Open Source AI/ML for infectious disease research

Open Research Forum, University of Reading
09.03.2022

Gemma Turon, gemma@ersilia.io
Ersilia Open Source Initiative
https://ersilia.io

Ersilia

I'm here today…

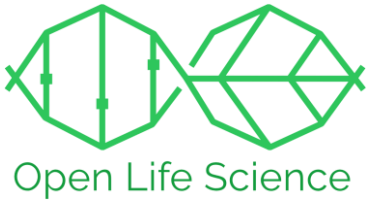
- Software Sustainability Fellowship
  - Improve computational practices in research software

    Reusability of AI/ML models for biomedical research

Prof. Al Edwards

- Open Life Sciences Program
  - Mentoring for Open Science Ambassadors

    Improve the documentation and accessibility of Ersilia's tools



- Digital Infrastructure Incubator
  - Supporting open source project leaders implementing best practices in sustainability, governance, and community health

    Community building tools and governance models

Ersilia

Edoardo Gaude, PhD

Co-founder & Trustee
Trained as molecular biologist at Cambridge University, UK
Co-founder of PockIt

Miquel Duran-Frigola, PhD

Co-founder & CSO
Trained as a computational chemist at IRB Barcelona, Spain

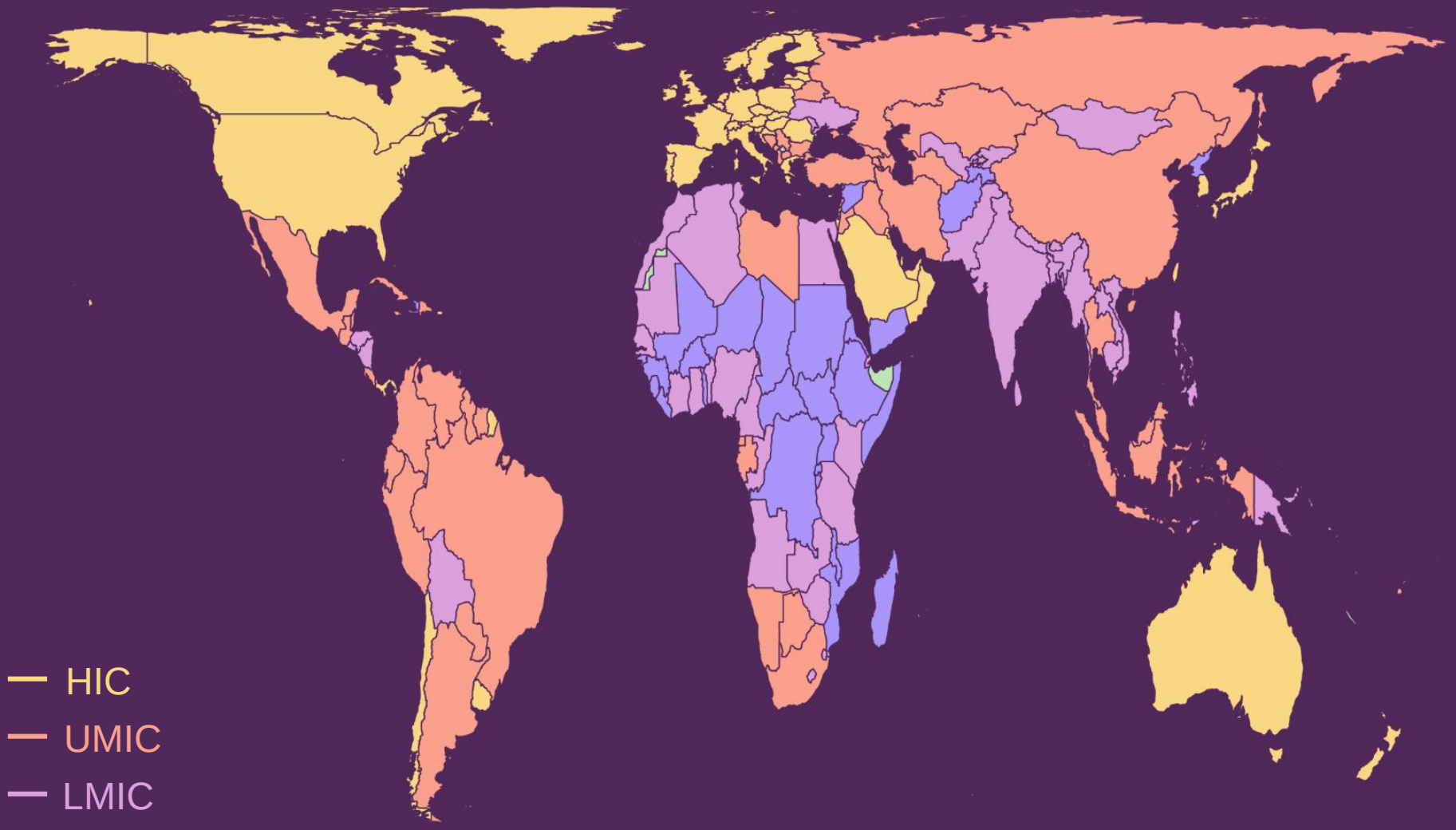
Gemma Turon, PhD

Co-founder & CEO
Trained as molecular biologist at IRB Barcelona, Spain

Our Mission

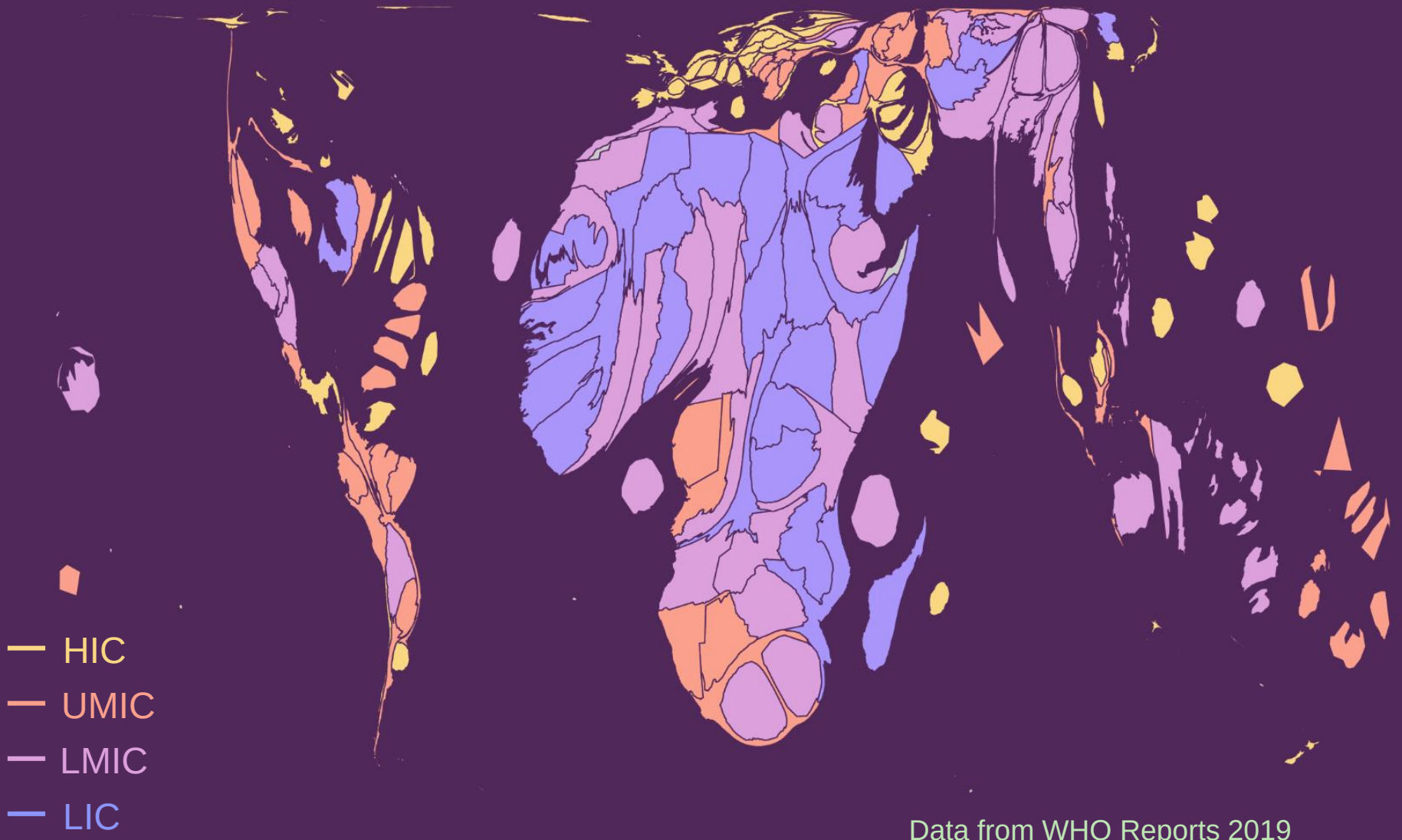
Strengthen the research capacity in Low and Middle Income Countries

Land area



Data from WHO Reports 2019

DALY – Communicable Diseases

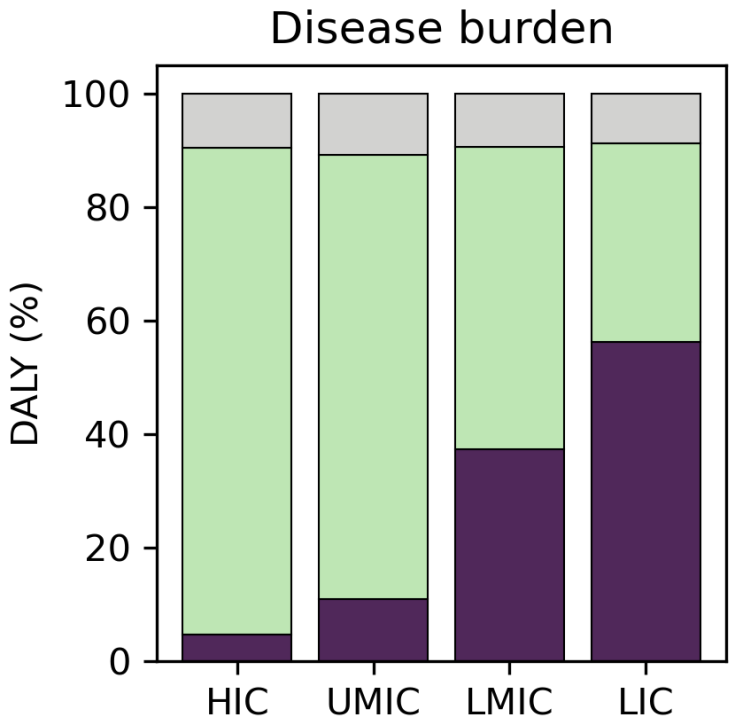


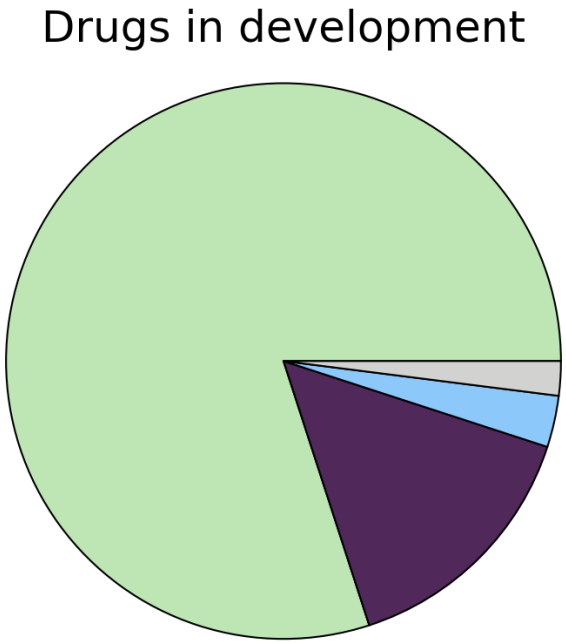
Data from WHO Reports 2019

Scientific Publications

Data from WHO Reports 2019

Western bias in Biomedical Research



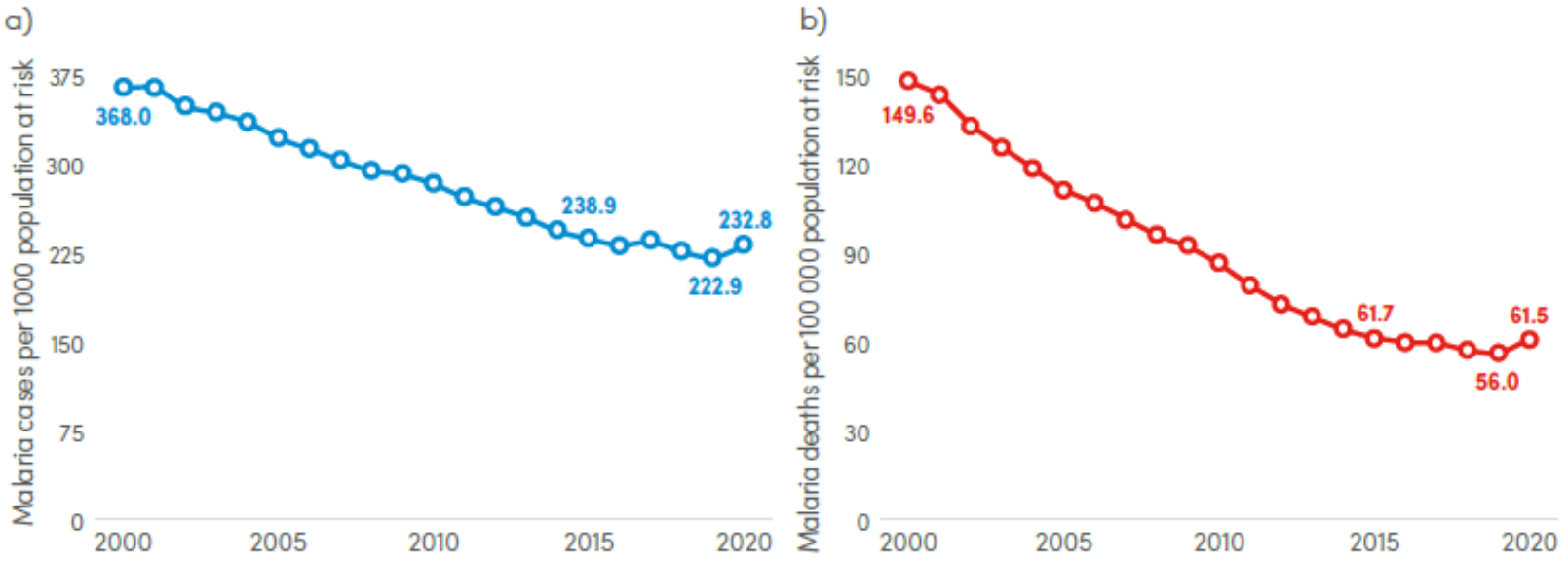


Non-communicable Communicable, maternal & neonatal Injuries Other

Western bias in Biomedical Research

Malaria

- Causes 0.5 million deaths a year (mostly amongst children)
- 95% of the new malaria cases are detected in Africa
- Resistance to front-line treatment (Artemisinin-combination therapies) is widespread in South-East Asia and first reports in Africa

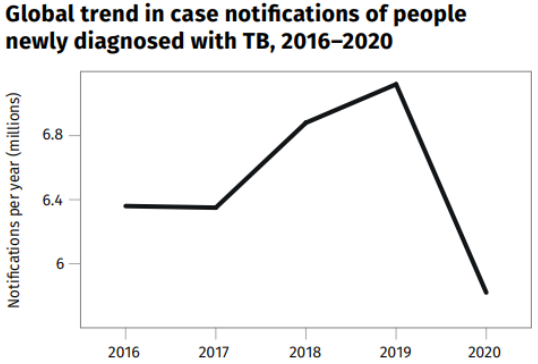


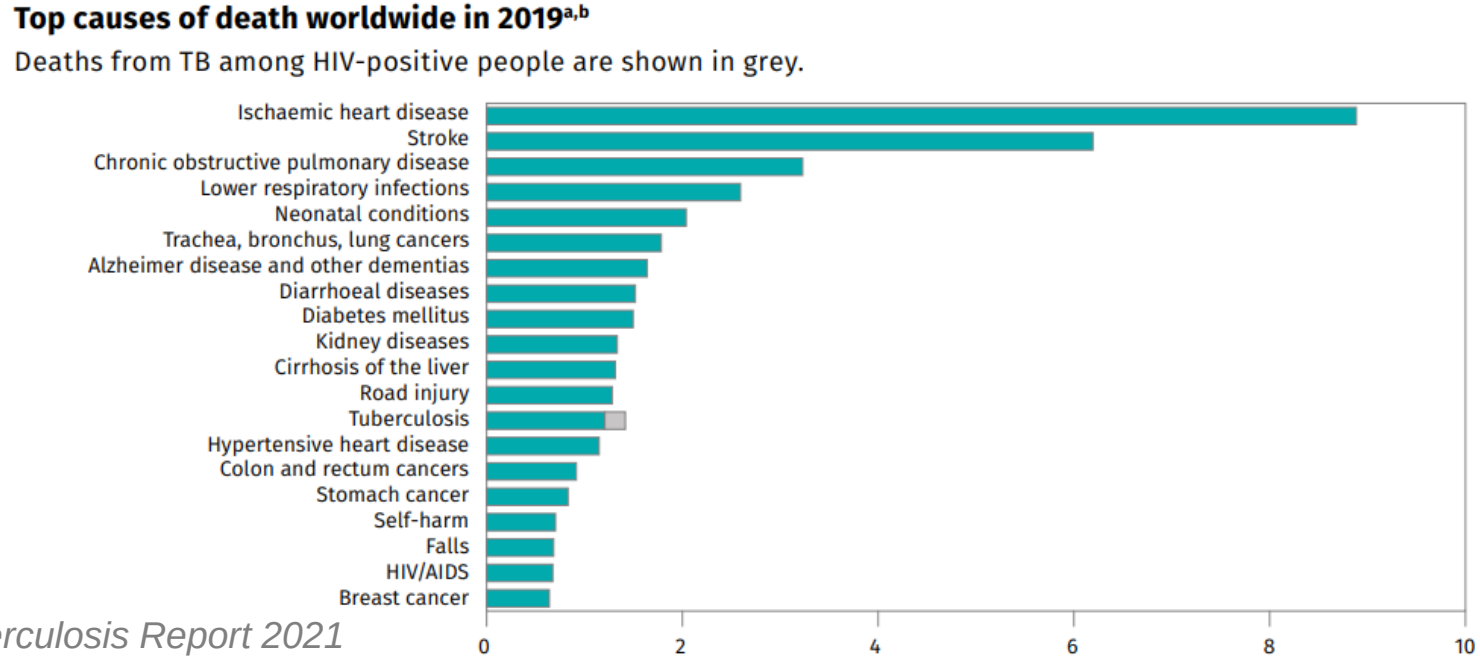
a) Malaria case incidence (per 1000 population at risk) b) mortality rate (deaths per 10000 population at risk) *WHO Malaria report 2021*

Western bias in Biomedical Research

Tuberculosis

- Leading cause of death from a single infectious agent (before COVID) – around 1, 1.5 million deaths per year (including HIV+ patients)
- 85% of the deaths occur in the WHO Africa and South-East Asia regions
- TB death incidence is back to levels of 2017





WHO Tuberculosis Report 2021

Free & Open Source

Real-time code sharing

Permissive licenses

No patents

Reproducibility

In-Country Research

Avoid “helicopter research”

Science led by local institutes

Implementation *in situ*

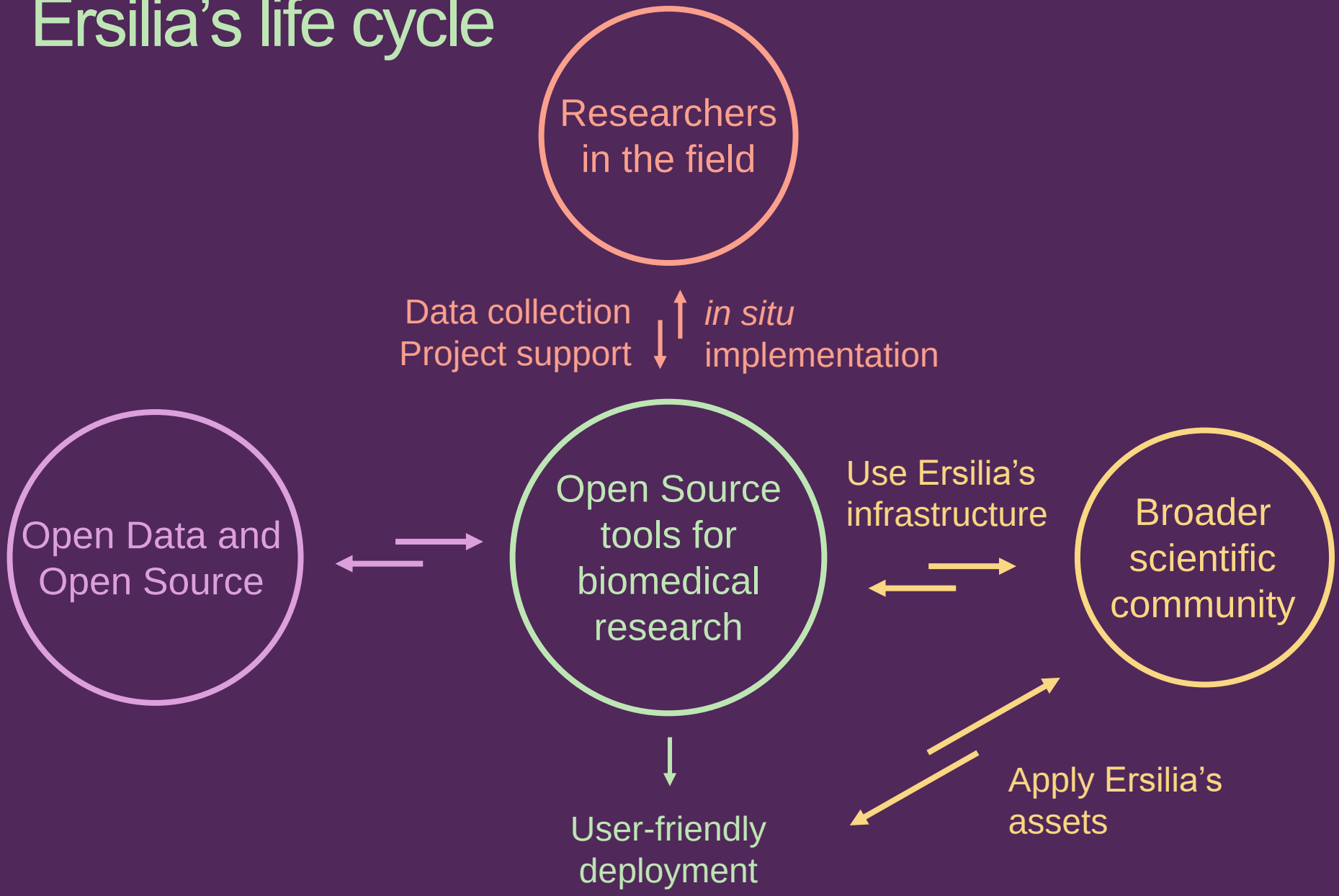
Sustainable Collaborations

Capacity building activities

Identify & train local champions

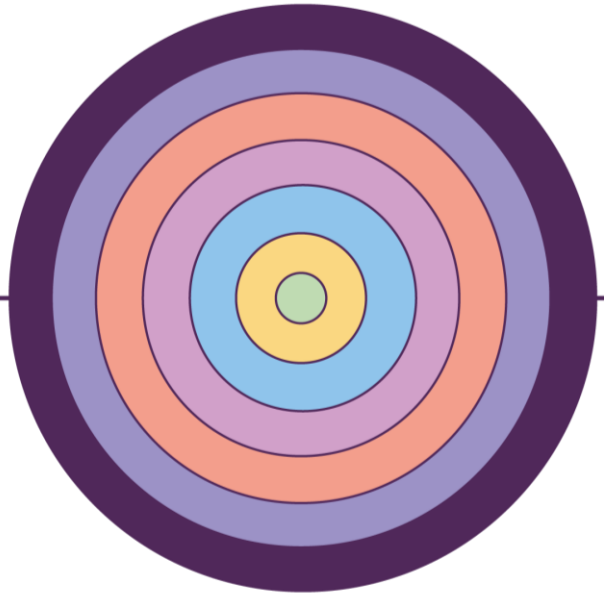
AI/ML with low resources

Ersilia's life cycle



Our goal: ready to use AI

Input:
e.g natural products



Output:
e.g activity against a pathogen

AI/ML from the literature

○ Ersilia “bundles” a model developed by others

Cell

Article

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes,[1,2,3] Kevin Yang,[3,4,10] Kyle Swanson,[3,4,10] Wengong Jin,[3,4] Andres Cubillos-Ruiz,[1,2,5] Nina M. Donghia,[1,6] Craig R. MacNair,[6] Shawn French,[5] Lindsey A. Carfrae,[7] Zohar Bloom-Ackerman,[3,7] Victoria M. Tran,[2] Anush Chiappino-Pepe,[3,7] Ahmed H. Badran,[1] Ian W. Andrews,[1,2,5] Emma J. Chory,[1,2] George M. Church,[3,7,8] Eric D. Brown,[6] Tommi S. Jaakkola,[3,4] Regina Barzilay,[3,4,9,*] and James J. Collins[1,2,5,8,9,11,*]

[1]Department of Biological Engineering, Synthetic Biology Center, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]Machine Learning for Pharmaceutical Discovery and Synthesis Consortium, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA
[6]Department of Biochemistry and Biomedical Sciences, Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON L8N 3Z5, Canada
[7]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[8]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA
[9]Abdul Latif Jameel Clinic for Machine Learning in Health, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)
https://doi.org/10.1016/j.cell.2020.01.021

SUMMARY

Due to the rapid emergence of antibiotic-resistant bacteria, there is a growing need to discover new antibiotics. To address this challenge, we trained a deep neural network capable of predicting molecules with antibacterial activity. We performed predictions on multiple chemical libraries and discovered a molecule from the Drug Repurposing Hub—halicin—that is structurally divergent from conventional antibiotics and displays bactericidal activity against a wide phylogenetic spectrum of pathogens including *Mycobacterium tuberculosis* and carbapenem-resistant Enterobacteriaceae. Halicin also effectively treated *Clostridioides difficile* and pan-resistant *Acinetobacter baumannii* infections in murine models. Additionally, from a discrete set of 23 empirically tested predictions from >107 million molecules curated from the ZINC15 database, our model identified eight antibacterial compounds that are structurally distant from known antibiotics. This work highlights the utility of deep learning approaches to expand our antibiotic arsenal through the discovery of structurally distinct antibacterial molecules.

INTRODUCTION

Since the discovery of penicillin, antibiotics have become the cornerstone of modern medicine. However, the continued efficacy of these essential drugs is uncertain due to the global dissemination of antibiotic-resistance determinants. Moreover, the decreasing development of new antibiotics in the private sector that has resulted from a lack of economic incentives is exacerbating this already dire problem (Brown and Wright, 2016; PEW Trusts, 2019). Indeed, without immediate action to discover and develop new antibiotics, it is projected that deaths attributable to resistant infections will reach 10 million per year by 2050 (O'Neill, 2014).

Historically, antibiotics were discovered largely through screening soil-dwelling microbes for secondary metabolites that prevented the growth of pathogenic bacteria (Clardy et al., 2006; Wright, 2017). This approach resulted in the majority of clinically used classes of antibiotics, including β-lactams, aminoglycosides, polymyxins, and glycopeptides, among others. Semi-synthetic derivatives of these scaffolds have maintained a viable clinical arsenal of antibiotics by increasing potency, decreasing toxicity, and sidestepping resistance determinants. Entirely synthetic antibiotics of the pyrimidine, quinolone, oxazolidinone, and sulfa classes have also found prolonged clinical utility, and continue to be optimized for the same properties.

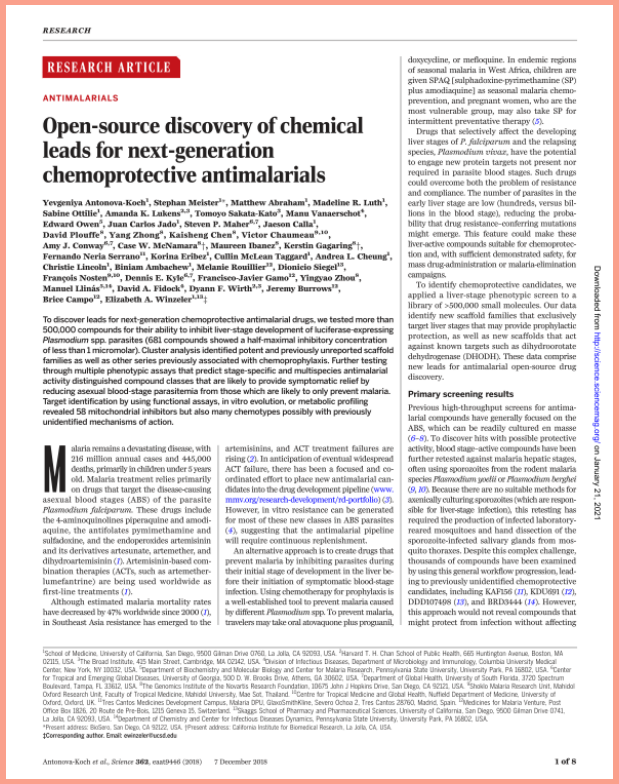
Unfortunately, the discovery of new antibiotics is becoming increasingly difficult. Natural product discovery is now plagued by the dereplication problem, wherein the same molecules are being repeatedly discovered (Cox et al., 2017). Moreover, given the rapid expansion of chemical spaces that are accessible by the derivatization of complex scaffolds (Ortholand and Ganesan, 2004), engineering next-generation versions of existing antibiotics results in substantially more failures than leads. Therefore, many antibiotic discovery programs have turned to screening large synthetic chemical libraries (Tommasi et al., 2015). However, these libraries, which can contain hundreds of thousands

Antibiotic activity *E.coli*
Stokes et al, 2020

Halicin | Active

In-House AI/ML

Ersilia trains an AI/ML model based on data



RESEARCH

RESEARCH ARTICLE

ANTIMALARIALS

Open-source discovery of chemical leads for next-generation chemoprotective antimalarials

Yevgeniya Antonova-Koch[1], Stephan Meister[2,*], Matthew Abraham[1], Madeline R. Luth[1], Sabine Ottilie[1], Amanda K. Lukens[3,†], Tomoyo Sakata-Kato[3], Manu Vanaerschot[4], Edward Owen[5], Juan Carlos Jado[1], Steven P. Maher[6,7], Jaeson Calla[1], David Plouffe[8], Yang Zhong[8], Kaisheng Chen[8], Victor Chaumeau[9,10], Amy J. Conway[6,7], Case W. McNamara[6,‡], Maureen Ibanez[2], Kerstin Gagaring[6,‡], Fernando Nerla Serrano[11], Korina Eribez[1], Caitlin McLean Taggard[1], Andrea L. Cheung[1], Christie Lincoln[1], Biniam Ambachew[1], Melanie Rouillier[12], Dionicio Siegel[13], François Nosten[9,10], Dennis E. Kyle[6,7], Francisco-Javier Gamo[14], Yingyao Zhou[8], Manuel Llinás[3,15], David A. Fidock[4], Dyann F. Wirth[2,†], Jeremy Burrows[12], Brice Campo[12], Elizabeth A. Winzeler[1,16,‡]

To discover leads for next-generation chemoprotective antimalarial drugs, we tested more than 500,000 compounds for their ability to inhibit liver-stage development of luciferase-expressing *Plasmodium* spp. parasites (681 compounds showed a half-maximal inhibitory concentration of less than 1 micromolar). Cluster analysis identified potent and previously unreported scaffold families as well as other series previously associated with chemoprophylaxis. Further testing through multiple phenotypic assays that predict stage-specific and multispecies antimalarial activity distinguished compound classes that are likely to provide symptomatic relief by reducing asexual blood-stage parasitemia from those which are likely to only prevent malaria. Target identification by using functional assays, in vitro evolution, or metabolic profiling revealed 58 mitochondrial inhibitors but also many chemotypes possibly with previously unidentified mechanisms of action.

Malaria remains a devastating disease, with 216 million annual cases and 445,000 deaths, primarily in children under 5 years old. Malaria treatment relies primarily on drugs that target the disease-causing asexual blood stages (ABS) of the parasite *Plasmodium falciparum*. These drugs include the 4-aminoquinolines piperaquine and amodiaquine, the antifolates pyrimethamine and sulfadoxine, and the endoperoxides artemisinin and its derivatives artesunate, artemether, and dihydroartemisinin (*1*). Artemisinin-based combination therapies (ACTs, such as artemether-lumefantrine) are being used worldwide as first-line treatments (*1*).

Although estimated malaria mortality rates have decreased by 47% worldwide since 2000 (*1*), in Southeast Asia resistance has emerged to the artemisinins, and ACT treatment failures are rising (*2*). In anticipation of eventual widespread ACT failure, there has been a focused and coordinated effort to place new antimalarial candidates into the drug development pipeline (www.mmv.org/research-development/rd-portfolio) (*3*). However, in vitro resistance can be generated for most of these new classes in ABS parasites (*4*), suggesting that the antimalarial pipeline will require continuous replenishment.

An alternative approach is to create drugs that prevent malaria by inhibiting parasites during their initial stage of development in the liver before their initiation of symptomatic blood-stage infection. Using chemotherapy for prophylaxis is a well-established tool to prevent malaria caused by different *Plasmodium* spp. To prevent malaria, travelers may take oral atovaquone plus proguanil, doxycycline, or mefloquine. In endemic regions of seasonal malaria in West Africa, children are given SPAQ [sulphadoxine-pyrimethamine (SP) plus amodiaquine] as seasonal malaria chemoprevention, and pregnant women, who are the most vulnerable group, may also take SP for intermittent preventative therapy (*5*).

Drugs that selectively affect the developing liver stages of *P. falciparum* and the relapsing species, *Plasmodium vivax*, have the potential to engage new protein targets not present nor required in parasite blood stages. Such drugs could overcome both the problem of resistance and compliance. The number of parasites in the early liver stage are low (hundreds, versus billions in the blood stage), reducing the probability that drug resistance–conferring mutations might emerge. This feature could make these liver-active compounds suitable for chemoprotection and, with sufficient demonstrated safety, for mass drug-administration or malaria-elimination campaigns.

To identify chemoprotective candidates, we applied a liver-stage phenotypic screen to a library of >500,000 small molecules. Our data identify new scaffold families that exclusively target liver stages that may provide prophylactic protection, as well as new scaffolds that act against known targets such as dihydroorotate dehydrogenase (DHODH). These data comprise new leads for antimalarial open-source drug discovery.

Primary screening results

Previous high-throughput screens for antimalarial compounds have generally focused on the ABS, which can be readily cultured en masse (*6–8*). To discover hits with possible protective activity, blood stage–active compounds have been further retested against malaria hepatic stages, often using sporozoites from the rodent malaria species *Plasmodium yoelii* or *Plasmodium berghei* (*9, 10*). Because there are no suitable methods for axenically culturing sporozoites (which are responsible for liver-stage infection), this retesting has required the production of infected laboratory-reared mosquitoes and hand dissection of the sporozoite-infected salivary glands from mosquito thoraxes. Despite this complex challenge, thousands of compounds have been examined by using this general workflow progression, leading to previously unidentified chemoprotective candidates, including KAF156 (*11*), KDU691 (*12*), DDD107498 (*13*), and BRD3444 (*14*). However, this approach would not reveal compounds that might protect from infection without affecting

[1]School of Medicine, University of California, San Diego, 9500 Gilman Drive 0760, La Jolla, CA 92093, USA. [2]Harvard T. H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA. [3]The Broad Institute, 415 Main Street, Cambridge, MA 02142, USA. [4]Division of Infectious Diseases, Department of Microbiology and Immunology, Columbia University Medical Center, New York, NY 10032, USA. [5]Department of Biochemistry and Molecular Biology and Center for Malaria Research, Pennsylvania State University, University Park, PA 16802, USA. [6]Center for Tropical and Emerging Global Diseases, University of Georgia, 500 D. W. Brooks Drive, Athens, GA 30602, USA. [7]Department of Global Health, University of South Florida, 3720 Spectrum Boulevard, Tampa, FL 33612, USA. [8]The Genomics Institute of the Novartis Research Foundation, 10675 John J Hopkins Drive, San Diego, CA 92121, USA. [9]Shoklo Malaria Research Unit, Mahidol Oxford Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand. [10]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [11]Tres Cantos Medicines Development Campus, Malaria DPU, GlaxoSmithKline, Severo Ochoa 2, Tres Cantos 28760, Madrid, Spain. [12]Medicines for Malaria Venture, Post Office Box 1826, 20 Route de Pre-Bois, 1215 Geneva 15, Switzerland. [13]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive 0741, La Jolla, CA 92093, USA. [14]Department of Chemistry and Center for Infectious Diseases Dynamics, Pennsylvania State University, University Park, PA 16802, USA. [†]Present address: BioSero, San Diego, CA 92122, USA. [‡]Present address: California Institute for Biomedical Research, La Jolla, CA, USA. *Corresponding author. Email: ewinzeler@ucsd.edu

Chemoprotective antimalarials
Antonova-Koch et al, 2018

Atovaquone analog | Active

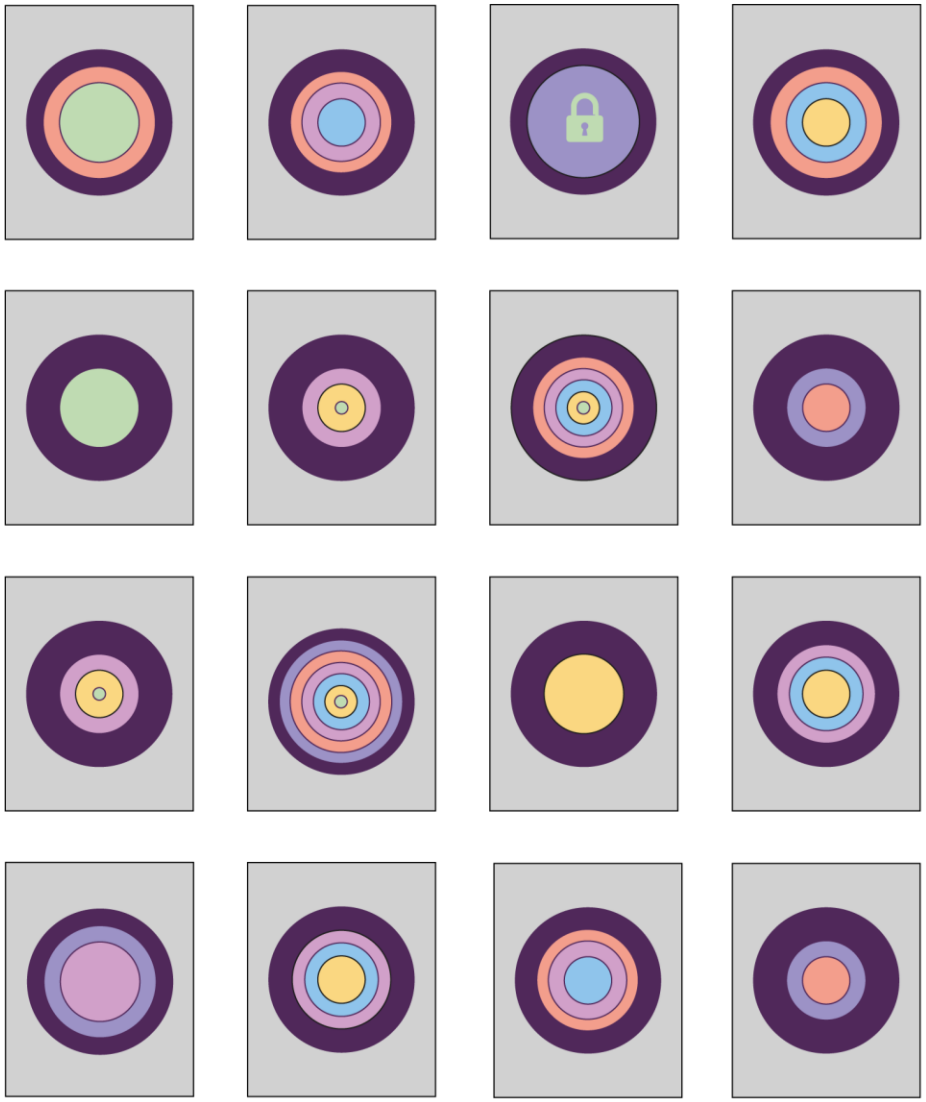
AI/ML in collaboration

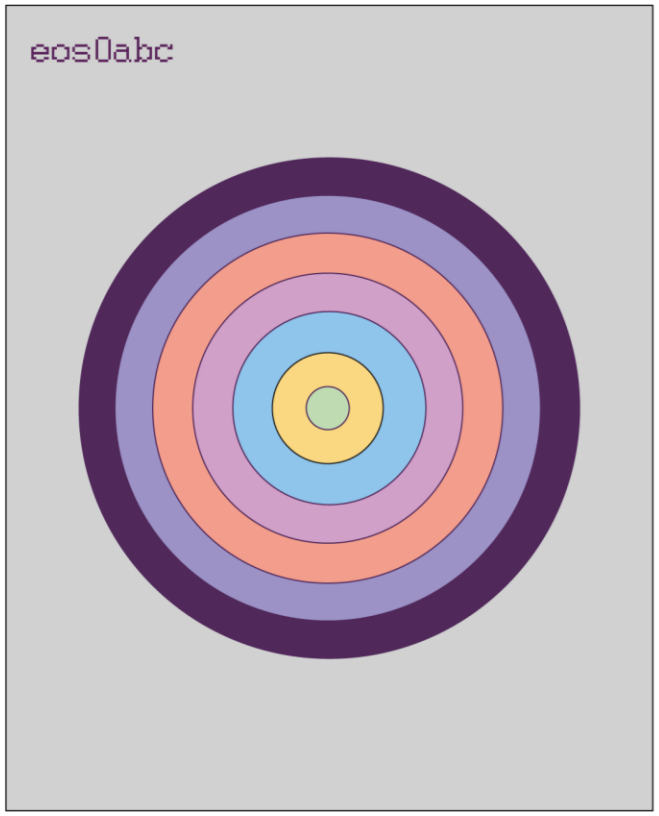
- Ersilia trains an AI/ML model based on partner's data

Your awesome project
You and Ersilia, 202

Your question | Our answer

The Ersilia Model Hub





User Query

Active

Author
GitHub repository
Summary / Applicability

The Ersilia Model Hub – How to

https://github.com/ersilia-os/ersilia

1. Ersilia installation in local computer
2. Selection of model of interest:

   40 publicly available models – browsable catalog
3. Use a command line interface to download model from our repository
4. Select the model api (predict, calculate…) and input the molecule (or list of molecules) of interest
5. Close model

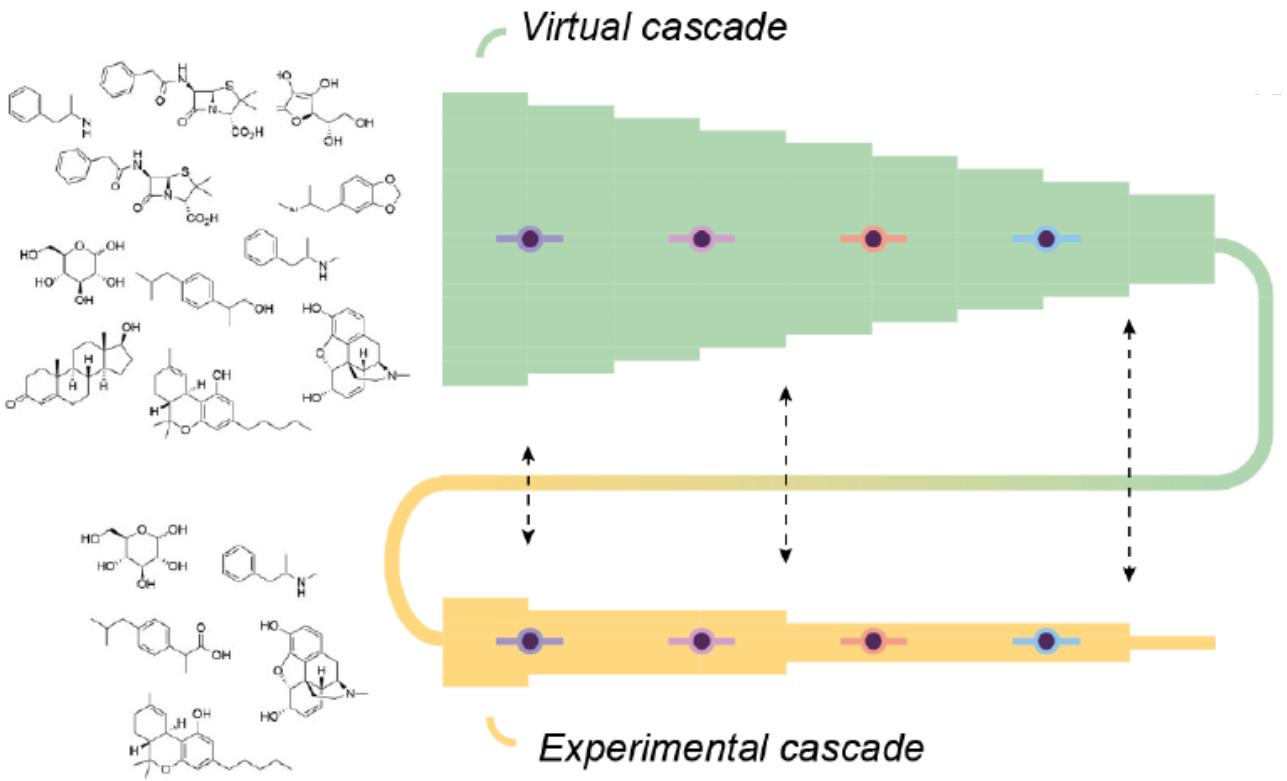
```
ersilia fetch chemprop-antibiotic
ersilia api predict -i "C1=C(SC(=N1)SC2=NN=C(S2)N)[N+](=O)[O-]"
ersilia close
```

**Disclaimer: the EMH is in testing mode*

Applications of the Ersilia Model Hub

Implementation of a virtual drug screening cascade

- Where: H3D Centre, Cape Town (South Africa)
- What: AI/ML modelling of drug screening assays for antimalarial and antituberculosis drug discovery



Compound prioritization for synthesis

Reduction of attrition rates

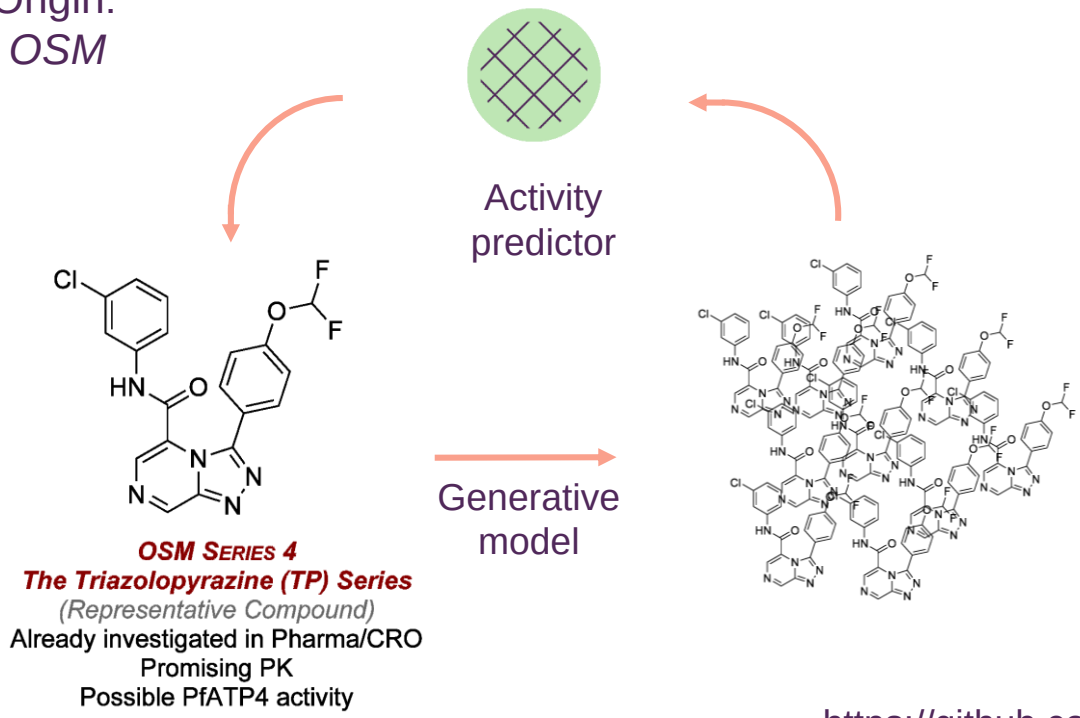
De novo design of new candidates

Applications of the Ersilia Model Hub

Generation of new antimalarial leads

- Where: Open Source Malaria Consortium (Prof. Todd, UCL)
- What: computer-based optimization of a chemical series with potent activity against malaria

Chemical Series Origin:
Pfizer → MMV → OSM



https://github.com/opensourcemalaria

Open Source Malaria Consortium

Ersilia's Roadmap

November 2020

Charity Incorporation

Today

First release of the EMH

Early adopters

End of 2022

Stable version of EMH

Users in several institutions

OS Contributors

End of 2023

Serving >1000 beneficiaries

Established OS repository for biomedicine

Application our our tools to other fields

User experience

- Testing, debugging and improving CLI
- Offer online predictions via cloud services
- Adding a user interface

Scalability

- Zairachem: an automated end-to-end ML pipeline for chemistry
- Model encryption to leverage IP-sensitive datasets (Merck funded)
- Facilitating third party model deposition

Dissemination

- Seminars and conferences
- Scientific publications
- Implementation *in situ* with our partners
- Training and workshops in underserved countries

Ersilia and Openness

Strengthen the research capacity in Low and Middle Income Countries by developing and deploying AI/ML tools in collaboration with scientists in low-resourced settings

- Make assets already developed more accessible
- Re-use published data & encourage sharing of private datasets
- Avoid reinventing the wheel
- Escaling collaborations between scientists from different institutes
- We are also working with low resources – Open Access, Open Source
- Find alternatives to the traditional drug discovery models, particularly in diseases with low revenue (MMV, DNDi, OSM, M4iD)

https://opencollective.com/ersilia https://twitter.com/ersiliaio https://github.com/ersilia-os

https://ersilia.io hello@ersilia.io https://medium.com/ersiliaio

Ersilia and Openness

How do we try to ensure we stay in the Open Science domain?

- Incorporated as a non-profit organisation
- We accompany our projects with training and dissemination activities to ensure open means accessible
- At the organizational level:
  - Open Code via repositories
  - Real-time financial status
  - Grant applications disclosed
  - Governance and strategic decision making – wip
  - Identifying avenues to work with proprietary data for the public benefit - wip

https://opencollective.com/ersilia https://twitter.com/ersiliaio https://github.com/ersilia-os

https://ersilia.io hello@ersilia.io https://medium.com/ersiliaio

Take Home Messages

- We are a Non Profit Organisation with the mission to democratise access to AI/ML tools for biomedical research
- We are building an international network of collaborators
- We combine remote working and on-site project development and capacity building
- All our assets are open-source
- We work at the intersection between academia, start-ups and pharmaceutical companies
- We **welcome** new contributors and collaborators

Open Source AI/ML for infectious disease research

Open Research Forum, University of Reading
09.03.2022

Gemma Turon, gemma@ersilia.io
Ersilia Open Source Initiative
https://ersilia.io

Ersilia