An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials


Submitted date: 05/11/2020 • Posted date: 06/11/2020
Licence: CC BY-NC-ND 4.0
Citation information: Tse, Edwin; Aithani, Laksh; Anderson, Mark; Cardoso-Silva, Jonathan; Cincilla, Giovanni; Conduit, Gareth J.; et al. (2020): An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13194755.v1

The discovery of new antimalarial medicines with novel mechanisms of action is key to combating the problem of increasing resistance to our frontline treatments. The Open Source Malaria (OSM) consortium has been developing compounds (“Series 4”) that have potent activity against Plasmodium falciparum in vitro and in vivo and that have been suggested to act through the inhibition of PfATP4, an essential membrane ion pump that regulates the parasite’s intracellular Na⁺ concentration. The structure of PfATP4 is yet to be determined. In the absence of structural information about this target, a public competition was created to develop a model that would allow the prediction of anti-PfATP4 activity among Series 4 compounds, thereby reducing project costs associated with the unnecessary synthesis of inactive compounds. In the first round, in 2016, six participants used the open data collated by OSM to develop moderately predictive models using diverse methods. Notably, all submitted models were available to all other participants in real time. Since then further bioactivity data have been acquired and machine learning methods have rapidly developed, so a second round of the competition was undertaken, in 2019, again with freely-donated models that other participants could see. The best-performing models from this second round were used to predict novel inhibitory molecules, of which several were synthesised and evaluated against the parasite. One such compound, containing a motif that the human chemists familiar with this series would have dismissed as ill-advised, was active. The project demonstrated the abilities of new machine learning methods in the prediction of active compounds where there is no biological target structure, frequently the central problem in phenotypic drug discovery. Since all data and participant interactions remain in the public domain, this research project “lives” and may be improved by others.
<table>
<thead>
<tr>
<th>File Name</th>
<th>Size (MiB)</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>NatComms Predict_V6.docx</td>
<td>2.57</td>
<td>view on ChemRxiv » download file</td>
</tr>
<tr>
<td>SI_V3.docx</td>
<td>6.12</td>
<td>view on ChemRxiv » download file</td>
</tr>
</tbody>
</table>
An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials

Edwin G. Tse1, Laksh Aithani2, Mark Anderson3, Jonathan Cardoso-Silva4, Giovanni Cincilla5, Gareth J. Conduit6,7, Mykola Galushka8, Davy Guan9, Irene Hallyburton3, Benedict W. J. Irwin7,10, Kiaran Kirk11, Julia C. R. Lindblom11, Raymond Lui11, Slade Matthews9, James McCulloch12, Alice Motion13, Ho Leung Ng14, Mario Öeren10, Murray N. Robertson15, Vito Spadavecchio16, Vasileios A. Tatsis2, Willem P. van Hoorn2, Alexander D. Wade’, Thomas M. Whitehead6, Paul Willis17 and Matthew H. Todd*1

Address: 1School of Pharmacy, University College London, London WC1N 1AX, United Kingdom; 2Exscientia Ltd., The Schrödinger Building, Oxford Science Park, Oxford OX4 4GE, United Kingdom; 3Drug Discovery Unit, Division of Biological Chemistry and Drug Discovery, School of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom; 4Department of Informatics, Faculty of Natural and Mathematical Sciences, King’s College London, London WC2B 4BG, United Kingdom; 5Molomics, Barcelona Science Park, Barcelona, Spain; 6Intelligens Ltd., Eagle Labs, Chesterton Road, Cambridge CB4 3AZ, United Kingdom; 7Theory of Condensed Matter Group, Cavendish Laboratories, University of Cambridge, United Kingdom; 8Auromind Ltd, 126 Eglantine Avenue, Belfast BT9 6EU, United Kingdom; 9School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia; 10Optibrium Ltd. Blenheim House, Denny End Road, Cambridge, United Kingdom; 11Research School of Biology, Australian National University, Canberra, ACT 2601, Australia; 12Kellerberrin, 6 Wharf Rd, Balmain, Sydney, NSW 2041, Australia; 13School of Chemistry, The University of Sydney, Sydney, NSW 2006, Australia; 14Department of Biochemistry and Molecular Biophysics, Kansas State University, Manhattan KS 66506, USA; 15Strathclyde Institute Of Pharmacy And Biomedical Sciences, University of Strathclyde, Glasgow G4 ORE, United Kingdom; 16Interlinked Therapeutics LLC, Portland, OR 97214, USA; 17Medicines for Malaria Venture, PO Box 1826, 20 rte de Pre-Bois, 1215 Geneva 15, Switzerland.

Email: Matthew H. Todd – matthew.todd@ucl.ac.uk
* Corresponding author

Abstract

The discovery of new antimalarial medicines with novel mechanisms of action is key to combating the problem of increasing resistance to our frontline treatments. The Open Source Malaria (OSM) consortium has been developing compounds (“Series 4”) that have potent activity against Plasmodium falciparum in vitro and in vivo and that have been suggested to act through the inhibition of PfATP4, an essential membrane ion pump that regulates the parasite’s intracellular Na’ concentration. The structure of PfATP4 is yet to be determined. In the absence of structural information about this target, a public competition was created to develop a model that would allow the prediction of anti-PfATP4 activity among Series 4 compounds, thereby reducing project costs associated with the unnecessary synthesis of inactive compounds.

In the first round, in 2016, six participants used the open data collated by OSM to develop moderately predictive models using diverse methods. Notably, all submitted models were available to all other participants in real time. Since then further bioactivity data have been acquired and machine learning methods have rapidly developed, so a second round of the competition was undertaken, in 2019, again with freely-donated models that other participants could see. The best-performing models from this second round were used to predict novel inhibitory molecules, of which several were synthesised and evaluated against the parasite. One such compound, containing a motif that the human chemists familiar with this series would have dismissed as ill-advised, was active. The project demonstrated the abilities of new machine learning methods in the prediction of active compounds where there is no biological target structure, frequently the central problem in phenotypic drug discovery.
Since all data and participant interactions remain in the public domain, this research project “lives” and may be improved by others.

**Keywords**
PfATP4; predictive modelling; Open Source Malaria; drug discovery; machine learning

**Introduction**

Efficiency in the early stages of the drug discovery pipeline, from hit identification to lead optimisation, is key to the development of new drugs. The initial identification of a hit compound is typically carried out using one of two approaches. In **target-based drug discovery** the molecular target of interest is known \[1\]. With this knowledge, libraries containing many compounds are screened (experimentally or computationally) against the known target to identify promising candidates or chemical scaffolds for further development. Through testing these chemicals, the key binding interactions may be identified and more directed structure activity relationship (SAR) studies can be conducted to optimise activity.

Alternatively, if the biological target is not known, **phenotypic drug discovery** may be undertaken \[2\]. This process involves the initial identification of potent compounds that give rise to the desired effect (e.g. inhibition of cell growth), with target determination performed thereafter. The lead optimisation phase in this type of drug discovery is less streamlined than that in the former method as it is conducted without guidance from target binding interactions and often relies upon the intuition of the medicinal chemist to design and synthesise compounds to explore the SAR. There are a number of obvious limitations to this approach, including the personal bias/imagination of the scientist or the availability/cost of resources. As a result, good hypotheses or key insights may be overlooked, which can lengthen the time taken to identify a lead candidate and increase costs associated with synthesising complex molecules that are later revealed to be inactive. Nevertheless, the advantage of phenotypic drug discovery, which underpins its popularity, is that hit or lead compounds are already known to be effective in their overall role (e.g., the killing of a pathogen).

To aid this latter approach and overcome the absence of knowledge of the target or its structure, computational models may be developed using artificial intelligence (AI) and machine learning (ML) \[3,4\]. Such approaches allow the activities of new compounds in a phenotypic-screening program to be predicted. For instance, matched molecular pair analysis \[5\] and quantitative structure activity relationship (QSAR) \[6\] models are commonly used in medicinal chemistry campaigns to determine the relationships between the physical and biological properties of a series of compounds. This information can then be used to guide the design of new active compounds. In those cases in which a target has been identified but its structure is not yet determined, a structural model may be developed based on a known close homolog of the target \[7\]. This method allows for docking studies to be conducted to examine potential binding interactions that may occur in the actual target, thus guiding the lead optimisation process more effectively. Recent years have seen the increased use of computational methods such as these to aid the drug discovery process \[8,9,10,11\]. For instance, there have been successes in the **in silico** target prediction of small molecules with activity against *Mycobacterium tuberculosis* \[12,13\].

In the case of the malaria parasite, the development of resistance to frontline treatments is an ever-present problem. Since the isolation of artemisinin from the plant *Artemisia annua* in 1971 by Tu Youyou and colleagues \[14\], this natural product and its derivatives have been used in some of the most effective treatments for malaria. The artemisinin-based combination therapies (ACTs) utilise a short-acting artemisinin derivative in combination with one or more complementary antimalarials that are long-acting and possess a different mechanism of action (MoA). The use of these combinations has, in part, been responsible for the slow development of resistance to ACTs, yet in recent years increasing numbers of
cases have emerged of reduced efficacy \cite{15}. There is an urgent need for new medicines that possess novel MoAs \cite{16}.

One promising biological target in *Plasmodium falciparum* is the essential P-type ATPase PfATP4, which localises to the plasma membrane of the intraerythrocytic parasite and exports Na\(^+\) while importing H\(^+\) equivalents \cite{17,18}. The structure of this membrane-bound protein remains unsolved. Evidence for the involvement if PfATP4 in the mechanism of action of antiplasmodial compounds comes from several sources, including parasite Na\(^+\) and pH assays that implicated PfATP4 as the target for the spiroindolone cipargamin \cite{17,19} (currently in Phase III clinical development), the dihydroisoquinoline (+)-SJ733 \cite{20}, and 28 compounds from the Medicines for Malaria Venture (MMV) Malaria Box \cite{21} as well as 11 compounds from the MMV Pathogen Box \cite{22}. These compounds represent a strikingly diverse range of chemotypes (Fig. 1) \cite{23}. A homology model of PfATP4 was developed using crystal structures from the closest mammalian homolog, a sarco/endoplasmic reticulum Ca\(^{2+}\)-ATPase (SERCA) \cite{20}. However, in the absence of a solved structure of PfATP4, ideally bound to small molecule inhibitors, it remains unclear how it is possible for such diverse molecules to share the same target. Indeed, a challenge to understanding such data is that structurally different molecules generating the same phenotype may be interacting with the biological target differently.

![Fig. 1 Examples of the diverse chemotypes that have been linked to PfATP4. Each of the compounds give rise to effects on the parasite's internal Na\(^+\) concentration and pH that are consistent with PfATP4 inhibition \cite{20,21,24}.](image-url)
Since 2011, contributors to Open Source Malaria (OSM) have been evaluating several series of compounds originating from high-throughput screens (HTS) performed by pharmaceutical companies [25]. The recent focus of OSM has been on a class of triazolopyrazine-based compounds (“Series 4”) that emerged from a screen carried out at Pfizer. There are currently more than 200 compounds in Series 4, with in vitro potencies against *P. falciparum* ranging from single-digit nanomolar to inactive. The highly promising nature of this series derives from several members having been found to be effective in the in vivo mouse model of the disease [26]. Based on preliminary investigations against *PfATP4*-resistant mutant strains (generated from the parent Dd2 strain by exposure to hits from the Malaria Box against *PfATP4* [21]), Series 4 compounds are thought to target *PfATP4* [27]. The intra-series similarity of their structures ought to imply a similarity in the way that the compounds interact with the target, but the interaction may differ from other compounds with the same phenotype.

The OSM Series 4 project is at the lead optimisation stage, with minor structural modifications being made in the search for improved solubility, potency and metabolic clearance. As is typical in such a search, analogs are being made that possess low potency, and these represent expensive “failures” (ca. $2K per compound for one postdoc-week per analogue). Better predictions of compound potency would save valuable resources and accelerate the science, so a predictive model was high on the list of priorities for the OSM consortium.

For the best means to develop such a model, we maintained an open mind. Available to us was a dataset of analogues with their associated activities, whether against the parasite or derived from biochemical (ion-regulation and/or ATPase) assays. Many of these compounds were from OSM Series 4, and there were also candidate antimalarials from other, structurally unrelated, series. It was possible to include “presumed inactives”: randomly-selected molecules from commercial catalogues that were unlikely to display activity. A homology model (vide supra) was available that might permit a more target-based approach. Acknowledging these varied resources, we opted not to prescribe the approach to be taken and instead, in 2014, approached the scientific community simply with the need for a model that would allow us to predict the activity of hypothetical compounds. All data from OSM research projects are freely available to anyone online, representing an ideal starting point for such an open competition.

Between then and now there has been an explosion of interest in machine learning and AI methods in drug discovery [28,29]. While these new methods had the potential to be game-changing, there is the ever-present challenge in this sector of hype, in the sense that the actual capabilities of some of the newer technologies, outside of marketing statements, are sometimes not clear. In OSM the openness extends to the research process itself, allowing contributors to share what they are doing, rather than what they have done. The use of competitions to progress scientific research is not novel in itself, with previous examples of this in data analysis for drug discovery [30], but it is uncommon for competitions to be accompanied by the next crucial step: benchmarking by chemical synthesis and biological evaluation of predicted molecules. It is rarer still for science competitions to run completely openly, where everyone can see, and potentially incorporate, other entrants’ solutions as they are submitted. We felt we could achieve two things by running this competition with OSM’s open source ethos, in which those submitting entries would reveal their predictions in real time and, ideally, provide full methods (within the boundaries of commercial sensitivities). We would be able to approach the scientific problem along multiple paths, but we would also be able to provide a clear case study of the current effectiveness of predictive modelling in phenotypic drug discovery.
Results and Discussion

Round 0
An initial attempt by a single OSM contributor to develop a pharmacophore model was based around the known PfATP4 active compounds from the MMV Malaria Box \cite{31,32}. By using Discovery Studio from Accelrys (now BIOVIA) to screen 28 active compounds with the Common Feature Pharmacophore Generation protocol, 10 four-feature models were produced. These were then narrowed down based on poses and score to one model that was developed further (Fig. 2A).

The 28 active compounds were mapped to the model and a shape feature was created (Fig. 2B). It was thought that this could give a general idea of the shape of the active site (Fig. 2C). Exclusion features were next added in areas where high scoring, inactive ligands penetrated outside of the shape figure. Unfortunately, when this model was applied in 2014 to a set of compounds that were evaluated for their ability to dysregulate ion homeostasis, the predictions were found to correlate poorly with the experimental potency results (Fig. 3). It was suggested that this lack of correlation could be due to factors not being taken into account by this first model (overlapping binding sites and compound chirality); a pharmacophore model explains aspects of the geometry of the interaction but not the details of the thermodynamics of the protein-small molecule contacts.

Fig. 2 Model creation workflow. A) The four-feature pharmacophore model chosen for further development with MMV006429 mapped. B) All 28 active compounds used in Round 0 superimposed onto the four-feature model. C) Shape feature added based on poses in B. D) Inactive molecules from the dataset mapped. E) Exclusion spheres added.
Fig. 3 Poor correlation was seen between the first model's predictions and experimental data. While there is excellent correlation between in vitro parasite killing potency and the ability to dysregulate parasite ion homeostasis ("PFATP4") activity, the majority of the model predictions did not correlate well with the experimental data. The compounds were tested for their effects on Na$^+$ concentration similar in magnitude to that of 50 nM cipargamin and a cytosolic alkalinisation, ‘No’: indicates that the compound did not affect the resting Na$^+$ concentration or pH, ‘Moderate’: indicates that the compound gave rise to an increase in Na$^+$ concentration similar to that observed on addition of 50 nM cipargamin.

This model was also used to screen[32] the Maybridge library of compounds[33] to identify a small and diverse selection of molecules to evaluate in biochemical assays. The results were filtered manually to give a final selection of 18 compounds that were subsequently evaluated for their effects on the parasite’s internal Na$^+$ concentration (at 1 μM) and pH (at 5 μM). None of the compounds were found to increase the parasite's Na$^+$ concentration or pH, which confirmed that the model required further optimisation and led to the start of a crowdsourced attempt to solve this challenge.

**Round 1**

The first full round of the predictive modelling competition was run between 2016 and 2017, and was intended to elicit the participation of members of the wider scientific community with expertise in computational chemistry[34]. The competition adhered to the open science principles underpinning the OSM consortium. Specifically, all participants were required to work openly for the duration of the competition, with working and data posted on open Electronic Laboratory Notebooks (ELN) that were made publicly available[35]. The participants were tasked with developing a predictive model using data provided by OSM that included a list of compounds with activity data for both in vitro whole cell potency and PFATP4 ion assays[36], along with the entire dataset of OSM compounds from previous series ((mostly presumed) inactives). Once the models were developed and deposited, the participants were provided with the molecular identifiers (e.g., SMILES strings) for the 400 compounds contained within the MMV Pathogen Box and were required to rank them in
order of predicted activity in the ion assays. The Pathogen Box compounds were at the same time screened for their effects on parasite Na⁺ concentration and pH and the data held back until the models had been submitted. A small cash prize inducement was employed to stimulate interest, despite the risk this brings of making the intrinsic reward for participation more extrinsic.\(^{[37]}\)

Six diverse, fully-fledged entries were submitted from individuals working in both public and private sectors, with all working shared online (Table 1)\(^{[38]}\). These submissions were reviewed by a panel of four judges (Prof. Matthew Todd, A/Prof. Alice Motion (University of Sydney), Dr. Murray Robertson (University of Strathclyde and creator of the previous model in Round 0) and Prof. Alexander Tropsha (University of North Carolina, Chapel Hill)) that evaluated the top twenty ranked compounds from each model against the undisclosed Pathogen Box data. Two entrants developed models that were able to predict correctly two active compounds within their top twenty rankings, with a further model a close third place\(^{[39]}\).

Table 1: Summary of the results from Round 1 of the predictive modelling competition.

<table>
<thead>
<tr>
<th>Entrant</th>
<th>Description of Model</th>
<th>Correctly Predicted Actives</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jonathan Cardoso-Silva</td>
<td>Gradient boosting model (using XGBoost) to predict actives and nonactives.</td>
<td>B just outside top 20</td>
<td>Runner-up</td>
</tr>
<tr>
<td>Giovanni Cincilla</td>
<td>(PfATP4) Ion Regulation Activity classification model using: CDK descriptors(^{[40]}), ECFC4 fingerprints and Random Forest.</td>
<td>(B, D)</td>
<td>Runner-up</td>
</tr>
<tr>
<td>Davy Guan</td>
<td>Semi-supervised machine learning, used to</td>
<td>(B, F)</td>
<td>Runner-up</td>
</tr>
</tbody>
</table>
construct QSAR models. Molecules were featurised by either Graph convolutional techniques or with 1024 Bit ECFP4 descriptors.

<table>
<thead>
<tr>
<th>Entrant (Affiliation)</th>
<th>Description of Modela</th>
<th>Precision of</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>James McCulloch</td>
<td>Deep Neural Network ML using a vector of the chemo-physical properties of the target molecules.</td>
<td>B, D, I F just outside top 20</td>
<td>Winner</td>
</tr>
<tr>
<td>Ho-Leung Ng</td>
<td>QSAR model based on homology modelling of PfATP4 -Cresset Forge.</td>
<td>K, D J just outside top 20</td>
<td>Winner</td>
</tr>
<tr>
<td>Vito Spadavecchio</td>
<td>Library of 'common' transformations' as seen in CHEMBL.</td>
<td>B</td>
<td>Runner-up</td>
</tr>
</tbody>
</table>

Compounds A-K shown to be active from the MMV Pathogen Box screen against PfATP4 [22].

While this first round of the competition was successful in demonstrating the capabilities of the community to work openly and provide quality data, the models, though obtained with diverse methods, were not yet highly predictive. Of note was, again, the striking diversity of chemotypes (A–K, Table 1) sharing a target.

**Round 2**

Given the diverse, spontaneous inputs from the initial round of the open competition, and the high quality of the associated dialogue that had taken place on the relevant project website, GitHub, it was decided that a second round would be run in 2019 since “expensive failure analogs” were still arising in the experimental programme. The aim for this round was not only to allow for the entrants from Round 1 to improve upon the original models, but for new participants to get involved with inputs from larger companies that specialised in artificial intelligence and machine learning (AI/ML) approaches. Since the series had moved on in the interim (with further compounds being evaluated), the community had access to an expanded dataset, including all the data used as the test set for the previous round [22].

The competition’s second round was launched in July 2019 [41]. In this new phase of the competition it was the intention to use the best-performing models to perform the most important task of all: to predict new chemical matter that would be active (rather than merely look at the fit of retrospective data). Synthesis and evaluation of these predictions would then serve as model validation in a “real” case. A small, new dataset of activity from recently-synthesised analogs was kept back to serve as the basis for judging model fitness.

By the conclusion of Round 2 (a period of ~10 weeks), ten entries had been submitted, five of which were from returning participants (Table 2). In a similar fashion, the submissions were reviewed by a panel of four judges (Prof. Matthew Todd, Dr. Edwin Tse (UCL), Dr. Murray Robertson (Strathclyde) and Prof. Robert Glen (Cambridge)) who compared the predicted potencies against the experimentally-derived blood stage potency values for thirty-four compounds. The precision of each model was calculated according to:

\[
\text{precision} = \frac{x}{x+y}, \quad \text{where} \quad x \text{ is the number of correct predictions (active and inactive combined) and} \quad y \text{ is the number of false positive predictions [42].}
\]

**Table 2: Summary of the results from Round 2 of the predictive modelling competition.**
<table>
<thead>
<tr>
<th>Name and Institution</th>
<th>Methodology</th>
<th>Accurate Predictions (Active and Inactive)(^b)</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jonathan Cardoso-Silva (King's College London)</td>
<td>Network-based piecewise linear regression for QSAR modelling (^{\cite{43}}).</td>
<td>36%</td>
<td>Runner-up</td>
</tr>
<tr>
<td>Giovanni Cincilla (Molomics)</td>
<td>(P. falciparum) inhibition classification model using: CDK descriptors (^{\cite{40}}), ECFC4 fingerprints and logistic regression (with: stochastic average gradient as solver, uniform regularisation and learning step size = 0.01).</td>
<td>91%(^c)</td>
<td>Winner (company)</td>
</tr>
<tr>
<td>Mykola Galushka (Auromind)</td>
<td>SMILES variational auto-encoder to generate chemical compounds fingerprint and cascade models Naive Bayes classifier with Multi-layer perceptron regressor for filtering active components and identifying a specific potency value.</td>
<td>58%</td>
<td>Runner-up</td>
</tr>
<tr>
<td>Davy Guan (The University of Sydney)</td>
<td>Automated machine learning method with 21 quantum mechanical descriptors using the Hartree Fock with 3 corrections method (^{\cite{44}}) and JClLogP, optimised for Mean Absolute Error.</td>
<td>82%</td>
<td>Winner (non-company)</td>
</tr>
<tr>
<td>Ben Irwin/Mario Öeren/Tom Whitehead (Optibrium/Intellegens)</td>
<td>Deep imputation (^{\cite{45,46,47}}) with quantum mechanical StarDrop6.6 Automodeller and pKa descriptors (^{\cite{48}}).</td>
<td>81%</td>
<td>Second place</td>
</tr>
<tr>
<td>Raymond Lui (The University of Sydney)</td>
<td>Automated machine learning method using 59 permutation feature importance selected Mordred and quantum mechanical descriptors optimised for Mean Absolute Error.</td>
<td>58%</td>
<td>Runner-up</td>
</tr>
<tr>
<td>Slade Matthews (The University of Sydney)</td>
<td>Random forest model using 200 Mordred descriptors based on optimised 3D structures. Training RMSE = 0.805.</td>
<td>N.A.</td>
<td>Runner-up</td>
</tr>
<tr>
<td>Ho-Leung Ng (Kansas)</td>
<td>QSAR model based on</td>
<td>71%</td>
<td>Runner-up</td>
</tr>
</tbody>
</table>
It was originally intended for each of the four winning entrants (first and second place winners) to generate two new structures that were predicted to be active using their models: one possessing the Series 4 triazolopyrazine core and the other being structurally distinct. This would give a total of eight molecules to be synthesised and validated experimentally. In addition to optimising potency, model generators were tasked with keeping good solubility in mind as a design criterion. It became evident that certain suggested compounds were synthetically inaccessible, or would take major resources to pursue. The former is often an issue when predictive models do not take into account known synthetic pathways, though there is significant activity at present to improve the incorporation of synthetic planning into library suggestion \[49,50\]. The initial list was narrowed to focus on five predicted triazolopyrazine compounds (Fig. 4). The five compounds were successfully synthesised and subsequently evaluated for \textit{in vitro} (growth-inhibition) activity against \textit{P. falciparum} along with the previously reported positive control for the series \[51\]. In addition to the standard potency (\textit{in vitro} growth) assay, these compounds were evaluated for their ability to inhibit \textit{PfATP4} in biochemical (\textit{Na⁺} regulation) assays to confirm that the MoA had not changed following these structural changes.
Fig. 4 Examples of the suggested compounds predicted by the winning entrants from Round 2 and the five chosen for experimental validation. The predictions were synthesised (see SI) and their potencies and MoAs (Fig. S9) experimentally validated. Three compounds were found to be active. *PfATP4 activity was not obtained for this compound.

Three of the six compounds were found to be active (<1 µM) or moderately active (1–2.5 µM) in in vitro growth assays with asexual blood-stage *P. falciparum* (3D7) parasites, representing a hit rate of 50% on a small sample size. Up to this point a total of 398 compounds had been made and evaluated for in vitro activity in OSM Series 4, with the design of these compounds driven entirely by the intuition of medicinal chemists. By setting a potency cut-off of 2.5 µM (the upper limit of reasonable activity), the tally of active compounds discovered in this series stands at 165, representing a comparable human intuition-derived hit rate of 41% on a larger sample size. Most of the compounds were tested (blind) for their ability to disrupt Na⁺ regulation in isolated asexual blood-stage parasites,
which confirmed an unchanged mechanism of action: two of the compounds found to be active in in vitro growth assays disrupted Na⁺ regulation whereas the three compounds inactive in growth assays did not, at the concentrations tested (Fig. S9).

It is interesting to compare these results with the intuition of the chemists who have deep experience of this series and who are familiar with the SAR. A recurring observation was the sensitivity of the length of the ether linker between triazolopyrazine core and northwest phenyl group, with a spacer of two methylene units (between phenyl ring and oxygen) leading to far higher potencies than other lengths. The Davy Guan prediction involving the shorter linker, and the Molomics 1 prediction without the pendant phenyl ring, lie in the class of inactive compounds subject to human retrospective wisdom (i.e. the “I could have told you that” class). In contrast, the Exscientia compounds were thought by the human team to be likely to be potent, but only one performed well (i.e. the “that’s odd” class). Lastly the Optibrium/Intellegens suggestion that included the tert-butyl pendant was thought by the human team to be a certain inactive, given what was known of variation in that part of the molecule, yet this compound displayed good potency and is a particularly useful outcome (i.e. the “I welcome our machine overlords” class).

To gain more insight, and to improve these potential antimalarials, further iterations of these models are needed. The open nature of the competitions and of the over-arching consortium is that anyone may work on improvements since everyone has access to all the data, making this a “living” research project. A potential explanation for the predicted hit rate not being higher is the relatively small dataset (~400 compounds) from which each model was developed, potentially compromising perfectly reasonably computational approaches yet representing a fairly typical situation for lead optimisation. Two further points are of particular note: 1) It was possible to involve leading experts from the private sector in an open competition to solve a public health challenge without those participants needing to compromise their competitive business advantage; indeed success in such an endeavour has been used as an open demonstration of capabilities [52]. 2) The private sector participants displayed high and sustained levels of collaborative working and commitment to a public good, in what is counter to the public's perception of the secretive nature of the modern pharmaceutical industry; indeed the “winning” and “losing” of the competition was less important than the extent to which entrants worked together openly to improve the underlying research [41].

Conclusion
With hit identification and lead optimisation being key steps in the development of any new drug, the continued advancements in machine learning and artificial intelligence approaches possess significant promise to streamline this process, which would result in more efficient medicinal chemistry campaigns. In the absence of target structural information, a crowdsourced approach was used to develop predictive models for a promising antimalarial series. Importantly, the winning models of the most recent competition round were used to generate novel compounds, which were then synthesised and evaluated for experimental validation of each model leading to a new counterintuitive “active”. The simple open science and crowdsourcing principles used throughout this campaign are applicable to many medicinal chemistry projects, whereby the community’s combined efforts can be used to accelerate the early stages of drug discovery and involve participants from public and private sectors. The work conducted here has been designed to be “living”, in that all methods and results are publicly available and contributions can continue to be made by anyone because everyone has access to all data and ideas.

References


[36] Ion Regulation Data for OSM Competition. Available at: https://docs.google.com/spreadsheets/d/1WWP8fE3X2BLzZ7jOo6bRWpnHZqVJBf8XgIYEb7hshXU/edit?usp=sharing. (Accessed: 26th September 2019).


Acknowledgements
This work has been supported by the Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery Network+, which has been funded by EPSRC (EP/S000356/1), the Australian Research Council (LP150101226), the Medicines for Malaria Venture (RD/11/0040) and University College London for funding (to MHT), the US National Science Foundation (1833181) (to HLN), the EPSRC Centre for Doctoral Training in Computational Methods for Materials Science (EP/L015552/2) (to ADW), the EPSRC Future Continuous Manufacturing and Advanced Crystallisation (CMAC) Research Hib (EP/L015552/1) (to MNR), the Royal Society (to GJC).

Author contributions
EGT, AM, MNR carried out the synthetic chemistry. AML, JCRL and KK planned and/or carried out the Na+ and pH assays and/or advised on the results. IH developed the Pf assay, manually prepared and analysed compounds, collated all data and contributed to the paper (methods section and data table). MA carried out routine Pf assay and analysis. MNR, DG, HLN, JC-S, BWJI, MÖ, TMW, GJC, ADW, LA, VAT, WPvH, JM, VS, RL, SM, GC and MG contributed models to the competitions and collaborated on the outputs during the competitions. PW helped conceive the study. MHT founded OSM, conceived the project and secured (AI3SD-FundingCall1_029) or co-secured (LP150101226, with MMV and KK) the funding. All authors contributed to the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at XXX.
Supporting Information

for

An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials


Round 2 predictive models, experimental procedures, characterisation data and copies of $^1$H and $^{13}$C NMR spectra for all synthesised compounds.
# Table of Contents

**GENERAL**.................................................................................................................................................. 3

**AUROMIND’S MODEL**........................................................................................................................................... 5

**DAVY GUAN’S MODEL**...................................................................................................................................... 7

**EXSCIENTIA’S MODEL**.................................................................................................................................... 10

**HO-LEUNG NG’S MODEL**.................................................................................................................................. 12

**JONATHAN CARDOSO-SILVA’S MODEL**.............................................................................................................. 14

**MOLOMICS’ MODEL**....................................................................................................................................... 17

**OPTIBRIUM/INTELLEGEN’S MODEL**.................................................................................................................. 20

**RAYMOND LUI’S MODEL**................................................................................................................................... 23

**SLADE MATTHEWS’S MODEL**............................................................................................................................ 26

**VITO SPADAVICHO’S MODEL**........................................................................................................................... 27

**FIG. S8 OVERALL SYNTHETIC ROUTE FOR THE PREPARATION OF TARGET COMPOUNDS FROM FIG. 4**........ 29

**CHEMISTRY**...................................................................................................................................................... 30

**\(^1\)H AND \(^{13}\)C NMR SPECTRA**.......................................................................................................................... 44

**LCMS TRACES FOR THE INTERMEDIATES OF MOLOMICS 1**.................................................................................. 70

**EXPERIMENTAL BIOLOGICAL PROCEDURES**.................................................................................................. 72

**FIG. S9 MECHANISM OF ACTION EXPERIMENTS**............................................................................................. 73

**REFERENCES**.................................................................................................................................................... 74
General

Reagents were purchased from either Sigma–Aldrich, Alfa Aesar, Acros, Merck, Fischer Scientific, Matrix Scientific, Ajax or Fluorochem. Unless otherwise specified, the reagents were used without further purification. Anhydrous conditions: glassware was dried at >130 °C for >12 h, assembled hot and allowed to cool under a high vacuum where appropriate or purged with inert gas. Anhydrous solvents were obtained from the PureSolv system or by drying over activated 3 Å molecular sieves. Nitrogen gas was dried over silica and calcium chloride. Argon gas was used as acquired. The phrase in vacuo corresponds to ~1 mbar on a Schlenk line. Reduced pressure means under rotary evaporation at 40 °C from 900–50 mbar. Flash chromatography was performed on DaviSil Grace Davison 40–63 μm (230–400 mesh) silica gel or on a Biotage Isolera One. Analytical thin layer chromatography was performed on Merck Silica Gel 60 F254 precoated aluminium plates (0.2 mm) and visualised with UV irradiation (254 nm) and potassium permanganate, anisaldehyde or ninhydrin staining. High temperature reactions were carried out in silicone oil baths, controlled by temperature probe in the oil bath.

Melting points (m.p.) were recorded on a Stuart SMP10 at 2 °C min⁻¹ (capillaries ø = 1.8–1.9 mm, 100 mm). Infrared spectroscopy was carried out on a Bruker Alpha-E (attenuated total reflectance) without atmospheric compensation and processed using OPUS 7.0 software. Samples were analysed neat. Nuclear magnetic resonance spectroscopy was carried out at 300 K on Bruker spectrometers: either AVANCE III 400 (¹H at 400 MHz, ¹³C at 101 MHz) or AVANCE III 500 (¹H at 500 MHz, ¹³C at 126 MHz). Spectra were processed using Mestrelab Research Mnova. Deuterated solvents (CDCl₃, DMSO-d₆, CD₃OD) obtained from the Cambridge Isotope Laboratories. ¹H and ¹³C chemical shifts are reported in parts per million (ppm) with respect to TMS at 0.00 ppm. The chemical shifts of the spectra were calibrated to residual solvent peaks (¹H: CHCl₃ 7.26 ppm, DMSO 2.50 ppm, MeOH 3.31 ppm, TMS 0.00 ppm; ¹³C: CHCl₃ 77.16 ppm, DMSO 39.52 ppm, MeOH 49.00 ppm, TMS 0.00 ppm). ¹H signal multiplicity is reported as: singlet (s), doublet (d), triplet (t), quartet (q), pentet (p) and combinations thereof, or multiplet (m). Broad signals are designated broad (br). Coupling
constants \( J \) are reported in Hertz (Hz). Integrals are relative. \( \text{app} = \text{apparent} \) when the multiplicity was unexpected, e.g. coincidental or unresolved. Low resolution mass spectrometry (\( m/z \)) was carried out on a Finnigan quadrupole ion trap mass spectrometer using electrospray ionisation (ESI). High resolution mass spectrometry (HRMS) was performed on a Bruker 7T FT-ICR using ESI or APCI. Positive and negative detection is indicated by the charge of the ion, e.g. \([\text{M+H}^+\]) indicates positive ion detection. Analytical liquid chromatography-mass spectrometry (LCMS) was performed on an Agilent Infinity 1290 II system consisting of a quaternary pump (G7111A) and a diode array detector WR (G7115A) coupled to a InfinityLab LC/MSD (G6125B) using ESI. An Agilent Poroshell 120 EC-C18 column (2.7 µm, 3.0 × 50 mm) was eluted at a flow rate of 1.5 mL/min with a mobile phase of 0.05% formic acid in \( \text{H}_2\text{O} \) and 0.05% formic acid in MeCN.
Auromind’s Model

The model developed by Auromind for predicting "Potency vs Parasite (µMol)" for the Series 4 compounds consists of three modules integrated into a single pipeline (Fig. S1).

Fig. S1 The architecture of Auromind model for predicting "Potency vs Parasite (µMol)" for the Series 4 compounds.

The first module is represented by an encoder isolated out of the pre-trained variational auto-encoder [1] (VAE). It transforms an input SMILES into a chemical compound fingerprint consisting of a 1024 real values vector. Often this vector is called “latent”-vector and a space formed by these vectors a “latent”-space accordingly. VAE positions similar compounds nearby to each other in the “latent” space. It helps to design machine learning algorithms, which utilise similarity in the “latent” space to predict various chemical properties. The VAE [2] used in this study has been trained on the 1.7m compounds selected from the ChEMBLv24 database [3]. The accuracy (of SMILES reconstruction) obtained as a result of 10 cross-validations was equal to 0.837±0.006. In addition to accuracy, the VAE performance was measured by Hamming [4] (0.682±0.051) and Levenshtein [5] (0.677±0.114) distances. Both
belong to a family of editing distances and gives a different perspective on the assessment of
the VAE performance.

The second module in the developed pipeline is a Gaussian Naive Bayes \(^6\) (GNB) classifier. It performs a selection of active compounds (which potency value less or equals 2.5 thresholds). To achieve better accuracy the initial features space was reduced to 100 using the Chi-squared test. GNB has been selected as the best performing model (with ROC-AUC equals to 0.741±0.024) using a grid-search among other classification models.

The third module in the developed pipeline is a Multi-layer Perceptron \(^7\) (MLP) regressor, which receives chemical compounds already classified as “active” and predicts an actual potency value in a range between 0 and 2.5. The approach of filtering compounds helped to deal with experimental samples where IC\(_{50}\) values were greater than the max concentration that was tested in the assay. MPL was selected as the best using a grid-search performed among other regression models. To achieve better performance were selected 40 features using a univariate linear regression test. It was observed that all of the tested models were suffering from overfitting and demonstrated relatively poor performance. MPL showed the best result R\(^2\) (coefficient of determination) equals 0.141±0.017 among others with the regularization parameter equals 1. Such poor performance caused by an extremely limited number of training samples (205), preventing adequately establishing a mechanism of predicting a real value of potency. These results can be potentially improved with an increasing number of training samples.
Davy Guan’s Model

The aim of phenotypic drug discovery in this competition to characterise the effects of chemical agents on *in vivo* assays instead of a specific biological target resonates with the effort in toxicology to improve predictive modelling of animal assays to reduce, refine or replace their use [8]. This meant we could utilise our experience and modelling methodology developments from that domain to address the same challenges arising from limited data availability and the nature of predicting *in vivo* outcomes in this project [9].

The provided 440 chemical OSM S4 dataset was deduplicated by averaging the potency values for each replicated ligand sharing the same OSM code to yield a 340 chemical training set in SMILES string format (Fig. S2). This training set was further curated with the removal of salts and solvents, charge neutralisation, and the addition of explicit hydrogens to output in 2D SDF using ChemAxon Standardizer 18.22.0, 2019, ChemAxon (https://www.chemaxon.com). These 2D structures were input into a KNIME workflow [10] consisting of the Open Babel [43] and RDKit [33] nodes to prepare initial 3D structures optimised using the Universal Force Field. The MaPhi descriptor program (manuscript in preparation, DOI: http://doi.org/10.5281/zenodo.1407646) acted as an interface for all subsequent quantum mechanical descriptor calculations. Geometry optimisation was conducted with the semiempirical PM7 method in gas phase implemented in MOPAC2016 [11] followed by a final geometry optimisation stage at the Hartree Fock with 3 corrections (Hf-3c) [12] level of theory with implicit aqueous solvation using the CPCM implementation in Orca 4.2.0 [13,14,15]. 21 quantum mechanical descriptors derived from the electronic properties were calculated at the Hf-3c level of theory in implicit aqueous solvation with 337 chemicals achieving convergence (99.1% yield). Water-octanol partition coefficients were predicted
using the JCLoGP model in JChem for Excel 19.7.26, 2019, ChemAxon (https://www.chemaxon.com), which was selected as an additional bioavailability descriptor after comparison to other partition coefficient models[16]. A 2.5 uM threshold was applied to the potency values in this dataset to produce a 22 descriptor training set for regression modelling.

The TPOT Python library[17] was used to automate model pipeline construction from dataset scaling and transformation to machine learning algorithm selection and hyperparameter optimisation. This library was initially configured to optimise 1000 models for the lowest Mean Absolute Error (MAE) from 10-fold cross validation over 50 generations, however, overfitting was detected for predicting the test dataset. The default setting of ten generations was selected instead with no overfitting detected. Additional dataset variants were also prepared with selected Mordred descriptors[18] in collaboration with Raymond Lui with no apparent performance improvement. The final OSM Round 2 predictive model found 0.548 uM MAE in 10-fold cross validation following TPOT pipeline optimisation with 0.82 precision for the unlabelled test set. This regression model achieved the best predictive performance in the academic category which led into the second phase of the Open Source Malaria Round 2 Predictive Challenge.

The aim of the second phase of the Open Source Malaria Round 2 Predictive Challenge was to generate two molecules, one of which with a triazolopyrazine scaffold and one with a novel, dissimilar scaffold. A 290,000 chemical generative modelling dataset was prepared with 250,000 ZINC molecules, 340 OSM S4 compounds, and 40,000 compounds screened
against *P. falciparum* in the ChEMBL database. This dataset only contained SMILES structures. The ChemGE program \[^{[19]}\] was selected to expediently generate novel compounds based on the representative triazolopyrazine scaffold. This was achieved by incorporating an additional calculated triazolopyrazine similarity index comparing generated molecules to the reference triazolopyrazine scaffold in addition to the default configuration calculating synthetic accessibility, lipophilicity, and an aromatic ring penalty for scoring all generated molecules. A 22,000 chemical dataset with known *P. falciparum* potency values was composed from the ChEMBL database to prepare an additional model based on Mordred descriptors in case the solvated quantum mechanical descriptors could not be calculated in time for the deadline. The dataset curation methodology and KNIME workflow to generate 3D structures from the first phase was applied followed with the calculation of 1825 descriptors in Mordred 1.1.1. This dataset was used for generating a regression model using TPOT to optimise 100 models over 10 generations. Both models were used to generate consensus predictions by averaging predicted potencies for generated molecules with subsequent ranking based on JCLogS solubility values calculated using JChem for Excel 19.7.26, 2019, ChemAxon (https://www.chemaxon.com).
Exscientia’s Model

Molecular features were generated with ECFP4 fingerprints \cite{20}. Due to the small sample size of the training dataset and the curse of dimensionality \cite{21} of chemical space (we might be working with 2000+ features), some featurisation preprocessing and model hyperparameter tuning was required to obtain a generalisable model.

Firstly, common bit fragments (fragments that appeared in more than 15 of the 34 test set molecules) were removed from the featurization. Since the test set lies in highly clustered chemical space defined by these common features, removing them prevents the model from overfitting to noise in these features.

Secondly, models were found to only be generalisable when subjected to heavy regularisation. In practice, this meant large regularisation constants in regressions, small tree depth and tree pruning in random forests, and early stopping in boosting models. These results are presented in the Table S1.

\textbf{Table S1: Coefficient of determination (R^2) scores for various models}

<table>
<thead>
<tr>
<th>Cross-validation split</th>
<th>Random Forest</th>
<th>AdaBoost</th>
<th>Ridge Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.381</td>
<td>0.276</td>
<td>0.332</td>
</tr>
<tr>
<td>1</td>
<td>0.368</td>
<td>0.246</td>
<td>0.404</td>
</tr>
<tr>
<td>2</td>
<td>0.303</td>
<td>0.273</td>
<td>0.190</td>
</tr>
<tr>
<td>3</td>
<td>0.368</td>
<td>0.368</td>
<td>0.353</td>
</tr>
<tr>
<td>4</td>
<td>0.167</td>
<td>0.189</td>
<td>0.292</td>
</tr>
</tbody>
</table>
The final model chosen was a ridge regression with the regularisation parameter set to unity.
Ho-Leung Ng’s Model

We sought to build on our relatively successful predictions made in the first round of the competition which used a structure-based approach of modelling inhibitors binding to the presumptive target, *Plasmodium falciparum* ATP4 \(^{[22]}\), an ATP-driven sodium transporter. Our primary goal was to create a structural model that would be easily interpretable and actionable by chemists. Towards that goal, we refined our ATP4 homology and predicted a consensus ligand binding mode. With these models, we calculated 1D, 2D, and 3D features and predicted the activity of the untested molecules using gradient boosted ensembles \(^{[23]}\).

To improve the prediction of the drug binding site and poses, we docked the ten most potent Series 4 compounds into our 2016 homology model of ATP4 using smina \(^{[24]}\) with the Vinardo scoring function \(^{[25]}\). Docking was performed with residues within 4 Å of the ligand allowed to sample different rotamers and the smina completeness parameter set to 16. The top three poses for each docked molecule were kept for further analysis. When the top poses were overlaid, we observed one common pose occurred far more than any other. This pose was identified for eight of the ten ligands. It corresponds approximately a 180° rotation from the pose we predicted in 2016. The ligand pose we identified in 2016 was only observed for 2-4 ligands and was the second most frequently occurring pose. The pose for OSM-S-532 (Fig. S3) was chosen as the template for redocking and scoring the Series 4 molecules.

![Docked pose of OSM-S-532 into homology model of ATP4.](image)

**Fig. S3** Docked pose of OSM-S-532 into homology model of ATP4.
We improved the model of ATP4 bound to OSM-S-532 with multiple rounds of simulated annealing energy minimization in Yasara [26]. Models were energy minimized first with the Nova2 force field, with the docked ligand fixed and without solvent to relax side chains at the binding site. This was followed by two rounds of energy minimization using the Yamber3 force field [27] with explicit solvation and the ligand free to move.

The energy minimized model of OSM-S-532 bound to ATP4 was used as a reference for redocking the Series 4 molecules using POSIT (OpenEye Scientific Software). POSIT docks molecules favoring poses similar to a reference pose. SAR models were built using pIC_{50} and pEC_{50} values. A serious challenge was handling activity data for weakly active molecules, usually notated as IC_{50} >10 mM. The POSIT docking classification score, “Great”, “Good”, “Mediocre”, and “None” was used to assign IC_{50} values to weakly active molecules of 30, 90, 270, and 1000 mM respectively. This is a rather arbitrary assignment but was considered to be better than the initial IC_{50} assignment of >10 mM. Quantitative 3D docking scores from idock [28] (based on Autodock Vina [29]) and RF-Score v3 [30] were added as features for SAR scoring. Over 1800 additional conventional 1D, 2D, and 3D SAR features were calculated with Mordred [18].

To perform regression, we chose to use gradient boosted tree ensembles rather than neural networks. In our experience, neural networks have a higher tendency to overfit when training data is limited and can require extensive trial and error optimization for best performance. Gradient boosted ensemble regression between molecular features and measured activities was performed with XGBoost [31]. The 10-fold cross validation R^2 was 0.33 averaged over ten independent runs.
Jonathan Cardoso-Silva’s Model

A recent two-stage algorithm that combines network analysis and piecewise linear regression modelling, called modSAR \[32\], was used to generate modular QSAR models for OSM Series4 data. Fingerprint features are used to create a network representation of the compounds and to split compounds into disjoint modules. Each module is then modelled by a piecewise linear regression model using molecular descriptors.

The original dataset provided containing 440 compounds was deduplicated by their OSM-ID, the median potency value was taken for each duplicated compound. A threshold of 2.5 µM was applied to potency values. Circular topological fingerprints (ECPF4 1024 bits) were generated in RDKit \[33\] and used to create a network representation of the data while approximately 200 descriptors from Chemistry Development Kit (CDK) version 2.0 \[34\] were used in the regression models. Highly correlated CDK descriptors as well as those with near zero variance were removed, resulting in 95 molecular descriptors which were then centered and put on a scale from 0 to 1.

The network of OSM Series 4 compounds can be seen in Fig. S4 below. Similarity between compounds was calculated using Tanimoto coefficient (Tc) on their ECPF4 fingerprints, a link is drawn between a pair of molecules if their similarity is above the threshold Tc >= 0.20. Four modules were identified and are pictured in the figure along with their most representative compound (the compound with the largest number of neighbours in their community).
Fig. S4 Network of OSM Series 4 compounds showing linked similarities calculated using Tanimoto coefficient on their ECPF4 fingerprints.

A piecewise linear regression model was trained for each module, further splitting the molecules into subgroups which are modelled independently by linear equations \[^{[35]}\]. The models were trained in 10-fold Cross validation schema as an example, these were the breakpoints and equations found for compounds in module m01:

\[
pIC_{50} = \begin{cases} 
0.15 \text{khs.d} s\text{CH} + 0.10 \text{khs.s} \text{OH} + 5.60 & \text{if khs.dssC} \leq 0.74 \\
0.56 \text{ATS} m1 - 0.29 \text{BCUT} w-11 \\
+ 0.15 \text{BCUT} w-11 + 0.43 \text{C} 2\text{SP3} - 2.16 \text{MDEO}-12 & \\
+ 1.80 \text{khs.s} \text{CH} 3 - 2.07 \text{khs.s} \text{F} + 3.77 \text{khs.s} \text{NH} \\
- 0.76 \text{khs.tsC} - 0.37 \text{nAtomP} - 2.48 \text{nHBDon} + 4.91 & \text{if } 0.74 < \text{khs.dssC} \leq 0.77 \\
2.37 \text{khs.s} \text{F} - 0.92 \text{khs.s} \text{NH} + 7.03 & \text{if khs.dssC} > 0.77 
\end{cases}
\]

Despite the descriptive capabilities of the method, this model didn’t achieve a high precision on the hidden test set. A possible cause is the choice of descriptors, a different set of
molecular descriptors, fingerprints or maybe pseudo-features generated by techniques such as autoencoder could probably yield better results. We could also try to use a different regression technique on the modules such as Random Forest.
Molomics’ Model

As the objective of the requested model was to predict activity of Series 4 compounds, only molecules of this series were used for model development. The data was curated according to Molomics standard protocol. A modelling set and an external set were generated for both regression and classification modelling.

General 2D descriptors and ECFC4 structural fingerprints were calculated. Constant or correlated descriptors were removed. Z-score normalization was applied to all the descriptors.

Many regression and classification models were developed (Fig. S5) using different Machine Learning methods (i.e. Random Forest, Gradient Boosted Tree, Logistic Regression, Multi-layer Perceptron Neural Network, Naïve Bayesian). The most important parameters of each of these models were tuned using response-stratified 10-fold Cross Validation. According to QSAR Best Practices [36], only models fulfilling strict acceptability criteria were considered to be eligible. No regression model resulted eligible. 3 of the most promising classification models were selected and submitted to a statistical significance analysis using bootstrapping (n=100) and Y-randomization (n=100) sampling. The final model was selected on the basis of the statistical significance analysis together with internal & external validation performance, resulting in a logistic regression model using a stochastic average gradient as solver [37], a uniform regularisation and a learning step size = 0.01.
Fig. S5 Performance metrics of Molomics’ classification models. GBTC: Gradient Boosted Trees; LRC: Logistic Regression; MLPNNC: Multi-Layer Perceptron Neural Network; NBC: Naïve Bayesian; RFC: Random Forest.

Nine possible Applicability Domain metrics were evaluated for the selected model according to a published unified approach.[38]

The best of them resulted to be a structure-based minimum distance towards the modelling set. Using this Applicability Domain allows to provide a confidence for each prediction. Compound suggestions were generated using a Collective Intelligence approach through Molomics Technology (not published). This allowed to design molecules with optimal in silico Pfal activity, solubility and Caco-2 cell permeability (Fig. S6). The 2 most desirable compounds (M295, M4250), selected through a consensus decision-making process, were discarded by OSM due to their difficulty in synthesis. Alternatively, compounds M2726 and M4625 were proposed for synthesis and experimental testing.
According to OSM rules of transparency and collaboration, the model was finally deployed through Molomics Technology and made publicly available (molomics.com/osm) and exploitable for the design of additional Series 4 candidates through Collective Intelligence.

**Fig. S6** Molecules generated by Molomics technology through Collective Intelligence.
Optibrium/IntelliGen’s Model

Machine learning was performed by the Alchemite tool (developed by Intellegens and available through Optibrium) [39,40,41] which performs multiple deep imputation on the entire dataset using an iterated deep neural network. Additional pKa descriptors (most basic site pKa, most acidic site pKa) were calculated using the StarDrop 6.6 pKa model [42] (Predicting pKa Using a Combination of Quantum Mechanical and Machine Learning Methods), this uses quantum mechanical descriptors. The master chemical dataset was kept partitioned as individual assays from different labs because mixing results from multiple assays is not desirable due to multiple definitions of inactivity cutoff and scales of resolution. A single Alchemite model was trained comprising each assay type over different isoforms from different labs as well as single shot inhibition and ion regulation activity columns. No hyperparameter optimisation was performed as the dataset was quite small, default general parameters were used with 3 iteration layers, iterative mixing fraction of 0.95, 50 input columns (from both descriptors or other endpoints) per output column. 330 StarDrop AutoModeller descriptors were generated for each compound.

Models were built using pIC\textsubscript{50} and pEC\textsubscript{50} values which were log-transformed from the original data. The coefficient of determination \( R^2 \) values for the training and test sets are shown in Table S2. Some of the assays apparently had very consistent data and both training and test models showed excellent performance, e.g. \textit{Pfal} (K1) IC\textsubscript{50} (Avery), \textit{Pfal} IC\textsubscript{50} (Syngene), \textit{Pfal} IC\textsubscript{50} (GSK) with test \( R^2 \) values as high as 0.95. Unfortunately the model for \textit{Pfal} IC\textsubscript{50} (Dundee) was less good, which was the final assay used to evaluate the test compounds. The data suggest there could be more inconsistency in the Dundee assays that other labs results, or rather that the GSK, Syngene labs were high precision results, or that the activity of compounds tested in those labs are more easily modelled.

Alchemite specialises in estimating error bars and produced these for both training and test results. This meant it was possible to discard predictions which were known to be inaccurate.
This was essential at the compound generation stage, where only confident predictions about compounds were trusted. We believe this is the reason why the Alchemite predicted pIC\textsubscript{50} value of 6.4 for the tert-butyl Optibrium suggested compound matched the experimental value of 6.2 so closely and led to a successfully active compound where other methods which could not consider uncertainty struggled (a small known variance versus a potentially large unknown variance), however further data would be needed to be confident in this hypothesis.

Table S2: Coefficient of determination score for independent assay columns which had enough test data to create a meaningful value (>4 points). Dundee was the assay conditions used for the final measurements.

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Train R\textsuperscript{2}</th>
<th>Test R\textsuperscript{2}</th>
<th>Endpoint</th>
<th>Train R\textsuperscript{2}</th>
<th>Test R\textsuperscript{2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC\textsubscript{50} Chembl (uM)</td>
<td>0.72</td>
<td>0.16</td>
<td>Pfal IC\textsubscript{50} (Dundee)</td>
<td>0.59</td>
<td>0.32</td>
</tr>
<tr>
<td>Ion Regulation Activity</td>
<td>0.76</td>
<td>0.25</td>
<td>Pfal IC\textsubscript{50} (GSK)</td>
<td>0.95</td>
<td>0.8</td>
</tr>
<tr>
<td>Pfal EC\textsubscript{50} (Inh)</td>
<td>0.73</td>
<td>0.21</td>
<td>Pfal IC\textsubscript{50} (Ralph)</td>
<td>0.79</td>
<td>0.17</td>
</tr>
<tr>
<td>Pfal (K1) IC\textsubscript{50} (Avery)</td>
<td>0.95</td>
<td>0.93</td>
<td>Pfal IC\textsubscript{50} (Syngene)</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>Pfal IC\textsubscript{50} (Avery)</td>
<td>0.86</td>
<td>0.70</td>
<td>Single Shot Inhibition %</td>
<td>0.79</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Compound suggestions were generated in two ways:

1. A “bottom up” method using med chem expansions around promising compounds in the original series to generate around 90,000 ideas. This expansion was performed using the Nova module in StarDrop. All solutions were passed through the Alchemite model and the best were assessed using a multi-parameter score function from StarDrop which attempted to maximise activity in all assays and confidence in those assays, as well as increase solubility.

2. A “top-down” approach, in which a recurrent neural network (RNN) was trained on most of ChEMBL to generate SMILES from a vector input of descriptors. This vector
comes from inverse solving the Alchemite model for vector which is predicted by the model to be confidently active. This vector was then turned into SMILES by the RNN, and the outputs were screened for synthesis and reactivity problems by hand.

The final compounds were suggested (Fig. S7), two from each generation method, with compound b) being selected for synthesis due to a) and c) having potential HF and OCO reactivity problems respectively.

![Chem Transforms](image1)

![RNN + Inverse Vector](image2)

**Fig. S7** Final compounds suggested from the two generation methods.
Raymond Lui’s Model

A machine learning regression model approach was used to generate quantitative structure-activity relationship (QSAR) correlations between physicochemical/electronic descriptors and PfATP4 inhibitory potency.

Dataset curation. A training set of 340 unique molecules with known PfATP4 inhibitory potencies was curated from the provided OSM data; duplicates containing the same OSM ID were consolidated by taking their average potency, and any final values greater than 2.5 µM were set to 2.5 µM to define an active target range for the regression model. Molecules were expressed in SMILES and further curated in ChemAxon Standardizer v18.22.0 (ChemAxon Ltd., 2019; https://www.chemaxon.com) by removing salts and solvents, neutralising any charged fragments, and adding explicit hydrogens.

Molecular descriptor calculation. Three-dimensional structural geometries were initially constructed in OpenBabel [43] and RDKit [33] using the Universal Force Field (UFF), then further optimised in MOPAC2016 [11] using the PM7 method in the gas phase. 1,572 two- and three-dimensional physicochemical descriptors were then calculated in Mordred v1.1.1 [18] for Python v3.6. Another 21 quantum mechanically-derived three-dimensional electronic descriptors were calculated in Orca v4.2.0 [14,15] after further structural optimisation using the Hartree Fock with 3 corrections (HF-3c) method [12] solvated in water with the Conductor-like Polarizable Continuum Model (CPCM) [13]. Structural curation, optimisation, and descriptor calculation was performed in collaboration with Davy Guan.

Baseline QSAR model optimisation. A genetic search algorithm was used to develop two baseline models correlating the provided PfATP4 inhibitory potencies with each of the physicochemical and electronic descriptor sets. A population of modelling pipelines each consisting of components including data preprocessors, machine learning algorithms, and their hyperparameters is firstly initiated. The genetic search algorithm then iteratively evolves
these pipelines over multiple generations by randomly mutating the values and combinations of the aforementioned components to minimise the 10-fold cross-validation (10-fCV) mean absolute error (MAE), eventually returning the lowest MAE pipeline model. The Tree-based Pipeline Optimisation Tool (TPOT) v0.10.2 \cite{17} for Python v3.7 was used to perform searches up to 50 generations with population sizes up to 1,000 pipelines.

**Descriptor ranking analysis.** A permutation feature importance approach was used to quantify the importance of each physicochemical and electronic descriptor to its respective QSAR model when predicting PfATP4 inhibitory potency. The importance weight of each descriptor was calculated as the MAE gained (i.e. performance lost) by the model when it predicts with that descriptor column converted to noise via random permutations. Feature importance analyses were performed using eli5 v0.10.1 (https://www.github.com/TeamHG-Memex/eli5; accessed October 2019) for Python v3.7, with importance weights computed as the average ΔMAE over 50 permutations for each descriptor.

**Serial descriptor selection.** The dimensionality of the 1,572 physicochemical descriptor training set was reduced in order to obtain a smaller and optimal descriptor-to-molecule ratio and improve model generalisability. A serial descriptor selection protocol was developed by iteratively performing the aforementioned genetic search algorithm and permutation feature importance sequence, with each iteration taking the top 50% ranked descriptors from the previous iteration to form a new QSAR model - similar to a serial dilution process. Five iterations were performed resulting in five smaller physicochemical models with 786, 393, 197, 99, and 50 descriptors.

**Final QSAR model ensemble.** The 50 physicochemical descriptor subset was chosen as it returned the lowest 10-fCV MAE of 0.364 μM. In order to diversify the molecular representations available for QSAR generation, the top 9 electronic descriptors from the original descriptor ranking analysis were also included. Seven final models, three using the
50 physicochemical descriptors and four using the 59 physicochemical and electronic descriptors, were each optimised using TPOT for up to 50 generations with population sizes up to 500 pipelines. The PfATP4 inhibitory potency predictions of the seven pipelines (10-fCV MAE ranging from 0.334-0.385 µM) were ensembled by computing the average.

All descriptor data, modelling code, pipeline details, and experimental outputs from this submission are available at https://github.com/luiraym/PfATP4-Potency-Predictor.
**Slade Matthews’s Model**

A series representative model types was selected from the different classes of classifier from the Weka stable to cover a wide variety of model types in the search for the best performer. Simple linear regression from the Functions category, Additive regression from the Meta classifier category, M5Rules from the Rules category, and Random Forest from the Trees category were chosen for performance evaluation using 10-fold cross validation. Target values were thresholded to 2.5 µM. Weka preprocessing consisted of removing useless and string type attributes followed by attribute selection using the Pearson correlation to the class in the CorrelationAttributeEval filter selecting for 200 attributes followed by a random sort. The model with the lowest root mean square error (RMSE) and the highest Person product moment ($r$) in 10-fold CV on the training data was a Random Forest model (Table S3). This model was saved for evaluation on the test set. Test set predictions on the Random Forest model were submitted to the OSM competition for assessment against held out potency data. All of the predicted potency values were very high, range (1.696 – 2.016), and subsequently the model was not able to predict potent ligands from within the test set. This may be due to the use of a training set with many values set to a ceiling of 2.5 as the maximum relevant value to the modelling problem accounting for 66.4% of the training data. This skewed the model toward predicting higher values. These results show that more complex modelling strategies are required to build a model capable of predicting the binding potency for the PfATP4 ion pump target.

**Table S3** Training performance for Weka models built on 200 Mordred descriptors calculated from 3D structures optimized at the semi-empirical PM7 level of theory with implicit water solvation (bold indicates lowest RMSE in training).

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Mean Absolute Error</th>
<th>RMSE</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear Regression</td>
<td>0.7289</td>
<td>0.863</td>
<td>0.3576</td>
</tr>
<tr>
<td>Method</td>
<td>AUC</td>
<td>F1</td>
<td>Precision</td>
</tr>
<tr>
<td>---------------</td>
<td>------</td>
<td>------</td>
<td>-----------</td>
</tr>
<tr>
<td>Additive Regression</td>
<td>0.6943</td>
<td>0.8981</td>
<td>0.3934</td>
</tr>
<tr>
<td>M5Rules</td>
<td>0.7059</td>
<td>0.9458</td>
<td>0.304</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.6659</td>
<td>0.8057</td>
<td>0.4958</td>
</tr>
</tbody>
</table>

**Vito Spadavecchio’s Model**

An ensemble machine learning model was created to predict the classification (active or inactive) of compounds in an *in vitro* assay for the presumptive target *Plasmodium*.
falciparum ATP4. Compounds were represented as SMILES patterns and transformed to 2048-bit fingerprints (ECPF4) using RDKit [33]. The ensemble model was constructed from a classification and regression model. The classification model utilised logistic regression (l2 penalty, lbfgs solver) and used the scikit-learn library [33]. The feed-forward neural network regression model (ReLU activation, 4 layers) written with the Keras API (https://keras.io/) and was trained using the Adam optimiser and mean squared error loss function. For both models, compounds in the training set had input IC\textsubscript{50} values transformed to Log(IC\textsubscript{50}). All compounds with \textit{in vitro} activity <1 μM were labelled as ‘active’ and all other compounds labelled as ‘inactive’. A total of 440 compounds were used in the course of training of both models. The classification model was tuned to minimise the Matthews correlation coefficient (MCC), while the regression model minimised the mean unsigned error (MUE) between prediction and test set. Both models were refined using a 85/15 train/test split across a 10-fold cross validation. The mean MCC for the classification model 0.45 +/- 0.07, and the mean MUE for the regression model was 0.463 +/- 0.043 (95% CI for both errors). For the unlabelled test set, all compounds which were both predicted as ‘active’ by the classifier and had a linear regression prediction less than 0.35 μM were labelled as being ‘active’ in the test set. Both models were written using Python 3.6.
Fig. S8 Overall Synthetic Route for the Preparation of Target Compounds from Fig. 4
Compounds 2-Chloro-6-hydrazinylpyrazine (1), (E)-2-chloro-6-(2-(4-(difluoromethoxy)benzylidene)hydrazinyl)pyrazine (5), 5-chloro-3-(4-(difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-a]pyrazine (9) and 3-(4-(difluoromethoxy)phenyl)-5-phenethoxy-[1,2,4]triazolo[4,3-a]pyrazine (\textbf{+ve Control} compound in Fig. 4) were previously synthesised according to literature procedures \cite{44}.

**General Procedure 1: Condensation of hydrazinylpyrazine with aldehyde.** Compound 1 (1 equiv.) was dissolved in EtOH (112 mM). Aldehyde (1 equiv.) was added and the reaction stirred at rt overnight. The suspension was filtered and washed with cold EtOH to give the corresponding hydrazone that was used without further purification.

**General Procedure 2: Cyclisation of hydrazone to triazolopyrazine core.** The product from General Procedure 1 (1 equiv.) was dissolved in CH$_2$Cl$_2$ (112 mM). PhI(OAc)$_2$ (1 equiv.) was added and the reaction stirred at rt overnight. The reaction was quenched with sat. NaHCO$_3$ solution, diluted with CH$_2$Cl$_2$ and the organic layer was separated. The aqueous layer was extracted with CH$_2$Cl$_2$ (2 x) and the combined organic layers were washed with sat. NaHCO$_3$ solution, brine, dried (MgSO$_4$), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica to give the corresponding triazolopyrazine core.

**General Procedure 3: Reduction of esters to alcohols.** Ester (1 equiv.) was dissolved in anhydrous THF (566 mM) and cooled to 0 °C. LiAlH$_4$ (1 M in THF, 2 equiv.) was added dropwise and the reaction mixture stirred for 10 min at 0 °C, then at rt. Upon completion, the
reaction was diluted with THF and cooled to 0 °C. H₂O (1 mL/1 g of LiAlH₄) was added followed by 15% aq. NaOH (1 mL/1 g of LiAlH₄) and H₂O (3 mL/1 g of LiAlH₄). The mixture was allowed to warm to rt and stirred for 15 min. MgSO₄ was added and the reaction mixture was filtered through a pad of celite and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica to give the corresponding alcohol.

**General Procedure 4: Nucleophilic displacement of triazolopyrazine core chlorine with alcohol.** Alcohol (1.0 equiv.) was added to PhMe (168 mM) along with triazolopyrazine core (1.0 equiv.), KOH (3.0 equiv.) and 18-crown-6 (0.1 equiv.). The reaction was stirred at rt until completion as indicated by TLC (100% EtOAc). The reaction was diluted with H₂O, then extracted with EtOAc (3 x). The combined organic layers were washed with H₂O until the aqueous layer became neutral, followed by brine, dried (MgSO₄), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica to give the corresponding ether-linked product.

**(E)-2-(2-(4-(tert-Butoxy)benzylidene)hydrazinyl)-6-chloropyrazine (2).**
Prepared according to General Procedure 1 from: compound 1 (750 mg, 5.19 mmol, 1 equiv.) and 4-(tert-butoxy)benzaldehyde (904 µL, 5.19 mmol, 1 equiv.) to give 2 as a pearlescent pale yellow powder (1.29 g, 82%). m.p. 207–211 °C; ¹H NMR (500 MHz, DMSO-d₆) δ: 11.47 (s, 1H), 8.53 (s, 1H), 8.03 (s, 1H), 8.02 (s, 1H), 7.65 (d, J = 8.6 Hz, 2H), 7.02 (d, J = 8.6 Hz, 2H), 1.33 (s, 9H); ¹³C NMR (126 MHz, DMSO-d₆) δ: 156.5, 152.4, 145.6, 142.5, 132.1, 129.1, 128.7, 127.6, 123.5, 78.6, 28.6; m/z (ESI+) 305 ([M+H]⁺, 100%).

**(E)-2-Chloro-6-(2-(4-isopropylbenzylidene)hydrazinyl)pyrazine (3).**
Prepared according to General Procedure 1 from: compound 1 (1.00 g, 6.92 mmol) and 4-isopropylbenzaldehyde (1.05 mL, 6.92 mmol) to give 3 as a
yellow powder (1.49 g, 78%). m.p. 231–233 °C; $^1$H NMR (500 MHz, DMSO-$d_6$) δ: 11.51 (s, 1H), 8.54 (s, 1H), 8.04 (s, 2H), 7.65 (d, $J = 8.2$ Hz, 2H), 7.30 (d, $J = 8.2$ Hz, 2H), 2.92 (p, $J = 6.9$ Hz, 1H), 1.21 (d, $J = 6.9$ Hz, 6H); $^{13}$C NMR (126 MHz, DMSO-$d_6$) δ: 152.4, 150.1, 145.6, 142.8, 132.2, 132.1, 128.7, 126.8, 126.7, 33.4, 23.7; m/z (ESI+) 275 ([M+H]$^+$, 100%).

(E)-2-Chloro-6-(2-(4-chlorobenzylidene)hydrazinyl)pyrazine (4).

Prepared according to General Procedure 1 from: compound 1 (1.00 g, 6.92 mmol) and 4-chlorobenzaldehyde (972 mg, 6.92 mmol) to give 4 as a yellow powder (1.53 g, 82 %). m.p. 230–233 °C; $^1$H NMR (500 MHz, DMSO-$d_6$) δ: 11.64 (s, 1H), 8.58 (s, 1H), 8.07 (s, 1H), 8.05 (s, 1H), 7.77 (d, $J = 8.5$ Hz, 2H), 7.48 (d, $J = 8.5$ Hz, 2H); $^{13}$C NMR (126 MHz, DMSO-$d_6$) δ: 152.2, 145.5, 141.3, 133.8, 133.4, 132.6, 128.9, 128.3 (one obscured signal); m/z (ESI+) 267 ([M+H]$^+$, 100%).

3-(4-((tert-Butoxy)phenyl)-5-chloro-[1,2,4]triazolo[4,3-a]pyrazine (6).

Prepared according to General Procedure 2 from: compound 2 (750 mg, 2.46 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 6 as a pale yellow powder (595 mg, 80%). m.p. 143–147 °C; $^1$H NMR (500 MHz, CDCl$_3$) δ: 9.31 (s, 1H), 7.85 (s, 1H), 7.52 (d, $J = 8.7$ Hz, 2H), 7.12 (d, $J = 8.7$ Hz, 2H), 1.43 (s, 9H); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 158.2, 148.4, 147.3, 143.1, 132.4, 129.8, 123.0, 122.2, 121.0, 79.7, 29.0; m/z (ESI+) 303 ([M+H]$^+$, 100%).

5-Chloro-3-(4-isopropylphenyl)-[1,2,4]triazolo[4,3-a]pyrazine (7).

Prepared according to General Procedure 2 from: compound 3 (750 mg, 2.73 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 7 as a pale orange powder (662 mg, 89 %). m.p. 133–137 °C; $^1$H NMR (500 MHz, CDCl$_3$) δ: 9.32 (s, 1H), 7.85 (s, 1H), 7.54 (d, $J = 8.3$ Hz, 2H), 7.37 (d, $J = 8.1$ Hz, 2H), 3.02 (p, $J = 6.9$ Hz, 1H), 1.32 (d, $J = 6.9$ Hz, 6H); $^{13}$C NMR (126 MHz, CDCl$_3$) δ:
5-Chloro-3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-a]pyrazine (8).

Prepared according to General Procedure 2 from: compound 4 (750 mg, 2.81 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 8 as an orange powder (634 mg, 85%). m.p. 180–183 °C; \(^1\)H NMR (300 MHz, CDCl\(_3\)) \(\delta\): 9.29 (s, 1H), 7.86 (s, 1H), 7.55 (d, \(J = 8.6\) Hz, 2H), 7.48 (d, \(J = 8.5\) Hz, 2H); \(^{13}\)C NMR (75 MHz, CDCl\(_3\)) \(\delta\): 147.2, 142.9, 137.2, 132.6, 129.8, 128.3, 125.1, 121.8 (one obscured signal); m/z (ESI+) 265 ([M+H]\(^+\), 100%).

Methyl 2-(3,4-difluorophenyl)-2-hydroxyacetate (10).

2-(3,4-Difluorophenyl)-2-hydroxyacetic acid (850 mg, 4.52 mmol, 1 equiv.) and \(\text{p-TsOH monohydrate (17.2 mg, 0.09 mmol, 0.02 equiv.)}\) were dissolved in MeOH (3.12 mL, 1.45 M) and the reaction heated to reflux (80 °C). The reaction was cooled to rt and the solvent removed. EtOAc was added to the residue and the organic layer washed with H\(_2\)O, sat. NaHCO\(_3\) solution, brine, dried (MgSO\(_4\)), filtered and concentrated under reduced pressure to give 10 as clear colourless oil that solidified on standing (738 mg, 81%). No further purification required. m.p. 44–49 °C; \(^1\)H NMR (500 MHz, CDCl\(_3\)) \(\delta\): 7.30 – 7.24 (m, 1H), 7.23 – 7.02 (m, 2H), 5.14 (d, \(J = 5.0\) Hz, 1H), 3.78 (s, 3H), 3.52 (d, \(J = 5.1\) Hz, 1H); \(^{13}\)C NMR (126 MHz, CDCl\(_3\)) \(\delta\): 173.6, 151.5 (d, \(J = 12.6\) Hz), 149.5 (d, \(J = 12.6\) Hz), 135.1, 122.8, 117.5 (d, \(J = 17.5\) Hz), 115.8 (d, \(J = 18.4\) Hz), 71.8, 53.5; m/z (ESI+) 225 ([M+Na]\(^+\), 100%).

Methyl 2-(3,4-difluorophenyl)-2-((tetrahydro-2\(H\)-pyran-2-yl)oxy)acetate (11).

Compound 10 (600 mg, 2.97 mmol, 1.0 equiv.) was dissolved in CH\(_2\)Cl\(_2\) (10.7 mL, 277 mM), \(\text{p-TsOH (102 mg, 0.59 mmol, 0.2 equiv.)}\) and 3,4-dihydro-2\(H\)-pyran (0.3 mL, 3.26 mmol, 1.1 equiv.) were added and the reaction stirred at rt. The reaction was quenched with ice cold H\(_2\)O and the organic layer separated. The aqueous layer was
extracted with CH₂Cl₂ (3 x) and the combined organic layers washed with H₂O, brine, dried (MgSO₄), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica (6–50% ethyl acetate in hexanes) to give 11 as a viscous dark orange oil (395 mg, 46%). ¹H NMR (500 MHz, CDCl₃, present as a mixture of diastereomers) δ: 7.38 – 7.28 (m, 2H), 7.24 – 7.10 (m, 4H), 5.27 (s, 1H), 5.18 (s, 1H), 4.86 (t, J = 3.0 Hz, 1H), 4.57 (t, J = 3.4 Hz, 1H), 3.72 (s, 6H), 3.55 – 3.44 (m, 4H), 1.93 – 1.36 (m, 12H); ¹³C NMR (126 MHz, CDCl₃, present as a mixture of diastereomers) δ: 171.3, 170.7, 150.7 (dd, J = 249.4, 12.0 Hz), 150.6 (dd, J = 249.3, 13.1 Hz), 150.5 (dd, J = 249.3, 13.4 Hz), 150.4 (dd, J = 248.5, 12.6 Hz), 133.8 (dd, J = 5.4, 4.2 Hz), 133.7 – 132.9 (m), 123.8 (dd, J = 6.6, 3.6 Hz), 123.4 (dd, J = 6.4, 3.7 Hz), 117.5 (t, J = 18.1 Hz, 2C), 116.6 (d, J = 18.0 Hz), 116.4 (d, J = 18.4 Hz), 97.4, 97.0, 75.8, 74.5, 62.5, 62.3, 52.62, 52.58, 30.3, 30.2, 25.3 (2C), 19.1, 18.8; m/z (ESI⁺) 309 ([M+Na]⁺, 100%).

2-(3,4-Difluorophenyl)-2-((tetrahydro-2H-pyran-2-yl)oxy)ethan-1-ol (12).

Prepared according to General Procedure 3 from: compound 11 (300 mg, 1.05 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 12 as a viscous pale yellow oil (150 mg, 55%). ¹H NMR (500 MHz, CDCl₃, present as a mixture of diastereomers) δ: 7.28 – 6.92 (m, 6H), 4.92 – 4.81 (m, 1H), 4.81 – 4.72 (m, 1H), 4.68 (dd, J = 6.8, 4.6 Hz, 1H), 4.50 (dd, J = 5.6, 2.8 Hz, 1H), 4.00 (dt, J = 11.0, 5.2 Hz, 1H), 3.73 – 3.62 (m, 4H), 3.56 (tt, J = 10.2, 4.6 Hz, 2H), 3.32 (dt, J = 10.9, 4.7 Hz, 1H), 3.05 – 2.96 (m, 1H), 2.18 – 2.07 (m, 1H), 1.91 – 1.36 (m, 12H); ¹³C NMR (126 MHz, CDCl₃, present as a mixture of diastereomers) δ: 150.5 (dd, J = 248.8, 12.8 Hz), 150.4 (dd, J = 248.2, 12.7 Hz), 150.1 (dd, J = 248.3, 12.7 Hz), 149.9 (dd, J = 247.6, 12.6 Hz), 137.7 – 136.7 (m), 136.3 – 135.5 (m), 122.9 (dd, J = 6.3, 3.6 Hz), 122.7 (dd, J = 6.3, 3.6 Hz), 117.4 (d, J = 17.3 Hz), 117.2 (d, J = 17.2 Hz), 115.9 (d, J = 17.7 Hz), 115.8 (d, J = 17.8 Hz), 99.6, 98.1, 79.6, 78.8, 67.5, 66.5, 63.9, 62.9, 31.1, 30.7, 25.30, 25.26, 20.3, 19.6; m/z (ESI⁺) 281 ([M+Na]⁺, 100%).
(6-Methylpyridin-3-yl)methanol (13).

Prepared according to General Procedure 3 from: 6-methylnicotinic acid (750 mg, 5.47 mmol); purified by automated flash chromatography on silica (1–10% MeOH in CH₂Cl₂) to give 13 as a yellow oil (103 mg, 15%). ¹H NMR (500 MHz, CDCl₃) δ: 8.42 (d, J = 2.0 Hz, 1H), 7.61 (dd, J = 7.9, 2.2 Hz, 1H), 7.14 (d, J = 7.9 Hz, 1H), 4.66 (s, 2H), 2.53 (s, 3H) (alcohol OH signal not seen); ¹³C NMR (126 MHz, CDCl₃) δ: 157.6, 147.8, 135.8, 133.6, 123.3, 62.8, 30.1; m/z (ESI+) 146 ([M+Na]⁺, 100%). Spectroscopic data matched those in the literature [45].

tert-Butyl (3-(2-hydroxyethyl)benzyl)carbamate (14).

Prepared according to General Procedure 3 from: 2-(3-((tert-butoxycarbonyl)amino)methyl)phenyl)acetic acid (500 mg, 1.88 mmol); purified by automated flash chromatography on silica (1–10% MeOH in CH₂Cl₂) to give 14 as a viscous clear colourless oil (332 mg, 70%). ¹H NMR (500 MHz, CDCl₃) δ: 7.27 (t, J = 7.8 Hz, 1H), 7.18 – 7.05 (m, 3H), 4.90 (br s, 1H), 4.29 (br d, J = 5.5 Hz, 2H), 3.84 (t, J = 6.6 Hz, 2H), 2.84 (t, J = 6.6 Hz, 2H), 1.46 (s, 9H) (alcohol OH signal not seen); ¹³C NMR (126 MHz, CDCl₃) δ: 156.1, 139.3, 139.1, 129.0, 128.2, 128.1, 125.7, 79.7, 63.7, 44.7, 39.2, 28.5; m/z (ESI+) 274 ([M+Na]⁺, 100%).

Methyl 2-((tert-butoxycarbonyl)amino)phenyl)acetate (15).

Boc₂O (363 mg, 1.66 mmol, 1.1 equiv.) was added to a solution of 3-aminophenylacetic acid methyl ester (250 mg, 1.51 mmol, 1.0 equiv.) in CH₂Cl₂ (4 mL) and the reaction was stirred at rt overnight. Sat. NH₄Cl solution (6 mL) was added and the organic layer separated. The aqueous layer was extracted with CH₂Cl₂ (2 x 8 mL) and the combined organic layers were dried (Na₂SO₄), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica (12–50% ethyl acetate in hexanes) to give 15 as a clear colourless oil (174 mg, 43%). ¹H NMR (500 MHz, CDCl₃) δ: 7.33 (br s, 1H), 7.23 (d, J = 4.4 Hz, 2H), 6.99 – 6.90 (m,
1H), 6.51 (br s, 1H), 3.68 (s, 3H), 3.59 (s, 2H), 1.51 (s, 9H); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 172.0, 152.8, 138.7, 135.0, 129.3, 124.0, 119.4, 117.4, 80.7, 52.2, 41.2, 28.4; m/z (ESI+) 288 ([M+Na]$^+$, 100%).

**tert-Butyl (3-(2-hydroxyethyl)phenyl)carbamate (16).**

Prepared according to General Procedure 3 from: compound 15 (130 mg, 0.49 mmol); purified by automated flash chromatography on silica (12–100% EtOAc in hexanes) to give 16 as a viscous clear colourless oil (71.3 mg, 61%). $^1$H NMR (500 MHz, CDCl$_3$) δ: 7.30 (br s, 1H), 7.22 (t, $J = 7.7$ Hz, 1H), 7.16 (d, $J = 8.1$ Hz, 1H), 6.90 (d, $J = 7.3$ Hz, 1H), 6.55 (br s, 1H), 3.84 (t, $J = 6.5$ Hz, 2H), 2.83 (t, $J = 6.5$ Hz, 2H), 1.51 (s, 9H) (alcohol OH signal not seen); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 152.9, 139.7, 138.7, 129.2, 123.9, 119.2, 116.8, 80.7, 63.6, 39.3, 28.5; m/z (ESI+) 260 ([M+Na]$^+$, 100%). Spectroscopic data matched those in the literature [46].

**2-(Furan-2-yl)ethan-1-ol (19).**

Prepared according to General Procedure 3 from: 2-(furan-2-yl)acetic acid (500 mg, 3.96 mmol); purified by automated flash chromatography on silica (12–100% EtOAc in hexanes) to give 16 as a yellow oil (291 mg, 65%). $^1$H NMR (500 MHz, CDCl$_3$) δ: 7.69 – 7.19 (m, 1H), 6.30 (dd, $J = 2.9$, 2.0 Hz, 1H), 6.09 (d, $J = 3.1$ Hz, 1H), 3.85 (t, $J = 6.3$ Hz, 2H), 2.88 (t, $J = 6.3$ Hz, 2H) (alcohol OH signal not seen); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 153.0, 141.6, 110.4, 106.5, 61.1, 31.6.

**3-(4-(tert-Butoxy)phenyl)-5-(2-(3,4-difluorophenyl)-2-((tetrahydro-2H-pyran-2-yl)oxy)ethoxy)-[1,2,4]triazolo[4,3-a]pyrazine (20).**

Prepared according to General Procedure 4 from: compound 12 (100 mg, 0.39 mmol) and compound 6 (117 mg, 0.39 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 20 as an orange powder (95.9 mg, 47%). m.p. 72–75 °C; $^1$H NMR (500 MHz, CDCl$_3$, present as a
mixture of diastereomers) δ: 9.02 (s, 1H), 9.01 (s, 1H), 7.62 (dd, \( J = 8.6, 1.5 \) Hz, 4H), 7.38 (s, 1H), 7.29 (s, 1H), 7.14 (dd, \( J = 8.6, 1.5 \) Hz, 4H), 7.09 – 7.01 (m, 1H), 7.00 (dt, \( J = 9.9, 8.2 \) Hz, 1H), 6.95 (ddd, \( J = 10.6, 7.6, 2.1 \) Hz, 1H), 6.80 (dq, \( J = 6.4, 2.0 \) Hz, 1H), 6.73 (ddd, \( J = 10.2, 7.6, 2.1 \) Hz, 1H), 6.60 (dq, \( J = 6.1, 1.9 \) Hz, 1H), 4.89 (t, \( J = 5.5 \) Hz, 1H), 4.70 (dd, \( J = 6.7, 4.8 \) Hz, 1H), 4.44 (dd, \( J = 9.9, 5.3 \) Hz, 1H), 4.37 (q, \( J = 5.1, 4.3 \) Hz, 2H), 4.29 (dd, \( J = 9.8, 6.9 \) Hz, 1H), 4.22 (ddd, \( J = 19.1, 9.9, 5.3 \) Hz, 2H), 3.77 (ddd, \( J = 10.7, 7.1, 3.3 \) Hz, 1H), 3.47 – 3.40 (m, 1H), 3.37 (ddd, \( J = 11.7, 9.1, 3.0 \) Hz, 1H), 3.25 (dt, \( J = 11.0, 4.6 \) Hz, 1H), 1.87 – 1.45 (m, 12H), 1.44 (s, 9H), 1.43 (s, 9H); \(^{13}\)C NMR (126 MHz, CDCl\(_3\), present as a mixture of diastereomers) δ: 157.8, 157.7, 150.5 (dd, \( J = 249.7, 12.6 \) Hz), 150.42 (dd, \( J = 250.0, 12.5 \) Hz), 150.36 (dd, \( J = 249.0, 13.0 \) Hz), 150.2 (dd, \( J = 248.7, 12.5 \) Hz), 147.88, 147.3, 147.2, 144.04, 143.99, 137.1, 137.0, 136.2 – 135.5 (m), 134.6 (t, \( J = 4.2 \) Hz), 131.7, 131.6, 123.3 (dd, \( J = 6.3, 3.5 \) Hz), 123.1, 123.0, 122.6 (dd, \( J = 6.2, 3.6 \) Hz), 122.3, 122.1, 117.6 (d, \( J = 17.4 \) Hz), 117.4 (d, \( J = 17.4 \) Hz), 116.1 (d, \( J = 17.7 \) Hz), 115.8 (d, \( J = 17.9 \) Hz), 109.1, 108.7, 99.3, 96.6, 79.6, 74.0, 73.7, 73.4, 63.2, 62.4, 30.6, 30.5, 29.04, 29.03, 25.3, 25.2, 19.7, 19.1 (two obscured signals); \( m/z \) (ESI+) 525 ([M+H]\(^+\), 100%).

** tert-Butyl (3-(2-((3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)benzyl)carbamate (21).**

Prepared according to General Procedure 4 from: compound 14 (150 mg, 0.60 mmol) and compound 8 (158 mg, 0.60 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 21 as a light brown powder (209 mg, 73%). \(^1\)H NMR (500 MHz, CDCl\(_3\)) δ: 8.98 (s, 1H), 7.59 (d, \( J = 8.5 \) Hz, 2H), 7.39 (d, \( J = 8.5 \) Hz, 2H), 7.30 (s, 1H), 7.19 (t, \( J = 7.5 \) Hz, 1H), 7.14 (d, \( J = 7.6 \) Hz, 1H), 6.86 (br s, 1H), 6.75 (d, \( J = 7.4 \) Hz, 1H), 4.86 (br s, 1H), 4.43 (t, \( J = 6.6 \) Hz, 2H), 4.22 (d, \( J = 5.6 \) Hz, 2H), 2.94 (t, \( J = 6.5 \) Hz, 2H), 1.43 (s, 9H); \(^{13}\)C NMR (126 MHz, CDCl\(_3\)) δ: 156.0, 147.9, 146.3, 143.9, 139.7, 136.6, 136.5, 136.4, 132.1, 129.1, 128.1, 127.8, 127.6, 126.33, 126.26, 108.5, 79.7, 71.2, 44.6, 34.5, 28.5; \( m/z \) (ESI+) 480 ([M+H]\(^+\), 100%), 502 ([M+Na]\(^+\), 52%).
**tert-Butyl (3-(2-((3-(4-chlorophenyl)[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)phenyl)carbamate (22).**

Prepared according to General Procedure 4 from: compound 16 (35.0 mg, 0.15 mmol) and compound 8 (39.1 mg, 0.15 equiv.); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give 22 as a light brown powder (30.1 mg, 44%). $^1$H NMR (500 MHz, CDCl$_3$) δ: 9.02 (s, 1H), 7.58 (d, $J =$ 8.5 Hz, 2H), 7.42 (d, $J =$ 8.5 Hz, 2H), 7.31 (s, 1H), 7.18 – 7.06 (m, 3H), 6.55 (d, $J =$ 7.3 Hz, 1H), 6.43 (br s, 1H), 4.44 (t, $J =$ 6.5 Hz, 2H), 2.93 (t, $J =$ 6.5 Hz, 2H), 1.51 (s, 9H); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 152.7, 148.0, 146.4, 144.0, 138.9, 137.2, 136.6, 136.4, 128.2, 126.4, 123.1, 118.5, 117.2, 108.5, 80.9, 71.2, 34.6, 28.5; m/z (ESI+) 466 ([M+H]$^+$, 100%), 488 ([M+Na]$^+$, 45%).

**2-((3-(4-(tert-Butoxy)phenyl)[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)-1-(3,4-difluorophenyl)ethan-1-ol (Optibrium/Intellegens).**

Compound 20 (70.0 mg, 0.13 mmol, 1 equiv.) was dissolved in EtOH (1.63 mL, 82 mM). CuCl$_2$•2H$_2$O (1.14 mg, 6.67 µmol, 5 mol%) was added and the reaction heated at reflux. The solvent was removed and EtOAc was added. The mixture was washed with H$_2$O (3 x), then brine, and the organic layer was dried (MgSO$_4$), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give Optibrium/Intellegens as a pale orange powder (49.2 mg, 84%). m.p. 75–79 °C; $^1$H NMR (500 MHz, CDCl$_3$) δ: 9.05 (s, 1H), 7.66 (d, $J =$ 8.6 Hz, 2H), 7.17 (d, $J =$ 8.6 Hz, 2H), 7.17 – 7.06 (m, 1H), 7.03 (ddd, $J =$ 10.2, 7.4, 1.6 Hz, 1H), 6.93 – 6.87 (m, 1H), 4.78 (dt, $J =$ 7.4, 3.3 Hz, 1H), 4.26 (dd, $J =$ 9.3, 3.3 Hz, 1H), 4.19 – 4.08 (m, 1H), 2.00 (d, $J =$ 3.7 Hz, 1H), 1.42 (s, 9H) (alcohol OH signal not seen); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 158.0, 150.6 (dd, $J =$ 249.5, 12.7 Hz), 150.4 (dd, $J =$ 249.7, 12.6 Hz), 147.8, 146.9, 143.8, 137.3, 135.2 – 135.0 (m), 131.7, 123.2, 122.6, 122.2 (dd, $J =$ 6.4, 3.6 Hz), 117.8 (d, $J =$ 17.4 Hz), 115.4 (d, $J =$
18.2 Hz), 108.5, 79.9, 75.0, 70.6, 29.0; m/z (ESI+) 441 ([M+H]^+, 100%); HRMS (ESI+) found 441.1739 [M+H]^+ requires 441.1738.

3-(4-Isopropylphenyl)-5-((6-methylpyridin-3-yl)methoxy)-[1,2,4]triazolo[4,3-a]pyrazine (Davy Guan).

Prepared according to General Procedure 4 from: compound 13 (100 mg, 0.81 mmol) and compound 7 (221 mg, 0.81 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give a white powder (181 mg, 62%). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H₂O) to give a white powder (40.6 mg). Repurified by automated reversed-phase flash chromatography on silica (5–75% MeOH in H₂O) to give Davy Guan as a white powder (16.3 mg, 6%). m.p. decomposed >150 °C; ^1H NMR (400 MHz, CD₃OD) δ: 9.00 (s, 1H), 8.33 (d, J = 2.0 Hz, 1H), 7.65 (s, 1H), 7.54 – 7.48 (m, 3H), 7.23 (d, J = 8.0 Hz, 1H), 7.12 (d, J = 8.2 Hz, 2H), 5.32 (s, 2H), 2.87 (p, J = 6.9 Hz, 1H), 2.53 (s, 3H), 1.20 (d, J = 6.9 Hz, 6H); ^13C NMR (101 MHz, CD₃OD) δ: 160.0, 153.0, 152.3, 149.9, 148.9, 146.0, 139.2, 136.6, 131.9, 128.7, 126.6, 125.8, 124.8, 110.2, 71.3, 35.2, 24.2, 23.7; m/z (ESI+) 360 ([M+H]^+, 100%); HRMS (ESI+) found 360.1832 [M+H]^+ requires 360.1824.

(3-(2-((3-(4-Chlorophenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)phenyl)methanamine (Exscientia 1).

Compound 21 (150 mg, 0.31 mmol, 1.00 equiv.) was dissolved in CH₂Cl₂ (0.91 mL, 345 mM). TFA (0.27 mL, 3.50 mmol, 11.2 equiv.) was added and the reaction stirred at rt overnight. The solvent was removed and the residue directly purified by automated flash chromatography on silica (1–10% MeOH in CH₂Cl₂, then 100% MeOH) to give a sticky brown solid (135 mg). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H₂O) to give Exscientia 1 as a white powder (77.0 mg, 65%). m.p. 74–77 °C; ^1H NMR (400 MHz, CD₃OD) δ: 8.97 (s, 1H), 7.68 (d, J = 8.5 Hz, 2H), 7.54 (s, 1H), 7.50 (d, J = 8.5 Hz, 2H), 7.32 – 7.25 (m, 2H), 7.07 (s, 1H), 6.98
– 6.89 (m, 1H), 4.60 (t, J = 6.2 Hz, 2H), 4.03 (s, 2H), 3.02 (t, J = 6.2 Hz, 2H) (amine NH2 signal not seen); 13C NMR (126 MHz, CDCl3) δ: 149.0, 147.6, 145.9, 139.7, 137.5, 136.2, 134.7, 133.5, 130.6, 130.4, 129.2, 128.1, 127.6, 110.0, 72.4, 44.2, 35.2; m/z (ESI+) 380 ([M+H]+, 100%); HRMS (ESI+) found 380.1282 [M+H]+, C20H18ClN5OH+ requires 380.1278.

3-(2-((3-(4-Chlorophenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)aniline (Exscientia 2).

Compound 22 (21.0 mg, 0.05 mmol, 1.00 equiv.) was dissolved in CH2Cl2 (0.13 mL, 345 mM). TFA (0.04 mL, 3.50 mmol, 11.2 equiv.) was added and the reaction stirred at rt overnight. The solvent was removed and the residue directly purified by automated flash chromatography on silica (1–10% MeOH in CH2Cl2) to give a yellow film (23.5 mg). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H2O) to give Exscientia 2 as a white powder (11.4 mg, 69%). m.p. 80–84°C; 1H NMR (500 MHz, CD3OD) δ: 8.93 (s, 1H), 7.61 (d, J = 8.6 Hz, 2H), 7.52 (s, 1H), 7.46 (d, J = 8.6 Hz, 2H), 6.94 (t, J = 7.8 Hz, 1H), 6.55 (d, J = 9.3 Hz, 1H), 6.37 (s, 1H), 6.26 (d, J = 7.5 Hz, 1H), 4.51 (t, J = 6.4 Hz, 2H), 2.84 (t, J = 6.4 Hz, 2H) (amine NH2 signal not seen); 13C NMR (126 MHz, CDCl3) δ: 149.0, 149.0, 147.8, 145.9, 139.2, 137.5, 136.0, 133.5, 130.2, 129.1, 127.4, 119.5, 116.4, 114.9, 110.0, 72.8, 35.3; m/z (ESI+) 366 ([M+H]+, 100%); HRMS (ESI+) found 366.1122 [M+H]+, C19H16ClN5OH+ requires 366.1122.

1-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)propan-2-ol (Molomics 1).

To a mixture of lactic acid (0.83 mL, 11.1 mmol, 1.0 equiv.), 3,4-dihydro-2H-pyran (3.04 mL, 33.3 mmol, 3.0 equiv.) and CH2Cl2 (2.98 mL, 3.73 M) in an ice bath was added pyridinium p-toluenesulfonate (279 mg, 1.11 mmol, 0.1 equiv.) and pyridine (1 drop). The reaction was stirred at rt overnight. CH2Cl2 was added and the solution washed with 5% NaHCO3, H2O (2 x), dried (MgSO4), filtered and concentrated under reduced
pressure to give a mixture of 17a and 17b as a clear colourless oil (1.09 g, 38%). Used without further purification. m/z (ESI+) 209 (17b, [M+Na]⁺, 100%); m/z (ESI+) 281 (17a, [M+Na]⁺, 100%). The mixture of 17a and 17b (700 mg, 2.71 mmol) was subjected to General Procedure 3 and purified by automated flash chromatography on silica (12–100% EtOAc to hexanes) to give a ~1:2 mixture of 18a and 18b as a clear colourless oil (96.3 mg, 17%).

1H NMR (500 MHz, CDCl₃) δ: 4.72 (dd, J = 5.1, 2.8 Hz, 1H), 4.56 (dd, J = 4.3, 2.8 Hz, 2H), 3.96 – 3.90 (m, 1H), 3.86 (dtd, J = 11.2, 7.6, 7.1, 3.5 Hz, 3H), 3.74 (dt, J = 9.6, 6.7 Hz, 2H), 3.64 (t, J = 6.5 Hz, 4H), 3.58 (dd, J = 11.6, 3.5 Hz, 1H), 3.54 – 3.43 (m, 3H), 3.39 (dt, J = 9.6, 6.5 Hz, 2H), 1.92 – 1.33 (m, 33H), 1.21 (d, J = 6.4 Hz, 3H); 13C NMR (126 MHz, CDCl₃) δ: 99.2, 99.1, 75.1, 67.6, 66.3, 63.4, 63.0, 62.6, 32.7, 31.2, 30.9, 29.6, 25.6, 25.5, 22.6, 20.2, 19.8, 17.8; m/z (ESI+) 183 (18a, [M+Na]⁺, 100%); m/z (ESI+) 211 (18b, [M+Na]⁺, 100%). The mixture of 18a and 18b (70.0 mg, 0.44 mmol) was subjected to General Procedure 4 with compound 9 (130 mg, 0.44 mmol) and purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give a ~1:1 mixture of 23a and 23b as a brown powder (87.0 mg, 47%). 1H NMR (500 MHz, CDCl₃) δ: 9.05 (s, 1H), 9.03 (s, 1H), 7.73 (t, J = 9.0 Hz, 4H), 7.33 (s, 1H), 7.29 (s, 1H), 7.29 – 7.20 (m, 4H), 6.64 (t, J = 73.3 Hz, 1H), 6.63 (t, J = 73.3 Hz, 1H), 4.52 (t, J = 3.7 Hz, 1H), 4.24 – 4.14 (m, 3H), 4.16 – 4.09 (m, 2H), 3.97 – 3.88 (m, 1H), 3.82 (dtd, J = 11.7, 8.8, 8.2, 3.8 Hz, 2H), 3.66 (dt, J = 9.7, 6.6 Hz, 1H), 3.54 – 3.40 (m, 2H), 3.29 (dt, J = 9.5, 6.2 Hz, 1H), 1.85 – 1.30 (m, 18H), 1.07 (d, J = 6.5 Hz, 3H); 13C NMR (126 MHz, CDCl₃) δ: 152.7, 152.5, 147.91, 147.88, 146.3, 144.3, 144.2, 136.6, 136.3, 132.6, 132.5, 125.0, 124.9, 118.9, 118.5, 115.9 (t, J = 260.7 Hz), 115.7 (t, J = 261.1 Hz), 108.5, 108.2, 99.2, 98.7, 74.0, 71.0, 70.8, 67.2, 62.7, 62.4, 53.6, 30.85, 30.82, 29.2, 28.4, 25.5, 25.4, 22.8, 19.9, 19.3, 18.2; m/z (ESI+) 421 (23a, [M+H]⁺, 100%); m/z (ESI+) 449 (23b, [M+H]⁺, 100%). The mixture of 23a and 23b (56.0 mg, 0.13 mmol, 1 equiv.) was dissolved in EtOH (1.62 mL, 82 mM). CuCl₂•2H₂O (1.14 mg, 6.66 μmol, 5 mol%) was added and the reaction heated at reflux. The solvent was removed and EtOAc was added. The mixture was washed with H₂O (3 x), then brine, and the organic layer was dried (MgSO₄), filtered and concentrated under reduced pressure to give the crude product, which was purified by
automated flash chromatography on silica (1–15% MeOH in CH₂Cl₂) to give a brown oil (39.8 mg). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H₂O) to give Molomics 1 as a white powder (10.3 mg, 23%). m.p. 137–140 °C; ¹H NMR (400 MHz, CDCl₃) δ: 9.07 (s, 1H), 7.75 (d, J = 8.7 Hz, 2H), 7.32 (s, 1H), 7.29 (d, J = 8.6 Hz, 2H), 6.62 (t, J = 73.0 Hz, 1H), 4.18 – 4.12 (m, 1H), 4.04 – 3.90 (m, 2H), 1.06 (d, J = 6.3 Hz, 3H); ¹³C NMR (126 MHz, CDCl₃) δ: 152.5, 147.9, 146.1, 144.0, 136.9, 132.5, 125.3, 119.1, 115.6 (t, J = 26.2 Hz, 1H), 108.5, 75.6, 65.5, 18.7; m/z (ESI [+]) 337 ([M+H]⁺, 100%); HRMS (ESI [+]) found 337.1107 [M+H]⁺, C₁₅H₁₄F₂N₄O₃ requires 337.1112.

5-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)pentan-1-ol (Molomics 1').

Isolated from the same reaction as for Molomics 1 to give Molomics 1' as a white powder (16.1 mg, 36%). m.p. 104–108 °C; ¹H NMR (500 MHz, CDCl₃) δ: 9.01 (s, 1H), 7.70 (d, J = 8.7 Hz, 2H), 7.28 (s, 1H), 7.24 (d, J = 8.6 Hz, 2H), 6.65 (t, J = 73.2 Hz, 1H), 4.20 (t, J = 6.0 Hz, 2H), 3.54 (t, J = 6.3 Hz, 2H), 1.79 – 1.54 (m, 2H), 1.53 – 1.30 (m, 2H), 1.14 (ddd, J = 11.7, 4.6, 2.5 Hz, 2H); ¹³C NMR (126 MHz, CDCl₃) δ: 152.5, 147.9, 146.3, 144.3, 136.3, 132.6, 125.1, 118.6, 115.7 (t, J = 261.3 Hz), 108.2, 71.0, 62.5, 32.0, 28.5, 22.3; m/z (ESI [+]) 365 ([M+H]⁺, 100%); HRMS (ESI [+]) found 365.1424 [M+H]⁺, C₁₇H₁₈F₂N₄O₃ requires 365.1425.

3-(4-(Difluoromethoxy)phenyl)-5-(2-(furan-2-yl)ethoxy)-[1,2,4]triazolo[4,3-a]pyrazine (Molomics 2).

Prepared according to General Procedure 4 from: compound 19 (100 mg, 0.89 mmol) and compound 9 (265 mg, 0.89 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give Molomics 2 as a brown powder (213 mg, 64%). m.p. 120–123°C; ¹H NMR (500 MHz, CDCl₃) δ: 9.03 (s, 1H), 7.65 (d, J = 8.7 Hz, 2H), 7.33 (s, 1H), 7.31 – 7.27 (m, 1H), 7.17 (d, J = 8.7 Hz, 2H), 6.59 (t, J = 73.3 Hz, 1H), 6.26 (dd, J = 3.1, 1.9 Hz, 1H), 5.82 (d, J = 3.1 Hz, 1H), 4.47 (t, J = 6.3 Hz,
2H), 2.98 (t, J = 6.3 Hz, 2H); $^{13}$C NMR (126 MHz, CDCl$_3$) δ: 152.5 (t, J = 2.7 Hz), 150.1, 147.9, 146.5, 143.9, 142.0, 136.8, 132.5, 124.9, 118.7, 115.7 (t, J = 261.1 Hz), 110.6, 108.4, 107.1, 68.6, 27.5; m/z (ESI+) 373 ([M+H]$^+$, 100%); HRMS (ESI+) found 373.1118 [M+H]$^+$, $\text{C}_{18}\text{H}_{14}\text{F}_{2}\text{N}_{4}\text{O}_{3}\text{H}^+$ requires 373.1112.
$^{1}\text{H}$ and $^{13}\text{C}$ NMR Spectra

(E)-2-(2-(4-(tert-Butoxy)benzylidene)hydrazinyl)-6-chloropyrazine (2)

$^{1}\text{H}$ NMR (500 MHz, DMSO-$d_6$)

$^{13}\text{C}$ NMR (126 MHz, DMSO-$d_6$)
(E)-2-Chloro-6-(2-(4-isopropylbenzylidene)hydrazinyl)pyrazine (3)

\[^1\text{H} \text{NMR (500 MHz, DMSO-}d_6\text{)}\]

\[^{13}\text{C} \text{NMR (126 MHz, DMSO-}d_6\text{)}\]
(E)-2-Chloro-6-(2-(4-chlorobenzylidene)hydrazinyl)pyrazine (4)

$^1$H NMR (500 MHz, DMSO-$d_6$)

$^{13}$C NMR (126 MHz, DMSO-$d_6$)
3-(4-(tert-Butoxy)phenyl)-5-chloro-[1,2,4]triazolo[4,3-a]pyrazine (6)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
5-Chloro-3-(4-isopropylphenyl)-[1,2,4]triazolo[4,3-a]pyrazine (7)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
5-Chloro-3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-a]pyrazine (8)

$^1$H NMR (300 MHz, CDCl$_3$)

$^{13}$C NMR (75 MHz, CDCl$_3$)
Methyl 2-(3,4-difluorophenyl)-2-hydroxyacetate (10)

$^1$H NMR (500 MHz, CDCl$_3$)

![Chemical structure of methyl 2-(3,4-difluorophenyl)-2-hydroxyacetate (10)](image)

$^{13}$C NMR (126 MHz, CDCl$_3$)

![Chemical structure of methyl 2-(3,4-difluorophenyl)-2-hydroxyacetate (10)](image)
Methyl 2-(3,4-difluorophenyl)-2-((tetrahydro-2H-pyran-2-yl)oxy)acetate (11)

$^1$H NMR (500 MHz, CDCl$_3$, present as a mixture of diastereomers)

$^{13}$C NMR (126 MHz, CDCl$_3$, present as a mixture of diastereomers)
2-(3,4-Difluorophenyl)-2-((tetrahydro-2H-pyran-2-yl)oxy)ethan-1-ol (12)

\(^1\)H NMR (500 MHz, CDCl\(_3\), present as a mixture of diastereomers)

\(^{13}\)C NMR (126 MHz, CDCl\(_3\), present as a mixture of diastereomers)
(6-Methylpyridin-3-yl)methanol (13)

\(^1\)H NMR (500 MHz, CDCl\(_3\))

\(^{13}\)C NMR (126 MHz, CDCl\(_3\))
**tert-Butyl (3-(2-hydroxyethyl)benzyl)carbamate (14)**

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
Methyl 2-((tert-butoxycarbonyl)amino)phenyl)acetate (15)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
tert-Butyl (3-(2-hydroxyethyl)phenyl)carbamate (16)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
2-((Tetrahydro-2H-pyran-2-yl)oxy)propan-1-ol (18a) and 5-((tetrahydro-2H-pyran-2-yl)oxy)pentan-1-ol (18b)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
2-(Furan-2-yl)ethan-1-ol (19)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
3-(4-(tert-Butoxy)phenyl)-5-(2-(3,4-difluorophenyl)-2-((tetrahydro-2H-pyran-2-yl)oxy)ethoxy)-[1,2,4]triazolo[4,3-a]pyrazine (20)

\(^1\)H NMR (500 MHz, CDCl\(_3\), present as a mixture of diastereomers)

\(^{13}\)C NMR (126 MHz, CDCl\(_3\), present as a mixture of diastereomers)
tert-Butyl (3-((2-(3-(4-chlorophenyl)[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)benzyl)carbamate (21)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
**tert-Butyl** (3-((2-((3-(4-chlorophenyl)[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)phenyl)carbamate (22)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
3-(4-(Difluoromethoxy)phenyl)-5-(2-((tetrahydro-2H-pyran-2-yl)oxy)propoxy)-
[1,2,4]triazolo[4,3-a]pyrazine (23a) and 3-(4-(difluoromethoxy)phenyl)-5-((5-
(tetrahydro-2H-pyran-2-yl)oxy)pentyloxy)[1,2,4]triazolo[4,3-a]pyrazine (23b)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
2-((3-(tert-Butoxy)phenyl)[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)-1-(3,4-difluorophenyl)ethan-1-ol (Optibrium/Intelligen)

\[^{1}H\] NMR (500 MHz, CDCl\textsubscript{3})

\[^{13}C\] NMR (126 MHz, CDCl\textsubscript{3})
3-(4-Isopropylphenyl)-5-((6-methylpyridin-3-yl)methoxy)-[1,2,4]triazolo[4,3-a]pyrazine
(Davy Guan)

$^1$H NMR (400 MHz, CD$_3$OD)

$^{13}$C NMR (101 MHz, CD$_3$OD)
(3-(2-((3-(4-Chlorophenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)ethyl)phenyl)methanamine (Exscientia 1)

$^1$H NMR (400 MHz, CD$_3$OD)

$^{13}$C NMR (126 MHz, CD$_3$OD)
3-(2-(3-(4-Chlorophenyl)\-\-1,2,4\-triazolo\-4,3-\-a\-pyrazin-5-yl)oxy)ethyl)aniline (Exscientia 2)

$^1$H NMR (500 MHz, CD$_3$OD)

$^{13}$C NMR (126 MHz, CD$_3$OD)
1-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)propan-2-ol
(Molomics 1)

\[ \text{\textsuperscript{1}H NMR (400 MHz, CDCl} \textsubscript{3}) \]

\[ \text{\textsuperscript{13}C NMR (126 MHz, CDCl} \textsubscript{3}) \]
5-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-a]pyrazin-5-yl)oxy)pentan-1-ol (Molomics 1')

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
3-(4-(Difluoromethoxy)phenyl)-5-(2-(furan-2-yl)ethoxy)-[1,2,4]triazolo[4,3-a]pyrazine (Molomics 2)

$^1$H NMR (500 MHz, CDCl$_3$)

$^{13}$C NMR (126 MHz, CDCl$_3$)
LCMS Traces for the Intermediates of Molomics 1

17a → [M+Na]^+ → 17b

18a → [M+Na]^+ → 18b
Experimental Biological Procedures

In Vitro Antiplasmodial Activity (Drug Discovery Unit, University of Dundee)

Cultures of the widely-used malaria reference strain of chloroquine-sensitive *P. falciparum* strain 3D7 were maintained in a 1.25% or 5% suspension of human red blood cells cultured in RPMI 1640 medium supplemented with 0.5% Albumax II (available from Gibco Life Technologies, San Diego, CA, cat. no. 11021-037), 12 mM sodium bicarbonate, 0.2 mM hypoxanthine, (pH 7.3), and 20 mg/L gentamicin at 37 °C, in a humified atmosphere of 1% O₂, 3% CO₂ with a gas balance of nitrogen. Growth inhibition of the *P. falciparum* cultures was quantified in a 10-point dose response curve with a 1 in 3 dilution series. This 384 well plate based fluorescence assay utilises the binding of SYBRgreen I (Thermo Fisher Scientific/Invitrogen cat. no. S7585) to double stranded DNA, which greatly increases the fluorescent signal at 528 nm after excitation at 485 nm. Mefloquine was used as a drug control to monitor the quality of the assay (Z′ = 0.6 to 0.8, where Z′ is a measure of the discrimination between the positive and negative controls on a screen plate). Dose-response curves were determined from a minimum of 3 independent experiments. Compound bioactivity was expressed as IC₅₀, the concentration of compound causing 50% inhibition. IC₅₀ values were determined from a minimum of 3 independent experiments. All data was processed using IDBS ActivityBase® raw data was converted into per cent inhibition through linear regression by setting the high inhibition control as 100% and the no inhibition control as 0%. Quality control criteria for passing plates were as follows: Z′ >0.5, S:B >3, %CV(no inhibition control) <15. The formula used to calculate

\[
Z' = 1 - \frac{3 \times (\text{StDev}_{\text{high}} + \text{StDev}_{\text{low}})}{\| (\text{Mean}_{\text{high}} - \text{Mean}_{\text{low}}) \|}.
\]

All IC₅₀ Curve fitting was undertaken using XLFit version 4.2 using Model 205 with the following 4 parametric equation:

\[
y = A + \frac{B - A}{1 + \left( \frac{C}{x} \right)^d},
\]

where
\( A = \% \) inhibition at bottom, \( B = \% \) inhibition at top, \( C = IC_{50} \), \( D = \) slope, \( x = \) inhibitor concentration and \( y = \% \) inhibition. If curve did not reach 100\% of inhibition, \( B \) was fixed to 100 only when at least 50\% of inhibition was reached.

**P. falciparum** ATP4 Assay (Lehane and Kirk Labs, Australian National University)

To measure ion concentrations inside the parasite, *P. falciparum* trophozoites were isolated from their host erythrocytes by brief exposure to saponin (0.05\% w/v final concentration), then loaded with either the Na\(^+\)-sensitive dye SBFI (for measurements of intracellular Na\(^+\) concentration)\(^{47}\) or the pH-sensitive dye BCECF (for measurements of intracellular pH)\(^{48}\). Fluorescence measurements and calibrations were performed at 37°C essentially as described previously, either in 96 well plates using a Tecan Infinite M1000 PRO plate reader, or in individual 1 mL suspensions using a PerkinElmer LS 50B fluorescence spectrometer\(^{47,49,50}\).

**Fig. S9 Mechanism of Action Experiments**
The effects of the final selections shown in Figure 4 (5 µM), DMSO (0.1% v/v; solvent control) and cipargamin (50 nM; positive control) on the internal Na\(^+\) concentration in isolated 3D7 trophozoites loaded with the Na\(^+\)-sensitive dye SBFI. The data are from a single experiment and are representative of those obtained in three similar experiments performed on different days.

**References**


