Predicting Gene Regulatory Networks with Augusta

ENBIK 2025

Karel Sedlar, Jana Musilova et al.

10.06.2025

Augusta

#BioSys_BU

- an open-source Python package for Gene Regulatory Network (GRN) and Boolean Network (BN) inference from the high-throughput gene expression data
- designed for non-model bacteria

ELSEVIER

Computational and Structural Biotechnology Journal

Volume 23, December 2024, Pages 783-790

Augusta: From RNA-Seq to gene regulatory networks and Boolean models

Augusta Ada Byron
English mathematician
and the first programmer

Augusta and Jana

#BioSys_BU

- an open-source Python package for Gene Regulatory Network (GRN) and Boolean Network (BN) inference from the high-throughput gene expression data
- designed for non-model bacteria

ELSEVIER

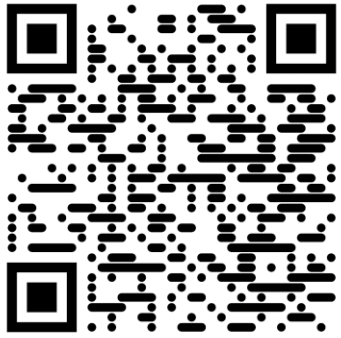
Computational and Structural Biotechnology Journal

Volume 23, December 2024, Pages 783-790

Augusta: From RNA-Seq to gene regulatory networks and Boolean models

Jana Musilova [a b], Zdenek Vafek [b c], Bhanwar Lal Puniya [b], Ralf Zimm

Karel Sedlar



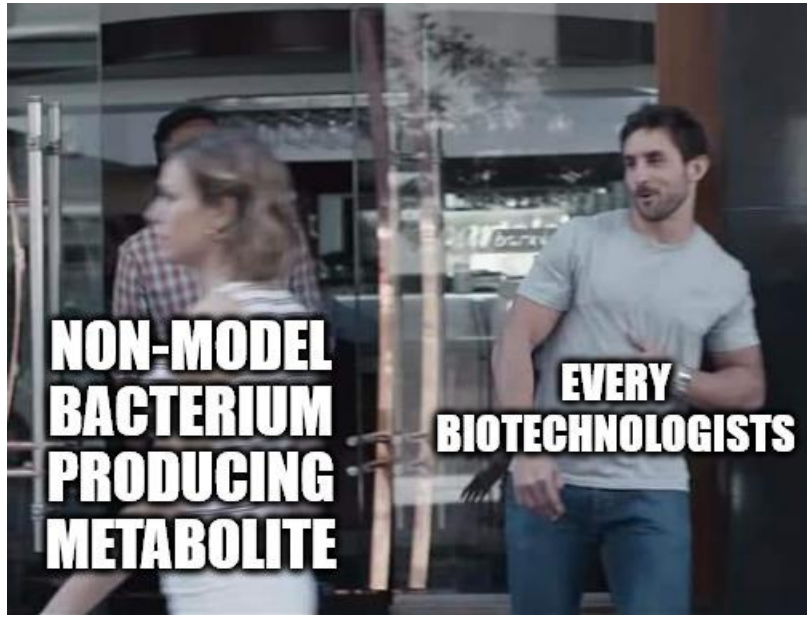
Augusta Ada Byron
English mathematician
and the first programmer

Jana Musilova
Czech bioinformatician
and my first PhD student

Non-conventional bacteria

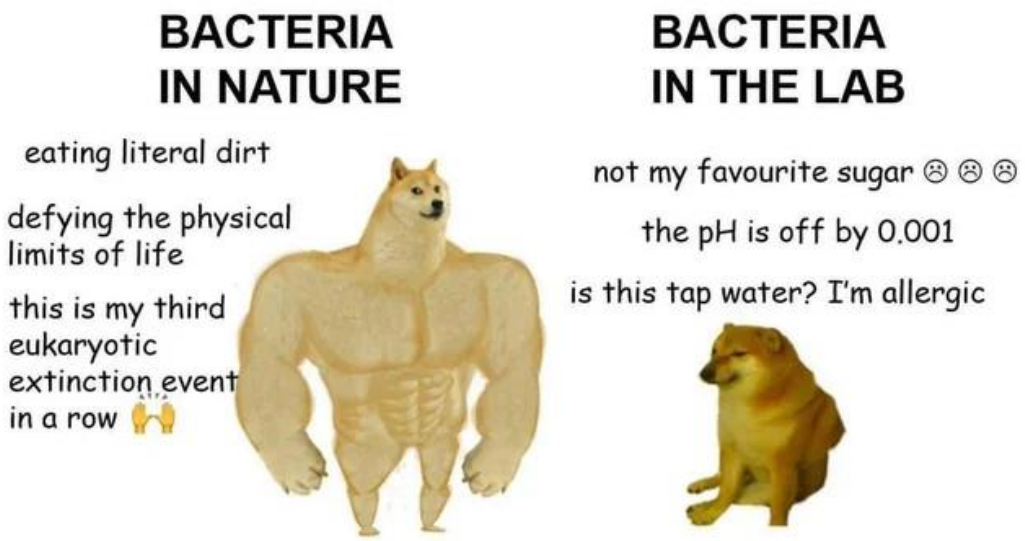
- omnipresent, literally everywhere

- the most abundant group of organisms on planet Earth: 5 nonilions (5.10^{30})
- 200 000 genera → several million species
- our cultivation capabilities: 99%
- available lab technique to study their genes:
  - bulk DNA- and RNA-Seq
  - single cell DNA- and RNA-Seq
  - ChIP-Seq
  - GRIL-Seq, RIL-Seq
  - WHATEVER-Seq



Non-conventional bacteria

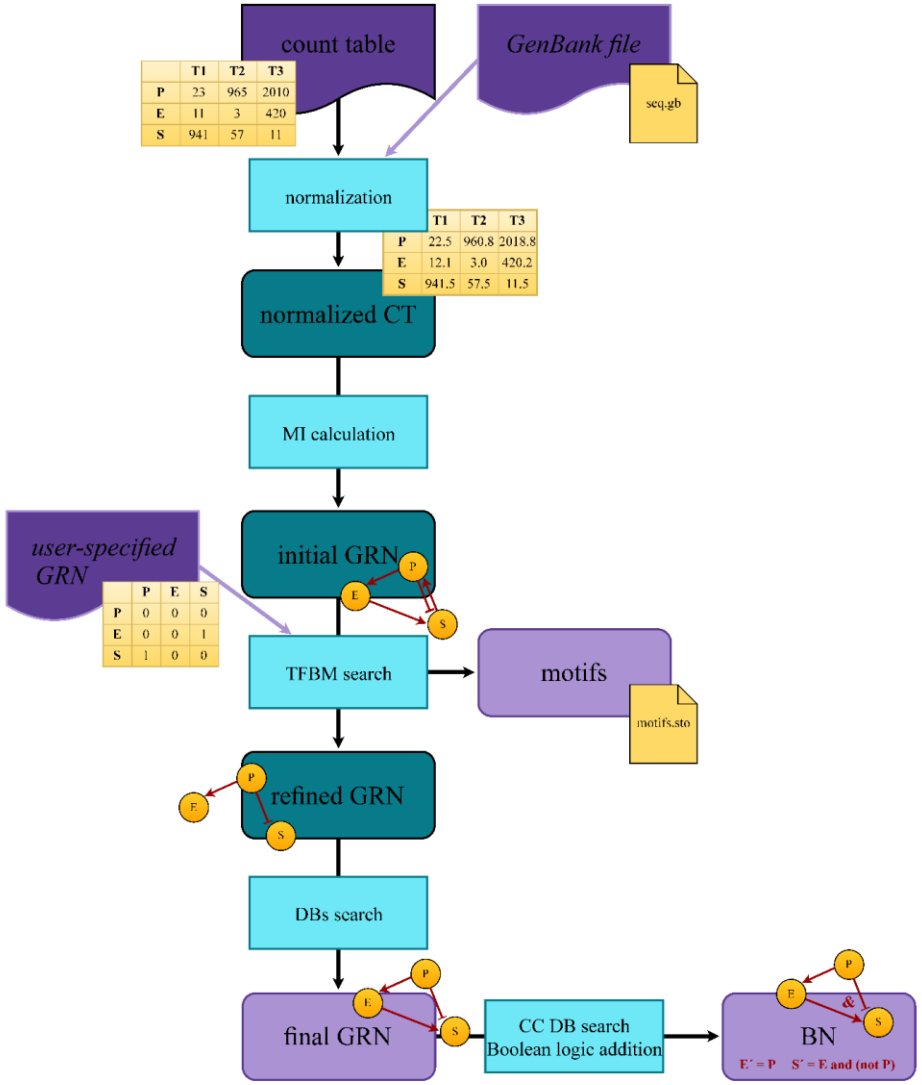
- omnipresent, literally everywhere except for the cultivation flask
- the most abundant group of organisms on planet Earth: 5 nonilions (5.10^{30})
- 200 000 genera → several million species
- our cultivation capabilities: 99% uncultivable!
- available lab technique to study their genes:
  - bulk DNA- and RNA-Seq
  - single cell DNA- and RNA-Seq
  - ~~Chip-Seq, ATAC-Seq~~
  - WHATEVER-Seq



Augusta

- presumes bulk RNA-Seq as input – normalized count table
- initial inference of a GRN with mutual information
- polishing the network with motif search
- polishing the network with database search

- optional: transformation into a Boolean network



Initial GRN

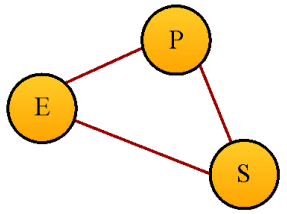
- entropy for a discrete random variable X, i.e., gene X in the input count table:

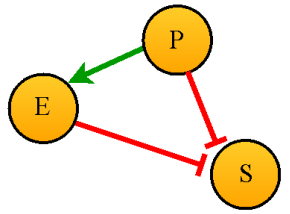
$$H(X) = -\sum_{X \in x} P(x) \log_b P(x) \ (1)$$

	T1	T2	T3	T4
P	40	960	980	999
E	25	60	280	272
S	322	321	348	12

→

	T2-T1	T3-T2	T4-T3
P	920	20	19
E	35	220	-8
S	-1	27	-336





- joint entropy:

$$H(X,Y) = -\sum_{X \in x} \sum_{Y \in y} P(x,y) \log_b P(x,y) \ (2)$$

- mutual information:

$$MI(X;Y) = \sum_{X \in x} \sum_{Y \in y} P(x,y) \log_b \frac{P(x,y)}{P(x)P(y)} = H(X) - H(X|Y) = H(Y) - H(Y|X) \ (3)$$
$$= H(X) + H(Y) - H(X,Y)$$

- binning of original expression levels:

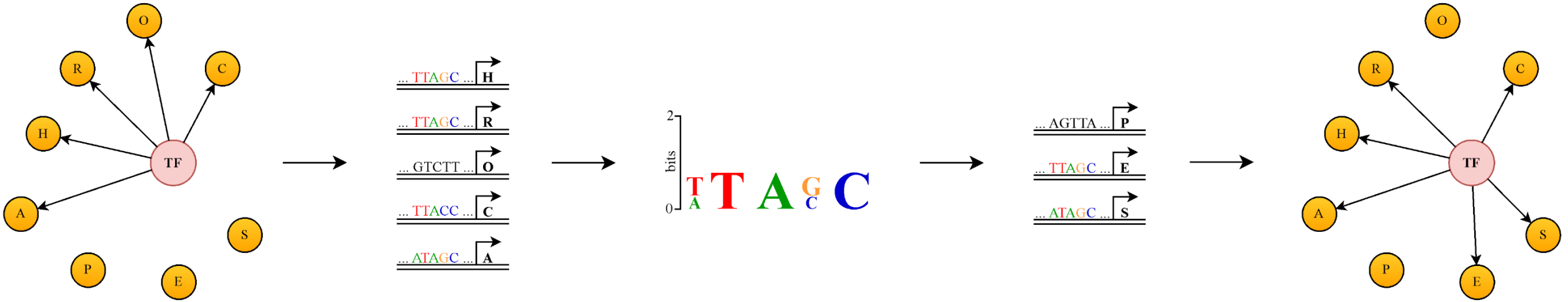
$$D = min\left(\left\lfloor \sqrt{n/5} \right\rfloor, 10\right) \ (4)$$

- edges:

$$e = \begin{cases} (v_1, v_2) \ if \ i < j \\ (v_2, v_1) \ if \ i > j \end{cases}, i = \underset{x \in (1,n)}{\arg\max}\left(\left|DM_{1,x}\right|\right), j = \underset{x \in (1,n)}{\arg\max}\left(\left|DM_{2,x}\right|\right) \ (5)$$

GRN refinement

- before searching for motifs in a network (statistically significant subgraphs), let‘s use sequece motifs in promotor
  - MEME suite

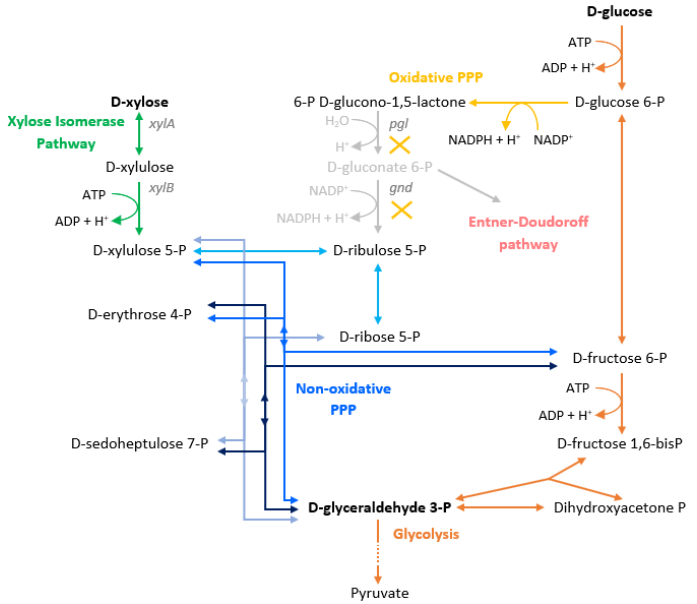


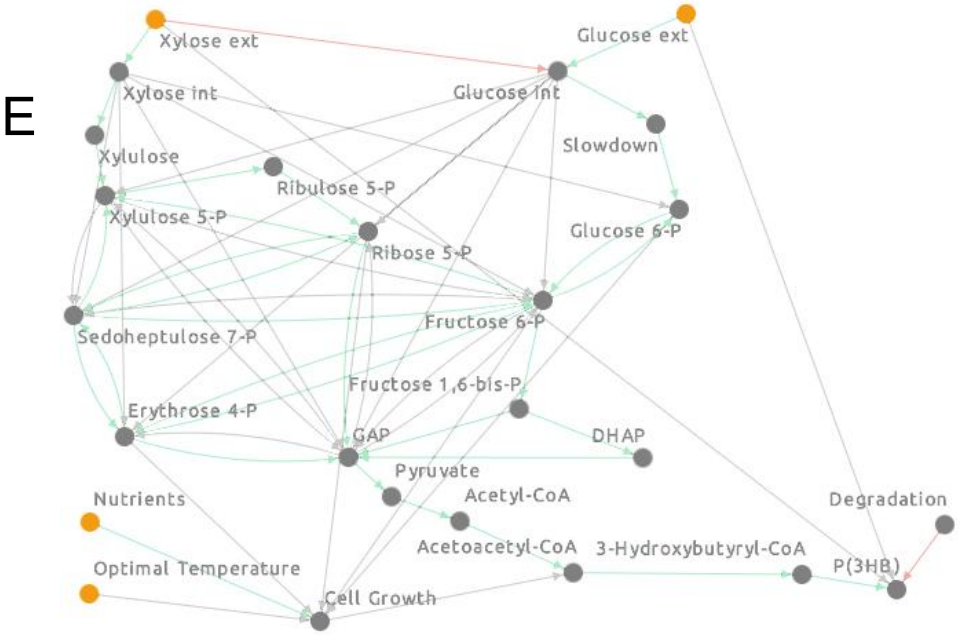
- then look into databases if interaction of particular genes are known:
  - OmniPath, Signor, SignaLink, and TRRUST

Boolean network (BN)

#BioSys_BU

- two ways:
  - by using already published Boolean functions in the Cell Collective (CC)
  - by creating new generic functions
- unlike GRN refinement, where transription fator (TF)-oriented, BN inference is target gebe (TG)-oriented
- follow most commonly observed regulation processes:
  - logical OR operator is applied if only negative/positive interactions influence the TG (e.g. A = B or C; D = not (E or F))
  - logical AND operator is applied if both negative and positive edges influence the TG to represent the dominance of the negative regulation (e.g. G = (not H) and I)





Is it precise?

- **not at all!** more like crystalball reading (but that is common for all GRN inference methods, benchmarking is available in Augusta paper)
- *Caldimonas thermodepolymerans* DSM 15344
  - 6 time points: 0h, 6h, 18h, 36h, 42h, 66h
  - 3,650 nodes; 1,623 connected
  - 208,507 edges
  - 61,7h computational time (Intel Xeon Gold 6128 @ 3.40 GHz, 8 cores, 64 GB RAM)

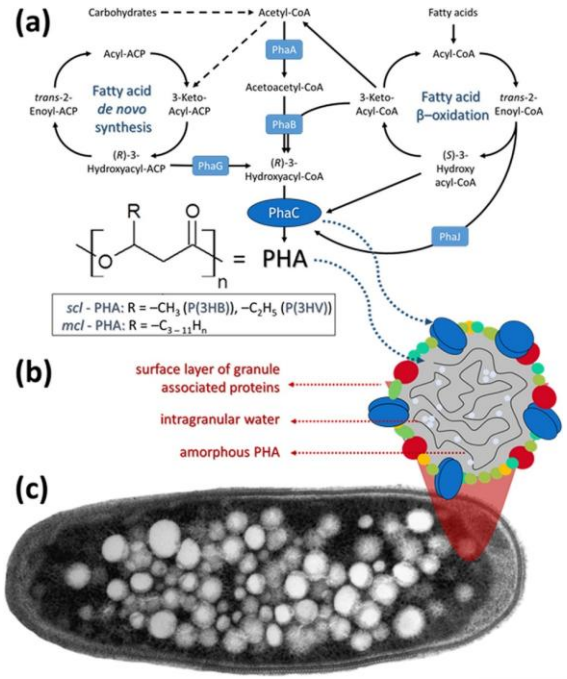
Is it uselful?

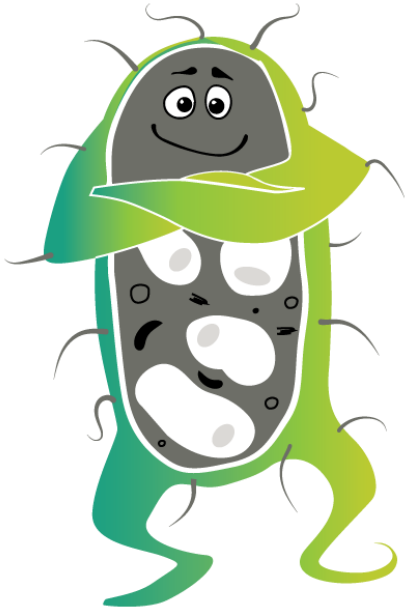
- **it is!** bias is expected in bulk RNA-Seq
  - cell cycle and other signalling processing are running all the time and are not synchronized among cells
  - on the other hand, significant regulations usable to produce value added chemicals are synchronized

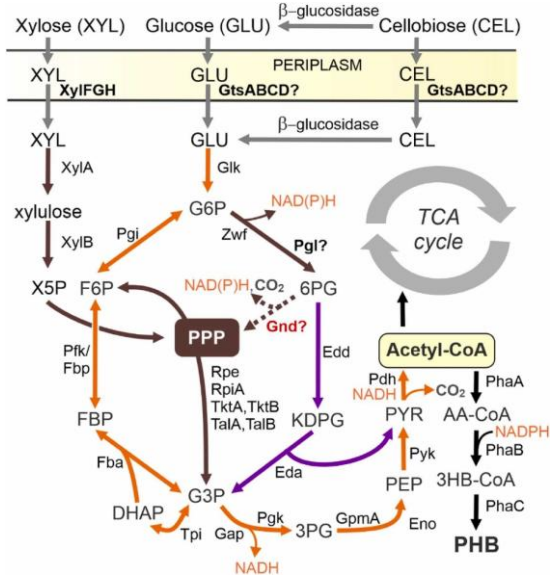
- *Caldimonas thermodepolymerans* DSM 15344

Czech collection strain
nodes: 3,650
  regulators: 1,665
  regulated: 543
edges: 184,514
    activation (+1): 107,030
    inhibition (-1): 77,484

German collection strain
nodes: 3,650
  regulators: 631
  regulated: 365
edges: 63,770
    activation (+1): 30,782
    inhibition (-1): 32,988







Is it FAIR?

- Augusta was FAIRified – to meet FAIR-RS metrics
- supported by FAIR-IMPACT grant under the Assessing the FAIRness of Software support action

eosc
FAIR-IMPACT

Home / Implementation & Adoption Stories /
FAIRification of Augusta, Research Software for Gene Regulatory Networks and Boolean Models Inference

- Findable:
  - improved readme, new metadata, creators, citation
- Accessible:
  - PyPi repository, GitHub
- Interoperable:
  - data formats are described and open
  - a reference to the schema is provided in extended documentation
- Reusable:
  - metadata according to community standards
  - improved machine-actionability

```
> pip install Augusta
```

Who We Are

#BioSys_BU

- a young dynamic team established in 2024
- building on more than 15 years of bioinformatics tradition at UBMI
- oriented mainly (not exclusively) to microbial world
- open to networking, collaboration, and joint projects

VŠCHT PRAHA
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA CHEMICKÁ
GeneProof Molecular diagnostics for your routine
BioVendor Group
MENDELU
Hes·so VALAIS WALLIS
MUNI
University of Nebraska Lincoln
FAKULTNÍ NEMOCNICE BRNO
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
PAQUES biomaterials

- successful (sometimes) in obtaining grant funding

eosc FAIR-IMPACT
GAČR CZECH SCIENCE FOUNDATION
Standard Projects
JUNIOR STAR
AZV

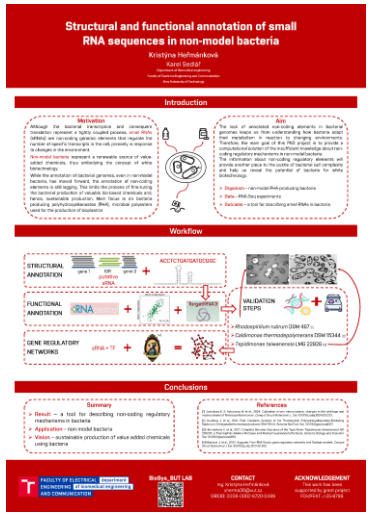
BRNO UNIVERSITY OF TECHNOLOGY
BioSys_BUT

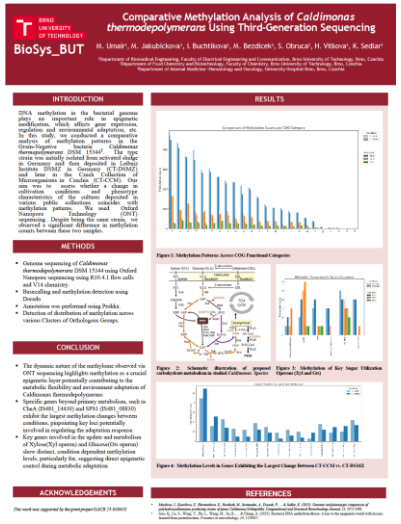
Check our work at ENBIK

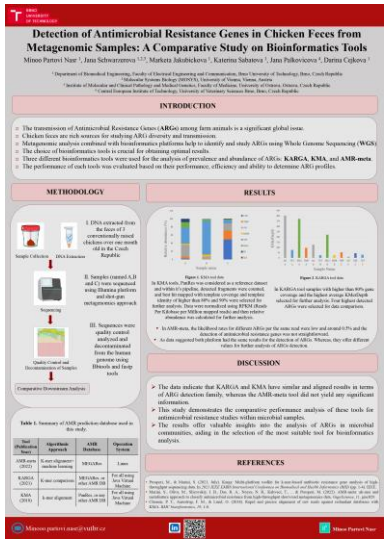
#BioSys_BU

- presentations:
  - **Darina Čejková:** Plasmid-Mediated Antibiotic Resistance Dynamics in Broiler Chickens Revealed by Long-Read Sequencing
  - **Markéta Jakubíčková:** Streamlined workflow for bacterial methylation analysis using nanopore data

- posters:
  - **Inderjeet Bhogal:** Probing the interactions between Ipragliflozin with RAGE for treating Alzheimer's disease: An in-silico drug repurposing approach
  - **Mohammad Umair:** Methylome Profiling Using Third-Generation Sequencing: A Comparison of PacBio and ONT in a PHA-Producing Bacterium
  - **Vaishali Pankaj :** An integrated computational strategy to identify selective HDAC6 inhibitors against breast cancer
  - **Kateřina Šabatová:** User-friendly web tool for typing and characterization of ESKAPEE pathogens
  - **Helena Vítková:** Consensus-Based Detection of Biosynthetic Gene Clusters with Application to RiPPs from Antarctic Bacteria
  - **Jana Musilová:** FAIRification of Augusta, a Python package for RNA-Seq-Based Inference of Gene Regulatory and Boolean Networks







Predicting Gene Regulatory Networks with Augusta

BRNO UNIVERSITY OF TECHNOLOGY
BioSys_BUT

JUNIOR STAR

Supported by the grant project GACR 25-17459M