Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Predicting the Activity of Drug Candidates when There is no Target

Interim Report
Project Dates: 01/07/2019 - 28/02/2020
UCL

Professor Matthew H. Todd
UCL

Report Date: 19/11/2019

AI3SD-Project-Series:Report-3_Todd_Interim

Predicting the Activity of Drug Candidates when There is no Target
AI3SD-Project-Series:Report-3_Todd_Interim
Report Date: 19/11/2019
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
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://github.com/OpenSourceMalaria/Series4_PredictiveModel

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 Other 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.

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 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 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 intervening period 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 are asking those winners to predict new molecules, molecules that have never existed before, that will be effective at killing the malaria parasite. The final part of our project is to validate these predictions: to make the molecules in the lab, and measure how effective they are at killing the parasite in blood.

The end result of this work, aside from a new and predictive approach to the synthesis of antimalarial drug candidates, will 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.

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 are being used to predict novel analogs in Series 4 that will be synthesised and evaluated against the parasite. As such the project will openly demonstrate the abilities of new machine learning algorithms in the prediction of active compounds where there is no confirmed target, frequently the central problem in phenotypic drug discovery.

5 Aims and Objectives

We aim 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 will be 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 are determined and modest prizes awarded.
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 would be synthesised in the laboratory at UCL using (largely) methods that have been discovered in the OSM consortium prior to this project. Potency evaluation would be measured in a standard blood stage assay performed regularly at the Drug Discovery Unit in the University of Dundee.

6.2 AI Methodology

The methods employed to generate a predictive model for the series would depend on the contributors and would 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 are being assembled here: https://docs.google.com/document/d/1aD29GjC8RjqrSDcWcEUptS04Z2v10deReRp0eB3kcp4/edit?usp=sharing

7 Interim 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 have been 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 is now underway in the laboratory.
- Meeting is being planned to complete the project (https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/21).

8 Outputs

- Draft paper being written: https://docs.google.com/document/d/1aD29GjC8RjqrSDcWcEUptS04Z2v10deReRp0eB3kcp4/edit?usp=sharing
- Update talk at the AI3SD meeting, Winchester, Nov 2019. 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

9 Progress Summary

The first part of the project is complete. We successfully ran the competition, securing inputs from both amateurs and professionals working in the AI/ML field. These models have already been used (by the entrants) to create suggestions for new molecules to synthesise that are predicted to be active. The entries were more numerous than previous rounds of this exercise and are, it seems, quite predictive. We have successfully elicited useful inputs from project outsiders throughout. The predictions are being made in the laboratory and will be evaluated, all being well, before Christmas so that we have the data in hand for the post mortem meeting at the end of January.

10 Next Steps

Once the deadline has been hit for the suggestion of new chemical entities, the target list for the molecules will be complete. Synthesis will continue until mid-Dec 2019, then the compounds will be shipped to Dundee for evaluation.

With the data in hand, we will be able to perform the project post mortem – why the models were successful, or not, and which models performed best. This will take place in the meeting in London at the end of January 2020.

The project paper will be completed with the aid of all entrants, and will be submitted for publication e.g. to a Special Issue of the Beilstein Journal of Organic Chemistry as part of an open data in drug discovery special issue, with a submission to a preprint server prior to peer review, in February 2020.

11 References

No specific references, but the current reference list of relevant work for the paper may be found at the end of that paper: https://docs.google.com/document/d/1aD29GjC8RjqrSDcWcEUptS04Z2v10deReRp0eB3kcp4/edit?usp=sharing.

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
- **Repo:** https://github.com/OpenSourceMalaria/Series4_PredictiveModel