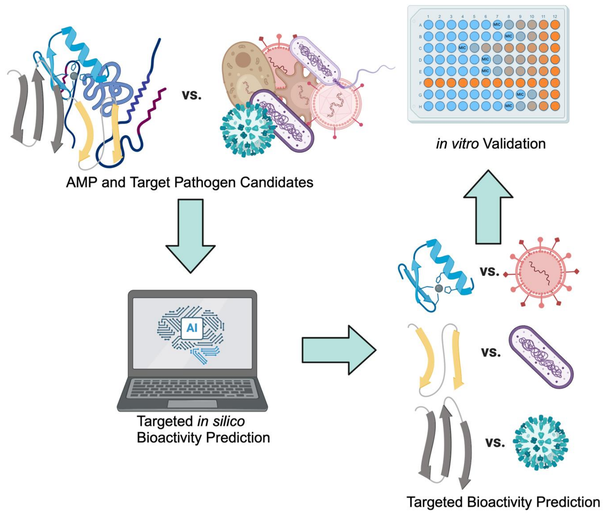
Pathogen-specific antimicrobial activity prediction with biological large language model-based methods

Berke Uçar[1,2], İnanç Birol[1,3,4], PeptAid Consortium[1,2,3,4,5]

[1]BC Cancer Research Institute, Vancouver, BC V5Z 1L3, Canada
[2]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
[3]Amphoraxe Life Sciences Inc., BC V6L 3C9, Canada
[4]Department of Medical Genetics, University of British Columbia, Vancouver, BC, V6H 3N1, Canada
[5]The School of Molecular Life Sciences, University of Victoria, Victoria, BC, V8P 5C2, Canada

BC CANCER RESEARCH
Provincial Health Services Authority

Introduction

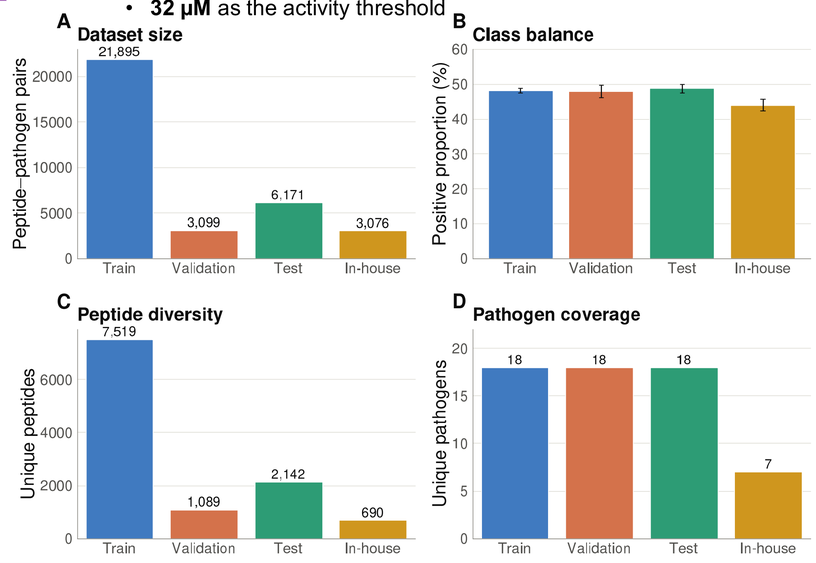


Background

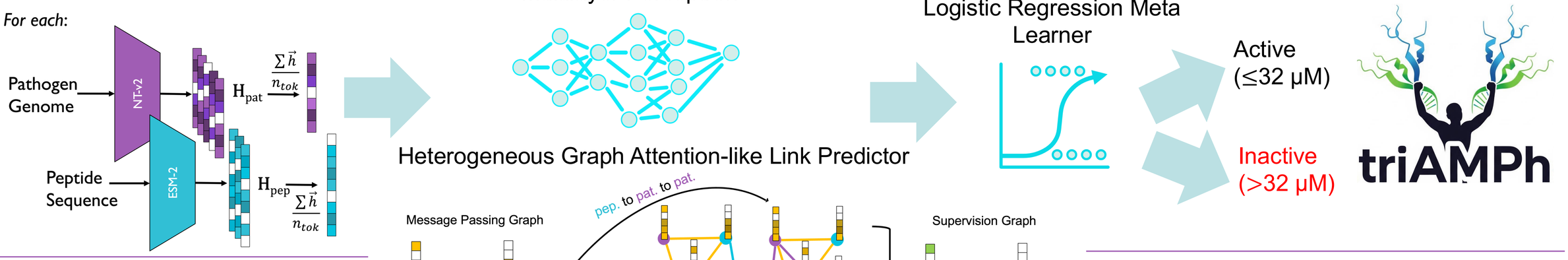
Approach	Caveat
Multiclass classification	Restricted pathogens, **no genomic content**
Functional group prediction	Coarse resolution, **no genomic content**
Oligonucleotide composition	Loses genome's **positional context**
Phylogenetic distance	**Proximity ≠ similar susceptibility**, point mutations change susceptibility

Data

- **Species-level** granularity
- Mined from **DBAASP**, **DRAMP**, **dbAMP** and generated by **PeptAid**
- **32 μM** as the activity threshold



Methods



Contact

bucar@bccrc.ca

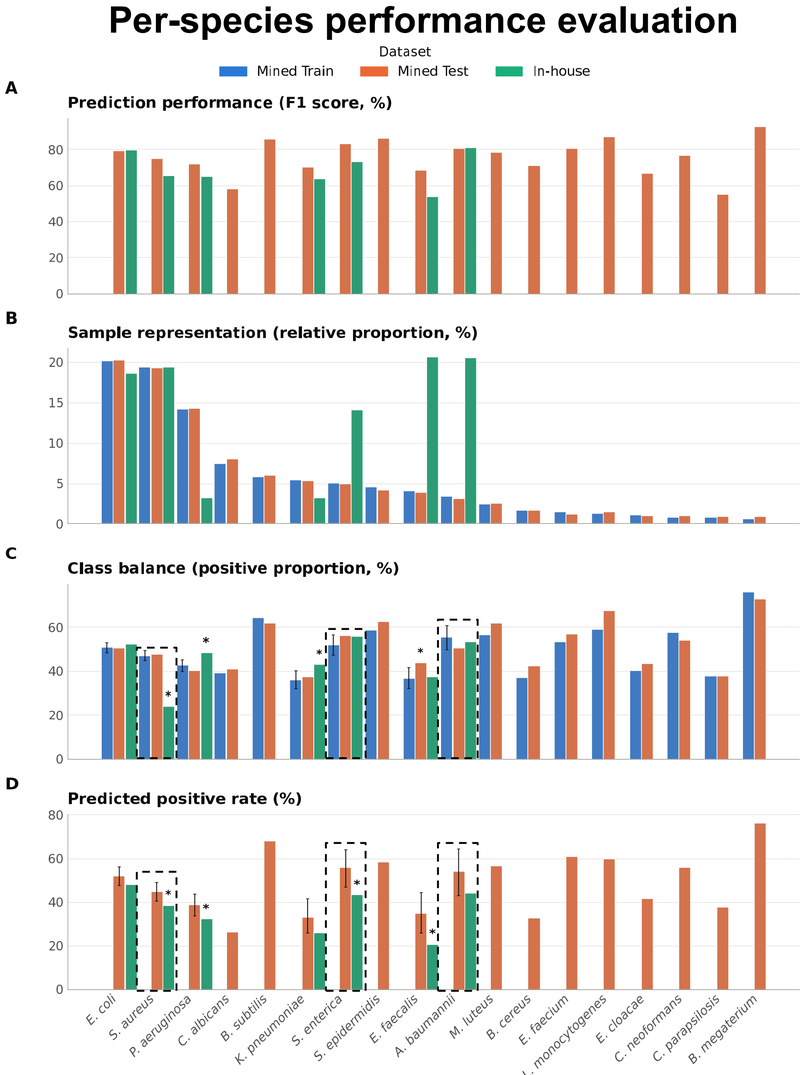
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Acknowledgements

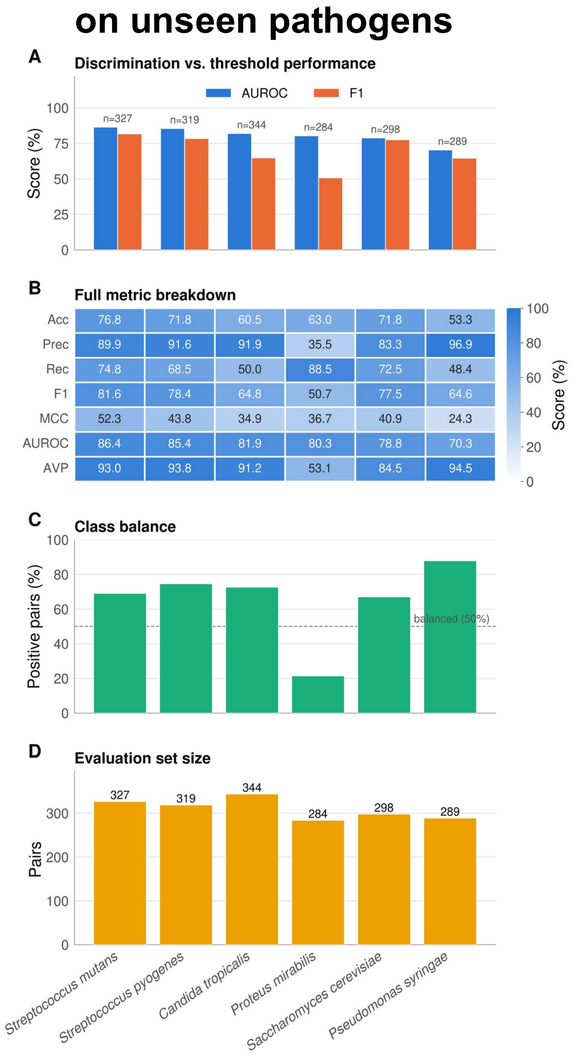
Genome BritishColumbia
Genome Canada
PeptAID
NSERC CRSNG

Results

Per-species performance evaluation



Zero-shot testing on unseen pathogens



Benchmarking against state-of-the-art methods

Mined Dataset

Model	Dataset Details	Accuracy	Precision	Recall	Micro F1-Score	Macro F1-Score	MCC	AUROC	Average Precision
AMPSpecies Specific	3,910 pairs, 5 species	61.10	56.56	**87.75**	68.79	70.98	27.30	71.86	71.32
triAMPh		**78.67**	**78.29**	77.96	**78.12**	**79.62**	**57.31**	**87.13**	**87.10**
Vishnepolsky et al.	1,065 pairs, 8 strains	83.47	82.68	84.25	83.46	84.32	66.96	-	-
triAMPh		81.41	79.85	83.49	81.63	80.46	62.89	89.58	89.92

In-house Dataset

Model	Dataset Details	Accuracy	Precision	Recall	Micro F1-Score	Macro F1-Score	MCC	AUROC	Average Precision
AMPSpecies Specific	1,272 pairs, 3 species	52.99	44.88	**94.51**	60.86	63.40	26.69	80.92	74.35
triAMPh		**78.54**	**70.32**	77.03	**73.52**	70.15	**55.70**	**85.57**	80.24
Vishnepolsky et al.	881 pairs, 4 strains	72.87	65.35	52.37	58.14	53.91	38.95	-	-
triAMPh		**83.09**	**80.00**	**70.66**	**75.04**	**76.64**	**62.60**	**88.53**	**84.57**

- *Datasets were subsetted based on the constraints of the tools being compared.*
- All results are presented as percentages.
- The significantly ($p<0.001$) leading metric for each comparison is highlighted in bold.