABC"s mirror image (interchanging K and D positions), may not result in a trimer that is orthogonal to "ABC". This is because even though the good stability of this

mirror image trimer is guaranteed, its 3 composite-chain association specificity in the presence of "ABC" trimer is not. Thus a computational design approach would be more appropriate to solve the protein design problem at hand, than a rational design approach.

The best stability scores obtained at the end of the simulation run for "GHI" and "JKL" were '-136' and '-135' respectively. The best energy gap between "GHI" and one of the next most stable competing states "GLF", was '-12' and for "JKL" the gap was '-11' between "JKL" and "JIC" (**Fig. 4**). The {stability, specificity} scores of "ABC" and "DEF" remained unchanged at {-135, -9} and {-133,-7} respectively, since "CEF" (stability score = '-127') continued to remain the strongest ***competing*** state when moving from an orthogonal level two '6 peptide ensemble' to orthogonal level four '12 peptide ensemble' (**Fig. 3**). It had previously been shown that even with as small a specificity gap as '-7', "ABC" and "DEF" trimer molecules could be bio-tin tagged and 'fished' out of a solution mixture consisting of the composite 6 monomer peptide ensemble[35]. It was thus hypothesized that similarly, due to the existing minimum individual target state timer gap with corresponding gap score of '-7', for a 12 peptides ensemble, the monomers to trimer association folding specificity in solution would be such that only the most stable designed for 4 states would be formed in large quantities, while the rest of the states would, if present, be present in negligible amounts. Therefore a case for good specificity in a level 4 orthogonality, 12 peptide ensemble, would be established. Also, since "GHI" and "JKL" each had 22 and 21 salt-bridges, similar in number to that of ABC's 21, and DEF's 19, it was expected that the designed trimers would demonstrate sufficient stability as well.



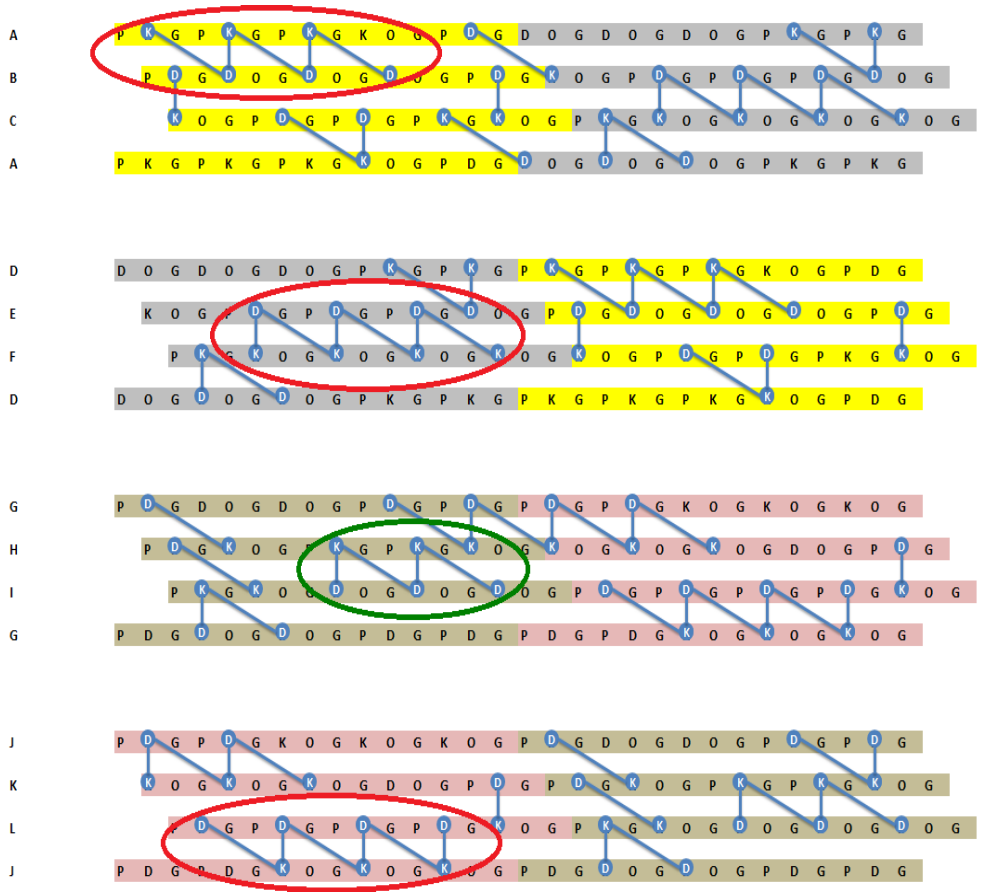


Fig. 3: The circular dichroism spectroscopy signal of the pre-existing heterotrimers ABC and DEF, along with its energy scores, is shown on the top panel. Also shown is the CD

spectrum of the competing states “CEF” and “BCF” within a 3 peptide ecosystem. CD spectrum data from paper reference: [35]. The computationally designed orthogonal heterotrimer “GHI” and its circularly permutated orthogonal counterpart “JKL”, is shown below the “ABC”, “DEF” sequence in the bottom panel. The orthogonal positioning of the salt bridges is shown in red and the orthogonality in positioning of residues ‘K’ and ‘D’ between trimers “DEF” and “GHI” is shown in green.

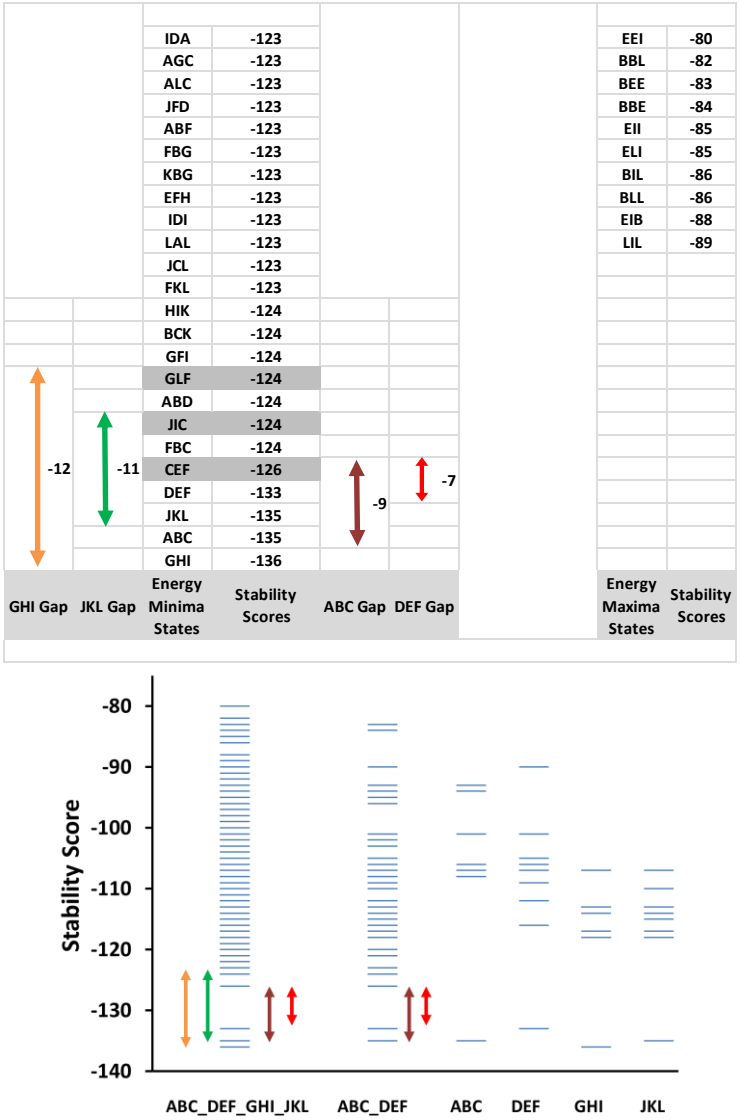


Fig. 4: Low energy score values indicate good stability. Of the 364 unique trimer association states possible (= $^{12}C_3$ with repetition) , the top 25 good stability states and the bottom 10 low stability states are shown along with their energy score on the table. Higher the number of salt bridges, lower is the energy score and higher is the stability of the corresponding association state. The energy bar diagram when there are ensembles of 3, 6 and 12 monomers in solution is also shown on the right. Note that the energy gap between the target energy state and the next immediate competing state in the energy

landscape diagram decreases as the number of peptide monomers in the peptide ensemble increases (moving from right column to left column).

Section II: Stability of GHI:

The Circular Dichroism (CD) plots of "GHI" are shown in **Fig. 5** and **Fig. s3**. The stability of "GHI" along with its 9 stoichiometric states was tested. The following observations were made: **i)** In 10mM phosphate buffer: The melting temperature of 0.2mM "GHI" (0.2mM = total peptide concentration) was 25.5 ^{o}C with no salt and 22.5^{o}C with 100mM NaCl. The corresponding "ABC"s and "DEF"s melting temperature (Tm) in no salt had been 29 ^{o}C and 24 ^{o}C respectively.

With salt, "ABC"s and "DEF"s melting temperature (Tm) remained more or less the same, with a slight drop in the MRE signal. In contrast, upon the addition of salt, while the Tm remains the same, the MRE drops considerably for "GHI" (**Fig. 5b**). This, along with the fact that the stability of GHI is lesser than "ABC" in-spite of having an additional salt-bridge, could be attributed to the net neutral charge of "ABC" and "DEF", in comparison to the net charge of '-4' for "GHI". The discrepancy of lack of increase in stability, corresponding to an increase in the number of salt bridges, could also be due to the following: During the cost function minimization simulation - only the gap existing between "GHI" and the rest of its competing states in a 12 peptide ensemble and not its composite 3 monomer chain peptide ensemble, had been optimized. Thus the 'internal trimer gap' – the gap between a trimer and the rest of its competing 3 chain association states within a composite 3 monomer chain ensemble, is considerably less for "GHI" than it is for "ABC" [Table 1]. Thus, we see states such as "HHG" (Tm = 11^{o}C), "GHG" (

Tm = 10^{o}C) and a couple of other states, with less than 5^{o}C Tm, when no salt is used (**Fig. 5a**). Addition of salt eliminates these states in the second set of experiments (**Fig. 5b**). The 3 peptide ensemble stability scores and their corresponding experimental Tm's in no salt are tabulated in **Fig. s3.1**.

Previously our lab had characterized a pair of acid tectons that had charges of '-7' and '-10', Tm's of 19 ^{o}C and 18.5 ^{o}C and number of salt-bridges equal to '18' and '17' respectively [27] . Since the '-10' acid tecton was marginally stable, we were hopeful that "GHI", with its high number of salt bridges and a net charge of only '-4', would demonstrate sufficient stability. Further simulations and experiments had to be carried out to study the effect of charge and trimer gap to ascertain their contribution to the overall stability.

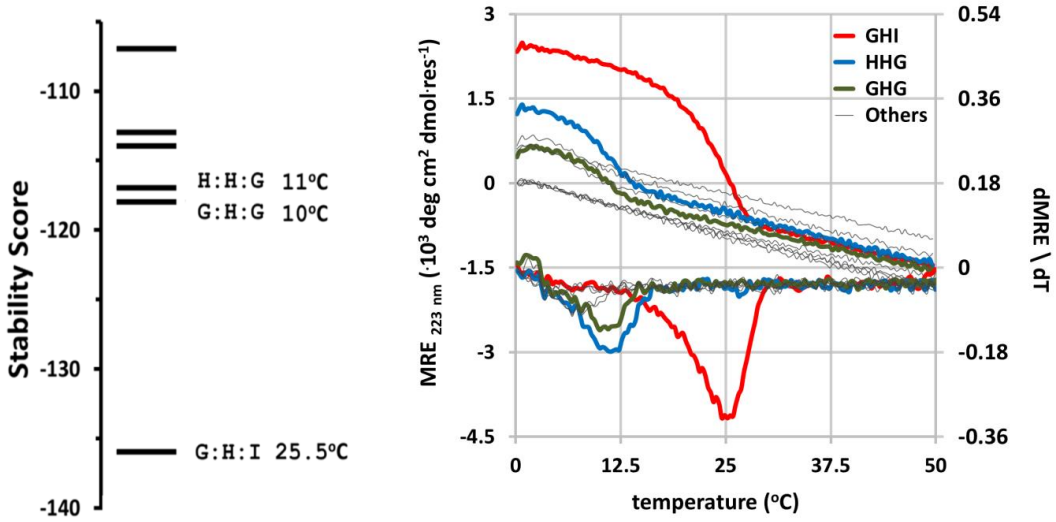


Fig. 5a: GHI Characterization: Circular dichroism temperature melting curves monitored at 223 nm of 0.2 mM "GHI" heterotrimer peptide in 10 mM phosphate buffer. pH = 7, along with the first derivative graph indicating the melting temperature of computationally designed "GHI" (25.5 ^{o}Celcius) is shown. The 10 possible

combinations of the monomers ‘G’, ‘H’ and ‘I’, were characterized. The energy bar diagram is shown to the left.

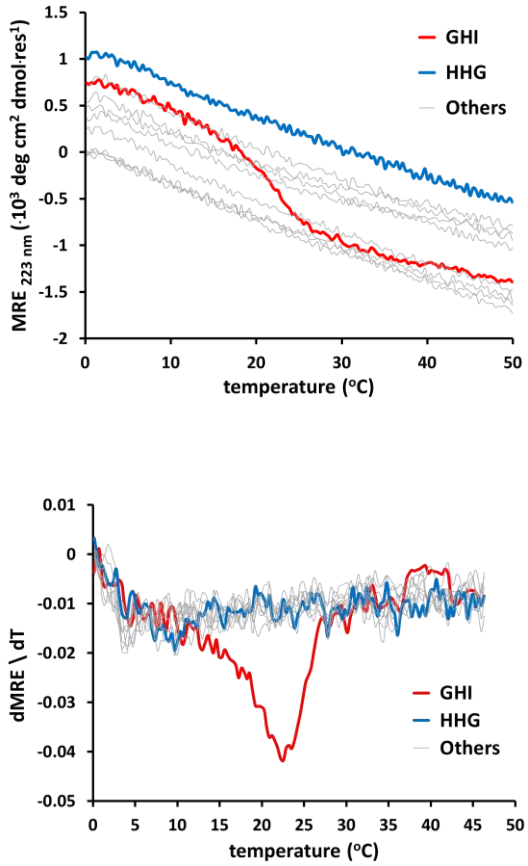


Fig. 5b: Circular dichroism temperature melting curves monitored at 223 nm of 0.2 mM “GHI” heterotrimer peptide in 10 mM phosphate buffer, pH = 7, 100 mM NaCl. The 10 possible combinations of the monomers ‘G’, ‘H’ and ‘I’, were characterized. The first derivative graph indicating the melting temperature of computationally designed “GHI” (22.5 ^{o}Celcius) is shown.

Section III: Orthogonality in a 9 peptide ensemble: Next, the specificity of “GHI” was tested. A series of combinatorial CD experiments was carried out to test the resistance to chain association specificity of the 3 target states “ABC”, “DEF” and “GHI”, by their respective competing states, in a ‘A to I’ 9 peptide ensemble. The hypothesis was that by assessing the strength of the cross-association states, we would

gain a general perspective of how well the monomers separate out to associate with their respective, designed for, chain association partners.

There were 2 scenarios to be considered: **Scenario A:** If all 9 monomers were to be present *in vitro*, only the target states should be favored to be formed. This is essentially due to their higher stability and the considerable difference between stabilities of the target state versus the set of next most stable competing state or states in the ensemble. If the desired "ABC", "DEF" and "GHI" trimers' constituent monomers { A to I } are used up from the reservoir pool to form the desired higher stability target state heterotrimers, then there's less likelihood that the competing states are being formed in solution due to the scarcity of their constituent monomer peptide chains. This feature was proven to hold true for level 2 orthogonality, 6 peptide ensemble [35]. We hoped to check if this possibility held true for a higher, level 3 orthogonality, 9 peptide ensemble as well. **Scenario B:** Under ideal conditions, in the absence of all the 3 chains that are required for formation of one or more target states, the competing states should form with little or no stability. This property follows from the fact that there are present, a relatively lesser number of attractive electrostatic interactions than accounted for in the target trimer states, along with the presence of a greater number of repulsive interactions between the competing association state chains. We decided to first test if the latter ‘Scenario B’ held true for our computational design.

The total number of heterotrimer stoichiometries possible for a 9 peptide ecosystem = $^{9}C_{3}$ = 84. Excluding our target state "ABC", "DEF" and "GHI" trimers, we would be left with 81 possible competing chain association states. By performing a 'leave one peptide out from each target trimer species', we tested for the level of collective strength of the

competing states through circular dichroism experiments. For example, say if we were to exclude the lagging strand monomer from each of the 3 target state trimers, we would get a six peptide ensemble of "ABDEGH". This would constitute 'experiment set one' and would account for a total of ${}^{6}C_{3}$ = 20 trimers out of the 81 possible competing chain association states. In-order to account for the remaining 81-20 = 61 states, we needed to find out what was the minimum number of such 6 monomer set groupings that would ensure that all the 81 competing states are accounted for. The lesser the total number of experiments to be conducted, the lesser would be the amount of peptides consumed for running the experiments. Such a problem (of deciding upon the least set of experiment or groupings that cover a given set) is known as the 'minimum set cover' problem in Combinatorics. The algorithm to arrive at this number is discussed in the methods section. For our experimental study, the minimum of such 6 peptide grouping was calculated to be 7. The 6 peptide monomer groupings of the 7 experiment sets are shown below along with their experiment set number.

1) "ABDEGH"

2) "ABDEHI"

3) "ABEFGH"

4) "BCDEGH"

5) "CAFDHI"

6) "BCFDIG"

7) "CAEFIG"

A CD wavelength scan and temperature melt of equi-molar concentrations of the 6 peptide group mixtures was measured (**Fig. 6)**. The total peptide concentration of the solution mixtures was 0.2mM and the buffer used was 10mM phosphate buffer with 100mM NaCl. The following observations were made: **i)** The first five experiment sets were almost indistinguishable, demonstrating low MRE (**Fig. 6b**) with unclear folding to unfolding transition and Tms were less than 18 ^{o}C in the melt profile (**Fig. 6b, d**). **ii)** The last 2 sets of the 7 set experiment sets containing the trimers "BCF" (experiment set 6) and "CEF" (experiment set 7), showed both higher helicity in the wavelength scan as well as a clear folding to unfolding transition along with a higher than expected melt of 29.5 ^{o}C and 16.6 ^{o}C respectively (**Fig. 6c**). "BCF" and "CEF", were already present in the 6 peptide {A, B, C, D, E, F} ensemble as strong competing trimers with a low energy score (high stability) and melting temperatures of 11.5 ^{o}C and 15.5 ^{o}C respectively (**Fig. 3**). However it was interesting to note that while the 3 peptide ecosystem containing trimer “CEF” had a greater Tm than the 3 peptide ecosystem containing trimer “BCF” (**Fig. 3**), the situation was reversed when these trimers were present in a 6 peptide ecosystem (**Fig. 6c**). **iii)** Next, the following 3 more experiment sets were added to the super-group of experiments:

8) "CFDIG"

9) "AEFIG"

10) "ABCDEFGHI"

The two experiment sets numbered 8 and 9 were formed as follows: The strand “B” from the 6 peptide ensemble “BCFDIG” was removed to eliminate the contribution of “BCF” and the strand “C” from the 6 peptide ensemble “CAEFIG” was removed to eliminate the

contribution of “CEF”. The total peptide concentration of the 5 peptide ensemble was kept at 0.2 mM. These mixtures did not show any triple helix transition. There may still be trimer states such as {“BCI”, “BFD”, “BFI”, “BFG”, “BIG”} that are unique to experiment set “BCFDIG” - that could be contributing to the high stability triple helicity signal. However, since their stability scores are high, (low number of attractive electrostatic interactions) it’s unlikely that they are independently or solely responsible for the strong CD signal. A study of the density of all the energy states unique to each of the 7 experiment sets will have to be carried out to ascertain the reason for the difference in CD signal pattern for the 3 versus 6 peptide ecosystem. For now it suffices to conclude that as the heterogeneity of the mixture increases, a higher probability of triple helix folding occurs, resulting in lowering of the difference in energy between the target state and the competing states.

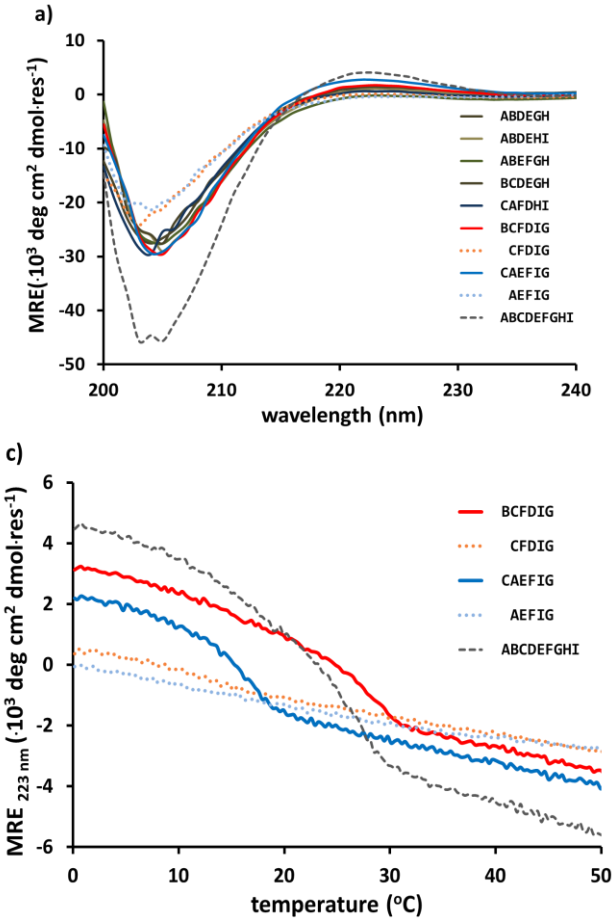
It was known previously that even if the concentration of the competing species is kept constant, the specificity of the target state decreases as the heterogeneity of the solution mixture increases [40]. However, it was not clear to what extent this phenomenon effects a in a collagen based system. Through our experiments, we now have a good set of data to model the underlying specificity-stability tradeoff in a 6 peptide ensemble. Increasing the salt concentration in all the 7 experiments resulted in no folding transition, implying that the helicity signal was due to the presence of ionic salt bridges holding the trimer strands together. **iv)** When all the 9 peptides A-I were mixed in equi-molar solutions, keeping the total peptide concentration the same at 0.2 mM, the CD melt signal is lower than that of the individual ABC signal (**Fig. 6c,d**). This difference is likely due to the following fact: The underlying constituent total monomer peptide concentration = 0.2

mM in both 3 peptide ensemble and 9 peptide ensemble. Three monomers get used up to form one trimer molecule. Since only “ABC” has been shown to form with good stability, the corresponding concentration of “ABC” trimer molecules in a 3 peptide ecosystem would be 0.2/3 = 0.66 mM of trimer “ABC”. In the case of the 9 peptide ensemble too, “ABC” has the most designed for stability other than “GHI” and “DEF”. However, its individual contribution to the CD signal is lesser as the timers formed with good stability under this scenario include “ABC”, “DEF” and “GHI” with the specific timer “ABC”’s concentration being (0.2/3)/3 = 0.02mM, provided all the monomer chains exist as a part of a target state trimer or a competing state triple helix, inspite of the overall trimer concentration being the same as that of the corresponding 3 peptide ecosystem. **v)** Both “BCFDIG” and "ABCDEFGHI" show multiple peaks in the first order differential of the melt curve, indicating there are more than one kind of trimer species being formed. **vi)** We also note here that, even the first 5 experiment sets, had a CD signal better than that of the “BCF” signal of a 3 peptide ensemble (**Fig. 3, 6a**). The next best stable state within the “GHI” versus rest of the possible timers using up one or two of “GHI”s constituent chains, was “GLF”, and its stability score ‘-124’ was the same as that of “BCF” ‘-124’ stability score(**Fig. 4**).

According to our current computational design, the level of any competing state’s CD signal in a 6 peptide ensemble should have been atleast as strong as that of the trimer “BCF”, but not greater, especially considering that an independent competing state trimer’s concentration is atleast halved if not further reduced when moving from a 3 peptide ecosystem to a 6 peptide ecosystem. Thus we see a dramatic decrease in

specificity, as the number of heterogeneous monomer components in the solution increases.

In light of these results, we would need an improved computational model wherein not just the gap between the target and one of the next strongest competing states is taken into account, but the entire density of competing states ensemble is considered as well.



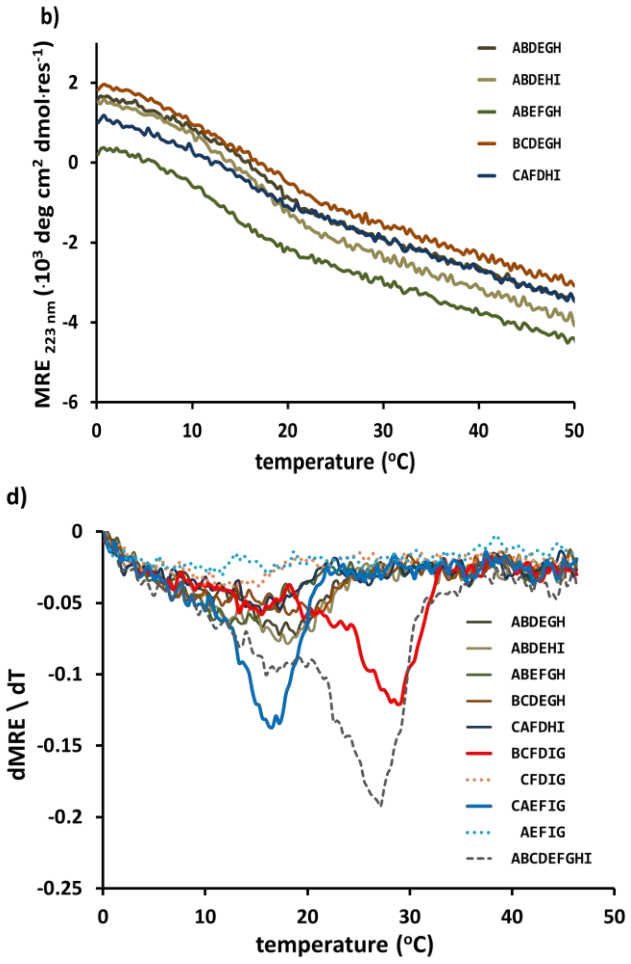


Fig. 6: Combinatorial experiments for estimating barrier to good specificity. **(a)** Circular Dichroism wavelength scan and **(b-c)** temperature melting curves monitored at 223 nm of a series of 10 experiment sets. 10mM phosphate buffer, pH = 7 was used. The total peptide concentration in each experiment was 0.2mM. The 7 sets consisting of 6 monomers in solution contain only 2 monomers from each target state trimer species, so a stable heterotrimer from among a single heterotrimer species cannot be formed. The 8th and 9th experiment set consisting of only 5 monomers in solution, eliminate the pre-existing "BCF" and "CEF" cross-talk association states. The 10th experiment set has all the 9 monomers under consideration. **(d)** First derivative plots of CD signal of all 10 sets.

Section IV. Stability of JKL:

"JKL" was characterized experimentally to see if it would provide further insights into the factors contributing to a charged trimer's stability. In phosphate buffer with no salt, "GHI" folded with a Tm of 25.5 ^{o}C. However, its circular permuted version "JKL", with one lesser salt bridge, showed little or no a transition from the folded to an unfolded state at the specific 0.2mM peptide in 10 mM phosphate buffer concentration. Upon increasing the concentration to 0.4mM, a Tm of 16.5 ^{o}C was recorded (**Fig. s6a**). Previously our lab had used the presence of copper in the buffer to increase the Tm of "ABC" by 5 ^{o}C [40]. It was shown with a degree of certainty that Copper provided a stabilizing effect by binding in a sequence-independent manner to the backbone amines of the N-terminus. We wondered if a similar strategy could be used to strengthen the helicity of "JKL" and stabilize it, in order to provide for a better 'folding to unfolding' melting transition CD signal. Since copper precipitates in phosphate buffer, we switched over to a 10mM Tris buffer, retaining the total peptide concentration at 0.2mM. The series of CD experiments conducted with varying concentrations of Copper are shown in **Fig. s8**. The observations made from the experiments are enlisted next : **i)** the very first observation made was that unlike in the case of the 0.2mM phosphate buffer, "JKL" showed a clear folding to unfolding transition at 16.5 ^{o}C in Tris (**Fig . 7**). This implied that the Tris buffer had a stabilizing effect on net negatively charged peptides. A similar observation had been made by the Hartgerink's group, wherein $(DOG)_{10}$ at 0.2mM concentration and pH 7, did not fold in the phosphate buffer but folded well in Tris buffer, giving a Tm of 39.5 ^{o}C and 37.5 ^{o}C with and without 150mM salt respectively [42]. It was noted that since the Tris buffer was cationic in nature, it could be interacting

with the negatively charged aspartic acid residue, reducing side chain charge repulsions and enabling the three chains of the trimer coming together to fold into a stable triple helix. We further hypothesize here that since phosphate buffer has an anionic nature to it, it may further have the opposite destabilizing effect on net negatively charged peptides - as opposed to playing a neutral or even a stabilizing role. **ii) Fig. 7** also shows that none of the competing 9 stoichiometry of "JKL" fold in Tris buffer, indicating that the trimer is hetero-specific (self-orthogonal) within a 3 peptide ecosystem. **iii)** Addition of 0.5mM Cu to the buffer solution does have a further stabilizing effect on "JKL" and increases its Tm by 4 ^{o}C, while the homotimers "3J", "3K"and "3L" remain unfolded (**Fig. s6c**). Further increase in Cu concentration did not further increase the stability due to charge screening. Inclusion of salt into the buffer solution containing Cu destabilized the "JKL" triple helix, indicating that the stability of "JKL" was arising from both electrostatic interactions and Cu stabilization (**Fig. s6b**).

DISCUSSION

Section I : Effect of 'POG' triplet on stability :

Although "GHI" and "JKL" folded into a triple-helix, they were not as stable and well folded as we had designed them to be. Two previously designed net-negatively charged peptides, acid tecton I and acid tecton II [27], inspite of having a greater net negative charge than "GHI" and "JKL", had higher melting temperatures when the buffer contained no salt. Acid tecton II was comparable in its energy score to "GHI"s, had greater net charge and unlike "JKL", folded in phosphate buffer [Table 1]. Closer inspection of the Acid tecton II sequence showed that its stability may be arising from the

presence of a single ‘POG’ triplet in its sequence. The stabilizing effect of a ‘POG’ triplet has been well documented before – however the ratio of its relative contribution to the stability of the triple helix in comparison to the stability provided by the ionic salt bridges has not been ascertained yet. Hopefully the addition of “GHI” and “JKL” to the set of well characterized net negatively changed heterotrimers will make their respective relative contributions clearer for future designs. Acid tecton I sequence had no stabilizing ‘POG’ triplet. Yet, even with a greater net charge of ‘-10’, in comparison to a similar energy score trimer such as “JKL”, it had a greater melting temperature in phosphate buffer. This could be attributed to the fact that Acid tecton I had a much better energy gap between the target and competing states than that of “JKL” [Table 1]. Thus we see several factors that influence the stability for net negatively charged peptides.

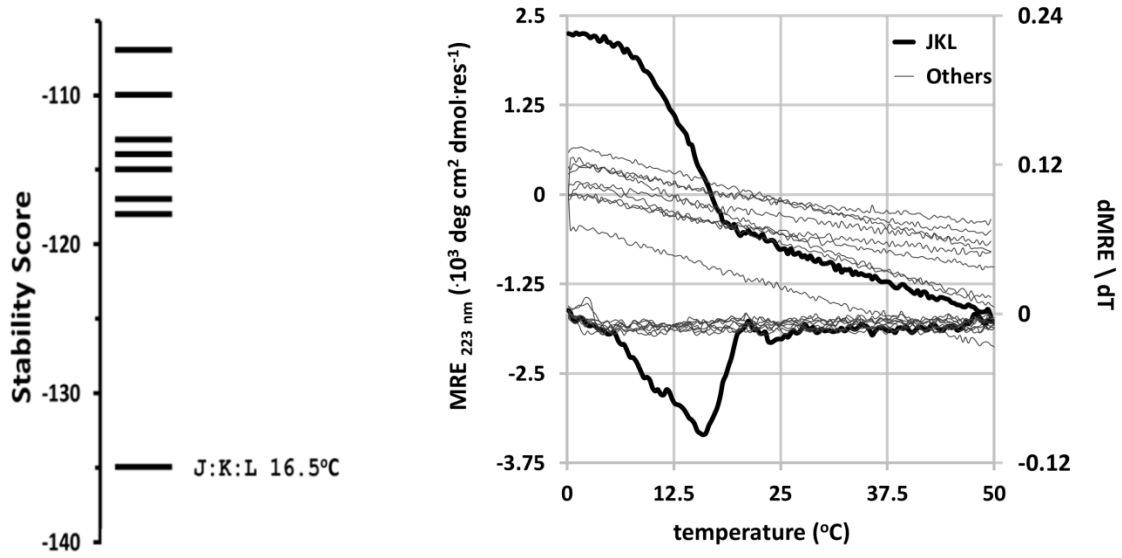


Fig. 7: Characterization of “JKL”. Temperature melting curves monitored at 223 nm of 0.2 mM “JKL” heterotrimer peptide in 10 mM Tris buffer. pH = 7.4. The 10 possible combinations of the monomers ‘J’, ‘K’ and ‘L’, were characterized. The first derivative graph indicating the melting temperature of computationally designed “JKL” (16.5 ^{o}Celcius) is also shown along with the energy bar diagram.

ABC SEQUENCE
gap : -27
energy_score : -135
num_of_favor : 21
num_of_pogs : 0

DEF SEQUENCE
gap : -17
energy_score : -133
num_of_favor : 19
num_of_pogs : 0

GHI SEQUENCE : (net charge -4 , Tm without salt 25.5°C)
gap : -18
energy_score : -136
num_of_favor : 22
num_of_pogs : 0

JKL SEQUENCE: (net charge -4 , Tm without salt 16.5°C)
gap : -17
energy_score : -135
num_of_favor : 21
num_of_pogs : 0

Acid Tecton_2: (net charge -7 , Tm without salt 19°C)
gap : -27.8
energy_score : -135.8
num_of_favor : 18
num_of_pogs : 1

Acid Tecton_1: (net charge -10 , Tm without salt 18.5°C)
gap : -29
energy_score : -131
num_of_favor : 17
num_of_pogs : 0

Table 1: Attributes of A to L trimer peptides along with Acid tecton 1 and Acid tecton 2 heterotrimers. The net charge and Tm's of charged peptides are listed next to the sequence name.

Section II : Effect of Net Negative Charge on Stability :

In order to further ascertain the exact contributions of the net charge on stability, a database of 125 previously designed trimers with similar sequence profiles was curated [**Fig. s8**]. The trimers were classified into 9 distinct sets. Set 1 and Set 8 consisted of

trimers from ‘A to L’ peptides that folded with Set 1 representing “ABC” and Set 8 representing the rest. Set 2 and 3 consisted of the 10 stoichiometric combinations of ‘Acid tecton I’ and ‘Acid tecton II’ peptides. Set 4 and 5 consisted of two heterotrimers that had been computationally designed to fold with the aid of either only axial or lateral salt bridges with the register being aided by cysteine knots. The effect of cysteine present in the sequences was subdued through the use of 2mM DDT (dichloro-diphenyl-trichloro-ethane). Set 6 peptides consisted of trimers from the Barbara Brodsky lab [43] and Set 7 consisted of trimers from Hartgerink’s lab [44]. Set 9 consisted of trimers from 1, 2, 3 and 4 triplet circular permutations of strand ‘A’ in combination with either {‘B’, ‘C’ } or { ‘E’ , ‘F’ } from Fei’s circular permutation paper [35]. Several attributes of the 125 trimers were computationally calculated and enlisted, including trimer gap in a 2 or 3 peptide ensemble (Internal trimer gap), No. of favorable attractive ionic salt bridges, No. of non-favorable repulsive ionic salt bridges, No. of favorable axial salt bridges, No. of favorable lateral salt bridges, individual net charges of the 3 composite monomer chains of the trimer, N-terminal pseudo-domain net charge, C-terminal pseudo-domain net charge and the number of ‘POG’ triplets in the trimer. Gap value for a homo-trimer was left blank and undefined.

Since “JKL” was a circular permutation of “GHI”, had only 1 salt bridge lesser than “GHI”, and yet showed considerable difference in stability compared to that of “GHI”, a new term called gradient was defined to get a better picture of the role of charge imbalance existing between N versus C terminal. Gradient denotes the charge distribution over the length of the trimer and is calculated as the sum of the absolute value of the net charge present over the left half and absolute value of the net charge present over the

right half of the trimer pseudo-domain. A high gradient indicates that the 2 halves of the terminal are highly charged. A low gradient such as in the case of "ABC" (and hence "DEF"), implies that the charges due to distribution of 'K's and 'D's, irrespective of whether they are involved in the formation of salt-bridges or not – are evenly distributed *within* the two halves of the trimer, resulting in an even spread of charges within the 2 halves of the trimer and in-turn an even spread over the whole trimer. For calculating the various enlisted attributes, in certain rare instances where the uncharged amino acid residues in the trimer were not from the set { P, O, G, K or D }, their contribution to stability was ignored (treated similar to glycine in the scoring function). For final comparison purposes, a smaller table consisting of 22 trimers from the initial 125 trimer set was generated by eliminating the following trimers: **a)** Trimers from all experiments that did not adhere to the 0.2mM total peptide concentration, **b)** Trimers that had 'POG' triplets in them, **c)** Trimers with absolute net charge greater than value 10 and **d)** Trimers with energy gap score greater than '-3' - as this implied an ambiguity in the chain association state. The Tm values for experiments conducted in no salt of trimers "CEF" and its circular permuted version "FBC" (= 'BCF' with a lower stability score), were not available. Their values were estimated from experiments conducted in the presence of 100mM salt. Both 'JKL' and Acid tecton I (named AT1_ABC) had very low or non-conspicuous MREs in phosphate buffer so their Tm values were approximated as well. The approximated values are highlighted in green in **Fig. 8a**. From the enlisted 22 trimers a final set of 13 timer's "Stability score versus experimental Tms" were plotted (Fig. 8b). The 13 trimers were chosen from the 22 initial trimers by further eliminating these

trimers of low stability: **a)** Trimers with Tm less than 5 ^{o}C and **b)** Trimers with extremely low MRE in phosphate buffer (= "JKL" and Acid tecton I).

Through linear regression, the data points on the graph were fit to a line and a squared value of the co-relation co-efficient R of '0.86' was arrived at. A new charge corrected stability score **Enew** was formulated where-in the old stability score **Eold** was corrected with the equation

Enew = Eold + α * (Absolute value of net charge of negatively charged trimers) + β * (Absolute value of net charge of positively charged trimers).

The values of $\alpha = 1$ and $\beta = 1$, improved the squared co-relation co-efficient R value the most to '0.89'. Even though the improvement from '0.86' to '0.89' may seem not too significant, we note that only 5 out of the 13 trimers under consideration were charged timers, contributing to the small change in the corrected R value. Also, since the 8 net-neutral charged peptides already adhered to a straight line with a high squared co-relation co-efficient R value of '0.86', there was little room for any further improvement in the available course-grained 'sequence only' based scoring model.

In summary, accounting for the net-charge would certainly provide for a better scoring function. An overall glance at the table in **Fig. 8a** also provided a few more useful insights such as **i)** a relatively good energy gap may also contribute to some folding (example trimer "EEF"). **ii)** Although the gradient factor (indicating change distribution) did not seem to play a role, it's note-worthy that "ABC" and "DEF" have the lowest possible gradient for a {'KOG', 'PDG', 'PKG', 'DOG' } triplet set 30 amino acids per chain trimer. This indicates uniform distribution of charge throughout the length of these

2 trimers. **iii)** The low stability of "JKL" in phosphate buffer seems to arise from a combination of factors, not excluding ones such as **a)** one less salt bridge than its similarly net charged circular permutation counter-part "GHI" and **b)** a drop in energy gap from '-27' to '-17' in comparison to a trimer with similar old charge uncorrected energy score (like '-135' of "ABC"'s).

Trimer Name	Tm no Salt	Energy	Net	Gap	Gradient	No_of_Fav	No_of_non	1st_strand_c	2nd_strand_c	3rd_strand_c	N_ter	C_ter
GHI	25.5	-136	-4	-18	4	22	0	-4	4	-4	-3	-1
ABC	29	-135	0	-27	2	21	0	2	-8	6	-1	1
JKL	16.5	-135	-4	-17	4	21	0	-4	4	-4	-1	-3
p4_EFA	24	-134	0	-22	6	20	0	-8	6	2	3	-3
DEF	24	-133	0	-17	2	19	0	2	-8	6	1	-1
AT1_ABC	18.5	-131	-10	-27.8	10	17	0	-2	-2	-6	-7	-3
p1_ABC	24	-131	0	-24	2	19	1	2	-8	6	1	-1
CEF	15.5	-126	4	-10	4	14	1	6	-8	6	3	1
p2_ABC	19.8	-125	0	-18	2	15	2	2	-8	6	1	-1
p3_EFA	17.6	-125	0	-16	10	17	2	-8	6	2	5	-5
FBC	11.5	-124	4	-7	4	14	2	6	-8	6	1	3
p3_ABC	9	-121	0	-14	2	13	3	2	-8	6	-1	1
GHG	10	-118	-4	-11	10	13	3	-4	4	-4	-7	3
IHI	4	-118	-4	-5	6	14	4	-4	4	-4	1	-5
HHG	11	-117	4	-10	4	11	3	4	4	-4	1	3
p4_ABC	4	-116	0	-9	6	11	4	2	-8	6	-3	3
HHI	4	-114	4	-7	6	10	4	4	4	-4	5	-1
p2_EFA	9	-112	0	-3	14	13	5	-8	6	2	7	-7
EFF	4	-109	4	-8	10	8	5	-8	6	6	7	-3
AT2_CCC	0	-108	0	NA	6	4	5	0	0	0	3	-3
EEF	4	-107	-10	-17	10	8	5	-8	-8	6	-1	-9
AT2_AAA	0	-107	-6	NA	12	4	5	-2	-2	-2	-9	3

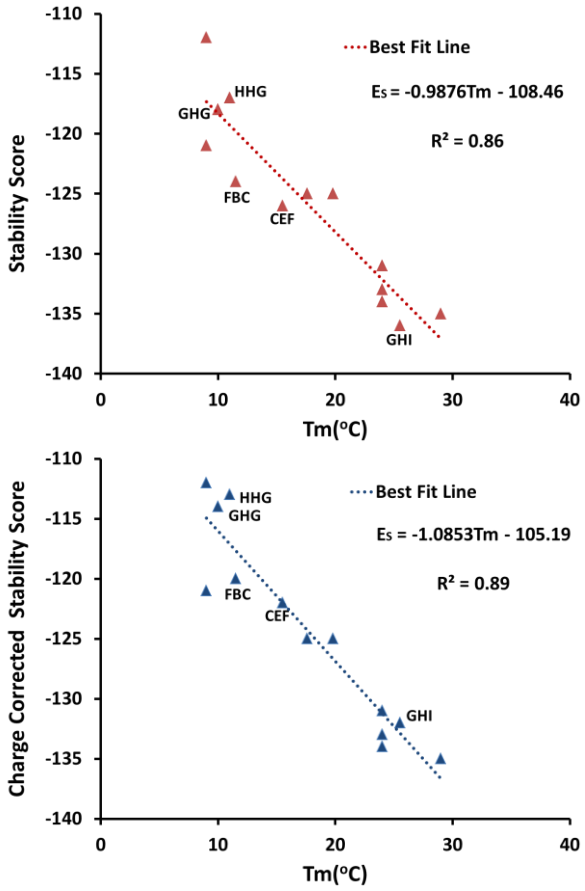


Fig. 8: Stability Scores list. Table enlist attribute of 22 different timers. Graphs show that Improved score incorporating net charge shows better fit for the 13 trimers under consideration. The two different Stability scores of each trimer versus their Experimental Tms are plotted. The 5 charged trimers are indicated by their letter code.

Section III : Effect of Density of competing states on Specificity :

To gain better insight into the Circular dichroism data of the '6 monomer peptide ensemble' experiments, an envelope of the frequency count of energy states that could potentially form from various 3 chain associations out of the 6 monomers in solution was plotted (**Fig. 9**). We notice that in **Fig 9a**, the 6 monomer set "CAEFIG" has more of its possible 6*6*6 = 216 chain association states with unique registries (leading, middle and lagging strand place-holders being treated as unique entities), towards the lower end of

the energy spectrum in comparison to that of "BCFDIG". This holds true even when the stability scores are corrected for only the trimers that are net negatively charged (**Fig 9c**). Positively charged trimers' stability score were not corrected and were retained as such. This was because the experiments were conducted in phosphate buffer and unlike the net negatively charged "JKL" trimer, the difference in behavior of positive tectons in phosphate versus tris buffer has not yet been experimentally catalogued. A worst case scenario was assumed where-in the positively charged 'competing state' trimers did not encounter a drop in stability in phosphate buffer. Thus the effective corrected energy score Enew used was :

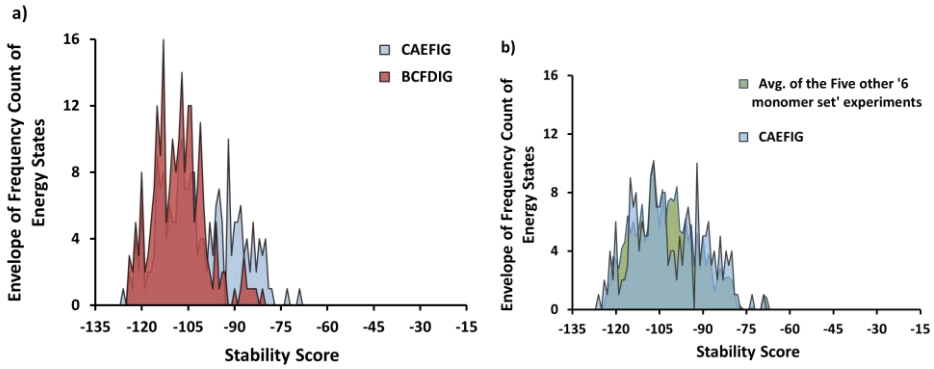
*Enew = Eold + α * (Absolute value of net charge of negatively charged trimers) + β * (Absolute value of net charge of positively charged trimers).*

with $\alpha = 2$ and $\beta = 0$.

In **Fig. 9b** the frequency counts of possible energy sates for "CAEFIG" versus the average of the frequency counts of the other five, '6 monomer ensemble' experiment sets : "ABDEGH", "ABDEHI", "ABEFGH", "BCDEGH" and "CAFDHI" is shown. It can been seen that "CAEFIG" has slightly more or equal spread of the energy states towards the lower end of the spectrum compared to the average of the rest of the experiment sets. **Fig. 9d** displays the same observation with respect to the stability score corrected for negatively charged trimers. Overall, the observation of the concentration of more states towards the lower end of the energy spectrum is in line with the **i)** stronger CD signal of "BCFDIG" versus "CAEFIG" and also **ii)** the stronger CD signal of "CAEFIG" versus the rest of the 6 monomer set experiments. Thus, we tend to get a rough estimate of the

strength of the underlying trimer states for a given peptide ensemble, and the corresponding CD signal strength can be used to inform us about the resistance offered to the formation of the designed target states “ABC”, “DEF”, “GHI” and “JKL”. At first glance, from among the seven 6 monomer ensembles, based on their energy scores and their individual signal strength alone, trimers “BCF” (rather “FBC” since “FBC” had a lower stability score than “BCF”) and “CEF”, seem to be the sole contributors to the strong signal of “BCFDIG” and “CAEFIG” respectively. However, the fact that “CEF” has a stronger CD signal than “BCF” (“CEF” > “BCF”) and “BCFDIG” has a stronger CD signal than “CAEFIG” (“BCFDIG” > “CAEFIG”) , implies that the energy density spectrum plays a significant role in the strength of the CD signal.

Since the designed target states {“GHI”, JKL”} were not as stable as the earlier {“ABC”, “DEF”} pair and since there were competing energy states that were as strong as atleast “BCF” (“FBC”, stability score = -124) if not “CEF” (“CEF” stability score = -126), we decided to not validate the specificity of level 4 heterotrimer orthogonality, ‘12 peptide ensemble’ design. Instead we choose to focus on modeling and elucidating the specificity barrier that we tend to run into, as the number of monomer components in a heterotrimer peptide ensemble increases, in the final section of our results discussion.



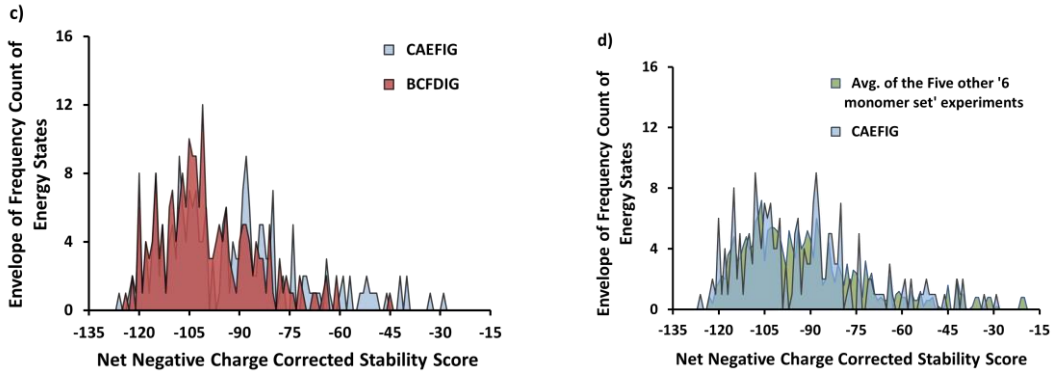


Fig. 9: An envelope of the histogram of the possible energy states present in the '6 monomer peptide ensemble' experiments. a-b) For old stability score **c-d)** for stability scores corrected for net negatively charged trimers.

Section IV : Effect of net 0 charge constraint on re-design of "GHI", "JKL" pair :

In-order to see if the design sequence attributes would have been any better if we had the additional constraint of "GHI" and "JKL" being net neutral, a graph of geometric mean of stability scores of the target states {"ABC", "DEF", "GHI", "JKL"} versus the geometric mean of the respective energy gaps in a 12 peptide ensemble was plotted. The G to L sequences were re-designed using a new more powerful protocol consisting of a combination of simulated annealing, genetic evolution and replica exchange algorithm. The details of the algorithm will be elaborated in later sections. The graph of the gap versus energy geometric mean plot is shown in **Fig. 10**. The graph indicates 2 simulation runs – one with a net 0 charge constraint and one without that was used for our synthesized sequences. Out of the 2 simulation runs, only six of the solution set sequences had net negative charges on "GHI", "JKL". This six set included the designed and currently synthesized G-L sequences for this paper. The 6 sequence sets are indicated by a "x" mark along with the rest of the possible 'A to L' sequences sets and their target

state scores corrected for net charge are also shown **Fig. 10**. Since the effect of the buffer on a positively charged trimer is not quantified, the following formulae for correction of the stability scores were used:

*Enew = Eold + α * (Absolute value of net charge of negatively charged trimers) + β * (Absolute value of net charge of positively charged trimers)*

The values of α = 2 and β = 1 were used for target state stability score calculation and α = 0 and β = 0 for competing state stability score calculation. This represents the worst case scenario wherein competing states with a high number of salt-bridges may still offer resistance to target state folding, regardless of their net charge or the buffer being used. Expect for the 6 charged sequence sets, the rest of the plotted sequence sets as are uncharged and the values of α and β are has no significance for the calculation of their target state energy score.)

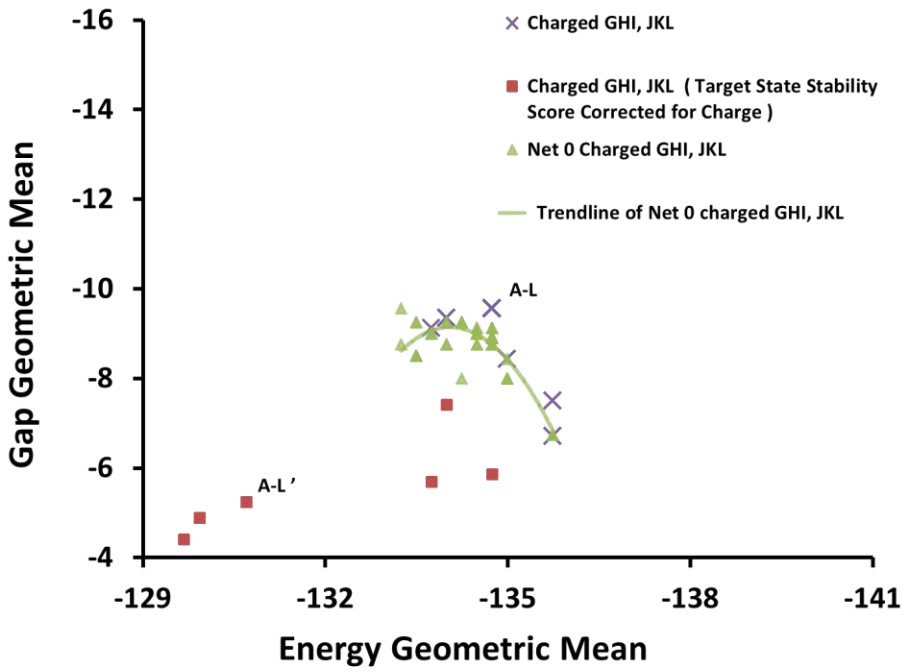


Fig. 10. Charged v/s Uncharged GHI, JKL : GHI, JKL pairs were regenerated with and without net 0 Charge constraint. The Pareto frontier for unchanged peptides is shown. The synthesized sequence is also included in the data and is labeled as A-L (A To L).

It can be seen from the graph, then when considering the old stability scores formulae which only took the number of salt bridges into account and not the net charge of the target state, a net charge of '-4' (the synthesized sequence) resulted in the best theoretical energy and gap for a 12 peptide ensemble. From the graph in **Fig. 10**, there does seem to be room for improvement by designing a net neutral "GHI", "JKL" pair that would result in a '12 peptide ensemble', level 4 orthogonality heterotrimer system. In-order to get a theoretical estimate of how well such a system would fare - and also to theoretically quantify the specificity barrier that one may run into as we add more and more orthogonality levels to the existing "ABC", "DEF" level 2 orthogonality system, a new set of in-silico simulations were run which will be discussed in the next chapter.

CHAPTER III

Stability versus Specificity Pareto Frontier Profile

Proteins and their interaction partners' co-evolution has been well documented in nature [29-33]. We wanted to estimate the extent to which we were losing out on attaining good specificity due to the presence of pre-existing "ABC", "DEF" trimers in our designed level 3 and level 4 orthogonal eco-system. To achieve this goal, the following 2 in-silico simulations were run : **i)** A new orthogonal "GHI" trimer for "ABC" trimer alone was designed to represent non-co evolved level 2 orthogonality specificity. **ii)** a separate new "GHI" trimer sequence was also designed as an orthogonal level 3 timer to the pre-existing "ABC", "DEF" pair. In both cases, simulations were run to generate a set of possible "GHI" sequences and the best sequence set having a considerably good stability and specificity score was selected. The sequences are shown in **Fig. s18**. For non-co evolved level 2 orthogonality specificity, the "GHI" stability and specificity scores were {-135, -14} and for level 3 it was {-135, -11}. This implied that in both the cases of lesser than level 4 non co-derived orthogonality/specificity, simulations showed an improvement in scores over the previously designed and tested level 4 specificity.

Section I: Pareto Definition:

For comparison of the previously designed non-co derived sequences with sequences specifically co-derived for specificity, a new set of ab-initio (from scratch) sequences for

level 1,2,3 and 4 orthogonality were computationally generated, where-in all 12 ‘A to L’ monomer chains were allowed to “co-evolve” for specificity while good stability was being targeted for through a series of triplet mutations. The simulations were run until a relatively good representative set of solutions with varying degrees of *trade-off between energy and gap* were obtained. The representative solution set – known as ***the Pareto optimal set***, *indicating that at least one objective is optimized while holding all other objectives constant*, is plotted in **Fig 1.** The sequences are shown in **Fig s18**.

Section II: Existing Search Algorithms:

For arriving at the Pareto set the problems encountered and the solutions suggested will be considered next. These procedures are common to all protein design problems where multiple objectives need to be optimized with a wide range of trade-off between the objectives under consideration. Traditionally, for locating the Pareto optimal set, multi-objective evolutionary algorithms have been used. The underlying sequence algorithms can be broadly classified as ***i) multi-objective genetic algorithms*** [45] and ***ii) non-genetic optimization methods*** [46]. We first focus on the second non-genetic approach, as it is less computationally intensive and provides for a faster approximation of the Pareto optimal set for orthogonality levels 1, 2, 3 and 4. Later, a more computationally intensive and well established algorithm that works on advancing the non-dominated front after the end of each genetic evolution generation, known as a ‘strength Pareto evolutionary algorithm 2’ (SPEA2) [47] is used, to compare our results for reference purposes and then used in combination with the previous methods. One of the unique features of this

algorithm is that it maintains an external archive of non-dominated solutions. A similar archive maybe implemented to further improve the convergence to minima speed.

Section III: A customized combination of search algorithms:

First, as mentioned before, a combination of simulated annealing, genetic evolution and replica exchange was tried. The algorithm flow chart is provided in **Fig. s10**. Briefly: 2 initial randomly generated sequence sets are chosen as the 2 parent strands. Each sequence set consists of a text block: 12 monomer sequences of 30 amino acid residues each written down in 12 lines (12 when considering level 4 heterotrimer orthogonality) . The two parent sequence sets are then subjected to alternate cycles of genetic evolution and replica exchange for every 50 'trials', with each trial defined to consist of a pre-defined number of generations. Every generation spawns two new children sequence sets from the parent set. Each generation consists of 500 iterations of simulated annealing cycles. The genetic evolution part is incorporated into the algorithm as follows: a) For every 3rd trials during the genetic evolution phase, the left half of the 1^{st} parent and the right half of the 2^{nd} parent sequence block is mixed to give rise to a child sequence set. Similarly, every 6^{th} trial, the right half of the 1^{st} parent and the left half of the 2^{nd} parent sequence block is mixed to give rise to another child. This is done in order to evolve towards a child sequence set that provides a better specificity profile as the salt bridge atteactions and disruptions of the left half and the right half of the parent sets are preserved. The overall top and bottom half of the sequence blocks are interchanged every 17^{th} and 34^{th} trial of the genetic evolution phase, to evolve towards children sequences

that provide better stability profile. Block mutation in incorporated into the algorithm every 11^{th} trial in order to help navigate the rough energy landscape. For the replica exchange part, the standard protocol is followed wherein the 2 parent chains are subjected to two different constant temperatures (High and low) every 500 cycles of simulated annealing (1 trial) and the temperatures are swapped at the end of each trial. Two sets of high and low temperature pairs were tried ({ 1e2, 1e5}, and {1e1, 1e20}) and no significant difference or pattern was observed in the way the solution sequences were scattered along the Pareto frontier – indicating that selection of the constant temperatures to run the simulated annealing was not too much of a concern in obtaining a good solution set. The convergence of solutions towards the Pareto front after a pre-defined number of iterations for Simulated annealing, multi-start simulated annealing (multiple starting points in the possible parent sequence space to overcome difficulties in navigating a rough landscape), Genetic evolution algorithm and Replica exchange both alone and in combination are shown in **Fig. s11**. The convergence graph indicates that a combination of the methods outperforms the individual methods in estimating the Pareto optimal set. This could primarily be attributed to the fact that replica exchange allows for a breadth search (more conformations searched) while genetic evolution allows for a depth search (the child generation has a better fine-tuned parameter than the parent generation). Also, replica-exchange might overcome the problem of disruption arising from the cross-over operation of the genetic evolution algorithm.

Section IV: Cost Function as a weighted sum of Multiple Objectives

Protein and peptide design consists of defining not just good search algorithms that enable efficient sampling of the rough energy protein folding landscape, but also the definition of a suitable cost function that needs to be minimized. In most cases, a straight forward implementation would be a weighted sum of all the objectives that need to be minimized. In our case the objectives to be optimized for a well foldable single hetero-trimer were mainly energy and gap. For multiple orthogonal heterotrimers, instead of optimizing the individual energies and the gap between the target state trimer and its respective nearest competing state or states, the mean of the individual energies and the mean of the individual gaps was considered in-order to reduce the number of parameters that require optimization. For the synthesized peptides A to L, the arithmetic mean of the energies and the arithmetic means of the gap had been considered. However, for the co-derived set, the geometric mean was used in place of the arithmetic mean. This was done in order to ensure that for a given cost function's value, there would be lesser variance amongst the individual energies and gap values obtained for the set of individual target state trimers. We note here that for that for the synthesized A-L sequence, only GHI and its permuted version JKL's parameters were being minimized. However, for the co-derived case, all four trimer energies need to be optimized simultaneously, leading to a lot of variance in the obtained set of energies and gaps for a given cost function's value. Thus, the new cost function to be considered was:

$$C1 = w1 * E_{GEOMETRIC\_MEAN} + w2 * Gap_{GEOMETRIC\_MEAN}$$

The Pareto frontier was plotted with w2 = 9 – w1 and w1 varying from 0 to 9 in steps of 1. The resulting plot is shown in **Fig. s15.** Orthogonal level 4 data has a more scattered appearance than Orthogonal level 1 data indicating that the cost function landscape dependent on the energy and gap is more fragmented or rough with a lot of local energy minima's spread out and appearing near the Pareto front region, separated by high energy barriers. Thus, unlike barrier trees [48] and disconnectivity graphs and a few other methods [49-50] which are used as the traditional means to visualize the cost function landscape, in our case, the Pareto frontier also provide a good picture of the clustering of local minima that are separated by high barriers leading to a non-convergence to a fewer set of minima points. In order to further truly differentiate the minima, two more objectives were added to the cost function. The first was the geometric mean of internal trimer gaps of each target state of an orthogonal set. As can be seen from **Fig. s12**, better internal trimer gaps lead to worse overall gaps and vice-versa. The second additional objective was the energy density factor. The definition of energy density factor (EDF) is provided next.

Section V: Redefining the Specificity Score

In order to decouple the Target state energy from the strength of the competing states, in the Boltzmann Specificity factor, the target state was replaced by the known lowest possible energy minima state, which corresponds to an stability score of -139 and it is a constant for our system. The '-139' score corresponds to the trimer state with the maximum number of stabilizing salt bridges that can be squeezed into the fixed length trimer peptide.

Now specificity is defined by the Boltzmann factor, given by

$$\textbf{Boltzmann Specificity factor} = e^{-ETarget\_state} / \sum e^{-ECompeting\_state}$$

$$= 1 / (1+ \sum e^{-EGap})$$

Minimizing the specificity factor is equivalent to minimizing the quantity $1 / \sum e^{-EGap}$

Thus EDF can be defined as, Energy_Density_Factor, EDF = $- 1 / \sum e^{-EGap}$ with the gaps being defined as the gaps between the *known* lowest possible energy/stability score and the respective competing state for each competing state in the ensemble of possible association states. The lowest possible energy/stability score for our system is '-139'. The EDF score tends to be exponential in nature and has a large range of values. For easier search of the rough cost function landscape, the landscape can be smoothened/flattened by using negative log negative EDF. Although the resolution is lost, this change still incorporates the cumulative effect of the density of states allowing for easier transitioning of barriers between local minima. Intuitively, there are 2 ways to interpret the EDF quantity: 1) Just as the Boltzmann factor, it can be interpreted as the projection of histogram of energy of competing association states over an exponential function (the dot product of 2 histograms indicates the measure of similarity between 2 distributions – with the 2 histograms in our case being that of the competing states distribution and the discrete exponential function). 2) That it is similar in aspect to the explanation provided by Sarel J. Fleishman's group in their fuzzy logic paper [51], and that it represents a way of incorporating the all the individual contributions of different competing states to the target state specificity in one quantity. EDF and Boltzmann quantity are proportional. Using EDF instead of Boltzmann factor for objective 3 for measuring specificity keeps the objective to be of the right "sign" (among positive,

negative, zero sign options), allows it to be in a similar format, resolution and range as that of objective 1 (between -100 and 0 instead of 0.999 and 0) and improves resolution by eliminating dependence on score of Target state stability (objective 1). This independence of score is achieved despite the fact that the underlying peptide sequences used for calculation of objective 1 and objective 3, is the same (In other words, mutating the sequence to alter one objective would affect the other).

Thus our new cost function is:

$$\mathbf{C1 = w1 * E_{GEOMETRIC\_MEAN} + w2 * Gap_{GEOMETRIC\_MEAN}}$$

$$\mathbf{+ w3 * Internal\_Trimer\_Gap_{GEOMETRIC\_MEAN}}$$

$$\mathbf{+ w4 * -log\ (-Energy\_Density\_Factor_{GEOMETRIC\_MEAN})}$$

Section VI: Simulation Results

The resulting Pareto frontier graph using the new cost function and the search algorithm not including SPEA2 component is shown in **Fig s15a**. This was compared with the standard "SPEA2" algorithm results (**Fig s15b**). The SPEA2 code was downloaded from author's website [45]. Replica exchange method seemed to work better for the same number of mutation cycles – most likely because the landscape is rough with high energy barriers that are difficult to navigate through genetic evolution alone or traversing down a gradient descent alone moving from one non-dominated front to the next. Besides constraint handling [52] for SPEA2 hasn't been explicitly discussed and it was not easy to implement a quick solution as one would need to cross high energy barriers that would in-turn necessitate staying in the infeasible solutions regions for long periods of simulation time. Instead, the SPEA2 algorithms concept of using a separate non-

dominated solutions archive was incorporated for obtaining a well distributed Pareto optimal solution set. **Fig. 1, 2** shows the final Pareto frontier graphs using a combination of replica exchange, genetic evolution and simulated annealing and spea2. As can be seen from the graphs, increasing the orthogonality level leaves very little room for improving specificity and stability.

Plotting the energy density spectrum of the competing states shows that synthesized peptide had the same spectrum as that of an example level 4 orthogonal sequence set lying on the Pareto frontier that has a similar energy and gap geometric mean profile (**Fig. 3**) .

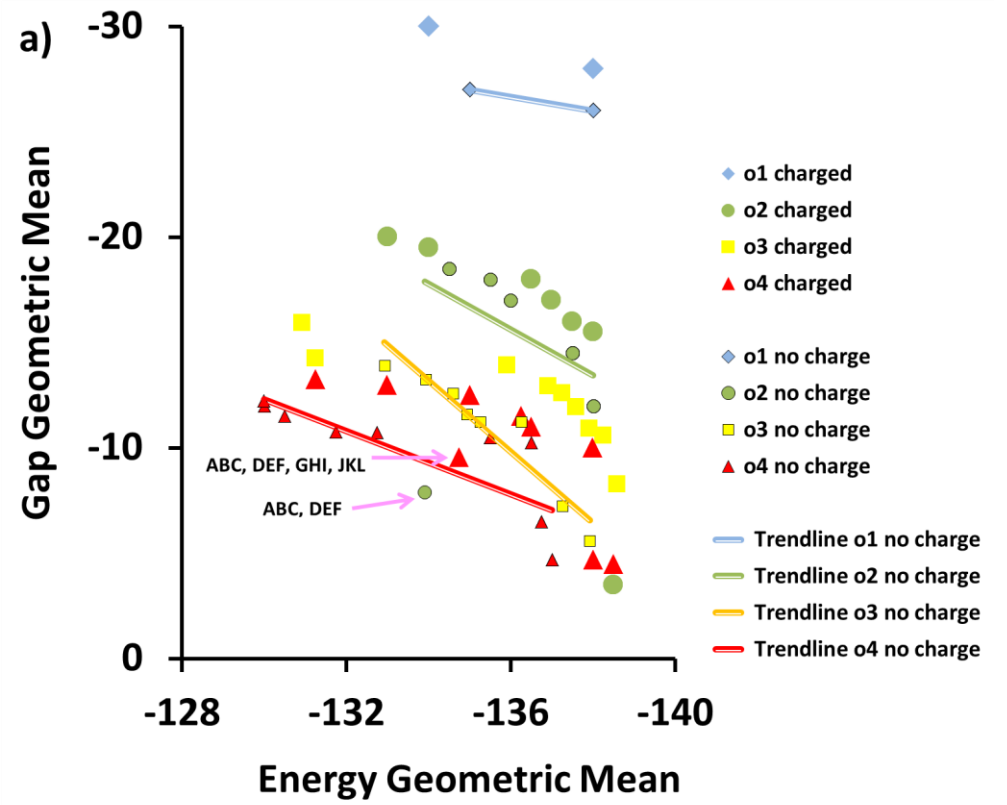
SPEA2 picks non-dominated fronts without using weights or a single aggregated objective cost function. This results in sequence data that may also include values that lie outside the range of solutions of practical interest - solution sets that do not fold and only hold a theoretical significance (such as competing state being more stable than the target state as would the case for Gap score < 0). Only Pareto points less than E < -127 lie in the region of interest as anything else would not fold into a trimer in solution. The data point having E = 21, G = 120 extreme scores was not generated in **Fig. s15b** since those extreme minima sequence points are probably not reachable by SPEA2 alone due to high energy barriers.

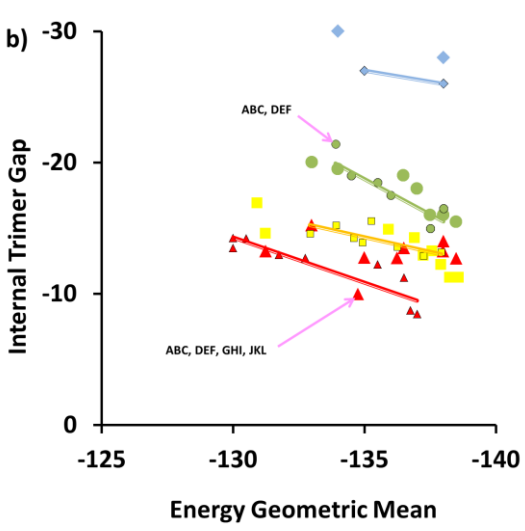
The final code contained a combination of SPEA2 and replica exchange. Constraint handling for SPEA2 was done by archiving points generated by Replica Exchange every

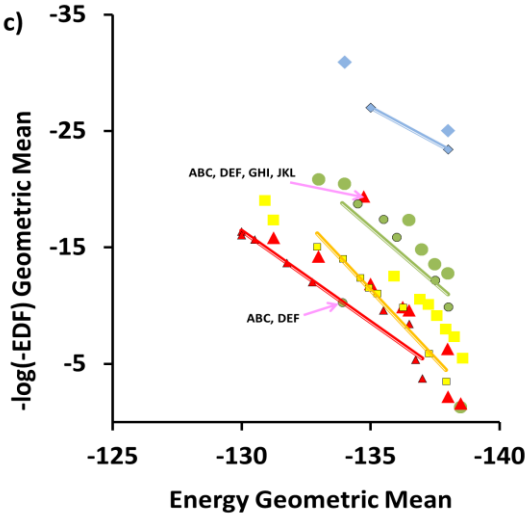
few predefined number of mutation cycles. The cost function used in replica exchange was C = w1*Energy + w2*Gap + w3*Internal_Trimer_Gap + w4*(-log(-EDF)) with weights { w1, w2, w3, w4 } = { 9 - α, α, 0.7 α, 0.7 α }, α varying from 0 to 9 in steps of 1.

Constraint handling can also be done by discarding points that are not of interest.

Replica exchange is known to work well for landscapes that have large basins of attraction [53] and it works well here. Improvements to Replica exchange has been suggested in literature and these improvements may influence the quality of the results in the future for a more than 2 objective minimization problem [54].


a)
Gap Geometric Mean
-30
-20
-10
0
-128
-132
-136
-140
Energy Geometric Mean
ABC, DEF, GHI, JKL
ABC, DEF
o1 charged
o2 charged
o3 charged
o4 charged
o1 no charge
o2 no charge
o3 no charge
o4 no charge
Trendline o1 no charge
Trendline o2 no charge
Trendline o3 no charge
Trendline o4 no charge


b)
Internal Trimer Gap
-30
-20
-10
0
-125
-130
-135
-140
Energy Geometric Mean
ABC, DEF
ABC, DEF, GHI, JKL


c)
-log(-EDF) Geometric Mean
-35
-25
-15
-5
-125
-130
-135
-140
Energy Geometric Mean
ABC, DEF, GHI, JKL
ABC, DEF

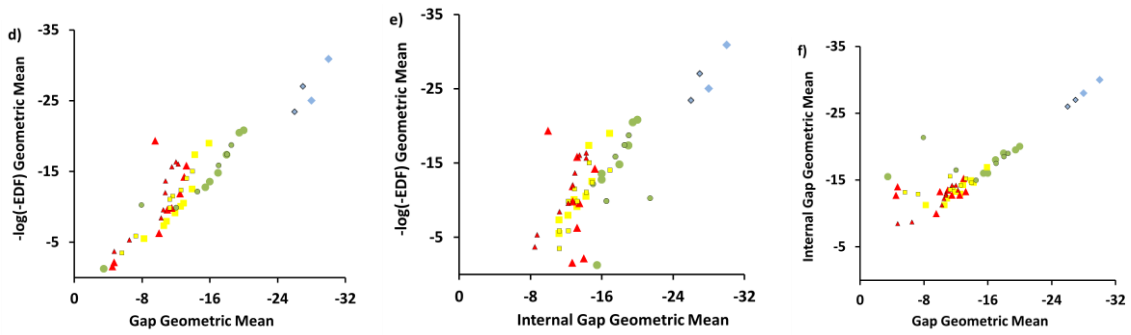


Fig. 1a: Pareto frontier 2 objectives energy and gap geometric mean optimized: Two sets of simulations are run, one with no charge constraint and another with a net zero charge constraint on the target states. The results of the simulation are shown for 4 orthogonal levels. The first 3 graphs show **a)** Gap v/s energy geometric mean **b)** Internal trimer gap v/s energy geometric mean and **c)** Negative log negative Energy density factor v/s energy geometric mean. The simulation sequence sets containing net zero charged target states have a black border and are seen to have worse trade-offs than the sequence sets containing the net zero charge target states. To plot the true Pareto front, solutions from SPEA2 archive, with archive size 20 were used. The effective archive size was 10 since the lowest point visited up to the current iteration was added twice at the end of each generation in-place of the spawned 2 offspring sequences for faster convergence (lesser diversity – see flowchart). Replica exchange was used to move from one non-dominated front to the next with the SPEA2 algorithm archiving Pareto optimal points lying on the Pareto front at regular intervals. The cost function used in replica exchange was C = w1*Energy + w2*Gap + w3*Internal_Trimer_Gap + w4*(-log(-EDF)) with weights { w1, w2, w3, w4 } = { 9 - α, α, 0.7 α, 0.7 α }, α varying from 0 to 9 in steps of 1. Two trials of replica exchange were run with all trial outputs being archived to the same external SPEA2 archive. Two additional data points indicating "ABC, DEF" and "ABC,DEF,GHI,JKL" were manually included in the graph output as labeled. Two data points that were net neutral changed but were generated during "no net neutral charge constraint on target state simulation run" were excluded inorder to better visualize non-dominated fronts.

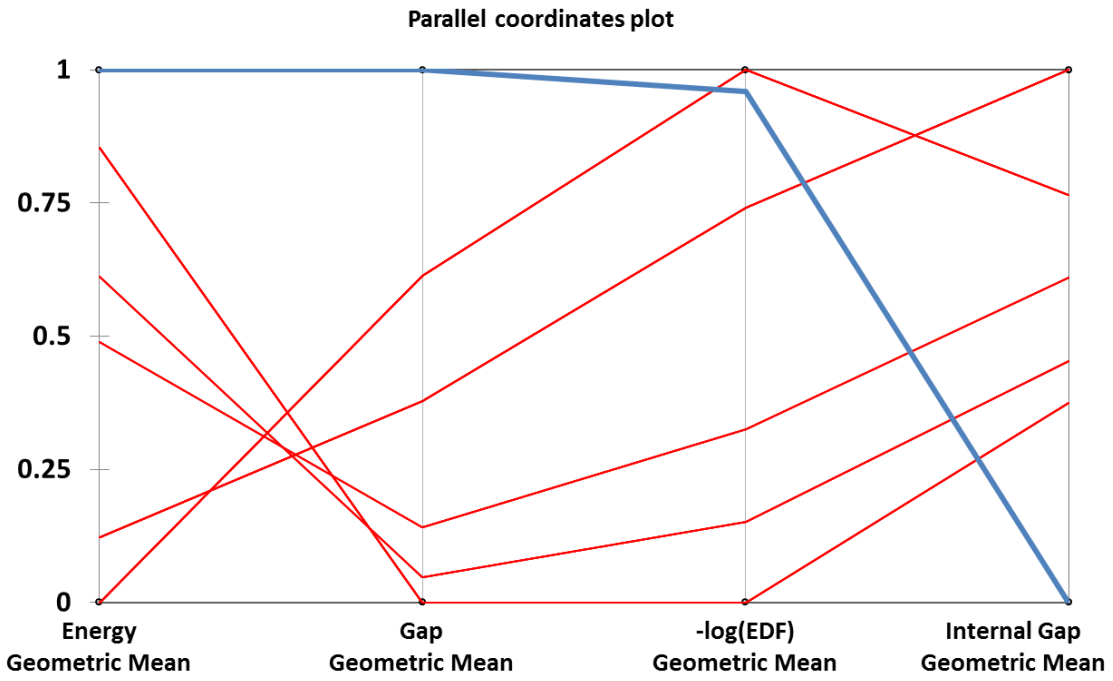


Fig. 1b: Scaled parallel coordinates plot for orthogonal level 2 for net zero charge constraint when only 2 objectives energy and gap geometric mean are optimized. The one indicated in blue (with energy score at 1) is the synthesized orthogonal level 2 "ABC, DEF" pair.

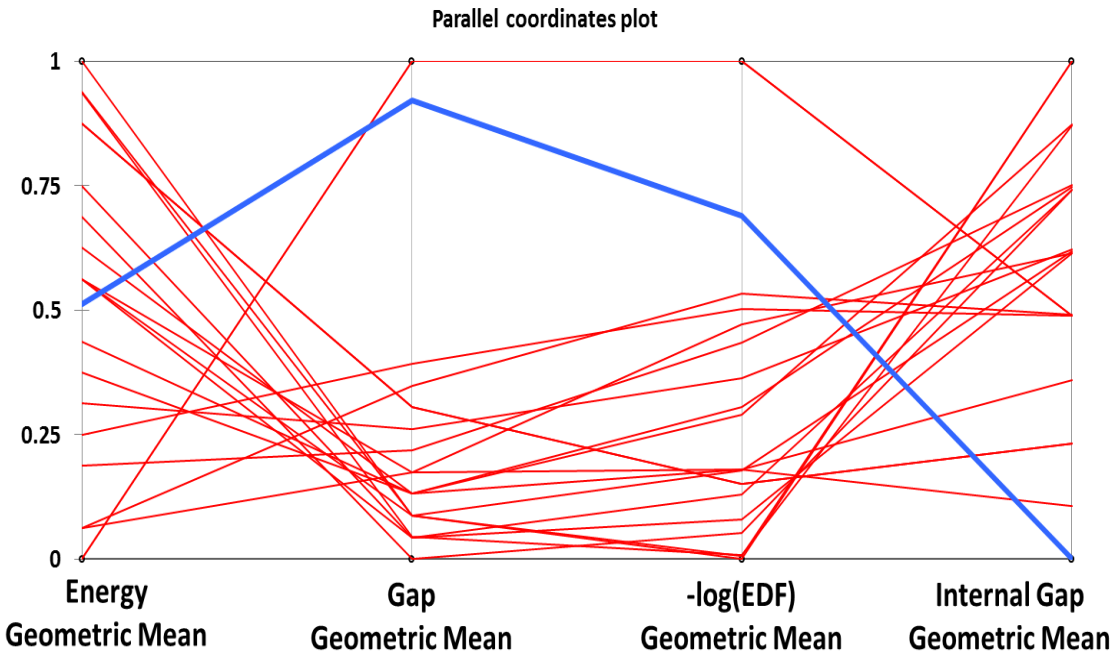


Fig. 1c: Scaled parallel coordinates plot for orthogonal level 2 for net zero charge constraint when 4 objectives are optimized. The one indicated in blue (with energy score at 0.5) is the synthesized orthogonal level 2 "ABC, DEF" pair. The distribution/spread of the 4 objectives' values with respect to one another can be visualized easily by the plot.

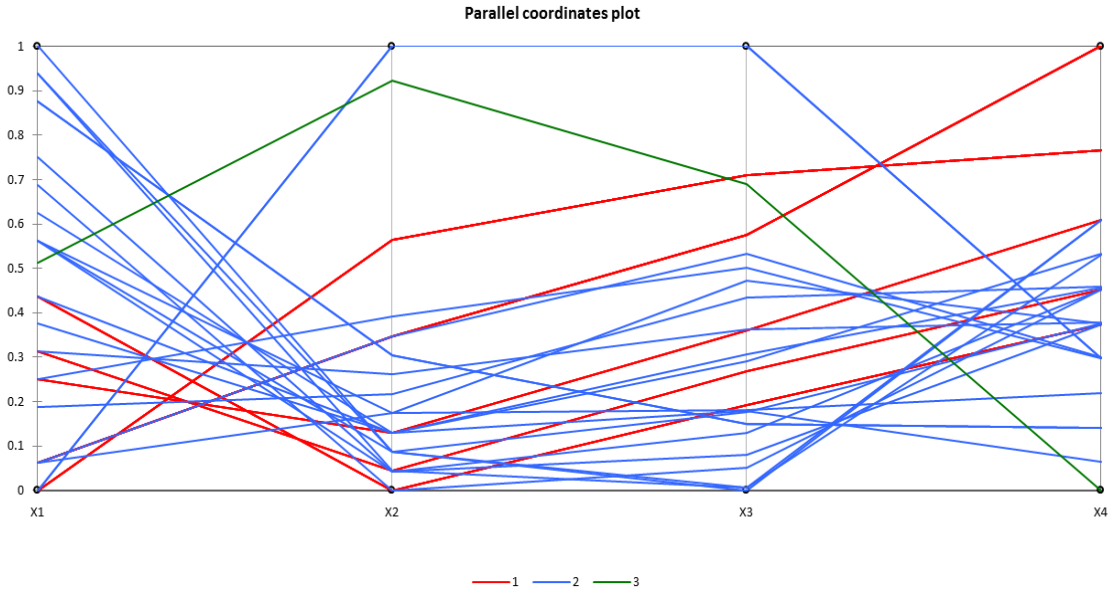


Fig. 1d: Scaled parallel coordinates plot for orthogonal level 2 for net zero charge constraint when 2 objectives are optimized (red) *versus* when 4 objectives are optimized (blue). The one indicated in green (with energy score at 0.5) is the synthesized orthogonal level 2 "ABC, DEF" pair.

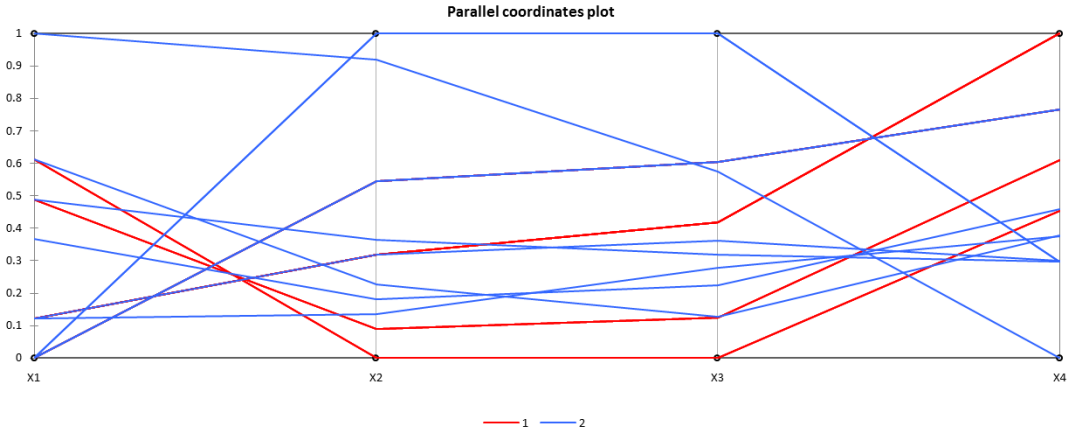
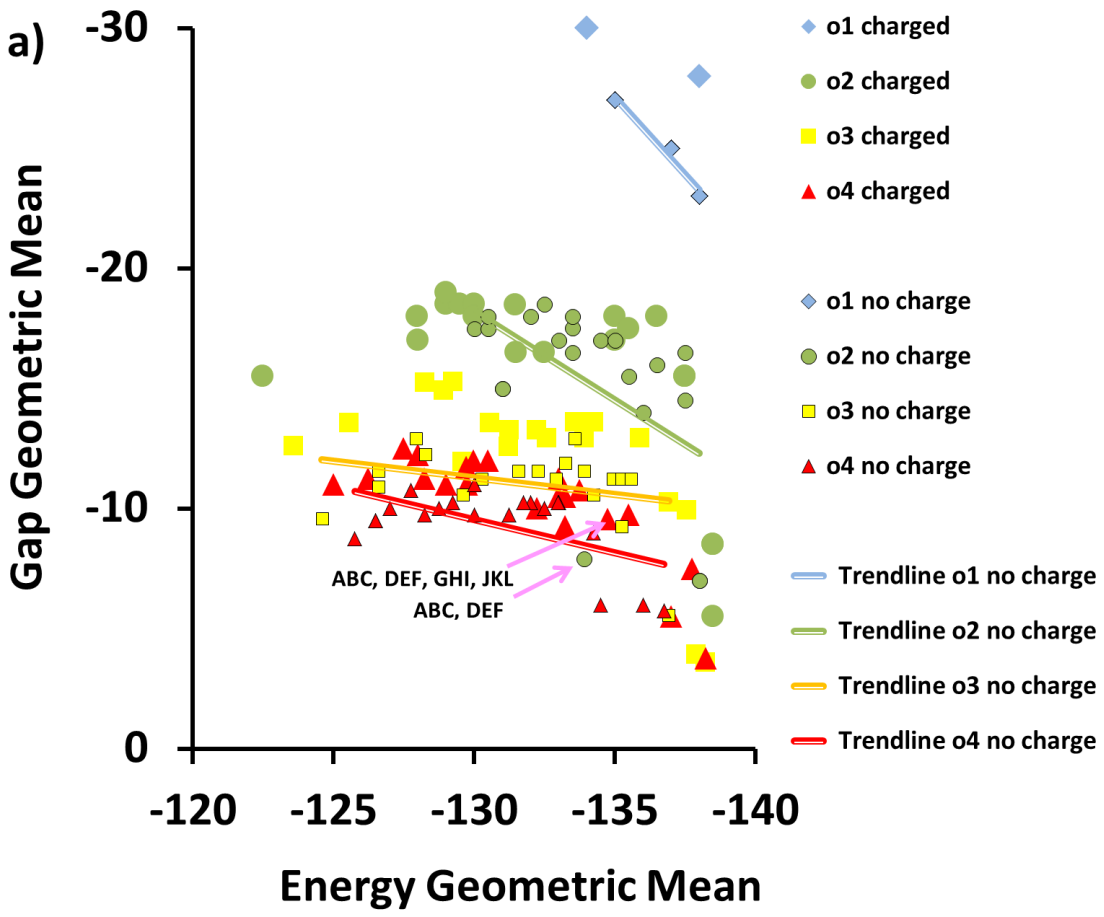
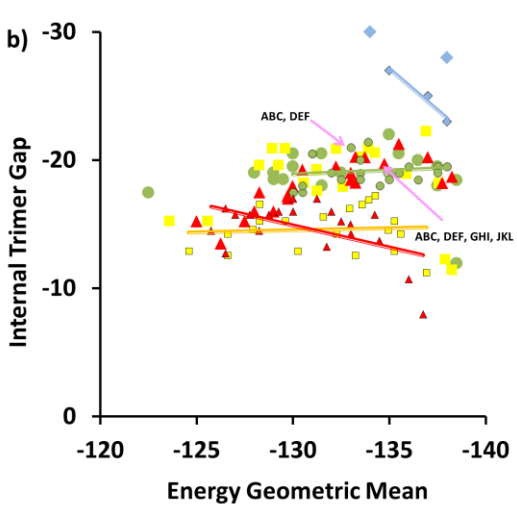
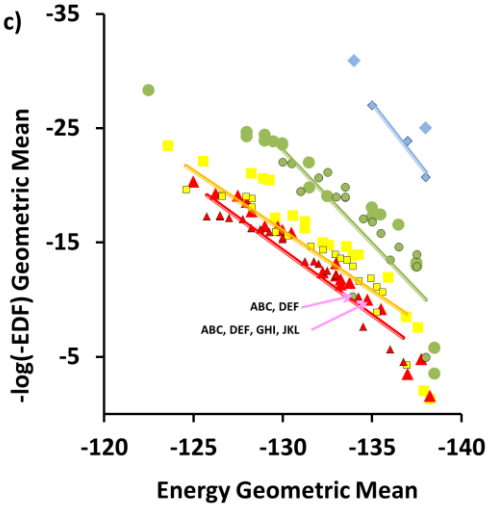


Fig. 1e: Scaled parallel coordinates plot for orthogonal level 2 for net zero charge constraint when 2 objectives are optimized (red) *versus* when 4 objectives are optimized (blue) : Only points with Energy Mean < -135 shown.


a)
Gap Geometric Mean
-30
-20
-10
0
-120
-125
-130
-135
-140
Energy Geometric Mean
o1 charged
o2 charged
o3 charged
o4 charged
o1 no charge
o2 no charge
o3 no charge
o4 no charge
Trendline o1 no charge
Trendline o2 no charge
Trendline o3 no charge
Trendline o4 no charge
ABC, DEF, GHI, JKL
ABC, DEF


b)
Internal Trimer Gap
-30
-20
-10
0
-120
-125
-130
-135
-140
Energy Geometric Mean
ABC, DEF
ABC, DEF, GHI, JKL


c)
-log(-EDF) Geometric Mean
-35
-25
-15
-5
-120
-125
-130
-135
-140
Energy Geometric Mean
ABC, DEF
ABC, DEF, GHI, JKL

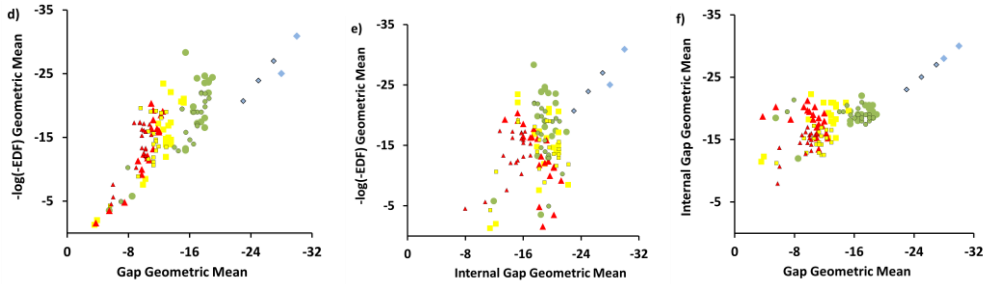


Fig. 2a: Pareto frontier all 4 objectives optimized: Two sets of simulations are run, one with no charge constraint and another with a net zero charge constraint on the target states. The results of the simulation are shown for all 4 orthogonal levels. The first 3 graphs show a) Gap v/s energy geometric mean b) Internal trimer gap v/s energy geometric mean and c) Negative log negative Energy density factor v/s energy geometric mean. The simulation sequence sets containing net zero charged target states have a black border and are seen to have worse trade-offs than the sequence sets containing the net zero charge target states. To plot the true Pareto front, solutions from SPEA2 archive, with archive size 20 were used. The effective archive size was 10 since the lowest point visited up to the current iteration was added twice at the end of each generation in-place of the spawned 2 offspring sequences for faster convergence (lesser diversity – see flowchart). Replica exchange was used to move from one non-dominated front to the next with the SPEA2 algorithm archiving Pareto optimal points lying on the Pareto front at regular intervals. The cost function used in replica exchange was C = w1*Energy + w2*Gap + w3*Internal_Trimer_Gap + w4*(-log(-EDF)) with weights { w1, w2, w3, w4 } = { 9 - α, α, 0.7 α, 0.7 α }, α varying from 0 to 9 in steps of 1. Two trials of replica exchange were run with all trial outputs being archived to the same external SPEA2 archive. Two additional data points indicating "ABC, DEF" and "ABC,DEF,GHI,JKL" were manually included in the graph output as labeled. Two data points that were net neutral changed but were generated during "no net neutral charge constraint on target state simulation run" were excluded in order to better visualize non-dominated fronts.

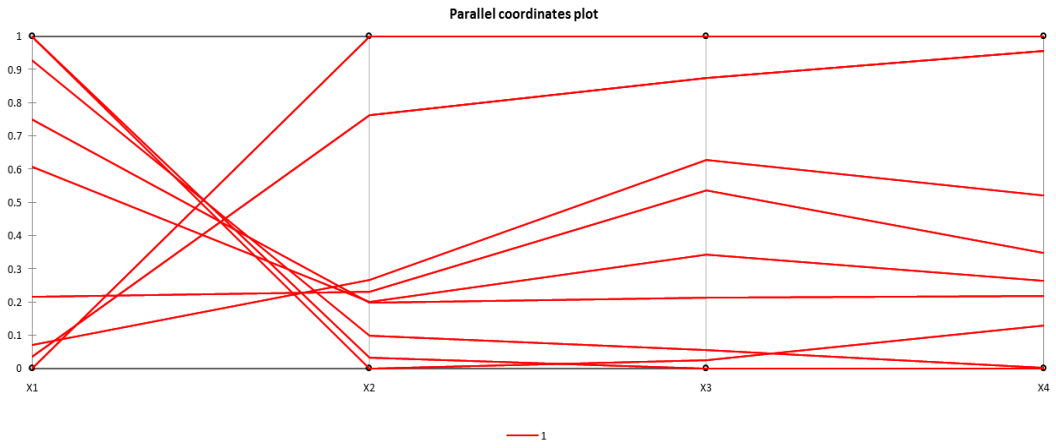


Fig. 2b: Scaled parallel coordinates plot for orthogonal level 4 for net zero charge constraint when only 2 objectives - energy and gap geometric mean - are optimized.

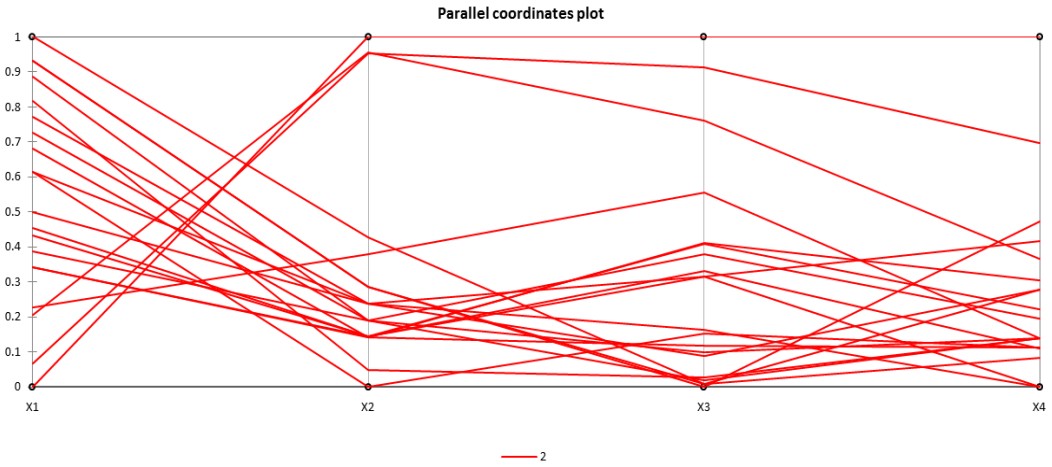


Fig. 2c: Scaled parallel coordinates plot for orthogonal level 4 for net zero charge constraint when 4 objectives are optimized. The distribution/spread/trade-off of the 4 objectives' values with respect to one another can be visualized easily by the plot.

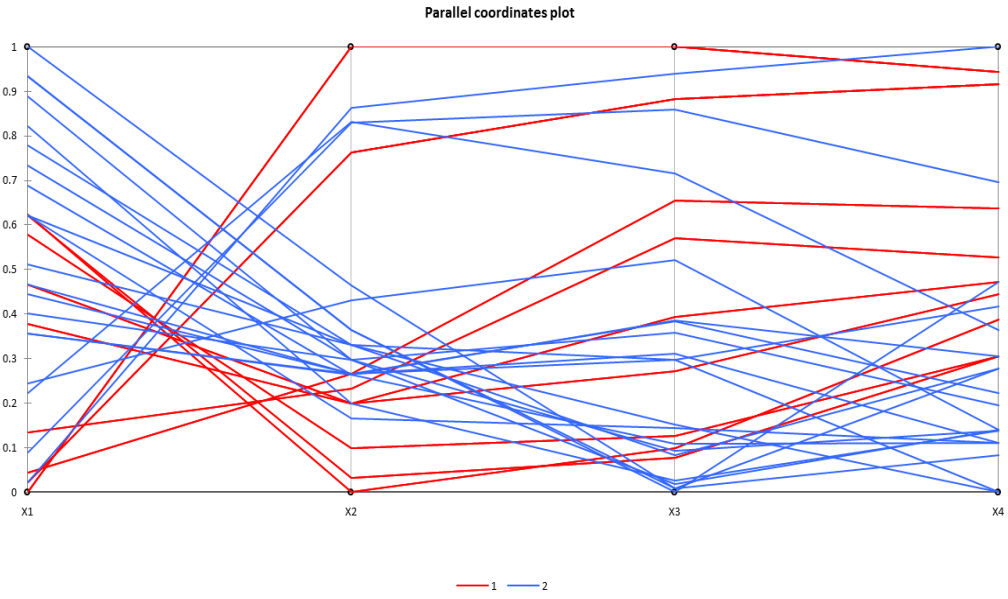


Fig. 2d: Scaled parallel coordinates plot for orthogonal level 4 for net zero charge constraint when 2 objectives are optimized (red) *versus* when 4 objectives are optimized (blue).

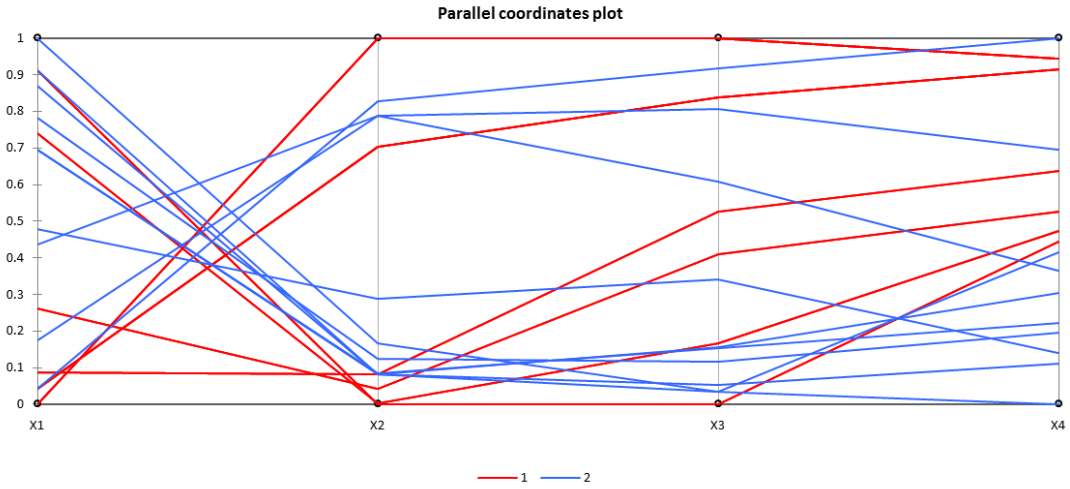
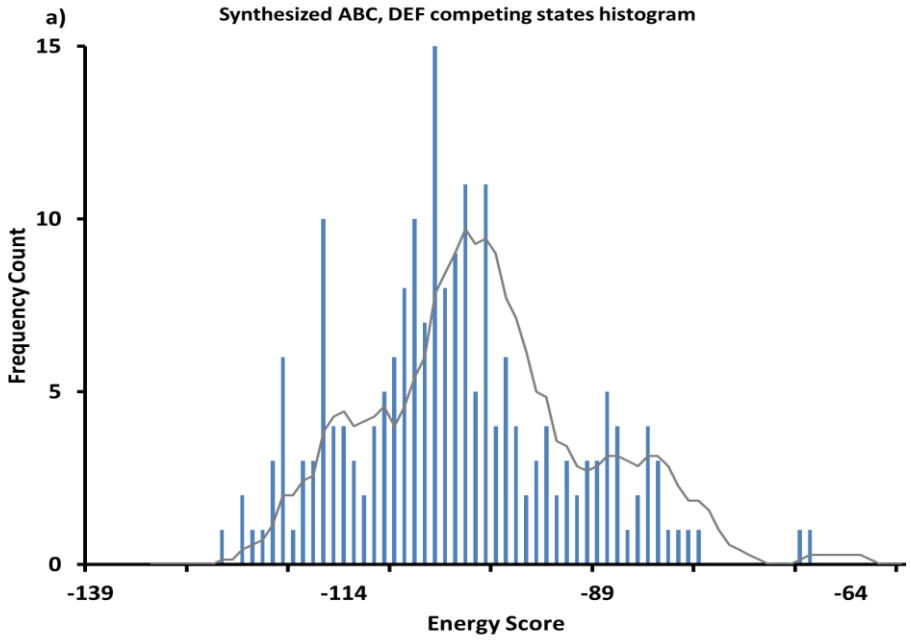


Fig. 2e: Scaled parallel coordinates plot for orthogonal level 4 for net zero charge constraint when 2 objectives are optimized (red) *versus* when 4 objectives are optimized (blue) : Only points with Energy Mean < -131 shown.

X1 = Energy, X2 = Gap, X3 = -log(EDF), X4 = Internal Gap geometric mean.



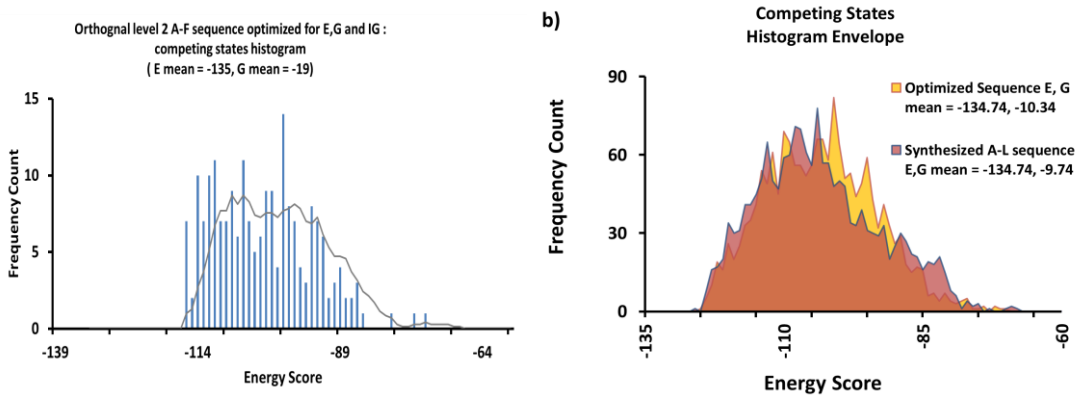


Fig 3: Competing energy density spectrum a) For level 2 orthogonality: "ABC", "DEF" pair versus co-derived pair. **b)** For level 4 orthogonality: Synthesized A-L versus a 4 objective co-optimized sequence with comparable internal trimer gap. It can be seen that we are running into the specificity barrier for level 4 orthogonality.

Section VII : Heterotrimer orthogonality v/s Homotimers Orthogonality :

Homo-trimers allowed for better gap – heterotrimers allowed for better energy. See **Fig. s16.**

Section VIII: Using POG to stabilize the timers

A Pareto frontier graph was plotted to see if have a 'POG' at each end of the timer would provide for raising the stability profile. However, as can be seen from **Fig. s17,** the specificity of the timers falls considerably. Only 'POG's at the terminal end of the peptides was considered as the effect of 'POG' at any other location along the monomer chain is as yet unknown.

Section IX : Specificity limitations :

Although we suspect the designed GHI or JKL would be orthogonal to ABC or DEF trimers alone, we did not feel the necessity to test it as it had already been shown previously that orthogonality level 2 can achieved through the ABC and DEF trimers alone. For similar reasons, the other possible level 3 orthogonality obtainable through the combinations of {“ABC”, “DEF”, “JKL”}, {“ABC”, “GHI”, “JKL”} and {“DEF”, “GHI”,”JKL”} were not tested.

Section X: Conclusion:

Based on our discussion, the following conclusions can be drawn:

1) Boltzmann factor is necessary to be optimized for large ensembles in-place of only the gap that was considered previously.

Cost function = Energy or Stability Score + Boltzmann Specificity Score.

Therefore on decoupling Energy and Specificity scores,

Cost function = Energy + Energy density Factor of Competing states

For a given energy, how would “distribution spread” of the competing states spectrum effect the specificity? For a given energy and gap, wider spread (larger standard deviation) should result in better specificity.

Overall, the CD signal MRE signal is directly proportional to the underlying stability of individual trimers. Heterogeneity appears to be forcing more monomers to exist in a trimer state for the same total peptide concentration. Thus the increase in CD signal could be a result of both individual trimer stability and the ensemble state stability.

2) Through analyzing the Pareto frontier simulation – conclusion can be drawn that there is room for improved orthogonality level 2 trimers for design. Design may further be also improved by inclusion of POGs to stabilize marginally stable charged trimers.

3) Design of orthogonal homotimers may be easier than heterotimers due to lesser heterogeneity (smoother folding landscape). There are also lesser competing states so the Boltzmann specificity is higher in the latter situation.

CHAPTER IV

Fitness Landscape Analysis:

The study of predictability of evolution and the study of mutational fitness landscape have been of considerable interest [55]. A study was conducted to assess the differences in the resulting solution sequence sets from among the different algorithms being employed for sequence search. Specifically, the results from sequence conformational sampling using replica exchange (parallel tampering) simulation versus the results from approaches that employ the method of deriving children sequences from archived parent generation sequences was studied. The analysis approach and its results are described next.

Section I: Analysis Method

The value of the one dimensional cost function when moving from one Pareto minima to the next through a series of triplet mutations was plotted (**Fig. 1a**). This movement path was traced a 1000 times over and the average of the path traced was plotted (**Fig. 1b**). By observing the minimum energy barrier height that was required to be overcome when moving from one point to the next from during these 1000 mutational paths traces generation, a barrier tree was constructed. The resulting barrier trees and fitness landscapes were studied.

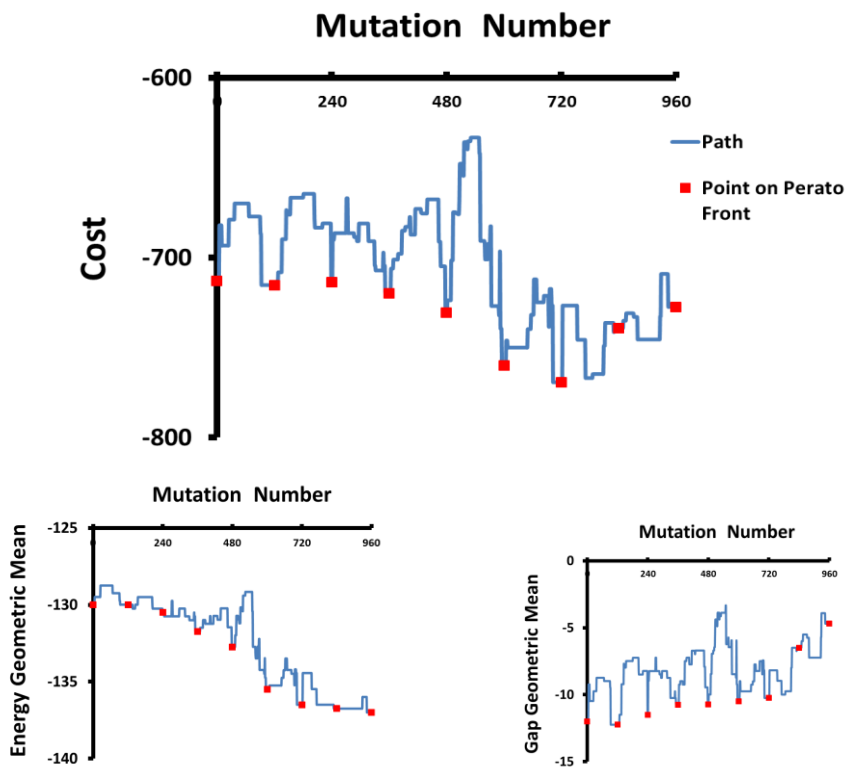


Fig 1a: A Single Path traced from one Pareto optimal point to the next through a consecutive series of 120 triplet mutations. Pareto data points of 2 Objective, orthogonality level 4 with a net zero charge constraint was chosen for analysis. Cost = 0.5*Stability Score value + 0.5*Specificity Score value. Energy geometric mean represents the Stability Score value and Gap geometric mean represents the Specificity Score value.

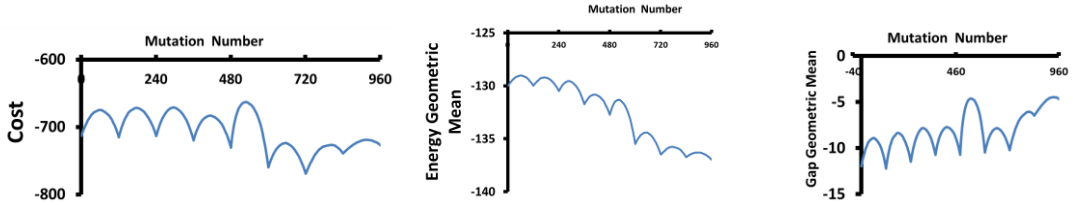


Fig 1b: Search function Path trace. The average trace of 1000 'triplet mutational' paths is indicated. Each search path starts at one Pareto point and ends at the adjacent Pareto point. Each Pareto point represents a 120 triplet sequence and is a minimum in the search landscape. A randomly generated sequence containing all the numbers 0 to 119 is used to make 120 mutations to hop from one Pareto point to the next Pareto point. This process is repeated 1000 times and the average of the 1 dimensional cost value trace is plotted.

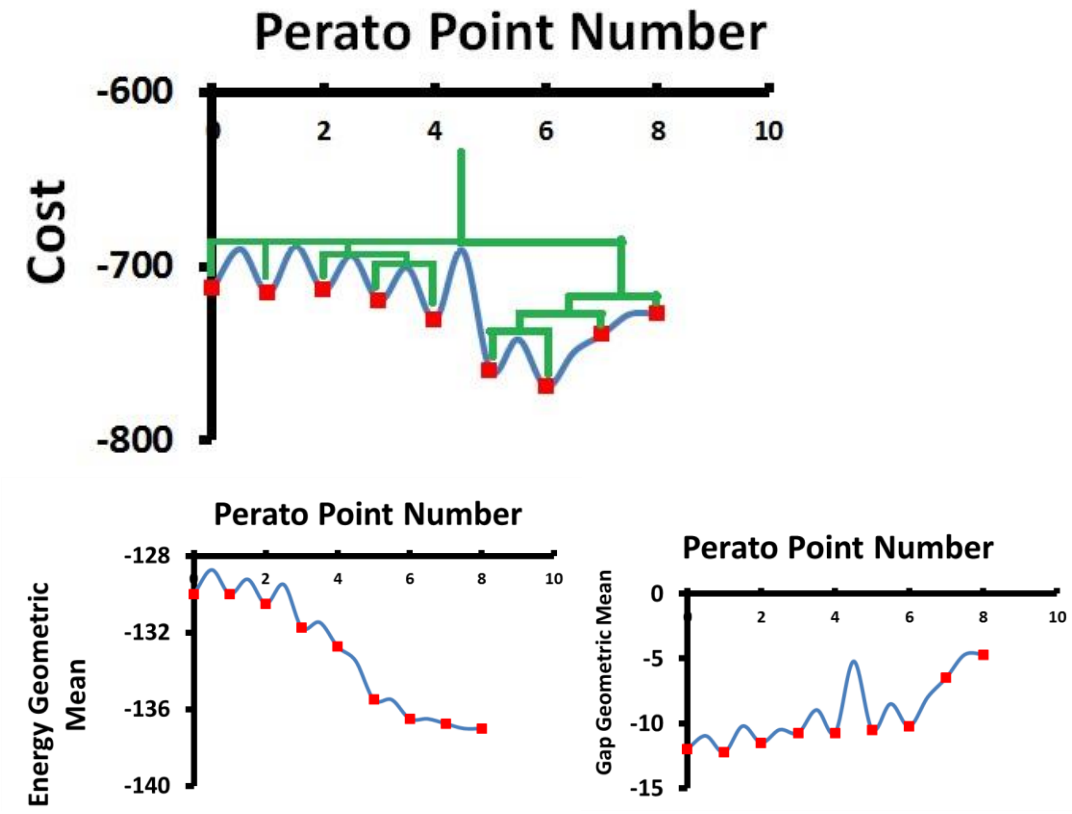


Fig. 2: Barrier tree. The Barrier tree is shown in green. The Barrier tree branches at the cost function's energy barrier height that exists between 2 minima. All minimas are points lying on the true Pareto front and are collected from the simulations previously run. Shown here are the Pareto data points from 2 Objectives optimized, orthogonality level 4 with a net zero charge constraint simulation run.

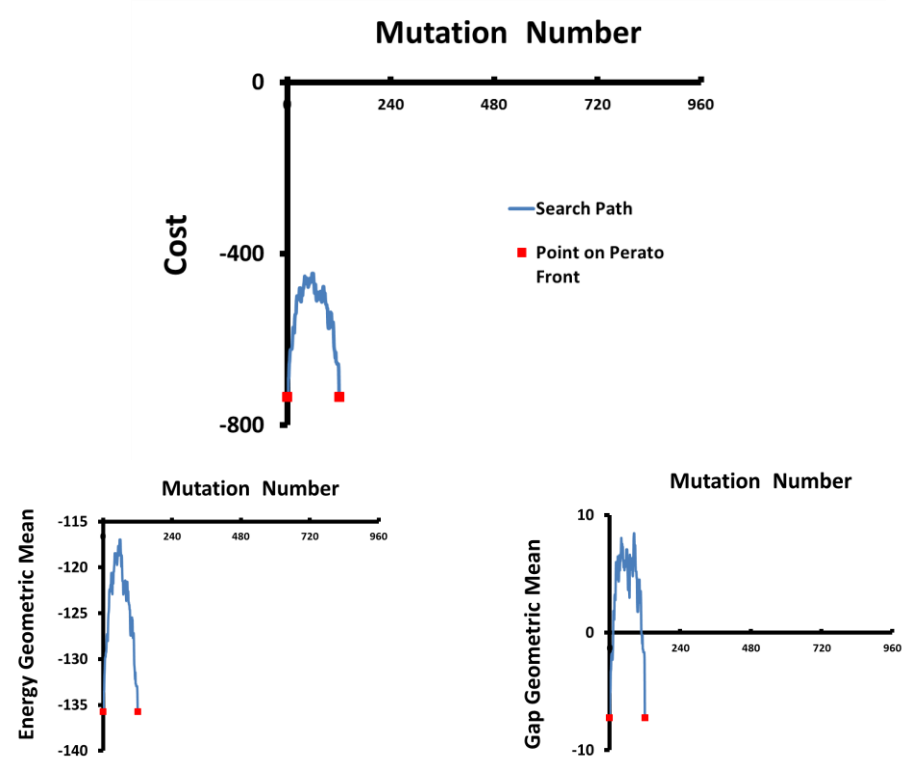


Fig. 3 : 2 Pareto points from 2 different simulation runs that were overlapping in objective space - located at the same minima point in the Pareto front of the objective values – but very different in salt bridge pattern or sequence pattern (orthogonal in sequence space) or both.

Section II: Analysis Results

Upon studying the individual collagen heterotrimer sequences sets representing each Pareto point, it was observed that high sequence orthogonality between 2 Pareto points led to high energy barriers. The high energy barrier implies that the search algorithm encountered a frustrated landscape when moving from one Pareto point to the next – this scenario occurred even when 2 Pareto points were overlapping in objective space -

located at the same minima point in the objective values but located at different places in sequence search space (**Fig. 3**). However, Pareto points that had low sequence orthogonality or low salt bridge position orthogonality led to low energy barriers (**Fig. 1a** … sequences not shown).

Children sequences derived from same the parent gene sequence during evolution tend to have low sequence orthogonality. This results in low energy barriers when transitioning between minima. It's likely that natural evolution results in deep energy basins having several clustered Pareto points with low energy barriers between them like in SPEA2, and other Pareto based archiving search algorithms. The sequences corresponding to a Pareto front are derived from the sequences of the previous Pareto front (parent generation).

The initial hypothesis was that the height of energy barriers between Pareto points is dependent on the underlying algorithm used. Replica exchange or parallel tampering during high temperature phase is similar to random walk - which in turn is similar to any mutation operation (performed for increasing diversity) in GA algorithm - both approach result in similar energy barriers (high energy barriers due to high sequence mutation rate). SPEA2 Pareto points have low sequence diversity/ low sequence orthogonality/ low energy barriers - this is because all Pareto points are all derived from the same parent strands that have been previously archived

Section III: Analysis Conclusion

Non evolved synthetic peptide sequence landscapes are highly frustrated in general. 2 overlapping Pareto points may have high energy barriers between them due to high sequence dis-similarity. High sequence dis-similarity is unlikely if the 2 overlapping Pareto point sequences have been derived from the same parent strand - as is the case in natural evolution/genetic evolution or SPEA2 and other archiving algorithms.

In contrast, non-archived Replica exchange and search algorithms with high rates of mutation tend to have high energy barriers between the Pareto points. These search algorithms are designed to overcome high energy barriers. Replica exchange is fast and provides good ‘breadth’ coverage of conformational sequence search sampling. SPEA2 is more computationally intensive, not dependent on weights and provides less genetically diverse yet more accurate minimized objective solutions.

CHAPTER V

CONCLUSION

A proof of principle concept for computationally designing an orthogonal collagen peptide ensemble was proposed. We show that even with a coarse grained sequence based model, it is possible to refine upon the existing level 2 orthogonality collagen trimer ensemble. Orthogonal peptide systems can form the foundational templates for building higher order modular systems in the synthetic biology arena of collagen mimetic peptide fibers. Although, nature uses a more sophisticated approach of molecular chaperons assisted folding and unfolding of complex molecular systems, it is possible to attain specificity through sequence orthogonality alone. Design of orthogonal systems will enable us to attach unique payloads to these orthogonal molecular tags. It will also provide for a finer control over the ease of use of molecular building blocks.

It may be true that in nature too, proteins have evolved to sit on the Pareto frontier. One may emulate this process *in-silico,* map out the representative Pareto front orthogonal sequences sets that have a good bi-objective specificity-stability trade-off and reduce the number of peptides that need to be synthesized and tested *in vitro.* This enables a less cumbersome and faster mechanism of studying protein-protein, protein-peptide and peptide-peptide interactions for understanding natural processes.

Control using the knowledge of the destabilizing effect of net charge, the stabilizing effect of "POG" triplets and the stabilizing effect provided by metal ions in solution, will all help in advancing the research of modular design of collagen mimetic peptides.

APPENDIX

MATERIALS AND METHODS

I. Peptide Synthesis and Sequences: Synthesis of the peptides was carried out at LifeTein to get >95% purity peptides. N- and C- termini were uncapped. Uncapped ends have minor effects on collagen model peptide stability at neutral pH(*16*). Peptides were dialyzed, lyophilized and end-products were verified by HPLC and by mass spectrometry.

Unless otherwise specified, peptide solutions were prepared in 10mM Phosphate or Tris buffer pH=7.4, with or without 100mM NaCl. Peptide concentrations in solution were measured by obtaining the absorbance at 214nm using $\square_{214}$ = 2200 $M^{-1}cm^{-1}$. After preparing mixtures at room temperature, they were heated to 75°C for 30 minutes and stored at 4°C for 48 hours.

III. Circular dichroism (CD)

CD measurements were conducted using the Aviv model 420SF spectrophotometer equipped with a Peltier temperature controller. Wavelength scans were conducted from 190 to 260nm at 5°C. Measurements were recorded at intervals of 0.5nm steps with an averaging time of 10s at each wavelength. The obtained ellipticity was converted to molar ellipticity by dividing the ellipticity with the peptide concentration, number of residues, and cell path length. For the temperature melt scans wherein the peptide was

slowly unfolded from its folded state, the ellipticity was measured at 223nm. A total peptide concentration of 0.2mM was maintained in all experiments. CD melt were smoothed using the Savitsky-Golay algorithm with nineteen points with a second-order polynomial [41]. First derivatives plots of the melt curves were drawn to assess the melting temperature of the peptides.

III. Minimum set cover algorithm:

To come up with the minimum number of experiments required to cover all the possible competing states in a 9 peptide ensemble or a 12 peptide ensemble, the following approach was used: We needed to set up the composition of the combinatorial peptide monomer group sets such that each individual set would not result in the formation of a target state and also ensure that the lest number of sets or experiments were created while managing to cover all the competing states amongst them. The protocol formulated was as follows: First, a combination of a only 2 out of the 3 peptides were picked from each of the designed 3 target heterotrimer species set {'A', 'B', 'C'}, {'D', 'E', 'F'} and {'G', 'H', 'I'} and grouped together, resulting in a group of a maximum 6 peptides . Each set covers ${}^{6}C_{3}$ = 20 possible heterotrimers, there are 81 competing trimer states that need to be accounted for and there exists a large redundancy of timers covered amongst the sets. The total number of such 6 peptide groupings is ${}^{3}C_{2}$ ${}_{+}{}^{3}C_{2+}{}^{3}C_{2}$ = 27. These sets are numbered from 1 to 27. The first 3 sets can be arranged into a super-group of 3 sets in 6 ways : [{1,2,3},{1,3,2},{3,1,2},{2,3,1},{2,1,3},{3,2,1}]. For each of these arrangements, we check if any of the sets in the order can be eliminated. A set can be eliminated if 1) the group of sets occurring before the set in the set order, cover

all the states covered by the set and 2) if the sets left out in the in the parent super-group after the set is eliminated manage to cover the required 81 states. So groups of 3 sets from the 27 sets are picked, ordered and checked to see if a set/sets can be eliminated. If a ‘super-group’ of 3 sets is insufficient to cover all the required sets, we move on to the next super-group of 4 and the process is repeated. After applying this protocol, it was found out that a super-group of 7 had the minimum required number of sets to cover all the 81 states for a 3 peptide ensemble.

SUPPORTING INFORMATION:

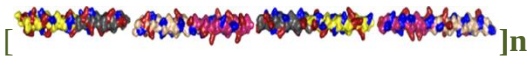


Fig. s1: Domain tempting of a futuristic long recombinant bacterial collagen trimer chain. Four orthogonal heterotrimers are laid out in tandem - as a series of swapped domain pairs that are half of full length of a trimer. The positive (red) and negative (blue) charges contributing to the formation of the salt bridge are shown.

CEF	-126
FBC	-124
ABD	-124
GFI	-124
IDA	-123
AGC	-123
ABF	-123
FBG	-123
EFH	-123
IDI	-123

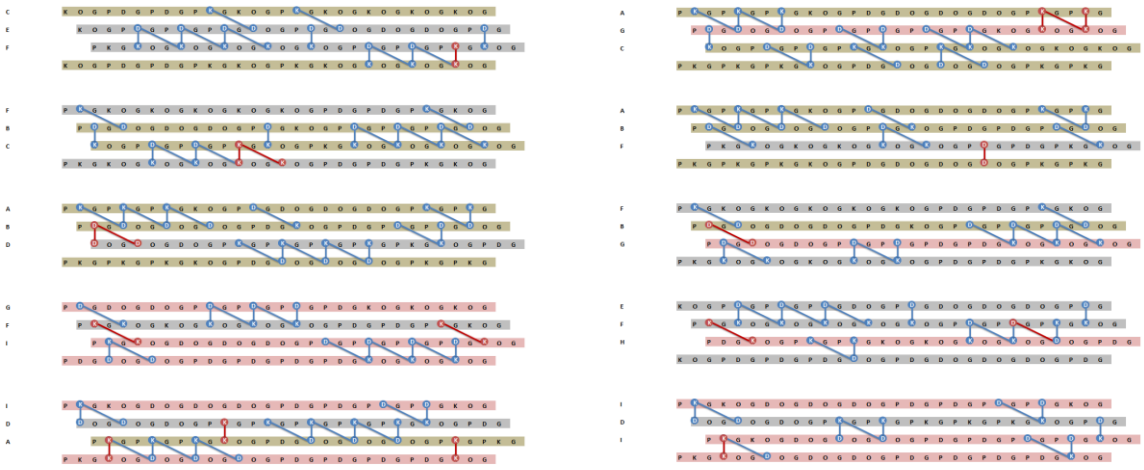


Fig. s2 : ABC + DEF + GHI competing states. The sequences and stability scores of the top 10 competing states are shown.

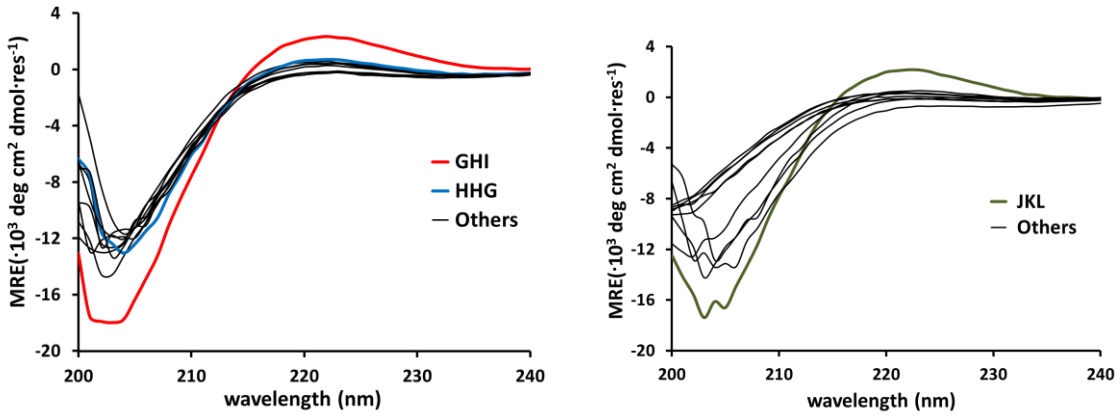


Fig. s3: Wavelength scans. a) Wavelength scan of GHI and its competing stoichiometry's, within a 3 peptide ensemble. B) Wavelength scan of JKL and its competing stoichiometry's, within a 3 peptide ensemble. 10 mM phosphate buffer, pH = 7, with total peptide concentration = 0.2 mM.

	Score	Tm	Charge			Score	Tm	Charge
GHI	-136	25.5	-4		JKL	-135	16.5	-4
GHG	-118	10	-4		KJK	-118		
IHI	-118	< 4	-4		LKL	-118		
HHG	-117	11	+4		JKJ	-117		
IIG	-114	< 4	-12		KLK	-115		
HHI	-114				LLJ	-114		
IGG	-113				LJJ	-113		
III	-113				LLL	-113		
GGG	-107				KKK	-110		
HHH	-107				JJJ	-107		

Fig. s4 : GHI and JKL 3 peptide ensemble scores.

CEF

CEF	-126
CCE	-116
CEE	-114
CCF	-110
CFF	-110
EFF	-109
EEF	-107
CCC	-101
FFF	-101
EEE	-90

BCF

FBC	-124
FFB	-117
FBB	-116
CFF	-110
CCF	-110
BCC	-107
BBC	-106
FFF	-101
CCC	-101
BBB	-93

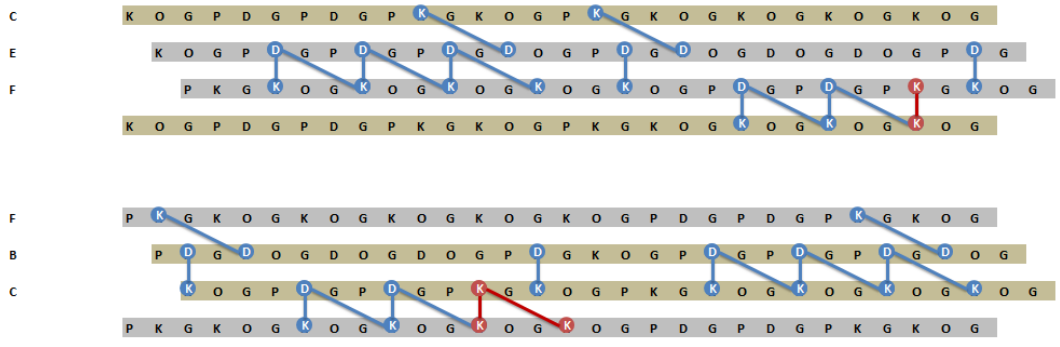


Fig. s5 : CEF and BCF stability : <u>gaps : -10 , -7</u> , net charge : +4

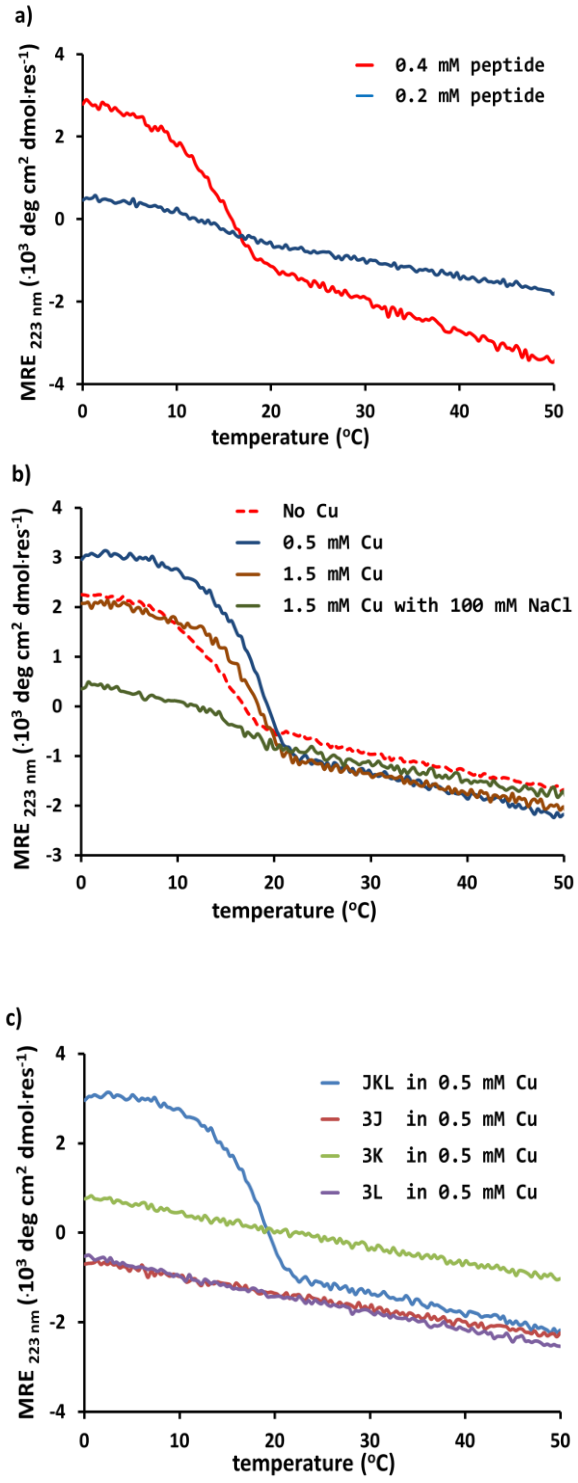


Fig. s6 : JKL characterization. a) JKL in 10mM phosphate buffer **b)** JKL stabilized by Copper in 10 mM Tris buffer. **c)** Comparison of heterotrimer versus homotimers folding in Cu. JKL Heterotrimer stabilized by Copper in 10 mM Tris buffer. The homotimers do not fold. Total peptide concentration = 0.2 mM.

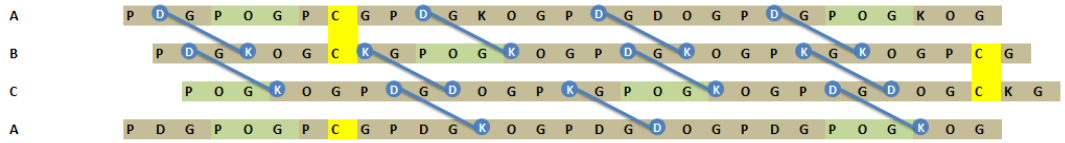
Fig s7 : Charge pattern of A-L. Each pseudo-domains charge is indicated.

Set1: Fei_ABC = A:B:C peptides

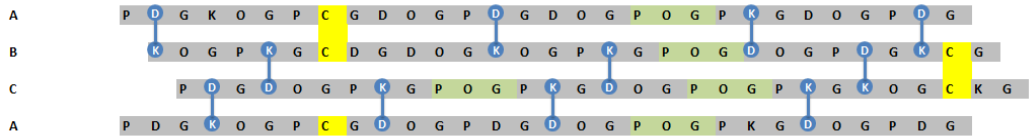
Set2: Acid_Tecton1

Set 3: Acid_Tecton2

Set 4: Axial_Trimer in 2mM DTT



Set 5: Lateral_Trimer in 2mM DTT

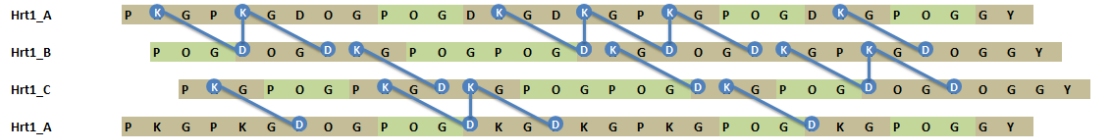


Set 6: Barbara Brodsky paper peptides*

Brb1_A: GPOGPOGPOGMOGVGEKGEOGKOGPOGPOGY

Brb1_B: POGDOGPOGPOGISLKGEEGPOGPAGPOGYOG

Set 7: Hartgerink paper peptides*



*For computational stability score calculation – all non P,K,D,G,O residues were mutated to Alanine.

Set 8: A to L peptides that fold

Set 9: The 1, 2, 3 and 4 triplet circular permutations of strand A in combination with either {B,C} or { E, F}.

Fig. s8a: List and names of Peptide sets considered for study. The set number, name and in some cases the corresponding sequences are indicated.

The table on the next page can be read as follows:

3rd Column : Concentration of peptide
4th Column : Trimer Association state in a 3 peptide ensemble
6th Column : Tm with no salt
5,7th Column : Tm with 100mM Nacl
8th Column : Color Green => Folds, Red => does not fold, Blank => no data available

Rest of the columns are as follows:

Gap : Gap in a 3 peptide ensemble
No_of_Fav : No of salt bridges (attraction)
No_of_non : No of salt bridges (repulsion)
No_Axi : No of axial salt bridges (attraction)
No_Lat : No of lateral salt bridges (attraction)
1st_strand_c : Leading strand net charge
2nd_strand_c : Lagging strand net charge
3rd_strand_c : Middle strand net charge
N_ter : N_terminal pseudo-domain net charge
C_ter : C_terminal pseudo-domain net charge
Net : Net charge
Gradient : Absolute value of N_terminal pseudo-domain net charge + C_terminal pseudo-domain net charge .. Indicates charge distribution.
Pogs : No of POGs
NPogs : No of N terminal POGs
CPogs : No of C terminal POGs

Values highlighted in light green color are approximate values.

Set No.	Set Name	Conc	Trimer Name	Energy	Tm no Salt	Tm no Salt	Tm With Salt	Tm With Salt	Gap	No_of_Fav	No_of_non	No_Axi	No_Lat	1st_strand_c	2nd_strand_c	3rd_strand_c	N_ter	C_ter	Net	Gradient	Pogs	NPogs	CPogs
1	Fei_ABC	0.2mM	AAA	-94	0		0		NA	0	8	0	0	2	2	2	9	-3	6	12	0	0	0
			AAB	-108	0		0		0	8	5	5	3	2	2	-8	1	-5	-4	6	0	0	0
			AAC	-107	0		0		1	6	6	3	3	2	2	6	7	3	10	10	0	0	0
			ABB	-108	0		0		0	9	5	6	3	2	-8	-8	-7	-7	-14	14	0	0	0
			ABC	-135	29		29		-27	21	0	11	10	2	-8	6	-1	1	0	2	0	0	0
			ACC	-108	0		0		0	6	6	3	3	2	6	6	5	9	14	14	0	0	0
			BBB	-93	0		0		NA	3	8	2	1	-8	-8	-8	-15	-9	-24	24	0	0	0
			BBC	-106	7		0		1	9	6	4	5	-8	-8	6	-9	-1	-10	10	0	0	0
			BCC	-107	8		6		0	8	6	3	5	-8	6	6	-3	7	4	10	0	0	0
			CCC	-101	0		0		NA	1	7	1	0	6	6	6	3	15	18	18	0	0	0
2	Acid_Tecton1	0.2mM	AT1_AAA	-98	0		0		NA	4	8	3	1	-2	-2	-2	-3	-3	-6	6	0	0	0
			AT1_AAB	-101	0		0		1.2	5	7	3	2	-2	-2	-2	-7	1	-6	8	0	0	0
			AT1_AAC	-102	0		0		0	8	7	5	3	-2	-2	-6	-3	-7	-10	10	0	0	0
			AT1_ABB	-100	0		0		1.4	4	7	2	2	-2	-2	-2	-11	5	-6	16	0	0	0
			AT1_ABC	-131	18.5				-27.8	17	0	8	9	-2	-2	-6	-7	-3	-10	10	0	0	0
			AT1_ACC	-101	0		0		0	7	7	4	3	-2	-6	-6	-3	-11	-14	14	0	0	0
			AT1_BBB	-91	0		0		NA	0	9	0	0	-2	-2	-2	-15	9	-6	24	0	0	0
			AT1_BBC	-102	0		0		0.4	8	7	4	4	-2	-2	-6	-11	1	-10	12	0	0	0
			AT1_BCC	-102	0		0		0.2	8	7	4	4	-2	-6	-6	-7	-7	-14	14	0	0	0
			AT1_CCC	-93	0		0		NA	0	8	0	0	-6	-6	-6	-3	-15	-18	18	0	0	0
3	Acid_Tecton2	0.2mM	AT2_AAA	-107	0		0		NA	4	5	3	1	-2	-2	-2	-9	3	-6	12	0	0	0
			AT2_AAB	-105.8	0		0		0	7	7	4	3	-2	-2	-5	-6	-3	-9	9	1	1	0
			AT2_AAC	-108	0		0		0	7	5	4	3	-2	-2	0	-5	1	-4	6	0	0	0
			AT2_ABB	-105.6	0		0		1	6	8	3	3	-2	-5	-5	-3	-9	-12	12	2	2	0
			AT2_ABC	-135.8	19				-29	18	0	9	9	-2	-5	0	-2	-5	-7	7	1	1	0
			AT2_ACC	-108	0		0		1	7	5	4	3	-2	0	0	-1	-1	-2	2	0	0	0
			AT2_BBB	-104.4	0		0		NA	0	8	0	0	-5	-5	-5	0	-15	-15	15	3	3	0
			AT2_BBC	-107.6	0		0		0	6	7	3	3	-5	-5	0	1	-11	-10	12	2	2	0
			AT2_BCC	-107.8	0		0		0	7	6	4	3	-5	0	0	2	-7	-5	9	1	1	0
			AT2_CCC	-108	0		0		NA	4	5	3	1	0	0	0	3	-3	0	6	0	0	0

Set No.	Set Name	Conc	Trimer Name	Energy	Tm no Salt	Tm no Salt	Tm With Salt	Tm With Salt	Gap	No_of_Fav	No_of_non	No_Axi	No_Lat	1st_strand_c	2nd_strand_c	3rd_strand_c	N_ter	C_ter	Net	Gradient	Pogs	NPogs	CPogs
4	Axial_Trimer in 2mM DTT	0.2mM	Axi_AAA	-131.8	0		0		NA	4	3	3	1	-3	-3	-3	-3	-6	-9	9	6	3	3
			Axi_AAC	-130	12.6				1.8	6	3	6	0	-3	-3	0	-2	-4	-6	6	6	3	3
			Axi_ABA	-129.2	12.3				2.6	6	2	5	1	-3	4	-3	0	-2	-2	2	5	3	2
			Axi_ABC	-136.4	17.21				-4.6	11	0	11	0	-3	4	0	1	0	1	1	5	3	2
			Axi_ACC	-122.2	13.29				9.6	5	4	5	0	-3	0	0	-1	-2	-3	3	6	3	3
			Axi_BBA	-122.6	0		0		9.2	6	2	2	4	4	4	-3	3	2	5	5	4	3	1
			Axi_BBB	-112	0		0		NA	6	4	4	2	4	4	4	6	6	12	12	3	3	0
			Axi_BBC	-121.8	18.85				0	8	2	7	1	4	4	0	4	4	8	8	4	3	1
			Axi_BCC	-119.6	16.8				2.2	6	3	5	1	4	0	0	2	2	4	4	5	3	2
			Axi_CCC	-105.4	3				NA	0	7	0	0	0	0	0	0	0	0	0	6	3	3
5	Lat_Trimer in 2mM DTT	0.2mM	Lat_AAA	-122.4	0		0		NA	6	3	4	2	-4	-4	-4	-6	-6	-12	12	3	0	3
			Lat_AAB	-118.8	16.5		6		3.6	7	2	3	4	-4	-4	1	-3	-4	-7	7	3	0	3
			Lat_AAC	-120.4	20		18		2	6	4	2	4	-4	-4	2	-4	-2	-6	6	4	1	3
			Lat_ABB	-110.2	17.5		17.5		-10.6	6	2	2	4	-4	1	1	0	-2	-2	2	3	0	3
			Lat_ABC	-128.8	19.5		15		-6.4	11	0	0	11	-4	1	2	-1	0	-1	1	4	1	3
			Lat_ACC	-117.4	20		20		-4	5	5	1	4	-4	2	2	-2	2	0	4	5	2	3
			Lat_BBB	-99.6	0		0		NA	6	3	5	1	1	1	1	3	0	3	3	3	0	3
			Lat_BCB	-113.2	19		19		-13.6	10	3	4	6	1	2	1	2	2	4	4	4	1	3
			Lat_CBC	-115.8	19.4		17		-2.4	7	4	1	6	2	1	2	1	4	5	5	5	2	3
			Lat_CCC	-113.4	12		12		NA	3	6	2	1	2	2	2	0	6	6	6	6	3	3

Set No.	Set Name	Conc	Trimer Name	Energy	Tm no Salt	Tm no Salt	Tm With Salt	Tm With Salt	Gap	No_of_Fav	No_of_non	No_Axi	No_Lat	1st_strand_c	2nd_strand_c	3rd_strand_c	N_ter	C_ter	Net	Gradient	Pogs	NPogs	CPogs
6	Barbara	0.1mM	Brb1_AAA	-151.2	16.5				NA	5	6	2	3	0	0	0	0	0	0	0	6	0	6
			Brb1_AAB	-151.2	14.5				0	5	6	2	3	0	0	0	0	0	0	0	6	0	6
			Brb1_BBB	-151.2	3.5				NA	5	6	2	3	0	0	0	0	0	0	0	6	0	6
7	Hartgerink	0.3mM	(PKG)10 : 2*(POGDOG)5	-162	44				8.9	10	0	5	5	10	-5	-5	1	-1	0	2	10	6	4
			2*(PKGPOG)5 : (DOG)10	-162	46				8.9	10	0	5	5	5	5	-10	1	-1	0	2	10	4	6
			PKGPOG5	-171	0				NA	0	0	0	0	5	5	5	9	6	15	15	15	6	9
			POG10	-228	67.5				NA	0	0	0	0	0	0	0	0	0	0	0	30	15	15
			DOG10	-114	0				NA	0	0	0	0	-10	-10	-10	-15	-15	-30	30	0	0	0
			PKG10	-114	0				NA	0	0	0	0	10	10	10	15	15	30	30	0	0	0
			POGDOG5	-171	35.5				NA	0	0	0	0	-5	-5	-5	-6	-9	-15	15	15	9	6
			Hrt1_AAA	-128	0				NA	14	0	6	8	2	2	2	3	3	6	6	9	3	6
			Hrt1_AAC	-135.6	42				9.2	14	0	8	6	2	2	0	4	0	4	4	10	4	6
			Hrt1_ABC	-140.6	58				4.2	19	0	14	5	2	-2	0	2	-2	0	4	10	6	4
			Hrt1_ACC	-141.2	42				3.6	12	0	7	5	2	0	0	5	-3	2	8	11	5	6
			Hrt1_BAA	-132	30				0	18	0	6	12	-2	2	2	1	1	2	2	9	5	4
			Hrt1_BBA	-132	30				0	18	0	6	12	-2	-2	2	-1	-1	-2	2	9	7	2
			Hrt1_BBB	-128	0				NA	14	0	6	8	-2	-2	-2	-3	-3	-6	6	9	9	0
			Hrt1_BBC	-135.6	42				9.2	14	0	8	6	-2	-2	0	0	-4	-4	4	10	8	2
			Hrt1_BCC	-141.2	42				3.6	12	0	7	5	-2	0	0	3	-5	-2	8	11	7	4
			Hrt1_CCC	-144.8	32				NA	8	0	3	5	0	0	0	6	-6	0	12	12	6	6

Set No.	Set Name	Conc	Trimer Name	Energy	Tm no Salt	Tm no Salt	Tm With Salt	Tm With Salt	Gap	No_of_Fav	No_of_non	No_Axi	No_Lat	1st_strand_c	2nd_strand_c	3rd_strand_c	N_ter	C_ter	Net	Gradient	Pogs	NPogs	CPogs
8	A-L	0.2mM	CEF	-126			15.5		-10	14	1	7	7	6	-8	6	3	1	4	4	0	0	0
			DEF	-133	24		24		-17	19	0	9	10	2	-8	6	1	-1	0	2	0	0	0
			EEF	-107	4		0		-17	8	5	3	5	-8	-8	6	-1	-9	-10	10	0	0	0
			EFF	-109	4		0		-8	8	5	3	5	-8	6	6	7	-3	4	10	0	0	0
			FBC	-124			11.5		-7	14	2	7	7	6	-8	6	1	3	4	4	0	0	0
			GHG	-118	10		0		-11	13	3	7	6	-4	4	-4	-7	3	-4	10	0	0	0
			GHI	-136	25.5		22.5		-18	22	0	12	10	-4	4	-4	-3	-1	-4	4	0	0	0
			HHG	-117	10		0		-10	11	3	4	7	4	4	-4	1	3	4	4	0	0	0
			HHI	-114	4		0		-7	10	4	5	5	4	4	-4	5	-1	4	6	0	0	0
			IHI	-118	4		0		-5	14	4	7	7	-4	4	-4	1	-5	-4	6	0	0	0
			JKL	-135	16.5		0		-17	21	0	11	10	-4	4	-4	-1	-3	-4	4	0	0	0

Set No.	Set Name	Conc	Trimer Name	Energy	Tm no Salt	Tm no Salt	Tm With Salt	Tm With Salt	Gap	No_of_Fav	No_of_non	No_Axi	No_Lat	1st_strand_c	2nd_strand_c	3rd_strand_c	N_ter	C_ter	Net	Gradient	Pogs	NPogs	CPogs
9	Circular_Permuted	0.2mM	p1_AAA	-94	0		0		NA	0	8	0	0	2	2	2	15	-9	6	24	0	0	0
			p1_AAB	-105	0		0		0	8	6	3	5	2	2	-8	5	-9	-4	14	0	0	0
			p1_AAF	-85	0		0		16	1	13	0	1	2	2	6	15	-5	10	20	0	0	0
			p1_ABB	-105	0		0		0	9	6	4	5	2	-8	-8	-5	-9	-14	14	0	0	0
			p1_ABC	-131	24				-24	19	1	8	11	2	-8	6	1	-1	0	2	0	0	0
			p1_AEF	-109	4				0	9	5	4	5	2	-8	6	7	-7	0	14	0	0	0
			p1_AFF	-86	0		0		15	1	13	0	1	2	6	6	15	-1	14	16	0	0	0
			p1_CAA	-106	0		0		1	3	5	2	1	6	2	2	11	-1	10	12	0	0	0
			p1_CCA	-107	0		0		0	3	5	2	1	6	6	2	7	7	14	14	0	0	0
			p1_EAA	-103	0		0		0	3	5	2	1	-8	2	2	7	-11	-4	18	0	0	0
			p1_EEA	-102	0		0		1	3	5	2	1	-8	-8	2	-1	-13	-14	14	0	0	0
			p2_AAA	-94	0		0		NA	0	8	0	0	2	2	2	15	-9	6	24	0	0	0
			p2_AAB	-99	0		0		0	7	8	3	4	2	2	-8	5	-9	-4	14	0	0	0
			p2_AAC	-104	0		0		1	3	6	2	1	2	2	6	11	-1	10	12	0	0	0
			p2_ABB	-99	0				0	8	8	4	4	2	-8	-8	-5	-9	-14	14	0	0	0
			p2_ABC	-125	19.8		0		-18	15	2	7	8	2	-8	6	1	-1	0	2	0	0	0
			p2_ACC	-105	0		0		0	3	6	2	1	2	6	6	7	7	14	14	0	0	0
			p2_EAA	-104	0		0		0	4	5	2	2	-8	2	2	7	-11	-4	18	0	0	0
			p2_EEA	-103	0		0		1	4	5	2	2	-8	-8	2	-1	-13	-14	14	0	0	0
			p2_EFA	-112	9				-3	13	5	6	7	-8	6	2	7	-7	0	14	0	0	0
			p2_FAA	-84	0		0		17	2	14	1	1	6	2	2	15	-5	10	20	0	0	0
			p2_FFA	-85	0		0		16	2	14	1	1	6	6	2	15	-1	14	16	0	0	0
			p3_AAA	-94	0		0		NA	0	8	0	0	2	2	2	9	-3	6	12	0	0	0
			p3_AAB	-99	0		0		0	6	8	3	3	2	2	-8	1	-5	-4	6	0	0	0
			p3_AAC	-103	0		0		1	2	6	1	1	2	2	6	7	3	10	10	0	0	0
			p3_ABB	-99	0		0		0	7	8	4	3	2	-8	-8	-7	-7	-14	14	0	0	0
			p3_ABC	-121	9				-14	13	3	5	8	2	-8	6	-1	1	0	2	0	0	0
			p3_ACC	-104	0		0		0	2	6	1	1	2	6	6	5	9	14	14	0	0	0
			p3_EAA	-108	0		0		0	6	4	4	2	-8	-8	2	-3	-11	-14	14	0	0	0
			p3_EEA	-107	0		0		1	8	5	3	5	-8	-8	6	-1	-9	-10	10	0	0	0
			p3_EFA	-125	17.6				-16	17	2	7	10	-8	6	2	5	-5	0	10	0	0	0
			p3_FAA	-97	0		0		4	5	10	2	3	6	2	2	11	-1	10	12	0	0	0
			p3_FFA	-98	0		0		3	5	10	2	3	6	6	2	13	1	14	14	0	0	0
			p4_AAA	-97	0		0		NA	0	7	0	0	2	2	2	3	3	6	6	0	0	0
			p4_AAB	-97	0		0		0	0	7	0	0	2	2	2	3	3	6	6	0	0	0
			p4_AAC	-99	0		0		-2	2	8	1	1	2	2	6	3	7	10	10	0	0	0
			p4_AAE	-112	0		0		0	6	3	4	2	2	2	-8	-1	-3	-4	6	0	0	0
			p4_ABB	-97	0		0		0	4	8	1	3	2	2	-8	-3	-1	-4	4	0	0	0
			p4_ABC	-116	4				-9	11	4	4	7	2	-8	6	-3	3	0	6	0	0	0
			p4_ACC	-100	0		0		1	2	8	1	1	2	6	6	3	11	14	14	0	0	0
			p4_AEE	-111	0		0		1	6	3	4	2	2	-8	-8	-5	-9	-14	14	0	0	0
			p4_EFA	-134	24				-22	20	0	10	10	-8	6	2	3	-3	0	6	0	0	0
			p4_FAA	-105	0		0		1	6	7	3	3	6	2	2	7	3	10	11	0	0	0
			p4_FFA	-106	0		0		0	6	7	3	3	6	6	2	11	3	14	14	0	0	0

Fig. s8b : Peptide sets tabulated. Description of the table headers is indicated at the top of the table.

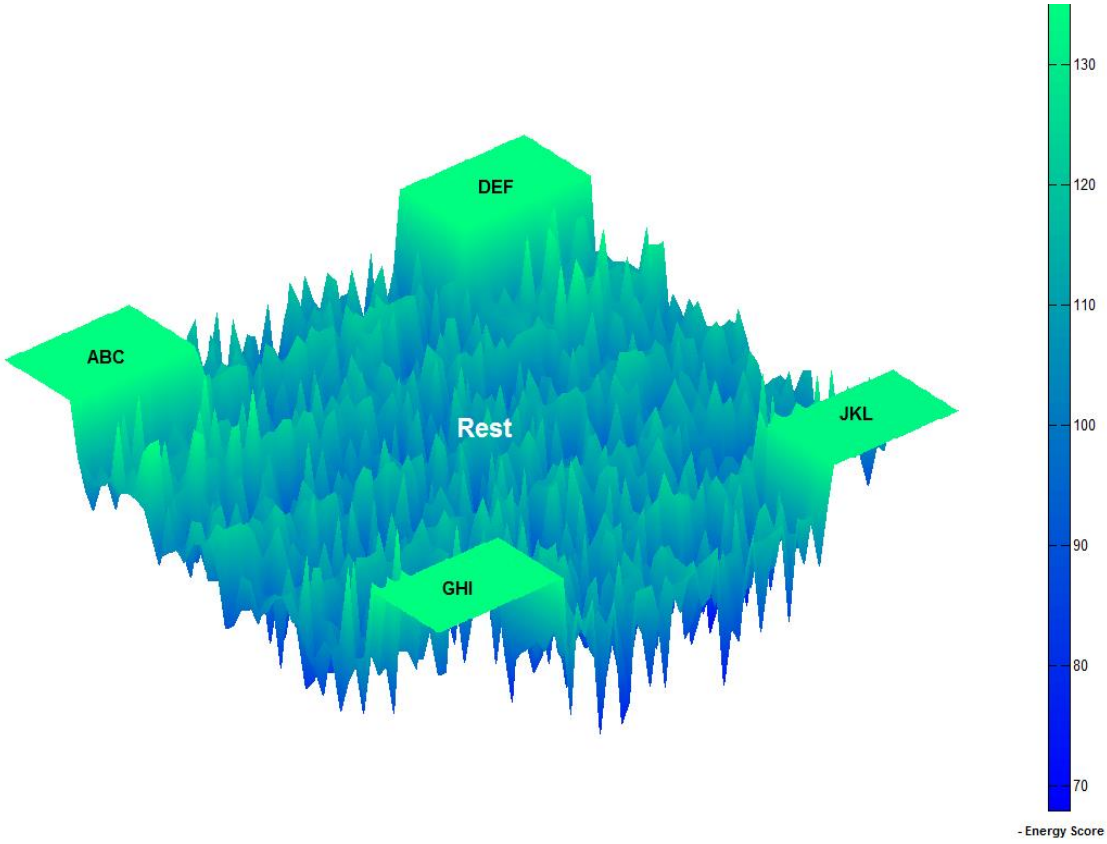


Fig. s9: Fitness landscape. (Negative of energy score)

Part A : Pseudo code :

#define NO_OF_TRIALS 7500 // 5000

#define NO_OF_ITERATIONS 500

#define NO_OF_STARTS 2

#define NO_OF_STEPS 9// = 10 including step 0

Therefore, total number of function evaluations = 5000 * 500 * 2 * 10 = 50E6

1 Iteration = 1 function evaluation.

1 TRIAL = 500 Iterations.

Outer Pseudo Start :

> *The code branches or alternates between 3 sections for every 50 TRIALS:*
>
> *First 50 TRIALS: Replica exchange*
>
> *Next 50 TRIALS : GA + SA*
>
> *Next 50 TRIALS : archive the output from previous 100 TRIALS ONCE using SPEA2 and SKIP next 49 TRIALS.*

Repeat Until Stop criteria.

Part B : Flowchart

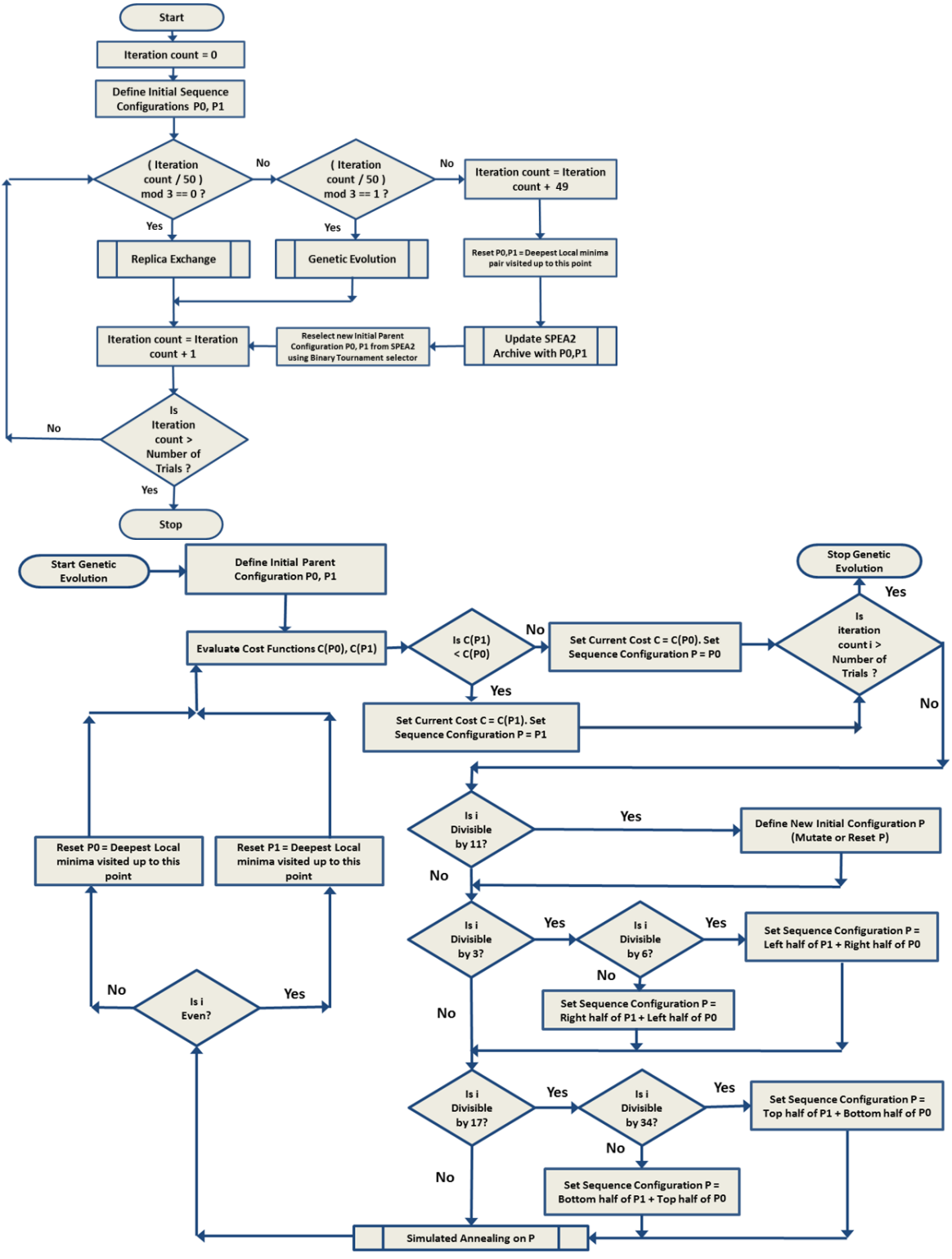


Fig. s10: A) Pareto Frontier pseudo code and B)code flowchart. Top and bottom half sequence configuration imply the first pair (“ABC”, “DEF”) and second pair (“GHI”, “JKL”) respectively for 4 orthogonal heterotrimers.

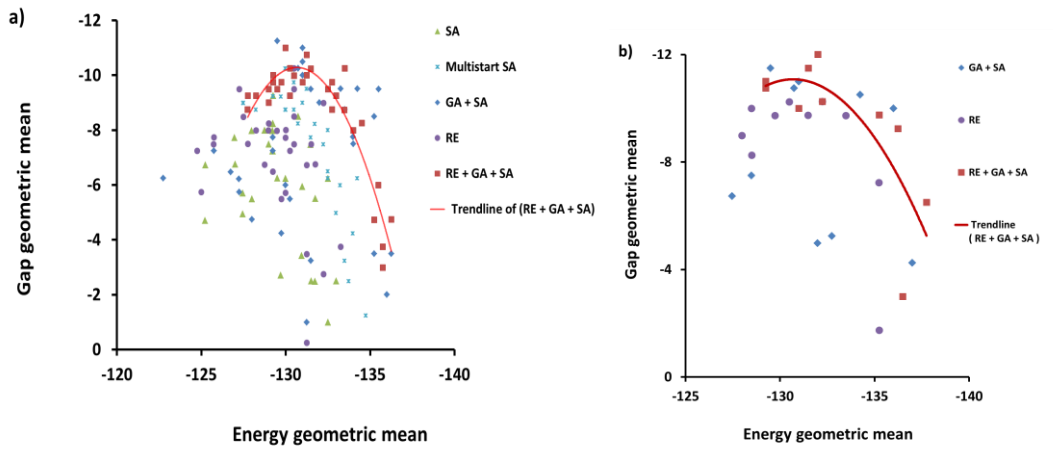


Fig s11 : Convergence of Algorithms a) For 30 data points per algorithm. Each data point was recorded after 0.2 million iterations. SA = Simulated annealing. GA = Genetic evolution algorithm. RE = Replica exchange. **b)** For 10 data points per algorithm over a total of 2.5 million iterations each.

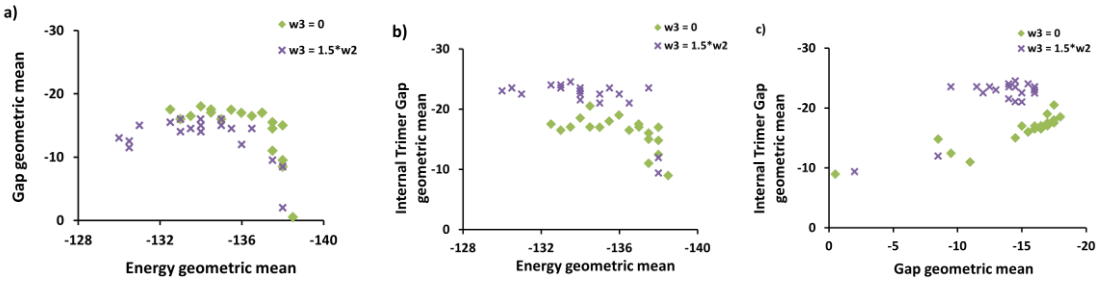


Fig s12 : Trade-off between Gap and Internal Trimer Gap for 2 orthogonal heterotrimers. The cost function used was C = w1*Energy + w2*Gap + w3*Internal_Trimer_Gap with weight pairs { w1, w2 } = { 1 - α, α }, α varying from 0 to 1 and { w3 } = ({ 0 } or { 1.5 * w2 }).

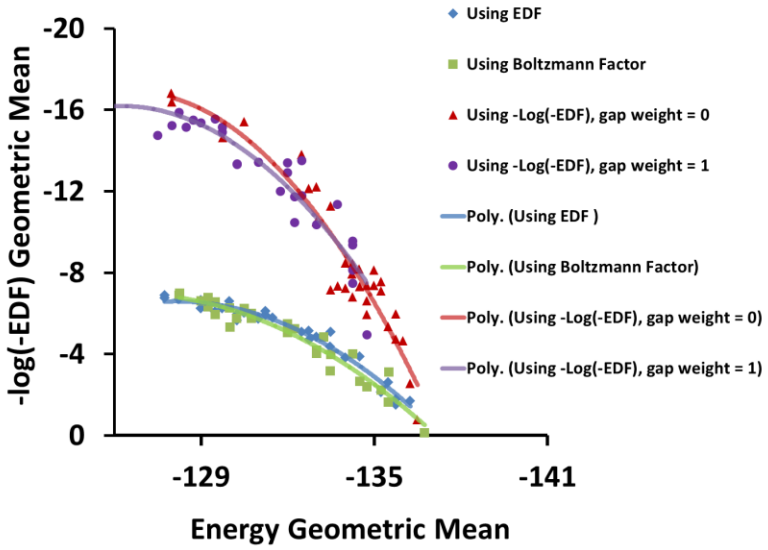
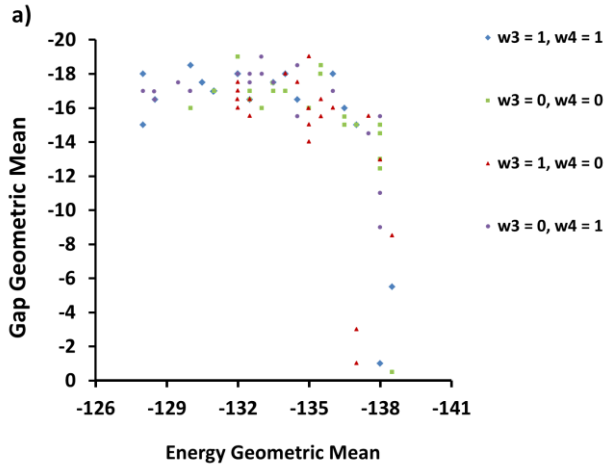


Fig s13 : Comparison of convergence of search to Pareto frontier for using Boltzmann factor or energy density factor or -log (negative energy density factor) for 2 orthogonal heterotrimers. with weight pairs { w1, w2 } = { 1 - α, α }, α varying from 0 to 1 and { w3 } = ({ 1E-25 } or { 1E-25 } or { 0.7 * w2 }).



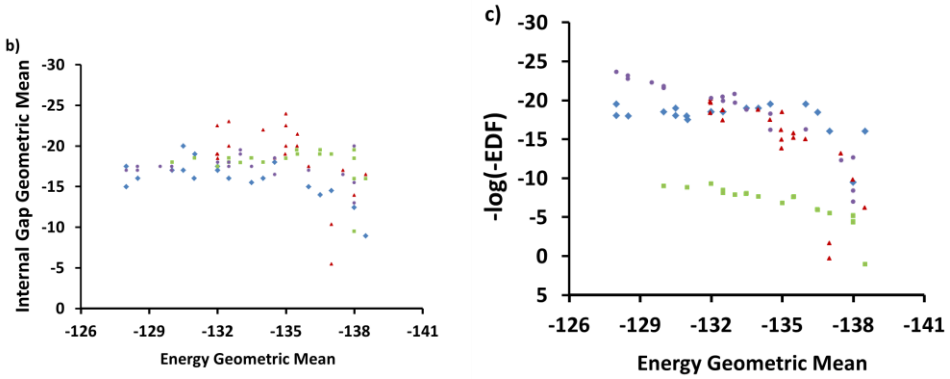
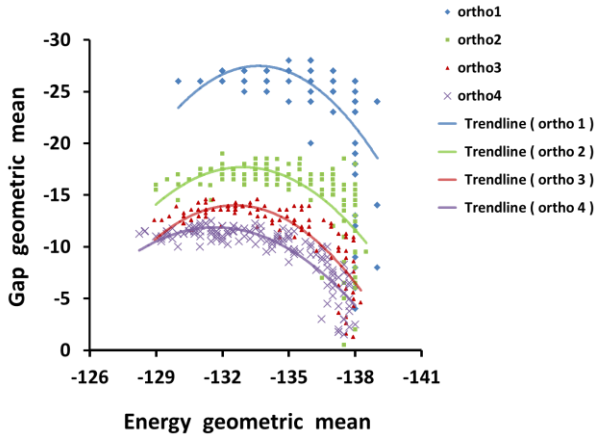


Fig s14 : Optimizing internal peptide set gap and energy density factor. The cost function used was C = w1*Energy + w2*Gap + w3*Internal_Trimer_Gap + w4*log(-EDF) with weight pairs {w1, w2} = {1,1} and {w3,w4} = ({0,0},{0.7,0.7},{0.7,0} or {0,0.7}). The 4 different {w3, w4} weight combinations explore different regions of the Pareto frontier for the 4 orthogonal heterotrimers.



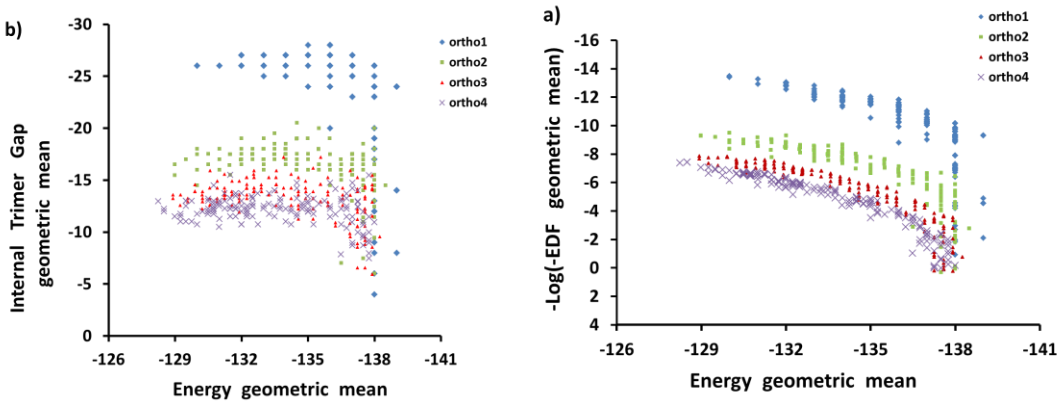
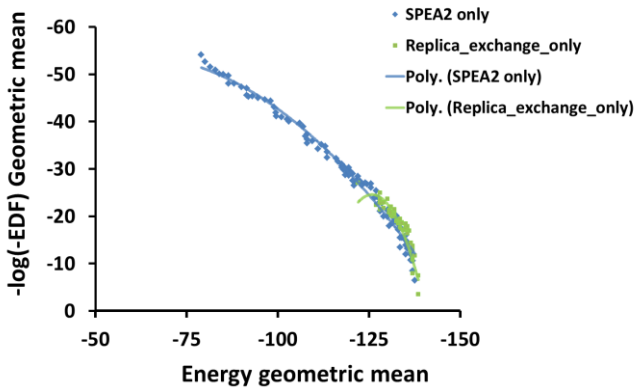


Fig s15a : Using Replica exchange + GA + SA : Pareto frontier : Using 2 objectives a) of -log(-EDF geometric mean). **b)** of Internal trimer gap geometric mean and **c)** of gap versus the energy geometric mean. 150 data points per orthogonality with 50 from including Energy Density Factor, 50 from Replica Exchange with high and low temperature pairs = { 10, 1E5} and 50 from Replica Exchange with high and low temperature pairs = {1, 1E20}. All points converge close to the same frontier and are scattered within in the Pareto optimal set. C = w1*Energy + w2*Gap + w3*Internal_Trimer_Gap + w4*EDF with weight pairs {w1, w2} = {1,1} and {w3,w4} = ({1E-25,1E-25},{0,0},{0,0}) for the three 50 data points simulation set.



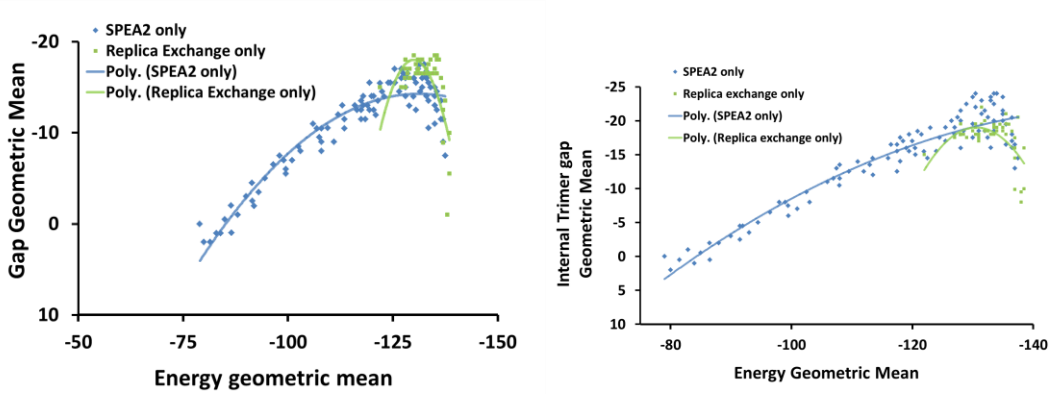


Fig s15b : Using Replica exchange v/s SPEA2 : for Ortho 2 for the same number of mutation cycles for each simulation. The population/archive size of SPEA2 used was 100.

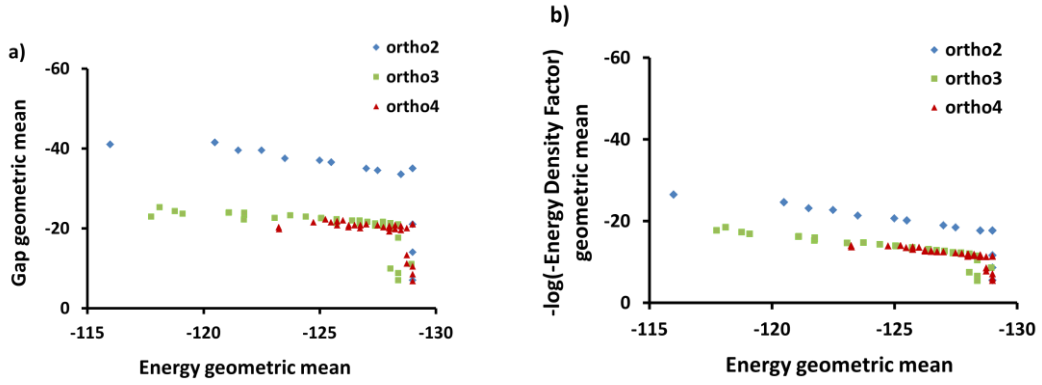


Fig s16 : Homotrimer Pareto frontier : a) of Gap geometric mean **b)** of -log(-EDF geometric mean)

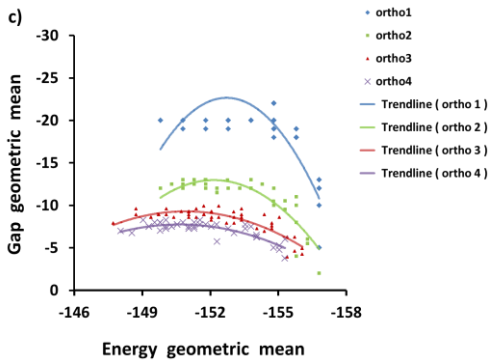
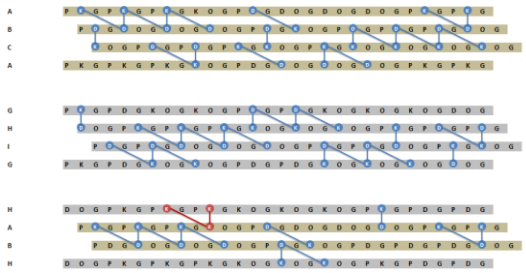


Fig s17 : Pareto frontier : gap v/s energy while considering sequences with one 'POG' triplet at each end of a 24 amino acid sequence long monomer.

Fig s18: (below) Example of designed non co-derived sequence sets lying on the Pareto frontier for each orthogonality level along with one of its next most stable competing state sequence.

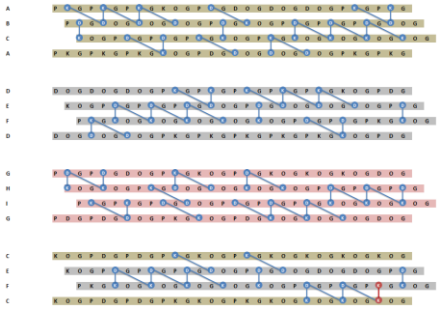
Ortho 2:



gap_0 :-14
energy_0 :-135
charge_0 :0
gradient_0 :2
gap_internal_0 :-27
boltzmann_factor_0 :-0.999998
energy_density_factor_0 :
-2.7988e+06

gap_1 :-14
energy_1 :-135
charge_1 :0
gradient_1 :2
gap_internal_1 :-14
boltzmann_factor_1 :-0.999997
energy_density_factor_1 :
-1.41319e+06

Ortho 3:



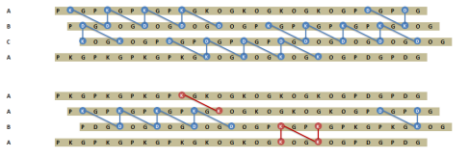
gap_0 :-9
energy_0 :-135
charge_0 :0
gradient_0 :2
gap_internal_0 :-27
boltzmann_factor_0 :-0.999785
energy_density_factor_0 :-22857.2

gap_1 :-7
energy_1 :-133
charge_1 :0
gradient_1 :2
gap_internal_1 :-17
boltzmann_factor_1 :-0.998517
energy_density_factor_1 :-24438.7

gap_2 :-11
energy_2 :-135
charge_2 :0
gradient_2 :2
gap_internal_2 :-14
boltzmann_factor_2 :-0.99994
energy_density_factor_2 :-81888.7

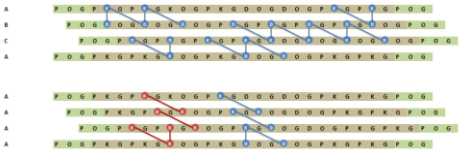
Fig s19: (below) Example of one designed co-derived (*ab initio)* sequence set lying on the Pareto frontier for each orthogonality level along with one of its next most stable competing state sequence.

Ortho 1 :



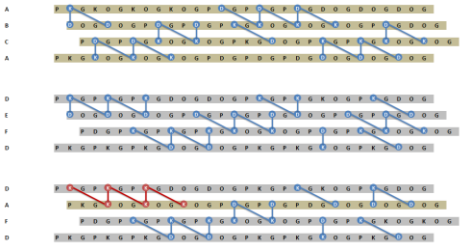
gap_0 :-24
energy_0 :-139
charge_0 :0
gradient_0 :2
gap_internal_0 :-24
boltzmann_factor_0 :-1
energy_density_factor_0 :
-2.0752e+09

Ortho 1 POG:



gap_0 :-22
energy_0 :-154.8
charge_0 :0
gradient_0 :0
gap_internal_0 :-22
boltzmann_factor_0 :-1
energy_density_factor_0 :-8.58897

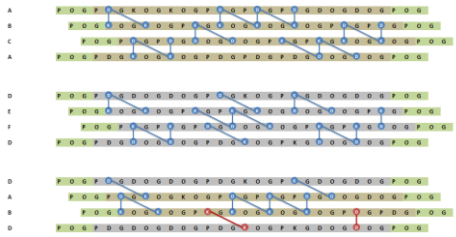
Ortho 2 :



gap_0 :-19
energy_0 :-135
charge_0 :0
gradient_0 :2
gap_internal_0 :-19
boltzmann_factor_0 :-1
energy_density_factor_0 :-
9.4678e+07

gap_1 :-19
energy_1 :-135
charge_1 :0
gradient_1 :2
gap_internal_1 :-21
boltzmann_factor_1 :-1
energy_density_factor_1 :-
1.19401e+08

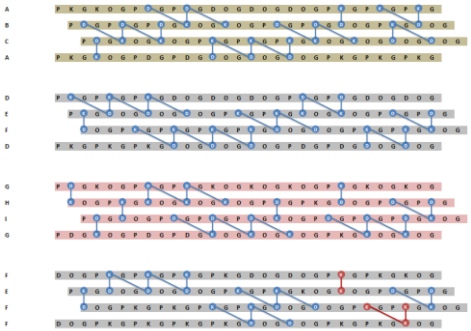
Ortho 2 POG :



gap_0 :-14
energy_0 :-154.8
charge_0 :0
gradient_0 :0
gap_internal_0 :-15
boltzmann_factor_0 :-0.999991
energy_density_factor_0 :
-0.00131485

gap_1 :-12
energy_1 :-152.8
charge_1 :0
gradient_1 :0
gap_internal_1 :-14
boltzmann_factor_1 :-0.999932
energy_density_factor_1 :
-0.00135172

Ortho 3 :



gap_0 :-14
energy_0 :-135

charge_0 :0
gradient_0 :2
gap_internal_0 :-14
boltzmann_factor_0 :-0.99999
energy_density_factor_0 :-489929

gap_1 :-15
energy_1 :-137
charge_1 :0
gradient_1 :6
gap_internal_1 :-15
boltzmann_factor_1 :-0.999998
energy_density_factor_1 :-419777

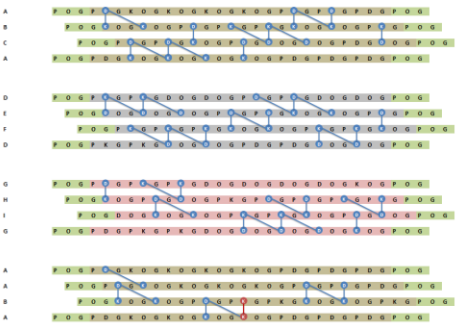
gap_2 :-13
energy_2 :-134
charge_2 :0
gradient_2 :2
gap_internal_2 :-13
boltzmann_factor_2 :-0.999975
energy_density_factor_2 :-526430

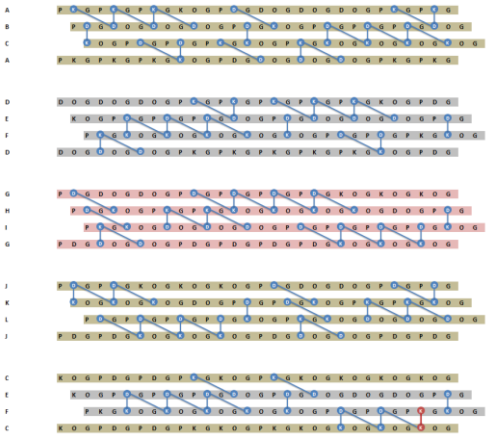
gap_1 :-12
energy_1 :-154.8
charge_1 :0
gradient_1 :0
gap_internal_1 :-12
boltzmann_factor_1 :-0.999922
energy_density_factor_1 :
-0.000159023

gap_2 :-9
energy_2 :-150.8
charge_2 :0
gradient_2 :4
gap_internal_2 :-9
boltzmann_factor_2 :-0.9979
energy_density_factor_2 :
-0.000321486

Ortho 4 Synthesized sequence :

Ortho 3 POG:





gap_0 :-9
energy_0 :-151.8
charge_0 :0
gradient_0 :4
gap_internal_0 :-9
boltzmann_factor_0 :-0.99872
energy_density_factor_0 :
-0.00019422

gap_0 :-9
energy_0 :-135
charge_0 :0
gradient_0 :2
gap_internal_0 :-27
boltzmann_factor_0 :-0.999711
energy_density_factor_0 :-16988.6

gap_1 :-7
energy_1 :-133
charge_1 :0
gradient_1 :2
gap_internal_1 :-17
boltzmann_factor_1 :-0.997781
energy_density_factor_1 :-16321.4

gap_0 :-10
energy_0 :-135
charge_0 :0
gradient_0 :2
gap_internal_0 :-20
boltzmann_factor_0 :-0.999886
energy_density_factor_0 :-42987.4

gap_2 :-4 (charge corrected)
energy_2 :-128 (corrected)
charge_2 :-4
gradient_2 :4
gap_internal_2 :-10 (corrected)
boltzmann_factor_2 :-0.999933
energy_density_factor_2 :-26938

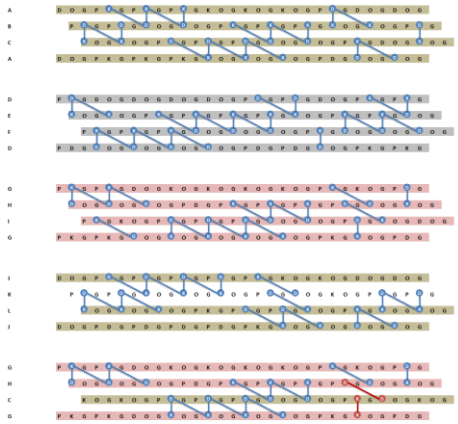
gap_1 :-13
energy_1 :-137
charge_1 :0
gradient_1 :2
gap_internal_1 :-13
boltzmann_factor_1 :-0.999978
energy_density_factor_1 :-29856.6

gap_3 :-3 (charge corrected)
energy_3 :-127 (corrected)
charge_3 :-4
gradient_3 :4
gap_internal_3 :-9 (corrected)
boltzmann_factor_3 :-0.999837
energy_density_factor_3 :-30080.1

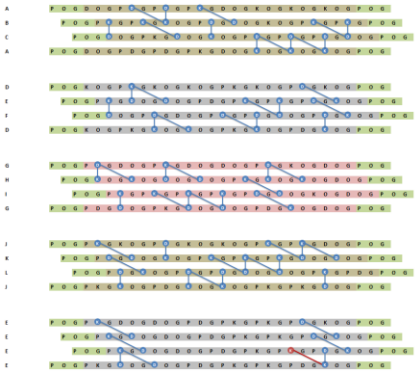
gap_2 :-8
energy_2 :-133
charge_2 :0
gradient_2 :2
gap_internal_2 :-17
boltzmann_factor_2 :-0.998874
energy_density_factor_2 :-32205.1

gap_3 :-11
energy_3 :-134
charge_3 :0
gradient_3 :2
gap_internal_3 :-15
boltzmann_factor_3 :-0.999854
energy_density_factor_3 :-91676.2

Ortho 4 :



Ortho 4 POG :



gap_0 :-8
energy_0 :-149.8
charge_0 :0
gradient_0 :0
gap_internal_0 :-11
boltzmann_factor_0 :-0.994651
energy_density_factor_0 :
-0.000341879

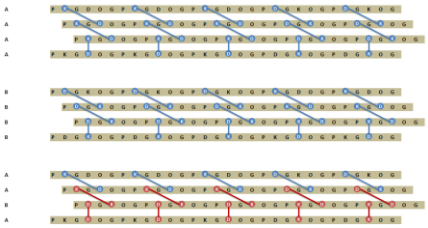
gap_1 :-7
energy_1 :-148.8
charge_1 :0
gradient_1 :4
gap_internal_1 :-7
boltzmann_factor_1 :-0.989594
energy_density_factor_1 :
-0.000475196

gap_2 :-8
energy_2 :-149.8
charge_2 :0
gradient_2 :4
gap_internal_2 :-9
boltzmann_factor_2 :-0.995223
energy_density_factor_2 :
-0.000383009

gap_3 :-10
energy_3 :-151.8
charge_3 :0
gradient_3 :0
gap_internal_3 :-10
boltzmann_factor_3 :-0.999568
energy_density_factor_3 :
-0.000576124

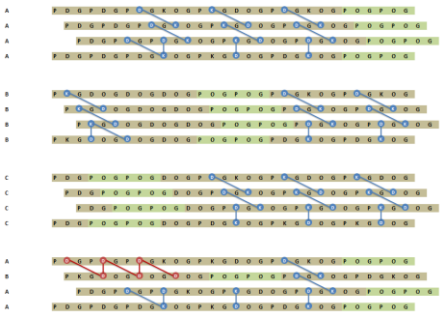
Fig s20: (below) Examples of designed *ab initio* sequence set for level 3 and 4 homotimer orthogonality along with one of their next most stable competing state sequence.

Ortho 2:

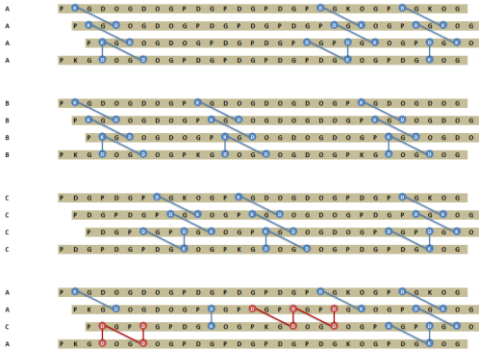


gap1: -35
gap2: -35
energy1: -129
energy2: -129

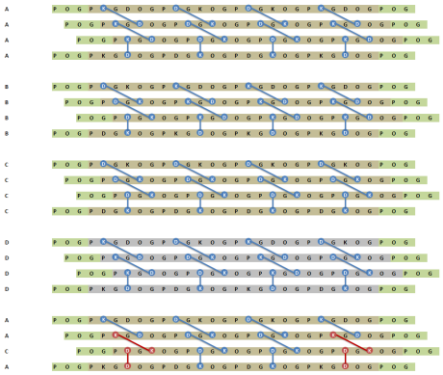
Ortho 3:



gap1: -19
gap2: -19
gap3: -18
energy1: -146.8
energy2: -146.8
energy3: -145.8

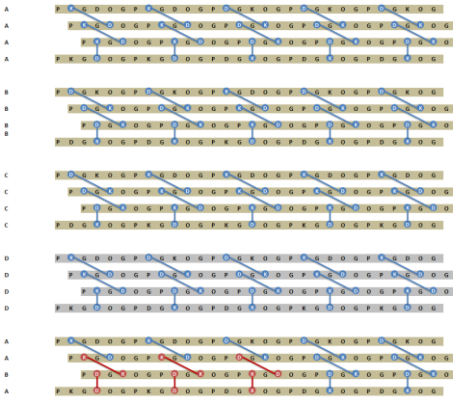


gap1: -25
gap2: -27
gap3: -26
energy1: -125
energy2: -126
energy3: -126



gap1: -14
gap2: -21
gap3: -14
gap4: -14
energy1: -148.8
energy2: -148.8
energy3: -148.8
energy4: -148.8

Ortho 4:

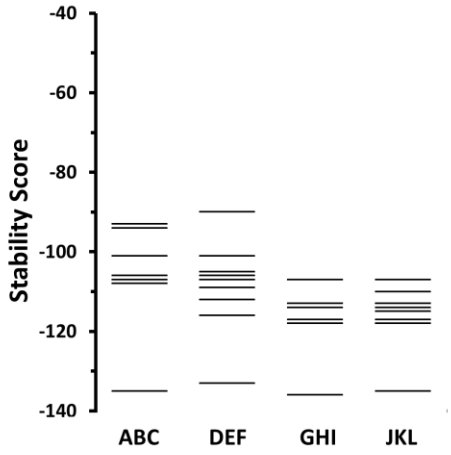


gap1: -21
gap2: -21
gap3: -21
gap4: -21
energy1: -129
energy2: -129
energy3: -129
energy4: -129

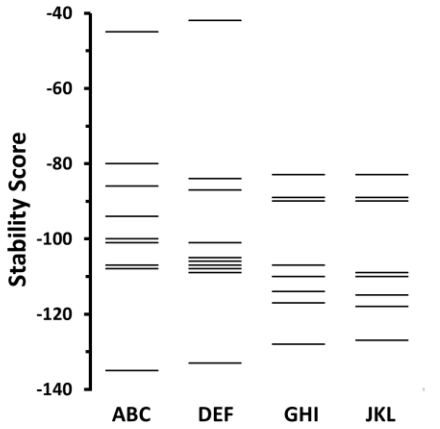
Fig s21 (below): Energy bar diagrams with possible alternate scores. Corrections to the old score include incorporating a small penalty for the effect of a triplet stagger or a net charge on the trimer states.

Part A: (10 states each)

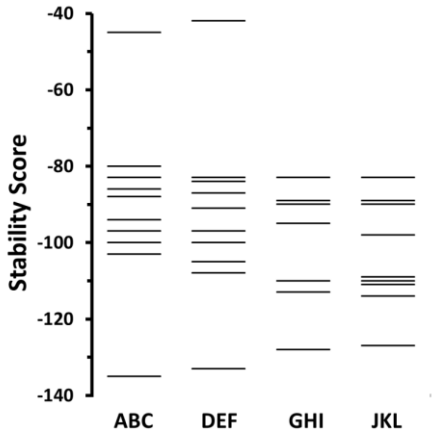
1) Old Score:



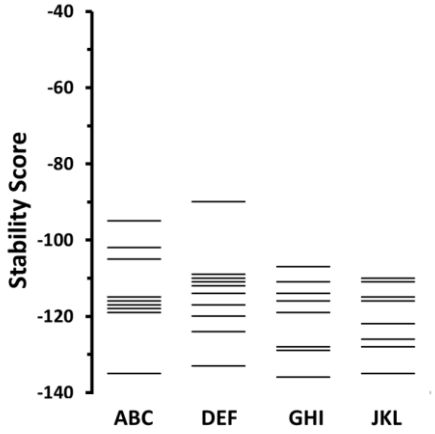
2) Old score Corrected for neg charge only .



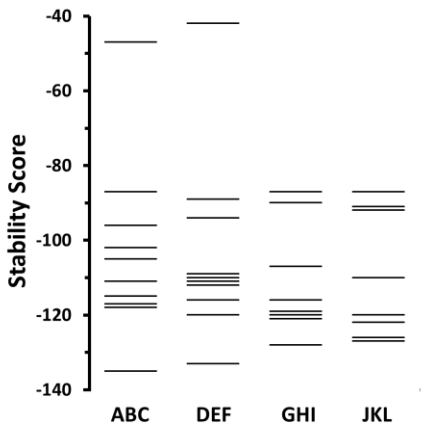
3) Old Score corrected for neg and pos charges on trimer .

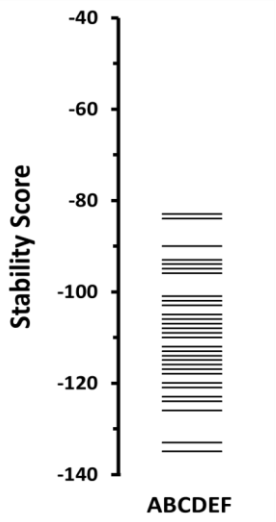


4) 1 Triplet Stagger left or right on any 1,2 or 3 strands, old score .

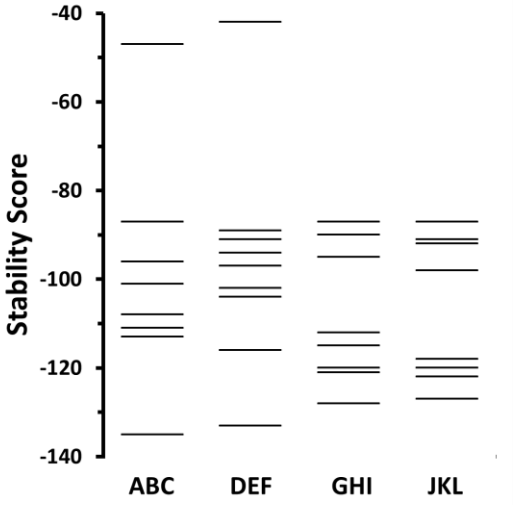


5) 1 Triplet Stagger, corrected for neg only score .

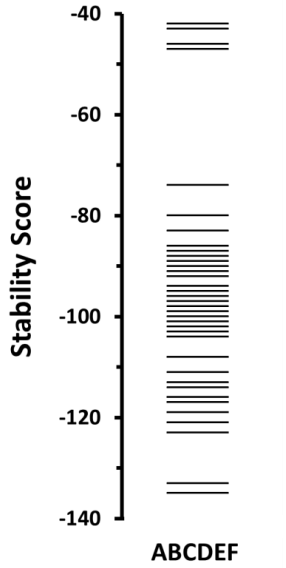




6) 1 Triplet Stagger, corrected for pos and neg score .



2) A to F, 1 triplet stagger, corrected for pos and neg score .



<u>Part B: (56 states each)</u>

1) **A to F, Old Score.**
   Most stable states in the bar diagram are representative of states = ABC, DEF, CEF, BCF .

Funding Sources

ACKNOWLEDGMENT

We gratefully acknowledge support from NSF DMR-0907273 and NIH DP2-OD-006478-1 to carry out this work.

ABBREVIATIONS

CMP collagen mimetic peptides

REFERENCES

1. X.H. Wang, D.P. Li, W.J. Wang, Q.L. Feng, F.Z. Cui, Y.X. Xu, X.H. Song, Mark van der Werf, Crosslinked collagen/chitosan matrix for artificial livers, Biomaterials, Volume 24, Issue 19, August 2003, Pages 3213-3220.
2. Ma, L., Gao, C., Mao, Z., Zhou, J., Shen, J., Hu, X., & Han, C. (2003). Collagen/chitosan porous scaffolds with improved biostability for skin tissue engineering. Biomaterials, 24(26), 4833-4841.
3. M. Koulikovska, M. Rafat, G. Petrovski, Z. Vereb, S. Akhtar, P. Fagerholm, N. Lagali, Enhanced regeneration of corneal tissue via a bioengineered collagen construct implanted by a non-disruptive surgical technique, Tissue Eng. Part A 21 (2015) 1116e1130.
4. Yu, S. M., Li, Y., & Kim, D. (2011). Collagen Mimetic Peptides: Progress Towards Functional Applications. Soft Matter, 7(18), 7927–7938.
5. Reyes, C. D., & García, A. J. (2003). Engineering integrin-specific surfaces with a triple-helical collagen-mimetic peptide. Journal of Biomedical Materials Research Part A, 65(4), 511-523.
6. Wright, C. F., Teichmann, S. A., Clarke, J., and Dobson, C. M. (2005) The importance of sequence diversity in the aggregation and evolution of proteins. Nature 438, 878-881.
   Xu, F.; Zahid, S.; Silva, T.; Nanda, V. Computational design of a collagen A:B:C-type heterotrimer. J. Am. Chem. Soc. 2011, 133, 15260–15263.
7. Lukatsky, D. B., Shakhnovich, B. E., Mintseris, J., and Shakhnovich, E. I. (2007) Structural Similarity Enhances Interaction Propensity of Proteins. J. Mol. Biol. 365, 1596-1606.
8. Han, J. H., Batey, S., Nickson, A. A., Teichmann, S. A., and Clarke, J. (2007) The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol 8, 319-330.
9. Yu, Z., An, B., Ramshaw, J. A., & Brodsky, B. (2014). Bacterial collagen-like proteins that form triple-helical structures. Journal of structural biology, 186(3), 451-461.
10. Yoshizumi, A., Yu, Z., Silva, T., Thiagarajan, G., Ramshaw, J. A., Inouye, M., & Brodsky, B. (2009). Self-association of streptococcus pyogenes collagen-like constructs into higher order structures. Protein Science, 18(6), 1241-1251.
11. Rutschmann, C., Baumann, S., Cabalzar, J., Luther, K. B., & Hennet, T. (2014). Recombinant expression of hydroxylated human collagen in Escherichia coli. Applied microbiology and biotechnology, 98(10), 4445-4455.
12. Boudko, S.P., J. Engel, and H.P. Bachinger, The crucial role of trimerization domains in collagen folding. Int J Biochem Cell Biol, 2012. 44(1): p. 21-32.
13. Ishida, Y., Kubota, H., Yamamoto, A., Kitamura, A., Bächinger, H. P., & Nagata, K. (2006). Type I Collagen in Hsp47-null Cells Is Aggregated in Endoplasmic Reticulum and Deficient in N-Propeptide Processing and Fibrillogenesis. Molecular Biology of the Cell, 17(5), 2346–2355.
14. Helseth, D. L., & Veis, A. (1981). Collagen self-assembly in vitro. Differentiating specific telopeptide-dependent interactions using selective enzyme modification and the addition of free amino telopeptide. Journal of Biological Chemistry, 256(14), 7118-7128.

15. Fallas, J. A., Dong, J., Tao, Y. J., & Hartgerink, J. D. (2012). Structural insights into charge pair interactions in triple helical collagen-like proteins. Journal of Biological Chemistry, 287(11), 8039-8047.
16. Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., ... & Xin, X. (2008). A specificity map for the PDZ domain family. PLoS Biol, 6(9), e239.
17. Tonikian, R., Xin, X., Toret, C. P., Gfeller, D., Landgraf, C., Panni, S., ... & Yu, H. (2009). Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. PLoS Biol, 7(10), e1000218.
18. Zarrinpar, A., Park, S. H., & Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature, 426(6967), 676-680.
19. Reinke, A. W., Grant, R. A., & Keating, A. E. (2010). A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. Journal of the American Chemical Society, 132(17), 6025-6031.
20. Bromley, E. H., Channon, K., Moutevelis, E., & Woolfson, D. N. (2008). Peptide and protein building blocks for synthetic biology: from programming biomolecules to self-organized biomolecular systems. ACS chemical biology, 3(1), 38-50.
21. Xu, F., Zahid, S., Silva, T., & Nanda, V. (2011). Computational design of a collagen A: B: C-type heterotrimer. Journal of the American Chemical Society, 133(39), 15260-15263.
22. Fiori, S., Saccà, B., & Moroder, L. (2002). Structural properties of a collagenous heterotrimer that mimics the collagenase cleavage site of collagen type I. J. Mol. Biol., 319(5), 1235-1242.
23. Ottl, J., Battistuta, R., Pieper, M., Tschesche, H., Bode, W., Kühn, K., & Moroder, L. (1996). Design and synthesis of heterotrimeric collagen peptides with a built-in cystine-knot Models for collagen catabolism by matrix-metalloproteases. FEBS letters, 398(1), 31-36.
24. Gauba, V., & Hartgerink, J. D. (2007). Self-assembled heterotrimeric collagen triple helices directed through electrostatic interactions. Journal of the American Chemical Society, 129(9), 2683-2690.
25. Dong, H., Paramonov, S. E., & Hartgerink, J. D. (2008). Self-assembly of α-helical coiled coil nanofibers. Journal of the American Chemical Society, 130(41), 13691-13695.
26. Gauba, V., & Hartgerink, J. D. (2007). Surprisingly high stability of collagen ABC heterotrimer: evaluation of side chain charge pairs. Journal of the American Chemical Society, 129(48), 15034-15041.
27. Parmar, A. S., Zahid, S., Belure, S. V., Young, R., Hasan, N., & Nanda, V. (2014). Design of net-charged abc-type collagen heterotrimers. Journal of structural biology, 185(2), 163-167.
28. Birk, D. E., Fitch, J. M., Babiarz, J. P., and Linsenmayer, T. F. (1988) Collagen type I and type V are present in the same fibril in the avian corneal stroma. The Journal of Cell Biology 106, 999-1008

29. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. J. Mol. Biol., 299(2), 283-293.
30. Jothi, R., Cherukuri, P. F., Tasneem, A., & Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. J. Mol. Biol., 362(4), 861-875.
31. Pazos, F., & Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. The EMBO journal, 27(20), 2648-2655.
32. Goh, C. S., & Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein–protein interactions. J. Mol. Biol., 324(1), 177-192.
33. Ramani, A. K., & Marcotte, E. M. (2003). Exploiting the co-evolution of interacting proteins to discover interaction specificity. J. Mol. Biol., 327(1), 273-284.
34. Chen, Z., & Zhao, H. (2005). Rapid creation of a novel protein function by in vitro coevolution. J. Mol. Biol., 348(5), 1273-1282.
35. Xu, F., Silva, T., Joshi, M., Zahid, S., & Nanda, V. (2013). Circular permutation directs orthogonal assembly in complex collagen peptide mixtures. Journal of Biological Chemistry, 288(44), 31616-31623.
36. Rich, A., and Crick, F. H. (1961) The molecular structure of collagen. J. Mol. Biol. 3, 483–506.
37. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9A ° resolution. Science 266, 75–81.
38. Baum, J., and Brodsky, B. (1997) Real-time NMR investigations of triple-helix folding and collagen folding diseases. Folding Des. 2, R53–R60.
39. Persikov, A. V., Xu, Y., and Brodsky, B. (2004) Equilibrium thermal transitions of collagen model peptides. Protein Sci. 13, 893–902.
40. Janin, J. (1996). Quantifying biological specificity: the statistical mechanics of molecular recognition. Proteins: Structure, Function, and Bioinformatics, 25(4), 438-445.
41. Parmar, A. S., Xu, F., Pike, D. H., Belure, S. V., Hasan, N. F., Drzewiecki, K. E., Shreiber D. I., & Nanda, V. (2015). Metal stabilization of collagen and de novo designed mimetic peptides. Biochemistry, 54(32), 4987-4997.
42. O'Leary, L. E., Fallas, J. A., & Hartgerink, J. D. (2011). Positive and negative design leads to compositional control in AAB collagen heterotrimers. Journal of the American Chemical Society, 133(14), 5432-5443.
43. Madhan, B., Xiao, J., Thiagarajan, G., Baum, J., & Brodsky, B. (2008). NMR monitoring of chain-specific stability in heterotrimeric collagen peptides. Journal of the American Chemical Society, 130(41), 13520-13521.
44. Fallas, J. A., & Hartgerink, J. D. (2012). Computational design of self-assembling register-specific collagen heterotrimers. Nature communications, 3, 1087.
45. Konak, A., Coit, D. W., & Smith, A. E. (2006). Multi-objective optimization using genetic algorithms: A tutorial. Reliability Engineering & System Safety, 91(9), 992-1007.

46. Calvo, F. (2009). Non-genetic global optimization methods in molecular science: An overview. Computational Materials Science, 45(1), 8-15.
47 Zitzler, E., Laumanns, M., & Thiele, L. (2001, May). SPEA2: Improving the strength Pareto evolutionary algorithm. In Eurogen (Vol. 3242, No. 103, pp. 95-100)
48. Wolfinger, M. T., Will, S., Hofacker, I. L., Backofen, R., & Stadler, P. F. (2006). Exploring the lower part of discrete polymer model energy landscapes. EPL (Europhysics Letters), 74(4), 726..
49. Smeeton, L. C., Oakley, M. T., & Johnston, R. L. (2014). Visualizing energy landscapes with metric disconnectivity graphs. Journal of computational chemistry, 35(20), 1481-1490.
50. De, S., Schaefer, B., Sadeghi, A., Sicher, M., Kanhere, D. G., & Goedecker, S. (2014). Relation between the dynamics of glassy clusters and characteristic features of their energy landscape. Physical Review Letters, 112(8), 083401.
51. Warszawski, S., Netzer, R., Tawfik, D. S., & Fleishman, S. J. (2014). A "fuzzy"-logic language for encoding multiple physical traits in biomolecules. Journal of molecular biology, 426(24), 4125-4138.
52. Michalewicz, Z. (1995). A Survey of Constraint Handling Techniques in Evolutionary Computation Methods. Evolutionary Programming, 4, 135-155.
53. Machta, J. (2009). Strengths and weaknesses of parallel tempering. Physical Review E, 80(5), 056706.
54. Vogel, T., & Perez, D. (2015). Towards an optimal flow: Density-of-states-informed replica-exchange simulations. Physical review letters, 115(19), 190602.
55. de Visser, J. A. G., & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics, 15(7), 480-490.