SeqScreen Software Overview

Todd Treangen, Rice University

IARPA Fun GCAT Program

Office of the Director of National Intelligence
Intelligence Advanced Research Projects Activity
IARPA
Creating Advantage through Research and Technology

Who We Are | Research | Engage With Us | Newsroom | COVID-19 Research

Home ▪ Research ▪ Office of Collection ▪ Fun GCAT

FUN GCAT

FUNCTIONAL GENOMIC AND COMPUTATIONAL ASSESSMENT OF THREATS

FunGCAT

INTELLIGENCE VALUE

The Fun GCAT is developing methods to rapidly assess the function of DNA sequences to determine if they pose a threat. Used to automatically process large datasets or to supplement subject matter expert review, Fun GCAT technology will enable improved detection of bio-error or bio-terror.

SUMMARY

Current screening methods to flag dangerous DNA sequences are inadequate—they do not consider DNA function, cannot process short or highly engineered sequences, and often require follow-up analysis by an expert. Fun GCAT is developing smart, AI-driven threat screening software to replace current look-up table-based screening. Fun GCAT researchers developed computational pipelines to analyze DNA and answer three questions per sequence: What organism does it come from? What biological functions does it have? How dangerous is it? By using neural networks and other powerful bioinformatic techniques to learn the common patterns of sequences with similar origins and functions, Fun GCAT tools are demonstrating high predictive accuracy against increasingly challenging test sets. Benchmarking has demonstrated significant performance increases beyond top winners in a closely related bioinformatic software development global challenge. It resulted in 500x improvement in computational efficiency over state-of-the-art and stable performance on even short (<50 base pairs) sequences. This enabled a range of Intelligence Community-relevant missions from rapid screening of very large datasets to field-based, targeted analysis.

Fun GCAT is also developing new approaches to meet the demand for rapid, relatively high-throughput experimental assessment of DNA function. The program has developed optical, cell-based tests to identify disruptors of the immune system hardwired within cells. New functions for dozens of virus genes have been discovered and the technologies are being leveraged to develop threat prediction capabilities for faster

CONTACT INFORMATION

PROGRAM MANAGER
Main Office
dni-iarpa-info@iarpa.gov
301-243-1995

RESEARCH AREA(S)

Machine learning, Synthetic biology, Threat determination, Taxonomic classification, Biosecurity, Bioinformatics , DNA screening, Gene function, Viral, Toxins, Innate immunity, Microscopy, DNA synthesis

BROAD AGENCY ANNOUNCEMENT (BAA)

LINK(S) TO BAA
IARPA-BAA-16-08

SOLICITATION STATUS
CLOSED

https://www.iarpa.gov/research-programs/fun-gcat

Disclaimer for our Fun GCAT work:

All of the co-authors were either fully or partially supported by the Fun GCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government.

RICE Acid signature science LLC

Software Developed under IARPA Fun GCAT Program

SeqScreen Fast and Sensitive Modes:
Input = Short Sequences
Processing Goal = Linux Server

SeqScreen ONT Mode:
Input = Nanopore Sequences
Processing Goal = Laptop

SEQSCREEN
Sequence of Interest Screening

https://gitlab.com/treangenlab/seqscreen

Machine Learning
(ML) Database

Open-Source ML-Based
Functional Annotation Software

Government Purpose Rights
Threat Detection Software

RICE Acid signature science LLC

SeqScreen Publication and GitLab Repo

Balaji *et al. Genome Biology* (2022) 23:133
https://doi.org/10.1186/s13059-022-02695-x

Genome Biology

SOFTWARE Open Access

SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning

Advait Balaji[1†], Bryce Kille[1†], Anthony D. Kappell[2], Gene D. Godbold[3], Madeline Diep[4], R. A. Leo Elworth[1], Zhiqin Qian[1], Dreycey Albin[1], Daniel J. Nasko[5], Nidhi Shah[5], Mihai Pop[5], Santiago Segarra[6], Krista L. Ternus[2*] and Todd J. Treangen[1*]

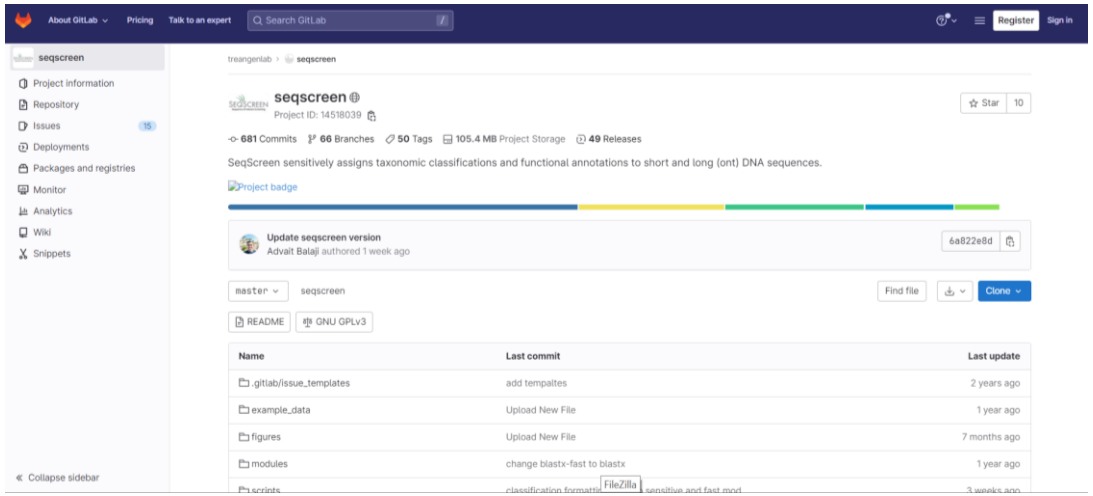
[†]Advait Balaji and Bryce Kille contributed equally.

* Correspondence: kternus@signaturescience.com; treangen@rice.edu

[2]Signature Science, LLC, 8329 North Mopac Expressway, Austin, TX, USA
[1]Department of Computer Science, Rice University, Houston, TX, USA
Full list of author information is available at the end of the article

Abstract

The COVID-19 pandemic has emphasized the importance of accurate detection of known and emerging pathogens. However, robust characterization of pathogenic sequences remains an open challenge. To address this need we developed SeqScreen, which accurately characterizes short nucleotide sequences using taxonomic and functional labels and a customized set of curated Functions of Sequences of Concern (FunSoCs) specific to microbial pathogenesis. We show our ensemble machine learning model can label protein-coding sequences with



https://gitlab.com/treangenlab/seqscreen

https://link.springer.com/article/10.1186/s13059-022-02695-x

Installation Tips

- SeqScreen is easiest to install via conda or mamba, or as a conda-pack or Docker/Singularity container if you are working on an air-gapped system
- There is a `--check_install` option to check that the required command line tools, python imports, and database files are present before you run the software
- Please see our wiki documentation for additional information: https://gitlab.com/treangenlab/seqscreen/-/wikis/Home

Home

Three Different SeqScreen Modes

- Fast mode (default)
  - Good choice for large datasets, uses DIAMOND instead of BLAST
  - Only one protein-coding region expected per read
- Sensitive mode
  - Ideal for deeply characterizing each read, uses BLASTX and optionally BLASTN
  - Only one protein-coding region expected per read
- ONT mode
  - Suitable when memory is limited, like in today's course
  - More than one protein-coding region is expected per read (e.g., nanopore sequences, contigs, vector sequences)
- All modes are run within Nextflow

SeqScreen-Nano Preprint (i.e., ONT Mode)

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBMIT | FAQ | BLOG | ALERTS / RSS | ABOUT | CHANNELS

Search

Advanced Search

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

Previous Next

New Results

Follow this preprint

Posted March 03, 2023.

SeqScreen-Nano: a computational platform for rapid, in-field characterization of previously unseen pathogens

Advait Balaji, Yunxi Liu, Michael G. Nute, Bingbing Hu, Anthony Kappell, Danielle S. LeSassier, Gene D. Godbold, Krista L. Ternus, Todd J. Treangen

doi: https://doi.org/10.1101/2023.02.10.528096

This article is a preprint and has not been certified by peer review [what does this mean?].

0 0 0 0 0 0 19

Download PDF

Print/Save Options

Revision Summary

Email

Share

Citation Tools

Tweet Like 0

COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

Subject Area

Bioinformatics

Abstract Full Text Info/History Metrics Preview PDF

Subject Areas

All Articles

Animal Behavior and Cognition

Biochemistry

Bioengineering

ABSTRACT

The COVID-19 pandemic forever underscored the need for biosurveillance platforms capable of rapid detection of previously unseen pathogens. Oxford Nanopore Technology (ONT) couples long-read sequencing with in-field capability, opening the door to real-time, in-field biosurveillance. Though a promising technology, streaming assignment of accurate functional and taxonomic labels with nanopore reads remains

Manuscript in prep

https://www.biorxiv.org/content/10.1101/2023.02.10.528096v2.abstract

RICE Acid signature science LLC

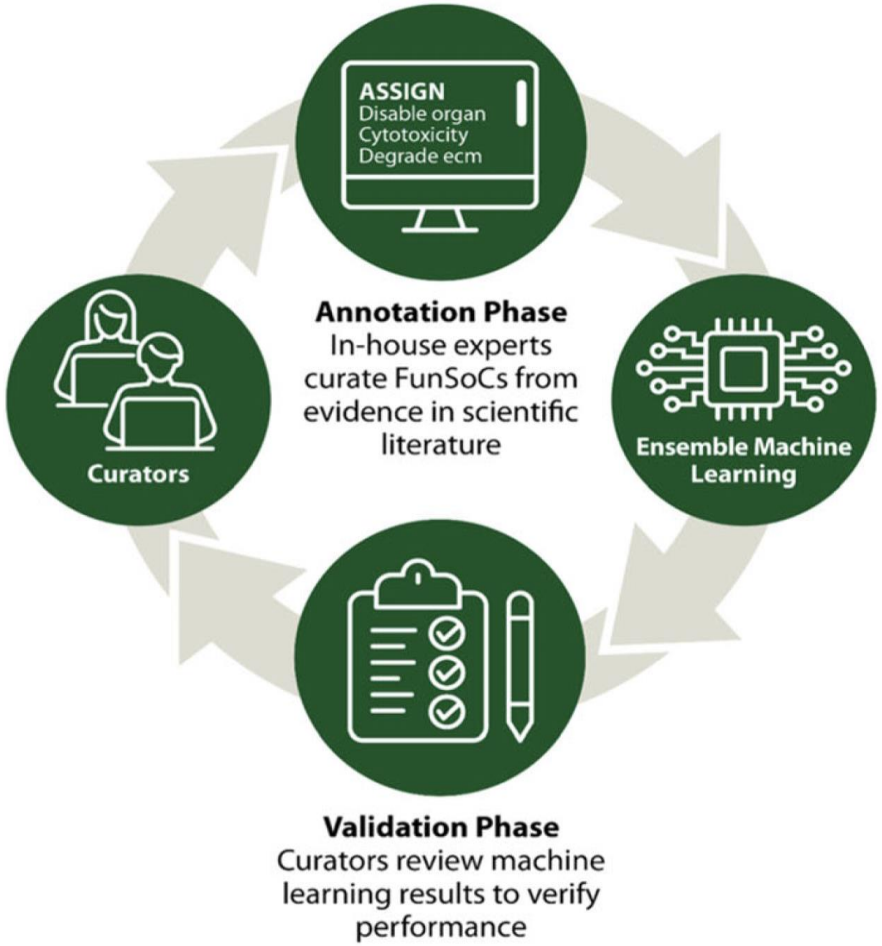
Many Other Execution Options

```
--fasta         Path to the input NT FASTA file
--databases     Path to the databases directory containing centrifuge, blast, go etc
--working       Path to the output working directory
--threads       Number of threads to use (Default=1)
--sensitive     Use SeqScreen sensitive mode (old default mode)
--ont           Enable SeqScreen to run ONT reads
--evalue        Cutoff to use for blastx/diamond
--hmmscan       Run hmmscan on input sequences
--bitscore      Tiebreak across all proteins within this % of the top bitscore
--ancestral     Include all ancestral GO terms in output
--splitby       Max number of sequences in an input chunk to diamond
--includecc     Include cellular component go terms
--blastn        Run blastn in addition to blastx (sensitive mode only)
--taxlimit      Maximum number of multi-taxIDs to output for a single query in fast mode
--slurm         Have pipeline modules run on SLURM execution nodes (Default = run locally)
--report_prefix Add prefix to beginning of seqscreen_report.tsv and seqscreen_html_report.zip. The prefix will
--skip_report   Skip report generation step and only generate intermediate files
--report_only   Remove intermediate output and only save the results in {output_dir}/report_generation
--format        Format type: [1] Original, [2] Hits only, [3] FunSoC only, [4] Gene-Centric,[5] Gene-Centric F
--online        Pull reference genomes from NCBI for reference_inference [Needs web access]
--filter_taxon  Filter comma separated list of taxon
--keep_taxon    Keep comma separated list of taxon
--taxonomy_confidence_threshold Confidence threshold for multi-taxids (Average) [Default 0.0]
--keep_html_ont Keep html report in ont mode [Takes additional memory]
--check_install Check for required command line tools, python imports, and database files
--version       Display the version and exit
--help      Print this help message out
```

You can see a full list of options with the `--help` command or in our software documentation

Human In-the-Loop Pathogen Prediction

Predict FunSoCs on targeted sets of sequences contained in database

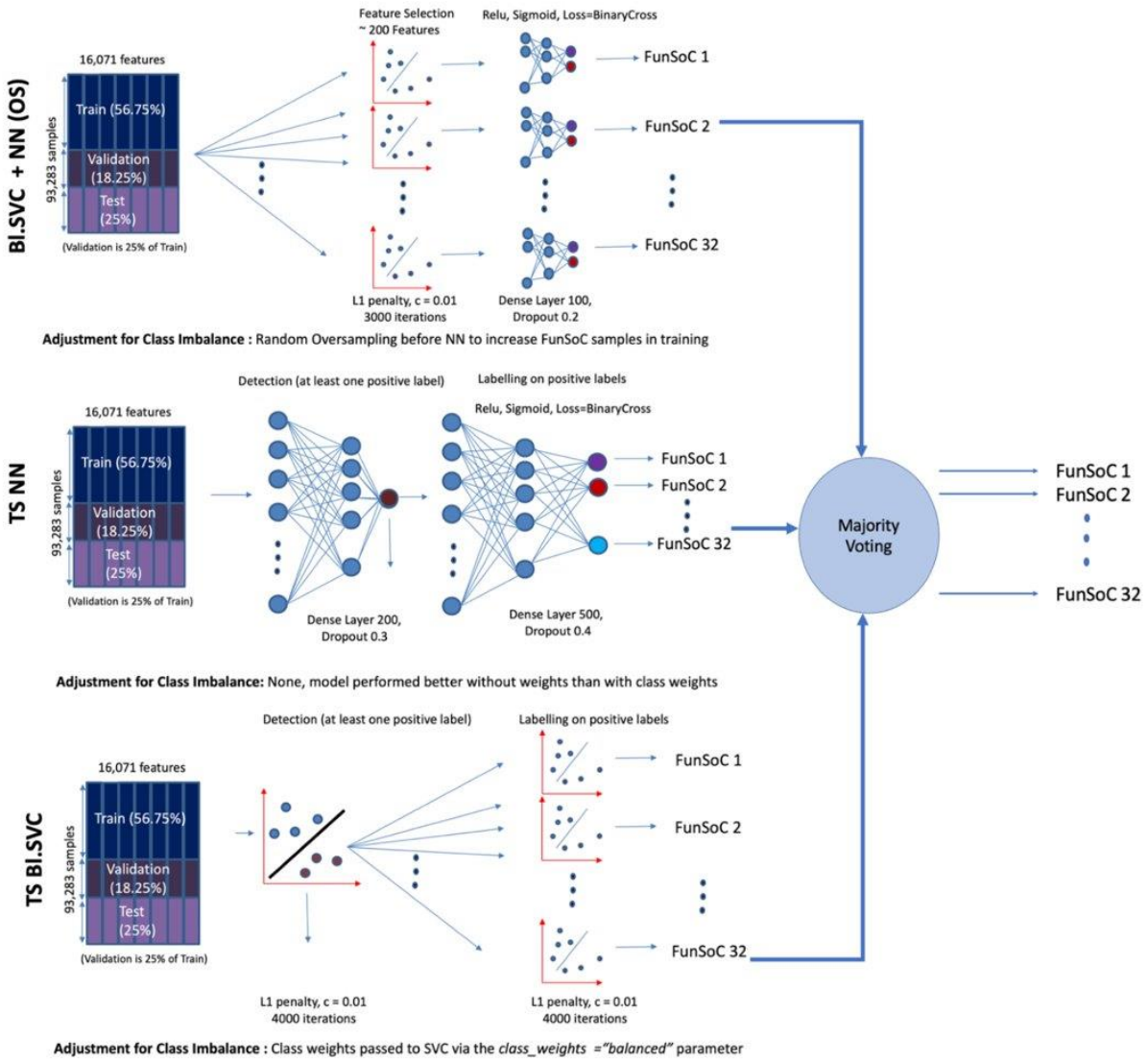


Have biocurators review, provide feedback on good predictions and bad ones

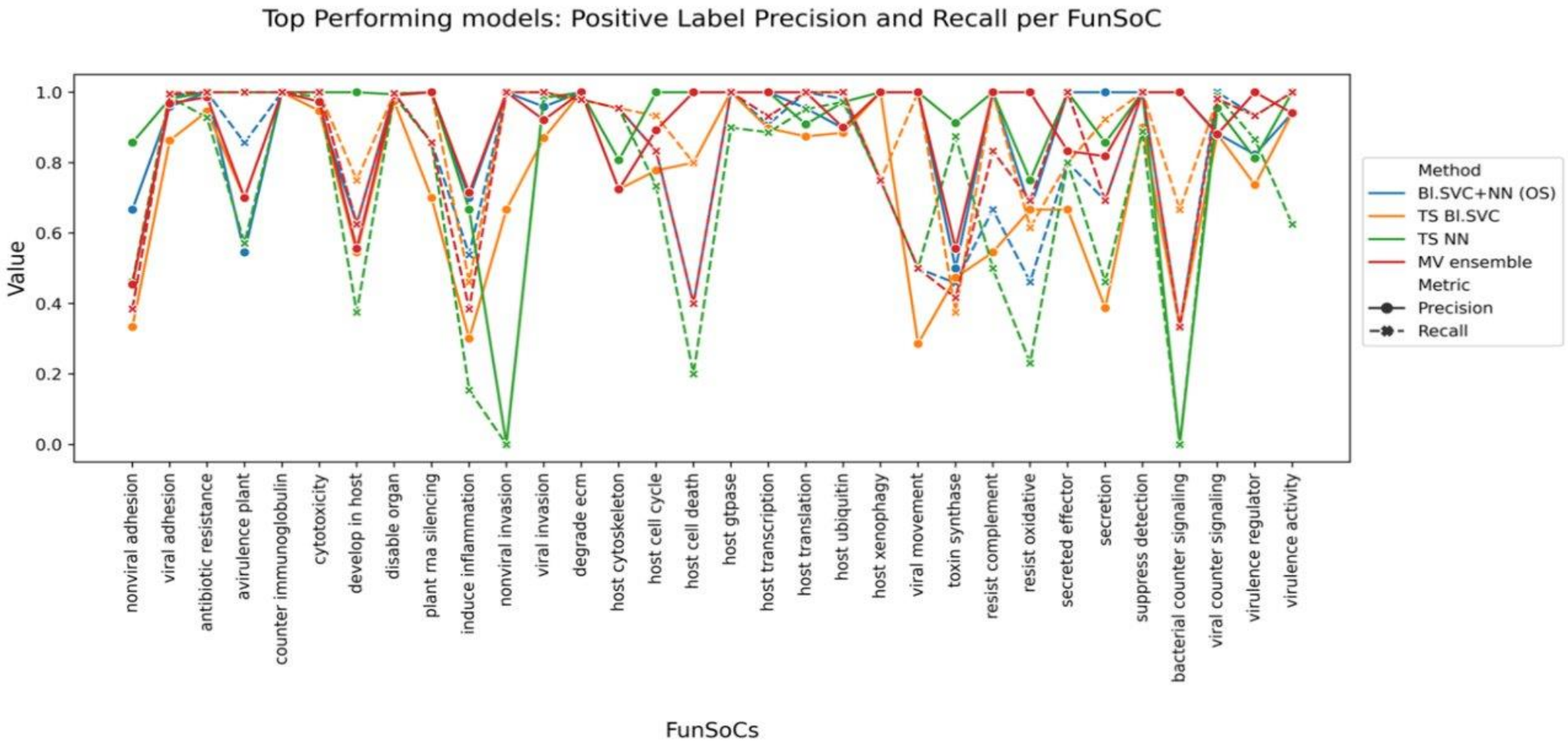
SEQSCREEN

Sequence of Interest Screening

Human In-the-Loop Pathogen Prediction



Peformance across FunSoCs



Advait Balaji
Rice University
PhD student

Bioinformatics, Graph Theory, Pathogen Detection and Analysis, Machine Learning, Microbial genomics and Metagenomics

Bryce Kille
Rice University
PhD student

String algorithms, Pangenomics, High performance computing, Cheminformatics

Kristen Curry
Rice University
PhD student

Computational Biology, Bioinformatics, Microbial genomics and Metagenomics

Michael Wang
Rice University
PhD student

Computational Biology, Bioinformatics, Microbial genomics and Metagenomics

Nicolae Sapoval
Rice University
PhD student

Computational Biology, Bioinformatics, Microbial genomics and Metagenomics, Graph algorithms and data structures, Deep learning

Yilei Fu
Rice University
PhD student

Yunxi Liu
Rice University
PhD student