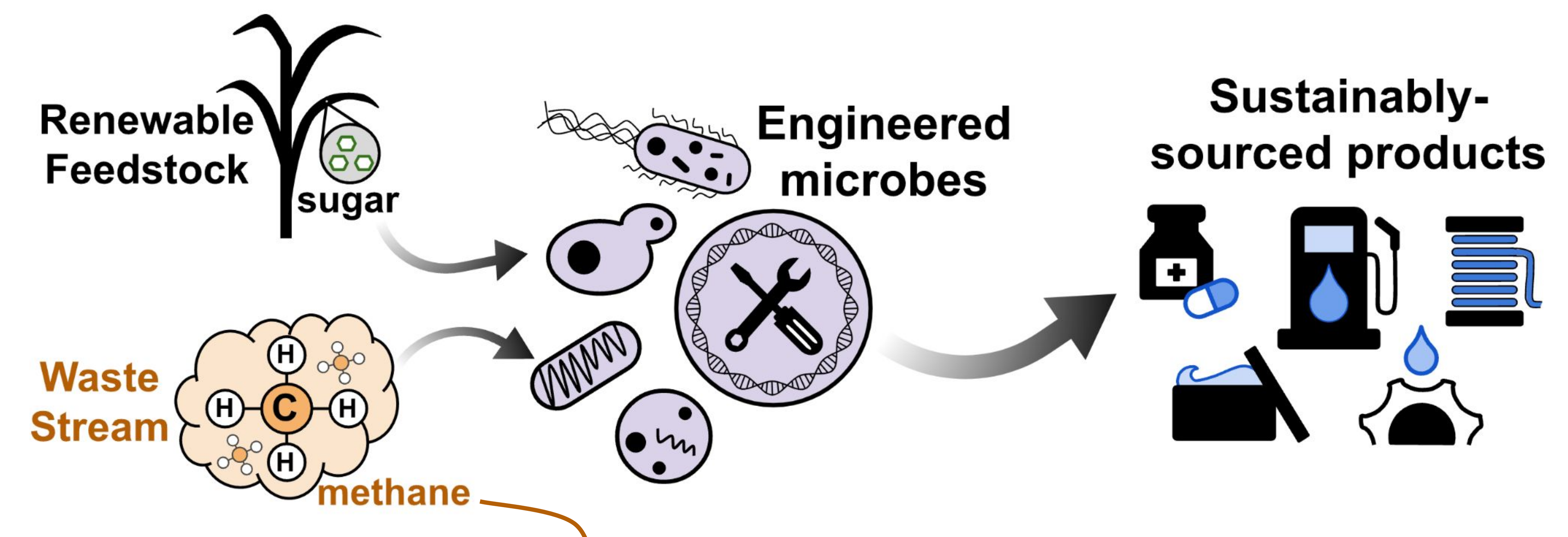
Methane, Microbes, and Machine Learning:

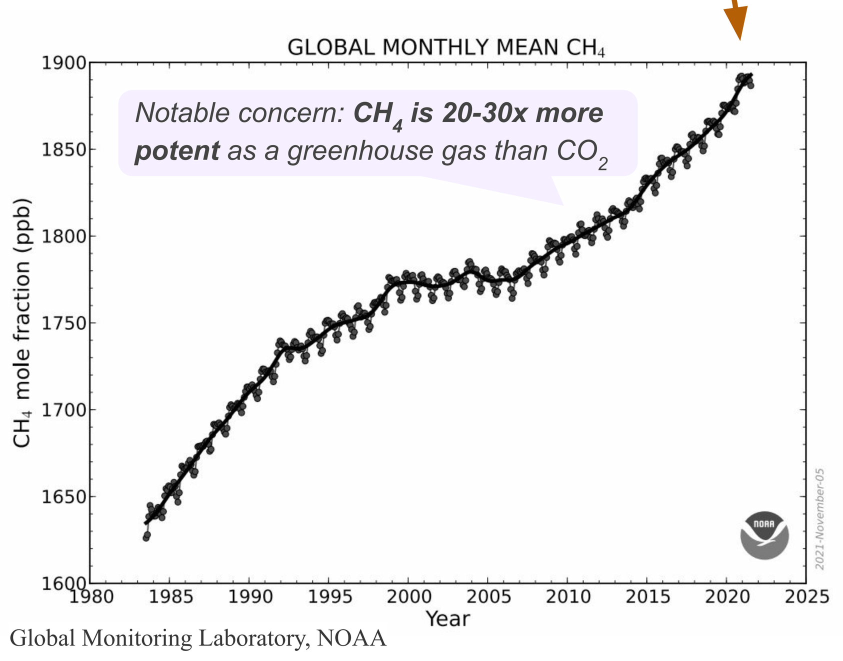
Engineering biology to combat climate change

Erin H. Wilson, Mary E. Lidstrom, David A. C. Beck
ewilson6@cs.washington.edu

1) A promising paradigm for methane mitigation

- Metabolic engineering: a field that aims to **engineer microorganisms into biological factories** that convert renewable feedstocks into valuable biomolecules.
  - ➔ Provides a more **sustainable alternative** to sourcing many materials, especially petroleum-based products
- Much progress with model organisms (baker's yeast and *E. coli*) to produce **malaria medicine, jet fuel,** fragrances



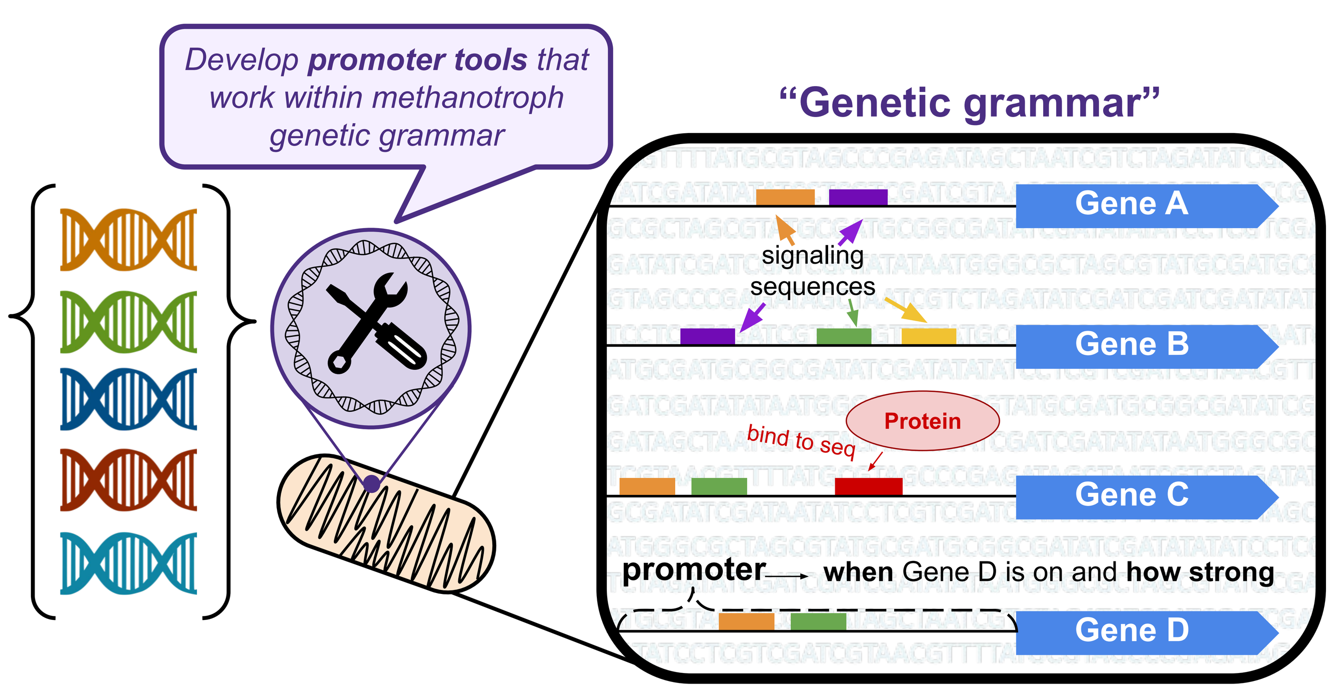


- Methanotrophs - **bacteria** that can **survive on methane** as their sole carbon source - are promising microbial hosts for industrial biomolecule production

★ Opportunity to **divert methane waste streams** into **valuable everyday materials**

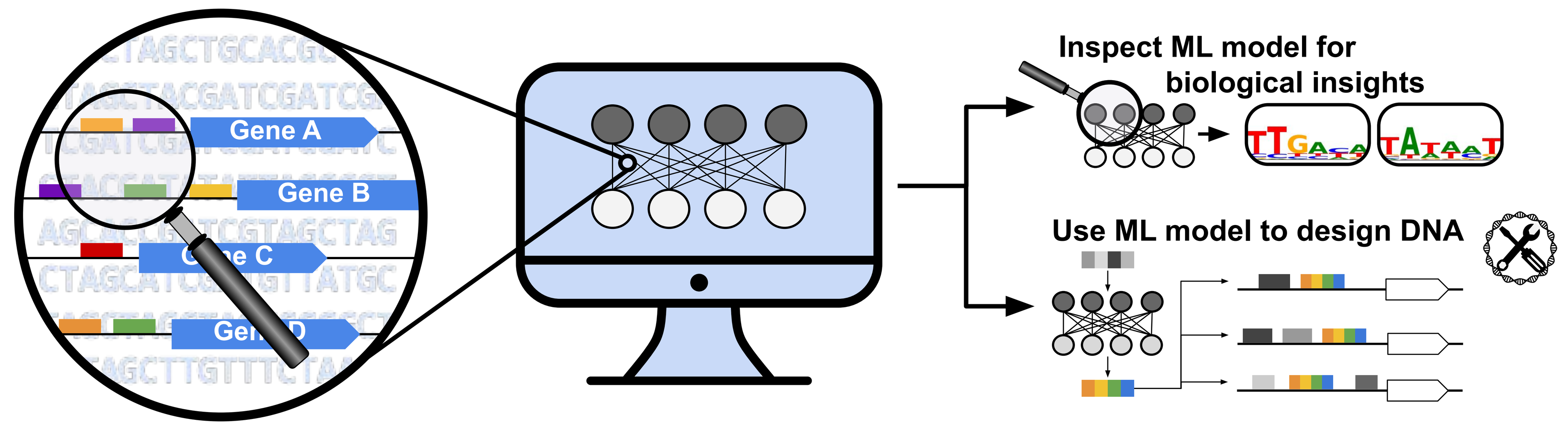
2) Regulatory DNA is a complex language to decipher

Expression of ***newly installed genes*** *must be carefully* ***controlled*** *in the host microbe*



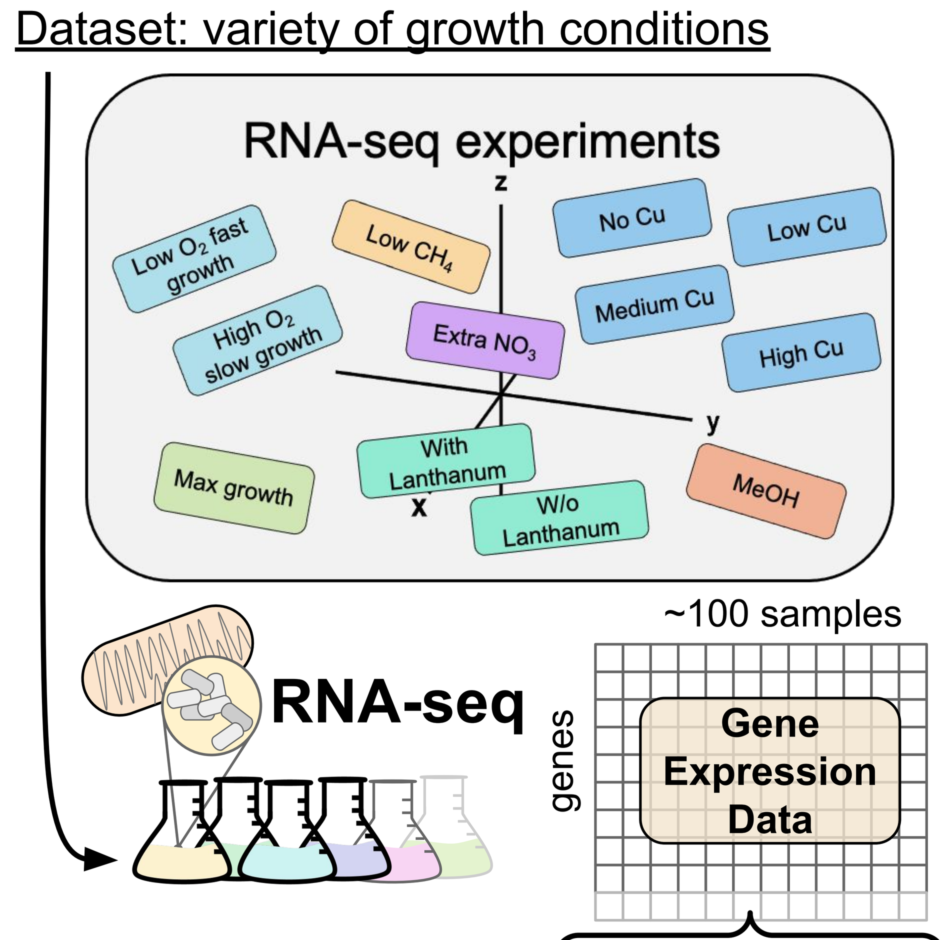
- Every microbe has evolved a different **genetic grammar**: a series of **signaling sequences** and logic patterns it uses to control its genes
  - ➔ *Promoter = sequence region containing many signals that influence when genes turn ON or OFF ("expression")*
- We must understand this grammar in order to **efficiently reprogram cells** for biomolecule production
  - ➔ *Research goal: develop methanotroph promoter tools*

3) Machine learning to automatically detect patterns in DNA



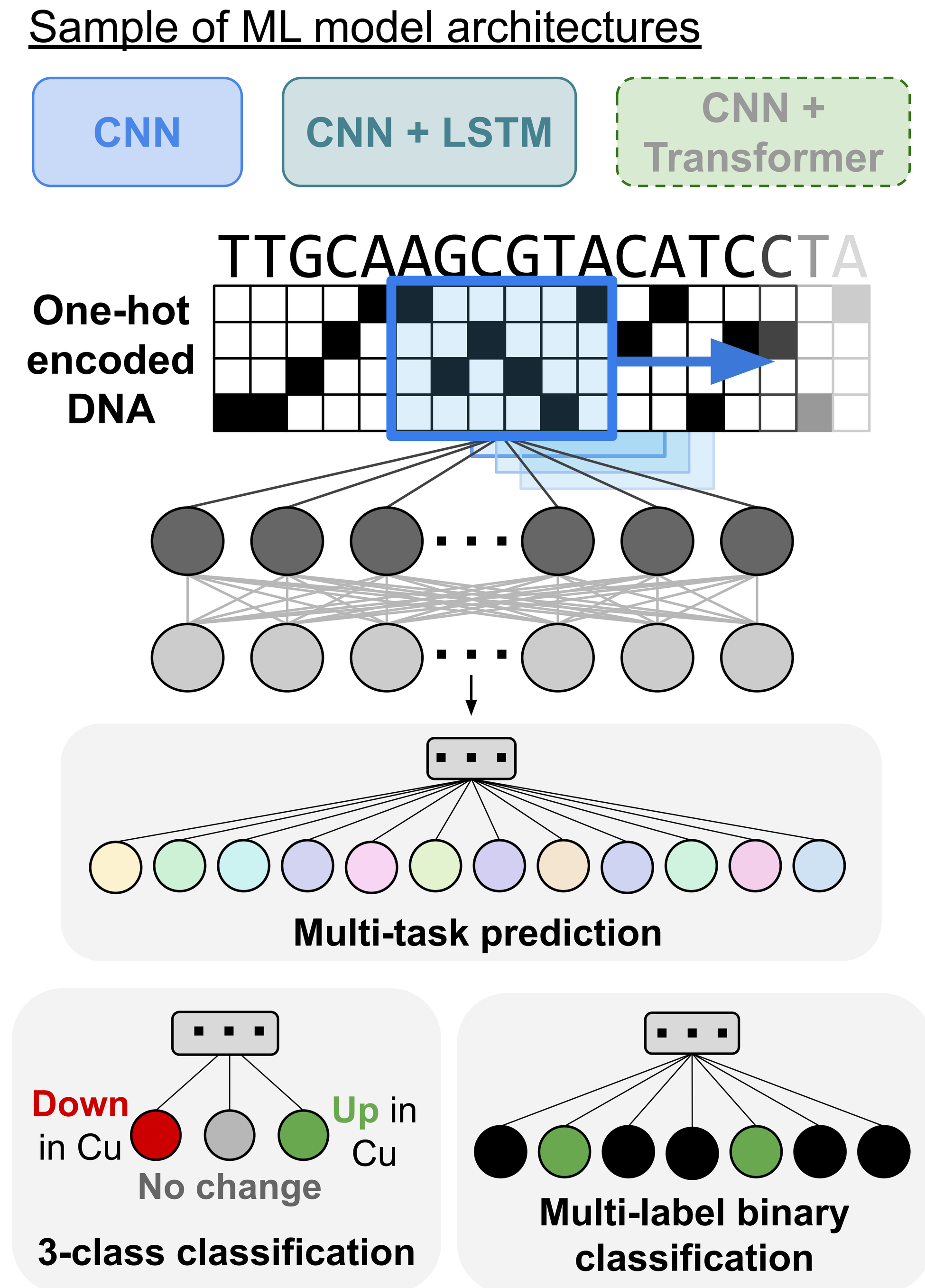
- Most DNA sequence signals are still unknown in methanotrophs
- Deep learning approaches can **learn relevant features directly from the data** without explicit encoding
  - ➔ Use deep learning models to **find patterns within methanotroph promoter sequences**
- **Biological insights:** what DNA patterns has the model learned?
- **Novel DNA:** freeze model and use for forward **DNA design**

Dataset: variety of growth conditions



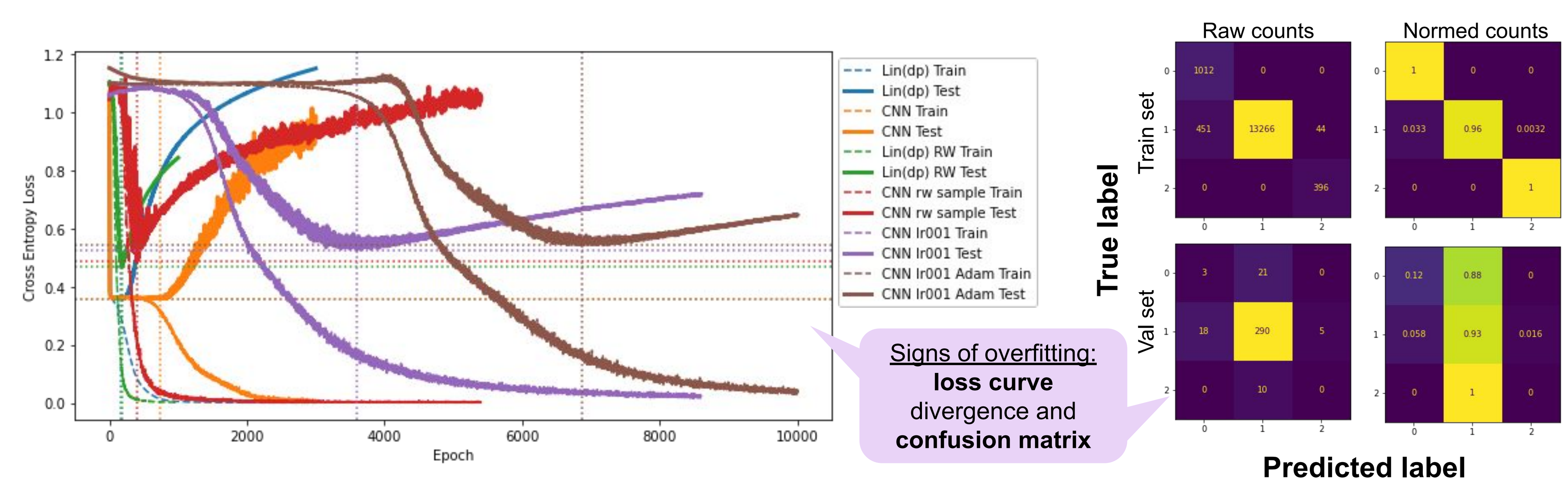
	X examples	Y labels			
Gene	Upstream seq	NoCu	lowCu	highCu	...
Gene 1	AGTACGTGTCCAGAT...	221	34	53	
Gene 2	GGCTGATGCTATACG...	32	73	3	
Gene 3	AGGGCAACGCGATCA...	893	90	8	
Gene 4	CCCGTGCAACTCACA...	6	359	78	
Gene N	GTGTGTACCGCCGAA...	209	8	7	

Sample of ML model architectures



4) Addressing key challenges: overfitting, dataset size, imbalance

- Current models are **overfitting** to the training data, despite initial strategies to address **class imbalance** and **limited data**



- Future work: **self-supervised pre-training** on general sequence tasks; fine tune model to methanotroph RNA-seq data

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING