AI 4 Science Discovery Network+

AI4SD Interview with Egon Willighagen
30/11/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

21/07/2022

Humans-of-AI4SD:Interview-25

AI4SD Interview with Egon Willighagen
Humans-of-AI4SD:Interview-25
21/07/2022
DOI: 10.5258/SOTON/AI3SD0227
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	AI4SD Interview with Egon Willighagen
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	EW: Egon Willighagen - Maastricht University
Interview Location	Online Interview
Dates	30/11/2021

2 Biography

Figure 1: Egon Willighagen

Egon Willighagen: "Up until five years ago, I was aware of all the open science in chemistry... I'm happy I can't keep up with it anymore."

Egon Willighagen is Assistant Professor at Maastricht University where he studies the role of machine representation of knowledge and hypothesis in life sciences, metabolomics, drug discovery, and toxicology, involving cheminformatics, chemometrics and semantic web technologies.

In this Humans of AI3SD interview he discusses the growth of open science over the past 20 years and the continued need for it, the challenge of deep learning methods, and his advice for early career researchers.

3 Interview

MP: What's been your path to where you are today?

EW: I started in 1993 as a student in organic chemistry at what is now the Radboud University Nijmegen in The Netherlands. I initially wanted to go into wet lab chemistry, but I couldn't for practical reasons around my health. The alternative was something like theoretical or quantum chemistry, but I didn't find them diverse enough, so I went into chemometrics and cheminformatics. I did my PhD in chemometrics, which involved data analysis of multivariate statistics on chemical data.

During my PhD I began working on the problem of how we represent the chemical knowledge we have and the data we use to measure it in a way which benefits AI. You can consider models as isolated studies with input and output data, but I don't think that approach is right. My hypothesis involves linking the model to independent data. For example, if you make a prediction of log P for some compound, that means that the predicted log P should itself provide valid conclusions using other models. The idea is that these models are only valid if they can be used in reality.

I was also concerned with the general availability of data, and FAIR (findable, accessible, interoperable, reusable) research output in general. At this moment, if I want to make property predictions for a chemical compound, for instance, I don't have access to all the models ever created to predict them, because that information is not available in a machine-readable way. This problem severely limits the validations of technology we're able to do.

MP: Apart from the unavailability of usable data, what other challenges have you faced in your research?

EW: There are social challenges and there are technological challenges. At this moment, the social challenges are the biggest hurdles. There are many reasons why data is not shared, but a lot of the time, it's simply that people don't care enough to share it properly. This is a result of rewarding the wrong views of what quality science is; some even claim that the journal you publish in can determine whether it's quality science. But we know what good quality science is, but we just don't care about tracking it anymore.

There are also problems around biases in science, where we see an overrepresentation of white male scholars. This is something we've been discussing for at least 20 years, but change is very slow. The current systems favour certain roles, and people of certain backgrounds are overrepresented in those roles. None of these biases are related to quality science, but we repeatedly find that they determine what is rewarded and recognised.

Most of the technological challenges are already solved. We know how to represent chemical problems very accurately, from the detailed models used in quantum chemistry to looser ones using a knowledge approach. The more detailed the data, the more the machine learning system can study, and the more precise your predictive model can be. We have these technological solutions, but we're not using all of them yet. If you look at the experimental data, the speed at which our technologies advance is faster than any computational system can capture

MP: What barriers would need to be removed in order to make the most of these technologies?

EW: If we had more accurate data, it would be easier to see which studies are wrong and why they're wrong. At the moment, we know that most studies are wrong in some way or another. We don't make perfect predictions, we're often able to measure the quality of models empirically, but we don't always know how to improve them. This is partly because if the model is opaque, we can't find the answers that indicate how it should be improved. One way of approaching this is to look at things that haven't been predicted well, and figure out why they weren't. From there, you can start to explore how to make better predictions, but I don't see that approach being taken much at the moment.

Also, we don't report when articles or data are bad, and we don't peer review that well either. In chemistry, there are so many datasets that have not been explored in detail, or adequately reported on. We see that in biology, for example, the issue with spreadsheets where certain gene names were converted to dates. The equivalent in chemistry, for small molecules, is SMILES (simplified molecular-input line-entry system) representations that have missing information. But none of the QSAR (quantitative structure activity relationship) models have a routine of reporting outcomes of quality analysis of the SMILES strings for those compounds, even if those SMILES are correct. One important change in our thinking should be to value the accuracy of our processes more: our notebook keeping, our design, the inclusion and exclusion rules in our workflows. These need to be more accurate and precise.

MP: What are the more structural changes that would help alongside these individual changes?

EW: At a higher level, we need open science. That is in motion, and more people are seeing that it requires investment, procedural changes, and, ultimately, a reevaluation of who they are as a researcher and why they do their research. One thing we have been seeing a lot is that to get funding, you need a competitive edge: that can be good research or good data. In the past, people have sat on data and basically said, "You can get access to my data, but I want to be a co-author on the results." That's still an ongoing practice, and while it's not horribly bad, it does slow things down. Open science generally makes it more inclusive; the research is more independent and allows for better peer review. Although not everyone is on board, this is a high-level and ongoing change to how we think about doing science.

MP: What is the key to getting more people on board with open science? Is it more of a stick or carrot approach?

EW: I started in open science probably around 1995, which was before I learned about copyright and licences, and could just put knowledge on the internet in an open way. This was a carrot approach: simply showing how it helps people. One of my first co-author publications was about the JChemPaint campaign — my contribution was only possible because of open science. In that article, we describe how in an open community, a collaborative approach allowed us to do research that was previously only done by commercial companies. People were convinced but, because of the recognition and rewarding system, we have found that in the last 10 years, the stick is necessary as well. You have to compensate for not complying to the existing system.

There are pretty clear sticks against researchers for not publishing in the right journals. We're trying to get rid of those, but in doing so, we learned that mandates are needed to compensate

for the other sticks. This is something we've seen painfully in The Netherlands, where there has been discussion about whether journal impact factors are a good measure of the quality of research or a researcher; they have no scientific basis but people feel strongly about them.

MP: How far has open science come, and how far does it have to go?

EW: Open science chemistry was an obscure, niche area 20 years ago. I remember a professor asking me, "Why are you doing open source cheminformatics? No one will cite that." That was the sentiment at the time. But today, the Journal of Cheminformatics expects open science approaches, and it's a very successful journal. Up until five years ago, I was aware of all the open science in chemistry, but now it's impossible to keep up. There's a flow of open science, and that's awesome. I'm happy I can't keep up with it anymore."

MP: What has surprised you in your research?

EW: Deep learning is surprising me a lot right now. The principles of deep learning aren't entirely different from neural networks, but I'm really surprised by how deep learning approaches seem to handle noise. Just by introducing noise on identities, you can duplicate identical observations. I don't find it very intuitive because, in the end, it's the exact same data that you're replicating. I guess the trick here is that artificial noise allows the model to recognise noise better, but that is quite surprising to me and I still haven't completely got my head around it! If we're representing a chemical structure within a SMILES string, we think of it as a well-defined graph, but in these deep learning approaches, it's no longer a static thing. A SMILES string is discrete, it's one thing. But with machine learning, it's no longer a discrete whatsoever. Even if you provide it with static information, the introduction of random noise adds additional dimensions that allow it to come up with wonderful, predictive models. It's a challenge for the field, we have to keep up with these deep learning methods.

MP: What advice would you give to early career researchers?

EW: That's difficult because I seriously don't know how I made it to Assistant Professor, and next year I go up for Associate Professor! I can give my personal experience, which involved following what felt good. When I start looking at something, I really want to solve it, and apparently my skills and my wits are good enough to do things that surprise other people. That means I get cited a lot, which means I get funding.

The thing that determines a scientific career at the moment is having plenty of everything: articles, citations, grant funding, teaching. That's what works at the moment: work 80 hours a week and then, if you're lucky, your career might go in the direction you like. That's the most realistic advice I can give, but it's not the advice I like to give because it doesn't make sense. There are a lot of problems with the current system that pushed us in this direction. If you look at the output of the average scholar right now, 50 years ago this person would have been considered excellent. But excellence is not absolute, it's relative; it's the top 1% compared to the other 99%. There is no such thing as everyone being better than average, because the average is not static.

The only wisdom I have is, try to find a job where you're happy. If you're not happy in your position then it's impossible to reach the output expected of you, or that you're going to be assessed on. If you don't have the motivation to reach where you want to be, then you'll be giving yourself a hard time. One thing I see people doing is applying to study in a field that is established, because if you want to secure a position in academia, then going into a niche field

is not the wisest choice. There aren't a lot of prizes or a lot of research funding in niche fields, so even if you do excellently, it's very hard. Cheminformatics is one of those fields, so I really don't know how I've managed to survive!.

Disclaimer: The opinions raised by Egon Willighagen are his own and do not necessarily reflect the opinions of his employers.