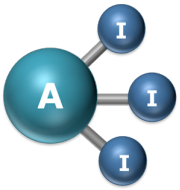

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Predicting the Activity of Drug Candidates when There is no Target
Final Report
Project Dates: 01/07/2019 - 28/02/2020 (plus NCE)
UCL

Professor Matthew H. Todd
UCL

Report Date: 15/03/2021

Predicting the Activity of Drug Candidates when There is no Target
AI3SD-Project-Series:Report-3_Todd_Final
Report Date: 15/03/2021
DOI: 10.5258/SOTON/P0042

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Predicting the Activity of Drug Candidates when There is no Target
Funding reference	AI3SD-FundingCall1_029
Lead Institution	University of UCL
Project Dates	01/07/2019 - 28/02/2020 (plus NCE)
Website	https://github.com/OpenSourceMalaria/Series4_PredictiveModel
Keywords	Drug Discovery, Malaria, Open Science

2 Project Team

2.1 Principal Investigator

Name and Title: Professor Matthew Todd
Association: University College London, School of Pharmacy
Work Email: matthew.todd@ucl.ac.uk
Website Link: https://todd-lers.github.io/about/

2.2 Project Partners

Name and Title: Dr Mykola Galushka
Association: Auromind Ltd
Work Email: mm.galushka@auromind.org
Website Link: https://www.auromind.org/

Name and Title: Dr Willem van Hoorn
Association: Exscientia Ltd
Work Email: wvanhoorn@exscientia.co.uk
Website Link: https://www.exscientia.ai/

Name and Title: Dr Tom Whitehead
Association: Intellegens Ltd
Work Email: tom@intellegens.ai
Website Link: https://www.intellegens.co.uk/

2.3 Researchers & Collaborators

Entrants (besides the co-investigators): Benedict Irwin (Optibrium), Vito Spadaveccio (independent), Ho Leung Ng (Kansas State University), Jon Cardoso (independent), Giovanni Cincilla (Molomics), Davy Guan (PhD student, Sydney Uni), Raymond Lui (research assistant), Slade Matthews (PhD student, Sydney University).

Other contributors: Girinath G. Pillai, Jacob Silterra.

Edwin Tse, completing PhD student (University of Sydney, visitor at UCL), has acted as the paid coordinator of the competition and will carry out the synthesis of all predicted molecules.

New contributors, since the close of the competition, include Evariste Technologies, and Jan Jensen (University of Copenhagen).

3 Publicity Summary

This project aims to harness artificial intelligence and machine learning approaches to improve the discovery of new medicines.

One of the most common situations in drug discovery is to know a molecule that possesses a desirable property and yet not to know how the molecule achieves that effect. The molecule needs improvement (often for things like how well it does its job, or how soluble it is in water) for it to be a realistic drug candidate and we must consider what changes need to be made. In this AI3SD project we know of molecules that efficiently kill the malaria parasite. The molecules need to be improved via small changes to their structures, yet we do not know the molecular biological target of these molecules. Without knowledge of the target it is impossible to design the improvements rationally. In such a project (known as "phenotypic drug discovery") it is typically the case that the scientist will apply rules of thumb and intuition acquired over many related projects, in order to alter the structure of the molecule in search of those improvements.

Yet it is also frequently the case that at such a stage of the project changes can be made to the molecule that accidentally obliterate the desired properties; most typically the molecules lose their potency against the pathogen. We will then have wasted time and resources making inactive molecules. On average each "fail" costs about two thousand pounds to make. It would be much more efficient if we could become more accurately predictive about which molecules need to be made.

In our research consortium, Open Source Malaria (OSM), we have twice tried and failed to generate predictive models. There have occurred, in the period since, major new advances in AI and ML, particularly in the private sector. Since all of OSM's data and ideas are freely shared in the public domain, it is possible for us to work with anyone in the generation of new models. We have therefore in this project used the AI3SD funds to run a predictive modelling competition and elicited contributions from amateurs and leading AI companies. Several models were better than the others and so, in a crucial part of this project, we asked those winners to predict new molecules, molecules that have never existed before, that they predict will be effective at killing the malaria parasite. We then went to the lab to validate these predictions by making the molecules suggested, and measuring how effective they are at killing the parasite in blood. The result of this was that three of the six predictions were active, a "hit rate" of about the same as the human hit rate across the rest of the project. Interestingly, these actives included a couple of molecules that the human chemists would probably not have tried.

The end result of this work, aside from a new and predictive approach to the synthesis of antimalarial drug candidates, is be a case study of the actual capabilities of new AI/ML technologies in drug discovery: what works, what does not and an examination of why. Notably the project is still ongoing: since all the data and details of the approaches taken are in the public domain, others can try out their own predictive algorithms to see if they can do better.

4 Executive Summary

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 PfATP4, an essential ion pump in the parasite membrane that regulates intracellular Na^+ and H^+ concentrations. 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 us to predict when compounds in Series 4 are likely to be active.

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 performed, with 10 models submitted. The best-performing models from this second round were used to predict novel analogs in Series 4 that were synthesised and evaluated against the parasite. Several new active molecules were found via the disparate approaches used. This clearly demonstrates the potential for AI/ML methods to accelerate progress in phenotypic drug discovery projects where the molecular details of the biological target are not known.

The project continues, since all data and previous attempts are in the public domain, and indeed two new entrants have decided to predict further structures for synthesis. Experimental validation of AI/ML methods is likely to be a centrally important feature in this field in the coming years.

5 Aims and Objectives

We aimed to develop a general AI-enabled approach to solving the prediction of biological activity in phenotypic drug discovery, through the use of a public competition.
This was achieved via the following objectives:

1. Data curation and sharing of everything that is known of the activity of molecules in a particular research project, in this case OSM Series 4.
2. Invitation to our co-applicant partners to submit models, as well as an invitation for the public to contribute.
3. Evaluation of the predictive models, including via synthesis of several of the predicted compounds. Winners were determined and modest prizes offered (but declined).
4. Post-mortem collaborative discussion of the outcomes at a physical workshop.
5. Publication of a summary article. Description of how the project can be continued.

The two over-arching motivations are:

1. To improve drug discovery
2. To provide a public domain case study of the real capabilities of AI/ML in drug discovery

6 Methodology

6.1 Scientific Methodology

Open science methods
Data shared using Google sheets. Collaboration enabled using Github (discussion and file sharing). Recruitment of new participants via social media outreach (Twitter, LinkedIn) and traditional cold-call emails.

Chemistry/biology methods
The molecules proposed were synthesised in the laboratory at UCL using (largely) methods that have been discovered in the OSM consortium prior to this project. Potency evaluation was measured in a standard blood stage assay performed regularly at the Drug Discovery Unit in the University of Dundee. Verification that the active molecules operate via inhibition of PfATP4, the suspected target, was performed pro bono in the laboratory of Professor Kiaran Kirk (ANU, Canberra) using published methods.

6.2 AI Methodology

The methods employed to generate a predictive model for the series depended on the contributors and were not be defined prior. The means of analysis and comparison of the data were also to be devised upon receipt of the entries, since the form of the predictions would also be kept loosely defined.
The methods actually used were:

Entrant (Affiliation)	Description of Model[a]	Precision of Accurate Predictions (Active and Inactive)[b]	Result
Jonathan Cardoso-Silva (King's College London)	Network-based piecewise linear regression for QSAR modelling.[43]	**36%**	Runner-up
Giovanni Cincilla (Molomics)	P. falciparum inhibition classification model using: CDK descriptors,[40] ECFC4 fingerprints and logistic regression (with: stochastic average gradient as solver, uniform regularisation and 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 using the Hartree Fock with 3 corrections method[44] and JCLogP, optimised for Mean Absolute Error.	**82%**	**Winner (non-company)**
Ben Irwin, Mario Öeren, Tom Whitehead (Optibrium/ Intellegens)	Deep imputation[45,46,47] with quantum mechanical StarDrop6.6 Automodeller and pKa descriptors.[48]	**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**

[a] See paper for full details. [b] Based on regression prediction. [c] Based on classification prediction

7 Results

Dataset was provided and improved by contributors:
(https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/4)

Rules were defined and timescales set:
(https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/1).

Ten entries were submitted by a mixture of amateur and professional teams (https://github.com/OpenSourceMalaria/Series4_PredictiveModel/tree/master/Submitted%20Models).

These models were evaluated against data that had been kept back, with winners announced following judgement by a panel of four people. Of note is that the models generated possess a much higher degree of predictive ability than previous rounds of this competition. There has been discussion of the relative merits of different means of evaluating the entries (https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/18).

The winners were tasked with prediction of two new chemical structures, with aqueous solubility decided as a guiding criterion
(https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/19).

Molecule synthesis was undertaken in the laboratory (impacted by COVID).

Meeting was held to examine project outcomes (Jan 2020)
(https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/21).

Paper was written up (https://chemrxiv.org/articles/preprint/13194755) and is undergoing revision for publication in the *Journal of Medicinal Chemistry*.

Project has successfully secured new inputs from the private sector after the competition ended (https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/29)

8 Outputs

Paper was written openly: https://docs.google.com/document/d/1aD29GjC8RjqrSDcWcEUptS04Z2v10deReRp0eB3kcp4/edit?usp=sharing

And posted to a preprint server before submission to the Journal of Medicinal Chemistry.

Update talk at the AI3SD meeting, Winchester, Nov 2019.
Interview: https://eprints.soton.ac.uk/443554/

Talk at AI3SD Winter Webinar Series, Dec 2nd 2020:
https://www.ai3sd.org/ai3sd-event/02-12-2020-ai3sd-winter-seminar-series-robots-ai-and-nlp-in-drug-discovery/

Github repository of all activity:
https://github.com/OpenSourceMalaria/Series4_PredictiveModel

Curated biological dataset refined by entrant:
https://docs.google.com/spreadsheets/d/1Cgpx5aF-

HcJuE3jWjDRh_hhaAcVjzbskpdS1Y9x0wwU/edit#gid=952560208

Meeting held at the Royal Society of Chemistry Jan 2020, report: https://eprints.soton.ac.uk/438123/

Medicine Maker article about the project
https://themedicinemaker.com/discovery-development/data-finders

Use of the "winning" of the competition to advertise success, which may have contributed to the company securing new investment: https://www.kestercapital.com/all-news/kester-capital-invests-in-optibrium.html

Grant application to extend the work was submitted by one of the entrants, to continue to work with Open Source Malaria, unfortunately unsuccessful (ca 4% success rate).

9 Conclusions

With hit identification and lead optimization 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 synthesized 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.

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 this endeavor has already been used as an unvarnished demonstration of capabilities (https://themedicinemaker.com/discovery-development/data-finders)
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 were less important than the extent to which entrants worked together openly to improve the underlying research.—

10 Future Plans

1. Synthesise new molecule variants to attempt to find improved potency.
2. Continue to seek new entrants attracted to the project by virtue of its openness.

3. Apply for new funding with AI/ML companies interested in experimental validation of approaches, or work with such companies directly. Models can be developed and applied to malaria, or another neglected/tropical infectious disease, yet be used to build a company's bottom line when applied to other areas.

11 References

These may be found at the end of the preprint for this project (https://chemrxiv.org/articles/preprint/13194755), and a selection follows:

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

[2] 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.* **2017**, *16*, 531–543.

[3] Sellwood, M. A.; Ahmed, M.; Segler, M. H. S.; Brown, N. Artifical Intelligence in Drug Discovery. *Future Med. Chem.* **2018**, *10*, 2025–2028.

[4] Griffen, E. J.; Dossetter, A. G.; Leach, A. G. Chemists: AI is Here; Unite to Get the Benefits. *J. Med. Chem.* **2020**, *In press.*

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

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

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

[8] 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.* **2019**, *59*, 947–961.

[9] Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.

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

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

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

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

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

[15] Hamilton, W. L. *et al.* Evolution and Expansion of Multidrug-Resistant Malaria in Southeast Asia: A Genomic Epidemiology Study. *Lancet Infect. Dis.* **2019**, *19*, 943–951.

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

[17] Spillman, N. J. *et al.* Na^+ Regulation in the Malaria Parasite *Plasmodium falciparum* Involves the Cation ATPase PfATP4 and Is a Target of the Spiroindolone Antimalarials. *Cell Host Microbe* **2013**, *13*, 227–237.

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

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

[20] 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.* **2014**, 111, E5455–E5462.

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

[22] Dennis, A. S. M.; Rosling, J. E. O.; Lehane, A. M.; Kirk, K. Diverse Antimalarials from Whole-Cell Phenotypic Screens Disrupt Malaria Parasite Ion and Volume Homeostasis. *Sci. Rep.* **2018**, *8*, No. 8795.

[23] 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.* **2015**, *5*, 149–162.

[24] Vaidya, A. B. *et al.* Pyrazoleamide Compounds are Potent Antimalarials that Target Na^+ Homeostasis in Intraerythrocytic *Plasmodium falciparum*. *Nat. Commun.* **2014**, *5*, No. 5521.

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

[26] In Vivo Efficacy. https://github.com/OpenSourceMalaria/Series4/wiki/In-Vivo-Efficacy (accessed Sept 26, 2019).

[27] Evaluation of Series 4 Compounds vs ATP4-Resistant Mutants. http://malaria.ourexperiment.org/biological_data/11448/post.html (accessed Oct 14, 2019).

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

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

[30] Bentzien, J.; Muegge, I.; Hamner, B.; Thompson, D. C. Crowd Computing: Using Competitive Dynamics to Develop and Refine Highly Predictive Models. *Drug Discov. Today* **2013**, *18*, 472–478.

[31] Pharmacophore Modelling of the Malaria Box PfATP4 Active Compounds. http://malaria.ourexperiment.org/pharmacophore_modelling_/7971/post.html (accessed Sept 26, 2019).

[32] Using the Pharmacophore Model to search Commercial Compounds for new leads. http://malaria.ourexperiment.org/pharmacophore_modelling_/12498/post.html (accessed: Sept 26, 2019).

[33] Maybridge Screening Collection. https://www.maybridge.com/portal/alias__Rainbow/lang__en/tabID__146/DesktopDefault.aspx (accessed Sept 26, 2019).

[34] COMPETITION: A Predictive Model for Series Four. https://github.com/OpenSourceMalaria/OSM_To_Do_List/issues/421 (accessed Sept 26, 2019).

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

[36] Ion Regulation Data for OSM Competition. https://docs.google.com/spreadsheets/d/1WWP8fE3X2BLzZ7jOm6bRWpnHZqVJBf8XgIYEb7hshXU/edit?usp=sharing (accessed Sept 26, 2019).

[37] Hars, A.; Shaosong O. "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 Jun 18, 2020.

[38] Summary of Competition Entries. https://docs.google.com/spreadsheets/d/1pY6sYXIw66jnzUO3CoP8HceYdDjLRvwg5_pLkBY1Wek/edit?usp=sharing (accessed Sept 26, 2019).

[39] VS Results. https://docs.google.com/spreadsheets/d/1uaQ_mSVY6vQSbnDHD5gf232ySH-KVjUg--dQCsilGWI/edit?usp=sharing (accessed Sept 26, 2019).

[40] Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

[41] COMPETITION ROUND 2: A Predictive Model for Series 4. https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/1 (accessed Sept 26, 2019).

[42] Round 2 Results. https://docs.google.com/spreadsheets/d/1ZPJlMM7znFa056yjOuIycLgzV7-ENt1mMSUJKgH9TxY/edit?usp=sharing (accessed Sept 12, 2019).

[43] Cardoso-Silva, J.; Papageorgiou, L. G.; Tsoka, S. Network-Based Piecewise Linear Regression for QSAR Modelling. *J. Comput. Aided Mol. Des.* **2019**, *33*, 831–844.

[44] Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

[45] Whitehead, T. M.; Irwin, B. W. J.; Hunt, P.; Segall, M. D.; Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1197–1204.

[46] Irwin, B. W. J.; Levell, J. R.; Whitehead, T. M.; Segall, M. D.; Conduit, G. J. Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data. *J. Chem. Inf. Model.* **2020**, *60*, 2848–2857.

[47] Irwin, B. W. J.; Mahmoud, S.; Whitehead, T. M.; Conduit, G. J.; Segall, M. D. Imputation versus Prediction: Applications in Machine Learning for Drug Discovery. *Future Drug. Discov.* **2020**, *2*, No. 2.

[48] Hunt, P. A. *et al.* Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods. *J. Chem. Inf. Model.* **2020**, *Just Accepted.*

[49] InfoChem ICSYNTH. https://www.infochem.de/synthesis/ic-synth (accessed Apr 21, 2020).

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

[51] Tse, E. G. *et al.* Non-Classical Phenyl Bioisosteres as Effective Replacements in a Series of Novel Open Source Antimalarials. *J. Med. Chem.* **2020**, *63*, 11585–11601.

12 Data & Software Links

Dataset: https://docs.google.com/spreadsheets/d/1Cgpx5aF-HcJuE3jWjDRh_hhaAcVjzbskpdS1Y9x0wwU/edit#gid=952560208

Models: https://github.com/OpenSourceMalaria/Series4_PredictiveModel/tree/master/Submitted%20Models

Repository: https://github.com/OpenSourceMalaria/Series4_PredictiveModel