**FIGURES AND TABLES**

Table 1. Experimentally validated gene clusters harboring a self-resistance gene

Figure 1. Identifying pks clusters harboring a putative self-resistant target gene

Table 2. Number of pks clusters harboring a putative self-resistant gene and more than one copy per nucleotide record.

Figure 2. Phylogenetic tree of KS domains closest to a self-resistant gene (150 KSs), ref clusters annotated only TODO:redo tree; annotate manually positive set

Figure 3. Coevolution of KS and target (upper left: positive set, upper right: 12 targets, 5kb, lower left: 10kb, lower right: >50kb)

Table 3. Top 20 ranked clusters harboring a putative **underexplored and novel** antibacterial target

Table 4. Table of putative novel antibacterial targets

Figure SXa. Coevolution of KS and UDP-3-O-acyl-N-acetylglucosamine deacetylase (Lpxc) target. Left – within **5kb** of a KS. Right: at least **50kb** away From a KS, **redundant set**

Figure SXb. Coevolution of KS and ClpP, DegP, and FtsH protease targets within **5kb** of a KS, **redundant set**

FigureS1a. Phylogenetic tree of KS domains closest to a self-resistant gene 12 targets, 10kb, annotated

FigureS1b. Phylogenetic tree of KS domains closest to a self-resistant gene 92 targets, 10kb

FigureS1c. Phylogenetic tree of KS domains closest to a self-resistant gene 609 targets, 10kb

Figure S2a,b,c. Coevolution of KS and target 12 targets, 10kb, one plot, subplots, individual plots

Figure S2d,e,f. Coevolution of KS and target 92 targets, 10kb, one plot, subplots, individual plots

Figure S2g,h,i. Coevolution of KS and target 609 targets, 10kb, one plot, subplots, individual plots

TableS1. Catalog of 150 clusters harboring a potential self-resistant gene (12 targets, 10kb)

TableS2. Catalog of 572 clusters harboring a potential self-resistant gene (92 targets, 10kb)

TableS3. Catalog of 3431 clusters harboring a potential self-resistant gene (609 targets, 10kb)

Supplementary file 1. Targets.12.fasta

Supplementary file 2. Targets.92.fasta

Supplementary file 3. Targets.609.fasta

**ABSTRACT**

With the increase in antibacterial resistant pathogens and the decline in discovery of novel antibiotics, the need for accelerating their discovery is urgent. More than two-thirds of the known antibiotics are natural products or natural-products-derived, as polyketides and non-ribosomal peptides being one of the most prolific classes. They are synthesized by large multimodular enzymes called polyketide synthases (PKSs) and nonribosomal peptide synthetases (NRPSs), respectively [1]. In microbial genomes, PKS and NRPS genes are often co-localized with all other genes required for the production of a given compound in a biosynthetic gene cluster. With the increasing ease of DNA sequencing, hundreds of novel PKS and NRPS gene clusters have been identified in recent years [2, 3]. However, accessing novel and useful PKS activities remains complicated, as many host organisms are uncultivable or do not express PKS genes in lab conditions. Furthermore, a significant challenge in the natural product discovery field is how to identify useful biosynthetic gene clusters, without a-priori knowledge of the biological target of the produced compound.

To identify useful PKS genes, we developed an automated method to identify and catalog clusters that harbor potential self-resistance genes. These genes provide the cluster-harboring microorganisms a defense against the antibiotics encoded by the clusters. For example, the genes encoding the fatty acid synthase inhibitor thiotetronic acid are clustered with a resistant copy of the fatty acid synthase gene (fabB/F) [11].

With this approach, we generated a non-redundant catalog of 152 PKS clusters. Manually curated lists of known antibacterial target genes were used to mine all NCBI nucleotide and genome databases. This method is generalizable since it can be applied to extract not only PKS-related gene clusters, but also other types of natural products. It can also be used to identify potential clusters not only from bacterial origin, but also from fungal and plant native hosts. Most importantly, it can be used to prioritize gene clusters harboring putative novel targets and encoding compounds with new mechanisms of action.

**INTRODUCTION**

* Resistance in pathogens occurs quickly after a drug is introduced in the clinic. Need for new antibacterials due to overuse and misuse of current antibiotics. Need for novel mechanisms of action, novel antimicrobial targets, no. Special interest in gram-negative bacteria (*Escherichia coli, K. pneumoniae, P. aeruginosa and A. baumannii)*, since they are harder to kill (drug hard to penetrate double cell wall and reach its intracellular target). *No new class of antibiotic has been approved for Gram-negative pathogens in over fifty years (REF?).*
* 2/3 of drugs – natural products, most of which polyketides – great track record for the clinic. Traditional discovery- bioactivity based, cultivating organism, isolating compound, screening for bioactivity, sequencing genes encoding the compound. Problems with the traditional approach: (1) producer not cultivable under standard lab conditions only about 1% of organism grow in lab conditions; (2) genes encoding the compound are silent and thus not expressed; (3) compound hard to chemically synthesize due to its complex structures. 64 steps needed to synthesize erythromycin.
* Traditional discovery vs sequence-guied discovery. Advent in dna sequencing, number of modular clusters (2013 PKS paper), and typeII clusters (2016 PNAS paper)
* Computational prediction of clusters – trivial now (Automated genome mining tools like antismash) but accessing novel and useful PKS activities remains the bottleneck.
* One way to look for self-resistant genes: Many antibacterial compounds are produced by bacteria and, in order to avoid self-toxicity, these antibiotic-producing microorganisms have developed several self-resistance mechanisms. One resistance mechanism is target modification, where a resistant copy of the target gene is co-localized and co-expressed with the biosynthetic genes [10]. There are several examples in literature mining for specific targets of interest and specific class of microorganisms. Mining self-resistant genes to discover clusters with potential antibacteria properties – examples, narrow focus on fatty acid inhibitors, actinomycetes, etc. (ARTS, Bradly Moore paper)
* The antibiotic resistant target seeker (ARTS) is a great web tool to detect known-resistance mechanism in a biosynthetic gene cluster, based on BCG proximity, duplications, and HGT. However, it has its limitations: 1) limited to Actinobacterial genomes and thus it cannot provide info for HGT or presence of housekeeping self-resistant genes for non-Actiobacterial species; (2) limited to known-resistant mechanisms; (3) one criteria: BGC proximity; it relies on Antismash to predict the boundaries of the BGC and antismash over-predicts boundaries to include all possible biosynthetic genes, very often the self-resistant target would be far away from the cluster.

We have developed an automated method to identify and catalog clusters that harbor a potential antibacterial target protein, generating a non-redundant catalog of 152 PKS clusters. Manually curated list of known antibacterial target genes was used to mine the NCBI database. The algorithm takes into account (1) the distance of the potential target to a core enzyme in the biosynthetic gene cluster, (2) the presence of a duplicated housekeeping copy of the target gene, (3) the presence of close homologs in diverse species, and (4) evidence for coevolution of target and core biosytnthetic gene. We have ranked the clusters based on the above-mentioned critera. We also applied this method to search for clusters harboring novel antibacterial targets and here we present the highly ranked top 20 hits. This method is generalizable since it can be applied to extract gene clusters from any class of natural products, which can be characterized by a core enzyme (DMATS (alkaloid) Trichodiene synthase (terpene) GGPPS (terpene)). It can also be used to identify potential clusters not only from bacterial origin, but also from fungal and plant native hosts. Most importantly, it can be used to prioritize gene clusters harboring novel targets and encoding compounds with new mechanisms of action.

**Materials and Methods**

**Curating of t1pks gene clusters harboring a putative self-resistant gene**

*Step 1. Blast search for KS homologs. Blast 8 diverse KS against 11 NCBI nucleotide databases.*

All ncbi nucleotide and genome databases were searched for KS homologs using tblastn. The tblastn algorithm searches a protein query against nucleotide databases by translating the nucleotide records into all 6 possible open reading frames. This allows search into unannotated databases, such as Assembly database. 8 diverse KS from modular type1 pks (erythromycin), cisat pks/nrps (curacin, epothilone, guadinomine, rapamycin) and transat pks/nrps hybrids (leinamycin, disorazol, chivosazol) polyketide classes were used as query sequences against the major ncbi nucleotide and genome databases (nt, wgs, refseq\_genomic, other\_genomic, env\_nt, patnt, htgs, tsa\_nt, sts, gss, est\_others (all updated April 25, 2018, except for wgs, Jan-27-2016). Since we were unable to update wgs, we downloaded all Bacterial Assemblies database (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702866/>). As of Oct-26-2018, there are 172 642 Bacterial Assemblies.

An initial relaxed blast search (e value < 1) identified non-redundant 199,894 protein records, of which 110,174 unique NCBI nucleotide records/genomes (<99% similar).

*Step 2. Run Antismash on all 100174 genbank files (89k genbank ids and 21k assembly ids).*

Download all 89449 genbank files from the traditional databases. The genbank files from the Assembly database (21080 total) are already stored locally. Antismash 4 with the minimal functionalities was run on all 100k sequences. While running antismash on assembly ids, the server was running out of memory because (1) genbank files were very big and (2) Antismash is also parallel. To avoid breaking the server, we split the assembly genbank files into smaller ones by extracting sequence 150kb upstream and downstream of a KS, resulting in a total of 26575 assembly files, and then ran antismash with 5 parallel processes on all sequences. Total run time for all 110k sequences was 95h.

Antismash ran on 79053 genbank files total (52480 traditional ncbi gebank files and 26573 assembly files). The rest 21121 genbank files were below 1000b and antismash does not run on shorter than 1000bp sequences.

*Step 3: Extract KS and gene sequences from all PKS-labeled clusters found by antismash and make a database.*

Otut of the 79k gbids, there are 29987 clusters annotated as type 1 pks (this includes cisat pks, transat pks, pks/nrps hybrids, and hybrids of them). We extracted all KS sequences, as well as all genes from all t1pks-annotated clusters and we made two databases: there are 244 196 sequences from nucleotide records in NCBI Bacterial assemblies database and 664,336 sequences from nucleotide records in the traditional NCBI databases. 78 clusters don’t have predicted KS domains (details\_data empty).

*Step 4. Blast search for a self-resistant target gene in a PKS cluster*

We searched the above-made antismash database for 12 experimentally verified targets using a relaxed e-value <1, which resulted in 4,404 protein hits. Filtering blast hits by e-value <1e-8 and identity > 0.3 (and FabB/F identity > 0.6) reduced the set to 806 proteins.

*Step 5. Require both KS and target to be less than 10kb apart*

In most experimentally characterized clusters harboring a self-resistant target gene (Table 1), the target gene was within a 10kb-distance of a core KS gene, with the exception of FabB/F in platencimycin (13kb away) and threonyl-tRNA synthetase in borrelidin (14.5kb). We chose an initial maximum allowed distance between a KS and a putative self-resistant gene as 10kb. This filter reduced the set to 252 KS + target proteins, in 250 unique nucleotide records.

[X gbids had a target but its partner KS is not annotated as a KS by antismash, skipped]

**Phylogenetic analyses**

To remove redundant gene clusters, we selected a standard redundancy threshold of 90% KS sequence identity, which further reduced the number of pks gene clusters to 152. Multiple sequence alignment was performed using MAFFT and phylogenetic tree was generated using FastTreePMP. Tree was visualized using the APE package in R. Tree was rooted on E.coli FabB/F and colored either by phyla or self-resistant gene.

**Housekeeping copy**

To count the number of copies per genbank record, the 12 self-resistant targets were used as queries in blastp search of all genbanks records harboring clusters with these 12 targets using e-vaule cutoff of of 1e-8 and identity threshold of 0.3 (and FabB/F identity of 0.6). The number of copies per nucleotide record were counted and the status of the genbank record – complete genome vs incomplete was noted.

[*There are some empty files because there is no sequence in genbank file. Ignored]*

**Coevolution**

We plotted KS1-KS2 pairwise amino acid identities vs. pairwise identities of the corresponding target within 5, 10, or 20kb of a core KS gene. We fist plotted pairwise identities from positive set, including clusters with putative self-resistance gene (clarexpoxin, eponemycin, cinnabaramide: target co-localized with cluster, compound known to inhibit target homolog). We also calculated a coevolution score for each biosynthetic gene cluster:

Score = sum of all distances to diagonal abs(x-y) divided by the number of pairs this cluster occurs

**Scoring scheme**

We devised a scoring scheme to rank the curated pks clusters harboring a potential self-resistance gene, based on several parameters: (1) distance of self-resistant gene to the closest KS, (2) presence of homologs in diverse species (>80% pairwise KS identity, positive control: niddamycin and spiramycin produce identical molecules and they have X% identity of M1KS; also thyolactomycin clade); (3) presence of a housekeeping copy; (3) Coevolution score; (4) Target ubiquity – how many times a target has been identified in the data set. If many times: higher chance of false positives, lower score.

**RESULTS**

**Positive set of 12 known self-resistant targets in 18 biosynthetic gene clusters (Table 1)**

In order to curate bacterial BGC harboring a self-resistant copy of the target gene, we first curated a list of known gene targets colocalized with the BGC by (1) and extensive literature search, and a computational search of the MibiG database. We downloaded all sequences from the well annotated MiBig database. For all of the molecules for which the Molecular target is known and annotated in Mibig, we downloaded all molecular targets gene sequences from uniprot and blasted each target against all genes in all BGC. After discovering hits, we read the literature for experimental evidence for self-resistance of that particular gene in that particular cluster. In total we have identified 16 clusters harboring 12 experimentally verifies self-resistant target genes, spanning several classes of natural products: PKS, NRPS and their hybrids, terpenes, RiPPs, aminocumarins. As of that search, MibiG consisted of ~1800 clusters as of 2018, ~700 of which are polyketides. This is a total discovery rate of 0.88% (16/1800) and 1.22% (9/737) for pks and its derivatives.

There are several additional BGC with a known molecular target, which is also co-colocalized with the biosynthetic genes, but it was not experimentally shown (although very likely), that the presence of that target is the self-resistant mechanism of the producer (additional table?)

Identifying clusters harboring a potential self-resistant target gene by mining MIBIG (target of molecule known, target copy in clusrter not tested if it is the resistant

**Curating of t1pks gene clusters harboring a putative self-resistant gene**

Out of the ~30, 000 annotated-as-pks clusters by Antismash, 152 non-redundant clusters had a co-localized putative self-resistant gene within 10kB of a KS domain. Increasing the distance cutoff to 20kB increased the number of non-reduntant clusters to 306. These are discovery rates of 152/29987= 0.5% and 306/29987=1.02%, respectively, which are in a good agreement with the discovery rate from known clusters in MiBiG (1.22%). The lower discovery rate for the 10kb cutoff suggests that this cutoff is very stringent, to increase the discovery of true positives and to reduce the discovery of false positives.

Chance of having a target gene next to a biosynthetic gene cluster = Average cluster size x Average Number of pks clusters per genome / average genome size = 10 000 \* 5 / 6 000 000 = 0.83%

**Target abundance**

In Table 1 are listed the number of clusters harboring one of these 12 targets. Clusters harboring ACC and Ile-tRNA synthetases are the most abundant ones.

For ACC the true positive discovery rate is 1/1800 = 0.05%, whereas by running this pipeline it is 49/29987= 0.16%, which is 3 times higher. This suggests that many of these clusters are false positives. One explanation is that the acc is the biosynthetic gene providing the extender unit for the polyketide chain. Similarly, Ile-tRNA synthetase could be a biosynthetic gene involved in the biosyhtesis of amino acids incorporated in the pokyletide/nrps molecule. In this case, the cluster harboring Ile-tRNA synthetase does not contain nprs modules, this could suggest that the target is a self-resistant gene, or if the A domain has specificity for another amino acid (although aa predictions are not great). Even if there are no nrps modules, the gene could be involved in regulation of translation. Similarly, for most of the target genes which are metabolic enzymes, we could not distinguish between being self-resistant genes or biosynthetic genes. For other non-metabolic enzymes-like targets, such as GyraseB, it is less likely that they are biosynthetic.

Other targets are underrepresented, like DNA polymerase sliding clamp- harboring ones. It is possible that some self-resistant targets are only specific to a certain class of natural products.

**Discovery of PKS clusters from the positive set:**

There are 15 clusters harboring a self-resistant gene. Of them, 9 are PKS or PKS/NRPS hybrids. 8 out of 9 were identified in the pipeline. We could not identify andrimid, because it was misannotated in antismash as a non-pks cluster (arylopropyene-nrps). 7/8 were identified using a KS-target tandem distance cutoff of 10kb, and 8/8 we identified using a KS-target tandem distance cutoff of 20kb (increasing the cutoff to 20kb included the discovery of borrelidin, which thr-tRNA self-resistant target is 12kb form a KS gene).

**Housekeeping copy**

Out of the 150 clusters, there are 27 clusters with more than one copy per nucleotide record. There are 24 complete genomes, 12 of which harbor at least 2 copies of the self-resistant gene. In comparison with the positive set of clusters, there is an additional housekeeping copy of the target on the genome of at least 4 out of 9 clusters-harboring producers (Table 1). We identified 3 out of 4 (the Isoleucyl tRNA synthetase housekeeping copy of the mupirocin producer has lower than 30% sequence identity with the self-resistant copy nd thus was not identified by our pipeline (23%). Thus, having an additional copy is a good metric to prioritize clusters to express. However, it is not required for the producer’s growth to keep both copies of the target, since harboring just the self-resistant one is sufficient for its survival, and it could even be argued that is is evolutionary more advantageous to lose the sensitive copy. from the positive set there is 3 out of the 252 clusters harbor. Indeed, for the ansamycin rifamycin, there is only one copy of RNA polymerase beta (ropB), which is resistant to the compound. Thus, harboring a housekeeping copy is not necessary characterisitic of self-resistant-gene-harboring clusters. Not much weight could be put on this parameter also because there is no reference copy number per genome. Could use ARTS for reference number, but it would be specific to Actinobacteria. Even in that case, isolates from the same species might have different number of copies of a gene of interest and it would be difficult to determine what is the reference copy number for that gene for that species.

**Phylogeny**

Most are Actinobacteria (2/3rds), but there are also other prominent phyla, such as Proteobacteria, Firmicutes, …

We hypothesize that if the target is the self-resistant gene, there would be homologs of that cluster, which would also harbor the target. The presense of homologs would speak to the importance of that cluster and thus the encoded molelcue, especially if it could be found in distant species. Indeed, we see several clades, clustered by target: there is an Ile-tRNA synthetase clade, there is a FabF clade, ACC clade, etc.

Many acc homologs, ile-trna synthetase homologs. Looking more closely at the architecture of the pks modules, we identified the true homologs on a module and even cluster-level. Interestingly, there are no gyrB-harboring cluster homologs. Reasons: 1) We might have not covered the whole sequence space and that’s why they are missing, or (2) this target is not found in pks-derived clusters and all we see are false positives, or (3) this is newly-evolved self-resistance mechanism and not picked up by other bacteria yet.

We have mapped all positive examples on a tree – a roadmap - looking for analogs or thiolactomycin or orphan clusters.

There are some very distantly related KS genes on the tree. Looking more closely, it seems that they are misannotated by antismash and are actually from type II PKS clusters.

**Coevolution**

We used co-evolution analyses to detect gene swap events in domains of type II pks clusters (REF). Here, we extended that approach to test if a KS gene co-evolved with a potential self-resistance gene (target) by comparing pairwise protein identity scores between pairs of homologs. If two KSs from different genomes have high sequence identity, and the corresponding targets pair also have high sequence identity, this could suggest that the KS and target co-evolved. Similarly, if the KSs have low sequence identity and the targets also have low sequence identity, this would also suggest that the KS and target co-evolved. However, there are several cavetas of using KS-target coevolution as a metric for presence of a self-resistance gene in a pks cluster: the closest to a target KS might have been horizontally gene transferred, which could (1) either lead to an off-diagonal group, even if the target is part of the cluster, or (2) lead to an on-diagonal group, even though the target is not part of the pks cluster. With that caveat in mind, we first tested whether we could apply co-evolution analyses to our positive data set. We build a co-evolution plot for pairs from the positive set (beta protesome inhibitors: salinosporamide, cinnabaramide, clarexpoxcin, eponemycin; Isoleucyl tRNA synthetase inhibitors: mupirocin and thiomarinol, and FabB/F inhibitors: thioteramide and thiolactomycin), extending it to clusters harboring self-resistant genes, which were not experimentally verified to be the self-resistant mechanism.

As you can see in Figure 3a, all pairs are located close to the diagonal. To further test this approach, we selected a KS-target disctance cutoff of 5kb, 10kb, 20kb, and >50kb (to capture random KS- target pairs) and constructed a co-evolution plot for each data set (Figure 3b-e, respectively). Pairwise identities are colored by target. Several interesting features can be observed: (1) most of the points on the 5kb plot follow on the diagonal, which suggests that the KS and targets pairs are not random; (2) for some targets, such as FabB/F, all points are on the diagonal, and most of these clusters are close homologs to thiolactomycin and thioteramide, which serves as another confirmation that coevolution plots can be used as a proxy for discovering clusters harboring self-resistant genes. Increasing the distance cutoff to 10 and then to 20kb increased both the number of points on the diagonal, but it also increased the number of off-diagonal points, suggesting that using a more stringent KS-target cutoff could reduce the false-discovery rate. Most importantly, increasing the cutoff to more than 50kb generated a plot with complete random set of pairwise identities (Fig 3e).

For each cluster, we calculated a coevolution score based on averaging all scores of all clusters this one pairs with. The lower the coevolution score, the more partners this cluster has on the diagonal and thus the higher the chance that the target coevolved with the cluster and is thus the self-resistant gene. Clusters from positive dataset are exclusively on diagonal.

Another interesting characteristic is the stripes pattern for some of the targets, like ACC – diverse KS1-KS2 pairwise identities over the same value for target1-target2 identities. This would indicte that the target gene does not coevolve with the KS gene. One should be careful when picking an ACC-harboring cluster based on coevolution score only.

Although points on diagonal, they are a little below diagonal: higher KS-KS pairwise identity vs lower target1-target2 identitiy. This could be explained due to the fact that the KS is just to domain (the most conserved part of the pks module, excluding variable linkers, etc), whereas the target is the whole gene, which could be much more variable.

**Ranking clusters**

We developed a scoring scheme based on distance, housekeeping copy, coevolution score, homologs, ubiquity of target (rare targets more valuable than common targets), and ranked all clusters. The top –ranked clusters, including the clusters from pos set are in Table 3.

**Applying pipeline to interesting drug targets (but hard to find inhibitors for,or pathway known, specific enzymes not):**

We performed an extensive literature search to identify interesting antibacterial targets. Furthermore, we searched MiBiG database for clusters harboring a potential self-resistance gene (See Materials and Methods section). We curated a list of 92 targets from pathways of interest but not explored as self-resistant genes. We applied our pipeline by mining these 92 genes and identified 639 clusters with 5kb threshold and Y clusters with 10kb threshold.

Interesting targets:

**Aminoacyl-trna synthetases as targets.**

We mined for all 20 aminoacyl trna synthetases but identified clusters harboring only a few of them. Name them. Examples from positive set, there is a precedent. New targets:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307249/>

*Pseudomonic acid or mupirocin is one example, yet is one of the most effective topically applied antibiotics used to combat methicillin resistant S. aureus. This antibiotic is a naturally occurring isoleucyl-tRNA synthetase inhibitor produced by Pseudomonas fluorescens strains and works by docking onto the enzyme catalytic active site and competing with the isoleucine and ATP substrates for binding. Rationally designed stable mimics of aminoacyl-adenylate reaction intermediates have also been studied extensively and prove to be potent inhibitors of their respective aaRS in vitro. However, these inhibitors have issues with bioavailability and bacterial uptake and often do not sufficiently discriminate between bacterial and eukaryotic aaRSs for use as an effective antibiotic.*

<https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#!divAbstract>

*One class of such natural product aaRS inhibitors consist of the polyketides (shown in*[*Fig. 2*](https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#fig2)*) mupirocin A (****3****) (targets IleRS), reveromycin A (****4****) (targets IleRS), borrelidin (****5****) (targets ThrRS), and granaticin (****6****) (targets LeuRS).*[*5*a](https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#cit5a)*Mupirocin is currently the world's leading topical antibiotic for the treatment of MRSA infections,*[*6*](https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#cit6)*and borrelidin was shown to be effective against*Plasmodium*, the causative agent of malaria.*[*7*](https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#cit7)*Successes have also been achieved with synthetic aaRS inhibitors, such as benzoxaboroles, a new class of broad-spectrum antifungals that inhibit LeuRS.*[*8*](https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#cit8)*However, this class of aaRS inhibitors do not generally represent ideal lead compounds because, (a) in spite of their potent inhibitory properties*in vitro*, their biological activity*in vivo*is compromised by their inability to effectively penetrate bacterial cell envelopes, and (b) some of these compounds show cross-reactivity against targets other than aaRSs.*[*9*](https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#cit9)

So there is a need for naturally occurring tRNA synthetase inhibitors.

methionyl-tRNA synthetase inhibitor REP8839 not a natrula product

<https://pubs-rsc-org.stanford.idm.oclc.org/en/content/articlelanding/2012/md/c2md20032e#cit86>

*However, a recently identified novel prokaryotic AspRS and GluRS activity may provide an as yet unexplored target for antibiotic development. Several prokaryotes do not contain the genes for AsnRS and GlnRS. In these organisms, AspRS and GluRS mischarge the tRNAAsn and tRNAGln with Asp and Glu respectively, to form Asp-tRNAAsn and Glu-tRNAGln. Amidotransferases then convert these mischarged tRNA molecules to Asn-tRNAAsn and Gln-tRNAGln which are subsequently used for translation*

One of the high scoring clusters CP011420 harbors indeed an Asp-tRNA synthetase and is very similar architecturally to the known antibiotic bacillaene. It is not known what is the bacillaene target (It could be asp-tRNA synhtetase). Mining of Mibig could prove useful for a hypothesis-driven research to identify the unknown targets of known clusters. But the cluster is somewhat different, so the compound might be different enough.

Other clusters – several trnaRS genes. False positives or even more potent antibacterials?

Methionine-tRNA ligase clusters – only 10 pts. No homologs, no coevolution score

Phenylalanine--tRNA ligase alpha subunit

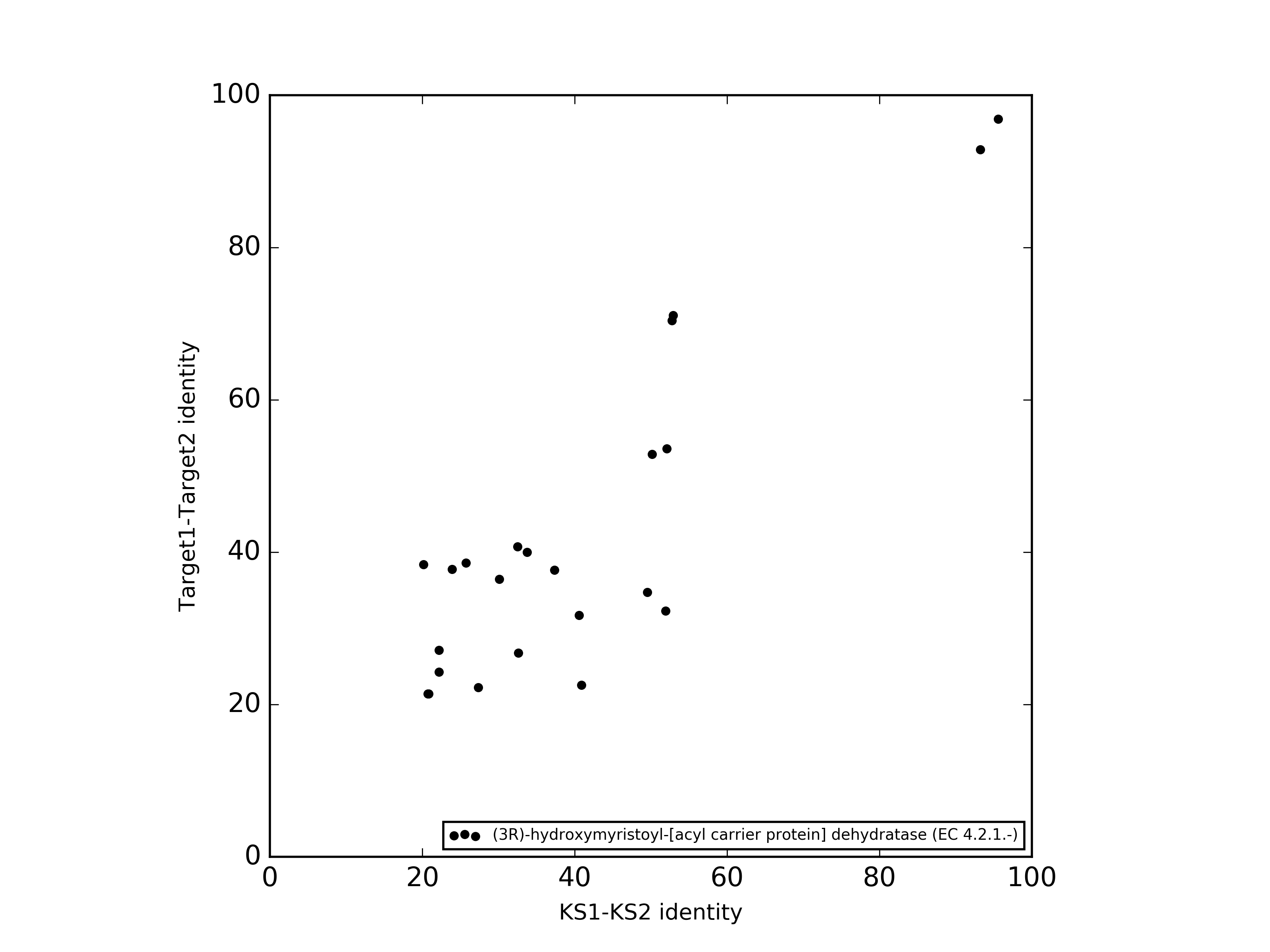
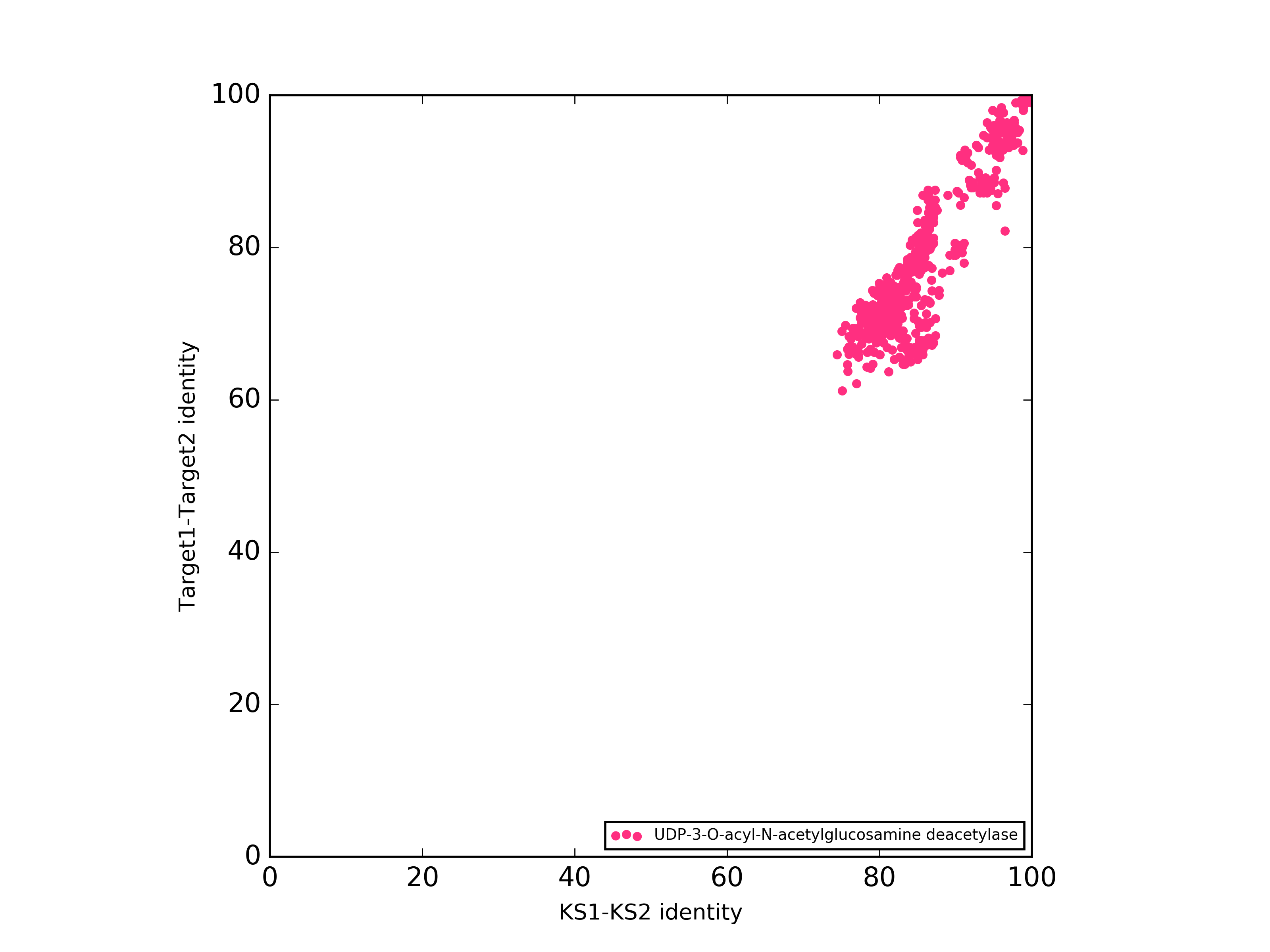
Proline--tRNA ligase- 50pt

Serine--tRNA ligase – 60pt

Threonine--tRNA ligase – 55pt

Tyrosine--tRNA ligase – 55pt

**Lipid A biosynthesis -** UDP-3-O-acyl-N-acetylglucosamine deacetylase (Lpxc) and FabZ coevolution plot:



Make a plot for >50kb

This enzyme catalyzes the first committed step in lipid A biosynthesis.Conserved in all gram negative bacteria (except Acinetobacter baumannii). Most common resistance mechanism was a mutation in FabZ. *LpxC programs have been much more successful than other target-directed antibiotic discov- ery efforts of the past several decades although no compound has yet been tested in human infections.* The coevolution plots suggest that LpxC and FabZ co-evolved with their corresponding KS genes, and thus could prove to be the self-resistant genes in these biosynthetic gene clusters. It is striking that there are no off-diagonal posint on the LpxC plot, suggesting that LpxC is exclusively close to pks genes and there are most probably no false positives. Hypotheses as to how this gene might be a biosynthetic gene?

*For both E. coli and K. pneumoniae, the most common finding was point mutation of fabZ, which encodes R-3-hydroxymyristoyl acyl carrier protein dehydrase (Clements et al. 2002; Zeng et al. 2013; Tomaras et al. 2014). As dis- cussed below, it is thought that reducing the rate of phospholipid synthesis allows the cell to tol- erate a reduction in LPS synthesis*. *These observations suggest that in fabZ mutants, the mechanism of resistance to LpxC inhibition is restoration of the balance between LPS synthesis and synthesis of phospholipids, reducing flux through both pathways while maintaining the permeability barrier of the outer membrane. Overexpression of fabZ was shown to increase the amount of LpxC protein in cells, supporting the idea that FabZ and LpxC activities are coregulated.*

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/>

*In recent years, infections caused by antibiotic-resistant bacteria have emerged as major threats to human communities worldwide.*[*1*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/#ref1)*However, the late-stage clinical development pipeline for antibacterials has been unacceptably lean in recent decades.*[*2*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/#ref2)*In particular, no drugs have reached advanced stages of development for the treatment of infection due to multidrug-resistant Gram-negative bacteria. Thus, there is a great need to develop new mechanisms by which Gram-negative antibacterial agents can combat bacterial antibiotic resistance.*[*3*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/#ref3)*,*[*4*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/#ref4)

*One of the emerging targets in Gram-negative bacteria is LpxC, an essential enzyme in the lipid A biosynthetic pathway. Because LpxC does not show homology to any mammalian protein, it is a promising antibiotic target for developing novel therapeutics against multidrug-resistant Gram-negative pathogens.*[*5*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/#ref5)*LpxC inhibitors have drawn much attention in new entities for Gram-negative antibacterial agents.*[*6*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4904260/#ref6)

**Proteases**

*Although their perturbation clearly offers the potential for antimicrobial drug development, both as traditional antibiotics and anti-virulence drugs, they are not yet the target of any clinically used therapeutics. One of the reasons for the lack of new antibiotics is a feeling that the traditional molecular target for antibiotics, cell wall biosynthesis, protein and DNA/RNA synthesis have perhaps been over-mined.1 Interfering with these enzyme and regulatory activities offers a route to untapped antibiotic targets.*

*In mammals, proteases have diverse, clinically relevant roles and are the target for an estimated 5–10% of all drugs being developed.14 Finally, the success of drugs targeting mammalian* ***proteases validates their druggability.*** *Indeed, proteases have complex structures with potential drug binding pockets in active sites, protein–protein interaction sites, cofactor-binding sites or other allosteric sites.*

*Despite these advantages, there are* ***currently no approved antimicrobial agents targeting bacterial proteases****.*

*he history of antibiotic development suggests that such methods are unlikely to be fruitful against bacteria, which have mechanisms to sense and avoid small molecules including efflux pumps and complex membrane structures such as those of Gram-negative bacteria.16 Furthermore, they are unlikely to have the structural complexity required to specifically bind a bacterial protease and not its mammalian homolog. In contrast to synthetic drug-like molecules, natural products are highly complex chiral molecules that have been sculpted by millions of years of evolution to enter bacterial cells.16 These may be the critical leads required to propel bacterial proteases into the realm of bona fide antimicrobial drug targets.*

*There are four families of intracellular proteolytic complexes ubiquitous in eubacteria: Lon, HslUV (ClpQY), ClpXP and FtsH. In addition to these five, HtrA (DegP) is a periplasmic/secreted proteolytic complex, whereas the prokaryotic proteasome is found only in actinomycetes. Several of these complexes have been investigated as potential antibacterial targets, including Lon,18 ClpXP,19 HtrA20 and the proteasome.21 Of these, ClpXP has been most extensively investigated and is the only complex for which natural product inhibitors have been found thus far.*

*The pioneering efforts of Böttcher and Sieber44 to target ClpP led to the development of a series of β-lactone inhibitors. Ultimately, however, low plasma stability due to rapid hydrolysis of the cyclic ester precluding further clinical development. Although ClpP inhibition shows promise as a mechanism of action, further development and discovery of novel scaffolds is clearly required.*

*Finally, the unprecedented mode of action by activation rather than inhibition of its target may allow such a drug to be effective against dormant persister cells.*

*This finding suggested that ADEPs activate ClpP, causing inappropriate protein degradation.*

*ADEP activation does not allow stably folded proteins to be degraded, as they are still too large to enter ClpP’s axial lumen, but does allow unstable proteins and nascent chains emerging from the ribosome to be degraded, especially if they fold slowly.57*

*Although ADEPs are promising leads for drug candidates, they have unfavorable pharmacological properties including poor water solubility, rapid systemic clearance and chemical instability.*

*Despite the increased potency of these ADEP derivatives, they still only have limited activity against Gram-negative bacteria and are efficiently removed from the cell by active efflux, especially in M. tuberculosis.*

*Lavey et al.68 instead screened 420 450 fungal and bacterial extracts or metabolites and identified a single ClpP activator, sclerotiamide (Figure 4a). This paraherquamide-related indolinone was 73-fold less potent than ADEP1 at activating EcClpP and failed to inhibit growth of efflux deficient E. coli or Pseudomonas aeruginosa.*

Natural products inhibitors of proteases: Cyclomarin *CymA, ecumicin and lassomycin are all bactericidal against replicating M. tuberculosis, a range of other mycobacterial species, and multidrug-resistant M. tuberculosis. Importantly, they are also active against nonreplicating M. tuberculosis.* Despite the relatively high potency of cymA, ecumicin and lassomycin against M. tuberculosis, optimization of pharmacological properties is required.

*The hybrid polyketide-peptide macrolactone antibiotic TA (also known as myxovirescin; Figure 7b) is also an inhibitor of SP-II.98*

<https://mibig.secondarymetabolites.org/repository/BGC0001104/index.html#cluster-1>

It has a SPII in the cluster! Not mentioned in the paper what is the self-resistant mechanism.

<https://onlinelibrary-wiley-com.stanford.idm.oclc.org/doi/full/10.1002/cbic.200600075>

Sclerotiamide – not in mibig

Cyclomarin – nrp <https://mibig.secondarymetabolites.org/repository/BGC0000333/index.html#cluster-1>

[Ecumicin - nrp https://mibig.secondarymetabolites.org/repository/BGC0001582/index.html#cluster-1](%20Ecumicin%20-%20nrp%20https://mibig.secondarymetabolites.org/repository/BGC0001582/index.html#cluster-1)

Lassomycin – not in mibig

**No pks, pks/nrps hybrids inhibitors of proteases**

***FtsH protease****, the product of the essential ftsH gene, is a membrane-bound ATP-dependent metalloprotease of Escherichia coli that has been shown to be involved in the rapid turnover of key proteins, secretion of proteins into and through the membrane, and mRNA decay. FtsH protease is the only ATP-dependent protease of E. coli that has been shown to be essential for bacterial growth (24). The ftsH homolog in Helicobacter pylori has also been shown to be essential (7), and while its homolog in the Gram-positive Bacillus subtilis is not essential, cells lacking ftsH grow poorly (6).*

**DegP – potential antibacterial target. No known inhibitors**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5223389/>

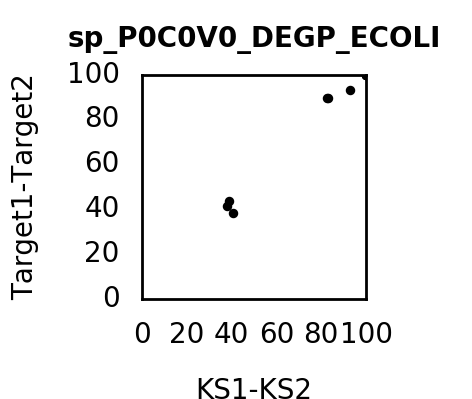
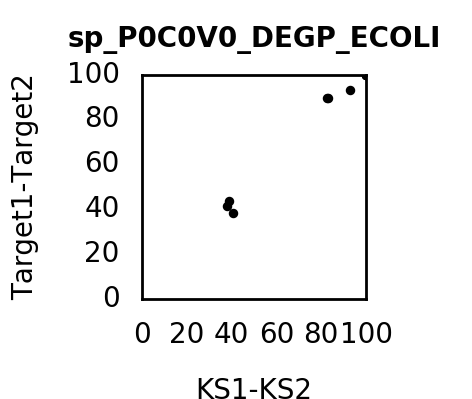
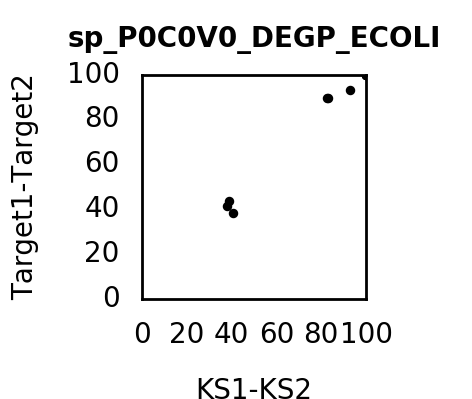
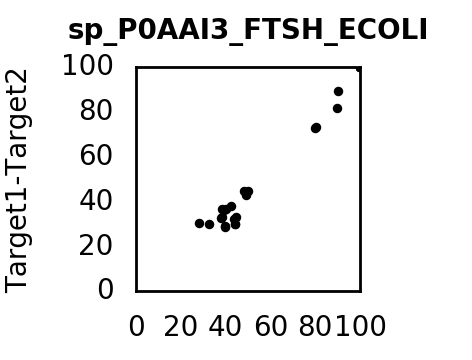
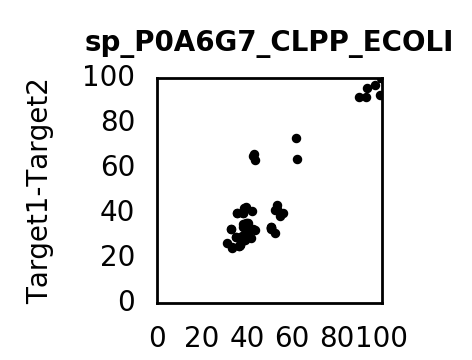
### *Why is HtrA(DegP) inhibition a step forward in the fight against pathogens?*

*With the exception of*Mycoplasma genitalium*and*Methanococcus janaschii*, it seems that all bacterial pathogens and commensals in the microbiota express HtrA proteins; a fact that evades the classical and precise definition of virulence or pathogenic factors [*[*42*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR42)*]. Consequently, this observation leads to the question if such a factor might also serve as a potent macromolecular drug target? In fact, targeting HtrA offers some potential advantages:*

1. *(i.)it is secreted into the extracellular micro-milieu or presented on the bacterial cell surface and therefore accessible to drug compounds [*[*43*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR43)*,*[*44*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR44)*],*
2. *(ii.)it has a defined enzymatic active site and substrate recognition [*[*19*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR19)*,*[*20*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR20)*,*[*45*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR45)*,*[*46*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR46)*],*
3. *(iii.)it cleaves E-cadherin, proteoglycans and fibronectin as host factors with important functions for bacterial pathogenesis [*[*19*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR19)*,*[*20*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR20)*,*[*21*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR21)*,*[*47*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR47)*], and*
4. *(iv.)it is an essential enzyme in*H. pylori*physiology [*[*40*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR40)*,*[*41*](https://biosignaling.biomedcentral.com/articles/10.1186/s12964-017-0162-5#CR41)*].*

*These characteristics make HtrA a potentially attractive candidate for novel therapeutic approaches to treat bacterial pathogenesis.*

**Lon protease – in 92.10kb**



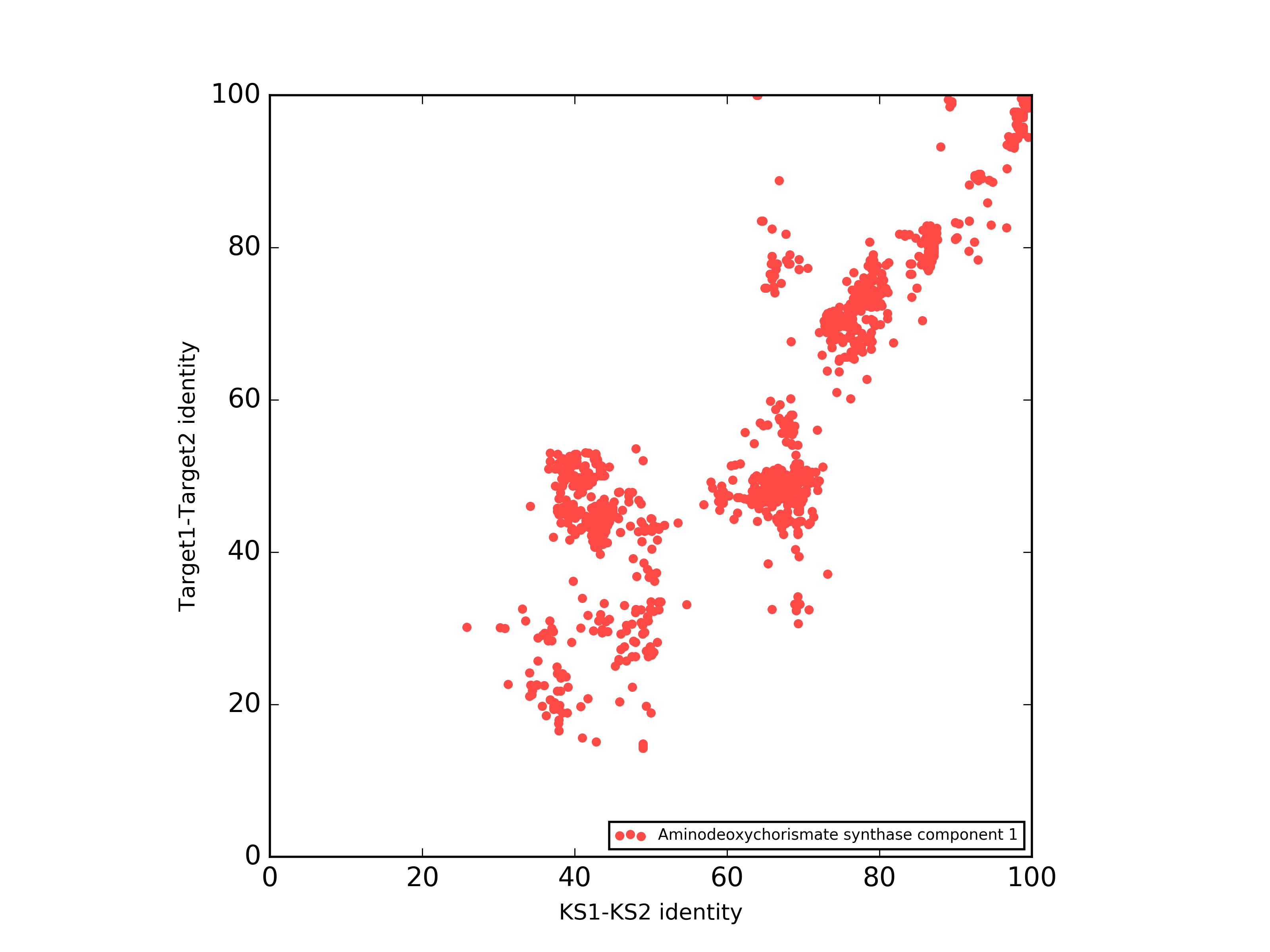
Good target- high evolution scores for all clusters.

**Tetrahydrofolate biosynthesis ( in 54 targets paper: 4-amino-4-deoxychorismate mentioned)**

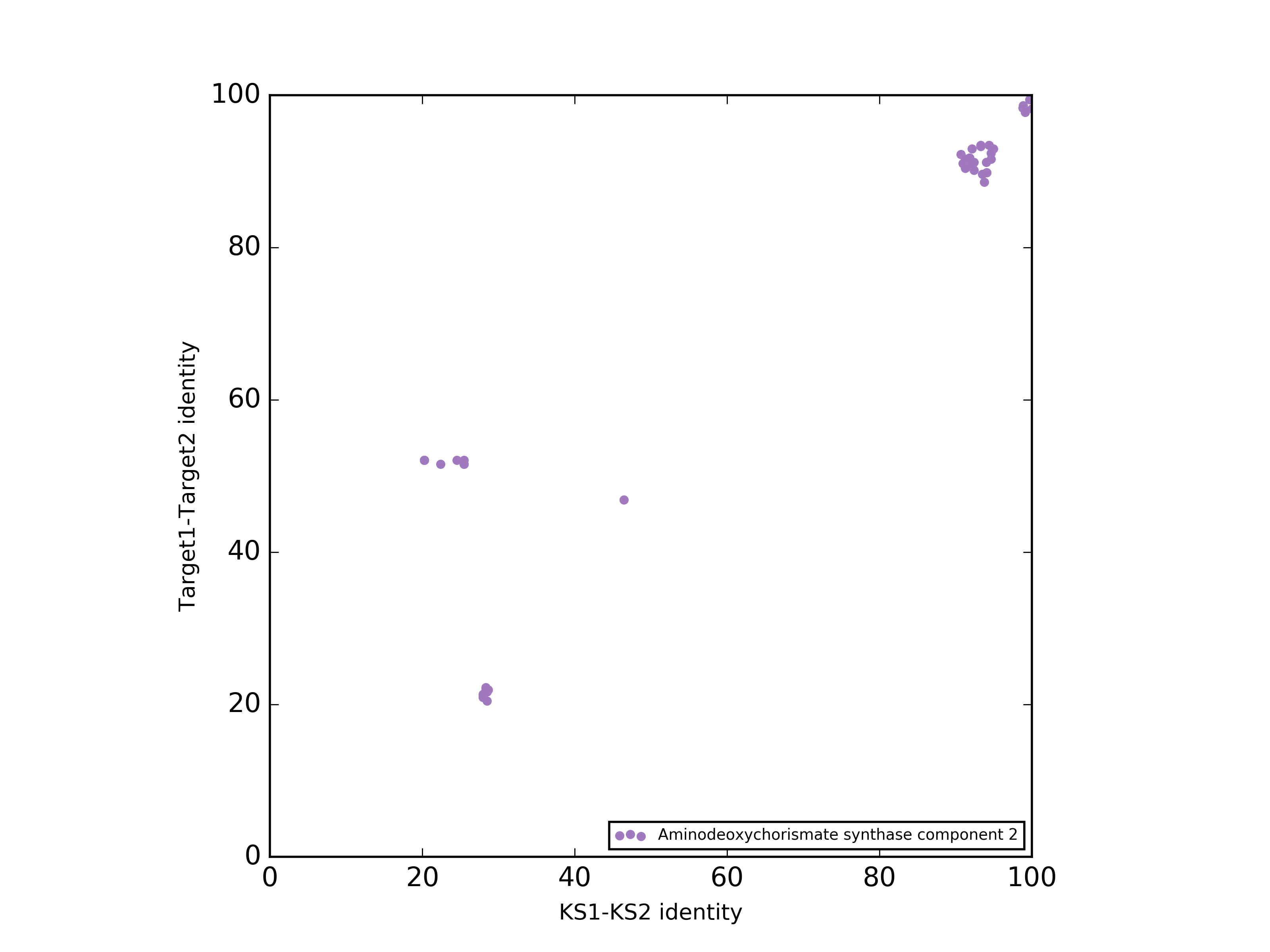
**Natural products inhibiting the** chorismate pathway, but self-resistant gene is a pump.

<https://www.mdpi.com/1420-3049/23/6/1371/htm#B21-molecules-23-01371>

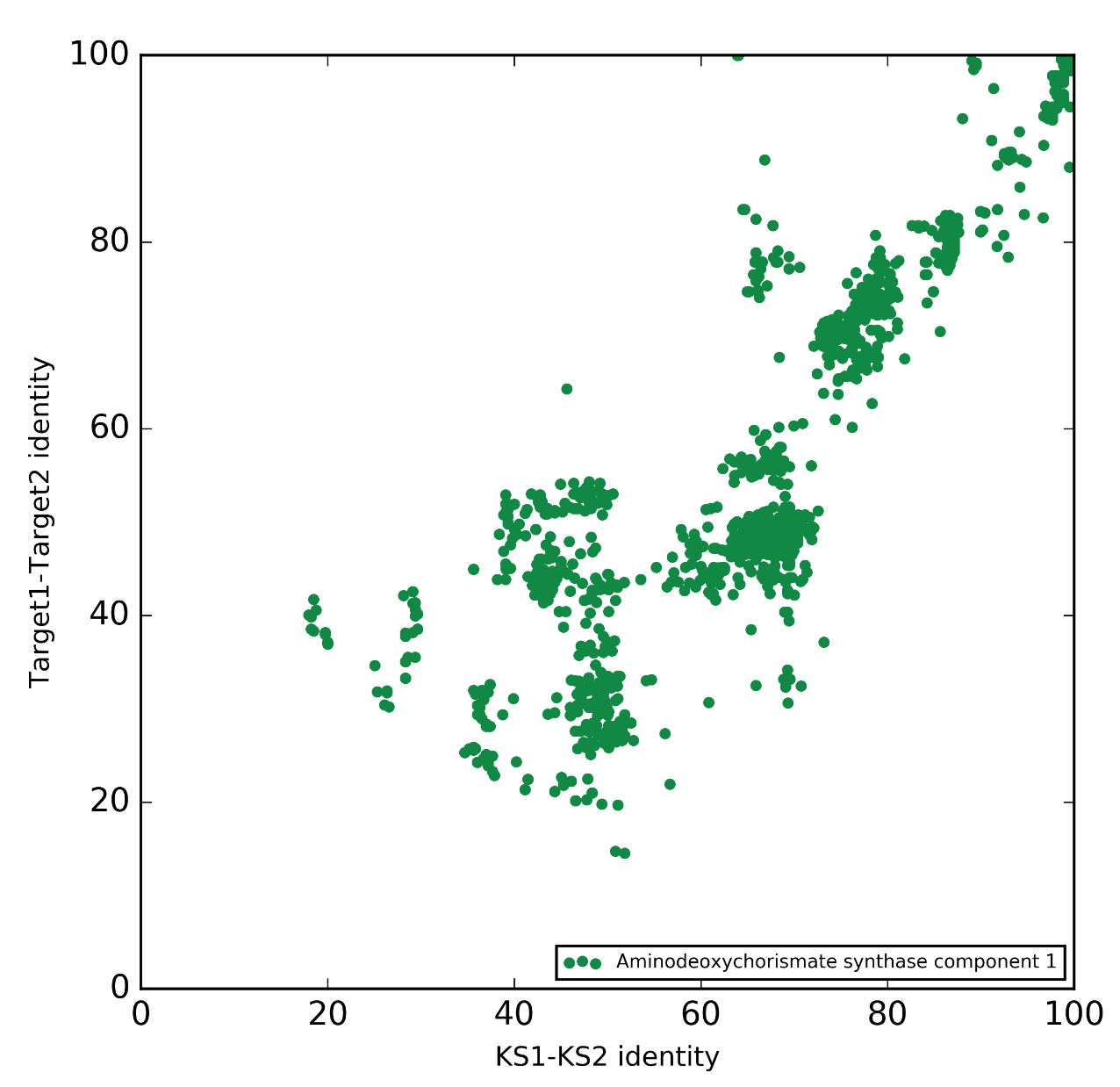
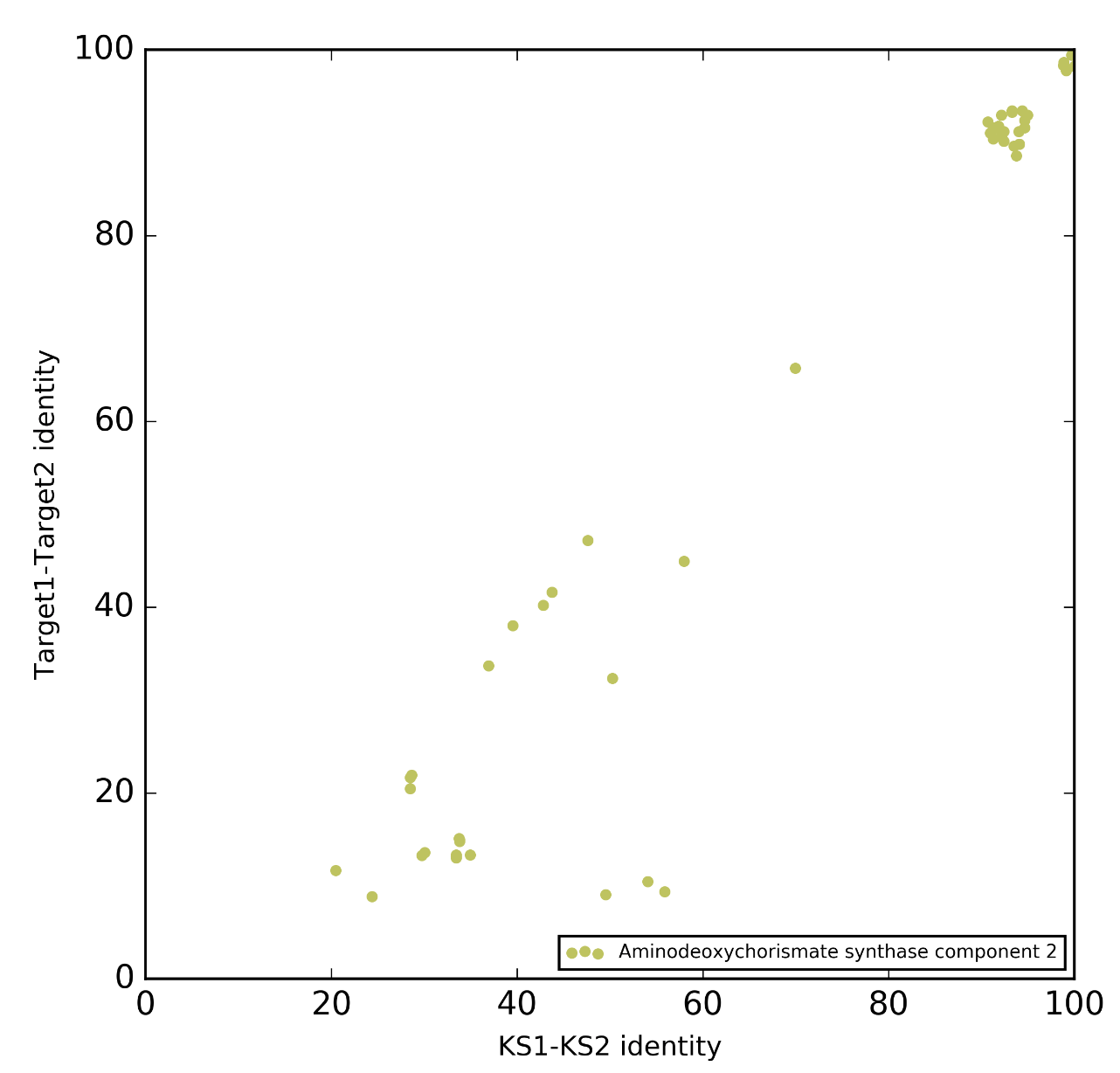
Abyssomicins are antifolates inhibiting the synthesis of pABA within the chorismate pathway.



Aminodeoxychorismate synthase component 1 (pabb)



Aminodeoxychorismate synthase component 2 (paba)



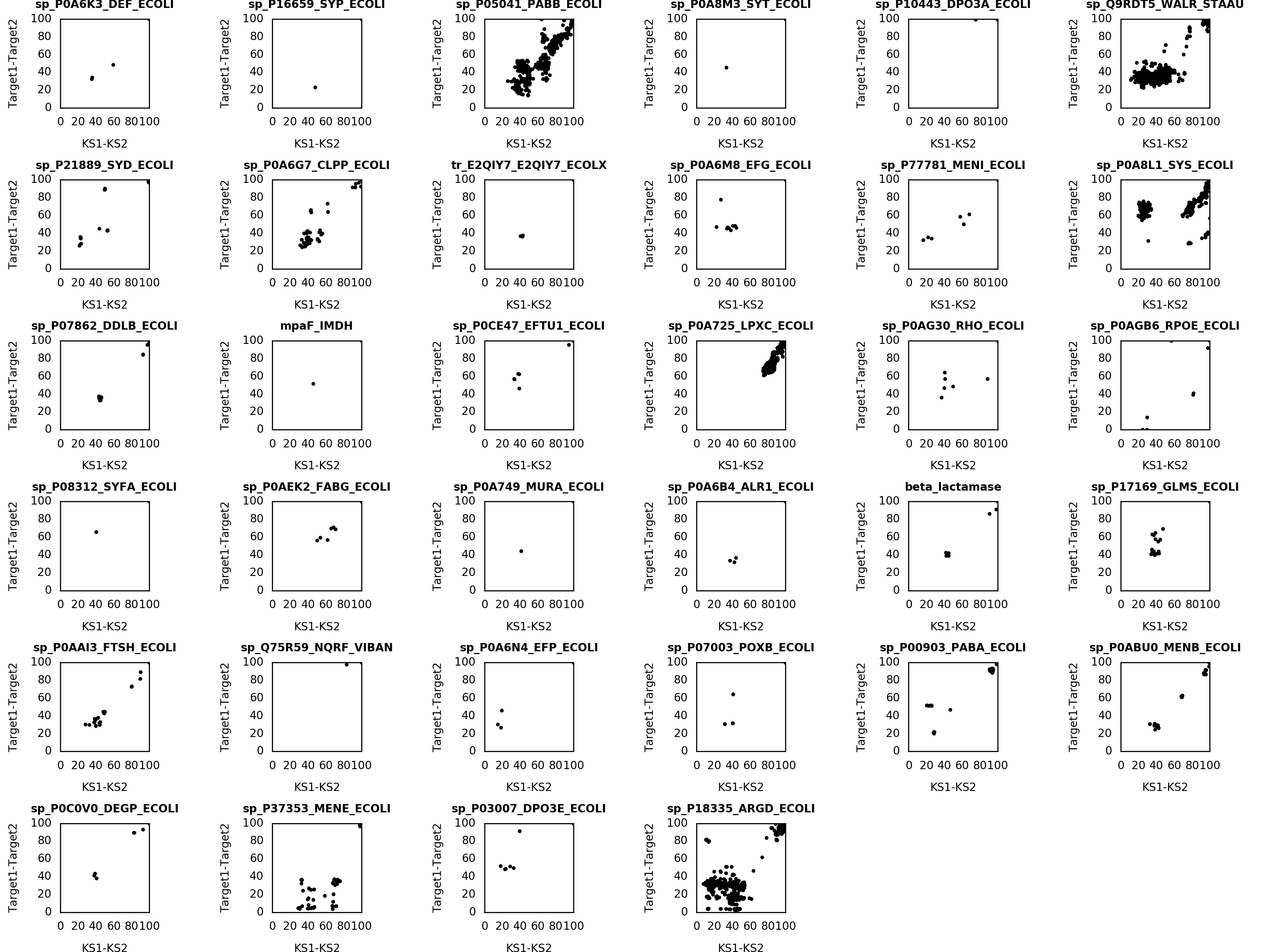
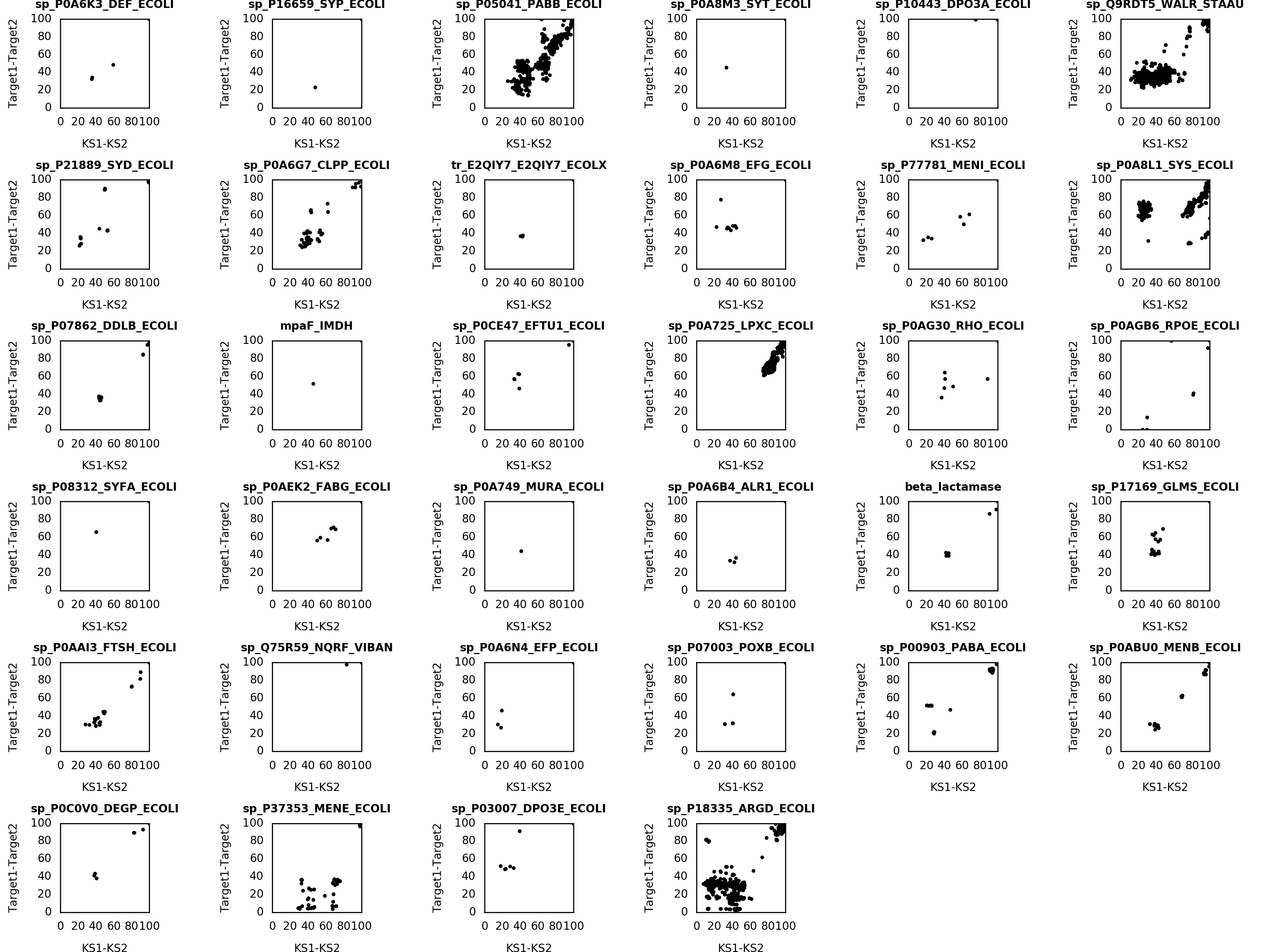
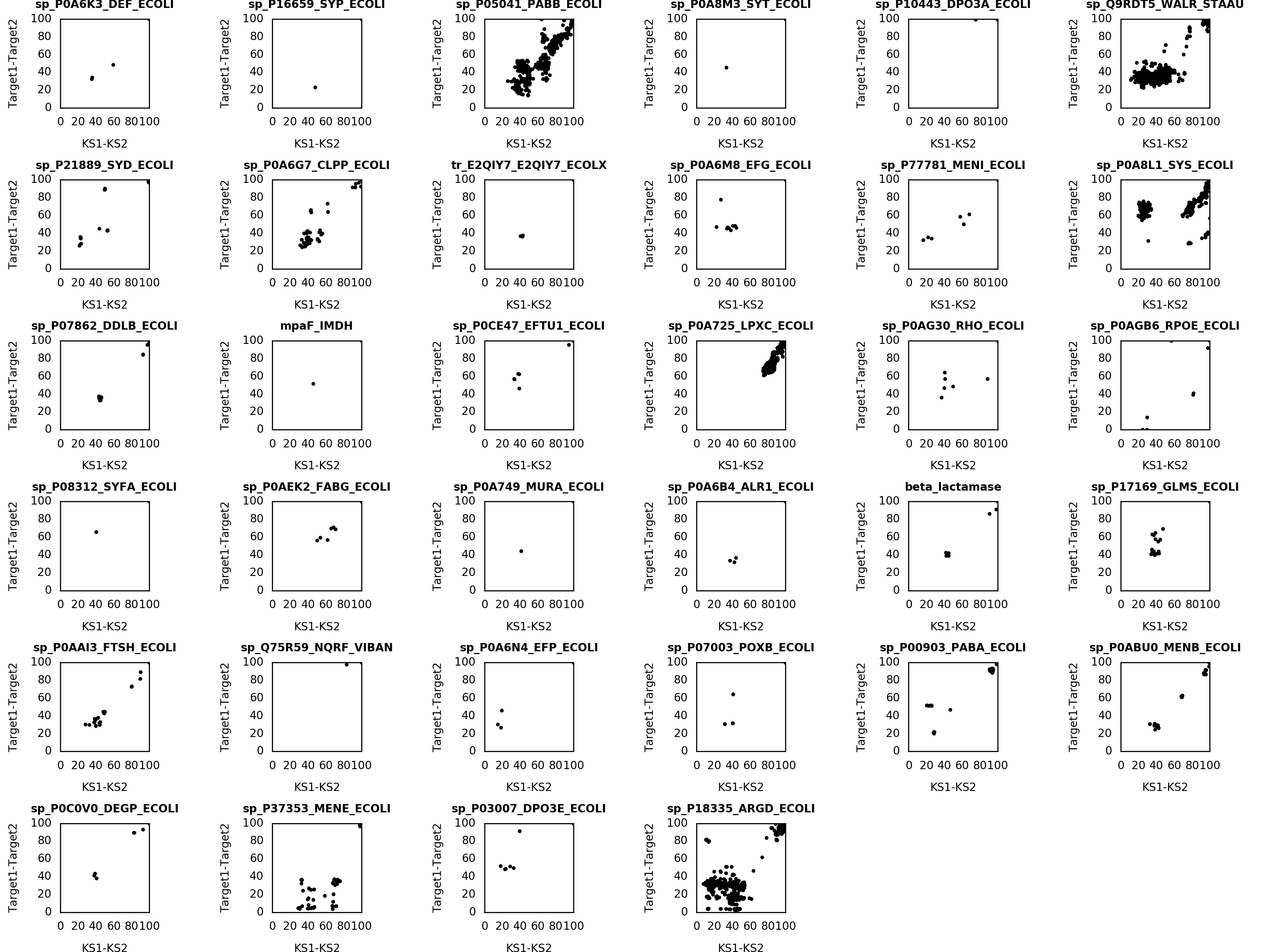
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3381196/>

Among potential targets, the enzymes involved in the pABA branch of the pathway are of great interest ([13](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3381196/#B13)). Indeed, they are absent in animals, and the only known metabolic fate of pABA is its commitment in folate synthesis.  Based on an automated screening to search for new inhibitors of folate biosynthesis, we found that rubreserine was able to inhibit the glutamine amidotransferase activity of the plant GAT-ADCS with an apparent IC50 of about 8 μm. The growth rates of *Arabidopsis thaliana*, *Toxoplasma gondii,*and *Plasmodium falciparum* were inhibited by rubreserine with respective IC50 values of 65, 20, and 1 μm.

Two main conclusions can be drawn from this study: first, our data validate for the first time the use of the bifunctional GAT-ADCS as an efficient drug target in eukaryotic cells, and second, we identified a new scaffold that inhibits plant growth and proliferation of apicomplexan parasites.

Menaquinone as a potential target of antibacterial agents.

<https://www.jstage.jst.go.jp/article/ddt/10/3/10_2016.01041/_pdf/-char/en>



Peptide deformylase

D-Ala-D-ala ligase alpha

Alanine racemase – 92.5kb vs 92.10kb – more points on diagonal

n-acetylglucosamine deacetylase

DNA-pol III subunit epsilon

Novel targets based on coevolution plot (E. coli)

Calculate the coevolution score for each target or the highest scoring ones

1. Dihidrolipoamide dehydrogenase –novel? regulation of raffinose transport

2. Lipoamide dehydrogenase is a component of the glycine cleavage system as well as of the alpha-ketoacid dehydrogenase complexes.

3. (3R)-hydroxymyristoyl-ACP dehydrase – lipid A biosynthesis (FabZ, mentioned above, phospholipids synthesis), novel target?

4. ACCA – ACCB known (andrimid

5. RNA 5-methylaminomethyl-2-thiouridylate methyltransferase - nvolved in the biosynthesis of a hypermodified uridine in the wobble position of tRNA – novel target?

6. UTP—glucose-1-phosphate uridylyltransferase (key player in [glycogenesis](https://en.wikipedia.org/wiki/Glycogenesis) and [cell wall](https://en.wikipedia.org/wiki/Cell_wall) synthesis)

7. hypothetical protein yeaN

8. Serine hydroxymethyltransferase - – antimalarial, anticancer, but no known antibiotic activity

(SHMT) is a [Pyridoxal phosphate](https://en.wikipedia.org/wiki/Pyridoxal_phosphate) (PLP) [(Vitamin B6)](https://en.wikipedia.org/wiki/Vitamin_B6) dependent enzyme ([EC](https://en.wikipedia.org/wiki/Enzyme_Commission_number)[2.1.2.1](https://enzyme.expasy.org/EC/2.1.2.1)) which plays an important role in cellular one-carbon pathways by catalyzing the reversible, simultaneous conversions of L-[serine](https://en.wikipedia.org/wiki/Serine) to [glycine](https://en.wikipedia.org/wiki/Glycine) and [tetrahydrofolate](https://en.wikipedia.org/wiki/Tetrahydrofolate) (THF) to [5,10-methylenetetrahydrofolate](https://en.wikipedia.org/wiki/5,10-methylenetetrahydrofolate) (5,10-CH2-THF).[[1]](https://en.wikipedia.org/wiki/Serine_hydroxymethyltransferase#cite_note-Raoetal2003-1)This reaction provides the largest part of the one-carbon units available to the cell.[[2]](https://en.wikipedia.org/wiki/Serine_hydroxymethyltransferase#cite_note-StoverSchirch1990-2)

9. Cell division ATP-binding protein FtsE - Part of the ABC transporter FtsEX involved in cellular division. Important for assembly or stability of the septal ring. Could be false positive if very similar to other pumps.

<https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full>

*The MacB architecture is also found in FtsEX which is required for efficient cell division in Gram-negative bacteria (*[*Schmidt et al., 2004*](https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full#B104)*;*[*Yang et al., 2011*](https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full#B136)*;*[*Du et al., 2016*](https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full#B25)*), sporulation in Bacillus (*[*Garti-Levi et al., 2008*](https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full#B30)*) and survival of mycobacteria (*[*Mavrici et al., 2014*](https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full#B75)*) and Streptococcus (*[*Sham et al., 2011*](https://www.frontiersin.org/articles/10.3389/fmicb.2018.00950/full#B106)*). In these organisms, the FtsEX complex is proposed to regulate the activity of extracytoplasmic cell wall amidases in the final stages of cell division. Inhibitors of MacAB have not been isolated, but screening of chemical libraries identified two different inhibitors that target the MacB homolog LolCDE*

*Indeed, targeting MacB-like ABC transporters for inhibition is particularly attractive, not only because FtsEX and LolCDE are essential, but also because this class of transporters is absent in humans.*

10. Phenylacetic degradation protein

11. Hypothetical protein yfch

12. Hypothetical protein ygaP?

13. Ala-tRNA synthetase

Octulosonate

deoxy d xylyose

L19

**Gram-negative bacteria inhibitors**

E. coli essential genes : targets for gram negative bacteria

targeting the cell wall in gram positive and negative bacteria:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5039524/>

**Metazoan clusters – probably pigments**

Metazoan and Fungal sequences

<https://www.ncbi.nlm.nih.gov/pubmed/17207587>

# A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs.

To our knowledge, no assembly-line PKS has been functionally characterized in eukaryotes. However, assembly-line PKSs are not thought to exist in metazoans. We were therefore surprised that the above analysis revealed the existence of an orphan clade that spanned a range of nematode species.

Parakeets PKS – pigment, but other functions in ther etazoans

<https://www.sciencedirect.com/science/article/pii/S0092867417309418#mmc2>

TODO: python seaborn