Royal Botanic Gardens

Kew

Mining Medicinal Molecules – an AI approach

Harnessing Plant Traits and AI to Predict Plant Bioactivity

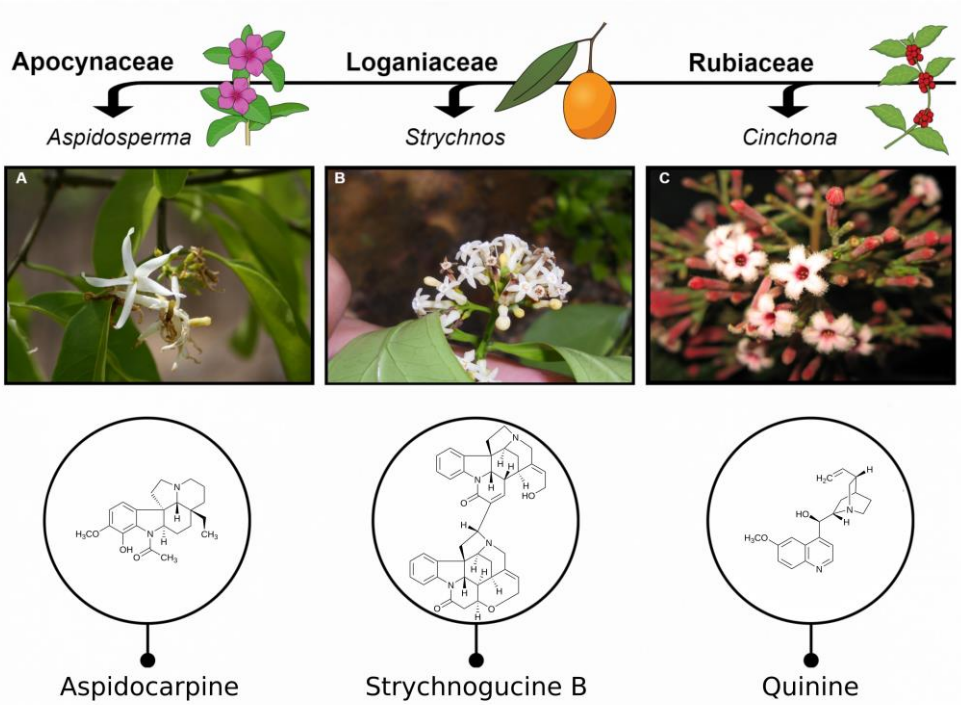
Adam Richard-Bollans a.richard-bollans@kew.org

Background

Malaria affected **247 million** people globally in 2021, with an estimated **619,000 deaths. Resistance** to existing antimalarial drugs is an escalating challenge for eliminating malaria.

Plants have the potential to provide new malaria treatments – **chloroquine** and **artemisinin** are derived from plants (*Cinchona* L. and *Artemisia annua* L. respectively).

We investigate three flowering plant families, **Apocynaceae**, **Loganiaceae** and **Rubiaceae**, selected based on numerous taxa being sources of chemically diverse **alkaloids.**



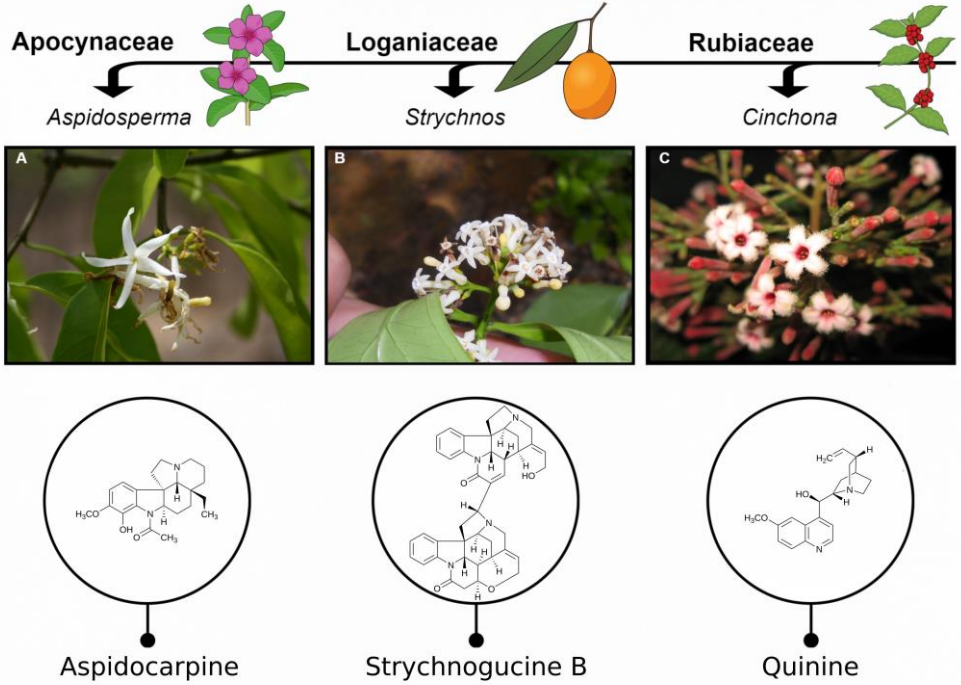
Examples of active antiplasmodial compounds in Apocynaceae, Loganiaceae and Rubiaceae. Photos by Cássia Bitencourt (A), Lucas Marinho (B) and Alexandre Antonelli (C).

Antiplasmodial Potential (Q1)

Of the **c. 21100 species** in Apocynaceae, Loganiaceae and Rubiaceae, we only know the antiplasmodial activity of **282** (and counting!).

Of these 282, approximately **half are active*** against *Plasmodium* species and merit further investigation.

Assuming no issues with the data sample, this would suggest c. 10,000 species in the three families are active – indicating a valuable untapped resource.



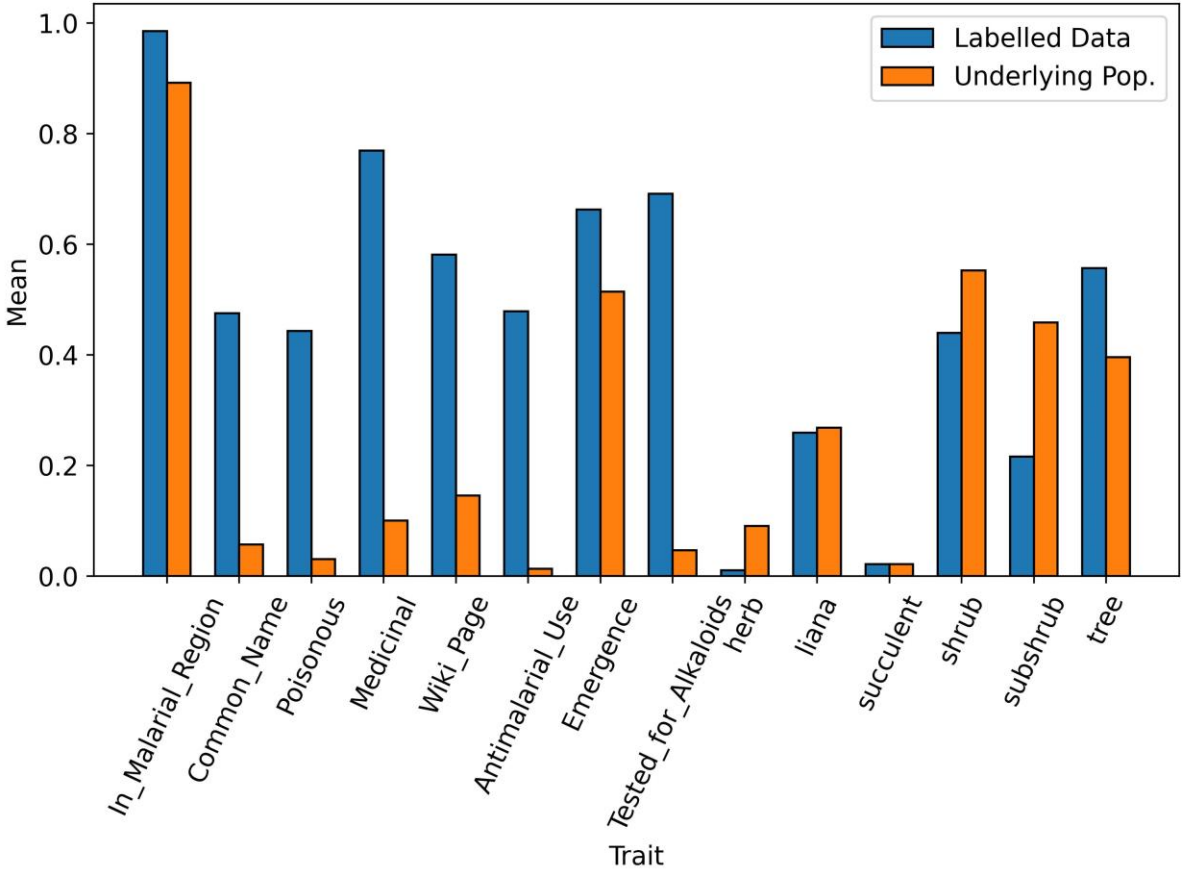
Examples of active antiplasmodial compounds in Apocynaceae, Loganiaceae and Rubiaceae. Photos by Cássia Bitencourt (A), Lucas Marinho (B) and Alexandre Antonelli (C).

Selecting Plants to Investigate (Q2&3)

How/why have the tested plants been selected?

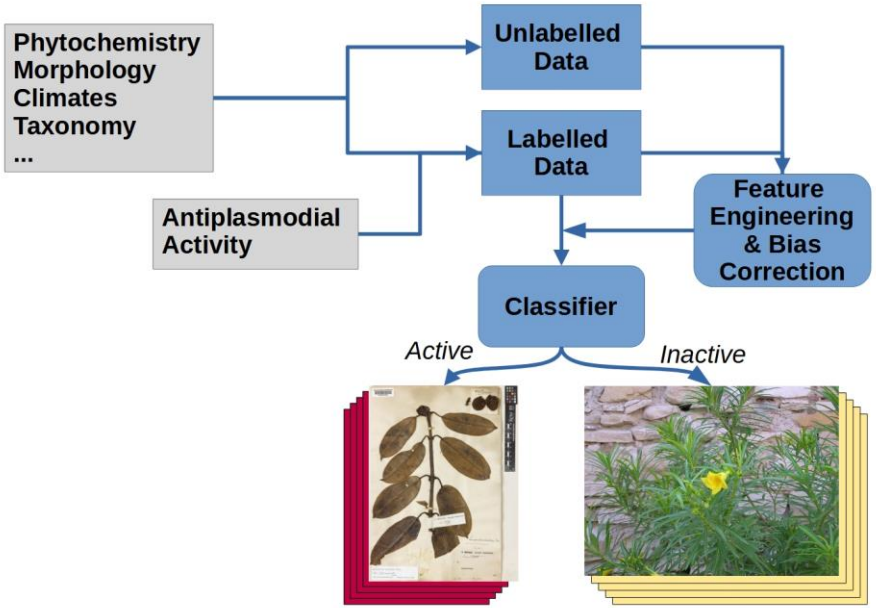
- Strongly influenced by **ethnobotany**
- (Almost!) exclusive to regions with high incidence of **malaria**
- **Phytochemical** knowledge of taxa
- **Taxonomic** relations

Can we improve on these strategies?



Predicting Bioactivity

Can we use machine learning methods to predict which plants are (strongly) antiplasmodial?



Overview of Machine Learning Approach

frontiers

Machine Learning Enhances Prediction of Plants as Potential Sources of Antimalarials

Adam Richard-Bollans [1,*], Conal Aitken [1,2], Alexandre Antonelli [1,3,4], Cássia Bitencourt [1], David Goyder [1], Eve Lucas [1], Ian Ondo [1], Oscar A. Pérez-Escobar [1], Samuel Pironon [1,5], James E. Richardson [6,7,8,9], David Russell [1], Daniele Silvestro [10], Colin W. Wright [11] and Melanie-Jayne R. Howes [1,12]

https://doi.org/10.3389/fpls.2023.1173328

Data Collection

Ethnobotany	**Topography**
Medicinal Use	Elevation
Antimalarial Use	Slope
Phytochemistry	**Soil**
Poisonous	pH
Alkaloids*	Water capacity
Morphology	Nitrogen
Lifeforms	Depth
Emergences	Carbon
Common Knowledge	**Geographic location**
Common names	Latitude + Longitude
Wikipedia pages	In region w/ malaria*
Climate	**Taxonomy**
Mean temperatures	Genus
Mean precipitations	Family
Köppen–Geiger	**Antiplasmodial Activity**

Trait data compiled from literature, expert input and existing datasets. Data collection has been focused on traits with reputed relation to bioactivity.

Some traits rely purely on standard methods and databases, e.g. environmental traits based on GBIF occurrence records and CHELSA.

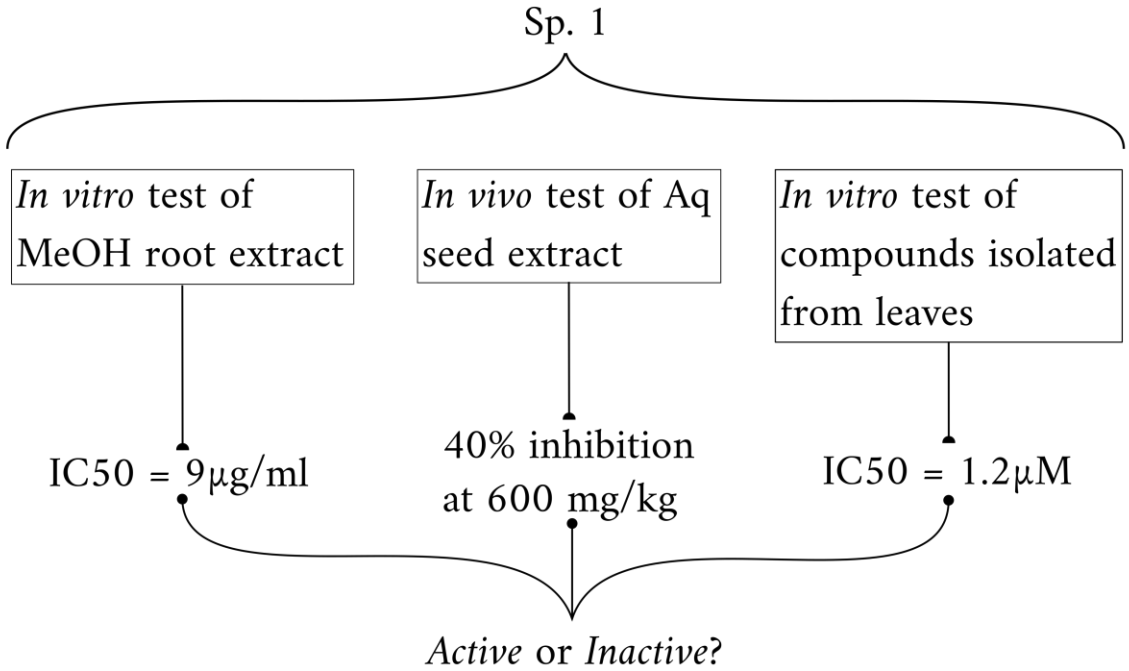
Developed Python library to standardise scientific names to WCVP: github.com/alrichardbollans/automatchnames

Challenges: Classifying Activity

Ideally a regression problem (predicting degree of activity), but standardising data from bioassays is a challenge.

Sources of Heterogeneity:

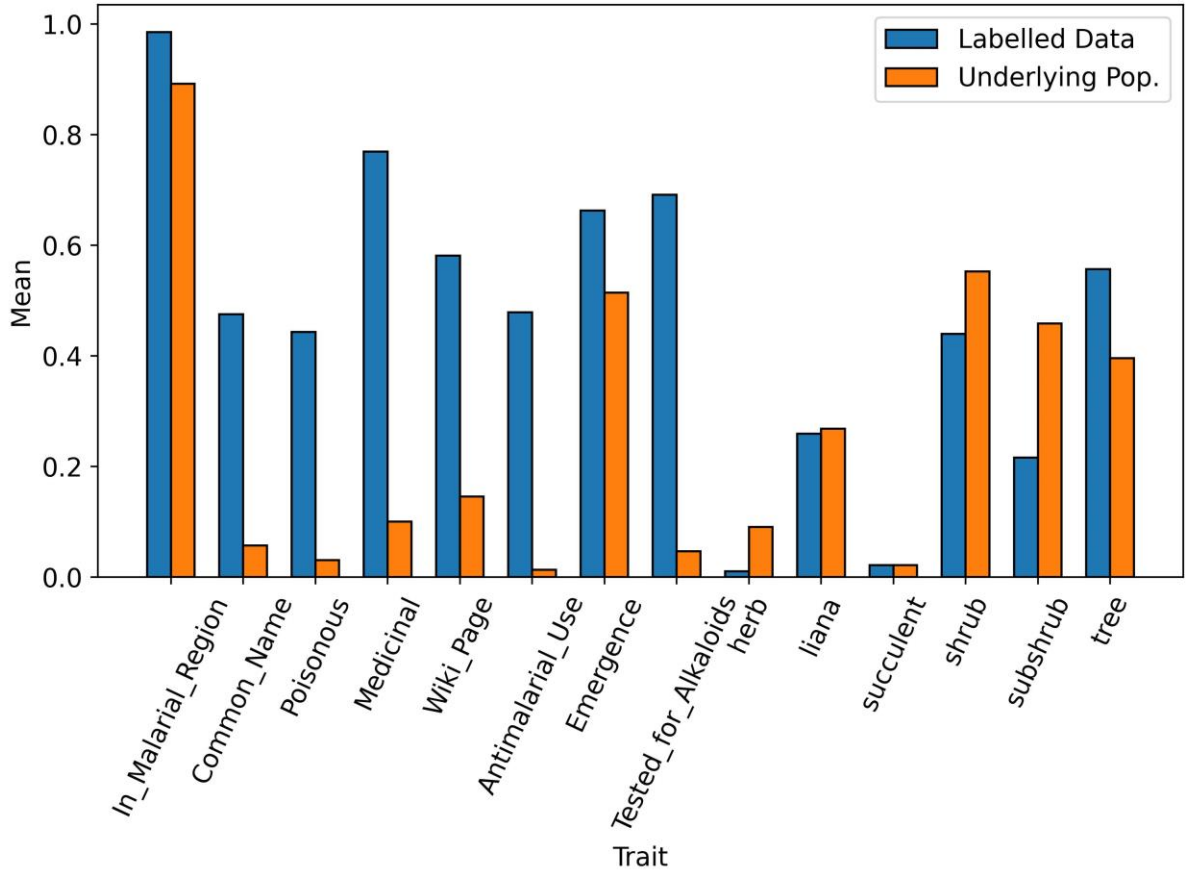
- Plant parts
- Conditions of plant growth and storage
- Extraction methods
- Type of test (*vivo* or *vitro*)
- *Plasmodium* strains
- …



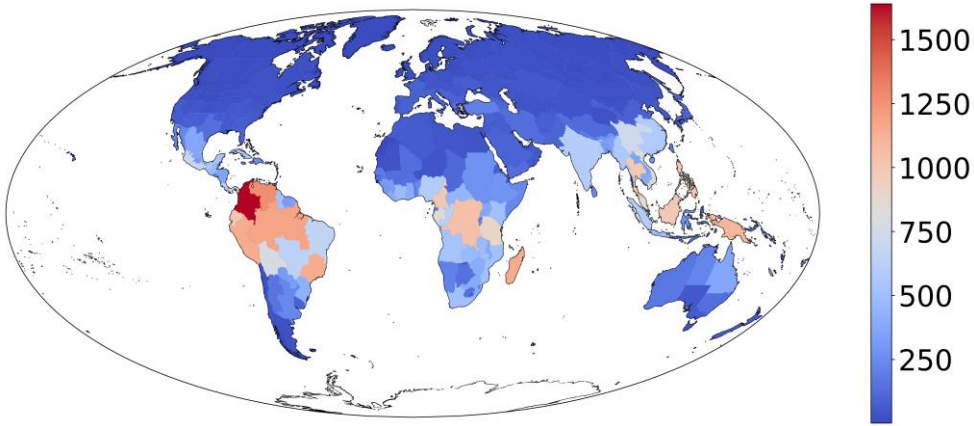
Is this plant worth investigating further?

Challenges: Sampling Biases

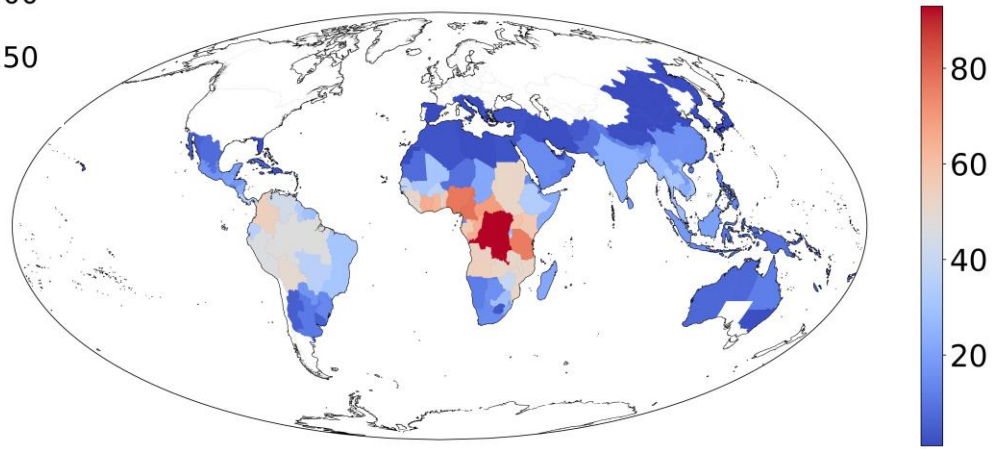
Species with specific properties are overrepresented in the available training data.



Challenges: Sampling Biases

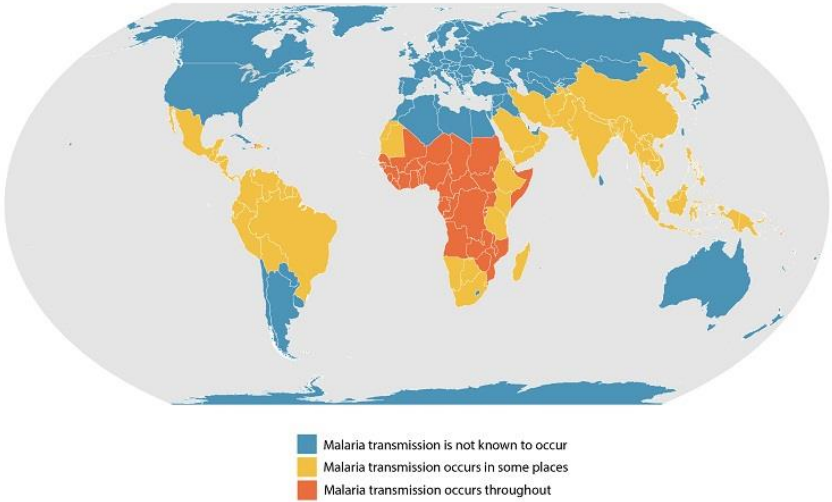


Global distribution of species in Apocynaceae, Loganiaceae and Rubiaceae.

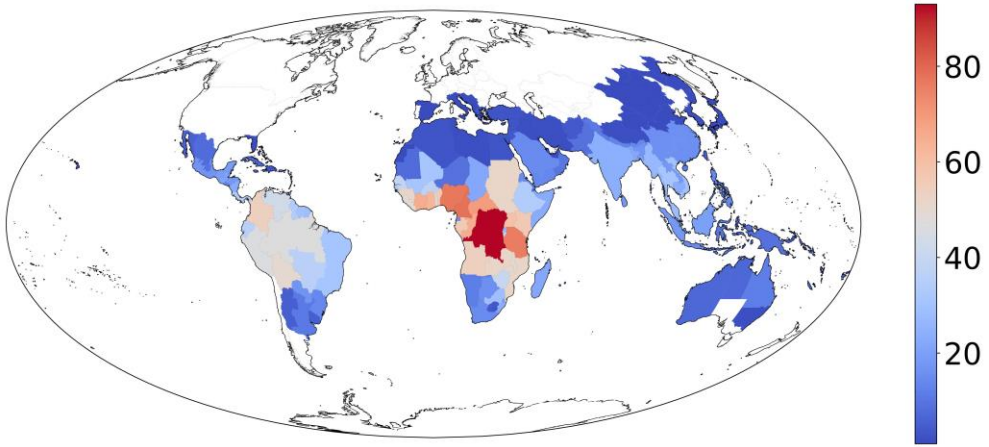


Global distribution of **labelled** species in Apocynaceae, Loganiaceae and Rubiaceae.

Challenges: Sampling Biases



Where Malaria Occurs: Centers for Disease Control and Prevention https://www.cdc.gov/malaria/about/distribution.html

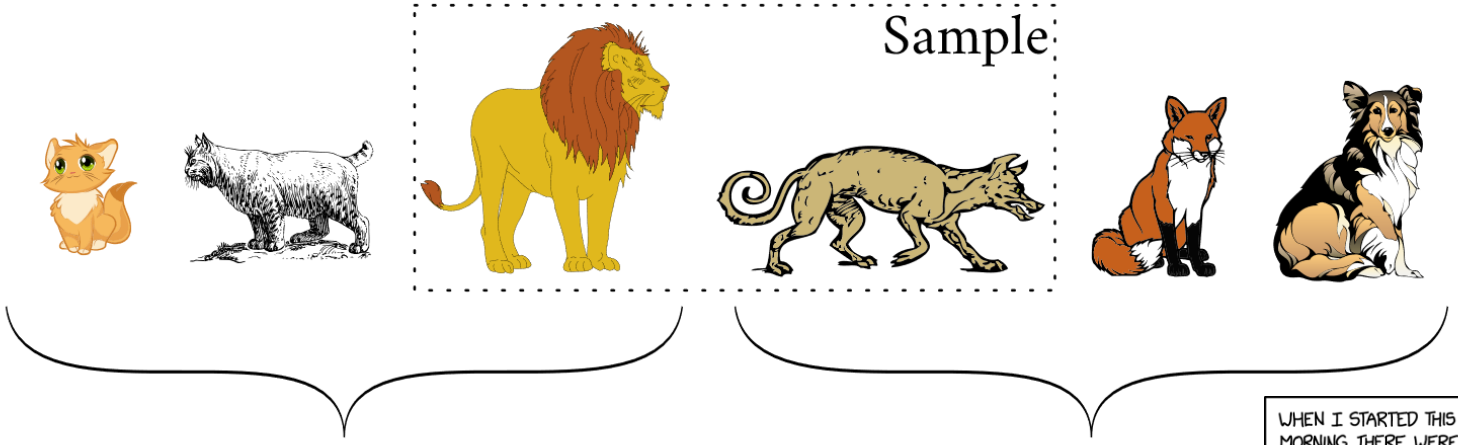


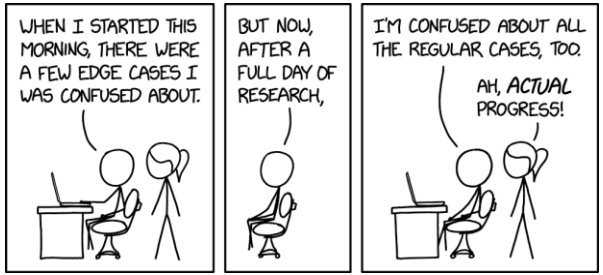
Global distribution of **labelled** species in Apocynaceae, Loganiaceae and Rubiaceae.

Challenges: Sampling Biases

Royal Botanic Gardens Kew

A toy example:





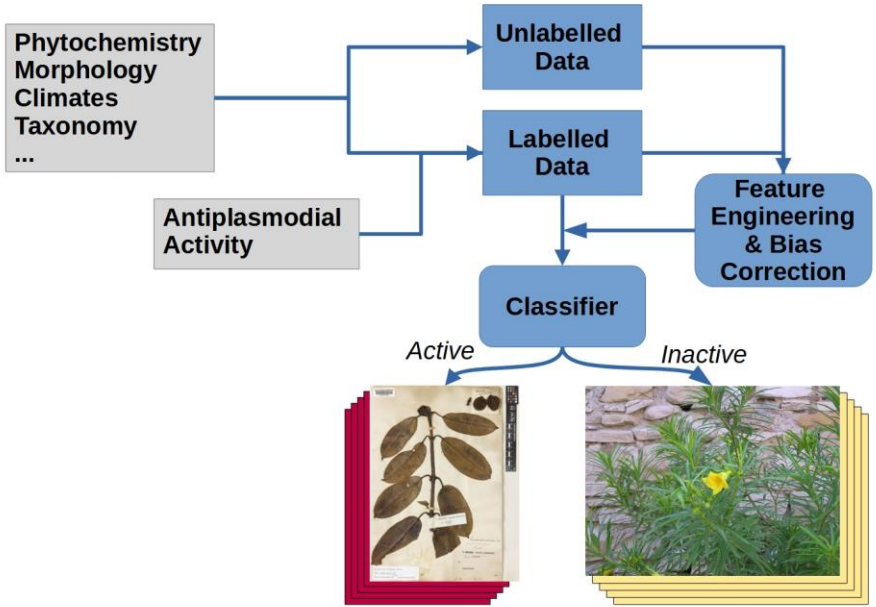
https://xkcd.com/2797/

AI Models

Compared a collection of different models (Support Vector Machines, Logistic Regression, Gradient Boosted Trees and Bayesian Neural Networks)

Compared in biased and biased correction scenarios

Compared to selection based on general medicinal use and selection based on use for malaria

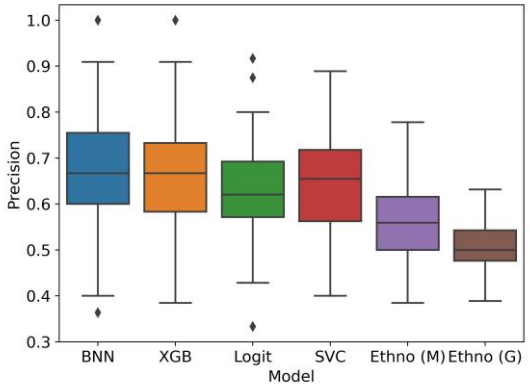


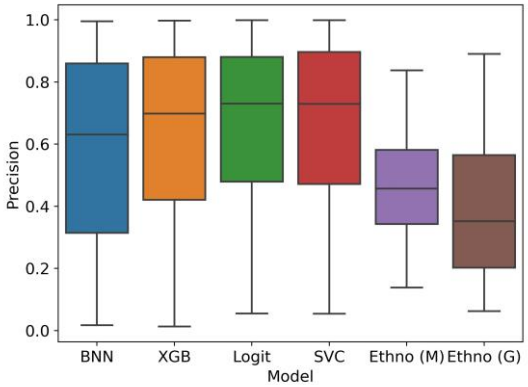
Overview of Machine Learning Approach

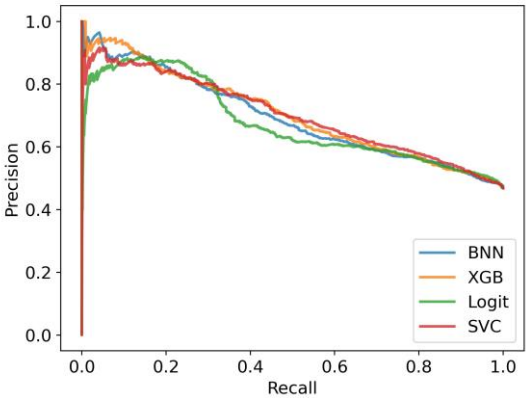
Results

Summary stats:
Estimate that ~**7600** species in Apocynaceae, Loganiaceae and Rubiaceae warrant further investigation

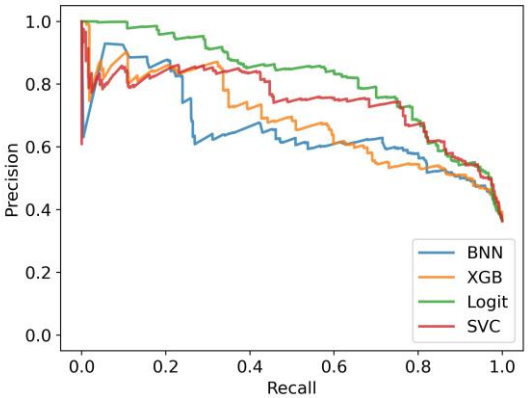
Estimate at least **1300** active antiplasmodial species are highly unlikely to be investigated by conventional approaches







Results with given data



Results with bias correction

Conclusions

Vast untapped potential of plants to provide new antiplasmodial drug leads

Machine learning provides additional tools for selecting plants to investigate

Further improvements to the data (random samples, more traits, more families etc..) can help enhance model accuracy

Acknowledgements

3M Team & Collaborators

Melanie-Jayne R. Howes, Conal Aitken, Alexandre Antonelli, Cássia Bitencourt, David Goyder, Eve Lucas, Ian Ondo, Oscar Alejandro Pérez Escobar, Samuel Pironon, James E. Richardson, David Russell, Daniele Silvestro, Colin W. Wright…

Thank you!