Clone Less, Know More: Efficient Expression Optimization of Proteins and Pathways

Howard Salis

Penn State University

De Novo DNA LLC

PENNSTATE 1855

De Novo DNA

Design software for engineering genetic systems

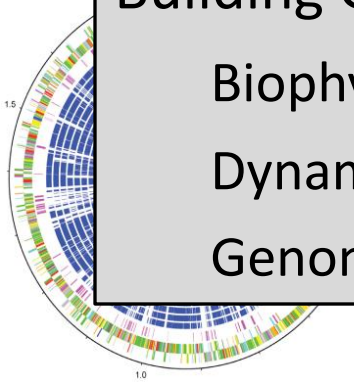
GenScript Webinar, April 30th, 2015

The Salis Lab

Biophysical Modeling to Predict Expression from DNA Sequence

Building Genetic Systems to **Challenge our Knowledge**:

- Biophysics of Gene Expression
- Dynamics of Enzymatic Multi-Protein Systems
- Genome-Scale Predictions vs. Phenotype



rates

Optimization of Synthetic DNA to Achieve a Targeted Function

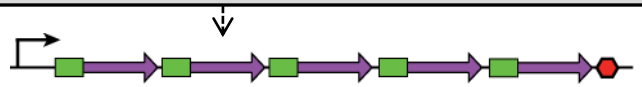
Building Genetic Systems to **Solve Problems**:

- Biodetoxification Pathway for Lignocellulose
- Versatile RNA-based Sensors for Diagnostics and Detection
- Re-engineering Central Metabolism to Over-produce Products

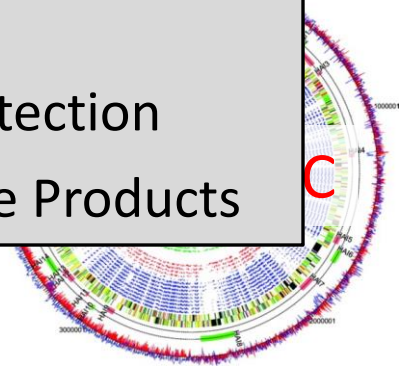
Design S

If substr
if pH >
call

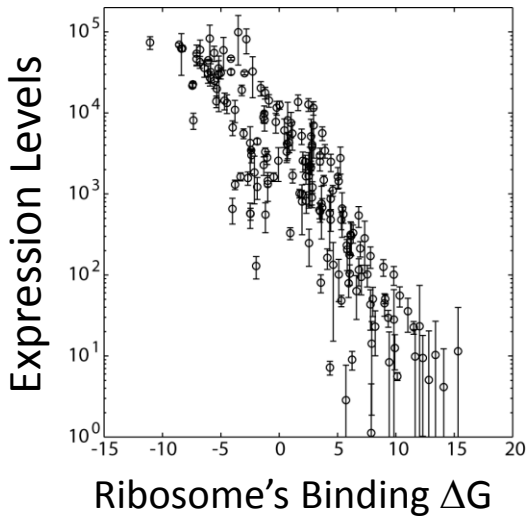
call shiftMetabolism(glycolysis = 1000,
TCA cycle = 100, biomass = 0)



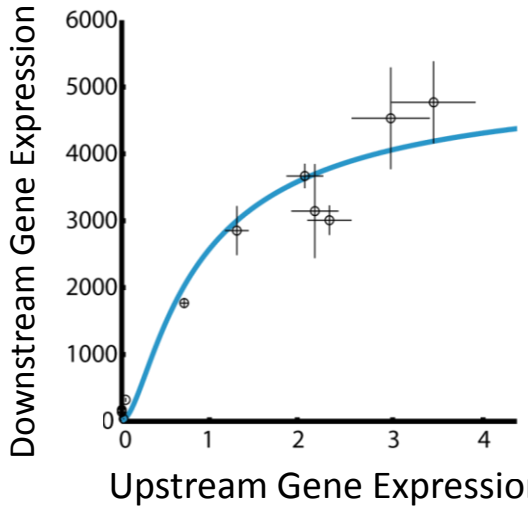
genes, operons, & pathways



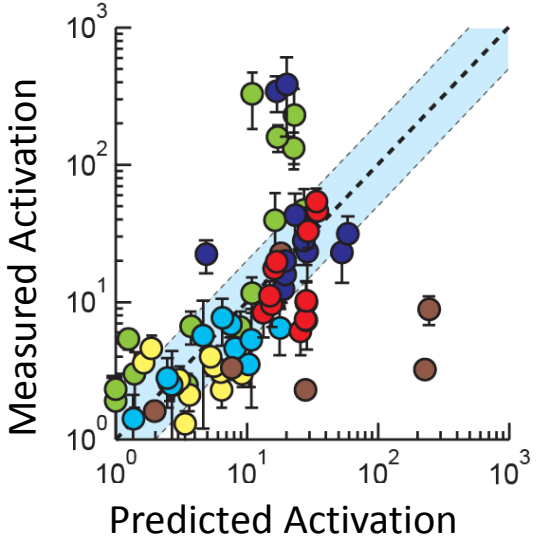
Biophysical Models of Gene Expression and Regulation
Sequence to Function



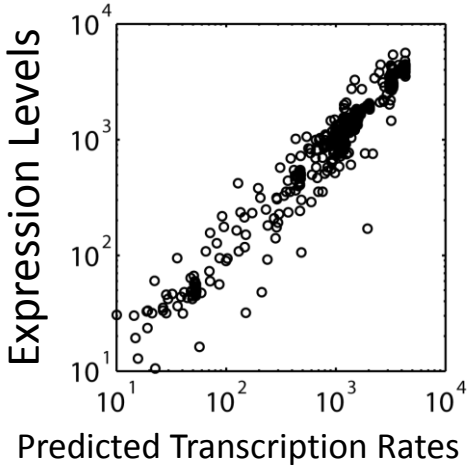
Translation Initiation



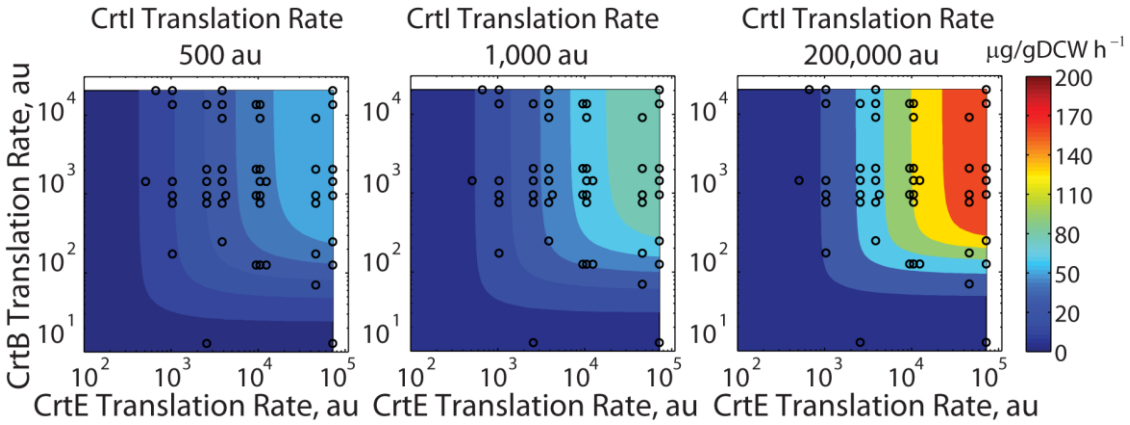
Translational Coupling



Riboswitch Regulation



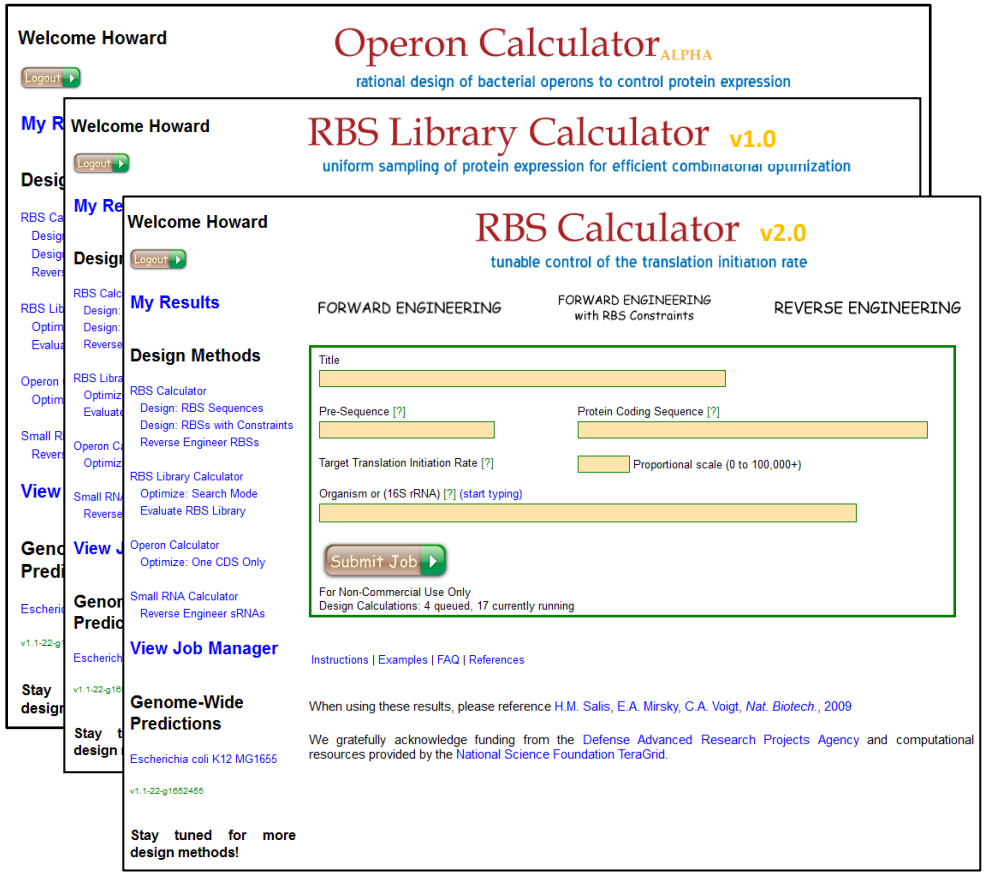
Transcriptional Regulation

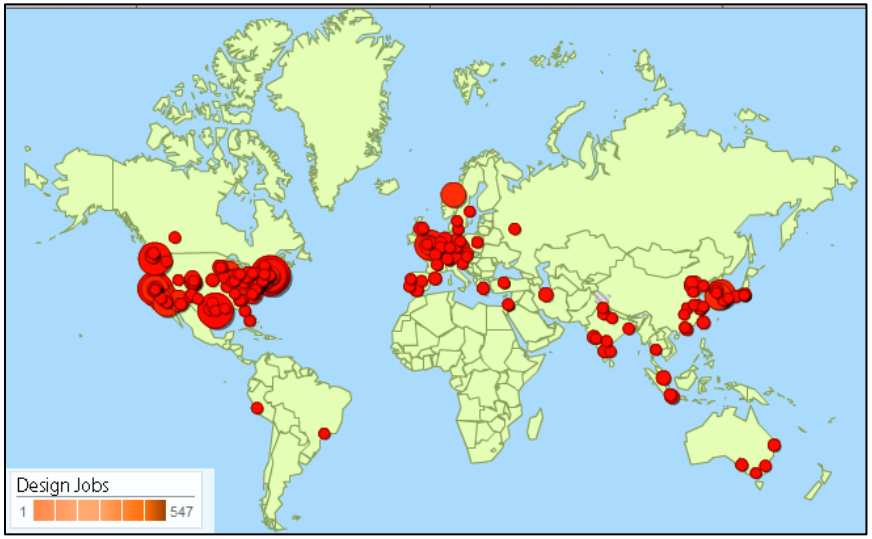


Multi-Enzyme Pathways

De Novo DNA LLC

Founded by Howard Salis to distribute our automated design methods for engineering large genetic systems





Over 70,000 DNA sequences have been designed since 2010

Freely available for non-commercial applications

http://www.denovodna.com/software

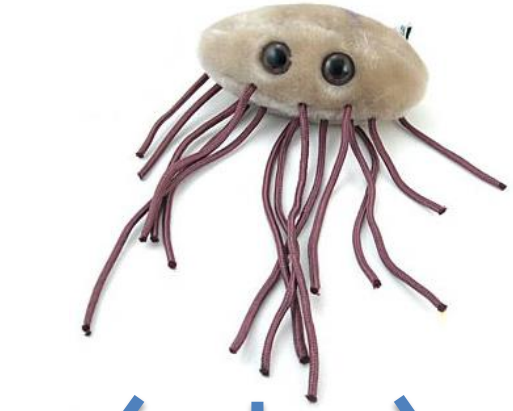
Clone Less, Know More

Objective: **Predict the *specific* DNA sequences that will maximize a product's titer**

Do it reliably, for diverse recombinant products.
And efficiently, with less cloning and fewer experiments.

We should also ***learn something new*** so the next product is developed even faster

Engineering Genetic Systems
to Manufacture Products



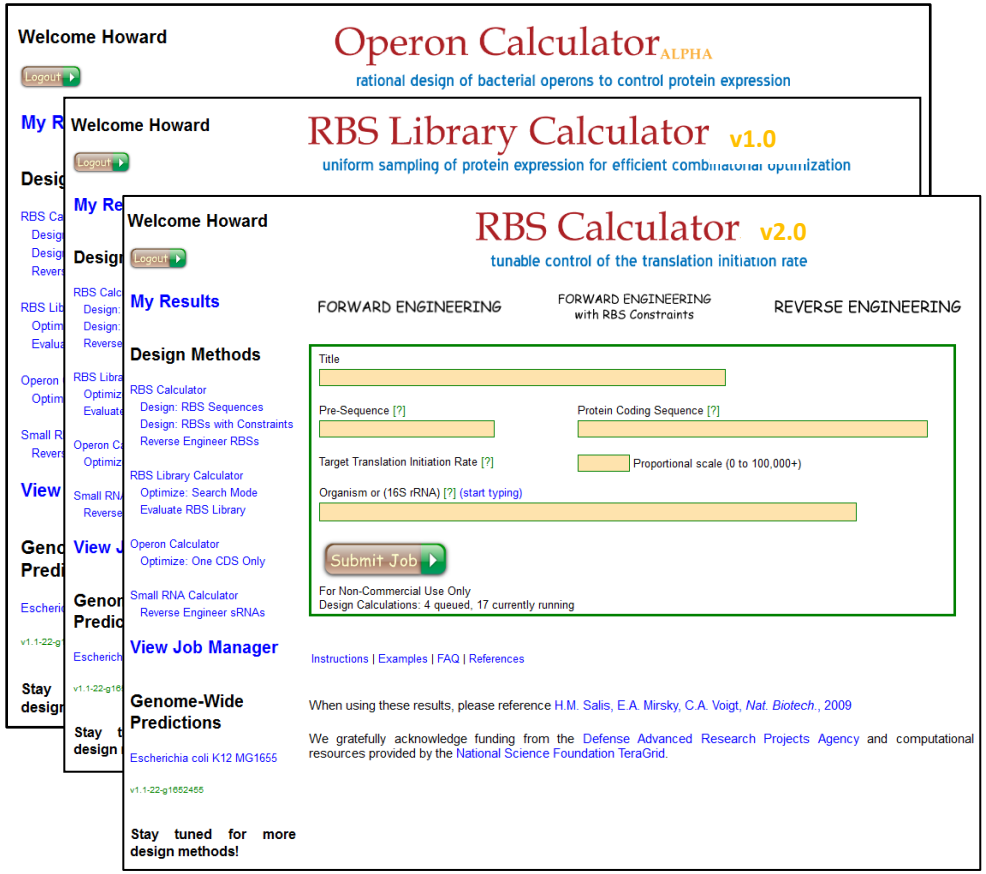
Recombinant Enzymes **Recombinant Biologics**

Natural Products

Applications of our Models & Algorithms

Recombinant Protein Expression

Antibody expression (Makino et. al., 2011)
P450 cytochrome expression (Chang et al., 2014)
Salmonella protein expression (Medina et. al., 2011)
Xylanase and glucanase expression (Liu et. al., 2012)
Nitrile hydratase expression (Kang et. al., 2014)
Secreted protein expression (Heggeset et. al. 2013)



Pathway Optimization & Circuit Engineering

Riboflavin biosynthesis (Lin et. al., 2014)

Terpenoid biosynthesis (Zhou et. al., 2014)

Amorphadiene biosynthesis (Nowroozi et. al., 2014)

2,3-butanediol biosynthesis (Oliver et. al., 2014)

Butanol biosynthesis (Lim et. al., 2013)

Isoprene biosynthesis (Zurbriggen et. al. 2012)

Nitrogen fixation (Temme at. al., 2012)

Fatty acid biosynthesis (Lennen et. al., 2013)

Engineering memory genetic devices (Yang et. al., 2014)

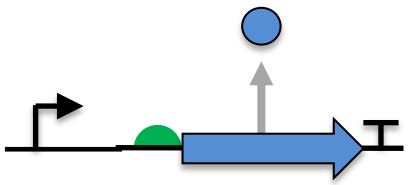
Third–party validation (Biggs et. al., 2014)

Today's Topics

#1: Designing DNA to Control Protein Expression
#2: Optimization of Multi-Protein Genetic Systems
#3: Case Studies: Engineered Metabolic Pathways

More recombinant products are requiring the expression of multiple proteins simultaneously
Our methods allow you to efficiently engineer & optimize multi-protein genetic systems

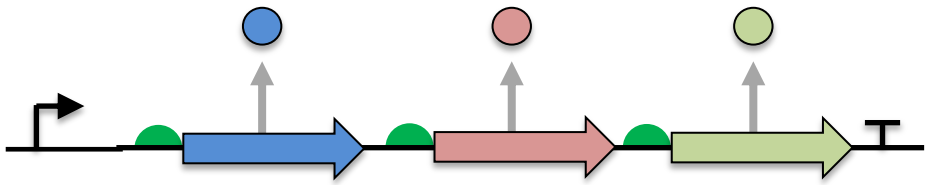
A Single-Protein Genetic System



expression & purification
of a single protein

Examples: single-chain antibodies,
single-subunit enzymes

A Multi-Protein Genetic System



multiple proteins working together
to produce a product

Examples: recombinant vaccines, multi-subunit enzymes,
natural product biosynthesis (antibiotics, therapeutics)

Protein Expression in a Nutshell

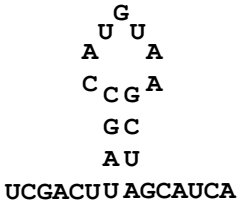
RATE OF Protein Production

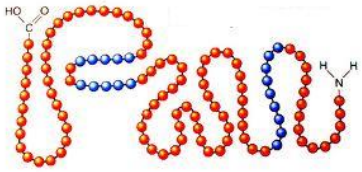
Transcription → Translation → **RATE OF** Protein Folding

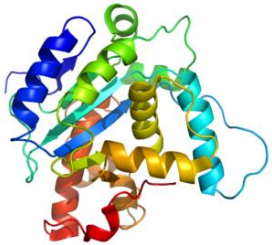
RATE OF Membrane Insertion

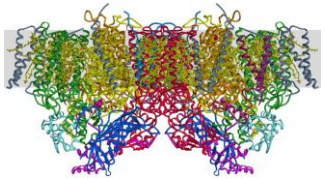
RATE OF Secretion

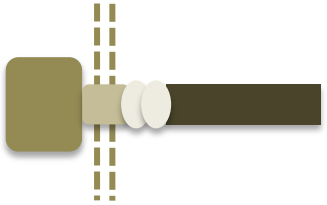
DNA











The Production **Rate** of Proteins must be balanced with other cellular processes

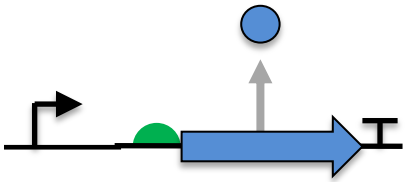
cell growth, secretion, membrane insertion,
glycosylation, protease cleavage, enzyme activity

Designing DNA to Control Protein Expression

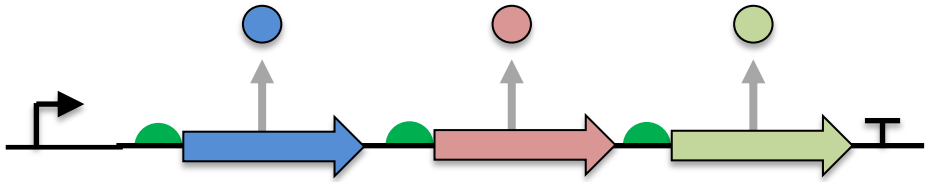
Target Protein Expression Levels → An Engineered DNA sequence

Low-cost DNA Synthesis allows you to custom-design
every genetic part to control transcription rate & translation rate
and to rapidly optimize the expression of all proteins in your system

Promoters
(controls transcription initiation)



Ribosome Binding Sites
(controls translation initiation)



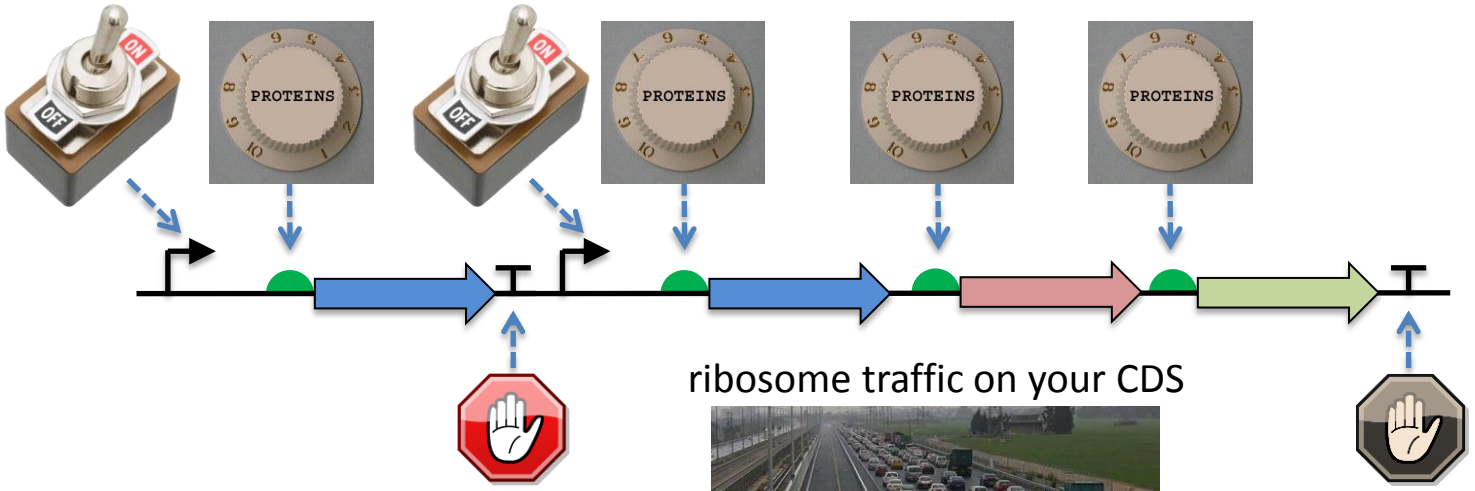
Protein Coding Sequences
(controls translation elongation rate)

Transcriptional Terminators
(controls termination efficiency)

Designing DNA to Control Protein Expression

Target Protein Expression Levels → An Engineered DNA sequence

use **Promoters** for dynamic/inducible control
use **Ribosome Binding Sites** to statically increase/decrease expression by 10,000-fold
all **Protein Coding Sequences** must be codon-optimized
all **Transcriptional Terminators** must be >95% efficient and *non-repetitive*

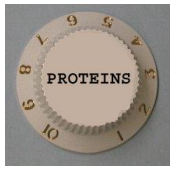


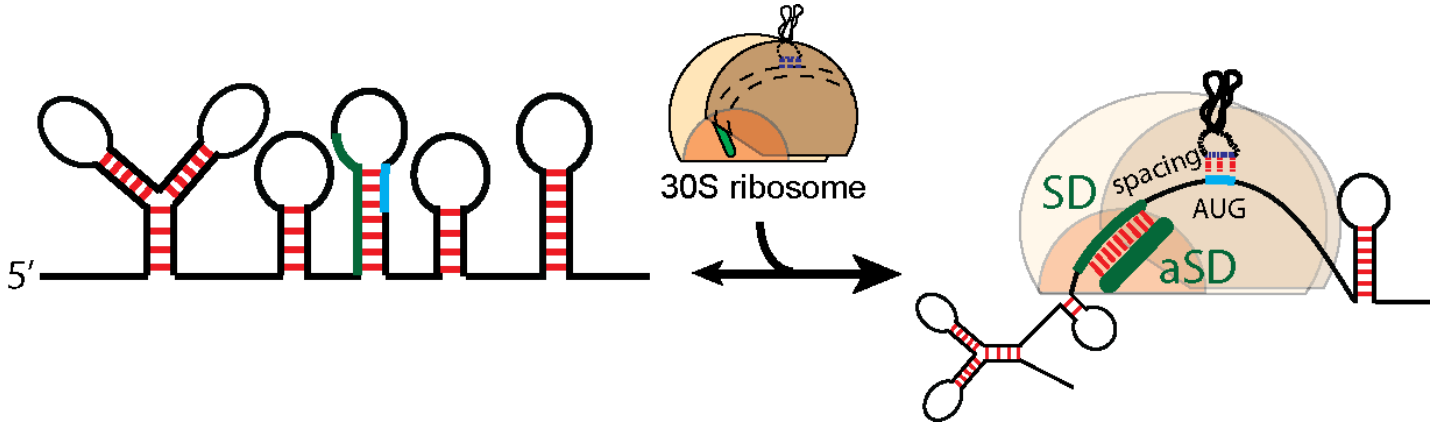
ribosome traffic on your CDS

CDSs optimized

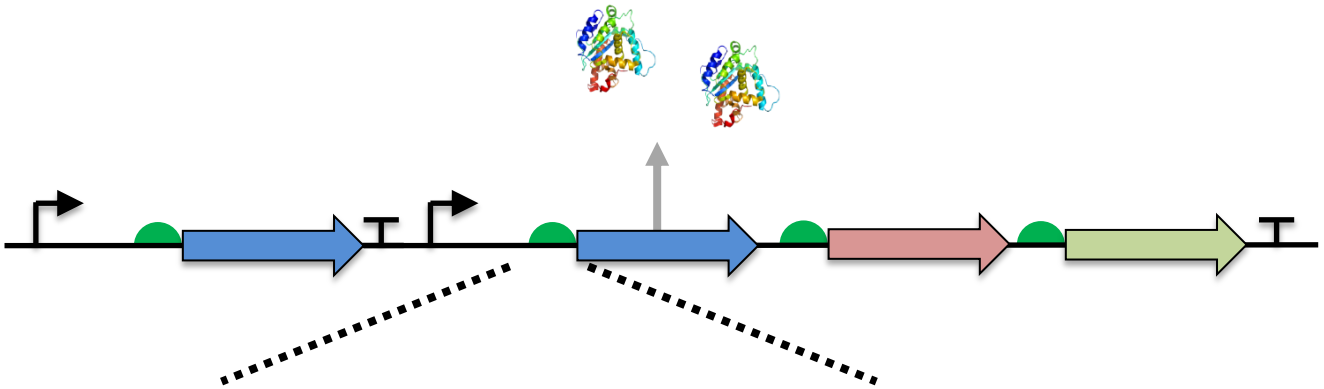
CDSs NOT optimized!

Designing DNA to Control Protein Expression





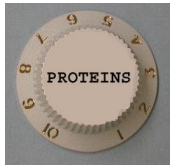
a RBS Sequence → Low translation initiation rate → production rate of protein: **Low**



CAGUACACAACUUCGCUCGUAGUUC ~35 nucleotides

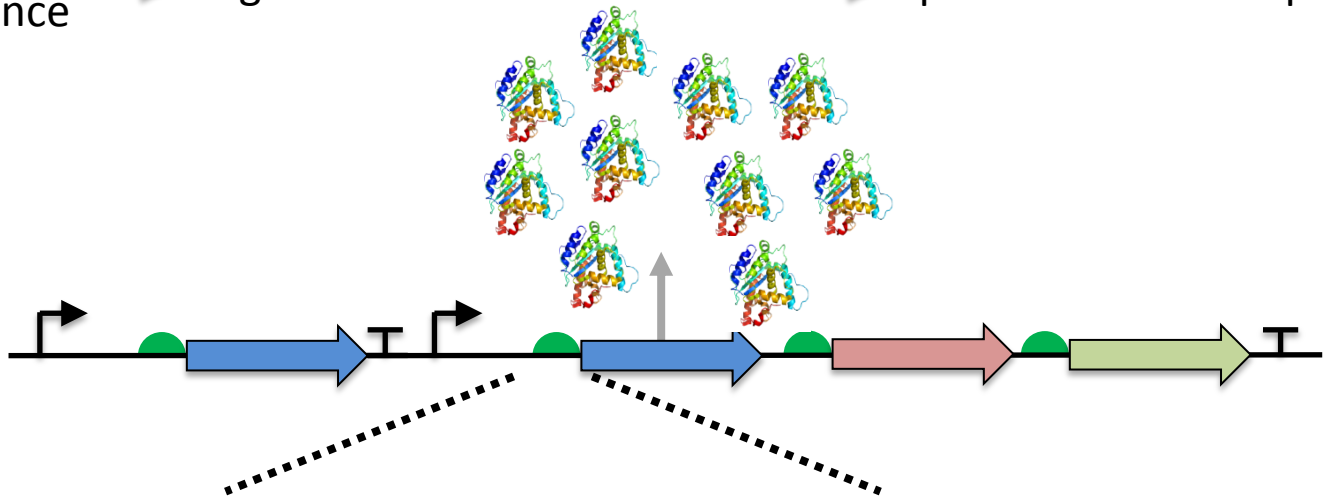
ribosome binding site sequence

Designing DNA to Control Protein Expression



By designing synthetic ribosome binding sites,
we may **Control**, **Coordinate**, and **Optimize** the Expression of *Multiple* Proteins

a different
RBS Sequence → High translation initiation rate → production rate of protein: **High**



AAUAUACACAAAGGAGGUUACAACG ~35 nucleotides

a different ribosome binding site sequence

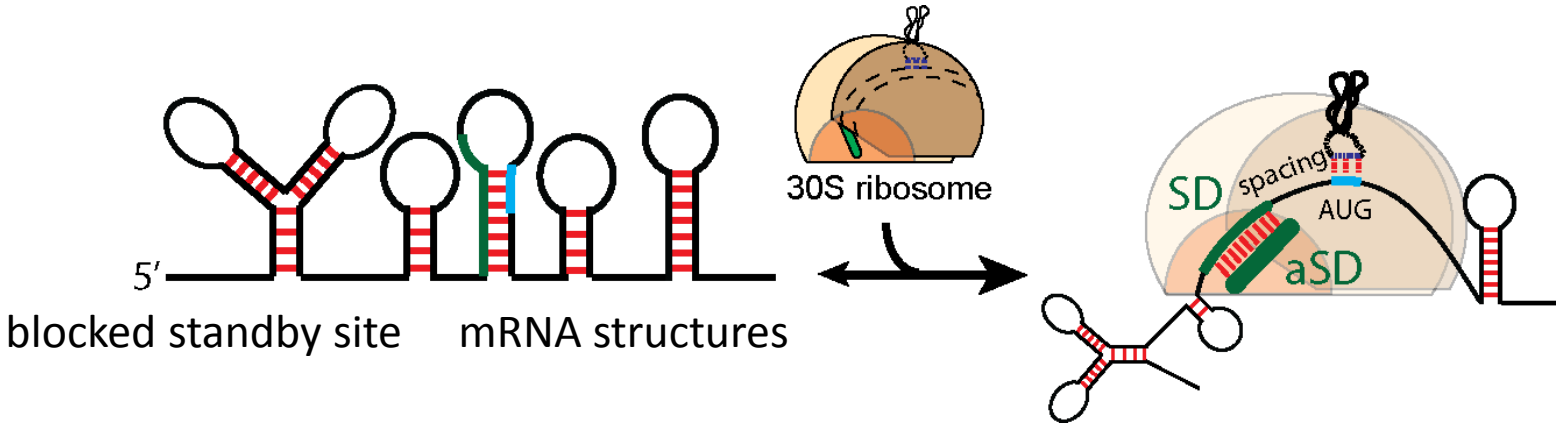
Factors that Affect Translation Initiation

RBS sequence **CDS sequence**

AUCCGAUAACUAGUCACACAGUAAAAAUUUAGUU<u>**AUG**</u>ACCACCUUUCACAAAGCAAGCGGA
GUGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC . . . **UAA**

<u>Molecular interactions controlling translation initiation</u>

1. Hybridization between the mRNA and 3' end of the 16S rRNA @ the "Shine-Dalgarno"
2. The unfolding of mRNA structures that overlap with the ribosome's footprint
3. Hybridization between the start codon and $tRNA^{fMet}$
4. Ribosome stretching or compression, due to long or short spacer regions
5. mRNA structures in standby sites that block ribosome binding
6. The time-scale of RNA folding kinetics vs. ribosome assembly kinetics



Factors that Affect Translation Initiation

RBS sequence **CDS sequence**

AUCCGAUAACUAGUCACACAGUAAAAAUUUAGUUAUGACCACCUUUCACAAAGCAAGCGGA
GUGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC . . . **UAA**

Let's insert a consensus Shine-Dalgarno sequence for maximum mRNA-rRNA hybridization
Uh Oh! We created a new mRNA structure that will inhibit translation rate.

AUCCGAUAACUAGUCACACAUAAGGAGGUAAGUUAUGACCACCUUCACAAAGCAAGCGGAG
UGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC . . . **UAA**

Many overlapping causes = difficult to design RBS sequences "by eye"

Cause: mRNA-rRNA hybridization
Cause: unfolding mRNA structures
Cause: non-optimal spacing
Cause: blocked standby sites
Cause: different start codons

Effect: Translation Initiation Rate

Factors that Affect Translation Initiation

RBS sequence **CDS sequence**

AUCCGAUAACUAGUCACACAGUAAAAAUUUAGUUAUGACCACCUUUCACAAAGCAAGCGGA
GUGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC . . . **UAA**

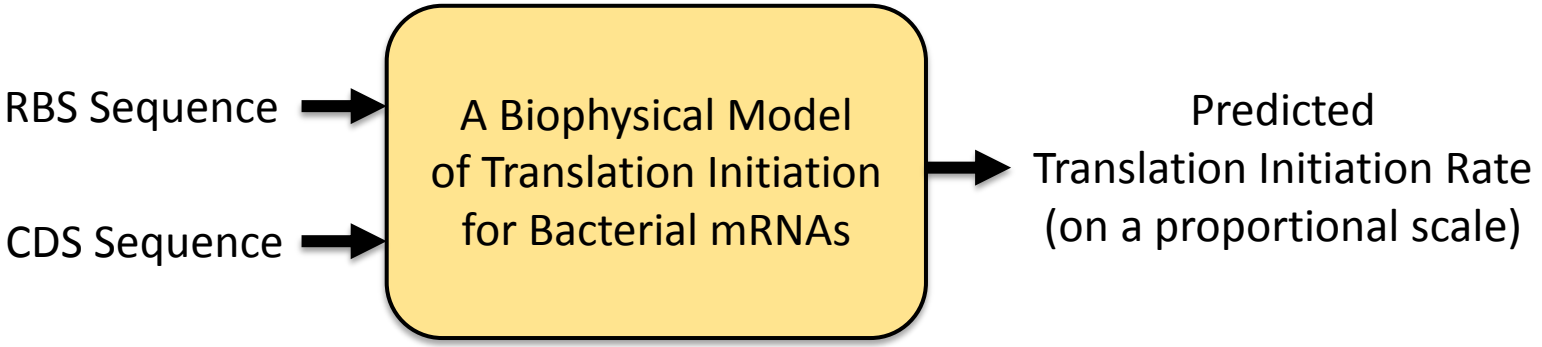
Let's insert a consensus Shine-Dalgarno sequence for maximum mRNA-rRNA hybridization
Uh Oh! We created a new mRNA structure that will inhibit translation rate.

↓

AUCCGAUAACUAGUCACACAUAAGGAGGUAAGUUAUGACCACCUUCACAAAGCAAGCGGAG
UGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC . . . **UAA**

Many overlapping causes = difficult to design RBS sequences "by eye"

We developed a Quantitative Model that Calculates Causes & Predicts their Effect



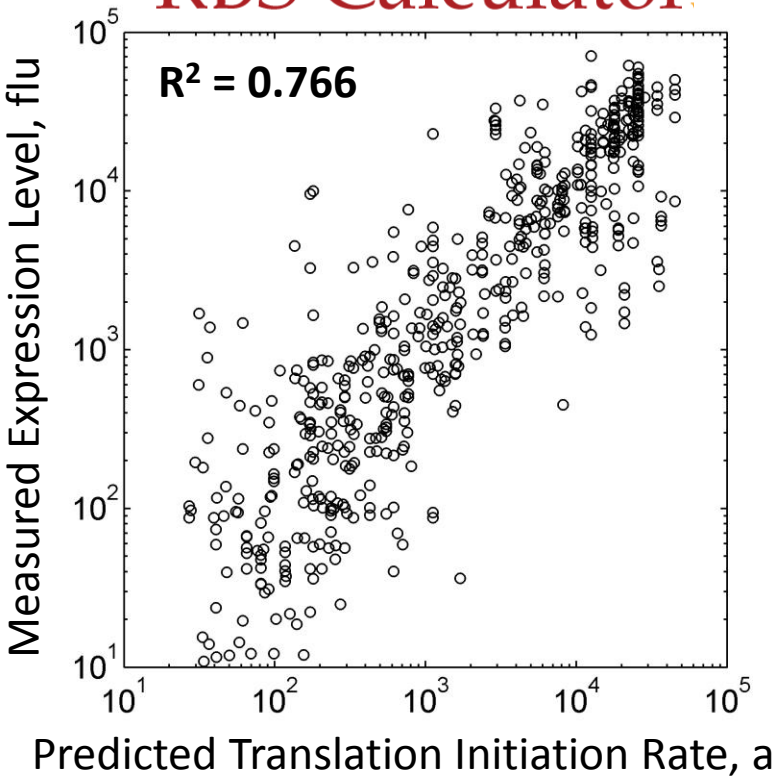
The Ribosome Binding Site Calculator

RBS Sequence →
CDS Sequence →

A Biophysical Model of Translation Initiation for Bacterial mRNAs

→ Predicted Translation Initiation Rate (on a proportional scale)

RBS Calculator



We can predict & control translation rate

Validation data-set

624 RBS sequences with different
- CDS reporters
- Shine-Dalgarno like sequences
- mRNA structures
- standby site regions
- spacer regions

measured in *E. coli* DH10B on different days

Relevant papers

Salis et. al., *Nature Biotechnology*, 2009
Salis et. al., Methods in Enzymology, 2011
Espah Borujeni et. al., *Nucleic Acid Research*, 2013
Farasat et. al., Molecular Systems Biology, 2014
and other unpublished work from the Salis Lab

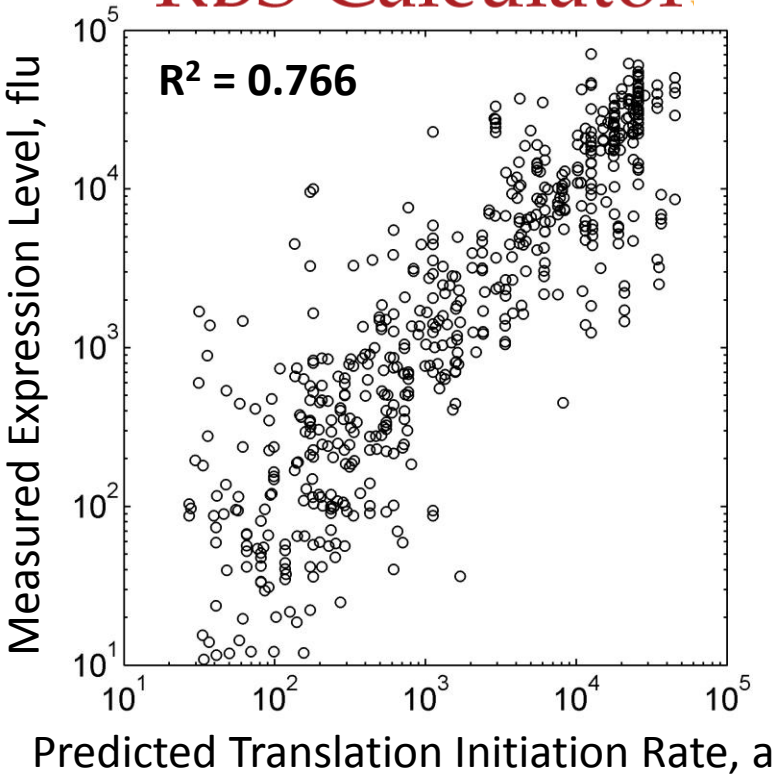
The Ribosome Binding Site Calculator

RBS Sequence →
CDS Sequence →

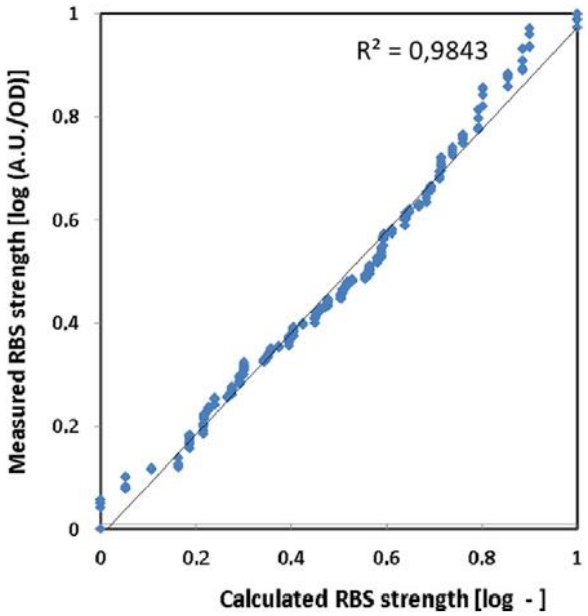
A Biophysical Model of Translation Initiation for Bacterial mRNAs

→ Predicted Translation Initiation Rate (on a proportional scale)

RBS Calculator



Third-party Validation $R^2 = 0.9843$



Coussement et. al. "One step DNA assembly for combinatorial metabolic engineering", *Metabolic Engineering*, 2014

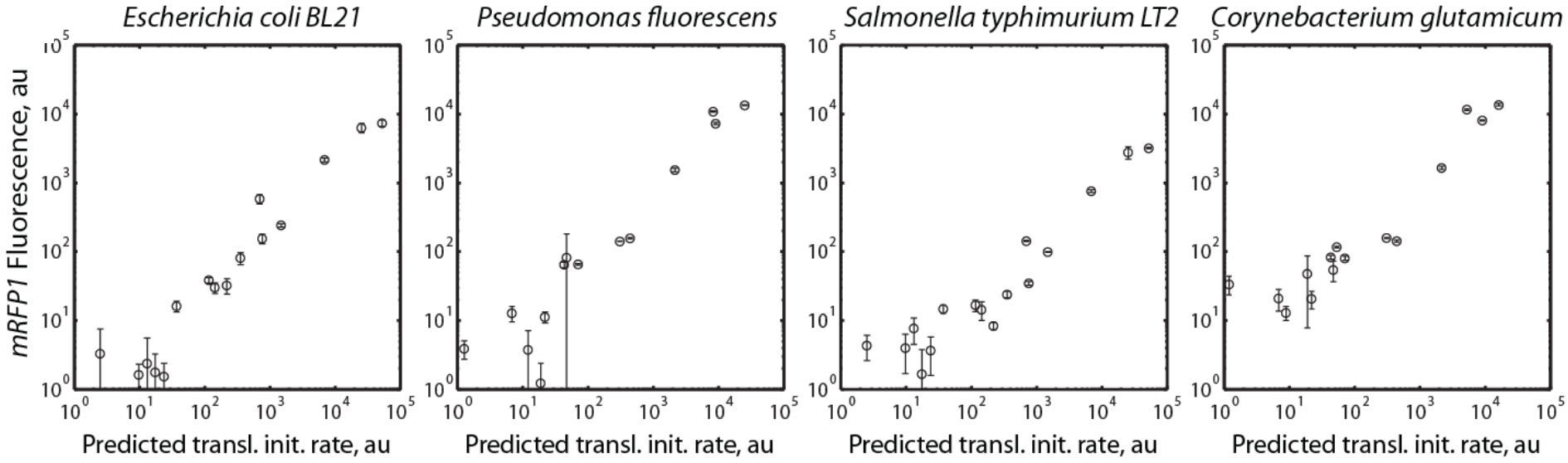
The RBS Calculator in Different Hosts

Industrial biotechnology uses many organisms to manufacture products

Different organisms, same RNA biophysics

Slightly different ribosomes

The biophysical model of translation initiation controls protein expression in diverse bacterial hosts



Farasat et. al., *Molecular Systems Biology*, 2014

Automated Design to Optimize Expression

Objective: design an RBS sequence
to increase production of a specific protein *by 20-fold*

Your initial sequence

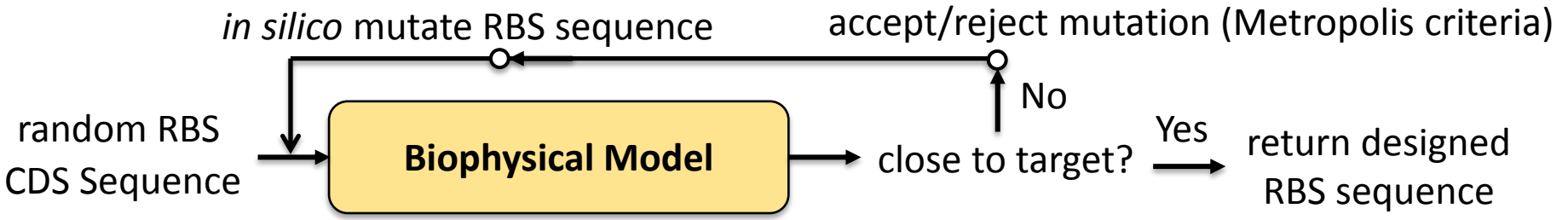
RBS sequence **CDS sequence**

AUCCGAUAACUAGUCACACAGUAAAAAUUUAGUUAUGACCACCUUUCACAAAGCAAGCGGA
GUGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC . . . **UAA**

Odds of answering this question correctly = like picking a winning NCAA bracket
How many choices? 35 nt 5' UTR … $4^{35} = 1.1 \times 10^{21}$

Computational Optimization

a target Protein production rate → ***a target*** translation initiation rate → ***a designed*** RBS Sequence



Automated Design to Optimize Expression

Objective: design an RBS sequence
to increase expression of a specific protein *by 20-fold*

Your initial sequence

RBS sequence **CDS sequence**

AUCCGAUAACUAGUCACACAGUAAAAAUUUAGUUAUGACCACCUUUCACAAAGCAAGCGGA
GUGCUUCUUGUGCCGACUUACGAGCAUCUAGCGAGCAUCUAGCGACUACUGAC ... **UAA**

RBS Calculator — predict translation init. rate → 1000 au

CGAACCGCUAUUCUAGAUAGUUCAAAAACAGAAC	100 au
UAUGCCUUCACAUUCACCAUUCAGAGACCGGUCG	300 au
AUCCGAUAACUAGUCUUUAAGUAAAAAUUUAGUU	3 000 au
UACCACUAAAACUAACCUAACGAGUAGGUUAUAA	10 000 au
AAAAUUUUCAUAAACAAGGUCGGGGGAUAUCAAG	**20 000 au**
UAUCAUUAUAUUCAUCGAAUAAGGGGAAUCUACU	100 000 au

1000 au → 20 x 1000 au = 20 000 au → **design a new RBS**

Optimizing Protein Expression

Q: Why do some proteins express better than others?

In this study, researchers from the Serrano Lab expressed 25 proteins with different RBS sequences, CDS sequences, and protein solubilities/folding

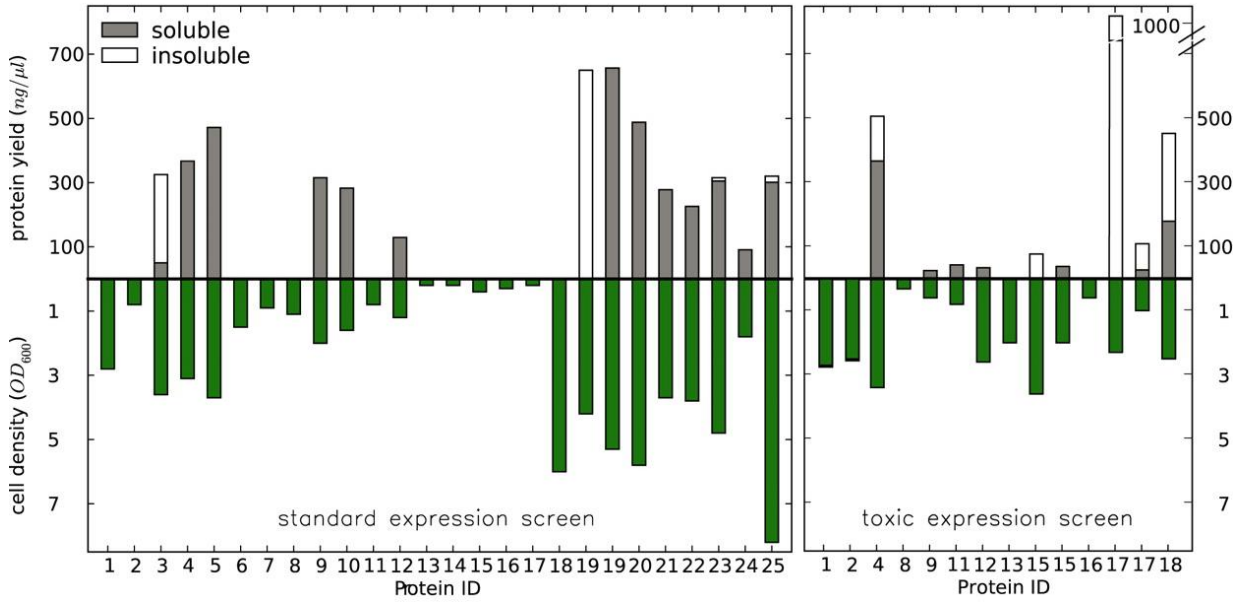
2 Expression modes:
Standard: T7pr + BL21-DE3 at 37°C
Toxic: T7pr + BL21-DE3-pLysE at 20 °C

all CDSs codon-optimized

Engineered leucine zipper	T7 promoter, RBS, start codon for expression in *E. coli*
Engineered leucine zipper	T7 terminator
FKBP12 (FK506-binding protein)	FLAG epitope tag (DYKDDDDK)
Engineered FKBP12-rapamycin-binding domain FRB(T2098L)	3-repeat FLAG epitope tag
Arabidopsis phototropin 1 LOV2 domain	Hexahistidine affinity tag
TEM-1 β-lactamase fragment 1	StrepII affinity tag
TEM-1 β-lactamase fragment 2	Glutathione S-transferase tag
Gaussia luciferase fragment 1	TEV protease cleavage site
Gaussia luciferase fragment 2	PreScission protease cleavage site
Engineered cyan fluorescent protein	2 aa flexible Glycine–Serine linker
Engineered yellow fluorescent protein	6 aa flexible Glycine–Serine linker
Engineered red fluorescent protein	10 aa flexible Glycine–Serine linker

the 25 Proteins expressed at greatly varying yields and host toxicities

Why?



Grunberg et. al., "Building blocks for protein interaction devices", Nucleic Acid Research, 2010

Optimizing Protein Expression

Q: Why do some proteins express better than others?

In this study, researchers from the Serrano Lab expressed 25 proteins with different RBS sequences, CDS sequences, and protein solubilities/folding

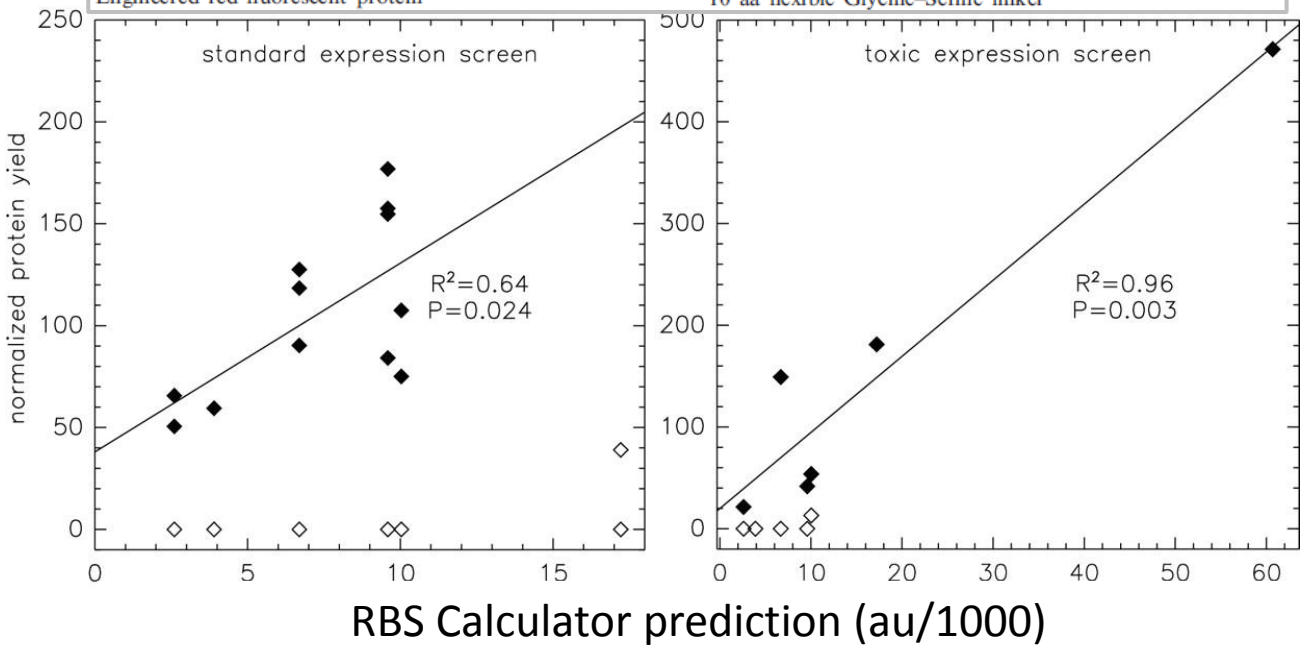
2 Expression modes:
Standard: T7pr + BL21-DE3 at 37°C
Toxic: T7pr + BL21-DE3-pLysE at 20 °C

all CDSs codon-optimized

Engineered leucine zipper	T7 promoter, RBS, start codon for expression in *E. coli*
Engineered leucine zipper	T7 terminator
FKBP12 (FK506-binding protein)	FLAG epitope tag (DYKDDDDK)
Engineered FKBP12-rapamycin-binding domain FRB(T2098L)	3-repeat FLAG epitope tag
Arabidopsis phototropin 1 LOV2 domain	Hexahistidine affinity tag
TEM-1 β-lactamase fragment 1	StrepII affinity tag
TEM-1 β-lactamase fragment 2	Glutathione S-transferase tag
Gaussia luciferase fragment 1	TEV protease cleavage site
Gaussia luciferase fragment 2	PreScission protease cleavage site
Engineered cyan fluorescent protein	2 aa flexible Glycine–Serine linker
Engineered yellow fluorescent protein	6 aa flexible Glycine–Serine linker
Engineered red fluorescent protein	10 aa flexible Glycine–Serine linker

A: Higher expressed proteins had higher predicted translation initiation rates

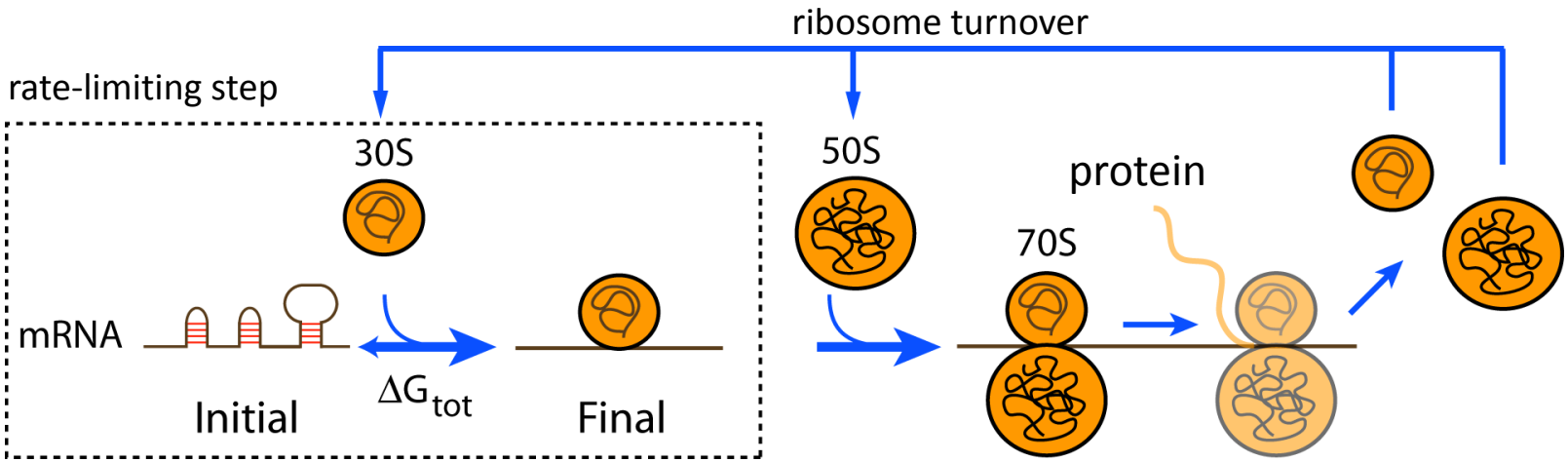
For soluble protein



Grunberg et. al., "Building blocks for protein interaction devices", Nucleic Acid Research, 2010

How does the RBS Calculator Work?

Translation is a multi-step process
Translation initiation is often the rate-limiting step



$$m + R \leftrightarrow m::R$$

RBS sequence on mRNA

30S subunits

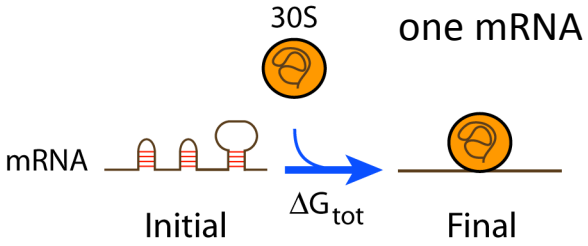
bound 30S subunits

$$\frac{m::R}{m\ R} = \exp\left(-\beta \Delta G_{tot}\right)$$

Boltzmann constant

Gibbs free energy change of ribosome binding

A Statistical Thermodynamic Model

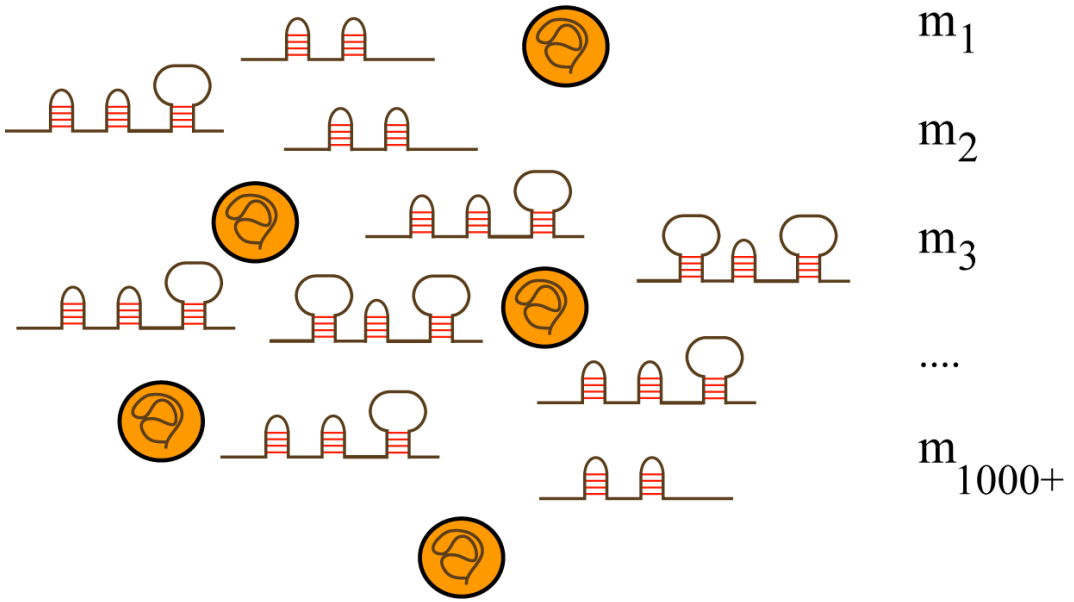


$$m_i + R \leftrightarrow m_i::R$$

$$\frac{m_i R}{m_i::R} = \exp\left(\beta \Delta G_{tot}^{(i)}\right)$$

$$r_i \propto \frac{m_i \exp\left(-\beta \Delta G_{tot}^{(i)}\right)}{1 + \sum_j m_j \exp\left(-\beta \Delta G_{tot}^{(j)}\right)}$$

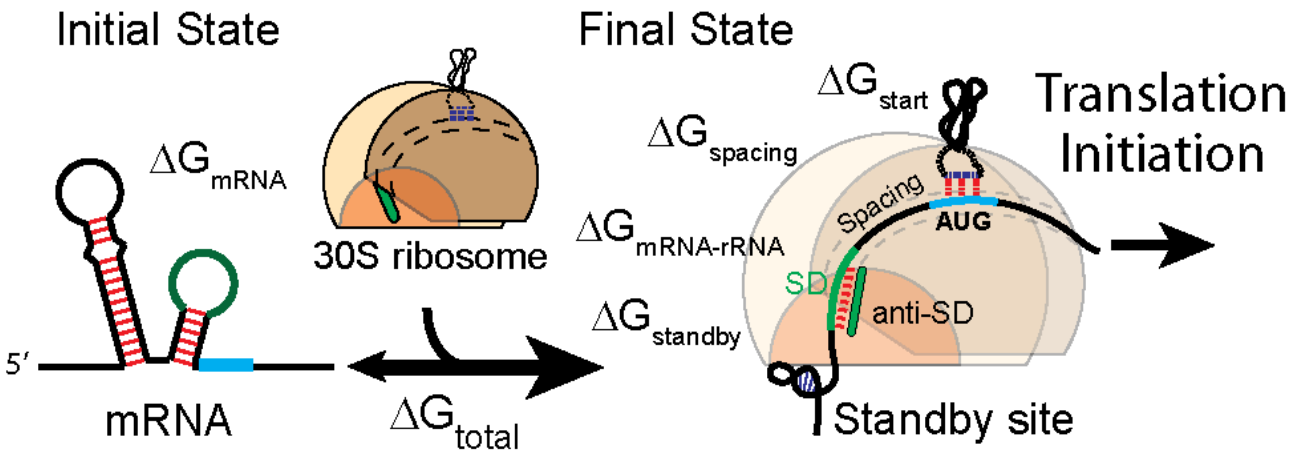
The "Ribosome Ensemble"
many mRNAs



$$r_{translation} \propto \exp\left(-\beta \Delta G_{tot}\right)$$

The Free Energy Model

We quantify the strengths of the molecular interactions controlling the ribosome's binding free energy to a mRNA sequence



$$\Delta G_{total} = \Delta G_{mRNA:rRNA} + \Delta G_{spacing} + \Delta G_{start} + \Delta G_{standby} - \Delta G_{mRNA}$$

$$\Delta G_{standby} = \Delta G_{distortion} + \Delta G_{unfolding} + \Delta G_{sliding}$$

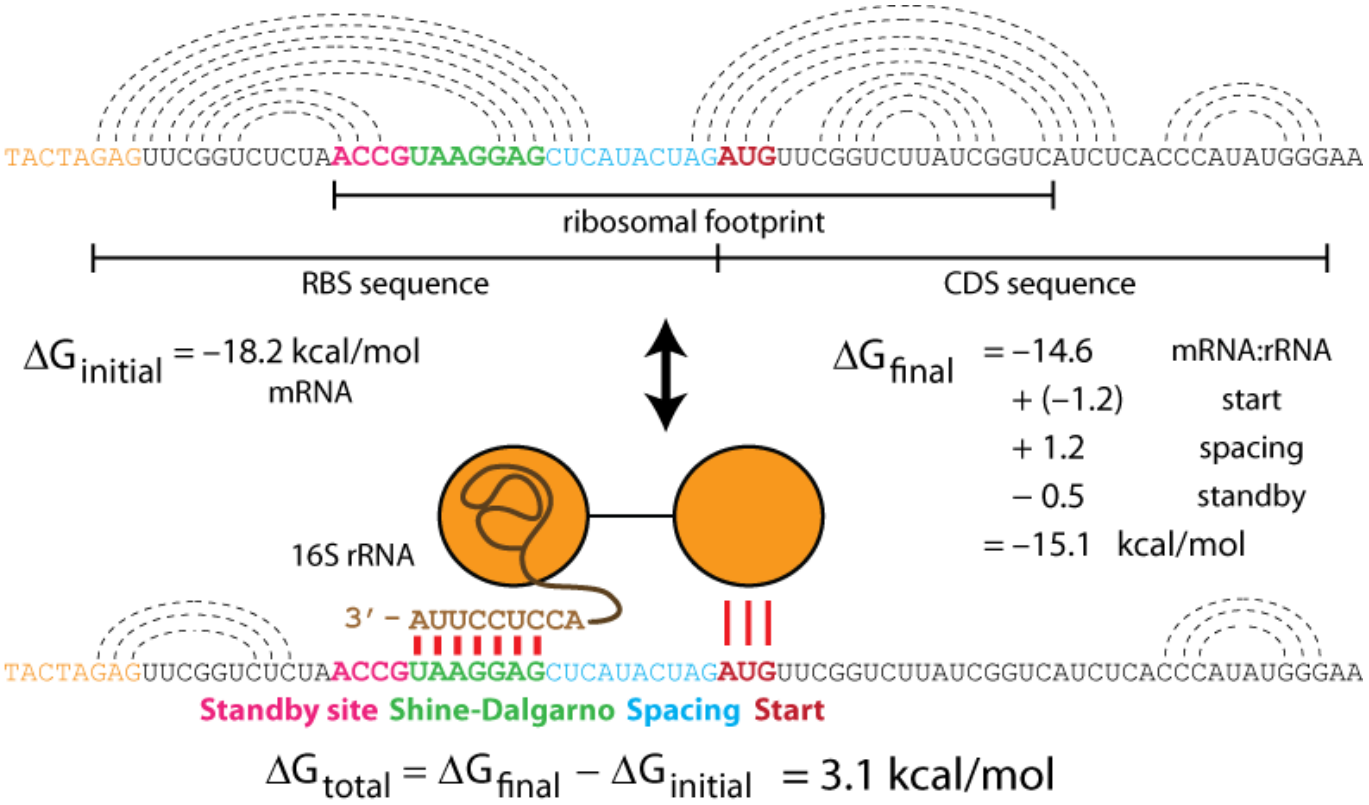
$$\rightarrow \quad r_{translation} \propto \exp\left(-\beta \Delta G_{tot}\right)$$

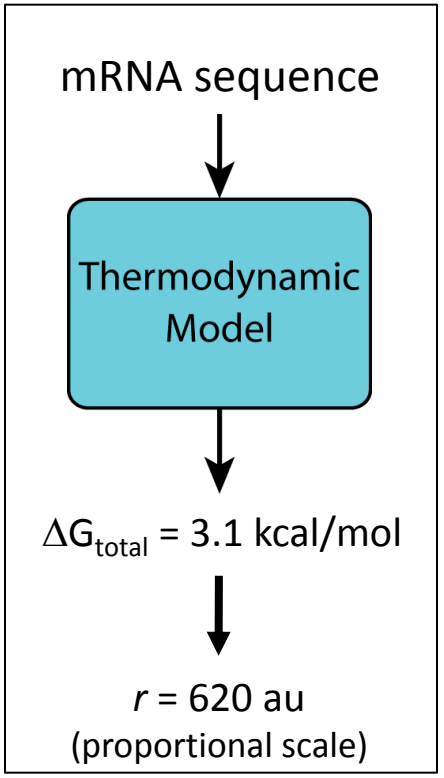
assumes thermodynamic equilibrium between a pool of ribosomes and a pool of mRNAs

An Example of Predicting Translation Rate

We can calculate the Gibbs free energy change when the ribosome binds to an mRNA and use statistical thermodynamics to predict its translation initiation rate

dotted lines = nucleotide base pairing

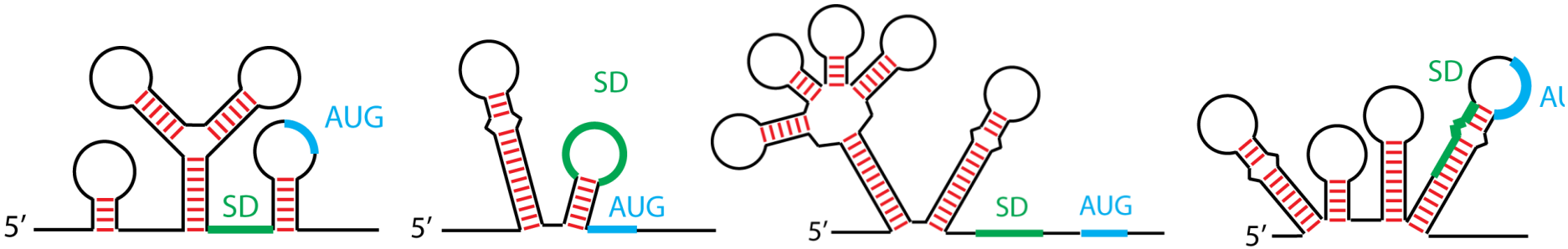




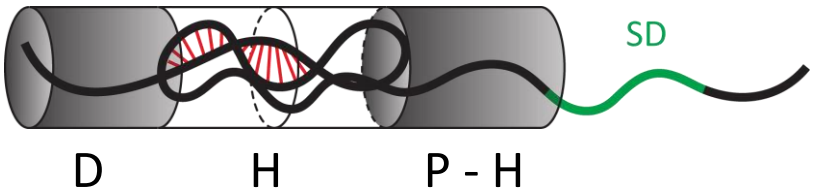
Important Questions, Systematic Answers

Q: How does the ribosome interact with highly structured 5' UTRs?

A: The ribosome's platform domain binds to single-stranded regions in the standby site



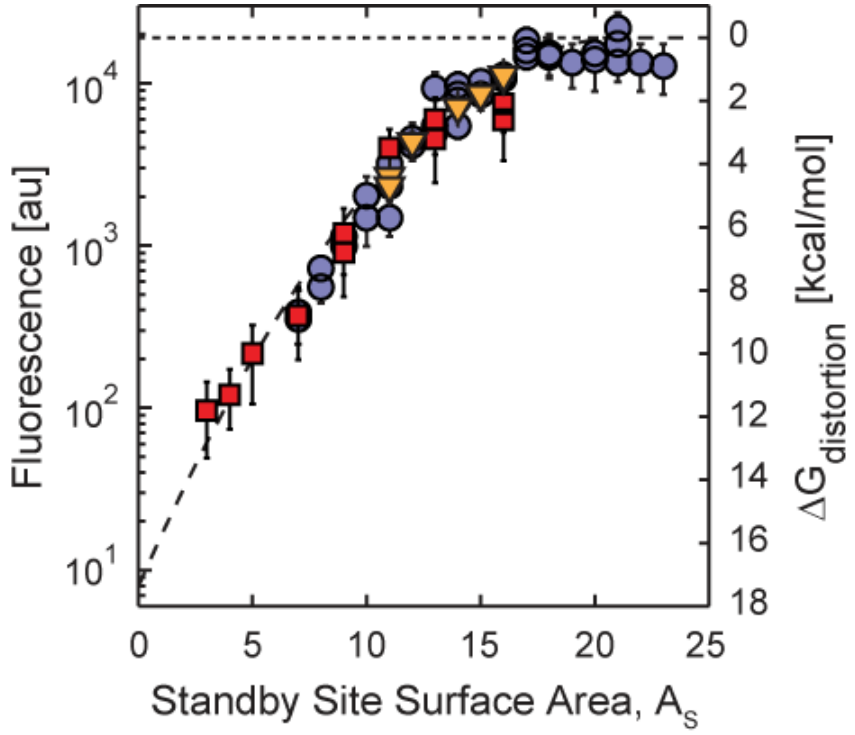
the standby site's single-stranded surface area controls the ribosome's ability to bind



$$A_s = 15 + P + D - H$$

D, P, and H were inter-changeable

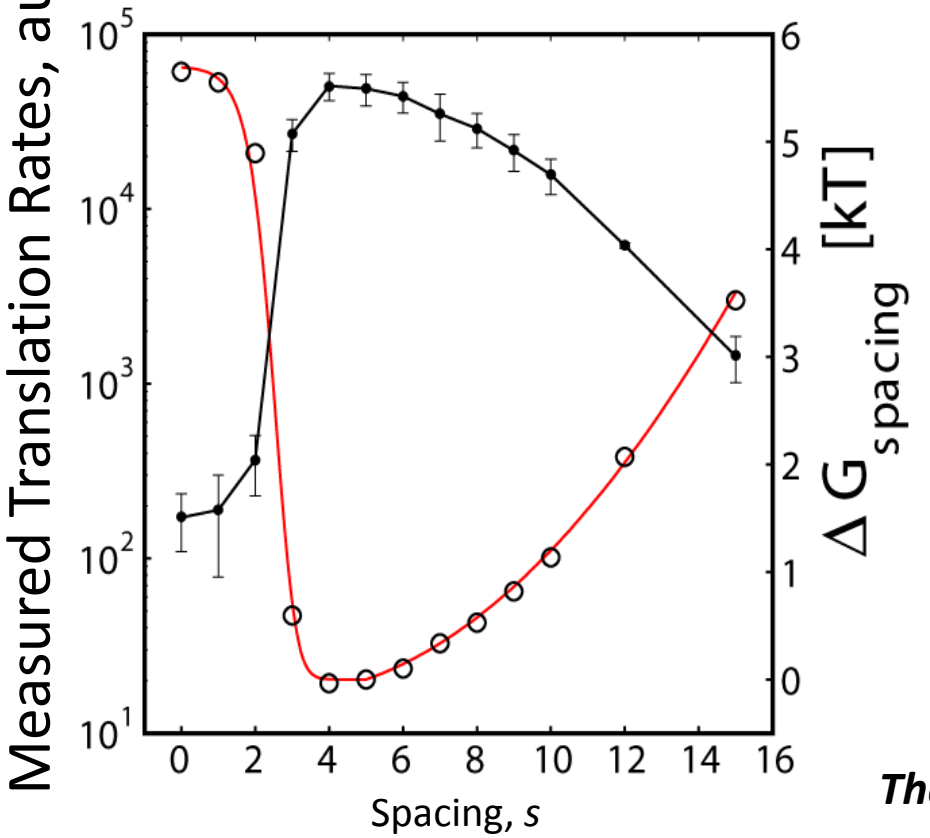
Only the standby site's surface area matters



Espah Borujeni et. al., *Nucleic Acid Research*, 2013

Important Questions, Systematic Answers

Q: What is the effect of the spacer region length?
A: If spacing is too long, then translation rate is reduced.
Too short? Translation rate is also reduced. Optimal spacer region is 5 nt.



ACACAUAAGGAGGUAAUG *s* = 1 nt

ACACAUAAGGAGGUACAUG

ACACAUAAGGAGGUACACAUG

ACACAUAAGGAGGUACACACAUG

ACACAUAAGGAGGUACACACACAUG

ACACAUAAGGAGGUACACACACACAUG

ACACAUAAGGAGGUACACACACACACAUG

s = 12 nt

The Ribosome is Stretched like a Rigid Spring
$\Delta G_{spacing}$ looks like Hooke's Law

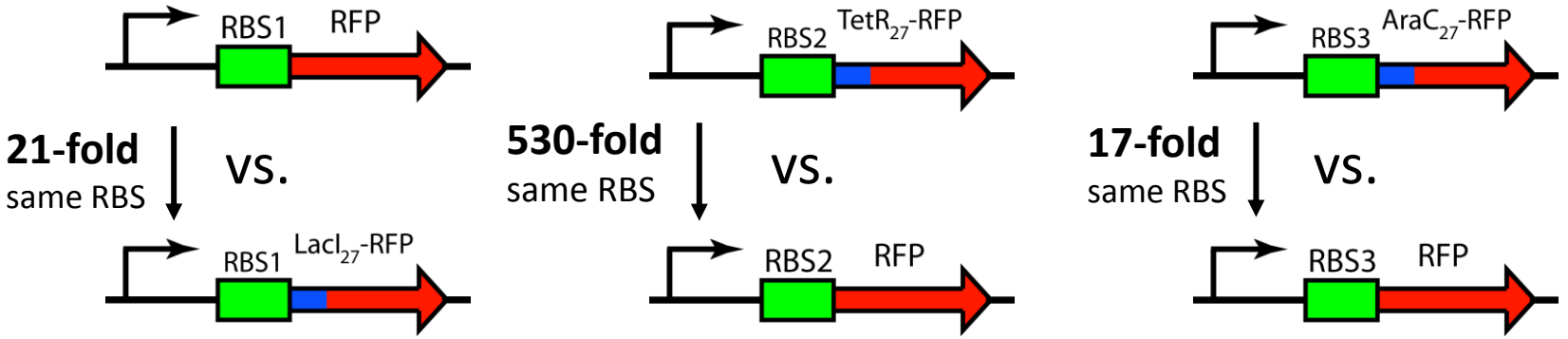
E. coli, flow cytometry, steady-state measurements, M9 media + 0.4% glucose

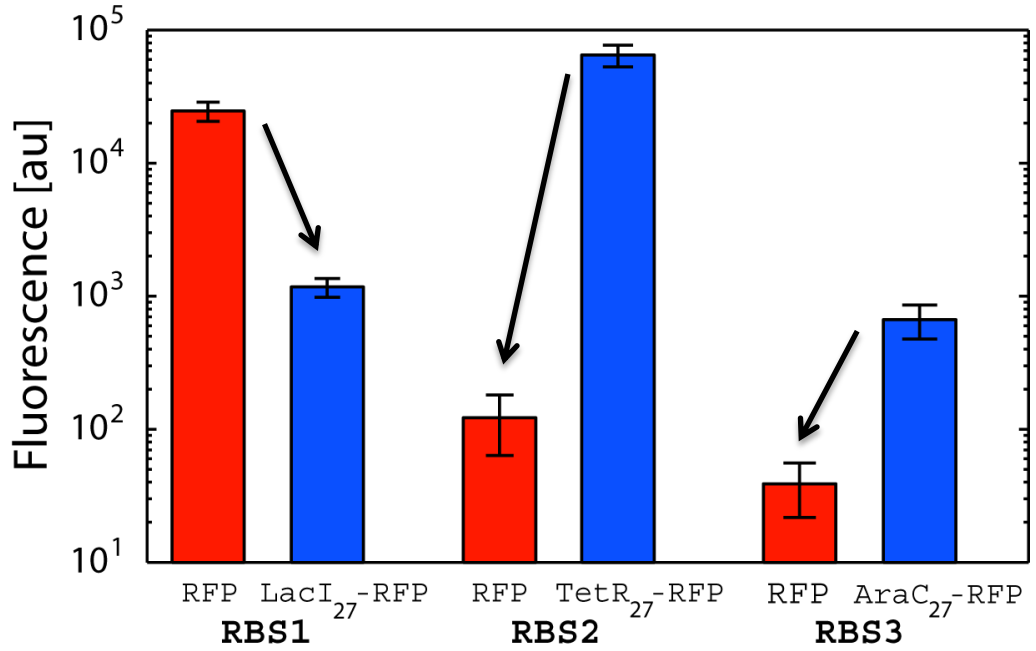
Salis et. al., *Nature Biotech.* , 2009

Important Questions, Systematic Answers

Q: Are Ribosome Binding Sites a "modular" genetic part?

A: No, re-using the same RBS sequence with different CDS sequences can affect translation





E. coli, flow cytometry, steady-state measurements, M9 media + 0.4% glucose

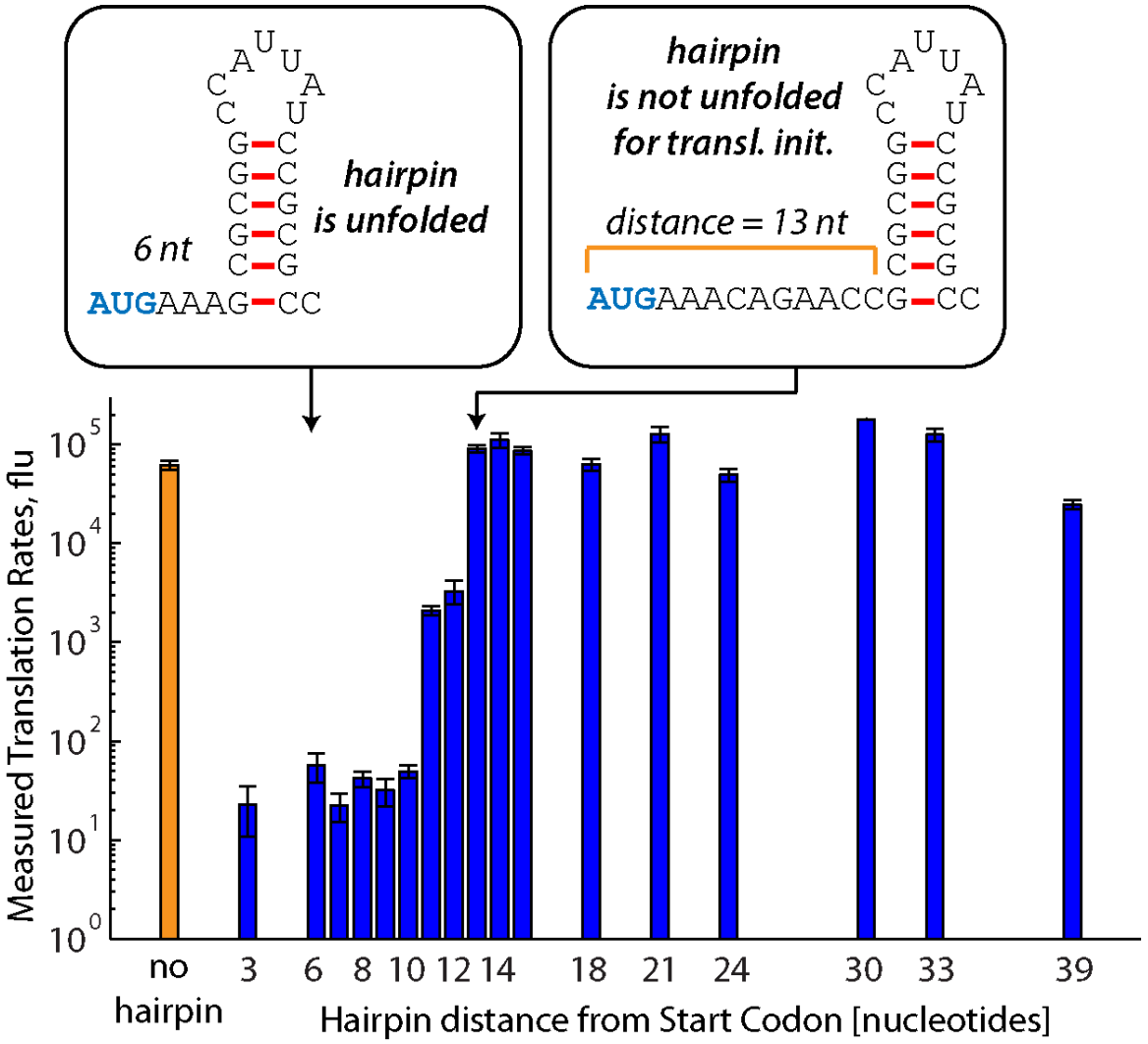
Salis et. al., *Nature Biotech.* , 2009

Important Questions, Systematic Answers

Q: Why does the CDS sequence affect translation rate?

***A:* mRNA structures form inside the CDS and between the RBS-CDS. These structures inhibit translation.**

We measured the ribosome's footprint to precisely predict which mRNA structures will inhibit translation.



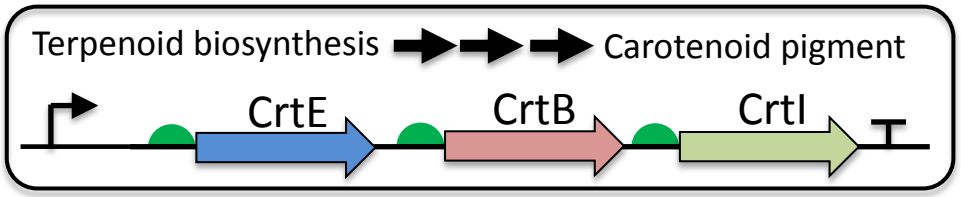
E. coli, flow cytometry, steady-state measurements, M9 media + 0.4% glucose

Espah Borujeni et. al., *in prep*

Optimizing Multi-Protein Genetic Systems

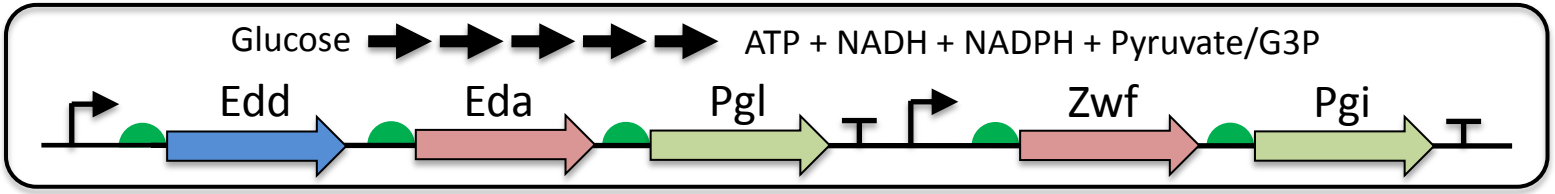
Multi-protein genetic systems are more difficult to engineer

Metabolic Pathway Engineering Examples from the Salis Lab

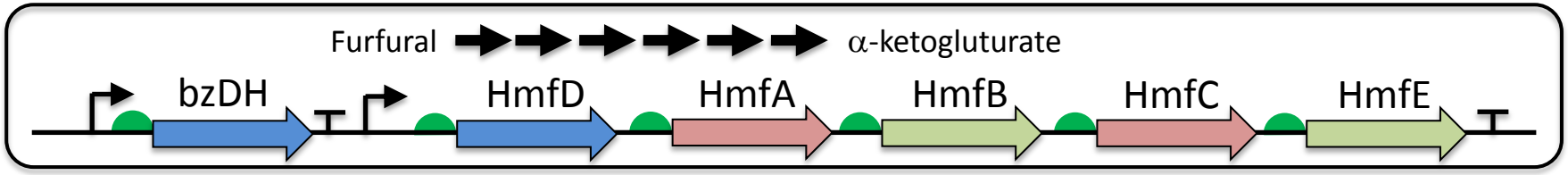


a 3-enzyme terpenoid biosynthesis pathway

a 5-enzyme Entner-Doudoroff pathway



a 6-enzyme furfural catabolic pathway for detoxifying lignocellulosic feedstock

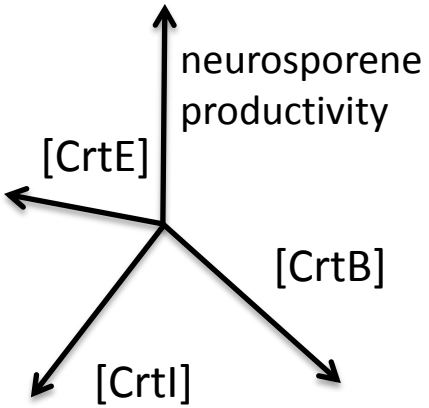


Farasat et. al. *Molecular Systems Biology*, 2014

Chiam Yu Ng et. al., *Metabolic Engineering*, 2015

Optimizing Multi-Protein Genetic Systems

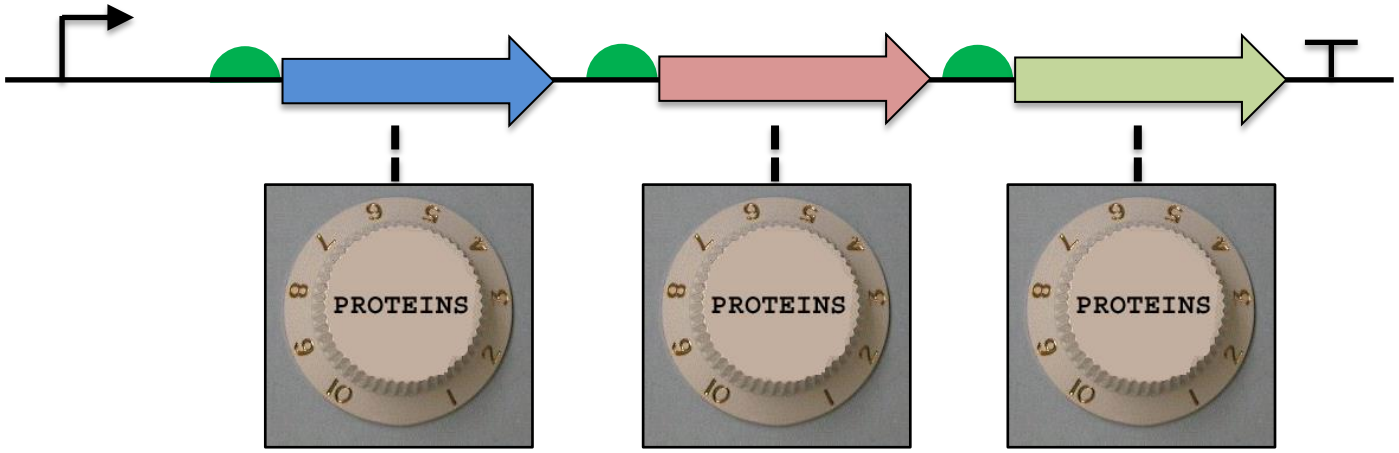
How do we efficiently search for optimal protein expression levels?



Q: What are the enzyme expression levels that will maximize the pathway's productivity?

Bigger Q: What is the relationship between RBS sequence, expression, and pathway productivity?

3 enzymes = 4-dimensional space



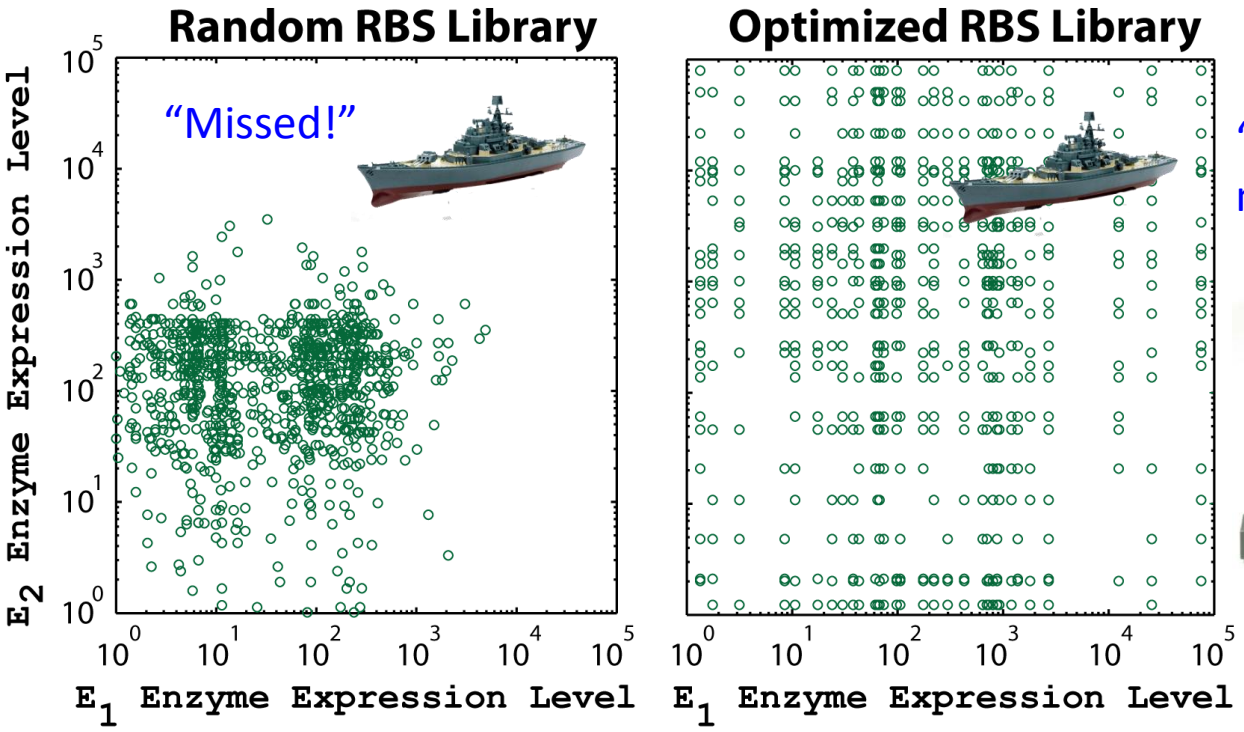
Farasat et. al. "Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria", *Molecular Systems Biology*, 2014

Optimizing Multi-Protein Genetic Systems

Systematic metabolic pathway optimization requires three ingredients:

1. the ability to quantitatively predict and control enzyme expression
2. an efficient way to search for optimal enzyme expression levels

Which search strategy is better?



"You've sunk my battleship!"

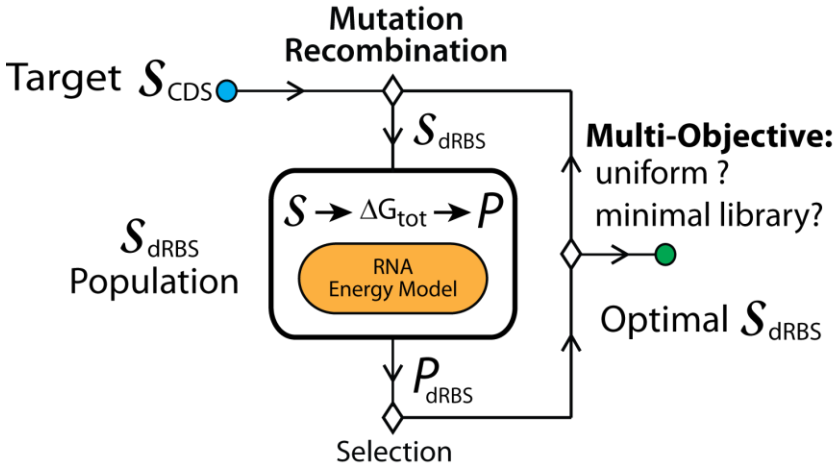
Optimizing Multi-Protein Genetic Systems

Systematic metabolic pathway optimization requires three ingredients:

1. the ability to quantitatively predict and control enzyme expression
2. an efficient way to search for optimal enzyme expression levels

RBS Library Calculator

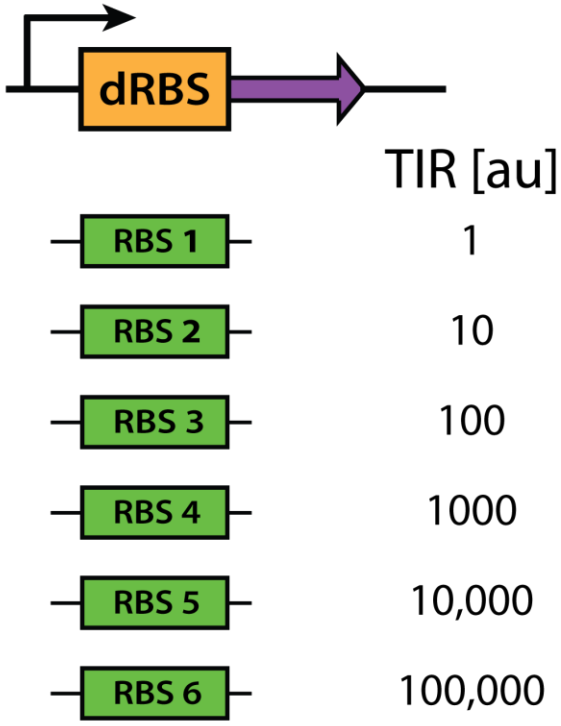
uniform sampling of protein expression levels



ACGAUCACAACUU**S**ATG**D**CGUAUUC

optimized, degenerate RBS: 6 variants

an Optimized RBS Library



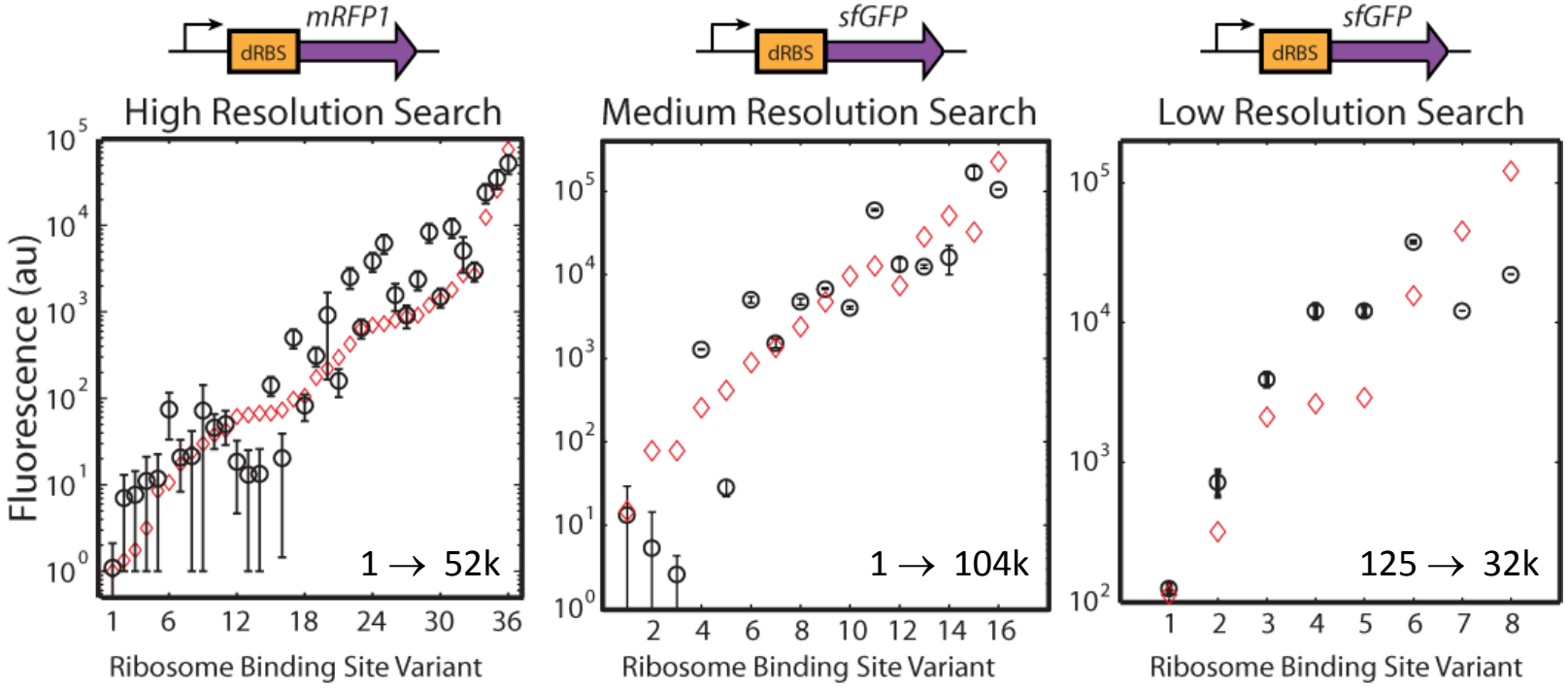
Optimizing Multi-Protein Genetic Systems

Systematic metabolic pathway optimization requires three ingredients:

1. the ability to quantitatively predict and control enzyme expression
2. an efficient way to search for optimal enzyme expression levels

RBS Library Calculator

uniform sampling of protein expression levels



Farasat et. al., *Molecular Systems Biology*, 2014

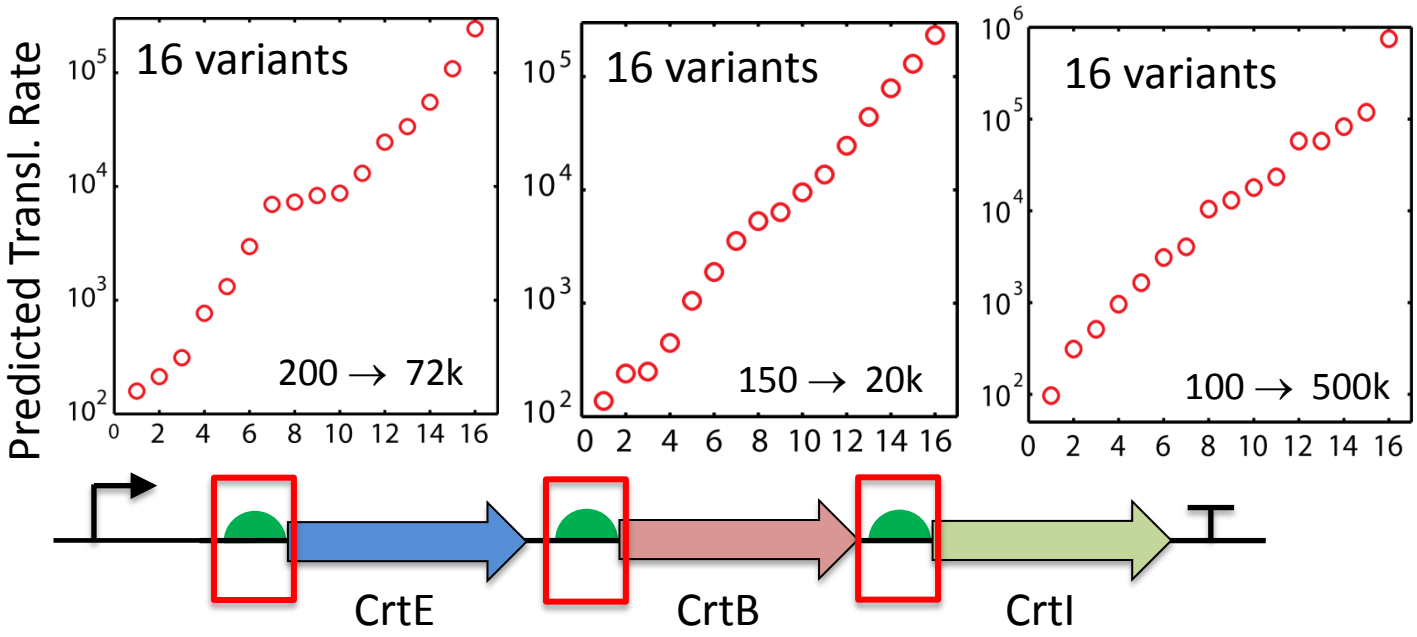
E. coli DH10B, M9 + 0.4% glucose, flow cytometry, n = 3+

Optimizing Multi-Protein Genetic Systems

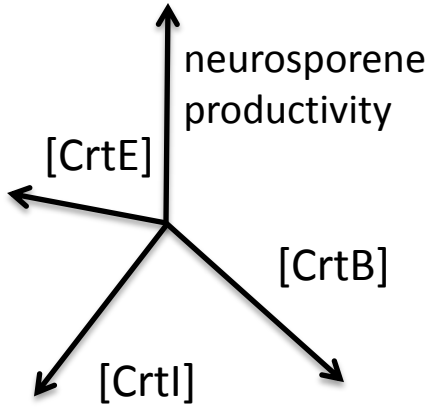
Systematic metabolic pathway optimization requires three ingredients:

1. the ability to quantitatively predict and control enzyme expression
2. an efficient way to search for optimal enzyme expression levels
3. the systematic mapping of the expression-activity relationship

optimized RBS libraries for efficient search



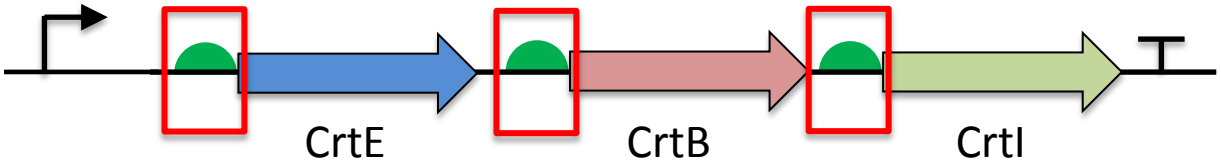
4-dimensional space

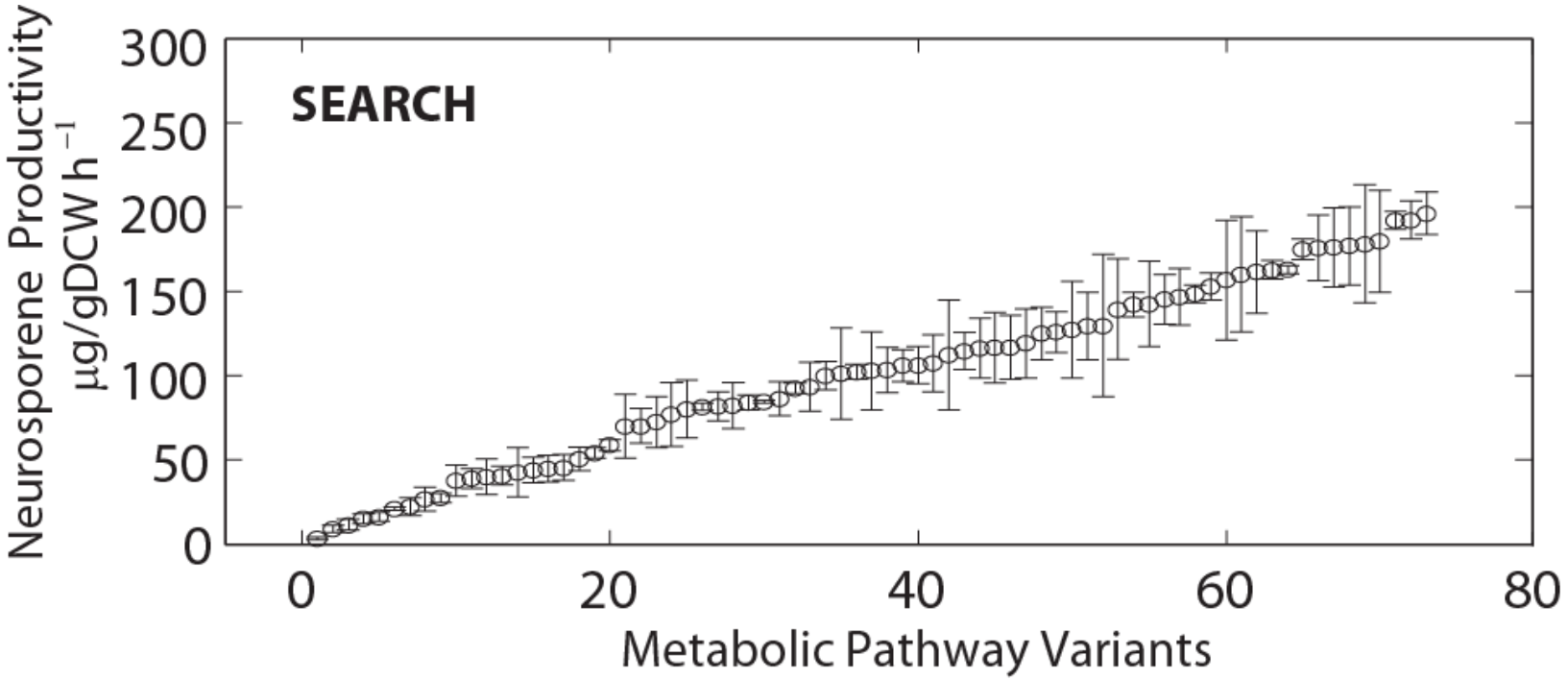


Farasat et. al., *Molecular Systems Biology*, 2014

Pathway Productivity Measurements

We randomly selected 73 pathway variants with different RBS sequences and found a wide range of pathway productivities

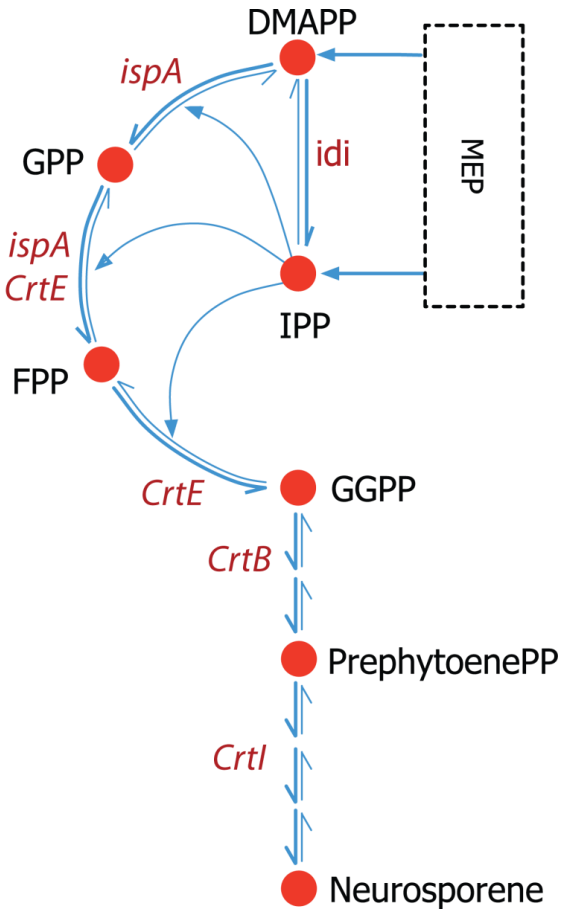




E. coli EcHW2f overnight, washed, and then grown in LB + 10 mM arabinose for **7 hours** in **stationary** phase neurosporene extracted with hot acetone and quantified at 470 nm, n = 3

Kinetic Modeling to Predict Pathway Activity

DNA sequence → Translation Rates → (Prediction?) → Pathway Productivity



what we have

73 measurements of predicted translation rates of CrtE, CrtB, and CrtI vs. measured pathway productivities (not a simple pattern)

what we want

a ***continuous model*** that relates
RBS sequences to enzyme expression levels,
and enzyme expression levels to pathway fluxes

If we have a model

we can predict optimal enzyme expression levels
to achieve maximal pathway fluxes
and apply other application-specific constraints

CM1, CM2, . . ., CM14: Enzyme complexes

Kinetic Modeling to Predict Pathway Activity

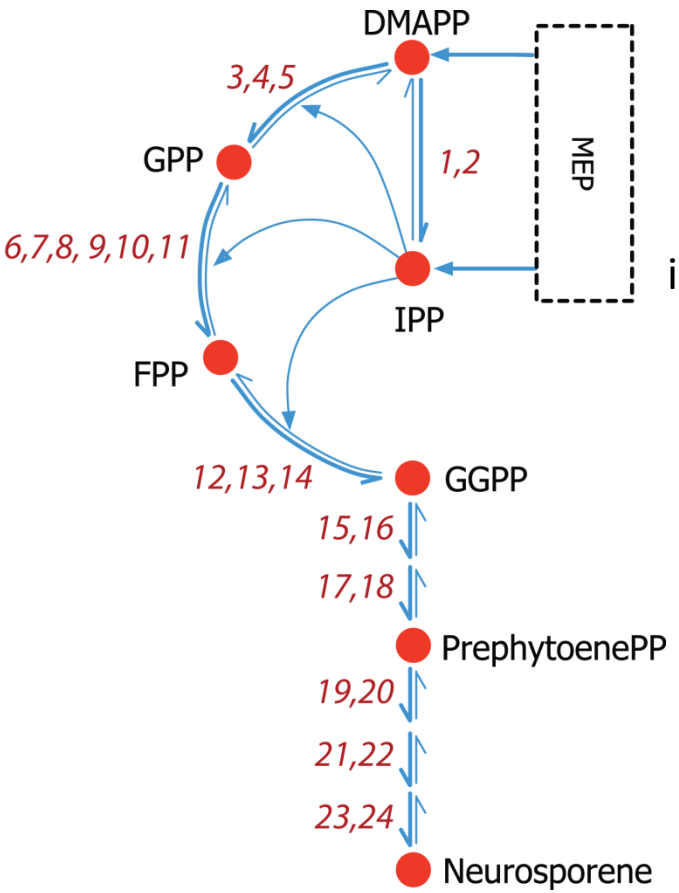
24 elementary reactions

48 unknown k

$$\frac{dFPP}{dt} = +k_1[GPP][ispA] + k_3[GPP][CrtE]$$

$$\frac{dGGPP}{dt} = k_3[crtE - FPP - complex] - k_4[GGPP][CrtE]$$

Mole balance equations



Model Reduction: 38 bounded DOFs, 73 measurements

identified kinetic parameter values using a genetic algorithm

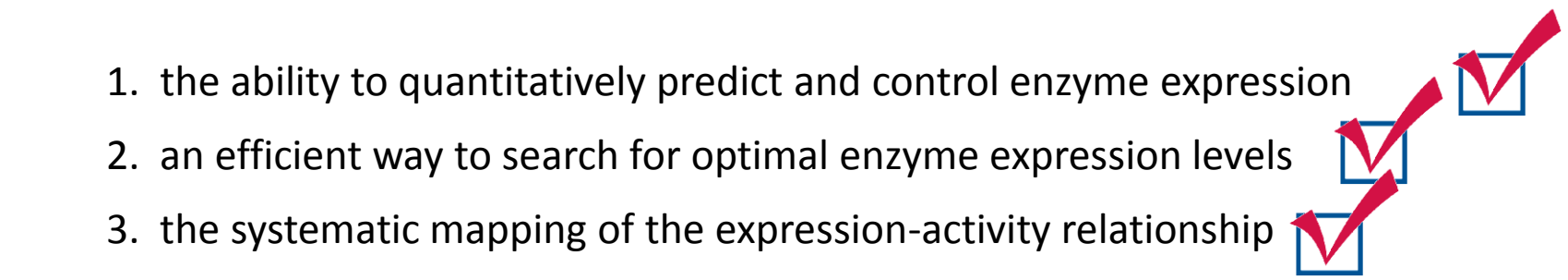
Rxn no.	$K_{+(forward)}$	$K_{-(reverse)}$	Rxn no.	$K_{+(forward)}$	$K_{-(reverse)}$
1	3.72	8.62E-06	13	233	6.44
2	0.92	0.25	14	237	0.84
3	1.46	5.12E-05	15	15.75	9007
4	81.84	272.48	16	2966	1.91
5	20.76	0.21	17	2.54	0.05
6	0.86	0.09	18	13.49	17.22
7	12569	3.18	19	2.54	44945
8	16.05	0.59	20	51793	0.71
9	4.56	0.79	21	4.13	1471
10	0.54	2.59E-04	22	1427	1.97
11	39.42	0.97	23	4.72	119
12	146	7008	24	1062	32.02

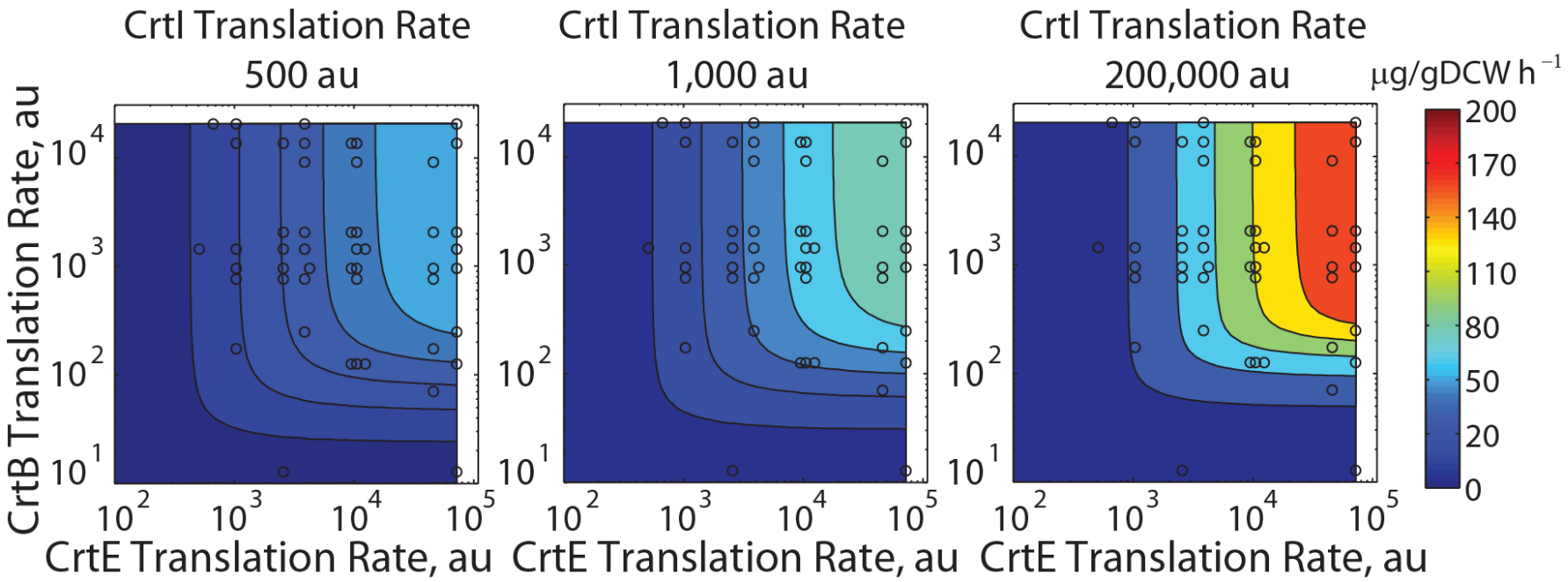
CM1, CM2, . . ., CM14: Enzyme complexes

A Sequence-Expression-Activity Map

Systematic metabolic pathway optimization requires three ingredients:

1. the ability to quantitatively predict and control enzyme expression
2. an efficient way to search for optimal enzyme expression levels
3. the systematic mapping of the expression-activity relationship

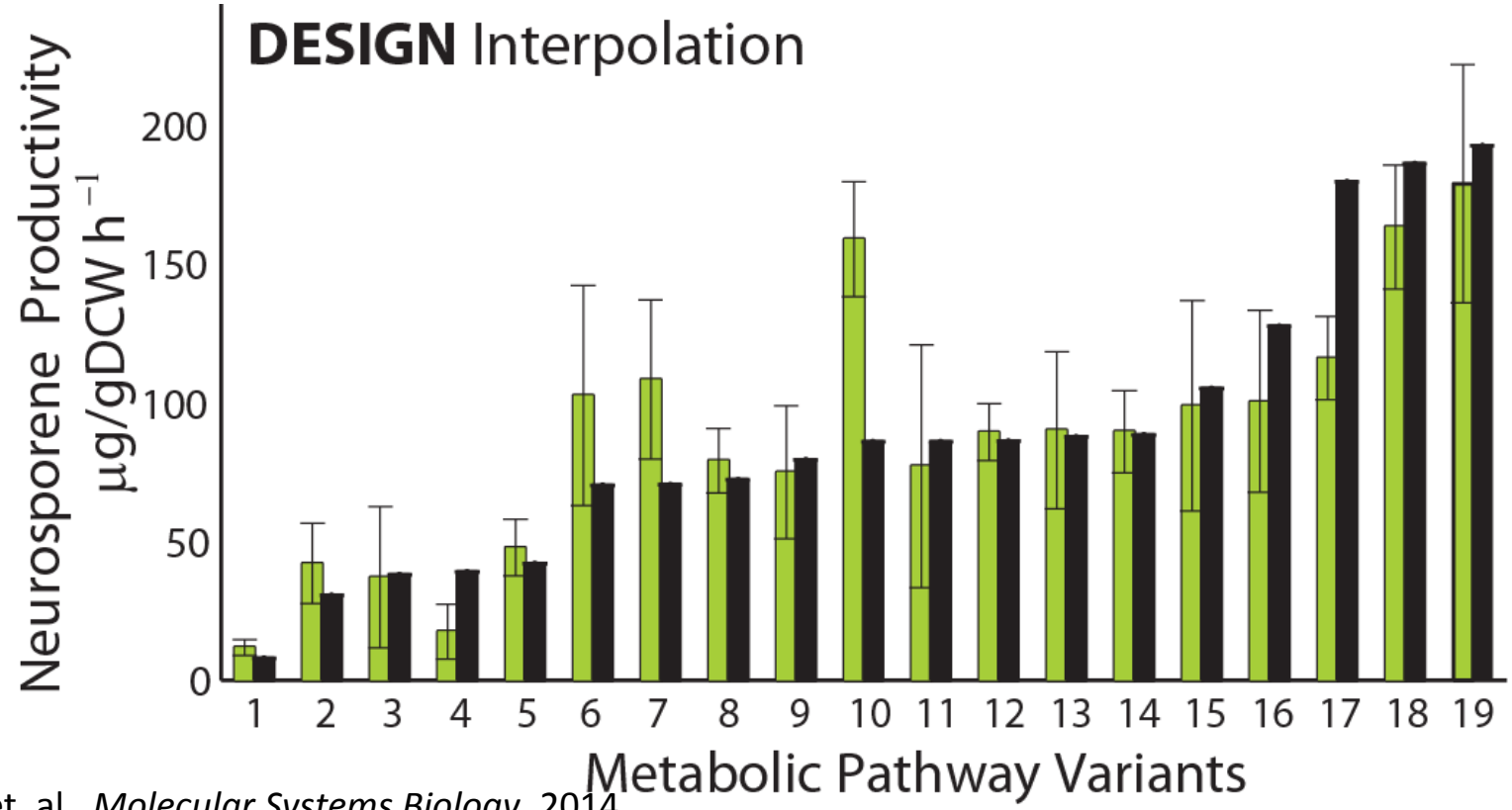




Sequence-Expression-Activity Map Accuracy

DNA sequence → (Thermodynamic model) → Translation Rates → (Kinetic model) → Pathway Productivity

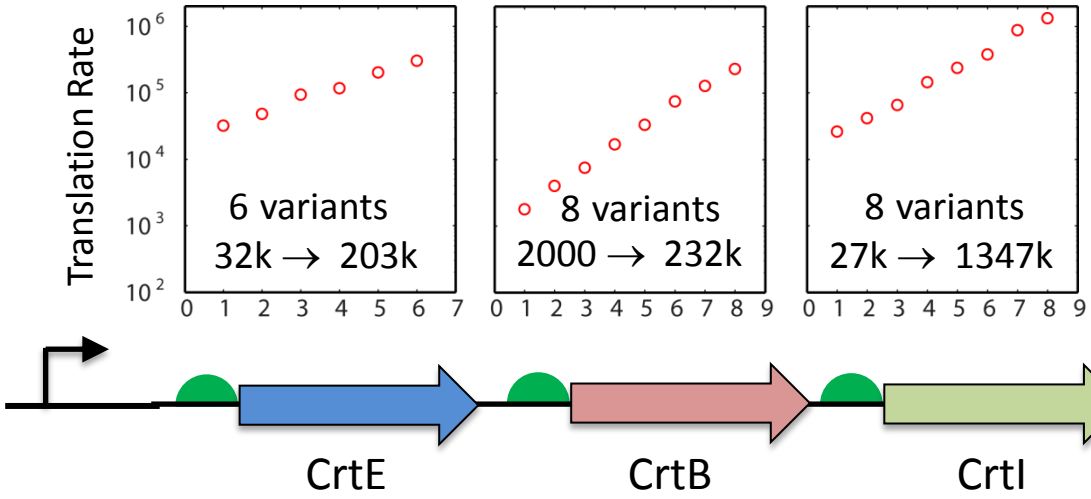
We used the SEA Map to design 19 new pathway variants
The predicted pathway activities were within 28% of the measurements, on average

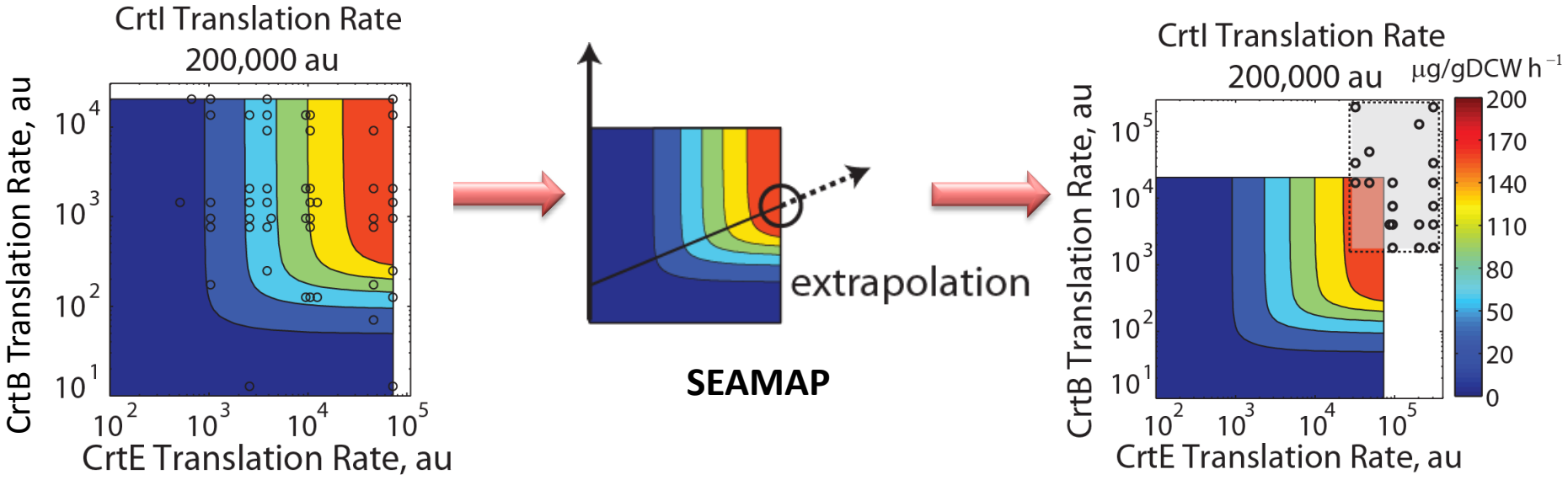


Farasat et. al., *Molecular Systems Biology*, 2014

Efficient Pathway Optimization

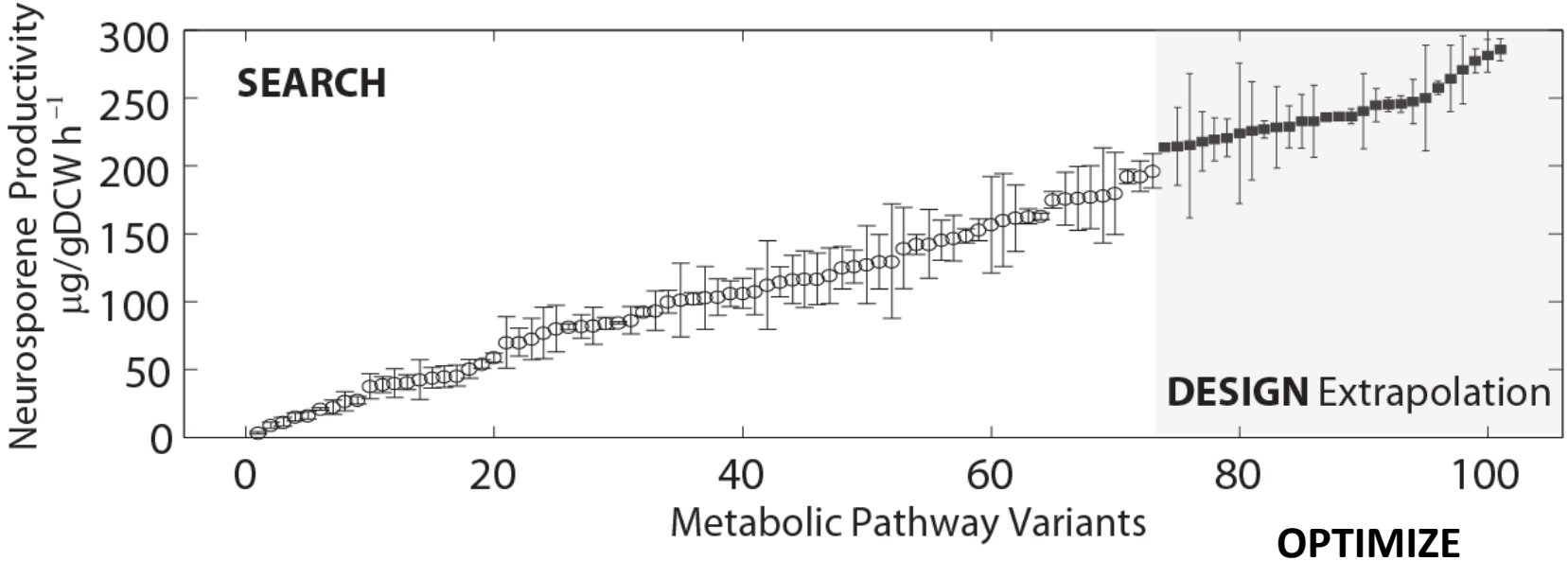
We used the SEA Map to predict the optimal enzyme expression levels
We applied the RBS Library Calculator again to "Zoom In" on the optimal levels



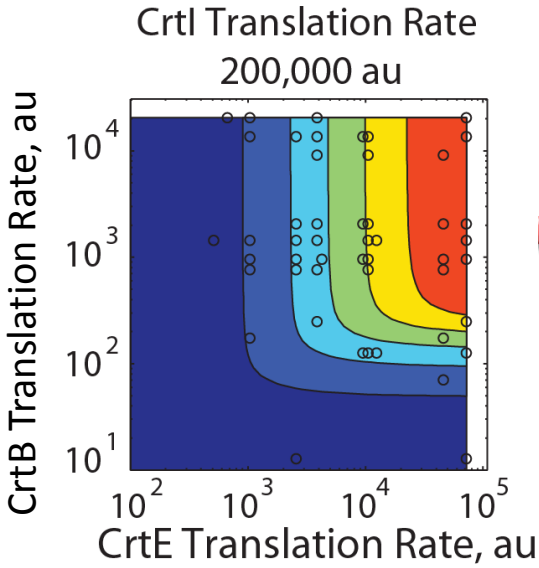


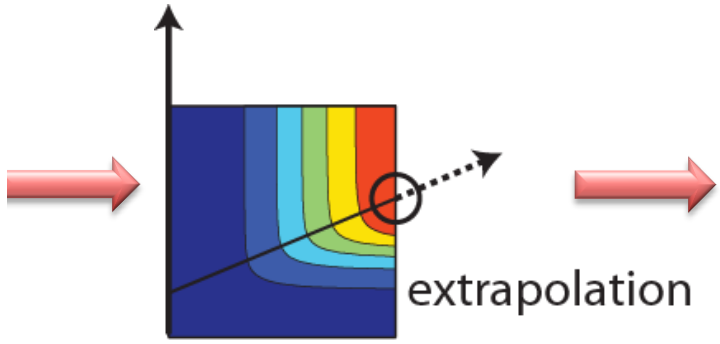
Efficient Pathway Optimization

1. SEARCH. 2. MAP. 3. OPTIMIZE.

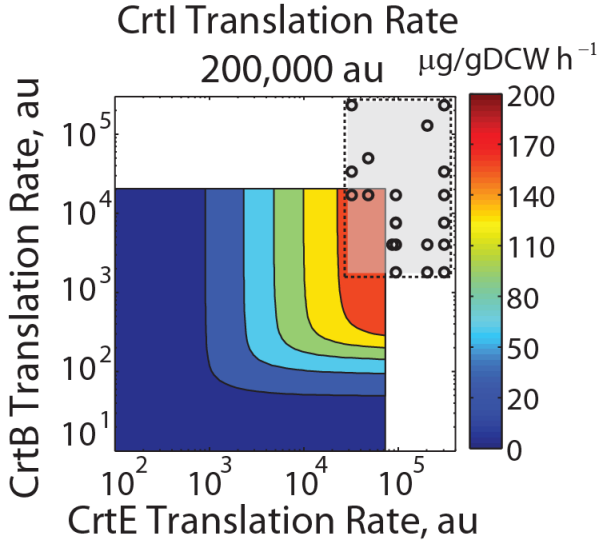


OPTIMIZE





MAP



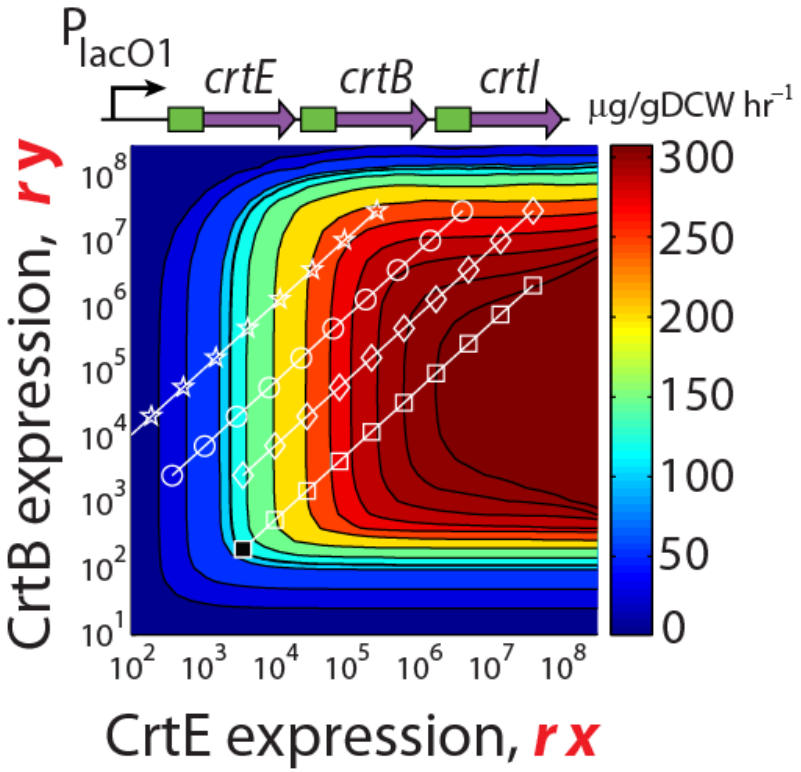
Using the Sequence-Expression-Activity Map

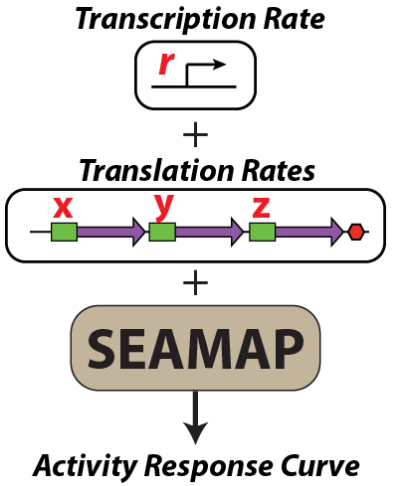
Finding the Best Transcription & Translation Rate with a small number of experiments

Promoter → Transcription rate

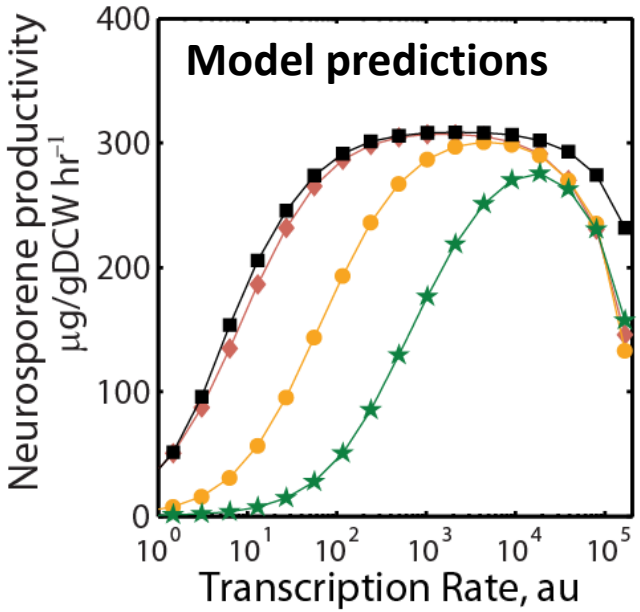
Ribosome Binding Site → Translation rate

SEAMAP → Pathway productivity





Activity Response Curve



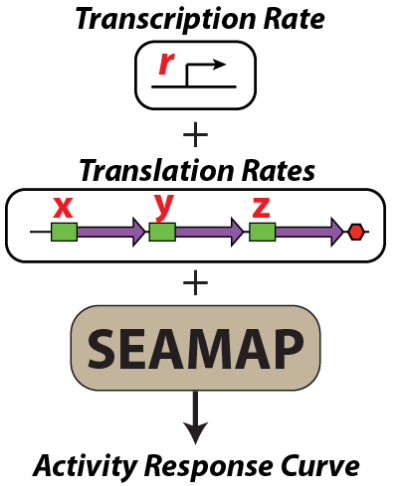
Using the Sequence-Expression-Activity Map

Finding the Best Transcription & Translation Rate with a small number of experiments

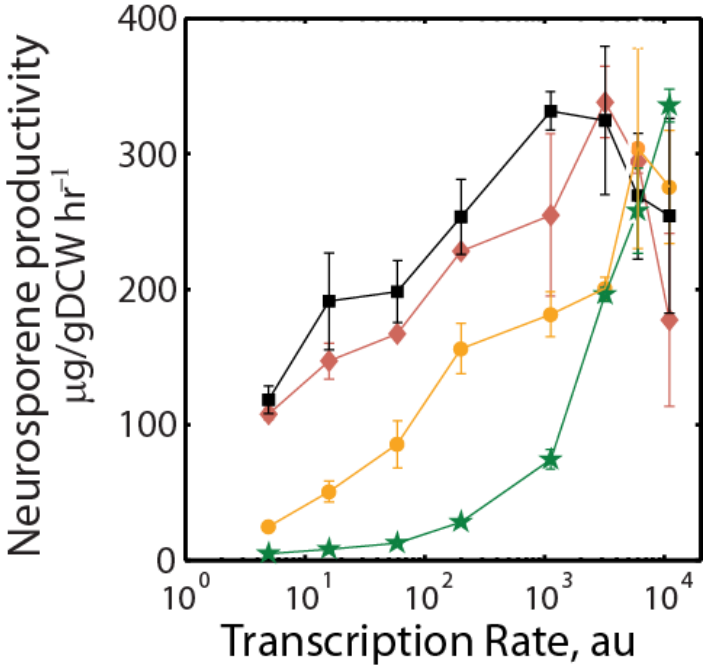
Promoter → Transcription rate

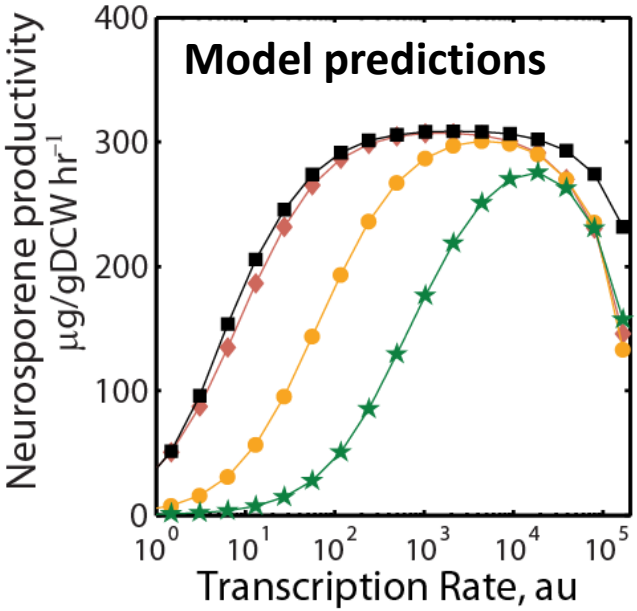
Ribosome Binding Site → Translation rate

SEAMAP → Pathway productivity



Experimental Measurements



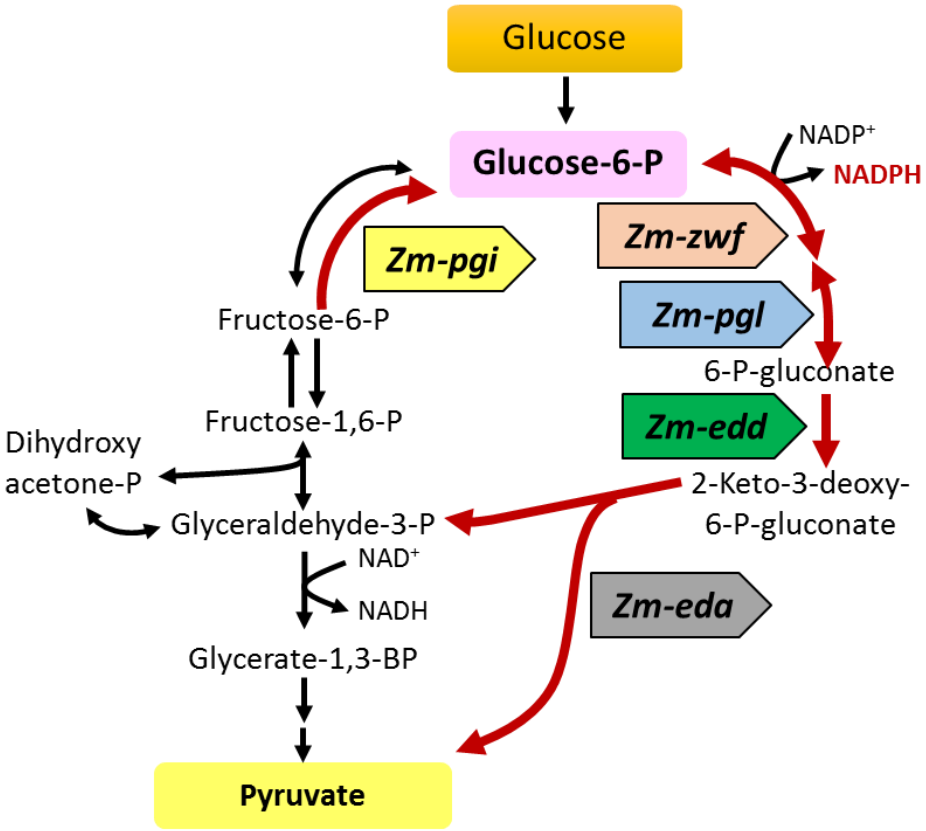


Pathway Engineering Example #2

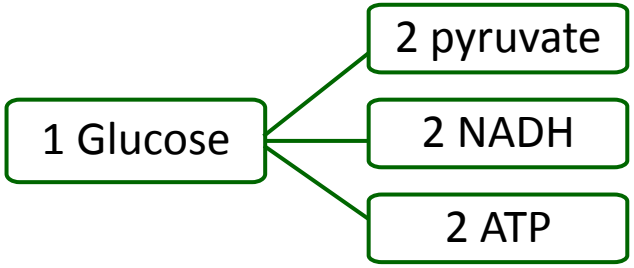
The Entner-Doudoroff pathway regenerates NADPH

Overcoming NADPH limitation is essential to over-producing fuels & materials

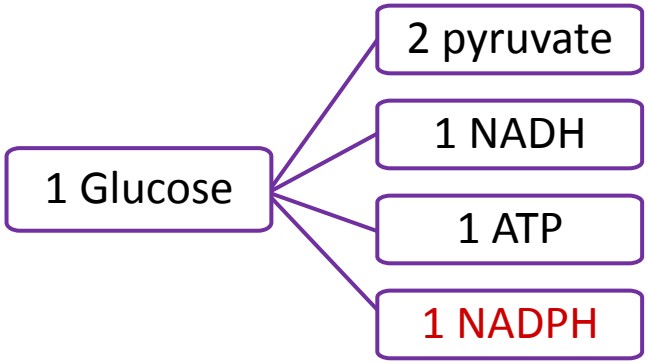
5 minimally allosteric enzymes from Z. mobilis now expressed in E. coli



Embden–Meyerhoff–Parnas (EMP) pathway



Entner-Doudoroff (ED) pathway

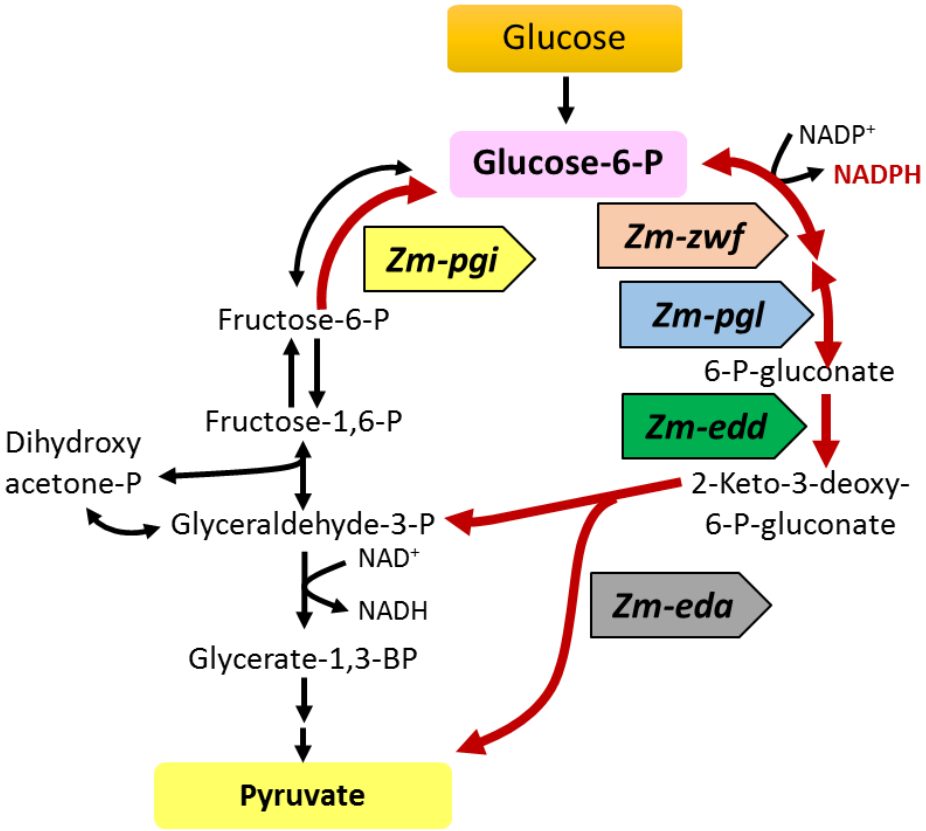


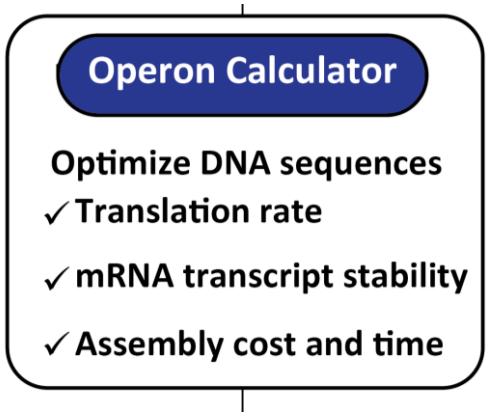
Pathway Engineering Example #2

The Entner-Doudoroff pathway regenerates NADPH

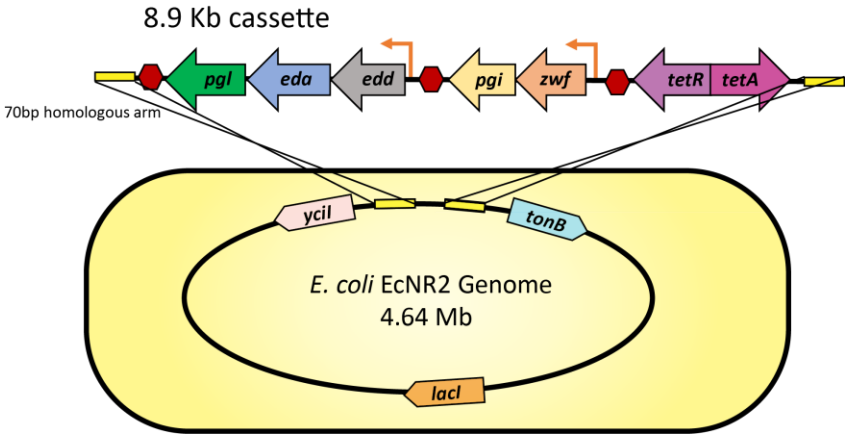
Overcoming NADPH limitation is essential to over-producing fuels & materials

5 minimally allosteric enzymes from Z. mobilis now expressed in E. coli





Rationally designed operons integrated into the E. coli genome



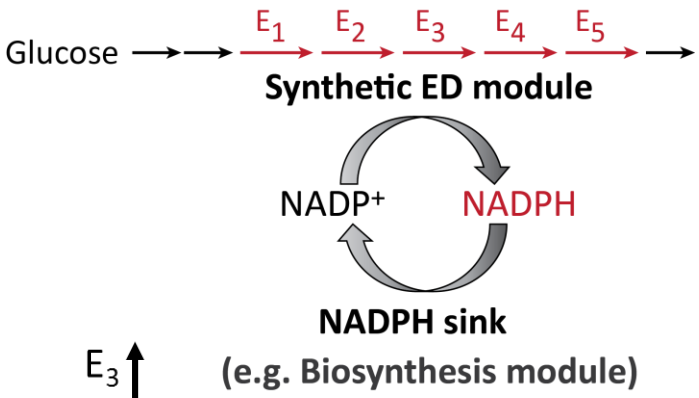
5-Dimensional Pathway Optimization

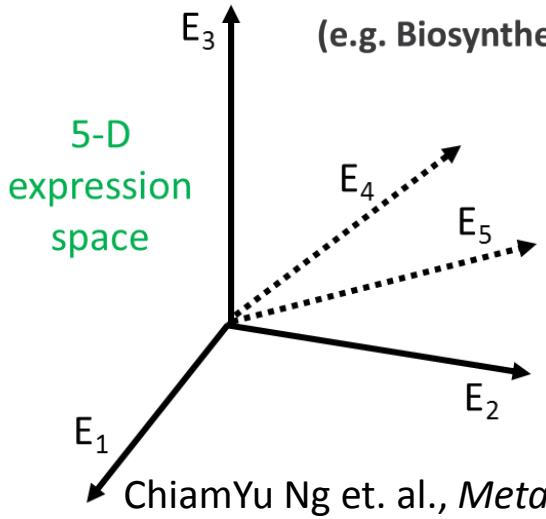
rational design + genome mutagenesis

RBS Library Calculator

+

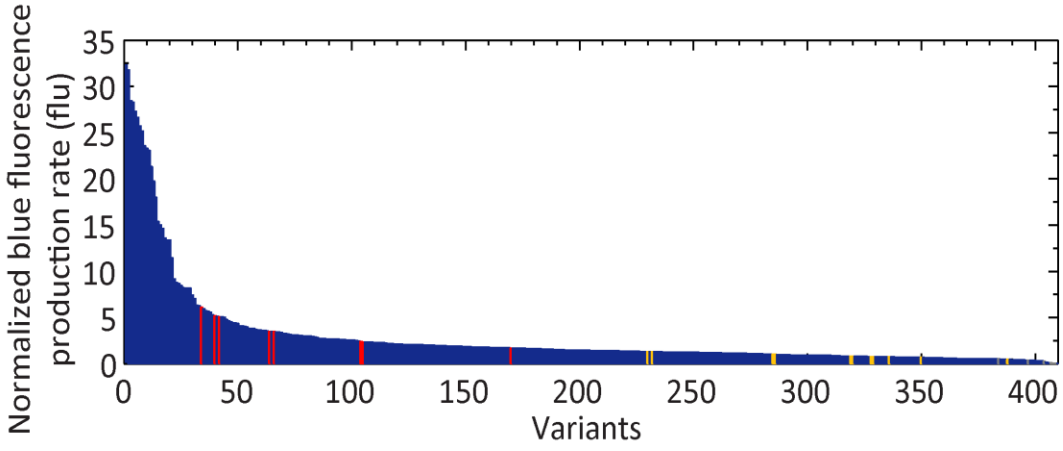
MAGE (Wang et. al, Nature, 2009)

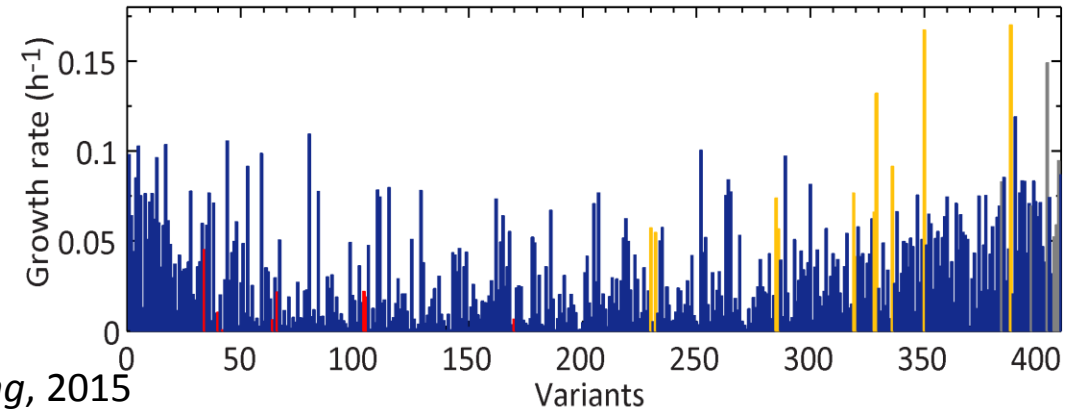




a high-throughput reporter for NADPH levels

The best pathway variant increased NADPH by 25-fold





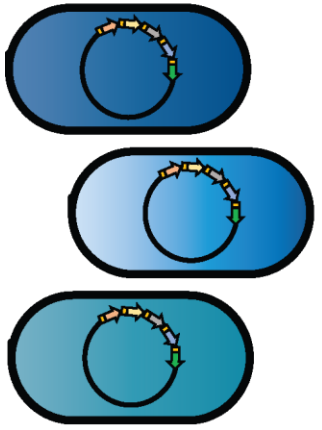
ChiamYu Ng et. al., *Metabolic Engineering*, 2015

Sequencing Pathway Variants from the Library

RBS Sequences → Predicted Translation Rates

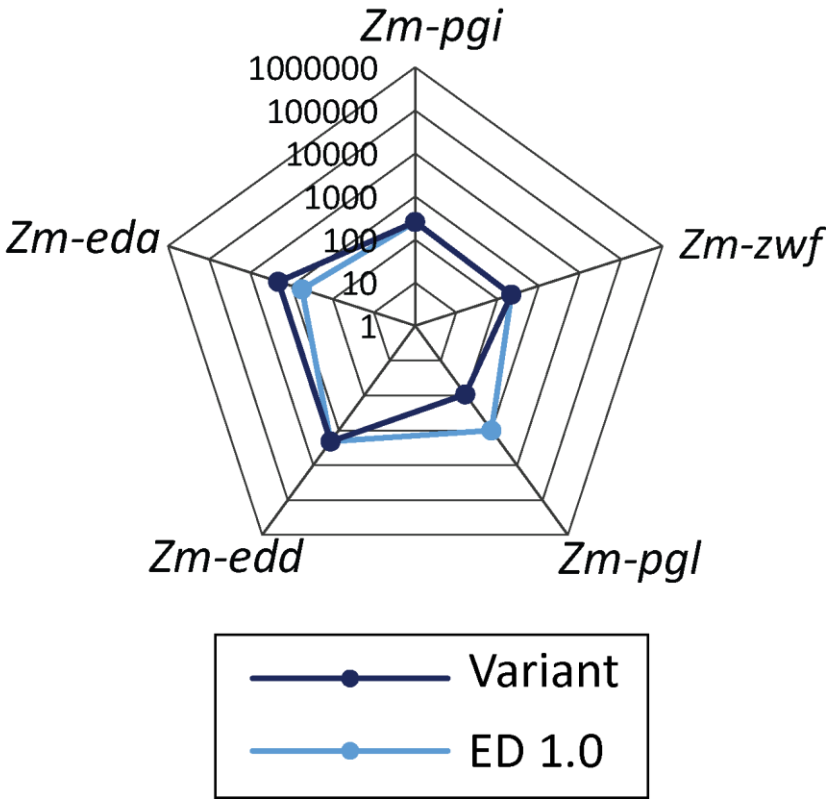
Selected 23 pathway variants with varying mBFP production rates

Selected variants



Sequencing

predicted translation rates for all five enzymes

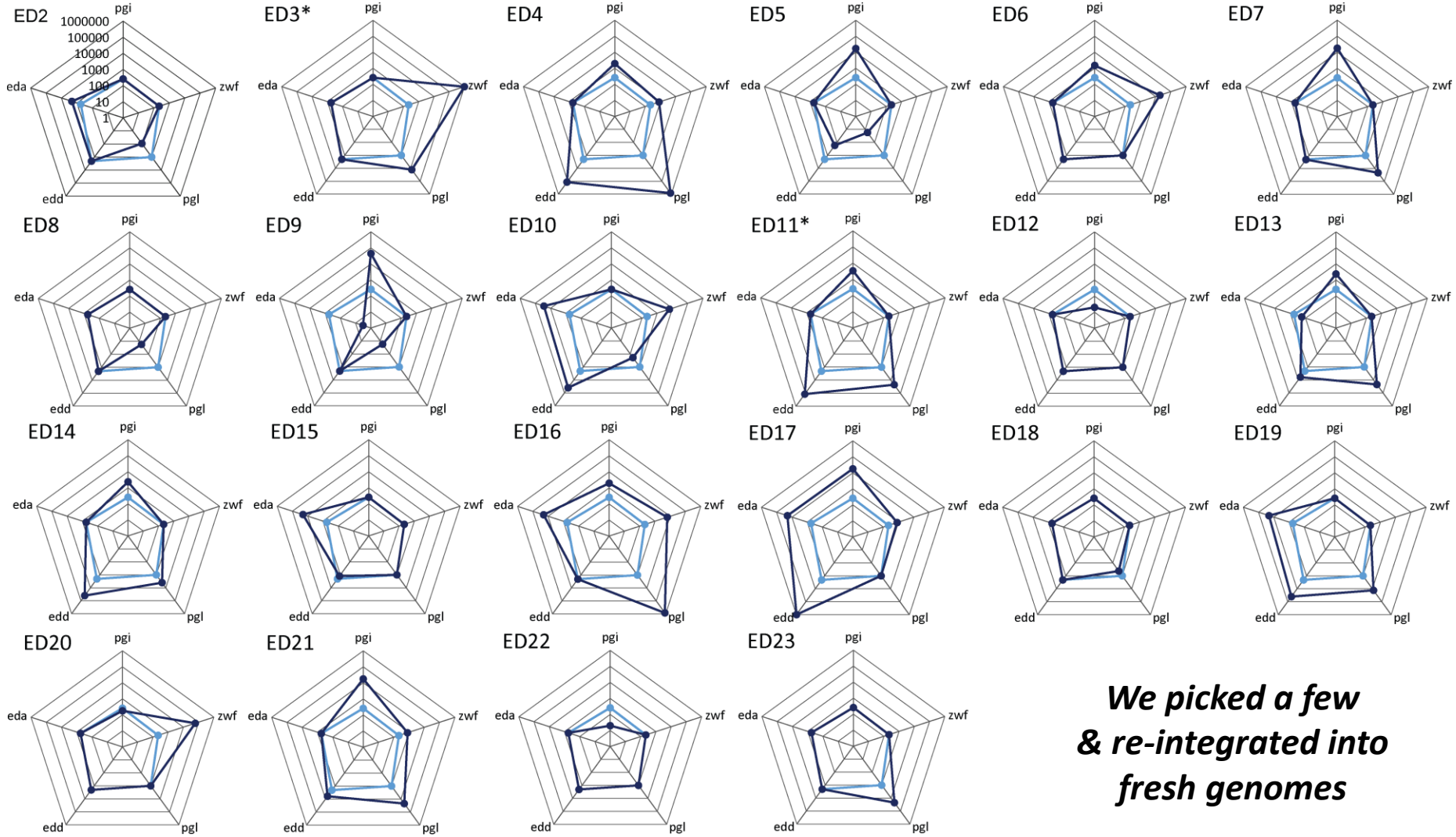


Now let's see what we find …

A Non-Linear SEA Map

Not a simple pattern

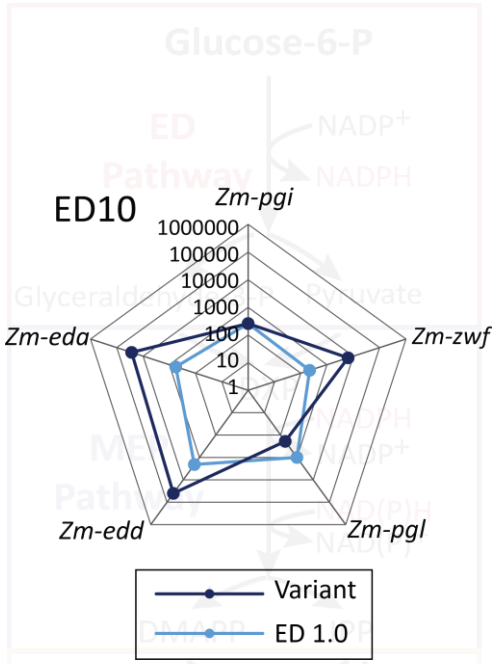
but “more of everything” did not result in more NADPH regeneration

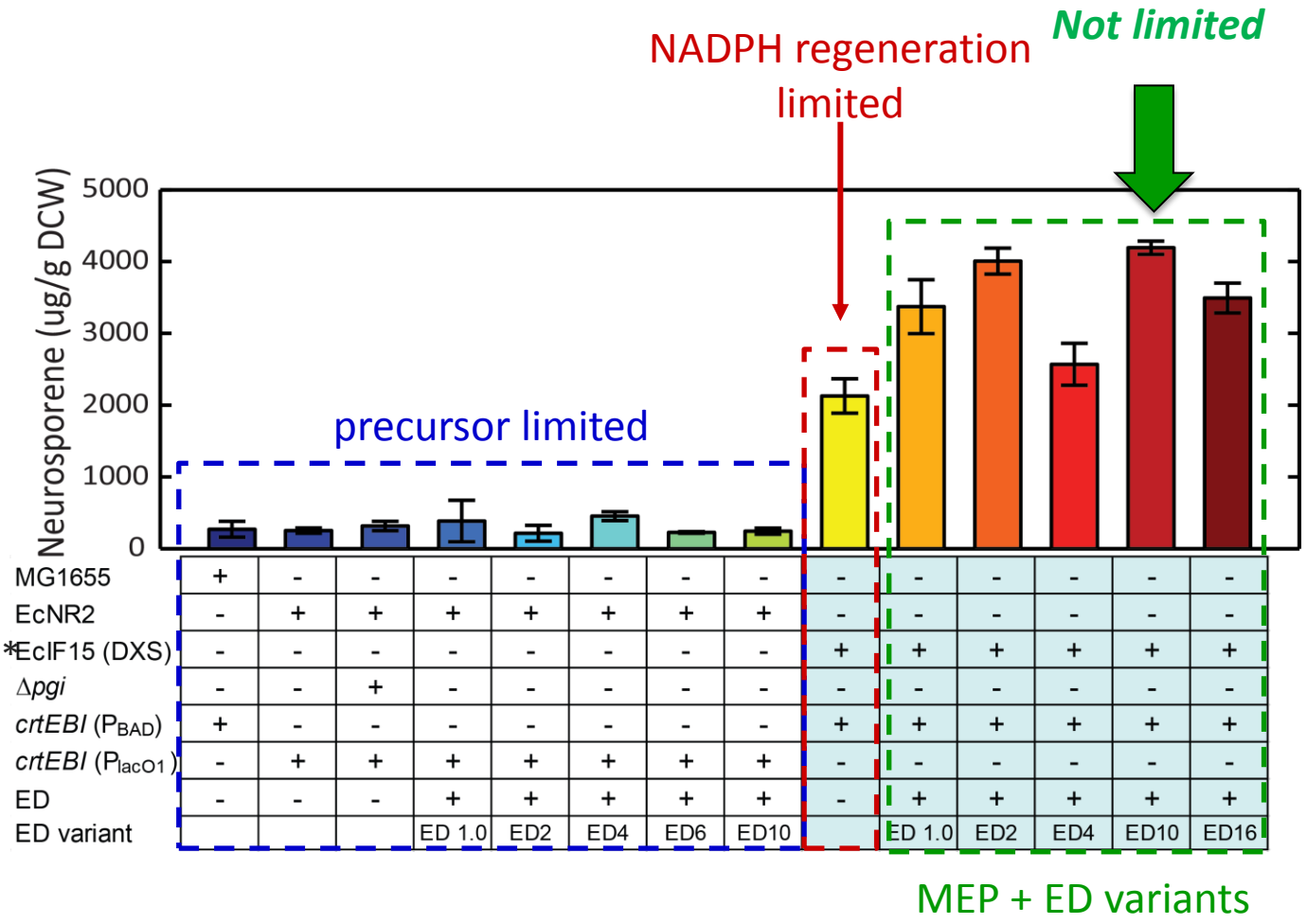


We picked a few & re-integrated into fresh genomes

More NADPH = More Product

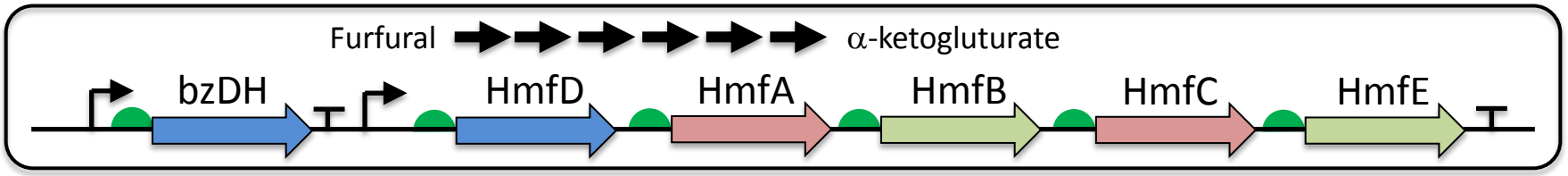
The best ED pathway variant increased natural product biosynthesis by 97%



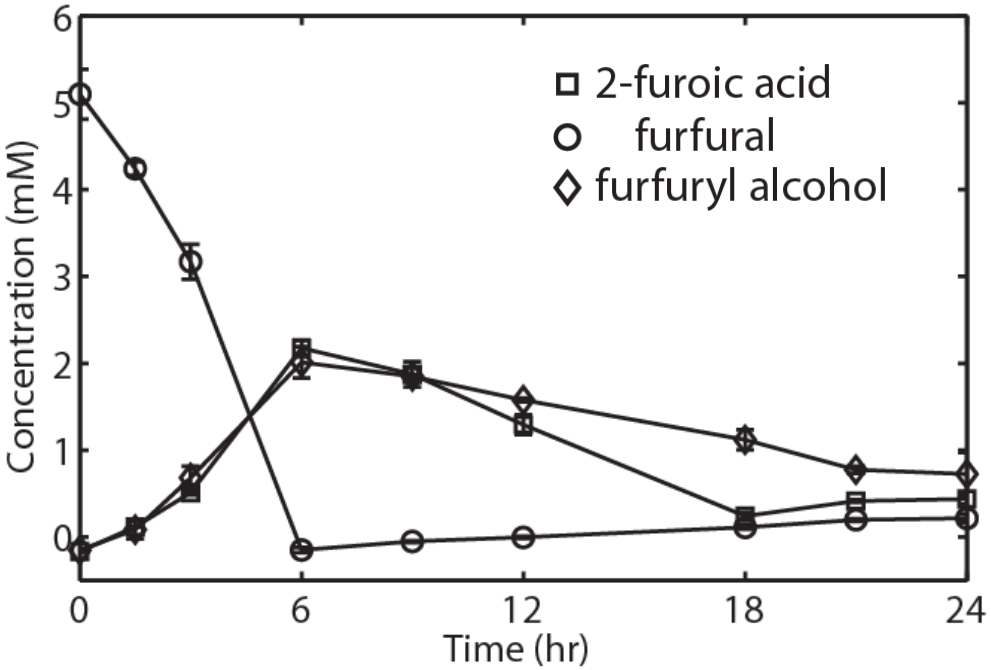


Pathway Example #3

a 6-enzyme furfural catabolic pathway for detoxifying lignocellulosic feedstock



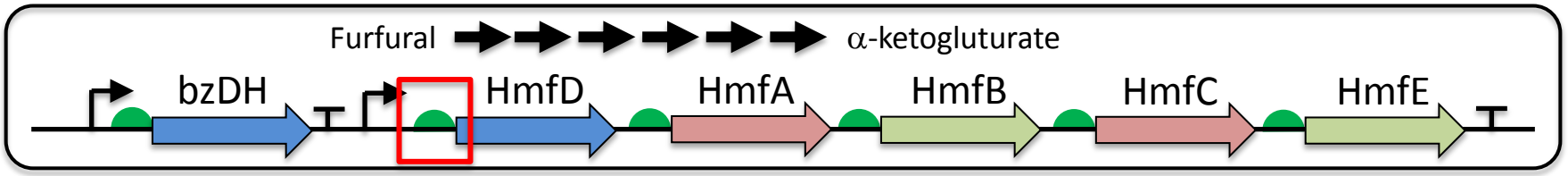
Hydrolyzed lignocellulose contains cheap sugars (glucose, xylose, arabinose)
but also several microbial inhibitors (furfural, hydroxy-methyl furfural, acetate)



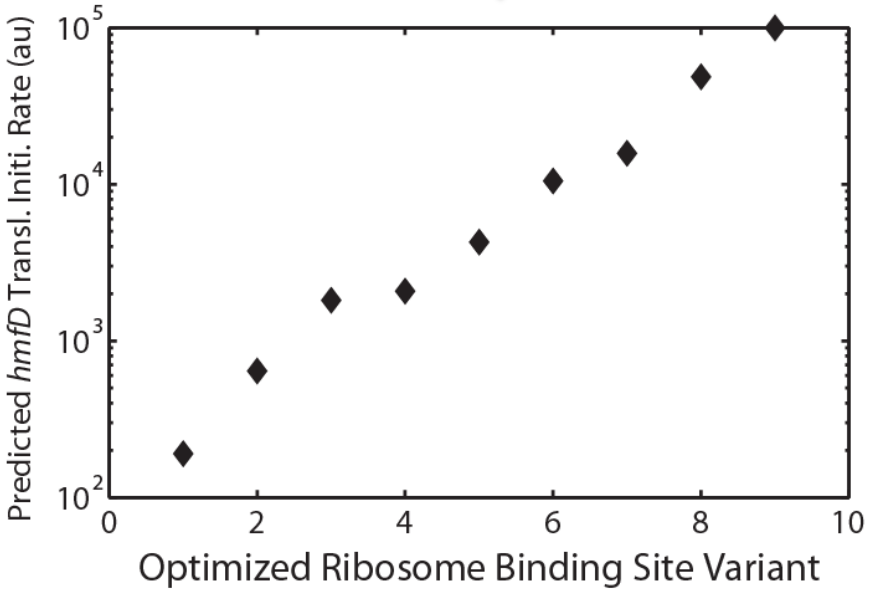
This 6-enzyme pathway catabolizes
5 mM furfural within 6 hours
and converts to TCA cycle metabolites

Pathway Example #3

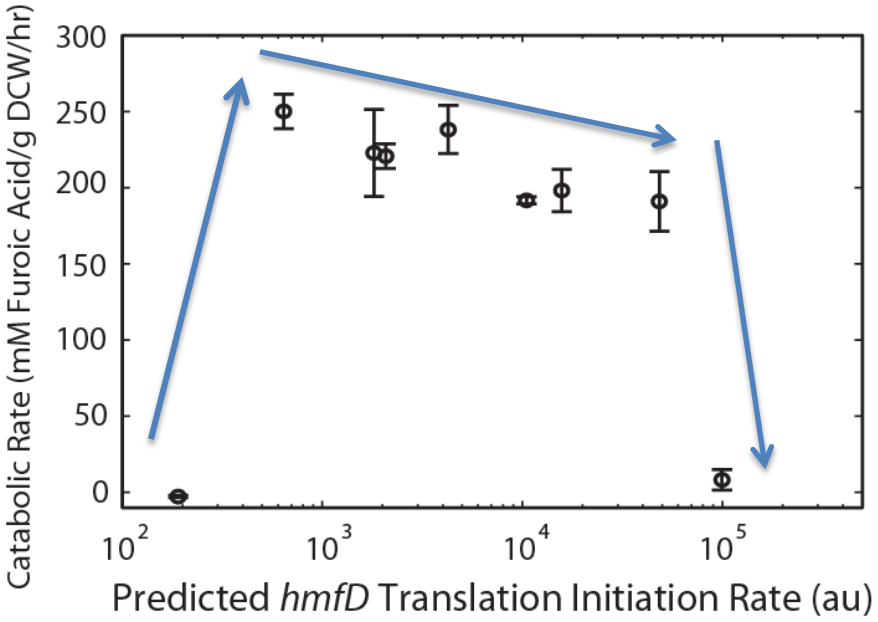
a 6-enzyme furfural catabolic pathway for detoxifying lignocellulosic feedstock



RBS Library Calculator



Non-linear relationship between hmfD expression and pathway activity

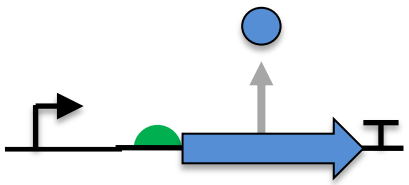


Today's Topics

#1: Designing DNA to Control Protein Expression
#2: Optimization of Multi-Protein Genetic Systems
#3: Case Studies: Engineered Metabolic Pathways

More recombinant products are requiring the expression of multiple proteins simultaneously
Our methods allow you to efficiently engineer & optimize multi-protein genetic systems

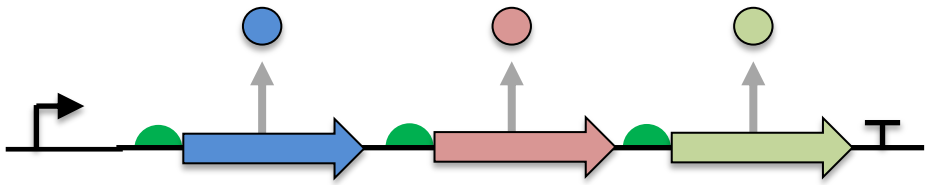
A Single-Protein Genetic System



expression & purification
of a single protein

Examples: single-chain antibodies,
single-subunit enzymes

A Multi-Protein Genetic System



multiple proteins working together
to produce a product

Examples: recombinant vaccines, multi-subunit enzymes,
natural product biosynthesis (antibiotics, therapeutics)

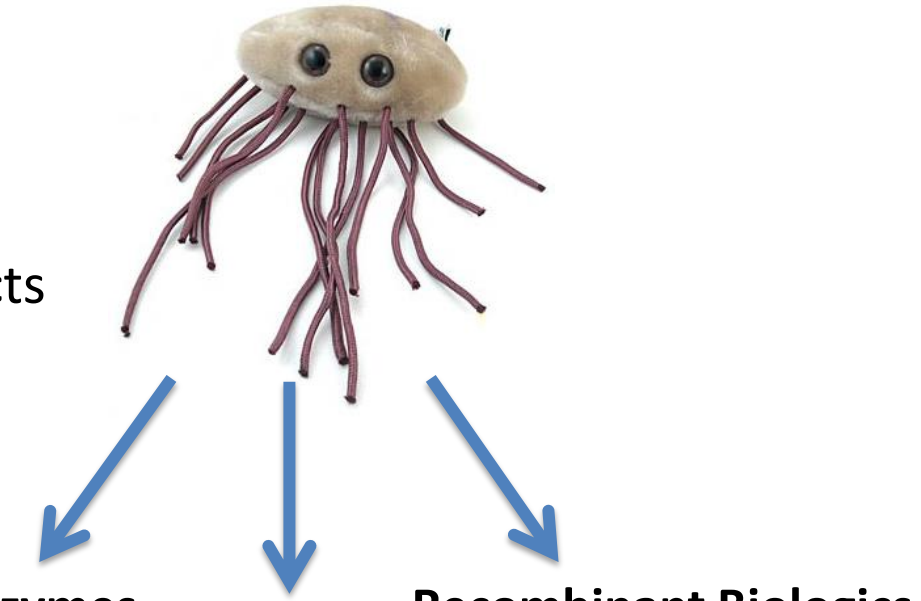
Clone Less, Know More

Predict the *specific* DNA sequences that will maximize a product's titer

Do it reliably, for diverse recombinant products.
And efficiently, with less cloning and fewer experiments.

We quantitatively identify **optimal expression levels & protein stoichiometries**, so the next (similar) product is developed even faster

An **Engineering Science** for Manufacturing Recombinant Products



Recombinant Enzymes **Recombinant Biologics**

Natural Products

Acknowledgements

Postdocs

Manish Kushwaha

Graduate students

Iman Farasat

Amin Espah Borujeni

Tian Tian

Long Chen

Chiam Yu Ng

Grace Vezeau

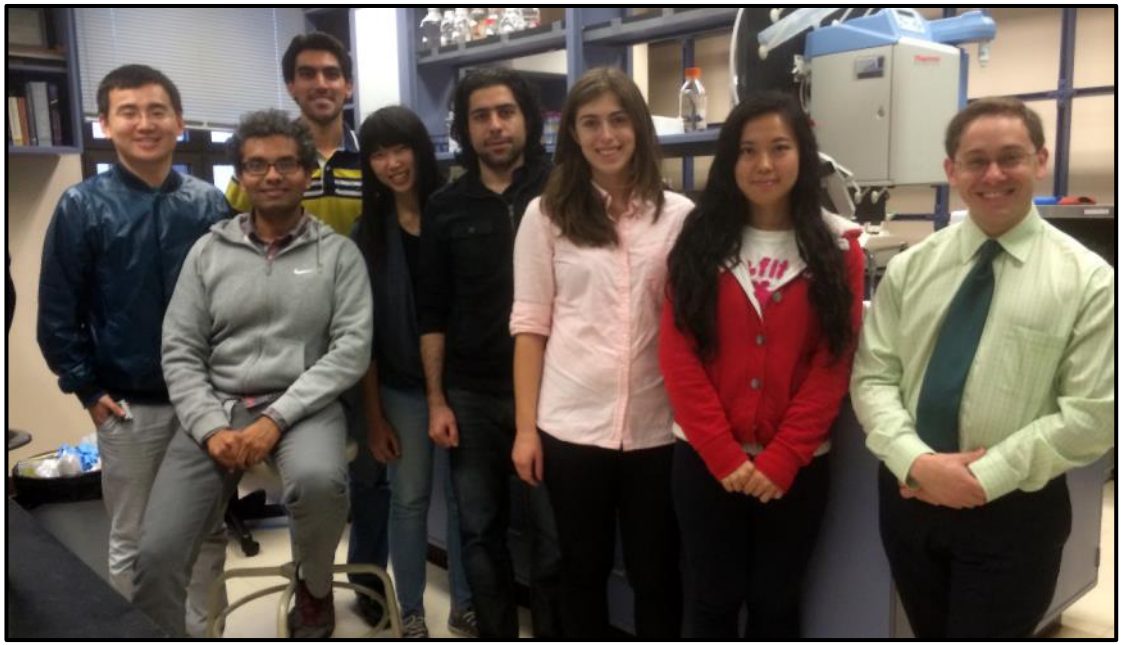
Sean Halper

Alex Reis

Undergraduate students

Ashlee Smith, Walker Huso,
Sam Klug, Justin Miller
Clay Swackhammer, Emily Sileo
Selina Roman-White

Questions?



Funding Sources

National Science Foundation

Air Force Office of Scientific Research

Office of Naval Research

DARPA

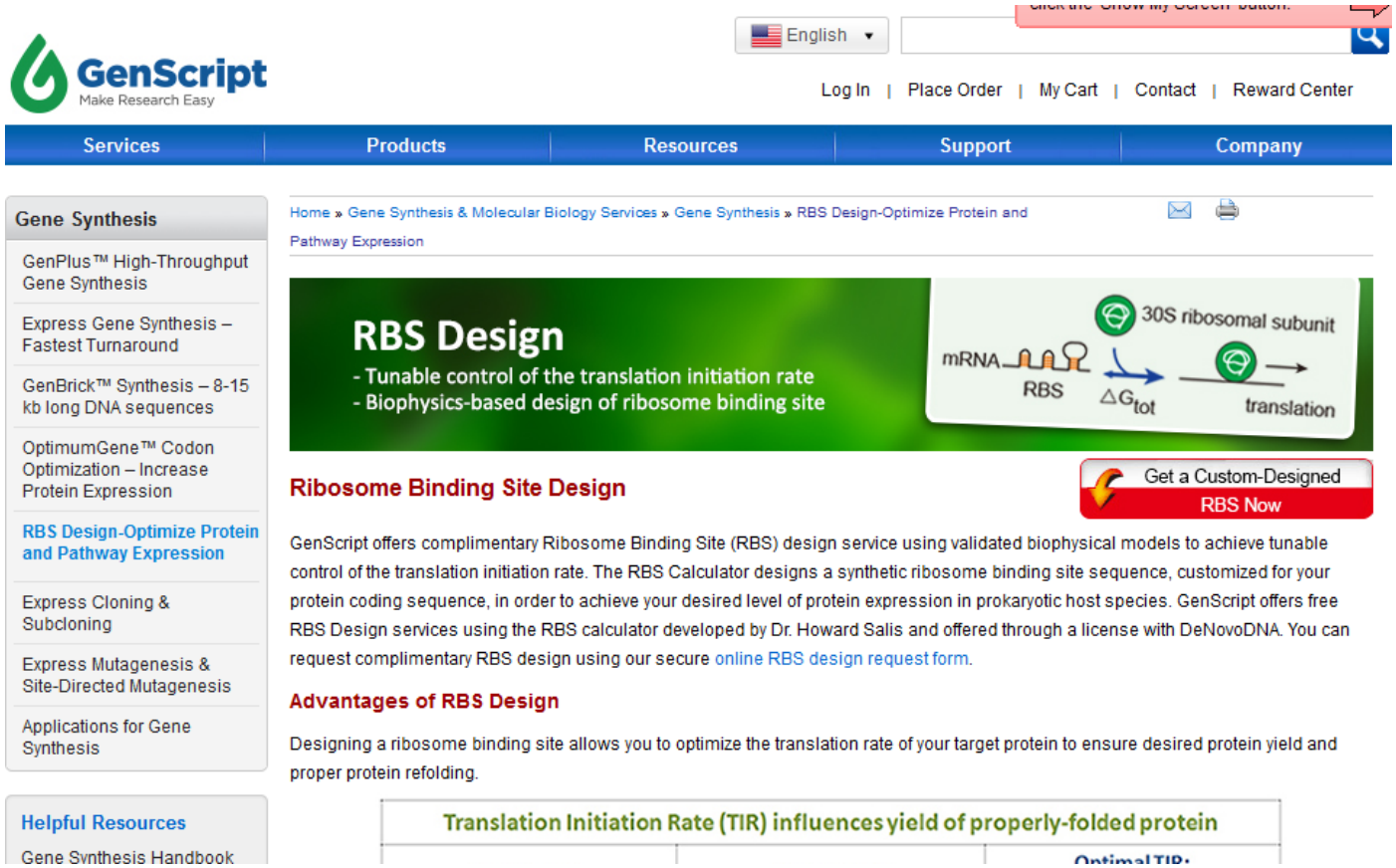
Penn State Institute for the Energy & Environment

PENNSTATE

1855

De Novo DNA

Visit GenScript website for more information: http://www.genscript.com/ribosome-binding-site-design.html

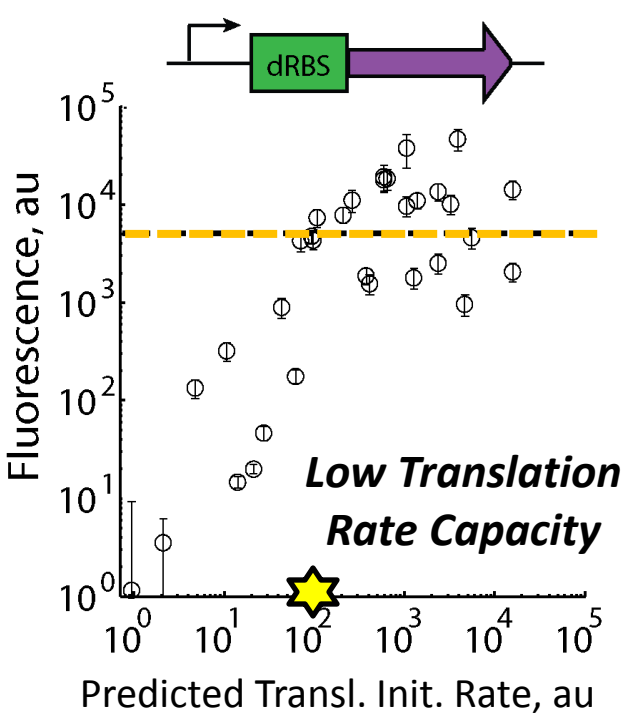


RBS Calculator to Verify Codon Optimization

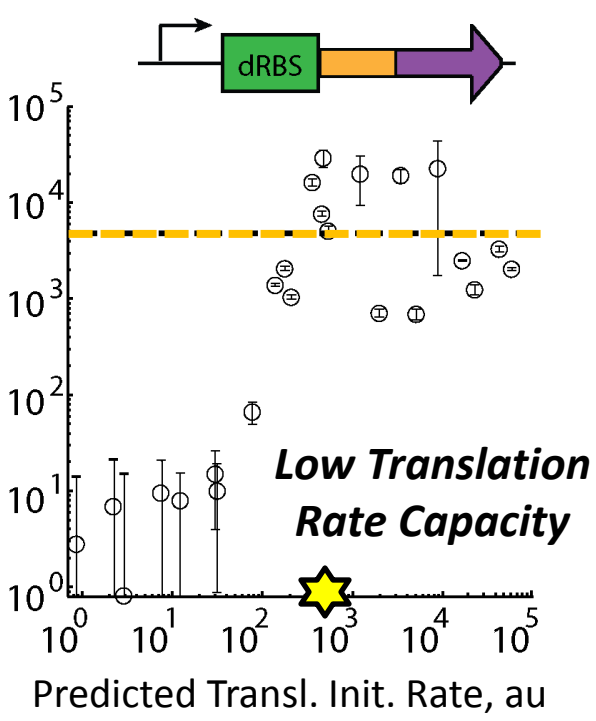
How do we know if a protein's coding sequence has been truly codon-optimized?
Systematically increase the translation initiation rate, and check it.

A "codon-optimized" GFP using frequent codons had a Low Translation Rate Capacity
But a GFP using fast codons had a High Translation Rate Capacity

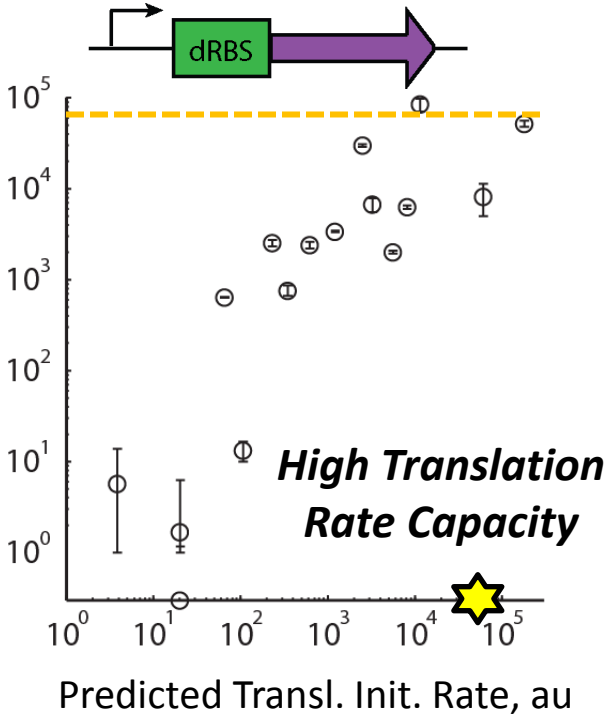
GFP with frequent codons



GFP fusion with frequent codons



*GFP with **fast** codons*



Quantifying Pathway Optimality

Q: How do we know if the pathway has been fully optimized?

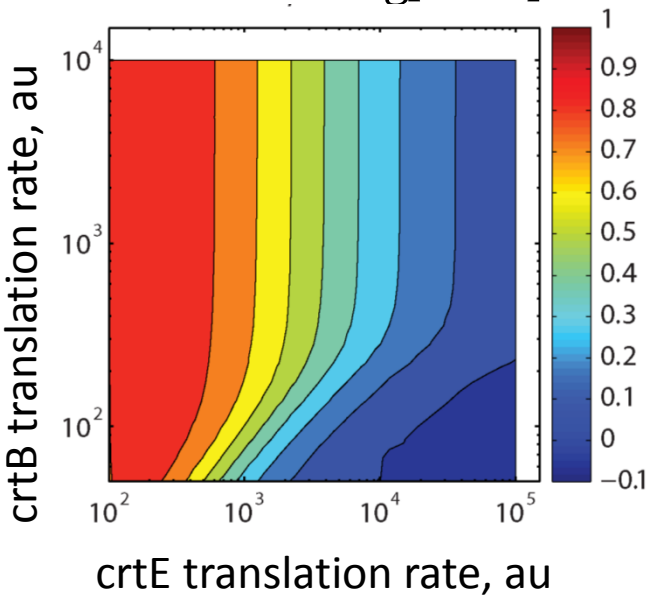
A: An optimally balanced pathway has zero-valued flux control coefficients

Flux control coefficients (FCCs) quantify the "rate-limiting-ness" of each enzyme's reactions

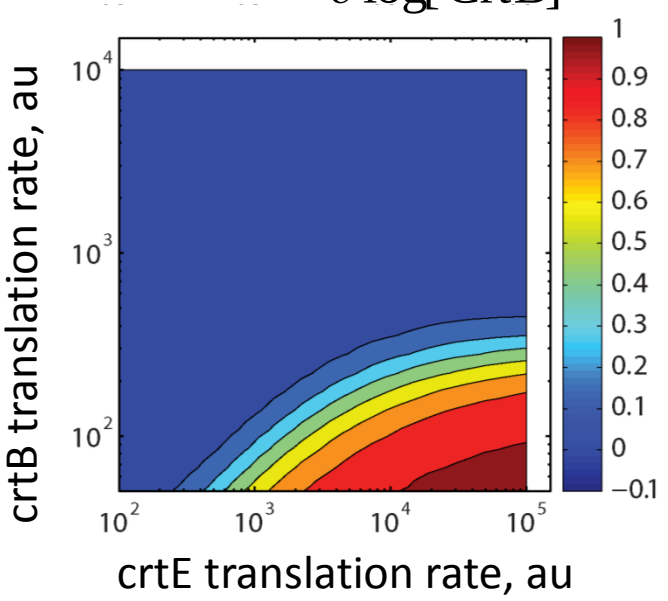
0: It is not rate-limiting . 1: it's the only rate-limiting step.

When a pathway's enzymes have FCCs = 0, then ***precursor biosynthesis is limiting***.

$$FCC_{crtE} = \frac{\partial \log P}{\partial \log[\mathrm{CrtE}]}$$



$$FCC_{crtB} = \frac{\partial \log P}{\partial \log[\mathrm{CrtB}]}$$



FCCs: Fell, 1992; Kholodenko & Westerhoff, 1993

Quantifying Pathway Optimality

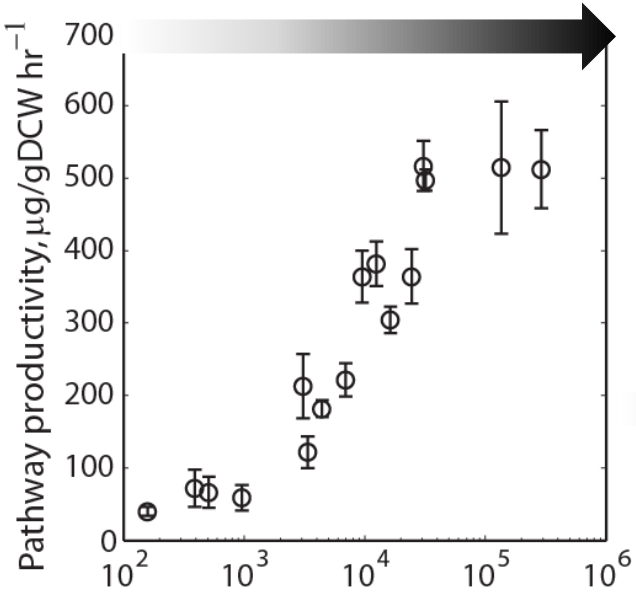
Q: How do we know when to increase precursor biosynthesis rates?

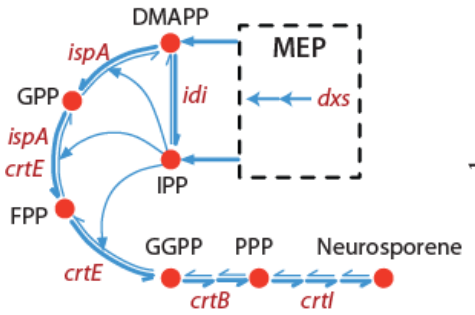
A: Only when the downstream pathway has been optimally balanced.

We used the RBS Library Calculator to systematically vary precursor biosynthesis rates

Pathway productivity only increased when the pathway was optimally balanced

optimally balanced pathway variant (FCCs near zero)





increasing precursor biosynthesis rates

rate-limited pathway variant (FCC for CrtE = 0.6)