Bringing data science and AI/ML tools to infectious disease research

Session 4: Open Science and Generative Models

Event Sponsors

Event Fund

Open Science

DeepMind, 2021

Ersilia

Mission

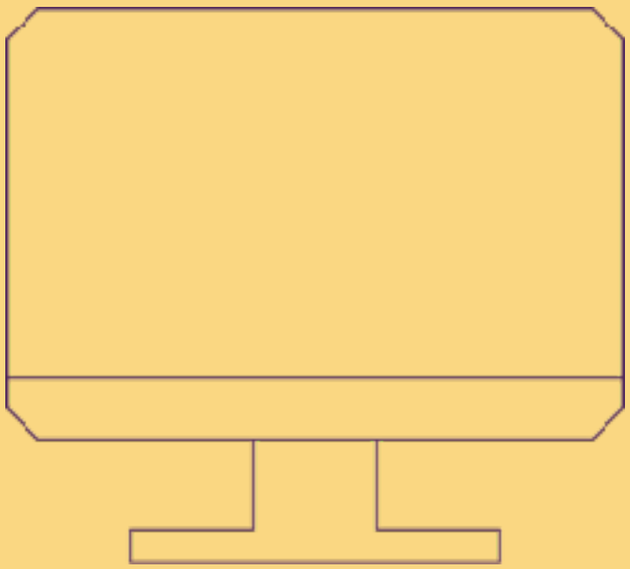
“To strengthen research capacity against infectious and neglected diseases in LMICs”

Vision

“A world with egalitarian access to healthcare”

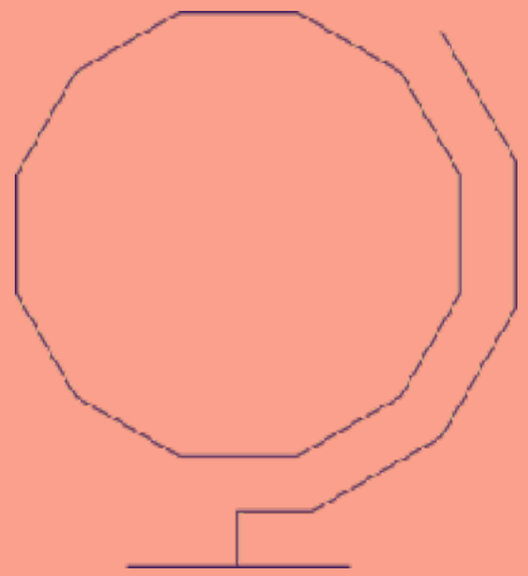
Free & Open Source

Real-time code sharing
Permissive licenses
No patents
Reproducibility



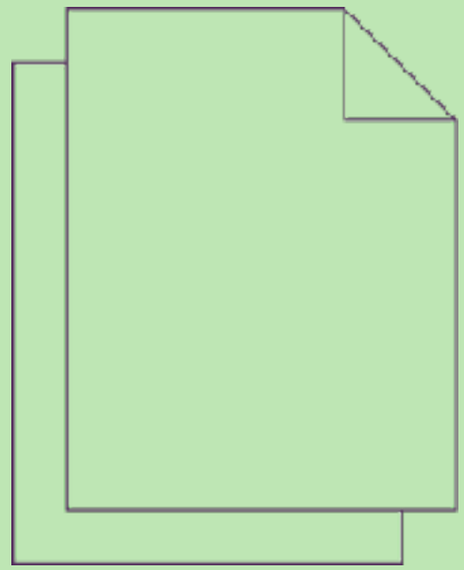
In-Country Research

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



Thanks!

The Fore Fellowship
Event Fund (CS&S)
SeedCorn Award (Rosetrees Trust)
Biopharma Speed Grant (Merck KGaA)
Calestous Juma (Gates Foundation)

A4ID
Airtable
Amazon Web Services
The Cranfield Trust
Atlassian Foundation
GitBook
Monday-dot-com
Outreachy
Red Española de Supercomputación
Software Sustainability Institute
Slack
Forma & Co

Astor Foundation
Digital Lift Software Grant
Black Rock Employee Award
Fast Forward
FundOSS campaign
GitHub
GSF Lab
Oakdale Trust
Okta Employee Giving
Splunk Pledge

Edoardo Gaude
Akash Rungta
Alacia Armstrong
Alice Chen
Raphael Brosular
Samuel Volk
Riyesh Nath
Ifeoluwa Ojumoro
Amna Ali
Su Yen
Lazar Deretic
Tomasz Tokarczuk
Núria Camí

University of Buea, Cameroon

H3D-University of Cape Town, South Africa

CIDRZ, Zambia

Matthew Todd
Edwin Tse
Kelly Chibale
Jason Hlozek
John Woodland
Fidele Ntie-Kang
Albert Manasyan
Kati Taghavi
Julia Bohlius
Patrick Aloy
Adrià Fernández-Torras
Martino Bertoni
Aleix Gimeno
Arnau Comajuncosa
Jake Pry
...

Ersilia Open Source Initiative
Website: https://ersilia.io
E-mail: hello@ersilia.io
Twitter: @ersiliaio

Miquel Duran-Frigola, PhD
E-mail: miquel@ersilia.io
Twitter: @mduranfrigola

Gemma Turon, PhD
E-mail: gemma@ersilia.io
Twitter: @TuronGemma

Ersilia

After the workshop

- Tell us about your scientific problem in more detail, and feel free to put us in contact with your supervisors ☎️
- Perhaps we can help you with modelling your data 🙌
- Do not use the Ersilia Model Hub on your own: let us know and we will offer support 💁‍♀️
- Keep in touch with your new colleagues 🌍

Open Science

– Open Access
– Open Data
– Open Source
– Open Hardware
– Open Infrastructure
– Open Methodology
– Open Education
– Open Peer Review
– Open Social Engageemnt

Can we bring Open Science to Drug Discovery?

– What are the advantages of Open Science for Drug Discovery?
– How can Artificial Intelligence support Open Science for Drug Discovery?

Logos in this slide do not denote endorsment by Ersilia

Ersilia Open Source Initiative

Ersilia is a charity developing open source tools to facilitate global health drug discovery, with a focus on neglected diseases, for equal healthcare

44 followers · United Kingdom · http://ersilia.io · @ersiliaio · hello@ersilia.io

Unfollow

README.md

Welcome to Ersilia! 👋

Pinned

Customize pins

ersilia Public
The Ersilia Model Hub, a platform featuring models for infectious and neglected disease research.
Python ☆ 68 77

eos-template Public template
Template repository to add new models to the Ersilia Model Hub
Python ☆ 1 1

olinda Public
A model distillation library
Python ☆ 2

chemxor Public
Privacy Preserving AI/ML for Drug Discovery
Python ☆ 4 1

Repositories

Find a repository... Type Language Sort New

event-fund-ai-drug-discovery Public
Coding and data materials for the Event Fund AI for Drug Discovery Course
Jupyter Notebook ☆ 3 GPL-3.0 0 0 0 Updated 41 minutes ago

eos4b8j Public
GDBChEMBL Similarity Search
Python ☆ 0 GPL-3.0 0 0 0 Updated 6 hours ago

ersilia Public

View as: Public
You are viewing the README and pinned repositories as a public user.

Get started with tasks that most successful organizations complete.

People

Invite someone

Top languages

Python · Jupyter Notebook · HTML · JavaScript · Vue

Most used topics

Manage

ersilia · model-hub

Developer Program Member

Ersilia Open Source Initiative

Ersilia is a charity developing open source tools to facilitate global health drug discovery, with a focus on neglected diseases, for equal healthcare

44 followers · United Kingdom · http://ersilia.io · @ersiliaio · hello@ersilia.io

Unfollow

README.md

Welcome to Ersilia! 👋

Pinned

Customize pins

ersilia Public
The Ersilia Model Hub, a platform featuring models for infectious and neglected disease research.
Python ☆ 68 77

eos-template Public template
Template repository to add new models to the Ersilia Model Hub
Python ☆ 1 1

olinda Public
A model distillation library
Python ☆ 2

chemxor Public
Privacy Preserving AI/ML for Drug Discovery
Python ☆ 4 1

Repositories

Find a repository... Type Language Sort New

event-fund-ai-drug-discovery Public
Coding and data materials for the Event Fund AI for Drug Discovery Course
Jupyter Notebook ☆ 3 GPL-3.0 0 0 0 Updated 41 minutes ago

eos4b8j Public
GDBChEMBL Similarity Search
Python ☆ 0 GPL-3.0 0 0 0 Updated 6 hours ago

ersilia Public

View as: Public
You are viewing the README and pinned repositories as a public user.

Get started with tasks that most successful organizations complete.

People

Invite someone

Top languages

Python · Jupyter Notebook · HTML · JavaScript · Vue

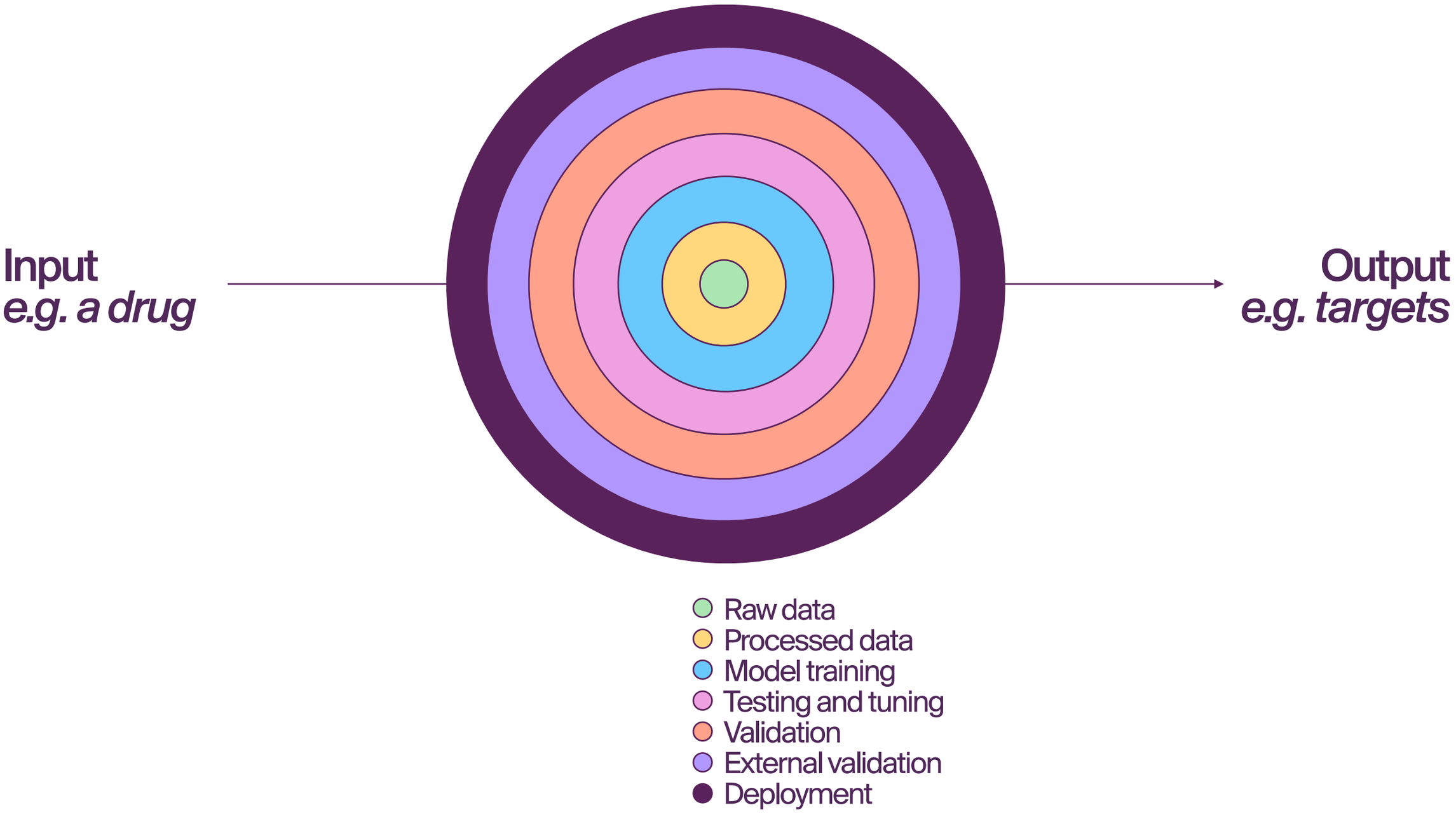
Most used topics

Manage

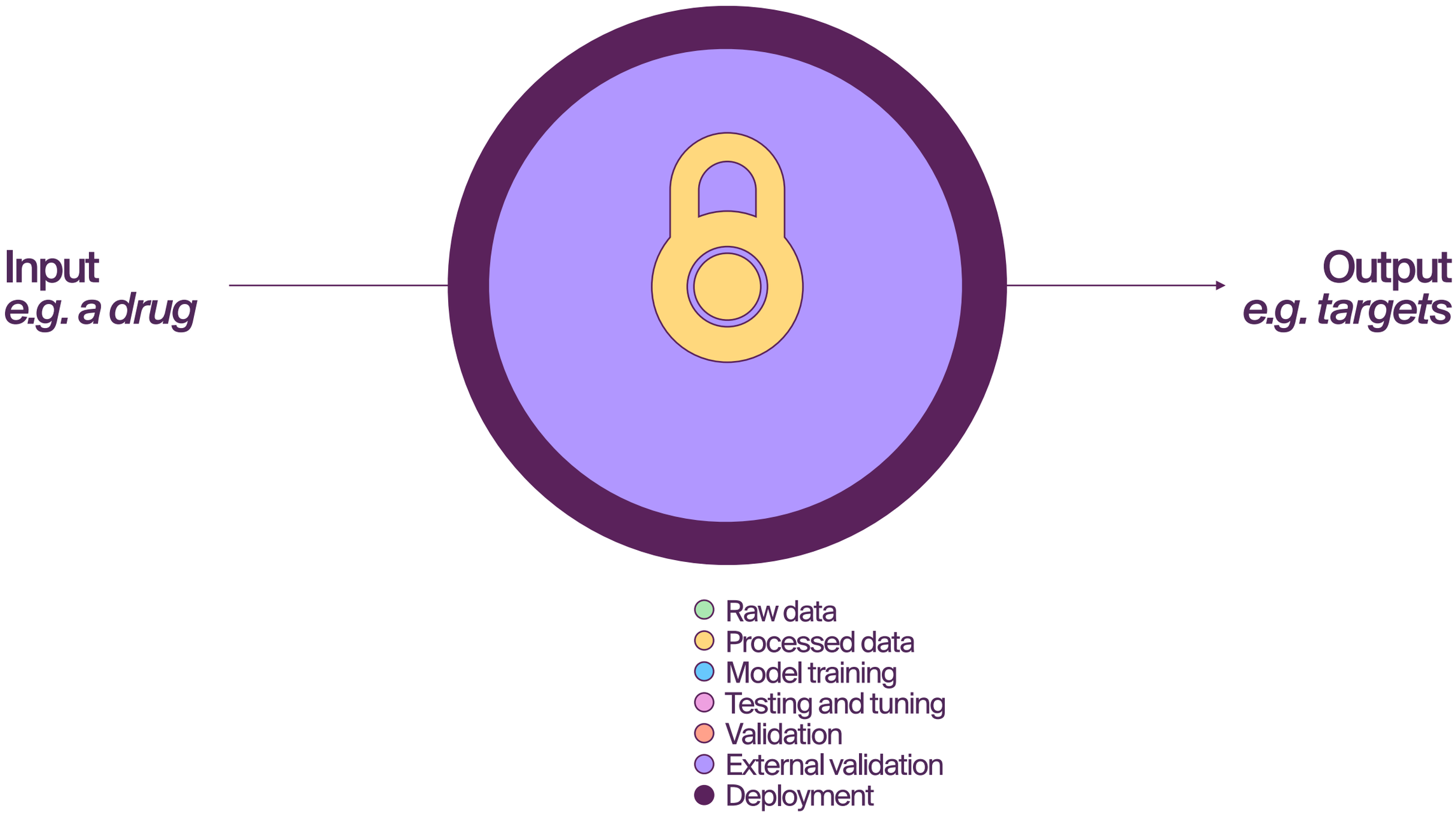
ersilia · model-hub

Developer Program Member

ChemXOR: Privacy-Preserving AI/ML for Drug Discovery



ChemXOR: Privacy-Preserving AI/ML for Drug Discovery



Decentralised machine learning

Article

Swarm Learning for decentralized and confidential clinical machine learning

https://doi.org/10.1038/s41586-021-03583-3

Received: 8 July 2020

Accepted: 26 April 2021

Published online: 26 May 2021

Open access

Check for updates

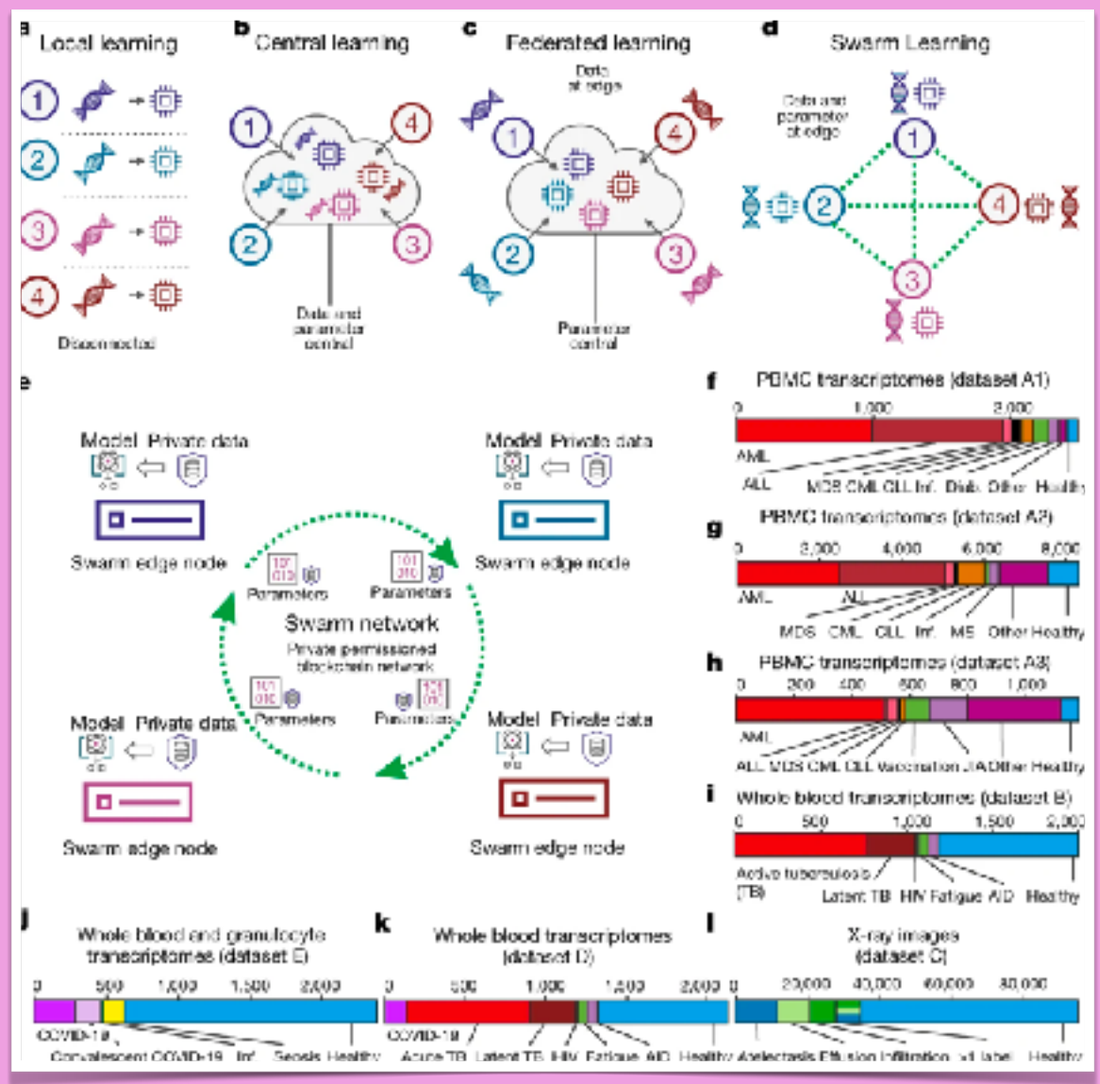
Stefanie Warnat-Herresthal[1,2,17], Hartmut Schultze[3,17], Krishnaprasad Lingadahalli Shastry[4,17], Sathyanarayanan Manamohan[4,17], Saikat Mukherjee[4,17], Vishesh Garg[4,17], Ravi Sarveswara[4,17], Kristian Händler[1,2,17], Peter Pickkers[5,17], N. Ahmad Aziz[1,2,17], Sofia Ktena[6,17], Florian Tran[7,17], Michael Bitzer[8], Stephan Ossowski[9,10], Nicolas Casadei[11], Christian Herr[12], Daniel Petersheim[13], Uta Behrends[14], Fabian Kern[15], Tobias Fehlmann[15], Philipp Schommers[16], Clara Lehmann[16,18,19], Max Augustin[16,18,19], Jan Rybniker[16,18,19], Janine Altmüller[20], Neha Mishra[21], Joana P. Bernardes[7], Benjamin Krämer[22], Lorenzo Bonaguro[1,2], Jonas Schulte-Schrepping[1,2], Elena De Domenico[1,2], Christian Siever[1], Michael Kraut[1,2], Milind Desai[23], Bruno Monnet[1], Maria Saridaki[24], Charles Martin Siegel[1], Anna Drews[1,2], Melanie Nuesch-Germano[1,2], Heidi Theis[1,2], Jan Heyckendorf[1,2], Stefan Schreiber[1,2], Sarah Kim-Hellmuth[1,2], COVID-19 Aachen Study (COMMUNITY)[25], Jacob Nattermann[26,27], Dirk Skowasch[28], Ingo Kurth[21], Andreas Keller[6,17], Robert Bals[7], Peter Nürnberg[29], Olaf Rieß[6,17], Philip Rosenstiel[30], Mihai G. Netea[31,32], Fabian Theis[33], Soumya Mukherjee[34], Michael Backes[35], Anna C. Aschenbrenner[1,2,36], Thomas Ulas[1,2], Deutsche COVID-19 Omics Initiative (DeCOI)[37], Monique M. B. Breteler[38,39], Evangelos J. Giamarellos-Bourboulis[5,40], Matthijs Kox[5,40], Matthias Becker[41,42], Sorin Cheran[41,42], Michael S. Woodacre[43,44], Eng Lim Goh[43,44] & Joachim L. Schultze[1,2,45]✉

Fast and reliable detection of patients with severe and heterogeneous illnesses is a major goal of precision medicine[1,2]. Patients with leukaemia can be identified using machine learning on the basis of their blood transcriptomes[3]. However, there is an increasing divide between what is technically possible and what is allowed, because of privacy legislation[4,5]. Here, to facilitate the integration of any medical data from any data owner worldwide without violating privacy laws, we introduce Swarm Learning—a decentralized machine-learning approach that unites edge computing, blockchain-based peer-to-peer networking and coordination while maintaining confidentiality without the need for a central coordinator, thereby going beyond federated learning. To illustrate the feasibility of using Swarm Learning to develop disease classifiers using distributed data, we chose four use cases of heterogeneous diseases (COVID-19, tuberculosis, leukaemia and lung pathologies). With more than 16,400 blood transcriptomes derived from 127 clinical studies with non-uniform distributions of cases and controls and substantial study bias, as well as more than 95,000 chest X-ray images, we show that Swarm Learning classifiers outperform those developed at individual sites. In addition, Swarm Learning completely fulfils local confidentiality regulations by design. We believe that this approach will notably accelerate the introduction of precision medicine.

Identification of patients with life-threatening diseases, such as leukaemia, tuberculosis or COVID-19[1,2], is an important goal of precision medicine[1]. The measurement of molecular phenotypes using omics technologies and the application of artificial intelligence (AI) approaches[3–6] will lead to the use of large-scale data for diagnostic purposes. Yet, there is an increasing divide between what is technically possible and what is allowed because of privacy legislation[4,5]. Particularly in global crises[7,8], reliable, fast, secure confidentiality- and privacy-preserving AI solutions can facilitate answering important questions in the fight against such threats[7–10]. AI-based concepts range from drug target prediction[11] to diagnostic software[12,13]. At the same time, we need to consider important standards relating to data privacy and protection, such as Convention 108+ of the Council of Europe[14].

AI-based solutions rely intrinsically on appropriate signal intensity between more or less large training datasets[15]. As medicine is inherently decentral, the volume of local data is often insufficient to train reliable classifiers[16–18]. As a consequence, centralization of data is a model that has been used to address these local limitations[19]. While beneficial from an AI perspective, centralized solutions have inherent disadvantages, including increased data traffic and concerns about data ownership, confidentiality, privacy, security and the creation of 'data monopolies' that favour data aggregators[20]. Consequently, solutions

A list of affiliations appears at the end of the paper.



Warnat-Herresthal, Nature, 2021

Event Fund

Generative models

DALL·E 2

An astronaut riding a horse

DALL·E 2

A bowl of soup with a whool monster

DALL·E 2

A nomad city full of threads that go from door to door

DALL·E 2

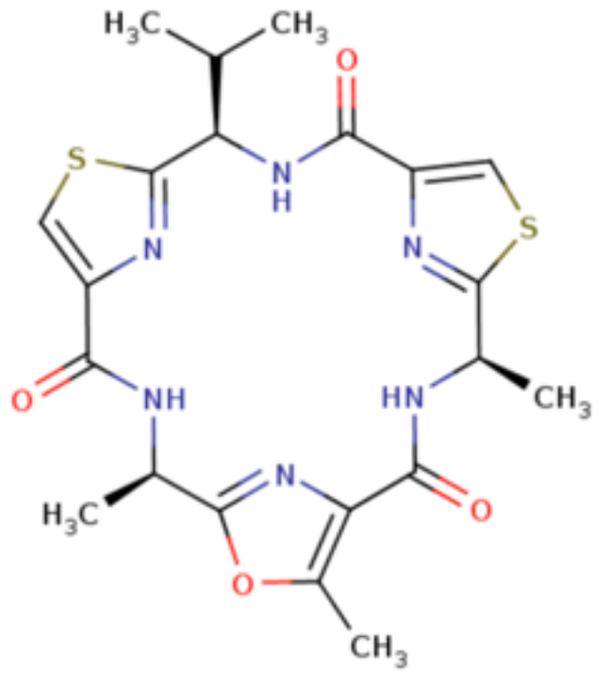
“Whispering molecular secrets”, modern art

DALL·E 2

A novel, cheap, safe and potent antimalarial compound

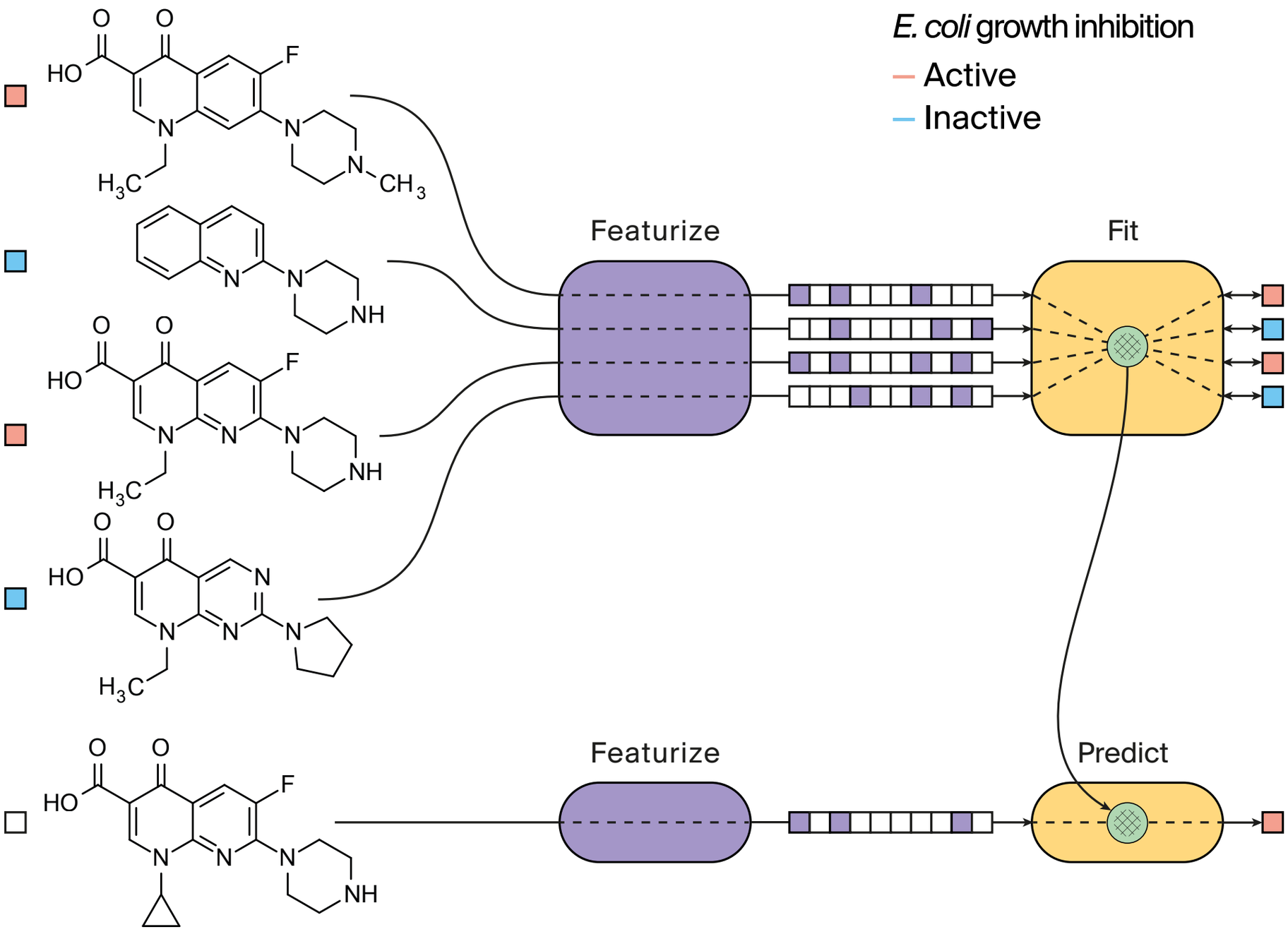
Translation between molecules and natural language

The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.

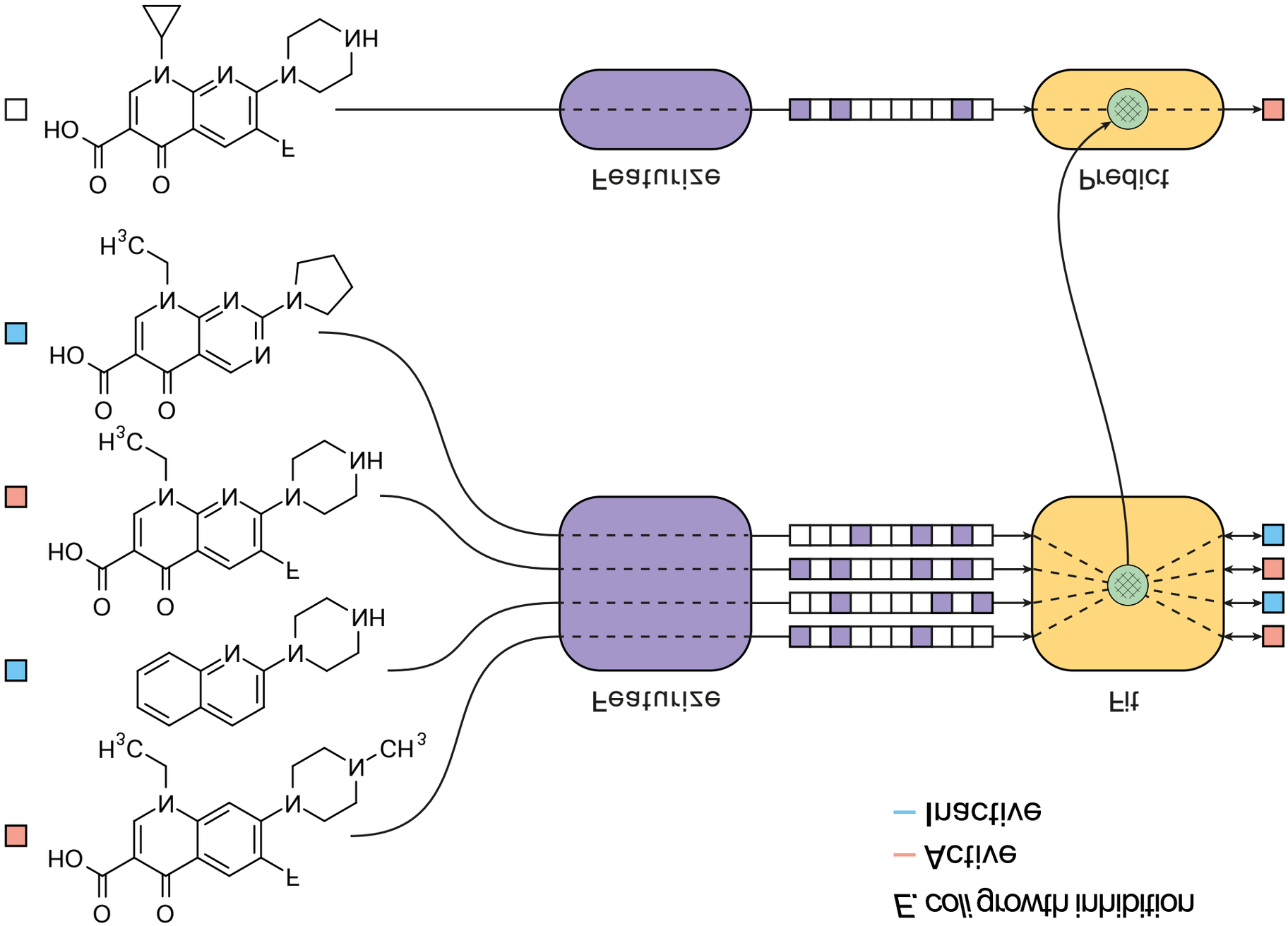


Edwards, ArXiV, 2022

Activity-guided generative models

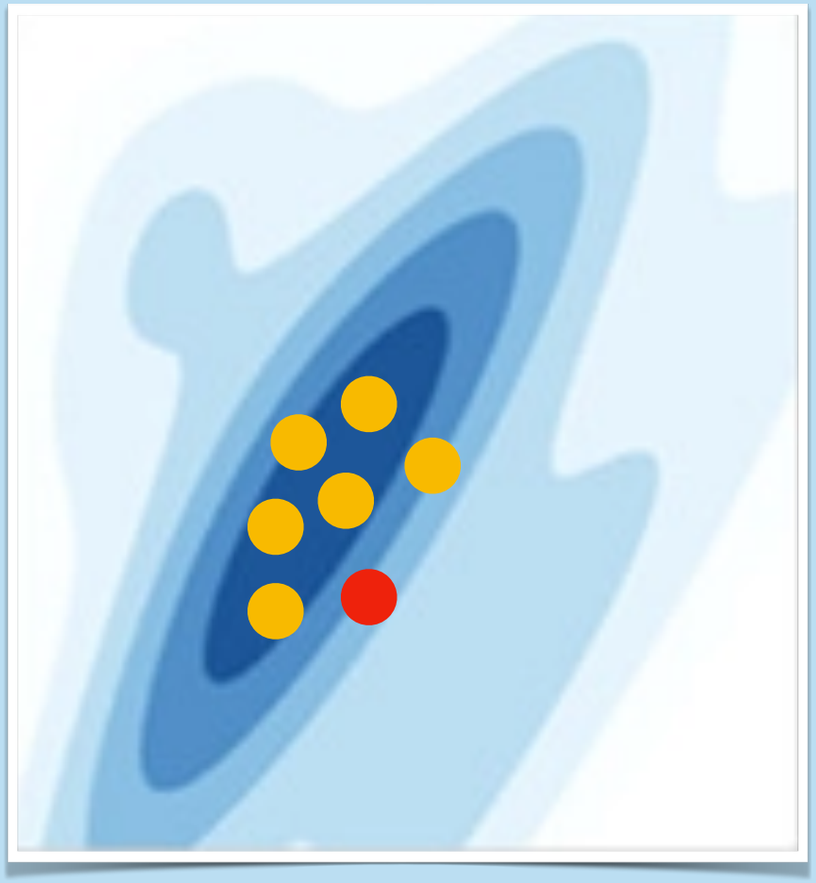


Activity-guided generative models



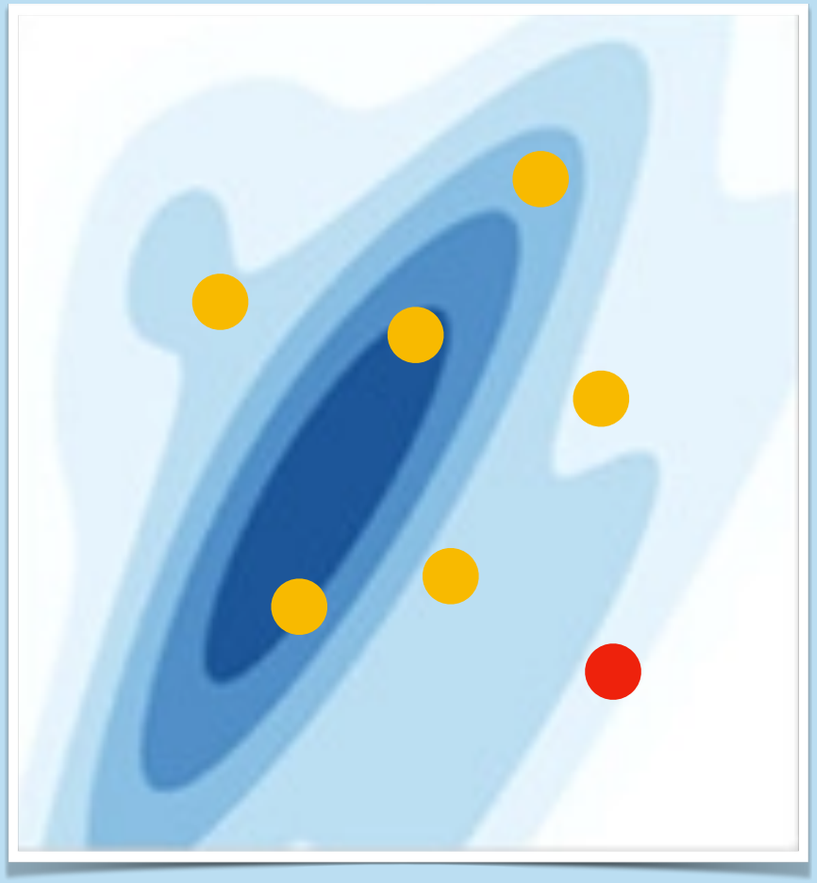
Exploitation mode

— Hit-to-lead optimisation



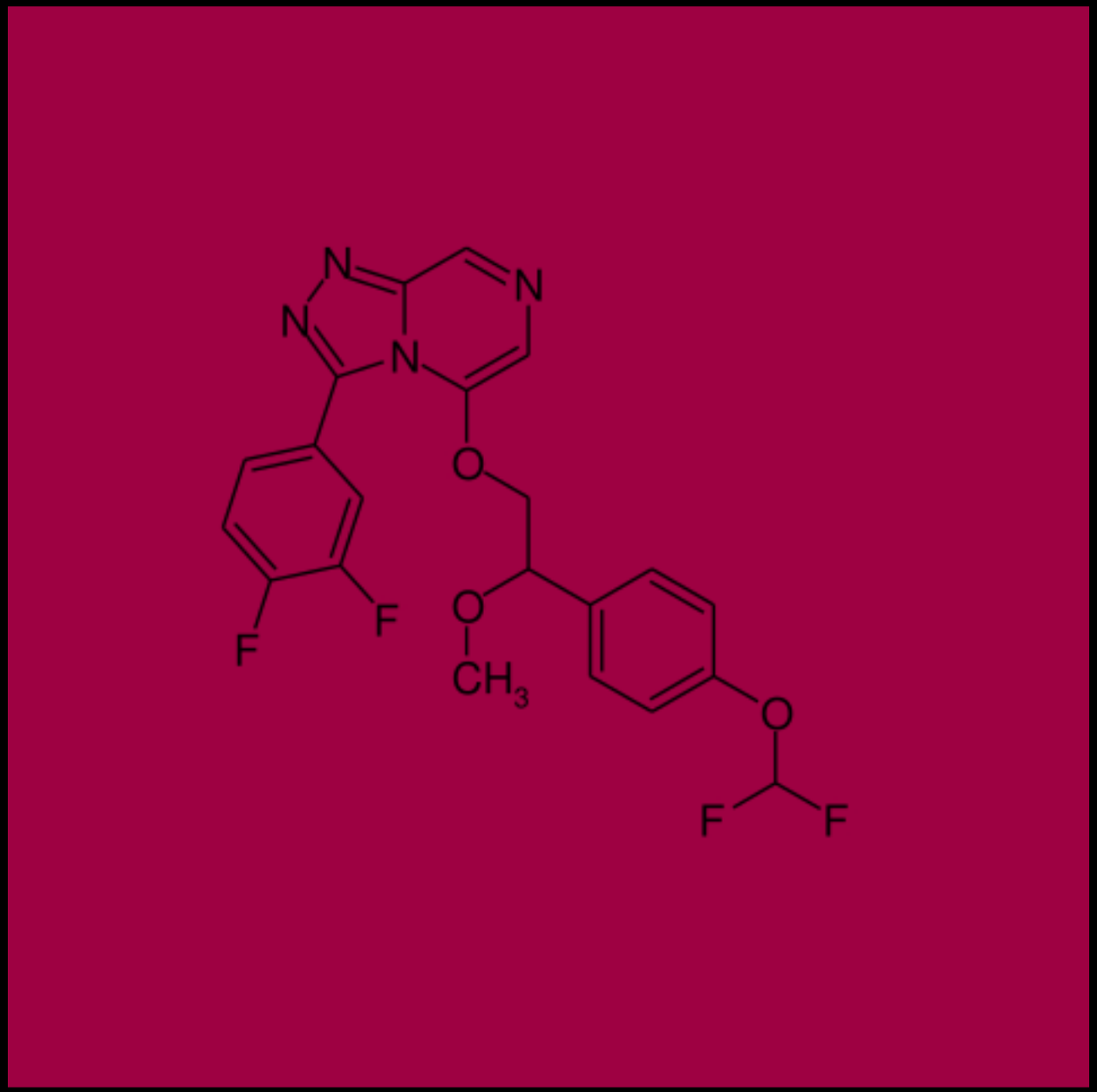
Exploration mode

— Library design



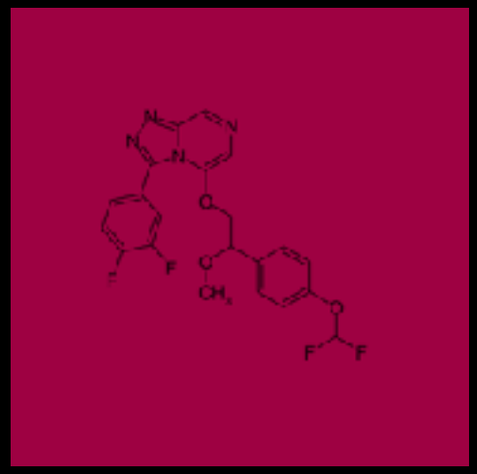
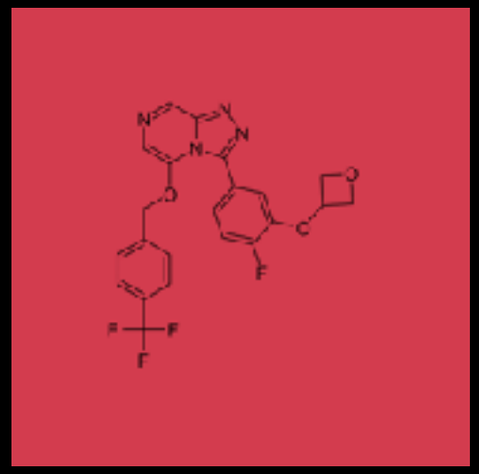
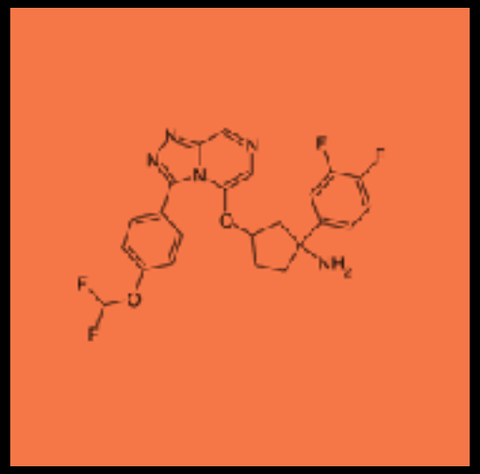
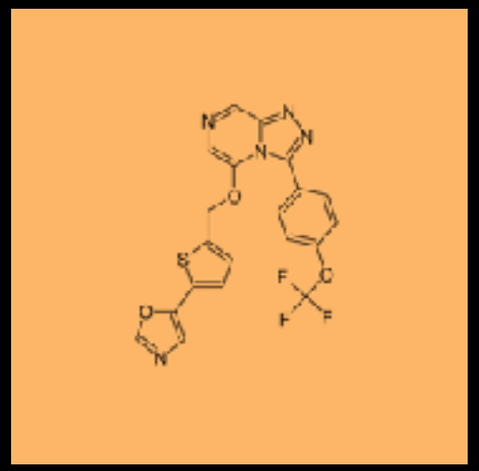
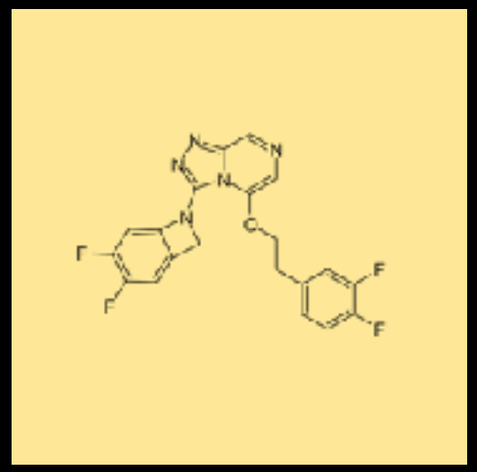
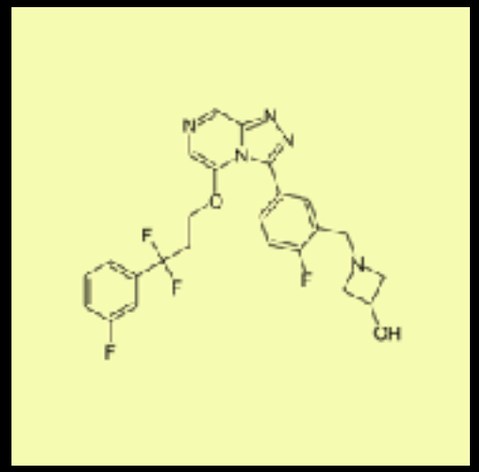
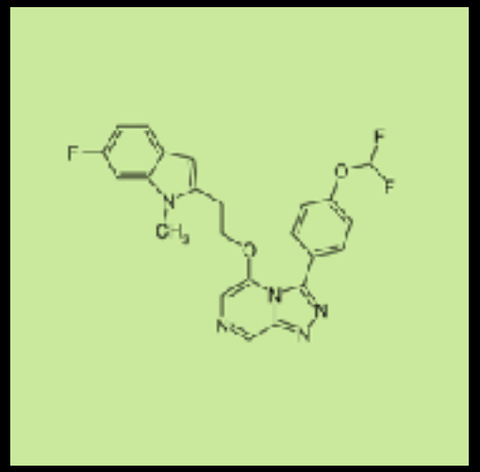
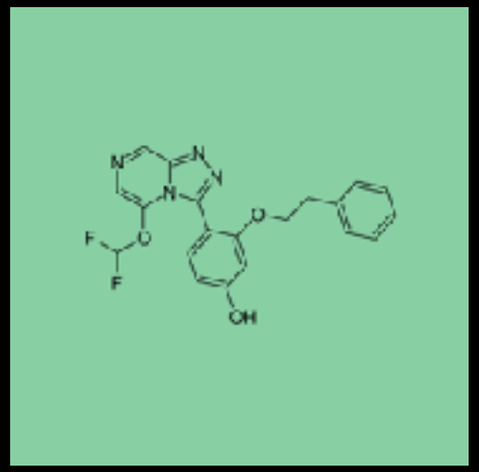
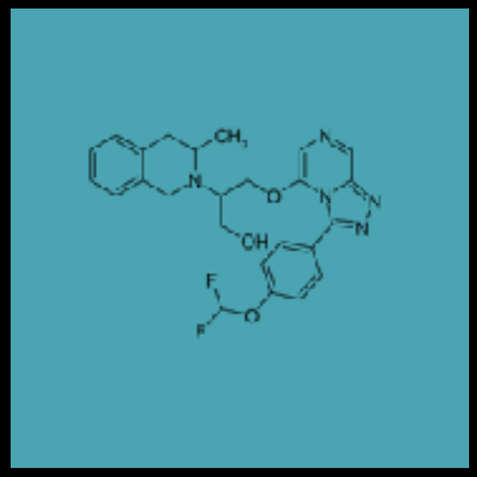
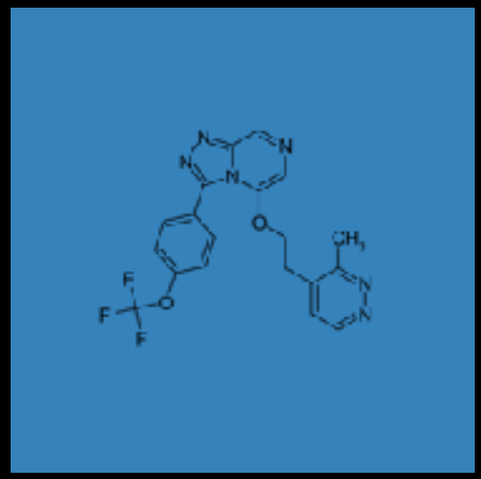
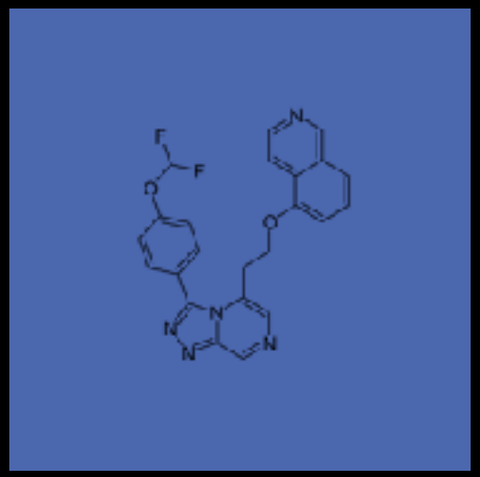
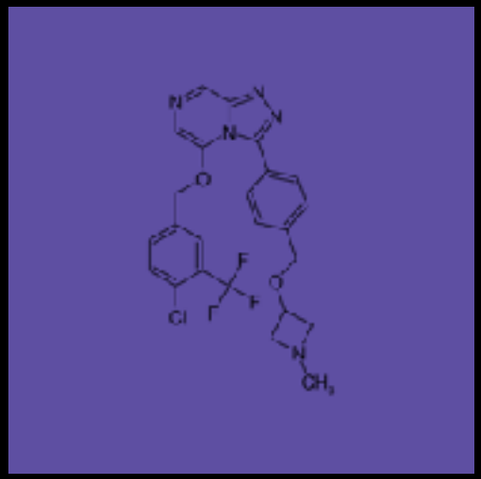
Event Fund

Drug candidates against the *P. falciparum* parasite



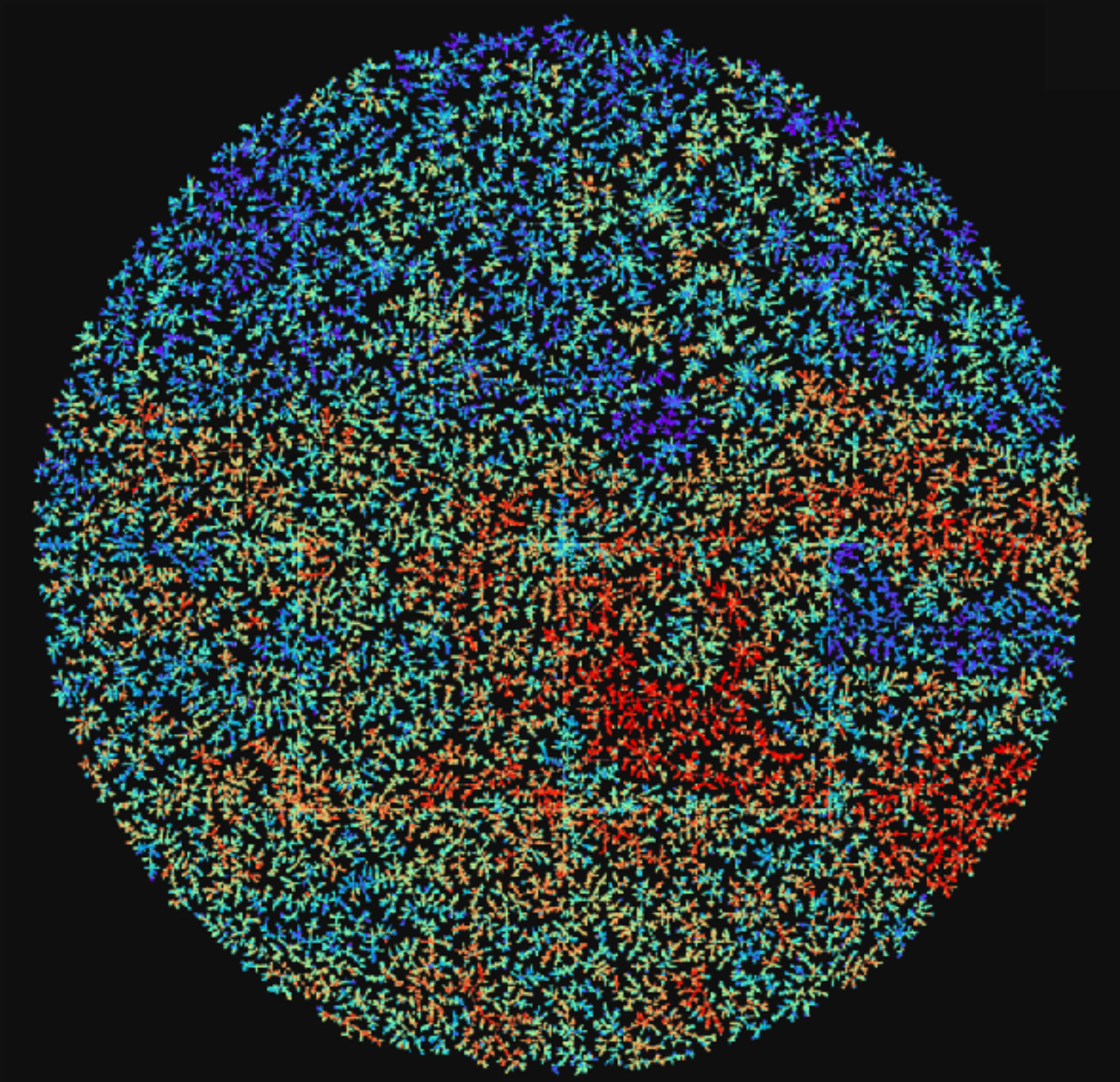
Contribution to the Open Source Malaria consortium

Event Fund

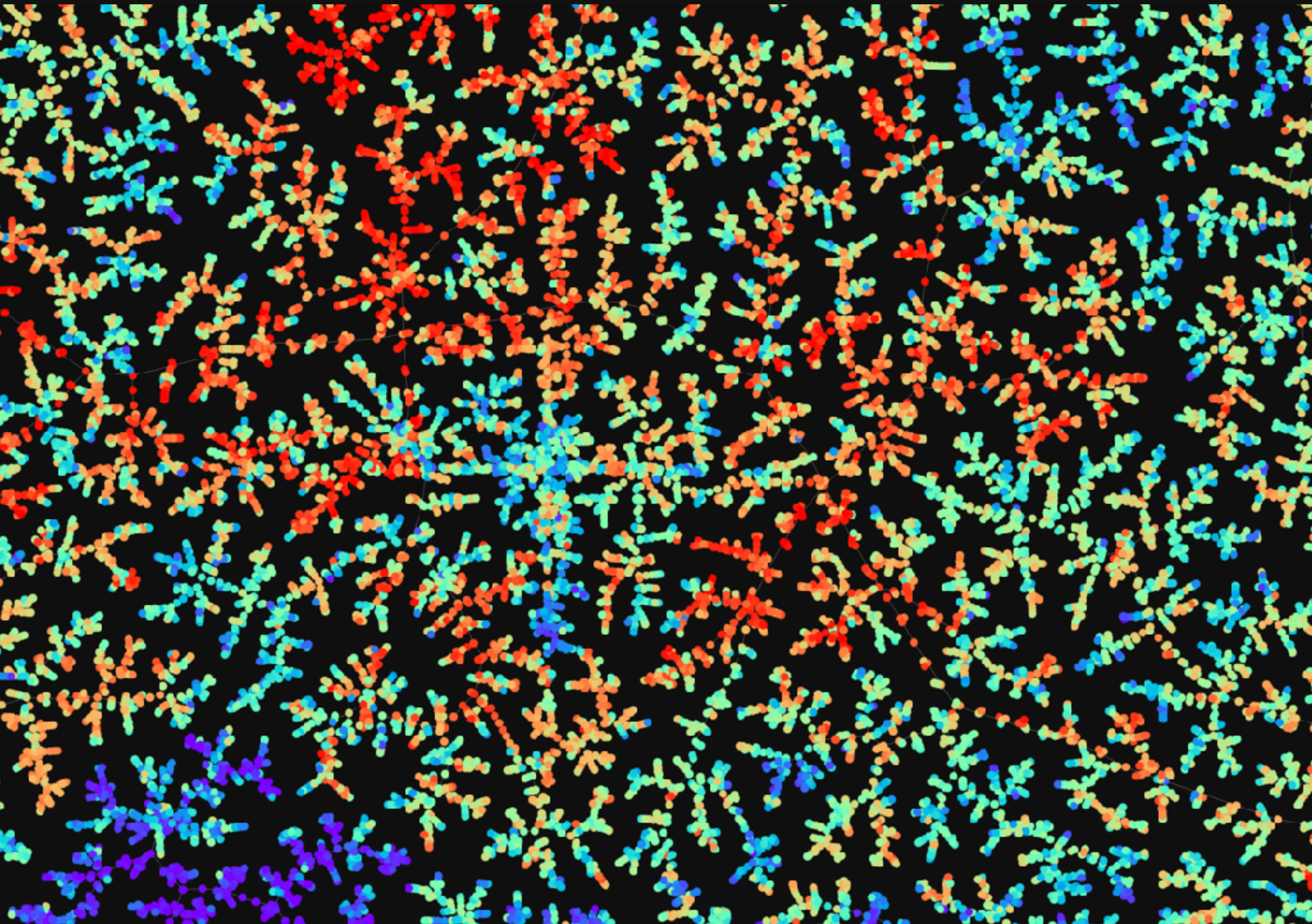













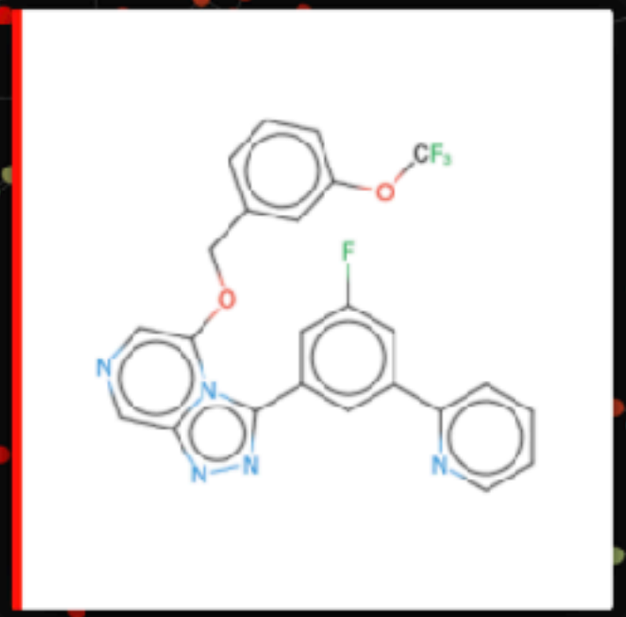
Event Fund

Drug candidates against the *P. falciparum* parasite

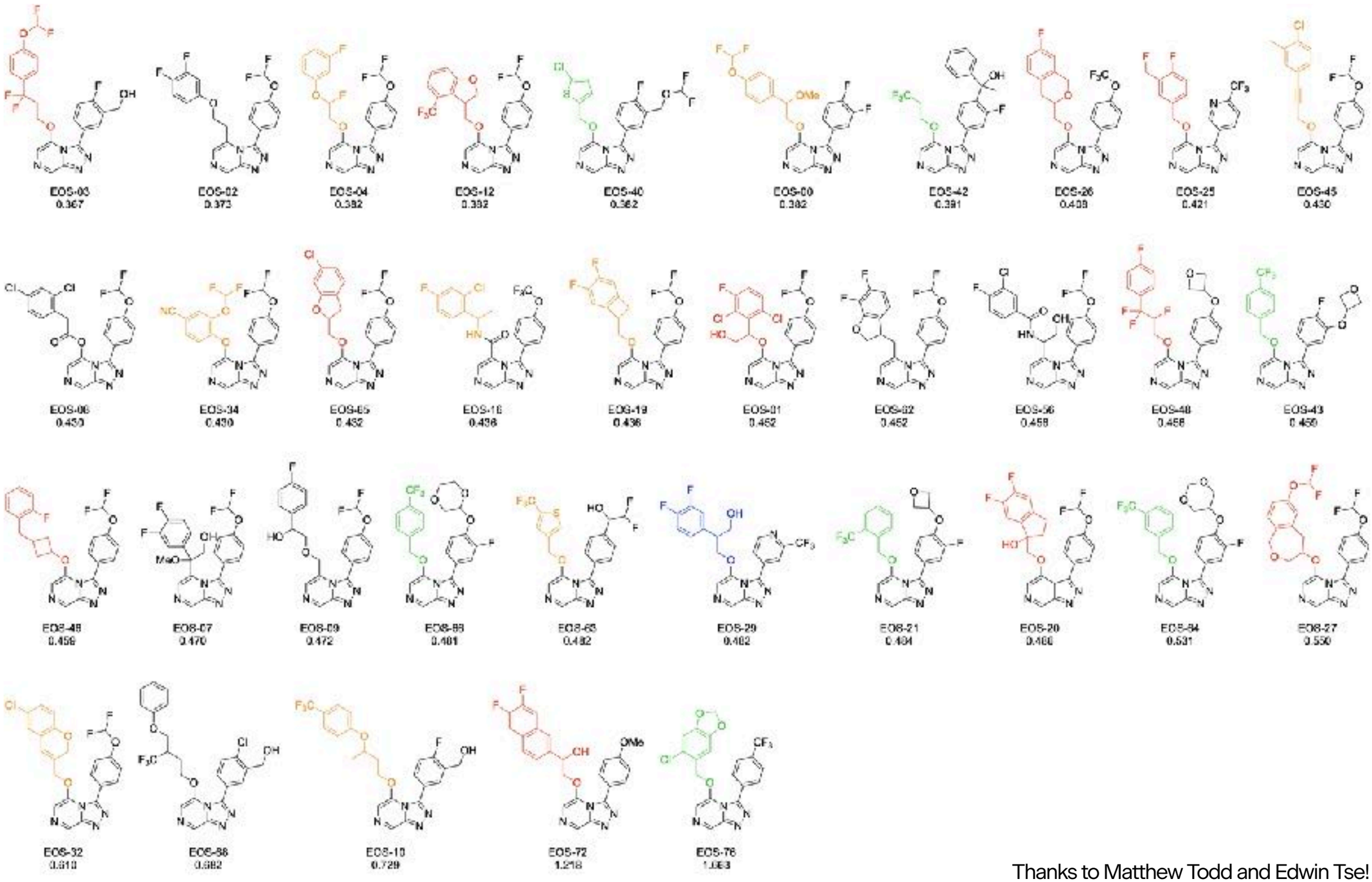


Contribution to the Open Source Malaria consortium




CF3
O
F
O
N
N
N
N
N

Open Source Malaria



Thanks to Matthew Todd and Edwin Tse!

https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/34

New Series 4 candidates based on generative model - EOSI #34

Edit New issue

Open miquelduranfrigola opened this issue on 1 Jun 2021 · 87 comments

miquelduranfrig... commented on 1 Jun 2021

Hello @mattodd @edwintse,

At @ersilia-os we have tried to generate new Series 4 candidates. In short, we provide two tables:

- A list of >100k molecules obtained with a generative model: download 100k
- A relatively diverse selection of 1k molecules: download 1k

For a first assessment of the results, you can check this dynamic visualization of the selected 1k candidates. If a cluster is of particular interest, please refer to the full results to discover other similar molecules. You can also check a tree map of all molecules.

Our generative model approach is based on Reinvent 2.0. We have implemented several reinforcement-learning agents, aimed at optimizing activity and other desirable properties. This GitHub Repository contains more detailed information and source code.

This is the first time we run a generative model, so please bear with us. We will be more than happy to optimize further runs based on your feedback.

Thanks!
@GemmaTuron @miquelduranfrigola

1

edwintse commented on 8 Jun 2021 Collaborator

Hi @miquelduranfrigola, thanks for all these great suggestions!

- Ideally we would be looking for suggestions that are a bit different from our existing compounds (e.g. we're less interested in single-point changes like adding halogens). Would it be possible to optimise to narrow down the list further?
- We were also wondering if, in the dynamic visualisation page, you were able to add a sliding filter for LogP? That would be helpful for us as a guide for solubility.

Assignees
mattodd
edwintse

Labels
Cheminformatics Collaboration Request enhancement

Projects
None yet

Milestone
No milestone

Development
No branches or pull requests

Notifications Customize
Unsubscribe
You're receiving notifications because you were mentioned.

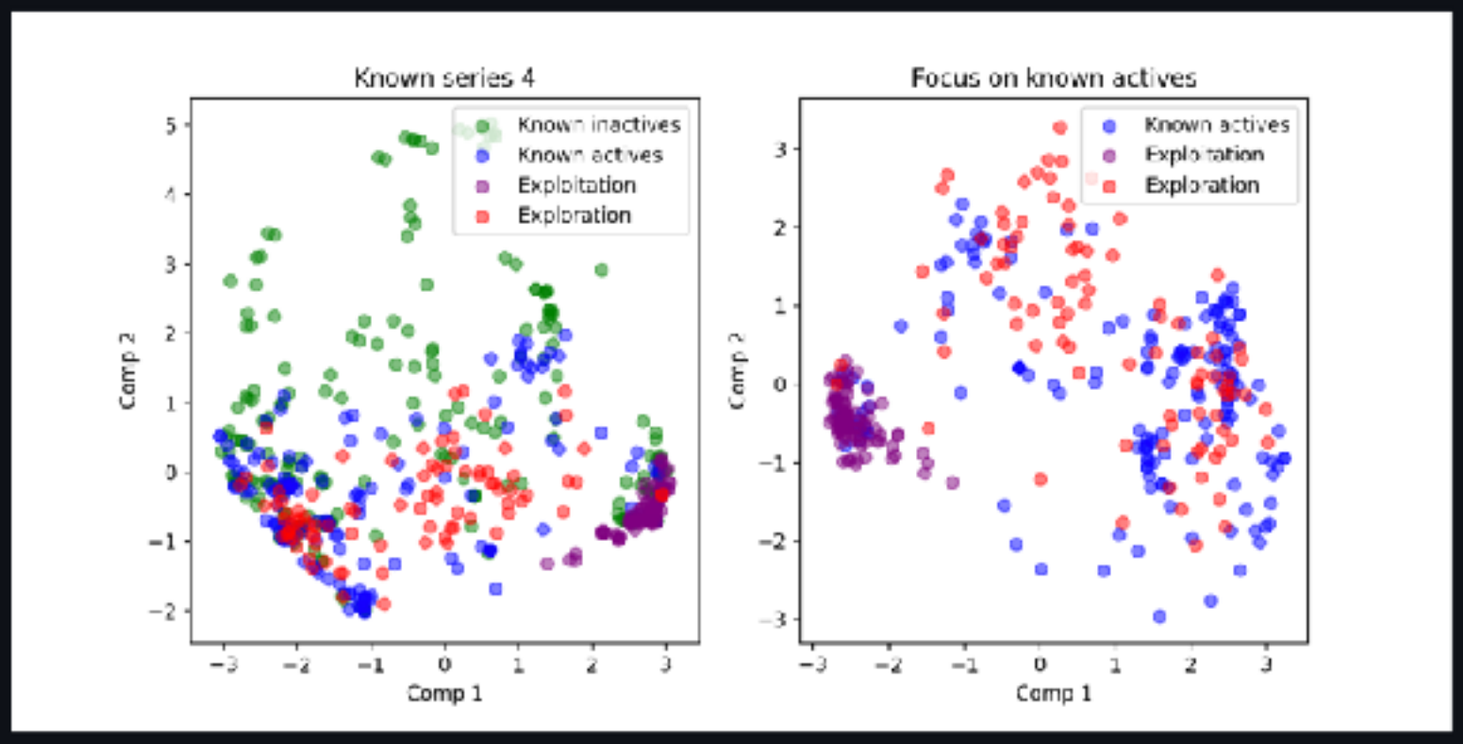
6 participants

miquelduranfrig... commented on 24 Jul 2021 • edited — Author

> Hi **@GemmaTuron**, I've been trying to make some compounds suggested by Evariste recently (#29) and was wondering if you guys ever generated any structures containing structures similar to those in this comment with indole/benzimidazole type groups on the RHS (or even any of the other structures that they predicted)? It would be interesting to see if there was any overlap between your suggestions and those from Evariste.

Hi @edwintse as you can see in the comment above by @GemmaTuron we have done a second round of generative models. To (sort of) answer your question, here two quick-and-dirty PCA plots (done with Morgan fingerprints) comparing:

1. **Known inactives (only left plot)**
2. **Known actives**
3. **Compounds in issue #29 (i.e. done in "exploitation" mode)**
4. **Our 90 selected compounds (i.e. done in "exploration" mode).**



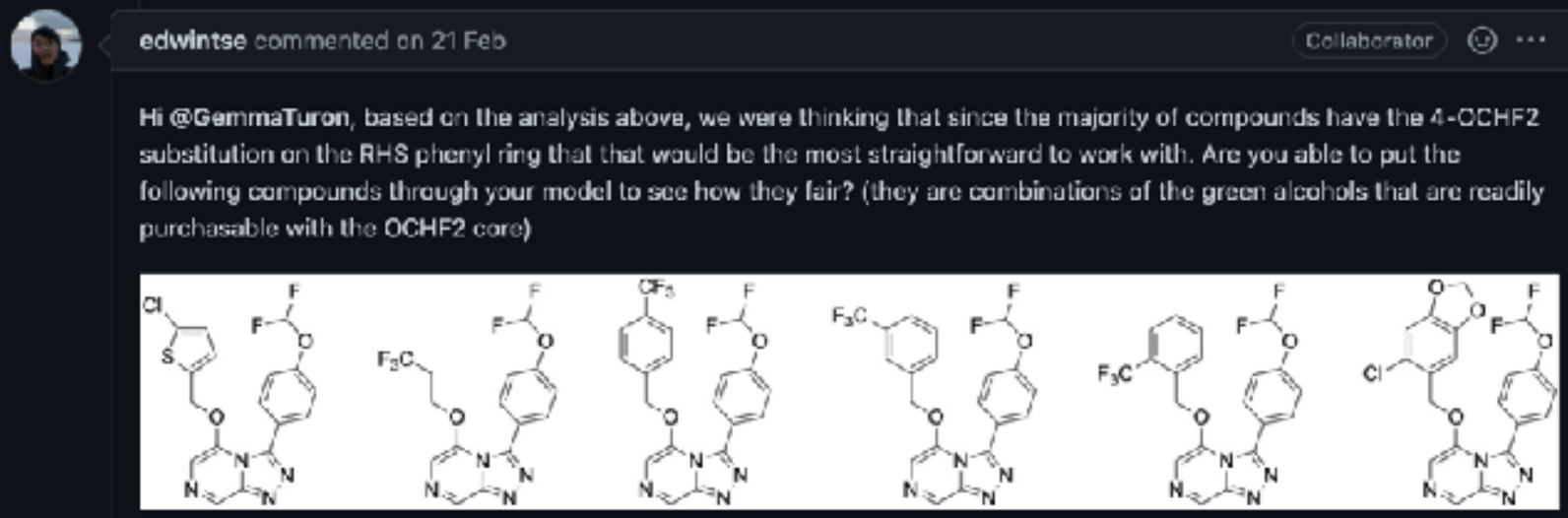
As you can see, we have a couple of compounds that cluster together with Evariste's compounds.

1

mattodd commented on 28 Jul 2021 — Member

OK @miquelduranfrigola @GemmaTuron this is most interesting. To make sure I understand:

The "exploitation" compounds are compounds you're predicting to be active that are derived fairly directly from other actives. The "exploration" compounds are those where you're intentionally trying to stay within the clusters of actives, and away from the inactives, yet which are sampling different areas of chemical space. So, in the left hand plot above we see no red Exploration compounds in regions where there are green inactives. In the right hand plot we're seeing exploration



edwintse commented on 21 Feb — Collaborator

Hi @GemmaTuron, based on the analysis above, we were thinking that since the majority of compounds have the 4-OCHF2 substitution on the RHS phenyl ring that that would be the most straightforward to work with. Are you able to put the following compounds through your model to see how they fair? (they are combinations of the green alcohols that are readily purchasable with the OCHF2 core)

ClC(S1)=CC=C1COC2=CN=CC3=NN=C(C4=CC=C(OC(F)F)C=C4)N32
FC(F)OC(C=C1)=CC=C1C2=NN=C3C=NC=C(OCCC(F)(F)F)N32
FC(F)OC(C=C1)=CC=C1C2=NN=C3C=NC=C(OCC4=CC=C(C(F)(F)F)C=C4)N32
FC(F)OC(C=C1)=CC=C1C2=NN=C3C=NC=C(OCC4=CC=CC(C(F)(F)F)=C4)N32
FC(F)OC(C=C1)=CC=C1C2=NN=C3C=NC=C(OCC4=CC=CC=C4C(F)(F)F)N32
ClC1=CC2=C(OCO2)C=C1COC3=CN=CC4=NN=C(C5=CC=C(OC(F)F)C=C5)N43

GemmaTuron commented on 21 Feb

Hi @edwintse! Thanks for checking the availability of purchasable compounds. Your suggestions seem very interesting, the second is the "less" active (still, predicted IC50 between 1 and 2.5 uM) and the rest are all < 1uM
I am attaching a .csv file containing the probability of a molecule of being active (1) when:

- using the model trained with a cut-off of 2.5 uM (probability_2.5)
- using a more restricted cut-off of 1 uM (probability_1)
- binarized activity prediction (0, inactive, 1, active) corresponds to the bin_activity for each cut-off.

You will see that the second molecule is the lowest scoring in the model that uses the 2.5 uM cutoff but is still predicted active, whereas the more restrictive cut-off predicts it as inactive. The rest all look highly active.
I am also adding the predicted activity in uM.

osm_sugg.csv

mattodd commented on 22 Feb — Member

Very interesting - that obviously makes some of the synthesis more straightforward. @edwintse is also looking into whether some of the interesting building blocks might be available through synthesis/purchase. I like some of the (sadly red)

Enamine €179/100 mg	FC £245/250 mg Enamine €179/100 mg	FC £281/250 mg	Chemspace in stock AA Blocks $153/100 mg Mcule awaiting quote

10.1016/j.bmc.2015.04.013

miquelduranfrig... commented on 23 Feb — Author

This looks great @edwintse and @mattodd ! Many thanks. Please let us know if you need further feedback on our side. In particular, if you come up with easier-to-synthesize/cheaper analogues, we will be happy to run predictions for them as @GemmaTuron did a few days ago.

mattodd commented on 25 Feb — Member

Great stuff @edwintse.

Top row - order the lot, except the last one.
Second row: don't know. Quite a lot of effort for a phenol that I bet won't be active. If the reagent in the first step is not nasty, then maybe it's worth making.
Third row - nice, make
Fourth row - nice, make, but we do lose some of the logP advantage of benzylic OHs etc (like EOS-12), but still a nice idea.
Fifth row - nice, make
Alkyne - would be nice to buy. Any cheaper with F rather than Cl (available from Biosynth - get if predicted to be OK by @GemmaTuron)? Or smaller bottle size?
Final thiophene - buy!

1

GemmaTuron commented on 21 Mar

Hi!
Just wanted to post here that we have been awarded a small grant by the Rosetrees Trust to test and optimize some of the compounds proposed here, so looking forward to the first experimental results!
@edwintse @mattodd any other test you want us to run before that?

edwintse commented on 22 Mar — Collaborator

Hi @GemmaTuron @miquelduranfrigola @mattodd, just updating on where I'm at currently with all the synthesis.

- There's a total of 11 compounds to make, the top row of which only requires commercially available alcohols. I've made 2 of them and have just received the reagents for the last 3. I will do those couplings tomorrow.
- The rest require the synthesis of the alcohol fragments. I've just done the difluoromethylation in the 2nd row. Need to convert it to the alcohol then do the coupling, perhaps next week.
- Synthesis of the fused ring fragment in the 4th row wasn't as nice as expected. I've carried through the first two steps crude and need to check if the ester is there.
- The final coupling for the 3rd row is done. Just need to purify.

Enamine €179/100 mg | FC £245/250 mg
Enamine €179/100 mg | FC £281/250 mg | Chemspace in stock
AA Blocks $153/100 mg
Mcule awaiting quote

10.1016/j.bmc.2015.04.013

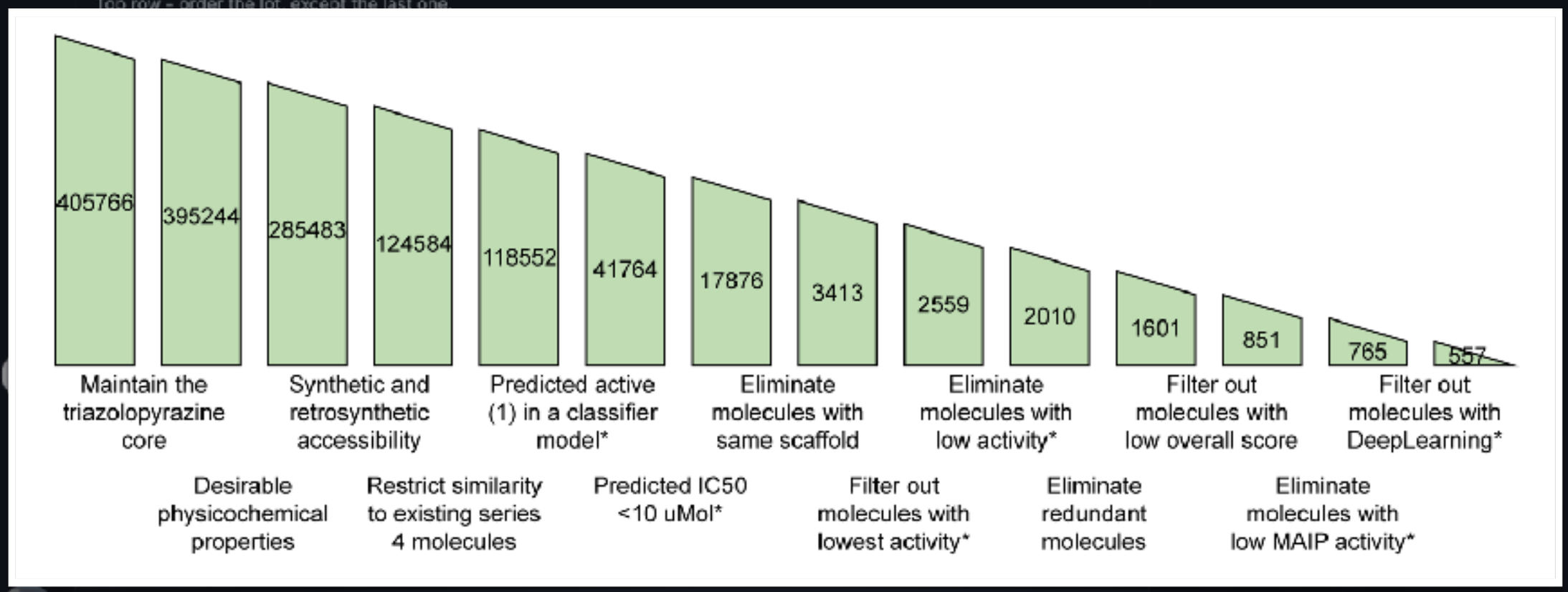
miquelduranfrig... commented on 23 Feb — Author

This looks great @edwintse and @mattodd ! Many thanks. Please let us know if you need further feedback on our side. In particular, if you come up with easier-to-synthesize/cheaper analogues, we will be happy to run predictions for them as @GemmaTuron did a few days ago.

mattodd commented on 25 Feb — Member

Great stuff @edwintse.

Top row - order the lot, except the last one.

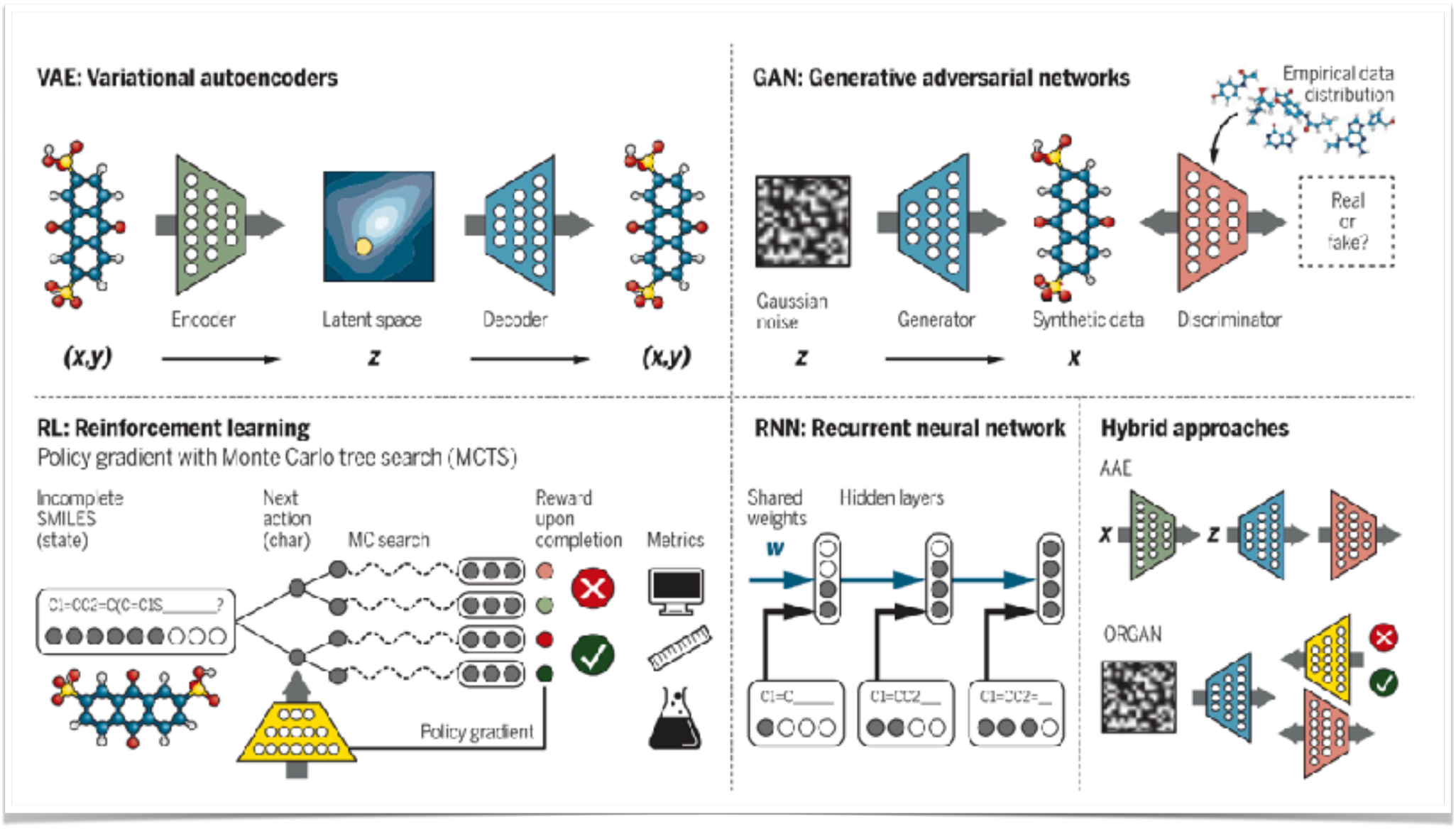


edwintse commented on 22 Mar — Collaborator

Hi @GemmaTuron @miquelduranfrigola @mattodd, just updating on where I'm at currently with all the synthesis.

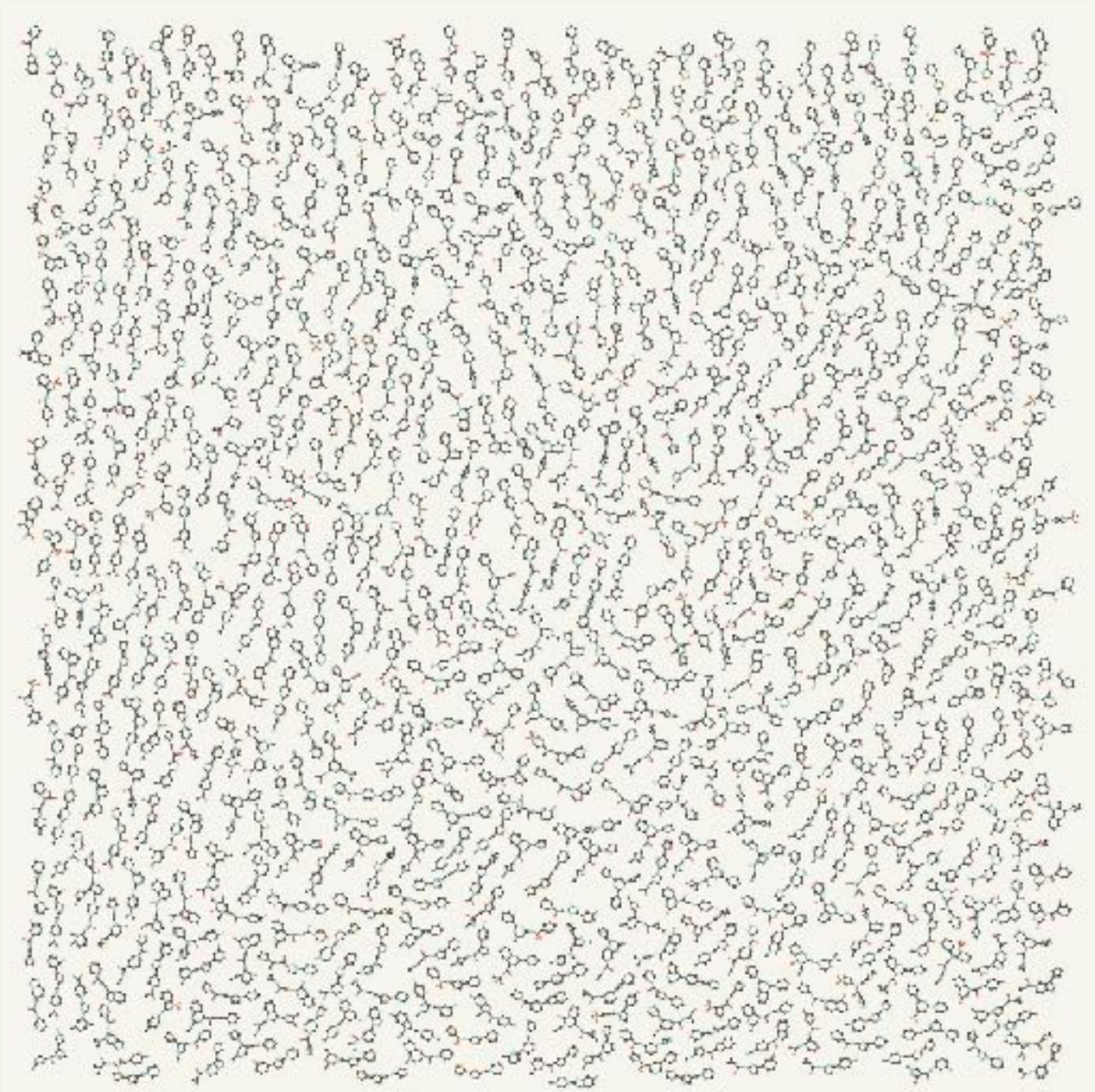
- There's a total of 11 compounds to make, the top row of which only requires commercially available alcohols. I've made 2 of them and have just received the reagents for the last 3. I will do those couplings tomorrow.
- The rest require the synthesis of the alcohol fragments. I've just done the difluoromethylation in the 2nd row. Need to convert it to the alcohol then do the coupling, perhaps next week.
- Synthesis of the fused ring fragment in the 4th row wasn't as nice as expected. I've carried through the first two steps crude and need to check if the ester is there.
- The final coupling for the 3rd row is done. Just need to purify.

Generative models

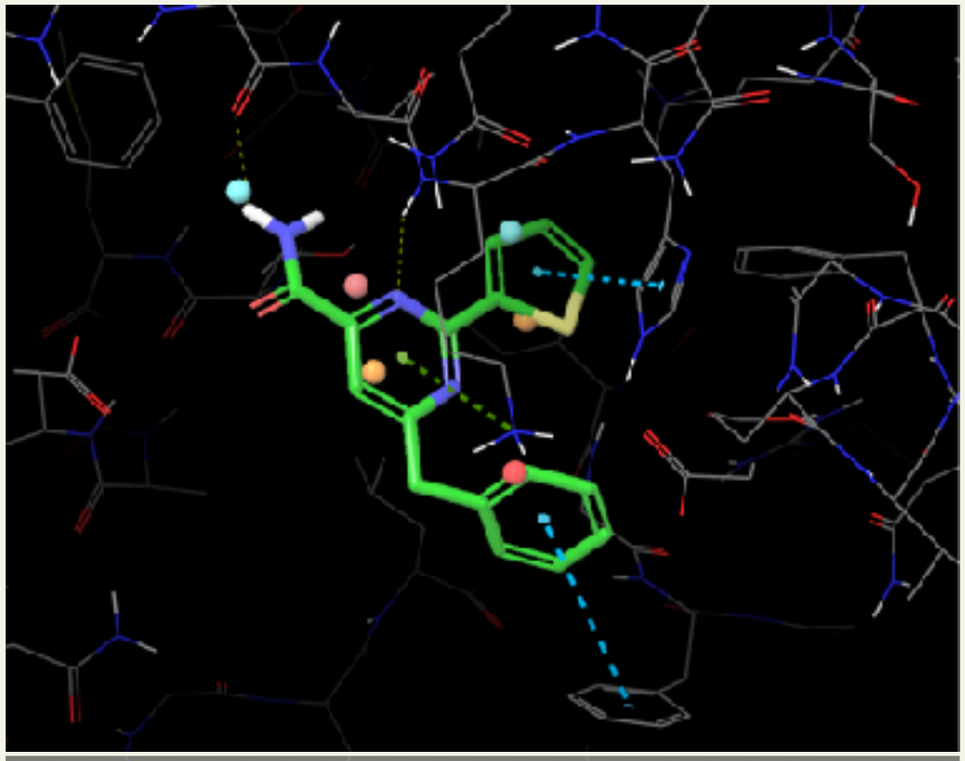


Sanchez-Lengeling, Science Advances, 2019

Open Source Antibiotics



– Structure-based + ML-based
– MurD ligase (*S. aureus*)
– 678 selected compounds
– 170 are purchasable



Thanks Arnau Comajuncosa!
In collaboration with the Structural Bioinformatics & Network Biology laboratory

Automated invention of compounds

SCIENCE ADVANCES | RESEARCH ARTICLE

CHEMISTRY

Combining generative artificial intelligence and on-chip synthesis for de novo drug design

Francesca Grisoni[1,2,3]†, Berend J. H. Huisman[1†], Alexander L. Button[1,3], Michael Moret[1], Kenneth Atz[1], Daniel Merk[1,4]*, Gisbert Schneider[1,5]*

Automating the molecular design–make-test-analyze cycle accelerates hit and lead finding for drug discovery. Using deep learning for molecular design and a microfluidics platform for on-chip chemical synthesis, liver X receptor (LXR) agonists were generated from scratch. The computational pipeline was tuned to explore the chemical space of known LXRα agonists and generate novel molecular candidates. To ensure compatibility with automated on-chip synthesis, the chemical space was confined to the virtual products obtainable from 17 one-step reactions. Twenty five de novo designs were successfully synthesized in flow. In vitro screening of the crude reaction products revealed 17 (68%) hits, with up to 60-fold LXR activation. The batch resynthesis, purification, and retesting of 14 of these compounds confirmed that 13 of them were potent LXR agonists. These results support the suitability of the proposed design–make-test-analyze framework as a blueprint for automated drug design with artificial intelligence and miniaturized bench-top synthesis.

INTRODUCTION

Rapid iteration of the molecular design–make-test-analyze (DMTA) cycle has the potential for making "better decisions faster" (*1*, *2*), with numerous applications in drug discovery and related fields (*3*, *4*). Recent advances in chemical reaction monitoring and optimization, computing hardware, and algorithms have boosted the automation of several parts of the drug discovery process, such as robotic synthesis (*5*–*8*), computational molecular design (*9*–*12*), and synthesis planning (*13*–*15*). Standardized experimental procedures with robotic assistance increase the reproducibility of results, reduce errors, and decrease the consumption of materials, thereby contributing to "green chemistry" (*16*). Furthermore, reasoning with machine intelligence supports the discovery of novel drug-like molecules by freeing the molecular design and optimization process from personal biases (*1*). Pioneering studies combined microfluidics platforms with machine intelligence for synthesis planning (*7*, *17*) as well as automated hit finding and hit-to-lead optimization in combinatorial libraries (*8*, *18*). Computer-assisted molecular design is a critical element of this automation process. Molecular structure generation is often performed in a "rule based" manner, i.e., by using algorithms for molecule assembly from predefined virtual reactions and reactants (*19*). Generative deep learning models extend the capabilities of rule-based de novo molecule generators by sampling new molecules from almost chemical space representation (*20*–*23*), without the need for human-crafted molecule construction rules. Recently, the prospective applicability of "rule-free" generative deep learning for de novo molecular design has been demonstrated in combination with batch synthesis (*9*, *10*, *24*–*26*).

This study aims to pioneer the integration of generative molecular design with automated synthesis. Here, a recently published generative deep learning model (*27*) was adapted to generate compounds that are at the same time (i) bioactive on a selected macromolecular target and (ii) synthesizable on a bench-top microfluidic synthesis platform (*16*, *28*). We challenged this automated DMTA pipeline to design liver X receptor (LXR) agonists from scratch, with minimal human interference. LXRs have emerged as promising drug targets because of their regulatory role in lipid metabolism and inflammation, thereby causing increased reverse cholesterol transport and reduction of atherosclerosis (*29*–*32*). With 28 molecules successfully synthesized and 13 fully validated for LXR activation in vitro, this present study pioneers the integration of generative artificial intelligence and automated synthesis by designing and experimentally testing the highest number of molecules reported thus far. The proposed modular framework has the potential to accelerate the DMTA cycle, thereby addressing one of the main bottlenecks of the preclinical drug discovery process (*33*).

RESULTS AND DISCUSSION

Modular DMTA platform

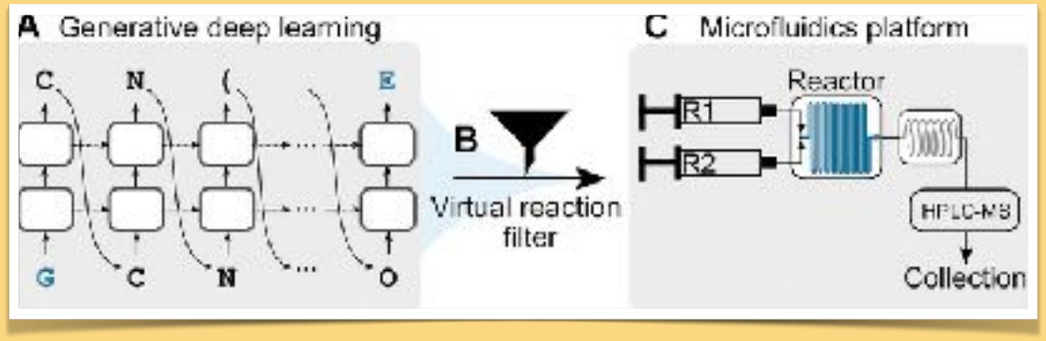
The automated molecular design pipeline was composed of three modules (Fig. 1):

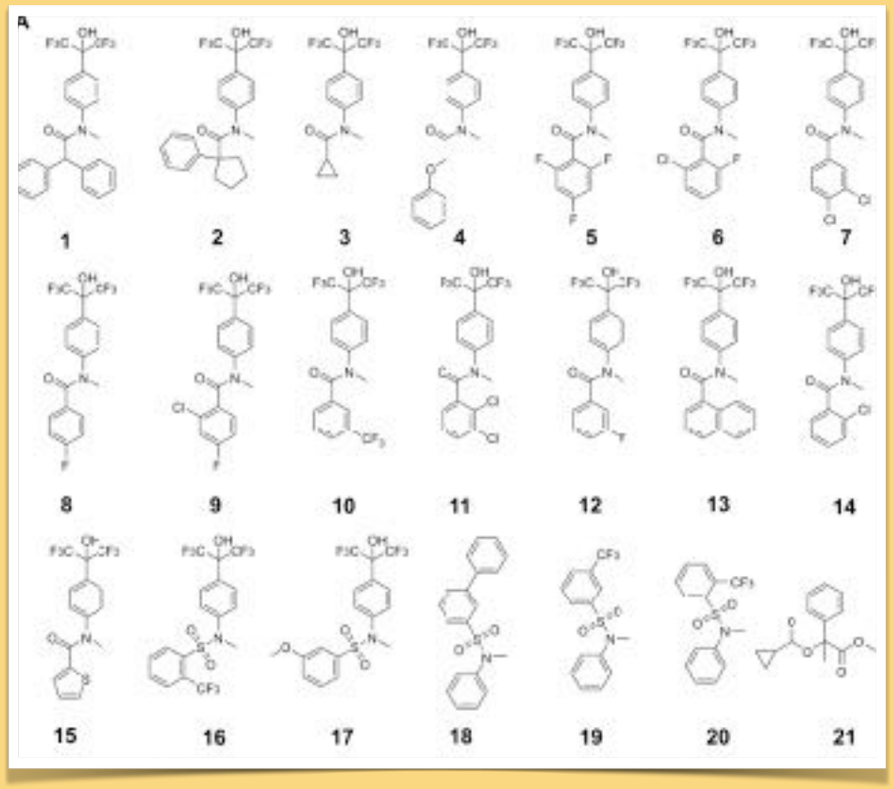
1) Module 1: A generative deep learning model (*27*) based on a recurrent neural network with long short-term memory (LSTM) cells (*34*). LSTM models were used for the design of new molecules represented as simplified molecular input line entry systems (SMILES) (*35*) strings (*20*, *21*, *36*). This LSTM-based "chemical language model" served as the de novo structure generator (Fig. 1A).

2) Module 2: A virtual reaction filter that captured 17 one-step reactions that were compatible with the microfluidics system (module 3). These reactions were encoded as SMILES arbitrary target specification (SMARTS) strings (*37*) (table S1). This filtering module selected those generated molecules that were synthetically compatible within the microfluidics platform (Fig. 1B).

3) Module 3: A microfluidics platform designed to minimize the amount of manual labor needed to optimize reaction conditions and synthesize focused compound libraries via one-step reactions. This compact bench-top system combined the automated retrieval of

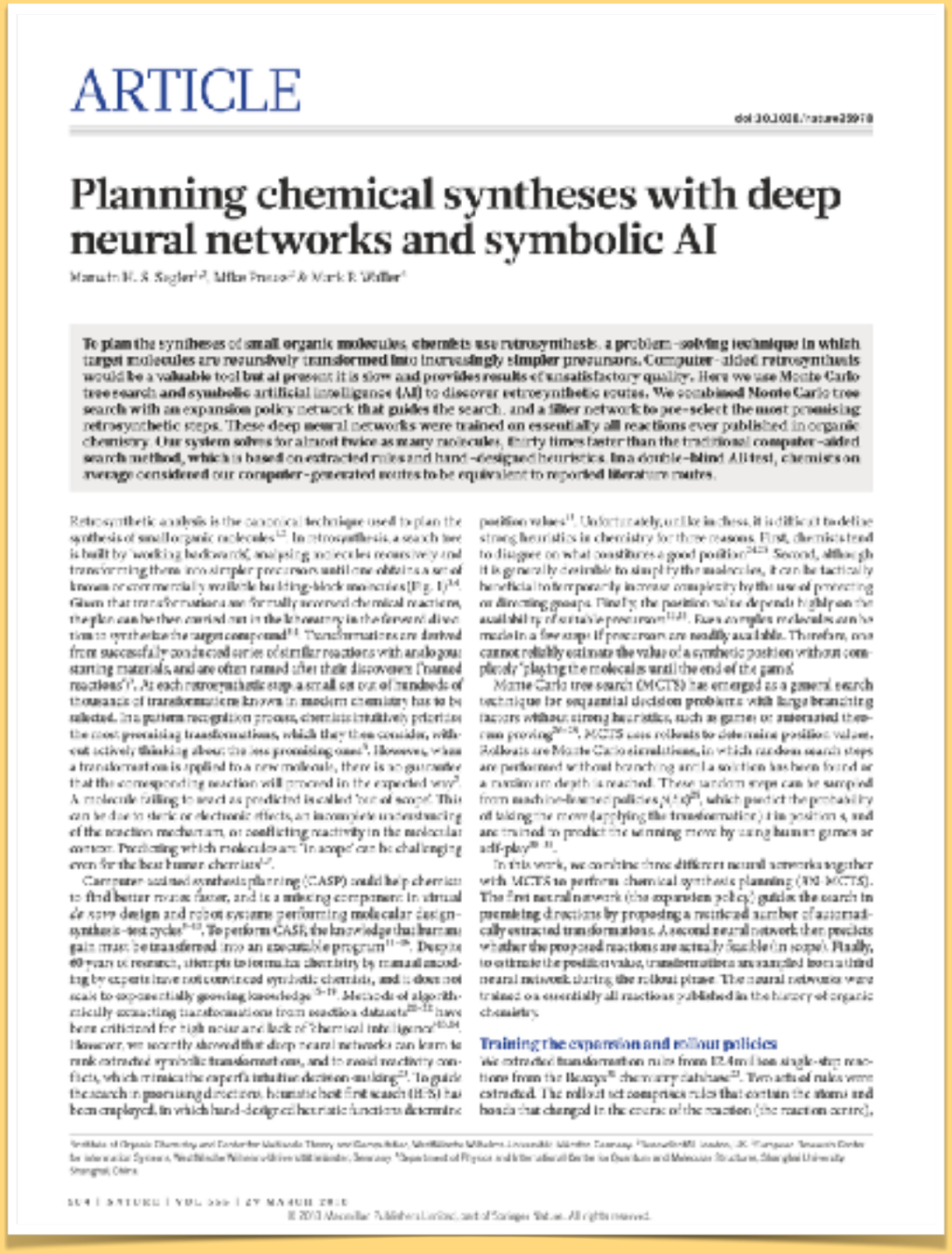
[1]ETH Zurich, Department of Chemistry and Applied Biosciences, RETHINK, Zurich, Switzerland. [2]Eindhoven University of Technology, Department of Biomedical Engineering, Eindhoven, Netherlands. [3]University of Lausanne, Department of Computational Biology, Lausanne, Switzerland. [4]Goethe University, Institute of Pharmaceutical Chemistry, Frankfurt, Germany. [5]ETH Singapore SEC Ltd, Singapore, Singapore.
*Corresponding author. Email: gschneider@ethz.ch (G.S.); merk@pharmchem.uni-frankfurt.de (D.M.)
†These authors contributed equally to this work.





Grisoni et al, Science Advances, 2021

Automated synthesis planning



ARTICLE

doi:10.1038/nature25978

Planning chemical syntheses with deep neural networks and symbolic AI

Marwin H. S. Segler[1,2], Mike Preuss[2] & Mark P. Waller[1]

To plan the syntheses of small organic molecules, chemists use retrosynthesis, a problem-solving technique in which target molecules are recursively transformed into increasingly simpler precursors. Computer-aided retrosynthesis would be a valuable tool but at present it is slow and provides results of unsatisfactory quality. Here we use Monte Carlo tree search and symbolic artificial intelligence (AI) to discover retrosynthetic routes. We combined Monte Carlo tree search with an expansion policy network that guides the search, and a filter network to pre-select the most promising retrosynthetic steps. These deep neural networks were trained on essentially all reactions ever published in organic chemistry. Our system solves for almost twice as many molecules, thirty times faster than the traditional computer-aided search method, which is based on extracted rules and hand-designed heuristics. In a double-blind A/B test, chemists on average considered our computer-generated routes to be equivalent to reported literature routes.

Retrosynthetic analysis is the canonical technique used to plan the synthesis of small organic molecules[1,2]. In retrosynthesis, a search tree is built by working backwards, analysing molecules recursively and transforming them into simpler precursors until one obtains a set of known or commercially available building-block molecules (Fig. 1)[3,4]. Given that transformations are formally reversed chemical reactions, the plan can be then carried out in the laboratory in the forward direction to synthesize the target compound[1,3]. Transformations are derived from successfully conducted series of similar reactions with analogous starting materials, and are often named after their discoverers ('named reactions')[1]. At each retrosynthetic step, a small set out of hundreds of thousands of transformations known in modern chemistry has to be selected. In a pattern recognition process, chemists intuitively prioritize the most promising transformations, which they then consider, without actively thinking about the less promising ones[5]. However, when a transformation is applied to a new molecule, there is no guarantee that the corresponding reaction will proceed in the expected way[2]. A molecule failing to react as predicted is called 'out of scope'. This can be due to steric or electronic effects, an incomplete understanding of the reaction mechanism, or conflicting reactivity in the molecular context. Predicting which molecules are 'in scope' can be challenging even for the best human chemists[1,2].

Computer-assisted synthesis planning (CASP) could help chemists to find better routes faster, and is a unifying component in virtual *de novo* design and robot systems performing molecular design-synthesis-test cycles[6–12]. To perform CASP, the knowledge that human gain must be transferred into an executable program[11–16]. Despite 60 years of research, attempts to formalize chemistry by manual encoding by experts have not convinced synthetic chemists, and it does not scale to exponentially growing knowledge[3–10]. Methods of algorithmically extracting transformations from reaction datasets[10–12] have been criticized for high noise and lack of 'chemical intelligence'[10,14]. However, we recently showed that deep neural networks can learn to rank extracted synthetic transformations, and to avoid reactivity conflicts, which mimics the expert's intuitive decision-making[13]. To guide the search in promising directions, heuristic best-first search (BFS) has been employed, in which hand-designed heuristic functions discriminate position values[11]. Unfortunately, unlike in chess, it is difficult to define strong heuristics in chemistry for three reasons. First, chemists tend to disagree on what constitutes a good position[14,15]. Second, although it is generally desirable to simplify the molecules, it can be tactically beneficial to temporarily increase complexity by the use of protecting or directing groups. Third, the position value depends highly on the availability of certain precursors[11,14]. Even complex molecules can be made in a few steps if precursors are readily available. Therefore, one cannot reliably estimate the value of a synthetic position without completely 'playing the molecules until the end of the game'.

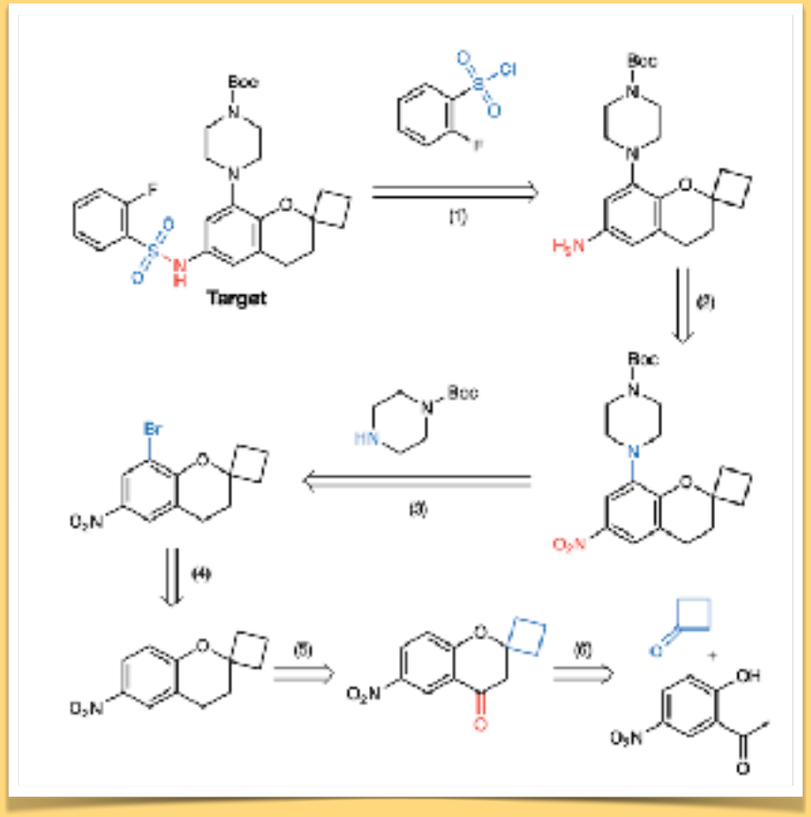
Monte Carlo tree search (MCTS) has emerged as a general search technique for sequential decision problems with large branching factors without strong heuristics, such as games or automated theorem proving[20–28]. MCTS uses rollouts to determine position values. Rollouts are Monte Carlo simulations, in which random search steps are performed without branching until a solution has been found or a maximum depth is reached. These random steps can be sampled from machine-learned policies[24,29], which predict the probability of taking one move (applying the transformation) i in position s, and are trained to predict the winning move by using human games or self-play[30–33].

In this work, we combine three different neural networks together with MCTS to perform chemical synthesis planning (Fig. 2|MCTS). The first neural network ('the expansion policy') guides the search in promising directions by proposing a restricted number of automatically extracted transformations. A second neural network then predicts whether the proposed reactions are actually feasible ('in scope'). Finally, to estimate the position value, transformations are applied from a third neural network during the rollout phase. The neural networks were trained on essentially all reactions published in the history of organic chemistry.

Training the expansion and rollout policies

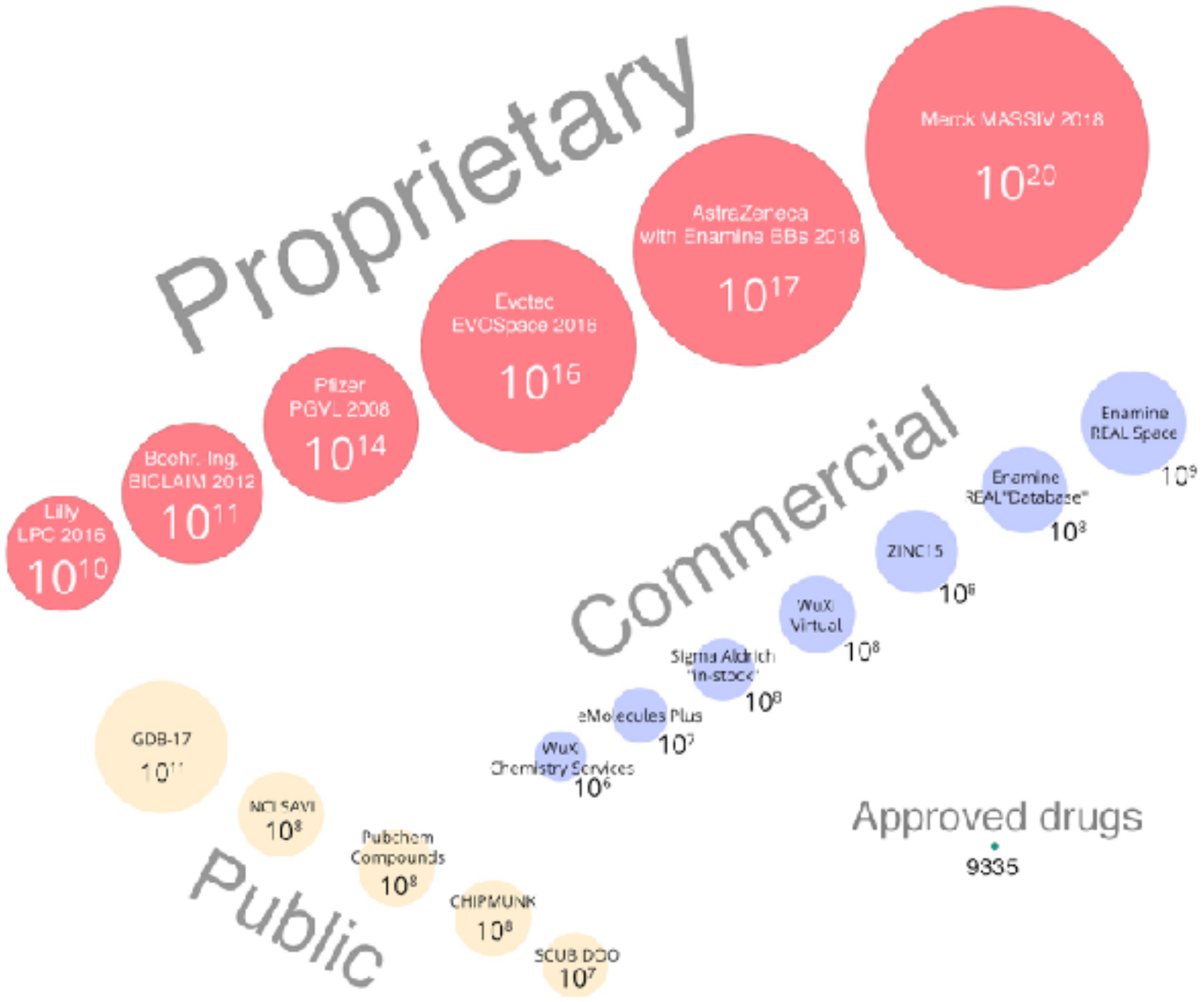
We extracted transformation rules from 12.4 million single-step reactions from the Reaxys® chemistry database[22]. Two sets of rules were extracted. The rollout set comprises rules that contain the atoms and bonds that changed in the course of the reaction (the reaction centre),

Institute of Organic Chemistry and Center for Multiscale Theory and Computation, Westfälische Wilhelms-Universität Münster, Corrensstraße 40, 48149 Münster, Germany. [2]Computer Science Department, TU Dortmund, 44221 Dortmund, Germany. [3]Department of Physics and Astronomy and Center for Multiscale Theory and Computation, Westfälische Wilhelms-Universität Münster, Corrensstraße 40, 48149 Münster, Germany. [4]Computer Science Center for Molecular Systems, Westfälische Wilhelms-Universität Münster, Corrensstraße 40, 48149 Münster, Germany. [5]Department of Physics and Center for Quantum and Molecular Structures, Shanghai University, Shanghai, China.



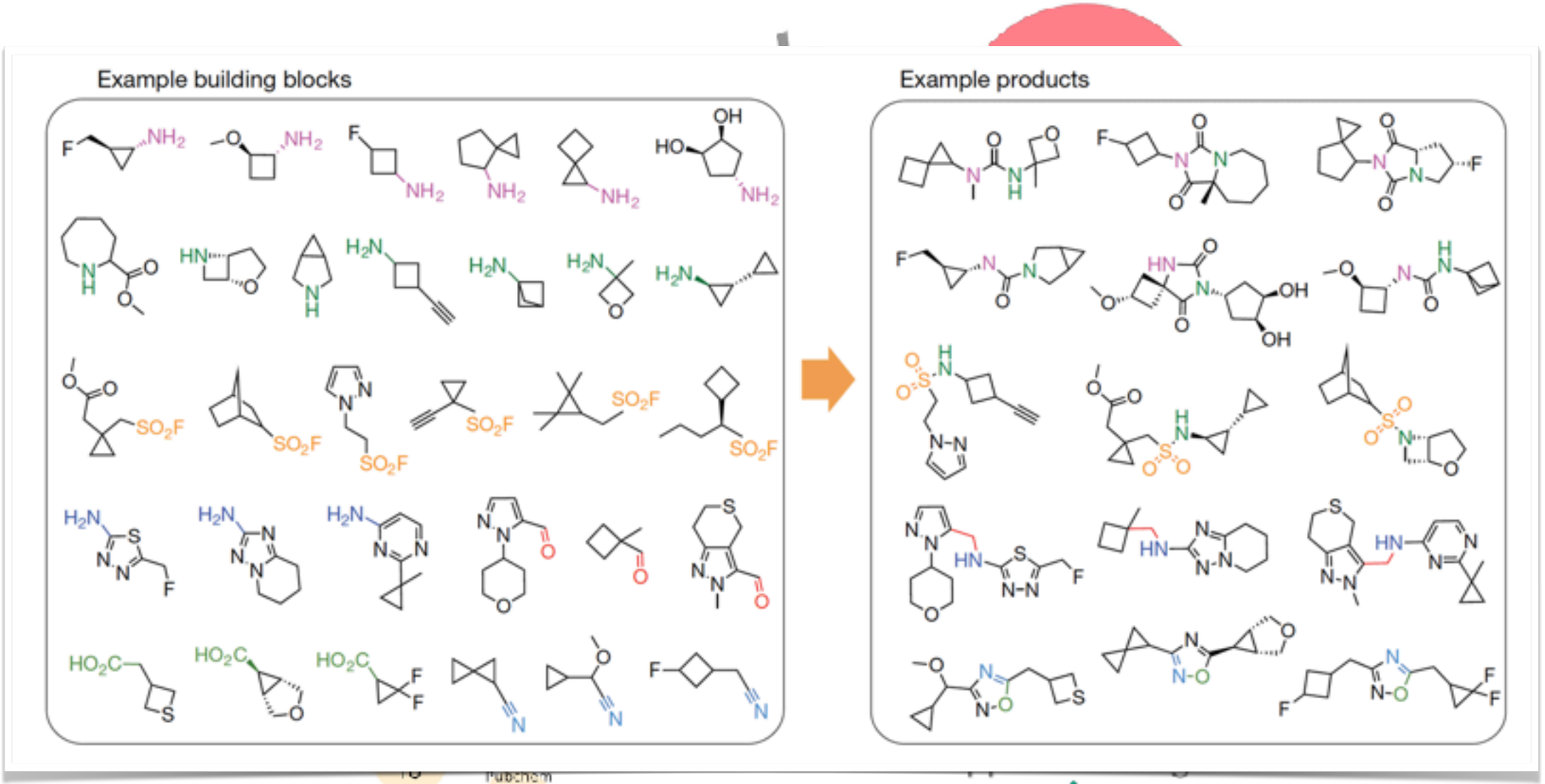
Segler et al, Nature, 2018

Ultra-large scale chemical libraries



Hoffmann et al, Drug Discovery Today, 2019

Ultra-large scale chemical libraries



Hoffmann et al, Drug Discovery Today, 2019

Ultra-large scale virtual screening

Article

Synthon-based ligand discovery in virtual libraries of over 11 billion compounds

https://doi.org/10.1038/s41589-021-04220-8
Received: 17 February 2021
Accepted: 8 November 2021
Published online: 16 December 2021
Check for updates

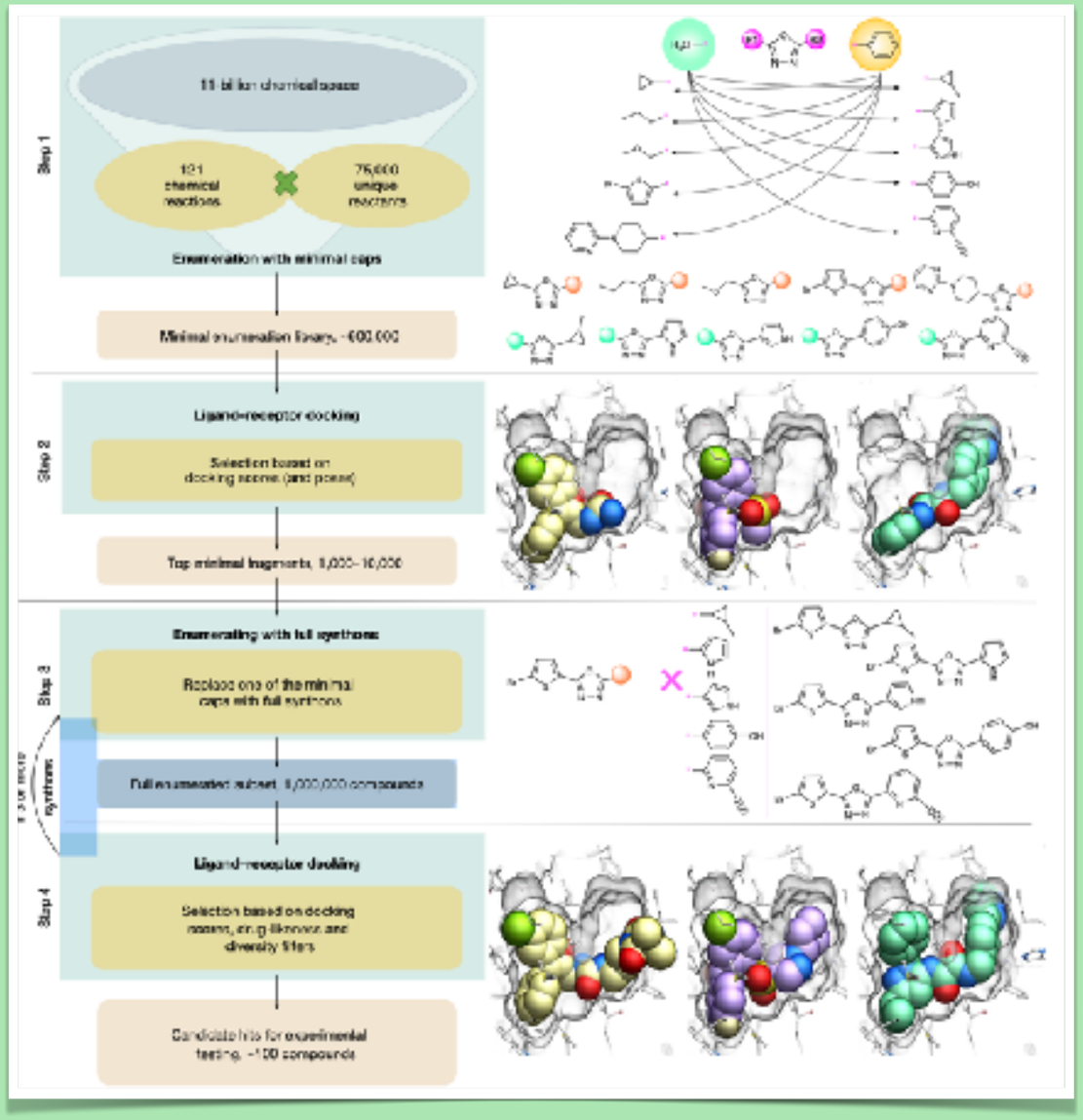
Arman A. Sadybekov[1,10], Anastasiia V. Sadybekov[2,10], Yongfeng Liu[3,4,10], Christos Iliopoulos-Tsoutsouvas[5,6], Xi-Ping Huang[7,8], Julie Pickett[9], Blake Houser[1], Nilkanth Patel[1], Ngan K. Tran[1], Fei Tong[1], Nikolai Zvonok[1], Manish K. Jain[1], Olena Savych[2], Dmytro S. Radchenko[2,11], Spyros P. Nikas[5], Nicos A. Petasis[5], Yuri S. Moroz[2,12], Bryan L. Roth[7,8,13], Alexandros Makriyannis[3,4,14] ✉, Vsevolod Katritch[1,14] ✉

Structure-based virtual ligand screening is emerging as a key paradigm for early drug discovery owing to the availability of high-resolution target structures[1–4] and ultra-large libraries of virtual compounds[5,6]. However, to keep pace with the rapid growth of virtual libraries, such as readily available for synthesis (REAL) combinatorial libraries[7], new approaches to compound screening are needed[8,9]. Here we introduce a modular synthon-based approach—V-SYNTHES—to perform hierarchical structure-based screening of a REAL Space library of more than 11 billion compounds. V-SYNTHES first identifies the best scaffold–synthon combinations and subsequently, for further growth, and then iteratively elaborates these seeds to select complete molecules with the best docking scores. This hierarchical combinatorial approach enables the rapid detection of the best-scoring compounds in the gigascale chemical space while performing docking of only a small fraction (<0.1%) of the library compounds. Chemical synthesis and experimental testing of novel cannabinoid antagonists predicted by V-SYNTHES demonstrated a 33% hit rate, including 14-subnanomolar ligands, substantially improving over a standard virtual screening of the Enamine REAL diversity subset, which required approximately 100 times more computational resources. Synthesis of selected analogues of the best hits further improved potencies and affinities (best inhibitory constant (K_i) = 0.9 nM) and CB_2/CB_1 selectivity (50–200-fold). V-SYNTHES was also tested on a kinase target, ROCK1, further supporting its use for lead discovery. The approach is easily scalable for the rapid growth of combinatorial libraries and potentially adaptable to any docking algorithm.

Standard libraries for high-throughput screening (HTS)[10] and virtual ligand screening (VLS)[11–13] have been historically limited to fewer than 10 million available compounds, which is a small fraction of the enormous chemical space, estimated to be 10^{30} to 10^{60} drug-like compounds[14,15]. This limitation of standard HTS and VLS slows the pace of drug discovery, usually yielding initial hits with modest affinities, poor selectivity and ADME/T profiles that require elaborate multistep optimization to grow lead- and drug-like candidate properties. Recently, ultra-large libraries of more than 100 million readily accessible (REAL) compounds have been developed and used in docking-based VLS, yielding high-quality hits for lead discovery[5,6]. The Enamine REAL library, which now comprises 1.4 billion compounds, and the REAL Space expansion with more than 11 billion drug-like compounds, take coverage of modular parallel synthesis with a large set of optimized reactions and building blocks (synthons)[7]. This makes the synthesis of potential hit compounds fast (less than 4–6 weeks), reliable (>80% success rate) and affordable.

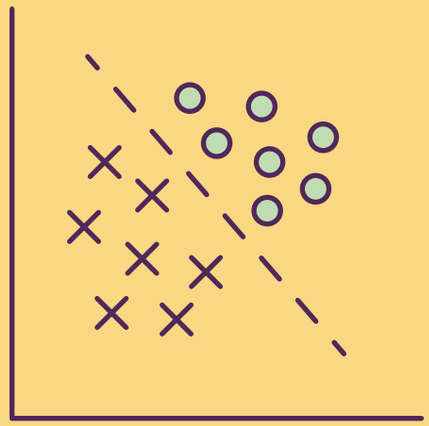
The modular nature of REAL libraries supports their further rapid growth way beyond 20 billion drug-like compounds[16]. However, with increasing library sizes, the computational time and cost of docking-based VLS itself become the most bottleneck in screening, even with massively parallel cloud-computing capacities. For example, the docking of 20 billion compounds at a standard rate of 10 s per compound would take more than 3,000 years on a single CPU core, or cost over US$80,000 on a computing cloud. The ability to substantially reduce the computational burden of VLS without compromising the

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [2]Enamine Ltd, Kyiv, Ukraine. [3]Center for Drug Discovery, Northeastern University, Boston, MA, USA. [4]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA. [5]Department of Pharmaceutical Sciences, University of Southern California, Los Angeles, CA, USA. [6]Department of Chemistry, University of Southern California, Los Angeles, CA, USA. [7]Department of Pharmacology, School of Medicine, University of North Carolina, Chapel Hill, NC, USA. [8]National Institute of Mental Health, Psychoactive Drug Screening Program, School of Medicine, University of North Carolina, Chapel Hill, NC, USA. [9]Chemistry Department, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. [10]These authors contributed equally: Arman A. Sadybekov, Anastasiia V. Sadybekov, Yongfeng Liu, Christos Iliopoulos-Tsoutsouvas. [11]Present address: Chemspace LLC, Kyiv, Ukraine. [12]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA. [13]Present address: Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA. [14]These authors jointly supervised this work: Alexandros Makriyannis, Vsevolod Katritch. ✉e-mail: vkatritch@usc.edu; amakriyannis@northeastern.edu; bryan_roth@med.unc.edu



Sadybekov, Nature, 2022

Supervised



Labeled data
Classification
Regression

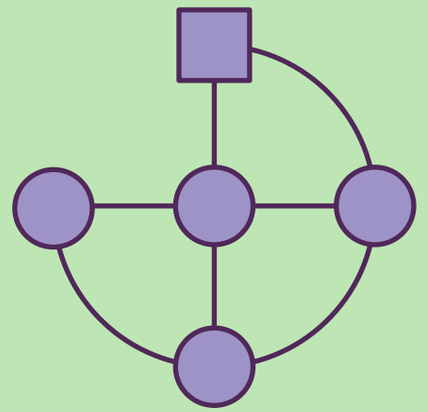
Bioactivity prediction
ADMET properties

Unsupervised

Unlabeled data
Clustering
2D projection
Similarity search

Visualisation of
chemical libraries

Reinforcement



Interaction with
environment / agent
Generative models

Library design
Hit-to-lead optimization

Take-home message

— Stay in touch after the workshop! 🤗

Bringing data science and AI/ML tools to infectious disease research

Session 4: Open Science and Generative Models

Event Sponsors