Antibiotic resistance detection using foundational language models

Assessment of antibiotic resistance gene distribution within different environments in territory of Latvia

Edgars Liepa
Faculty of Biology
University of Latvia
Latvian Biomedical Research and Study Centre
PhD advisor dr. Dāvids Fridmanis

Motivation

- Bioinformatics masters in CS.
- Apply CS for biology research.
- Researcher assistant at biomedical research.
- AI is leading current bio-revolution.
- Antibiotic resistance is a disaster for modern medicine
- Assessment of antibiotic resistance gene distribution within different environments of territory of Latvia
- Monitoring System.

Content

- Antibiotic resistance (AR) research?
  - Antibiotic resistant bacteria and Antibiotic resistance genes (ARG)
  - ARG detection
  - Machine learning for AR research.
  - Large language models (LLMs) in life sciences.
  - Foundational language models
- National research program.
- Current progress.
- Planned activities.

Antibiotic resistance (AR)

- Ability of these pathogenic bacteria to survive and multiply in the presence of antibiotics, which are drugs designed to kill or inhibit their growth. This resistance occurs primarily through two mechanisms: (a) genetic mutations that alter the bacterial target of the antibiotic, and (b) acquisition of resistance genes from other bacteria through horizontal gene transfer.
- An ARG, or Antibiotic Resistance Gene, is a gene that enables bacteria to develop resistance against antibiotics. These genes provide bacteria with the ability to survive and multiply in the presence of antibiotics that would otherwise be lethal or inhibitory.

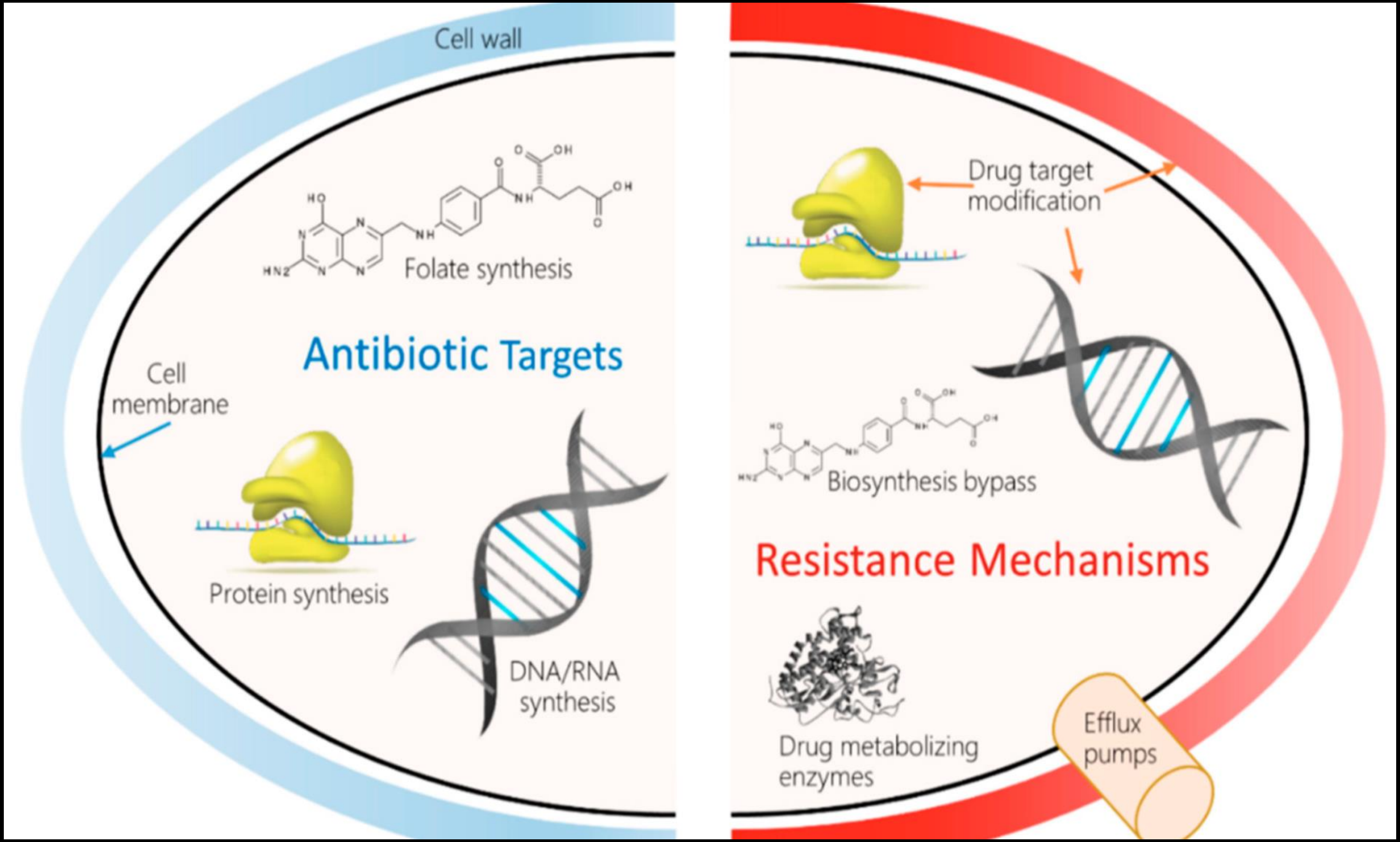


Figure 1. *Antimicrobial drug targets and molecular mechanisms of antimicrobial resistance (AMR). Left: The most common classes of AB currently in use impede bacterial growth by inhibiting the biosynthesis of peptidoglycan, a main constituent of cell wall; disrupting the bacterial cell membrane; and inhibiting DNA replication, gene transcription and translation, and folate biosynthesis. Right: In turn, bacteria have developed many resistance mechanisms to these attacks, such as pumping the AB out of the cell, inactivating the drug using specialized enzymes, modifying the target structures to prevent interference, and bypassing the affected metabolic pathway.* [2]

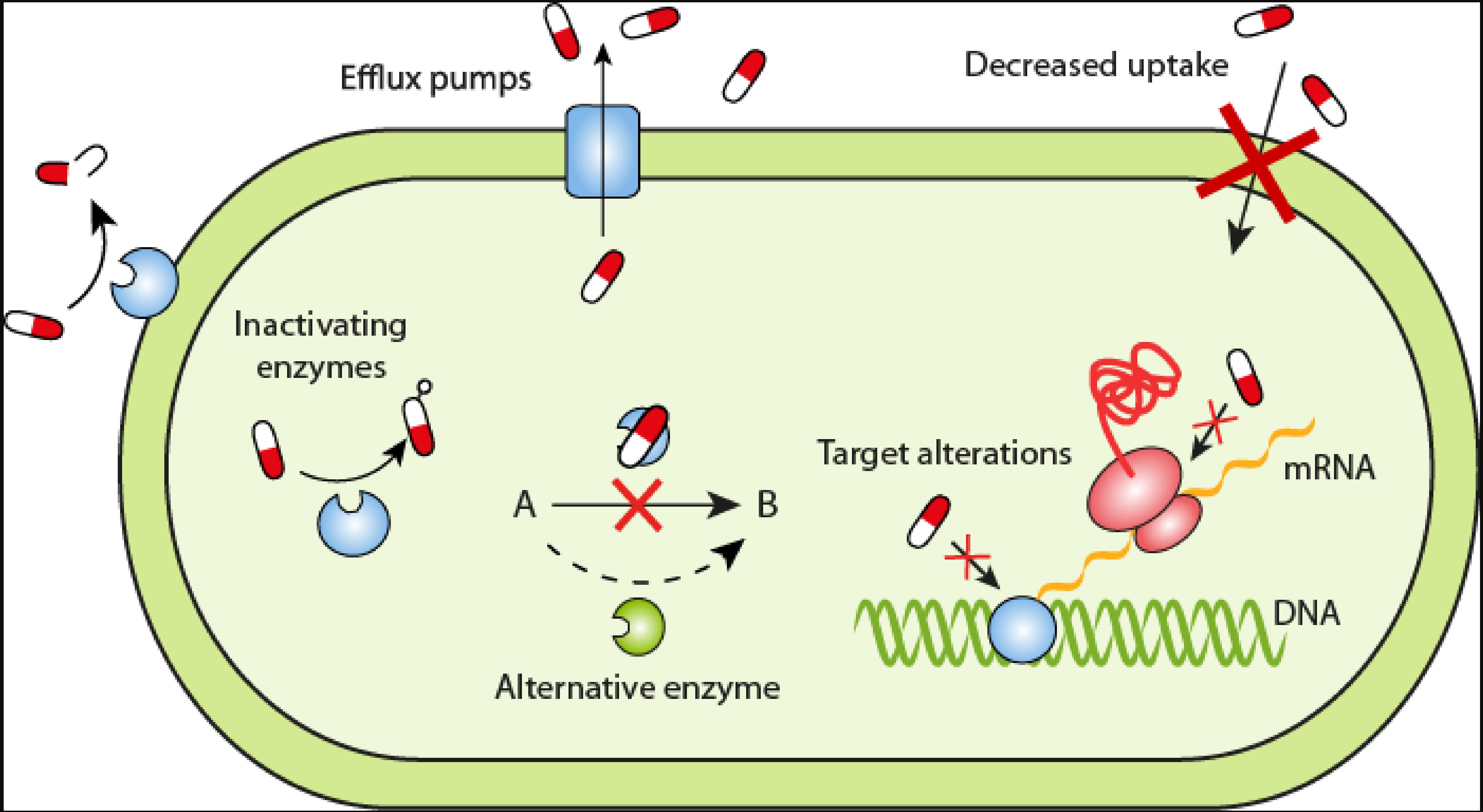
How resistance is acquired

Natural resistance - always expressed in the species or expressed to resistance levels after exposure to an antibiotic

Intrinsic resistance - trait that is shared universally within a bacterial species, is independent of previous antibiotic exposure , and not related to horizontal gene transfer

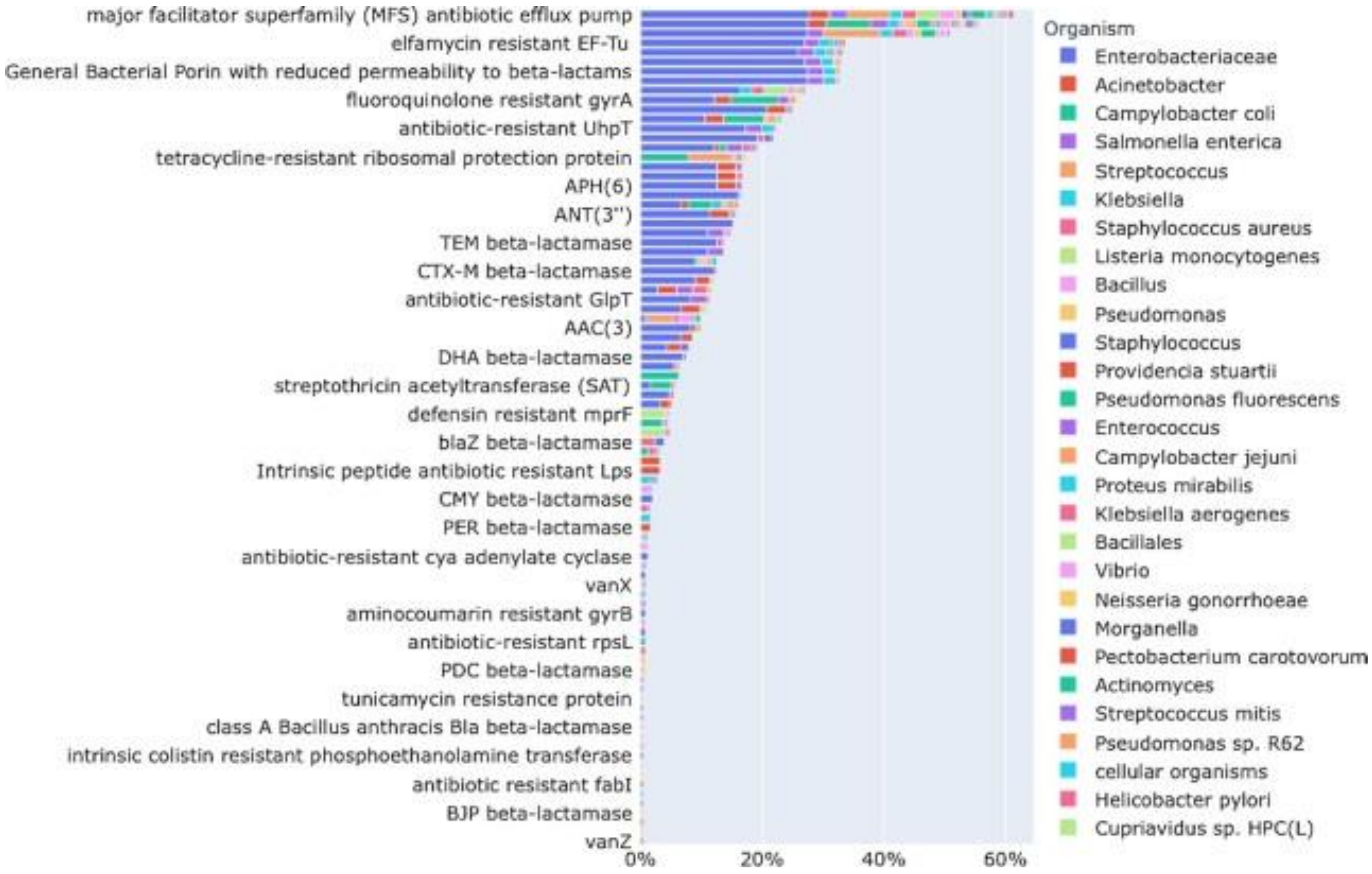
Acquired resistance:

- Acquisition of genetic material that confers resistance through- transformation
- Transposition (change position in a genome)
- conjugation (all termed horizontal gene transfer—HGT)
- mutations to its own chromosomal DNA.

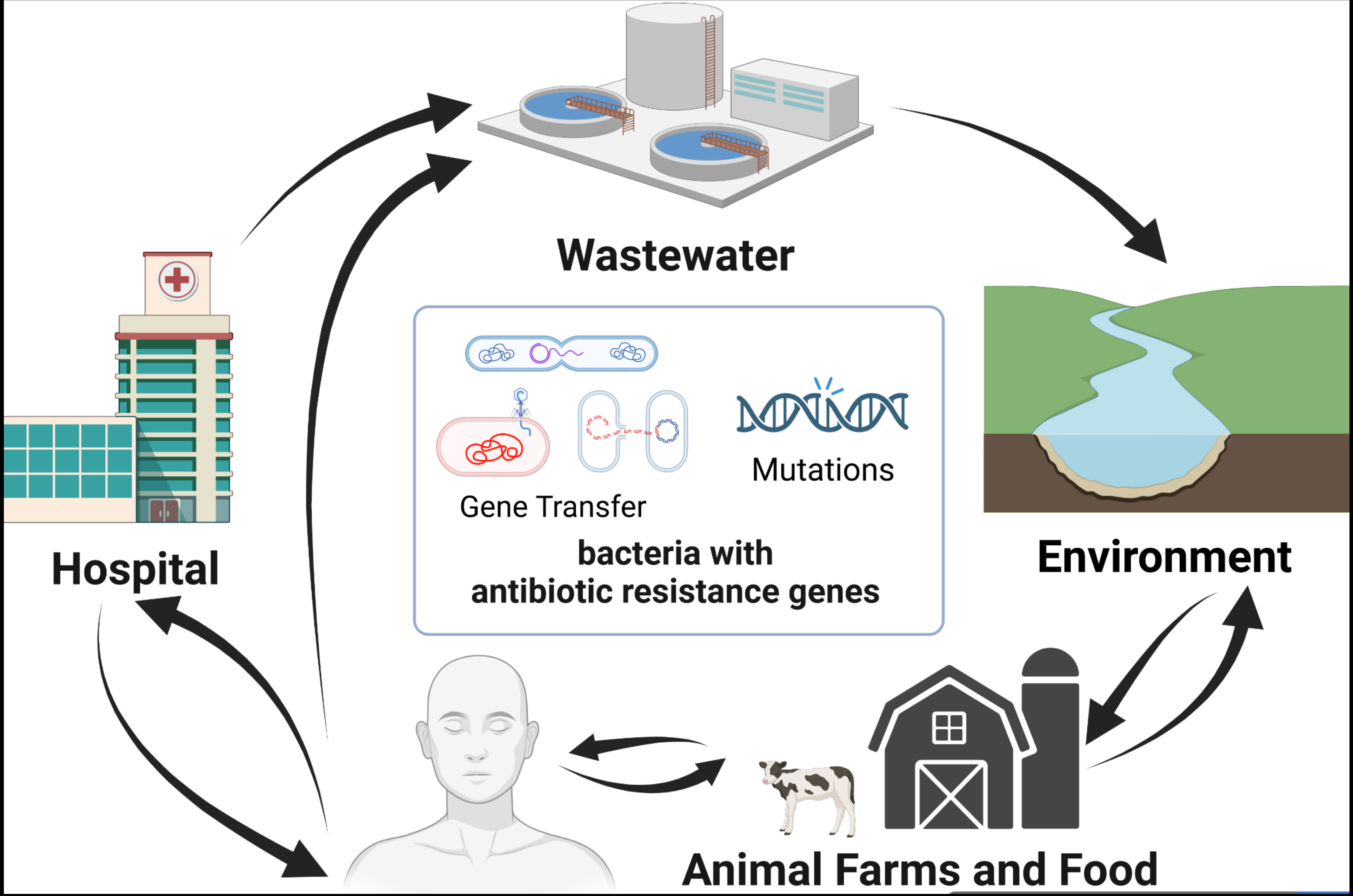


Antibiotic resistance mechanisms (reproduced from Gullberg et al., 2014) [5]

1. The first approach to predicting the AMR is to **analyze WGS** (e.g., Illumina sequencing) data by identifying the presence of known ARG or gene variants.
2. The second approach to the AMR analysis and prediction is to study changes **in gene expression** of the isolate upon drug treatment.
3. The third approach is gene agnostic and based on **global genomic comparison** of multiple strains with various susceptibility to different drugs.
4. The last approach, orthogonal to those described above, takes advantage **of metabolic profiling**. [2]



CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 2023



Antibiotic resistant genes are spread through environment and can be observed the most in ARG hotspots. In these hotspots there is incresed risk of new ARG's appearing.

Methods for AR detection

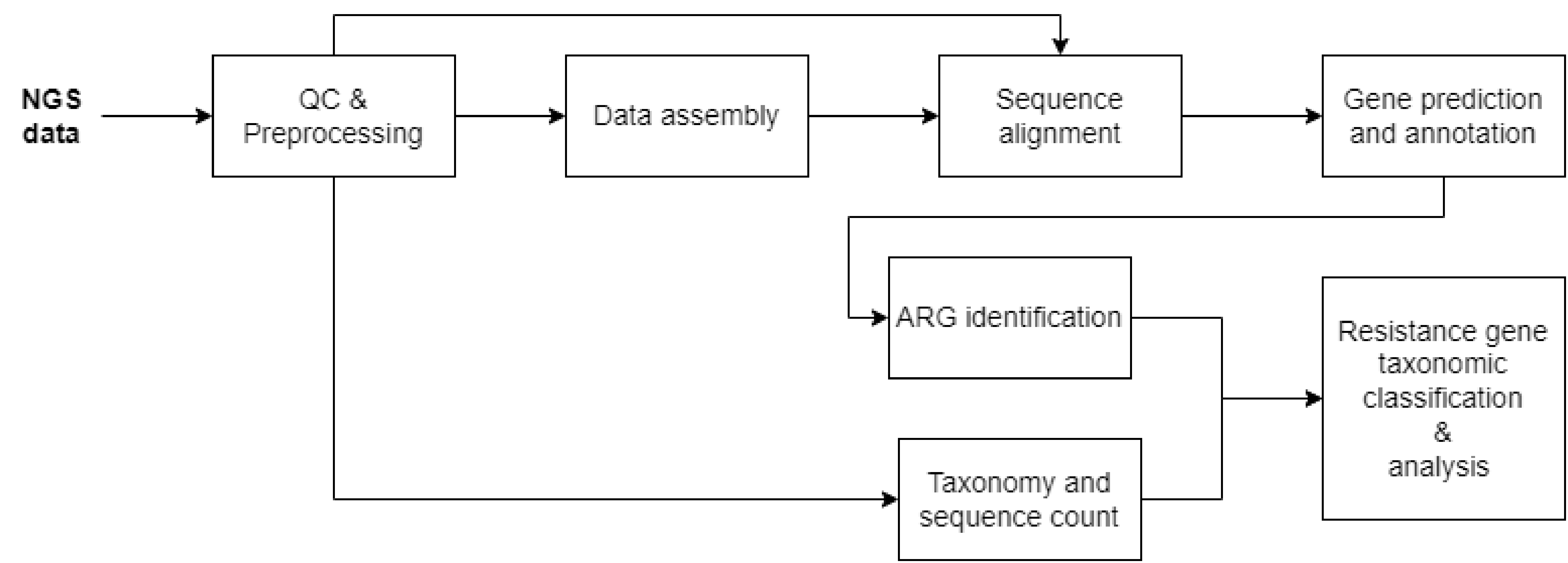
Phenotypic methods

- Antimicrobial susceptibility testing AST

Genotypic methods:

- DNA sequencing
- ARG from genome or metagonem data
  - ARG matching to database
    - Resfinder, AMRfinder, Comprehensive Antibiotic Resistance Database (CARD)
  - Machine learning and classification
    - ARG-ML
  - Gene function prediction
    - GEN-LMS

Next generation sequencing data processing pipeline



NGS data

QC & Preprocessing

Data assembly

Sequence alignment

Gene prediction and annotation

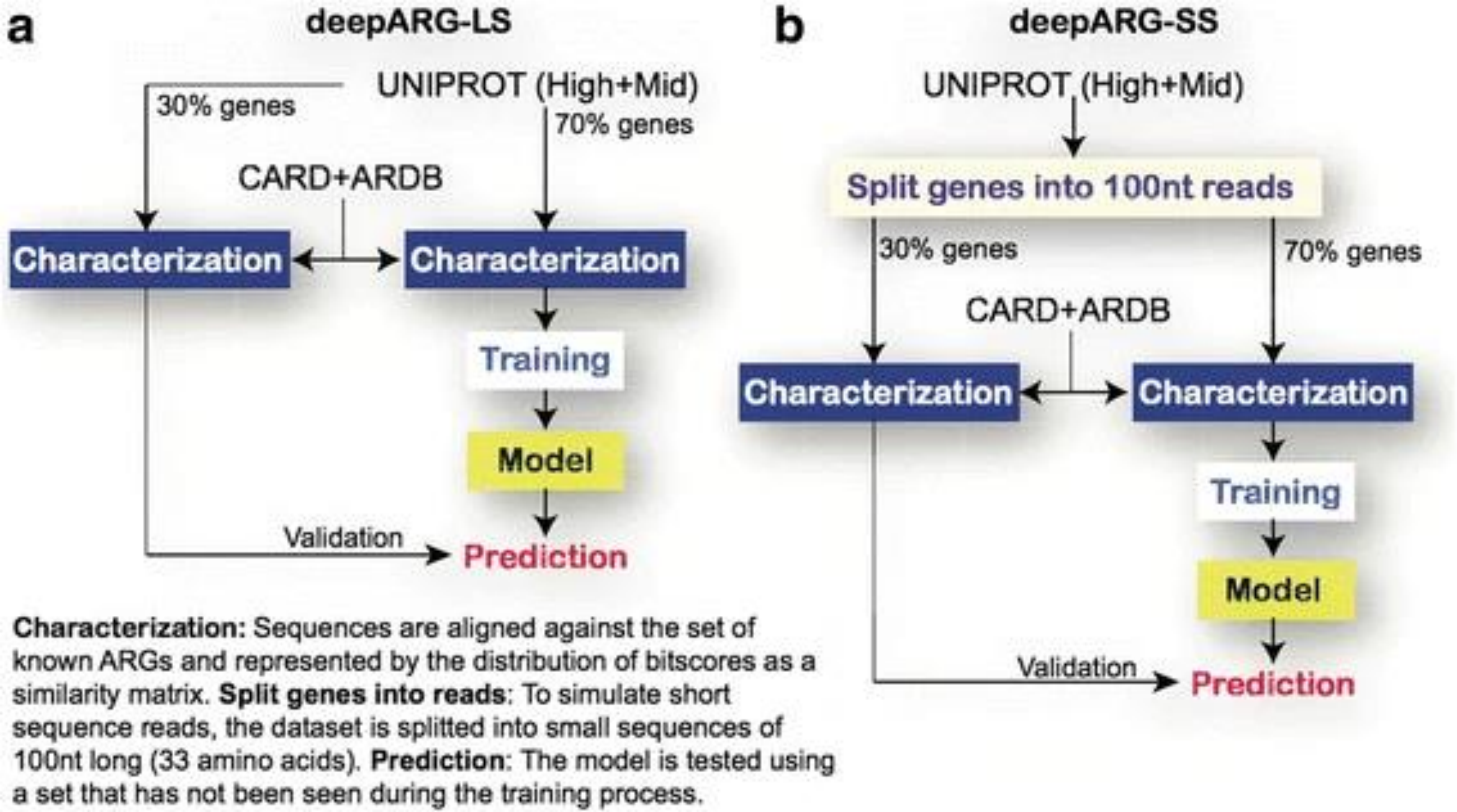
ARG identification

Taxonomy and sequence count

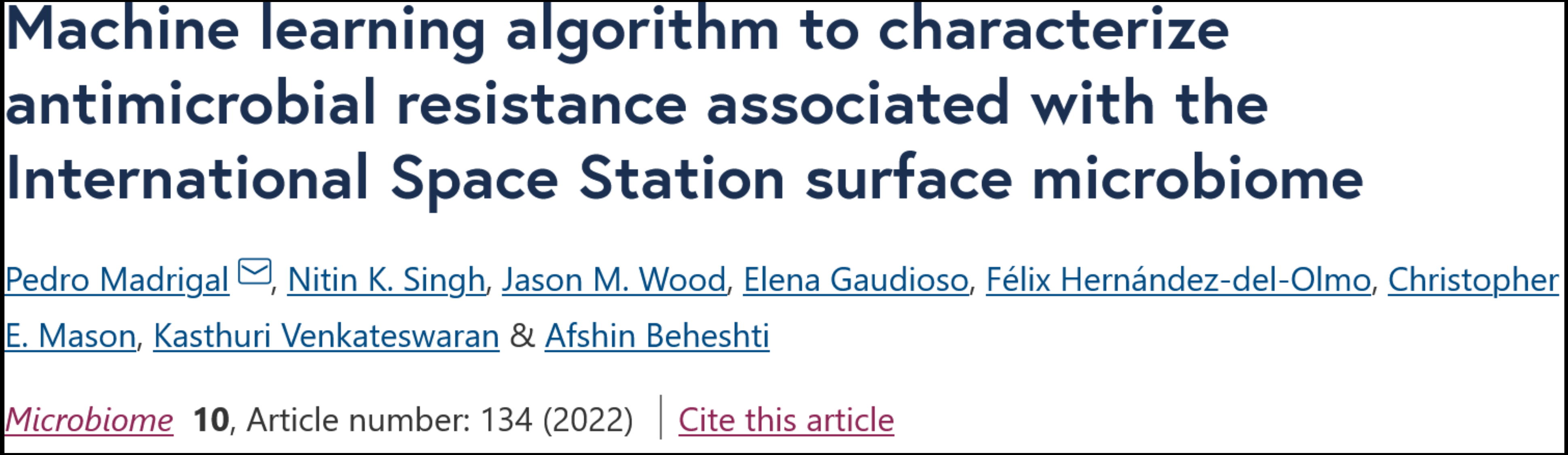
Resistance gene taxonomic classification & analysis

Commonly used ML algorithms for AMR

- Naïve Bayes (NB),
- Decision trees (DT),
- Random forests (RF),
- Support vector machines (SVM),
- Artificial neural networks (ANN).



Characterization: Sequences are aligned against the set of known ARGs and represented by the distribution of bitscores as a similarity matrix. **Split genes into reads**: To simulate short sequence reads, the dataset is splitted into small sequences of 100nt long (33 amino acids). **Prediction**: The model is tested using a set that has not been seen during the training process.



Machine learning algorithm to characterize antimicrobial resistance associated with the International Space Station surface microbiome

Pedro Madrigal, Nitin K. Singh, Jason M. Wood, Elena Gaudioso, Félix Hernández-del-Olmo, Christopher E. Mason, Kasthuri Venkateswaran & Afshin Beheshti

Microbiome **10**, Article number: 134 (2022) | Cite this article

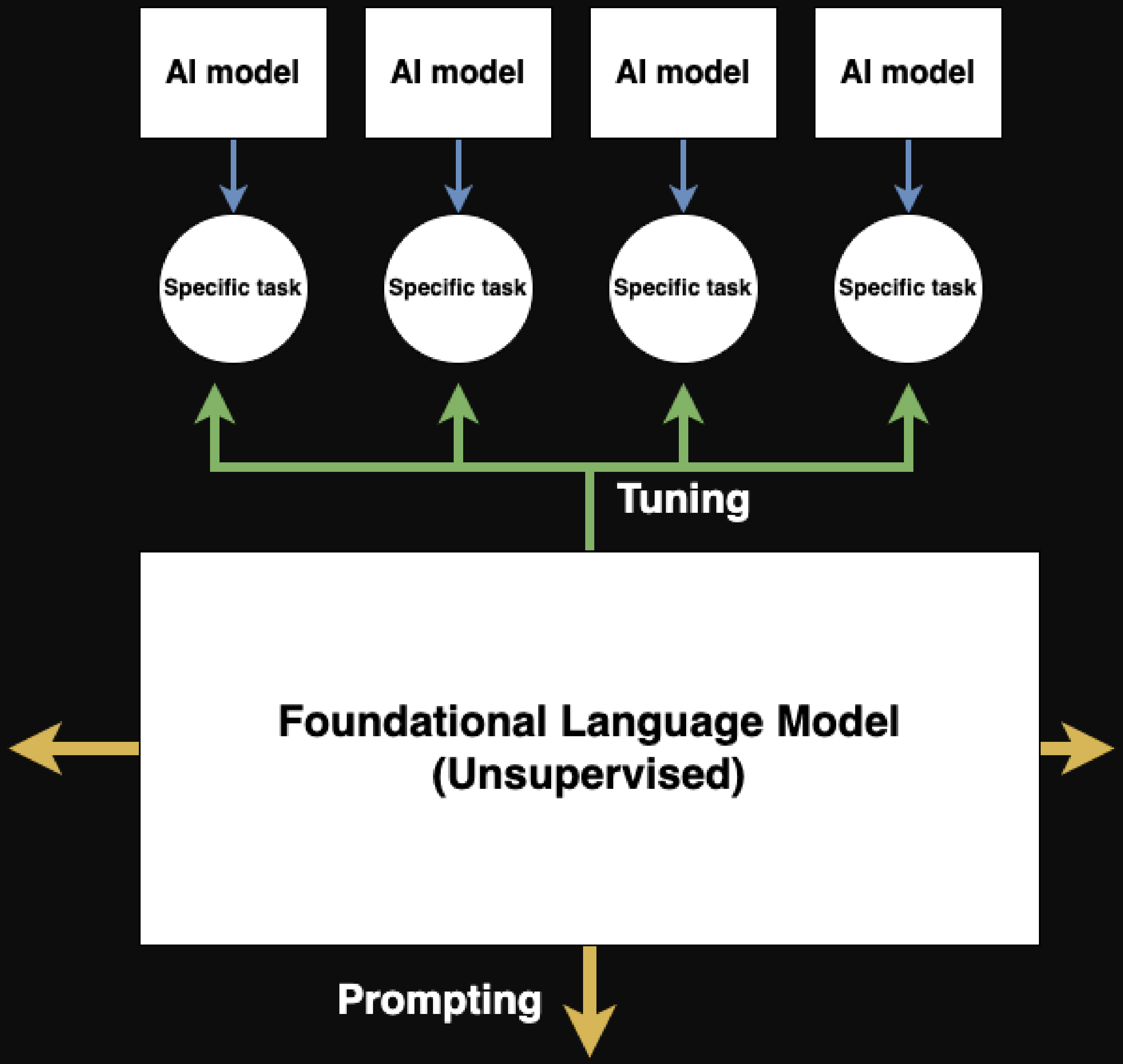
Results

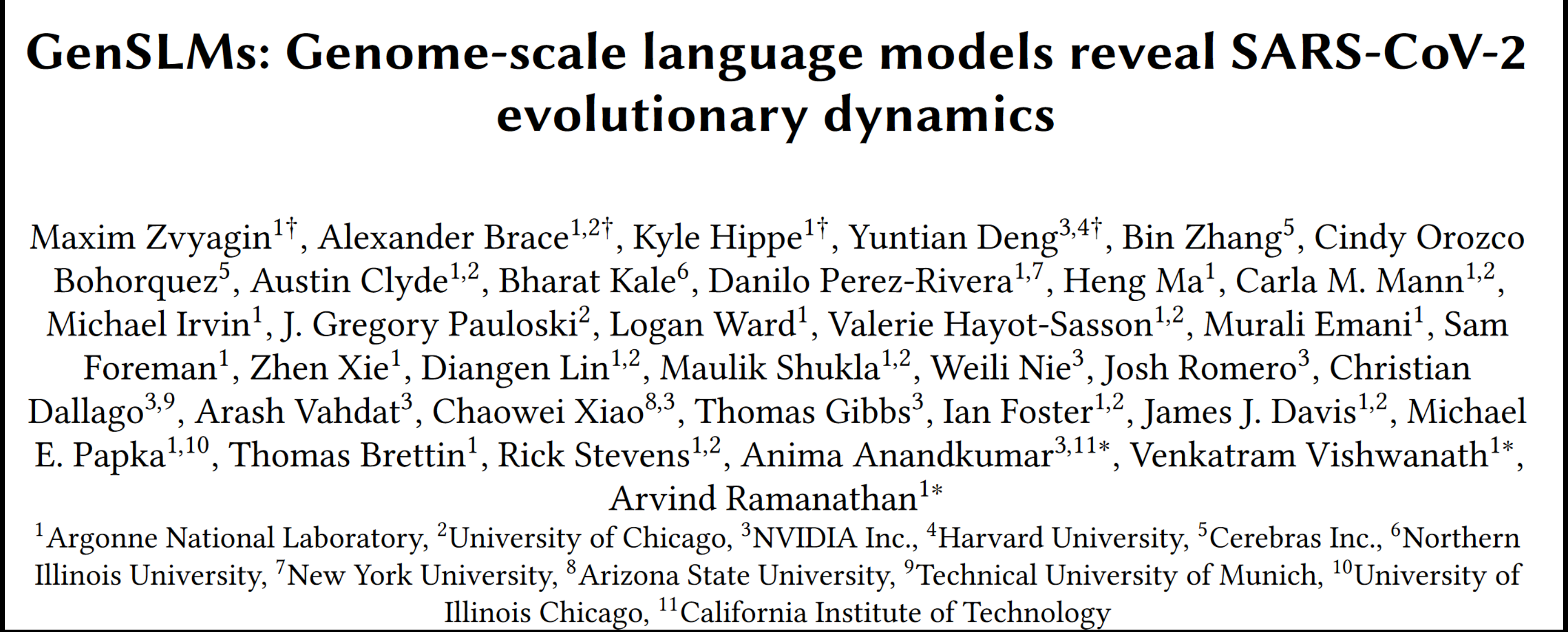
We have analyzed the data using a deep learning model, allowing us to go **beyond traditional cut-offs** based only on high DNA sequence similarity and extending the catalog of AMR genes. Our results in PMA treated samples revealed AMR dominance in the last flight for Kalamiella piersonii, a bacteria related to urinary tract infection in humans. **The analysis of 226 pure** strains isolated from the MT-1 project **revealed hundreds of antibiotic resistance genes from many isolates**, including two top-ranking species that corresponded to strains of Enterobacter bugandensis and Bacillus cereus. Computational predictions were **experimentally validated** by antibiotic resistance profiles in these two species, showing a high degree of concordance. Specifically, disc assay data confirmed the high resistance of these two pathogens to various beta-lactam antibiotics.

Conclusion

Overall, our computational predictions and validation analyses demonstrate the advantages of machine learning to uncover concealed AMR determinants in metagenomics datasets, expanding the understanding of the **ISS environmental microbiomes** and their pathogenic potential in humans.

- Advantage
  - Performance
  - Generalization
  - Adabyability
- Disadvantage
  - Compute heavy while training
  - Trust
  - Inference at the moment is quite compute heavy as well.





GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics

Maxim Zvyagin[1†], Alexander Brace[1,2†], Kyle Hippe[1†], Yuntian Deng[3,4†], Bin Zhang[5], Cindy Orozco Bohorquez[5], Austin Clyde[1,2], Bharat Kale[6], Danilo Perez-Rivera[1,7], Heng Ma[1], Carla M. Mann[1,2], Michael Irvin[1], J. Gregory Pauloski[2], Logan Ward[1], Valerie Hayot-Sasson[1,2], Murali Emani[1], Sam Foreman[1], Zhen Xie[1], Diangen Lin[1,2], Maulik Shukla[1,2], Weili Nie[3], Josh Romero[3], Christian Dallago[3,9], Arash Vahdat[3], Chaowei Xiao[8,3], Thomas Gibbs[3], Ian Foster[1,2], James J. Davis[1,2], Michael E. Papka[1,10], Thomas Brettin[1], Rick Stevens[1,2], Anima Anandkumar[3,11*], Venkatram Vishwanath[1*], Arvind Ramanathan[1*]

[1]Argonne National Laboratory, [2]University of Chicago, [3]NVIDIA Inc., [4]Harvard University, [5]Cerebras Inc., [6]Northern Illinois University, [7]New York University, [8]Arizona State University, [9]Technical University of Munich, [10]University of Illinois Chicago, [11]California Institute of Technology

We seek to transform how **new and emergent variants** of pandemic-causing viruses, specifically SARS-CoV-2, **are identified and classified**. By adapting large language models (LLMs) for genomic data, we build **genome-scale language models** (GenSLMs) which can **learn the evolutionary landscape** of SARS-CoV-2 genomes. By pretraining on over **110 million prokaryotic gene sequences** and fine-tuning a SARS-CoV-2-specific model on **1.5 million genomes**, we show that GenSLMs can accurately and **rapidly identify** variants of concern. Thus, to our knowledge, GenSLMs represents one of the first whole genome scale foundation models which can **generalize to other prediction tasks**. We demonstrate scaling of GenSLMs on GPU-based supercomputers and AI-hardware accelerators utilizing 1.63 Zettaflops in training runs with a sustained performance of 121 PFLOPS in mixed precision and peak of 850 PFLOPS. We present initial scientific insights from examining GenSLMs **in tracking evolutionary dynamics** of SARS-CoV-2, paving the path to realizing this on **large biological data**.

Nr. VPP-EM-BIOMEDICĪNA-2022/1-0001

Goal: Create national research infrustructure platform for biomedicine, which is oriented to improve health issues in specilised fields “Biomedicine, med tech, biopharma and biotechnology”.

UNIVERSITAS LATVIENSIS
SCIENTIAE ET PATRIAE

LV
BMC

RĪGAS STRADIŅA
UNIVERSITĀTE

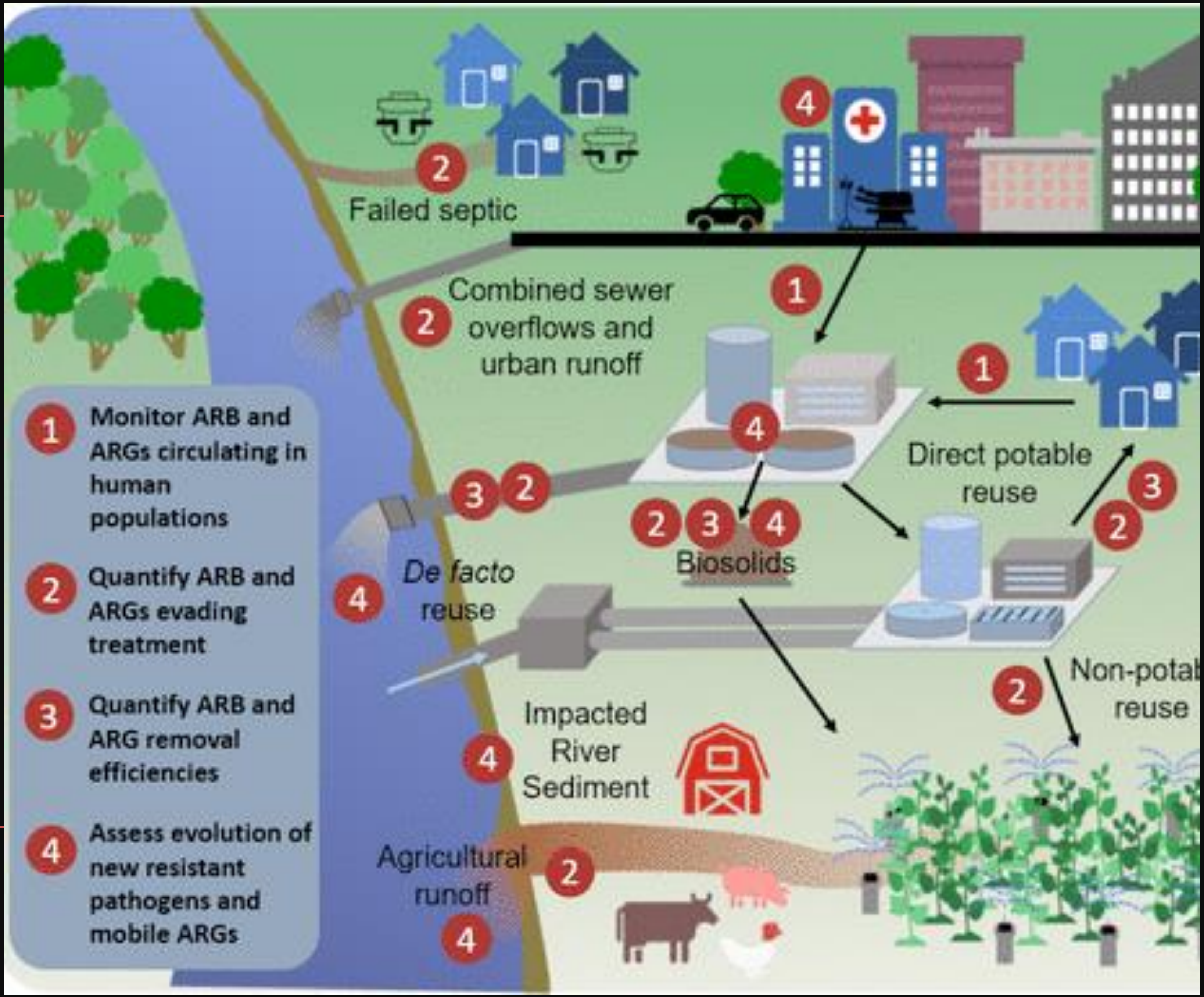
VPP
Valsts pētījumu
programma

BIOR
PĀRTIKAS DROŠĪBAS, DZĪVNIEKU VESELĪBAS
UN VIDES ZINĀTNISKAIS INSTITŪTS

1862
RĪGAS TEHNISKĀ
UNIVERSITĀTE

Experimental setup for AR monitoring

- 3 samples every season from 3 years (108 samples a year)
  - Hospitals.
  - City wastewater.
  - Environmental samples.
  - At least 6 animal farms.
- NGS sequencing
- 90 million sequencing reeds per sample required
- ARG detection
- Future variant prediction
- Report on ARG prevalence for healthcare and government
- Support new technology development


Failed septic
Combined sewer overflows and urban runoff
Direct potable reuse
Biosolids
De facto reuse
Non-potab reuse
Impacted River Sediment
Agricultural runoff
1 Monitor ARB and ARGs circulating in human populations
2 Quantify ARB and ARGs evading treatment
3 Quantify ARB and ARG removal efficiencies
4 Assess evolution of new resistant pathogens and mobile ARGs

Stakeholders

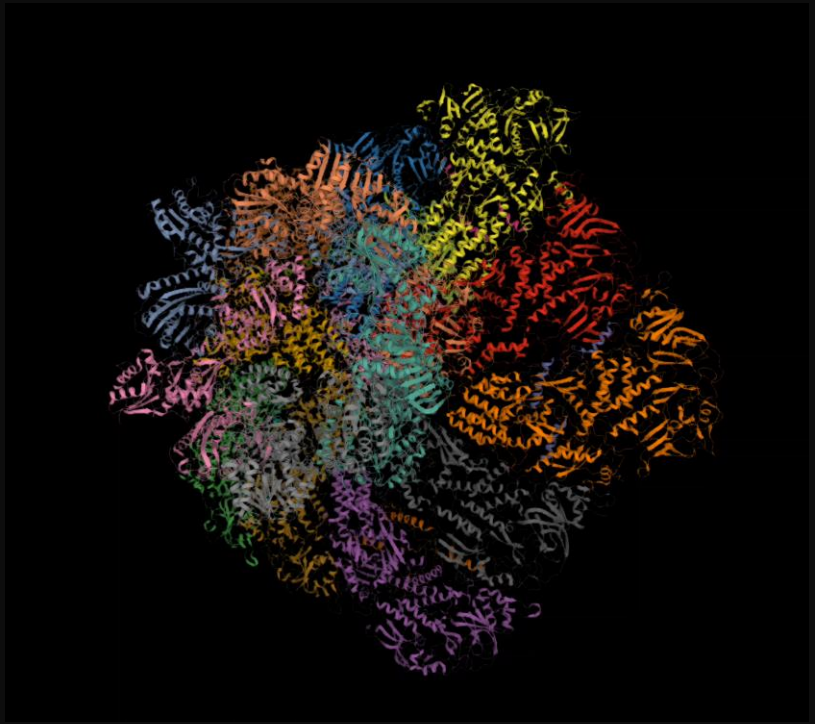
- Academia (LU, RSU, RTU, BMC,)
- Pharma (Pfizer, Roche, Novartis)
- Computing (RTU HPC, LU, EU HPC, Nvidia, Google, Amazon)
- Healthcare (RSU, local hospitals, testing labs)
- Governmental (Latvia gov., SPKC, EU, WHO)

Progress

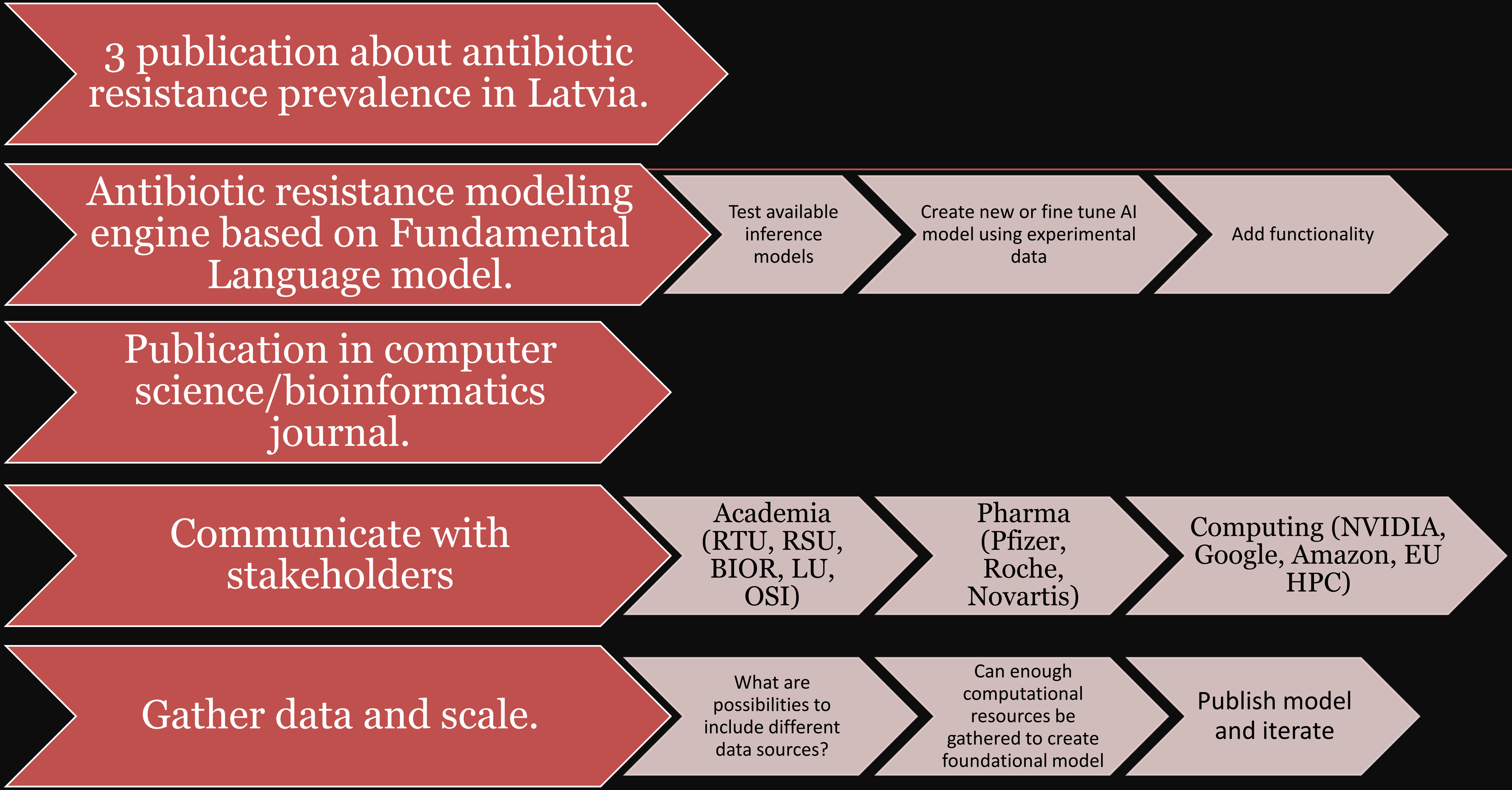
- Collaborated on a submitted publication: *Application of machine learning methods in prognosis of type 2 diabetes mellitus clinical outcomes in Latvian population* ***Journal of Diabetes Science and Technology`***
- Overview of a literature
- First testing done with ML tools
- Resistance detection from a historical data,

wastewater samples.

- Hospital sequencing data in next 1-2 weeks.
- Computational resources (RTU-HPC, Nvidia)
- Protein structures and metagenomes.



Plan



3 publication about antibiotic resistance prevalence in Latvia.

Antibiotic resistance modeling engine based on Fundamental Language model.

- Test available inference models
- Create new or fine tune AI model using experimental data
- Add functionality

Publication in computer science/bioinformatics journal.

Communicate with stakeholders

- Academia (RTU, RSU, BIOR, LU, OSI)
- Pharma (Pfizer, Roche, Novartis)
- Computing (NVIDIA, Google, Amazon, EU HPC)

Gather data and scale.

- What are possibilities to include different data sources?
- Can enough computational resources be gathered to create foundational model
- Publish model and iterate

References

1. An overview of the antimicrobial resistance mechanisms of bacteria https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6604941/
2. Van Camp, P.-J.; Haslam, D.B.; Porollo, A. Bioinformatics Approaches to the Understanding of Molecular Mechanisms in Antimicrobial Resistance. Int. J. Mol. Sci. 2020, 21, 1363. https://doi.org/10.3390/ijms21041363
3. GenSLMs Genome-scale language models reveal SARS-CoV-2evolutionary dynamics doi: https://doi.org/10.1101/2022.10.10.511571
4. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0401-z
5. Pal, Chandan. (2017). Effects of biocides and metals on antibiotic resistance: a genomic and metagenomic perspective. 10.13140/RG.2.2.27592.72967