**Predicting Bioactivity when there is No Target: Performance of Methods in an Open, Crowdsourced Competition**

Edwin G. Tse1, Adele Lehane,2 Kiaran Kirk2, Davy Guan3, Ho Leung Ng4, Alice Motion1, Murray N. Robertson5, Jonathan Cardoso-Silva6, Benedict W. J. Irwin7,10, Mario Öeren7, Thomas M. Whitehead8, Gareth J. Conduit8,10, Alexander D. Wade10, Laksh Aithani11, Willem P. van Hoorn11, James McCulloch15, Vito Spadavecchio14, Raymond Lui3, Slade Matthews3, Giovanni Cincilla12, Mykola Galushka13 and Matthew H. Todd\*9

Address: 1School of Chemistry, The University of Sydney, Sydney, NSW 2006, Australia; 2Research School of Biology, Australian National University, Canberra, ACT 2601, Australia; 3School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia; 4Department of Biochemistry and Molecular Biophysics, Kansas State University, Manhattan KS 66506, USA; 5Strathclyde Institute Of Pharmacy And Biomedical Sciences, University of Strathclyde, Glasgow G4 ORE, United Kingdom; 6Department of Informatics, Faculty of Natural and Mathematical Sciences, King's College London, London WC2B 4BG, United Kingdom; 7Optibrium Ltd. Blenheim House, Denny End Road, Cambridge, United Kingdom; 8Intellegens Ltd., Eagle Labs, Chesterton Road, Cambridge CB4 3AZ, United Kingdom; 9School of Pharmacy, University College London, London WC1N 1AX, United Kingdom; 10Theory of Condensed Matter Group, Cavendish Laboratories, University of Cambridge, United Kingdom; 11Exscientia Ltd., The Schrödinger Building, Oxford Science Park, Oxford OX4 4GE, United Kingdom; 12Molomics, Barcelona Science Park, Barcelona, Spain; 13Auromind Ltd, 126 Eglantine Avenue, Belfast BT9 6EU, United Kingdom; 14Interlinked Therapeutics LLC, Portland, OR 97214, USA; 15Kellerberrin, 6 Wharf Rd, Balmain, Sydney, NSW 2041, Australia.

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 increasing reports of resistance to our frontline treatments. The Open Source Malaria (OSM) consortium have been developing compounds ("Series 4") which possess potent activity against *Plasmodium falciparum* *in vitro* and *in vivo* and have been suggested to act through the inhibition of *Pf*ATP4, an essential ion pump in the parasite membrane that regulates the concentration of intracellular Na+ and H+. This pump has not yet been crystallised, so in the absence of structural information about this target, a public competition was created to develop a model that would allow the prediction of when compounds in Series 4 are likely to be active, thereby reducing project costs associated with the unnecessary synthesis of inactives.

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, again with freely-donated models that other participants could see. The best-performing models from this second round were used to predict novel chemical matter in Series 4, 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 openly demonstrated the abilities of new machine learning methods in the prediction of active compounds where there is no confirmed biological target, 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**

*Pf*ATP4; 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 with one of two methods. In *target-based drug discovery* the molecular target of interest is known [Croston2017]. 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. From these observations, the key binding interactions may be identified and more directed structure activity relationship (SAR) studies can be conducted to optimise the small molecule.

The biological target may alternatively not be known or its structure not determined. In these cases, *phenotypic drug discovery* is undertaken [Moffat2017]. This process involves the initial identification of potent, target non-specific compounds, with target determination performed thereafter. The lead optimisation phase in this type of drug discovery is less streamlined than 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 turn out to be inactive. Yet 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 structure, computational models may be developed using artificial intelligence (AI) and machine learning (ML) [Sellwood2018], which allows for the activities of new compounds in a phenotypic program to be predicted. For instance, matched molecular pair analysis (MMPA) [Tyrchan2017] and quantitative structure activity relationship (QSAR) [Neves2018] 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 where 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 [Muhammed2018].) 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 these kinds of computational methods to aid the drug discovery process [Sieg2019,Walters2020,Stokes2020]. For instance, there have been successes in the *in silico* target prediction of small molecules with activity against *Mycobacterium tuberculosis* [Mugumbate2015,Homeyer2019].

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 [Qinghaosu1979], this natural product has 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. New medicines that possess novel MoAs must be discovered [Tse2019].

One promising biological target is the *Plasmodium falciparum* P-type Na+-ATPase (*Pf*ATP4) transporter, which is an essential part of the parasite that allows for the regulation of intracellular ions when inside a red blood cell [Kirk2015]. The structure of this membrane-bound protein remains unsolved, due to its size and complexity. Evidence for its involvement in parasite killing comes from several sources, including an ion regulation assay that implicated this target for a number of compounds from the MMV Malaria Box [Lehane2014]: In total, 28 compounds were found to exhibit ion regulation activity consistent with *Pf*ATP4 inhibition. Several other promising compounds in development are thought to inhibit the same target [Rottmann2010,Jiménez-Díaz2014], including KAE609, currently in Phase III clinical development. These compounds were found to possess a strikingly diverse range of chemotypes (Fig. 1) [Spillman2015].A homology model was developed which utilised crystal structures from the closest mammalian homolog, sarco/endoplasmic reticulum Ca2+-ATPase (SERCA) [Jiménez-Díaz2014], but without the solved structure of the protein, ideally bound to these small molecules, it is unclear how it is possible for such diverse molecules all to inhibit 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 shown to target *Pf*ATP4.**

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 [Williamson2016].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 over 200 compounds in Series 4, with *in vitro* potencies against *Plasmodium falciparum* ranging from inactive to single-digit nanomolar. 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 [invivo]. Based on preliminary investigations against *Pf*ATP4-resistant mutant strains (generated from the parent Dd2 strain by exposure to three hits from the Malaria Box against *Pf*ATP4, including KAE609 [Lehane2014]), Series 4 compounds are thought to target *Pf*ATP4 [ATP4resistance]. 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 analog). 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 the ion regulation assay. Many of these compounds were from OSM Series 4, yet there were also candidate antimalarials from other, structurally unrelated, series. It was possible to include “presumed inactives”: randomly-selected molecules from commercial catalogs 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 preferred 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, and awaited the outcome. 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 [Vamathevan2019,Chan2019]. 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. We felt we could achieve two things by running this competition with OSM’s open source ethos, in which submitters would reveal their predictions in real time and, ideally, provide full methods (within the boundaries of commercial sensibilities). 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 *Pf*ATP4 active compounds from the MMV Malaria Box [Murray1,Murray2]. By using Discovery Studio from Accelrys (now BIOVIA) to screen the 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).



**Fig. 2 Model creation workflow.** A) The four-feature pharmacophore model chosen for further development with MMV006429 mapped. B) All 28 active compounds mapped to the query. C) Shape feature added based on poses in B. D) Inactive molecules from the dataset mapped. E) Exclusion spheres added.

The 28 active compounds were mapped to the model and a shape feature was created. 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 to a set of compounds that was evaluated for ion regulation activity in 2014, 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 does not address the thermodynamics of the process.



**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 experimental ion regulation (“*Pf*ATP4”) activity, the majority of predictions did not correlate well with experiment.

This model was also used to screen [Murray2] the Maybridge library of compounds [Maybridge] to identify a small and diverse selection of molecules to evaluate in the ion regulation assay. The results were manually filtered to give a final selection of 18 compounds that were subsequently evaluated in the ion regulation assay. None were found to exhibit activity, 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 [Round1]. The competition adhered to the open science principles underpinning the OSM consortium, meaning 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 [Todd2019]. 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 *Pf*ATP4 ion regulation activity [IonRegData], 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 regulation assay. The pathogen box was at the same time screened in this assay and the data held back until the models had been submitted. A small cash prize inducement was employed to stimulate interest, despite the risk of making intrinsic reward for participation more extrinsic.[opensourcecash]

Six diverse, fully-fledged entries were submitted from individuals working in both public and private sectors, with all working shared online (Table 1) [Round1Entries]. These submissions were reviewed by a panel of four judges (Prof. Matthew Todd, Dr. 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 [VSResults].

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



|  |  |  |  |
| --- | --- | --- | --- |
| **Entrant** | **Description of Model** | **Correctly Predicted Actives** | **Result** |
| Jonathan Cardoso-Silva | Gradient boosting model (using XGBoost) to predict actives and nonactives. | **B** just outside top 20 | Runner-up |
| Giovanni Cincilla | PfATP4 Ion Regulation Activity classification model using machine learning. | **B**, **D** | Runner-up |
| Davy Guan | Semi-supervised machine learning, used to construct QSAR models. Molecules were featurised by either Graph convolutional techniques or with 1024 Bit ECFP4 descriptors. | **B**, **F** | Runner-up |
| James McCulloch | Deep Neural Network ML using a vector of the chemo-physical properties of the target molecules. | **B**, **D**, **I**  F just outside top 20 | **Winner** |
| Ho-Leung Ng | QSAR model based on homology modelling of PfATP4 -Cresset Forge. | **K**, **D**  **J** just outside top 20 | **Winner** |
| Vito Spadavecchio | Library of 'common' transformations' as seen in CHEMBL. | **B** | Runner-up |

Compounds A-K shown to be active from the MMV Pathogen Box screen against *Pf*ATP4.

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 phenotype.

**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 campaign. 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, the community would have access to an enhanced dataset, including all the data used as the test set for the previous round [Dennis2018].

The competition’s second round was launched in July 2019 [Round2]. 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 experimental values for thirty-four compounds. The precision of each model was calculated according to: , where is the number of correct predictions (active and inactive combined) and is the number of false positive predictions [Round2Results].

**Table 2: Summary of the results from Round 2 of the predictive modelling competition.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Entrant (Affiliation)** | **Description of Modela** | **Precision of Accurate Predictions (Active and Inactive)b** | **Result** |
| Jonathan Cardoso-Silva (King’s College London) | Network-based piecewise linear regression for QSAR modelling | 36% | Runner-up |
| Giovanni Cincilla (Molomics) | Logistic regression classifier model using a stochastic average gradient as solver, a uniform regularisation and a learning step size = 0.01. | 91%c | **Winner (company)** |
| Mykola Galushka (Auromind) | 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. | 58% | Runner-up |
| Davy Guan (The University of Sydney) | Automated machine learning method with 21 quantum mechanical descriptors at the Hartree Fock with 3 corrections method and JCLogP optimised for Mean Absolute Error. | 82% | **Winner (non-company)** |
| Ben Irwin/Mario Öeren/Tom Whitehead (Optibrium/Intellegens) | Deep imputation with quantum mechanical StarDrop6.6 Automodeller and pKa descriptors. | 81% | **Second place** |
| Raymond Lui (The University of Sydney) | Automated machine learning method using 59 permutation feature importance selected Mordred and quantum mechanical descriptors optimised for Mean Absolute Error | 58% | Runner-up |
| Slade Matthews (The University of Sydney) | Random forest model using 200 Mordred descriptors based on optimised 3D structures. Training RMSE = 0.805. | N.A. | Runner-up |
| Ho-Leung Ng (Kansas State University) | QSAR model based on detailed homology modeling of PfATP4 and docking. 3D features are combined with 1D/2D QSAR features using XGBoost (gradient boosted trees) to make a regression model. | 71% | Runner-up |
| Vito Spadavecchio (Interlinked TX) | Ensemble classification (logistic regression) and regression (MLP) using ECFP4 (Morgan radius 2) | 79%c | Runner-up |
| Laksh Aithani/Willem van Hoorn (Exscientia) | Ridge regression model with alpha = 1. ECFP4 fingerprints with (Morgan radius 2) were the input to the model. | 81% | **Second place** |

aSee SI for full experimental details. bBased on regression prediction. cBased on classification prediction.

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 [ICSYNTH,Nicolaou2020]. The initial list was narrowed to focus on four predicted triazolopyrazine compounds (Fig. 4). The four compounds were successfully synthesised and subsequently evaluated for *in vitro* activity against *P. falciparum*.



**Fig. 4 Examples of the suggested compounds predicted by the winning entrants from Round 2 and the four chosen for experimental validation.** The predictions were synthesised (see SI) and their potencies experimentally validated. Only one compound was found to be active (*Note here on data incoming from Kiaran Kirk confirming that all these still have the same phenotype*).

One of these four compounds was found to be active, a hit rate of 25% on a small sample size. To date a total of 398 compounds have been made and evaluated for *in vitro* activity in OSM Series 4. The design of these compounds has been driven entirely by the intuition of medicinal chemists. By setting a potency cut-off of 0.647 µM (the activity level of the predicted active), the tally of active compounds is 96, representing a comparable human hit rate of 24% on a larger sample size.

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 two predictions involving the shorter linker 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 compound was thought by the human team to be likely to be potent, but performed poorly (i.e. the “that’s odd” class). Lastly the 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 be used as an open demonstration of capabilities[PR]. 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 [Round2].

**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. With 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**

[Williamson2016] Williamson, A. E. *et al.* Open Source Drug Discovery: Highly Potent Antimalarial Compounds Derived from the Tres Cantos Arylpyrroles. *ACS Cent. Sci.* **2**, 687-701 (2016).

[invivo] In Vivo Efficacy (2018). Available at: <https://github.com/OpenSourceMalaria/Series4/wiki/In-Vivo-Efficacy>. (Accessed: 26th September 2019).

[Todd2019] Todd, M. H. Six Laws of Open Source Drug Discovery. *ChemMedChem* **14**, 1804–1809 (2019).

[Lehane2014] Lehane, A. M., Ridgway, M. C., Baker, E. & Kirk, K*.* Diverse Chemotypes Disrupt Ion Homeostasis in the Malaria Parasite. *Mol. Microbiol.* **94**, 327–339 (2014).

[Kirk2015] Kirk, K. Ion Regulation in the Malaria Parasite. *Annu. Rev. Microbiol.* **69**, 341–359 (2015).

[Spillman2015] Spillman, N. J. & Kirk, K. The Malaria Parasite Cation ATPase PfATP4 and its Role in the Mechanism of Action of a New Arsenal of Antimalarial Drugs. *Int. J. Parasitol. Drugs Drug Resist.* **5**, 149–162 (2015).

[Rottmann2010] Rottmann, M. *et al.* Spiroindolones, a Potent Compound Class for the Treatment of Malaria. *Science* **329**, 1175–1180 (2010).

[Jiménez-Díaz2014] Jiménez-Díaz, M. B. *et al.* (+)-SJ733, a Clinical Candidate for Malaria that acts through ATP4 to Induce Rapid Host-Mediated Clearance of *Plasmodium*. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5455–E5462 (2014).

[Dennis2018] Dennis, A. S. M. *et al.* Diverse Antimalarials from Whole-Cell Phenotypic Screens Disrupt Malaria Parasite Ion and Volume Homeostasis. *Sci. Rep.* **8**, No. 8795 (2018).

[Sellwood2018] Sellwood, M. A., Ahmed, M., Segler, M. H. S. & Brown, N. *Future Med. Chem.* **10**, 2025–2028 (2018).

[Sieg2019] Sieg, J., Flachsenberg, F. & Rarey, M. In Need of Bias Control: Evaluating Chemica Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model* **59**, 947–961 (2019).

[Walters2020] Walters, W. P. & Murcko, M. Assessing the Impact of Generative AI on Medicinal Chemistry. *Nat. Biotechnol.* **38**, 143–145 (2020).

[Stokes2020] Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688–702 (2020).

[Qinghaosu1979] Qinghaosu Antimalaria Coordinating Research Group. Antimalaria Studies on Qinghaosu. *Chin. Med. J. (Engl.)* **92**, 811–816 (1979).

[Tse2019] Tse, E. G., Korsik, M. & Todd, M. H. The Past, Present and Future of Anti-Malarial Medicines. *Malar. J.* **18**, No. 93 (2019).

[Tyrchan2017] Tyrchan, C. & Evertsson, E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. *Comput. Struct. Biotechnol. J.* **15**, 86–90 (2017).

[Neves2018] Neves, B. J. *et al.* QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Front. Pharmacol.* **9**, No. 1275 (2018).

[Muhammed2018] Muhammed, M. T. & Aki-Yalcin, E. Homology Modelling in Drug Discovery: Overview, Current Applications, and Future Perspectives. *Chem. Biol. Drug Des.* **93**, 12–20 (2018).

[Croston2017] Croston, G. E. The Utility of Target-Based Discovery. *Expert Opin. Drug Discov.* **12**, 427–429 (2017).

[Moffat2017] Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov.* **16**, 531–543 (2017).

[Vamathevan2019] Vamathevan, J. *et al.* Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).

[Chan2019] Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H. & Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **40**, 592–604 (2019).

[Murray1] Pharmacophore Modelling of the Malaria Box PfATP4 Active Compounds (2015). Available at: <http://malaria.ourexperiment.org/pharmacophore_modelling_/7971/post.html>. (Accessed: 26th September 2019).

[Murray2] Using the Pharmacophore Model to search Commercial Compounds for new leads (2015). Available at: <http://malaria.ourexperiment.org/pharmacophore_modelling_/12498/post.html>. (Accessed: 26th September 2019).

[Maybridge] Maybridge Screening Collection. Available at: <https://www.maybridge.com/portal/alias__Rainbow/lang__en/tabID__146/DesktopDefault.aspx>. (Accessed: 26th September 2019).

[Round1] COMPETITION: A Predictive Model for Series Four (2016). Available at: <https://github.com/OpenSourceMalaria/OSM_To_Do_List/issues/421>. (Accessed: 26th September 2019).

[IonRegData] Ion Regulation Data for OSM Competition. Available at: <https://docs.google.com/spreadsheets/d/1WWP8fE3X2BLzZ7jOm6bRWpnHZqVJBf8XgIYEb7hshXU/edit?usp=sharing>. (Accessed: 26th September 2019).

[opensourcecash] Hars, Alexander, and Shaosong Ou. “Working for Free? Motivations for Participating in Open-Source Projects.” International Journal of Electronic Commerce, vol. 6, no. 3, 2002, pp. 25–39. JSTOR, www.jstor.org/stable/27751021. Accessed 18 June 2020.

[Round1Entries] Summary of Competition Entries (2017). Available at: <https://docs.google.com/spreadsheets/d/1pY6sYXIw66jnzUO3CoP8HceYdDjLRvwg5_pLkBY1Wek/edit?usp=sharing>. (Accessed: 26th September 2019).

[VSResults] VS Results. Available at: <https://docs.google.com/spreadsheets/d/1uaQ_mSVY6vQSbnDHD5gf232ySH-KVjUg--dQCsilGWI/edit?usp=sharing>. (Accessed: 26th September 2019).

[Round2] COMPETITION ROUND 2: A Predictive Model for Series 4 (2019). Available at: <https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/1>. (Accessed: 26th September 2019).

[Round2Results] Round 2 Results. Available at: <https://docs.google.com/spreadsheets/d/1ZPJlMM7znFa056yjOuIycLgzV7-ENt1mMSUJKgH9TxY/edit?usp=sharing>. (Accessed 12th September 2019).

[Mugumbate2015] Mugumbate, G. *et al.* Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and *In Vitro* Validation. *PLoS ONE* **10**, e0121492 (2015).

[Homeyer2019] Homeyer, N. *et al.* A Platform for Target Prediction of Phenotypic Screening Hit Molecules. *J. Mol. Graph. Model.* **95**, 107485 (2020).

[ATP4resistance] Evaluation of Series 4 Compounds vs ATP4-Resistant Mutants (2015). Available at: <http://malaria.ourexperiment.org/biological_data/11448/post.html>. (Accessed: 14th October 2019).

[ICSYNTH] InfoChem IC*SYNTH*. Available at: <https://www.infochem.de/synthesis/ic-synth>. (Accessed: 21st April 2020).

[Nicolaou2020] Nicolaou, C. A., Watson, I., LeMasters, M. A., Masquelin, T. & Wang, J. Context Aware Data-Driven Retrosynthetic Analysis. *J. Chem. Inf. Model.* Just Accepted (2020).

[PR] Data Finders: How can artificial intelligence and deep-learning platforms be used to combat malaria? Medicine Maker, May 15th 2020 <https://themedicinemaker.com/discovery-development/data-finders>. (Accessed June 18th 2020)

**Acknowledgements**

The authors thank the EPSRC Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery Network (AI3SD-FundingCall1\_029), the Australian Research Council (LP150101226), the Medicines for Malaria Venture (RD/11/0040) and University College London for funding (to M.H.T.), the US National Science Foundation (1833181) (to H.L.N.).

**Author contributions**

EGT, AM, MNR carried out the synthetic chemistry. AL and KK carried out the ion regulation assay and advised on the results. MNR, DG, HLN, JC-S, BWJI, MO, TMW, GJC, ADW, LA, WPvH, JM, VS, RL, SM, GC and MG contributed models to the competitions and collaborated on the outputs during the competitions. MHT founded OSM, conceived the project and secured 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.