Ersilia, a hub of AI/ML tools for neglected tropical diseases

FOSDEM, 5th February 2022

Gemma Turon, PhD, @TuronGemma
Ersilia Open Source Initiative, @ersiliaio
http://ersilia.io

Ersilia

Ersilia

Edoardo Gaude, PhD

Co-Founder & Trustee
Trained as a molecular biologist
at Cambridge University, UK.
Co-Founder of POCKiT.

Gemma Turon, PhD

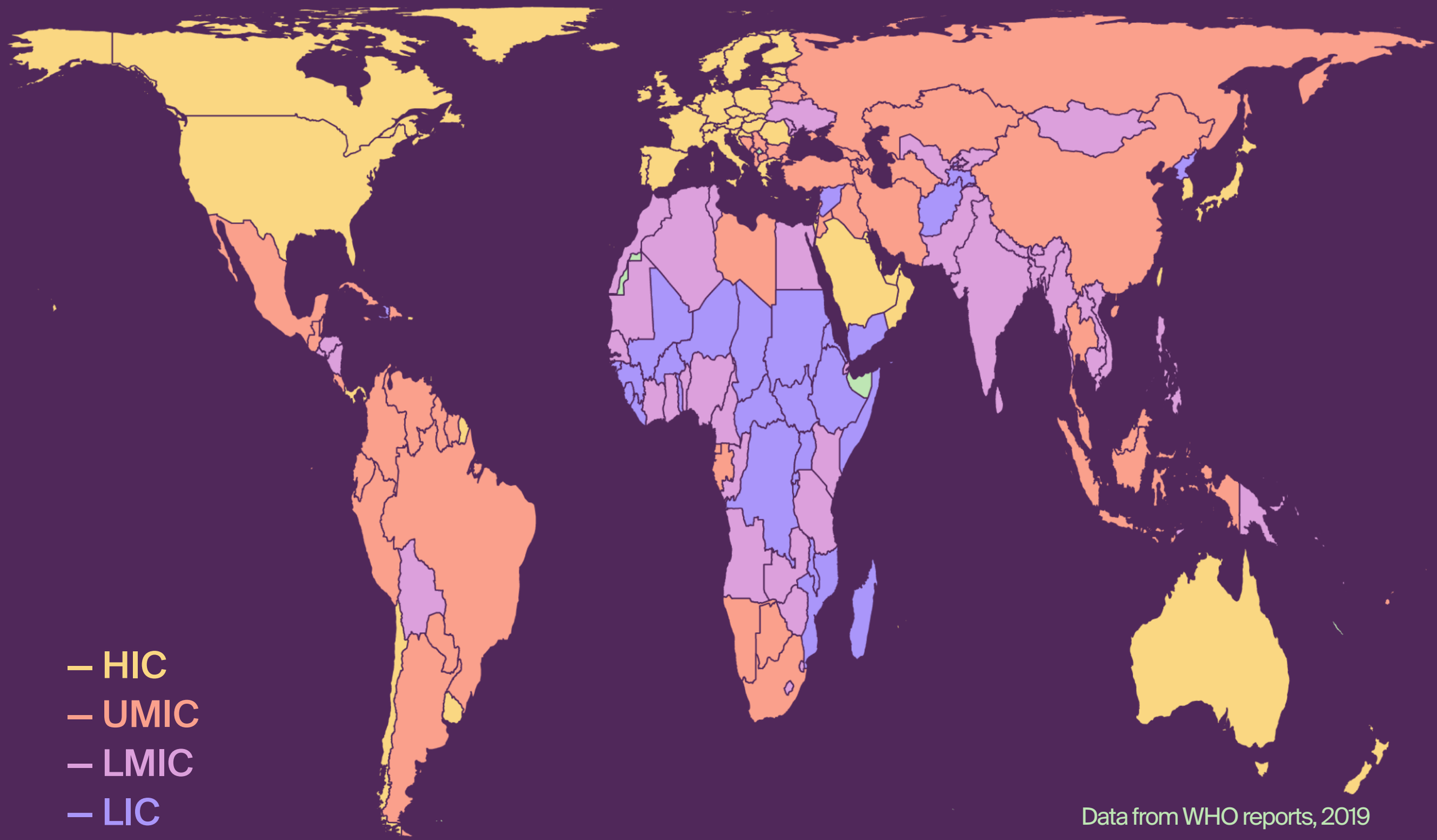
Co-Founder & CEO
Trained as a cell biologist
at IRB Barcelona, Spain.

Miquel Duran-Frigola, PhD

Co-Founder & CSO
Trained as computational chemist
at IRB Barcelona, Spain.

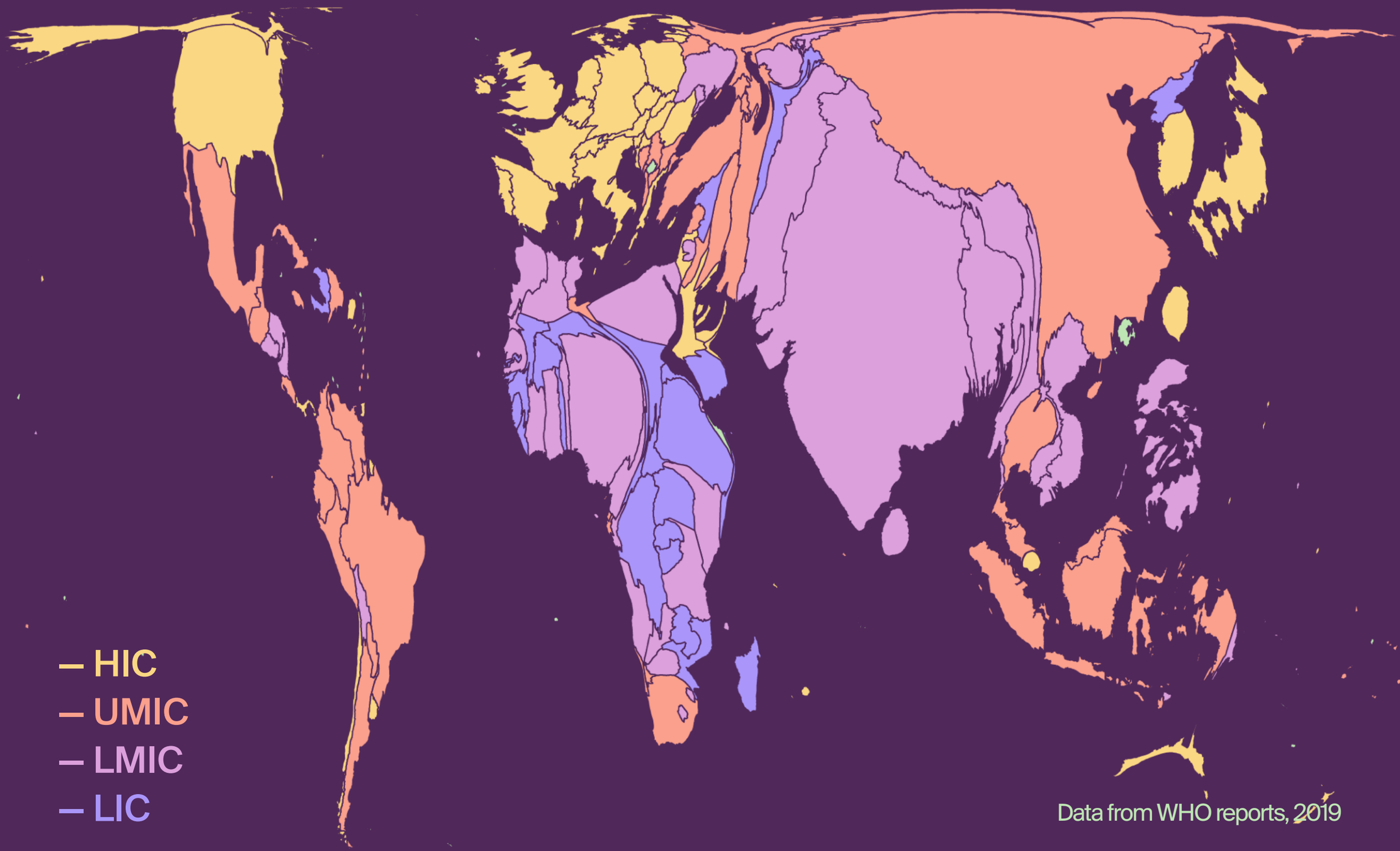
Charity Commission for England and Wales
November, 2020

Land area



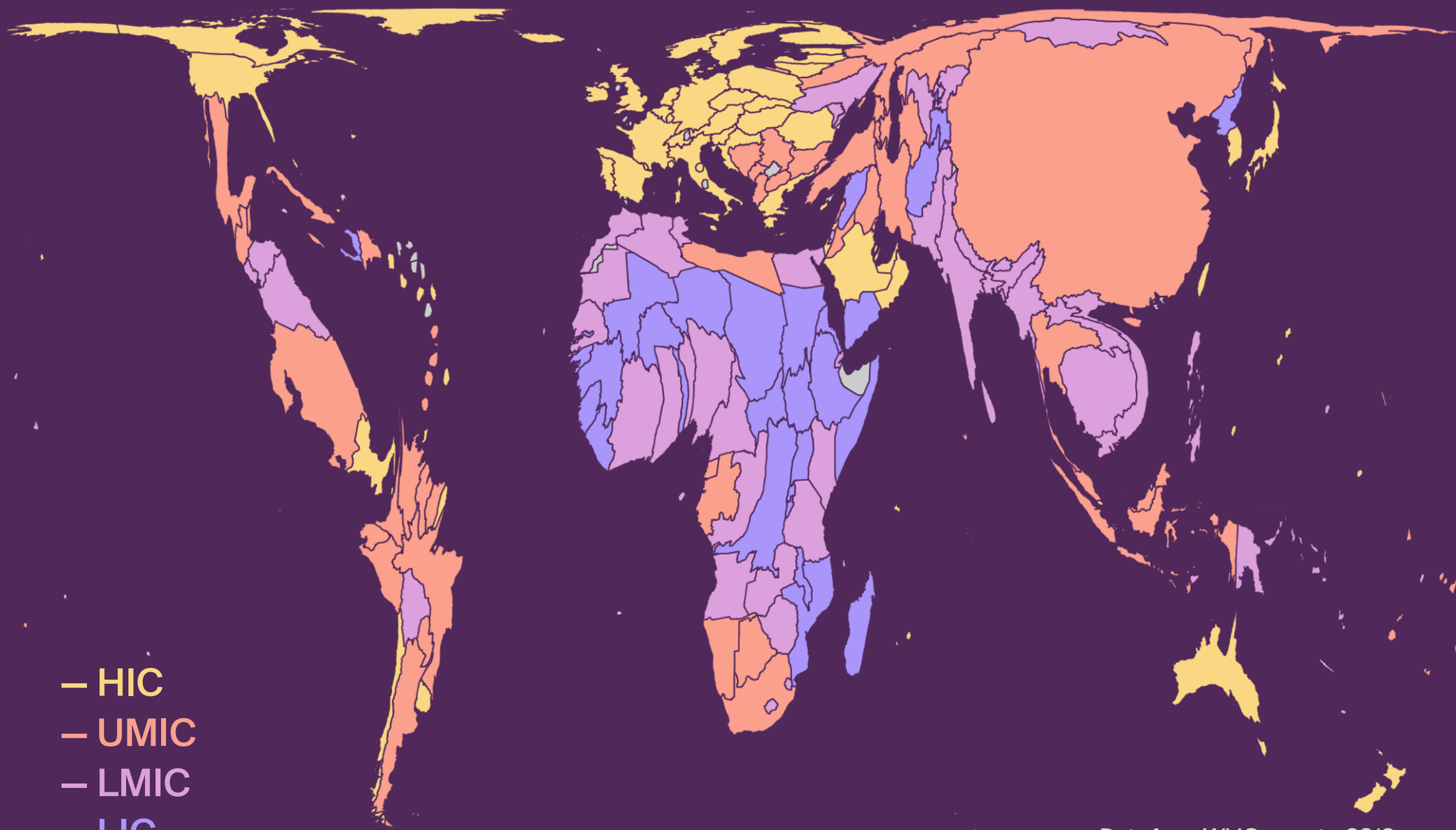
Data from WHO reports, 2019

Total population



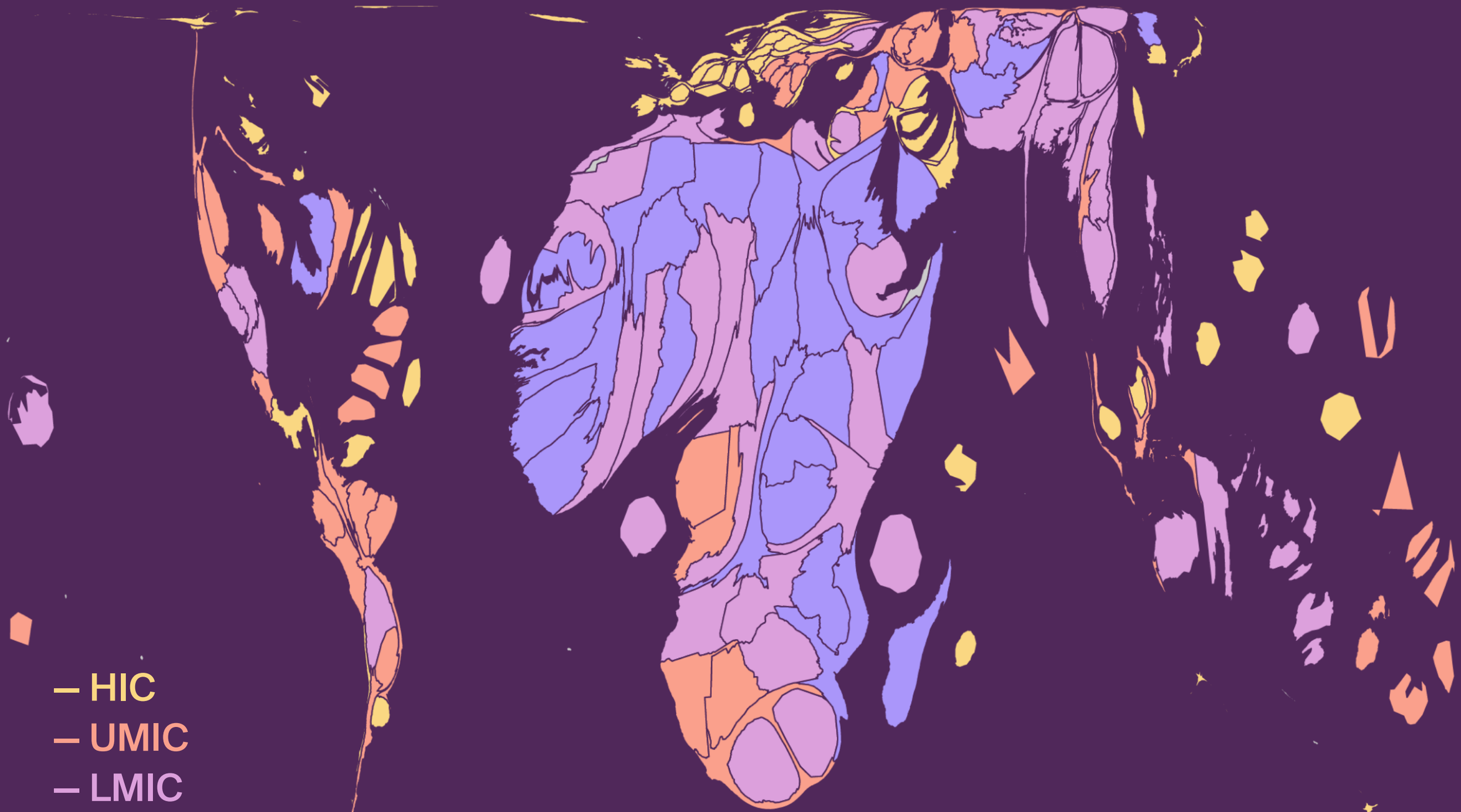
Data from WHO reports, 2019

DALY (premature death & years lived with disability)



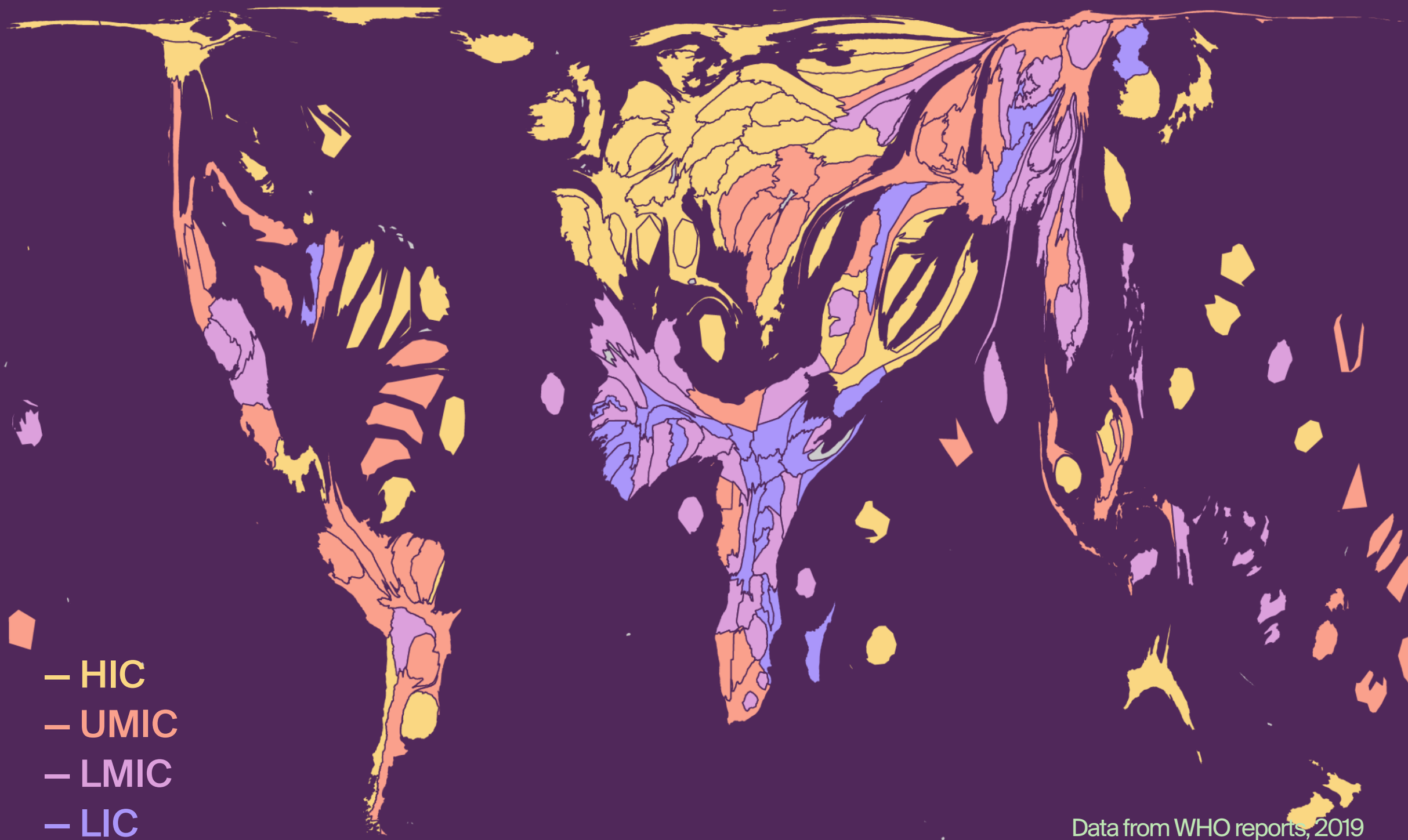
Data from WHO reports, 2019

Communicable disease DALY



Data from WHO reports, 2019

Non-communicable disease DALY



Data from WHO reports, 2019

Scientific publications

Free & Open Source

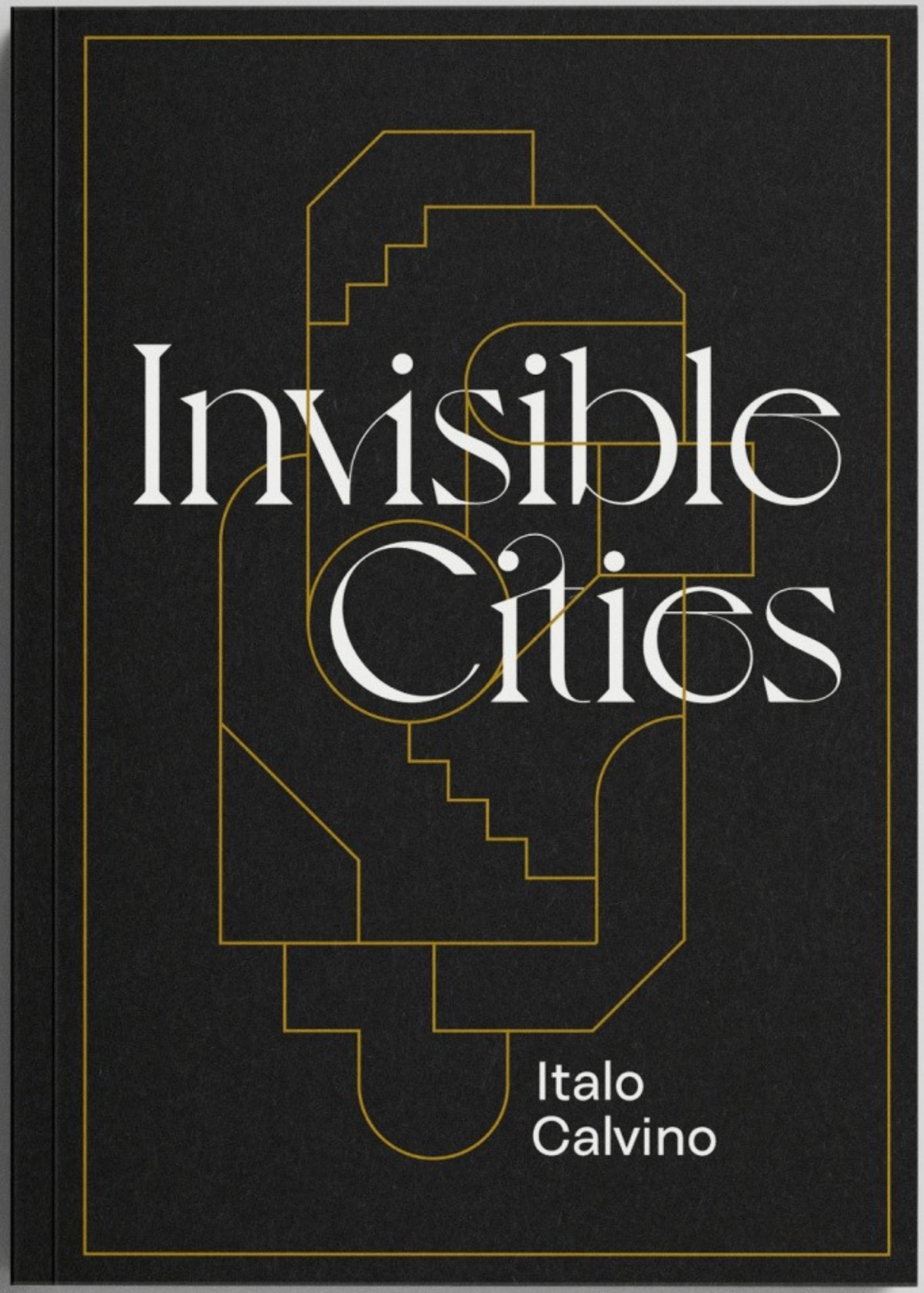
Real-time code sharing
Permissive licenses
No patents
Reproducibility

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



In Ersilia, to establish the relationships that sustain the city's life, **the inhabitants stretch strings from the corners of the houses**, white or black or gray or black-and-white according to whether they mark a relationship of blood, of trade, or authority, agency. **When the strings become so numerous that you can no longer pass among them, the inhabitants leave**: the houses are dismantled; only the strings and their supports remain. From a mountainside, camping with their household goods, Ersilia's refugees look at the labyrinth of taut strings and poles that rise in the plain. That is the city of Ersilia still, and they are nothing.

They rebuild Ersilia elsewhere. They weave a similar pattern of strings which they would like to be more complex and at the same time more regular than the other. Then they abandon it and take themselves and their houses still farther away.

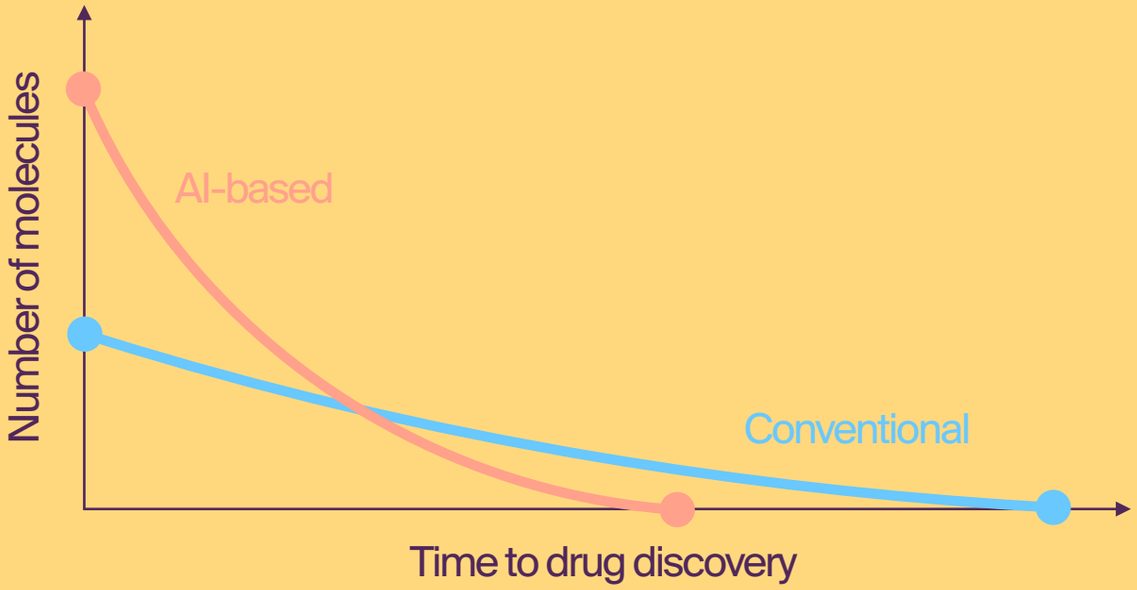
Thus, when travelling in the territory of Ersilia, you come upon the ruins of the abandoned cities, without the walls which do not last, without the bones of the dead which the wind rolls away: **spiderwebs of intricate relationships seeking a form**.

Invisible Cities by Italo Calvino

Open-source drug discovery

AI/ML models

Drug discovery



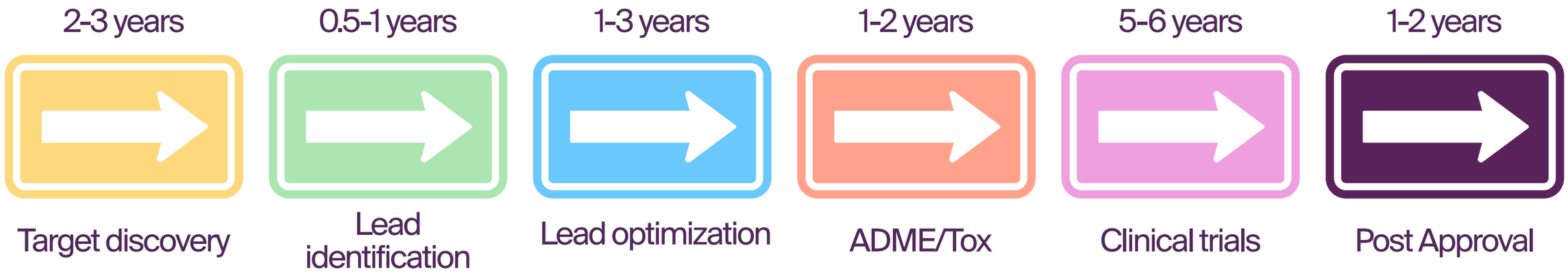
Less 10% of drugs target infections

AI is cheap, but...

— Only 6% of AI papers share code
— Unclear benchmarks
— No experimental validation
— AI models are ‘black boxes’
— Strong technical skills are needed

Haibe-Kains et al, 2020
Hudson et al, 2018
Lorica, 2019
Smith, 2018

Plugging A.I. to the drug discovery pipeline

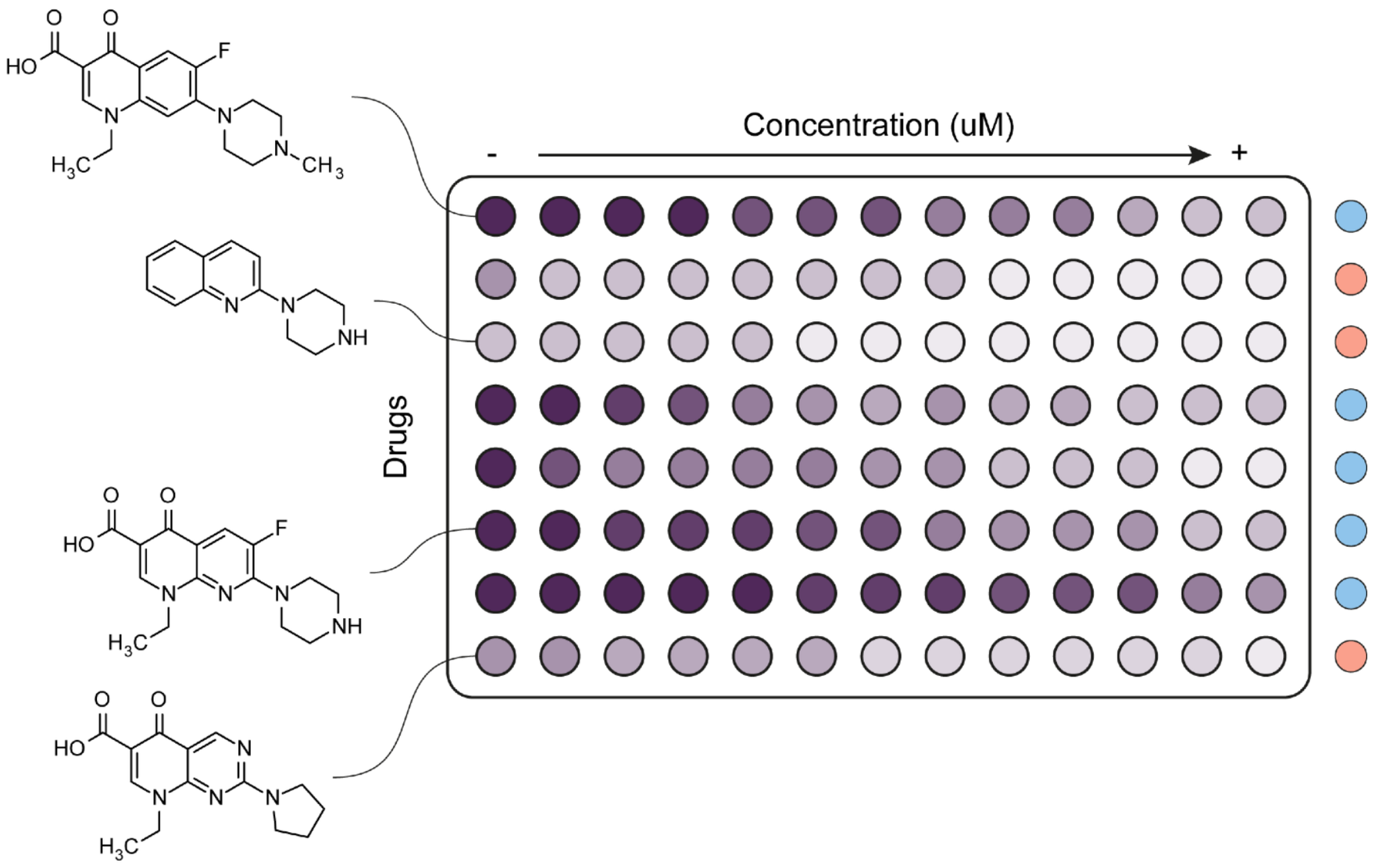


- Disease-gene prioritization
- Target druggability assessment
- Chemical library optimization
- Ligand-based compound screening
- Compound synthesis reaction plan
- Compound optimization with desirable properties
- Formulation prediction
- Tissue-specific biomarker identification
- Drug-drug interactions
- Clinical trial design
- Post-marketing ADR surveillance
- ...

Ekins et al, 2019
Vamathevan et al, 2019

The AI/ML pipeline

Starts with available experimental data (molecule & measured activity)...

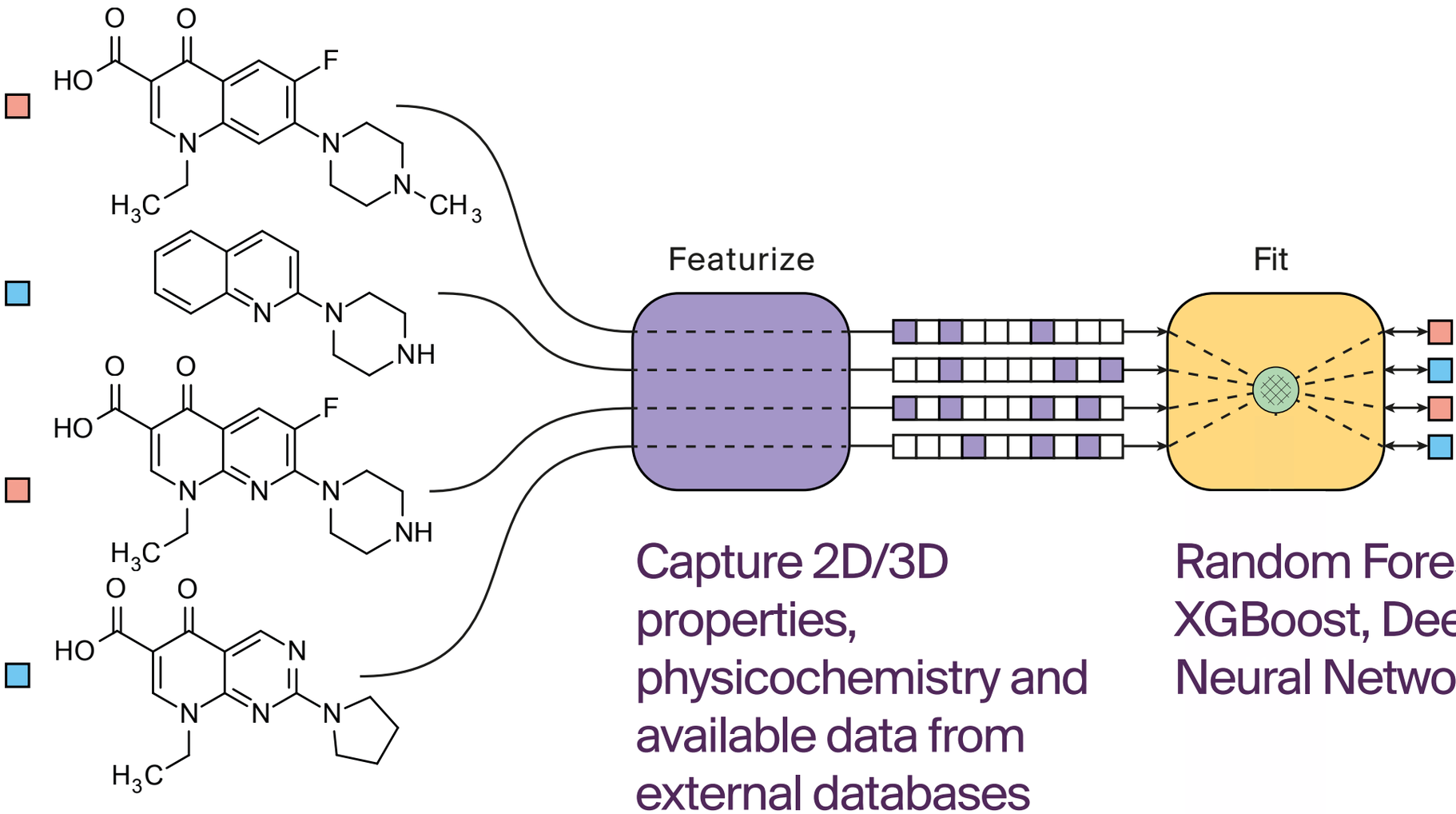


SMILES,
SDF...

Inactive
Active

The AI/ML pipeline

... and continues by calculating molecular descriptors and training an algorithm



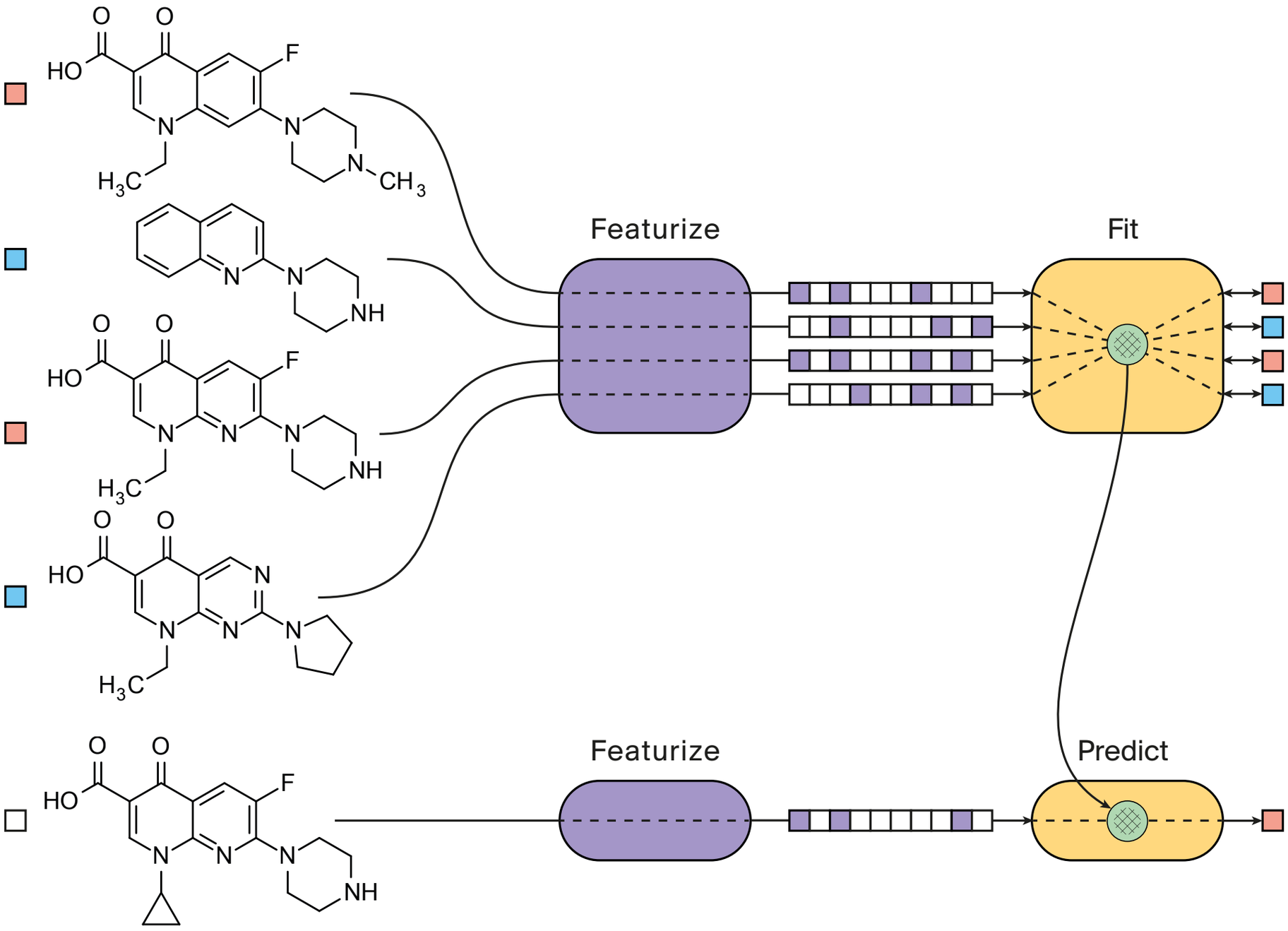
Capture 2D/3D properties, physicochemistry and available data from external databases

Moriwaki et al, 2018
Duran-Frigola et al, 2020
Capecchi et al, 2020
Li et al, 2021
...

Random Forest, XGBoost, Deep Neural Networks ...

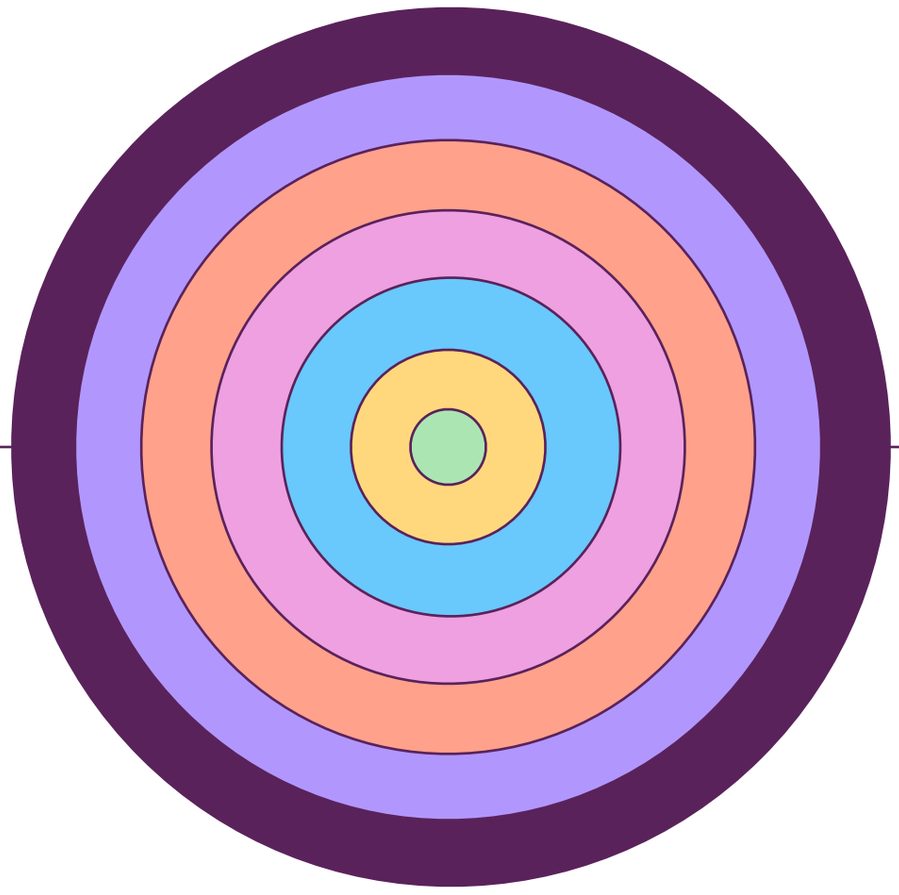
The AI/ML pipeline

Once the model is fitted to the training set, we can input new molecules



Ready-to-use AI/ML models

Input
e.g. a drug



Output
e.g. targets

- Raw data
- Processed data
- Model training
- Testing and tuning
- Validation
- External validation
- Deployment

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,[1,5] Craig R. MacNair,[6] Shawn French,[6] Lindsey A. Carfrae,[6] Zohar Bloom-Ackerman,[2,7] Victoria M. Tran,[2] Anush Chiappino-Pepe,[5,7] Ahmed H. Badran,[2] Ian W. Andrews,[1,2,5] Emma J. Chory,[1,2] George M. Church,[5,7,8] 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*

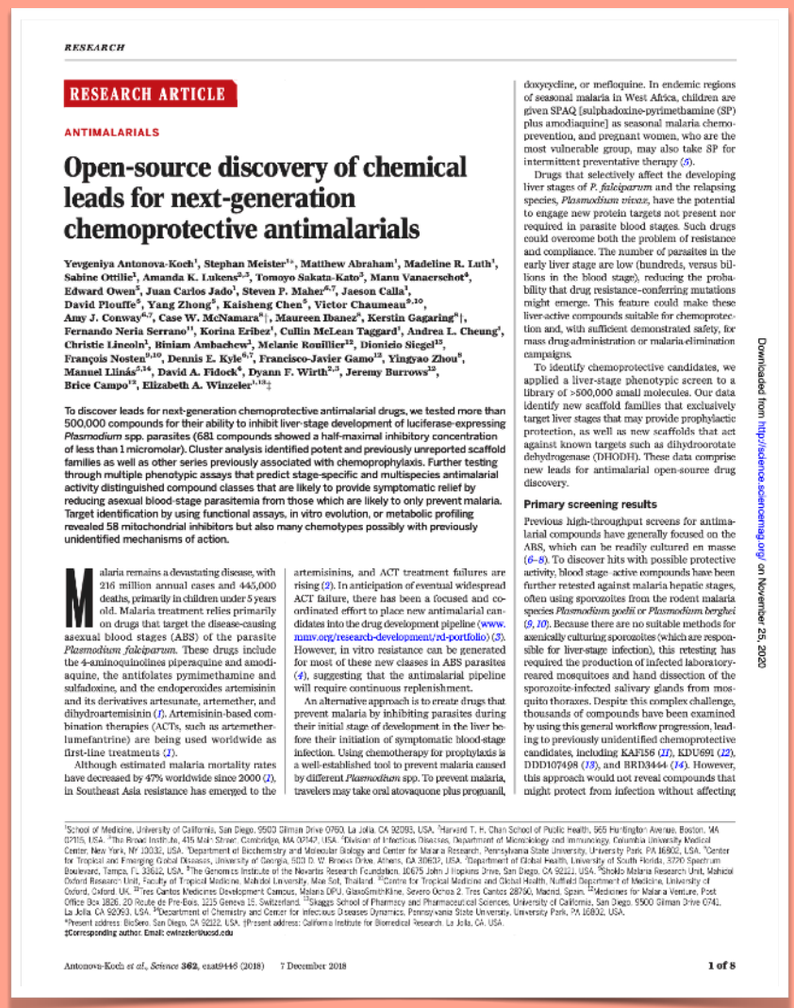
Stokes et al, 2020

Halicin

Active!

In-house AI/ML

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

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 (*2*).

Although estimated malaria mortality rates have decreased by 47% worldwide since 2000 (*3*), 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 IRR3444 (*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

Downloaded from http://science.sciencemag.org/ on November 25, 2020

Chemoprotective antimalarials

Antonova-Koch et al, 2018

Atovaquone analog

Active!

AI/ML in collaboration

Ersilia ‘trains’ a model based on partner’s data

Your awesome project
You and us, 2021

Your question | Our answer!

Zairachem

Ersilia Model Hub

100 models by Christmas 2021

o BACE
, 2020

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

Human cell-line sensitivity
Kuenzi et al, 2020

Toxicity panel Tox21
Mayr et al, 2018

Drug-drug interactions
Rezaul et al, 2019

Blood-brain barrier penetr.
Bertoni…, 2020

Cancer pharmacoge
Huanb et a

l binding arison
y…, 2020

Large-scale QSAR target profiles
Bosc et al, 2019

EGFR inhibitors
Ngoc et al, 2019

Design of focussed molecule libraries
Segler et al, 2017

Molecule generation
Panagiotis… 2020

Chemoprotective antimalarials
Antonova…, 2018

Retrosynth analysis
Thakkar…,

of HIV-1
es…, 2018

Biomedical text summarization
Beltagy…, 2020

Binding pocket descriptors
Zhang et al, 2020

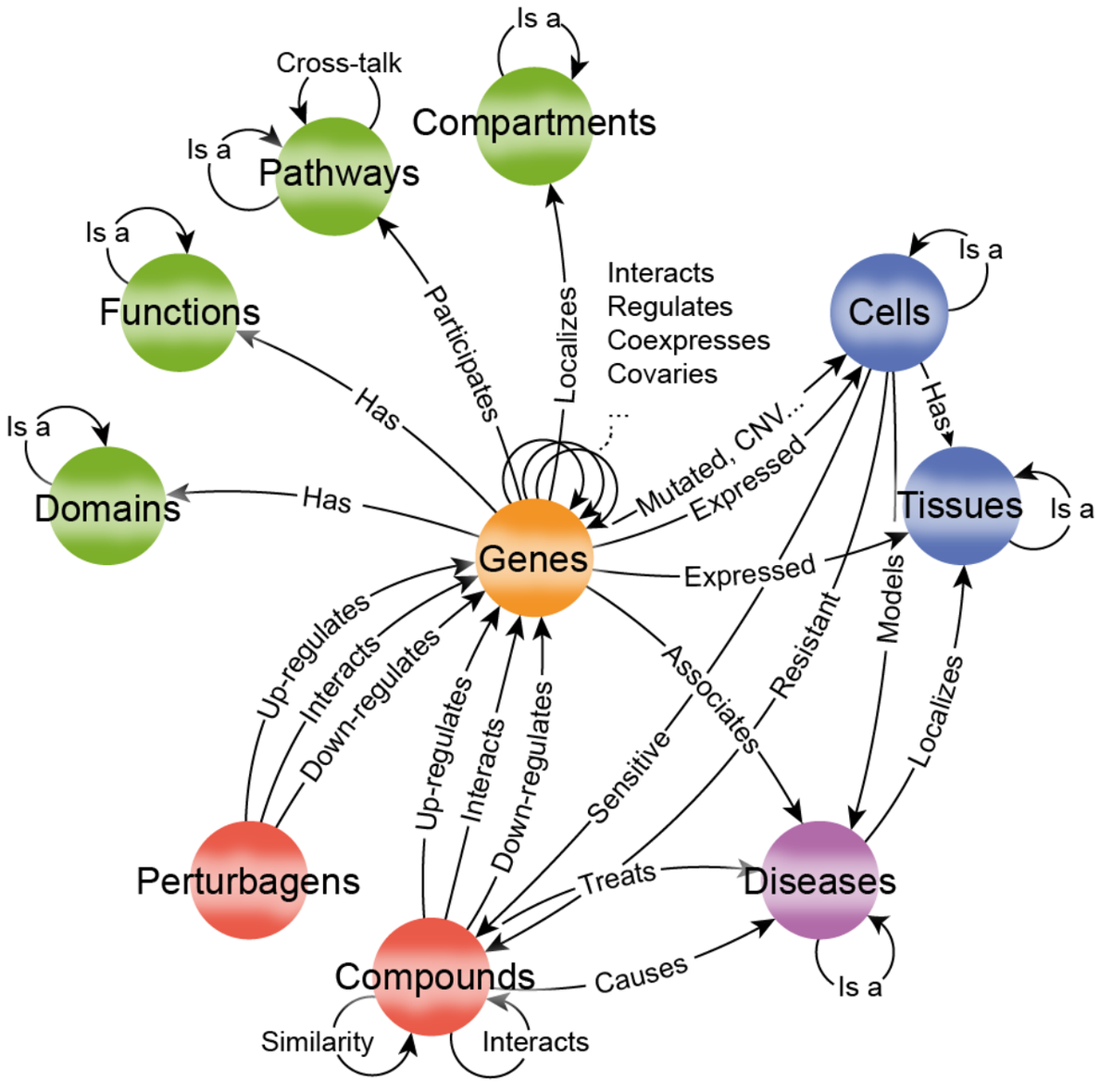
SARS-CoV-2 inhibition
Hofmarcher…, 2020

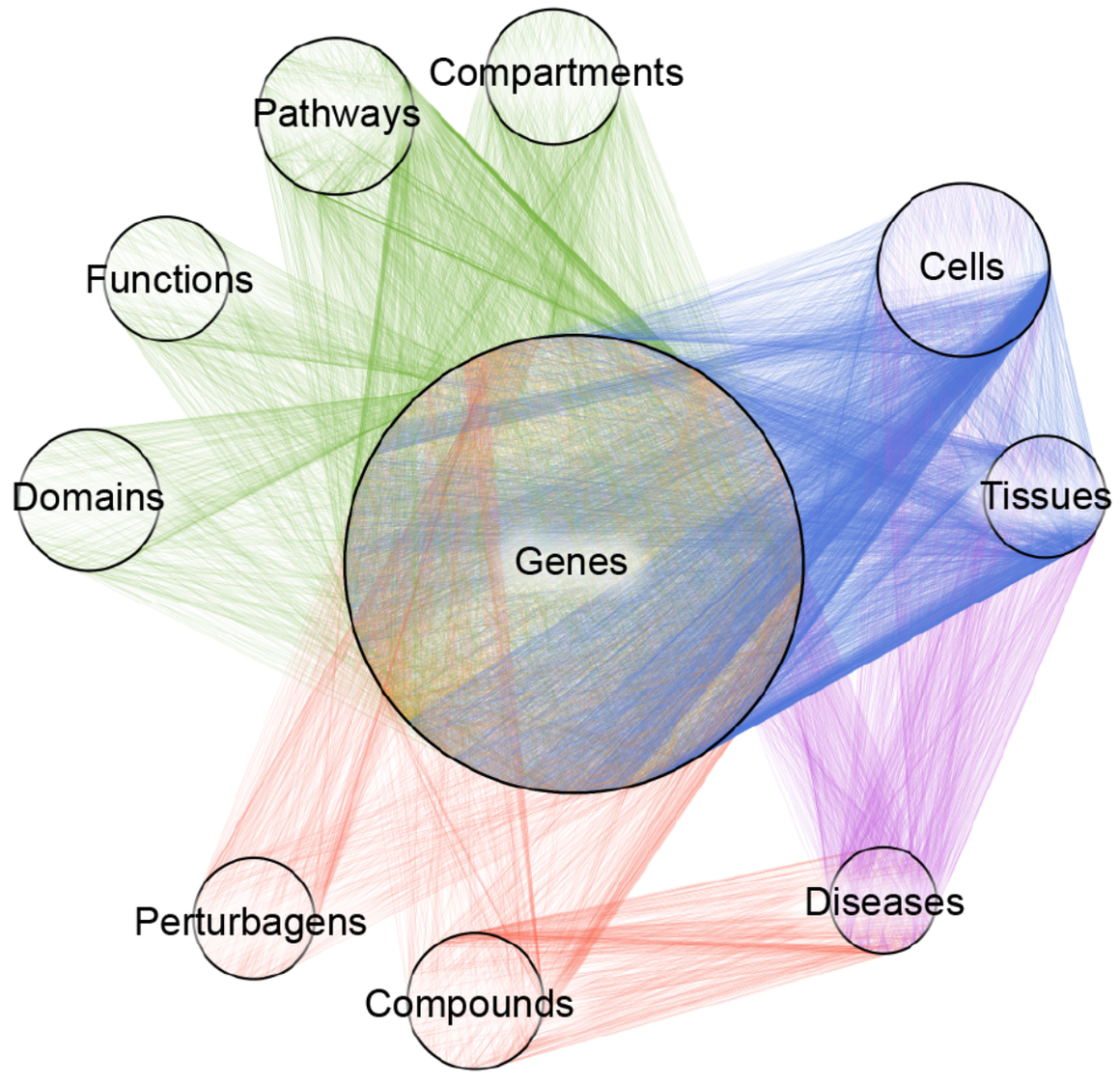
Adverse drug reactions
Ietswaart…, 2020

Transcriptomics-based repurposing
Carrella…, 2014

Physicoche properties
Zhang et a

Biomedical knowledge graph





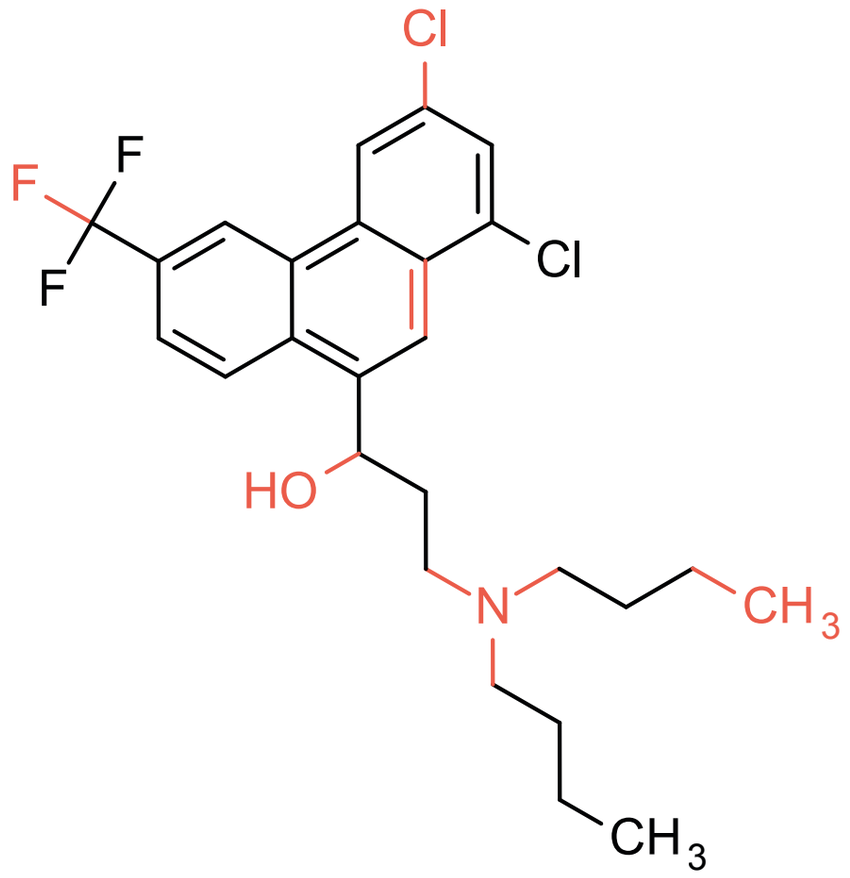
Duran-Frigola et al. WIREs Advanced Reviews, 2019

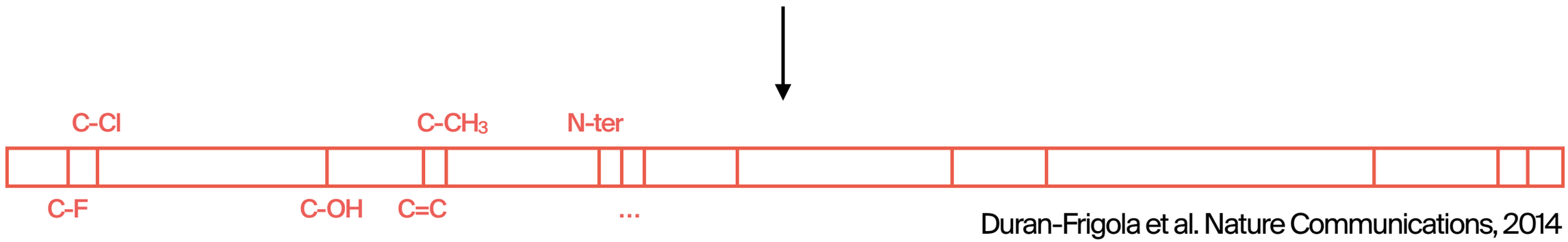
The Chemical Checker, Ersilia's backbone technology

— The largest repository of bioactivity signatures of small molecules
— Based on massive integration of chemoinformatics and bioinformatics tools
— Experimentally validated for Cancer and Alzheimer's research
— Currently being adapted to Infectious diseases

Duran-Frigola et al. Nature Biotechnology, 2020
Bertoni*, Duran-Frigola* et al. Nature Communications, 2021
🏆 Pan-Cancer Activity DREAM Challenge, 2020

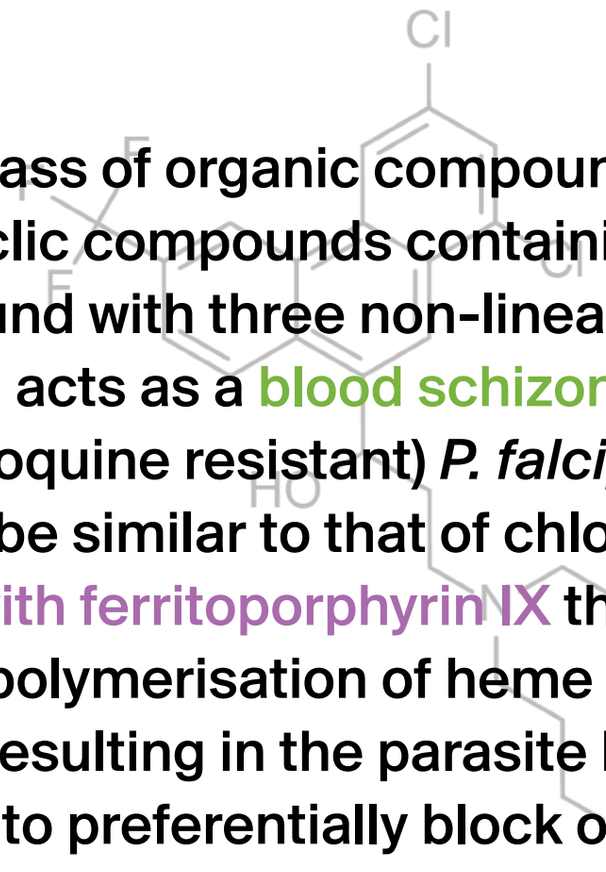
Classical chemical fingerprints





Duran-Frigola et al. Nature Communications, 2014

Chemical Checker signature



Halofantrine belongs to the class of organic compounds known as phenanthrenes and derivatives. These are polycyclic compounds containing a phenanthrene moiety, which is a tricyclic aromatic compound with three non-linearly fused benzene. Halofantrine is a synthetic antimalarial which acts as a blood schizonticide. It is effective against multi drug resistant (including mefloquine resistant) *P. falciparum* malaria. The mechanism of action of Halofantrine may be similar to that of chloroquine, quinine, and mefloquine; by forming toxic complexes with ferritoporphyrin IX that damage the membrane of the parasite. It appears to inhibit polymerisation of heme molecules (by the parasite enzyme 'heme polymerase'), resulting in the parasite being poisoned by its own waste. Halofantrine has been shown to preferentially block open and inactivated HERG channels leading to some degree of cardiotoxicity. Side effects include coughing noisy, rattling, troubled breathing, loss of appetite, aches and pain in joints, indigestion, and skin itching or rash, *et cetera, et cetera.*

Duran-Frigola et al. Nature Biotechnology, 2020

The bioactive chemical space, organised

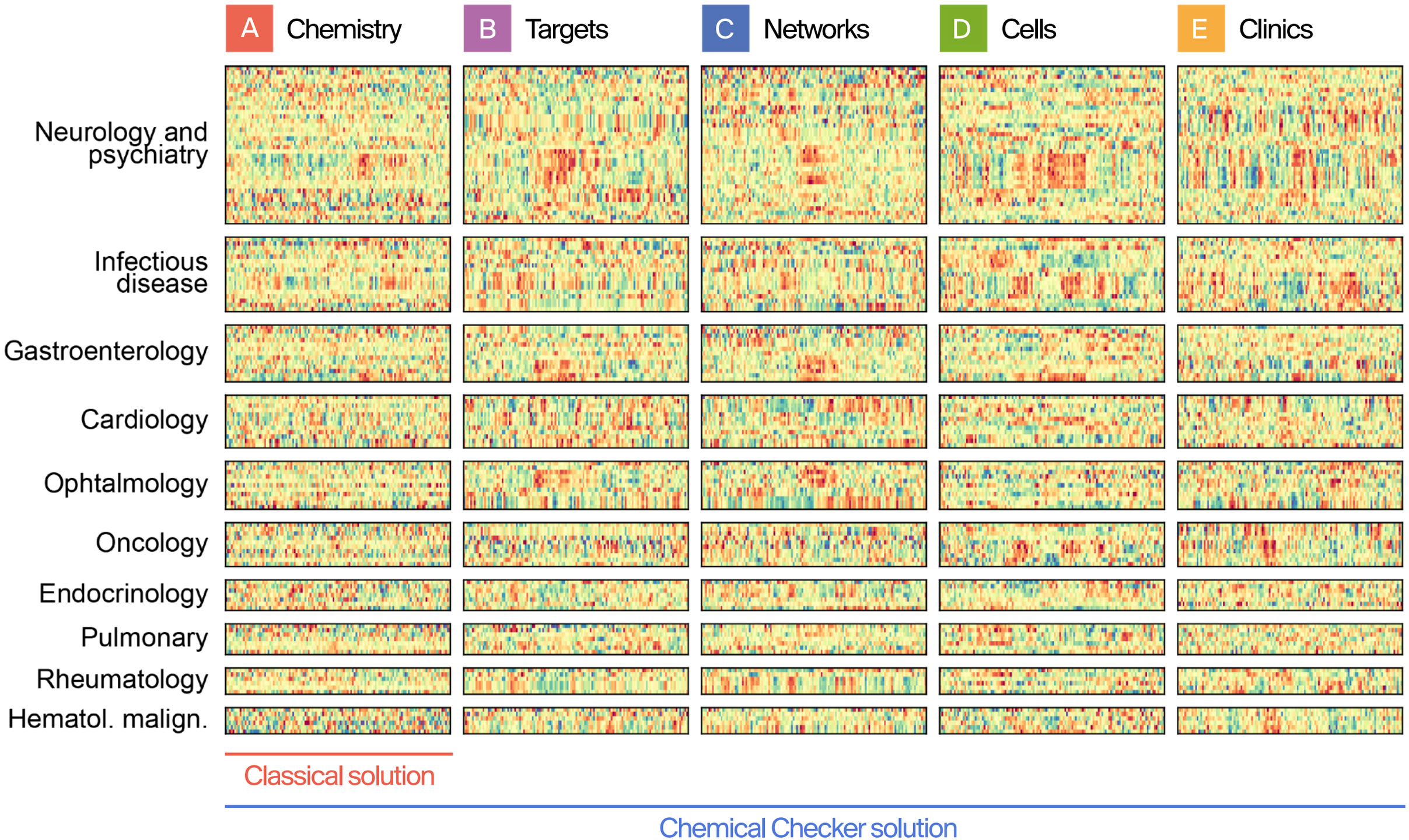
Chemistry	2D fingerprints	3D fingerprints	Scaffolds	Structural keys	Physico-chemistry
Targets	Mechanism of action	Metabolic genes	Crystals	Binding	HTS bioassays
Networks	Small mol. roles	Small mol. pathways	Signaling pathways	Biological processes	Interactome
Cells	Gene expression	Cancer cell lines	Chemical genetics	Morphology	Cell bioassays
Clinics	Therapeutic areas	Indications	Side effects	Diseases and toxicology	Drug-drug interactions

– 1M bioactive molecules
– 25 data types, from chemistry to the clinics
– Massive inference with siamese neural nets
– chemicalchecker.org
– bioactivitysignatures.org

Chemical Checker

Duran-Frigola et al. Nat Biotech, 2020
Pauls et al. Genome Med, 2021
Bertoni,* Duran-Frigola* et al. Nat Commun, 2021
🏆 CTD2-pancancer DREAM challenge, 2020

Chemical Checker signatures



Implementation of a virtual drug screening pipeline

- Where: H3D Centre, Cape Town (South Africa)
- What: AI/ML modelling of relevant *in vitro* assays for drug discovery against malaria and tuberculosis (QSAR)
- Assets:
  - Local mirror of the Ersilia Model Hub
  - Automated & active machine learning
  - Privacy-preserving models before release (Merck funded)

African natural products as antiviral agents

- Where: University of Buea, Buea (Cameroon)
- What: screening of african natural products to identify putative antiviral drugs
- Assets:
  - Setting up a computational research unit at UBuea
  - Largest database of African Natural Products
  - Scaffold hopping tools (easy to synthesize natural product analogs)

Record linkage of medical data

- Where: Centre for Infectious Disease Research Zambia (CIDRZ), Lusaka
- What: Link HIV database to cervical cancer database
- Assets:
  - Linkage tool tuned to Zambian names
  - Recommender system for cervical cancer testing
  - Workshops with Swiss TPH & co-supervision of 1 PhD student

Minimally invasive autopsies

- Where: ISGlobal & CISM, Maputo (Mozambique)
- What: Understanding the chain of infectious events before child's death
- Assets:
  - Dashboard of autopsy data

Open source antimalarials

– Where: Open Source Malaria consortium
– What: Automated generation of antimalarial candidates
– Assets:
  – Antimalarial activity prediction tool
  – Chemical space visualization
  – 500k ‘series 4’ candidates; 50 shortlisted

Pharmacogenomics of African ethnicities

– Where: H3D Center, Cape Town (South Africa)
– What: Dose optimisation of (antitubercular) treatments tailored to African subpopulations
– Assets:
  – Six AI/ML models tailored to tuberculosis treatment
  – Coupling with in-house PBPK modeling tools

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

Ersilia, a hub of AI/ML tools for neglected tropical diseases

FOSDEM, 5th February 2022

Gemma Turon, PhD, @TuronGemma
Ersilia Open Source Initiative, @ersiliaio
http://ersilia.io

Ersilia

Year One, summarised

– First commit to GitHub repository 🤓
– Recognised by the Charity Commission 👩‍💼
– Article published in *The Lancet Global Health* 🎉
– Passed Software Sustainability Institute health check 🩺
– 250 commits in the main GitHub repository 😩
– Top-supported project in the FundOSS crowd-funding campaign 💸
– Article published in *Nature Communications* 🎉
– Contributions to Open Source Malaria 🦟
– 30 funding applications submitted 🤞
– Probabilistic record linkage workshop with the Swiss TPH 👩‍🏫
– Research visit to H3D in South Africa ✈️
– First commit by external contributor 🤗
– Migration to the cloud (AWS Open Data Registry) ☁️
– Code for Science & Society incubator 🐣
– Board of Trustees and administrative help from the Cranfield Trust 🤝