מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

עבודת גמר (תזה) לתואר
מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

מאת
עדי גולדנצוייג

Thesis for the degree
Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

By
Adi Goldenzweig

תכנון אפיניות בין חלבונים והנדסה מחדש של מסלולי קיפול בחלבון מודל

Affinity design of protein interfaces and reengineering of kinetic folding pathways

מנחה:
ד"ר שראל פליישמן

Advisor:
Dr. Sarel Fleishman

חודש ושנה עבריים
אדר א' התשע"ד

Month and Year
February 2014

Abstract

Protein-protein interactions (PPI) underlie all biological processes and make two prerequisites on the interacting pairs: that they recognize one another with sufficient affinity and specificity, and that they fold spontaneously into highly organized three-dimensional structures. Experiments with thousands of natural proteins have led to breakthroughs in our understanding of the key principles governing both molecular recognition and protein folding. Computational protein design provides an excellent way to test and refine our understanding of these principles. It also holds great promise for research and therapeutic use. I present two design projects implementing similar design strategies but with a major difference between them: First I present the first synthetic PPI network comprising one native hub-protein and three *de novo* designed binders, where the designed proteins interact through overlapping surfaces on the hub. Currently all three pairs bind with low to medium affinities. To generate variants that bind with very high affinity to their binder I implemented an interface design strategy based solely on the physico-chemical properties of the binding components (without using evolutionary or functional data). I report preliminary results on 3 high-affinity hub-protein designs validated experimentally. Further characterization of these mutants is required. Second, I collaborated with the Haran lab of the Faculty of chemistry to reengineer a complex folding pathway of their model system, the adenylate kinase (AK). AK comprises three domains, and previous experimental and computational work suggested a temporal path by which the domains fold. Relying on Hammond's postulate, I hypothesize that thermostabilizing AK's last-to-fold domain, may stabilize the final intermediate in the folding pathway, and lead to a change in the folding pathway. In addition to stabilization, here I wished to maintain AK function. To overcome this challenge I incorporated rich sequence information available for AK, and report several promising designs that showed different order of folding events compared to the wt enzyme according to molecular dynamics simulations performed at Takada lab in the University of Kyoto. All together I hope to demonstrate how protein design provides an important way to deepen our understanding of various requirements for protein function in our case folding, thermostability and molecular recognition.

Table of contents

Introduction.……3

Research aims…….....9

Results……..11

Discussion…….....22

Materials and methods.…….25

Figures……..…35

Tables……..47

List of abbreviations.…….49

References …….….50

Introduction

Background

Protein-protein interactions (PPI) play a key role in a variety of biological processes. Native protein interactions exhibit a broad range of affinities, specificities and kinetics allowing their function in the crowded environment of the cell. Despite the exponentially growing number of PDB structures (and the data accumulating from many studies on protein complexes), the physico-chemical principles governing molecular recognition are only partially understood. Identification of the energy features that govern molecular recognition in native interfaces is especially challenging, since various selective pressures and random drift acting on these interfaces affected their shape and properties. The features of binding affinity to a specific partner are not the only information hidden in the small details of a protein structure. Additional information about allosteric mechanisms, binding to multiple partners, protein stability and protein folding is also reflected in protein structures as they evolved to function in a multi-constraint environment, which is very different from the isolated *in vitro* environment used to study the basic chemistry of proteins. Computational design of PPI interfaces is a rigorous test to our current understanding of the principles governing molecular recognition. It provides a complementary way to refine our understanding of these principles in isolation from evolutionary pressures acting upon native protein systems. From an applicative point of view, the ability to design very high affinity interactions has great potential for therapeutics, diagnostics and research use. A key advantage of computational design over current methods such as repertoire screening is the potential ability to target specific regions of interest on the protein surface.

De novo design of protein interfaces has seen remarkable progress in the last decade. *Jha et al.* designed a naïve scaffold to bind PAK1 kinase with binding affinity of $K_d = 100\mu M$. Their design method included rigid-body docking followed by iterations of sequence redesign and minimization[1]. According to the design model the binder had a relatively small buried surface area of $1000A^2$, non-polar residues were over-represented compared to native interfaces (82% vs. 69%), and calculations suggested that binding may have occurred through more than one orientation. *Liu et al.* identified 3 key residues on erythropoietin that mediate erythropoietin's binding to erythropoietin receptor. They grafted these residues onto an unrelated scaffold to create a nanomolar binder[2]. This

method is limited by definition, as it requires a crystal structure of the desired target with a native binder. In both works, substitutions of key residues abrogated binding in agreement with the models, but crystal structures were not available. *Karanicolas et al.* designed a high-affinity interaction pair with a k_d of $130nM$ and k_d of $180pM$ following *in vitro* affinity maturation[3]. Their method consisted of introducing hotspot residues to both sides of the interface, redesigning the residues surrounding the hotspot with aliphatic residues to form a hydrophobic core, and finally designing the outer shell of the interface with more polar residues. The crystal structure of the affinity-matured complex revealed that the binding occurs through the designed surface but is rotated by 180 degrees compared to the design model, due to conformational changes at the designed binding surface. A follow-up analysis showed that sidechain conformations are significantly more restricted in natural complexes than in designed ones, suggesting that design algorithms underestimate the importance of sidechain rigidity at binding sites. In addition, high probability sidechain conformations in native proteins tend to make several interactions within the host protein, thus stabilizing the desired conformation, while low probability sidechain conformations, over-represented in designed interfaces, tend to be exposed in the unbound state[4]. This observation together with the 180 degrees rotated structure[3] emphasized the need to include elements of negative design in the design methodology to disfavor alternative conformations of the unbound hotspots that would be incompatible with the designed binding mode. However, standard negative design using the multi-state design approach[5] requires explicit enumeration of each of the alternative states. In most cases this is too computationally demanding.

A novel *de-novo* design methodology overcomes shortcomings in previous methods

To overcome the challenge of precluding off-target states, *Fleishman et al.* presented a general method for PPI design involving three main steps[6,7]. In the first step a library of disembodied residues is docked against a desired target surface to find residues that form favorable interactions with both the target and among themselves. These are defined to be hotspot residues, akin to interaction hotspots in natural PPIs. In parallel, the target surface is docked against a set of small monomeric protein scaffolds from the PDB to find high shape-complementarity orientations. In a second step the hotspot residues are incorporated onto a scaffold protein that passes the shape-complementarity step and is compatible with

hotspot geometry. In a third step the binding interface undergoes iterations of sequence redesign and minimization. A key point in this algorithm is the incorporation of negative design elements without explicit modeling of the negative states: the designed hotspots form stable interactions among themselves, thus stabilizing the binding conformation while precluding off-target states. Implementing this strategy, *Fleishman et al.* designed two proteins that interact with a conserved region on the influenza hemagglutinin protein and, following *in vitro* affinity maturation, bound with low-nanomolar dissociation constant. One of the binders was found to inhibit the hemagglutinin conformational change required for cell invasion. In further studies this algorithm was applied successfully to design an enzyme inhibitor and a pH-sensitive binder of the Fc domain of human IgG1 antibodies to serve as a superior antibody purification reagent[8].

The first synthetic protein network to study affinity and specificity in isolation from evolutionary pressures

In this thesis I present the first synthetic PPI network designed at the Baker lab using the hotspot methodology. The network comprises a native hub-protein and three *de-novo* binders, all binding the same surface on the hub-protein using physico-chemically similar hotspot residues. The hub-protein is M. tuberculosis acyl carrier protein 2 (Mtb ACP2), which was selected for its high stability and expression levels in standard bacterial strains. Unlike previous design work, Mtb ACP2 has no structurally characterized binding partners, broadening the hotspot method to a challenge that was not tackled previously. The three binders, codenamed E35, E52, E58, bind with low to medium affinities. Though using similar hotspot residues, the scaffolds of the three binders arise from diverse evolutionary origins and each uses a different binding footprint to bind the Mtb ACP2. Experimentally, this system was expected to behave very well since all binders originate from scaffolds that were successfully expressed using standard bacterial systems and due to ACP's high stability. With all these properties, this multi-specific and low-affinity network is a perfect model system to study how the key principles in molecular recognition, affinity and specificity, are encoded.

Main project: design high affinity interacting pairs

In my thesis I attempted to redesign ACP2 to generate variants that bind with very high affinities to each of the three binders. Our working hypothesis is that two-sided design (on

the binders' side when generating the network and on ACP in this thesis) can achieve very high-affinity pairs since all peptide degrees of freedom are submitted to design, similar to natural co-evolved interacting proteins. The design method iterates between sequence redesign on ACP's interface and backbone minimization treating separately each of the three ACP2-binder complexes. In a complementary thesis at the lab, ACP2 was redesigned to create highly specific pairs in which each ACP variant binds one of the three binders but not the other two. At this point, I report preliminary results on 3 high-affinity ACP mutants validated by yeast surface display. Further analysis and characterization of these mutants is required.

Affinity and thermostability are related protein design problems

High-affinity design is similar in some respects to the thermostabilization of a protein, where the protein core is analogous to a protein-protein interface. In both applications better packed cores (of the individual protein in the case of thermostabilization and of the interface in the case of a PPI) would lead to tighter interactions. In other respects, however, there are differences; most significantly a PPI comprises two protein entities that must be stable and soluble in separation, unlike a single protein. In practice, the affinity design methodology I implemented on the ACP system was developed and first implemented to design a novel protein fold[9]. A very similar design strategy was later used to thermostabilize an enzyme[10]. In a second project in this thesis I thermostabilize an enzyme domain for the purpose of a kinetic research presented below.

Protein design to modulate protein folding pathways

The folding of a protein to a highly organized three dimensional structure is an essential process for protein functionality, and over the years many severe diseases were found to be related to misfolded proteins[11]. Until recently, little was known about the remarkably efficient folding process of natural proteins, partly since it involves short-lived transition and intermediate states that require advanced systems for their capture. Intensive research on protein folding especially of small, single-domain proteins was conducted for more than the past two decades[12–14]. The findings in many works led to the perception that proteins have evolved to have smooth, funnel-like energy landscapes that minimize the presence of intermediates and kinetic traps. However, more than 70% of the eukaryotic proteome comprises large, multiple-domain, slow folding proteins[15]. As I stated in the affinity

project, protein structures reflect a compromise between various requirements, making it even more difficult to decode the key features governing protein folding by solely observing the folding of completely native proteins. Protein redesign can serve as a complementary way that directly probes factors governing folding kinetics, as it allows us to suggest hypotheses and then redesign protein folding pathways to test these hypotheses. However, since structures of folding transition states or intermediate states are not available, and the energy functions used in protein design are not suited to model dynamics, it is not obvious how protein design can affect folding kinetics. As a working hypothesis for this project I extrapolate from Hammond's postulate, which argues that two consecutive states with similar free energies in a chemical reaction will also have similar molecular structures[16]. The postulate effectively relates kinetics and thermodynamics if one of the states is the reaction's stable (and therefore structurally characterized) product and the other is the pre-product intermediate (which is unstable and therefore is difficult to structurally characterize). Under this rationale, computational protein thermostabilization (or destabilization) may affect the last folding intermediate (closest to the native structure of the protein) and by that affect the folding kinetics. For example, thermostabilizing the last domain to fold in a protein's folding pathway may affect its folding kinetics.

In a second project in collaboration with Haran lab at the faculty of chemistry, I suggest a novel design method to selectively thermostabilize one of three domains of the adenylate kinase (AK) enzyme, a model enzyme for complex folding, to reorder folding pathways.

The AK enzyme is responsible for energy balance in the cell, converting AMP, a dead-end substrate in the cell, into ADP. The chemical reaction is: $AMP + ATP \leftrightarrow 2ADP$. Fundamental to cell function, homologs of this enzyme are present from bacteria to humans. AK comprises 3 domains that according to previous studies strongly interact with one another and cannot be taken as independently folding units[17]. This made AK a popular model enzyme to study the folding of complex proteins that do not show a simple two-state folding behavior, but rather a more complex folding pathway involving several metastable intermediates. *Pirchi et al* performed single-molecule FRET (sm-FRET) experiments to identify 6 metastable intermediates on AK folding landscape[18]. *Li et al.* characterized the features of these 6 intermediates and transition pathways between them[19]. They found dependence between the pathway and the temperature. In the main folding pathway, the

nucleotide mono-phosphate binding domain (NMP) is the last to fold. Haran's lab is now interested to deepen the understanding of AK folding by introducing mutations that will reorder the folding events. Relying on Hammond's postulate[16], I hypothesize that thermostabilizing the last-to-fold NMP binding domain, may stabilize the last intermediate in the main folding pathway in which the NMP domain is unstructured. I expect a thermostabilized folding intermediate to change the prevalence ratio of folding trajectories connecting between the various intermediates. In more extreme cases it may even result in the abrogation of an existing state or the addition of a new one.

Thermostabilization of enzymes

Enzyme structures represent a molecular compromise between several requirements that together allow their proper function in the living cell. These requirements include well-defined catalytic-site geometry, flexibility often critical to enzyme activity, suitable stability and appropriate degradation in vivo. As a reflection of this compromise, most enzymes evolved to be just as stable as necessary to function in the living cell. However, this modest thermostability sometimes hampers the use of enzymes for research and industrial purposes, especially in processes that involve elevated temperatures. Computational thermostabilization of enzymes will therefore have a large variety of applications. Though similar to affinity design of binding interfaces, computational thermostabilization of an enzyme presents an additional challenge. The thermostabilized enzyme needs to maintain the exact geometry of the catalytic site and also allow for dynamic motions necessary for catalysis. If the catalytic site is known or when a crystal structure with a substrate analogue is available, this challenge is relieved since we can fix the catalytic site during design or add a filter that takes into consideration the analogue-bound state. However, none of these captures the dynamic motions.

Second project: domain thermostabilization to reorder protein-folding events

I suggest a novel design methodology for enzyme thermostabilization, which circumvents the high risk for loss of enzyme function. I combined a standard physico-chemical energy function with evolutionary data derived from more than 100 homologs to thermostabilize the NMP domain of *E. Coli* Adenylate Kinase enzyme. Incorporating evolutionary data in the design method allows us to constrain sequence identities in positions important for function (or any of the other objectives mentioned above) to those already sampled by

hundreds of millions of years of evolution; by contrast, positions where the functional constraints are weaker also have a weaker conservation signal allowing design to rely more on physico-chemical parameters. Effectively, this is another way to incorporate some negative design elements without explicit modeling of alternative states. From an applicative point of view, this project demonstrates the potential use of computational thermostabilization for cutting-edge research purposes, relating thermostabilization with a kinetic question.

I performed two rounds of design each with a different weight given to the evolutionary data. Several promising substitutions were selected for experimental validation. Five designs containing 1-3 of the selected substitutions were sent for molecular dynamics (MD) analysis in the Takada lab at the University of Kyoto; four showed promising results (see Results). These designs together with the rest of the selected substitutions are tested these days for higher thermostability at Haran's lab. Thermostabilized mutants will be further analyzed by sm-FRET to study the effects on folding.

Research aims

1. **Design very high affinity binding pairs starting from non-interacting proteins:** redesign ACP2 against each of the three *de novo* binders to generate variants that bind with very high affinity to their matching binder. Our working hypothesis is that two-sided design together with *in vitro* affinity maturation can achieve very high affinity binders. Two-sided design makes available all peptide degrees of freedom during the full design process, similar to natural co-evolved protein pairs. *In vitro* affinity maturation identifies complementary mutations that did not come up during the computational design. These mutations typically reflect shortcomings in the current design methods or energy functions. One prominent example is peripheral interface mutations of long-range electrostatics that are known to be poorly modeled in the energy function. Another typical example is backbone-affecting mutations; backbone conformation sampling increases the computational complexity and protein design has focused on rigid scaffolds until now. The expected result is a structural-energetic trajectory starting from non-interacting scaffolds, going through

a multi-specific network of three designed scaffolds, all binding a single target protein with low affinities and culminating in high affinity pairs.

2. **Redesign Adenylate kinase NMP domain for higher thermostability:** according to a previous work, the NMP domain of AK is the last to fold[19]. Based on Hammond's postulate[16], I hypothesize that thermostabilization of the NMP domain will affect AK folding pathways, by changing the population ratios in its six intermediate states[18,19] or by changing the number of intermediate states in more extreme cases. To avoid loss of function I combine evolutionary data from 114 homologs of AK in the design process. The expected result is a set of conservative mutations that exhibit higher thermostability. Typical mutations are expected to improve the core packing in the NMP domain. Other potential mutations might be backbone stabilizing mutations such as N- or C- terminal helix stabilizing mutations. The folding of thermostabilized mutants will be then analyzed through sm-FRET experiments in Haran's lab. I expect a sub-set of thermostabilized mutants to show a significant change in AK's folding pathway, thus demonstrating the potential of thermostabilization for cutting-edge research purposes.

Results

Main project

Preliminary results (done at the Baker lab)

The first synthetic PPI network as a model system to study molecular recognition

Target selection: with the aim to design *de novo* binders of a protein with no known natural binders, a set of 865 non-redundant protein structures was scanned to select an appropriate target. The set is comprised of proteins that are likely to be experimentally well behaved (*e.g.* expressed in bacterial systems, contain no disulfide bonds or cofactors, and are between 80 and 350 residues long). It was first culled by *Fleishman et al.* in a previous work[6]. Structures in bound form were removed from the set and the rest were scanned using the pocket software[20] to find at least three spatially clustered long hydrophobic residues surrounding a concave surface. The scan yielded a list of 8 proteins with surfaces that are potentially suited for hotspot regions of interaction.

De-novo binders with similar hotspots: for each target a hotspot region was generated from scratch by manually docking disembodied residues against the concave surface to complement its shape and electrostatic properties, and to make favorable contacts among themselves. The 2-3 hotspot residues were then refined automatically. In parallel, the same set of structures described for target selection was docked against each target to find shape complementary configurations. These were later scanned to find configurations that can accommodate the hotspot residues, following the same procedure as in ref[6]. Out of the 8 targets, designs targeting the Mtb ACP2 were the only ones to show satisfying computed binding characteristics (*e.g.* high shape complementarity, native-like buried surface area and satisfying binding energy) and were pursued further.

Experimental validation: 30 designs against ACP2 with satisfying computed binding characteristics were tested for binding ACP2 using yeast surface display[21]. Three designs bound specifically. The original scaffolds from which the designs were derived did not bind ACP2. Mutations to alanine of hotspot residues on the binders significantly reduced binding in all three cases. These results indicate that the binders bind through the designed surface. *Network features:* the described results provide the first synthetic PPI network comprising one native hub-protein, Mtb ACP2, with no known native binders, and three *de novo* binders codenamed E35, E52, E58. The binders use similar hotspot residues to bind

the concave surface of ACP (Fig. 1). The core of the interaction is comprised of a salt-bridge between a hotspot arginine on each of the binders and Glu45 on ACP. For E35 and E52 this interaction is strengthened by an additional hydrogen bond between the hotspot arginine and Thr42 on ACP. E58 shows a slightly shifted mode of binding and has instead an additional hydrogen bond between its hotspot arginine and the backbone carbonyl of Ile52. The other two hotspot residues are hydrophobic (Phe and Leu). They interact with the hotspot arginine to support its desired rotamer and preclude non-binding conformations. They also form van der Waals interactions with hydrophobic residues inside the concave surface of ACP (Fig. 1B and C). The binders' scaffolds are derived from diverse evolutionary origins and utilize different secondary elements to graft the hotspot residues. E35 is derived from a fold similar to Ran binding domain found in the hyperthermophilic bacterium *Thermotoga maritima*[22]. It utilizes the short loops of a beta sheet to bind ACP2. E52 is derived from a porphyrin binding protein found in *Arabidopsis thaliana*, and utilizes a helix to graft the hotspot residues[23]. E58 is derived from a thioredoxin protein found in *Neisseria gonorrhoeae*[24] bacterium and utilizes an eight amino acid loop to bind ACP2 (Fig. 1). All three binders bind with low to medium affinities. The buried surface area upon binding is $1300Å^2$ on average which is quite small but within the range of native proteins ($1{,}600Å^2 \pm 400\ Å^2$). All the network characteristics are summarized in table 1. Being multi-specific and of low affinities, with binders derived from different evolutionary origins all successfully crystallized in bacteria, this network was expected to be an excellent model system to study how the key principles governing molecular recognition, affinity and specificity, are encoded. Since ACP is a flexible protein and considering the network's diversity we hypothesized that two sided design has a great potential to achieve affinities and specificity switches not seen in previous design works. In this thesis I focused on ACP redesign against each binder to generate high affinity pairs. Another student in the lab studies specificity using the same network.

Current results

Computational redesign of Mtb ACP2

To generate very high affinity pairs I computationally redesigned the interface of ACP2 implementing a design strategy that was previously developed to design a novel protein fold[9]. Two critical features characterize this design protocol. One is the cycling between

sequence redesign and backbone optimization. The latter is analogous to structure relaxation under a fixed sequence and allows for better packing. For example, the introduction of a favorable amino acid may not be possible under fixed backbone due to very small clashes that elevate the energy. Backbone optimization may relieve the clash and lead to a better solution that would have not been accepted under fixed backbone. A second feature is the temporal reduction of the atom radii used for energy calculations. The temporal reduction (from 2.0Å to 1.6Å) smoothes the binding energy landscape and by that increases the algorithm's probability to cross energy barriers and approach the optimal solution. Only amino acids within 10Å from the interface were redesigned. Amino acids interacting with the hotspot arginine were fixed. Each pair of E-binder and ACP was redesigned separately. I asked for 20 output structures per pair. The outputs contained between 3 and 13 substitutions on ACP. All structures were visually inspected to prioritize the substitutions and omit those that seem unimportant (*e.g.* substitutions that look like a result of the algorithm's shortcomings or biases, or substitutions that came up since they improve the total score of the pair but not due to higher binding affinity). The manual scan included several steps. Generally it involved the examination of all energy values of substituted residues and nearby residues, in order to detect the effect of the examined substitution according to the model (*e.g.* is it better packing, a new hydrogen bond or perhaps better solvation or lower sidechain conformation energy). I also visually examined each output structure to characterize why the mutations were selected (*e.g.* cavity filling substitutions allowing for tighter packing at the interface, a hydrogen bond addition, a mutation allowing a nearby residue to adopt a different rotamer and contribute to binding). The computational design followed by the manual selection yielded a set of promising substitutions for experimental validation.

Experimental validation

I originally planned to create ACP libraries containing different combinations of the computationally proposed substitutions, in order to test and select for higher affinity designs using yeast surface display[21] (YSD). The method requires the expression and purification of the 3 *de novo* binders E35, E52, E58. After all 3 were found to be extremely unstable and prone to aggregation (see below), I inverted the system and implemented a low throughput approach instead. The rationale behind libraries creation and all the results

that preceded this inversion are detailed here. *Library rationale:* though the number of single substitutions offered by the computational design was 10-20 per pair of binders, combinations of double, triple or more substitutions raise the number of variants to test to impractical levels. Testing combinations of substitutions is extremely important as a single mutation might not yield a detectable effect. In addition, we might overlook the effect of coupled mutations or synergistic effects. I use YSD[21] as a high-throughput system to measure binding affinity of up to 10^7 different protein variants, and select for higher affinity binders. The method requires the preparation in advance of the two components participating in binding. All the desired variants of the designed protein are expressed as a library in yeast. The designed proteins are fused to a host protein that enables their display on the yeast surface and contain a Myc-tag at theirs C-terminus. The binding protein is expressed and purified as a soluble fraction and is biotinylated on a specific tag (Avi-tag) added to its C-terminus. During YSD the binding protein is incubated with yeasts expressing the different designs on their surface. In a second incubation, the proteins are labeled with anti-Myc and Streptavidin (anti-Biotin) fluorophores. After washing, the binding is measured using a flow-cytometer or a cell sorter in case we want to select for higher affinity variants. In my case I need three libraries of ACP variants (one per each E-binder). I also need E35, E52 and E58 purified.

ACP library against E35: I successfully prepared a library of ACP designs against E35 in *EBY100* yeast species. The library was prepared using primers with multiple nucleotide types at selected positions (see materials and methods), comprising about 2,000 variants, a quarter of which contain different combinations of E35 substitutions, the rest contain non-related amino acid codons that result from the use of multi-nucleotide primers for the library design. This library was not used eventually due to inversion of the experimental system (see *Experimental system inversion)*.

ACP library against E52 and E58: against E52 and E58 the number of single substitutions I wanted to test was three-fold higher, making the combinatorics too high for the use of multi-nucleotide primers. Therefore, I decided in these cases to create an error-prone PCR (epPCR) based library with an average of 1-2 substitutions per protein. The epPCR was limited to a sub-region in the gene that contained all the positions I was interested in. For a preliminary rough calibration of the substitutions average, I started with three libraries

prepared under different epPCR conditions. All three libraries failed, probably due to temperature problems during transformations. Before I managed to retry this calibration, the experimental system had been changed and there was no need for this library anymore (see *Experimental system inversion*).

E-binders purification: we aimed to purify the E-binders for measurements in YSD. All three binders are derived from scaffolds that were successfully crystallized in bacterial systems, thus we assumed that this would be experimentally straightforward. Unexpectedly, all the three *de novo* binders were found to be extremely sensitive. All showed low to modest solubility and had a strong tendency to aggregate. E35 expressed well but aggregated when removed from imidazole after Nickel purification. Gradual dialysis steps prevented the formation of eye-visible aggregates, but the protein was still mostly oligomeric and aggregated according to gel filtration analysis. A small non-aggregated fraction was isolated. When tested on YSD it bound ACP non-specifically with similar binding affinities to other unrelated proteins (insulin, E52), indicating that the protein is unfolded or aggregated. E52 was initially insoluble. Different solubilization strategies were carried out and succeeded to increase its solubility, but like E35 it tended to aggregate. Eventually, when a monomeric fraction was successfully isolated it did not bind ACP in YSD even though it is the highest affinity binder among the three. E58 was initially insoluble. Expression under various conditions (temperature, time, IPTG levels, buffers, bacteria strain) did not increase solubility and it was finally rescued when fused to a solubility tag, Maltose binding protein (Mbp). When Mbp was removed after Nickel purification, E58 quickly formed eye-visible aggregates. Since Mbp was fused at the N-terminus of E58, far from the interface, I decided to skip the Mbp pruning. This fusion seemed to stay soluble but did not bind ACP on YSD. Attributing the negative YSD results to biotinylation levels of E58, I re-expressed the Mbp-E58 fusion and purified it on avidin beads to ensure that I further analyze a 100% biotinylated fraction. It was tested in YSD and again did not bind. When analyzed on analytic gel filtration it was found to be oligomeric (for detailed description of all attempts to purify the proteins, see materials and methods). Due to these challenges and the limited time of the MSc thesis I decided to change the experimental testing protocol.

Experimental system inversion: the difficulties I experienced with the E-binders purification, together with the fact the all E-binders showed high expression levels on yeast surface and bound ACP-wt, led us to invert the experimental system: E-binders will be presented on the yeast surface while selected redesigned ACP variants will be expressed and purified separately in bacterial systems (Fig. 2). This inversion is a last resort as the method becomes quite low throughput requiring the separate expression and purification of each individual variant of interest. *In vitro* affinity maturation (of ACP), an important step for this project, is not possible under the inversed system.

Reselection of ACP designs for experimental validation: the YSD inversion requires the separate expression and purification of each individual variant of interest. The original plan was to select from among $\sim 10^3$ variants for E35 and $\sim 10^5 - 10^6$ variants for E52 and E58. Obviously this complexity was impractical if each variant needs to be separately cloned, expressed, purified, and tested. I re-examined the substitutions that had high priority previously and selected for the most promising ones: 6 substitutions for E35 and 7 substitutions for E58 (2 of which are common to E35). I decided to leave aside E52 that has already a medium affinity and focus on E35 and E58.

42 designs cloning into bacteria: I planned to synthesize 54 ACP designs including all possibilities for a single-substitution designs and a portion of the double-, triple-, or quadruple mutant combinations. Out of 54 I managed to clone 42 designs. Part of the rest failed and potentially required another set of primers, but I pursued them no further. I used site directed mutagenesis to generate some of the designs and restriction-ligation cloning for others. A list of all cloned designs is available in table 2.

ACP designs expression and purification: to facilitate the procedure and allow the parallel purification of several designs, I decided to perform only one step of purification on a Nickel column. Designs that show higher affinity on a YSD preliminary experiment will be further purified on a gel filtration column and then re-tested. The single purification step on a Nickel column removed most of the impurities. One major impurity that was not removed is a protein with the size of $\sim 36 - 40\ kD$. I assumed it is the birA enzyme used for *in vivo* biotinylation in our system (see Fig. 6 left gel). The YSD inversion adds noise which does not exist under the normal conditions: individually synthesized ACP designs show variability in biotinylation levels. I assume that this noise is negligible with respect to the

affinity changes I am expecting. For accurate measures of promising designs only, biotinylation levels may be evaluated later.

Experimental validation reveals 3 higher-affinity designs

ACP designs validation on YSD: I tested every ACP design that was expressed and purified on YSD. I report 3 designs that show higher affinity to their E-binder relative to ACP wild type. The three are codenamed dACP2 (Fig. 3), dACP42 (Fig. 4) and dACP5 (Fig. 5). dACP2 and dACP5 bound with higher affinity to both E35 and E58 although dACP2 is a design against E35 while dACP5 is a design against E58. This result becomes reasonable when looking at the actual substitutions. dACP2 has 3 substitutions and dACP5 5 has 4 substitutions (table 2). They share a pair of coupled substitutions (V27F and L37A). dACP2 contains a third substitution that is specific against E35 (L35E) while dACP5 contains a third substitution that is specific against E58 (S34Q). When I visualized these mutations I saw that they neither open cavities nor form clashes with any of the E-binders, rationalizing the observation that they experimentally raise affinity to all E binders. dACP5 contains an additional K61F substitution which came up also in the computational design against E35. dACP2 and dACP5 also showed higher affinity when tested against E52. This is more surprising since none of the substitutions in both designs came up in the computational design of ACP against E52. When modeled on *pymol* against E52, all substitutions seemed reasonable. In future work it could be interesting to understand why these substitutions did not come up computationally for E52. The third design, dACP42, was designed against E58 and showed higher affinity towards it. It is a double mutant containing two substitutions (T42Y, A57S) that are not common to dACP2 or dACP5 (Fig. 4). It was not tested yet against the two other binders, E35 and E52. According to YSD preliminary results, dACP5 (Fig. 5) is the most promising of the 3 higher-affinity designs and was first to be further analyzed. *Further analysis of dACP5:* I purified a new batch of dACP5 on analytic gel filtration following Nickel purification. The protein runs as a dimer on gel filtration indicating it is folded. All impurities left after Nickel purification were removed on gel filtration and the fraction I collected for further experiments shows high degree of purity on SDS-PAGE gel (Fig. 6). The pure fraction of dACP5 bound to all 3 binders with higher affinity relative to ACP wild type (Fig. 7). While ACP-wt at $0.1\mu M$ does not show detectable binding to E58, dACP5 at the same concentration showed a binding signal

higher than the wt at $1\mu M$ (Fig. 7D-F). I assess that dACP5 affinity towards the binders is 1-2 order of magnitudes higher than ACP-wt. This can be quantitatively determined by measuring k_d of the wild type protein and of dACP5 and will be done in future work. *Designs expression levels:* designs containing the V27F (coupled with L37A) or K61F substitutions had lower expression levels than ACP wt or other designs. Designs that had both of them like dACP5 showed even lower expression. A promising design, dACP1, containing these mutations together with a L35E substitution, showed extremely low expression levels and was not further pursued. Being non conservative and very aggressive is common to these mutations that change the net charge (K61F, L35E) or introduce a bulky amino acid instead of a small one (V27F) and vice versa (L37A). *A negative control is binding:* already at the Baker lab in Seattle, E52, the most promising binder, was tested for binding of ACP-triple knockout mutant. The mutant contains 3 physico-chemically conservative substitutions in the ACP surface involved in binding to all binders (Fig. 8), and was expected to reduce affinity. E45Q eliminates the salt-bridge between the hotspot arginine and the original glutamate. T42V eliminates the additional hydrogen bond with hotspot arginine. A57V results in a delicate clash with the arginine guandinium group (Fig. 8). The triple knock-out was purified and a CD spectrum indicated it is folded. When tested on YSD it did not show detectable binding to E52 at 1uM. After I finally had some positive results I decided to repeat this experiment at our lab. I expressed and purified ACP triple knockout as all other designs and tested it on YSD against all three binders. The triple knock-out bound all three binders with similar binding signals to ACP wild type at 1uM. Surprised by the results I thought that this might be due to the impurities left after Nickel purification. I therefore purified the triple knockout on a gel filtration column. The pattern resembled that of dACP5 and ACP-wt with the protein running as a dimer, indicating it is folded. I then re-tested the isolated dimeric fraction but it still bound all three binders with similar affinities to ACP-wt at 1uM and at 0.1uM (the latter was tested for E52 only). The triple-knockout was meant to provide key evidence for the designed binding modes. However, since all 3 substitutions are very mild, it does not necessarily mean that the predicted binding mode is wrong. In fact, in most biochemical analyses of binding interfaces, alanine or tryptophan scanning are done to validate binding modes. This kind of negative control is much more aggressive. Our current hypothesis is that the triple knockout

is too subtle. Though E45Q eliminates the salt bridge, the hotspot arginine can still make hydrogen bonds with the glutamine. T42V is a very delicate mutation and A57V results in only a small clash that can be potentially relieved by a minor shift in the binding orientation. I plan to design a new triple knockout with more aggressive mutations. For example, Glu45 will be mutated to a hydrophobic residue or even to a positive residue which is expected to result in significant electrostatic repulsion. This will be the first thing to be done in future work since all other results stand under a serious question mark as long as I do not strengthen the evidences supporting the designed mode of binding.

Secondary project

Computational redesign of AK for higher thermostability

Based on Hammond's postulate[16] I hypothesized that thermostabilization of the last-to-fold 47 amino acids NMP domain, will affect the folding pathway of the AK enzyme. Computationally, affinity and thermostability are similar challenges: the same chemical interactions dominate them both and maximizing these interactions is the guiding principle in both cases. Specifically, hydrophobic interactions govern the tight packing of stable protein cores and similarly allow the tight packing of binding surfaces and are essential for high shape complementarity. Hydrogen bonds play a key role in stabilization of secondary motifs and in connections between near-by motifs. Similarly, hydrogen bonds in binding surfaces stabilize the binding mode as they force specific geometries. I implemented a similar design strategy to that in the main project to redesign the NMP domain for higher thermostability. However, I decided to combine the bio-physical method with evolutionary data from 114 homologs. In the affinity project, evolutionary data was not available since I was designing *de novo* interactions. AK on the other hand, is a critical enzyme common to many living organisms. Sequence conservation allows the preservation of residues that are potentially essential for enzyme function, even if mutations in these residues increase thermostability. In fact, thermostability and function can be contradicting in some cases. AK is a flexible enzyme that undergoes tremendous dynamic motions during catalysis. A thermostabilizing mutation which adds chemical contacts between residues may abrogate critical dynamic motions. The enzyme had naturally evolved as a compromise between these demands. When we attempt to increase thermostability, the evolutionary data helps us trace function essential residues without explicit knowledge about the catalytic site or

enzyme catalysis cycle. I worked on the bacterial version of unbound AK (PDB: 4AKE). I started the process with a search for amino acid sequence homologs of the bacterial AK using the following server *http://consurf.tau.ac.il*. The search yielded 114 non-redundant homologs for which the average degree of conservation is extremely high (Fig. 9). I used the search results to derive a position-specific scoring matrix (PSSM). A PSSM is a table of values indicating how evolutionary conserved is any of the 20 amino acids at any possible position of a protein. A positive value indicates that an amino acid is more probable in a specific position than it would have been in a random position in a random protein. A negative value indicates that an amino acid is less probable at a certain position than it would have been in a random protein. The PSSM file can be easily incorporated in the energy function as an extra energy term. Higher PSSM scores lower the total energy score and vice versa. The PSSM energy term can be given any weight. I computationally redesigned AK NMP domain for higher thermostability. I repeated the design process twice with different stringencies for sequence conservation. The first round was done with a standard PSSM weight and the second was done with a low PSSM weight. I also defined a PSSM cutoff (score of 0) to narrow the sequence space and eliminate very unlikely identities during design. Taking into account the high flexibility of AK when no substrate is bound, I added a harmonic coordinate restraint term to the energy potential that penalizes large changes in backbone atom coordinates. I manually scanned all the output designs to find substitutions that fill cavities within the core of the NMP domain or others that form hydrogen bonds or stabilize helices termini. The design process yielded 4 conservative and promising substitutions (Fig. 10). V39I and V64I improve packing according to the model. S43T is a back to consensus substitution and potentially provides better packing. The most intriguing substitution is E44P at the N-terminus of a helix which is also a back to consensus one. Proline is known as a helix breaker and for its dramatic effects on folding which is exactly what I was interested in. Not surprisingly, all 4 substitutions had a positive PSSM score and 3 out of the 4 are conservative. A second round of design with lower PSSM score yielded a slightly different set of less conservative substitutions. New outputs were again manually scanned and 8 promising mutations were selected for experimental characterization. Since in this round I used a low PSSM score, some mutations may potentially reduce catalysis. Reduction in catalysis can be tolerated since the main purpose

of the project is to study folding. However, the 8 most promising mutations were prioritized according to their probability to reduce function and the more harmful ones will be tested only if the others do not show satisfying change in the folding routes (Fig. 11).

MD simulations of representative designs

I selected together with our collaborators 5 designs containing 1-3 of the selected substitutions from both design rounds. We sent them as PDB files to Shoji Takada in the University of Kyoto for MD analysis (see Fig. 12). They previously performed MD simulations of AK folding. They characterized the coarse structures of each of its intermediates and suggested favorable pathways of folding under different temperatures[19]. For each design they counted the number of folding trajectories via a certain sub-state. Summing all the data allowed them to compare the relative populations in each sub-state for every design. Their analysis revealed that 4 out of 5 designs tend to follow trajectories in which the NMP domain is folded at earlier stages relative to wild type AK (Fig. 12), consistent with our hypothesis that a more thermostable NMP domain will fold earlier in the folding route. Designs 1 and 5 showed the highest tendency to follow earlier folding of NMP trajectories. I note that these two designs contain mostly substitutions from the first design round where stricter PSSM restraint was used. It is encouraging that the MD analysis, which is independent of the design procedure I carried out, nevertheless validates the design calculations.

Experimental validation

The experimental validation is done these days at our collaborators' lab. Selected combinations of substitutions are cloned into bacteria and will be soon expressed and purified. They will be tested for higher thermostability on CD. Those who show higher thermostability will be further analyzed by sm-FRET experiments to reveal effects on folding.

Discussion

Protein design of high affinity interactions has seen remarkable progress in the last decade. The results from different design works were impressive and encouraging, but more important; they reflected deficiencies in our understanding of the features governing binding interfaces, preparing the ground for future work. Early works implemented a rigid body dock-design strategy[1,4]. The yield of low affinity interfaces that differed significantly from native ones led later works to focus on the formation of hotspot key residues that contribute most of the binding energy[2,3]. Results in several works emphasized the importance of precluding off-targets[3,4]. *Fleishman et al.* aimed to overcome this challenge and presented the hotspot design strategy[7]. They designed two proteins that bind the conserved region of *influenza* hemagglutinin, and seemed to overcome the off-targets problem[6]. However, in that work as in all previous ones there was some information about native binders of the target protein and/or the designed scaffolds. As the basis for this thesis the hotspot design strategy was extended to design *de novo* binders of a target protein with no known native binders. Many indirect evidences indicate that three *de novo* binders were successfully designed to bind Mtb ACP2 generating the first synthetic PPI network. To study affinity I aimed to redesign Mtb ACP2 to generate variants that bind with higher affinity to the *de novo* binders. I report three higher affinity ACP variants. Most promising is dACP5 that possibly reaches k_d's in the low nano-molar range. Examination of the most contributing substitutions in dACP5 is indicative of the current strengths and shortcomings in affinity design. dACP5, designed against E58, has 3 binding-relevant substitutions: V27F, S34Q and K61F. V27F and K61F significantly increase the hydrophobic surface in contact with E58 (Fig. 5C and D). S34Q adds more van der Waals interactions but also makes two hydrogen bonds with backbone carbonyls on a helix on E58 (Fig. 5B). dACP5 bound with significantly higher affinity to all three binders (Fig. 7). Obviously, the hydrogen bonding between dACP5 and E58 through Gln34 is not playing a role in the case of E35 and E52 that are derived from completely different scaffolds. It is therefore reasonable to assume that the major contribution to high affinity of dACP5 is a result of newly introduced hydrophobic interactions. The domination of hydrophobic interactions is not surprising since hydrophobicity is the key driving force for all biomolecular interactions. S34Q on dACP5 and A57S on dACP42 are potentially successful examples

for affinity increase through more complicated hydrogen bonding. However, they seem to have a modest contribution and it remains to be seen whether these interactions form experimentally in crystal structures as in our models. Since high affinity was mainly accomplished by design of hydrophobic interactions, it is not surprising that specificity did not emerge from these designs as specificity often demands electrostatic and hydrogen bonding. The successful design of the first synthetic PPI network is accompanied by the revelation of current deficiencies in our understanding of binding interfaces, preparing again the ground for future developments in design methodologies. All 3 *de novo* binders were found to be unstable with large fractions in the pellet and high tendency to aggregate. None of them was successfully purified using conventional expression and purification systems even though all three are derived from scaffolds that were successfully crystallized using bacterial systems. This means that the perception that scaffolds can readily adopt different functions is at least partially wrong. The use of rigid scaffolds without significant backbone sampling is obviously limited. Even if binding is accomplished, the design process may have undesired effects on the scaffold's solubility, stability and folding. Sequence and structure data together with information about the scaffolds' native function and properties will probably help to reduce these undesired outcomes of the design process. For instance sequence constraints extracted from phylogenetic analysis are useful for guiding sequence design choices (see the adenylate kinase design project, above). However, for many of the proteins in the *de novo* binder scaffold set such information is not available and the design process is entirely dependent on the design energy function. The realization that fixed backbone design is inherently limited implies what may be the future directions in protein design methodologies. Nature exhibits several excellent scaffold families that are used over and over again for different functions. Among them are the immunoglobulin family that can bind a very wide range of epitopes, the TIM-barrels that catalyse a broad range of chemical reactions and the ankyrin family used as binding platform in a wide variety of contexts. Common to these families is their modularity as they are all built from a stable structural framework on which conformationally variable loops are anchored. In antibodies these are the 6 loops that are anchored to a highly stable backbone that can accommodate for them. TIM barrels and ankyrins are fully built from a repetitive module, a $\beta - \alpha$ motif or a $\alpha - loop - \alpha$ respectively. The repetitive modules comprising a scaffold

slightly differ in sequence and structure. The scaffolds' modularity enables their combinatorial reconstitution through evolution which allowed them to change quickly to bind a wide variety of targets or catalyse diverse reactions. The focus on this kind of scaffolds in protein design holds two main advantages. One is that while all previous methods were based on fixed backbone design, combinatorial reconstitution of nature's excelling scaffolds is effectively backbone design. The second advantage is the wide literature and structural data available for these families, a result of their importance and exceptional abilities. All together this approach may provide a broad range of design solutions that were not available under fixed backbone, using scaffolds which evolved to accommodate for large changes while staying stable and functional, giving the designer much information to guide the design process.

AK thermostability design also presented in this thesis is complementary in some aspects to the affinity project. Thermostability design shares many common features with affinity design but in this case, working on a model enzyme I enjoyed rich conservation data that indeed served as a guiding light for the design process. In the framework I implemented the designer can choose how to incorporate the conservation data by using more a less strict cutoffs and restraints, depending on the problem. According to MD analysis the most promising mutants are AK-mut1 and AK-mut5, which contain only substitutions with high PSSM scores. If these results will be experimentally supported, it might mean that sequence data is not just useful to minimize the chances of loss of function but can actually direct the design to more optimal solutions.

Materials and Methods

Computational protocol of the affinity project

The main principles in the design method were described in the Result section. The design method has been implemented through RosettaScripts[25] and the script is available as Supplemental Data. I asked for 20 outputs (-out:nstruct 20) for each pair of binders.

The command line is:

rosetta_scripts.static.linuxgccrelease -s PDB file -parser:protocol XML file -out:nstruct NUMBER of desired output -ex1 -ex2 -use_input_sc -extrachi_cutoff 5 -ignore_unrecognized_res –chemical:eclude_patches (see list of patches below)

The PDB entries of the original scaffolds are 1VR8 (E35), 1Y6I (E52), 2H30 (E58).

Computational protocols of the thermostability project

The main principles in the design method were described in the Result section. I started by relaxing the original structure (pdb code: 4ake). This step is essential to adjust the structure for further design with Rosetta. The relaxation was done with the same energy function, coordinate constraint and movers later used for design (movers were restricted to repacking). The coordinate constraint penalizes large changes in backbone conformation, and was given a weight of 0.1. I then did two design rounds with different PSSM weights and slightly different movers. The input PDB file was the relaxed one from the previous step. In the first design round the PSSM weight was 0.25, the PSSM cutoff was 0, and the design mover was PackRotamersMover followed by minimization with MinMover. In the second design round the PSSM weight was 0.1 with a PSSM cutoff of -2. The protocol contained two iterations with PackRotamersMover followed by the FastRelax mover. The latter performs some more extensive minimizations and allows for tighter packing. The design methods have been implemented through RosettaScripts[25], and all scripts (relaxation 1, relaxation 2, design 1, design 2) are available as Supplemental Data. The command line is identical to the one in the affinity project with the relevant pdb and xml files. I asked for 5 outputs in each design round.

List of patches for exclude (-chemical:exclude_patches)

LowerDNA UpperDNA Cterm_amidation SpecialRotamer protein_cutpoint_upper protein_cutpoint_lower VirtualBB ShoveBB VirtualDNAPhosphate VirtualNTerm CTermConnect sc_orbitals pro_hydroxylated_case1 pro_hydroxylated_case2

ser_phosphorylated thr_phosphorylated tyr_phosphorylated tyr_sulfated lys_dimethylated lys_monomethylated lys_trimethylated lys_acetylated glu_carboxylated cys_acetylated tyr_diiodinated N_acetylated C_methylamidated MethylatedProteinCterm

Plasmids utilized in the affinity project

The original pET29b bacterial expression vector encoding ACP and the pETCON yeast display vectors encoding the 3 binders were obtained from the Baker lab, where they were designed and tested by Timothy Whitehead (now at Michigan State University) and Sarel Fleishman. Later in our lab ACP was also transferred from pET29b to pETCON and the 3 binders were transferred from pETCON to pET29b by Yearit Friedman. All plasmids were sequence verified.

Cloning of ACP designs into bacteria

Cloning of ACP designs was done using a pET29b vector (kanamycin resistance) containing the ACP-wt gene between NdeI/XhoI restriction sites followed by Avi-tag and then His-tag. The described vector served as a template for PCR amplification in one of two methods described below. I used the most frequent codons in *E. coli class II* to insert mutations into the primers. *Method I:* Site directed mutagenesis with the QuickChange Lightning kit (Catalog # 210518 Agilent technologies) to synthesize single mutant designs mainly and part of the double mutant designs. I followed the kit's manual for the PCR reaction and enzyme digestion with DpnI but used different cells and mediums for the transformation part. *Method II:* PCR with mutation-containing primers followed by Restriction-ligation with NdeI and XhoI restriction enzymes to synthesize the rest of the double mutants and the more complex mutants. Both primers also contained 8-9 random bases followed by NdeI restriction site for the forward primer and XhoI restriction site for the reverse primers at their 5' prime. Plasmids generated in both methods were then transformed into *E. cloni* electrocompetent cells (electroporation was followed by one hour recovery and plating on LB-kanamycin plates). For each clone a single colony was selected and grown overnight. At the next morning plasmids were extracted and sequence verified.

Expression and purification of ACP designs

E. coli BL21 (DE3) competent cells were co-electroporated with two plasmids. One is pET29b expression vector (kanamycin resistance) that contains an ACP design followed by Avi-tag for specific biotinylation and a hexa-histidine (His)-tag for protein purification. The

second is pBirACm vector (chloramphenicol resistance) for *In vivo* biotinylation (avidity https://www.avidity.com). pBirACm encodes the birA enzyme, a biotin-ligase that biotinlyates proteins on a specific sequence (Avi-tag). Both birA enzyme and ACP designs expression is induced with IPTG. After one hour recovery the cells were seeded on LB-kanamycin-chloramphenicol plates and grown O/N in $37°c$. Few colonies taken from the fresh plates were incubated in $10ml$ LB + $50\mu g$ kanamycin + 30 μg chloramphenicol for O/N. In the next morning the starter culture was diluted 1:250 into LB + kanamycin (5 μg per 1 ml culture), no chloramphenicol is added at this stage. The cells were grown in LB media with extra 0.5% glucose (to avoid expression leaking) until they reached an $OD_{600} = 0.6$. At 0.6 cells were induced with 0.8mM IPTG for over expression of both the ACP design and birA, and d-biotin was added in excess (pre-dissolved to form 5mM d-biotin solution in 10mM bicine buffer pH 8.3, final concentration was $50\mu M$). Cells were further grown in $16°c$ O/N (about 18-20 hours). The next morning the cells were pelleted and stored at -$20°c$ for at least few hours. Cells were resuspended in cold HBS buffer (20mM Hepes and 150mM NaCl pH 7.4 at $4°c$) containing a nuclease for DNA and RNA removal and a protease inhibitor. Resuspension volume was 2ml HBS for 100ml cell culture. Cells were then disrupted by sonication. I gave 9-12 short pulses of 10 seconds separated by long intervals (45-50 seconds) to keep the temperature low. Cell debris were removed by a 1 hour centrifugation at 20,000g followed by syringe-filtration of the clear supernatant. The lysate was then purified on a gravity column containing Ni-NTA beads (0.3ml beads per 2ml lysate). To maximize yields and purification degree on this single step I incubated the supernatant with the beads on the column for 5 minutes. I then washed the beads with x30 excess of 50mM Imidazole in HBS. Proteins were eluted with 250mM Imidazole in HBS buffer (elution buffer was also incubated with the beads on the column for 5 minutes). Elution volume was 1ml per 2ml loaded lysate. The elution fraction was loaded on SDS-PAGE gel for validation and then dialyzed against HBS (1:200 fraction/buffer ratio) twice to remove imidazole leftovers. Protein concentration was measured using a spectrophotometer at OD_{280}. Since the proteins contained impurities, the measurements were probably over estimating the ACP designs concentration. The purified proteins were stored at $4°c$. dACP5 design and ACP triple knock-out were further purified

on an analytical Superdex75 gel filtration column, a service provided by the Israeli structural proteomics center (ISPC).

Designs screening methodology

I used yeast surface display to screen for higher affinity ACP designs. I followed the procedure in reference[21]. In our system I presented E35, E52 and E58 on yeast surface (separately) using the pETCON vector (a hybrid of the pCTCON2 yeast surface expression vector with cloning sites from the pET vector series). The vector creates a fusion of the ORF to the surface-expressed Aga2p yeast protein. It also incorporates an N-terminal hemagglutinin (HA) tag and a C-terminal Myc-tag (for fluorescent antibody staining). I chemically transformed *S. cerevisiae* EBY100 cells with the pETCON vectors. The transformed cells were seeded on SDCAA -Trp plates. I stabbed few colonies from each plate to inoculate 5ml SDCAA cultures and let them grow at $30°c$ overnight. At the next morning I pelleted 1ml of the overnight culture, resuspended in 5ml of fresh SDCAA and let the cells grow for another 6-9 hours. I then pelleted 1ml of culture and resuspended it in 5ml of SGRCAA (has the same content as SDCAA media[21] with extra 20g of galactose and 23.6g of rafinose but only 1.1g of glucose) but only and induction media. I induced overnight for at least 20 hours. Right before the experiments I measured OD to use $1x10^6$ cells per sample ($OD_{600} = 1$ means $1x10^7$ cells/ml). I then followed steps 11-20 in reference[21] but washed with 0.5ml of PBSF rather than 1ml. For each sample I incubated $1x10^6$ cells expressing one of the E-binders with a specific ACP design or ACP-wt at concentrations varying from $100nM - 10\mu M$. The fluorophores used for labeling were in most experiments FITC-anti-mouse-anti-myc with Streptavidin-APC or Alexa647-anti-myc with Streptavidin-PE in few of the others. Second incubations were 10 minutes. I used accuri C6 flow cytometer to analyze the samples.

All attempts for expression and purification of the E-binders (E35, E52, E58)

We used pET29b vectors for bacterial expression with ORFs encoding to E35, E52 and E58, optimized for yeast codon usage. In general, the expression and purification procedures followed those described for ACP designs (see expression and purification of ACP designs). Elements that differed from the basic procedure will be detailed here for each case.

E35

Expression and purification

The basic construct was soluble when expressed as described above (induced with 1mM IPTG). Ni-NTA gravity column purification was performed with two washing steps at 20mM imidazole and 50mM imidazole in HBS. Elution was performed with 300mM imidazole in HBS. Imidazole was removed by dialysis or centricon (Millipore) to HBS and resulted in protein aggregates visible by eye. OD measurements indicated lost of almost all protein. When imidazole was not removed, the fraction looked transparent.

Aggregates following imidazole removal troubleshooting

Several approaches aimed to remove the imidazole while avoid aggregation. 1) We skipped dialysis and continued directly to size exclusion on a gel filtration column (superdex200 and later superdex75) at 4^0C. However, E35 aggregated on the column and never came out. 2) Purification with monomeric-avidin beads loaded on a gravity column (Pierce http://www.piercenet.com/product/monomeric-avidin-resins-kit) instead of Ni-NTA purification. We followed the manufacturer's manual but washed and eluted with HBS based buffers rather than PBS. The protein bound the beads but remained bound when eluted with 5mM biotin. Only the regeneration buffer (pH 2.8) successfully eluted the protein but since it is very acidic we did not further characterized the sample. 3) Buffer exchange using gradual dialysis following Ni-NTA purification: in the first step the protein was dialyzed to HBS with 500mM NaCl instead of 150mM and with 10% glycerol. In the second step the protein was dialyzed to HBS with standard NaCl levels and 5% glycerol. The procedure yielded a soluble protein. We further purified the protein on gel filtration (Superdex75, GE healthcare) to remove residual impurities, oligomers and small aggregates. At this point we tested several fractions of E35 for binding in YSD: a fraction after Ni-NTA (with 300mM imidazole), a fraction with 500mM NaCl in HBS, a fraction in standard HBS before gel filtration and a fraction after gel filtration. The YSD was done a next day after the gel filtration. All fractions bound ACP-wt to some extent but they also bound unrelated controls (insulin and E52) **.**

E52

Solubility troubleshooting

The basic construct was reported to be insoluble in a small-scale lysis with BugBuster protein extraction reagent (Novagen) under different induction times and temperatures. In order to increase solubility several things had been done. 1) The N-terminal domain of E52, which does not participate in binding to ACP, was removed (since we were not sure where is the optimal site to prune, we built two constructs one with amino acids 1-83 removed and the other with amino acids 1-92 removed). Removal was done by PCR on the original vector. The product was cloned into pET29b vector. 2) Expression and purification of E52 fused to maltose binding protein (MBP) solubility tag and a His-tag in its N-terminus. The expression vector was pET28 with a C-terminal Avi-tag that was already available at the lab. 3) Expression in *E. coli* Rosetta strain (DE3) that contains additional tRNA genes for translation of rare codons. If insolubility is attributed to translation difficulties this approach might solve the problem (the pET29b vectors we had contained yeast codon optimized genes). 4) E52 gene codon-optimized for *E.Coli* class II (using DNAworks website) was ordered from Gen9 and cloned into pET29b vector. All constructs were sequence verified. The constructs were expressed in small-scale (10ml culture) under the following conditions: 30^0C for 3 hours 0.1mM IPTG, 16^0C O/N 0.1mM IPTG, 16^0C O/N 1mM IPTG. The frozen cell pellet was resuspended in a buffer containing 50mM Tris-HCl, 500mM NaCl, 10% v/v glycerol, 1% tween-20, 10mM beta mercaptoethanol, pH 8.9 at 4^0C. This is the buffer that was used in the purification of the original scaffold[23], but with higher pH that is at least 1.5 pH units above the calculated isoelectric point (pI) of all constructs. From this point we will refer to this buffer as E52-buffer. All constructs besides the single-domain E52 were found to be soluble when expressed at 16^0C with 0.1mM IPTG. We decided to continue with the E52 codon optimized for *E. coli* construct described in number 4.

Codon optimized E52 expression and purification

We expressed in large scale. Cell pellet was resuspended in E52-buffer. Ni-NTA column purification was performed with two washing steps with 20mM imidazole and 50mM imidazole in E52-buffer and then elution with 300mM imidazole in E52-buffer. The eluted fraction was dialyzed against E52-buffer with no imidazole. We then purified a small-scale

with monomeric-avidin beads following the manufacturer manual (Pierce http://www.piercenet.com/product/monomeric-avidin-resins-kit), but with E52-buffer instead of PBS. A large fraction bound the beads. Half of it eluted in the elution step (5mM biotin) and half only at the regeneration step. After dialysis a small sample was analyzed on an analytical gel filtration column (superdex75) at the ISPC and was shown to be mostly aggregated with a small monomeric pick. We decided to purify the protein on a Superdex75 26/600 gel filtration column (GE) to isolate the monomeric pick (eluted after180-210ml).

E58

Solubility troubleshooting

A construct of E58 followed by c-Myc-tag was reported to be modestly expressing, soluble, but with low binding to Ni-NTA beads. Our construct of interest, E58-Avi-tag, was mostly insoluble. Several approaches to increase the solubility of E58-Avi-tag had been implemented successively. 1) Expression under the following conditions in small scale (10ml LB): 30°*c* for 3 hours, 30°*c* for 5 hours, 30°*c* O/N, 16°*c* for 3 hours, 16°*c* for 5 hours, 16°*c* O/N the frozen cell pellets were lysed with BugBuster protein extraction reagent (Novagen), in all cases induction was with 1.5mM IPTG. 2) Similar to number 1 but with cell pellet resuspension in the following buffer (from now on referred as E58-buffer): 50mM hepes, 300mM NaCl, 5% glycerol and 0.1% triton X-100 pH 8.3. 3) Expression of E58-Avi-tag and E58-Myc-tag (for comparison) in *E. coli* Rosetta strain (see E52 for strain details) instead of *E. coli* BL21, under the following conditions: 30°*c* for 3 hours 1mM IPTG, 30°*c* for 3 hours 0.1mM IPTG, 30°*c* O/N 0.1mM IPTG, 16°*c* for 3 hours 1mM IPTG. 4) Buffer calibration: conditions that showed some increase in solubility in number 3 were selected for re-expression and resuspension in three different buffers: the original E58-buffer (pH 8.3), E58-buffer at pH 6.3, E52-buffer (pH 8.9). The encouraging results from experiments in number 3 did not reproduce. 5) Solubility tag addition: I used the original pET29b construct as a PCR template to amplify the E58 followed by the C-terminal Avi-tag sequence. I then inserted the amplified product to a pET28 vector containing a N-terminal Mbp solubility tag followed by a His-tag and a TEV (Tobacco Etch Virus) cleavage site. Insertion was done by restriction-ligation between BamHI and SalI restriction sites. The clone was sequence verified. I expressed in small-scale the Mbp-E58

fusion in various buffers, temperatures, times and induction levels (not detailed). In all cases Mbp-E58 was found soluble. Further characterization was done under the following expression conditions: 16°c, O/N 0.1mM IPTG.

Soluble Mbp-E58-Avi expression and purification

With Mbp pruning: I expressed the Mbp-E58 fusion in large scale (3L LB). The frozen pellet was resuspended in E58-buffer and disrupted using a sonicator (six 30s pulses with intervals of 45s in between). I then purified the protein on Ni-NTA beads (after a small scale calibration). The lysate was loaded twice on the column for maximum binding. I performed two washing steps with 20mM and 60mM imidazole in E58-buffer and then eluted the protein with 120mM imidazole in E58-buffer. At the next step I cleaved the fusion with TEV enzyme (kindly provided to us from the Fass lab) for O/N at 4c. Cleavage conditions were selected after a small scale calibration with various times and temperatures, none of which yielded close to 100% cleavage). The cleavage site splits the protein into Mbp-His-tag fusion and E58-Avi-tag fusion. The protein was again purified on Ni-NTA beads. The separation was not satisfying with E58-Avi-tag coming out in the flow-through (as expected) but also with large fractions in the 20mM and 60mM imidazole washes, while Mbp-His-tag and non-cleaved Mbp-E58 fusion were eluted in 60mM and above. I therefore continued to a third purification step on superdex75 gel filtration column. Most of the cleaved E58 eluted together with leftovers of Mbp-E58-fusion indicating that it is mostly trimeric or aggregative. I isolated a small fraction of cleaved E58 that eluted after the cleaved Mbp indicating that this fraction is probably monomeric. I concentrated the monomeric fraction with a centricon (millipore). Some protein was lost in this step. An ELISA assay to qualitatively validate biotinylation indicated that the sample is biotinylated. At the next day the concentrated fraction had visible aggregates. *No Mbp pruning:* I repeated the expression and purification in large scale of Mbp-E58 fusion but skipped the TEV cleavage and the second Ni-NTA purification. I purified on gel filtration and collected fractions from the latter part of a peak containing Mbp-E58 fusion (this time I could not determine if the isolated fraction is monomeric, since I did not cleave and no free Mbp was available as a reference). I carefully concentrated the fusion to $\sim 5\mu M$. *Alternative purification on avidin beads:* I expressed the Mbp-E58 fusion under the previous conditions. I purified the protein on avidin beads in a gravity column following the

manufacturer's manual but with E58-buffer instead of PBS (after a small scale calibration). Most of the protein eluted in 2mM biotin in E58-buffer and a smaller fraction which I did not further characterize in the regeneration buffer. After a day in the fridge the eluted fraction had visible aggregates.

ISPC acknowledgement

I wish to thank the valuable consultancy regarding binders' expression and purification troubleshooting, given to us by Shira Albeck, Tamar Unger, Yoav Peleg and Orly Dym. A special thanks to Shira Albeck who analyzed our samples on analytic gel filtration and kindly assisted us with large scale purifications.

ACP library against E35

To introduce specific substitutions to ACP-wt two libraries were generated. The first had position 37 fixed to the wt identity (Leu37) while all other positions of interest had different identities comprising the wt identity, the desired substitiution and in some cases extra identities (resulting from codon combinatroics), excluding a V27F substitution that was not covered intentionally. The second library was practically identical but contained fixed substitutions at positions 27 and 37 (V27F and L37A). For each library I designed 2 pairs of primers. The first pair had a forward primer with a ~30 bases overlap to the template plasmid (pETCON), and a reverse primer complementary to part of ACP interface, containing multiple-base identities as positions of interest. The second pair had a forward primer complementary to part of ACP interface, containing multiple-base identities as positions of interest, and a reverse primer with a ~30 bases overlap to the template plasmid. Multiple base identities were selected to allow substitutions that are codon optimized to *E.Coli II*. The reverse primer and the forward primer of the first and second pairs respectively had a ~30 bases overlap. Two standard PCR reactions were carried out to generate the overlapping products, followed by DpnI digestion. A yeast culture was chemically transformed with both PCR products and NdeI/BamHI digested pETCON plasmid.

ACP random mutagenesis library against E52 and E58

I used the GeneMorph II random mutagenesis kit (Catalog No. 200550) for epPCR to introduce mutations into ACP-wt. The template was pETCON containing ACP-wt. Aiming for 1-2 mutations per gene (at the interface only) I designed a pair of primers that flank the interface and then performed 3 reactions with different template concentrations (6, 60 and 600ng) to calibrate the mutation rate. From this point I followed steps 35-46 in reference[21] but skipped the gel-purification of the epPCR products (step 34). pETCON plasmid was digested BamHI and NdeI (step 35).

Figures

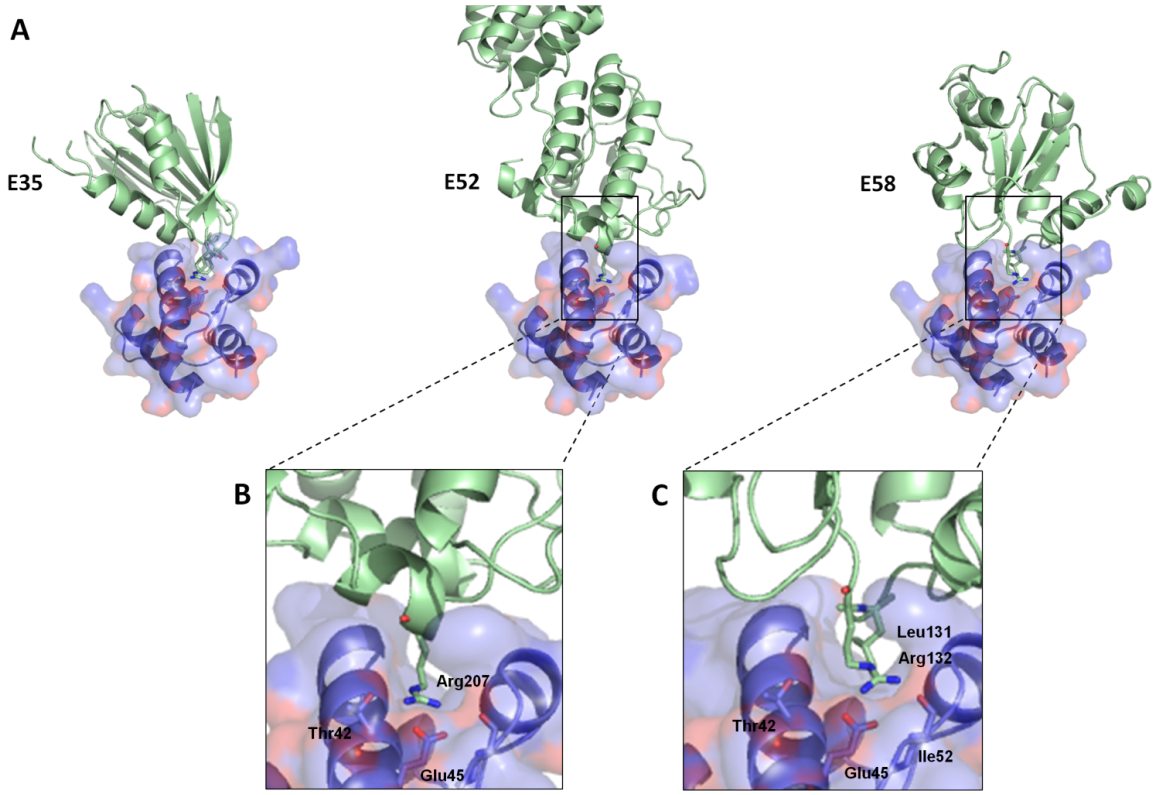

Figure 1: description of the first synthetic PPI network.

The network is comprised of a native hub-protein Mtb ACP2 shown in blue cartoon covered by semi-transparent surface, and three *de novo* binders shown in green cartoon. Residues that participate in the hotspot interaction are shown in sticks. (A) The 3 binders are derived from diverse scaffolds and use different binding footprints to bind the Mtb ACP2. E35 on the left binds through the short loops of a beta sheet, E52 in the middle binds through an alpha helix, and E58 on the right binds through an 8 residue loop. All binders use similar hotspot residues to bind the concave surface on ACP. (B) The core interaction between E52 and Mtb ACP2. The hotspot residue, Arg207 on E52, forms a salt bridge with Glu45 on ACP. Thr42 on ACP forms an additional hydrogen bond with the Guandinium group on the hotspot arginine. E35 interacts with ACP in the same mode of binding and forms the same contacts. (C) The core interaction between E58 and Mtb ACP2 is slightly shifted with respect to E52/E35. The hotspot Arg132 adopts a slightly different orientation relative to the concave surface. It forms a salt bridge with Glu45 on ACP and an additional hydrogen bond with the backbone carbonyl of Ile52 on ACP. Another hotspot

residue, Leu131, supports the binding conformation of Arg132 from behind and forms additional contacts with hydrophobic residues inside the ACP pocket.

Figure 2

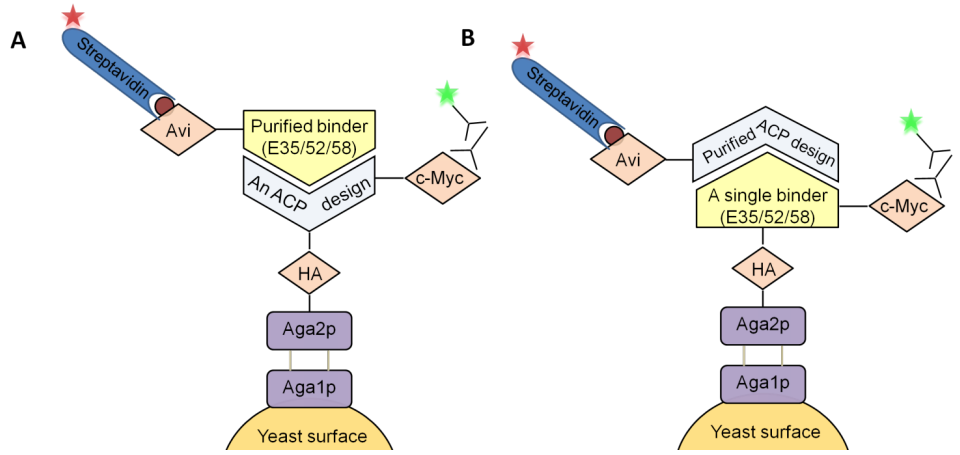

Figure2: Introduction of the originally planned (A) and inverted (B) YSD experimental systems.

The classic system allows high throughput screening of up to 10^7 designs. Yeast cells are expressing different designs on their surface. A single cell displays on its surface up to 50,000 copies of a single design. At the N-terminus the designs are fused to the Aga2p protein, naturally used for cell mating, followed by a HA-tag and a linker. At the C terminus the designs contain a c-Myc tag. During the main step the yeast culture is incubated with the protein to which the designs are expected to bind. The protein has been previously purified and has a biotin molecule on a specific tag at its C-terminus (Avi-tag). In a second incubation Streptavidin (anti biotin) conjugated to a fluorescent dye and an anti c-Myc antibody conjugated to another dye are added. After washing, cells are injected to a cell sorter to measure binding levels and select for binding designs. Correlated high levels of both dyes during measurement indicate a well expressed design that binds with high affinity to its partner. (A) A scheme of the classic system as I planned to use it. Libraries of ACP designs are displayed on yeasts in a culture. The classic system requires the purification of E35, E52 and E58 but we failed to purify all three. (B) A scheme of the inverted system which I eventually used. The E-binders are the ones presented on yeasts surface, a single binder per cell culture. Under this inversion each ACP design is purified

and tested separately. By using the inverted system I circumvent the purification problems I had but the method becomes quite low throughput.

Figure 3

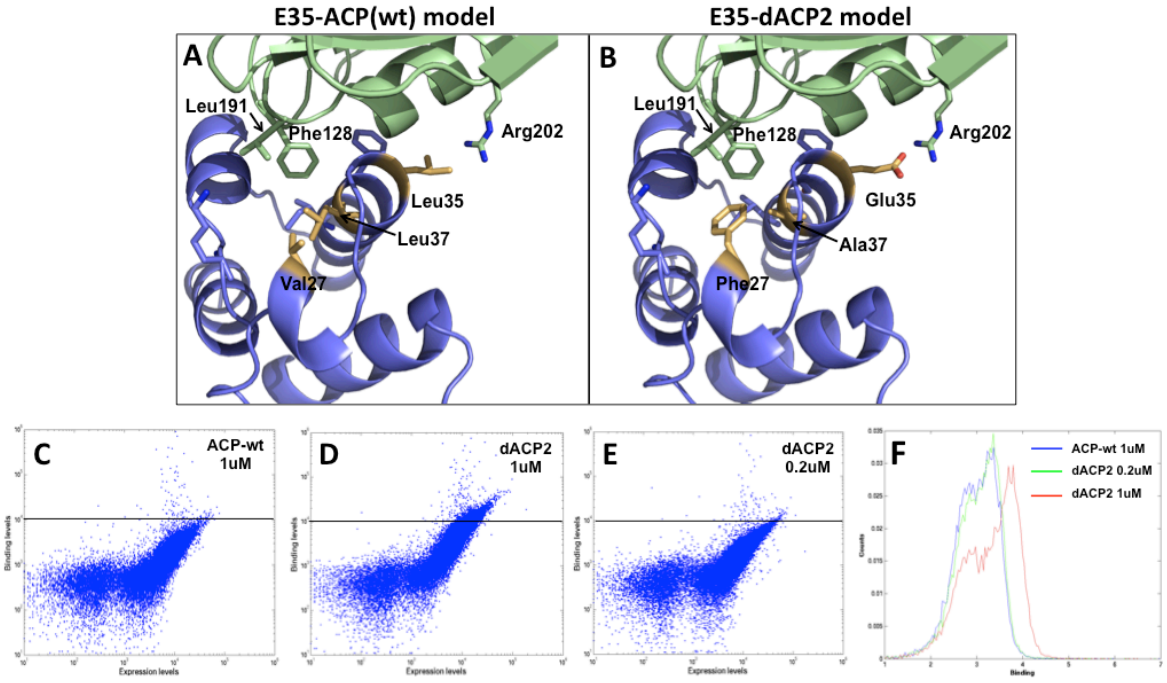

Figure 3: **Computational model and YSD results for dACP2 design against E35.** Computational models of ACPwt-E35 (A) and dACP2-E35 (B) complexes. E35 is shown in green cartoon, ACP-wt and dACP2 are shown in blue cartoons. Residues that differ between the wt and dACP2 are shown in orange sticks. V27F substitution partially seals a cavity and significantly increases the hydrophobic surface contact between dACP2 and E35. It forms interactions mainly with the hotspot residues Phe128 and Leu191 on E35. The latter residue is 'floating' and exposed when in complex with ACP-wt. L37A is required to allow the V27F substitution; otherwise the introduced phenylalanine would clash with the original leucine. L35E substitution forms a hydrogen bond with Arg202 on the binder. (C-E) YSD results for dACP2 and ACP wt. On the x-axis the expression levels of E35 on yeasts. On the y-axis the binding levels of ACP-wt $1\mu M$ (C), dACP2 $1\mu M$, (D) dACP2 $0.2\mu M$ (E). Both axes are in logarithmic scale which means that differences that look small by eye are actually quite significant. For convenience there is a black line indicating binding levels of 10^4. (F) A single histogram comparing ACP-wt and dACP2. On the x-axis the different binding levels, on the y-axis counts percentage for each binding

level. Figures C-F show that dACP2 binds E35 with higher affinity. At $0.2\mu M$ dACP2 binds roughly the same as ACP-wt at $1\mu M$. To be noted that dACP2 contains more impurities than ACP-wt, thus its actual concentration is probably lower making the results potentially more significant.

Figure 4

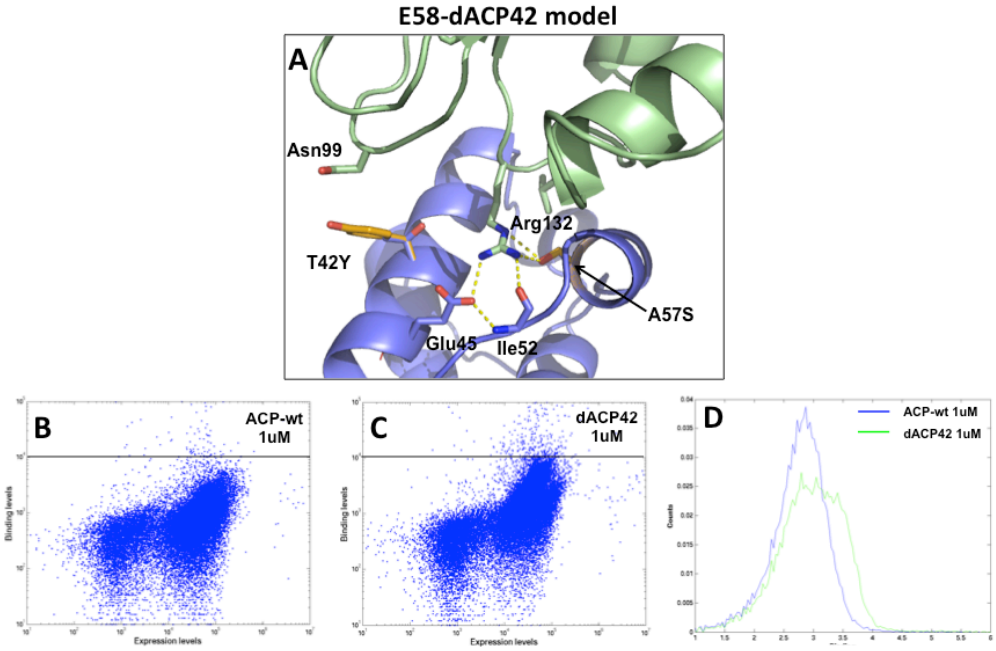

Figure 4: Computational model and YSD results for dACP42 **design against E58.**

(A) Computational models of ACPwt-E58 and dACP42-E58 complexes. E58 is shown in green cartoon, ACP-wt and dACP42 are shown in blue cartoon. Residues important for the interaction are shown as sticks (note that for Ile52 only some backbone atoms are shown). dACP42 residues, T42Y and A57S, are shown as orange sticks. Ser57 that replaced an alanine forms hydrogen bonds with the hotspot Arg132. T42Y substitution provides slightly better packing and potentially a hydrogen bond with Asn99. (B-C) YSD results for dACP42 and ACP-wt at $1\mu M$. On the x-axis the expression levels of E58 on yeasts. On the y-axis the binding levels of ACP-wt $1\mu M$ (B) and dACP42 $1\mu M$, (C). Both axes are in logarithmic scale which means that differences that look small by eye are actually quite significant. For convenience there is a black line indicating binding levels of 10^4. (D) A single histogram comparing ACP-wt and dACP42. On the x-axis the different binding levels, on the y-axis counts percentage for each binding level. Figures B-D show that dACP42 binds E58 with higher affinity than ACP-wt. To be noted that dACP42 contains

more impurities than ACP-wt, thus its actual concentration is probably lower making the results potentially more significant.

Figure 5

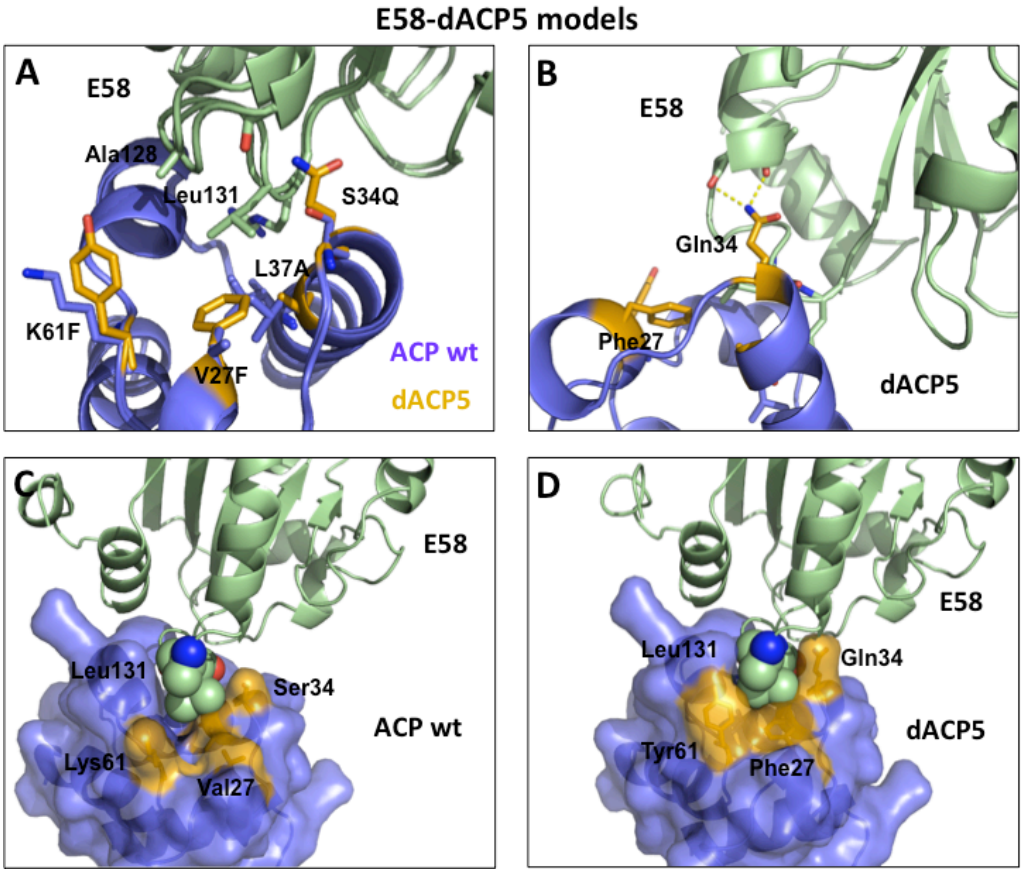

Figure 5: Computational model of the most promising design dACP5 **against E58.**

(A) Computational models of ACPwt-E58 and dACP5-E58 complexes. E58 is shown in green cartoon, ACP-wt and dACP5 are shown in blue cartoon. Interacting residues are shown as sticks. dACP5 residues, V27F/S34Q/L37A/K61F are shown as orange sticks. As for dACP2, here also V27F substitution partially seals a cavity and significantly increases the hydrophobic surface contact between dACP5 and E58. It forms interactions with the hotspot residue Leu131 on E58. L37A is required to avoid a clash between the newly introduced Phe27 and the original Leu37. S34Q substitution provides slightly better packing with Leu131 but more important, the longer glutamine forms hydrogen bonds with backbone carbonyls on a helix of E58 (B). K61F substitution has a similar effect to V27F as it increases the hydrophobic contact surface and completes the cavity sealing. (B) The polar contacts between Gln34 on dACP5 and backbone carbonyls on E58. (C) A surface

representation of ACP-wt reveals an exposed cavity on ACP binding tunnel which is not in contact with any residue on E58. (D) A surface representation of dACP5 shows how the cavity is sealed due to V27F and K61F substitutions. Leu131 on E58 now forms tight contacts from all sides. For experimental characterization of dACP5 see Figure 7.

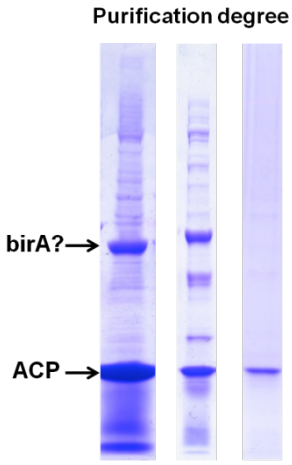

Figure 6: ACP designs and dACP5 purification degree.
ACP designs were purified in one step of Nickel purification. On the left gel the purification of a representative design. A significant impurity is left and is probably the enzyme birA. This pattern characterized most of ACP designs. In the middle gel is dACP5 after Nickel purification. The bend is thinner relative to the representative design reflecting lower *in vivo* expression levels. This phenomenon characterized several aggressive designs. On the right gel is dACP5 after size exclusion. This sample was used for further YSD analysis of dACP5 (see Fig. 7)

Figure 7

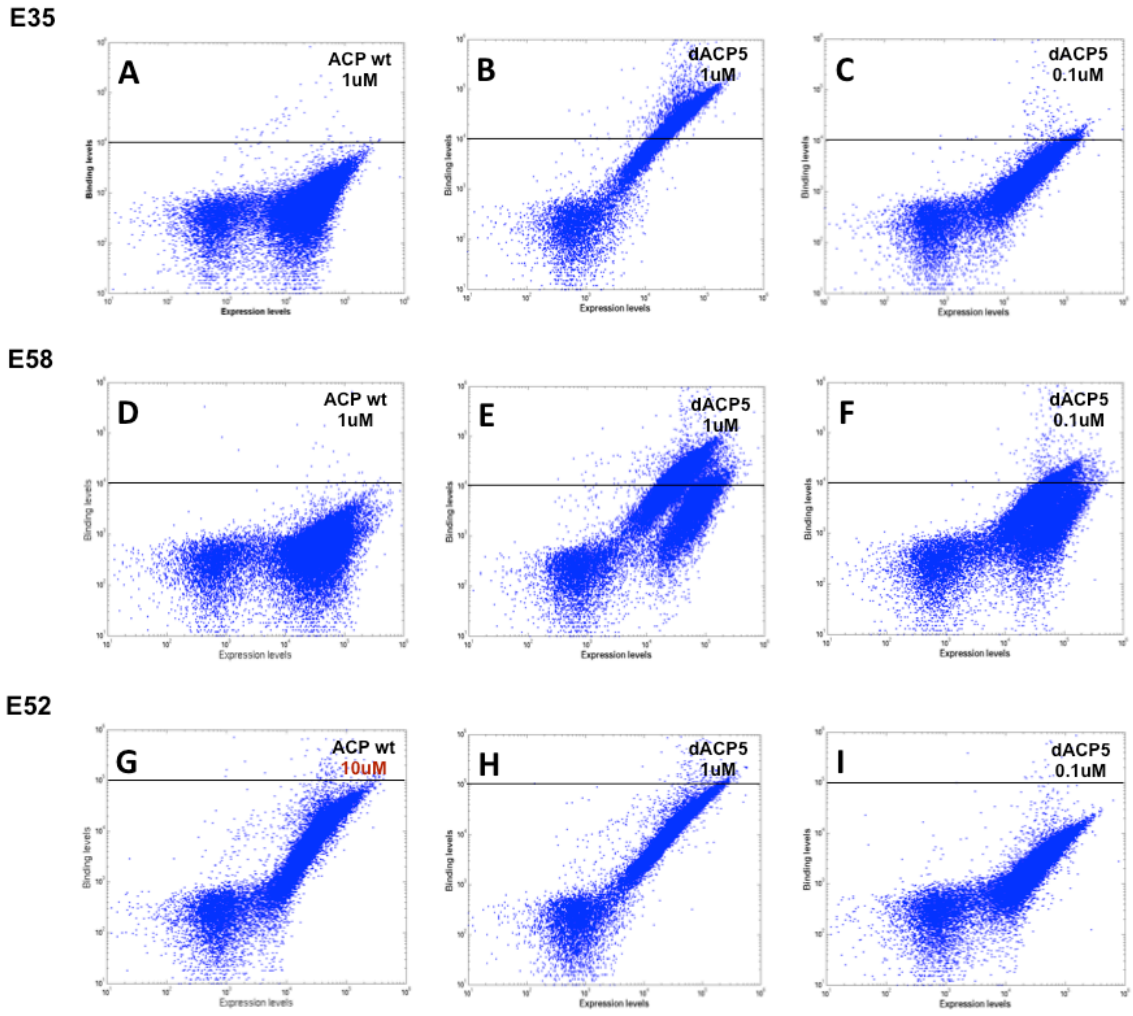

Figure 7: YSD results for dACP5 **after size exclusion**

dACP5 contains 4 substitutions, 3 of which are common to E35 and E58. None of the 4 came up in the computational design of E52. (A-C) E35 binding pattern. On the x-axis the expression levels of E35 on yeasts. On the y-axis the binding levels of ACP-wt $1\mu M$ (A), dACP5 $1\mu M$ (B) and dACP5 $0.1\mu M$ (C). A black line is indicative of binding levels of 10^4. At $0.1\mu M$ dACP5 still shows better binding to E35 than 10 times more concentrated ACP-wt. (D-F) E58 binding pattern, the binder against which dACP5 was designed. On the x-axis the expression levels of E58 on yeasts. On the y-axis the binding levels of ACP-wt $1\mu M$ (D), dACP5 $1\mu M$ (E) and dACP5 $0.1\mu M$ (F). A black line is indicative of binding levels of 10^4. At $0.1\mu M$ dACP5 shows significantly better binding to E58 than 10 times more concentrated ACP-wt. (G-I) E52 binding pattern. On the x-axis the expression levels

of E52 on yeasts. On the y-axis the binding levels of ACP-wt $10\mu M$ (G), dACP5 $1\mu M$ (H) and dACP5 $0.1\mu M$ (I). A black line is indicative of binding levels of 10^5. Both axes are in logarithmic scale which means that differences that look small by eye are actually quite significant. dACP5 binds with significantly higher affinity to all three binders. According to these results I expect affinity increase of 1-2 orders of magnitudes, potentially more for E58.

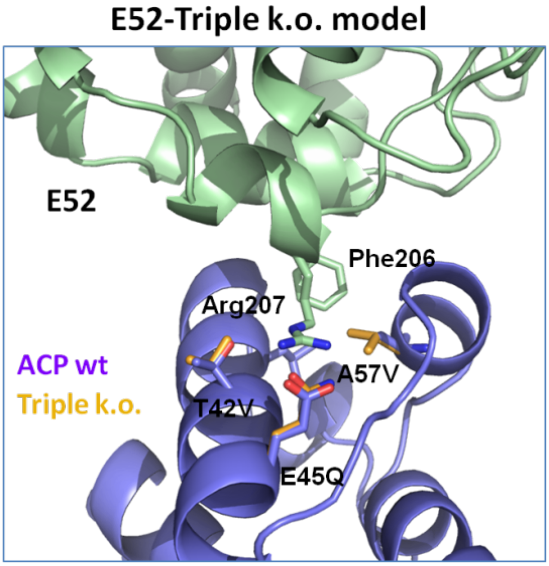

Figure 8: The triple k.o. design for a negative control.

The design contains 3 substitutions that are shown here as part of a complex with E52, but were expected to have similar effect on all 3 *de novo* binders. E52 is shown in green cartoon, ACP-wt and triple k.o. are shown in blue cartoon. The 3 substitutions are shown in blue sticks (wt) and orange sticks (triple k.o.). E45Q is assumed to abrogate the salt bridges between Glu45 and the hotspot arginine. T42V eliminates the hydrogen bond with the hotspot arginine (not relevant for E58). Valine instead of alanine in position 57 is expected to cause a small clash with the hotspot arginine. All together the triple k.o. was expected to significantly reduce binding. YSD experiments at the Baker lab in Seattle showed no binding of the triple k.o. to E52 at $1\mu M$ and lower. It was not tested there then against E35 and E58. When I repeated the experiment, the triple k.o. bound all 3 binders with similar affinities to ACP-wt (data not shown). The current hypothesis is that this control might be too mild.

Figure 9

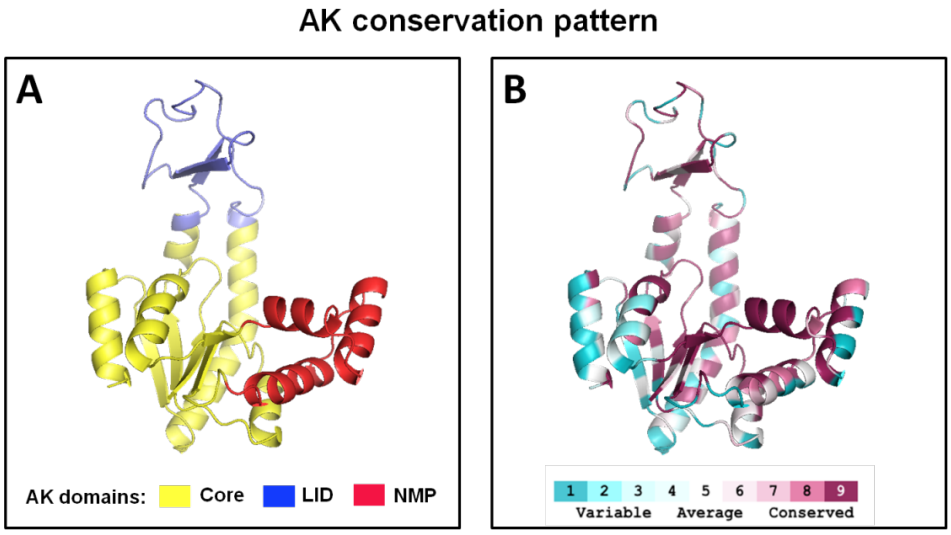

Figure 9: AK conservation pattern based on 114 homologs.

(A) AK is shown in cartoon and colored according to its 3 domains. The core domain which is not continuous in sequence is shown in yellow, the LID domain is shown in blue and the last to fold NMP domain in shown red. The NMP is a 47 amino acids long continuous domain. (B) The conservation pattern of AK according to sequence data from 114 homologs with the same length. Dark pink residues are highly conserved while light blue residues are variable. A more detailed conservation scale is shown on the figure. The enzyme exhibits a typical conservation pattern showing variability in surface residues while high degree of conservation at core residues or future being (at the bound state) core residues. The NMP domain which I aimed to redesign for higher thermostability is demonstrating extremely high conservation in residues that are not solvent exposed.

Figure 10

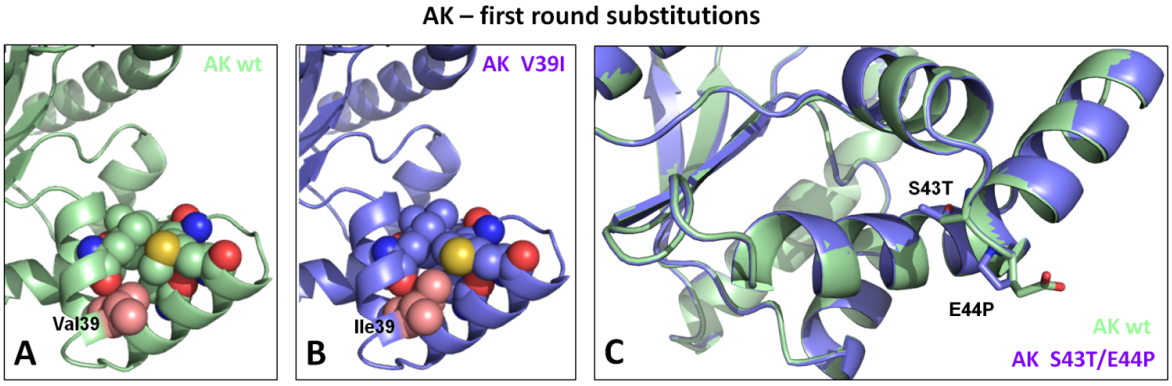

Figure 10: suggested substitutions for a more thermostable AK given high weight PSSM scores.

4 substitutions were selected manually after examining all the computational results under high PSSM weight. 3 of them are shown here (V64I is not shown). AK-wt is shown in green cartoon and AK designs are shown in blue cartoons. V39I provides better core packing of the NMP domain. Val39 (A) and Ile39 (B) are shown in pink spheres. Residues at the proximity of Val39 or Ile39 are shown in green or blue spheres respectively. There is a small cavity right in the core of the NMP domain (A). The extra methyl of Ile39 (B) seals the cavity and provides better packing in that region. (C) A model of S43T and E44P both shown as sticks. Threonine has the best PSSM score at position 43. A Proline at the N-terminus of a helix is a promising substitution especially with regard to folding.

Figure 11

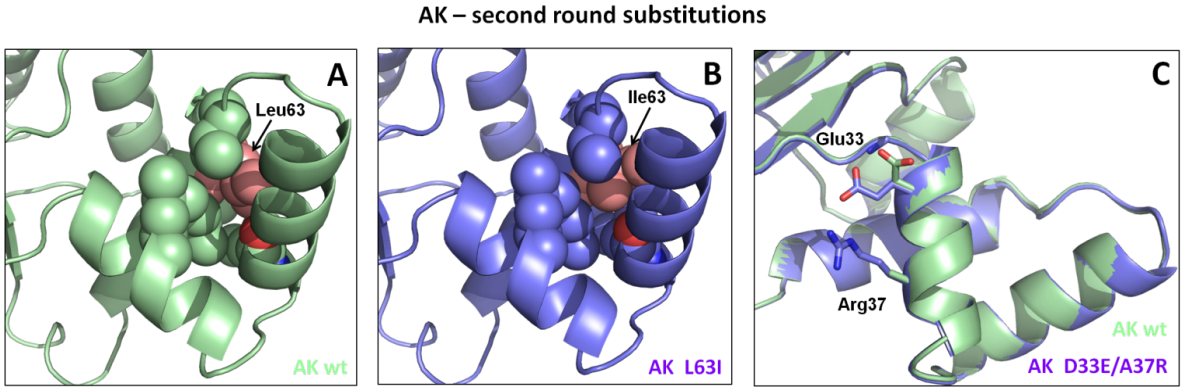

Figure 11: suggested substitutions for a more thermostable AK given low weight PSSM scores.

Two examples of substitutions offered under low weight PSSM scores and more extensive minimizations are shown. AK-wt is shown in green cartoon and AK designs are shown in blue cartoons. L63I provides better packing at the core of the NMP domain, similar to V39I shown in figure 10. Leu63 (A) and Ile63 (B) are shown in pink spheres. Residues at the proximity of Leu63 or Ile63 are shown in green or blue spheres respectively. AK-L63I (B) exhibits tighter packing at the core of the NMP comparing to the native enzyme (A), with the methyl on Ile63-C_β pointing to the hydrophobic core of the NMP domain, while the methyl on C_γ of the original Leu63 is more surface exposed. (C) A pair of substitutions, D33E and A37R enables the formation of a salt bridge between the two, potentially stabilizing the helix.

Figure 12

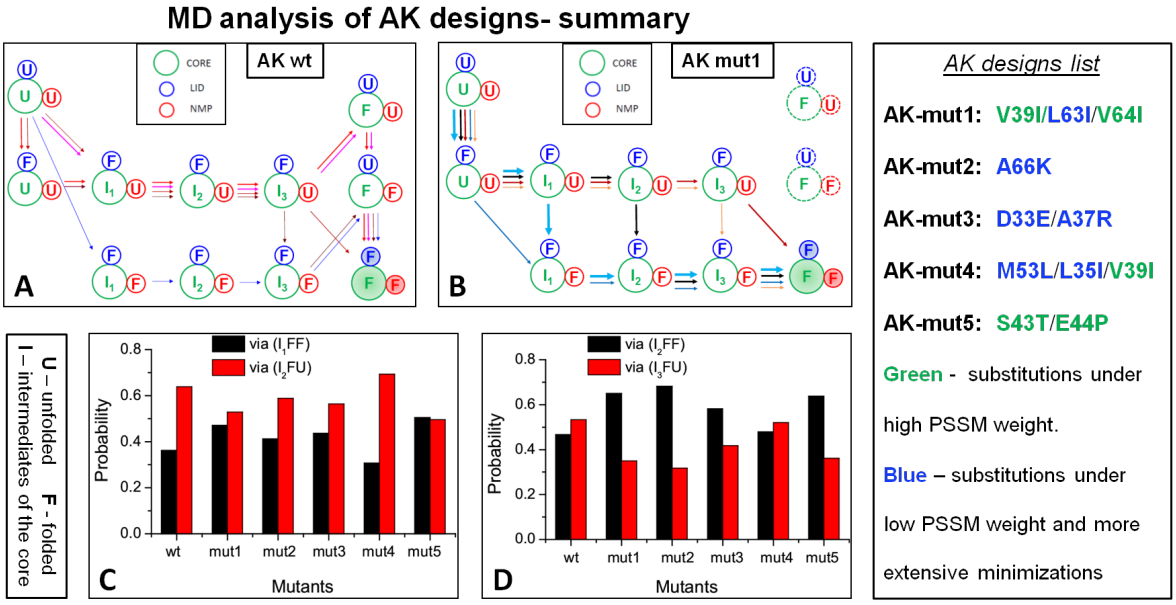

Figure 12: MD folding simulations of AK – summarized results (done at Takada lab in the University of Kyoto

We sent 5 ACP designs containing substitutions from the first and second rounds of computational design to MD analysis at Takada lab in Kyoto University. A detailed designs list is on the right. (A) Folding routes of AK-wt. green circles, blue circles and red circles represent the Core domain, LID domain and NMP domain respectively. The letters inside the circles designate how structured is the domain: U, I_1, I_2, I_3 and F represent the sub-states of the Core domain (unfolded, various intermediates, fully folded). Similarly, U and F within the blue and red circles represent the unfolded and folded sub-states of the LID and NMP domains respectively. The five most probable folding routs that reach the native state are represented by colored arrows with the line widths representing the abundance of each folding route. In most routes the NMP domain is unfolded until late stages of the folding route. (B) Folding routes of AK-mut1. This mutant contains the three most promising conservative mutations that are expected to provide better packing of the NMP domain. For this design, a folding route represented by a light blue arrow is significantly more populated. This route is characterized by the early folding of the NMP domain (progresses via sub-states I_1FF, I_2FF). Other folding routes with relatively early folding of the NMP domain also become more populated with respect to AK-wt. (C) Quantitative analysis of the number of folding trajectories via sub-states I_1FF (black bars) and I_2FU (red bars). I_1FF

sub state (core – partially folded, LID – folded, NMP folded) is populated with molecules that follow trajectories with early folding of the NMP domain, while I_2FU sub state is populated with molecules that follow trajectories showing middle or late folding of the NMP domain. For all designs except AK-mut4, the number of folding trajectories via sub state I_1FF increases with respect to AK-wt. Most prominent are Ak-mut1 and AK-mut5, both mainly contain substitutions from the first design round. (D) Quantitative analysis of the number of folding trajectories via sub-states I_2FF (black bars) and I_3FU (red bars). I_2FF sub state is populated with molecules that follow trajectories showing early or middle folding of the NMP domain, while I_3FU sub state is populated with molecules that follow trajectories with late folding of the NMP domain. Similarly to C, for all designs except AK-mut4, the number of trajectories via I_2FF increases with respect to AK-wt. Most prominent are AK-mut1, AK-mut2, AK-mut5.

Tables

Network characteristic	E35	E52	E58
Apparent k_d $[\mu M^3]$	1	0.1	>5
Buried surface area ($Å^2$)	1230	1420	1230
Hotspot binding foot print	Two β hairpins	α-helix	8 residue loop
Computed hotspot residues	R127, F128, L191	R207, F206	R132, L131
Number of substitutions compared to original scaffold	7	11	9

Table 1: Summary of the first synthetic PPI network characteristics

Design name	Substitutions	Against	Synthesized
dACP1	V27F (L37A), L35E, K61F	E35	Yes
dACP2	V27F (L37A), L35E	E35	Yes
dACP3	V27I, L35E, K61F	E35	Yes
dACP4	V27I, L35E	E35	Yes
dACP5	V27F (L37A), S34Q, K61F	E58	Yes
dACP6	V27F (L37A), S34Q	E58	Yes
dACP7	S34Q, L35D, K61F	E58	Yes
dACP8	S34Q, L35D	E58	Yes
dACP9	V27F (L37A), L35D, K61F	E58	Yes
dACP10	V27F (L37A), L35D	E58	Yes
dACP11	V27F (L37A), S34Q, L35D, K61F	E58	Yes
dACP12	V27F (L37A), S34Q, L35D	E58	Yes
dACP13	F54Y, A57S, K61F	E58	Yes
dACP14	F54Y, A57S	E58	Yes
dACP15	V27F (L37A), K61F	E35/E58	Yes
dACP16	V27F (L37A)	E35/E58	Yes
dACP17	V27F (L37A), F54Y, A57S	E58	Yes
dACP18	V27F (L37A), L35D, T42Y	E58	Yes
dACP19	V27F (L37A), L35E, I41L, K61F	E35	Yes
dACP20	V27F (L37A), L35E, I41L	E35	Yes
dACP21	V27I, L35E, I41L, K61F	E35	Yes
dACP22	V27I, L35E, I41L	E35	Yes

dACP23	V27F (L37A), I41L, K61F	E35	Yes
dACP24	V27F (L37A), I41L	E35	Yes
dACP25	L35E, I41L, K61F	E35	Yes
dACP26	L35E, I41L	E35	Yes
dACP27	V27I, K61F	E35	Yes
dACP28	V27E, K61F	E35	Yes
dACP29	L35E, K61F	E35	Yes
dACP30	I41L, K61F	E35	No
dACP31	V27I, I41L	E35	No
dACP32	V27E, I41L	E35	No
dACP33	V27F (L37A), T42Y	E58	No
dACP34	V27F (L37A), F54Y	E58	No
dACP35	V27F (L37A), A57S	E58	No
dACP36	S34Q, F54Y	E58	No
dACP37	S34Q A57S	E58	Yes
dACP38	S34Q, K61F	E58	Yes
dACP39	L35D, F54Y	E58	No
dACP40	L35D, A57S	E58	Yes
dACP41	L35D, K61F	E58	Yes
dACP42	T42Y, A57S	E58	Yes
dACP43	T42Y, K61F	E58	Yes
dACP44	V27I	E35	Yes
dACP45	V27E	E35	Yes
dACP46	S34Q	E58	Failed
dACP47	L35E	E35	Yes
dACP48	L35D	E58	Yes
dACP49	I41L	E35	No
dACP50	T42Y	E58	Yes
dACP51	F54Y	E58	Failed
dACP52	F54L	E58	No
dACP53	A57S	E58	Yes
dACP54	K61F	E35/E58	Yes

Table 2: a list of all ACP designs cloned into pET29b bacterial expression vector. I planned to synthesize 54 ACP designs containing a single substitution or a combination of substitutions selected from the computational affinity design. The designs are designated as dACP and a following number. The substitutions in each design are listed in the second column. The third column designates against which binder the designs were expected to increase affinity. Out of 54 designs, 42 were successfully cloned. Two designs failed and probably require another set of primers. The remained 10 designs require the failed designs as template for PCR amplification or clones synthesized very late as a template, and were not synthesized due to lack of time.

List of abbreviations

PPI	Protein-protein interaction
Mtb ACP2	M. tuberculosis acyl-carrier protein
AK	Adenylate Kinase protein
NMP	Nucleotide mono phosphate binding domain in AK protein.
sm-FRET	Single molecule Förster resonance efficiency transfer.
YSD	Yeast Surface Display
epPCR	error prone PCR
Mbp	Maltose binding protein
MD	Molecular dynamics
TEV	Tobacco etch virus
ISPC	Israeli structural proteomics center

References

1. Jha RK, Leaver-Fay A, Yin S, et al. Computational design of a PAK1 binding protein. *J Mol Biol*. 2010;400(2):257–270. doi:10.1016/j.jmb.2010.05.006.

2. Liu S, Liu S, Zhu X, et al. Nonnatural protein-protein interaction-pair design by key residues grafting. *Proc Natl Acad Sci U S A*. 2007;104(13):5330–5335.

3. Karanicolas J, Corn JE, Chen I, et al. A de novo protein binding pair by computational design and directed evolution. *Mol Cell*. 2011;42(2):250–260. doi:10.1016/j.molcel.2011.03.010.

4. Fleishman SJ, Khare SD, Koga N, Baker D. Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci*. 2011;20(4):753–757. doi:10.1002/pro.604.

5. Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. *Nat Struct Biol*. 2003;10(1):45–52. doi:10.1038/nsb877.

6. Fleishman SJ, Whitehead T a., Ekiert DC, et al. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science (80-)*. 2011;332(6031):816–821. doi:10.1126/science.1202617.

7. Fleishman SJ, Corn JE, Strauch E-M, Whitehead T a., Karanicolas J, Baker D. Hotspot-Centric De Novo Design of Protein Binders. *J Mol Biol*. 2011;413(5):1047–1062. doi:10.1016/j.jmb.2011.09.001.

8. Strauch E-M, Fleishman SJ, Baker D. Computational design of a pH-sensitive IgG binding protein. *Proc Natl Acad Sci U S A*. 2014;111(2):675–80. doi:10.1073/pnas.1313605111.

9. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science*. 2003;302(5649):1364–8. doi:10.1126/science.1089427.

10. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. *Science*. 2005;308(5723):857–60. doi:10.1126/science.1107387.

11. Chiti F, Dobson CM. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*. 2006;75:333–66. doi:10.1146/annurev.biochem.75.101304.123901.

12. Eaton WA, Mu V, Hagen SJ, et al. F AST K INETICS AND M ECHANISMS IN P ROTEIN F OLDING □. 2000.

13. Dobson CM. Protein folding and misfolding. *Nature*. 2003;426(6968):884–90. doi:10.1038/nature02261.

14. Fersht AR. folding : delights and scepticisms.

15. Han J-H, Batey S, Nickson A a, Teichmann S a, Clarke J. The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol*. 2007;8(4):319–30. doi:10.1038/nrm2144.

16. Hammond S. A Correlation of Reaction Rates. 1953;(3):334–338.

17. Schulz GE. Induced-fit movements in adenylate kinases. *Faraday Discuss*. 1992;(93):85–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1290942.

18. Pirchi M, Ziv G, Riven I, et al. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat Commun*. 2011;2:493. doi:10.1038/ncomms1504.

19. Li W, Terakawa T, Wang W, Takada S. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. *Proc Natl Acad Sci U S A*. 2012;109(44):17789–94. doi:10.1073/pnas.1201807109.

20. Liang J, Edelsbrunner H, Fu P, Sudhakar P V, Subramaniam S. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins*. 1998;33(1):1–17.

21. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD. Isolating and engineering human antibodies using yeast surface display. *Nat Protoc*. 2006;1(2):755–68. doi:10.1038/nprot.2006.94.

22. Xu Q, Krishna SS, Mcmullan D, et al. Crystal Structure of an ORFan Protein (TM1622) from Thermotoga maritima at 1 . 75 Å Resolution Reveals a Fold Similar to the Ran-Binding Protein Mog1p. 2006;782(August):777–782. doi:10.1002/prot.

23. Verdecia M a, Larkin RM, Ferrer J-L, Riek R, Chory J, Noel JP. Structure of the Mg-chelatase cofactor GUN4 reveals a novel hand-shaped fold for porphyrin binding. *PLoS Biol*. 2005;3(5):e151. doi:10.1371/journal.pbio.0030151.

24. Brot N, Collet J-F, Johnson LC, Jönsson TJ, Weissbach H, Lowther WT. The thioredoxin domain of Neisseria gonorrhoeae PilB can use electrons from DsbD to reduce downstream methionine sulfoxide reductases. *J Biol Chem*. 2006;281(43):32668–75. doi:10.1074/jbc.M604971200.

25. Fleishman SJ, Leaver-Fay A, Corn JE, et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One*. 2011;6(6):e20161. doi:10.1371/journal.pone.0020161.

Supplemental data

Contains all the design protocols in the following order:

1. Affinity protocol (for E35 as example)
2. AK protocols
    a. Relaxation of the input structure for the first design round.
    b. Relaxation of the input structure for the second design round.
    c. NMP domain design – round 1
    d. NMP domain design – round 2
3. AK additional files
    a. Resfile
    b. Severs used to create the PSSM files.

Affinity protocol for E35

```
<ROSETTASCRIPTS>

        <SCOREFXNS>
                <SFXN1 weights=score12_w_corrections>
                        <Reweight scoretype=atom_pair_constraint weight=1.0/>
                </SFXN1>
        </SCOREFXNS>

        <TASKOPERATIONS>
                <LayerDesign name=layer layer=core_boundary_surface/>
                <ProteinInterfaceDesign name=pido interface_distance_cutoff=10.0
                  design_chain1=1 design_chain2=0/>
                <PreventRepacking name=fix_42 resnum=42/> Thr on ACP
                <PreventRepacking name=fix_45 resnum=45/> Glu on ACP
                <PreventRepacking name=fix_R132 resnum=132/> Arg on E35
          <InitializeFromCommandline name=init/>
        </TASKOPERATIONS>

        <FILTERS>
                <PackStat name=pstat threshold=0.60 confidence=0/>
                <Ddg name=ddg confidence=0/>
                <Sasa name=sasa confidence=0/>
                Sc name=sc confidence=0/>
        </FILTERS>

        <MOVERS>
                <AtomTree name=tree anchor_res=45A connect_from="CD"
                  pdb_num=132B connect_to="CZ"/> Glu->Arg connection
                <FastRelax name=fastrelax
                  task_operations=pido,fix_42,fix_45,fix_R132,init>
                        <MoveMap>
                                <Jump number=0 setting=0/>
                                <Chain number=1 chi=1 bb=1/>
                                <Chain number=2 chi=1 bb=1/>
                        </MoveMap>
                </FastRelax>
                <PackRotamersMover name=design
                   task_operations=pido,fix_42,fix_45,fix_R132,init/>
                <ParsedProtocol name=aggregate>
                            <Add mover=design/>
                            <Add mover=fastrelax/>
                 </ParsedProtocol>
                 <LoopOver name=iter5 mover_name=aggregate iterations=5/>
        </MOVERS>

        <PROTOCOLS>
                <Add mover=tree/>
                <Add mover=iter5/>
                <Add filter=pstat/>
                <Add filter=ddg/>
                <Add filter=sasa/>
                Add filter=sc/>
```

```
	</PROTOCOLS>
</ROSETTASCRIPTS>
```

AKE relaxation prior to first round design

```
<ROSETTASCRIPTS>
	<SCOREFXNS>
		<score12_with_constraints weights = score12_w_corrections>
			<Reweight scoretype = coordinate_constraint weight = 0.1/>
			<Reweight scoretype = "res_type_constraint" weight = 0.1/>
		</score12_with_constraints>
	</SCOREFXNS>

	<SCOREFXNS> #used only with the filter stability_without_pssm to report
			actual physico-chemical stability without the pssm bias.
		<score12_without_pssm weights = score12_w_corrections>
			<Reweight scoretype = coordinate_constraint weight = 0.1/>
		</score12_without_pssm>
	</SCOREFXNS>

	<TASKOPERATIONS>
		<InitializeFromCommandline name=init/>
		<ReadResfile name = readResfile filename = FILENAME/>
	</TASKOPERATIONS>

	<MOVERS>
		<Prepack name=prepack scorefxn=score12_with_constraints
		  jump_number=0 min_bb=0 task_operations = init/>
		<MinMover name=min scorefxn=score12_with_constraints chi=1
		  bb=1 jump=0/>
		<TaskAwareCsts name=coordinate/>
		<FavorSequenceProfile name = PSSMrestraint scaling = "none"
		  weight = 1 pssm = FILENAME scorefxns =
		   score12_with_constraints />
	</MOVERS>

	<FILTERS>
		<ScoreType name=stability_score_full
		  scorefxn=score12_with_constraints score_type=total_score
		  threshold=0/>
		<ScoreType name = stability_without_pssm scorefxn =
		  score12_without_pssm score_type = total_score threshold = 0/>
		<Rmsd name=rmsd confidence=0/>
	</FILTERS>

	<PROTOCOLS>
			<Add mover_name = PSSMrestraint/>
			<Add mover_name=prepack/>
			<Add mover_name=coordinate/>
			<Add mover_name=min/>
			<Add mover_name=prepack/>
			<Add mover_name=min/>
			<Add mover_name=prepack/>
			<Add mover_name=min/>
```

```
                <Add filter_name = stability_score_full/>
                <Add filter_name = stability_without_pssm/>
                <Add filter_name=rmsd/>
        </PROTOCOLS>

</ROSETTASCRIPTS>
```

AKE relaxation prior to second round design

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
                <score12_with_constraints weights = score12_w_corrections>
                        <Reweight scoretype = coordinate_constraint weight = 0.1/>
                        <Reweight scoretype = "res_type_constraint" weight = 0.1/>
                </score12_with_constraints>
        </SCOREFXNS>

        <TASKOPERATIONS>
                <InitializeFromCommandline name=init/>
                <ReadResfile name = readResfile filename = FILENAME/>
        </TASKOPERATIONS>

        <MOVERS>
                <Prepack name=prepack scorefxn=score12_with_constraints
                jump_number=0 min_bb=0 task_operations = init/>
                <FastRelax name = fastrelax scorefxn = score12_with_constraints
               design = 0 task_operations = init,readResfile/> #prepacks,
         minimizes and relax.
                <TaskAwareCsts name=coordinate/>
                <FavorSequenceProfile name = PSSMrestraint scaling = "none"
                weight = 1 pssm = FILENAME  scorefxns =
                score12_with_constraints />
        </MOVERS>

        <FILTERS>
                <ScoreType name=stability_score_full
                  scorefxn=score12_with_constraints score_type=total_score
                  confidence = 0 threshold=0/>
                <ScoreType name = stability_without_pssm scorefxn =
                  score12_without_pssm score_type = total_score confidence = 0
                  threshold = 0/>
                <Rmsd name=rmsd confidence=0/>
        </FILTERS>

        <PROTOCOLS>
                        <Add mover_name = PSSMrestraint/>
                        <Add mover_name=coordinate/>
                        <Add mover_name=fastrelax/>
                        <Add mover_name=fastrelax/>
                        <Add filter_name = stability_score_full/>
                        <Add filter_name = stability_without_pssm/>
                        <Add filter_name=rmsd/>
        </PROTOCOLS>
```

```
</ROSETTASCRIPTS>
```

First round design protocol

```
<ROSETTASCRIPTS>
        <SCOREFXNS
                <score12_with_constraints weights = score12_w_corrections>
                        <Reweight scoretype = coordinate_constraint weight = 0.1/>
                        <Reweight scoretype = "res_type_constraint" weight = 0.25/>
                </score12_with_constraints>
        </SCOREFXNS>

        <SCOREFXNS> #used only with the filter stability_without_pssm to report actual
physico-chemical stability without the pssm bias.
                <score12_without_pssm weights = score12_w_corrections>
                        <Reweight scoretype = coordinate_constraint weight = 0.1/>
                        </score12_without_pssm>
        </SCOREFXNS>

        <TASKOPERATIONS>
                <InitializeFromCommandline name = init/>
                <ReadResfile name = readResfile filename = FILENAME/>
                <SeqprofConsensus name = pssm_cutoff filename = FILENAME
                  min_aa_probability = 0 probability_larger_than_current = 1
                  convert_scores_to_probabilities = 0/>
                <RestrictResidueToRepacking name = fix_73 resnum = 73/>  Cys.
                  attachment site for FRET dye
                <RestrictResidueToRepacking name = fix_203 resnum = 203/> Cys.
                  attachment site for FRET dye
        </TASKOPERATIONS>

        <MOVERS>
                <PackRotamersMover name = design scorefxn =
                  score12_with_constraints task_operations =
                  init,readResfile,pssm_cutoff,fix_73,fix_203/>
                <MinMover name = min scorefxn = score12_with_constraints chi = 1
                  bb = 1 jump = 0/>
                <TaskAwareCsts name = coordinate/>
                <FavorSequenceProfile name = PSSMrestraint scaling = "none"
                 weight = 1 pssm = FILENAME  scorefxns =
                 score12_with_constraints /> #here the weight is 1 and in
                                                        SCORFXNS is 0.25
        </MOVERS>

        <FILTERS>
                <ScoreType name = stability_score_full scorefxn =
                  score12_with_constraints score_type = total_score confidence = 0
                  threshold = 0/>
                <ScoreType name = stability_without_pssm scorefxn =
                score12_without_pssm score_type = total_score threshold = 0/>
                <Rmsd name = rmsd confidence = 0/>
```

```
            <DesignableResidues name = designableResidues task_operations =
              readResfile,pssm_cutoff,fix_73,fix_203/>
        </FILTERS>

        <PROTOCOLS>
                        <Add mover_name = PSSMrestraint/>
                        <Add mover_name = design/>
                        <Add mover_name = coordinate/>
                        <Add mover_name = min/>
                        <Add filter_name = stability_score_full/>
                        <Add filter_name = stability_without_pssm/>
                        <Add filter_name = rmsd/>
        </PROTOCOLS>

</ROSETTASCRIPTS>
```

Second round design protocol

```
<ROSETTASCRIPTS>
        <SCOREFXNS>
                <score12_with_constraints weights = score12_w_corrections>
                        <Reweight scoretype = coordinate_constraint weight = 0.1/>
                        <Reweight scoretype = "res_type_constraint" weight = 0.1/>
                </score12_with_constraints>
        </SCOREFXNS>

        <TASKOPERATIONS>
                <InitializeFromCommandline name = init/>
                <ReadResfile name = readResfile filename = FILENAME/>
                <SeqprofConsensus name = pssm_cutoff filename = FILENAME
                  min_aa_probability = 0 probability_larger_than_current = 0
                  convert_scores_to_probabilities = 0/>
                <RestrictResidueToRepacking name = fix_73 resnum = 73
                <RestrictResidueToRepacking name = fix_203 resnum = 203/>
        </TASKOPERATIONS>

        <MOVERS>
                <PackRotamersMover name = design scorefxn =
                 score12_with_constraints task_operations = init, readResfile,
                 pssm_cutoff,fix_73,fix_203/>
                <FastRelax name = fastrelax scorefxn = score12_with_constraints design =
0 task_operations = init,readResfile,pssm_cutoff,fix_73,fix_203/>
                <TaskAwareCsts name = coordinate/>
                <FavorSequenceProfile name = PSSMrestraint scaling = "none"
                  weight = 1 pssm = FILENAME  scorefxns =
                  score12_with_constraints />
                <ParsedProtocol name = DesignandFastRelax>
                        <Add mover = design/>
                        <Add mover = coordinate/>
                        <Add mover = fastrelax/>
                </ParsedProtocol>
                <LoopOver name = iter2 mover_name = DesignandFastRelax
                  iterations = 2/>
```

```
        </MOVERS>

        <FILTERS>
                <ScoreType name = stability_score_full scorefxn =
                  score12_with_constraints score_type = total_score confidence = 0
                  threshold = 0/>
                <ScoreType name = stability_without_pssm scorefxn =
                  score12_without_pssm score_type = total_score confidence = 0
                  threshold = 0/>
                <Rmsd name = rmsd confidence = 0/>
                <DesignableResidues name = designableResidues task_operations =
                readResfile,pssm_cutoff,fix_73,fix_203/>
        </FILTERS>

        <PROTOCOLS>
                        <Add mover_name = PSSMrestraint/>
                        <Add mover_name = iter2/>
                        <Add filter_name = stability_score_full/>
                        <Add filter_name = stability_without_pssm/>
                        <Add filter_name = rmsd/>
        </PROTOCOLS>

</ROSETTASCRIPTS>
```

Resfile (required to restrict the design to the NMP positions only)

```
NATAA
start
#residues to design in the NMP domain
30      A   ALLAA
31      A   ALLAA
32      A   ALLAA
33      A   ALLAA
34      A   ALLAA
35      A   ALLAA
36      A   ALLAA
37      A   ALLAA
38      A   ALLAA
39      A   ALLAA
40      A   ALLAA
41      A   ALLAA
42      A   ALLAA
43      A   ALLAA
44      A   ALLAA
45      A   ALLAA
46      A   ALLAA
47      A   ALLAA
48      A   ALLAA
49      A   ALLAA
50      A   ALLAA
51      A   ALLAA
52      A   ALLAA
53      A   ALLAA
```

54 A ALLAA
55 A ALLAA
56 A ALLAA
57 A ALLAA
58 A ALLAA
59 A ALLAA
60 A ALLAA
61 A ALLAA
62 A ALLAA
63 A ALLAA
64 A ALLAA
65 A ALLAA
66 A ALLAA
67 A ALLAA
68 A ALLAA
69 A ALLAA
70 A ALLAA
71 A ALLAA
72 A ALLAA
73 A ALLAA
74 A ALLAA
75 A ALLAA
76 A ALLAA

PSSM file

We identified all AK homologs using the following server: *http://consurf.tau.ac.il* with 4AKE pdb file as input structure. We then used the following website *http://fasta.bioch.virginia.edu/fasta_www2/chaps.cgi* to convert the sequence data into a PSSM matrix. The PSSM file was incorporated in the design protocols (see protocol above)