Establishing Synthesis Pathway-Host Compatibility via Enzyme Solubility

Sara A. Amin[1], Venkatesh Endalur Gopinarayanan[2], Nikhil U. Nair[2,*] and Soha Hassoun[1,2,*]

[1]Department of Computer Science, Tufts University, 161 College Avenue, Medford, MA, 02155

[2]Department of Chemical and Biological Engineering, 4 Colby Street, Medford, MA, 02155

*corresponding authors:

Science and Technology Center #276

4 Colby St, Medford, MA 02155

617-627-2582

nikhil.nair@tufts.edu

Halligan Hall #232

161 College Avenue, Medford, MA

617-627-5177

soha.hassoun@tufts.edu

Funding

This material is based upon work supported by the National Science Foundation under Grant No. CCF-1421972, and a Tufts Collaborates Grant.

Conflict of Interest: none declared.

Abstract

Current pathway synthesis tools identify possible pathways that can be added to a host to produce a desired target molecule through the exploration of abstract metabolic and reaction network space. However, not many of these tools do explore gene-level information required to physically realize the identified synthesis pathways, and none explore enzyme-host compatibility. Developing tools that address this disconnect between abstract reactions/metabolic design space and physical genetic sequence design space will enable expedited experimental efforts that avoid exploring unprofitable synthesis pathways. This work describes a workflow, termed Probabilistic Pathway Assembly with Solubility Scores (*ProPASS*), which links synthesis pathway construction with the exploration of the physical design space as imposed by the availability of enzymes with characterized activities within the host. Predicted protein solubility propensity scores are used as a confidence level to quantify the compatibility of each pathway enzyme with the host (*E. coli*). This work also presents a database, termed Protein Solubility Database (*ProSol DB*), which provides solubility confidence scores in *E. coli* for 240,016 characterized enzymes obtained from *UniProtKB/Swiss-Prot*. The utility of *ProPASS* is demonstrated by generating genetic implementations of heterologous synthesis pathways in *E. coli* that target several commercially useful biomolecules.

Keywords: metabolic engineering, synthetic biology, solubility, synthesis pathway, pathway implementation, pathway design

Availability: *ProSol DB* data and code for *ProPASS* are provided upon request.

Introduction

Synthesis pathways have been engineered within microbial hosts to produce commercially useful biomolecules including polyesters (Fidler & Dennis, 1992), building blocks for industrial polymers (Tong, Liao, & Cameron, 1991), biofuels (Nawabi, Bauer, Kyrpides, & Lykidis, 2011; Radakovits, Jinkerson, Darzins, & Posewitz, 2010; E. J. Steen et al., 2010), and therapeutic natural products derived from isoprenoids (Martin, Pitera, Withers, Newman, & Keasling, 2003; Pitera, Paddon, Newman, & Keasling, 2007; Watts, Mijts, & Schmidt-Dannert, 2005), polyketides (Peirú, Menzella, Rodríguez, Carney, & Gramajo, 2005; Pfeifer, Admiraal, Gramajo, Cane, & Khosla, 2001), and non-ribosomal peptides (Takahashi et al., 2007). Synthesis pathways consist of series of non-native enzyme-controlled reactions from metabolites within the host to a target molecule. Often, many possible biochemical routes to a target molecule exist such as in the production of 3-hydroxypropionic acid through glycerol, β-alanine, or acrolein pathways (Cheng, Jiang, Wu, Li, & Ye, 2016; Luo et al., 2016; Song, Kim, Cho, & Lee, 2016). Further, the genetic implementation of each pathway is not unique since genes can be sourced from various organisms. For example, production of thebaine and hydrocodone in *S. cerevisiae* required construction of a heterologous pathway with enzymes from a bacterium (*Pseudomonas putida*), a mammal (*Rattus norvegicus*), and several plants (*Papaver somniferum*, *Papaver bracteatum*, *Coptis japonica*, *Eschscholzia californica*) (Galanie, Thodey, Trenchard, Interrante, & Smolke, 2015).

Needless to state, experimental efforts to explore all possible pathways and all possible organism-specific enzyme selections are prohibitive. Unfortunately, current pathway synthesis computational tools provide solutions based on exploring the abstract metabolic space without regards to possible genetic implementation. Identifying best genetic implementations are

typically ad hoc and there is a lack of systematic design tools and workflows that support the co-exploration of the metabolic and genetic implementation design spaces. Pathway synthesis tools that consider genetic compatibility between biosynthetic pathways and heterologous hosts can significantly expedite the metabolic engineering cycle by focusing on design spaces with a greater fraction of profitable options.

In this light, *pathway synthesis tools* must conceptually perform two distinct tasks. The first task involves identifying possible synthesis pathways from the host to a target molecule and evaluating potential yield. This is known as the *pathway construction* or *pathway identification* problem, where sequences of reactions that synthesize the target molecule from a host metabolite are selected based on data from multi-organism databases such as KEGG (Kanehisa & Goto, 2000), MetaCyc (Caspi et al., 2008), and SEED (Overbeek et al., 2005). Pathway construction utilizes either rule-based techniques (e.g. BNICE (Wu, Wang, Assary, Broadbelt, & Krilov, 2011), PathPred (Moriya et al., 2010), Meta (Klopman, Dimayuga, & Talafous, 1994; Klopman, Tu, & Talafous, 1997), Meteor (Greene, Judson, Langowski, & Marchant, 1999; Marchant, Briggs, & Long, 2008), and UM-PPS (Ellis & Wackett, 2012; Hou, Wackett, & Ellis, 2003)) or graph-based techniques such as *ProPath* (Yousofshahi, Lee, & Hassoun, 2011), DESHARKY (Rodrigo, Carrera, Prather, & Jaramillo, 2008), Metaroute (Blum & Kohlbacher, 2008). Thus, there are currently multiple tools available for pathway construction.

The second synthesis task concerns the *parts selection problem* or, more broadly, the *pathway implementation* problem. This task involves identifying gene sequences from specific organisms that best realize reactions along the synthesis pathway. There are several factors to consider. One set of factors is associated with transcription and translation, including promoter strength and efficiency of ribosome binding. Another set of factors is dependent on the

interactions of the enzyme's protein coding sequence with the cellular metabolic machinery to ensure correct conformation and function while avoiding misfolding or aggregation, known as protein solubility. Overcoming protein solubility issues can be achieved by codon optimization, co-expressing molecular chaperones (Trésaugues et al., 2004; Xia et al., 2016), lowering culture temperature (Makrides, 1996) or modifying growth media (Makrides, 1996). These strategies can improve solubility by finding optimal (re-)folding conditions, but may also lead to activity loss of a protein (Singh & Panda, 2005). Further, when multiple enzymes in a pathway are insoluble, a single strategy may not be applicable. It is therefore desirable to select high-solubility enzymes to implement synthesis pathways.

To identify profitable synthesis pathway designs for experimental implementation, we couple in this paper the two synthesis tasks: *pathway construction* and *pathway implementation*. Using sequences from *UniProtKB* database (UniProt Consortium, 2017), we assemble a Protein Solubility Database (*ProSol DB*), that comprises confidence scores for the likelihood of an enzyme being soluble in *E. coli* and also allows for quick lookup using Enzyme Commission (EC) numbers. We used *ccSOL omics* (Agostini, Cirillo, Livi, Ponti, & Tartaglia, 2014) to compute solubility propensity scores for various proteins from *UniProtKB* in *E. coli.* Our database stores solubility confidence scores for 240,016 sequences, of which 34,046 sequences are associated with commonly used organisms. We develop a new workflow, termed Probabilistic Pathway Assembly with Solubility Scores (*ProPASS*), to couple *ProPath* (Yousofshahi et al., 2011), a method for constructing synthesis pathways, and *ProSol DB*. The workflow consists of first identifying synthesis pathways from a host to a target metabolite. Pathways are then ranked based on their length and yield. *ProPASS* then recommends sequences based on their solubility confidence scores in *ProSol DB* to implement reactions along each

pathway. We applied our workflow to identify synthesis pathways for several target molecules and analyzed three test cases in details. As far as we know, this is the first description of a workflow or method that explores using gene-specific sequence information to systematically identify host-compatible enzymes that can be used to implement synthesis pathways.

Methods

Assembling the Protein Solubility Database (*ProSol DB*)

To create a database of solubility confidence scores for various enzymes, we utilize annotated sequences in the *UniProtKB* database (UniProt Consortium, 2017). Protein sequences that have either experimentally validated or manually analyzed are considered "reviewed" within the *UniProtKB/Swiss-Prot* database. While *UniProtKB/Swiss-Prot* currently contains over 550,000 sequences, only 240,016 sequences are associated with 4,652 Enzyme Commission (EC) numbers. *ProSol DB* is designed such that the EC numbers serve as key for database lookups, and the returned value contains the *UniProtKB/Swiss-Prot* sequence IDs, solubility confidence scores, and associated organisms. Solubility confidence scores are computed only once, using *ccSOL omics*, and stored locally in the database, eliminating the need to predict scores repeatedly, thus speeding up enzyme selection.

ProPASS: Coupling probabilistic pathway construction with protein solubility prediction

With the ability to lookup protein solubility confidence scores in *ProSol DB* based on EC numbers, we developed *ProPASS*, a workflow that couples pathway construction with identification of organism-specific enzymes for each step. The workflow (**Figure 1**) consists of a design step followed by an implementation step. Additional details can be found in **Supplementary File 1**.

Step 1 – Synthesis pathway design.

Step1a. Synthesis pathway identification. Given a target molecule, *ProPath* synthesizes pathways. *ProPath,* a graph-based probabilistic search algorithm, probabilistically selects synthesis pathways, from a target molecule terminating at a molecule within the host, using all reactions extracted from the KEGG database. We selected *ProPath* for pathway construction as it is computationally efficient and was shown effective in generating synthesis pathways with yield distributions similar to those obtained using limited-in-depth exhaustive search. Additionally, *ProPath* reproduces validated pathways published in the literature.

Step 1b. Synthesis pathway ranking. Identified pathways are ranked by two metrics: pathway yield and pathway length. These metrics are reported to users to allow the examination of different designs and their tradeoffs when identifying suitable pathway implementations for experimenters.

Step 2 – Synthesis pathway implementation. Given a synthesis pathway, the EC number for each of its enzymatic reaction is utilized to look up the associated *UniProtKB/Swiss-Prot* protein IDs in *ProSol DB*. The retrieved data also contains solubility confidence scores and organisms associated with each protein ID. The scores provide a *confidence level* that helps determine if a protein is soluble in *E. coli*. The scores should not be taken as any measure of relative solubility of a given enzyme in *E. coli*. In the case of non-matching protein IDs in *ProSol DB*, we allow the option to retrieve non-reviewed sequences from *UniProtKB*, and solubility confidence score for each sequence using *ccSOL omics* on ad hoc manner. Each reaction will have one or more recommended implementation sequence, along with its solubility confidence score and source organism.

Results

Overview of Protein Solubility Database (*ProSol DB*)

ProSol DB is considered the primary source of solubility confidence scores for *ProPASS* and consists of 240,016 sequences. *ProSol DB* is built using all reviewed protein sequences in *UniProtKB/Swiss-Prot*. The protein sequences in *ProSol DB* are associated with 4,652 EC numbers out of 6,896 EC numbers reported in KEGG (**Figure 2a**). We divide protein sequences organisms into two sets. The first set consists of 20 commonly used and well-studied organisms (e.g. *E. coli, S. cerevisiae, B. subtilis*, etc.) from which protein sequences can be easily acquired for experimental validation (A-list) (**Supplementary File 2**). The second set, the B-list, contains all other organisms associated with the remaining protein sequences associated with EC numbers. Protein sequences associated with the A-list organisms cover 2,427, or 52%, of all unique EC numbers available in *ProSol DB*, and 13% of all sequences in the database even though it consists of only 0.33% of the organismal diversity (**Figure 2b**). Example of solubility confidence scores present in our database can be found in **Table I**. The first column lists the EC number. The second column shows the associated *UniProtKB/Swiss-Prot* IDs for protein sequences associated with the EC number in the first column. The sequences are identified by their *UniProtKB/Swiss-Prot* IDs to make it easier to obtain any information related to these sequences from *UniProtKB/Swiss-Prot*. The third column contains a solubility confidence score for each protein sequence. The fourth and fifth columns show organisms associated with each protein sequence. When an EC number is queried, a list of protein IDs associated with the A-list organisms is returned. If none are associated with the A-list, a list of protein IDs associated with the B-list is returned. Further examples of solubility confidence scores within *ProSol DB* can be found in **Supplementary File 3.**

Since *ProSol DB* is the largest database of solubility confidence scores in *E. coli*, we used it as a resource to investigate any potential correlation between propensity for soluble expression and EC numbers (**Figure 3**). For this analysis, we only considered EC numbers that are associated with 10 or more protein sequences in *ProSol DB*. The average predicted solubility confidence score for the 1,600 enzymes that meet this criterion is 54.3 (**Figure 3a**). For each EC number, the range of solubility confidence scores (solubility range) varied tremendously (**Figure 3b**), with over 96% of ECs having scores that varied by a difference of 75 or more. Analyzing the standard deviation of average solubility confidence score per EC number shows that they are also spread over similar wide ranges (**Figure 3c**). There was low correlation between the score ranges and average solubility confidence scores (**Figure 3d**) as analyzed using Spearman's Rank Test (Spearman's rank coefficient = 0.00897)

Further, we investigated the correlation between solubility confidence scores and the similarity of the source organisms to *E. coli* (**Figure 4**). The phylogenetic distance from *E. coli* was computed using phyloT, a tool for the annotation and visualization of phylogenetic trees (Letunic & Bork, 2016). We only considered organisms that had at least 100 solubility confidence scores in *ProSol DB*. The distances varied between 0 and 40 (**Figure 4a**), where a score of 0 indicated a specific strain of *E. coli*, and a score of 40 indicated a very distant species such as *Drosophila melanogaster* (fruit fly). The average solubility confidence score per organism was 53.43 (**Figure 4b**). The correlation between the phylogenetic distance and average solubility confidence score per organism (**Figure 4c**) showed low positive correlation as calculated using Spearman's Rank Test (Spearman's rank coefficient = 0.176919).

Validating *ProPASS* workflow

We used the *ProPASS* workflow to identify synthesis pathways and their potential implementation for several commercially useful biomolecules. The targets include industrial precursors used in the production of plastics, printing ink, pharmaceuticals (2,3-butanediol (Cho et al., 2015), *cis*-muconic acid (Weber et al., 2012)), biofuels (triacylglycerols (Radakovits et al., 2010) and fatty acid methyl esters (Nawabi et al., 2011)), a commodity chemical that is widely used in industry (3-hydroxypropanoic acid (Della Pina, Falletta, & Rossi, 2011)) and biological precursors (isopentenyl diphosphate (Kim & Keasling, 2001) and *myo*-inositol (Shiue & Prather, 2014)). The test cases show that 20% − 100% of the identified pathways have solubility confidence scores for enzymes catalyzing all their reactions from the reviewed sequences present in *ProSol DB*. (**Table II**). Three of the seven test cases were analyzed in detail of which two are presented below (3-hydroxypropanoic acid and 2,3-butanediol), and one in **Supplementary Files 4 and 5** (*cis*-muconic acid). In all cases, we constructed the pathway using *ProPath* and ranked the designs based on yield and length, and then explored the different implementation options as provided by the solubility confidence scores.

3-Hydroxypropanoic acid

We applied *ProPASS* to synthesize 3-hydroxypropanoic acid (3-HP). Since 3-HP production through the *pdu* pathway is the only known pathway from glycerol, we only explored those originating from glucose and identified 112 synthesis pathways (**Figure 5**). Pathways lengths varied from 2 to 14 and yield values varied from 0.271 to 0.41 mmol/mmol of glucose/g CDW/h. The identified pathways were ranked based on the two design criteria: pathway yield and pathway length. Next, we identified implementations for each of the top 5 ranked pathways in each category. We then compare the identified pathways with those published in the literature and contrast their implementations with those recommended by *ProPASS.*

When ranked based on yield (**Table IIIa**), the top pathway had five steps and yield of 0.408 mmol/mmol of glucose/g CDW/h. However, two of the five reactions had no solubility data in *ProSol DB*. Other pathways in the five top-ranked pathways had three or five steps. Two of the three-step pathways had a yield value of 0.407 mmol/mmol of glucose/g CDW/h with complete solubility confidence scores.

When ranked based on length (**Table IIIb**), the top five had only 2 reaction steps – from malonate semialdehyde to 3-HP. All five pathways had similar yields (0.365 – 0.373 mmol/mmol of glucose/g CDW/h); however, only the pathway that produced malonate semialdehyde through the β-alanine pathway had complete solubility data.

We compared the predicted solubility confidence scores of the enzymes used in the published literature to the highest solubility confidence score alternatives identified by *ProPASS* (**Figure 6**). Of the identified pathways, three have been experimentally used previously for 3-HP production either through the propionyl-CoA pathway (**Figure 6a**) (Luo et al., 2016), the malonyl-CoA pathway (**Figure 6b**) (Cheng et al., 2016), or β-alanine pathway (**Figure 6c**) (Song et al., 2016). In the case of the enzyme malonyl-CoA reductase, in the malonyl-CoA pathway (**Figure 6b**), *ProPASS* identified a solubility confidence score of 74, while the corresponding enzyme from Cheng et al. (Cheng et al., 2016) has a score of 97. *UniProtKB* did not catalogue the sequences used by Cheng et al., which precluded their enzyme from the *ProPASS* workflow results. However, our workflow identified sequences with high or similar solubility confidence scores when compared with sequences used in the literature for 3-HP synthesis through the propionyl-CoA and the β-alanine pathways.

2,3-Butanediol

ProPASS identified only one 2-step synthesis pathway that produces 2,3-butanediol using the (*R,R*)-butanediol dehydrogenase and acetolactate decarboxylase enzymes (**Table IV**). The highest solubility confidence scores for sequences identified by *ProPASS* from the A-list organisms were 87 and 83, respectively. Expanding the search to include sequences from the B-list did not yield higher scores. Further expanding the search to include reviewed and non-reviewed sequences from *UniProtKB* identified sequences with solubility confidence scores of 100 and 99, respectively. We compared our findings with those in the literature and found that this 2-step reaction is the only known synthesis pathway for 2,3-butanediol. Further, the selected enzymes match an experimentally validated pathway by Yan et al. (Yan, Lee, & Liao, 2009), where enzyme sequences were selected from two A-list organisms, *B. subtilis* and *Klebsiella pneumonia.* Both sequences have high-predicted solubility scores, comparable to those identified in the non-reviewed sequences. As these sequences were not catalogued in *UnitProtKB/Swiss-Prot*, *ProPASS* did not identify them as possible sequence options.

Discussion

Productive integration of heterologous pathways into model host organisms is vital for metabolic engineering. Key design steps where computational tools can significantly expedite the engineering cycle are *pathway construction* to identify reactions required to connect the host to a target metabolite, and *parts selection* to choose specific biochemical parts (such as enzymes, promoters, etc.) to express in the host. While current pathway construction tools can identify high-yielding pathways, they cannot be directly translated into experiments without the parts selection step. Next-generation synthesis tools must provide detailed biochemical and biophysical information about the biological parts that can be directly translated and integrated

into the engineering design-build-test-learn cycle. Such information includes, but is not limited to, the sequence of heterologous enzymes, their activity, solubility, stability, etc.

Aiming to create a next-generation synthesis tool, we expanded in this work the scope of conventional pathway construction to identify potential sources from which each of the pathway enzymes can be isolated using protein solubility as a criterion of parts selection and host-compatibility. A specific amino acid sequence of enzymes in pathway determines many outcomes such as kinetics, stability, and actual yield. Solubility however is a barrier in achieving these outcomes. When a recombinant enzyme is expressed as an insoluble aggregate (known as inclusion bodies), they retain none or a fraction of the catalytic activity, except in a few cases (Choi et al., 2011; Lin, Zhou, Wu, Xing, & Zhao, 2013; Nahalka & Nidetzky, 2007; Zhou, Xing, Wu, Zhang, & Lin, 2012). An expressed recombinant enzyme therefore must be soluble in the heterologous host to ensure product yield and titer.

Current metabolic engineering practices utilize an ad hoc approach to enzyme sourcing based on domain knowledge or practical considerations such as accessibility to organisms from which enzymes can be sourced. For example, *n*-butanol synthesis pathway is often sourced entirely from natural producers, *Clostridia* (Inui et al., 2008; Nielsen et al., 2009). However, when solubility and expression of heterologous enzymes (i.e., host compatibility) are taken into consideration, several groups have found that a "mix-and-match" approach is often more profitable than sourcing all enzymes from a single organism, as was demonstrated when implementing the *n*-butanol pathway in yeast (Eric J Steen et al., 2008) and *E. coli* (Bond-Watts, Bellerose, & Chang, 2011). More drastic examples of such sourcing paradigm were demonstrated in by both Keasling and Smolke groups when synthesizing complex natural products (Galanie et al., 2015; Ro et al., 2006). Assembly of such "mix-and-match" pathways

has been significantly aided by decreasing costs of gene synthesis. However, such empowerment also leads to choice overload. We posit that host-compatibility in the form of solubility can provide a powerful guide to prune down the options to most profitable choices. Enzyme engineering can be used strategy to overcome enzyme-host incompatibility. It can, however, be time-consuming, even when aided by computational protein engineering tools (e.g. (Hassanpour, Ullah, Yousofshahi, Nair, & Hassoun, 2017; Sachsenhauser & Bardwell, 2018)), especially if several enzymes within a pathway are insoluble. Thus, implementing pathways with knowledge that the enzymes selected have a high propensity for solubility in the host can reduce the traditional trial-and-error approach to enzyme selection.

Therefore, protein solubility is used as a rating and/or selection criterion before the implementation phase. In most cases, multiple implementation options based on solubility confidence scores are available to choose from. Since solubility confidence scores do not affect maximum theoretical yield, they do not provide a ranking criterion for pathway design. Rather, they provide a guideline for the implementation phase. Thus, *ProPASS* advances the state of pathway synthesis by exploring the underlying biophysical design space, leading to more profitable design implementation options when compared to using pathway synthesis tools that only explore the abstract metabolic space. Such incorporation of sequence information is distinct from previous descriptions where protein sequence and structure information has been used to inform putative promiscuous enzymatic activity to assemble non-natural synthetic pathways (Brunk, Neri, Tavernelli, Hatzimanikatis, & Rothlisberger, 2012; Erb, Jones, & Bar-Even, 2017).

To allow for enzyme selection, we created a database of predicted solubility confidence scores, *ProSol DB*, for reviewed protein sequences from the *UniProtKB/Swiss-Prot* database. *ProSol DB* contains 240,016 sequences and links EC numbers with protein sequences, their

source organism, and their solubility confidence scores in *E. coli*. The solubility confidence scores were calculated using *ccSOL omics*, with reported accuracy comparable with other available solubility prediction tools. Future improvements in solubility prediction algorithms can be used to update *ProSol DB*, thus providing more confidence regarding enzyme selection. Analysis of solubility confidence scores and enzymes revealed that sequences associated with a particular enzyme have a wide range of associated scores. Further analysis showed no correlation between phylogenetic distance from *E. coli* and solubility confidence scores. Our findings emphasize that solubility is a property of the encoding sequence, and not a function of the host organism or associated EC number. As far as we know, *ProSol DB* is the largest database of enzyme solubility confidence scores for heterologous enzymes in *E. coli* and serves as a resource for pathway synthesis tools as well as for meta-analysis of sequence independent factors that affect solubility of enzymes in *E. coli*.

We validated the *ProPASS* workflow by comparing predicted pathway construction and implementations with those published in literature for several target molecules. In all cases, we identified a significant number of pathway implementations for which solubility confidence scores were available through *ProSol DB*. The availability of data was dependent on the studied target molecule. In the case of *myo*-inositol, 61% of identified pathways had confidence scores for each enzyme, while in the case of *cis*-muconic acid only 20% of identified pathways had such solubility data. Several observations can be drawn based on examining the three detailed test cases. When considering solubility, longer pathways have a higher chance of having enzymes with low solubility confidence scores. For example, the *cis*-muconic acid synthesis pathway with the highest yield has a pathway length of 10, with one of the steps having a low solubility confidence score of 22. In contrast, the pathway with the highest confidence of having

a soluble enzyme for each step is a three-step pathway but has only half the yield of the highest-yield pathway. *ProPASS* can provide several implementation alternatives, from which one or more pathways can be selected for experimental testing. When the solubility confidence score for a given enzyme in a given pathways was low, it was possible to utilize non-reviewed sequences in *UniProtKB/Swiss-Prot* to identify higher score alternatives. This was the case when synthesizing 2,3-butandiol. As more reviewed sequences are added to *UniProtKB,* the reliance on non-reviewed sequences will diminish. Importantly, *ProPASS* can provide alternative implementations that would not have been easily discoverable otherwise. All test cases discussed here are limited to using *E. coli* as host. However, development of solubility prediction tools for other organisms would allow extending *ProPASS* to other hosts.

Strain optimization, the process by which opportunities for metabolic re-direction are identified to ensure maximal target formation, is a crucial step that ensures the profitable integration of heterologous synthesis pathways into model host organisms. While several computational tools (e.g., (Burgard, Pharkya, & Maranas, 2003), (Patil, Rocha, Förster, & Nielsen, 2005), and (Ranganathan, Suthers, & Maranas, 2010)) were proposed to identify the most profitable modifications, almost none take into account implementation concerns during optimization. One exception is *CCOpt* (Yousofshahi, Orshansky, Lee, & Hassoun, 2013), which addresses the uncertainty in precise tuning of enzyme levels and imposes probabilistic flux capacity constraints that capture the uncertainty in tuning enzyme levels during implementation. In this work, we treated the two design steps (pathway synthesis and strain optimization) as independent. A non-competitive (in terms of yield) synthesis pathway during the pathway synthesis step may provide higher yields than other synthesis options once strain optimization is performed. When developed judiciously to limit the design space, co-optimization tools that are

aware of design choices associated with the underlying implementation can potentially yield more profitable experimental guidance.

Acknowledgements

We thank Emily Chicklis, an REU student, for her help with some figure illustrations.

References

Agostini, F., Cirillo, D., Livi, C. M., Ponti, R. D., & Tartaglia, G. G. (2014). *cc*SOL *omics*: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics, 30*(20), 2975-2977.

Blum, T., & Kohlbacher, O. (2008). MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics, 24*(18), 2108-2109.

Bond-Watts, B. B., Bellerose, R. J., & Chang, M. C. (2011). Enzyme mechanism as a kinetic control element for designing synthetic biofuel pathways. *Nature chemical biology, 7*(4), 222-227.

Brunk, E., Neri, M., Tavernelli, I., Hatzimanikatis, V., & Rothlisberger, U. (2012). Integrating computational methods to retrofit enzymes to synthetic pathways. *Biotechnol Bioeng, 109*(2), 572-582. doi:10.1002/bit.23334

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering, 84*(6), 647-657.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., . . . Tissier, C. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research, 36*(suppl 1), D623-D631.

Cheng, Z., Jiang, J., Wu, H., Li, Z., & Ye, Q. (2016). Enhanced production of 3-hydroxypropionic acid from glucose via malonyl-CoA pathway by engineered *Escherichia coli*. *Bioresource technology, 200*, 897-904.

Cho, S., Kim, T., Woo, H. M., Kim, Y., Lee, J., & Um, Y. (2015). High production of 2, 3-butanediol from biodiesel-derived crude glycerol by metabolically engineered *Klebsiella oxytoca* M1. *Biotechnology for biofuels, 8*(1), 1.

Choi, S.-L., Lee, S. J., Ha, J.-S., Song, J. J., Rhee, Y. H., & Lee, S.-G. (2011). Generation of catalytic protein particles in *Escherichia coli* cells using the cellulose-binding domain from *Cellulomonas fimi* as a fusion partner. *Biotechnology and Bioprocess Engineering, 16*(6), 1173-1179.

Della Pina, C., Falletta, E., & Rossi, M. (2011). A green approach to chemical building blocks. The case of 3-hydroxypropanoic acid. *Green chemistry, 13*(7), 1624-1632.

Ellis, L. B., & Wackett, L. P. (2012). Use of the University of Minnesota Biocatalysis/Biodegradation Database for study of microbial degradation. *Microbial informatics and experimentation, 2*(1), 1.

Erb, T. J., Jones, P. R., & Bar-Even, A. (2017). Synthetic metabolism: metabolic engineering meets enzyme design. *Curr Opin Chem Biol, 37*, 56-62. doi:10.1016/j.cbpa.2016.12.023

Fidler, S., & Dennis, D. (1992). Polyhydroxyalkanoate production in recombinant *Escherichia coli*. *FEMS microbiology reviews, 9*(2-4), 231-235.

Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F., & Smolke, C. D. (2015). Complete biosynthesis of opioids in yeast. *Science, 349*(6252), 1095-1100.

Greene, N., Judson, P. N., Langowski, J. J., & Marchant, C. A. (1999). Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR and QSAR in environmental research, 10*(2-3), 299. doi:10.1080/10629369908039182

Hassanpour, N., Ullah, E., Yousofshahi, M., Nair, N. U., & Hassoun, S. (2017). Selection Finder (SelFi): A computational metabolic engineering tool to enable directed evolution of enzymes. *Metab Eng Commun, 4*, 37-47. doi:10.1016/j.meteno.2017.02.003

Hou, B. K., Wackett, L. P., & Ellis, L. B. M. (2003). Microbial Pathway Prediction: A Functional Group Approach. *Journal of chemical information and computer sciences, 43*(3), 1051-1057. doi:10.1021/ci034018f

Inui, M., Suda, M., Kimura, S., Yasuda, K., Suzuki, H., Toda, H., . . . Yukawa, H. (2008). Expression of Clostridium acetobutylicum butanol synthetic genes in Escherichia coli. *Applied microbiology and biotechnology, 77*(6), 1305-1316.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 28*(1), 27-30.

Kim, S. W., & Keasling, J. (2001). Metabolic engineering of the nonmevalonate isopentenyl diphosphate synthesis pathway in *Escherichia coli* enhances lycopene production. *Biotechnology and bioengineering, 72*(4), 408-415.

Klopman, G., Dimayuga, M., & Talafous, J. (1994). META. 1. A Program for the Evaluation of Metabolic Transformation of Chemicals. *Journal of chemical information and computer sciences, 34*(6), 1320. doi:10.1021/ci00022a014

Klopman, G., Tu, M., & Talafous, J. (1997). META. 3. A Genetic Algorithm for Metabolic Transform Priorities Optimization. *Journal of chemical information and computer sciences, 37*(2), 329. doi:10.1021/ci9601123

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res, 44*(W1), W242-245. doi:10.1093/nar/gkw290

Lin, Z., Zhou, B., Wu, W., Xing, L., & Zhao, Q. (2013). Self-assembling amphipathic alpha-helical peptides induce the formation of active protein aggregates *in vivo*. *Faraday discussions, 166*, 243-256.

Luo, H., Zhou, D., Liu, X., Nie, Z., Quiroga-Sánchez, D. L., & Chang, Y. (2016). Production of 3-Hydroxypropionic Acid via the Propionyl-CoA Pathway Using Recombinant *Escherichia coli* Strains. *PloS one, 11*(5), e0156286.

Makrides, S. C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiological reviews, 60*(3), 512-538.

Marchant, C. A., Briggs, K. A., & Long, A. (2008). *In Silico* Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicology Mechanisms and Methods, 18*(2-3), 177. doi:10.1080/15376510701857320

Martin, V. J. J., Pitera, D. J., Withers, S. T., Newman, J. D., & Keasling, J. D. (2003). Engineering a mevalonate pathway in *Escherichia coli* for the production of terpenoids. *Nat Biotech, 21*(7), 796-802.

Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., & Kanehisa, M. (2010). PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Research*, gkq318.

Nahalka, J., & Nidetzky, B. (2007). Fusion to a pull-down domain: a novel approach of producing *Trigonopsis variabilis* D-amino acid oxidase as insoluble enzyme aggregates. *Biotechnology and bioengineering, 97*(3), 454-461.

Nawabi, P., Bauer, S., Kyrpides, N., & Lykidis, A. (2011). Engineering *E. coli* for biodiesel production utilizing a bacterial fatty acid methyltransferase. *Applied and environmental microbiology*, AEM. 05046-05011.

Nielsen, D. R., Leonard, E., Yoon, S.-H., Tseng, H.-C., Yuan, C., & Prather, K. L. J. (2009). Engineering alternative butanol production platforms in heterologous bacteria. *Metabolic Engineering, 11*(4-5), 262-273.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., . . . Edwards, R. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research, 33*(17), 5691-5702.

Patil, K. R., Rocha, I., Förster, J., & Nielsen, J. (2005). Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC bioinformatics, 6*(1), 308.

Peirú, S., Menzella, H. G., Rodríguez, E., Carney, J., & Gramajo, H. (2005). Production of the Potent Antibacterial Polyketide Erythromycin C in *Escherichia coli*. *Applied and environmental microbiology, 71*(5), 2539-2547.

Pfeifer, B. A., Admiraal, S. J., Gramajo, H., Cane, D. E., & Khosla, C. (2001). Biosynthesis of Complex Polyketides in a Metabolically Engineered Strain of *E. coli*. *Science, 291*(5509), 1790-1792.

Pitera, D. J., Paddon, C. J., Newman, J. D., & Keasling, J. D. (2007). Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. *Metabolic Engineering, 9*(2), 193-207.

Radakovits, R., Jinkerson, R. E., Darzins, A., & Posewitz, M. C. (2010). Genetic engineering of algae for enhanced biofuel production. *Eukaryotic cell, 9*(4), 486-501.

Ranganathan, S., Suthers, P. F., & Maranas, C. D. (2010). OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput Biol, 6*(4), e1000744.

Ro, D.-K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., . . . Kirby, J. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature, 440*(7086), 940.

Rodrigo, G., Carrera, J., Prather, K. J., & Jaramillo, A. (2008). DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics, 24*(21), 2554-2556.

Sachsenhauser, V., & Bardwell, J. C. (2018). Directed evolution to improve protein folding in vivo. *Current opinion in structural biology, 48*, 117-123.

Shiue, E., & Prather, K. L. (2014). Improving D-glucaric acid production from *myo*-inositol in *E. coli* by increasing MIOX stability and *myo*-inositol transport. *Metabolic Engineering, 22*, 22-31.

Singh, S. M., & Panda, A. K. (2005). Solubilization and refolding of bacterial inclusion body proteins. *Journal of bioscience and bioengineering, 99*(4), 303-310.

Song, C. W., Kim, J. W., Cho, I. J., & Lee, S. Y. (2016). Metabolic Engineering of *Escherichia coli* for the Production of 3-Hydroxypropionic Acid and Malonic Acid through β-Alanine Route. *ACS synthetic biology*.

Steen, E. J., Chan, R., Prasad, N., Myers, S., Petzold, C. J., Redding, A., . . . Keasling, J. D. (2008). Metabolic engineering of Saccharomyces cerevisiae for the production of n-butanol. *Microbial cell factories, 7*(1), 36.

Steen, E. J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., . . . Keasling, J. D. (2010). Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature, 463*, 559-562.

Takahashi, H., Kumagai, T., Kitani, K., Mori, M., Matoba, Y., & Sugiyama, M. (2007). Cloning and characterization of a *Streptomyces* single module type non-ribosomal peptide synthetase catalyzing a blue pigment synthesis. *Journal of Biological Chemistry, 282*(12), 9073-9081.

Tong, I.-T., Liao, H. H., & Cameron, D. C. (1991). 1, 3-Propanediol production by *Escherichia coli* expressing genes from the *Klebsiella pneumoniae dha* regulon. *Applied and environmental microbiology, 57*(12), 3541-3546.

Trésaugues, L., Collinet, B., Minard, P., Henckes, G., Aufrère, R., Blondeau, K., . . . van Tilbeurgh, H. (2004). Refolding strategies from inclusion bodies in a structural genomics project. *Journal of structural and functional genomics, 5*(3), 195-204.

UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 45*(D1), D158-D169.

Watts, K. T., Mijts, B. N., & Schmidt-Dannert, C. (2005). Current and Emerging Approaches for Natural Product Biosynthesis in Microbial Cells. *Advanced Synthesis & Catalysis, 347*(7-8), 927-940. doi:10.1002/adsc.200505062

Weber, C., Brückner, C., Weinreb, S., Lehr, C., Essl, C., & Boles, E. (2012). Biosynthesis of *cis, cis*-Muconic acid and its aromatic precursors, catechol and protocatechuic acid, from renewable feedstocks by *Saccharomyces cerevisiae*. *Applied and environmental microbiology, 78*(23), 8421-8430.

Wu, D., Wang, Q., Assary, R. S., Broadbelt, L. J., & Krilov, G. (2011). A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *Journal of chemical information and modeling, 51*(7), 1634-1647.

Xia, P. F., Zhang, G. C., Liu, J. J., Kwak, S., Tsai, C. S., Kong, I. I., . . . Jin, Y. S. (2016). GroE chaperonins assisted functional expression of bacterial enzymes in *Saccharomyces cerevisiae*. *Biotechnology and bioengineering*.

Yan, Y., Lee, C.-C., & Liao, J. C. (2009). Enantioselective synthesis of pure (*R, R*)-2, 3-butanediol in *Escherichia coli* with stereospecific secondary alcohol dehydrogenases. *Organic & biomolecular chemistry, 7*(19), 3914-3917.

Yousofshahi, M., Lee, K., & Hassoun, S. (2011). Probabilistic pathway construction. *Metabolic Engineering, 13*(4), 435-444.

Yousofshahi, M., Orshansky, M., Lee, K., & Hassoun, S. (2013). Probabilistic strain optimization under constraint uncertainty. *BMC Syst Biol, 7*, 29. doi:10.1186/1752-0509-7-29

Zhou, B., Xing, L., Wu, W., Zhang, X.-E., & Lin, Z. (2012). Small surfactant-like peptides can drive soluble proteins into active aggregates. *Microbial cell factories, 11*(1), 10.

Tables:

Table I: Sample from the *ProSol DB* for sequences associated with the acetolactate decarboxylase enzyme.

KEY	VALUE			
EC #	**UniProt ID**	**Solubility confidence score**	**A-List organisms**	**B-List organisms**
4.1.1.5 (acetolactate decarboxylase)	P77880	89	*Lactococcus lactis*	–
	Q04518	6	N/A	*Raoultella terrigena*
	Q8PZ55	32	N/A	*Methanosarcina mazei*
	Q65E52	96	N/A	*Bacillus licheniformis*
	P95676	83	*Lactococcus lactis*	–
	Q04777	96	*Bacillus subtilis*	–
	P23616	34	N/A	*Brevibacillus brevis*
	Q8L208	68	N/A	*Streptococcus thermophilus*

N/A, no enzyme with corresponding EC# found in A-list,
–, EC# not queried in B-list since an A-list sequence is found.

Table II: A list of commercially useful molecules used to evaluate the *ProPASS* workflow.

Target molecule	# pathways identified by *ProPASS*	# pathways with solubility confidence scores for all reactions	% pathways with solubility confidence scores for all their reactions
2,3-Butanediol	1	1	100 %
cis-Muconic acid	147	29	20 %
Triacylglycerols	83	38	46 %
Fatty acid methyl ester	48	28	58 %
3-Hydroxypropanoic acid	88	32	36 %
Isopentenyl diphosphate	33	16	48 %
myo-Inositol	384	235	61 %

Table III: Implementation options for (a) five top-yielding pathways and (b) top-five shortest pathways. The columns list the pathway yield, length, reaction numbers (using KEGG reaction identification number) in the pathway, the number of sequences with solubility confidence score that were identified using *ProSol DB* for each reaction step, the maximum solubility confidence score value per reaction step, and bottleneck solubility for this path. Entries in parentheses indicate that multiple enzymes catalyze the reaction step, and in this case the number of sequences and the maximum solubility is reported per enzyme for that reaction step. A "–" indicates that no sequence was identified in both *ProSol DB* and non-reviewed sequences in *UniProtKB*.

(a)

Pathway yield	Pathway length	KEGG ID for reactions in pathways (R#)	# sequences per reaction	Max sol. conf. score per step
0.408	5	09286, 04919, 10947, 01607, 00743	2,350, 1, –, –	34, 100, 42, –, –
0.407	3	09286, 03045, 00919	2, (255,1), (5,163)	34, (100,37), (95,100)
0.407	3	03157, 03045, 00919	159, (255,1), (5,163)	100, (100,37), (95,100)
0.397	3	03157, 03045, 10161	159, (255,1), 1	100, 100, 37, 99
0.391	5	03158, 04919, 10947, 1604, 00743	14, 350, 1, 1, 1	100, 100, 42, –, –

(b)

Pathway yield	Pathway length	KEGG ID or reactions in pathways (R#)	# sequences per reaction	Max sol. conf. score per step
0.373	2	01608, 00907	1, 2	–, 90
0.373	2	01608, 00908	1, 120	–, 99
0.367	2	01608, 00740	1, 2	–, 74
0.365	2	09289, 00907	386, 2	100, 90
0.365	2	09289, 00908	386, 20	100, 99

Table IV: The pathway identified to produce 2,3-butanediol consists of two reaction steps, each catalyzed by a single enzyme. The table shows solubility confidence scores for enzymes from organisms used in Yan *et al*. and those for enzymes from organisms recommended by *ProPASS*.

Enzyme	Pathway identified by *ProPASS*				From Yan *et al.*	
	Reviewed A-list organisms	Predicted sol. conf. score	All sequences (reviewed & non-reviewed)	Predicted sol. conf. score	Organisms	Predicted sol. conf. score
(*R,R*)-butanediol dehydrogenase	*S. cerevisiae*	87	*L. lactis*	100	*B. subtilis*	96
acetolactate decarboxylase	*L. lactis*	83	*E. aerogenes*	99	*K. pneumoniae*	98

Figures Legends:

Figure 1: An overview of the steps used by *ProPASS* to identify synthesis pathways and their enzymatic implementations for a given target molecule. Solubility confidence scores from *ProSol DB* are retrieved for all sequences associated with a specific enzyme from the A-list or B-list organisms or computed on-the-fly using *ccSol omics* for non-reviewed sequences in *UniProtKB*.

Figure 2: Statistics related to *ProSol* DB. (a) Venn diagram showing sets of all EC numbers in KEGG (6,896), those covered by the A-list (2,427), and those covered by the B-list (4,350). The intersection between the A-list and the B-list consists of 2,125 EC numbers covering 46% of EC numbers available in *ProSol* DB. (b) Each stack shows information related to the A-list and B-list organisms. The first bar shows tally of EC numbers in A-list, B-list, and their intersection. The second bar shows that the A-list consists of 20 organisms, which is only 0.33% of the organism diversity yet covers 52% of all the EC numbers in the database. The third bar shows that 13% of all protein sequences in *ProSol DB* are covered by the A-list organisms. These statistics demonstrate that the number of verified enzyme sequences and EC numbers are overrepresented in the A-list, which comprises many model organisms.

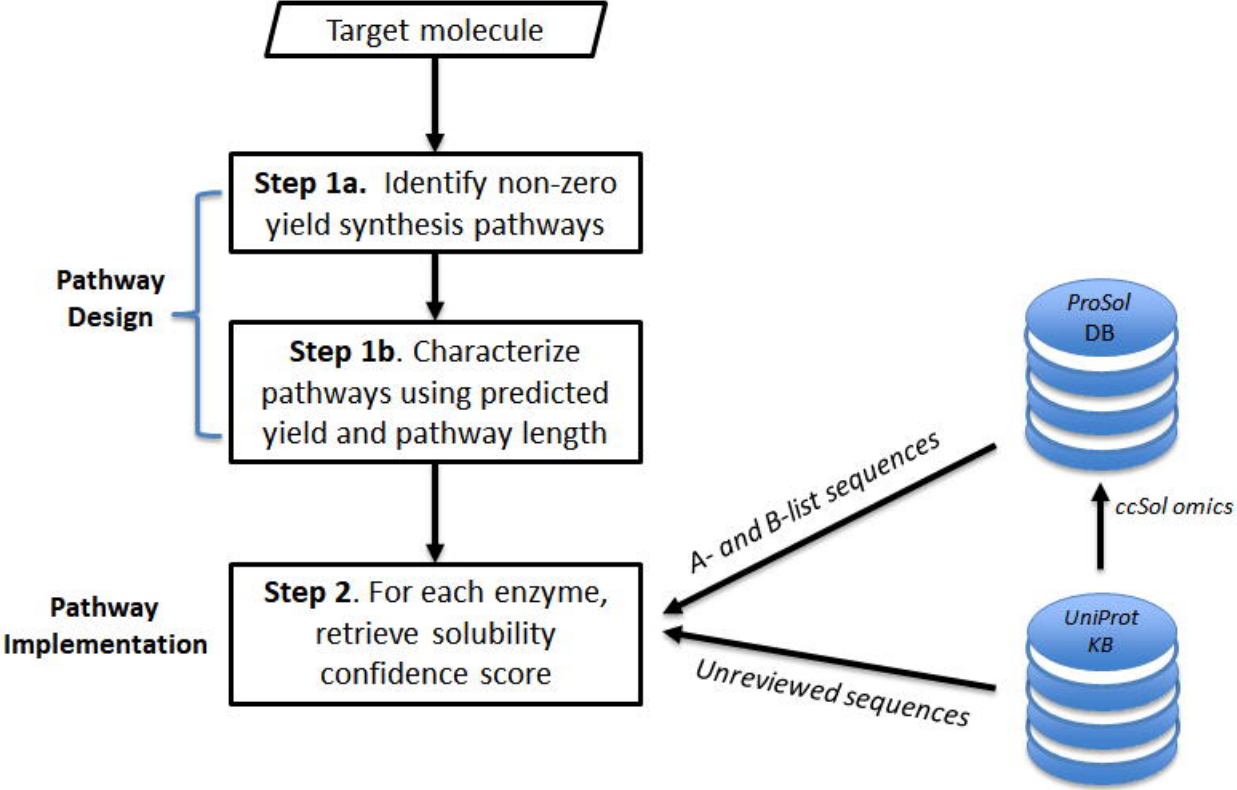
Figure 3: Correlation analysis between solubility confidence scores and EC numbers for enzymes with 10 or more sequences in *ProSol DB*. (a) Average solubility confidence scores for enzymes, sorted from highest to lowest average score. (b) Predicted solubility confidence score ranges for each enzyme, sorted from smallest to largest. (c) Standard deviation of solubility confidence scores for each enzyme, sorted as in panel (a). (d) Scatter plot of solubility

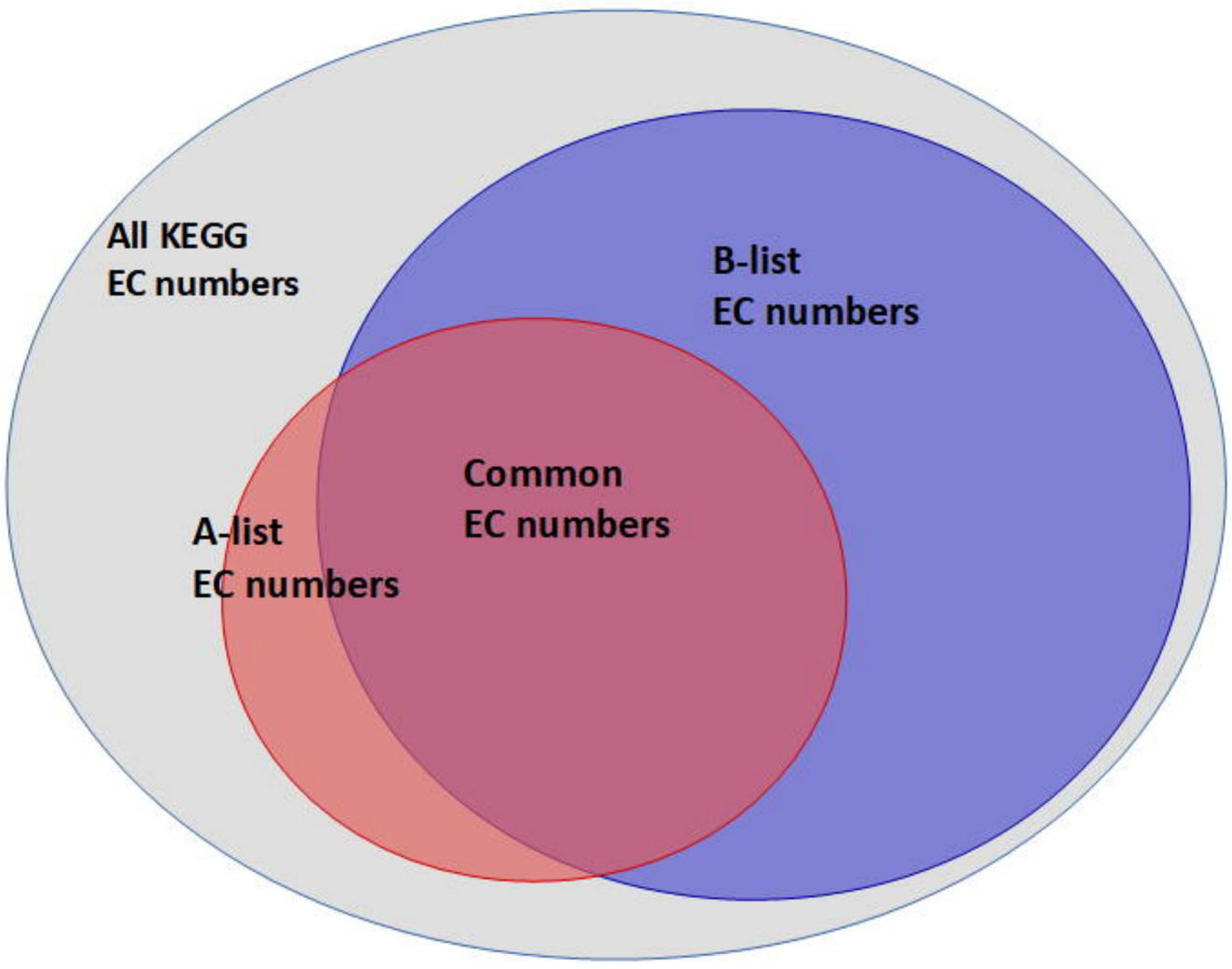
confidence score ranges and average scores. There is low correlation between the score ranges and average solubility confidence scores as determined using Spearman's Rank Test (Spearman's rank coefficient = 0.00897).

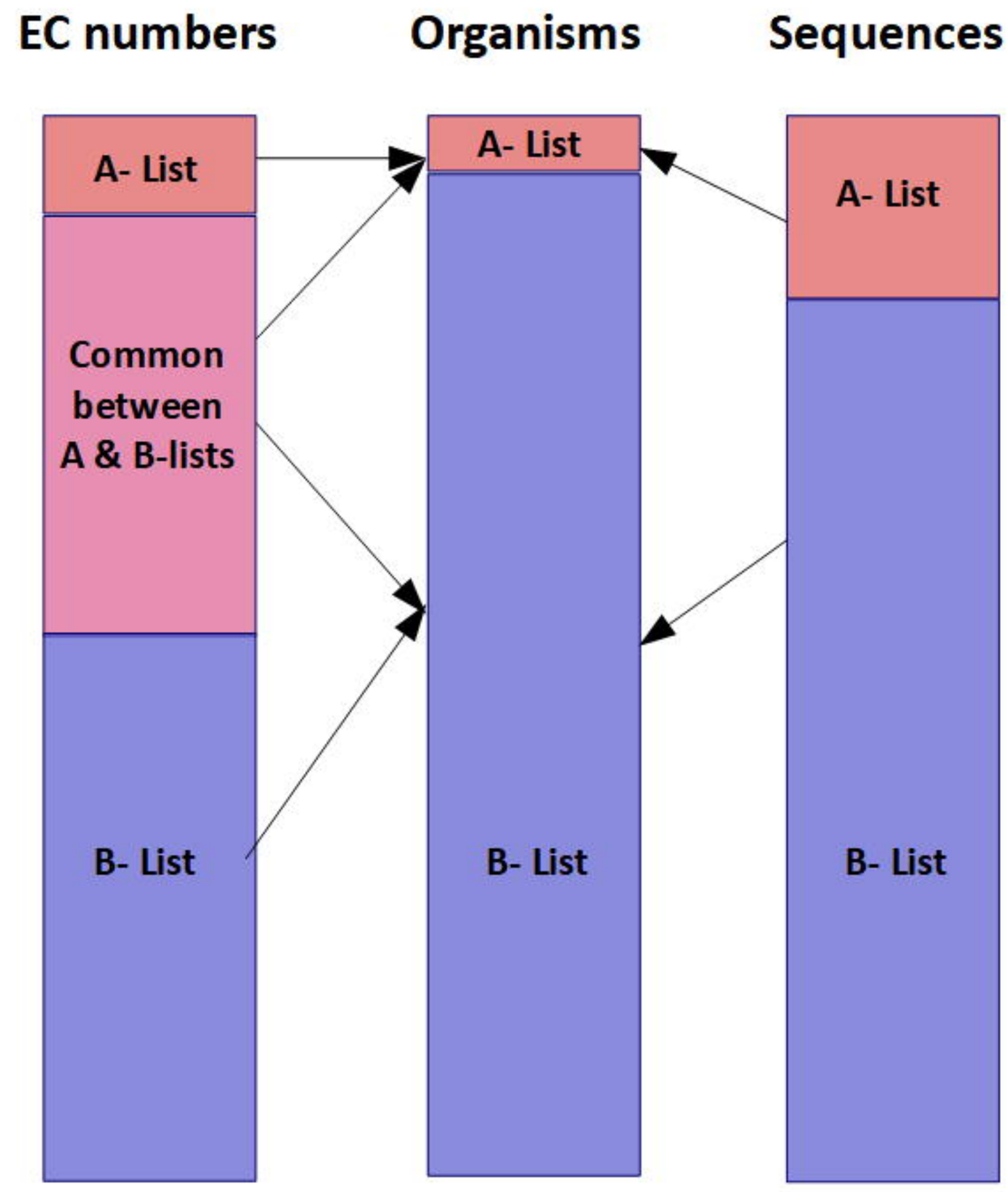
Figure 4: Correlation analysis between predicted solubility scores and sourcing organisms for those with 100 or more sequences in *ProSol* DB. (a) Histogram of phylogenetic distances from *E. coli.* (b) Average predicted solubility per organism, ranked from largest to smallest predicted solubility values. (c) Scatter plot of average predicted solubility per organism vs. phylogenetic distance from *E. coli* indicates low positive correlation as calculated using Spearman's Rank Test (Spearman's rank coefficient = 0.176919).

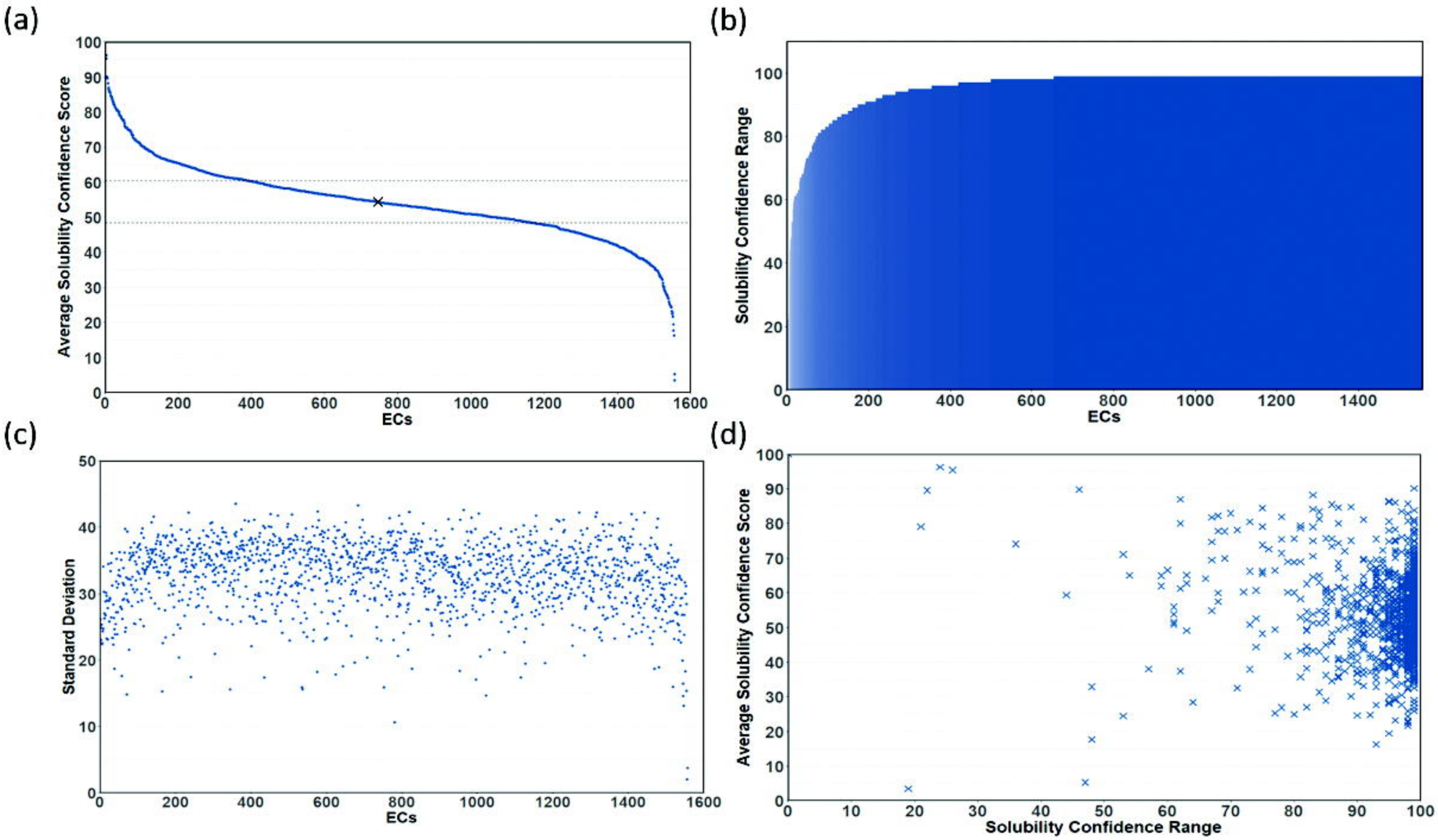
Figure 5: Predicted yields and corresponding lengths for pathways producing 3-hydroxypropanoic acid identified by *ProPASS*. Each point on the plot represents one pathway. The three orange triangles coincide with pathways reported in the literature.

Figure 6: Three pathways for producing 3-hydroxypropanoic acid (3-HP) in *E. coli* were identified by *ProPASS* and proposed by (a) Luo et al. (2016), (b) Cheng et al. (2016), and (c) Song et al. (2016). Two predicted solubility scores are provided for each enzymatic reaction. The black score is for a sequence from organism recommended by *ProPASS* and the red score is for the corresponding sequence that was used experimentally.

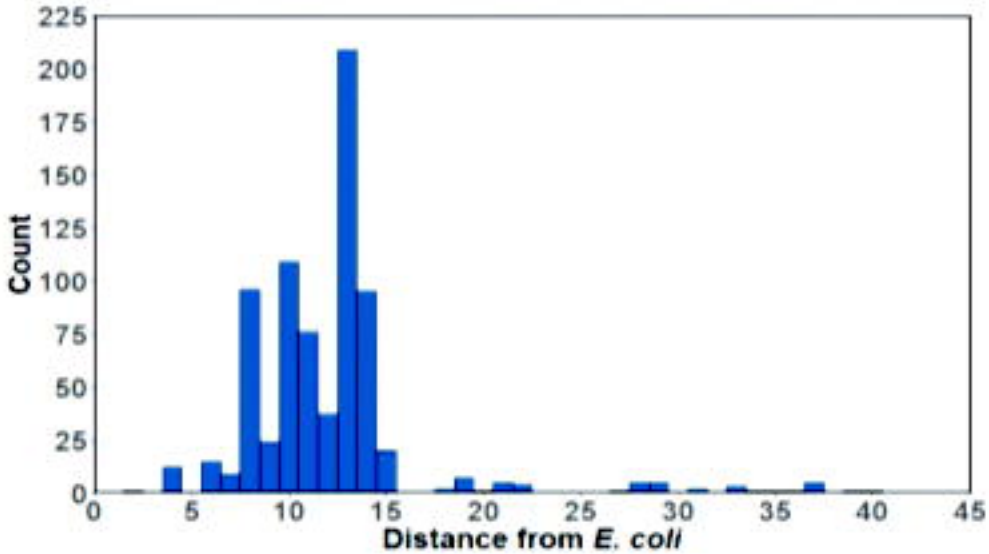

Target molecule
Step 1a. Identify non-zero yield synthesis pathways
Pathway Design
Step 1b. Characterize pathways using predicted yield and pathway length
Pathway Implementation
Step 2. For each enzyme, retrieve solubility confidence score
A- and B-list sequences
Unreviewed sequences
ProSol DB
ccSol omics
UniProt KB


All KEGG
EC numbers
B-list
EC numbers
Common
EC numbers
A-list
EC numbers

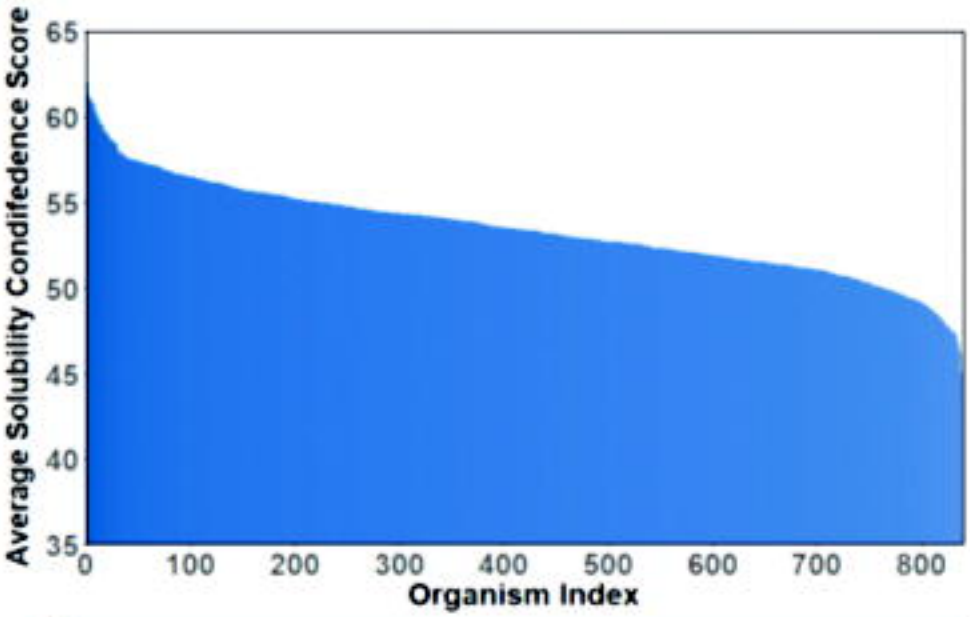

EC numbers
Organisms
Sequences
A- List
Common
between
A & B-lists
B- List
A- List
B- List
A- List
B- List


(a)
Average Solubility Confidence Score
ECs
(b)
Solubility Confidence Range
ECs
(c)
Standard Deviation
ECs
(d)
Average Solubility Confidence Score
Solubility Confidence Range

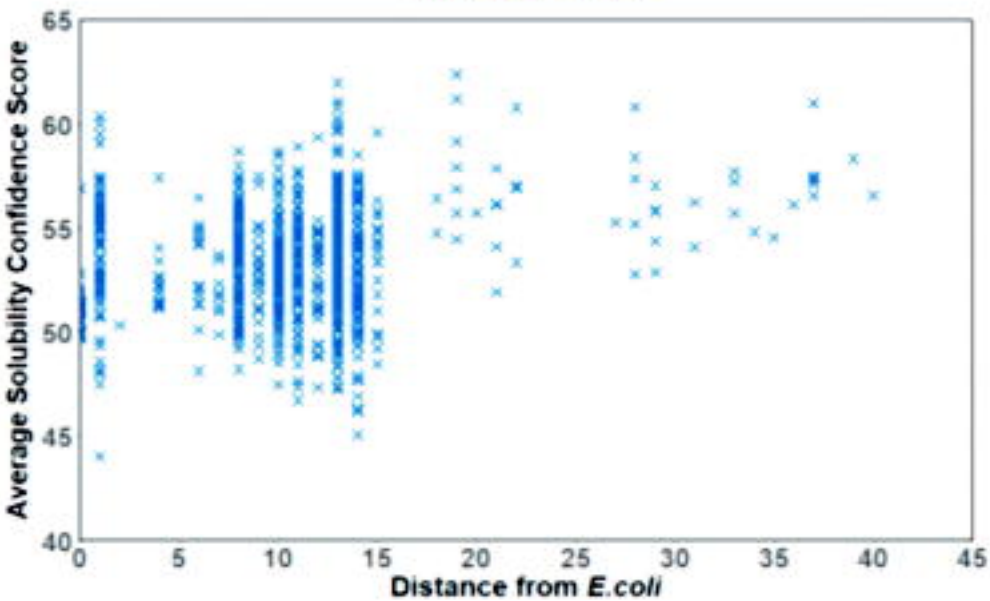
(a)

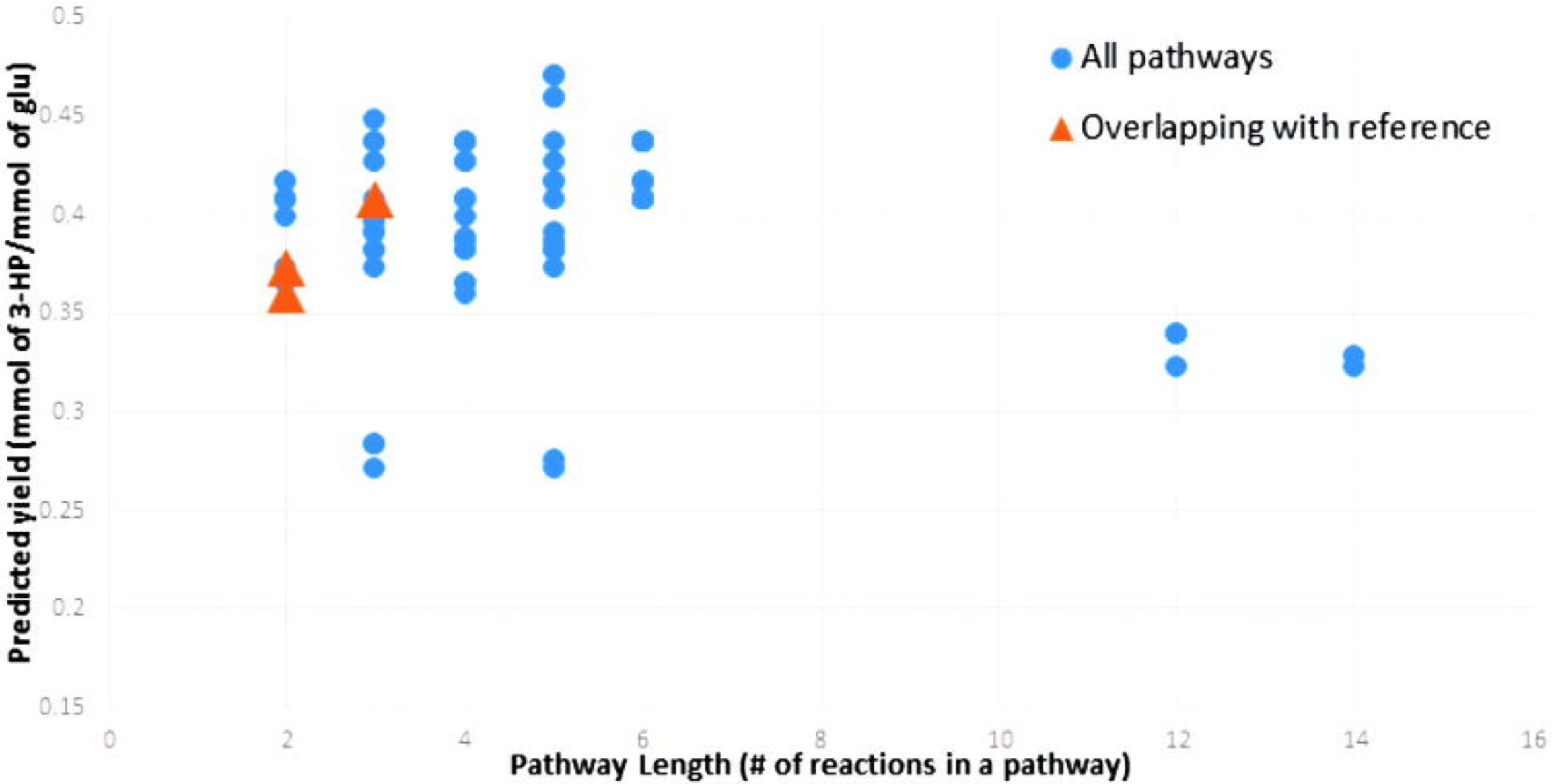

Count
Distance from E. coli

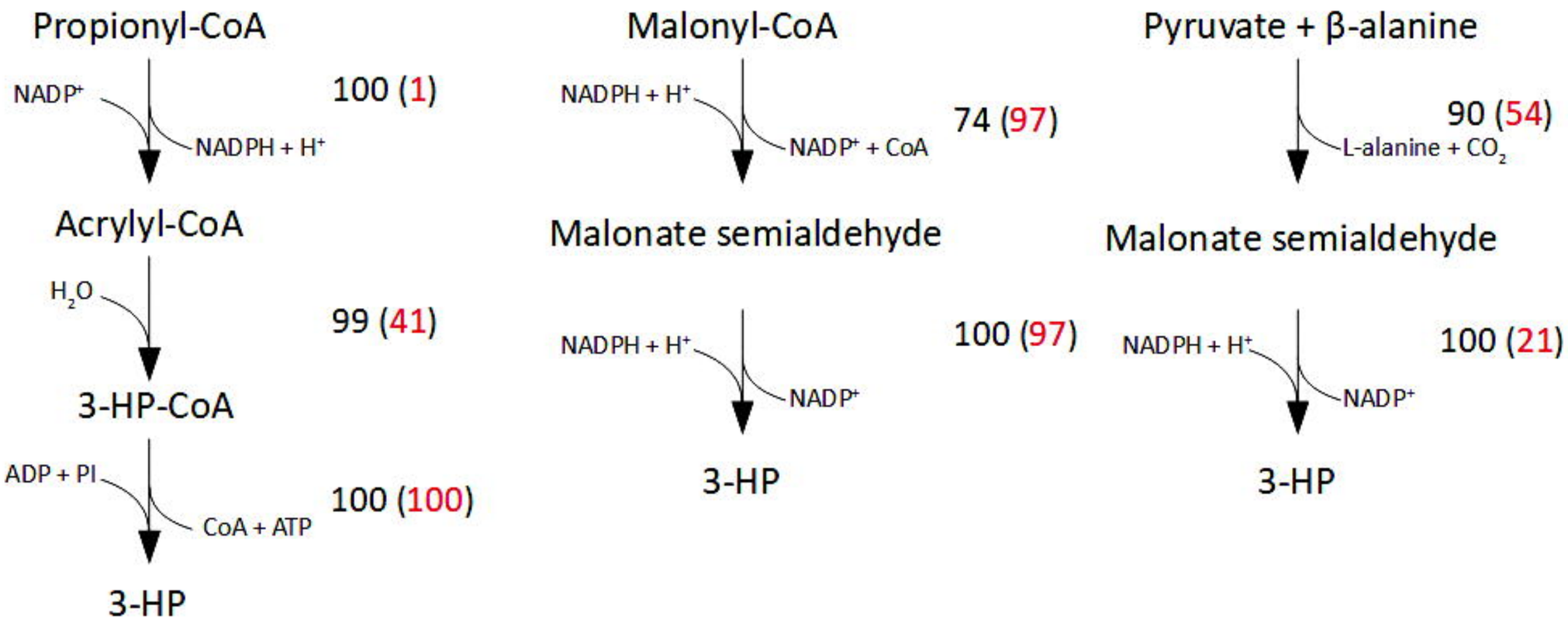
(b)


Average Solubility Condifedence Score
Organism Index

(c)


Average Solubility Confidence Score
Distance from E.coli


All pathways
Overlapping with reference
Predicted yield (mmol of 3-HP/mmol of glu)
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0
2
4
6
8
10
12
14
16
Pathway Length (# of reactions in a pathway)


Propionyl-CoA
NADP+
NADPH + H+
100 (1)
Acrylyl-CoA
H2O
99 (41)
3-HP-CoA
ADP + PI
CoA + ATP
100 (100)
3-HP
Malonyl-CoA
NADPH + H+
NADP+ + CoA
74 (97)
Malonate semialdehyde
NADPH + H+
NADP+
100 (97)
3-HP
Pyruvate + β-alanine
L-alanine + CO2
90 (54)
Malonate semialdehyde
NADPH + H+
NADP+
100 (21)
3-HP