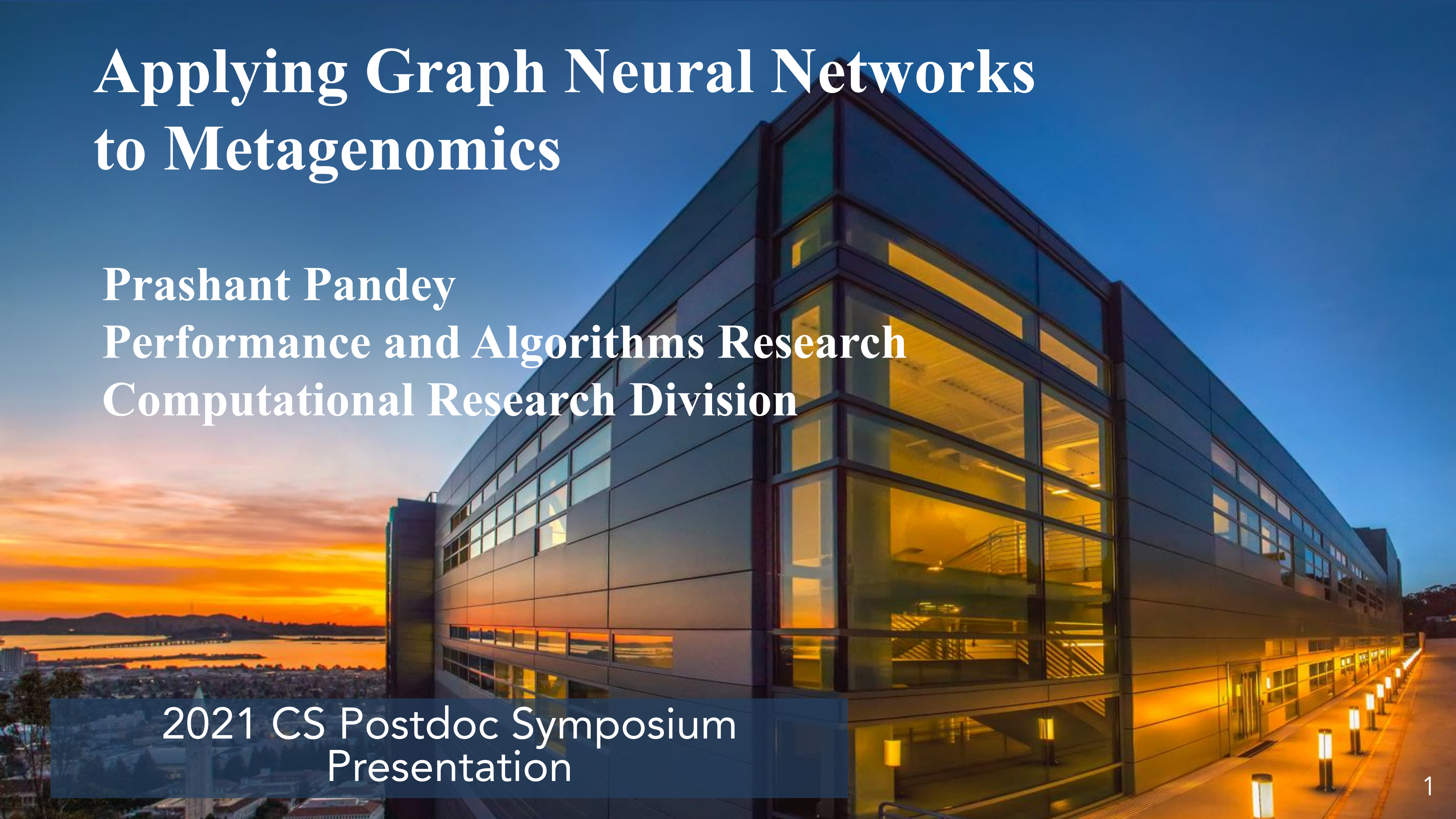
Applying Graph Neural Networks to Metagenomics

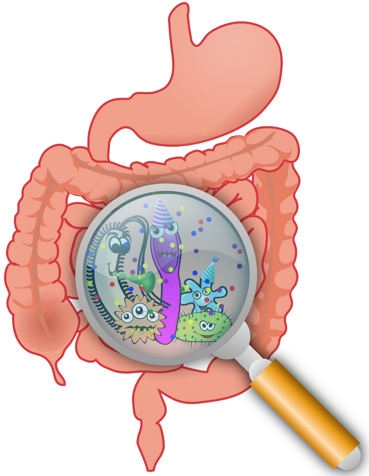
Prashant Pandey
Performance and Algorithms Research
Computational Research Division

2021 CS Postdoc Symposium Presentation



Metagenomics

Soil sample



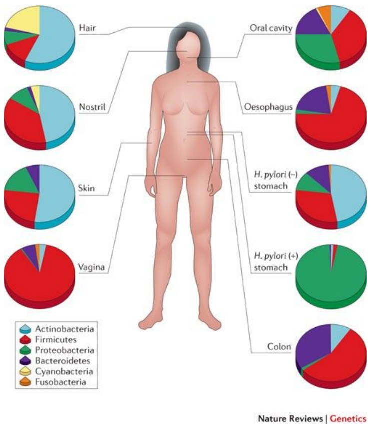
Human gut

Ocean sample

The ***study of microbes*** that inhabit an environment, such as soil, human gut, or ocean

Why study microbes in an environment?

Environmental science



Human health

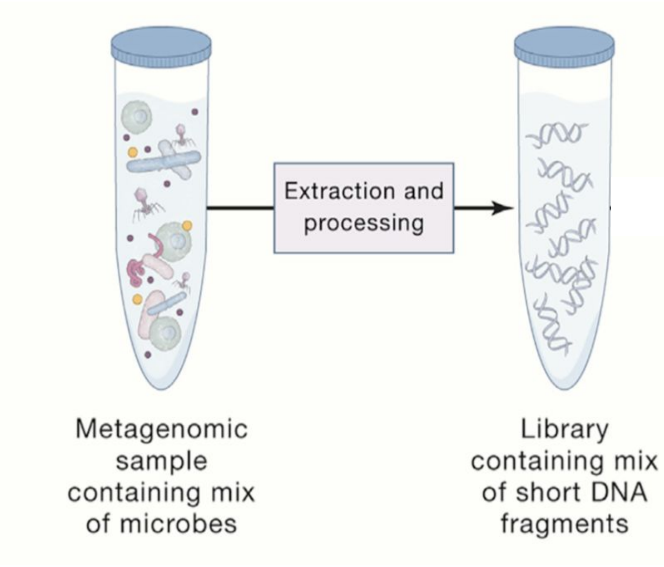
Industrial applications

Environment: elemental cycle, pollution control, cleanup, etc.
Human: protection from pathogens, immune systems regulation, etc.
Industrial: wastewater treatment, bioprospecting, fermentations, etc.

Classification is the critical first step in metagenomics

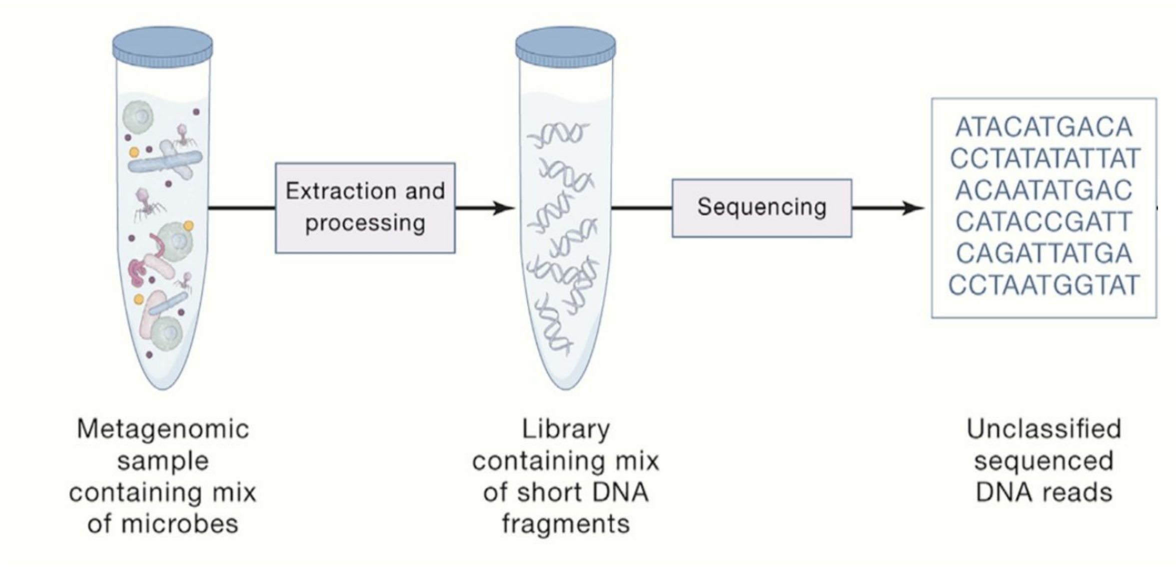
Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti **Cell 2019**

Classification is the critical first step in metagenomics


Extraction and processing
Metagenomic sample containing mix of microbes
Library containing mix of short DNA fragments

Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti **Cell 2019**

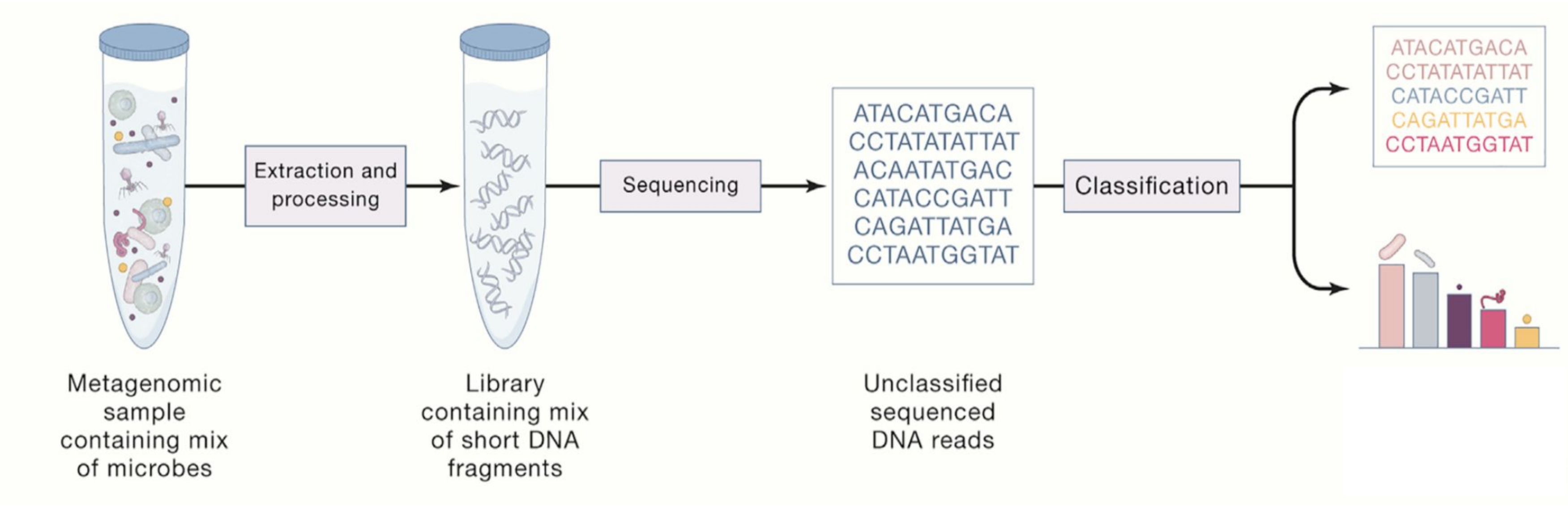
Classification is the critical first step in metagenomics


Extraction and processing
Sequencing
Metagenomic sample containing mix of microbes
Library containing mix of short DNA fragments
Unclassified sequenced DNA reads

ATACATGACA
CCTATATATTAT
ACAATATGAC
CATACCGATT
CAGATTATGA
CCTAATGGTAT

Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti **Cell 2019**

Classification is the critical first step in metagenomics


Extraction and processing
Sequencing
ATACATGACA
CCTATATATTAT
ACAATATGAC
CATACCGATT
CAGATTATGA
CCTAATGGTAT
Classification
ATACATGACA
CCTATATATTAT
CATACCGATT
CAGATTATGA
CCTAATGGTAT
Metagenomic sample containing mix of microbes
Library containing mix of short DNA fragments
Unclassified sequenced DNA reads

Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti **Cell 2019**

In this talk:

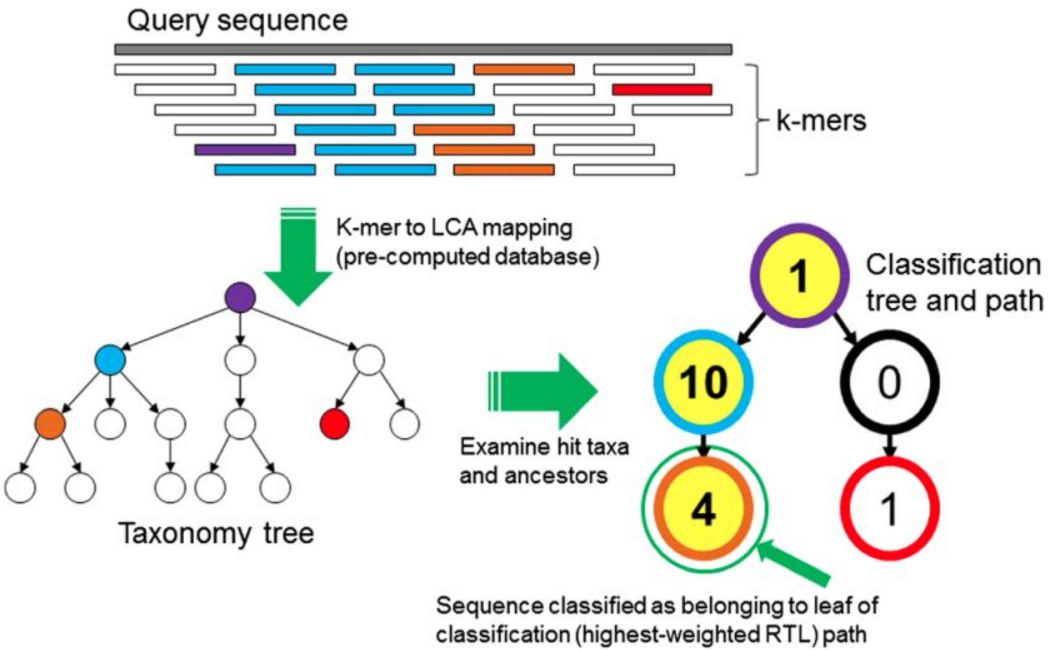
- MetaGNN: uses ***graph neural networks (GNN)*** to perform metagenomic classification
  - ○ Graph neural networks are employed in a ***semi-supervised*** manner
- MetaGNN uses both the ***sequence contents and connectivity information***
- Works for both ***short- and long-reads*** metagenomic data
- In our evaluation, compared to existing tools:
  - ○ Short reads: MetaGNN gets an ***order-of-magnitude higher*** accuracy
  - ○ Long reads: MetaGNN gets ***similar*** accuracy

Prashant Pandey, Giulia Guidi, Alok Tripathy, Aydın Buluc, and Katherine Yelick **2021**

Classification is computationally challenging

- High throughput sequencing generates millions of short sequences
- Aligning sequences to a databases of known genomes is not feasible
- Exponential growth of sequencing data makes the problem even more challenging

k-mer matching is the standard classification technique



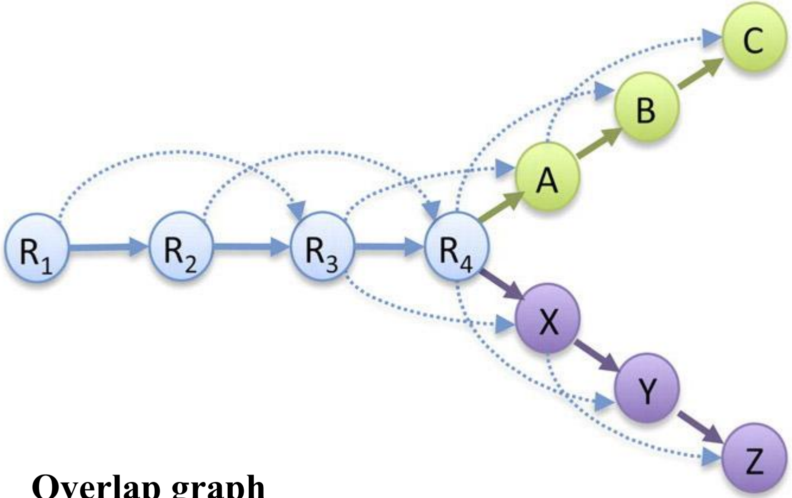
Derrick E Wood and Steven L Salzberg **Genome Biology 2014**
Derrick E Wood, Jennifer Lu, and Ben Langmead **Genome Biology 2019**

Using only nucleotide content is suboptimal

- Short sequences do not offer enough abundance information
- Low abundance species tend to be classified with similar species with high abundance
- Hard to distinguish between closely related species

Can we use connectivity information?

```
R1:  GACCTACA
R2:   ACCTACAA
R3:    CCTACAAG
R4:     CTACAAGT
A:       TACAAGTT
B:        ACAAGTTA
C:         CAAGTTAG
X:       TACAAGTC
Y:        ACAAGTCC
Z:         CAAGTCCG
```

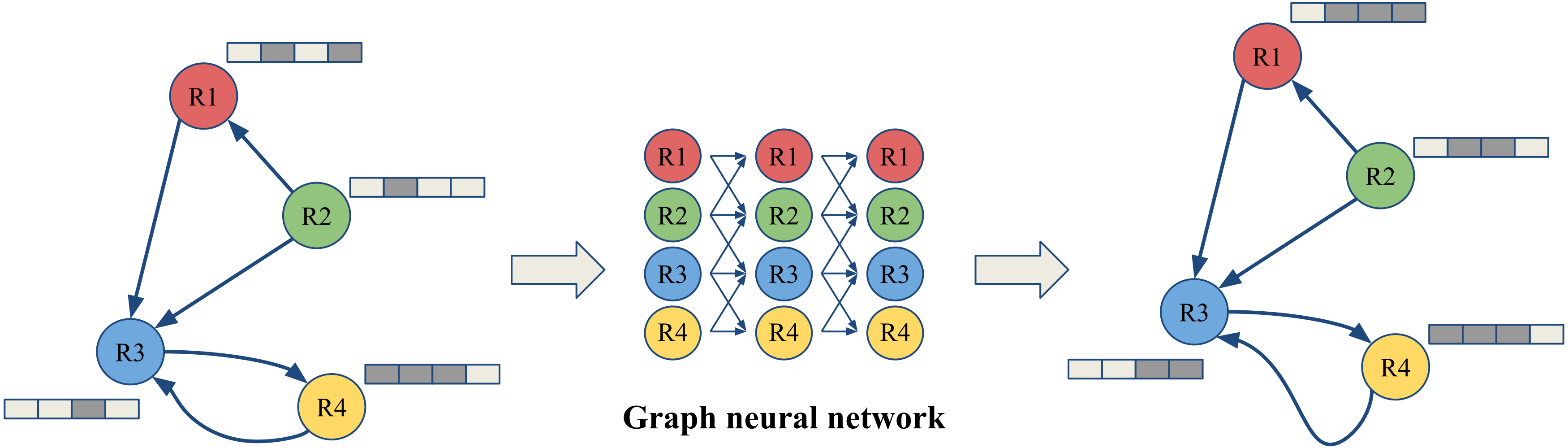


Overlap graph

Nodes: sequences, Edges: overlaps between sequences

Schatz MC, Delcher AL, Salzberg SL **Genome Research 2010**

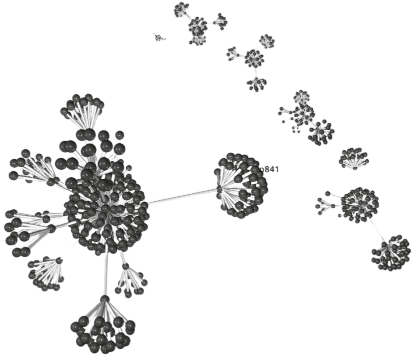
Applying GNN to overlap graphs



A GNN learns embeddings for each node in the graph using neighborhood aggregation

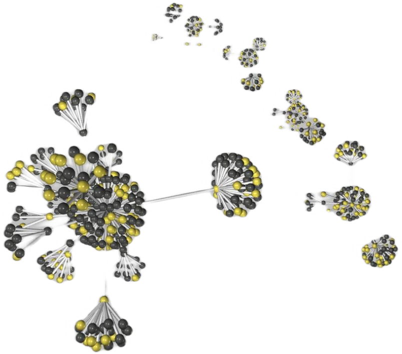
Node embeddings can be further used for node label prediction

MetaGNN pipeline



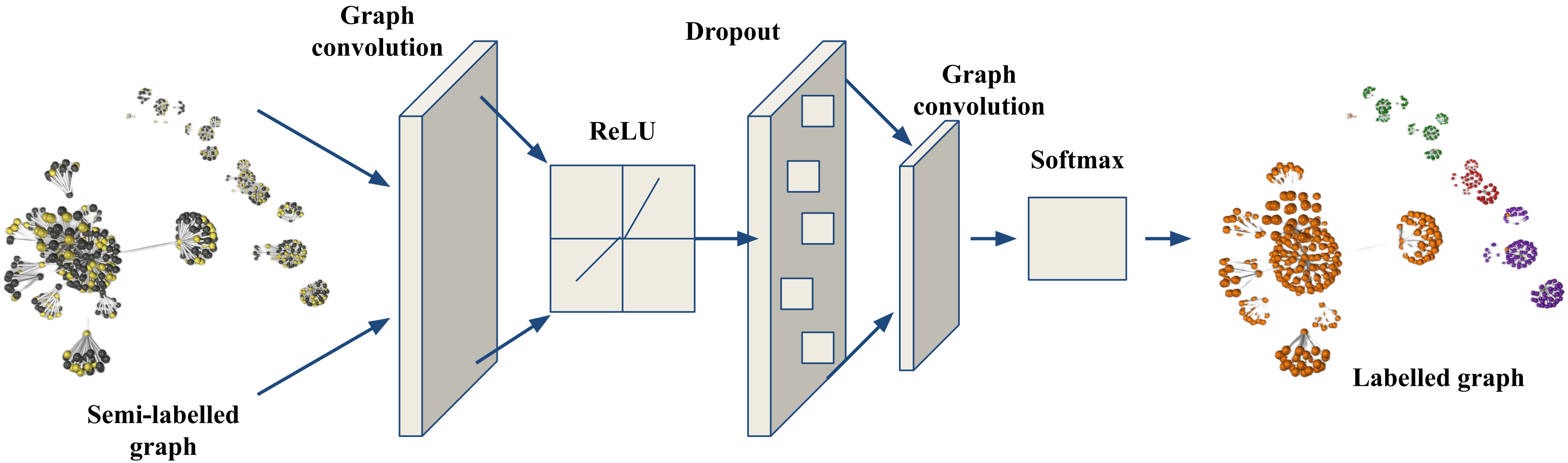
Overlap graph with ***tetra-nucleotide content*** of sequences as ***node embeddings***

MetaGNN pipeline



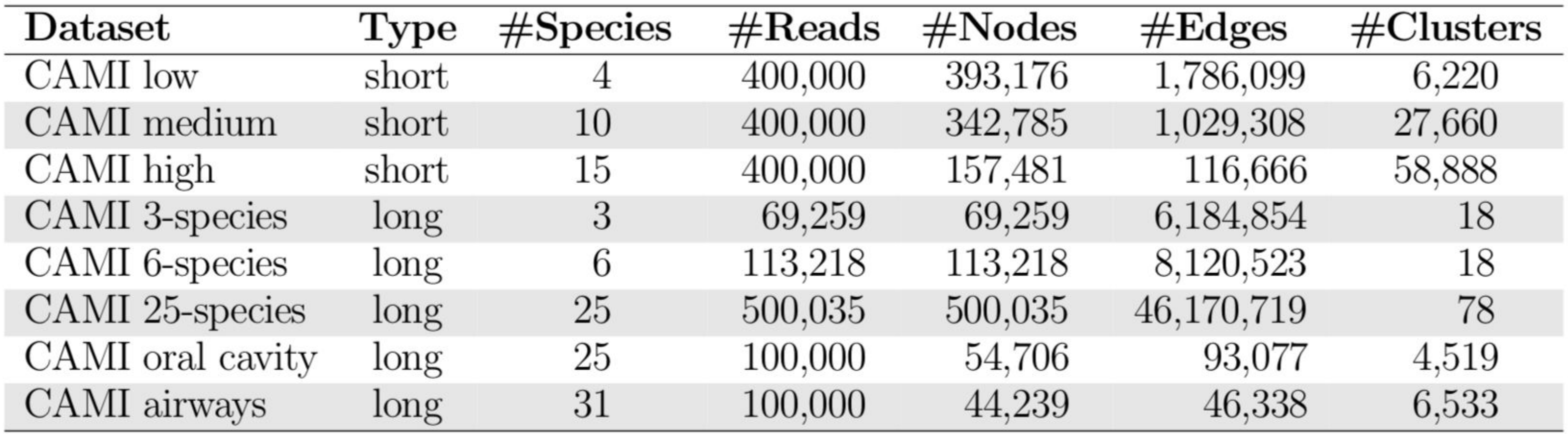
Overlap graph with ***ground-truth labels*** for training set nodes

MetaGNN pipeline



Semi-supervised node classification using a two-layer graph convolution network and ReLU

Metagenomic datasets



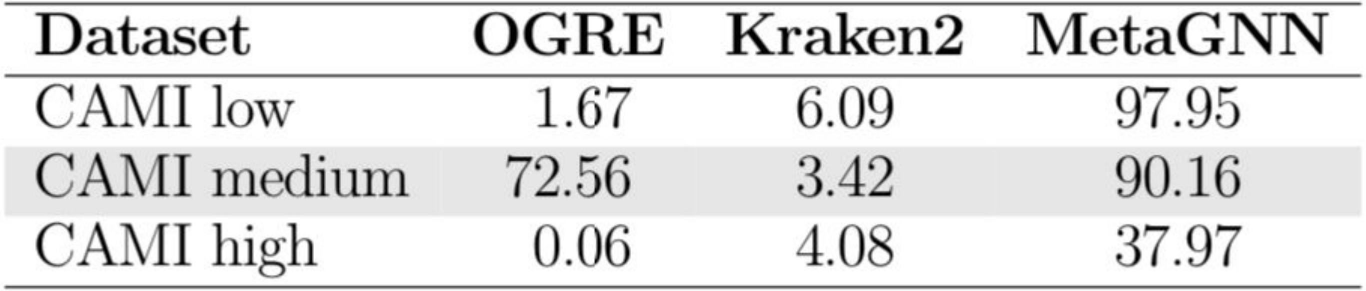
Dataset	Type	#Species	#Reads	#Nodes	#Edges	#Clusters
CAMI low	short	4	400,000	393,176	1,786,099	6,220
CAMI medium	short	10	400,000	342,785	1,029,308	27,660
CAMI high	short	15	400,000	157,481	116,666	58,888
CAMI 3-species	long	3	69,259	69,259	6,184,854	18
CAMI 6-species	long	6	113,218	113,218	8,120,523	18
CAMI 25-species	long	25	500,035	500,035	46,170,719	78
CAMI oral cavity	long	25	100,000	54,706	93,077	4,519
CAMI airways	long	31	100,000	44,239	46,338	6,533

Short and long read datasets

CAMI datasets are sampled based on the species

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software **Nature Methods 2017**

CAMI low (short reads)

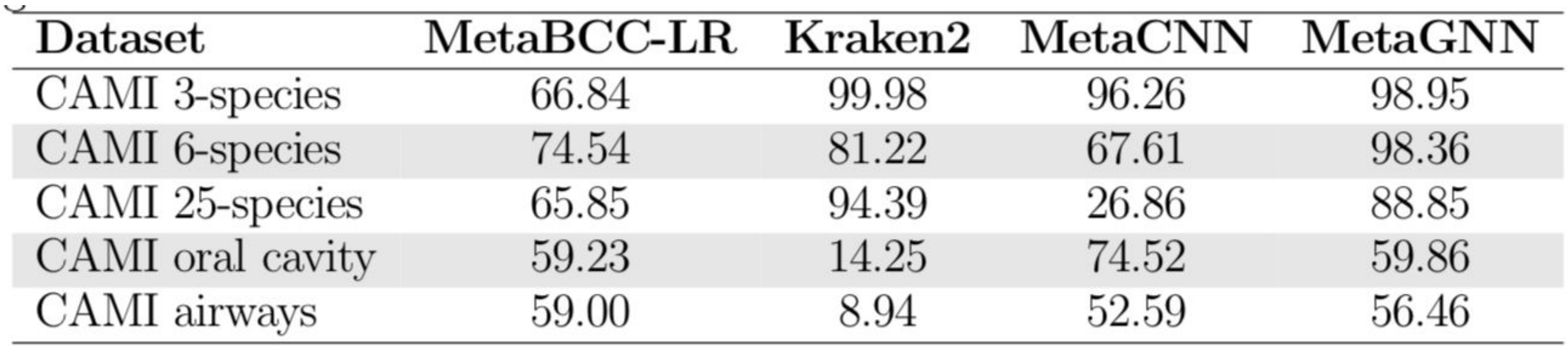


Dataset	OGRE	Kraken2	MetaGNN
CAMI low	1.67	6.09	97.95
CAMI medium	72.56	3.42	90.16
CAMI high	0.06	4.08	37.97

F1 score

MetaGNN ***improves*** the accuracy by an ***order-of-magnitude*** for short-read data compared to state-of-the-art binning/classification tools

CAMI airways (long reads)



Dataset	MetaBCC-LR	Kraken2	MetaCNN	MetaGNN
CAMI 3-species	66.84	99.98	96.26	98.95
CAMI 6-species	74.54	81.22	67.61	98.36
CAMI 25-species	65.85	94.39	26.86	88.85
CAMI oral cavity	59.23	14.25	74.52	59.86
CAMI airways	59.00	8.94	52.59	56.46

F1 score

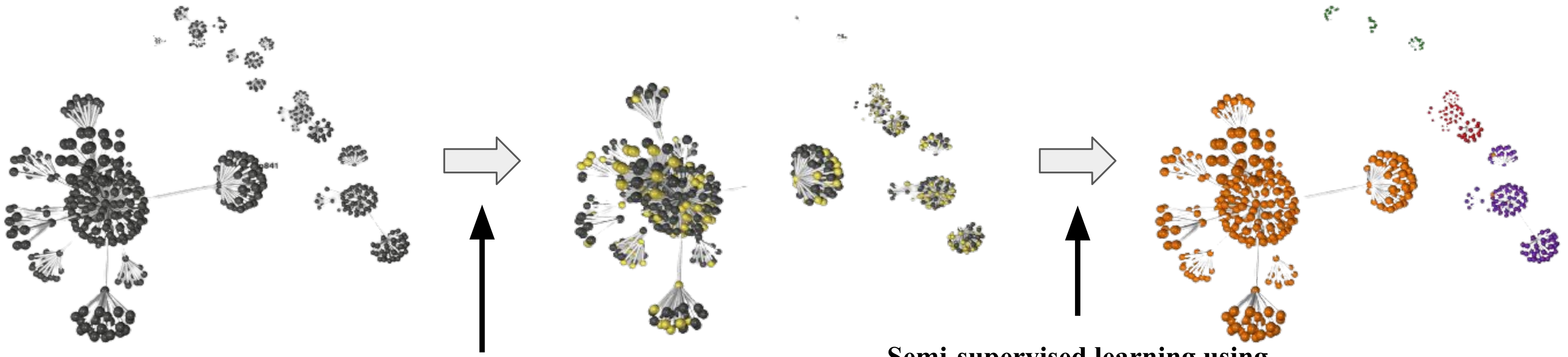
MetaGNN offers ***similar accuracy*** compared to the best binning/classification tools for long-read data

Conclusion

- We used GNN in a semi-supervised setting where ground truth is known for a small portion of reads
- MetaGNN shows that GNN can serve as a powerful classifier for metagenomic data
- **Future directions**
  - Perform unsupervised clustering of metagenomic data using GNN
  - Model a classifier for novel species found in real metagenomic data
  - Scale MetaGNN to larger metagenomic datasets

https://prashantpandey.github.io

MetaGNN pipeline



Overlap graph with no labels

Overlap graph with training labels

Overlap graph with learned labels

- **Assign ground truth labels to training nodes**
- **Assign tetra-nucleotide frequency as node vectors**

Semi-supervised learning using Graph Convolutional Network (GCN)

Alternate Camera space