Addressing bioinformatics challenges in pathogen agnostic detection from metagenomic sequencing

November 19, 2024

Jonathan Allen, Ph.D.
Informatics Scientist

LLNL-PRES-XXXXX

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Pathogen agnostic detection for human health

Input sample type

- Clinical samples
  - Blood, nasal swabs
- Air filters
  - Dirty (subway, outdoor)
  - Clean(er) (building/space station)
- Wastewater
- Clean rooms

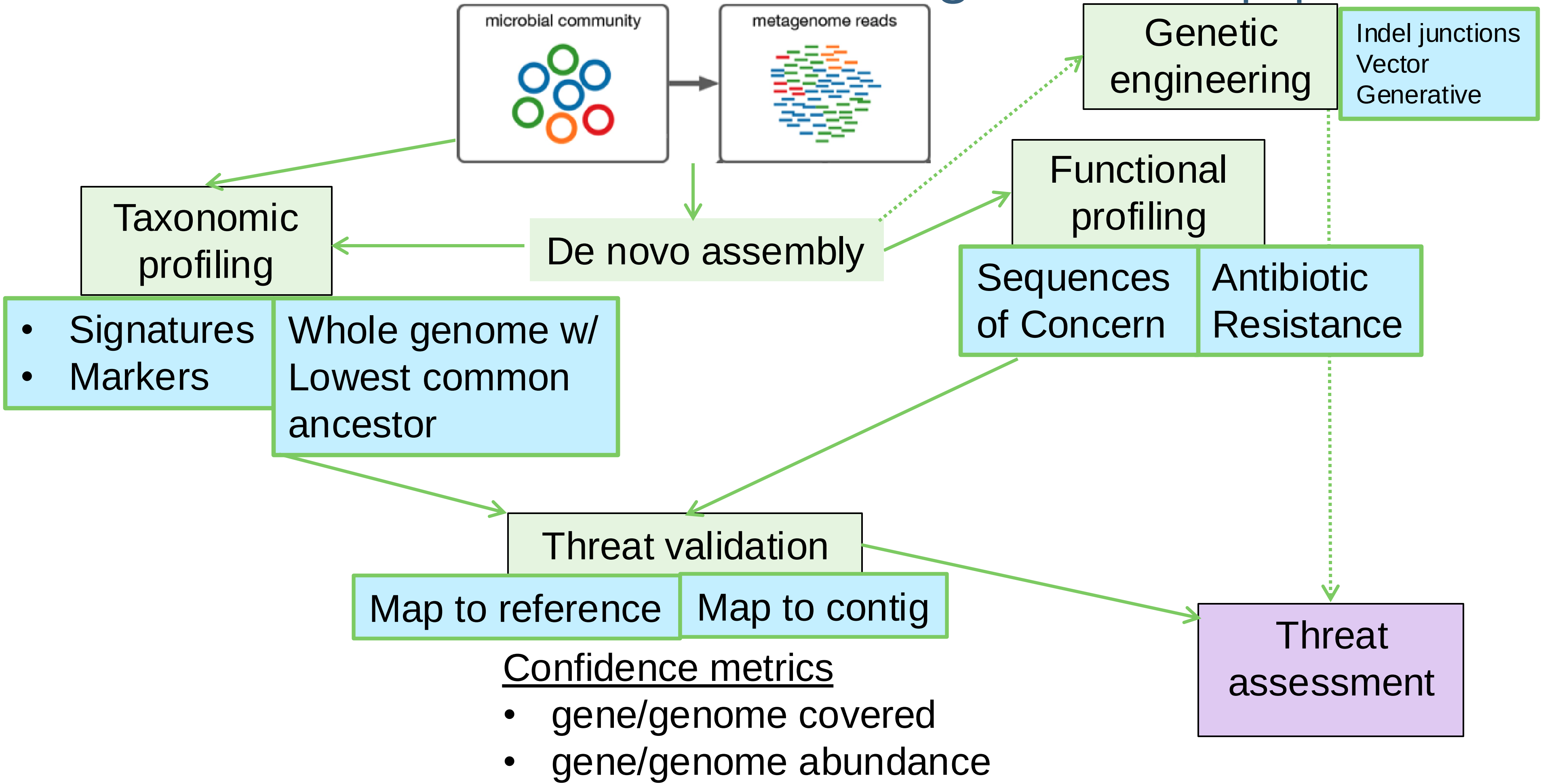
Organism detection needs

- All human infecting pathogens
- All animal infecting pathogens
- Species emerging from natural background
- Engineered or generative design
- Truly novel – extra terrestrial or otherwise highly divergent

Kingdom

- Fungi
- Bacteria
- Microeukaryote
- RNA virus
- DNA virus

Current state of the art metagenomics pipeline



Genetic engineering

Indel junctions
Vector
Generative

Functional profiling

Sequences of Concern

Antibiotic Resistance

De novo assembly

Taxonomic profiling

- Signatures
- Markers

Whole genome w/ Lowest common ancestor

Threat validation

Map to reference

Map to contig

Threat assessment

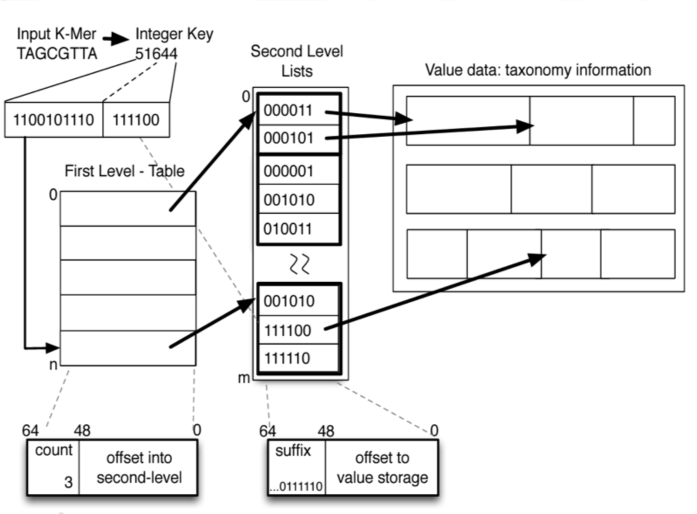
Confidence metrics

- gene/genome covered
- gene/genome abundance

Maintaining a reference database for *comprehensive* metagenomic classification

We published 1st scalable k-mer search index 2013 (LMAT)

Ames et al., Bioinformatics 2013

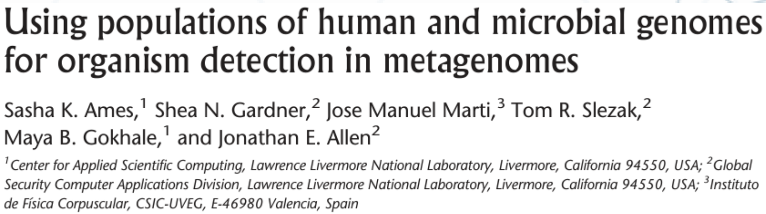


A key innovation was to include draft genomes

- Microbial
- Human

- DB size = 612 GB

We found human sequences in the human microbiome project in screen against 1000 human genomes

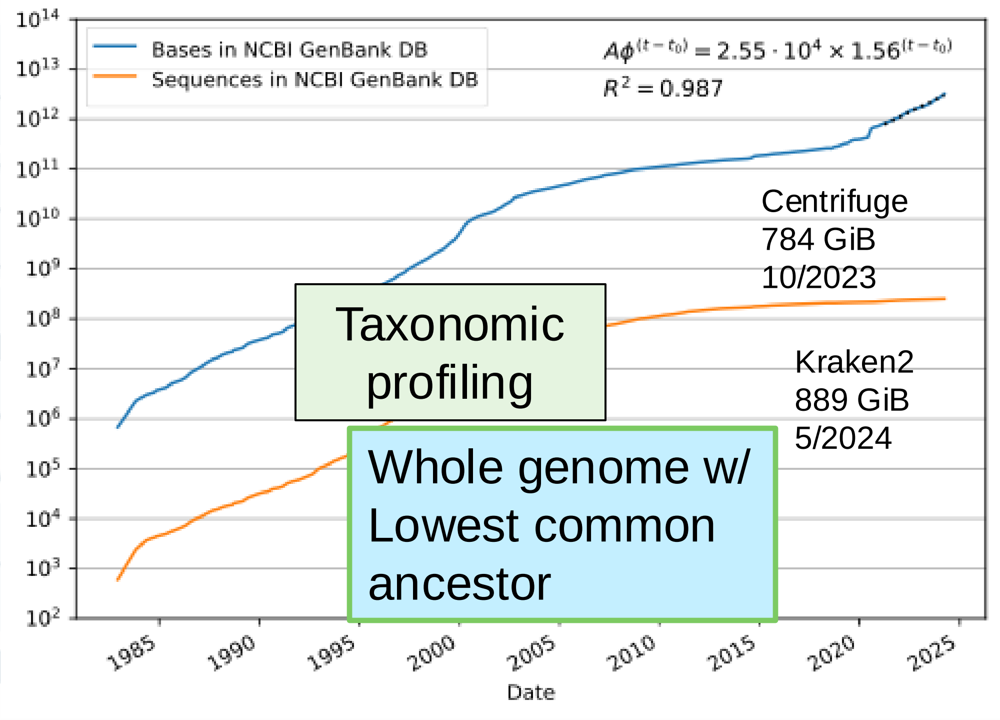


Using populations of human and microbial genomes for organism detection in metagenomes

Sasha K. Ames,[1] Shea N. Gardner,[2] Jose Manuel Marti,[3] Tom R. Slezak,[2] Maya B. Gokhale,[1] and Jonathan E. Allen[2]

[1]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; [2]Global Security Computer Applications Division, Lawrence Livermore National Laboratory, Livermore, California 94550, USA; [3]Instituto de Física Corpuscular, CSIC-UVEG, E-46980 Valencia, Spain

Using the right database for *comprehensive* metagenomic classification is a challenge



Centrifuge
784 GiB
10/2023

Kraken2
889 GiB
5/2024

Taxonomic profiling

Whole genome w/ Lowest common ancestor

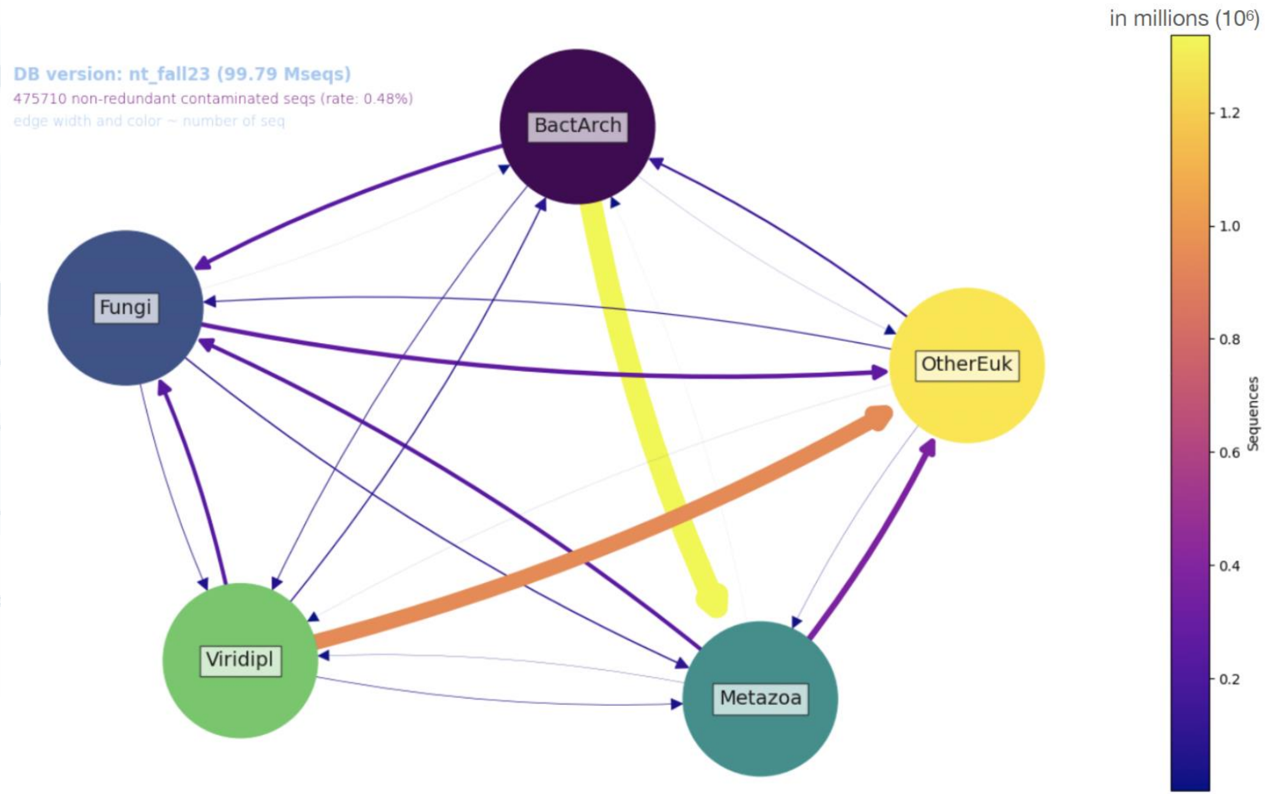
Marti JM, et al. bioRxiv 2024.

CZ ID stopped using full NT database in May 2024 and will use lossy compression

Kraken2’s database is not documented for decontamination

We developed the most complete decontaminated DB to date with a published protocol

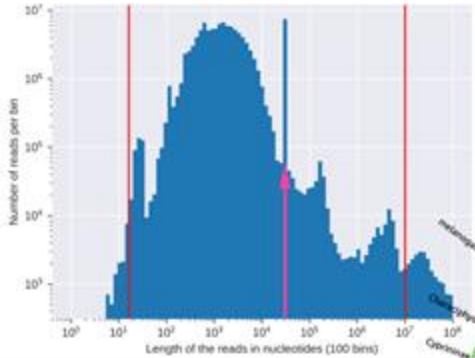
Decontamination of reference sequences is a growing computational burden


DB version: nt_fall23 (99.79 Mseqs)
475710 non-redundant contaminated seqs (rate: 0.48%)
edge width and color ~ number of seq
BactArch
Fungi
OtherEuk
Viridipl
Metazoa
in millions (10⁶)
1.2
1.0
0.8
0.6
0.4
0.2
Sequences

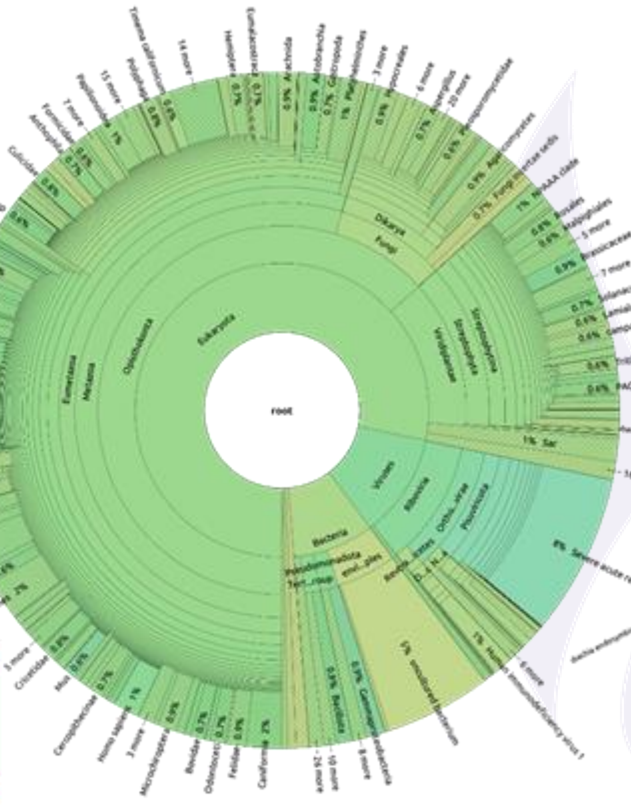
Now takes 2+ weeks to run on our cluster

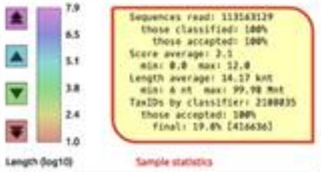
Conterminator software

nt DB content evolution over time



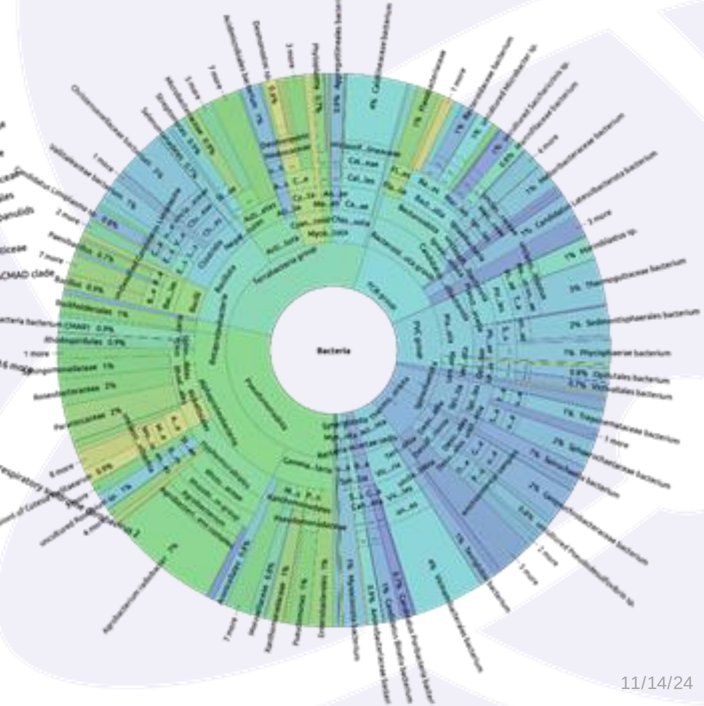
Full nt content
(by num/length of seqs)
[Jan 2024]



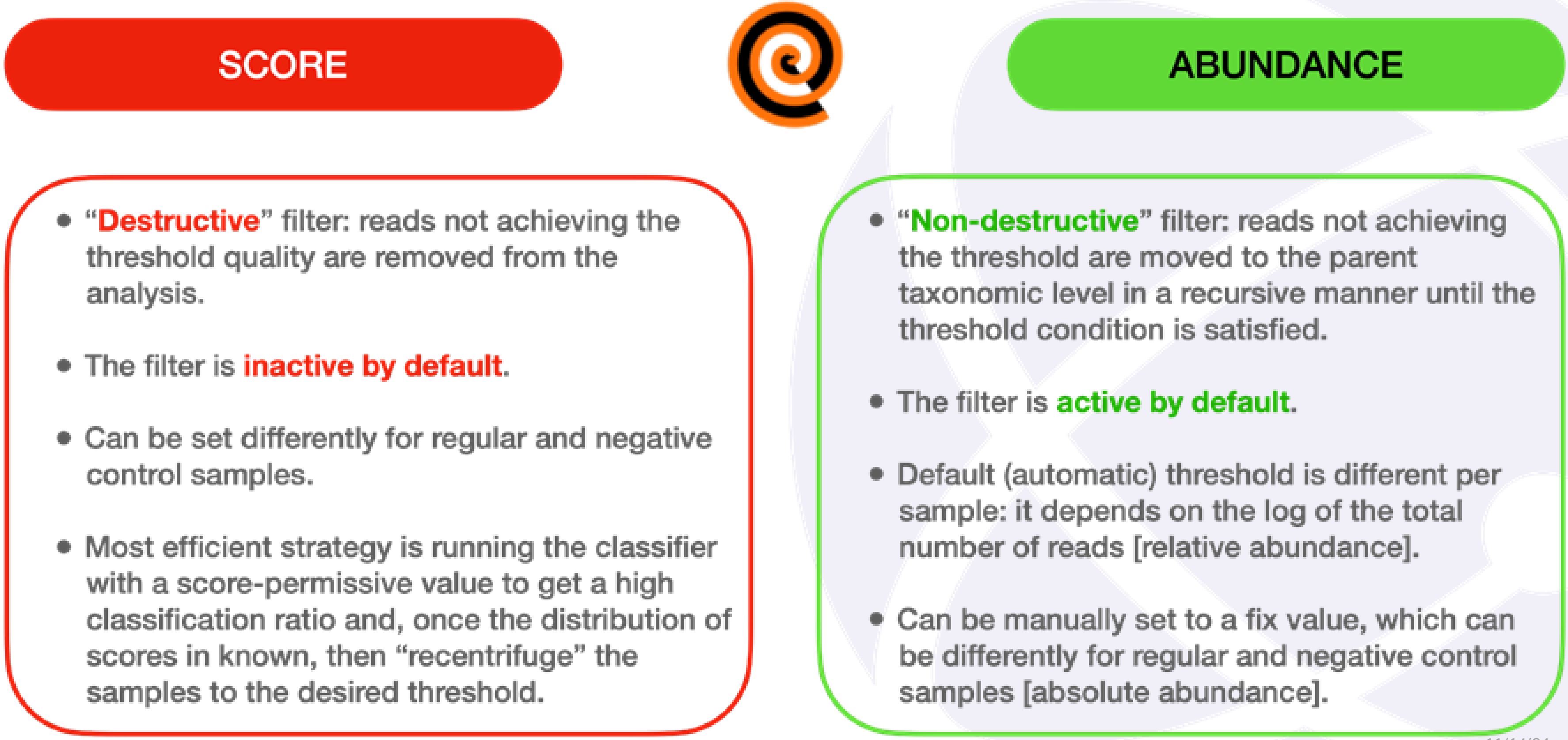


```
Sequences read: 113163129
  those classified: 100%
   those accepted: 100%
Score average: 2.1
  min: 0.0  max: 12.0
Length average: 14.17 knt
  min: 6 nt  max: 99.98 Mnt
TaxIDs by classifier: 2100035
  those accepted: 100%
   final: 19.0% [416636]
```

Novel bacterias added to nt
(by num/length of seqs)
[from Oct 2023 to Jan 2024]



Two different filtering strategies: by score and abundance



SCORE

- “**Destructive**” filter: reads not achieving the threshold quality are removed from the analysis.
- The filter is **inactive by default**.
- Can be set differently for regular and negative control samples.
- Most efficient strategy is running the classifier with a score-permissive value to get a high classification ratio and, once the distribution of scores in known, then “recentrifuge” the samples to the desired threshold.

ABUNDANCE

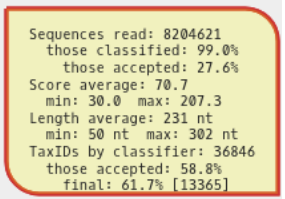
- “**Non-destructive**” filter: reads not achieving the threshold are moved to the parent taxonomic level in a recursive manner until the threshold condition is satisfied.
- The filter is **active by default**.
- Default (automatic) threshold is different per sample: it depends on the log of the total number of reads [relative abundance].
- Can be manually set to a fix value, which can be differently for regular and negative control samples [absolute abundance].

What is the source of current unknowns?

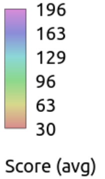
Waste water sample from contra costa county after human DNA removal

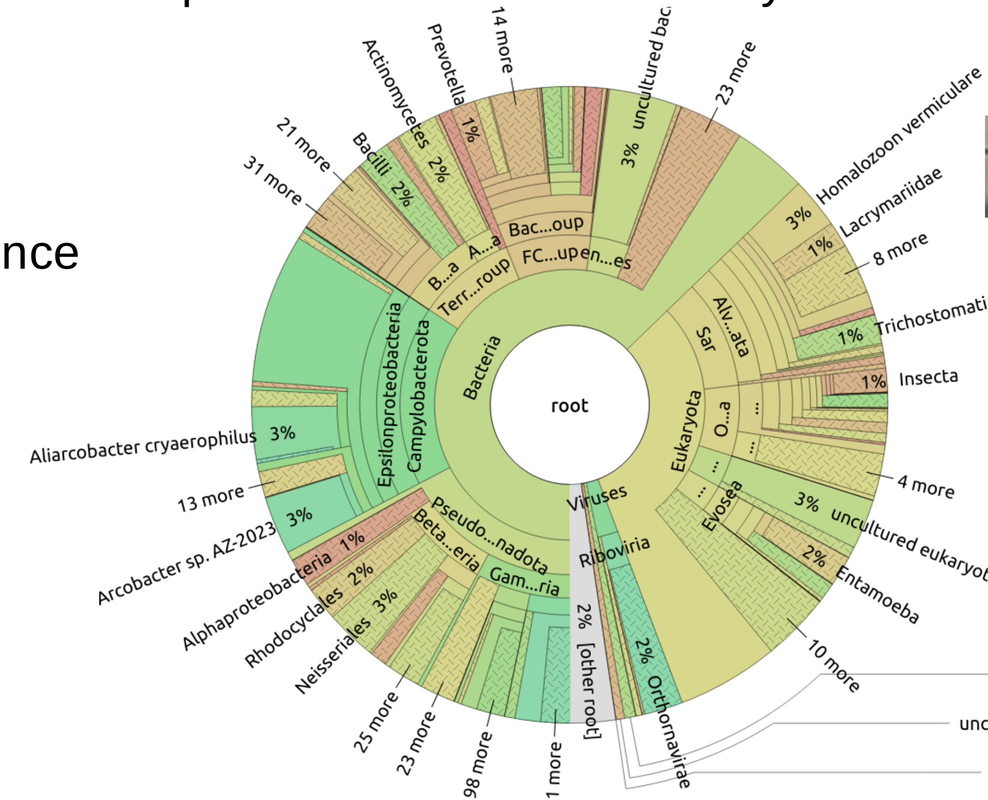
Source of unknowns

- Sequence divergence
- Eukaryotes



```
Sequences read: 8204621
  those classified: 99.0%
    those accepted: 27.6%
Score average: 70.7
   min: 30.0  max: 207.3
Length average: 231 nt
   min: 50 nt  max: 302 nt
TaxIDs by classifier: 36846
   those accepted: 58.8%
     final: 61.7% [13365]
```

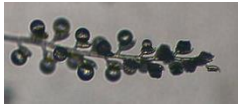




Archaea 0.6% ▨

unclassified sequences 0.4% ▨

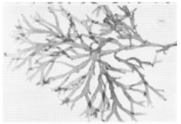
other sequences 0.004% ▨



Kickxellaceae

Homalozoon vermiculare

Obtectomera

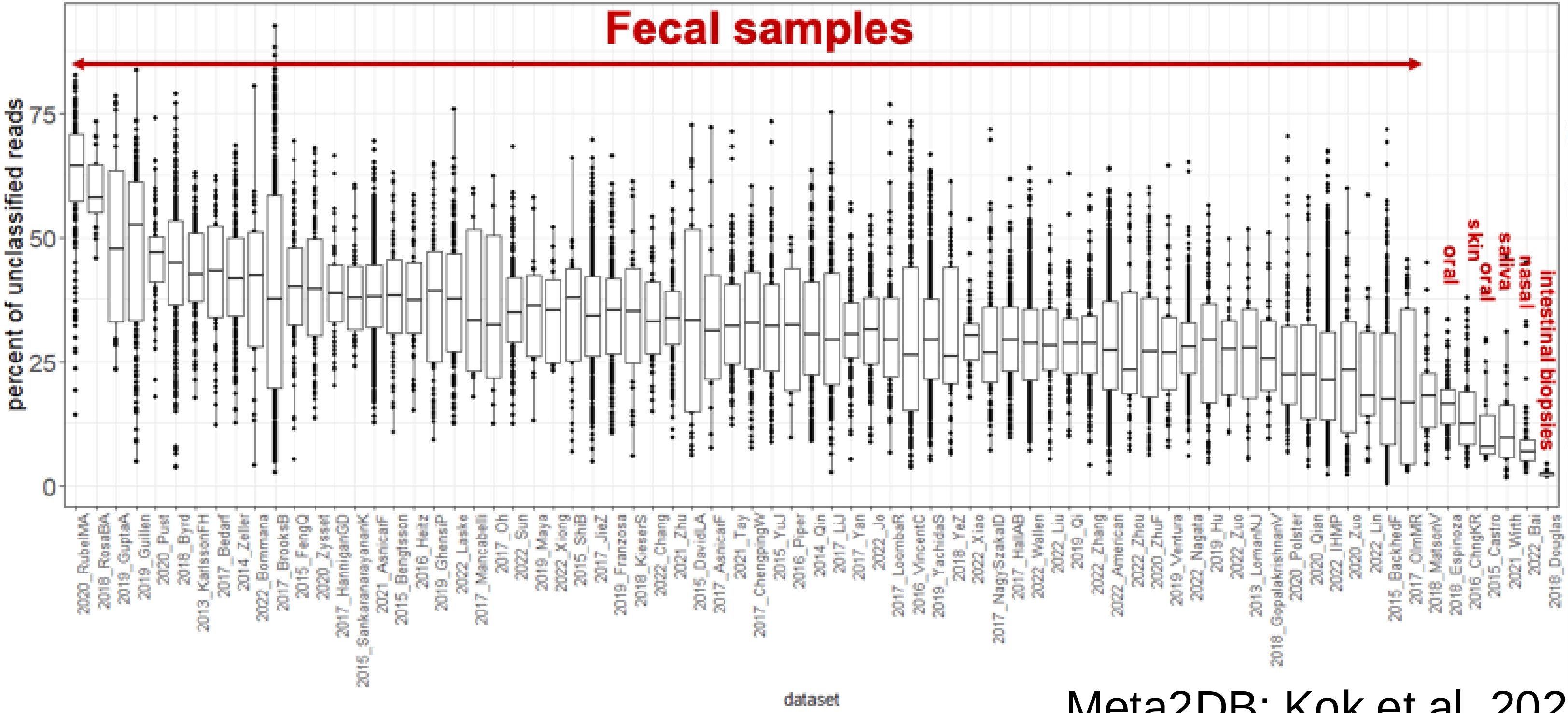


Acrosorium

Entamoeba

With subtraction of no template controls

Unclassified reads present challenges for detection of novel pathogens



Meta2DB: Kok et al, 2024 bioRxiv

Unclassified reads across ~12K human-associated microbiome

Path forward with traditional genomic search

Make more reference data searchable!!!!

- Metagenomic assembled contigs
- More eukaryotic draft genomic sequences

Pros

- Search algorithms are well understood and relatively scalable with sufficient compute
- Potential for explainability: matches can be assigned to individually sequenced organisms and environments

Cons

- Computational cost to build database AND every search is costly

Challenges

- Automated quality control of new reference sequences and incorporating uncertainty into taxonomy calls

How do we conduct threat assessment for unknowns?

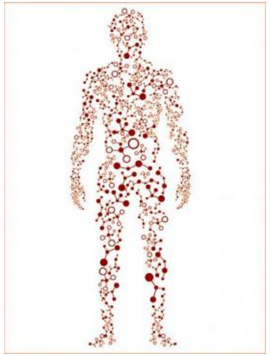
Model pathogen-host interactions
Predict protein-protein interactions
Predict physiologic impact of molecular interactions

Fold protein

Input: peptide or receptor

Match with existing protein-protein co-complexes

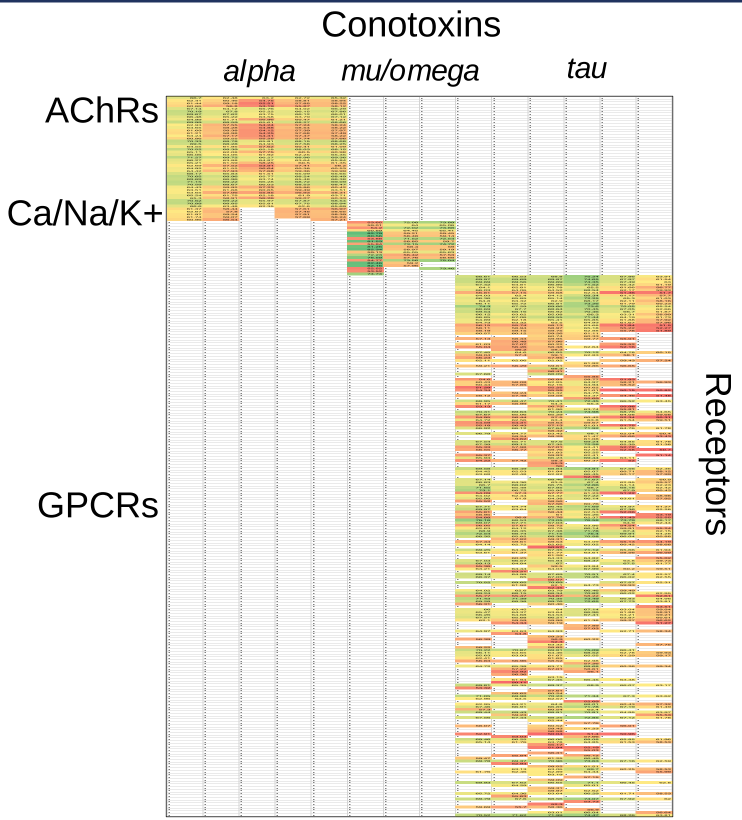
Structurally align matched receptor against **<u>entire human “proteome”</u>**

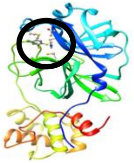


Use PDBSpheres to structurally align matching templates
Zemla et al., 2023

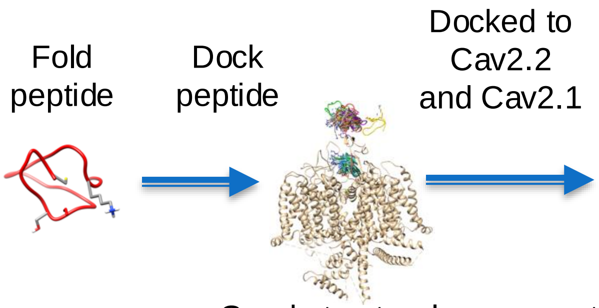
Structural alignment identifies receptors with potential functional interference

Selected peptide-protein co-complexes are searched against all human protein pockets with available structures

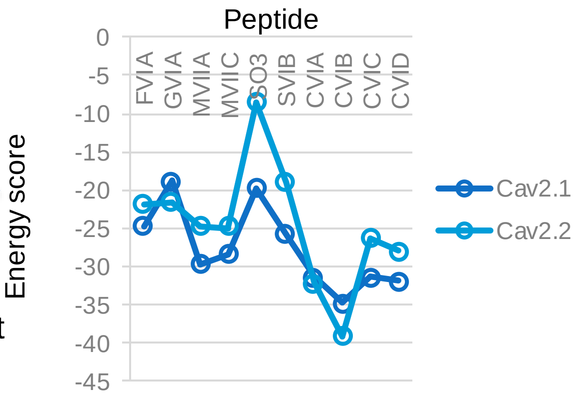


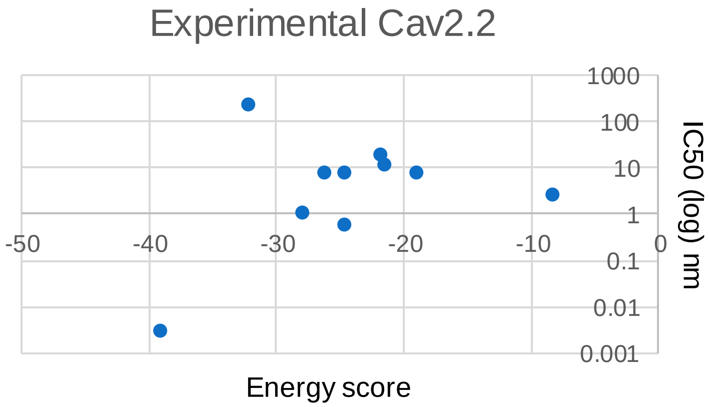


Docking simulations model peptide binding to capture molecular dynamics



Good structural agreement where data is available





We are evaluating docking methods for conotoxins against all human cell receptors with comparisons to experimental data (patch clamp and binding assays)

Current computational cost: ~13 minutes per peptide/receptor pair

Uncovering microbial dark matter

- Do more with traditional search by searching more reference material than currently done
  - Presents significant engineering challenges: contamination in reference sequences, matching to samples not just reference genomes
- Expand traditional search into 3D structure space and learned feature representations
  - Fundamental limitation to assessing protein function in context of a single protein
- Infer microbial pathogens through improved molecular interaction modeling
  - Mechanisms replication
  - Host cell attachment
  - Microbial transmission
- Major challenges with mapping molecular level mechanisms to physiologic impact:
  - Host to host transmission and disease outcomes

Acknowledgements

Jose Manuel Marti - Recentrifuge
Crystal Jaing - Genomics
Nick Be – Microbiome
Car Reen Kok - Bioinformatics
Ed Lau – MD simulation
Heesung Shim – MD simulation

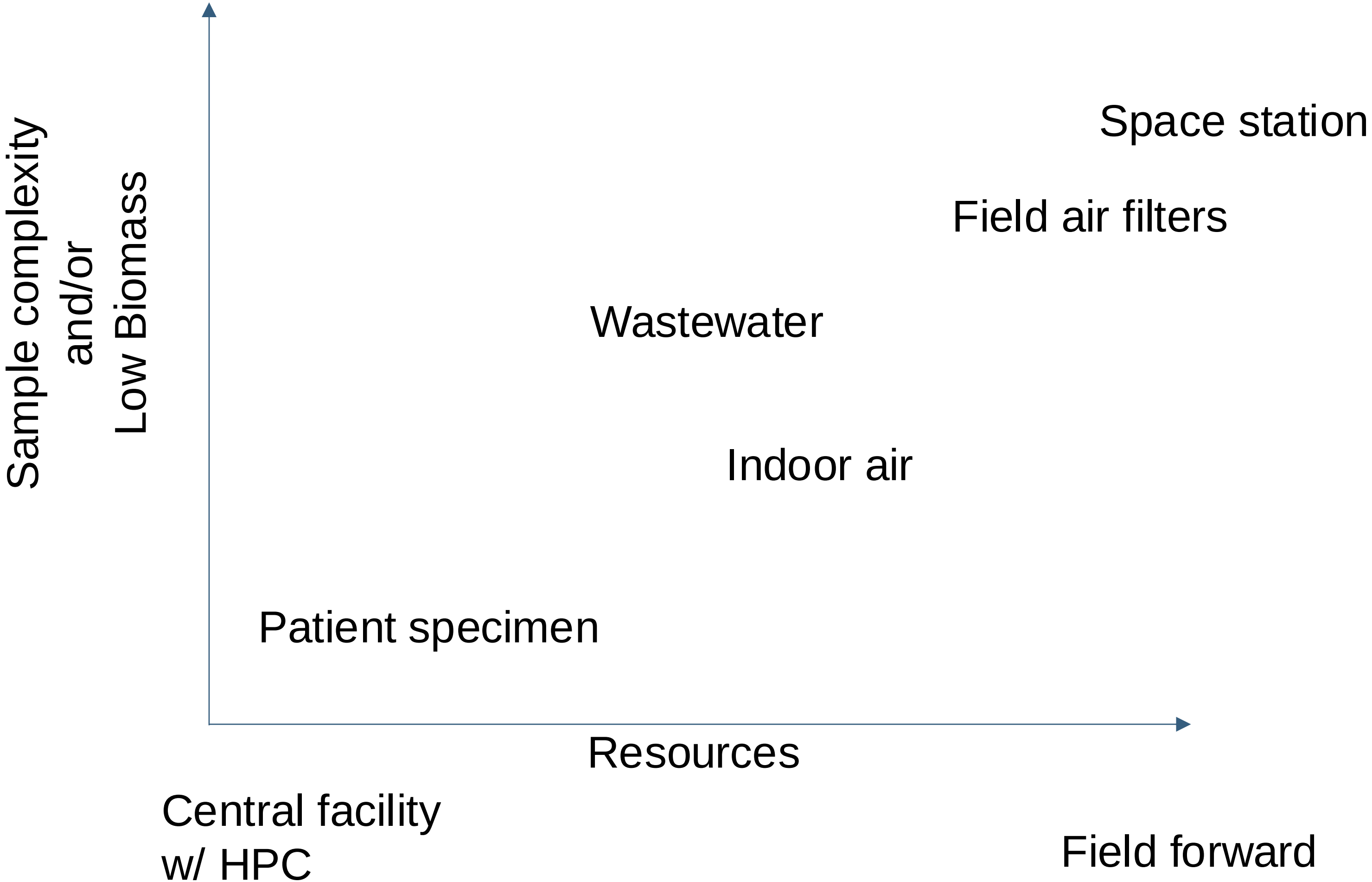
Bioinformatics and Genomics

Data science and Machine Learning

Funding: DOE-LLNL LDRD

Thank you

Limitations from operational setting



Sample complexity
and/or
Low Biomass

Space station

Field air filters

Wastewater

Indoor air

Patient specimen

Resources

Central facility
w/ HPC

Field forward

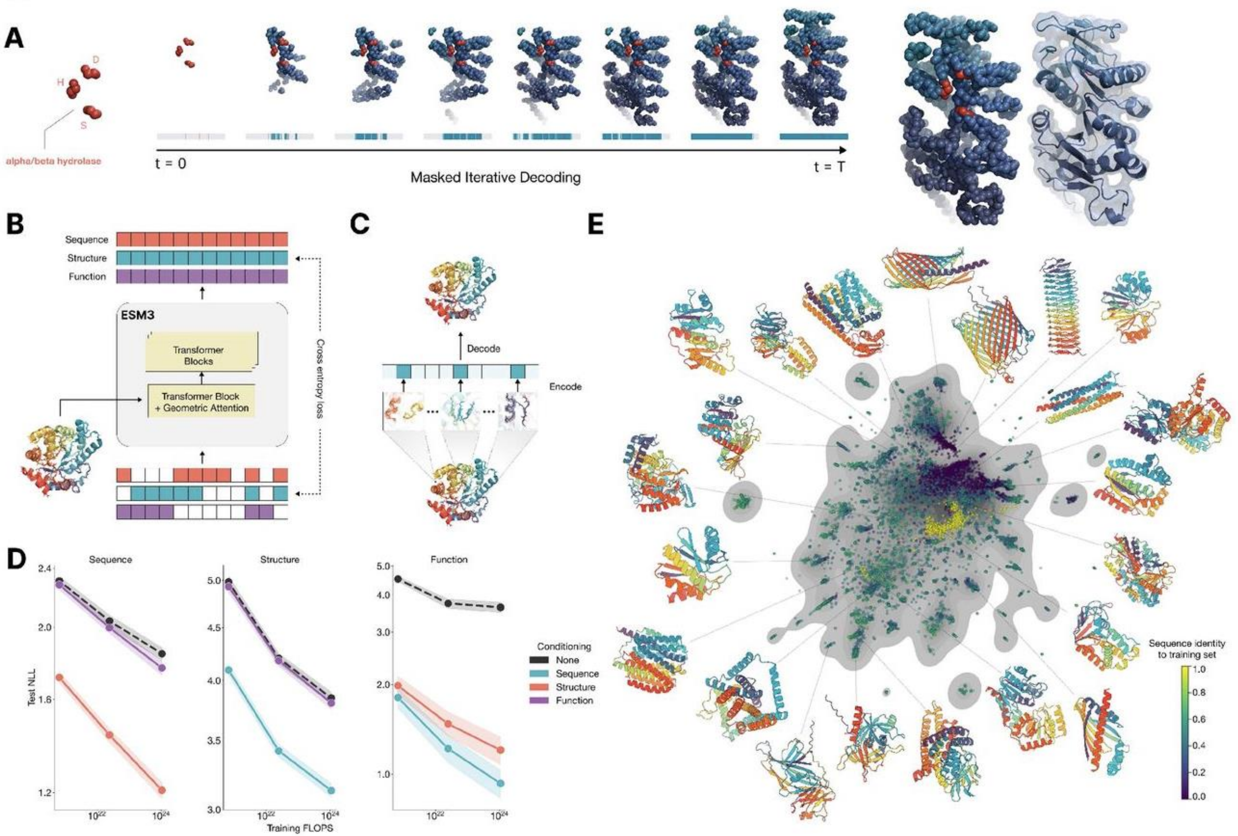
Path forward with new approaches: function from sequence

Pretrain model on large corpus of sequence data

- ESM3, AlphaFold3, Chia-1,
- LucaOne/Prot
- ProteInfer
- Evo

Some evidence for improvement in remote homology detection

Example protein language model ESM3


A
D
H
S
alpha/beta hydrolase
t = 0
Masked Iterative Decoding
t = T
B
Sequence
Structure
Function
ESM3
Transformer Blocks
Transformer Block + Geometric Attention
Cross entropy loss
C
Decode
Encode
E
D
Sequence
Structure
Function
Test NLL
2.4
2.0
1.6
1.2
5.0
4.5
4.0
3.5
3.0
5.0
4.0
3.0
2.0
1.0
10^{22}
10^{24}
Training FLOPS
Conditioning
None
Sequence
Structure
Function
Sequence identity to training set
1.0
0.8
0.6
0.4
0.2
0.0