ARSENAL:
Antimicrobial resistance prediction by machine learning approach

Ulysse Guyet

Macha Nikolski & Alexis Groppi

Computational Biology and Bioinformatics team, IBGC, UMR 5095

Machine Learning for Microbial Genomics Workshop – 23/09/2022

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

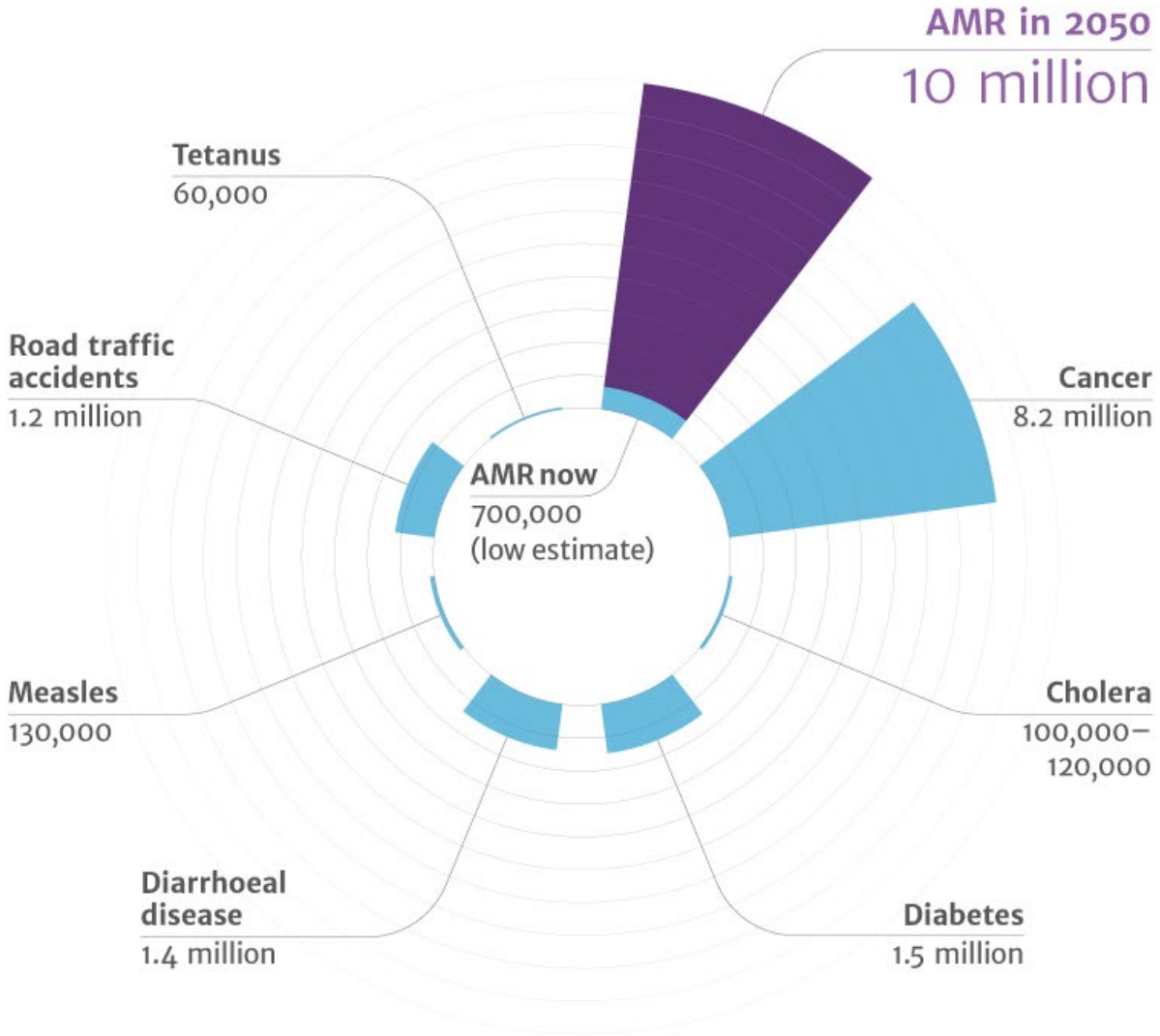
TABLE OF CONTENTS

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

TABLE OF CONTENTS

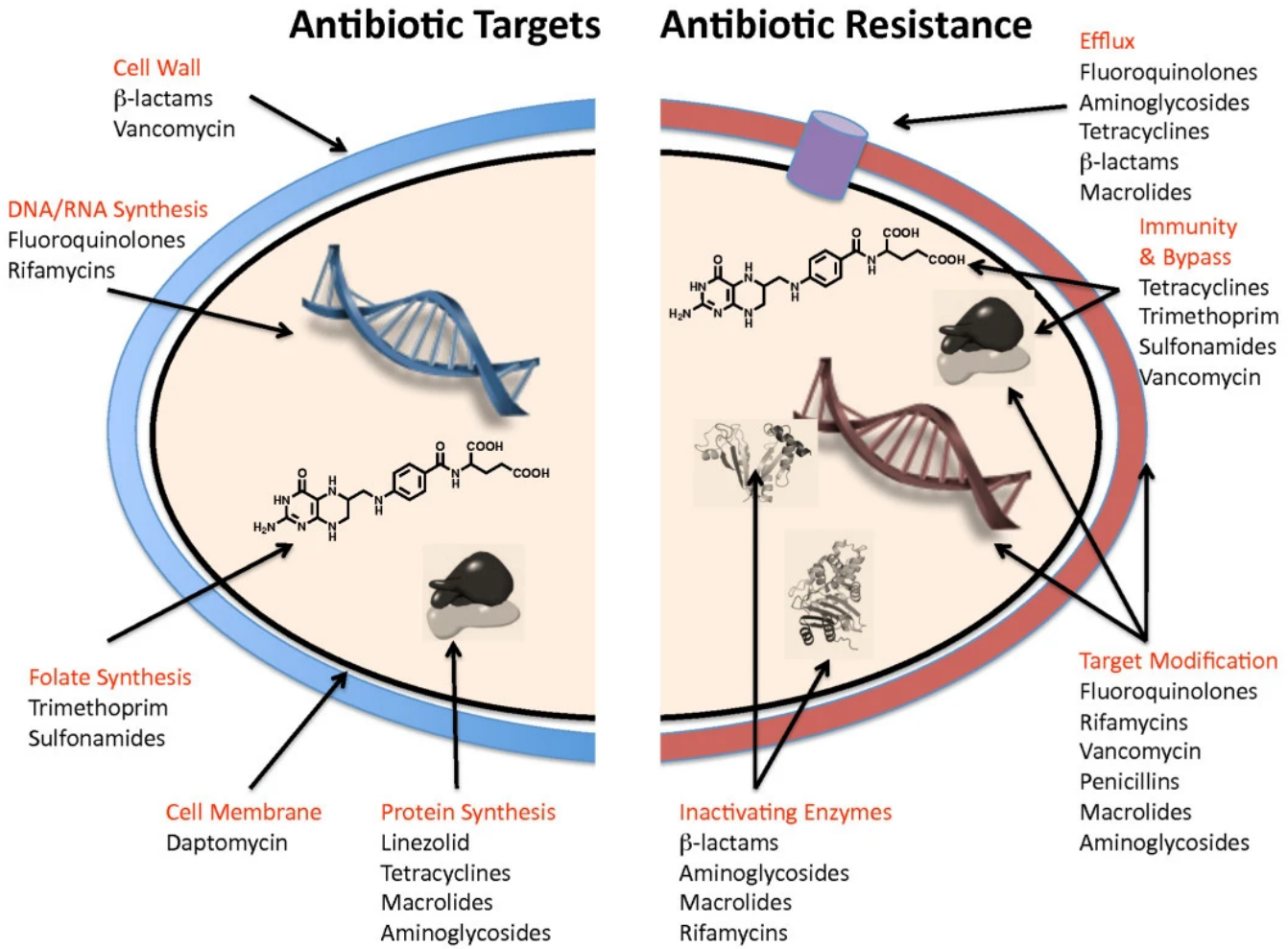
Antibiotic resistance: a major health problem

Major causes of death at the global scale



O'Neill et al., 2014

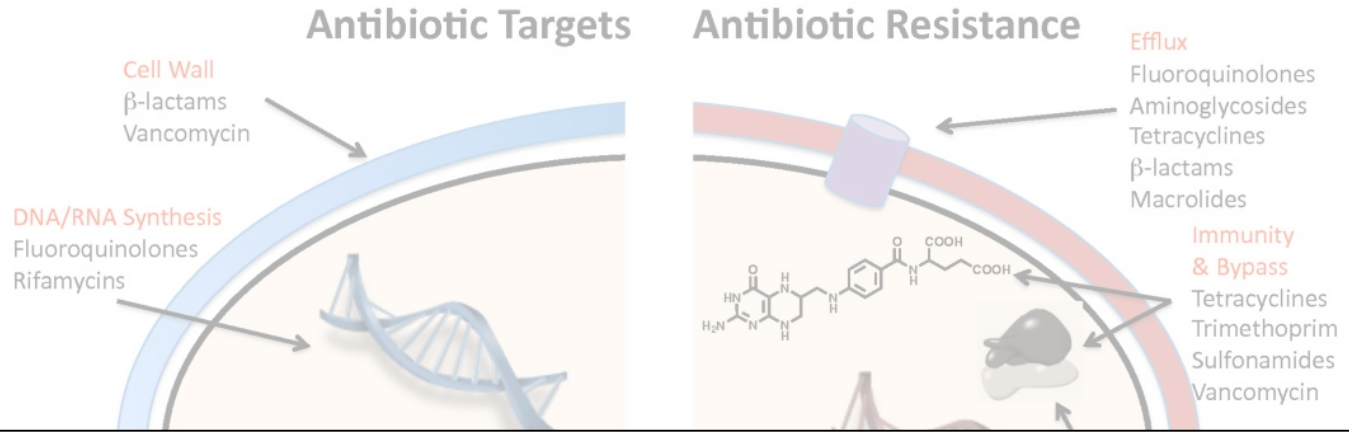
Antibiotic resistance mechanisms



- **Resistance mechanisms are mainly acquired by genome modifications: e.g. mutations, horizontal gene transfers**
- **Misuse or overuse of antibiotics increases the drug resistance by selection pressure**
- **Resistance is measured by MIC (Minimum Inhibitory Concentration) which is the minimum concentration of an antibiotic needed to inhibit bacterial growth**

Wright et al., 2010

Antibiotic resistance mechanisms



Predicting antibiotic resistance can allows:

- **Fast screening of antibiotics to determine the most effective antibiotic and the right dose against a specific bacterial infection**
- **Highlight new antibiotic resistance genes and biomarkers**

Macrolides
Aminoglycosides

Macrolides
Rifamycins

- Resistance mechanisms are mainly acquired by genome modifications: e.g. mutations, horizontal gene transfers
- Misuse or overuse of antibiotics increases the drug resistance by selection pressure
- Resistance is measured by MIC (Minimum Inhibitory Concentration) which is the minimum concentration of an antibiotic needed to inhibit bacterial growth

Wright et al., 2010

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

TABLE OF CONTENTS

Antimicrobial resistant bacteria models

Streptococcus pneumoniae

- Genome size: ~2 Mbp
- Gram-positive
- WHO antimicrobial (AMR) priority pathogen
- Resistance acquired by either **mutations** or by DNA **transformation** leading to mosaic genes

Pseudomonas aeruginosa

- Genome size: 5.5-7 Mbp
- Gram-negative
- WHO AMR priority pathogen
- Resistance mainly mediated by **mutations** or horizontal gene transfer (mostly plasmid **conjugation**)

Van der Poll, & Opal, 2009; Tacconelli et al. 2018

Antimicrobial resistant bacteria models

Streptococcus pneumoniae

- Genome size: ~2 Mbp
- Gram-positive
- WHO antimicrobial (AMR) priority pathogen
- Resistance acquired by either **mutations** or by DNA **transformation** leading to mosaic genes

Pseudomonas aeruginosa

- Genome size: 5.5-7 Mbp
- Gram-negative
- WHO AMR priority pathogen
- Resistance mainly mediated by **mutations** or horizontal gene transfer (mostly plasmid **conjugation**)

Van der Poll, & Opal, 2009; Tacconelli et al. 2018

Datasets

- **1312 *Streptococcus pneumoniae* newly sequenced genomes of strains isolated from patients**
  - Different sequence types (ST) and 64 distinct serotypes
  - many strains are multidrug resistant (MDR)
  - Geographical origin (isolated from 28 hospitals in 18 provinces of China between 2007 and 2020)

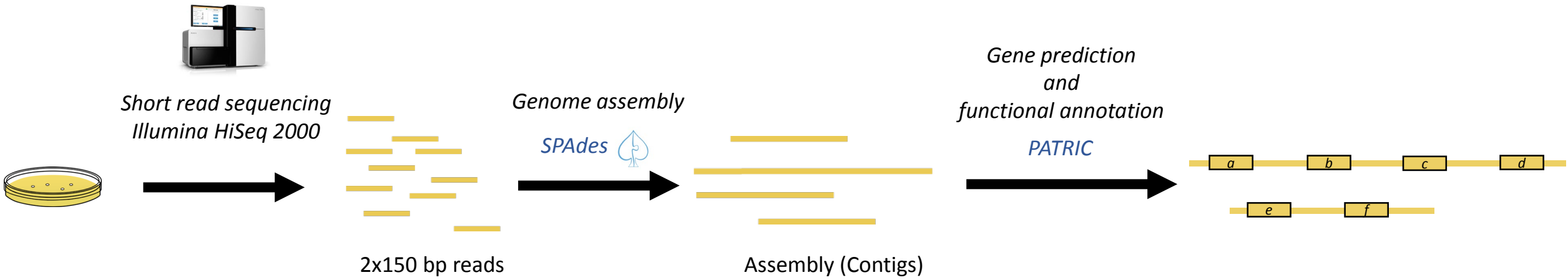
- **MIC values (Minimum Inhibitory Concentration) of each strain for 11 various antibiotic classes :**

  - β-lactams
  - macrolides
  - Lincosamide
  - Fluoroquinolones
  - Oxazolidinone
  - glycopeptides
  - Rifampin
  - Sulfonamides
  - Chloramphenicol
  - Phosphonic
  - aminosides

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

TABLE OF CONTENTS

Genome assembly & annotation pipeline



PATRIC (Pathosystems Resource Integration Center) Database :

- Prediction of CDS and functional annotation
- Assignment of each gene to a PLFam: group of genus-specific genes, wich share the same function and high sequence homology

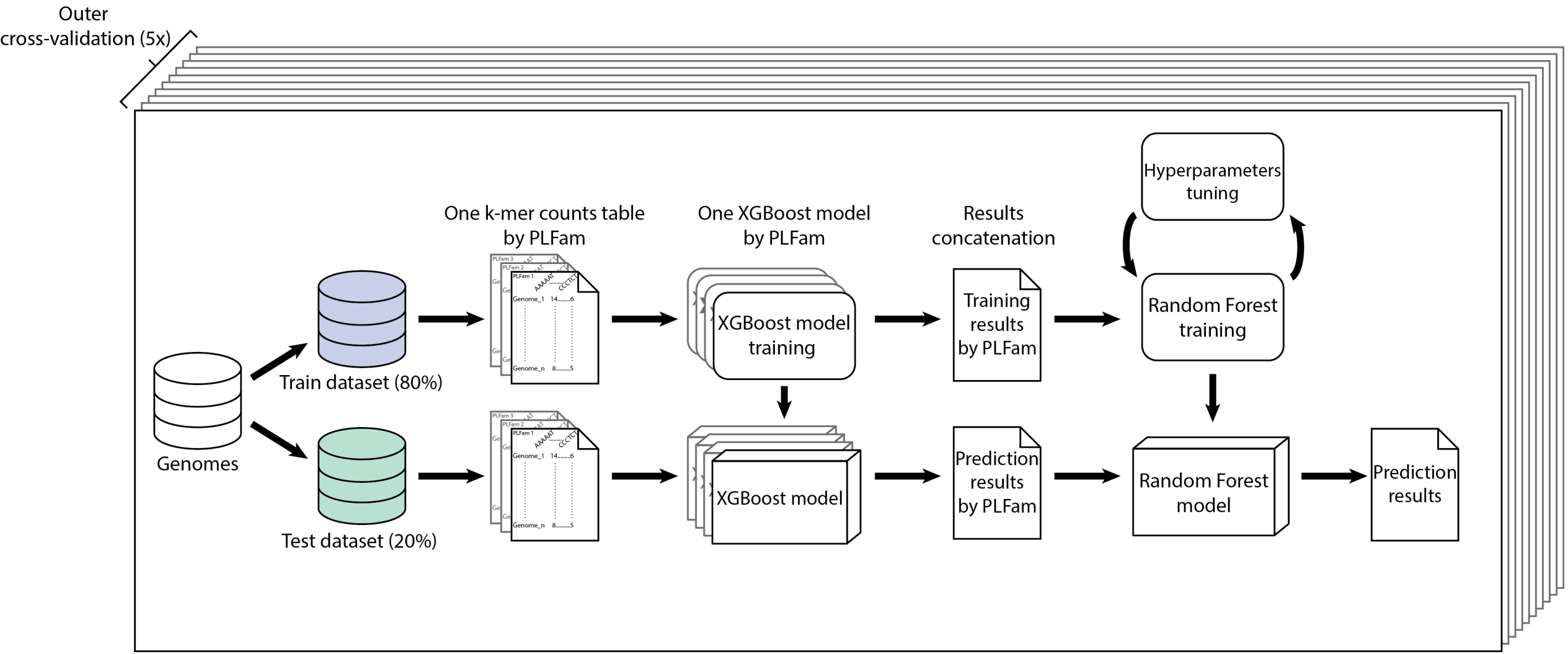
5770 PLFams by considering all genomes of *Streptococcus pneumoniae*

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

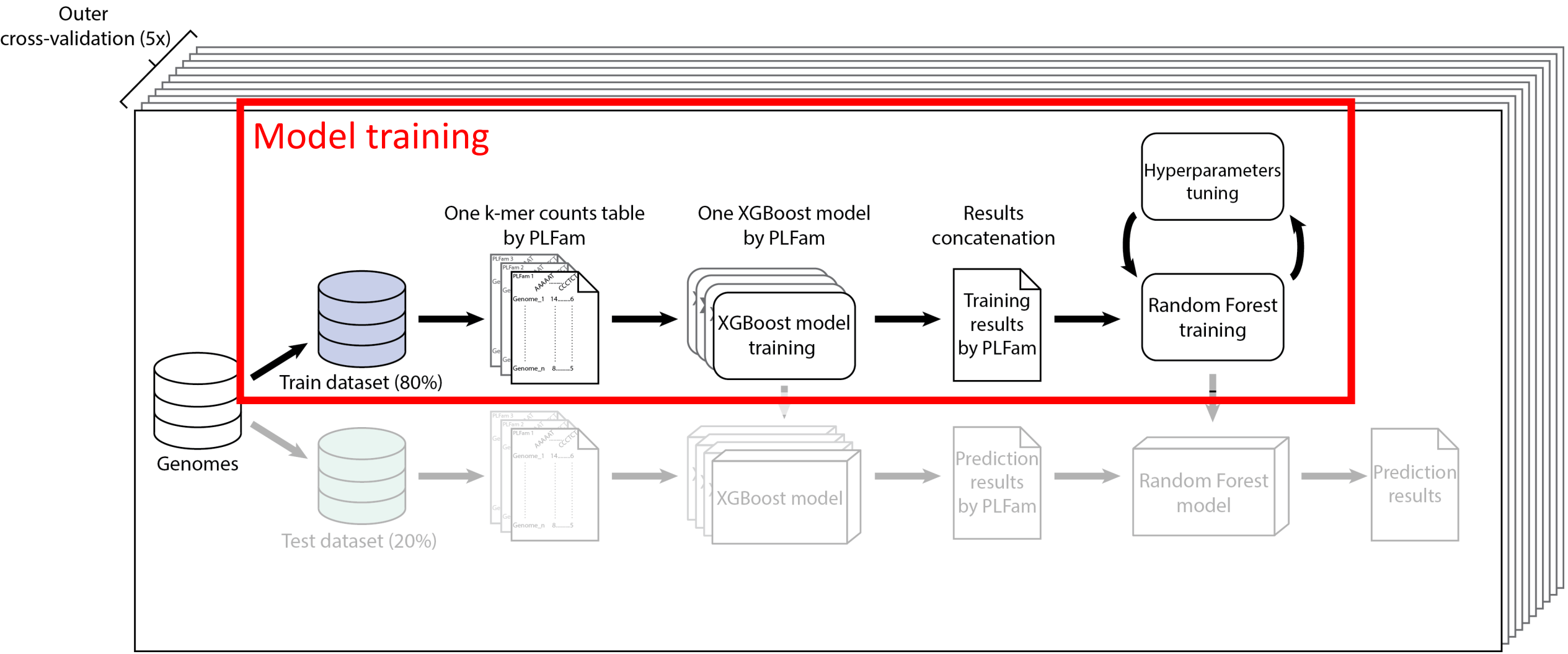
TABLE OF CONTENTS

ARSENAL prediction pipeline

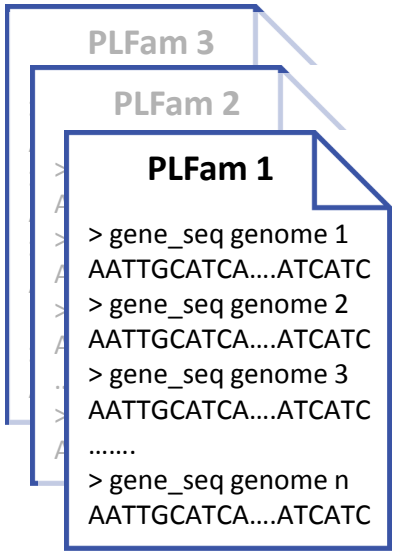
Machine learning pipeline for MIC prediction and linking phenotype and genotype



Machine learning pipeline for MIC prediction and linking phenotype and genotype



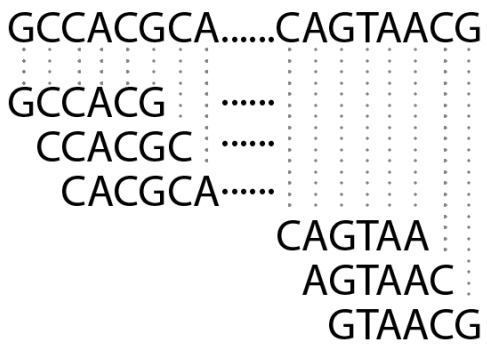
ARSENAL prediction pipeline



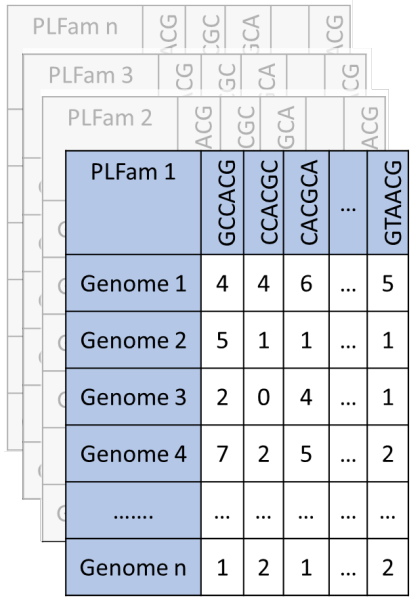
Fasta files containing all sequences of a specific PLFam

For each of the resulted 5770 PLFam:

Gene sequence fragmentation into k-mers (6 nucleotides lentgh)



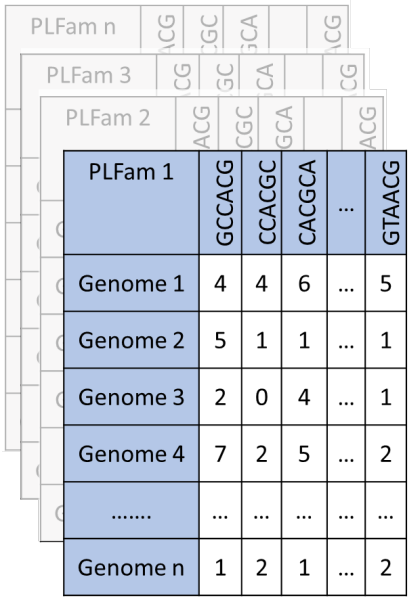
One k-mer occurrences table for each of the PLFam



PLFam 1	GCCACG	CCACGC	CACGCA	…	GTAACG
Genome 1	4	4	6	…	5
Genome 2	5	1	1	…	1
Genome 3	2	0	4	…	1
Genome 4	7	2	5	…	2
…….	…	…	…	…	…
Genome n	1	2	1	…	2

ARSENAL prediction pipeline

One k-mer occurrences table for each of the PLFam



PLFam 1	GCCACG	CCACGC	CACGCA	...	GTAACG
Genome 1	4	4	6	...	5
Genome 2	5	1	1	...	1
Genome 3	2	0	4	...	1
Genome 4	7	2	5	...	2
.....	...	...	...	...	...
Genome n	1	2	1	...	2

Antibiogram containing MIC values

Genomes	MIC values
Genome 1	1
Genome 2	8
Genome 3	2
Genome 4	0.5
.....	...
Genome n	2

One XGBoost model for each of the PLFam

Predicted MIC values	PLFam 1	PLFam 2	PLFam 3	...	PLFam n
Genome 1	1	1	0.5	...	1
Genome 2	8	4	16	...	4
Genome 3	2	4	4	...	1
Genome 4	0.5	1	0.5	...	1
.....	...	...	...	...	...
Genome n	2	2	2	...	2

Table gathering each MIC prediction for every PLFam

ARSENAL prediction pipeline

Predicted MIC values	PLFam 1	PLFam 2	PLFam 3	...	PLFam n
Genome 1	1	1	0.5	...	1
Genome 2	8	4	16	...	4
Genome 3	2	4	4	...	1
Genome 4	0.5	1	0.5	...	1
.....	...	...	...	...	...
Genome n	2	2	2	...	2

Table gathering each MIC prediction for every PLFam

Antibiogram containing MIC values

Genomes	MIC values
Genome 1	1
Genome 2	8
Genome 3	2
Genome 4	0.5
.....	...
Genome n	2

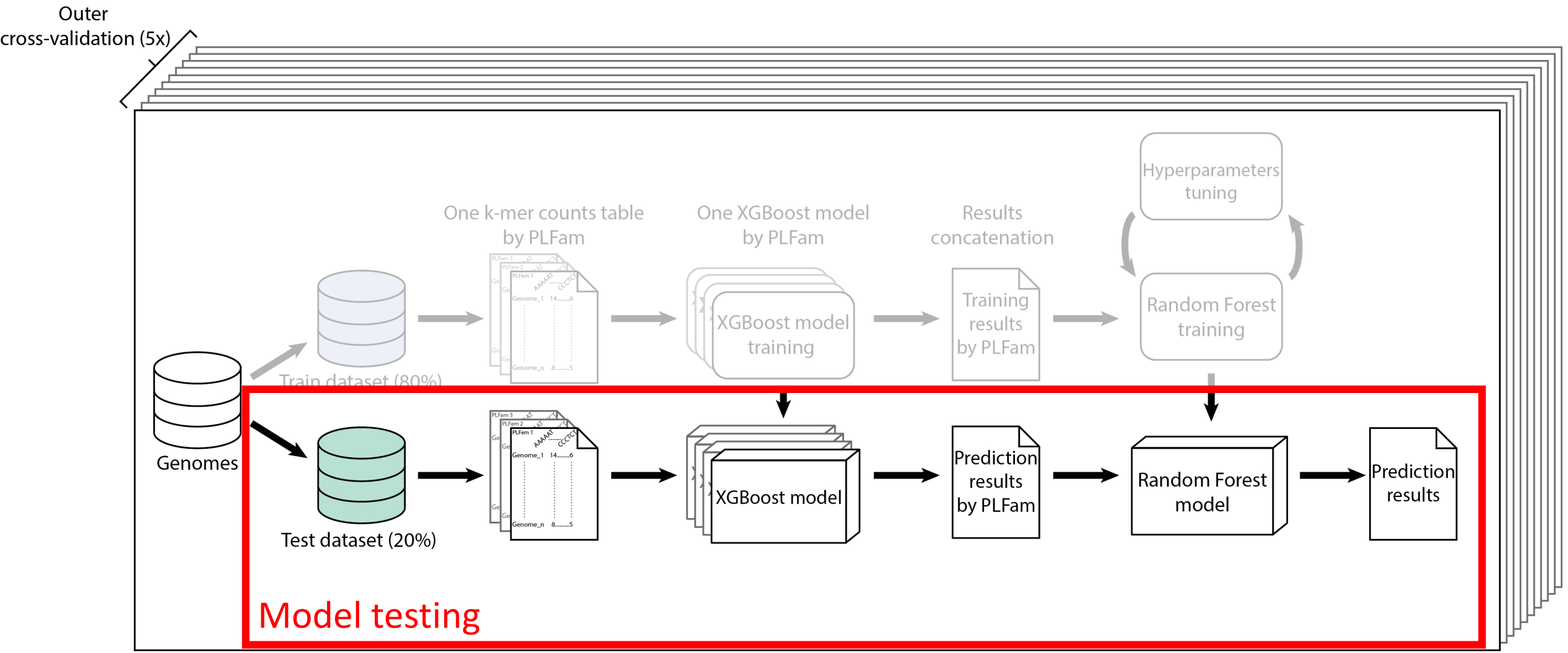
Random Forest model training

Tuning hyperparameters using Bayesian sampler

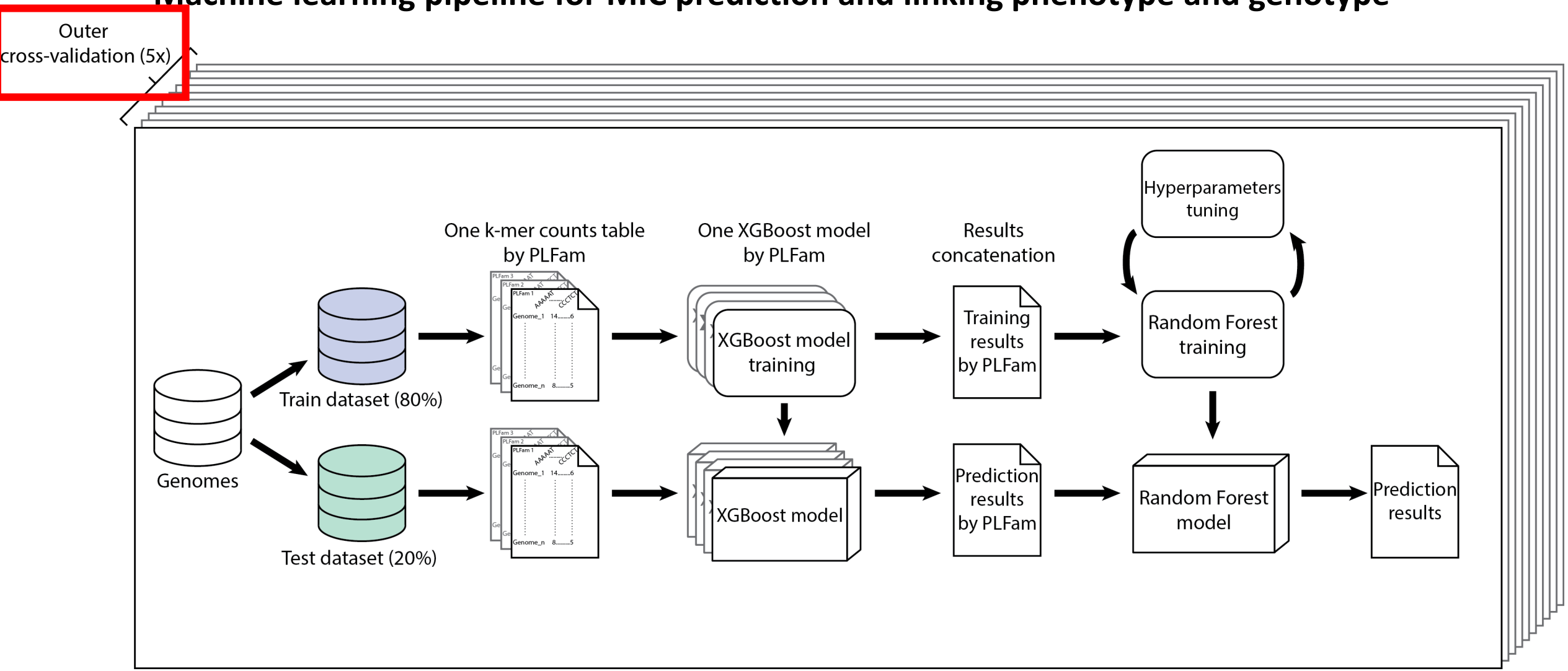
Genomes	Predicted MIC values
Genome 1	1
Genome 2	8
Genome 3	2
Genome 4	0.5
.....	...
Genome n	2

One predicted MIC value by genome

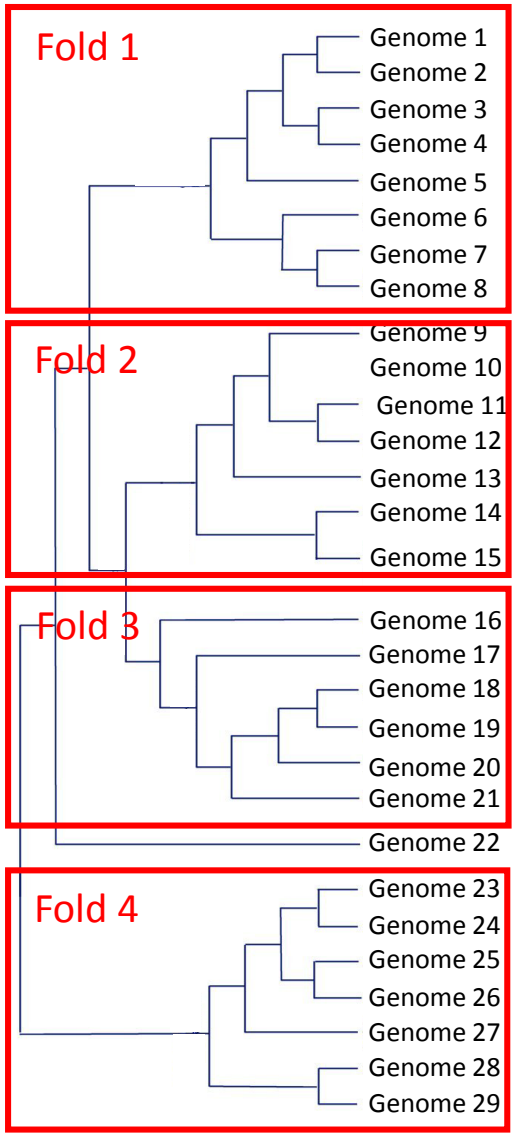
Machine learning pipeline for MIC prediction and linking phenotype and genotype



Machine learning pipeline for MIC prediction and linking phenotype and genotype



ARSENAL prediction pipeline



Genome-distance-based cross-validation

- Maximize genome distance between the test sets of cross-validation folds
- Improve independence of test sets by segregating samples based on a known dependence structure in the data (here genomic distance)

→ mirrors the application of the trained model towards independently sampled datasets

Lüftinger et al., 2021

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

TABLE OF CONTENTS

Model precisions of MIC prediction

Streptococcus pneumoniae

β-lactams:

Antibiotic Class	Antibiotic	Accuracy (±1 cat.)
Penicillins	Penicillin	0.832
	Amoxicillin	0.897
Cephalosporins	Cefuroxime	0.829
	Cefepime	0.848
	Ceftriaxone	0.859
Carbapenem	Imipinem	0.935
	Meropenem	0.898
	Ertapenem	0.939

Antibiotic Class	Antibiotic	Accuracy (±1 cat.)
Fluorquinolones	Ciprofloxacin	0.848
	Levofloxacin	0.949
	Moxifloxacin	0.909
Rifampin	Rifampicin	0.838
phosphonic	phosphonomycin	0.838

Importance ranked genes - penicillin

SHAPley value (SHapley Additive exPlanations):
Determine the top genes that have a positive effect on the model construction

Among genes of interest :

- The 3 Penicillin-binding proteins of *Streptococcus pneumoniae* are found in 10th, 14th and 16th position
- Significant number of mobile elements (1st, 2nd; 12 genes in the first 100 most important genes)
- choline binding proteins- associated with drug and cell death (4th, 13th, 31th position)
- Cell wall surface anchor family protein (15th)

→ Identification of these resistance-related genes provides confidence in model predictions

ARSENAL: **A**ntimicrobial **r**esi**s**tance pr**e**dictio**n** by m**a**chine **l**earning approach

TABLE OF CONTENTS

Conclusion and future work

Conclusions

- The model is accurate for β-lactams antibiotics in the case of *Streptococcus pneumoniae*
- Validation of the method by the identification of genes known to be involved in resistance mechanisms
- Highlight some genes without functional annotation potentially linked with resistance

  ➡ Functional characterization via CRISPR/Cas9

Future computational work

- Test the method for other antibiotics classes and for *Pseudomonas aeruginosa*
  (data from V. Dubois lab, Bordeaux & J. Feng lab, Tianjin - China)
- Add tRNAs and promoters to the model since they may influence some resistance mechanisms

Long-term perspectives

- Rapid identification of resistance markers in patients based on genomic & metabolic markers

  ➡ Integration of metabolic data (in process of generation)

Acknowledgements

MAGITICS project collaborators

Jacques Corbeil (Université Laval, Québec, Canada)

Jie Feng (Institute of microbiology, Tianjin, China)

Véronique Dubois (CNRS, Bordeaux)

Léa Bientz (CNRS, Bordeaux)

CBiB/IBGC

Macha Nikolski

Alexis Groppi

Aurélien Barré

Emmanuel Bouilhol

Edgar Lefevre

Grigorii Sukhorukov

Thanks for your attention !