SysBioBig

Thesis Proposal

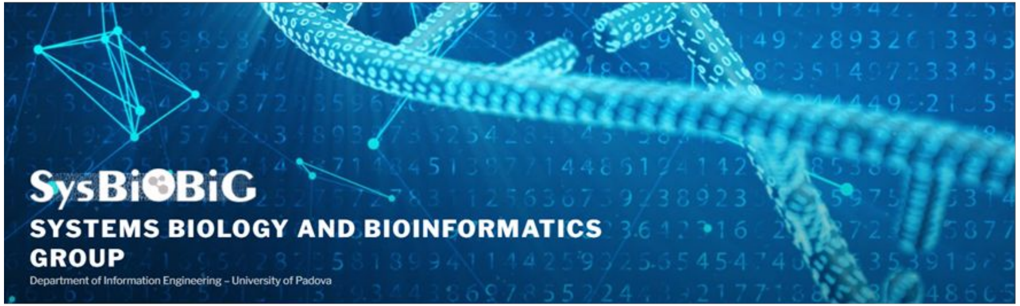
2nd March, 2026

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

SysBiOBiG

The research group



We work in the field of predictive ***modelling***, ***data mining***, ***machine learning***, ***computer science*** methods for the analysis of high-throughput ***omics*** and ***clinical data***, dealing with data ***complexity***, ***heterogeneity***, and ***dynamics***.

PI: **Barbara Di Camillo**
Contact:
barbara.dicamillo@unipd.it

https://sysbiobig.dei.unipd.it/

GitLab https://gitlab.com/sysbiobig

SysBioBiG - System Biology and Bioinformatics Group at UNIPD

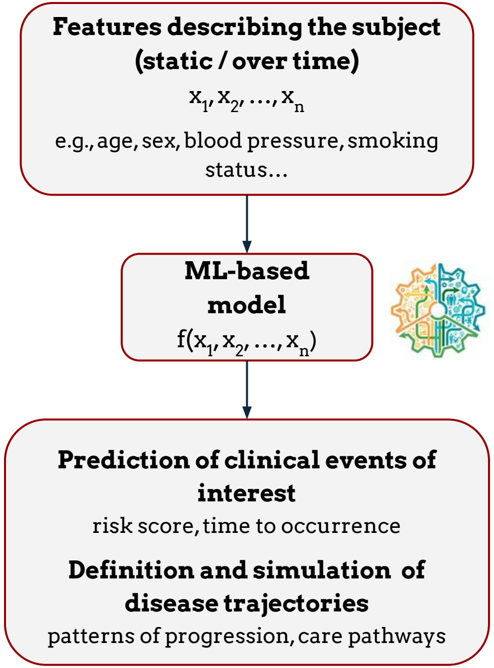
@sysbiobigunipd

Modeling and simulation of clinical trajectories using AI

Objective

Develop data-driven models to **simulate disease progression trajectories** from longitudinal clinical data and **generate** realistic synthetic clinical populations.

Our main expertise lies in **dynamic Bayesian networks**, probabilistic graphical models able to capture relationships among variables over time. We aim to explore alternative approaches for clinical time-series simulation, such as **Generative Adversarial Networks (GANs)**.



Features describing the subject (static / over time)

$x_1, x_2, \ldots, x_n$

e.g., age, sex, blood pressure, smoking status…

ML-based model

$f(x_1, x_2, \ldots, x_n)$

Prediction of clinical events of interest

risk score, time to occurrence

Definition and simulation of disease trajectories

patterns of progression, care pathways

Skill requirements:

- Machine learning and deep learning fundamentals
- Basic programming skills in R and/or Python

Preferred background:

- Computer engineering, Bioengineering
- AI, Data science

References/useful resources:

Tavazzi et al, 2022: Predicting functional impairment trajectories in amyotrophic lateral sclerosis: a probabilistic, multifactorial model of disease progression. Journal of Neurology. 10.1007/s00415-022-11022-0

Nasimov et al, 2024. GAN-Based Novel Approach for Generating Synthetic Medical Tabular Data. Bioengineering. 10.3390/bioengineering11121288

Contacts: barbara.dicamillo@unipd.it
erica.tavazzi@unipd.it

Process Mining for Modelling Complex Health & Industrial Processes

Objective

Use **Process Mining** to reconstruct, analyse, and model **real-world processes from event data**, capturing sequences, timing, and variability of complex trajectories.

Case studies

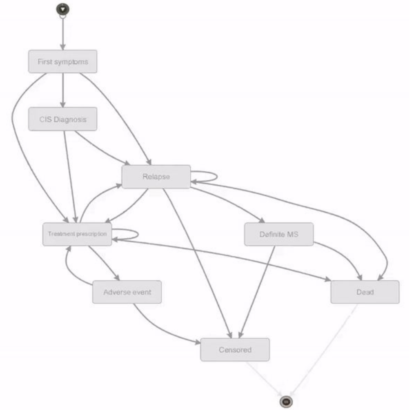
- **Clinical processes** - modeling of trajectories of disease evolution to identify typical progression patterns and dynamic phenotypes
- **Industrial processes** - modeling of laboratory workflows for antibiotic resistance detection

Skill requirements:

- Fundamentals of data analysis
- Interest in temporal data
- Basic programming skills in R and/or Python

Preferred background:

- Computer engineering, Bioengineering
- Data science



References/useful resources:

Rojas, Eric, et al., 2016: Process mining in healthcare: A literature review. Journal of biomedical informatics. 10.1016/j.jbi.2016.04.007

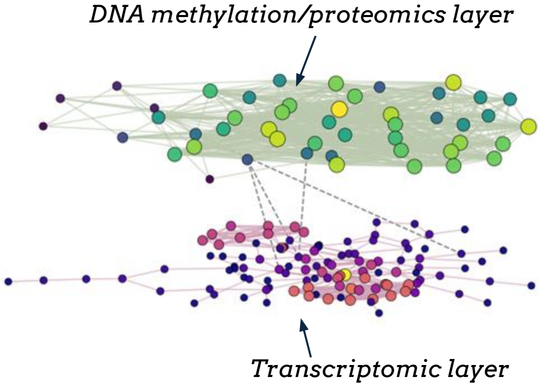
Trescato et al, 2024: DYNAMITE: Integrating Archetypal Analysis and Process Mining for Interpretable Disease Progression Modelling. IEEE JBHI. 10.1109/JBHI.2024.3453602

Contacts: barbara.dicamillo@unipd.it
erica.tavazzi@unipd.it

Integrating different -Omics into Patient-Specific Gene Correlation Networks through a Multilayer Approach

Objective: The project investigates how **integrating multi-omics data** into a pre-existing **gene-correlation network** modulates the weights of gene–gene interactions (edges) by constructing and comparing patient-specific networks through **optimal transport.**

Case study: Colorectal cancer data



Skills requirements:

- Proficiency in R or Python (Pandas, Scipy, NetworkX)
- Basic understanding of graph theory and correlation statistics

Preferred Background:

- Computer engineering
- Bioengineering
- Computational Biology

References:

- Hammoud, Z., Kramer, F. Multilayer networks: aspects, implementations, and application in biomedicine. Big Data Analytics 5, 2 (2020). https://doi.org/10.1186/s41044-020-00046-0
- Silverman EK, Schmidt HHHW, Anastasiadou E, et al. Molecular networks in Network Medicine: Development and applications. WIREs Syst Biol Med. 2020; 12:e1489. https://doi.org/10.1002/wsbm.1489

Contacts:
barbara.dicamillo@unipd.it
giulia.cesaro@unipd.it
elena.marinello.1@studenti.unipd.it

Structural analyses of large multilayer networks with MuxVizPy and Py3Plex

Objective:

We are **actively developing MuxVizPy**, a Python library for structural analyses over **large** multilayer networks.

You will **help extend** a functioning **profiling** framework to **benchmark** its performance against similar software libraries such as Py3Plex.

If interested, **you** will **help extend** the **features** of MuxVizPy.

Skill requirements:

- Proficiency in Python (numpy).
- Basic experience in Bash scripting.
- Being methodical.

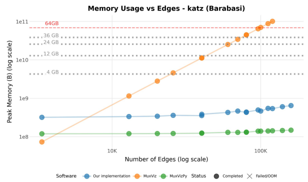
Preferred background:

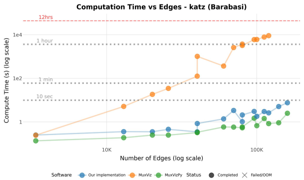
- Computer Engineering

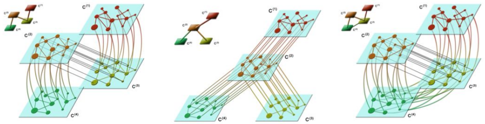
References/useful resources:

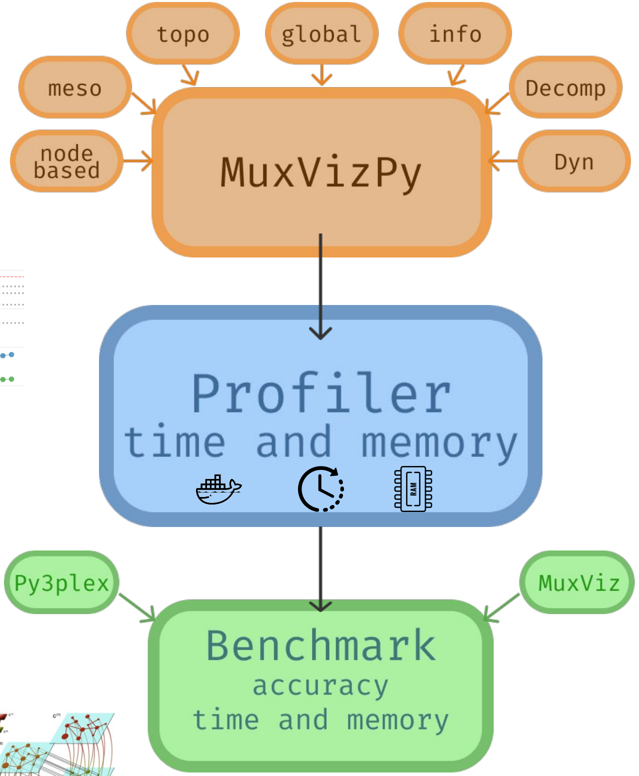
Škrlj, B., Kralj, J. & Lavrač, N. Py3plex toolkit for visualization and analysis of multilayer networks. [link]

Manlio De Domenico, Mason A. Porter, Alex Arenas, MuxViz: a tool for multilayer analysis and visualization of networks, Journal of Complex Networks, 2015. [link]









topo

global

info

meso

Decomp

node based

Dyn

MuxVizPy

Profiler

time and memory

Py3plex

MuxViz

Benchmark

accuracy

time and memory

Contacts: barbara.dicamillo@unipd.it
matteo.baldan.7@phd.unipd.it

Border-Like spatial niche detection with Cellular Communication and Multilayer Networks.

Objective:

Contribute to **explore** a mathematical framework to **detect border-like spatial niche** in biological tissues, based on the way **cells interact** with each other.

Goals:

You will understand what is **spatial transcriptomics**.

You will work on a **models** to describe **interactions** within a biological tissue.

You will work with **high-dimensional** and heavy data.

Skill requirements:

- Basics of biology.
- Proficiency in Python.
- Basics of ML/DL.

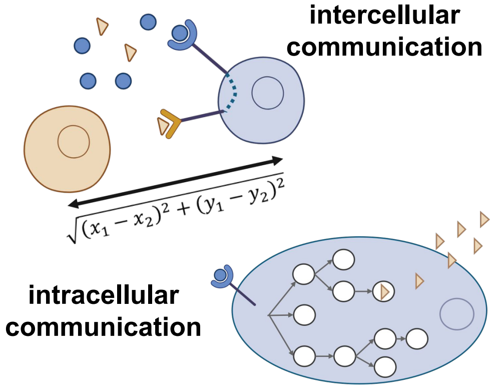
Preferred background:

- Bio/computer engineering

References/useful resources:

Birk, S., Bonafonte-Pardàs, I., Feriz, A.M. et al. Quantitative characterization of cell niches in spatially resolved omics data. Nat Genet 57, 897–909 (2025). Link

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., & Arenas, A. (2013). Mathematical Formulation of Multilayer Networks. Physical Review X. Link

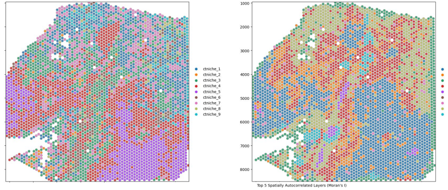

intercellular communication

$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

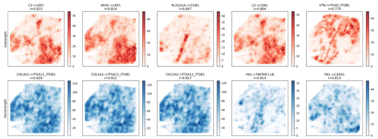
intracellular communication

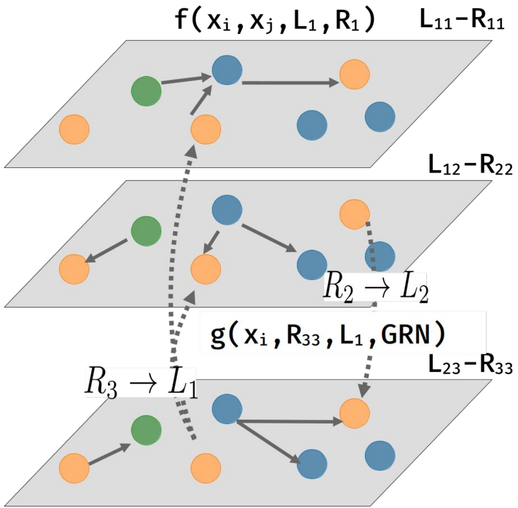
cell type

niche

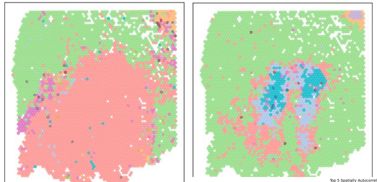


relevant spatial communication ways



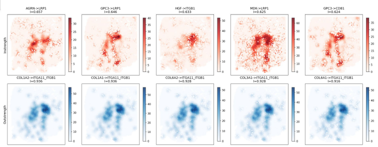


niche



cell type

relevant spatial communication ways



Contacts: barbara.dicamillo@unipd.it
matteo.baldan.7@phd.unipd.it

Candidate drug identification from scRNAseq data

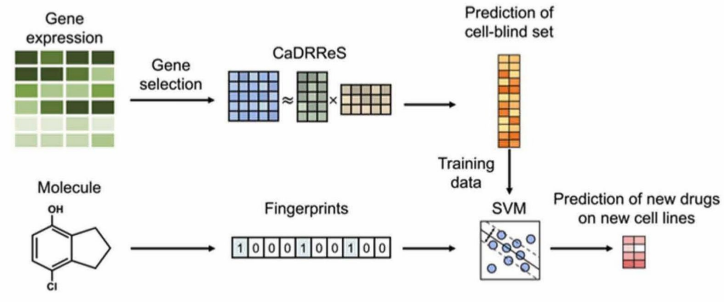
Objective:
In scRNAseq research a new promising research field is emerging: the development of computational methods for the identification of candidate drugs from scRNAseq data. This thesis project is focused on a detailed study of the literature on this subject and the application of the tools on one or more disease case studies.

Hard/soft skills requirements:

- Basic programming skills in Python and/or R
- Familiarity with scRNAseq, interest in molecular and chemical subjects

Preferred background:

- Computer engineering/science
- Bioengineering
- Computational Biology



References/useful resources:

Feng J, Goedegebuure SP, Zeng A, et al. sc2MeNetDrug: A computational tool to uncover inter-cell signaling targets and identify relevant drugs based on single cell RNA-seq data. *PLoS Comput Biol*. 2024;20(1):e1011785. Published 2024 Jan 5. doi:10.1371/journal.pcbi.1011785

Schäfer S, Smelik M, Sysoev O, et al. scDrugPrio: a framework for the analysis of single-cell transcriptomics to address multiple problems in precision medicine in immune-mediated inflammatory diseases. *Genome Med*. 2024;16(1):42. Published 2024 Mar 20. doi:10.1186/s13073-024-01314-7

Sun YY, Hsieh CY, Wen JH, et al. scDrug+: predicting drug-responses using single-cell transcriptomics and molecular structure. Biomed Pharmacother. 2024;177:117070. doi:10.1016/j.biopha.2024.117070

Dann E, Teeple E, Elmentaite R, et al. Estimating the impact of single-cell RNA sequencing of human tissues on drug target validation. Preprint. https://doi.org/10.1101/2024.04.04.24305313

Contacts: barbara.dicamillo@unipd.it
gaia.tussardi@phd.unipd.it

Bivariate statistic extension to account for multiple covariates in genotype-phenotype association

Objective:

Develop a **methodology to integrate multiple covariates** into the **NEBULA bivariate entropy-based statistic**, enabling covariate-adjusted **genotype–phenotype association** analysis and **improving statistical power** while **reducing false positives** in rare variant detection.

Skill requirements:

- Familiarity with **C / C++**
- Proficiency in **R** and/or **Python** for statistical modeling

Preferred background:

- Computer engineering/science
- Bioengineering

$$S_2(snp_A, snp_B) = \sum_{g=1}^{9} \frac{H_0 - \boxed{H_g}}{H_0} \cdot F_g$$

$$\boxed{H_g} = -\sum_{c} \frac{f_{cg}}{\sum_c f_{cg}} \log_2\left(\frac{f_{cg}}{\sum_c f_{cg}}\right)$$

CONTINGENCY TABLE FOR GENOTYPE CONFIGURATION OF A SNP PAIR.

	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
C_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}
C_2	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	f_{27}	f_{28}	f_{29}

References/useful resources:

Di Camillo B, Sambo F, Toffolo G, Cobelli C. ABACUS: an entropy-based cumulative bivariate statistic robust to rare variants and different direction of genotype effect. Bioinformatics. 2014 Feb 1;30(3):384-91. 10.1093/bioinformatics/btt697

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):59. 10.1038/s43586-021-00056-9

How can covariates influence the genotype entropy?

Contacts: barbara.dicamillo@unipd.it
mikele.milia@phd.unipd.it

Synthetic generation of phenotype-genotype contingency matrices for null model optimization

Objective:

Develop a framework that uses SNP parameters to generate **synthetic genotype–phenotype contingency tables**, avoiding direct computation from raw genetic data for **null model construction**. Evaluate it against **NEBULA** in terms of computational efficiency, memory usage, and accuracy on large-scale genomic data.

Allele frequencies
$p, q = 1 - p$

Genotype frequencies:
$P(AA)=p^2, P(Aa)=2pq, P(aa)=q^2$

Phenotype distribution
case / control ratios

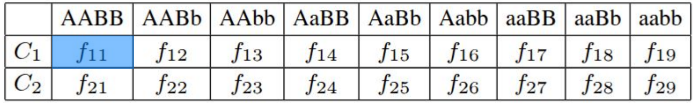
Linkage disequilibrium
D', r^2

Skill requirements:

- Familiarity with **C / C++**
- Proficiency in **R** and/or **Python** for statistical modeling

Preferred background:

- Computer engineering/science
- Bioengineering



CONTINGENCY TABLE FOR GENOTYPE CONFIGURATION OF A SNP PAIR.

	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
C_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}
C_2	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	f_{27}	f_{28}	f_{29}

How can we compute a null model using only few parameters? Is it possible to infer exact frequencies?

References/useful resources:

Di Camillo B, Sambo F, Toffolo G, Cobelli C. ABACUS: an entropy-based cumulative bivariate statistic robust to rare variants and different direction of genotype effect. Bioinformatics. 2014 Feb 1;30(3):384-91. 10.1093/bioinformatics/btt697

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):59. 10.1038/s43586-021-00056-9

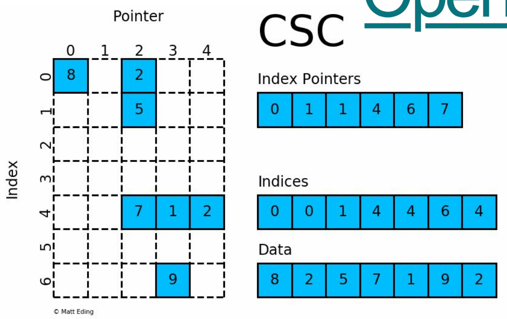
Contacts: barbara.dicamillo@unipd.it
mikele.milia@phd.unipd.it

Memory and time efficient representation and computation on sparse matrix

Objective:

Implement computationally **efficient operation** on large-scale **sparse matrices** for a R library, leveraging C++ for **efficient data structures** and **parallel computation**.

- Develop **efficient** C++ implementations for computationally intensive **operations on matrices**, possibly in **parallel (OpenMP)**, stored in compressed sparse column (CSC) object.
  - e.g. Distance and dissimilarity computations; Cross-product and dot-product matrix computation; Statistical tests
- Design and implement **additional sparse matrix data structures in C++**, including algorithms for conversion between formats
  - e.g. compressed sparse row (CSR), coordinate format (COO), …



Index	0	1	2	3	4
0	8		2		
1			5		
2					
3					
4			7	1	2
5					
6				9	

CSC

Index Pointers

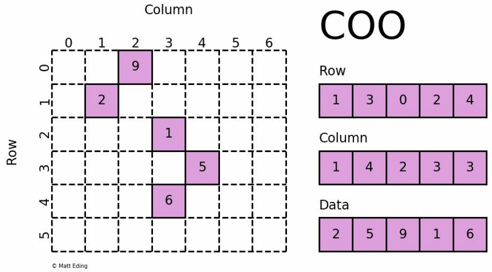
0	1	1	4	6	7

Indices

0	0	1	4	4	6	4

Data

8	2	5	7	1	9	2



Row	0	1	2	3	4	5	6
0			9				
1		2					
2				1			
3					5		
4				6			
5							

COO

Row

1	3	0	2	4

Column

1	4	2	3	3

Data

2	5	9	1	6

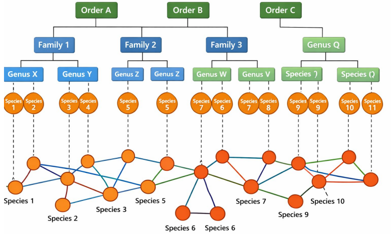
Contacts: giacomo.baruzzo@unipd.it
barbara.dicamillo@unipd.it
giulia.cesaro@unipd.it
piero.mariotto@phd.unipd.it

Memory and time efficient representation and computation on sparse matrix

Objective:

Implement computationally **efficient operation** on large-scale **sparse matrices** for a R library, leveraging C++ for **efficient data structures** and **parallel computation**.

- Redesign the package architecture to support **GPU-based** sparse computation
    - GPU-friendly data structure (Ellpack, Hybrid format ELL+COO) and operation
- Implement a **data structure** that jointly represents a **graph** structure and a **hierarchical dendrogram**, including fast queries and operations based on dendrogram cuts (e.g. TreeSummarizedExperiment package)



Skill requirements:

- Proficiency in C++/C programming

Preferred background:

- Computer engineering/science

References/useful resources:

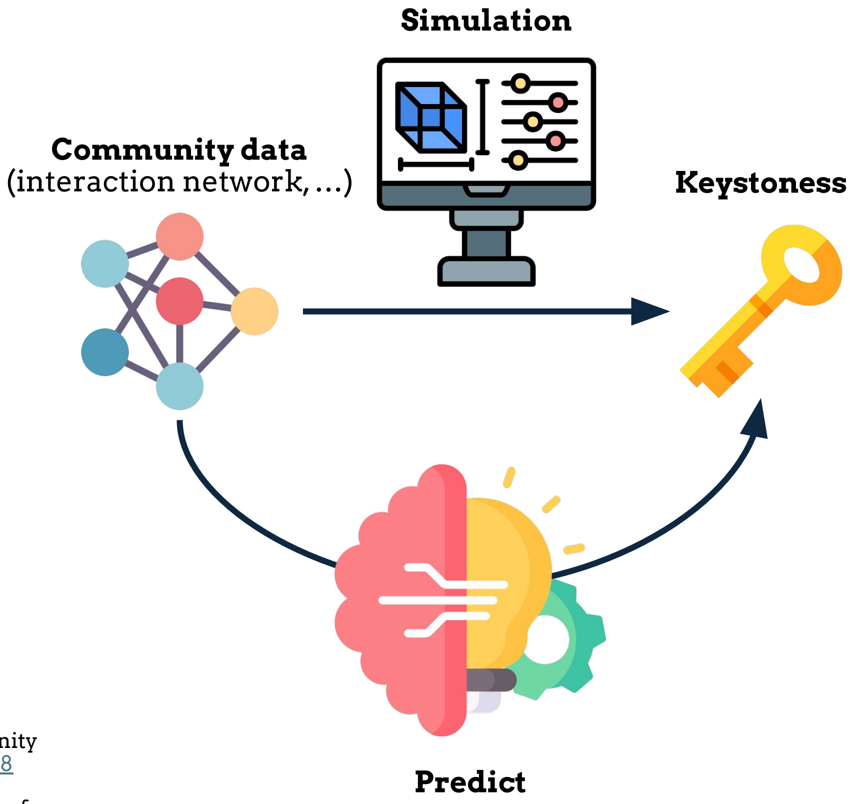
- *G. Baruzzo, G. Cesaro and B. Di Camillo, "quickSparseM: a library for memory- and time-efficient computation on large, sparse matrices with application to omics data," PDP, 2025,* https://doi.org/10.1109/PDP66500.2025.00045
- https://gitlab.com/sysbiobig/quicksparsem
- https://www.rcpp.org/

Contacts: giacomo.baruzzo@unipd.it
barbara.dicamillo@unipd.it
giulia.cesaro@unipd.it
piero.mariotto@phd.unipd.it

Linking topology to keystoneness on synthetic microbial communities

Objective :

Keystoneness measures how important is an individual to the community it inhabits. The thesis entails the generation of large amounts of data using a simulator and **predict (also using ML) the keystoneness** (a.k.a. impact) of the community members from available data



Simulation

Community data
(interaction network, ...)

Keystoneness

Predict

Skill requirements:

- Data analysis
- Programming skills (mainly Python)

Preferred background:

- Computer engineering/science
- Bioengineering

References/useful resources:

Garza, D.R., Gonze, D. & Faust, K. Keystone concept revisited: insights into microbial community dynamics and control. *Nat Rev Microbiol* (2026). https://doi.org/10.1038/s41579-025-01266-8

Marsland R, Cui W, Goldford J, Mehta P (2020) The Community Simulator: A Python package for microbial ecology. PLOS ONE 15(3): e0230430. https://doi.org/10.1371/journal.pone.0230430

Contacts: giacomo.baruzzo@unipd.it
barbara.dicamillo@unipd.it
piero.mariotto@phd.unipd.it

Inference of microbial genomic data: a benchmark of existing tools

Objective :

Sequencing of the whole genome of bacteria is costly and complex. For this reason, many tools have been developed to **infer genome sequences from 16S data**. Our objective is to perform a comprehensive **benchmark of existing methods**.

Skill requirements:

- Careful and systematic development of benchmark pipeline
- Attention to data handling
- Programming skills (Python & R)

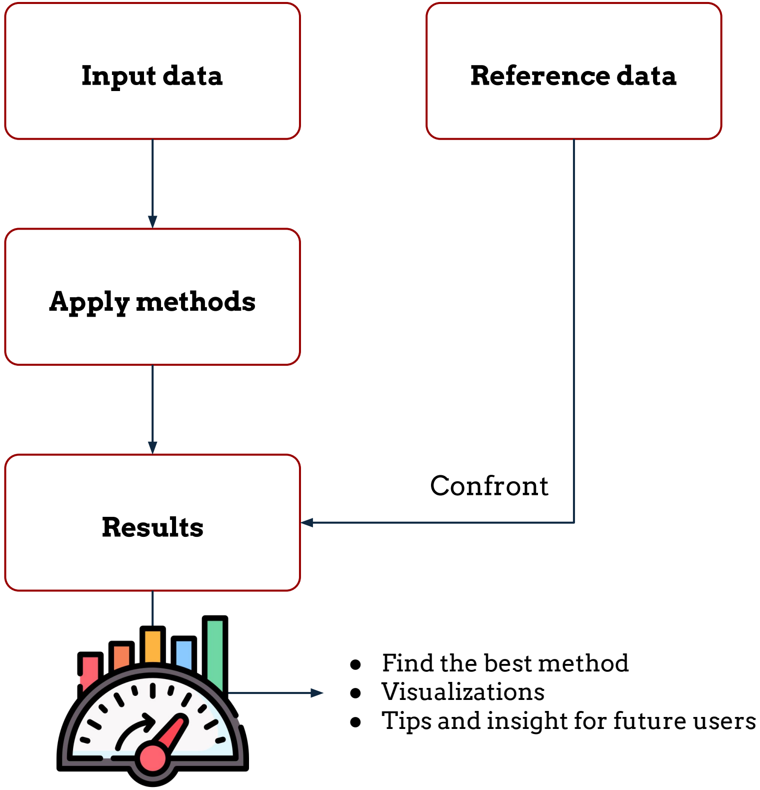
Preferred background:

- Computer engineering/science
- Bioengineering

References/useful resources:

Sun, S., Jones, R.B. & Fodor, A.A. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* **8**, 46 (2020). https://doi.org/10.1186/s40168-020-00815-y

Christophe Djemiel, Pierre-Alain Maron, Sébastien Terrat, Samuel Dequiedt, Aurélien Cottin, Lionel Ranjard, Inferring microbiota functions from taxonomic genes: a review, *GigaScience*, Volume 11, 2022, giab090, https://doi.org/10.1093/gigascience/giab090


Input data

Reference data

Apply methods

Results

Confront

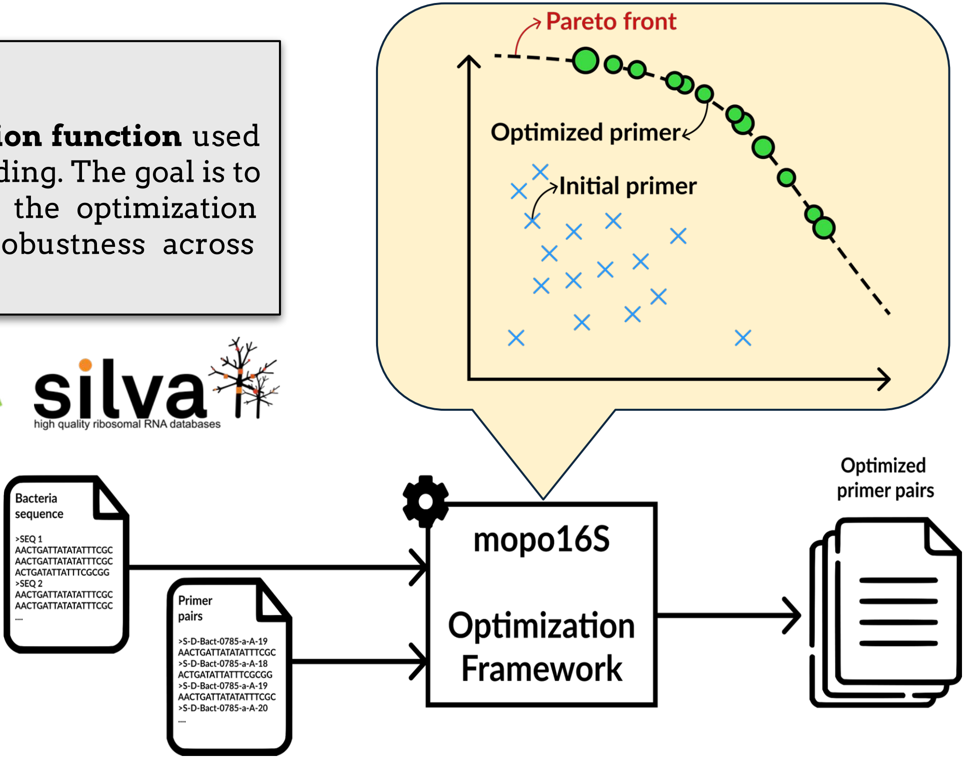
- Find the best method
- Visualizations
- Tips and insight for future users

Contacts: giacomo.baruzzo@unipd.it
barbara.dicamillo@unipd.it
piero.mariotto@phd.unipd.it

Multi-objective scoring for 16S primer design and evaluation (mopo16S) and its related webapp

Objective 1:

Extend and refine the current multi-objective optimization function used in mopo16S to better evaluate primer sets for 16S metabarcoding. The goal is to **integrate biologically meaningful quality metrics** into the optimization process, improving taxonomic coverage, specificity and robustness across microbial communities.



Skill requirements:

- Proficiency in C/C++ programming

Preferred background:

- Computer engineering/science

References/useful resources:

Sambo, F., Finotello, F., Lavezzo, E., Baruzzo, G., Masi, G., Peta, E., ... & Di Camillo, B. (2018). Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. BMC bioinformatics, 19(1), 343.
https://doi.org/10.1186/s40168-020-00815-y

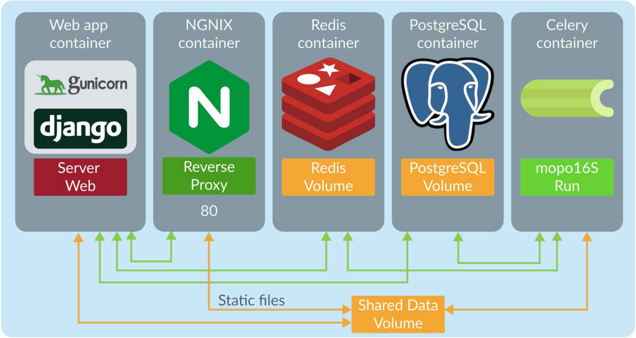
Contacts: giacomo.baruzzo@unipd.it
barbara.dicamillo@unipd.it
mikele.milia@phd.unipd.it

Multi-objective scoring for 16S primer design and evaluation (mopo16S) and its related webapp

Objective 2:

Modernize and extend mopo16Sweb to ensure compatibility with recent Django releases and improve maintainability, scalability, and reproducibility.

- Migrate to **Django 5.2 LTS**
- Improve async job management
- Design new page and functionalities



Skill requirements:

- Python / Django (backend)
- JavaScript (asynchronous frontend, AJAX/fetch)

Preferred background:

- Computer engineering/science


Select the score to display in the 3rd dimension (circles radius).

Forward primers: CCTACGGGNGGCWGCAG
Reverse primers: CGACARCCATGCASCACCT
Efficiency: 8.30331
Coverage: 0.57384
Matching-bias: 0.918295

References/useful resources:

Sambo, F., Finotello, F., Lavezzo, E., Baruzzo, G., Masi, G., Peta, E., ... & Di Camillo, B. (2018). Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. BMC bioinformatics, 19(1), 343. https://doi.org/10.1186/s40168-020-00815-y

Contacts: giacomo.baruzzo@unipd.it
barbara.dicamillo@unipd.it
mikele.milia@phd.unipd.it

Optimization of a Nupack pipeline to design novel cell free biosensors

Objective:

Integrate Nupack tools library to other bioinformatics tool to:

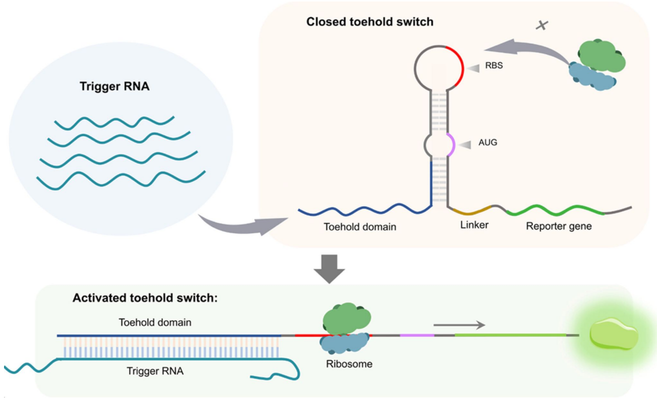
- Create a pipeline for Toehold switch design given a target microbe
- Create a GUI to simplify usage for non-programming people (inspired by the proprietary one)

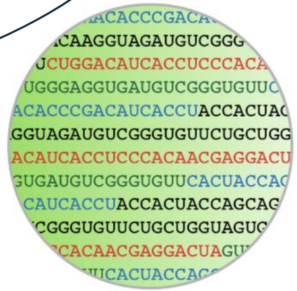
Skill requirements:

- Command line programming skills (Python, SLURM)
- GUI design

Preferred background:

- Computer engineering/science
- Bioengineering





References/useful resources:

Full list of useful papers and original tool available at https://www.nupack.org/cite

Contacts: massimo.bellato@unipd.it
barbara.dicamillo@unipd.it

Explore AI tool for unveiling phage genomics

Objective:

Evaluate different AI and deep learning tools to analyze phage genomic data for:

- protein annotation
- phage clustering
- phage-host interaction

Skill requirements:

- Database crawling
- Data integration
- Programming skills (Python, SLURM)

Preferred background:

- Computer engineering/science
- Bioengineering

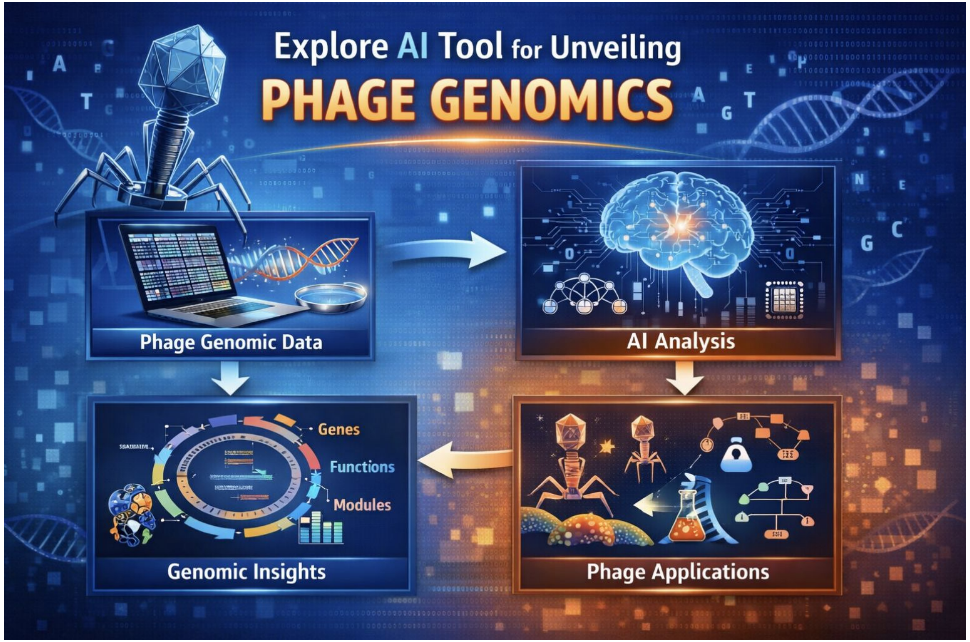


Image created with ChatGPT 5.2

References/useful resources:

Panigrahi, S., et al. HieVi: Protein Large Language Model for proteome-based phage clustering. Biorxiv, (2024) https://doi.org/10.1101/2024.12.17.627486

Chen Q, et al. MoEPH: an adaptive fusion-based LLM for predicting phage-host interactions in health informatics. *Front. Microbiol* **16,** (2025) https://doi.org/10.3389/fmicb.2025.1634705

Zeming Lin *et al.* ,Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, (2023) https://doi.org10.1126/science.ade2574

Contacts: massimo.bellato@unipd.it
barbara.dicamillo@unipd.it

Modeling and wet-lab activities for Synthetic biology

Objective:

Model genetic circuits, fit parameters, implement engineered bacteria to gather data for:

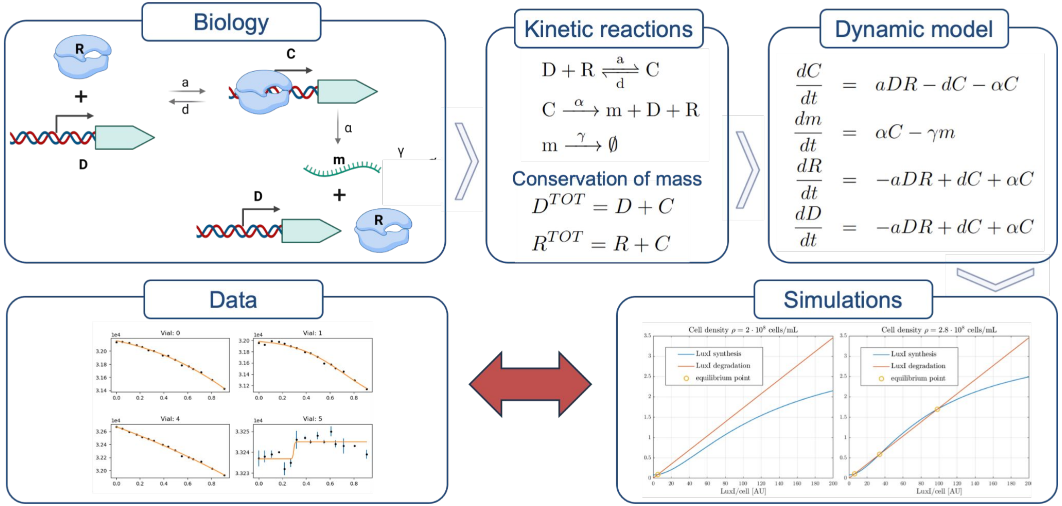
- Memory device to RNA exposure
- Urea biosensor
- Tuning bacterial growth

Skill requirements:

- Modeling
- Basics of biology
- Programming skills (Matlab)

Preferred background:

- Bioengineering
- Automation and Systems



Kinetic reactions

$$D + R \underset{d}{\overset{a}{\rightleftharpoons}} C$$

$$C \xrightarrow{\alpha} m + D + R$$

$$m \xrightarrow{\gamma} \emptyset$$

Conservation of mass

$$D^{TOT} = D + C$$

$$R^{TOT} = R + C$$

Dynamic model

$$\frac{dC}{dt} = aDR - dC - \alpha C$$

$$\frac{dm}{dt} = \alpha C - \gamma m$$

$$\frac{dR}{dt} = -aDR + dC + \alpha C$$

$$\frac{dD}{dt} = -aDR + dC + \alpha C$$

Cell density $\rho = 2 \cdot 10^8$ cells/mL

Cell density $\rho = 2.8 \cdot 10^8$ cells/mL

References/useful resources:

Brophy, J., Voigt, C. Principles of genetic circuit design. *Nat Methods* **11**, (2014). https://doi.org/10.1038/nmeth.2926

Barajas, C., *et al*. Feedforward growth rate control mitigates gene activation burden. *Nat Commun* **13**, (2022). https://doi.org/10.1038/s41467-022-34647-1

Pasotti L, et al. Re-using biological devices: a model-aided analysis of interconnected transcriptional cascades designed from the bottom-up. *J Biol Eng*. **14** (2017) https://doi.org/10.1186/s13036-017-0090-3

Contacts: massimo.bellato@unipd.it
barbara.dicamillo@unipd.it