Hypothesis Generation for Antibiotic Resistance using Machine Learning Techniques

Nicholas Joodi, Minseung Kim, Ilias Tagkopoulos

Tagkopoulos Lab

UC DAVIS genome center

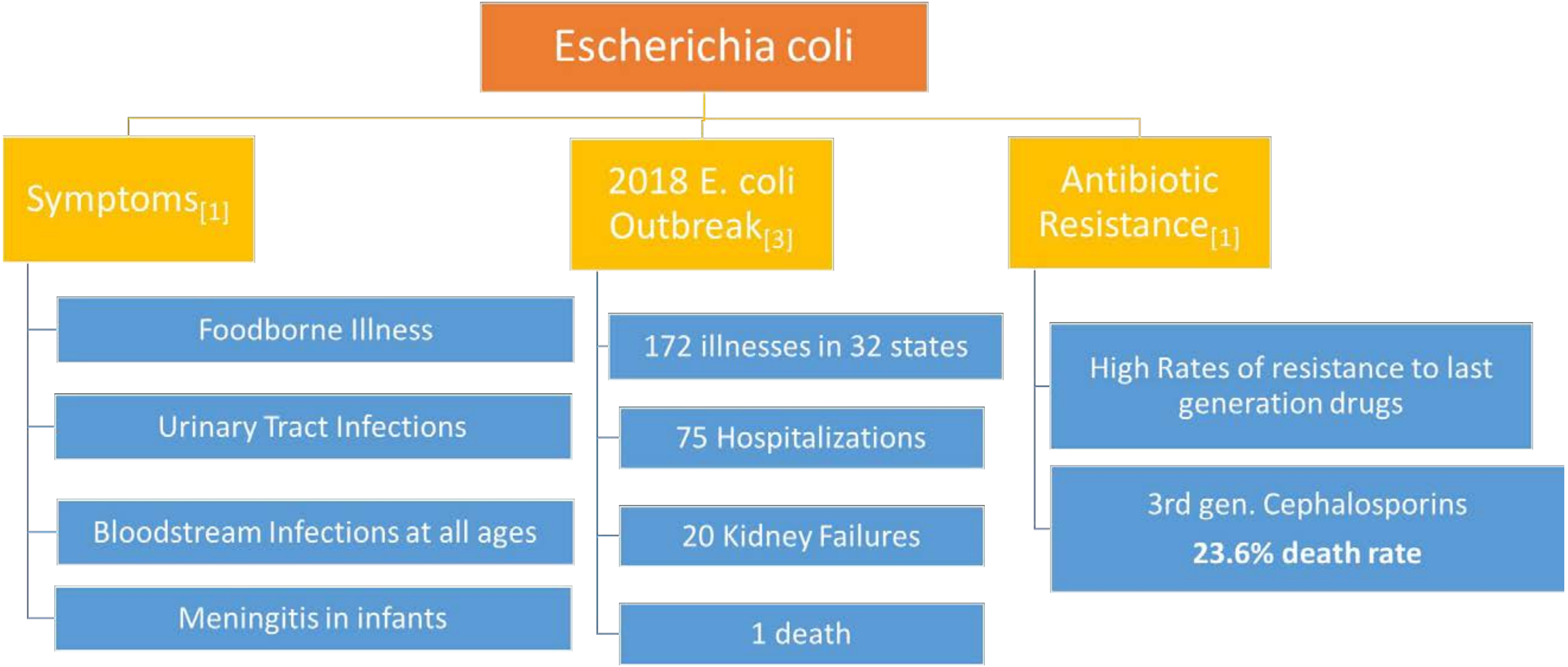
UC DAVIS UNIVERSITY OF CALIFORNIA

UC DAVIS COMPUTER SCIENCE

Antibiotic Resistance

- Medicines for treating infection lose effect because of Microbe change:
  - Mutation
  - Acquire new genetic information to develop resistance
- WHO: Antibiotic Resistance has reached alarming levels[1]
  - Study in the United States (CDC 2013)[2]
    - 2 million people infected by bacteria resistant to antibiotics
    - 23,000 deaths
  - Overall Societal costs[2]
    - Up to $20 billion direct
    - Up to $35 billion indirect

Escherichia coli



Related Work

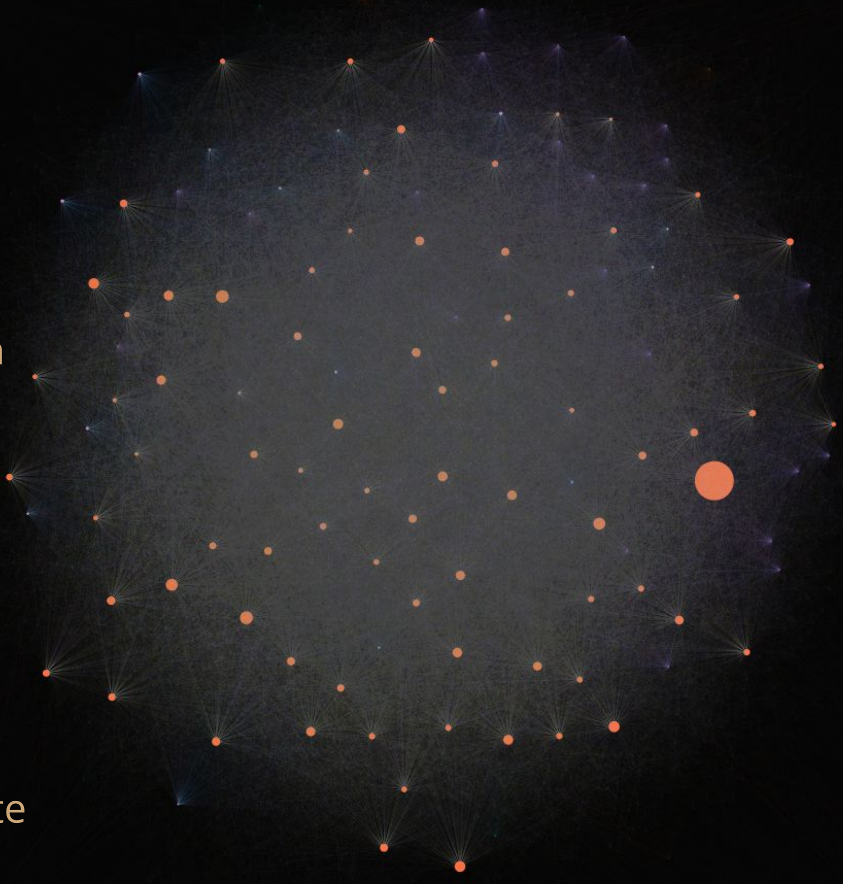
Predict the Antibiotic Resistant Genes (ARG)

- Existing Bioinformatics tools[4]
  - leverage known ARG sequences from within genomic or metagenomic sequence libraries
  - Commonly used approach: “Best Hit”
- DeepArg[5]
  - A machine learning approach over sequencing data
  - Improvements to the “Best Hit” approach
- Limited to sequence data

Approach

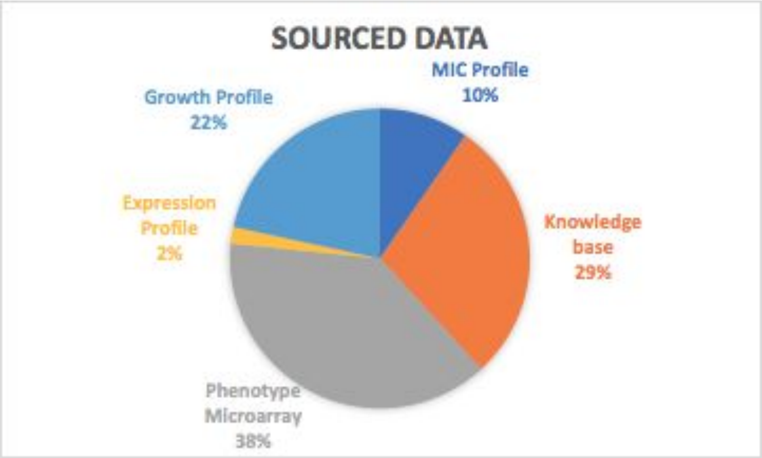
Graph Inference

- Leverage the relational data existing in an integrated/discrepancy resolved *E. coli* knowledge base to predict antibiotic resistance
- Knowledge graph:
  - Composed of entities (nodes) and relations between entities (edges)
- Inspired by Google Knowledge Vault[6]
  - Combine the powers of two disparate approaches to predict new facts
- **Predict whether a gene confers resistance to an antibiotic**



Knowledge Graph

- Pulled from 9 different sources
  - 5 groups



Entity Type	Node Count
gene	4769
antibiotic	109
cellular component	152
biological process	1522
Molecular Function	1782

Knowledge Graph

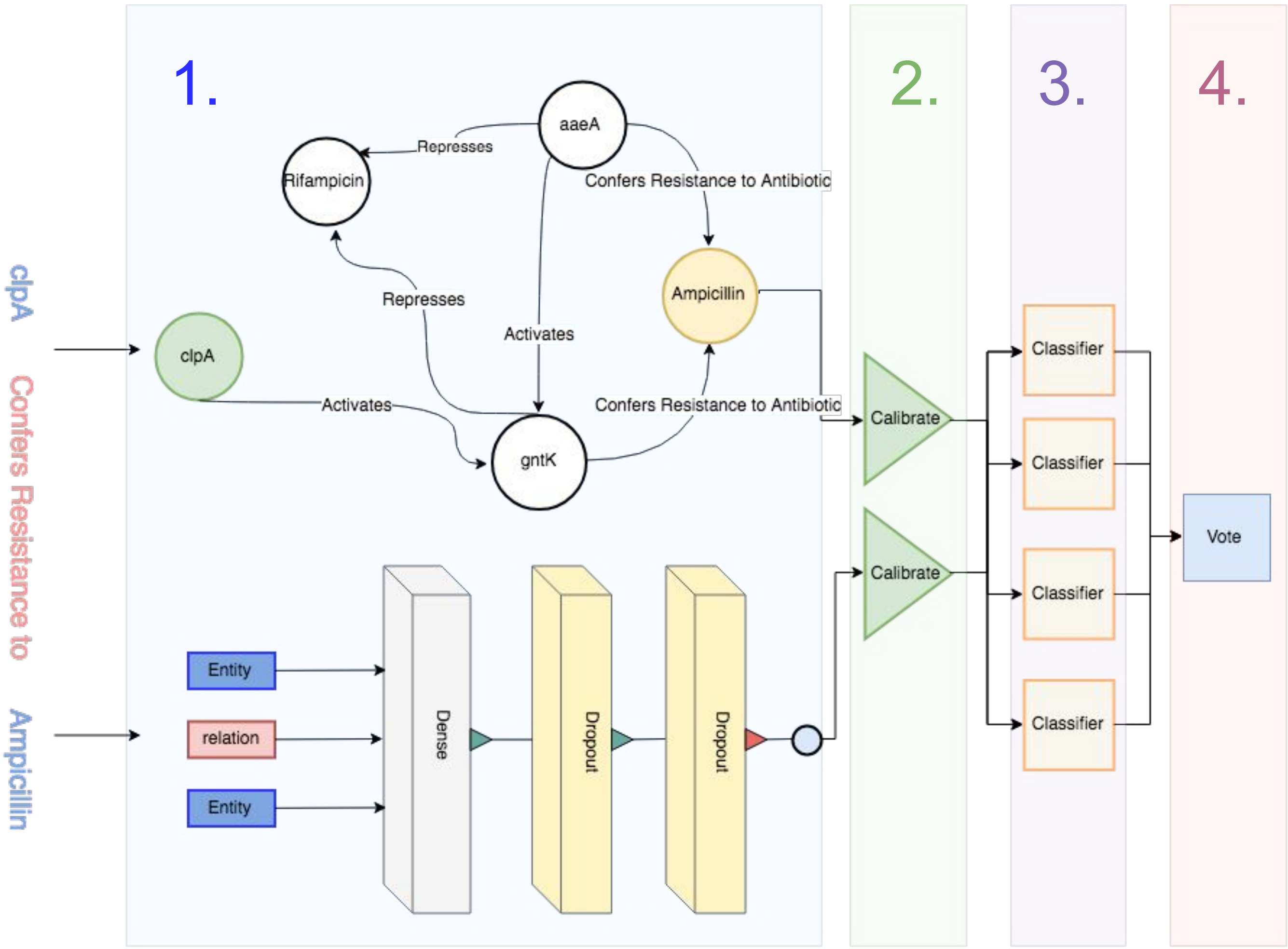
12 relation types

- 4 negatives

Domain	Relation Type	Range	Edge Count
Gene	activates	gene	2549
Gene	is	Cellular component	4325
Gene	represses	gene	2473
Gene	Is involved in	Biological process	6508
Gene	Upregulated by antibiotic	antibiotic	159
Gene	Confers resistance to antibiotic	antibiotic	902
Gene	has	Molecular function	7835
Gene	Targeted by	antibiotic	31
Gene	Not upregulated by antibiotic	antibiotic	338124
Gene	Not confers resistance to antibiotic	antibiotic	422899
Gene	Not activates	gene	48312
Gene	Not represses	gene	48544

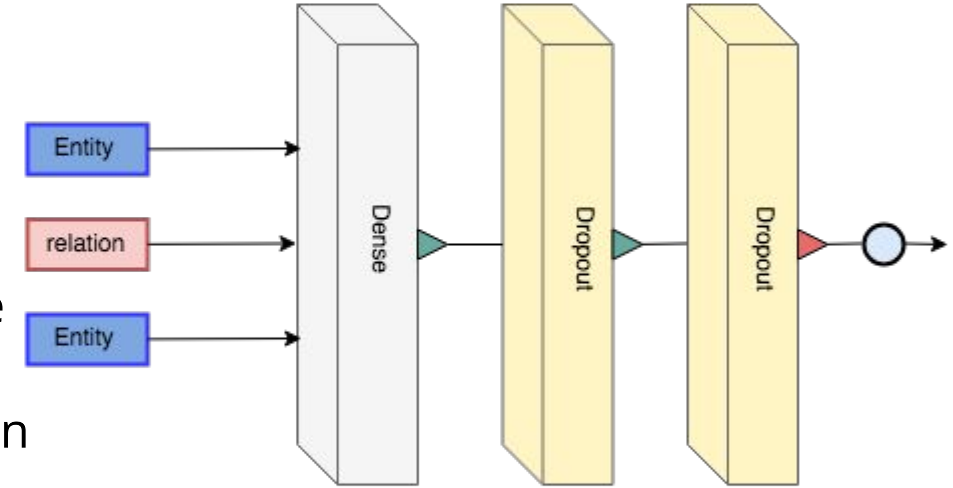
Architecture

1. Score edge using PRA and ER-MLP
2. Calibrate Scores
3. Majority vote using Boosted Decision Stumps
4. Boolean Prediction



Entity Relation Multilayered Perceptron

- Latent Feature Model
- Fully connected feedforward artificial neural network
- 150 inputs, matching the size of the concatenation of the two entity and relation embeddings
- 3 dense layers:
    1. With ReLU activation
    2. Dropout with ReLU activation
    3. Dropout with Sigmoid activation
- Single dense feature to produce the confidence score
- Trained on the 8 positive relation types

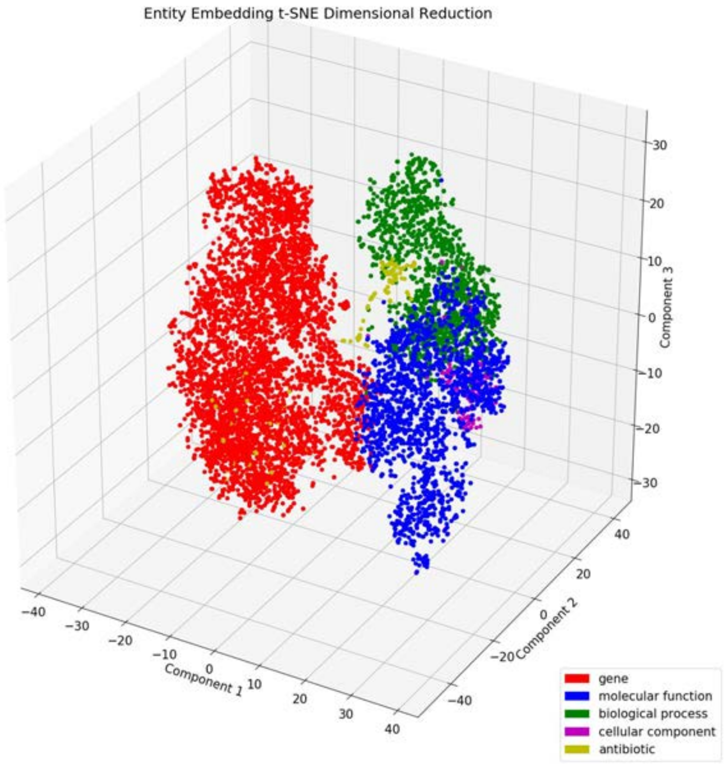


ER-MLP Training

- Trained using margin based ranking loss:

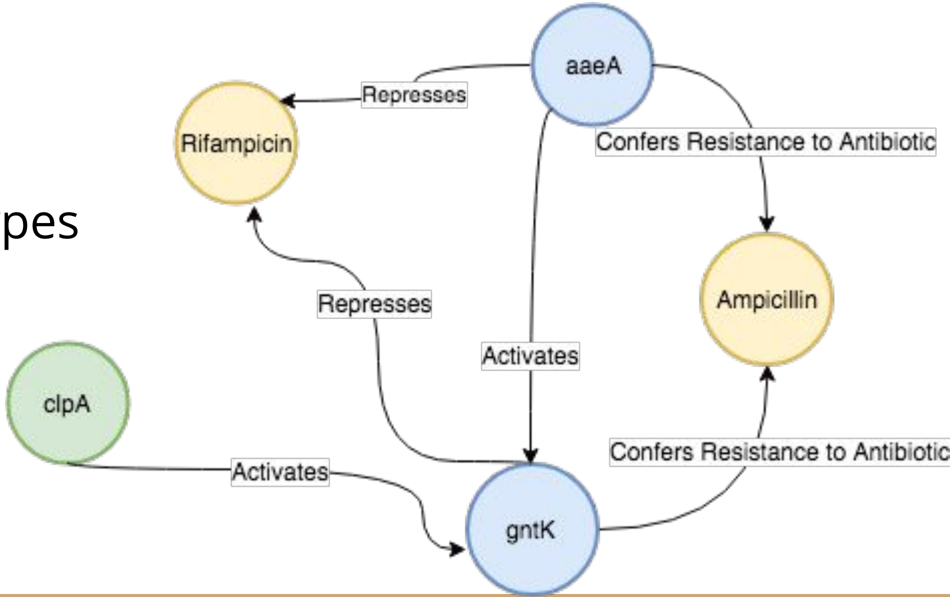
$$J(\omega) = \sum_{i=1}^{N}\sum_{c=1}^{C} \max\left(0,1 - g\left(T^{i}\right) + g\left(T_c^{i}\right)\right) + \lambda\|\omega\|_2^2$$

- The entities and relations are created by averaging the constituent word embeddings
- The word embeddings are initialized randomly
- Treated as learnable parameters by the model
- A noticeable semantic clustering of the types of entities is established after training



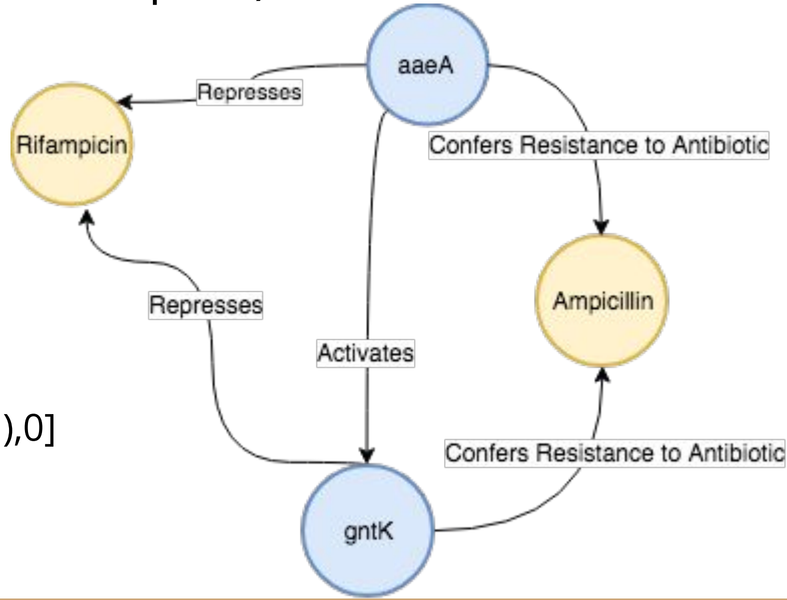
Path Ranking Algorithm

- Observable graph feature model
- A path is a sequence of relations linking two entities
- Classify the existence of an edge based on the paths between the subject and object entities
    - Paths are the features
- A model for every relation
- Trained on the 8 positive relation types



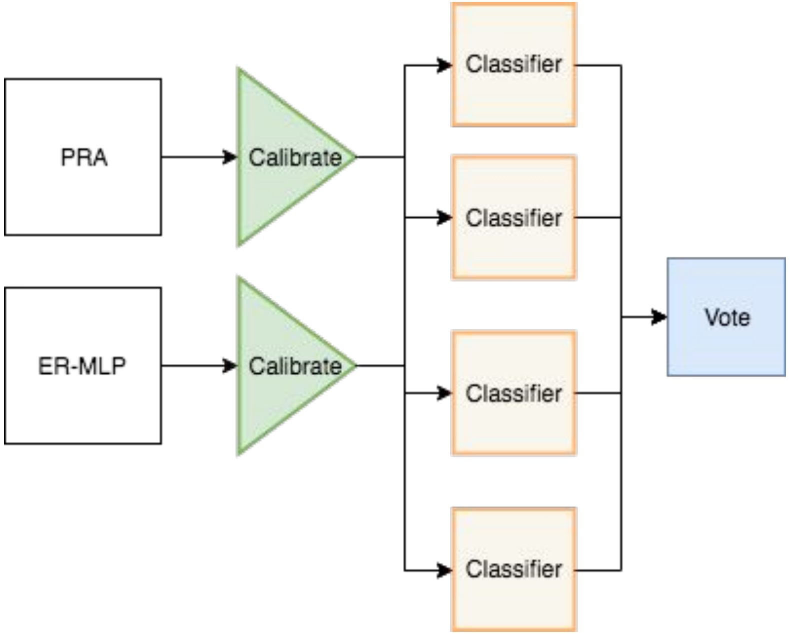
PRA - Training

- Relation: **Confers Resistance to Antibiotic**
- Positive Samples: (aaeA,Ampicillin), (gntK,Ampicillin)
- Negative Samples: (aaeA,Rifampicin), (gntK,Rifampicin)
- Features:
  - Activates → Confers Resistance to Antibiotic
  - Activates^{-1} → Confers Resistance to Antibiotic
  - Represses
  - Activates → Represses
  - Activates^{-1} → Represses
- Training Set:
  - [(1,0,0,0,0), 1], [0,1,0,0,0),1], [0,0,1,1,0),0], [0,0,1,0,1),0]
- Standard loss function used for training
  - Log Loss, Hinge Loss, Exponential Loss



Stacking

- Combining latent and observable graph feature models have shown to be superior in prediction
- Probability Calibration
  - Isotonic Regression
- Calibrate outputs of PRA and ER-MLP
- Train an ensemble of weak learners
  - Decision stumps with Adaboost

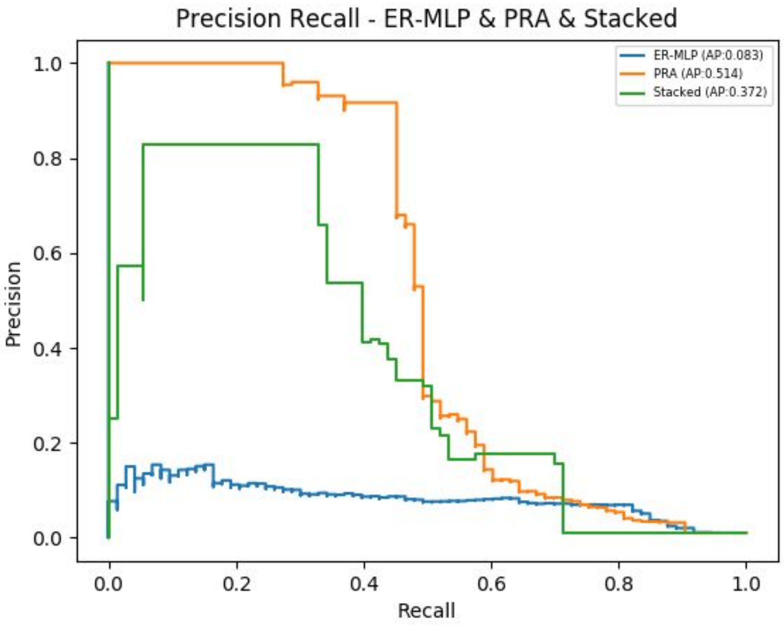


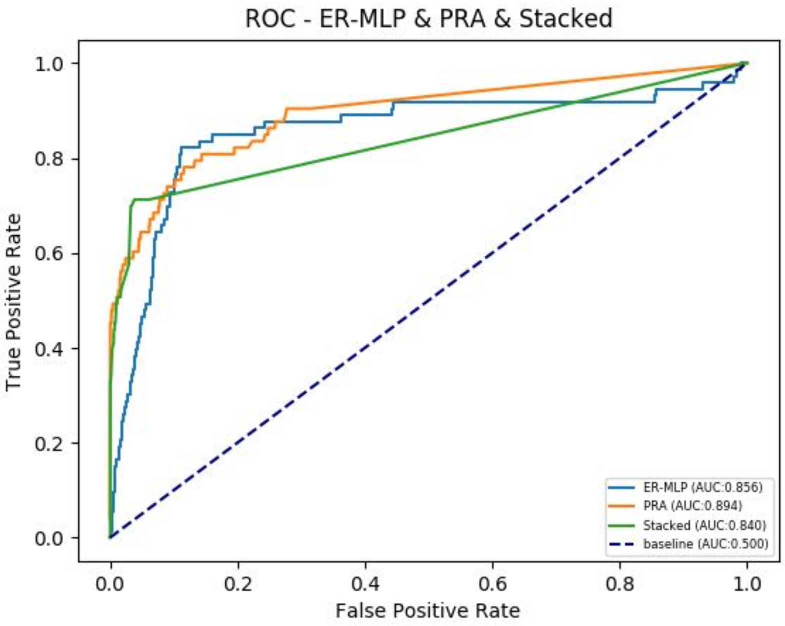
Method of Evaluation

- Test set includes 73 unique antibiotics
  - 100 samples of each
    - 1 positive edge of **confers resistance to antibiotic**
    - 99 negative edges of **confers resistance to antibiotic**
- 7300 samples total
- The goal is to predict the correct positive edge out of the 100 candidates

Results - ROC & PR

- All Models performed well in terms or Receiver Operating Characteristic
- PRA is superior in terms of Average Precision (Approximate baseline: 1%)





Results - Confusion Matrix

Preliminary results show that the PRA performed optimally while the Stacked had the highest recall

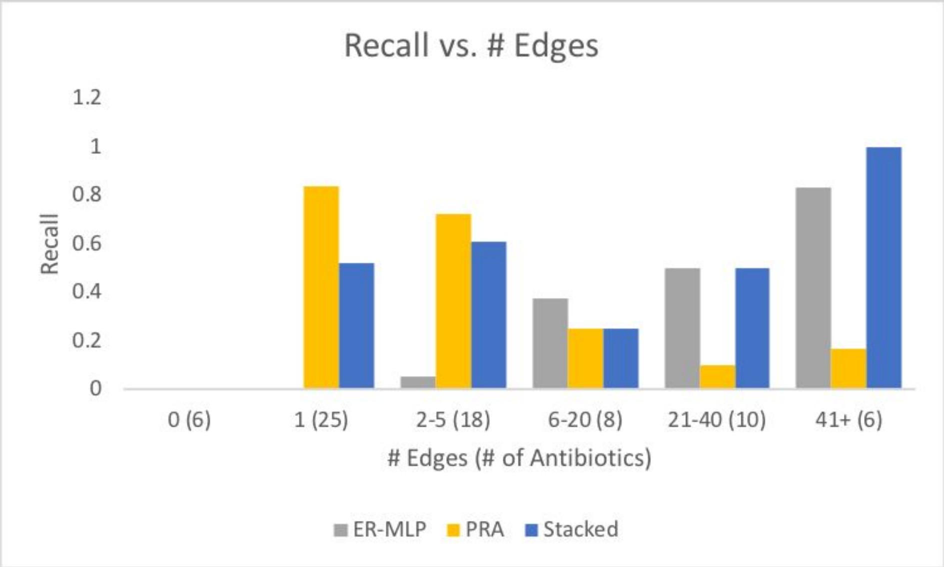
PRA		Known				
		Resistance	No Resistance	Total		
Prediction	Resistance	38	107	145	26.20%	Precision
	No Resistance	35	7120	7155	99.50%	NPV
	Total	73	7227			
		52.10%	98.50%	2.02%	98.10%	34.90%
		Sensitivity	Specificity	FDR	Accuracy	F1

ER-MLP		Known				
		Resistance	No Resistance	Total		
Prediction	Resistance	14	104	118	11.90%	Precision
	No Resistance	59	7123	7182	99.20%	NPV
	Total	73	7227			
		19.20%	98.60%	1.64%	97.80%	14.50%
		Sensitivity	Specificity	FDR	Accuracy	F1

Stacked		Known				
		Resistance	No Resistance	Total		
Prediction	Resistance	37	108	145	25.50%	Precision
	No Resistance	36	7119	7155	99.50%	NPV
	Total	73	7227			
		50.70%	98.50%	2.02%	98.00%	33.90%
		Sensitivity	Specificity	FDR	Accuracy	F1

Analysis

- At least one edge in the knowledge graph is necessary to predict for a particular antibiotic
- PRA performs very well when limited number of edges exist for the particular antibiotic
- ER-MLP performs very well when there are significantly more edges that exist for the particular antibiotic
- The stacked ensemble works well in both categories



Future Work

- Currently training ensemble on scores produced from **confers resistance to antibiotic** relation only
    - Training on the scores produced from the other edges could provide for more training data
    - Would reduce size of knowledge graph to include more edges in validation set
    - Would require the use of the local closed world assumption
- Incorporate the use of the negative relations during training of ER-MLP/PRA
- Experimentally validate in our wet lab

Thank you

- Blue Waters
- Lab Members
- Others

References

1. Organization, W.H., *Antimicrobial resistance: global report on surveillance*. 2014: World Health Organization.
2. Centres for Disease Control and Prevention (US). *Antibiotic resistance threats in the United States, 2013*. Centres for Disease Control and Prevention, US Department of Health and Human Services, 2013.
3. Achenbach, Joel. "CDC comes close to an all-clear on romaine lettuce as E. coli outbreak nears historic level." *The Washington Post.* The Washington Post Company, 16 May 2018. Web. 28 May 2018.
4. McArthur, Andrew G., and Kara K. Tsang. "Antimicrobial resistance surveillance in the genomic age." *Annals of the New York Academy of Sciences* 1388.1 (2017): 78-91.
5. Arango-Argoty, Gustavo, et al. "DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data." *Microbiome* 6.1 (2018): 23.
6. Dong, X., et al. *Knowledge vault: A web-scale approach to probabilistic knowledge fusion*. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014. ACM.