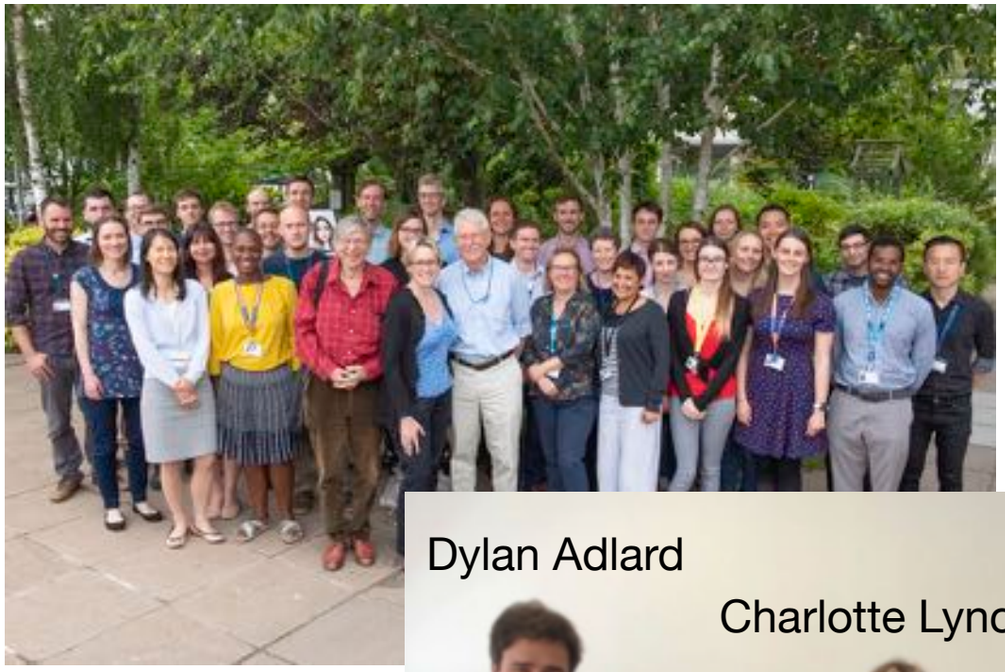
Predicting resistance to antituberculars in a global cloud-based platform.

Philip W Fowler
Experimental Medicine
John Radcliffe Hospital
University of Oxford

http://fowlerlab.org
@_fowlerlab_



Dylan Adlard

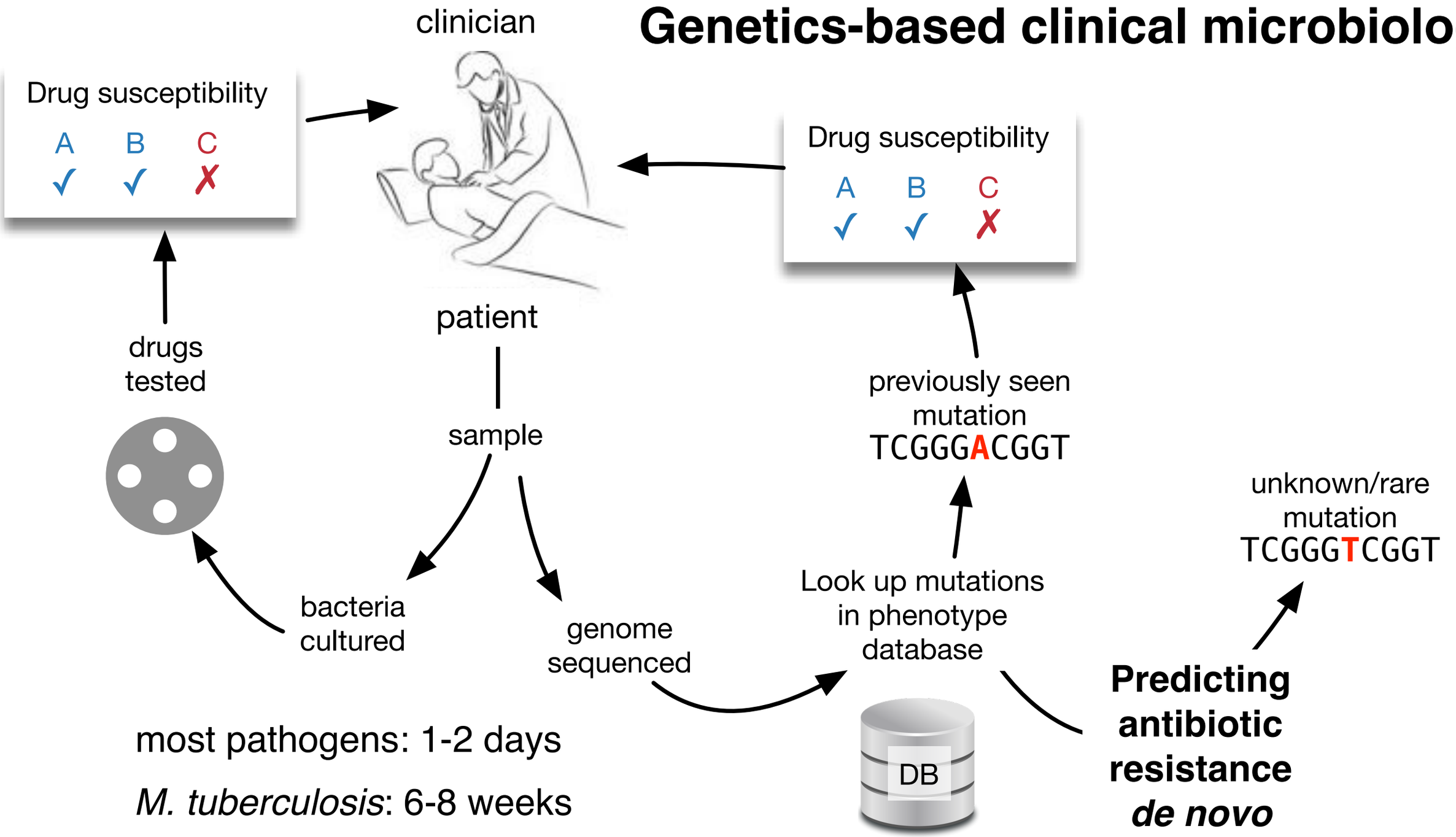
Charlotte Lynch

Alice Brankin Matty Johnson Philip Fowler

https://fowlerlab.org
@_fowlerlab_
https://github.com/fowler-lab

Culture-based clinical microbiology

Genetics-based clinical microbiology



Predicting antibiotic resistance *de novo*

Machine Learning to **rule-out** large number of no-effect mutations

Relative binding free energies to **rule-in** remaining mutations that confer resistance

May 2021

Press Release

Oxford University and Oracle Partner to Speed Identification of COVID-19 Variants

The partnership will enable global genomic sequencing and examination through a specialist platform developed on Oracle Cloud Infrastructure to help mitigate the impact of potentially dangerous COVID-19 variants

Oxford, England and Austin, Texas—May 14, 2021

The emergence of more infectious variants of the COVID-19 virus is threatening to slow the global recovery and potentially thwart current vaccine immunity. To help governments and medical communities identify and act on these variants faster, Oxford University and Oracle have created a Global Pathogen Analysis System (GPAS) combining Oxford's Scalable Pathogen Pipeline Platform (SP^3) with the power of Oracle Cloud Infrastructure (OCI). This initiative builds on the work of a Wellcome Trust-funded consortium including Public Health Wales, the University of Cardiff, and Public Health England.

"This powerful new tool will enable public health scientists in research establishments, public health agencies, healthcare services, and diagnostic companies around the world to help further understanding of infectious diseases, starting with the coronavirus," said Derrick Crook, Professor of Microbiology in the Nuffield Department of Medicine at the University of Oxford.

"The Global Pathogen Analysis System will help to establish a global common standard for assembling and analyzing this new virus, as well as other microbial threats to public health. This adds a new dimension in our ability to process pathogen data. We are excited to partner with Oracle to further our research using this cutting-edge technology platform," added Crook.

First used for tuberculosis, SP^3 has been repurposed to unify, standardize, analyze, and compare sequence data of SARS-CoV-2, yielding annotated genomic sequences and identifying new variants and those of concern. SP^3's processing capability has been enhanced with extensive new development work from Oracle, enabling high performance and security plus 7 by 24 worldwide availability of the SP^3 system in the Oracle Cloud. The SP^3 system will now deliver comprehensive and standardized results of COVID-19 analyses within minutes of submission on an international scale. The results will be shared with countries around the globe in a secure environment.

"The opportunity of applying systematic examination for genetic variants in a range of pathogens will have major benefits for global public health. This program, with Oracle as a partner, takes us a step closer to this goal," said Sir John Bell, Regius Professor of Medicine at the University of Oxford.

Coupled with the extensive machine learning capabilities in the Oracle Cloud, collaborating scientists, researchers, and governments worldwide can process, analyze, visualize, and act on a wide collection of COVID-19 pathogen data for the first time. This includes identifying variants of interest and their potential impact on vaccine and treatment effectiveness. For example, analytics dashboards in the system will show which specific strains are spreading more quickly than others and whether genetic features contribute to increased transmissibility and vaccine escape. Already, Oxford has processed half the world's SARS-CoV-2 sequences, more than 500,000 in total.

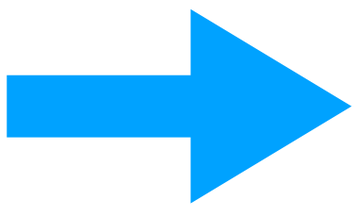
"There is a critical need for global cooperation on genomic sequencing and examination of COVID-19 and other pathogens," said Oracle

UNIVERSITY OF OXFORD

+

ORACLE

GPAS | Global Pathogen Analysis System



This isn’t HPC; a single SARS-CoV-2 sample requires about 0.3-0.5 CPUh

But there are >3 million SARS-CoV-2 raw genetic files deposited in the European Nucleotide Archive.

Free to low- and middle-income countries

Engaged in **tenders** with public health agencies in high-income countries

Organisations **own** their data

GPAS | Global Pathogen Analysis System

Organisations can choose to **share** samples with the **relatedness service**

Organisations can choose to **grant access** to samples so another organisation can view them.

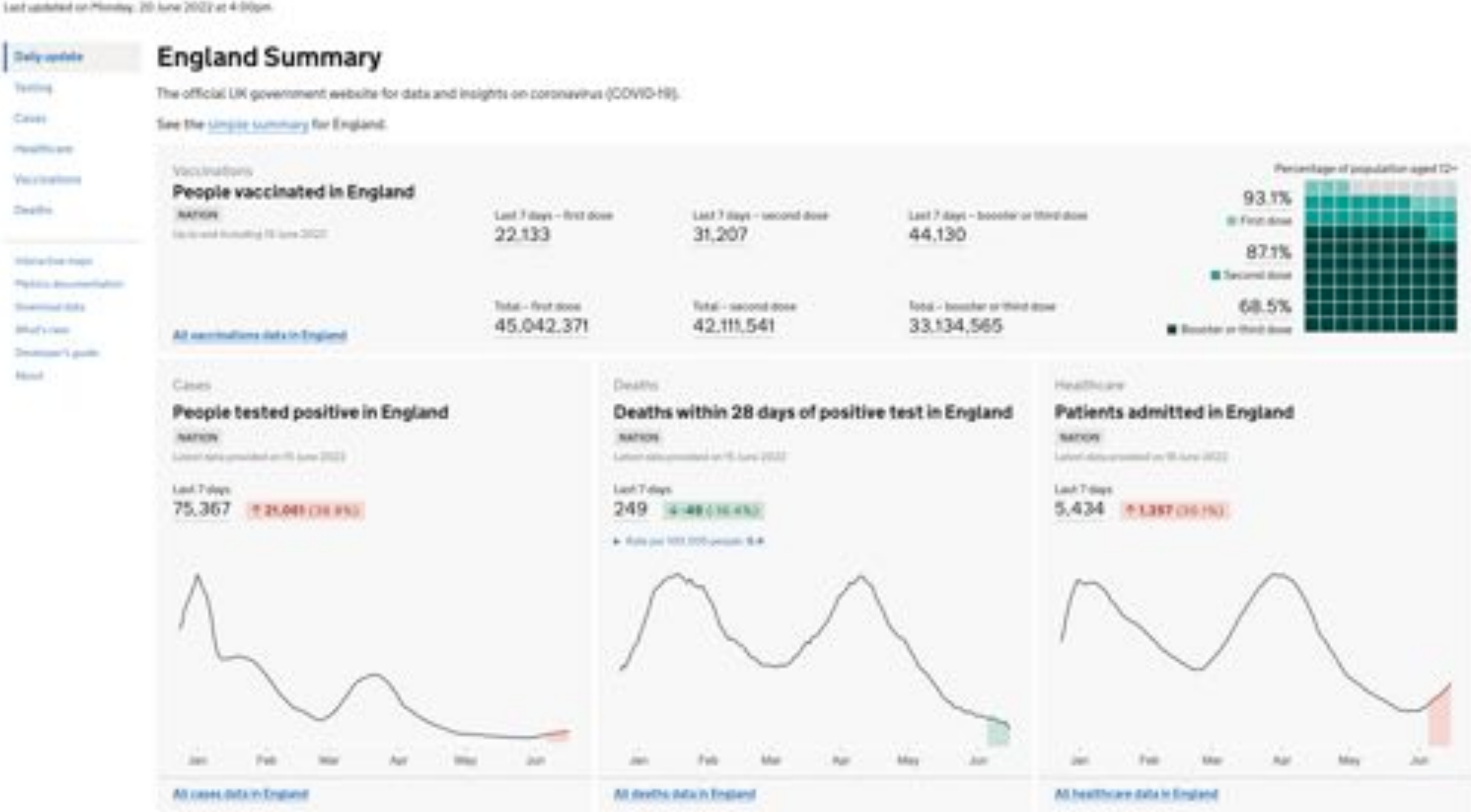
At its core, GPAS just **processes samples**; it doesn't replace GISAID or ENA or NextStrain, but hopefully makes it easier to use them

Use GPAS to (optionally with consent) collect additional data for research projects

Identical samples assumed to be part of the same outbreak

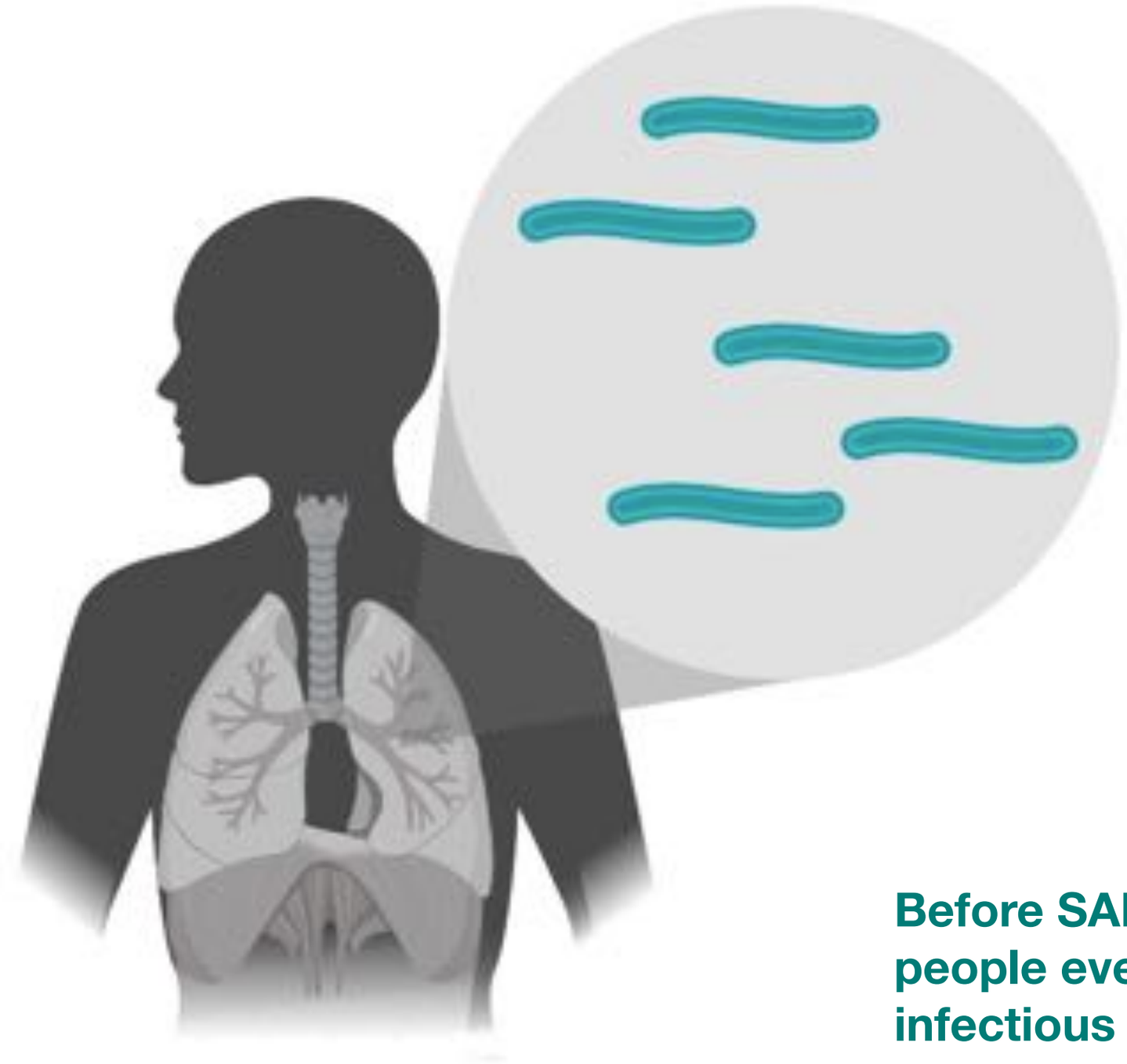
If enough countries share their samples, will be able to detect **importation events** (outbreaks across geographical borders)

Will also enable **novel variant detection** since will tell you that no-one has seen a related sample



BA.4, BA.5 on the rise in the UK + recombinants (“Deltacron”)

Mycobacterium tuberculosis



Grows slowly (≥24h doubling time)

~10 million develop TB each year and around 1.5 million die

Before SARS-CoV-2, TB killed more people every year than any other infectious disease

Tuberculosis

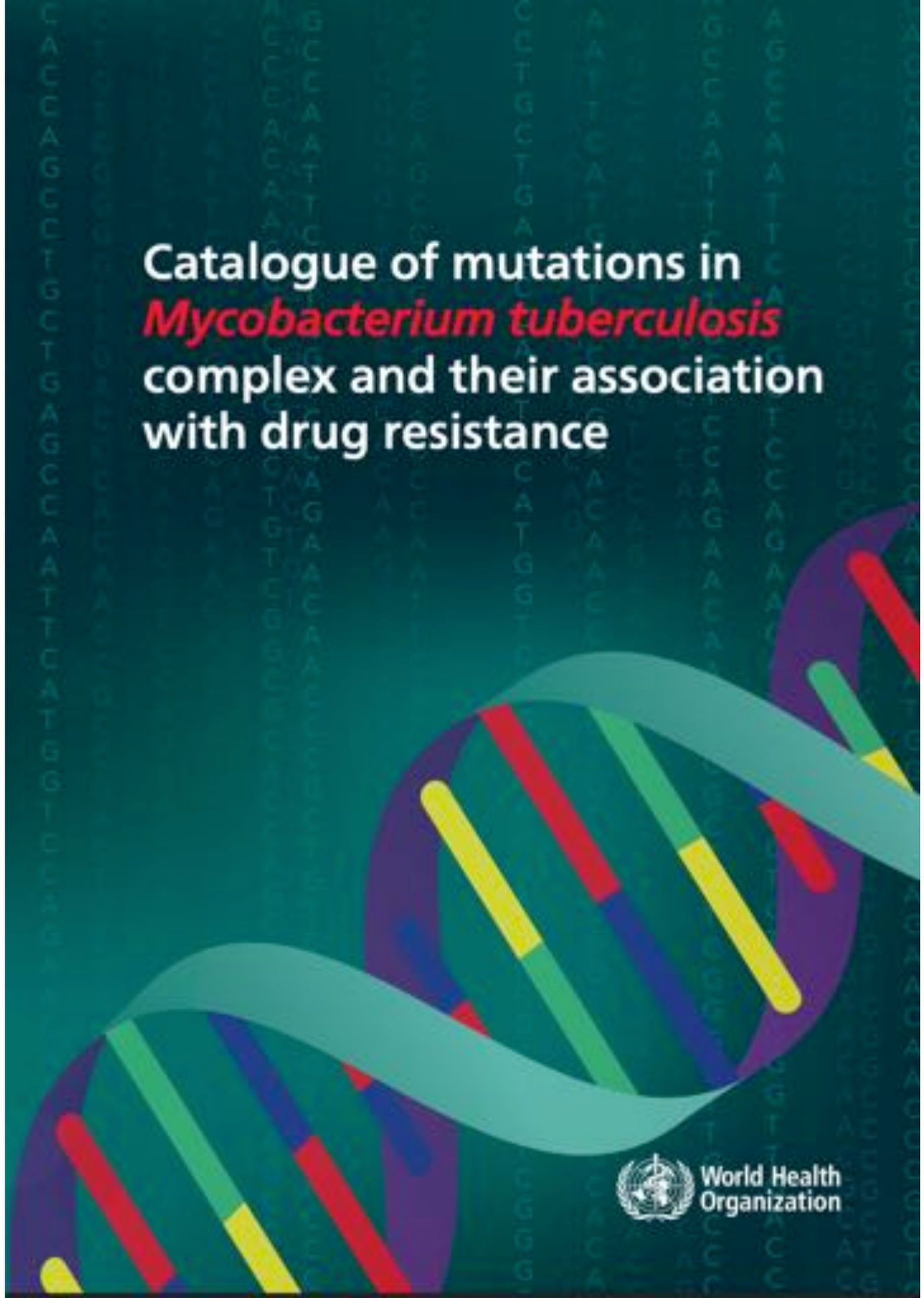
Most resistance-causing mutations in TB have now been identified

That will permit personalised treatment of the disease

Collected >20,000 TB samples from 11 countries on five continents. Each sequenced and tested with panel of 13 antituberculars.

Oct 19th 2021

UNTIL SARS-COV-2 emerged, the most destructive pathogen on the planet was *Mycobacterium tuberculosis*, the bug that causes TB. In 2020, according to the latest report from the World Health Organisation, published on October 14th, this organism cut short 1.5m lives—a figure 100,000 higher than the previous year (and the first annual rise since 2005), mainly because of disruptions to health services caused by covid-19. People whose immune systems have been wrecked by HIV are particularly at risk. Some 200,000 of the 680,000 annual AIDS deaths are a result of secondary TB infections.



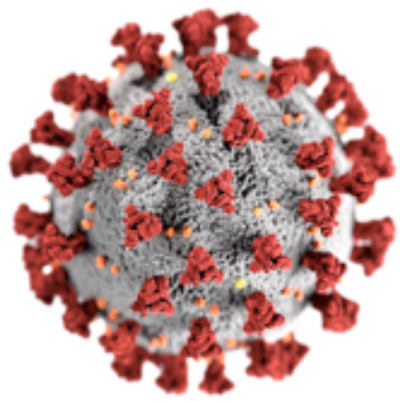
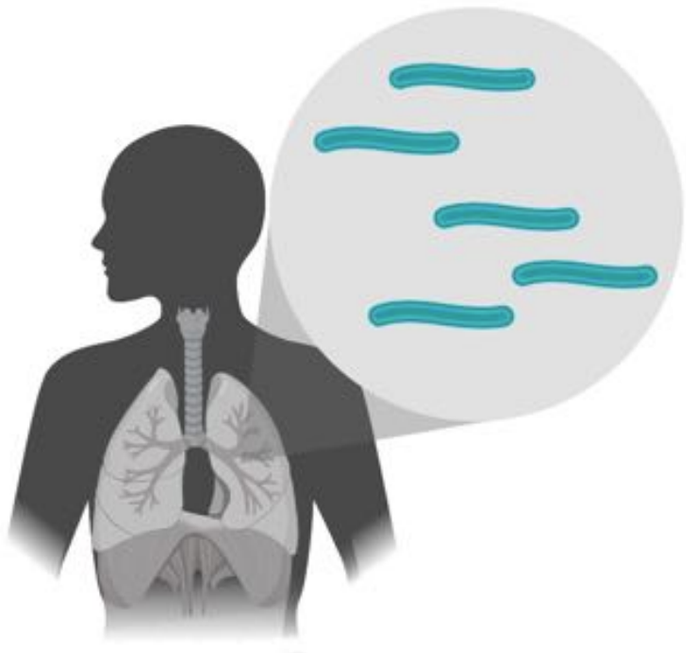
75,000 TB genomes
Most with some drug susceptibility testing data

+

GPAS | Global Pathogen Analysis System

=

cheap, rapid treatment plans to treat tuberculosis?

		relevant information	*where decisions made*
SARS-CoV-2		lineage	public health
M tuberculosis complex		antibiogram	clinician/ patient

Rifampicin U Isoniazid S Pyrazinamide S Ethambutol S

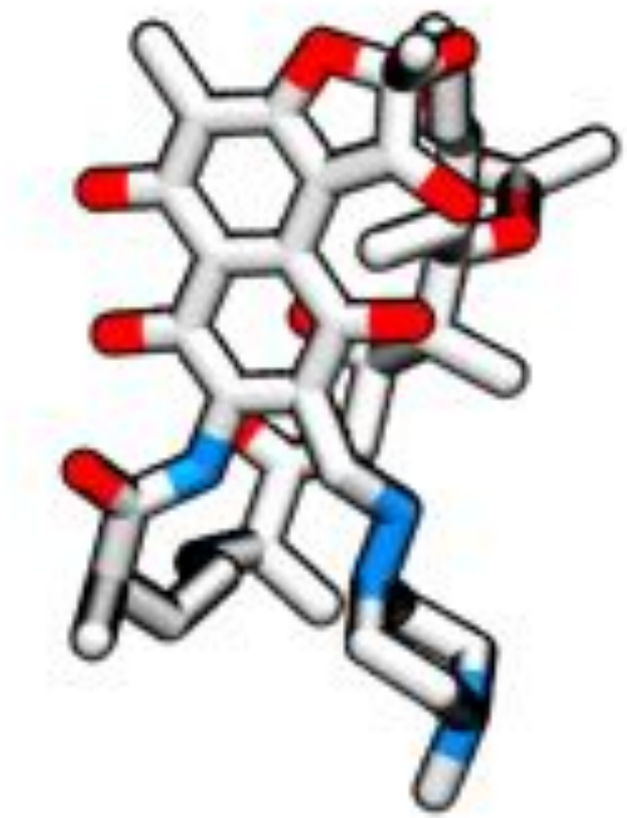
Predicting antibiotic resistance *de novo*

Machine Learning to **rule-out** large number of no-effect mutations

1

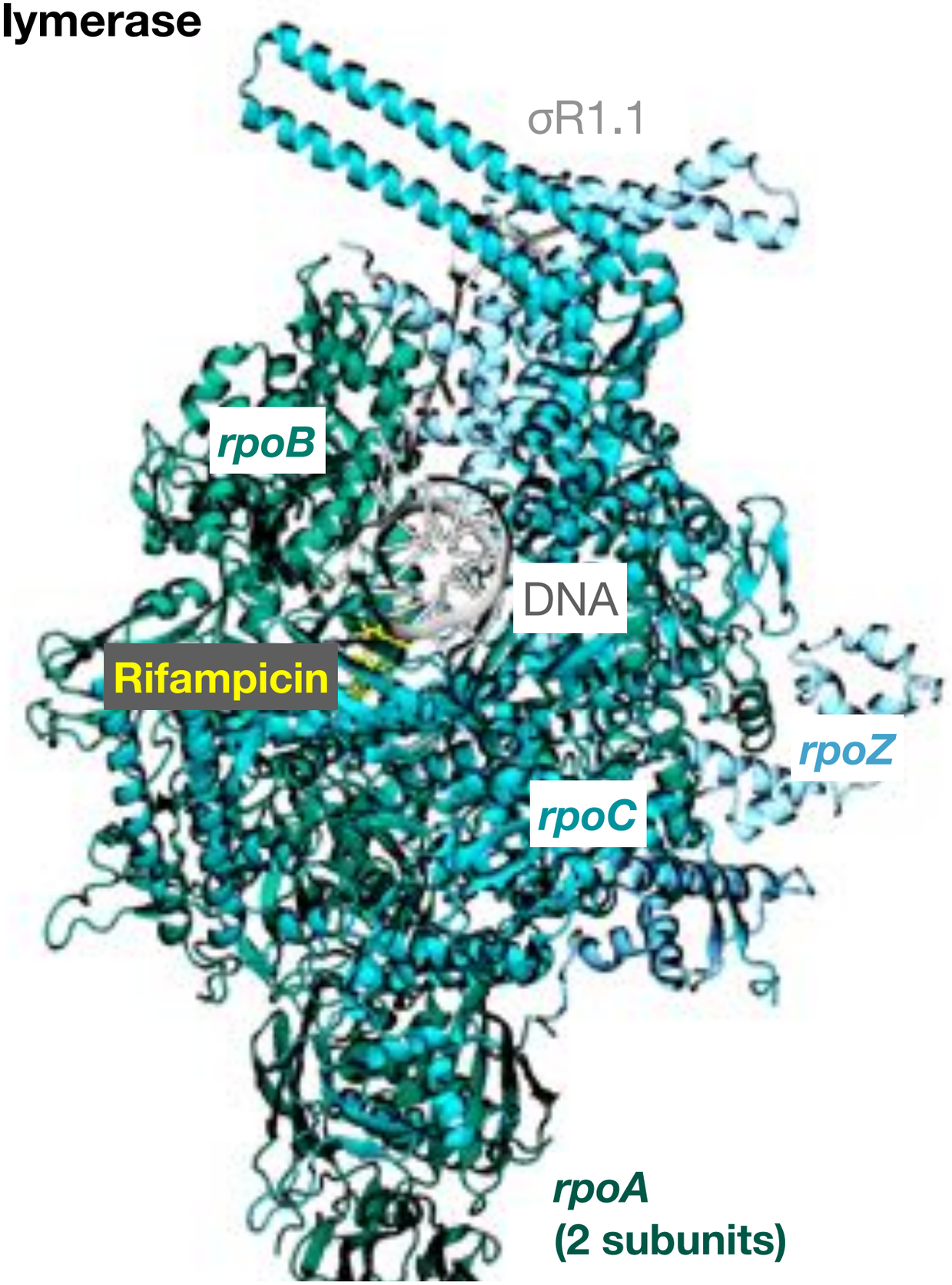
Relative binding free energies to **rule-in** remaining mutations that confer resistance

2

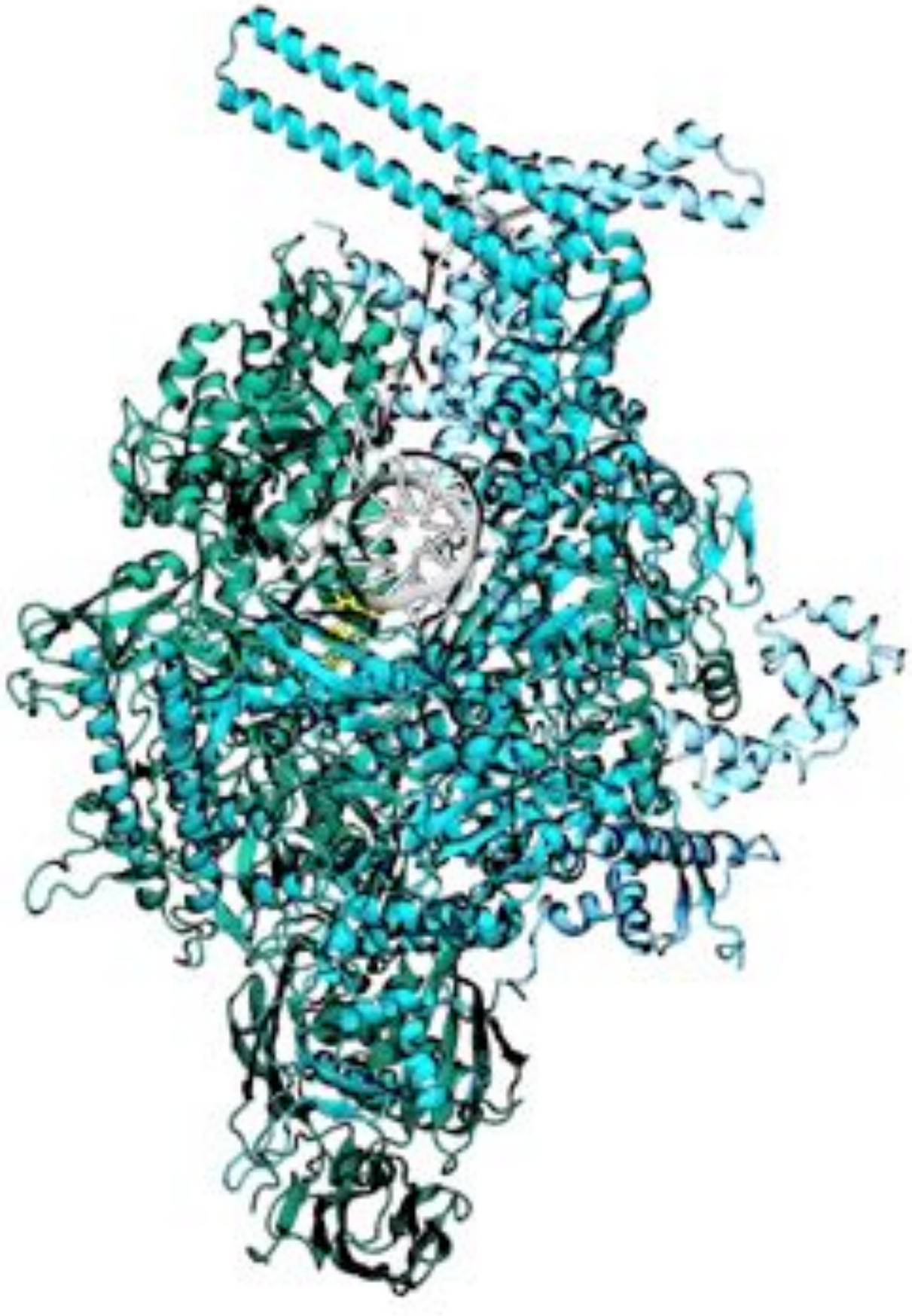


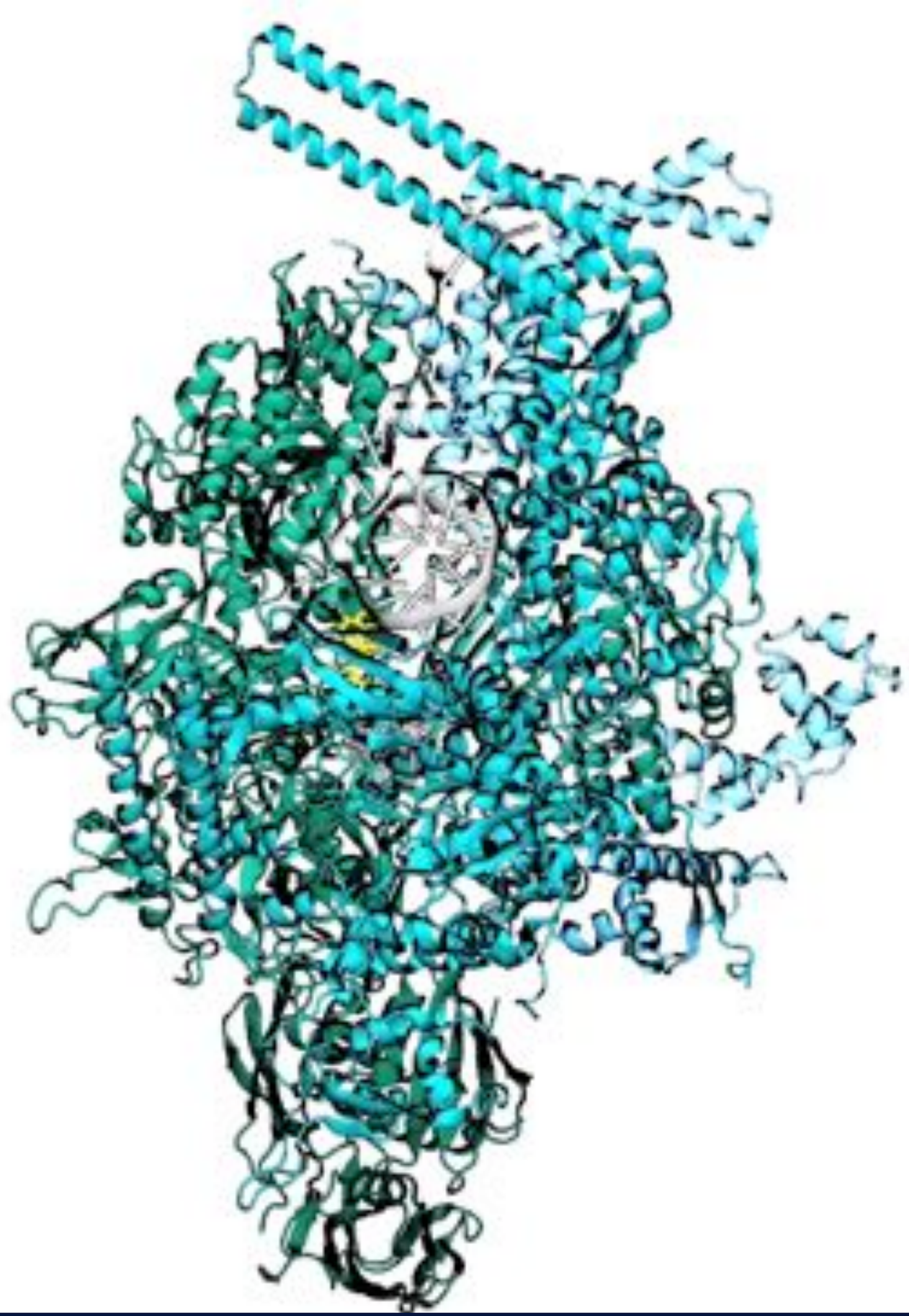
M. tuberculosis RNA polymerase

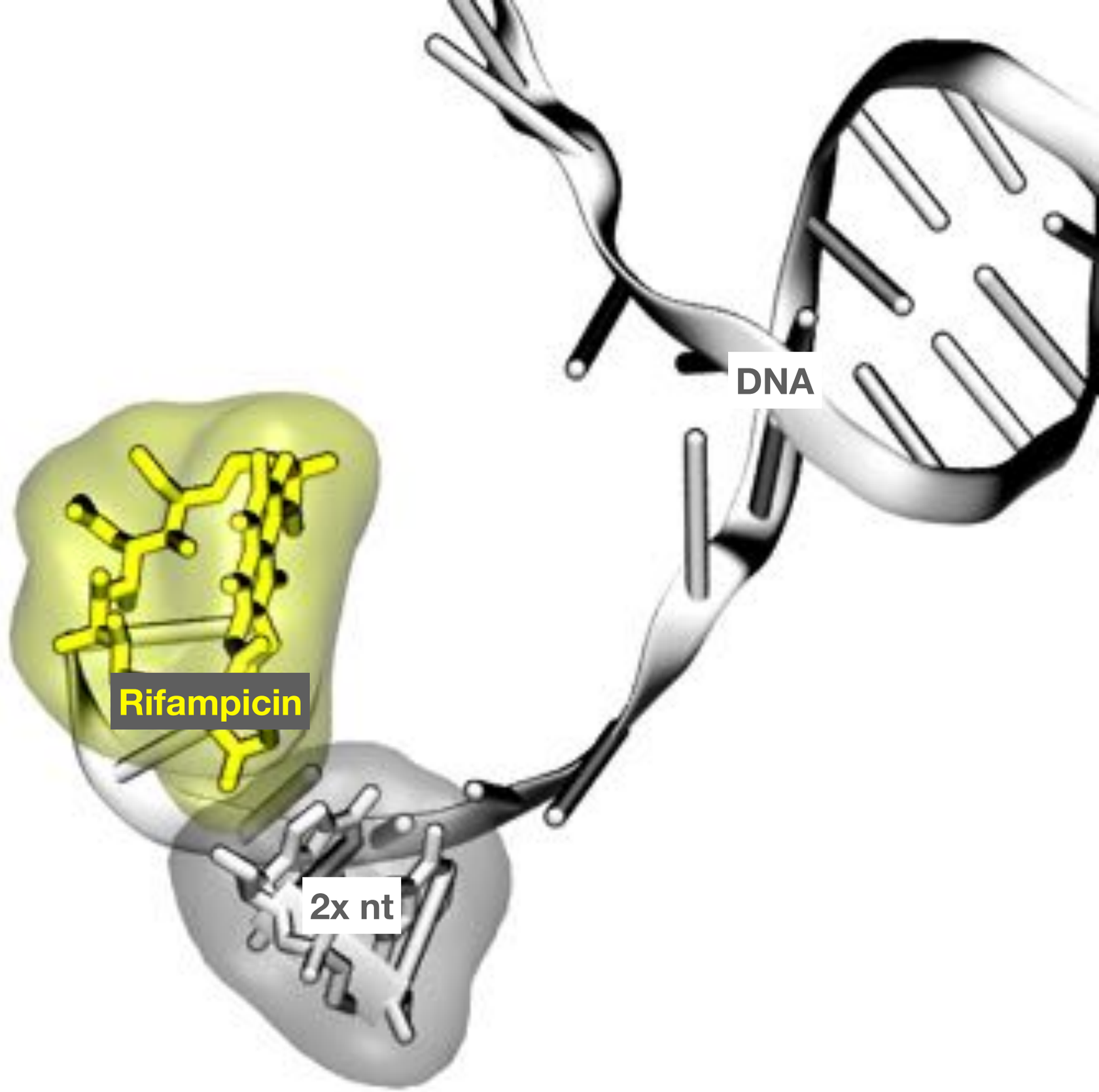
PDB: 5uh6



Lin W, Mandal S, Degen D, Liu Y, Ebright YW, Li S, Feng Y, Zhang Y, Mandal S, Jiang Y, Liu S, Gigliotti M, Talaue M, Connell N, Das K, Arnold E, Ebright RH. 2017. Structural Basis of Mycobacterium tuberculosis Transcription and Transcription Inhibition. Mol Cell 66:169–179.e8.






DNA
Rifampicin
2x nt

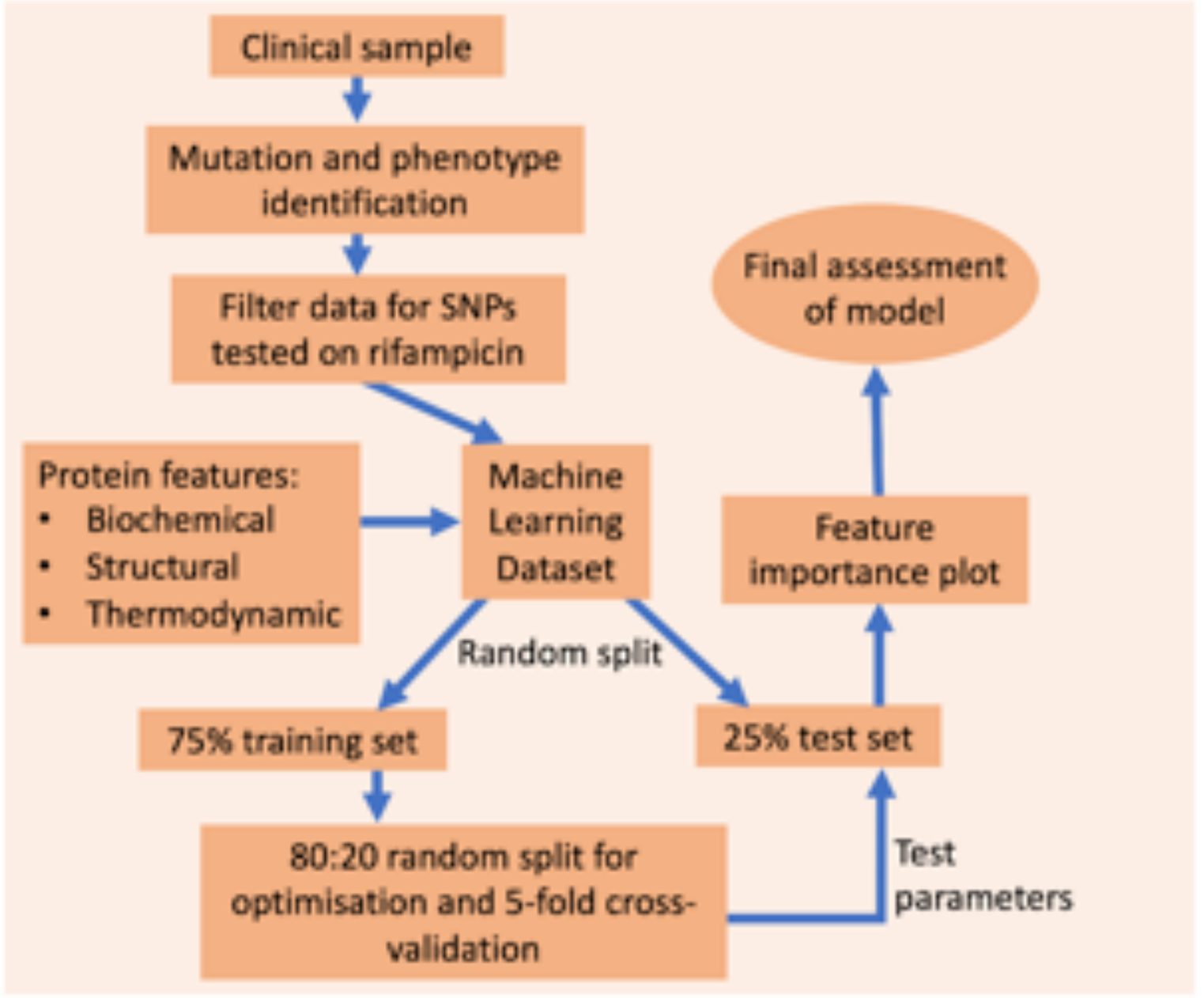
Pipeline

Dataset:

- CRyPTIC (Comprehensive Resistance Prediction for Tuberculosis: an International Consortium) project
- WHO-endorsed catalogue of mutations (binary and MICs)
- 17978 samples

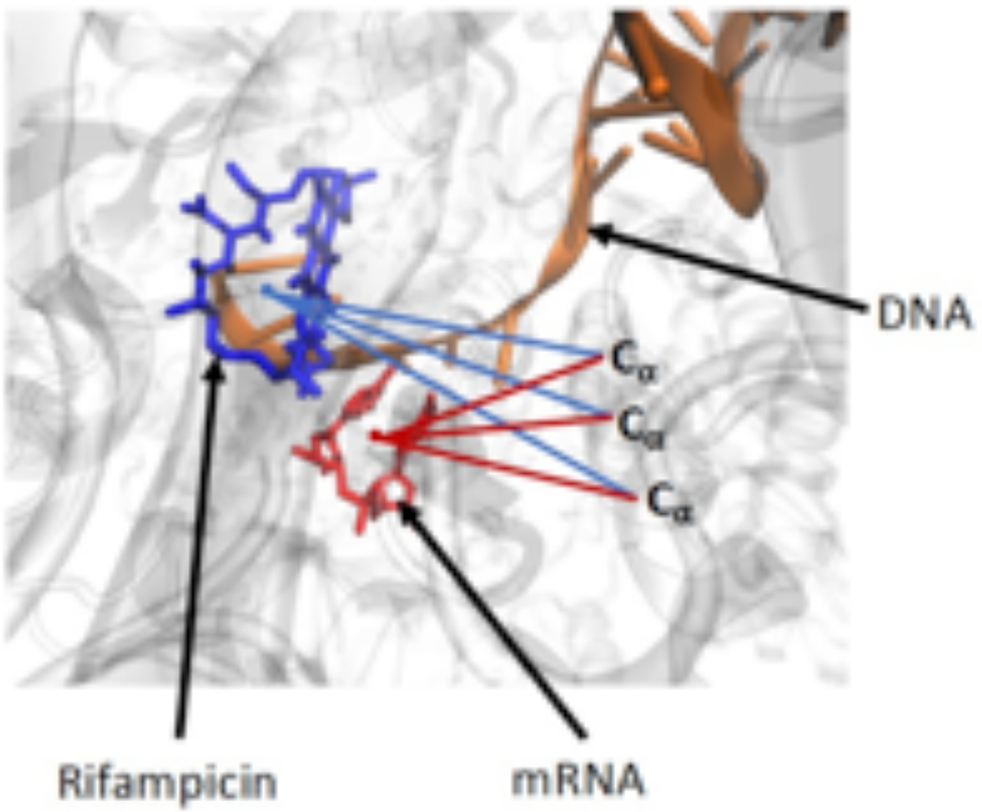
Filter for:

- *rpoB*
- unique mutations (mostly S450L)
- remove synonymous and phylogenetic mutations
- aggregate phenotypes
- 307 samples: 259 S, 48 R



Building the features dataset: sbmlcore

Python module of classes and functions to build a protein features dataset



AminoAcidProperties

Changes in:

- Volume
- Hydrophobicity (Kyte-Doolittle and Wimley-White)
- Molecular weight
- Isoelectric point

StructuralDistances

- COM of entity of interest to each Cα

DeepDDG

- Predicts change on protein stability

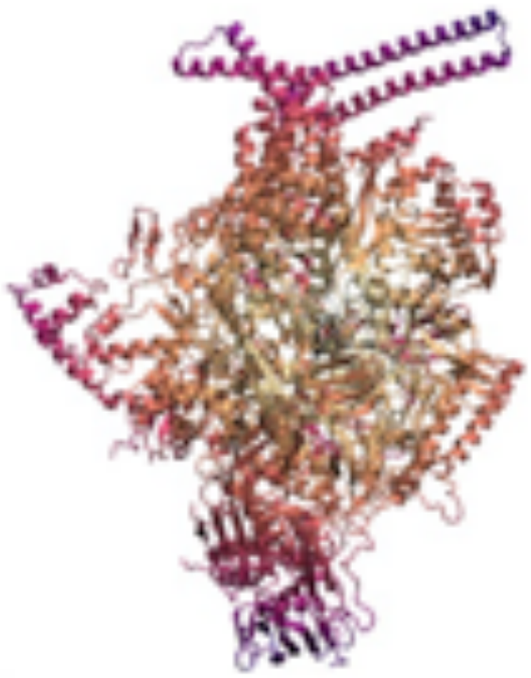
ExternalCode

- Secondary structure (Stride)
- Solvent accessible surface area (FreeSASA)
- Predictions in change to protein function (SNAP2)

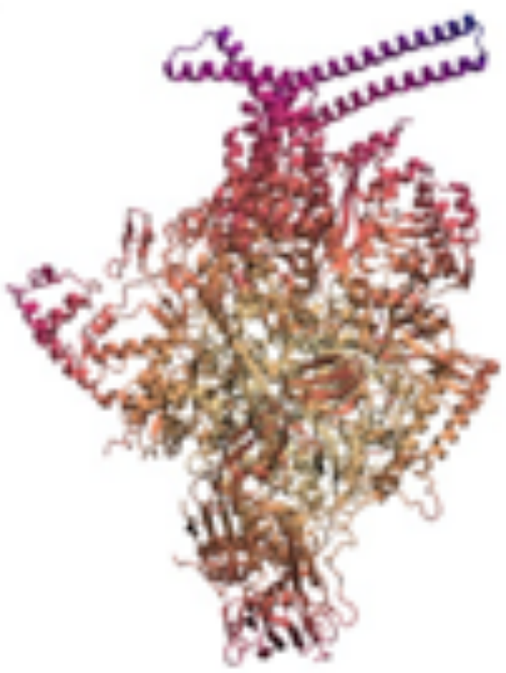
TrajectoryDistances

- Calculation of mean distances across MD simulation trajectories – in progress!

Structural distances in RNAP

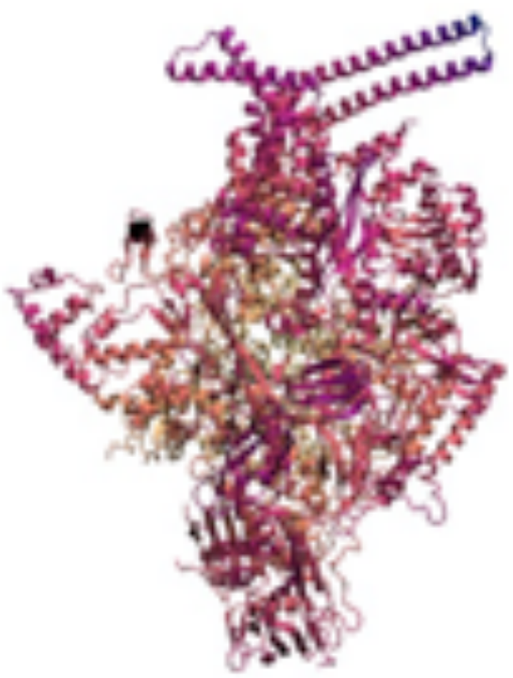


Distance from rifampicin
(very similar to mRNA)



Distance from Mg ion

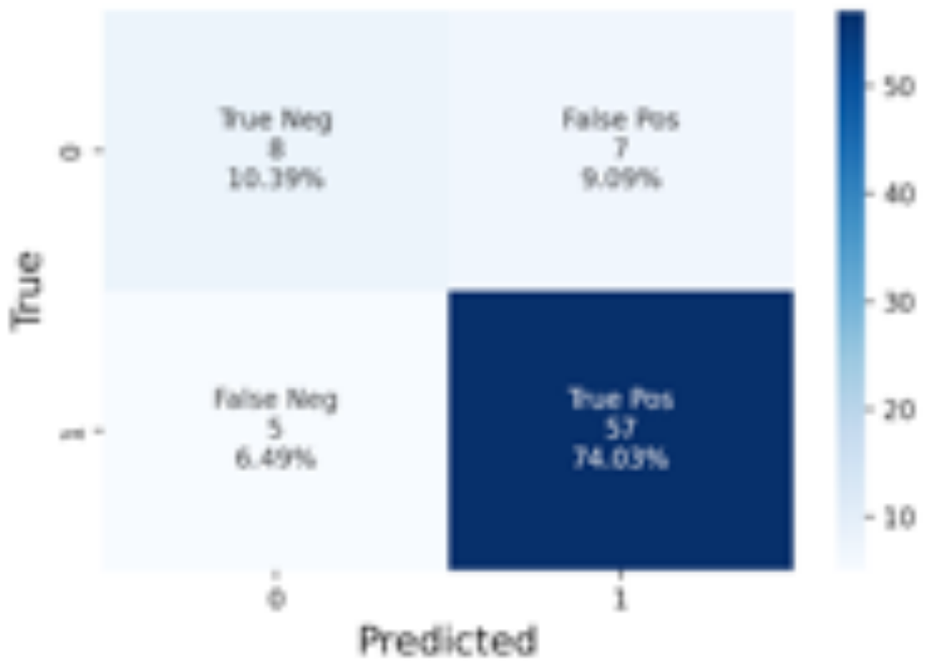
Distance from
peripheral zinc ion



Distance from
buried zinc ion

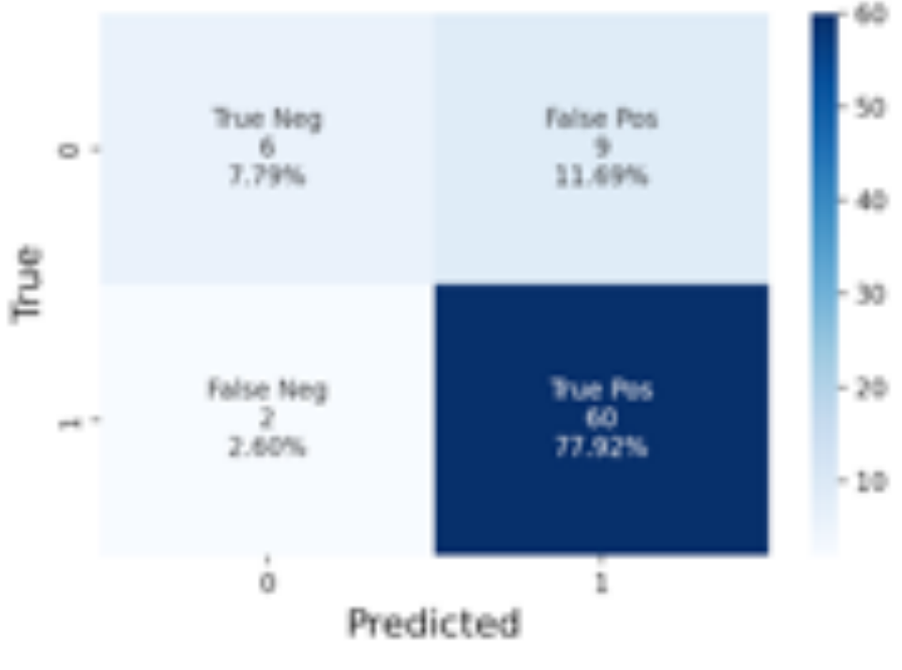
Logistic Regression

All features



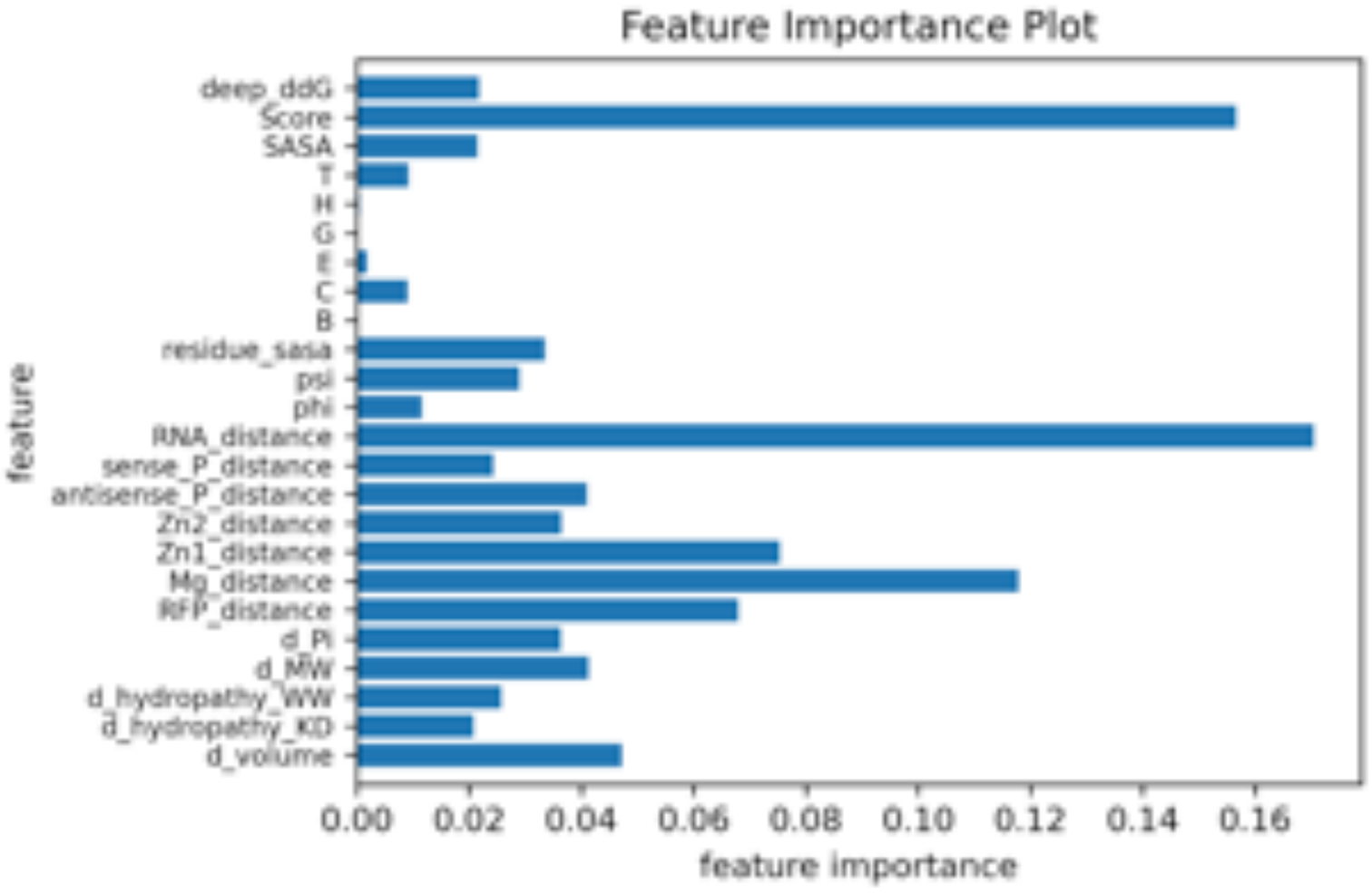
Precision: 0.89
Sensitivity: 0.92
Specificity: 0.53

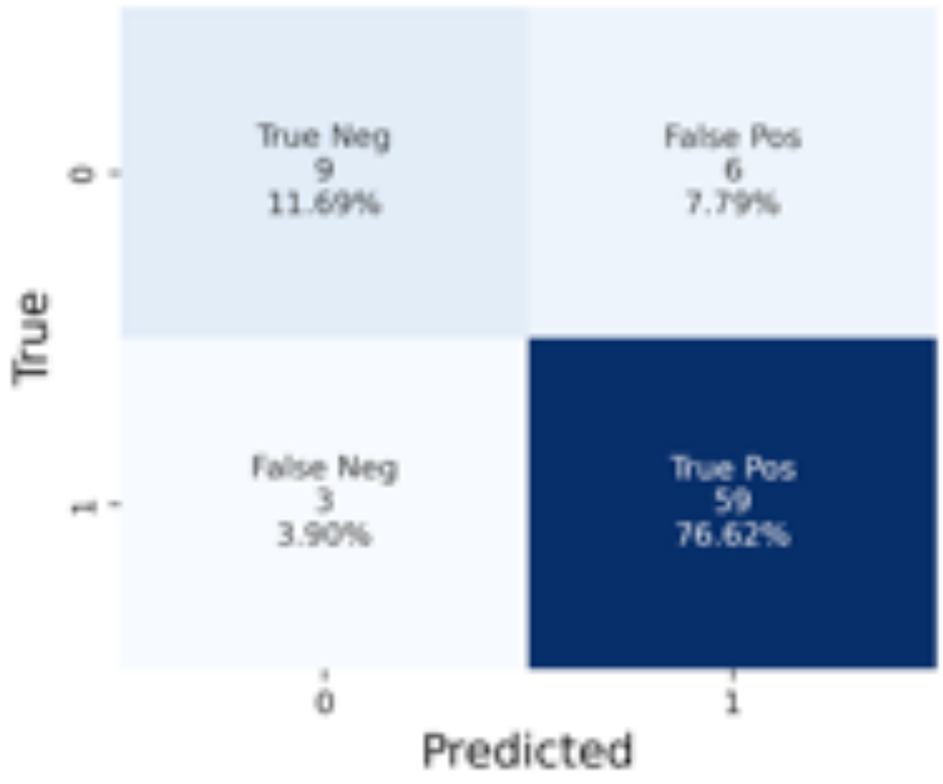
Reduced features



Precision: 0.87
Sensitivity: 0.97
Specificity: 0.40

Random Forest





Max depth: 6 Precision: 0.91 Sensitivity: 0.95 Specificity: 0.60

Predicting antibiotic resistance *de novo*

Machine Learning to **rule-out** large number of no-effect mutations

1

Relative binding free energies to **rule-in** remaining mutations that confer resistance

2

Ser388Leu

Rifampicin Resistance Determining Region

Val170Phe

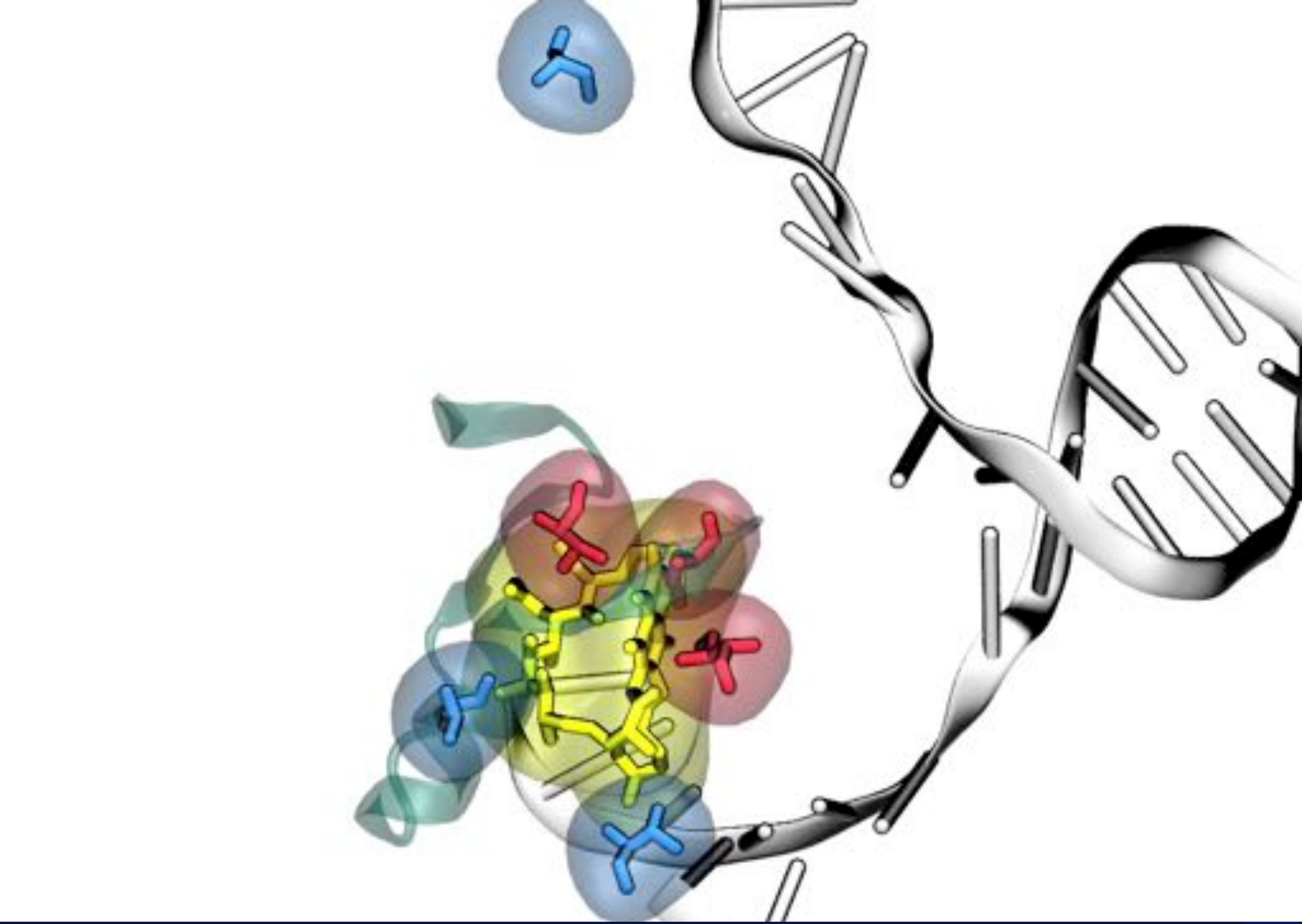
Ser450Leu

Leu443Phe

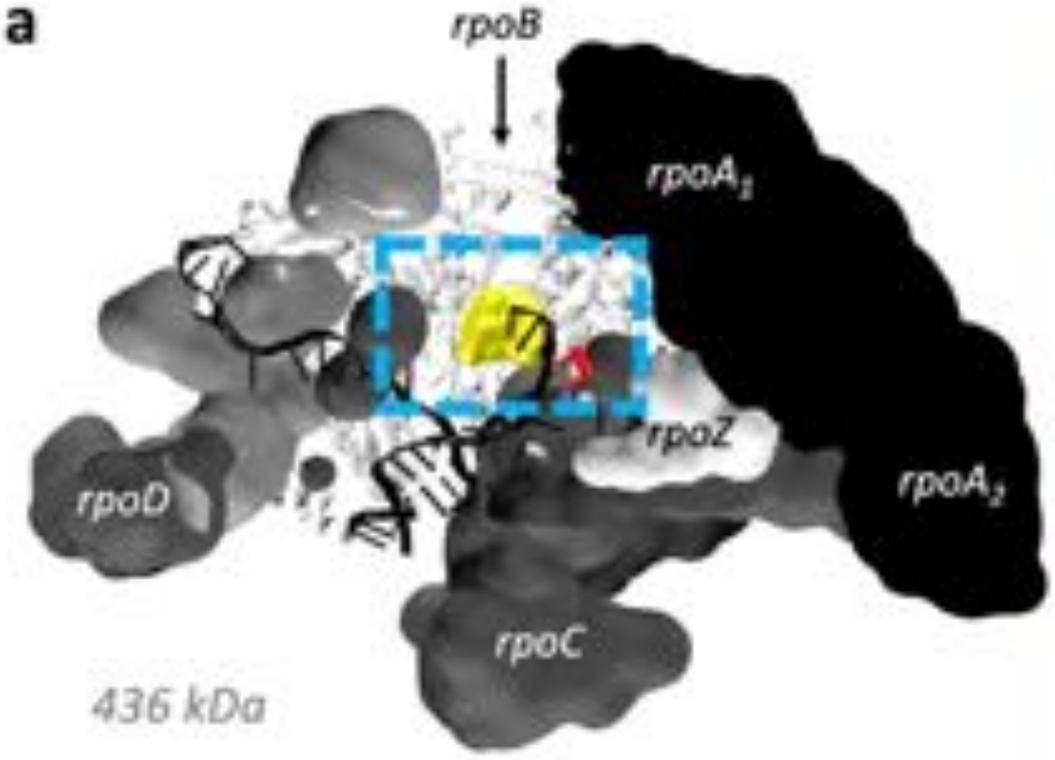
Rifampicin

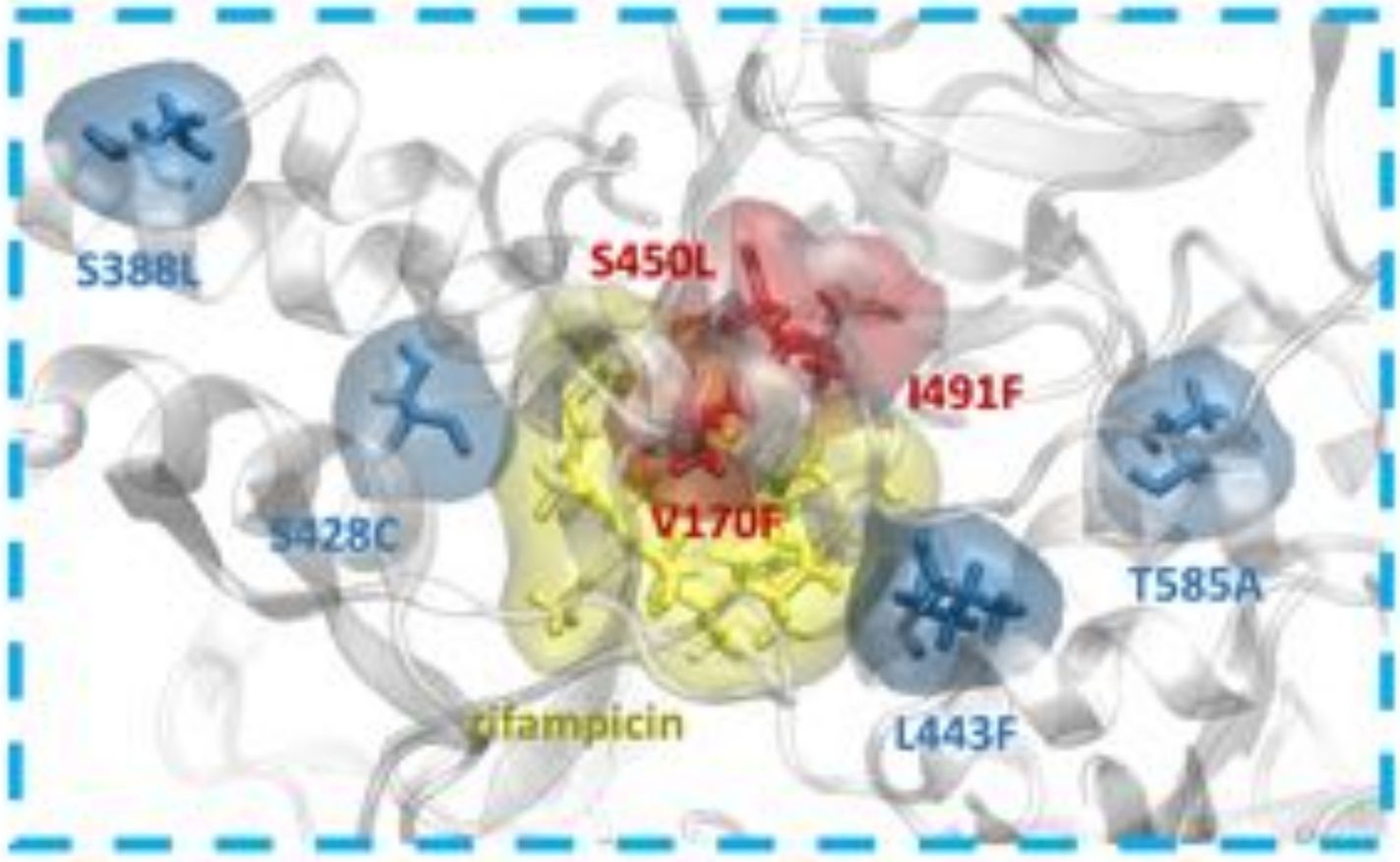
Ile491Phe

Thr585Ala

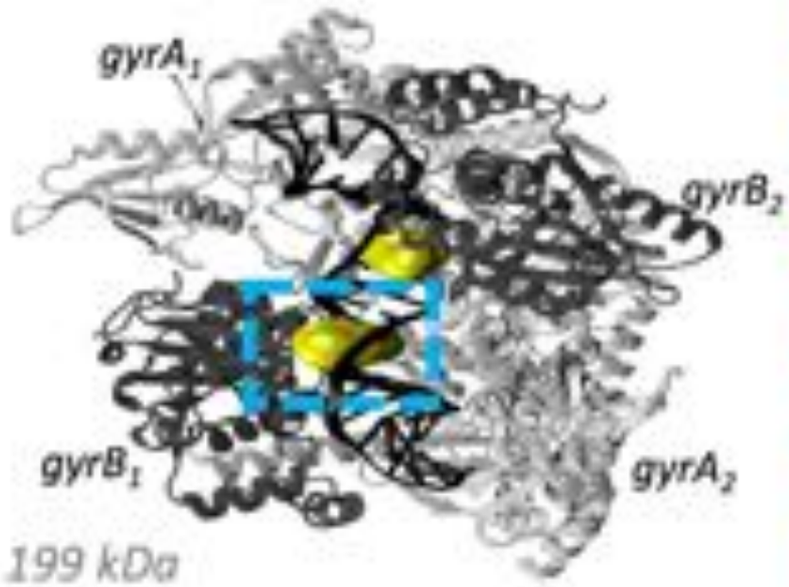


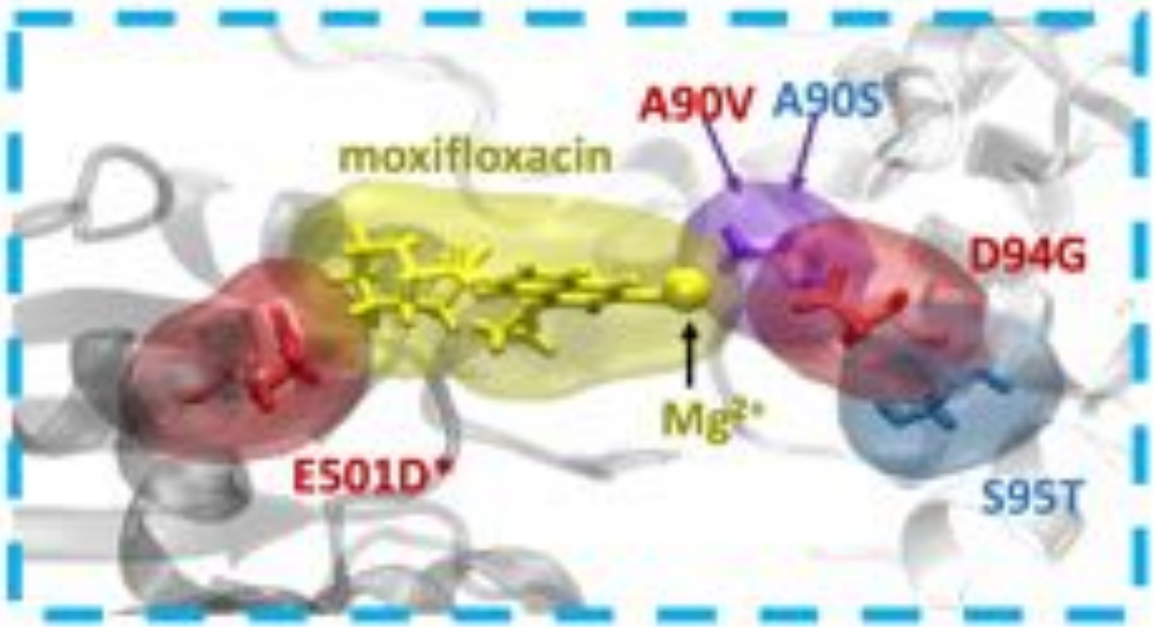
a


rpoB
rpoA1
rpoA2
rpoZ
rpoD
rpoC
436 kDa


S388L
S450L
I491F
S428C
V170F
T585A
rifampicin
L443F

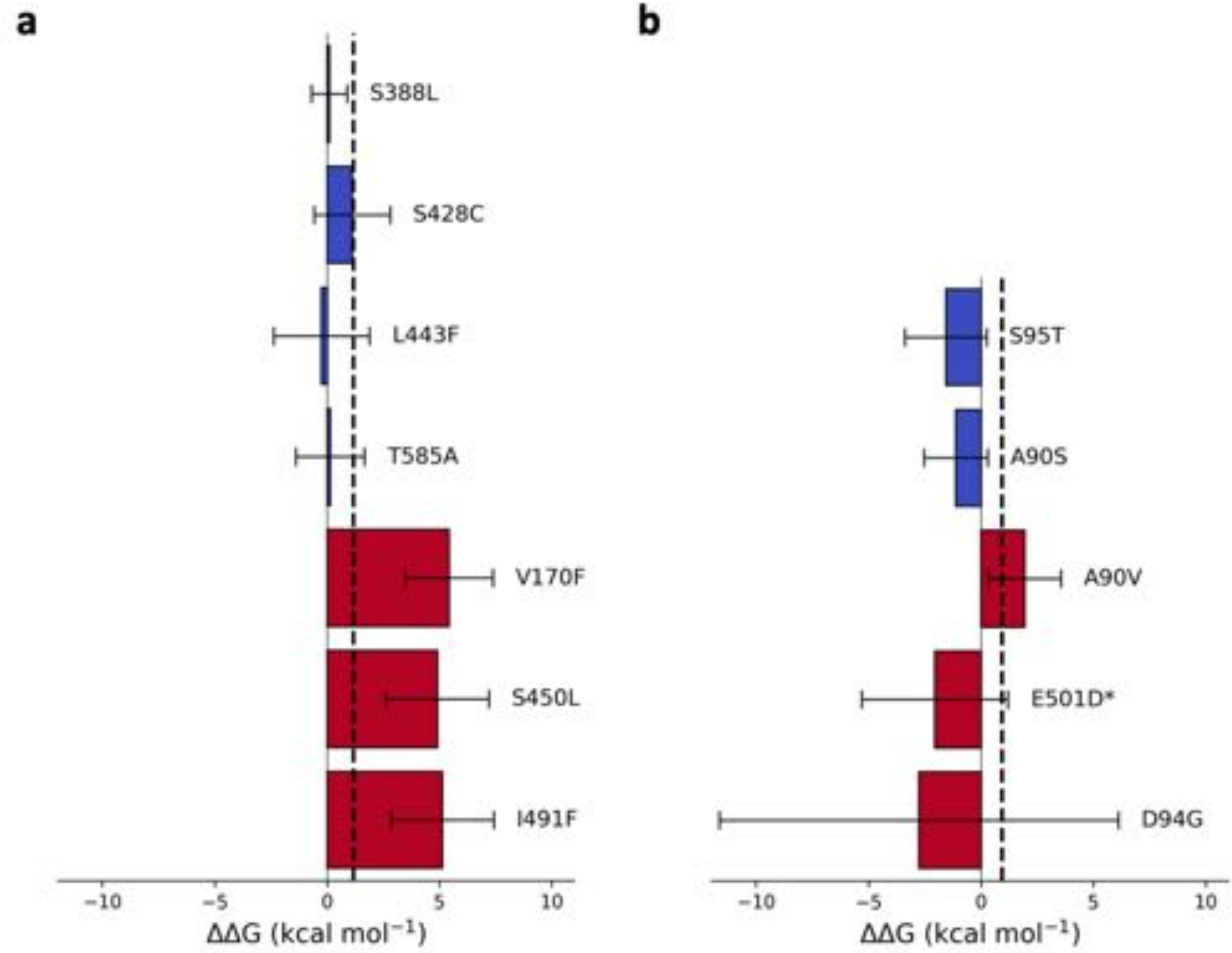
b

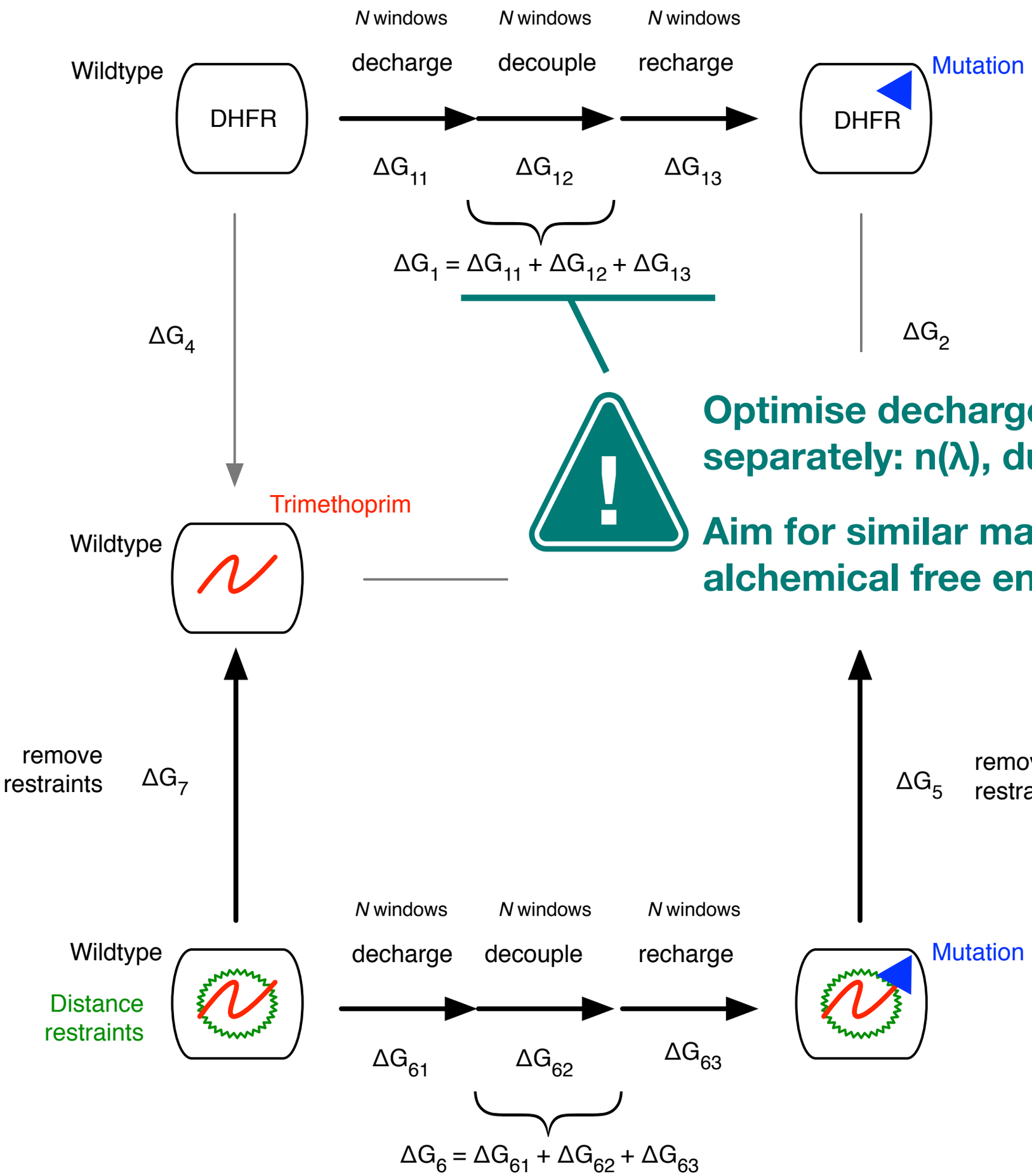

gyrA1
gyrB2
gyrB1
gyrA2
199 kDa


A90V
A90S
moxifloxacin
D94G
Mg2+
E501D
S95T

Rifampicin/RNA polymerase

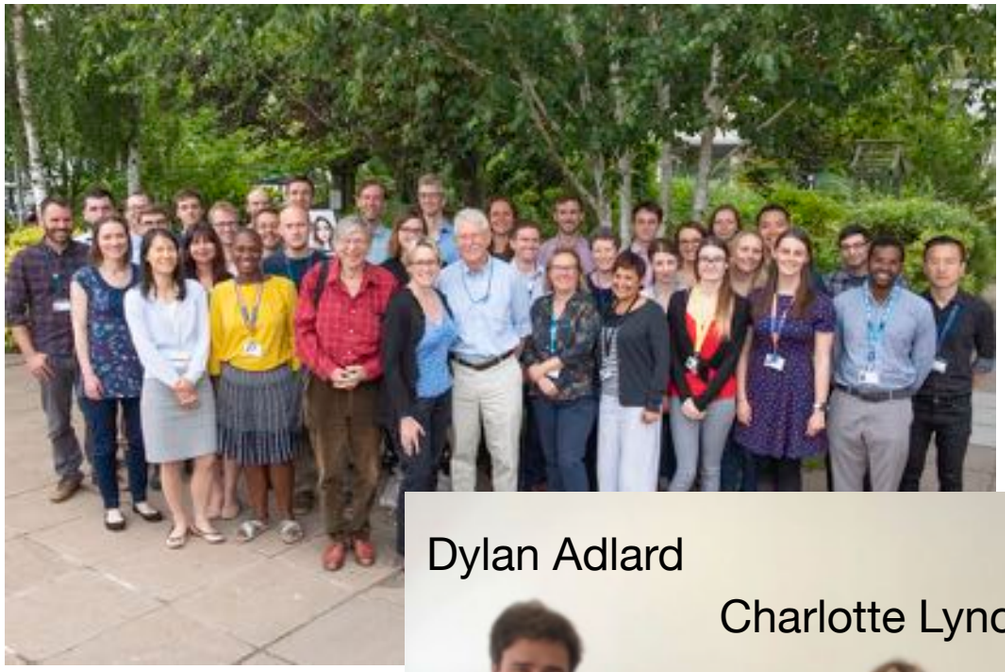
Fluroquinolones/DNA gyrase





Optimise decharge, decouple, recharge separately: n(λ), duration, number of sims

Aim for similar magnitude of error in all alchemical free energy calculations



Alice Brankin Matty Johnson Philip Fowler

https://fowlerlab.org
@_fowlerlab_
https://github.com/fowler-lab