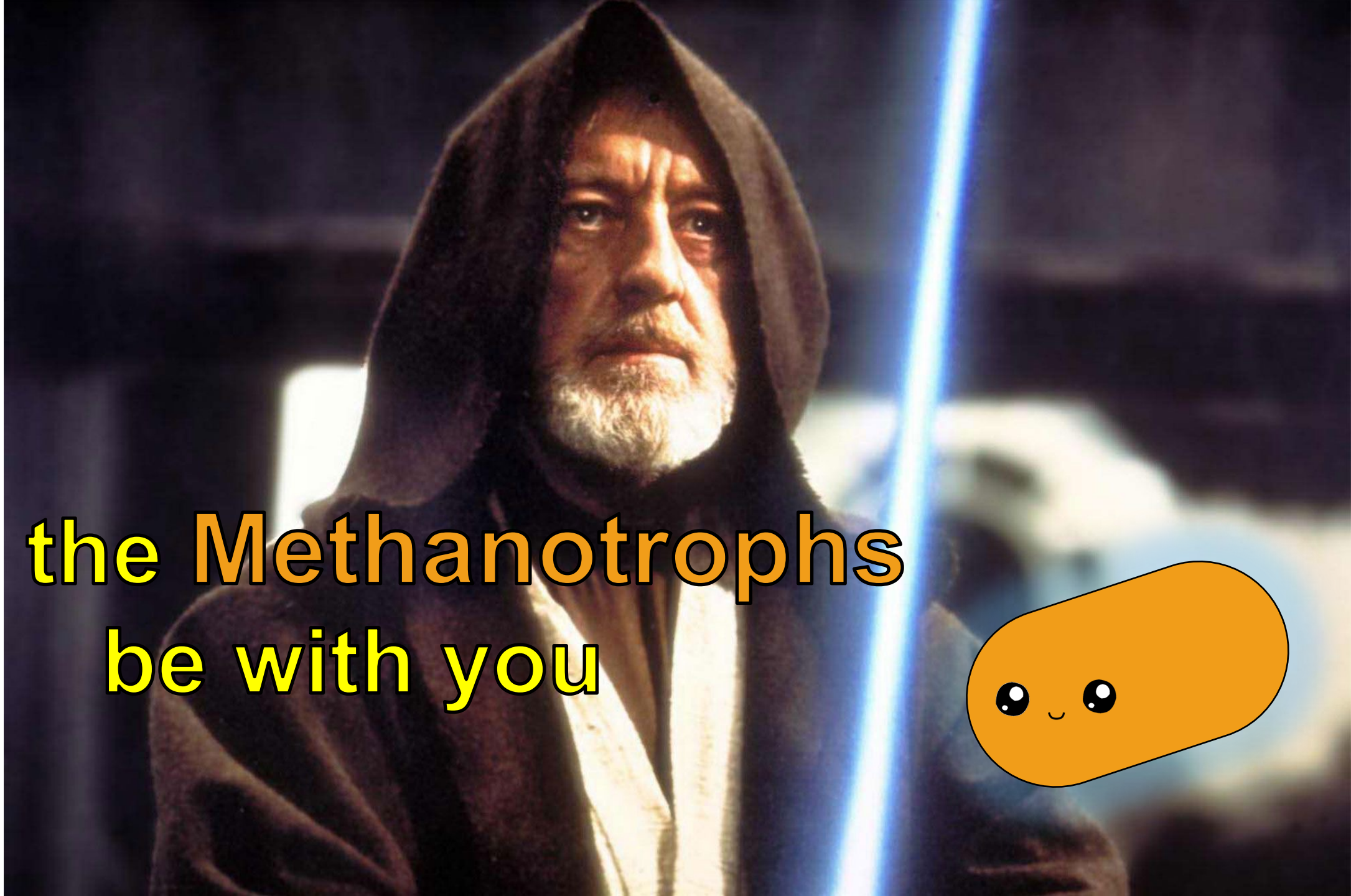
Erin Wilson

CompBio Seminar

May the 10th 2021



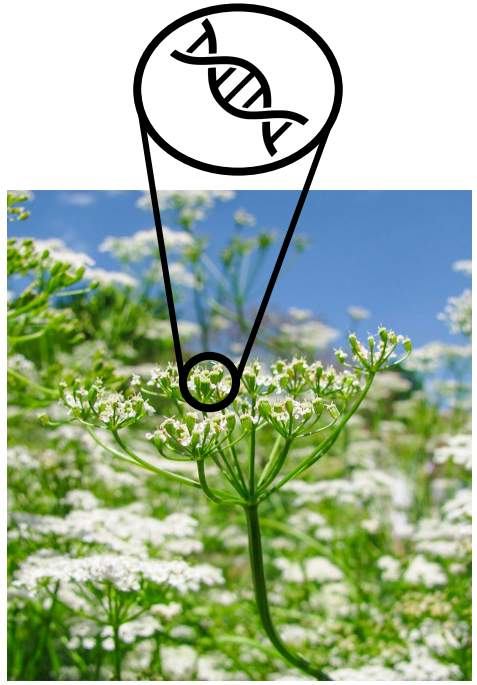
Overview

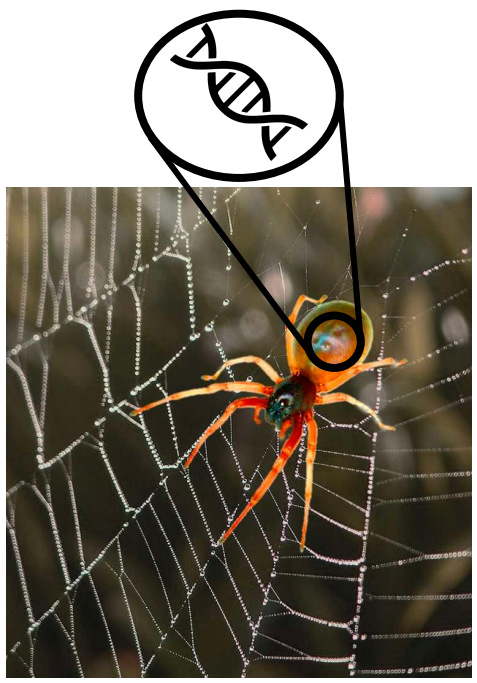
- Background
  - Microbes and Metabolism and Methane, oh my!
- Dataset and previous project
  - A hunt for strong promoters
- Idea/early work on new project
  - more nuanced promoter tools
  - ***Would love Feedback!***


Where does our
stuff
come from?

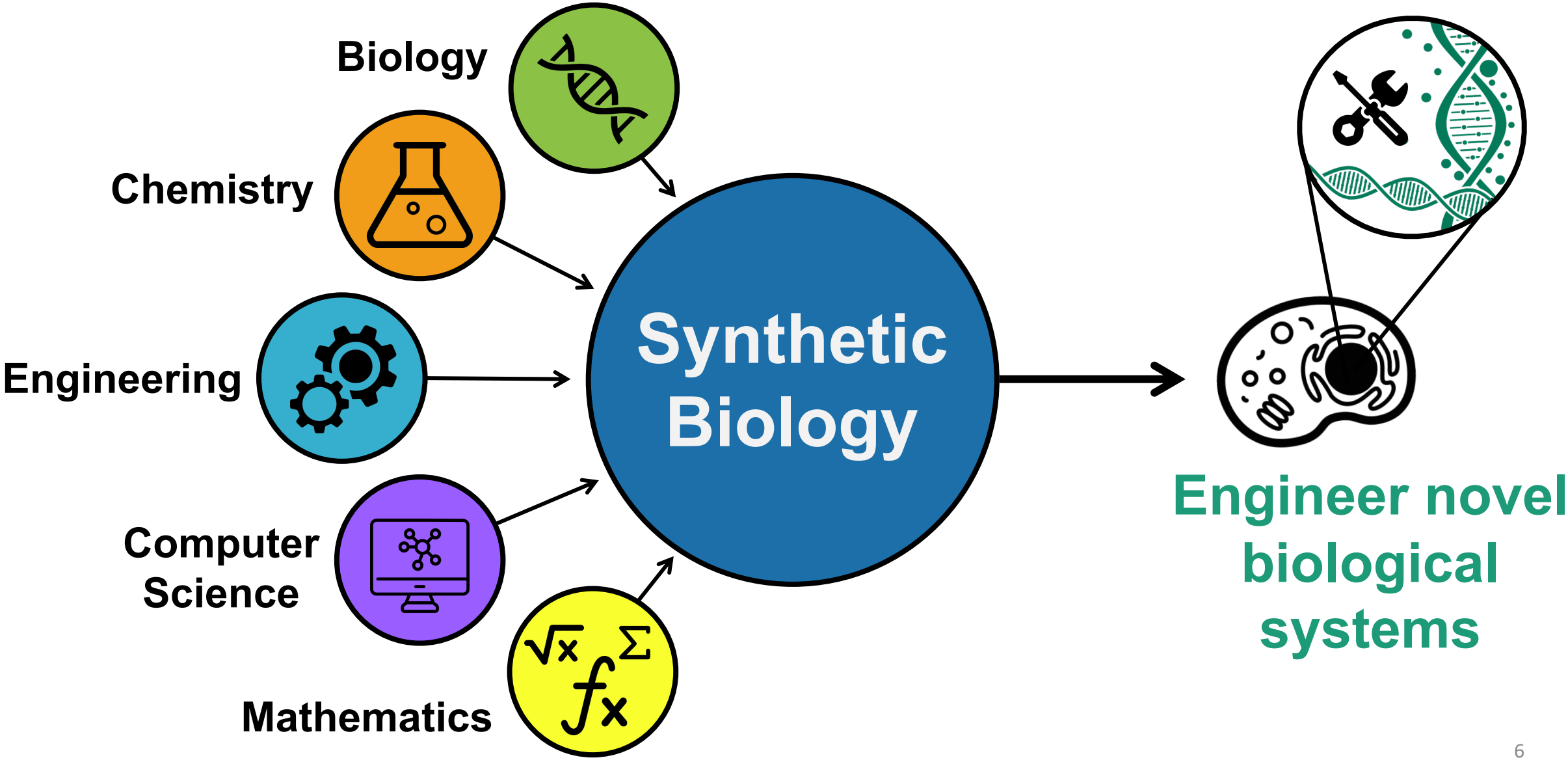
A lot of stuff comes from biology!

Nature has the instructions saved… in DNA!





Synthetic Biology: rewiring Nature's instructions to engineer **novel biological systems**

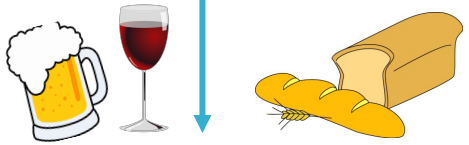


What is “Metabolic Engineering”…?

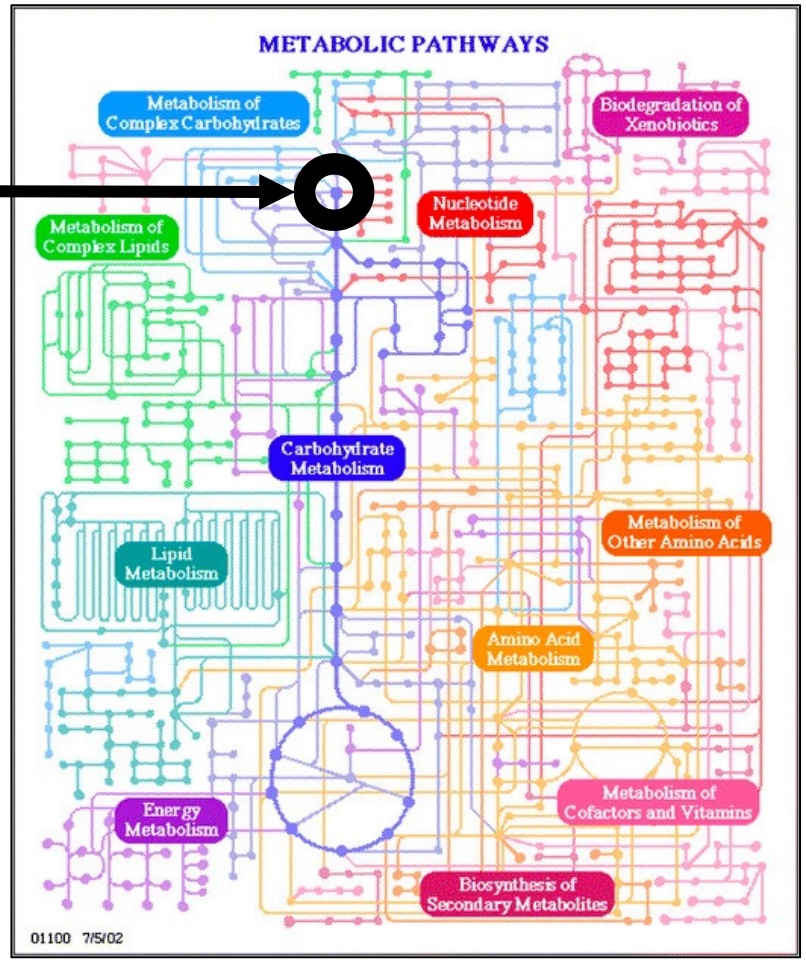
Glucose
(entry point)

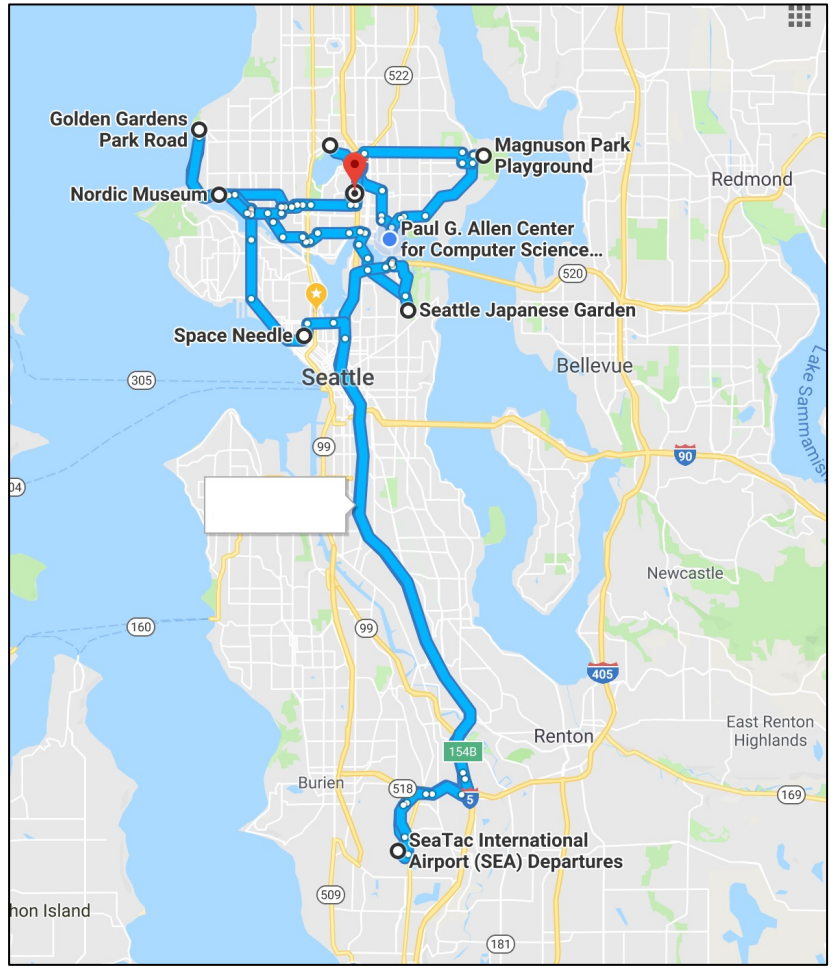
Sugar cane

Microorganism
factory
Yeast

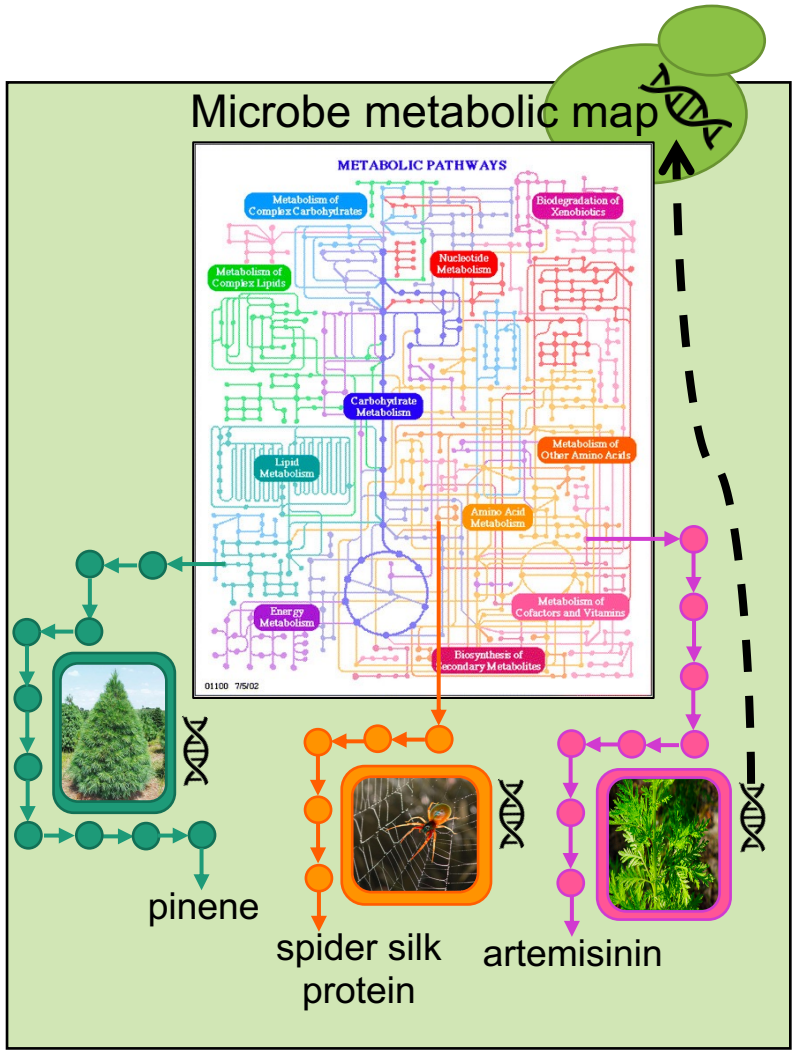


Useful molecules!

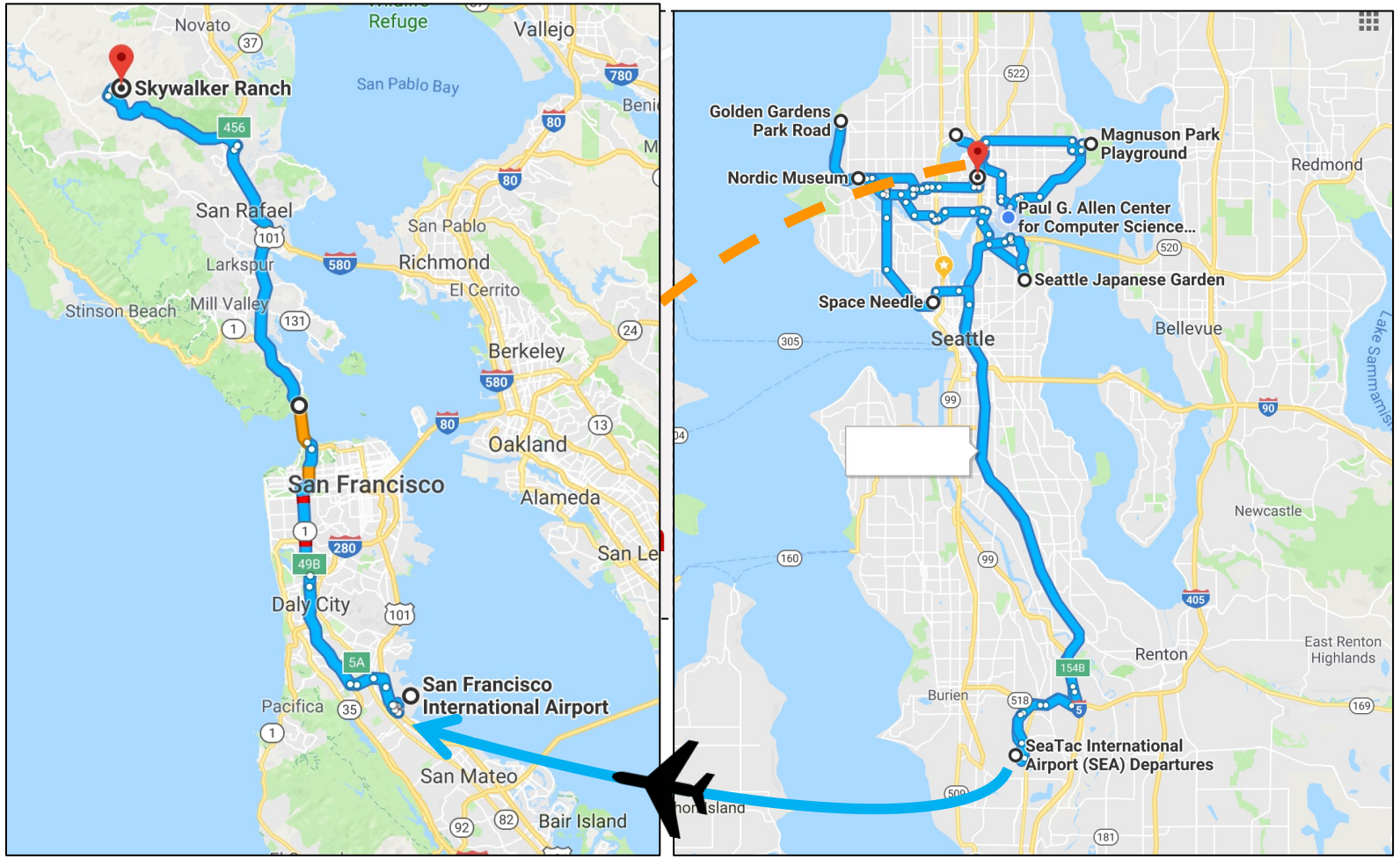




Connect metabolic maps between organisms!

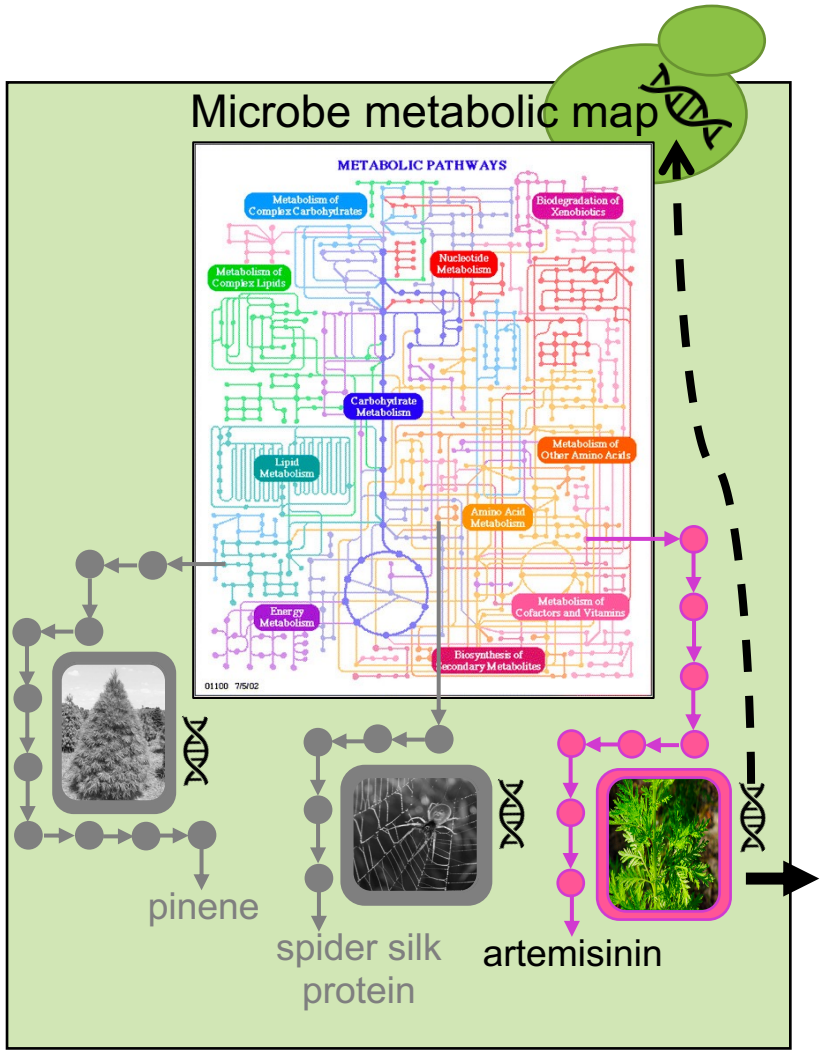


Other interesting molecules!



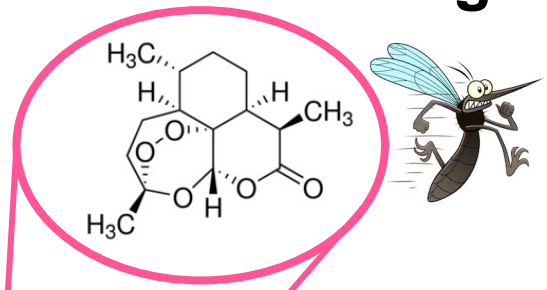
Expand local map by connecting to a node in another map!

Artemisinin: an early SynBio success story!



Other interesting molecules!

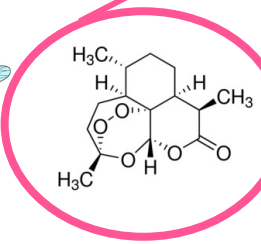
Artemisinin: anti-malaria drug



Sweet wormwood

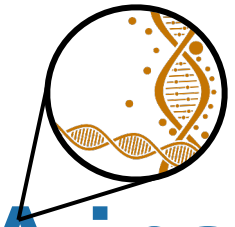
artemisinin genes

Sugar cane



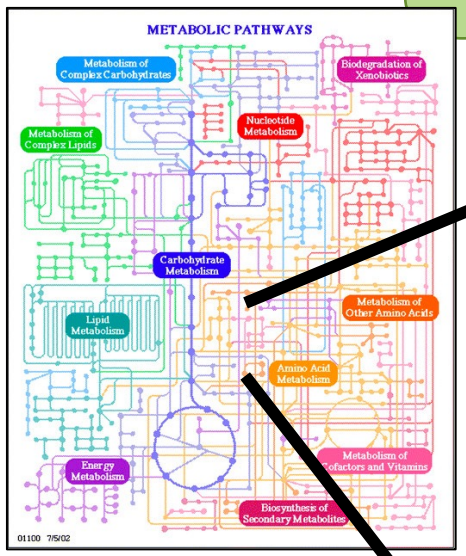
Metabolic Engineering: The Big Picture

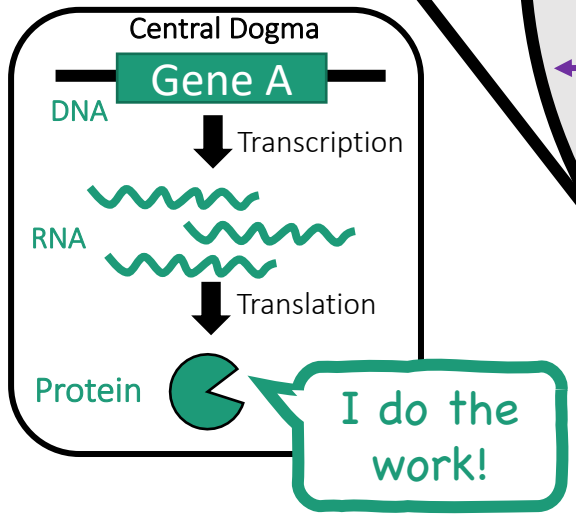
For any molecule made by any organism in Nature, there exists some metabolic pathway to get there…

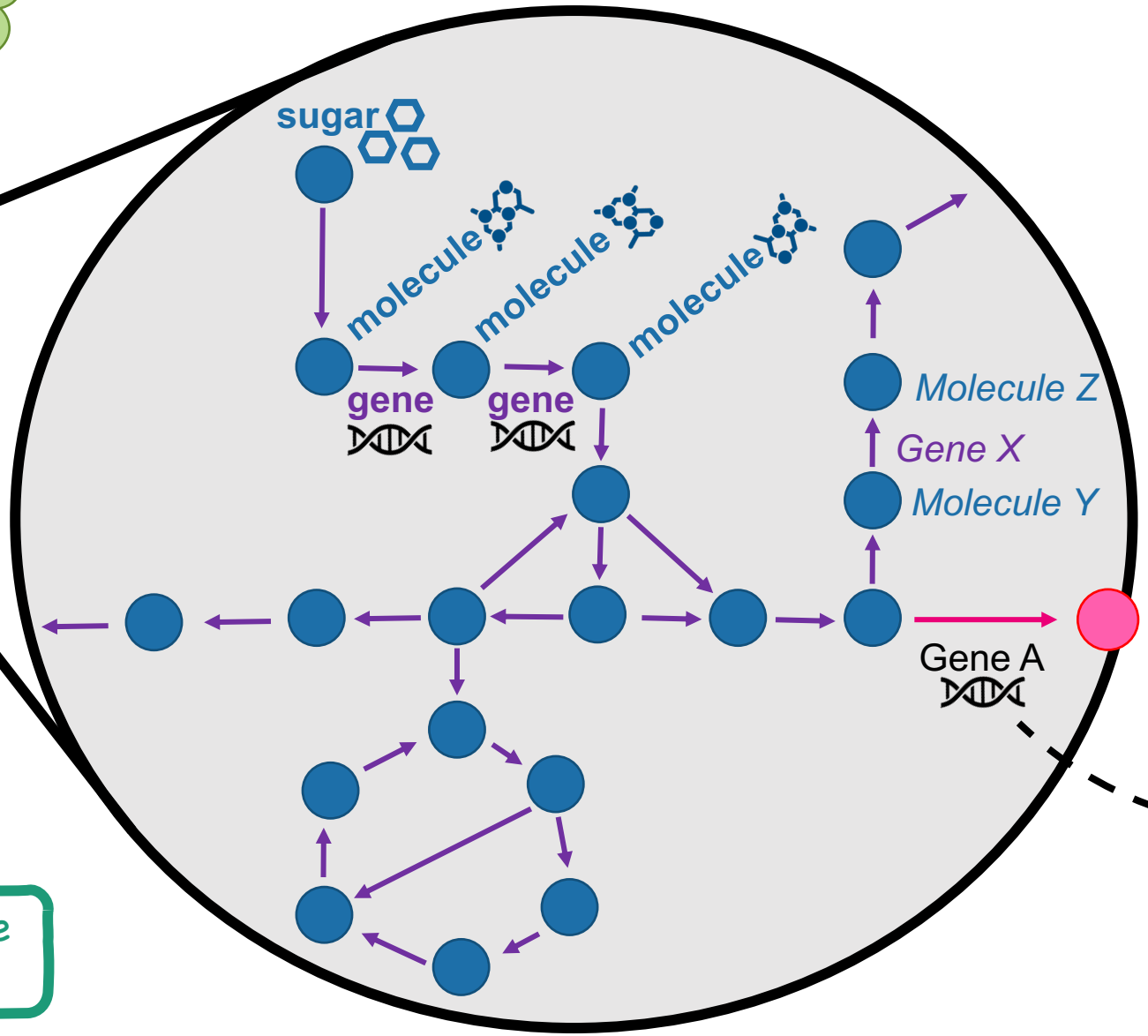


If you can identify the DNA instructions that encode that pathway, *hypothetically* you can try to put it in a microbe.

Metabolism is like a graph

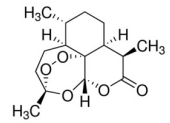






Sweet wormwood

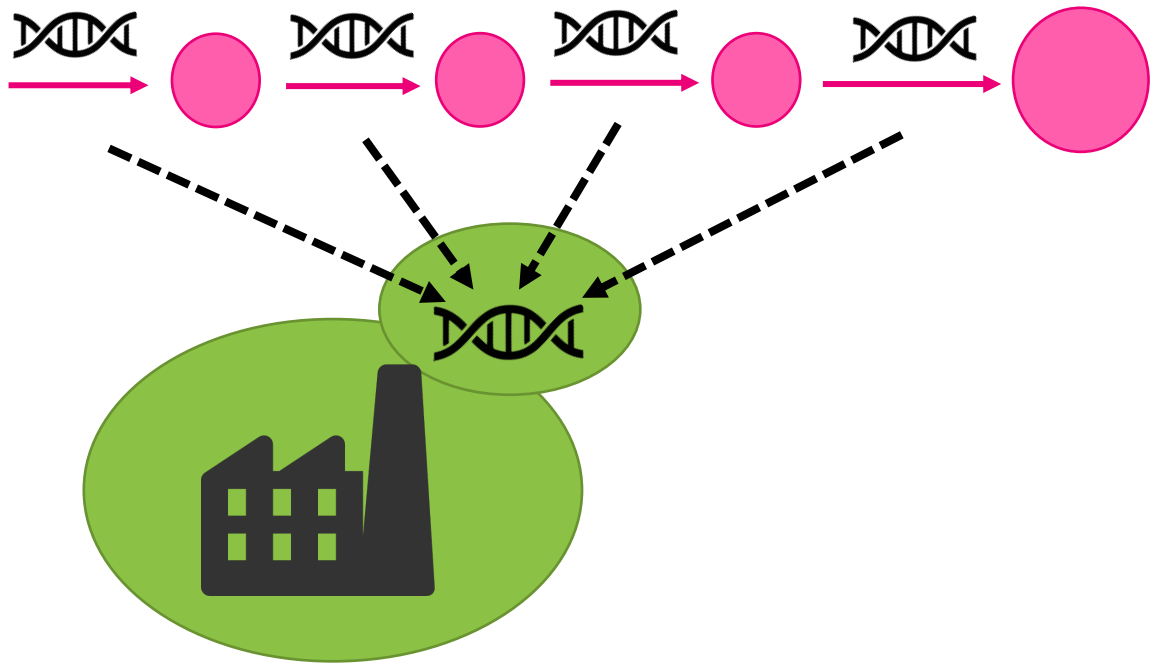
Artemisinin



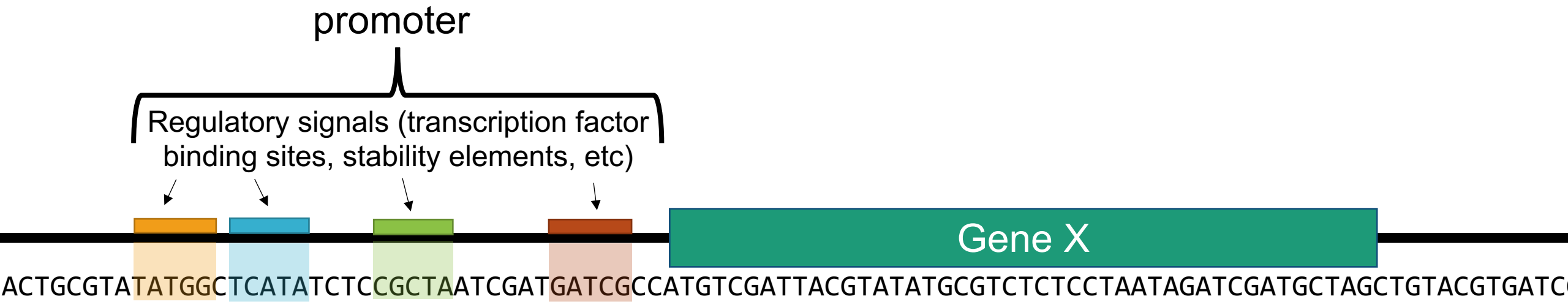
Host Microbe += [A, B, C, D, E]

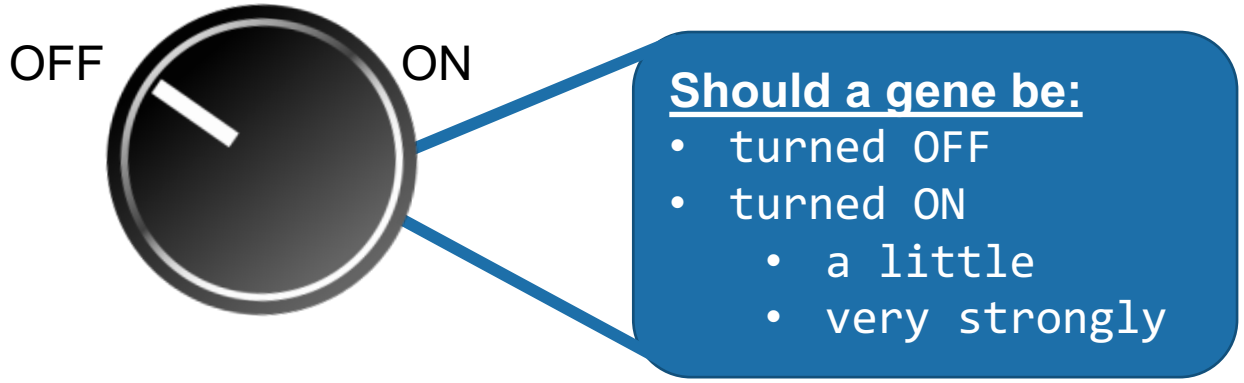
To install new pathways, insert new genes

Insert protein-coding DNA



DNA sequences encode many important signals for regulating genes!

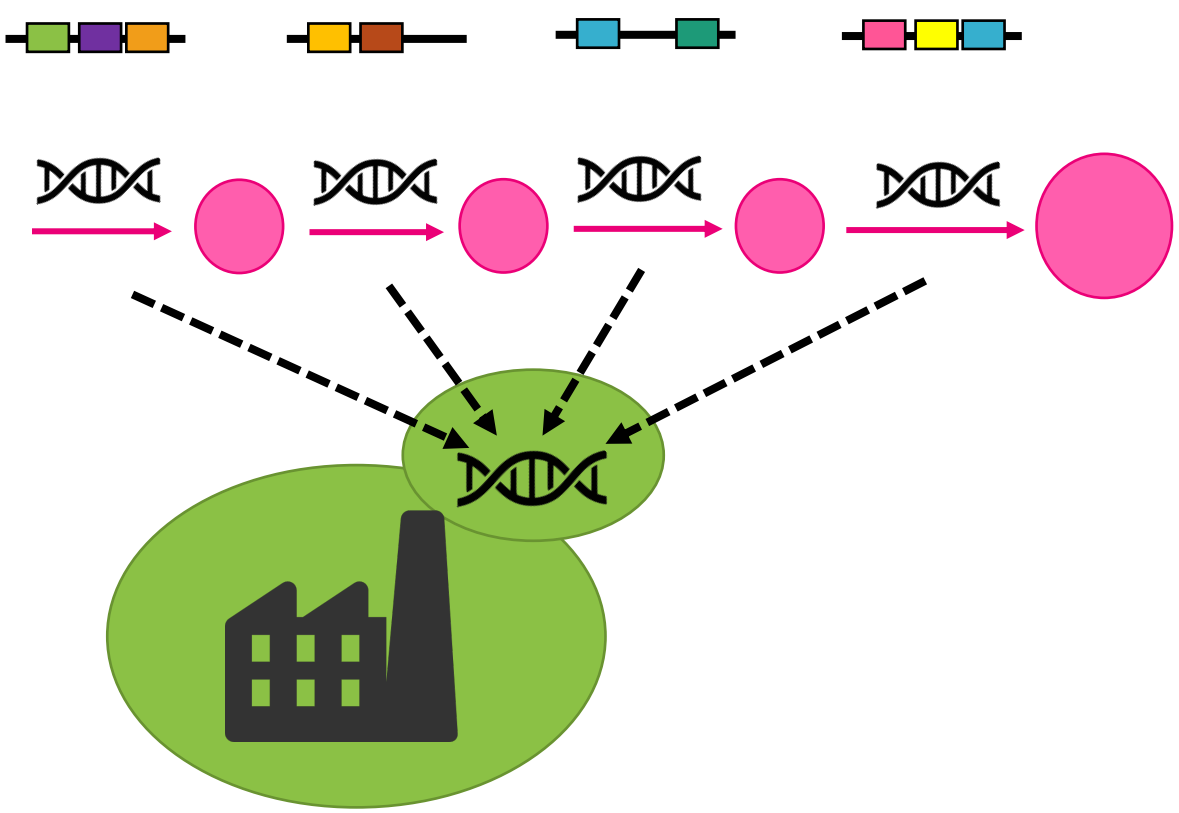




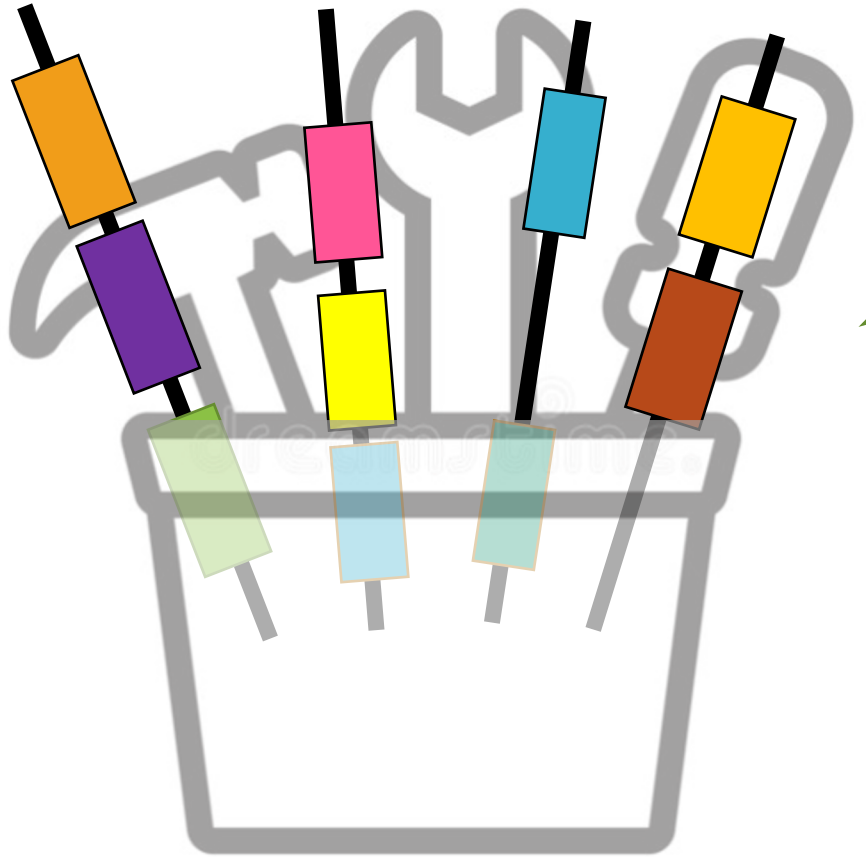
When installing new pathway genes, must also install regulatory signals

Drive gene expression with regulatory DNA

Insert protein-coding DNA



To engineer a microbe, build out a genetic engineering toolkit



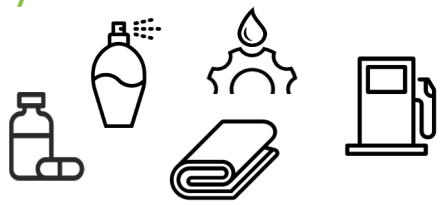
Variety of promoters

Sugar + microbe + science = sustainable products!

Renewable Feedstock
(Sugar cane)

Microorganism factory
(yeast)

Useful molecules!
(Jet fuel, medicine, flavors & fragrances
any molecule found in nature!)



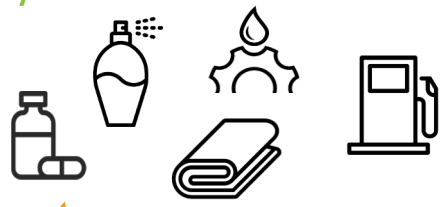
Bolt Threads
DEMETRIX
amyris
zymergen
LanzaTech
Arzeda.
GINKGO BIOWORKS™
THE ORGANISM COMPANY

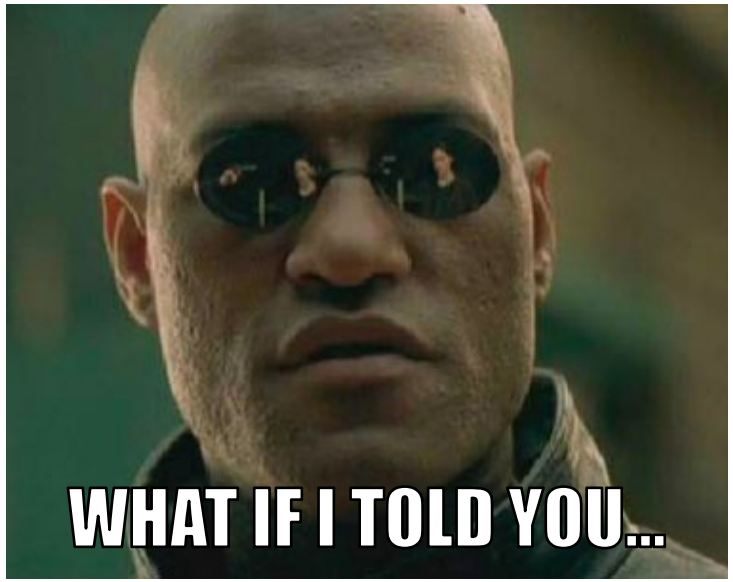
Sugar + microbe + science = sustainable products!

Renewable Feedstock
(Sugar cane)

Microorganism factory
(yeast)

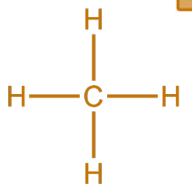
Useful molecules!
(Jet fuel, medicine, flavors & fragrances
any molecule found in nature!)





Waste Stream
(Methane)

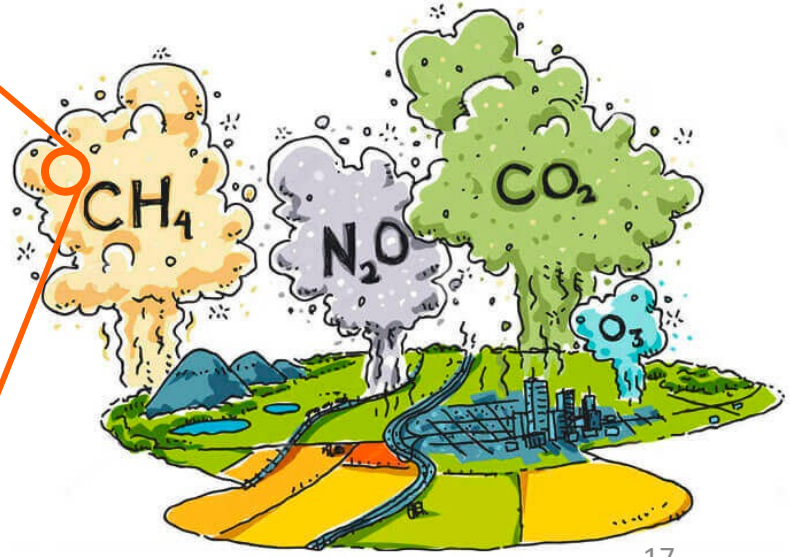
CH_4



Microorganism factory
(Methanotroph)

Methane (CH_4)

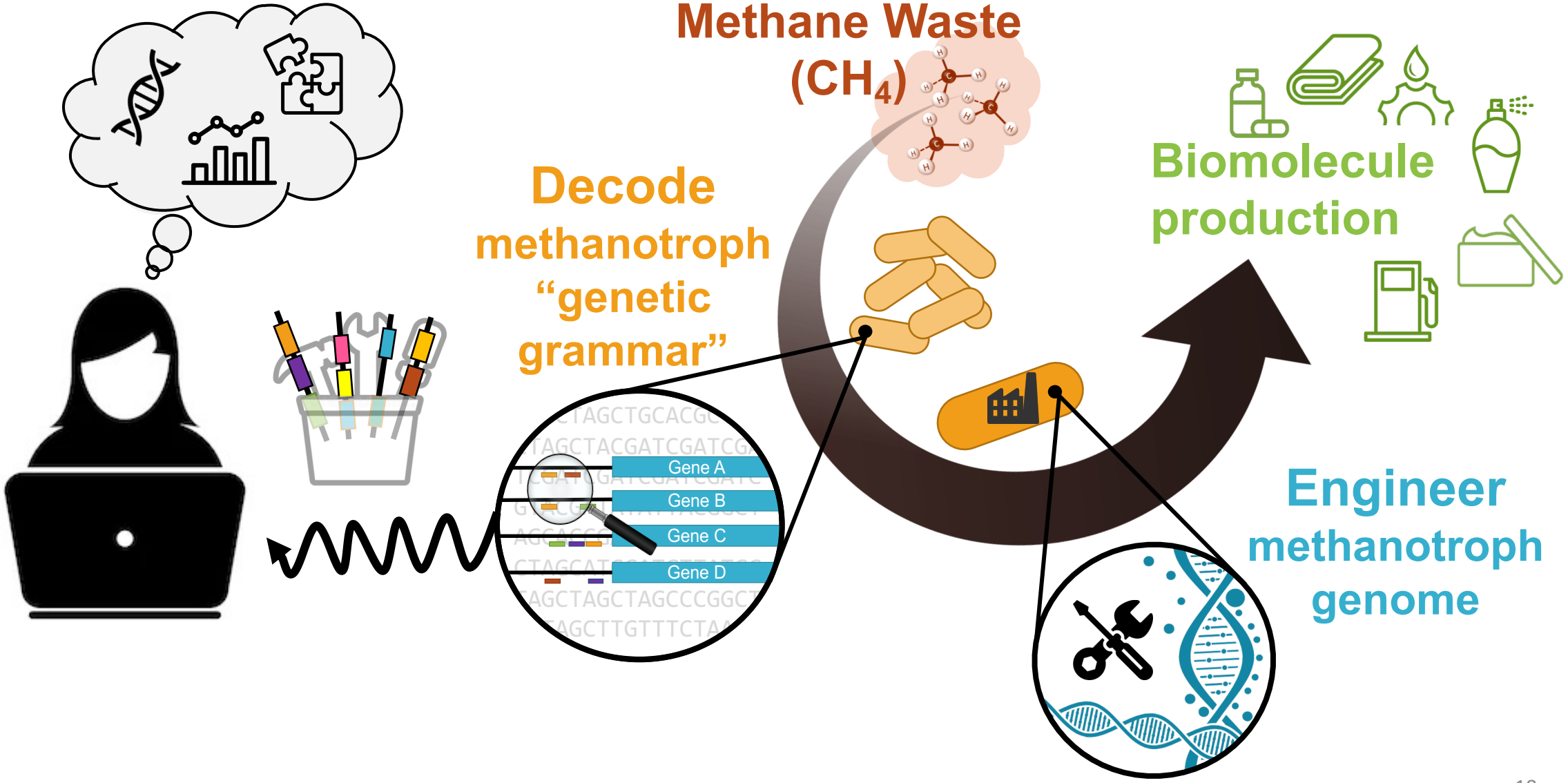
- Greenhouse gas from both natural and human activity
- 2nd greatest contributor to anthropogenic climate change behind CO_2
- 20-30x more potent than CO_2



https://www.globalcarbonproject.org

https://www.science-sparks.com/know-your-greenhouse-gases/

My research focus: Computationally decode the language that methanotrophs use to control their genes



Overview

- Background
  - Microbes and Metabolism and Methane, oh my!

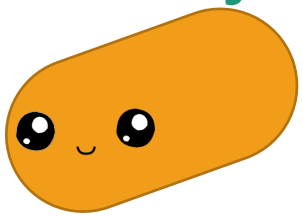
- **Dataset and previous project**
  - **A hunt for strong promoters**

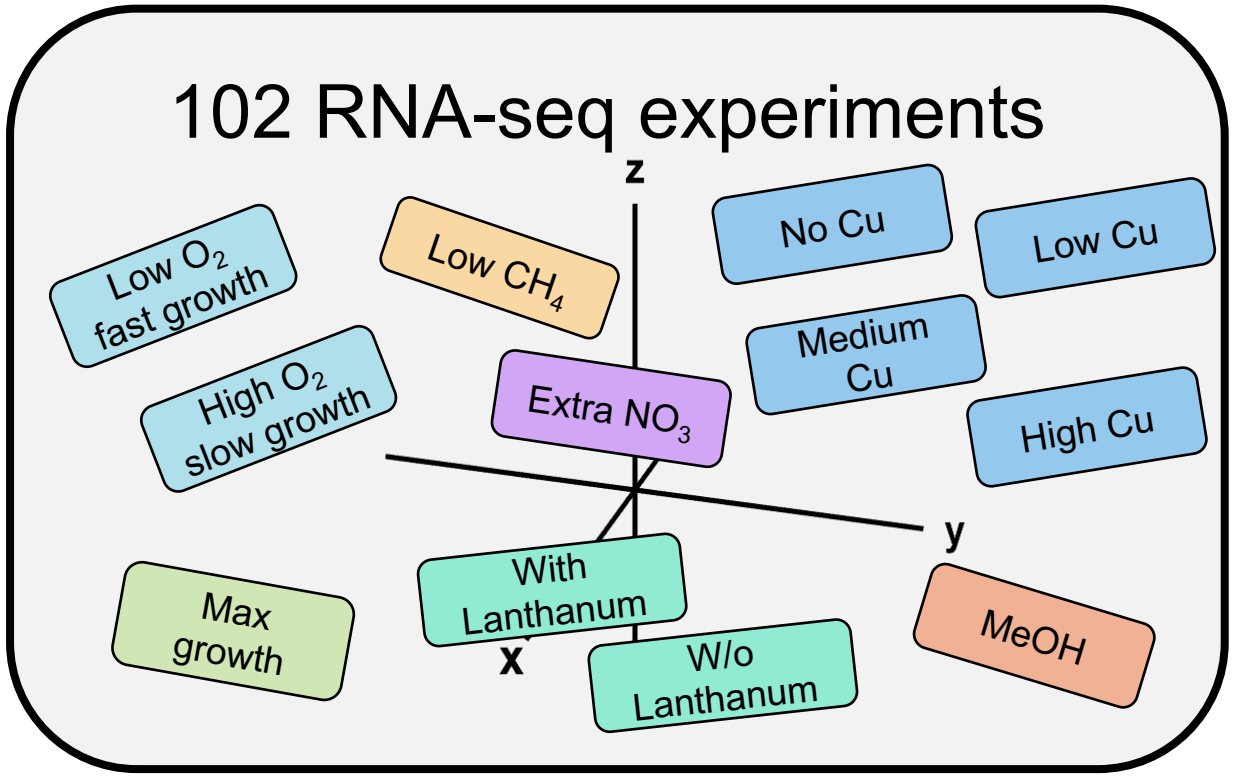
- Idea/early work on new project
  - more nuanced promoter tools
  - ***Would love Feedback!***

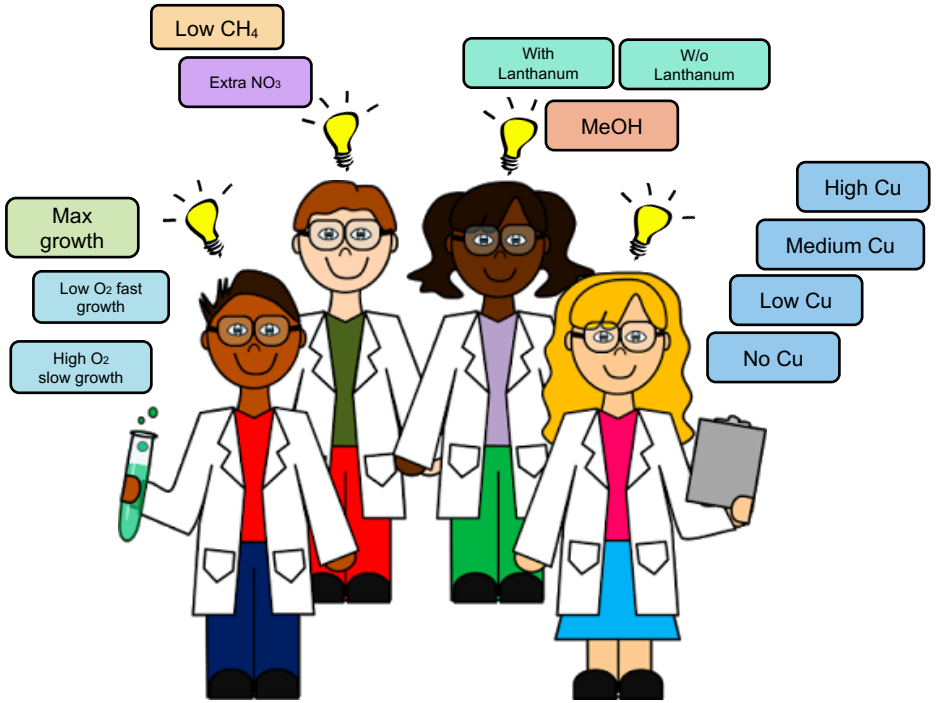
Lidstrom Lab: ~100 RNA-seq datasets
a unique opportunity?

ChemE
Microbiology

"5G"







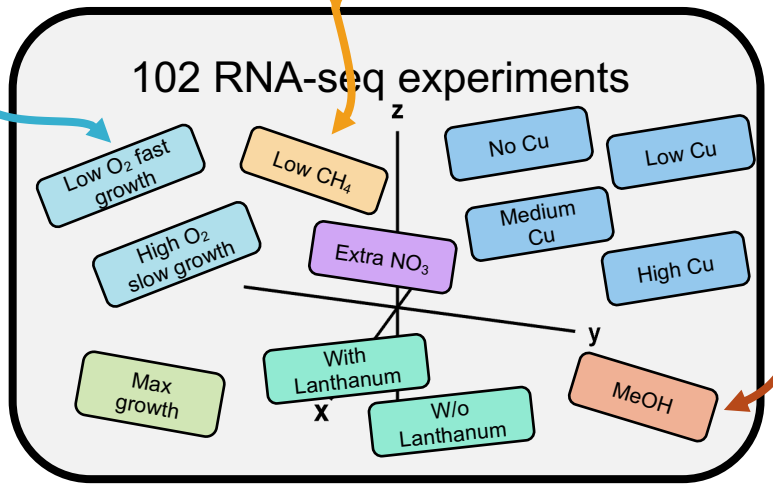
What do the data look like?

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	…
Gene A	7	213	44	55	456	25	77	124	
Gene B	12	574	78	65	227	413	158	71	
Gene C	201	99	5	374	87	145	352	47	
…									. . .

Count of RNA transcripts (TPM)

What do the data look like?

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	...
Gene A	7	213	44	55	456	25	77	124	
Gene B	12	574	78	65	227	413	158	71	
Gene C	201	99	5	374	87	145	352	47	
...									...



What do the data look like?

	Condition 1	Condition 2	Condition 3	…
Gene A	7	213	44	
Gene B	12	574	78	
Gene C	201	99	5	
…				. . .

Average TPMs by condition

~4,000 genes x 12 conditions

Previous project:

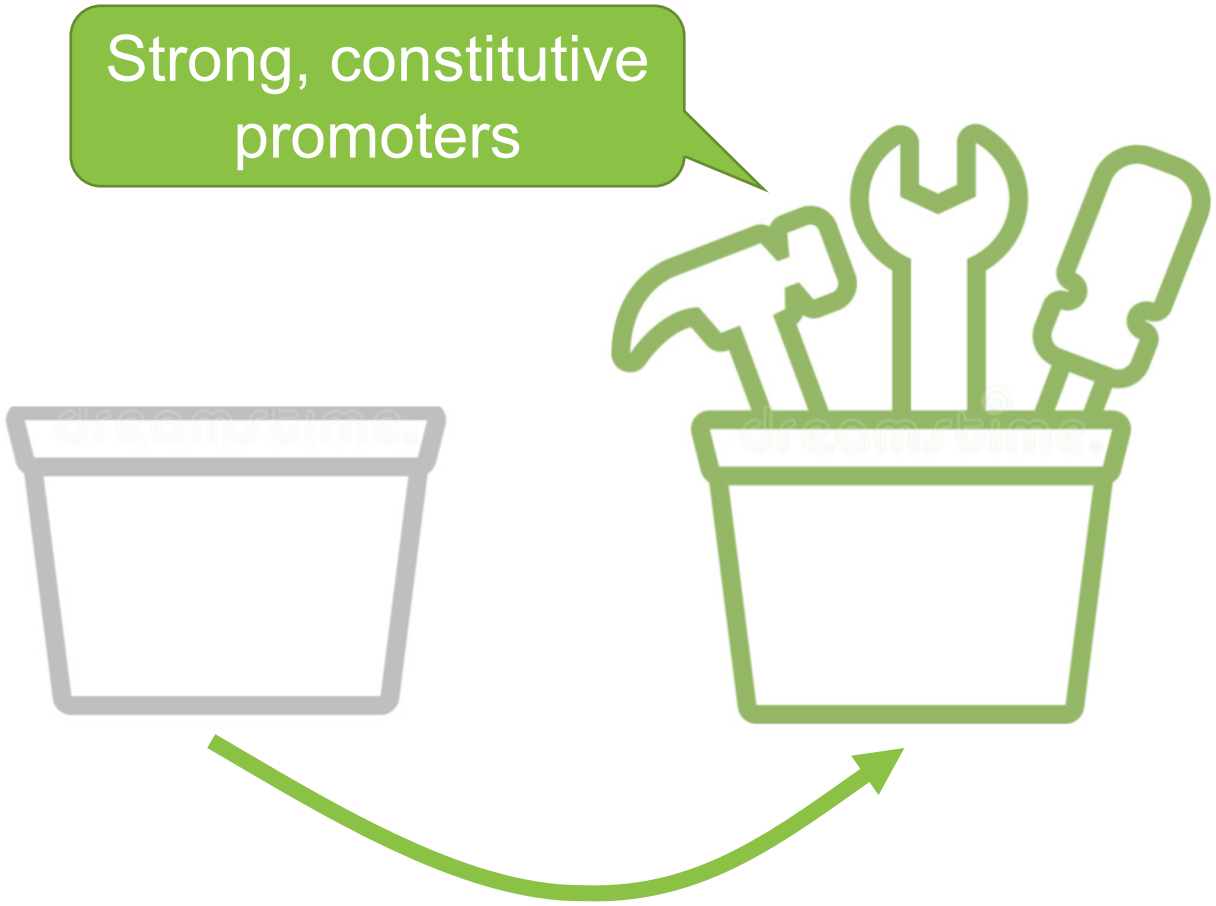
A computational framework for identifying promoter sequences in non-model organisms using RNA-seq datasets

Erin H. Wilson[1], Joseph D. Groom[2], M. Claire Sarfatis[3], Stephanie M. Ford[2,†], Mary E. Lidstrom[2,3], David A. C. Beck[2,4,*]

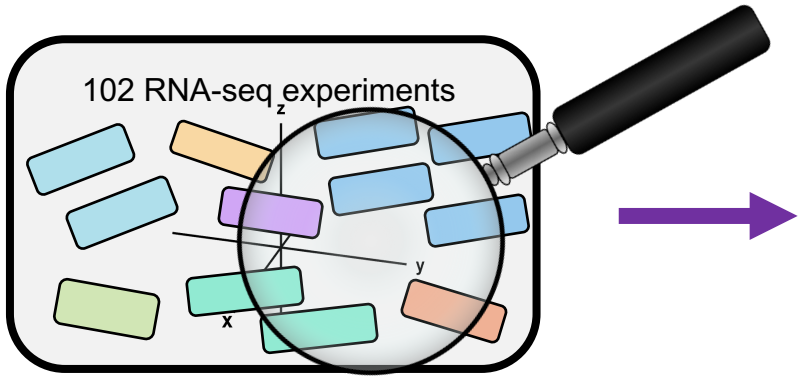
ACS Synthetic Biology

Accepted last week! 

Previous project main message:



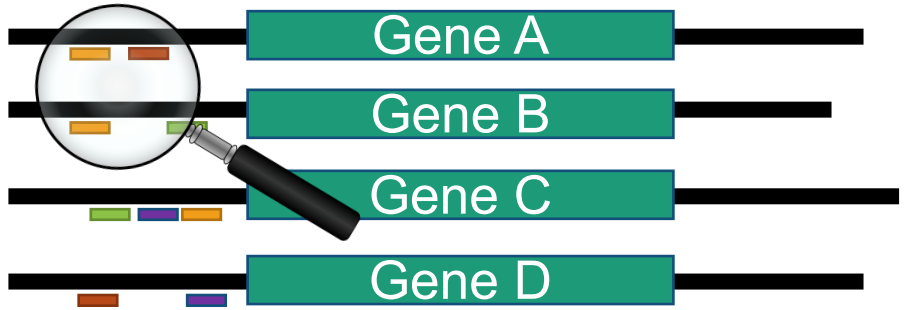
Previous project – the gist



Set of "top" genes

- Gene A
- Gene B
- Gene C
- …

Search upstream seqs for promoter signal

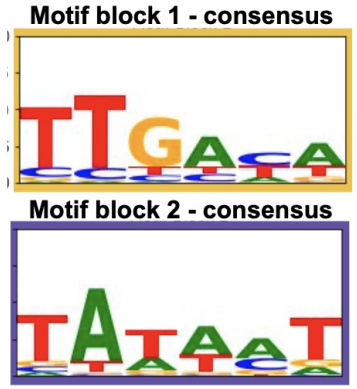


27 - 30bp

Predict best "core promoter" for each top gene

- Gene A → TTGACAGCTGCACGCTAGGTAGCTATA
- Gene B → TATACAGCGGTCAAGCTAACTGCTACT
- Gene C → TTTACAGCTGGCTAAGTCGTAGCTATT
- …

Compile "strong core promoter" consensus



Validate predictions

Computational

Search genome for consensus motif

Experimental

XylE reporter assay

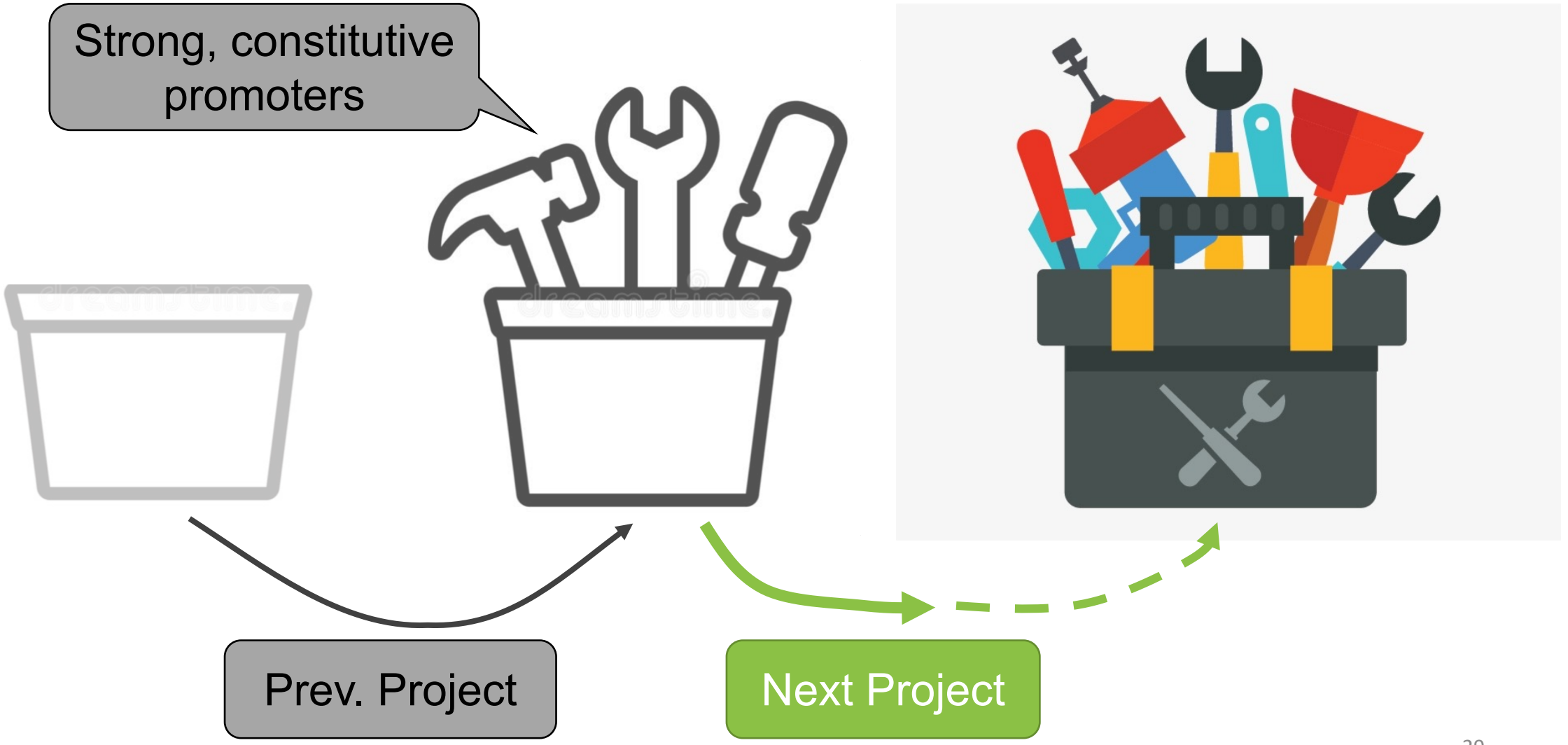
Validation

Overview

- Background
  - Microbes and Metabolism and Methane, oh my!
- Dataset and previous project
  - A hunt for strong promoters

- **Idea/early work on new project**
  - **more nuanced promoter tools**
  - ***Would love Feedback!***

New project – extending a genetic toolkit



New project direction:

a deep learning approach to identify useful sequences for creating more nuanced promoter tools

Why nuanced promoters?

<u>Strong promoters are a good start… but:</u>

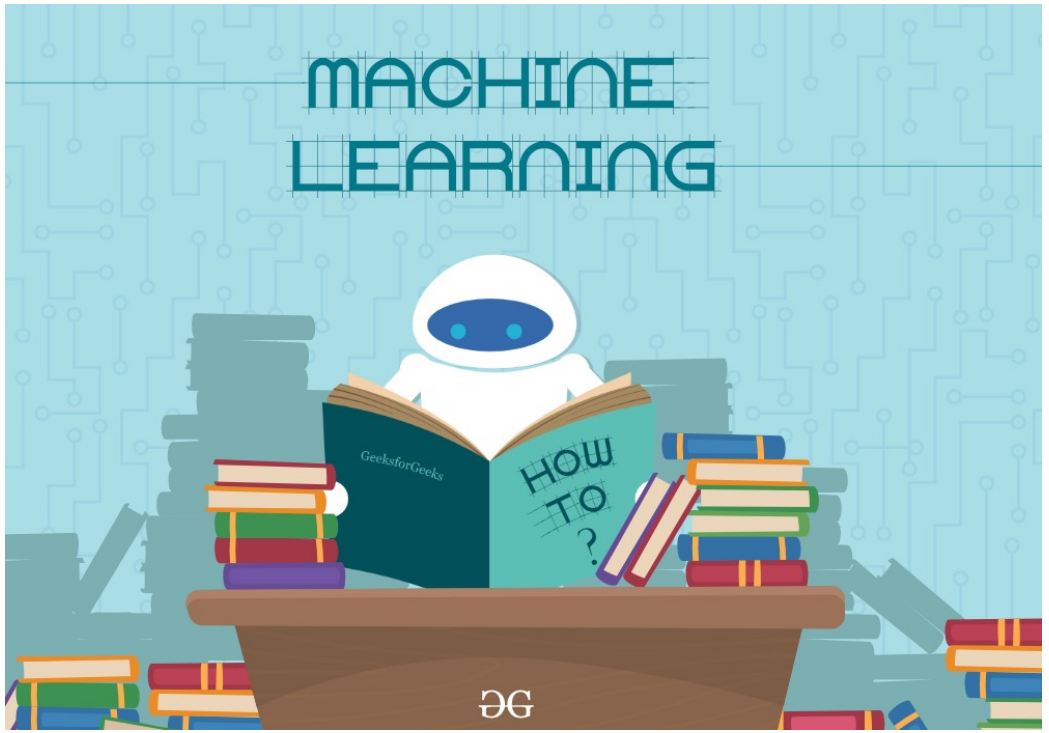
- "fire hose" expression approach has **biological limits**
- Range of promoter strengths → more fine-tuned expression control
- Inducible promoters → key innovation for producing molecules at large scales

Inducible promoters are way more useful than constitutive

Why deep learning?

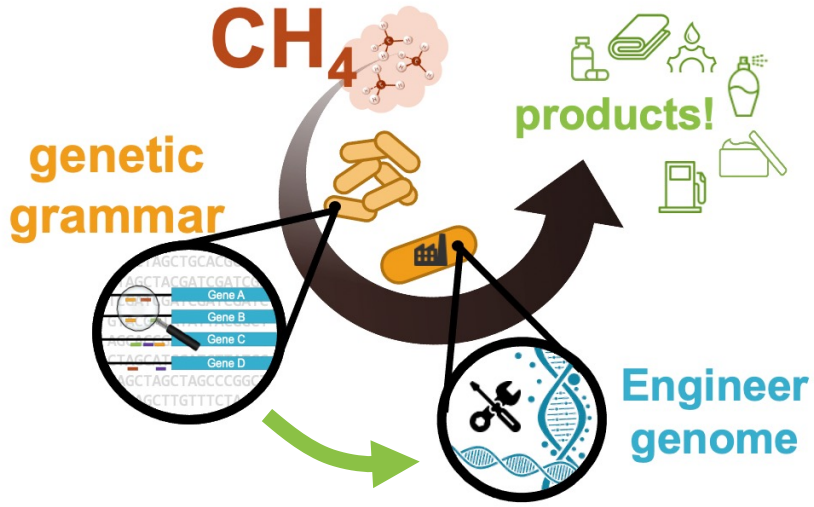
Deep learning is pretty good at:

- learning important features **without prior knowledge**
- finding **small, relevant patterns** in larger contexts
- learning **non-linear** combinations of features



Big Goal: identify important regulatory motif patterns

More specifically: sequence patterns that **promote** or **repress** gene expression in specific, controllable conditions
(can be used as an expression tool for met. engg.)

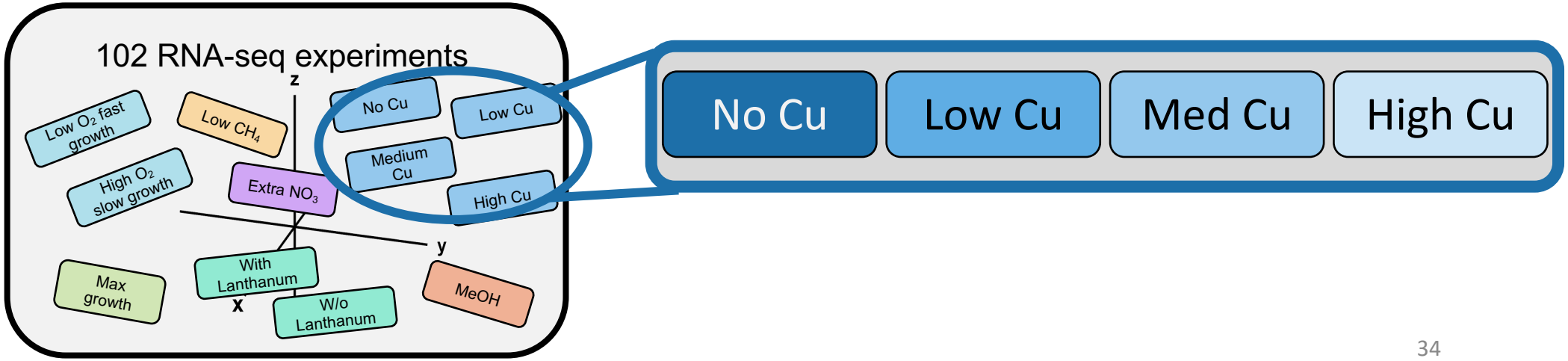


Most specifically (given our lab's data): a sequence pattern that promotes/represses expression in **response to Copper**
(useful as a metabolic switch tool?)

Recall: the data in play

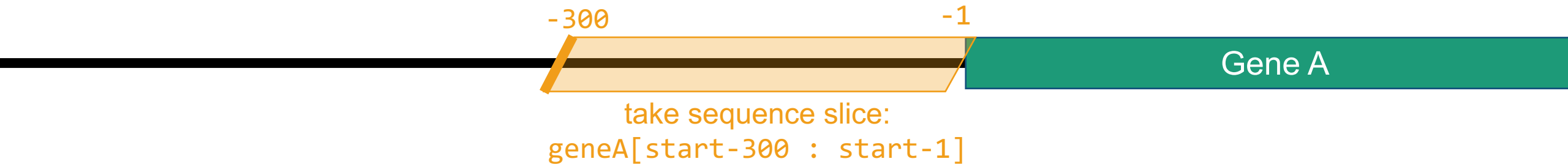
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	…
Gene A	7	213	44	55	456	25	77	124	
Gene B	12	574	78	65	227	413	158	71	
Gene C	201	99	5	374	87	145	352	47	
…									. . .

	Condition 1	Condition 2	Condition 3	…
Gene A	7	213	44	
Gene B	12	574	78	
Gene C	201	99	5	
…				. . .

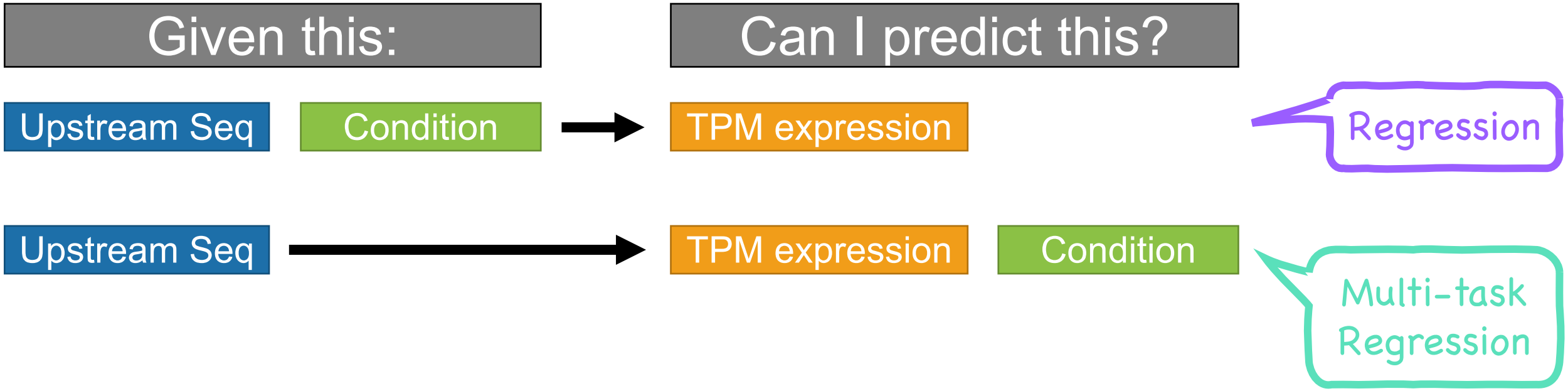


Extract upstream sequences as approximate promoter regions

Extract upstream sequences as approximate promoter regions



For each gene, we know:
TPM expression, experimental condition, upstream seq



Deep learning approaches on DNA inputs is not new

Zhou 2015

Basset/Basenji

Kelley 2016/2018

Promoter or NOT promoter

Typically **classification** tasks using ChIP-seq

Often eukaryotes

Use deep learning on DNA sequences to predict regulatory features

Presence/ absence of TFBS

Random seq library → predict strength

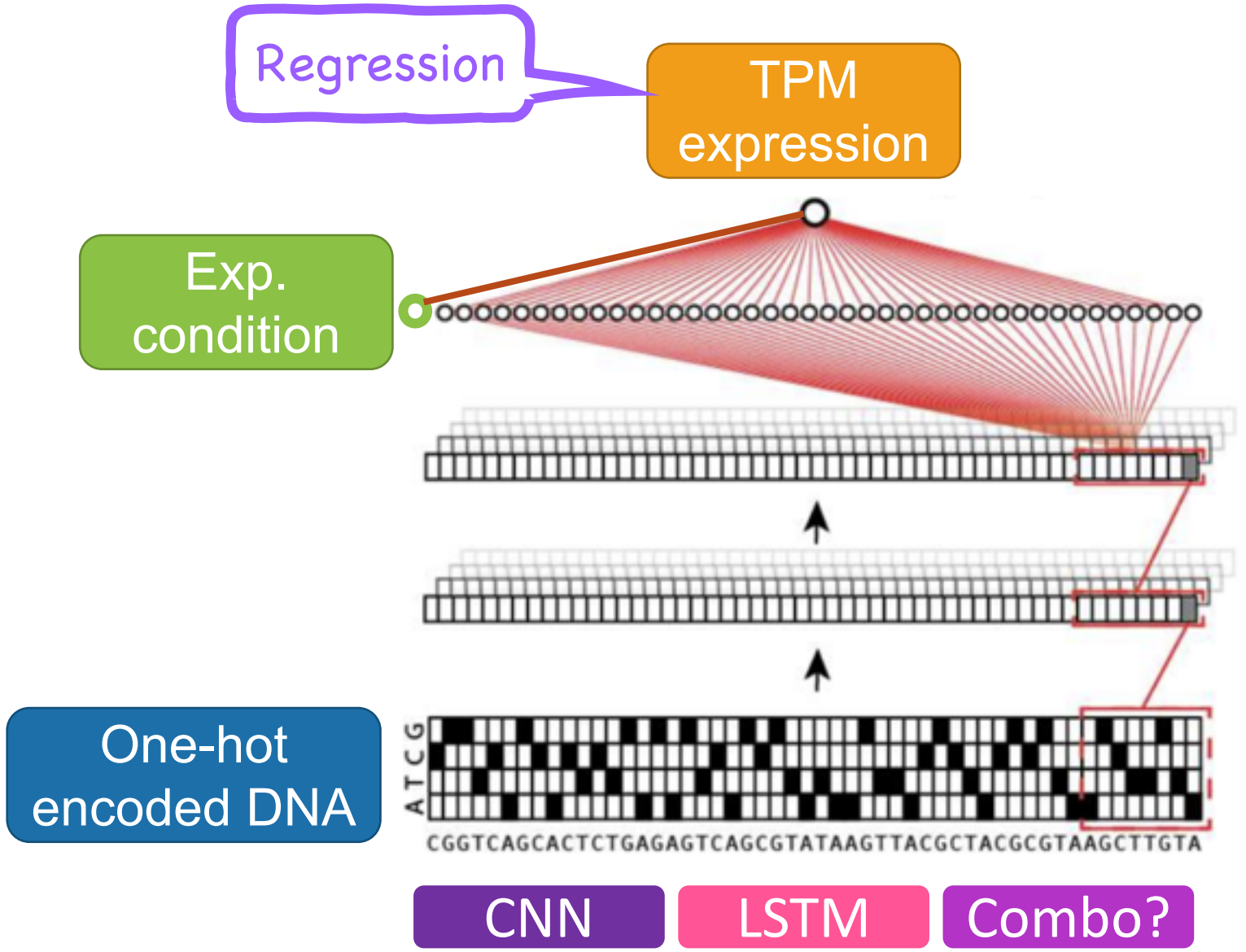
A basic DNA deep learning framework

Given this:

Upstream Seq | Condition

Can I predict this?

TPM expression



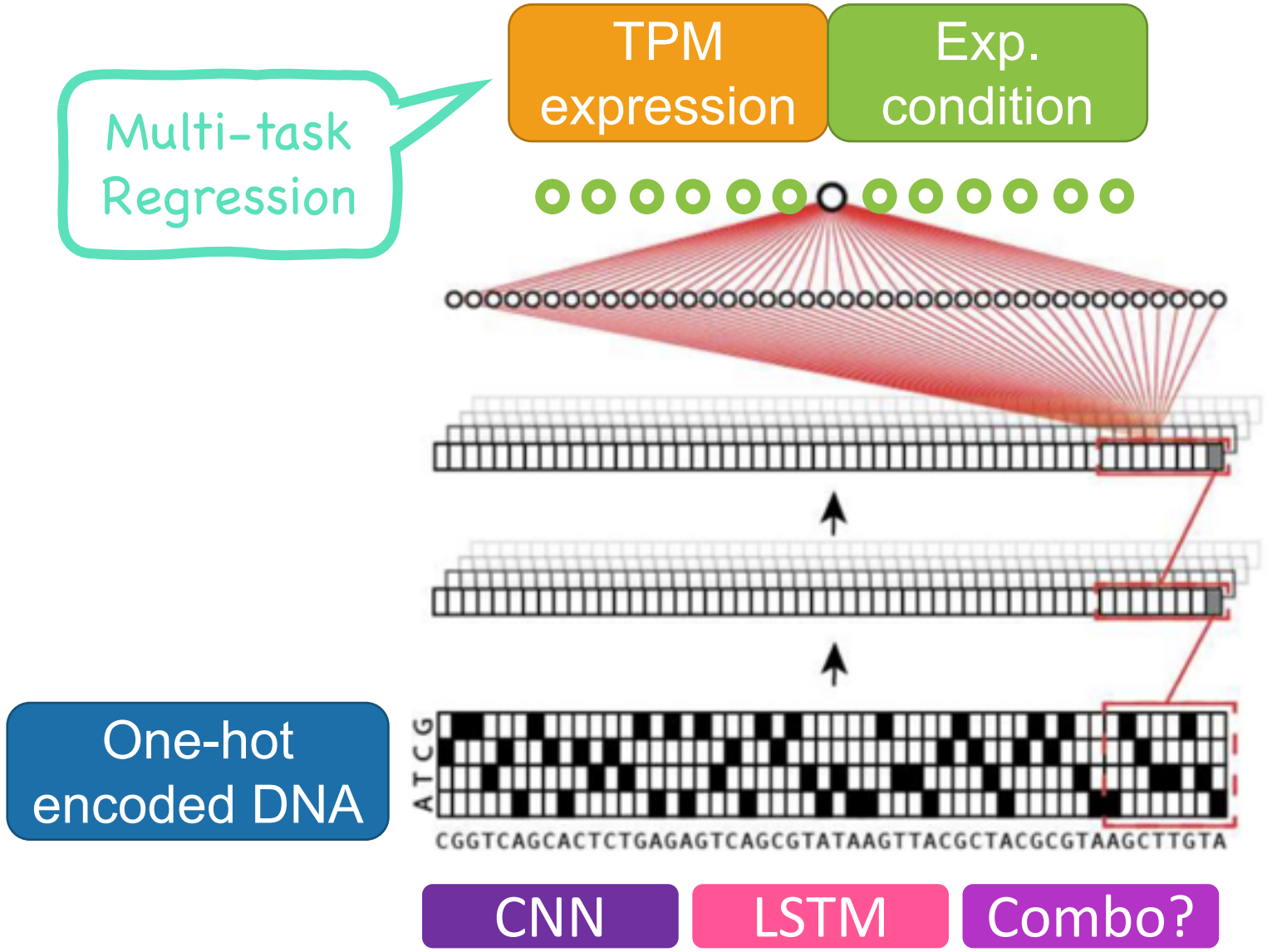
A basic DNA deep learning framework

Given this:

Upstream Seq

Can I predict this?

TPM expression | Condition



Final set of text-heavy slides

1. Rough sketch of project plan
2. What if this worked??
3. Expected challenges
4. Current status
5. Open questions!

A rough sketch:

1. Decide on a **suite of model types** to try and compare
   - Baselines: (probably should be bad?)
     - just predict average
     - linear regression by position
     - linear regression by k-mer counts
   - Deep learning:
     - CNN
     - LSTM
     - combo CNN+LSTM
     - others?

2. Model **evaluation**:
   - MSE? (error on TPM predictions?)

3. **Feature analysis** for bio insights (eg. ID important motifs)
   - CNN filter activations
   - Feature attribution
     - DeepLift
     - DeepShap
     - Scrambler Networks

What would this mean if it worked?

1. Given a **new upstream sequence**, now you can predict how it may influence gene expression across a range of conditions
    - *Use specific promoter sequence in front of heterologous genes when installing pathways?*

2. Once a model is trained and "good", go back and **analyze model features** for biological insights? (CNN filters? Feature Attribution methods?)
    - *Perhaps could reveal small, testable regulatory motifs?*

3. Once a model is trained and "good", could you freeze the parameters and use it to **design a sequence** for a particular objective?
    - *If you want a gene to express in a certain pattern across variable experimental set ups?*

Possible Challenges

Too few examples?

- ~2,000 genes (4,000 genes minus possible “in-operon genes” filtered out)

Some genes are **MEGA expression outliers** (orders of magnitude)

- MSE error could be GIANT during training?

Looking for **novel motifs**… but I don’t actually know what they look like… how to be confident enough to ask an experimentalist to test?

Similar papers seem to mostly use ChIP-seq or other “peak” related data… I’ve got **bulk RNA-seq**

Noisy Data – cobbled dataset, large promoter windows

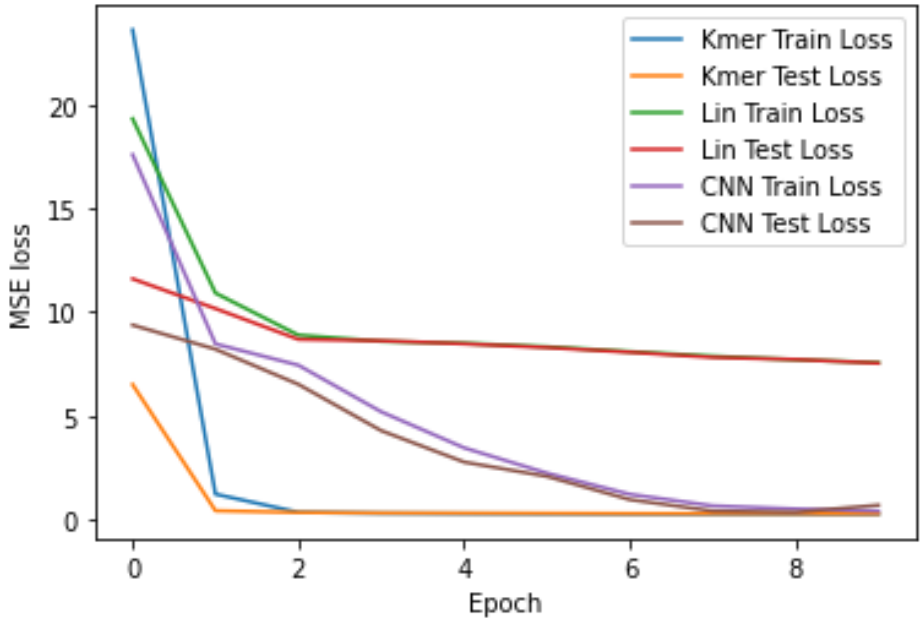
Current status

- **Learning PyTorch!**
  - Synthetic DNA seq dataset
  - Practice connecting tubes
- Plan to submit to “Proposal” track of **Climate Change AI** Workshop??

```
score_dict = {
    'A':20,
    'C':17,
    'G':14,
    'T':11
}

def model(seq):
    '''Return average score, T
    score = np.mean([score_dic
    if 'TAT' in seq:
        score += 10
    if 'GCG' in seq:
        score -= 10

    return score
```

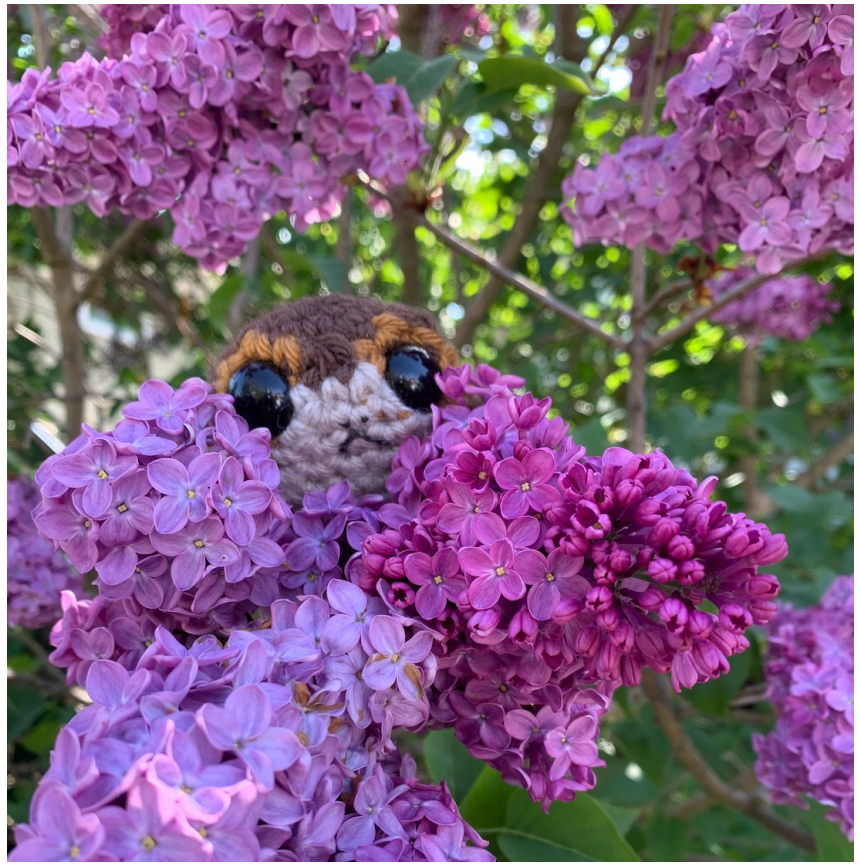
Try CNN??

```
class DNA_CNN_Multi(nn.Module):
    def __init__(self,
                 seq_len,
                 num_filters=32,
                 kernel_size=3,
                 lin_share_size=10
                ):
        super().__init__()
        self.seq_len = seq_len

        self.conv_share = nn.Sequential(
            nn.Conv1d(4, num_filters, kernel_size=
            nn.ReLU(inplace=True),
            nn.Flatten(),
            nn.Linear(num_filters*(seq_len-kernel_
            nn.ReLU(inplace=True),
        )

        # define the multi task objectives?
        self.obj0  = nn.Linear(lin_share_size,1)
        self.obj1  = nn.Linear(lin_share_size,1)
```



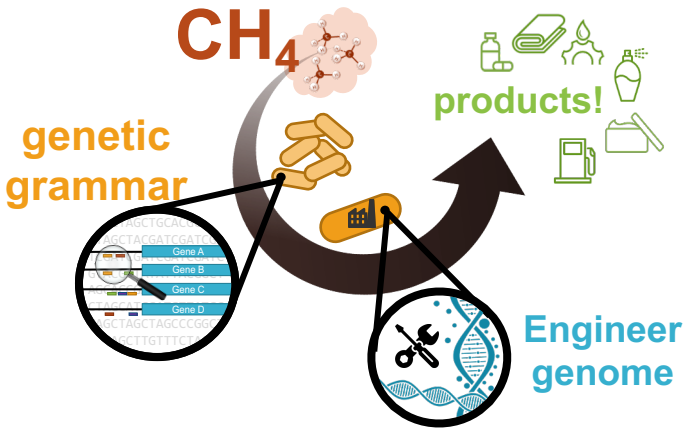
Open questions for you!

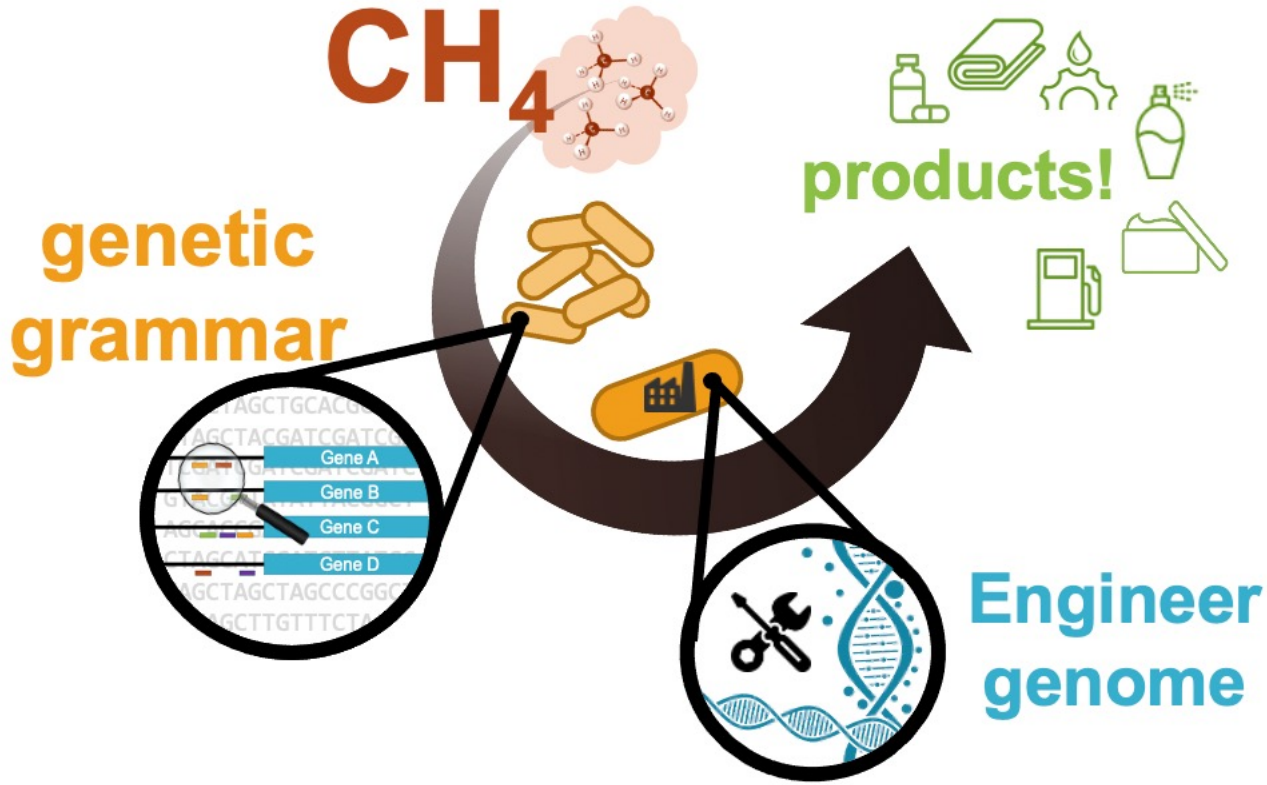
- **Given the data I have (type and amount), do these regulatory motif questions sound answer-able?**
- **Does this approach sound reasonable? Useful?**
- **Any big "gotchas" I'm missing?**




Thanks for listening!
Mojave Desert, CA

https://www.science-sparks.com/know-your-greenhouse-gases/


CH_4
genetic grammar
products!
Gene A
Gene B
Gene C
Gene D
Engineer genome


CH_4
genetic grammar
products!
Gene A
Gene B
Gene C
Gene D
Engineer genome