Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

AI3SD Interview with Professor John Overington
22/06/2020
Cheshire, U.K.

Dr. Wendy A. Warr
Wendy Warr & Associates

26/06/2020

AI3SD-Interview-Series:Report-2

AI3SD Interview with Professor John Overington
AI3SD-Interview-Series:Report-2
26/06/2020
DOI: 10.5258/SOTON/P0023
Published by University of Southampton

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 Interview Details

Title	AI3SD Interview with Professor John Overington
Interviewer	WAW: Dr Wendy Warr - Wendy Warr & Associates
Interviewee	JPO: Professor John Overington - Medicine Discovery Catapult
Interview Location	Cheshire, U.K.
Interview Date	22/06/2020

2 Biography

Figure 1: Professor John Overington

John Overington holds a first degree in chemistry and studied for his PhD on comparative protein modelling, drug design, and sequence-structure relationships with Sir Tom Blundell at Birkbeck College, London. After a postdoctoral fellowship with the Imperial Cancer Research Fund, London, he joined Pfizer Central Research in the United Kingdom and eventually became manager of the Molecular Informatics, Structure and Design Department, where he was responsible for cheminformatics, structural biology, target analysis, and molecular modelling. He joined Inpharmatica in 2000; Inpharmatica was acquired by Galapagos NV in December 2006. There he was Senior Director, Discovery Informatics of the Galapagos subsidiary BioFocus DPI. In October 2008 he was appointed Team Leader, Chemogenomics at the European Molecular Biology Laboratory (EMBL), working at the European Bioinformatics Institute (EMBL-EBI) in Hinxton, U.K. where he established and ran the ChEMBL group. Following that, he spent two years as VP Biomedical Informatics at BenevolentAI. He became Chief Informatics Officer at Medicines Discovery Catapult (MDC) in 2017.

3 Interview

WAW: I interviewed you for the Journal of Computer-Aided Molecular Design[1] in December 2008. You were at EMBL-EBI then. Where has your cheminformatics journey taken you since then?

JPO: I enjoyed building ChEMBL but there were limited opportunities to apply ChEMBL directly in drug discovery at EMBL-EBI due to the remit and focus of the Institute. It was a little frustrating in a way to see ChEMBL transform cheminformatics AI, as it's clearly done, from a distance. An interesting opportunity came up at Stratified Medical (now BenevolentAI) and I was excited by all of the progress and potential in AI and deep learning, so I became VP Biomedical Informatics in May 2015. It was a great experience learning about AI, and fun to be working on drug discovery again, but after two years there, I felt the need to have a role that was more outward facing and allowed collaborations and use of my experience with the U.K. SME sector.

So I then became Chief Informatics Officer at MDC. MDC is a non-profit, highly collaborative organisation, and is highly U.K.-focused, so life is now easier for me from the travel point of view! Now I have a job that is really fun, developing a variety of software and data solutions to support the life sciences sector. I have swapped a modest semi in London for a large country cottage in rural Cheshire with three and a half acres of gardens and woodlands. I am really enjoying the change in quality of life too. It's made me reflect on the large concentration of drug discovery activity in the London-Oxford-Cambridge triangle, but there are actually lots things happening elsewhere, and the potential of future investment in other regions I think will untap this latent opportunity. We should definitely have a more regionally distributed and robust sector than we do. I had reservations at first about recruitment, but it's gone well, and we recruit from a completely different demographic than if we were trying to recruit in the South East, where extra-sector competition from, for example, Amazon or Google distorts the market for life science AI and data-science researchers. To balance this, I'm only just over an hour and twenty minutes from London Euston on the train.

WAW: Tell us a bit about MDC and how it differs from the organisations you have worked in previously.

JPO: MDC is different because it is focused on translational interactions with SMEs and academia. In the informatics area for example, GSK and AstraZeneca, and many other large companies, will have the scale for large informatics and data groups, but smaller companies cannot invest in the full breadth of informatics techniques that are needed, so rely more on partnership and collaboration. It is really enjoyable to work with smaller companies and complement what they themselves are skilled at. They want to work with you; they want to be helped, and there's essentially none of the "Not Invented Here" attitude in our interactions. Another important role for us is to complement and support the U.K. SMEs, not directly compete, so this has led to a highly differentiated project portfolio, which is great to work on.

WAW: How has the field of cheminformatics changed?

JPO: Going back 10-20 years, there was very little in the way of public data. Nowadays, precompetitive, open science attitudes, open source software, and open data have emerged. In the past, tools were proprietary and expensive; now CDK[2,3] and RDKit[4] have come along and they provide a balance to some great, but essentially expensive commercial cheminformatics packages. Workflow tools such as KNIME[5] are available. TensorFlow[6] and PyTorch[7] are open source machine learning platforms. PubChem and ChEMBL are significant open data sources, alongside a large number of biomedical bioinformatics resources. In the past, small groups also did not have enough data or computing power to see these various factors come together, and essentially wipe out the barrier to entry for researchers. This has really transformed the level of innovation.

WAW: How do you see the importance of chemical information in the wider discipline of chemistry?

JPO: We are starting to see some really diverse and nice cheminformatics applications around synthesis planning, and then they are linked to robots and automation synthesis techniques. I am focused on drug discovery, but exciting computational work is going on in materials chemistry too, for example, in optoelectronics and semiconductors. The boundary between bioinformatics and cheminformatics is also blurring and that is really exciting too: fundamentally we need to consider the interactions of a cheminformatics object (a drug) with a bioinformatics object (a protein target). Another new technology is DNA-encoded libraries (DELs): this is a very powerful approach to exploration of chemical space and identification of novel leads.

WAW: What impact have machine learning and computational chemistry had on drug discovery?

JPO: Few things happen now in drug discovery without computational chemistry involved somewhere. Property prediction and multiparameter optimisation are “incremental” concepts that are now centrally adopted. But the area of AI, like many different technology areas before, is full of hype. This hype causes consideration of how technologies could impact science and business but also can lead to herd- type adoption of the fear of missing out, and inevitable bubbles in supply of skilled practitioners. Another historical example of this for me would be structure-based drug design, which has arguably progressed from a sole method, stand-alone solution to drug discovery to an integrated approach. The same is happening with AI and machine learning in drug discovery, but I don’t think there will be a distinct tipping point: as before, the best innovations will be integrated into the processes that already exist. Generative adversarial networks in computational chemistry are particularly fascinating when coupled to automation of synthesis. For me, this is a great future opportunity for U.K. investment and retention of value and jobs in the United Kingdom.

WAW: How has COVID-19 affected your work?

JPO: Working at home is very different for me (I like to “walk the beat” and talk to people) but it has gone better than I thought it would. You don’t need a physical lab to do informatics but I do miss seeing people: accidental meetings, left-field conversations, and serendipity matter. I think there has been a very big impact on the conference sector. Presenters of talks in virtual conferences don’t stay logged on and accessible for six or so hours: they just do their own talk then disappear. I feel bad that I do this. Laboratories at companies around the United Kingdom are now returning to work as normal, but I don’t know when office-based roles in general will return. Of course, one of the UK Lighthouse Labs, managed by MDC, is carrying out national COVID-19 testing at Alderley Park. It has been great to see the agility and effort that was involved in setting up a significant component of the U.K. response to Covid-19. Longer term, I’m reflecting on the future of flexible working. I think it has come to stay in some aspects, and it will be a better way to work for many jobs and people’s individual family responsibilities. It looks as if air travel will be significantly more expensive in the future, and it will be interesting to see what happens with conferences like the ACS National Meetings, and great networking and research conferences such as the Gordon and Keystone series.

WAW: Are international standards important for data and programs in machine learning and chemistry?

JPO: Clearly standards are important. There are credibility gaps when it comes to the reproducibility of some AI publications. Companies cannot be expected to give away the family silver in code and full datasets for a journal article. Peer review cannot realistically validate new methods either. A further confounding trend will be datasets becoming more proprietary. The concept of findability, accessibility, interoperability, and reusability (FAIR) is great but largely restates well established principles that have existed for some time. The General Data Protection Regulations (GDPR) is an even newer challenge: how will it affect text mining? Licensing is a pain, and there are ambiguities, in my view, on definitions of commercial use. So I think there are still many more standards to be defined and established.

WAW: What do you think the funding priorities should be for the United Kingdom for the next 10-20 years?

JPO: There will be big choices to make in the post-COVID-19 era across academic and commercial sectors. One of the areas I'd personally consider is that of CPD in life science. Lots of professions have integral continuing professional development (CPD) programmes and people are constantly refreshing knowledge, skills and learning. In science, CPD has been neglected a bit. Skills gaps are not necessarily solved by training and recruiting new entrants. People will generally be looking for sustainable careers and "sticky" jobs that allow long-term planning, for a family, or other reasons. On the positive side, it is great to see the current government investing in life sciences, innovation and the knowledge economy.

WAW: What are the job prospects for young people in this area for the next 10 years? What would you recommend them to study?

JPO: Robustness and flexibility matter. Get quantitative skills: mathematics, computing and statistics are useful for many jobs. They give you flexibilty. Look outside London and the South East. House ownership and commuting are significant challenges even for people like me at the late stage of a career. It is a hard call to recommend which discipline to follow. Medicine is always important, but the complexity and wonder of chemistry and biology are so seductive!

WAW: John, it is always a pleasure to speak to you. Thank you so much for taking the time to share your thoughts with me today.

References

(1) Warr, W. A. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput.-Aided Mol. Des.* **2009**, *24* (4), 195–198.

(2) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O. et al. The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* **2017**, *9*, 33.

(3) Chemical Develpment Kit (CDK). http://cdk.github.io/ (accessed June 24, 2020).

(4) RDKit: open-source cheminformatics. http://www.rdkit.org (accessed March 30, 2020).

(5) KNIME. http://www.knime.com/ (accessed June 24, 2020).

(6) TensorFlow. http://www.tensorflow.org/ (accessed June 24, 2020).

(7) PyTorch. http://pytorch.org/ (accessed June 24, 2020).