AI 4 Science Discovery Network+

AI4SD Interview with Dr Jan Jensen
09/11/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

11/08/2022

Humans-of-AI4SD:Interview-33

AI4SD Interview with Dr Jan Jensen
Humans-of-AI4SD:Interview-33
11/08/2022
DOI: 10.5258/SOTON/AI3SD0230
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 Dr Jan Jensen
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	JJ: Dr Jan Jensen - University of Copenhagen
Interview Location	Online Interview
Dates	09/11/2021

2 Biography

Figure 1: Dr Jan Jensen

Jan Jensen: 'Are our expectations of machine learning realistic? The community has yet to address this'

Jan H Jensen obtained his PhD in theoretical chemistry in 1995 from Iowa State University working with Mark Gordon, where he continued as a postdoc until he joined the faculty at the University of Iowa in 1997. In 2006 he moved to the University of Copenhagen, where he is now professor of computational chemistry.

In this Humans of AI4SD interview he discusses evolving molecules, the surprising simplicity of machine learning, improvements in openscience and his advice for early career researchers.

3 Interview

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

JJ: My interest has always been in chemistry. I grew up on a small Danish island close to Germany and at that time, there was a German TV channel which broadcast chemistry experiments. I was fascinated, and ended up setting up my own lab in the garage. I found, however, that I was more interested in reading about chemistry than actually doing it, which led me to the more theoretical side. I would buy textbooks and just make my way through them. I started with chemistry, and then moved into biochemistry and molecular biology.

At the age of 16, I went to the States as a foreign exchange student. The plan was to have one year in high school there, but it turned into 21 years, during which I completed a bachelors in chemistry and a PhD in theoretical chemistry before working as a postdoc and getting an academic position at the University of Iowa. I moved back to Denmark in 2006 to work at the University of Copenhagen, where I am now. Although my PhD looked at quantum mechanics, I've branched out into machine learning and its interface with quantum mechanics.

MP: What research are you working on at the moment?

JJ: I'm interested in looking at how we find molecules with new and interesting properties. That could be, for example, antibiotics or molecules that absorb light at a certain wavelength. Until now, a lot of these molecules have been found through serendipity: trial and error or chemical intuition. Our goal is to write software that can search through various possible molecules and find the ones of interest. That's important because it would cut down on the development time for making new materials. Trial and error is time-consuming and expensive, so if you can do things more efficiently, then we'd see much faster progression.

The techniques we use for searching for these molecules is quite unique. We use the genetic algorithm, which uses the principles of evolution for optimising properties: you try to evolve a molecule. That was popular for a while, before it got taken over by machine learning. Although machine learning has become the new hot thing, I've tried to revive interest in the genetic algorithm and show that, in some cases, it is actually more efficient than the fancy machine learning models. This is particularly the case when you don't have a lot of data, because it takes a long time to make new molecules and test them. Genetic algorithms have some key advantages here over machine learning.

The quantum mechanical methods we use are not used much by other groups. In quantum mechanics, you have a range of accuracy and computational expense and a lot of people work at either end. That means some people wait a long time to get the answer but it's relatively accurate, while others get the answer very fast but it's no longer quantum mechanics. What we've found is a sort of middle ground which is a compromise in both speed and accuracy. The challenge, however, is how to work with methods that are sometimes quite inaccurate. When you sometimes get errors, how do you work that into your strategy? Many people, when they see errors, throw it all out and go straight to the more expensive methods. But we've found that the initial methods can actually do a lot, as long as you're careful about how you apply them.

MP: Given the recent popularity of machine learning, do you think it's been overhyped?

JJ: There's two aspects to it. On the one hand, machine learning opens up a lot of new tools, which should be tried out. But there's a lack of comparisons to look at whether the new tools are actually better. Part of the scientific process is to develop a lot of new methods so, from that point of view, it's good that new things are being tried out, and every once in a while there's a method that's actually an improvement.

On the other hand, a lot of the hype around machine learning comes from people who are not experts, and it has to do with their expectations. For example, if a machine learning model predicts a new molecule, the reaction is usually one of two extremes: either "That'll never work," or "I could have thought of that, we don't need machine learning for it." So people expect new things, but they can't be too new, otherwise people think they'll never work. When machine learning is not able to deliver something in that middle ground, people can get disappointed, but what can one reasonably expect? That's a question that hasn't really been addressed by the community: what are our expectations of this field? Have our expectations been reasonable? Perhaps it's not the fault of machine learning, but with the realistic-ness of our expectations.

MP: What has surprised you in your research?

JJ: I'm always surprised by just how difficult it is to actually make molecules, and all the considerations that go into it. There are tons of practical things to consider: Which vendor can you trust when you buy your building blocks? Are they reliable? Will it come on time? There are also molecules that are, for all practical purposes, near-impossible to make. These things have surprised me, but it's mainly just ignorance.

I was also somewhat surprised at the simplicity of machine learning. I was never trained in it as such, so when it started coming up in the journals it looked very complicated, and I expected to never really understand it. At first, seeing a machine beat a chess champion or detecting that an image depicted a cat, it seemed like a miracle to me. The more I learned, I came to realise that these "miracles" are mostly driven by the data — if you don't have the data, no amount of clever software is doing to fix that. Some aspects of machine learning are still very complicated but, when you look under the hood and get down to the lines of code, it's actually not that bad. It's a long and steep learning curve, but I suppose you could say that about a lot of things.

MP: Where are we right now with open science?

JJ: I think it's getting better. With computational chemistry, more and more people are willing to share their code and their data. It's one thing that they're fairly good about in machine learning, because a lot of these practices come from computer science. I think it has to do with the amount of time one has to invest in doing these models. Because machine learning is faster than implementing the equations of quantum mechanics, you're perhaps less invested in the code and more willing to share it.

Open science is much better than it was, but it used to be pretty bad. We're not at a point where everyone does it, and that's holding back the field; it could progress much faster if everyone were doing it. We've certainly spent a lot of time trying to recreate others' work, and that's ultimately time wasted.

Funding agencies should be driving open science. There should be an absolute mandate that you share your data and code if you're getting certain funding. It's starting to happen with open access publishing. There are some funders who insist on that and it certainly creates a change because people want the money. They won't get fewer applications, and the worst that can happen is that the success rate goes up by a few percent, which wouldn't be bad either. Journals could also mandate open sharing of data and code. A lot won't because they're more concerned about turning a profit, frankly. But the journals run by professional societies have no excuse for not doing this already, in my opinion.

There are also things you can do on an individual level. I tend to refuse to review papers if the code is not shared. That's an example of one of the things you can do: to share openly that you're doing it to try to get others on board.

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

JJ: I'm a little hesitant to give advice because it really is kind of random. There's a danger of survivorship bias: I could say do this and that because I did, but it's actually a random path of things like luck and coincidence.

That being said, the number one thing is that you have to enjoy what you're doing. If you're doing something you dislike but you think will improve your chances in academia, you might as well make twice as much money doing something you don't like for a company. The other thing that's important is growing a thick skin. You will get rejected on, if not a daily basis, then on a monthly basis, be it your grant proposal or your paper. You have to learn to live a happy life with rejections and that involves not being emotionally invested in these things. Enjoying what you're doing and dealing well with rejection are both important.