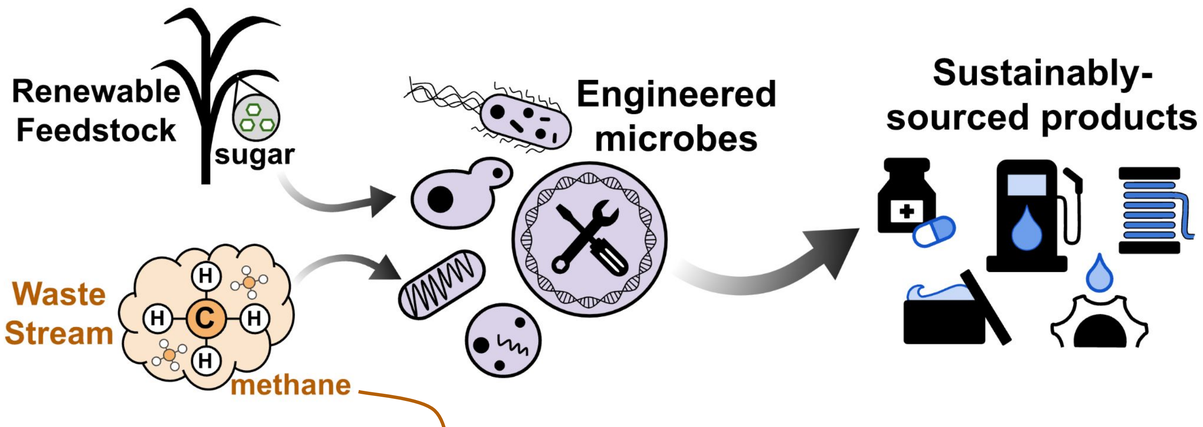
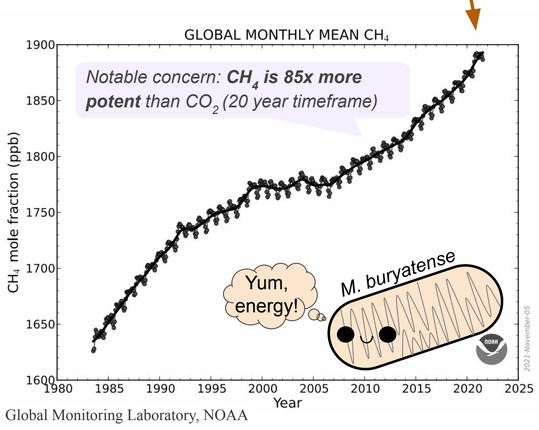
Probing the limits of deep learning methods for predicting gene expression in non-model microbes

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

1) A promising paradigm for mitigating methane emissions

- Metabolic engineering: a field that aims to **engineer microorganisms into biological factories** that convert renewable feedstocks into valuable biomolecules.

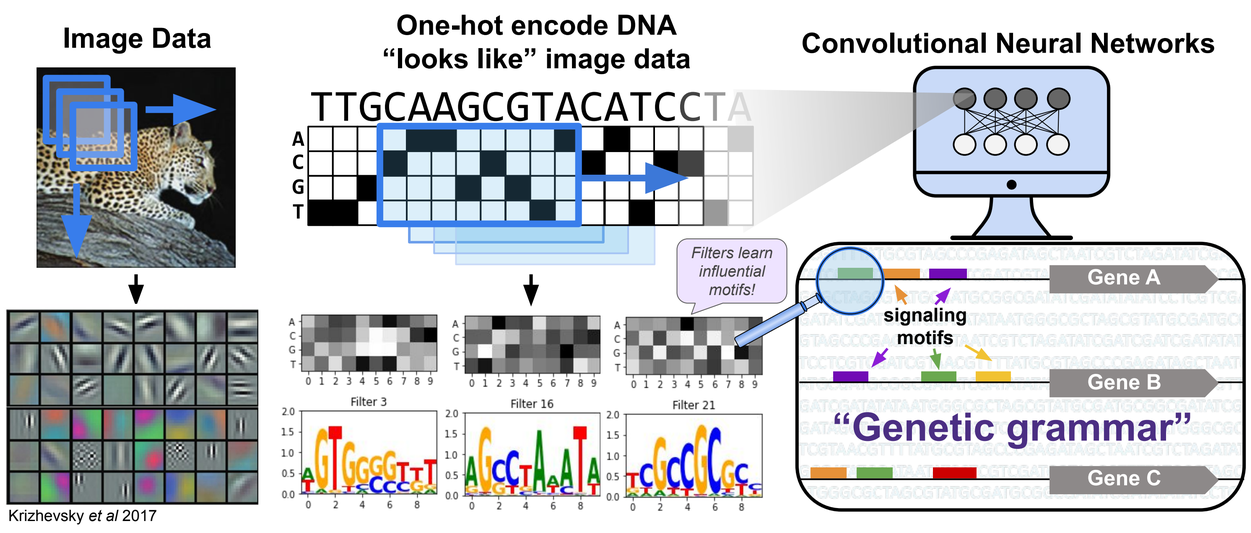




- Methanotrophs - bacteria that can **survive on methane** as their sole carbon source - are promising microbial hosts for industrial biomolecule production
- ★ Opportunity to **divert methane waste streams** into **valuable everyday materials**

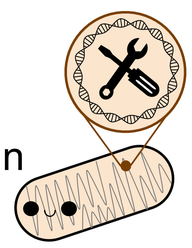
2) Machine learning approaches can automatically detect patterns in DNA

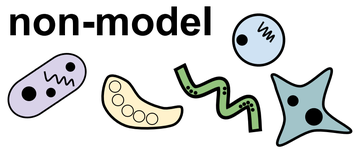
- Most regulatory signals are still unknown in *M. buryatense*
- Deep learning approaches can **learn relevant features directly from the data** without explicit encoding



Research goals:

- ➔ Use deep learning models to decode *M. buryatense* genetic grammar by **finding influential motifs** within promoter regions
- ➔ Expand **metabolic engineering tools** for *M. buryatense*
- ➔ Maintain general approach: apply to other **non-model organisms** with limited data





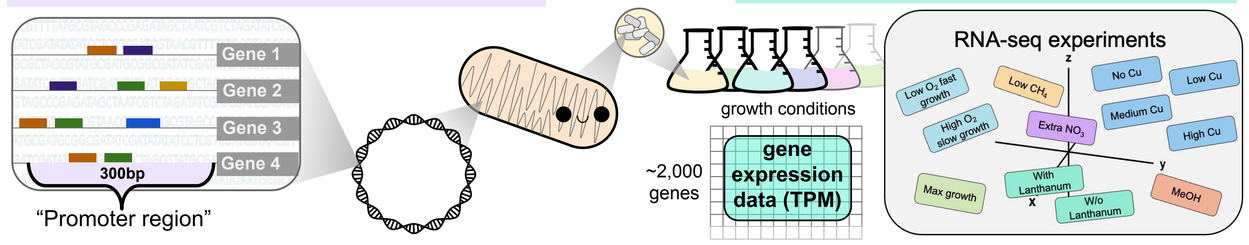
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

GRFP NSF GRADUATE RESEARCH FELLOWSHIP PROGRAM

3) Models struggle to predict RNA-seq expression from promoter regions

Input data: upstream promoter regions

Labels to predict: RNA-seq values



➔ **Learning objective:** predict gene expression outcomes from promoter DNA sequences

X examples

Gene	Promoter_region
Gene 1	AGTACGTGTCCAGAT…
Gene 2	GGCTGATGCTATACG…
Gene 3	AGGGCAACGCGATCA…
Gene 4	CCCGTGCAACTCACA…
Gene N	GTGTGTACCGCCGAA…

Attempted many model arch. + hyperparameters

CNN | LSTM | CNN + LSTM

Y labels

Gene	NoCu	lowCu	highCu	…
Gene 1	221	34	53	
Gene 2	32	73	3	
Gene 3	893	90	8	
Gene 4	6	359	78	
Gene N	209	8	7	

➔ ***Model results:*** *generally poor performance across many architectures and task formulations, despite strategies to mitigate class imbalance and limited data* (😖 → 🤔)

4) Probing performance across varying levels of motif information density

- If an expression response is controlled by a simple activation or repression event, how much information *would* be enough?

synthetic motif prediction experiment

Dataset: ***N*** random DNA sequences of length ***L***

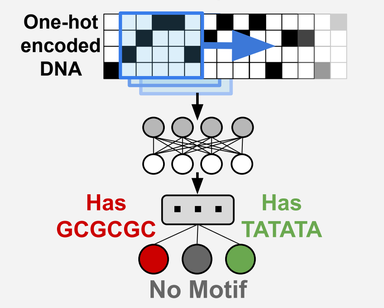
Objective: Train CNN to predict class

Class labels:
- 0 if **GCGCGC** present (minority)
- 1 if **No Motif** present (majority)
- 2 if **TATATA** present (minority)

→ *Control minority class imbalance at prevalence **P***

X / **Y**

ID	Random DNA	Class
1	AGTACGTGTCCAGAT…	1
2	GGCAC**TATATA**CCTG…	2
3	AGGGCAGCGAGCTCA…	1
4	CCCGTGCAACTCACA…	1
N	GT**GCGCGC**CGCCGAA…	0



Motif Information Density =

$$\texttt{motif\_len}/L * P * N$$

***M. buryatense* data insight:**

Even for expression responses as simple as a pair of activating and repressing motifs, these types of models are **unlikely to capture the signal** in a dataset this small.

>> More data are needed >>