Biosensor Guided Polyketide Synthase Engineering for Optimization of Domain Exchange Boundaries

Elias Englund
KTH - Royal Institute of Technology
Matthias Schmidt
Joint BioEnergy Institute
Alberto Nava
Joint BioEnergy Institute https://orcid.org/0000-0002-0172-4145
Qingyun Dan
Joint BioEnergy Institute
Leonard Katz
Joint BioEnergy Institute
Satoshi Yuzawa
Keio University
Jay Keasling (jdkeasling@lbl.gov)
Joint BioEnergy Institute

Article

Keywords:

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1528836/v1

Version of Record: A version of this preprint was published at Nature Communications on August 12th, 2023. See the published version at https://doi.org/10.1038/s41467-023-40464-x.

Abstract

Type I modular polyketide synthases (PKSs) are multi-domain enzymes functioning like assembly lines. Many engineering attempts have been made for the last three decades to replace, delete and insert new functional domains into PKSs to produce novel molecules. However, the resulting PKS hybrids typically have reduced catalytic activities and are often insoluble due to misfolding. Here, we have developed a fluorescence-based biosensor method for detecting engineered PKSs with high solubility. The biosensor has been used to sort through PKS hybrids that had acyltransferase (AT) domains from other PKSs exchanged for the native AT with randomly assigned linker junctions. Importantly, we observed a significant correlation between activity and solubility. Evaluation of highly soluble mutants *in vitro* revealed new boundaries for AT domain exchanges that give a wild-type level of catalytic activity. Together, we have successfully developed an experimentally validated high-throughput method to efficiently screen active engineered PKSs that produce target molecules.

Introduction

Type I modular polyketide synthases (PKSs) are large, assembly line-like enzymes that are capable of producing structurally complex natural products, including many important pharmaceuticals with antibiotic, antiparasitic, and immunosuppressive properties[1]. A modular PKS is comprised of several modules with each carrying out a single polyketide chain-extension and modification. Each module contains several catalytic domains. A ketosynthase (KS) catalyzes a decarboxylative Claisen condensation between a growing polyketide chain and an extension unit loaded by an acyltransferase (AT) onto an acyl carrier protein (ACP). In addition to the three core domains, a ketoreductase (KR) domain, a dehydratase (DH) domain, and/or an enoyl reductase (ER) domain may be included in a module to reduce a newly generated b-keto group after chain elongation to a hydroxyl group, a carbon-carbon (C-C) double bond, or a saturated C-C bond. The full-length polyketide is lastly offloaded from the assembly line by a thioesterase (TE) domain.

One of the most important goals in the field of PKS research is highly accurate rational protein engineering for generating novel molecules since the first gene cluster was discovered and the reaction mechanisms were proposed in the early 1990s[2,3]. The modular organization of the enzyme enables swapping in and out domains and modules to generate hybrid PKSs. Early studies demonstrated that combinatorial engineering of PKSs can generate new predicted polyketide structures, though with significantly reduced activities in most cases[4,5]. Since then, more precise protein engineering strategies have become the main thrust of the research to improve the activities of the hybrid PKS[6]. One of the most frequent targets for domain engineering has been the AT domain due to it being responsible for substrate incorporation in polyketide biosynthesis[7]. Precise AT engineering enables insertions of a large diversity of functional groups including linear and branched chains[8,9], terminal alkynes[10-12], and halogenated alkyl groups[13] into polyketide structures, and production of target molecules with not only a high success rate but also an increased titer.

Although there are many examples of AT domain swapping to date, the resulting chimeric PKSs often show significantly reduced catalytic activities compared to their wild-type counterparts[6]. One of the major barriers to generating functional and stable hybrid PKSs is a poor understanding of where in the interdomain linkers one domain exactly ends and another starts. Very recently, the first crystal structure of an entire PKS module containing a KS, an AT, a KR, and an ACP (module 7 of the lasalocid PKS) has been reported[14]. Using this structural information as well as model structures of the lasalocid PKS and the erythromycin PKS proposed by cryo-electron microscopy analysis[14], it could be possible to extract domain boundary information. However, since the first structural analysis of a PKS module by cryo-electron microscopy that has the same domain composition (module 5 of the pikromycin PKS) has challenged the before-mentioned structures in AT domain positioning[15,16], it is still unclear if one could use this structural information to unambiguously predict domain boundaries.

Here we describe a method for efficiently sorting through a PKS library engineered with randomized swap junctions using a solubility biosensor in *Escherichia coli*. To date, only three examples of randomized mutagenesis libraries of engineered PKSs have been described. Two of them used DNA shuffling and yeast homologous recombination to generate libraries of the pikromycin and the erythromycin PKS hybrids[17,18]. The third example serendipitously generated rapamycin analogs by recombination using plasmids with high sequence similarities in a *Streptomyces* strain[19]. Yet, these experiments still require screening of metabolite production of thousands of mutants to find a handful of active proteins[18]. Using a solubility biosensor, we screened through a library of AT-exchanged variants to find those with high solubility and thus with a high probability to be active. To our knowledge, this is the first high throughput screen based on protein stability that has been developed for PKSs.

Results

Development of solubility biosensor for PKS. In our previous AT domain swapping effort, we observed a correlation between solubilities and *in vitro* activities of hybrid PKSs[6]. We reasoned that engineered PKSs that maintain a stable conformation, thereby avoiding aggregation, have a higher probability of exhibit expected activities. To test that hypothesis, we sought to develop *E. coli* biosensor strains that could detect protein misfolding.

Several methods have already been developed for assaying *in vivo* protein stability with fluorometric outputs[20–23]. However, these methods have not been tested with large, multidomain proteins such as PKSs. Heat-shock genes *ibpA* and *fxs* are highly expressed when misfolded proteins accumulate inside *E. coli*[24]. The promoters of those genes (P_{ibp}[20,25] and P_{fxs}[26]) were used to drive expression of the green fluorescent protein (GFP) gene (P_{ibp} alone or P_{ibp} and P_{fxs} in tandem) and integrated into the genome of *E. coli* BL21 (DE3) in the *arsB* gene, thereby creating Δ*arsB*::P_{ibp} GFP and Δ*arsB*::P_{ibpfxs} GFP. The *arsB* site encodes an arsenic membrane pump, which should be a neutral site under standard laboratory conditions. In parallel, we made Δ*ibpA*::GFP by integrating the GFP gene in frame of the *ibpA* gene, consequently knocking out the native gene and appropriating the promoter. To assess how these

biosensors react to PKSs with different levels of solubility, we used the sixth module of the erythromycin PKS (6-DeoxyErythronolide B Synthase) with the neighboring TE domain (DEBSM6) and two engineered versions, D0 and D1, that contained an AT from the epothilone PKS module 4 (EposM4) in place of the native AT with different swap junctions; DEBSM6 was previously shown to be more soluble than D1, which was more soluble than D0[6].

The three PKS variants together with an empty vector control were expressed on pET plasmids in the *E. coli* strains, and fluorescence was measured (Fig. 1a). The *ibpA* promoter integrated in the *arsB* locus was not induced by the highly soluble DEBSM6 while it was similarly induced by D0 and D1. Meanwhile, the tandem promoter P_{ibpfxs} was more sensitive and became induced by DEBSM6 but to a lower degree than the other two PKSs. Integration of GFP into the *ibpA* locus gave high background GFP fluorescence with the empty vector, but it was lower than that of DEBSM6, which was lower than both DO and D1.

Next, we investigated the use of fluorescent fusion tags to be able to normalize biosensor activation to the amount of heterologous protein produced. Previous results from non-PKS proteins showed that the folding of a fluorescent fusion protein is affected by the folded state of the protein to which it is attached[21]. To see the response of fusion tags with multi-domain proteins like PKSs, mCherry was attached to the C-terminus of the three reference PKSs, thereby creating DEBSM6 mCherry, D0 mCherry, and D1 mCherry. Although DEBSM6 mCherry had higher fluorescence than the other proteins, measuring cellular abundance of the expressed PKSs by SDS-PAGE quantification revealed it was due to higher amounts of protein and not due to differences in solubility (Fig. 1b). To investigate why the less soluble D0 has proportionally the same fluorescence compared to DEBSM6, we separated soluble and insoluble fractions and measured the amounts of expressed PKSs using SDS-PAGE and fluorescence (Fig. 1c, d). We observed that mCherry fused to DEBSM6 did not change the solubility, which was still more than 95% soluble while D0 mCherry had more than half of its protein content in the insoluble fraction. D1 mCherry was 75% soluble. The fluorescence in the different fractions mirrored the protein content indicating that the chromophore is still fluorescing even when attached to an insoluble protein, showing that PKSs offer a unique set of challenges unlike smaller proteins[22,27].

To test biosensor activation across varying levels of expressed proteins, we combined the mCherry tagged PKSs with the *E. coli* strain that harbors Δ*arsB*::P_{ibp} GFP biosensor and induced protein expression by adding different concentrations of IPTG. Even at high IPTG concentrations, DEBSM6 mCherry only weakly induced the biosensor while D0 mCherry had high induction even at 50 μM (Fig. 1e). D0 mCherry showed a higher GFP fluorescence than D1 mCherry, indicating that the biosensor can discriminate the two proteins when appropriate IPTG concentrations are used. To simplify presentation of the solubility data, we define the solubility coefficient as mCherry fluorescence divided by GFP fluorescence, which is a measure of the PKS solubility and how well it expressed (Fig. 1f and Supplementary Fig. 1–4).

Constructing an AT swapped PKS library. We created a library of engineered DEBSM6 with EposM4 AT variants carrying randomly assigned swap junctions in the KS-AT and the post-AT linker regions. The

EposM4 AT natively uses either malonyl-CoA or methylmalonyl-CoA, unlike the native DEBSM6 AT which uses only methylmalonyl-CoA. To get a structural understanding of the linker regions around the DEBSM6 AT, we used the newly released AlphaFold[28] to generate a homodimeric structure model without the TE domain (Fig. 2a). The predicted DEBSM6 structure showed a high degree of similarity with the experimentally solved structures of KS-AT didomains from DEBSM3 and DEBSM5 (56% and 57% sequence identity with DEBSM6 KS-AT)[29,30] as well as the PKS module structures recently reported[14,31]. The predicted KS-AT linker consists of a disordered region starting from the KS which then forms three alpha helices surrounding three beta strands (Fig. 2b, d). The post-AT linker on the other hand wraps itself around the residues of the KS-AT linker and continues along the KS domain, interacting with several residues until it reaches the structural subdomain ψKR[32] (Fig. 2c, e).

Next, we developed an *in vitro* method for creating a randomized mutagenesis library where each AT-swapped DEBSM6 variant had one random upstream and downstream junction. This was done using oligo pools, up to 350 base pair long oligonucleotides designed to each carry the swap junction at a different amino acid position (Fig. 3a). In total, 72 unique swap junction oligos (duplicate oligos in homology regions were removed) were designed for the KS-AT linker and 73 oligos were created for the post-AT linker, thereby making 5,256 possible combinations when randomly combining an upstream and downstream junction.

The oligo pools were used to amplify the EposM4 AT sequence using PCR and the resulting fragment mix was cloned into the AT position of DEBSM6 mCherry. Many of the resulting colonies of the library carried small deletions in the linker regions, possibly due to synthesis errors caused by the length of the oligos. Roughly 40% of colonies were visibly red indicating the presence and in-frame expression of mCherry.

Biosensor-guided screening of soluble PKSs. Around 800 colonies with a visible red color were induced for protein expression and fluorescence was measured using flow cytometry (Fig. 3b). Forty colonies with high solubility were selected, and the corresponding plasmids were purified and sequenced to determine what positions were enriched in the high solubility set. The highly soluble PKSs had swap junctions in amino acids us10-45 and us82-86 in the KS-AT linker and in ds18-65 in the post-AT linker (Fig. 3c, Supplementary Table 1). The gap between us46-81 was predicted by AlphaFold to contain the alpha helix 2 (α2) and the beta strand 2 (β2), both of which are deeply embedded within the KS-AT linker structure (Fig. 2b). AlphaFold also predicted that the junctions in amino acids ds1-18 in the post-AT linker would be inside the AT structure (Fig. 3d), and those in ds 65–90 would be inside the ψKR domain, both of which would cause the PKS to be insoluble (Fig. 2c).

The 40 library colonies that showed high solubility were remeasured in triplicates and a subgroup of five were selected to assess protein solubility by SDS-PAGE and enzyme activity: o4 (us85/ds21), o8 (us13/ds43), o15 (us21/ds64), o17 (us28/ds62) and o33 (us17/ds25) (Fig. 4a). The most soluble of these, o33, had close to the wild type DEBSM6 solubility. To investigate the solubility of these proteins in the absence of mCherry, the C-terminal mCherry was removed from each variant and their protein amounts in the soluble and insoluble fractions were quantified by SDS-PAGE (Fig. 4b). Insoluble fractions

of the five selected variants (o4 10.4%, o8 8.6%, o15 13.5%, o17 13.1%, o33 13.7%) were all lower than the references (D0 63.6%, D1 19.4%) but still higher than the wild type DEBS M6 (7.5%). These data confirm that the selected variants have improved solubility.

In vitro **activity assay for highly soluble AT-swapped PKSs.** The five highly soluble variants, DEBSM6 and D1 (us1/ds41) were purified using nickel affinity resins (Supplementary Fig. 5a). *In vitro* enzymatic activity was measured using a synthetic starter substrate and either malonyl-CoA or methylmalonyl-CoA as extension substrates, resulting in a desmethyl or methyl triketide lactone (TKL), respectively. Three variants, o4, o8 and o33, and D1 produced methyl TKL in a turnover rate similar to the wild-type DEBSM6 indicating that protein structures of o4, o8 and o33 are not destabilized even with a heterologous AT domain (Fig. 4c). The positions in the KS-AT linker of the domain-swapped mutants were at the beginning of the linker (β0-α0) for o8 (us13) and o33 (us17), or in the middle of α3 for o4 (us85). These junctions either included the entire KS-AT linker from EposM4 or retained the counterpart of the parental PKS. The downstream junctions in the post-AT linker had the swap boundaries at the end of the AT domain for o4 and o33 (ds21 or ds25) or just before the residues interacting with the KS domain for o8 (ds43). Both o15 and o17 had high solubility but showed significantly lower activities. These non-functional variants had the downstream junction at ds62 and ds64, respectively, meaning KS interacting residues in the post-AT linker (ds44-56 in AlphaFold prediction) had the heterologous EposM4 AT sequence. This part of the linker is known to tightly interact with the KS in DEBS[29,30] and is critical for KS condensation reaction[34], indicating that the heterologous linker sequence in the o15/o17 variants is unable to complement that function and the quaternary structure is likely not retained[35].

When malonyl-CoA was used as a substrate, no production was observed in DEBSM6 as expected since the native AT cannot accept malonyl-CoA[6] (Fig. 4d). In contrast, o4, o8 and o33, and D1 showed product formation with malonyl-CoA albeit at a lower amount compared with methyl TKL production. This is consistent with previous results and thought to be due to the preference of DEBSM6 KS[6].

Investigating swap positions in KS-AT linker. Next, we selected eight swap junctions in the KS-AT linker and three in the post-AT linker and constructed all 24 combinations. Upon measuring fluorescence, results show that positions us85, us28 and us17 always led to the highest relative solubility, no matter if the downstream junction was in a position that contributed to high (ds25), medium (ds62) or low solubility (ds81) (Fig. 5a). These results indicate that swap junctions influence protein stability independently of each other. Therefore, it would be possible to assess an upstream junction's general influence on stability, independent of the paired downstream junction.

To do so, we selected the post-AT ds25 junction from the most soluble o33 variant and combined it with each possible upstream position using EposM4 AT. In total, 72 constructs were made, and solubility was assessed using the solubility coefficient. The measurements show that the α2 and β2 structure regions are not appropriate for recombination (Fig. 5b, c), which is consistent with our randomized library data (Fig. 3c). Surprisingly, most other variants showed relatively high solubilities; a notable exception was the

variant with one proline residue in β1 (us27), which destabilized the protein, while its neighboring positions did not.

To further investigate the solubility-activity relationship, four variant pairs were selected with nearby swap junctions and but with differences in solubility: us1/ds25, us3/ds25, us26/ds25, us27/ds25, us48/ds25, us52/ds25, us79/ds25, and us84/ds25. These PKSs were purified without the mCherry fusion tag and all proteins were isolated with high purity except us52/ds25 (Supplementary Fig. 5b). *In vitro* activity analysis of desmethyl and the methyl TKL production showed that the less soluble mutants were also less active in all examples but one in which the activity was equal (Fig. 6e, f). Because all tested variants have the same post-AT linker junction (ds25), the cause of the structural destabilization and low activities should be attributed to an unfavorable junction in the KS-AT linker. The ACP probably docks with the KS-AT linker surface while interacting with the KS for chain elongation[36], disruption of which may explain the observed reduction of activities.

To see how generalizable these results are, we repeated the experiment with a phylogenetically distant AT, the ethylmalonyl-CoA–specific AT from the tiacumicin PKS module4 (TiasM4)[37]. In this case, 69 unique constructs were made to cover the same linker region. Results from the fluorescence measurement showed that the solubility patterns from TiasM4 AT-swapped PKSs agreed well with the corresponding EposM4 AT results (Fig. 6b, d). Again, upstream junctions within α2 and β2 appear to destabilize the protein when ds25 is used as a downstream junction.

Discussion

Type I modular PKSs are multi-domain enzymes that produce various polyketide structures by combining several catalytic domains in a specific order in an understandable way. Recent studies suggest that tens of thousands of type I modular PKSs and hybrids with nonribosomal peptide synthetase genes are encoded in genomic sequences[38], indicating that nature utilizes these enzymes to produce diverse molecules on demand. Since the early 1990s, significant efforts have been made to rationally engineer type I modular PKSs by domain swapping. Although hundreds of natural polyketide analogs have been generated by rational protein engineering, unfortunately the resulting engineered PKSs generally have significantly reduced kinetics compared with their wild-type counterparts, producing low titers of desired molecules[4,5]. One significant barrier that slows PKS engineering progress is the lack of a general method to screen active versus inactive PKS variants in an engineered PKS library. Polyketide products from type I modular PKSs usually do not absorb light or fluoresce. Although a recently developed polyketide biosensor holds promise for high-throughput screening[39,40], it is not practical to design a biosensor for every target molecule.

In the present study, we report an experimentally validated biosensor that discriminate soluble and insoluble PKS variants. In the first screening from ~ 800 AT-swapped PKSs where swap junctions in the KS-AT linker and the post AT liner were randomized, we selected 40 variants that showed high solubilities (Fig. 4a). *In vitro* activity assays demonstrated that 60% of the highly soluble AT-swapped PKSs showed a

wild-type level of catalytic activity. As expected, these variants also produced a polyketide product from a non-native substrate. The most soluble variant from the initial screening was o33 where the junctions in the KS-AT linker and the post AT linker were us17 and ds25, respectively. The us17 junction in the KS-AT linker is located between β0 to α0, an unstructured region in the linker. We used AlphaFold to compare 15 additional PKSs and this region consists of a random coil in all models except for the bafilomycin PKS module 5 where there is an insertion of ~ 25 amino acid residues (Supplementary Table 2, Supplementary Fig. 6). These structural models also show that the triple alpha helix motif (α1-α3) in the KS-AT linker is conserved except for the aculeximycin PKS M7 and the rifamycin PKS M1. Both structures lack α3 between β2 and β3, which can be seen as a gap in the sequence alignment. The triple beta sheet motif is also conserved in all analyzed PKSs except for the borrelidin PKS M1 where β3 was missing. For the post-AT linker, all predicted modules exit the AT domain with α4, after which the linkers wrap around KS-AT linkers and continues along with KS domains where LPTY(A/P)FQ(H/R)xRYWL motif binds to the KS surface after which the models diverge depending on what domain that follows. These models explain why ds21-25 (o4 and o33) and ds41-43 (D1 and o8) variants retain activities, but ds62-64 variants (o15 and o17) lose activities. In the second screening using ds25 as a downstream junction, we observed that most variants in the KS-AT linker are highly soluble (solubility coefficient > 20) except for the ones that have the upstream junctions between α2 and β2 structure regions. Furthermore, *in vitro* assays demonstrated a clear correlation between solubility and activity.

In summary, we created a method to easily and rapidly assess solubility of hybrid PKSs, which correlates well with enzyme activity and increases the throughput of screening. Based on *in vivo* solubility data and *in vitro* activity data, as well as structure data predicted by AlphaFold, we revise our previously suggested boundaries for AT domain swapping[6]. The above observations suggest three types of AT domain swapping junctions, which are us1-17/ds41-43, us1-48/ds21-25, and us84-102/ds21-25, respectively. To locate these sites, the respective parental and donor PKS can be aligned with the primary sequence of DEBSM6. Secondary structure features of experimentally verified and AlphaFold models of KS-AT domains aligned well (Supplementary Fig. 7). This indicates that structure prediction or experimental data is not necessarily required to locate these secondary structural elements, but a sequence alignment should, in most cases, suffice. Our work may suggest that this system could also be used to select active domain-swapped PKSs with other PKS domains such as KS, KR, DH, and/or ER domains to fully optimize each PKS domain swap junctions.

Methods

Cloning and cell cultivation

Each plasmid was constructed using some combination of restriction enzyme digestion and Gibson assembly[41]. Positive colonies were confirmed using Sanger sequencing. All cell cultivation was done in lysogeny broth (LB, Miller) supplemented with 50 ug/ml kanamycin sulfate (Teknova), and plate selection was done on LB + kanamycin agar plates (Teknova).

Protein fractionation and SDS-PAGE quantification

Plasmids were transformed into BAP1[42] and selected on LB plates. Colonies were picked in triplicates and grown at 37°C. The next day 1% of cultures were inoculated into fresh LB and grown at 37°C for 2 hours until OD_{600} ~0.5. Cultures were put on ice for 30 min to stop growth, induced with 250 µM IPTG (Teknova), then grown overnight at 18°C. The following day, OD_{600} was measured and the same amount of cells was taken for each culture equal to around 1.5 ml of overnight culture. Cells were centrifuged and the pellet was resuspended in 500 ul phosphate-buffered saline (PBS). Samples were sonicated with a Q125 sonicator (Qsonica) at 30% amplitude for 5 seconds twice to lyse the cells and a small sample was saved as the “total protein fraction”. Samples were centrifuged at max speed (~21000 g) for 2 min and the supernatant was saved as the “soluble fraction”. The pellet was resuspended in an equal volume PBS as was removed and saved as the “insoluble fraction”. If proteins were fused with mCherry, fluorescence was determined by measuring at 587 excitation/617 emission with a 610 excitation cut off on a SpectraMax M2e (Molecular Devices). For SDS-PAGE, fractionated samples were mixed 1:1 with Laemmli Sample Buffer (Bio-Rad) supplemented with 100 mM DTT (VWR Scientific) and boiled at 95C. Samples were loaded onto Mini-Protean TGX 8-16% 12-well precast gel (Bio-Rad) and run for 30 minutes at 200 volts. Gels were washed twice in boiling water and then stained with GelCode Blue Safe Protein Stain (Thermo Scientific). Protein bands were quantified using the software VisionWorks (Analytik Jena) by first measuring the relationship between pixel intensity and loaded protein using a standard curve created using a series of 2-fold diluted samples. Relative protein content in each fraction was then calculated by diving by the total protein fraction for each individual replicate.

Integration of biosensor into genome

The biosensor strains were engineered using the λ red recombinase protocol described in Babe et al. 2016[43]. For the D*arsB*::P_{ibp} GFP and DarsB::P_{ibpfxs} GFP strain, the promoter(s), GFP, a kanamycin cassette and an upstream and downstream homology region of 1000 base pairs were amplified using PCR and combined into a single fragment with Gibson assembly. For the D*ibpA*::GFP strain, the same method was used except that the promoter sequence was already located on the upstream homology region (check Supplementary Table 3 for sequences). The combined DNA fragments were used to transform chemically BL21 (DE3) cells carrying pKD46 plasmid, induced with 0.1 % arabinose to express recombines genes. After positive colonies were verified by colony PCR screening, the kanamycin cassette was removed using FLP recombinase on the pCP20 plasmid.

Induction of biosensor + measurements

Purified plasmids were transformed into D*arsB*::P_{ibp} GFP and plated on LB agar plates. For the oligo pool library, the casted agar with colonies was carefully lifted and IPTG was added to the bottom of the plates to a concentration of 50 µM assuming a content of 20 ml LB agar. The plates were incubated at room temperature for a day until colonies turned visibly red. Either red colonies for the oligo pool library or uninduced colonies for the other experiments were grown in 300 ul LB in 96-well deep plates as seed cultures. The next day, 1% was seeded into 300 ul LB and grown at 37°C for 2 hours, put on ice for 30 min, induced with 250 µM IPTG (if not otherwise noted) and grown at 18°C, 250 rpm overnight. If several 96-well plates were grown at the same time, cultures were started with a 1 hour delay so that they could be measured after an equally long induction. The following day after ~20 hours of induction, samples were measured using flow cytometry (BD Accuri C6). The flow cytometer collected 50000 events larger than 2000 FSC-H per sample and GFP and mCherry fluorescence was measured. At 250 µM IPTG induction, mCherry fluorescence of cells showed a bimodal distribution with more than half expressing mCherry and the rest not (Supplementary Fig. 8). Gating was used to calculate the average mCherry and GFP fluorescence of the mCherry containing cell subpopulation. The solubility coefficient was calculated as mCherry divided by GFP fluorescence.

Design of oligo pools and construction of swap junction library

The oligo pool nucleotide sequences used to create the swap junction library can be divided into three parts: a shared 5’ region that was used to enrich for full length sequences and had an overlap with the DEBSM6 plasmid, a variable region that is unique for each oligo and encoded the swap junction somewhere along its sequence, and a shared 3’ region which anneals to the EposM4 AT sequence. For the variable region, DEBSM6 and EposM4 were aligned using Clustal Omega and homology sequences in the KS-AT and post-AT linkers were selected corresponding to roughly 300 base pairs in length each[44]. To make sure each oligo contained a unique junction, only one was selected in regions of homology between EposM4 AT and DEBSM6 AT where adjacent junctions resulted in identical protein sequences. In total, the “forward” KS-AT oligo pool library contained 72 oligos and the “reverse” post-AT library contained 74 oligos. The oligo pool libraries were synthesized by IDT. For a complete sequence list, see Supplementary File 1.

For the library construction, the EposM4 AT sequence was amplified using the two oligo pool libraries using standard PCR conditions. The correct size band was gel purified and further enriched by using it as PCR template with primers binding to the shared 5’ sequence. The resulting product was cloned into a digested DEBSM6 mCherry. The library was then transformed into XL1-blue, purified and then transformed into the D*arsB*::P_{ibp} GFP strain.

Protein structure prediction using AlphaFold

Structure prediction of the homodimeric DEBSM6 was performed utilizing an adapted version of the AlphaFold program[28]. Modifications to enable homooligomer modeling were adapted from the ColabFold project[45]. Final DEBSM6 structure prediction was performed without the thioesterase, as AlphaFold was unable to generate a structure with correct placement of the thioesterase, for a total of 3360 amino acids. The best structure achieved a plddt of 87.9 using 2626 seconds for feature generation, 200 seconds for feature processing, and 883 seconds for model compilation and prediction.

For the additional PKSs modeled, we selected 15 modules with a high degree of domain diversity. These included modules with varying amounts of reducing domains, AT substrate specificities and with and without KR dimerization elements. To get models with high confidence scores, the sequence of the KS, AT and subsequent domain was input into AlphaFold. See Supplementary Table 2 for a list of PKSs, Supplementary File 2 for PDB files.

Protein purification

All purified proteins had the C-terminally fused mCherry tag clonally removed and the resulting plasmid transformed into BL21 (DE3). Growth and induction were done the same way we previously described except cultures were grown in 1 L cultures. After growing overnight at 18°C, cultures were harvested by centrifuging at 5000 g for 10 min and pellets were frozen and stored at -80°C until future use. Proteins were purified as described before[6]. Thawed cell pellets were resuspended in lysis buffer (50 mM phosphate buffer pH 7.6, 50 mM NaCl, 10 mM imidazole), lysed by sonication and centrifuged at 8000 g for 15 min three times to remove cell debris. The supernatant was mixed with 2 mL HisPur Ni-NTA resin (ThermoFisher Scientific) and incubated at 4°C for 1 hour. Next, the protein-resin mixture was washed three times with lysis buffer and eluted with elution buffer (150 mM phosphate buffer pH 7.6, 50 mM NaCl, 150 mM imidazole). The eluted proteins were then injected into an Äkta Explorer (Cytival) and captured on the anion exchange column HiTrap Q HP 5 mL (Cytiva), washed with 5 column volumes of wash buffer (50 mM phosphate buffer pH 7.6, 8% glycerol) and gradually eluted with anion elution buffer (50 mM phosphate buffer pH 7.6, 500 mM NaCl, 8% glycerol). Collected fractions were concentrated using Pierce Protein Concentrator 100K (ThermoFisher Scientific), and then aliquoted, frozen and stored at -80C.

In vitro assay

The *in vitro* reaction was based on a previously described protocol[6]. The following compounds were added to the reaction: 400 µM methylmalonyl-CoA tetralithium salt or malonyl-CoA tetralithium salt (both Milipore-Sigma), 2 mM 3-hydroxy-2-methylpentanoyl-S-N-acetylcyste- amine thioester, 2 mM DTT, 100 mM phosphate buffer pH 7.2 and 1 µM PKS enzyme. Reactions were incubated at room temperature and samples were taken at 1 hour for the methylmalonyl-CoA reaction and 24 hours for both methyl- and

malonyl-CoA samples. In vitro products were extracted by adding two volumes of ethyl acetate and collecting the supernatant twice. After evaporating the ethyl acetate, products were resuspended in 50% methanol and filtered through 3K Amicon centrifugational filters (Merck).

LC separation of TKLs was conducted at room temperature with a Kinetex XB-C18 column (100 mm length, 3 mm internal diameter, 2.6 µm particle size; Phenomenex) using a 1260 Infinity II LC System (Agilent Technologies). The mobile phase was 0.1% formic acid in water (solvent A) and 0.1% formic acid in methanol (solvent B). Products were separated at a flow rate of 0.42 mL/min using the following gradients: 20% to 72.1% B in 6.5 min, 72.1% to 95% B in 1.3 min and held for 1 min. Then, at a flow rate of 0.65 mL/min, 95% to 20% B in 0.2 min and held for 1.2 min. The LC system was coupled to an Agilent InfinityLab LC/MSD iQ single quadruple mass spectrometer (Agilent Technologies) and ESI was conducted in the negative-ion mode. Desmethyl and methyl triketide lactones were quantified by comparing peak areas with authentic standards synthesized by Acme Bioscience.

Data availability

A list of all plasmids used in this paper can be found in Supplementary Table 3 and plasmid sequences and strain request is available at the Joint BioEnergy Institute's public Inventory of Composable Elements (https://public-registry.jbei.org).

References

1. Butler, M. S., Robertson, A. A. & Cooper, M. A. Natural product and natural product derived drugs in clinical trials. Natural product reports **31**, 1612–1661 (2014).
2. Cortes, J., Haydock, S. F., Roberts, G. A., Bevitt, D. J. & Leadlay, P. F. An unusually large multifunctional polypeptide in the erythromycin-producing polyketide synthase of Saccharopolyspora erythraea. Nature **348**, 176–178 (1990).
3. Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L. Modular organization of genes required for complex polyketide biosynthesis. Science **252**, 675–679 (1991).
4. Menzella, H. G. *et al.* Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. Nature biotechnology **23**, 1171–1176 (2005).
5. McDaniel, R. *et al.* Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proceedings of the National Academy of Sciences* **96**, 1846–1851 (1999).
6. Yuzawa, S. *et al.* Comprehensive in vitro analysis of acyltransferase domain exchanges in modular polyketide synthases and its application for short-chain ketone production. ACS synthetic biology **6**, 139–147 (2017).
7. Dunn, B. J. & Khosla, C. Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. Journal of The Royal Society Interface **10**, 20130297 (2013).

8. Takahashi, S. *et al.* Reveromycin A biosynthesis uses RevG and RevJ for stereospecific spiroacetal formation. Nature Chemical Biology **7**, 461–468 (2011).
9. Laureti, L. *et al.* Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in Streptomyces ambofaciens. *Proceedings of the National Academy of Sciences* **108**, 6258–6263 (2011).
10. Zhu, X., Liu, J. & Zhang, W. De novo biosynthesis of terminal alkyne-labeled natural products. Nature chemical biology **11**, 115–120 (2015).
11. Porterfield, W. B., Poenateetai, N. & Zhang, W. Engineered biosynthesis of alkyne-tagged polyketides by type I PKSs. Iscience **23**, 100938 (2020).
12. Musiol-Kroll, E. M. *et al.* Polyketide bioderivatization using the promiscuous acyltransferase KirCII. ACS synthetic biology **6**, 421–427 (2017).
13. Walker, M. C. *et al.* Expanding the fluorine chemistry of living systems using engineered polyketide synthase pathways. Science **341**, 1089–1094 (2013).
14. Bagde, S. R., Mathews, I. I., Fromme, J. C. & Kim, C.-Y. Modular polyketide synthase contains two reaction chambers that operate asynchronously. Science **374**, 723–729 (2021).
15. Whicher, J. R. *et al.* Structural rearrangements of a polyketide synthase module during its catalytic cycle. Nature **510**, 560–564 (2014).
16. Dutta, S. *et al.* Structure of a modular polyketide synthase. Nature **510**, 512–517 (2014).
17. Chemler, J. A. *et al.* Evolution of efficient modular polyketide synthases by homologous recombination. Journal of the American Chemical Society **137**, 10603–10609 (2015).
18. Kim, B. S., Sherman, D. H. & Reynolds, K. A. An efficient method for creation and functional analysis of libraries of hybrid type I polyketide synthases. Protein Engineering Design and Selection **17**, 277–284 (2004).
19. Wlodek, A. *et al.* Diversity oriented biosynthesis via accelerated evolution of modular gene clusters. Nature communications **8**, 1–10 (2017).
20. Zutz, A. *et al.* A dual-reporter system for investigating and optimizing protein translation and folding in E. coli. Nature Communications **12**, 1–15 (2021).
21. Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. Rapid protein-folding assay using green fluorescent protein. Nature biotechnology **17**, 691–695 (1999).
22. Wigley, W. C., Stidham, R. D., Smith, N. M., Hunt, J. F. & Thomas, P. J. Protein solubility and folding monitored in vivo by structural complementation of a genetic marker protein. Nature biotechnology **19**, 131–136 (2001).
23. Cabantous, S., Terwilliger, T. C. & Waldo, G. S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. Nature biotechnology **23**, 102–107 (2005).
24. Lesley, S. A., Graziano, J., Cho, C. Y., Knuth, M. W. & Klock, H. E. Gene expression response to misfolded protein as a screen for soluble recombinant protein. Protein engineering **15**, 153–160 (2002).

25. NIELSEN, A. T., Ariane, Z. & Lennen, R. (Google Patents, 2017).
26. Kraft, M. *et al.* An online monitoring system based on a synthetic sigma32-dependent tandem promoter for visualization of insoluble proteins in the cytoplasm of Escherichia coli. Applied microbiology and biotechnology **75**, 397–406 (2007).
27. Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D. & Davidson, A. R. A simple in vivo assay for increased protein solubility. Protein Science **8**, 1908–1911 (1999).
28. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021).
29. Tang, Y., Chen, A. Y., Kim, C.-Y., Cane, D. E. & Khosla, C. Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase. Chemistry & biology **14**, 931–943 (2007).
30. Tang, Y., Kim, C.-Y., Mathews, I. I., Cane, D. E. & Khosla, C. The 2.7-Å crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proceedings of the National Academy of Sciences* **103**, 11124–11129 (2006).
31. Cogan, D. P. *et al.* Mapping the catalytic conformations of an assembly-line polyketide synthase module. Science **374**, 729–734 (2021).
32. Keatinge-Clay, A. T. & Stroud, R. M. The structure of a ketoreductase determines the organization of the β-carbon processing enzymes of modular polyketide synthases. Structure **14**, 737–748 (2006).
33. Khosla, C., Tang, Y., Chen, A. Y., Schnarr, N. A. & Cane, D. E. Structure and mechanism of the 6-deoxyerythronolide B synthase. Annu. Rev. Biochem. **76**, 195–221 (2007).
34. Chen, A. Y., Cane, D. E. & Khosla, C. Structure-based dissociation of a type I polyketide synthase module. Chemistry & biology **14**, 784–792 (2007).
35. Yuzawa, S., Kapur, S., Cane, D. E. & Khosla, C. Role of a conserved arginine residue in linkers between the ketosynthase and acyltransferase domains of multimodular polyketide synthases. Biochemistry **51**, 3708–3710 (2012).
36. Kapur, S., Chen, A. Y., Cane, D. E. & Khosla, C. Molecular recognition between ketosynthase and acyl carrier protein domains of the 6-deoxyerythronolide B synthase. *Proceedings of the National Academy of Sciences* **107**, 22066–22071 (2010).
37. Xiao, Y. *et al.* Characterization of tiacumicin B biosynthetic gene cluster affording diversified tiacumicin analogues and revealing a tailoring dihalogenase. Journal of the American Chemical Society **133**, 1092–1105 (2011).
38. Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nature chemical biology **12**, 1007–1014 (2016).
39. Kasey, C. M., Zerrad, M., Li, Y., Cropp, T. A. & Williams, G. J. Development of transcription factor-based designer macrolide biosensors for metabolic engineering and synthetic biology. ACS synthetic biology **7**, 227–239 (2018).

40. Miller, C. A., Ho, J. M., Parks, S. E. & Bennett, M. R. Macrolide Biosensor Optimization through Cellular Substrate Sequestration. ACS Synthetic Biology **10**, 258–264 (2021).
41. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature methods **6**, 343–345 (2009).
42. Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E. & Khosla, C. Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. Science **291**, 1790–1792 (2001).
43. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2**, 2006.0008 (2006).
44. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology **7**, 539 (2011).
45. Mirdita, M., Ovchinnikov, S. & Steinegger, M. ColabFold-Making protein folding accessible to all. bioRxiv (2021).

Declarations

Acknowledgment

This work was performed as part of the US Department of Energy (DOE) Joint BioEnergy Institute (https://www.jbei.org) supported by the DOE, Office of Science, Office of Biological and Environmental Research, under contract DEAC02-05CH11231 between the DOE and Lawrence Berkeley National Laboratory. E.E. was supported by Formas Mobility Grant Nr. 2017-00335. A.A.N. was supported by a National Science Foundation Graduate Research Fellowship, fellow ID 2018253421.

The authors acknowledge the support of Alex Hexemer, Hari Krishnan and Peter Zwart via Center for Advanced Mathematics for Energy Research Applications (CAMERA), which is jointly funded by the Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) programs in the Department of Energy Office of Science, under Contract No. DE-AC02-05CH11231 and via the Artificial Intelligence and Machine Learning at DOE Scientific User Facilities program under Award Number 107514 MLExchange.

Author Contribution

E.E. performed most experiments. E.E. and M.S. analyzed *in vitro* production. A.N. performed structure modeling. E.E., L.K., S.Y., J.D.K., were responsible for experimental design. All authors contributed to the preparation of the manuscript.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/xxx.

Correspondence and requests for materials should be addressed to Jay D. Keasling (email: jdkeasling@lbl.gov).

Competing interests: J.D.K. has a financial interest in Amyris, Lygos, Demetrix, Maple Bio, Napigen, Apertor Pharma, Ansa Biotechnologies, Berkeley Yeast, and Zero Acre Farms. L.K. has a financial interest in Lygos. All other authors declare no competing interests.

Figures

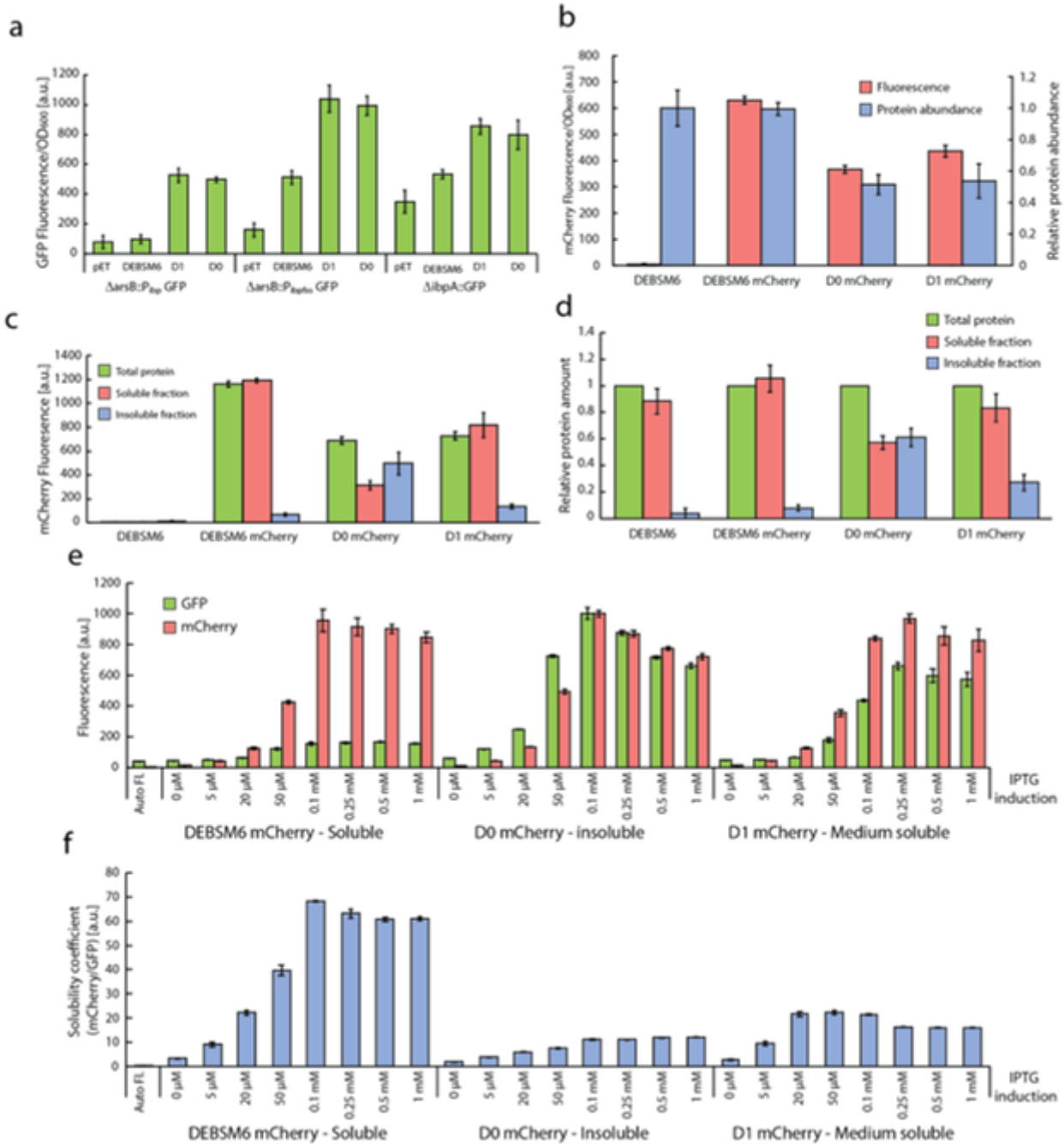

Figure 1

Activity of solubility biosensor and mCherry fusion proteins. (a) GFP fluorescence of *E. coli* solubility biosensor strains expressing PKSs with variable solubilities. pET = pET28a (empty vector control). (b) Fluorescence of mCherry tagged PKSs (left y-axis) and SDS-PAGE quantified abundance of the proteins relative to DEBSM6 amount (right y-axis). (c) mCherry fluorescence and (d) SDS-PAGE quantification of PKS proteins in different protein fractions of lysed cells, relative to "Total protein" for each replicate. Cells were induced with 250 μM IPTG in (a-d). (e) Fluorescence of D*arsB*::P_{ibp} GFP strain expressing mCherry tagged PKSs. (f) Same results as 1e with a simplified "solubility coefficient": the ratio of expressed protein (mCherry fluorescence) over activation of insolubility biosensor (GFP fluorescence). Error bars represent the standard deviation of three biological replicates.

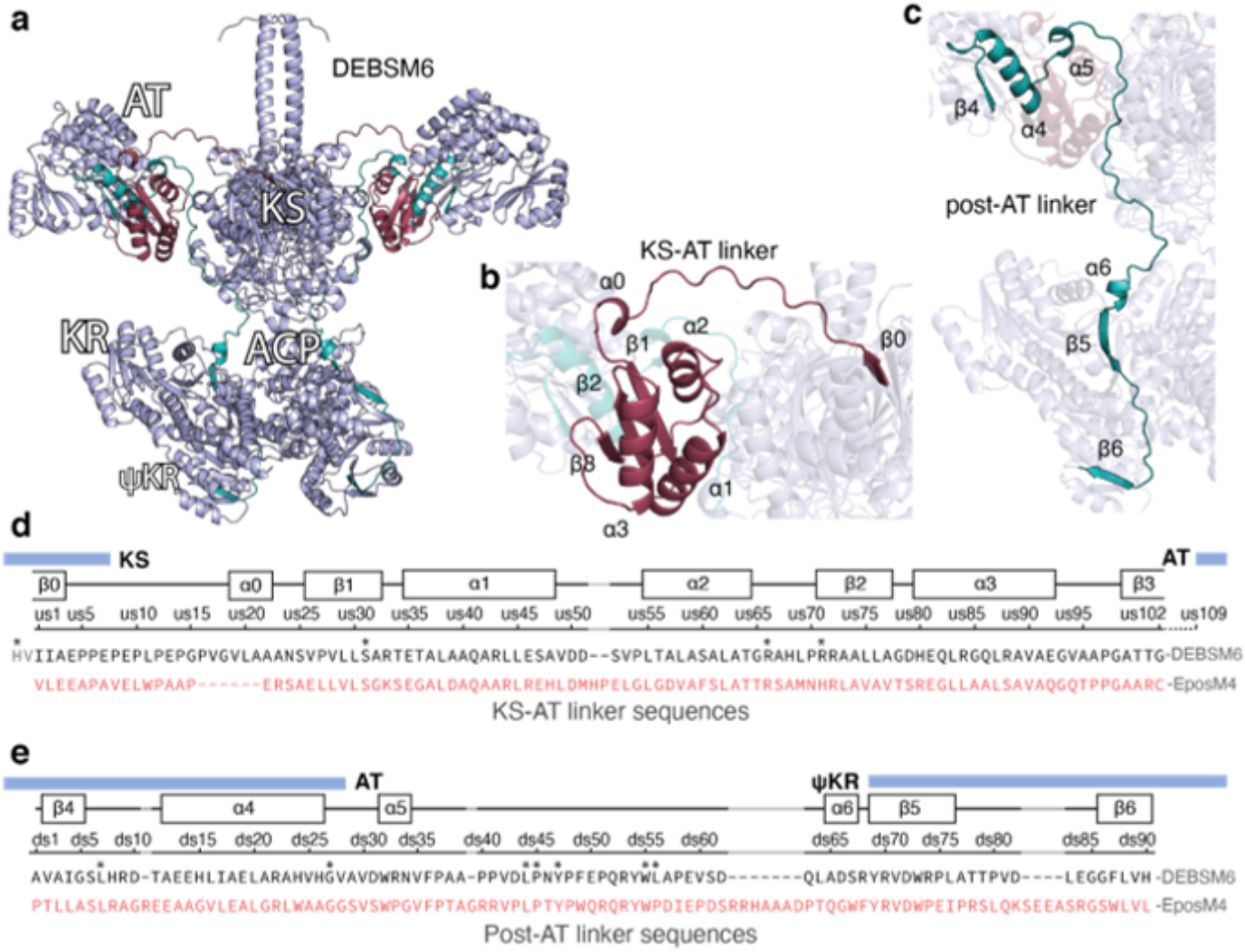

Figure 2

Predicted structure and sequence of DEBSM6 and KS-AT and post-AT linker. (a) AlphaFold structure prediction of homodimeric DEBSM6 without TE and a (b) highlighted structure of the KS-AT linker in dark red and the (c) post-AT linker in teal. (d) Alignment of DEBSM6 and EposM4 KS-AT and (e) post-AT linker with secondary structure elements predicted by AlphaFold and domain boundaries in blue boxes[33]. Each swap junction position in KS-AT linker is sequentially called us(upstream)1-102 and post-AT junctions called ds(downstream)1-90. Highly conserved residues are marked with asterisks. DEBSM6 KS-AT linker sequence starts with HV to denote where the conserved GTNAH sequence is positioned. Gaps in alignments are marked in grey.

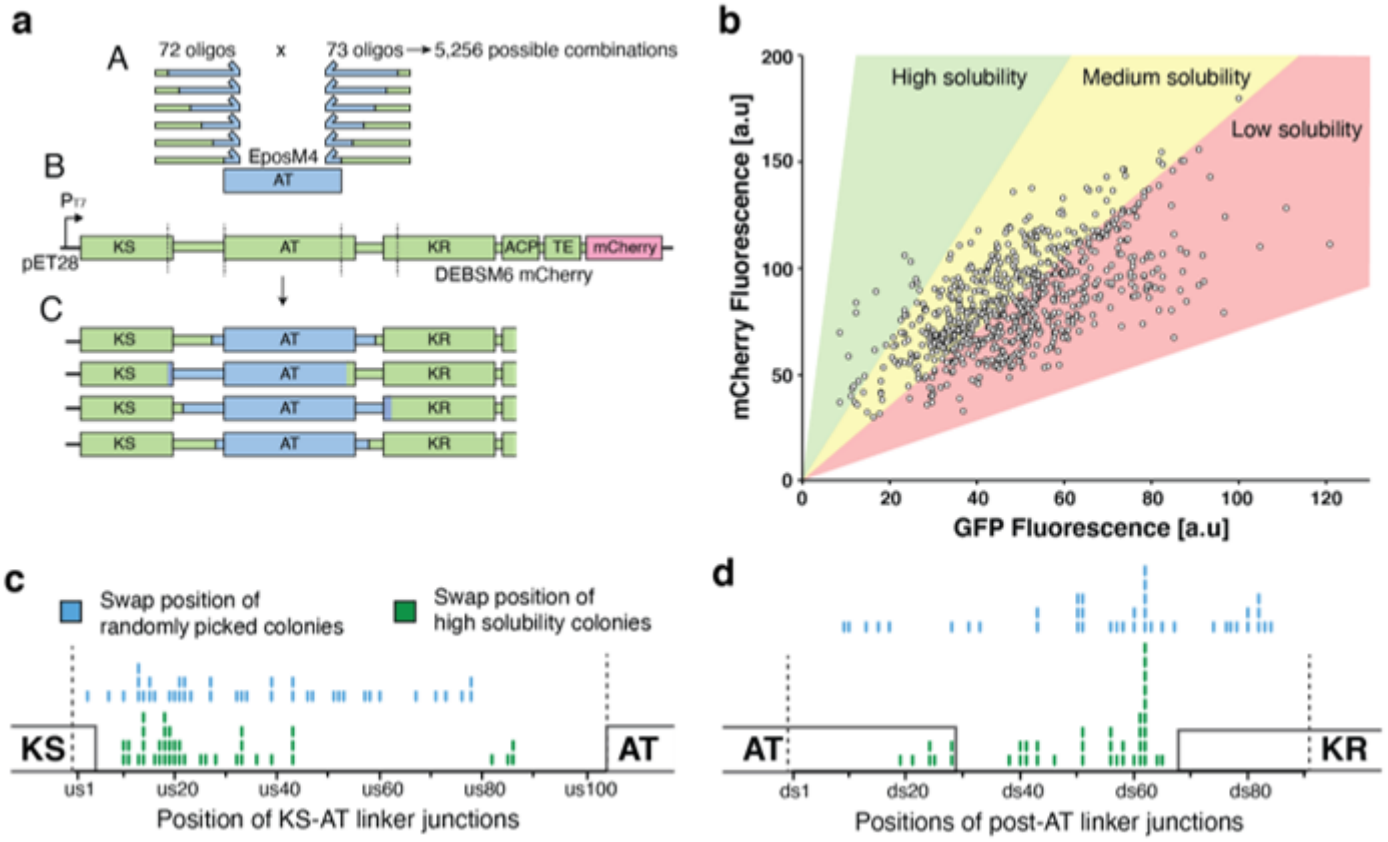

Figure 3

Creation of randomized swap junction library. (a) Strategy for creating library. (A) Oligo pool libraries each carrying a specific swap junction PCR amplifies EposM4 AT and is inserted into (B) DEBSM6 mCherry with the native AT excised. The resulting library (C) is comprised of a randomized upstream and downstream junction, with 5256 possible combinations. (b) Fluorescence measurement of swap junction library in D*arsB*::P_{ibp} GFP strain using flow cytometry. Each dot represents one measured colony. Colored areas are estimations where differently soluble variants would end up. (c) Comparing swap positions between the randomly picked colonies (blue sticks) with high solubility colonies (green sticks) in the KS-AT linker and (d) in the post-AT linker. Certain swap positions were overrepresented (e.g. ds62) due to gaps in the alignment leading to multiple library variants sharing the same swap junction. Dotted line denotes selected linker region.

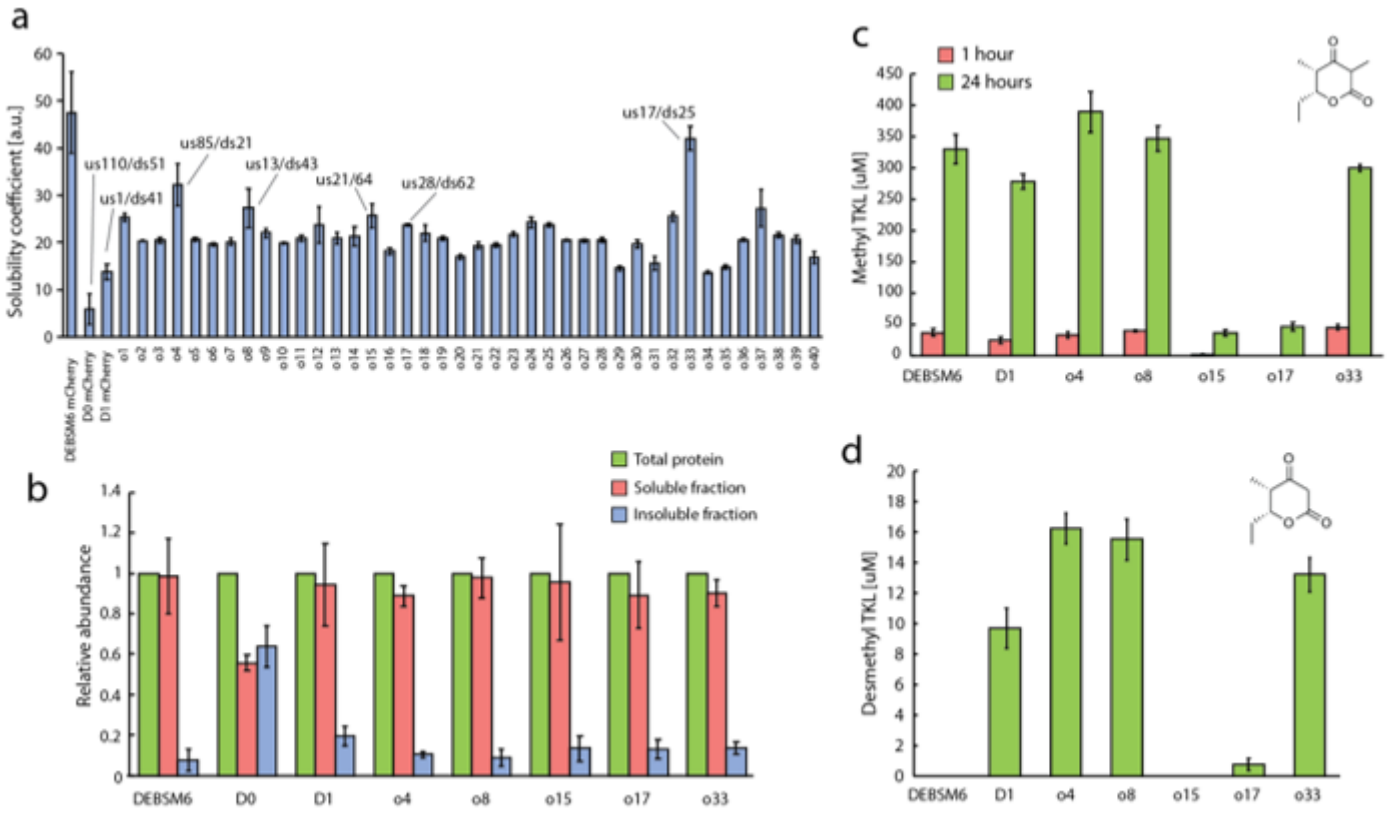

Figure 4

Solubility and *in vitro* activity of high solubility library variants. (a) Solubility measurement using D*arsB*::P_{ibp} GFP biosensor strain of 40 high solubility colonies, a subset of which have their swap positions marked and were selected for (b) SDS-PAGE quantification of PKS abundance in different protein fractions. (c) *In vitro* production after 24 hours of methyl TKL and (d) desmethyl TKL after 1 and 24 hours. DEBSM6 is the parental PKS, D1 is included as a reference to what is currently known as the optimal swap junctions. D0 was excluded due to it already been shown to be inactive[6]. All strains were induced with 250 μM IPTG. Error bars represent the standard deviation of three biological replicates.

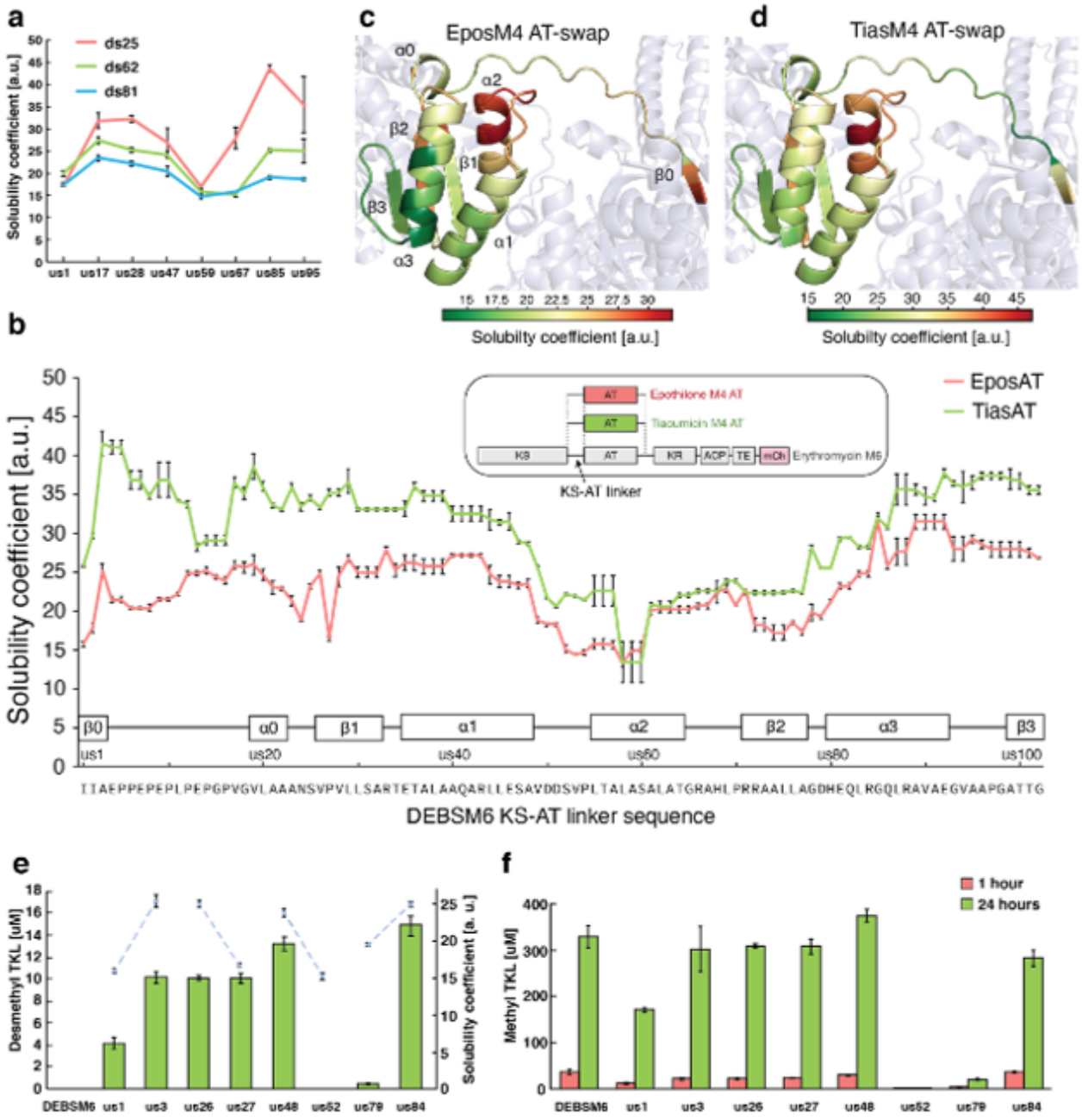

Figure 5

Effect of KS-AT linker junction positions on solubility and activity. (a) Testing amount of synergy between KS-AT and post-AT linker junctions on solubility. Solubility measured by D*arsB*::P_{ibp} GFP strain of DEBSM6 mCherry engineering with EposM4 AT with different swap junction combinations. (b) KS-AT swap junction effect on solubility measured by D*arsB*::P_{ibp} GFP strain for DEBSM6 mCherry engineered with EposM4 AT or TiasM4 AT with downstream junction at ds25. Data points in regions of homology that give identical polypeptide sequence are repeated. Induction was 250 μM IPTG. (c) Solubility data from 5b visualized on predicted protein structure of DEBSM6 for EposM4 AT domain swap or (d) TiasM4 AT swap. (e) *In vitro* production of desmethyl TKL (green bars, left axis) and solubility (blue dots, right axis) of engineered DEBSM6 with EposM4 AT swap with the ds25 downstream junction. Dotted lines

between samples indicate variants being compared with similar junction positions but differences in solubility. (f) *In vitro* production of metyhl TKL at 1 hour (green bars) and 24 hours (red bars). Error bars represent the standard deviation of three biological replicates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx
- SupplementaryFile1oligopoollibrarysequences.xlsx
- SupplementaryFile2pdbstructures.zip