Utrecht University

Sharing science,
shaping tomorrow

INAUGURATION

AI Technology for life

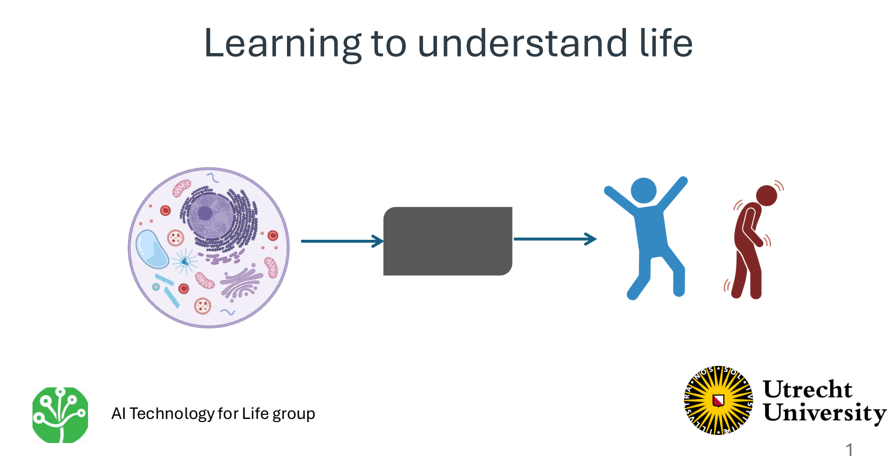
Prof. dr. Sanne Abeln

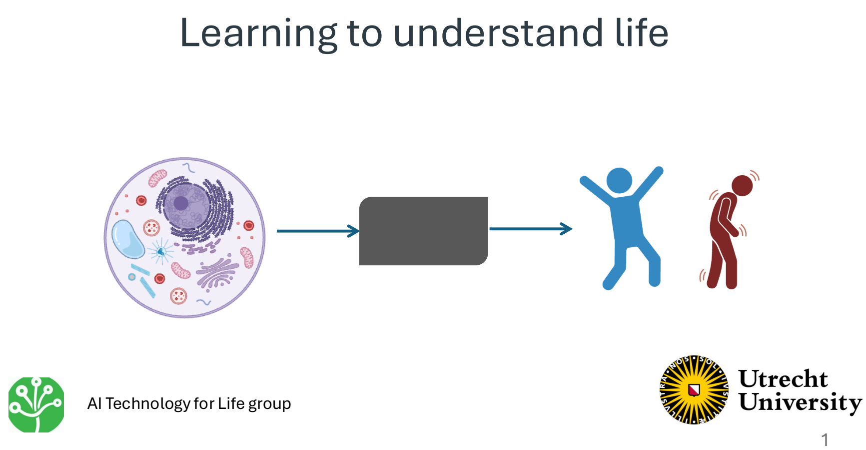
April, 2024

Inaugural Lecture AI Technology for Life

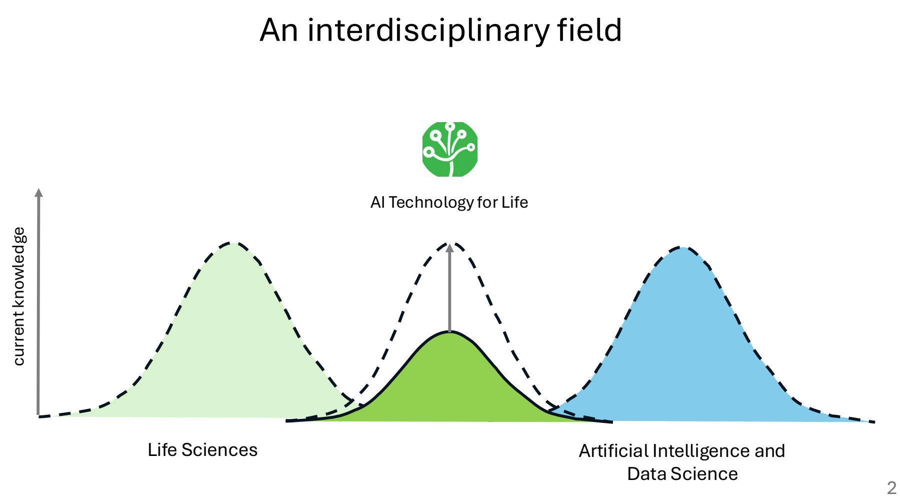
Sanne Abeln

April 2024

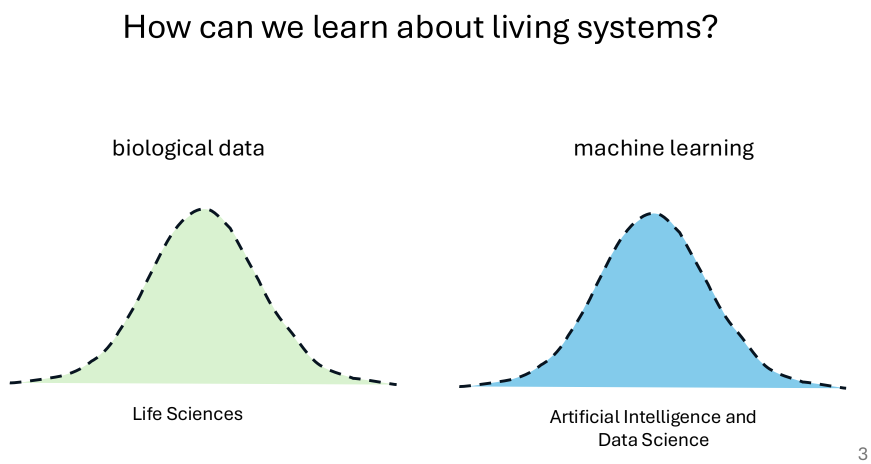




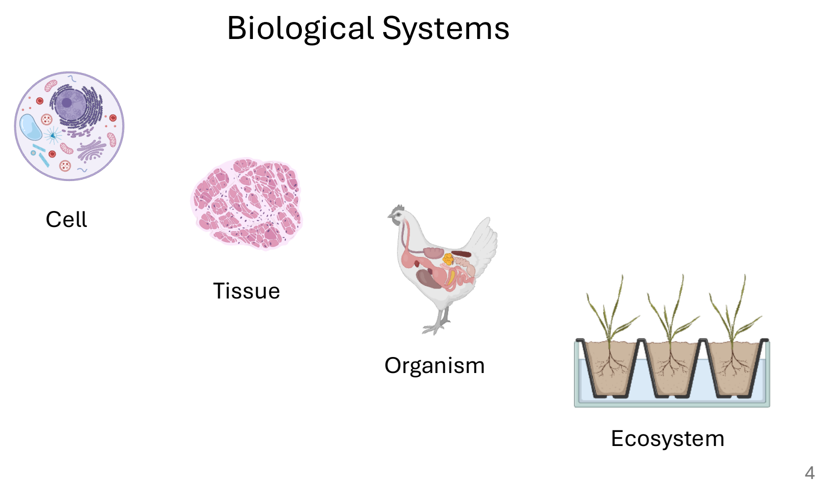
Mevrouw de Rector Magnificus, dear family, friends and colleagues, thank you for being here today. I see a very diverse audience with different backgrounds: my family and friends, who are perhaps wondering if they will be able to understand anything of this lecture; my colleagues from the biology department - who may wonder if they will understand the AI technology; my colleagues from Information and Computing Sciences - who may have similar concerns about understanding the biology; and then, there are the colleagues from my research field - who may wonder if they are going to hear anything new at all. With this lecture I hope to explain to you in simple terms how we can "learn to understand life" and to show you why I am so excited to work in this research area. I hope to show why this research is important and I will highlight some key challenges that we would like to tackle.



The diversity of the audience nicely reflects that my research field is interdisciplinary - meaning between two disciplines. Our new research group 'AI Technology for Life', at Utrecht University, is a bridge between the Life Sciences on the one hand, and AI and Data Science on the other hand. To reflect this, the group is affiliated with two different departments: Information and Computing Sciences, and Biology. In addition, there will also be a strong connection to the Pharmacy Department. Exactly because interdisciplinary fields lie at the intersection of other fields, there is a lot of new knowledge to be discovered, and developments go very fast. I will explain some of these key developments during this lecture.



Today, I would like to consider how we can learn about living systems. This is a complex question, and to simplify this I will split it into two parts: 1) what is the biological data from which we can learn? 2) What are the computational techniques that can help us to learn from data? In other words, I will split this question into two different domains: the life sciences and computing sciences. Even the first question is not a simple one; if we want to understand data about biological systems, I should explain what I mean with a biological system, and what I mean with data. Let us start there.



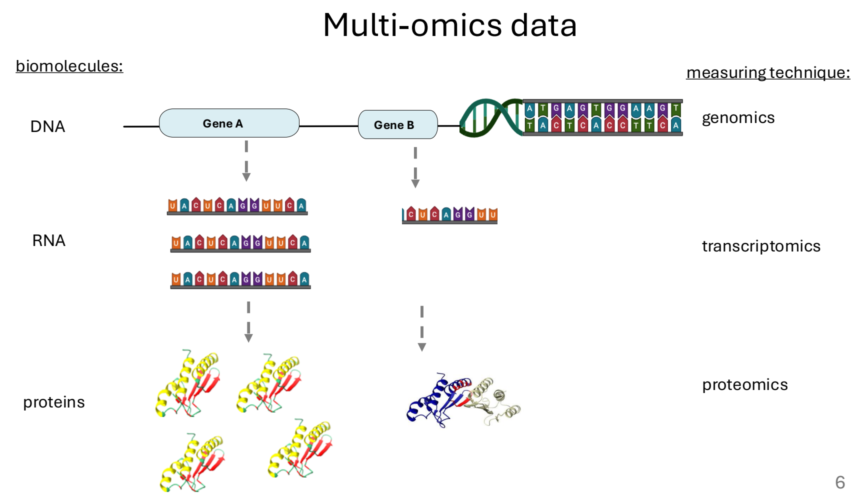
With a biological system, we can mean many things: for example, a cell, a tissue made of multiple cells, an organism made of several tissues (and organs), and if we go even higher we may consider entire ecosystems. Within this lecture, I will focus mostly on examples of the lower levels - of molecules and cells - and on human biology. Here, I would like to mention that in our research group we do also investigate other systems - for example how bacteria interact with the roots of a plant.

Advances in measuring biological systems

DNA sequencing

5

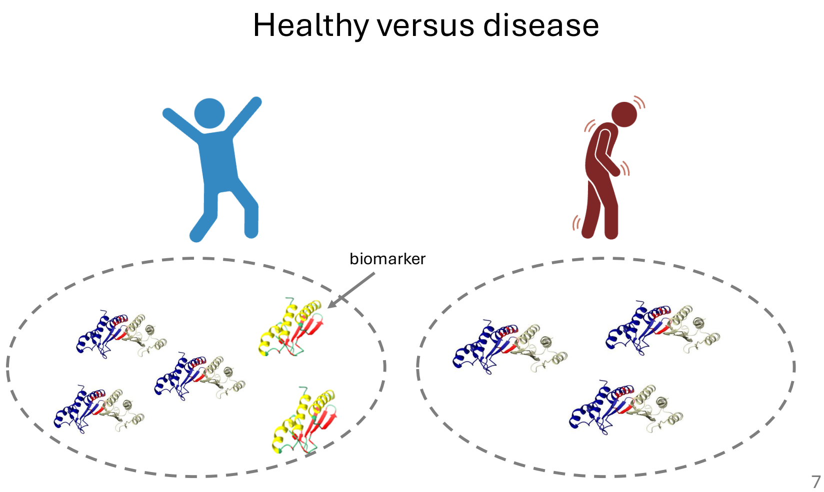
Over the past twenty years, huge advances have been made to measure these biological systems. One of the biggest breakthroughs is the machine depicted here. This machine is able to read out the DNA letter code of an entire genome. Note that these machines are currently present in all academic hospitals in the Netherlands, as well as many other research facilities This DNA sequencing provides us with a huge amount of data per sample. The human genome contains three billion (in Dutch 'drie miljard') different letters, providing us with a wealth of information on human cells.



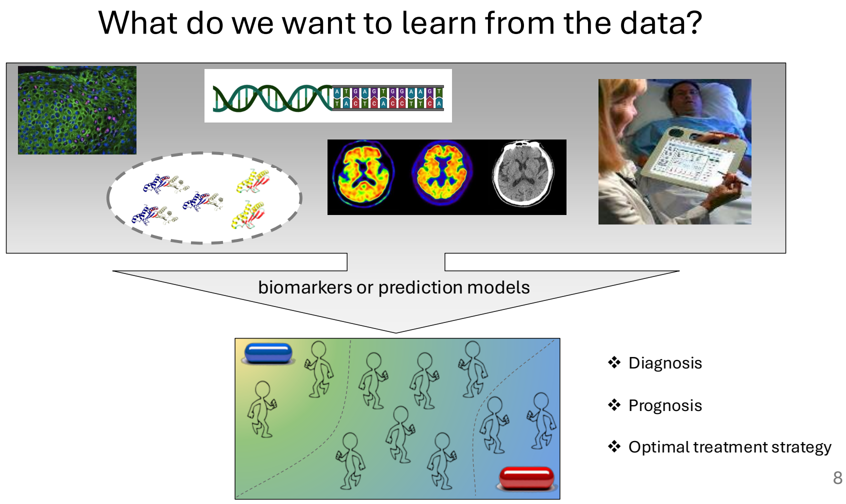
In the DNA of the genome our blueprint is encoded. This effectively provides a book with recipes for all bio-molecules that are produced in our cells. Within the genes on the DNA, the recipes are encoded. In the cell these recipes are first transcribed into ”orders”, which come in the form of RNA molecules. Subsequently, these ”orders” are translated into the actual ”dishes”: the protein molecules. These proteins are very active: they can be little motors within the cell, or they can help chemical reactions. For our computer scientists in the room: we can consider the DNA more like the CODE of a system, whereas the RNA and proteins reflect the STATE of the biological system.

Not only can we measure the entire DNA of a sample of cells, we can also measure the RNA molecules and a large fraction of the protein molecules in a sample. For DNA, we are mostly interested in the exact letters or code present and in changes that may have occurred. For RNA and protein molecules, we are also interested in how many molecules of each gene or protein type there are. We call this type of data - where we try to measure all the biomolecules in a sample - omics data. There are different types of omics data: genome data is genomics, RNA data is transcriptomics and protein data is proteomics.

We can perform these measurements over different systems: for example, on different tissue types, so that we understand which biomolecules are present in the skin and which biomolecules are present in the brain.



Perhaps an even more interesting difference to consider is comparing the measurements between samples of cells from people who are healthy and people who have a specific disease. By considering the differences between healthy and diseased samples, we can start to understand which biomolecules show changes. We can for example observe changes in the numbers of specific biomolecules. Note that in reality, such samples are extremely complex, as we can measure the levels of thousands of different types of RNA or protein molecules per sample. By comparing healthy people and people with the disease, we hope to observe changes in specific biomolecules. Specific biomolecules (e.g. DNA, RNA or proteins) that can indicate these changes are also referred to as biomarkers: markers that can indicate underlying processes of the disease.



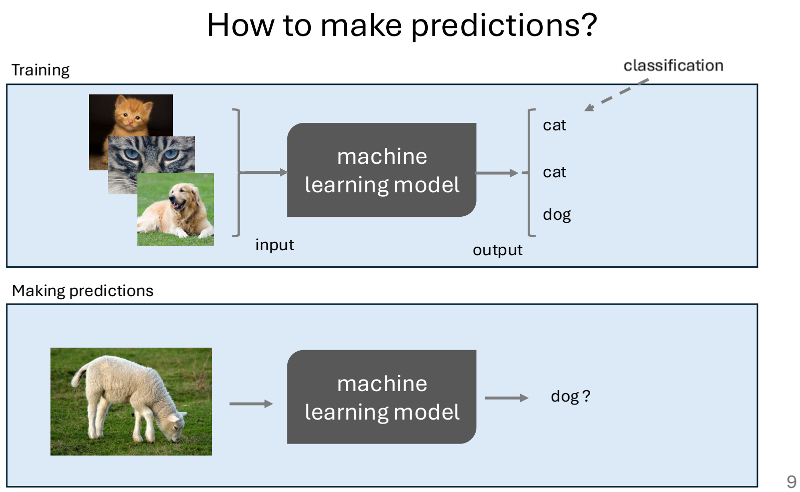
So what do we want to learn from this omics data? Typically, two things: 1) we would like to make predictive models that can help us make decisions, and 2) we want to learn about the biological system and what is going wrong in a disease. For now, I will focus on the first objective, prediction, but I will come back to the understanding later.

There are different prediction objectives, such as diagnosis: does someone have the disease? Prognosis: how will the disease further develop? And treatment: what treatment option will work best? To develop novel treatments, we also need an understanding of the system.

To make predictions about how a disease progresses, we typically need to combine at least two different sources of data. Classically, we may use clinical data, such as the symptoms of the patient, but we can now also include a wide range of omics measurements, as well as images from tissues or data from MRI or CT scans to make such predictions.

We can base our decisions on how to treat patients either on prediction models, or we can try to find specific biomarkers that can guide us in a decision. Note that I am now focusing on human disease, but that we could make similar models to predict how well a plant will grow in a given environment.

In the next section, I hope to explain how we can make predictions using large sets of data.

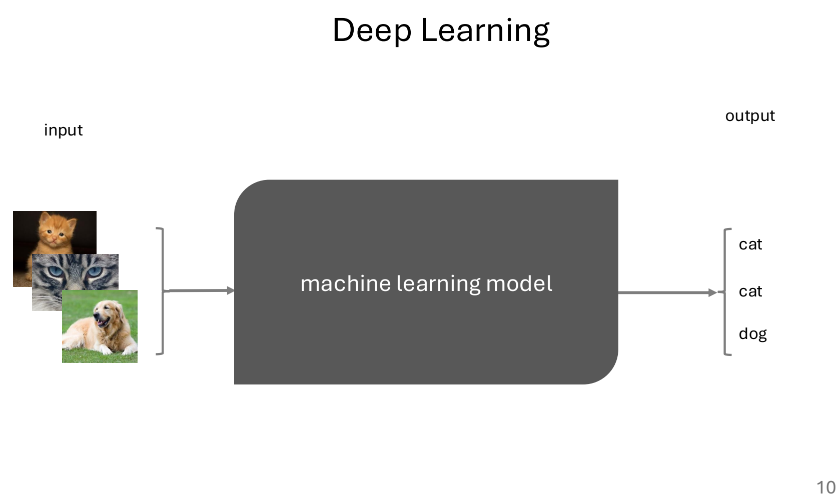


If we have a lot of data, we can use it to train a prediction model. This is called machine learning, which is a subfield of Artificial Intelligence (AI). A machine learning model is able to learn if you give it a lot of examples. For example, you can provide the machine learning model images of either a dog or a cat. With each image, we also tell the model whether it is a dog or a cat that is depicted. By learning from a lot of these examples, a machine learning model is able to make predictions.

Once the machine learning algorithm has been trained well, we can provide it with a new picture. It should now be able to predict whether the new image is a dog or cat. It is very important to realise that the machine learning model is very dependent on the type of data it has been given during training.

For now, I have depicted the machine learning model as a black box: for this lecture, it is not essential for you to understand how this box works on the inside. Inside the model there are a lot of parameters or weights; during the training, these weights are set so that the model provides accurate answers.

Now of course we could change the type of data and the type of prediction task that we want the machine learning algorithm to learn. We could for example provide the model with radiology images to predict whether a cancerous tumour is present. Similarly, we could provide it with genomics data to train it to predict whether a treatment would be beneficial. This does mean we need data - with a lot of examples - to train the model.



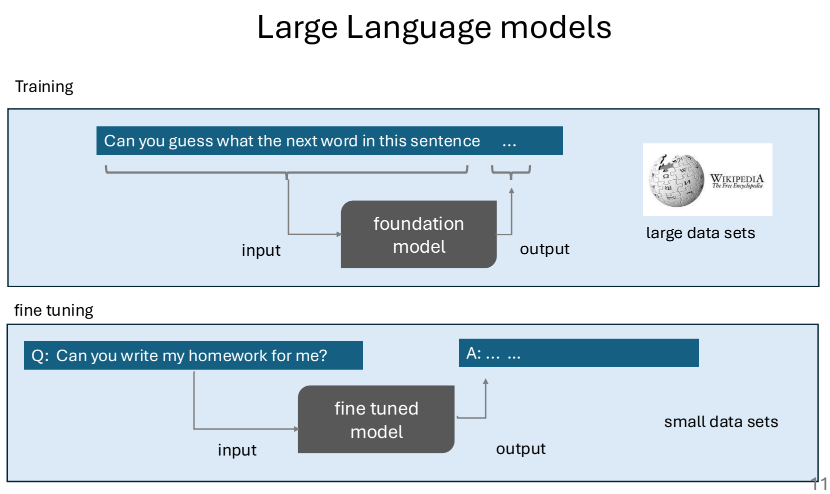
Over the last 10 years, major improvements in machine learning algorithms have been made. Most importantly, it is now possible to stack many machine learning models. We usually call these layers. In fact, each layer is a machine learning model in its own right: it takes in data as input and provides predictions as output that are passed onto the next layer.

If you look even deeper under the hood, you can see a complex network of connections that may have very different architectures.

There are very many different deep learning models. You may for example have heard of convolutional models, attention-based models and graphical neural nets, all with different architectures to set those weights. In fact, it is still not fully understood which types of architecture work best for which types of data and prediction problems. Although giving data the right attention seems to be important.

What these deep machine learning models have in common is that they can contain many layers that can be effectively trained. More layers typically mean that more weights need to be learned in the model. In order to set these weights, we also need more examples (more data) to learn over. The big breakthrough is that these deep architectures keep getting more accurate if you provide them with more data during the training process.

I now want to give you two examples of deep learning models that have made major impacts.

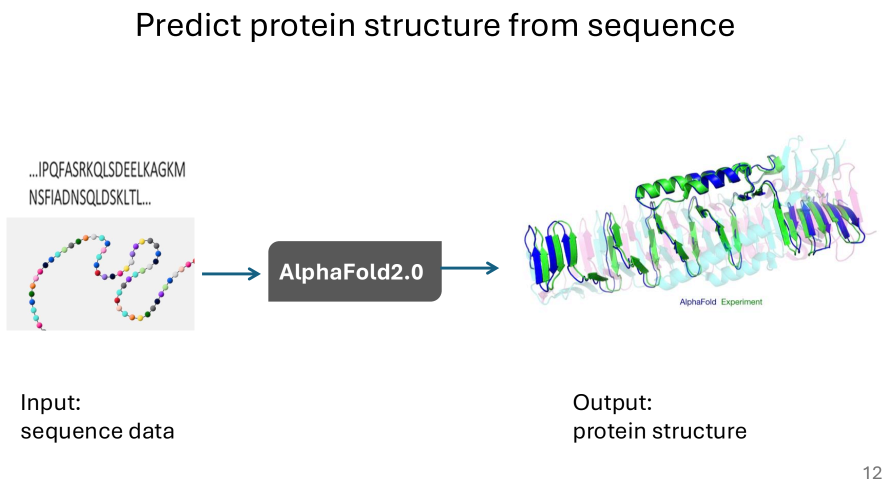


Now most of you will have heard of ChatGPT, the chat robot powered by the GPT deep learning model. ChatGPT can pass many types of high school and bachelor-level exams. Most AI experts, including me and many of my colleagues, were very surprised by the accuracy of the predictions made by GPT4.0[1] when it came out last year.

I will give you a brief sketch of how this may work. A machine learning model - GPT - is first trained to predict the next word in a sentence. A huge number of text sources are used (for example Wikipedia, but probably also many other sources) to train this model, which contains billions of parameters.

Subsequently, the model is fine-tuned to answer specific questions in the form of a chat. I cannot give you a full explanation of how ChatGPT4.0 works, as many of the details of how the method has been trained have not been published. So despite the name of GPT's developers - "OpenAI" - the training datasets and model architecture are not at all open.

I will come back to this point towards the end of this lecture sharing some ethical considerations in developing AI technology.



Another major breakthrough of deep learning models is the ability to predict the protein structure from a sequence. Let me first explain what the prediction task means: DNA contains genes that encode for proteins. Because we can sequence genomes, we also know the code for the proteins. We do not know, however, how these molecules fold up into a three-dimensional structure. It is this three-dimensional structure that makes the protein function in a cell. Predicting this structure was one of the major unsolved scientific problems of the last 30 years.

[In fact, almost 30 years ago, a scientific contest - CASP - was set up to evaluate protein structure prediction using experimentally validated structures. Many research groups had claimed to solve the prediction problem, but the contest clearly showed it remained an unsolved problem for many years.]

Less than 4 years ago the introduction of AlphaFold2.0[**?**] made a disruptive change, suddenly allowing scientists to predict protein structure. The deep and complex architecture of AlphaFold2.0, together with cleverly reusing the same examples to train the model, made this breakthrough possible.

However, it is unclear how and why AlphFold is able to make such good predictions. In fact, the largest improvement between AlphaFold1.0 to AlphaFold2.0 came by removing explicit physical constraints in the learning model: in the newest model atoms are allowed to overlap. In addition, we do not understand why specific protein folds are stable, or how long they take to fold. In other words, here the advance in prediction has not (yet) led to an advance in understanding.

AI for the Life Sciences – some examples

Can we predict ...

... how sticky protein surfaces are?

... which proteins may be captured inside vesicles?

... if a genomic alteration is important for cancer biology?

... how high RNA expression is, given the (starting) sequence?

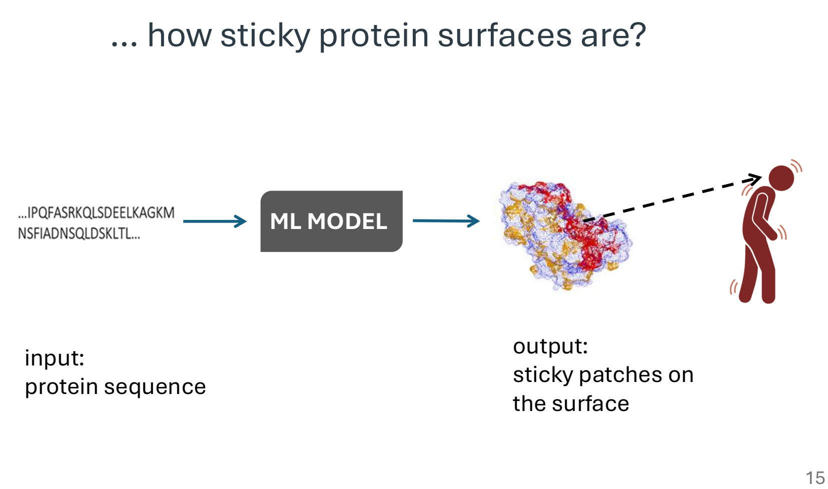
13

Now, I would like to show examples of AI prediction problems. These are all examples of active research projects in our group, carried out by PhD and master students, so all credit for this work should go to them.

I would also like to stress that many of these research projects have not been performed in isolation by our computational group, but are in fact collaborations between our group and other (experimental) research groups.

Firstly, I would like to introduce a prediction method developed by several people in our group: three PhD students - Juami van Gils, Dea Gogishvili and Erik van Dijk - as well as three master students - Jan van Eck, Robbin Bouwmeester and Henriette Capel. (Note that if the science gets a bit tricky here, you can also try and remember these faces and speak to them during the reception; they will probably do a much better job at explaining their research.)

Can we predict ... how sticky our protein surfaces are?

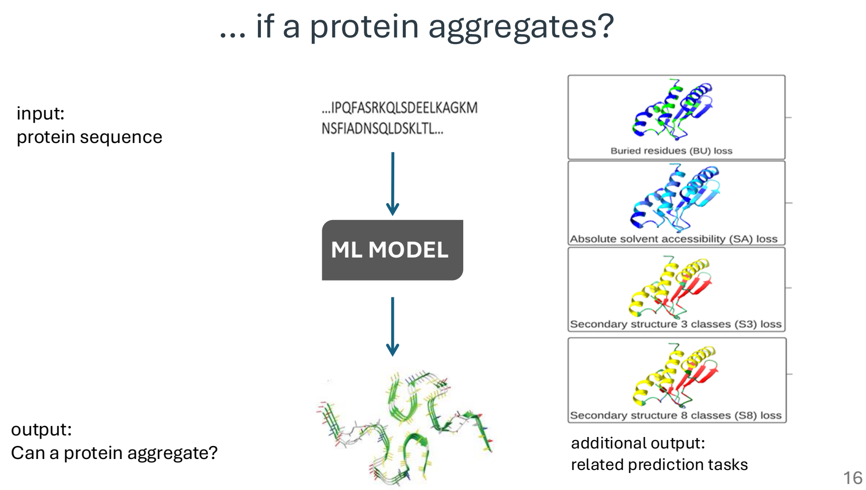


This task is about predicting which parts of the protein sequence will form sticky patches on the outside of the protein in its folded form. This is not an easy task because these sticky patches usually fold into the inside of the protein. In other words, in this task we are trying to predict the exceptions. Nevertheless, these sticky patches can be very important: they can have specific functional tasks. Proteins with large sticky patches on their surface are also more likely to stick to other proteins, forming unintended aggregates.

Protein aggregation is associated with several neurodegenerative diseases, such as Alzheimer's, Parkinson's, ALS and Prion disease. These are diseases, mostly dementias, that become progressively worse over time, resulting in loss of cognition and/or mobility because cells are dying in our nervous system: in the brain and other nerves. These diseases are thought to be caused by proteins that stick together; these proteins, stuck together, make up fibrils that can be observed in the tissue of patients with these diseases.

We successfully created machine learning models that could make these predictions[**?**], allowing us to have an estimate of the stickiness of the surface of each protein in the human body.

Now we have a prediction model - it is possible to make a prediction for each protein in the human proteome. Our research showed that the brain contains surprisingly many proteins with these sticky patches. This perhaps explains why the brain is known to be very susceptible to neurodegenerative disease.



Currently, we are developing a model that can predict aggregating proteins directly, but we have very few examples to train the model. Nevertheless, we still want to use a deep learning model, and to do so we use a trick to train the model on related tasks.

We made a model to predict aggregation, but alongside we also predict other related tasks, such as the sticky patches and the local structure. These other tasks help the model to learn the aggregation task for which we have very few examples.

This is not dissimilar to how we learn ourselves: if you speak Dutch, it is typically simpler to learn German than Chinese. In other words, it helps to know Dutch when learning German. They are related learning tasks. Hence, if we use related tasks we can use deep learning methods even if we have little annotated data available[**?**].

... which protein may be captured inside vesicles?

Dea Gogishvili

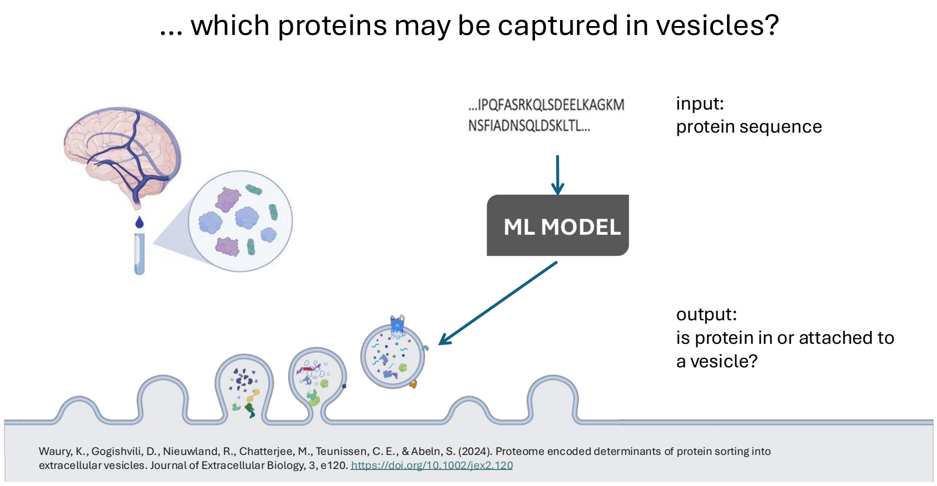
Katharina Waury

Charlotte Teunissen

Waury, K., Gogishvili, D., Nieuwland, R., Chatterjee, M., Teunissen, C. E., & Abeln, S. (2024). Proteome encoded determinants of protein sorting into extracellular vesicles. Journal of Extracellular Biology, 3, e120. https://doi.org/10.1002/jex2.120

17

Secondly, I would like to explain a prediction method developed by the PhD students, Katharina Waury and Dea Gogishvili in our group. This is a collaboration with the lab of Charlotte Teunissen at the AUMC. This research was also part of a European consortium to discover biomarkers for dementia, Miriade.

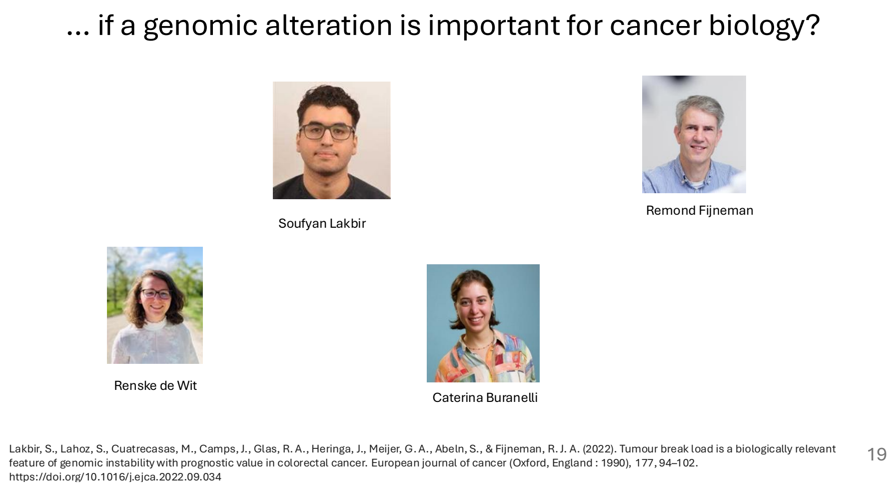


To diagnose the specific form of dementia a patient has, it is helpful to look at the level of specific proteins - in other words, protein biomarkers. However, we cannot measure these proteins in the brain directly but need to capture them in fluids such as blood or cerebrospinal fluid. Proteins that we can measure in bodily fluids have been discharged by the cells in which they are produced. Many proteins float freely in the fluid, but others are captured inside or on the surface of small vesicles (as depicted on the slide). If we want to measure proteins in these fluids, it is helpful to know if the biomarker of interest floats around freely or may be found in or on an extracellular vesicle.

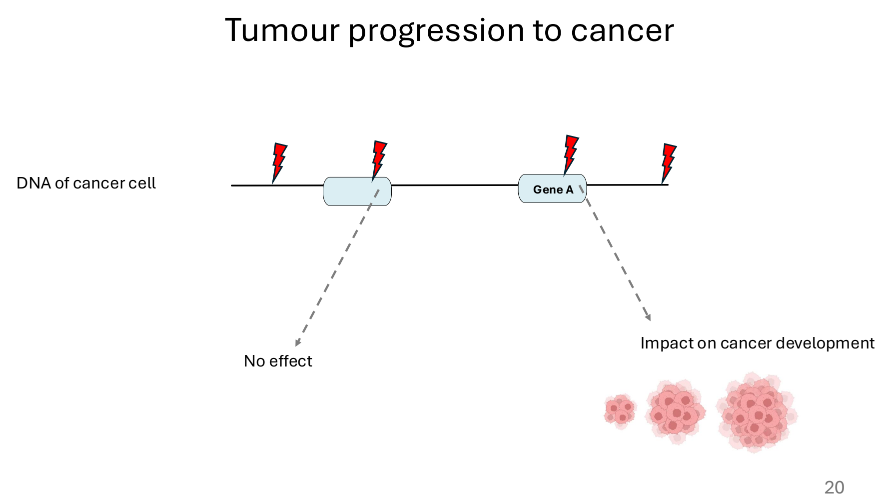
Can we predict ... which proteins may be captured in vesicles? Here we try to predict which proteins may end up outside the cell, and in vesicles. To solve the prediction task, we made a machine learning method that can predict which proteins end up in the vesicles, and which proteins do not. This was a hard task, as we had to correct for an experimental bias in the training data. Nevertheless, we were indeed able to train a model that could make these predictions[**?**], showing the mechanism of how proteins end up in a vesicle is encoded in their sequence.

We also wanted to explain the predictions. By reading out our machine learning model, we could understand that proteins with specific post-translational modifications - these are small changes that are made to the sequence after the protein is translated - were much more likely to be associated with these vesicles. Hence, the AI methodology has shown potential biological mechanisms, as well as giving leads for how we should handle specific protein biomarkers in the laboratory.

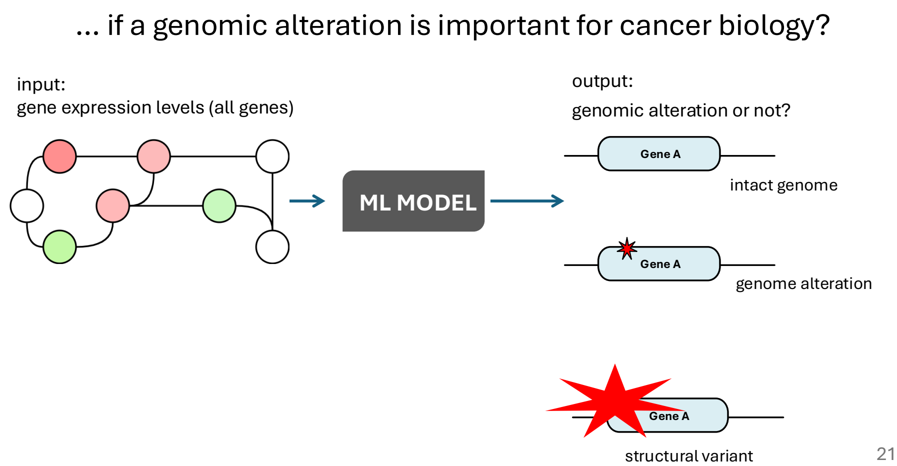
[We are now using the model to help several experimental groups that develop biomarkers for dementias to see whether a protein of interest is likely to be inside or attached to a vesicle. Note that there are experimental techniques to fish out these vesicles and open them up, so the experimental protocol may be adapted if a biomarker is likely to be found in a vesicle.]



Thirdly, I would like to explain a research area that two PhD students, Soufyan Lakbir and Renske de Wit, and a master's student, Caterina Buranelli, have been working on. This is a project in collaboration with Remond Fijneman and the Dutch Cancer Institute (NKI).



First, I will need to explain some essential details about cancer development. Cancer cells gain many more mistakes in their DNA than normal cells - we will call these changes genomic alterations. Many of these genomic alterations are random events that appear by chance - just because cancer cells acquire many mistakes. However, some of them really do change the biology of the tumour. For example, these DNA changes may block a signal that the cell needs to die or to stop dividing. These are changes that largely determine the characteristics of the cancer cells, how aggressive the cancer is, if it is likely to metastasize (spread to other parts of the body) and how sensitive the cancer cells are to specific drugs.



Our computational problem is to find those genomic alterations that do have an impact on the biology. We figured that if a genomic alteration affects the tumour biology, it will also largely disrupt the RNA gene expression in the cell. In our approach, we try to train a machine learning model to predict whether a genomic alteration is present by giving expression data as an input[**?**].

Here we use machine learning in a very different way. We are not actually interested in the predictions of the model. Note that the genomic alterations in a single gene are easier to measure than the gene expression patterns of all genes. However, it is not the prediction itself we are interested in. We are interested to see whether we can train the machine learning model at all.

If we are able to train the machine learning model effectively, this tells us that the information of the genomic alteration is also present in the gene expression profile. In other words, the genomic alteration affects the mechanisms in the cell and thus is important for the cancer cell. On the other hand, for genes with mutations that are not important, we see that it is not possible to train a model that makes accurate predictions.

We used this method to look at the importance of a specific type of genomic alterations, structural variants, which make large changes to a gene. However, until now it was not well understood if these larger changes were important. With our method, we discovered that in many cancer types - specifically in colorectal cancer - structural variants are at least as important for the mechanisms in the cancer cell as the better-known small point mutations. Together with the experimental group at the NKI, we are now in the process of exploring whether these structural variants also can tell us how patients with these specific genomic alterations should be treated. Initial results suggest that such cancer cells with specific structural variants are indeed less sensitive to certain drugs. Hence, the AI method has given us important leads for new genomic biomarkers in colorectal cancer.

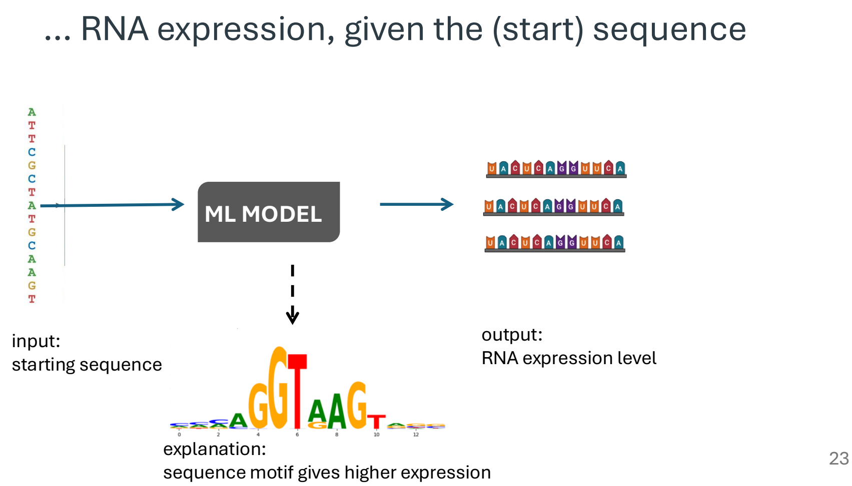
... RNA expression, given the (start) sequence


Laura Torroba Vicario


Hanneke Vlaming

22

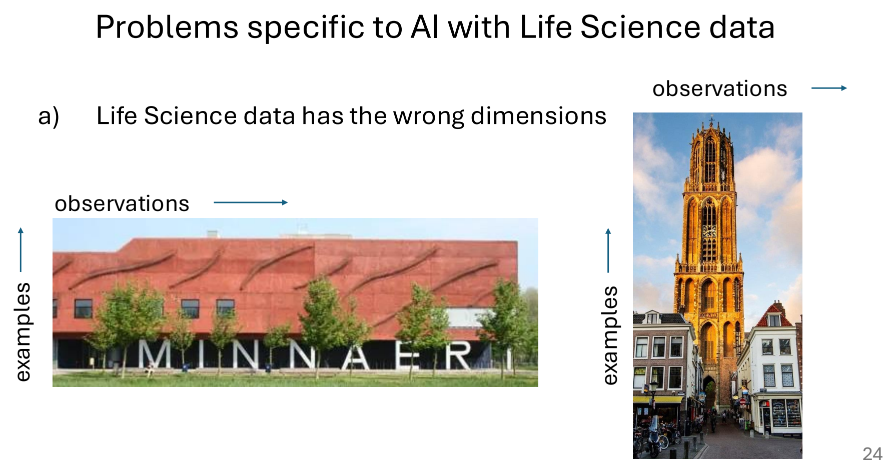
The last example is a recently started project in Utrecht performed by master student Laura Torroba Vicario in collaboration with Hanneke Vlaming from the Biology department



This example is a prediction problem that is set up the other way around as compared to the previous problem. Here we predict from the DNA sequence what the level of RNA expression is. The setup is a dedicated experimental dataset, with measurements of how high the RNA expression is, given a specific starting sequence.

The machine learning model tries to predict from the gene sequence how effective the transcription into RNA is. The model can learn this task quite well, meaning that part of the expression level is indeed encoded in the sequence.

Here we do not only want to make a prediction but also learn to understand the system. Hence we want to explain the predictions. We can read out the model by observing whether a prediction changes when we change the input sequence. Using this type of explainability, we were able to find two motifs that increase the expression: one was a known motif, but the other motif had not been previously found. Now Hanneke is doing experiments to validate these findings in the lab. Again, we see that AI may be used to explain a complex biological system.



Next, I want to explain where our research group is heading. First, we will consider some properties of Life Science data that make it tricky to use the new advances in deep learning models. I want to discuss three common problems:

a) Life Science data has the wrong dimensions

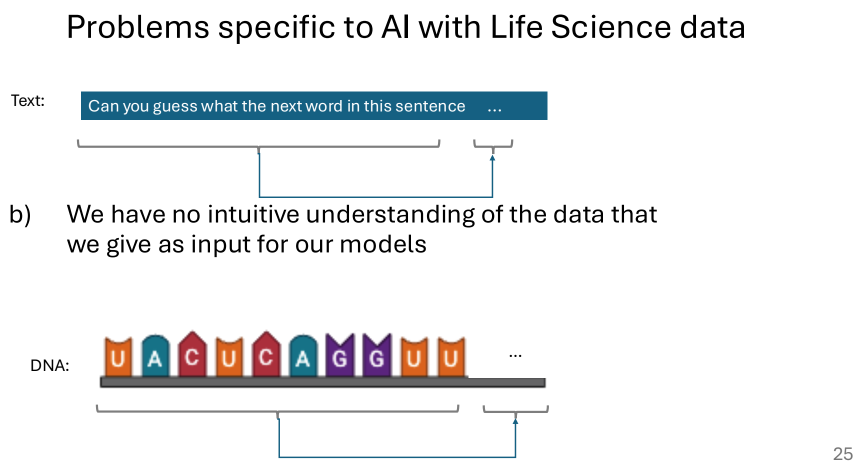
Firstly, typically data in the life sciences has the wrong dimensions for deep learning. Let me explain what I mean by this. Imagine we want a big building - because we want to look far away. Consider the Minaert building at the Utrecht Science Park - it is big, but it is broad but low. To look far, we would rather have a building shaped like the Dom in Utrecht, thin but tall. In other words, the red building is big, but it has the wrong shape to look far away.

Problem:

There is a similar problem with life science data: we typically have many observations for a sample (e.g. the number of genes), but very few examples with helpful (clinical) annotations (e.g. if treatment worked). To give some numbers, from the genomics data we may have ten-thousands to millions of observations per sample, but a typical cohort of patients is only 200–300 samples. This means that the machine learning algorithm only has 200 examples to learn over. Using standard techniques, this is too few to effectively train the huge number of weights in deep learning models.

Possible solutions:

In most of the questions we are working with, the lack of annotated examples is a real problem. However, we also see that by learning over related tasks we can reduce the need for labelled data; in addition - once we know a problem is predictable - we can often fall back on more advanced statistical methods to analyse datasets with smaller numbers. Our group developed statistical methods so that we could also look at the importance of genomic alterations if we have as little as 10 annotations.



b) We have no intuitive understanding of the data that we give as input for our models

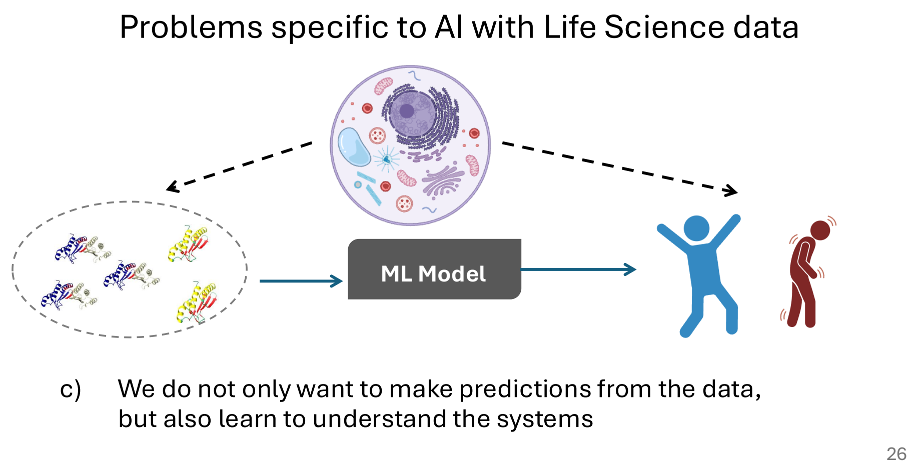
AI researchers have mostly developed their AI algorithms on text and images. For a human, it is relatively easy to assess whether the answer of an AI model is right or wrong. For example, when I showed you the picture of the lamb earlier - you would immediately recognise it was not a dog.

Problem:

For machine learning in the Life Sciences, the data we use is much less intuitive. I, for example, cannot read and understand a DNA or protein sequence. In fact, there are only very few people in the world that can. So the problem is not only to make predictions but even to start learning the language of the biological systems. This makes it not only harder to develop models but also harder to evaluate models to test if they are making useful and sensible predictions.

Possible solutions:

In order to work with life science data, we need an additional effort to label the data in such a way that we can understand it and correct for experimental bias in the data. We also need specialised methods to explain prediction models, such as techniques that can capture sequence motifs.



c) We do not only want to make predictions from the data but also learn to understand the systems

The mechanisms in the cell cause different biomolecules to be present in healthy people and patients. While it is helpful to train a machine learning algorithm that can predict if a sample is from a healthy person or a patient, this does not necessarily lead to an understanding of the underlying disease mechanisms.

Problem:

In biology, there are still many open questions: we do not fully understand how a cell works, how the cells interact together to form a functioning body or how proteins fold. If I ask my colleagues in biology, or my collaborators, what percentage of the existing interactions are currently known, their estimate is well below 50%.

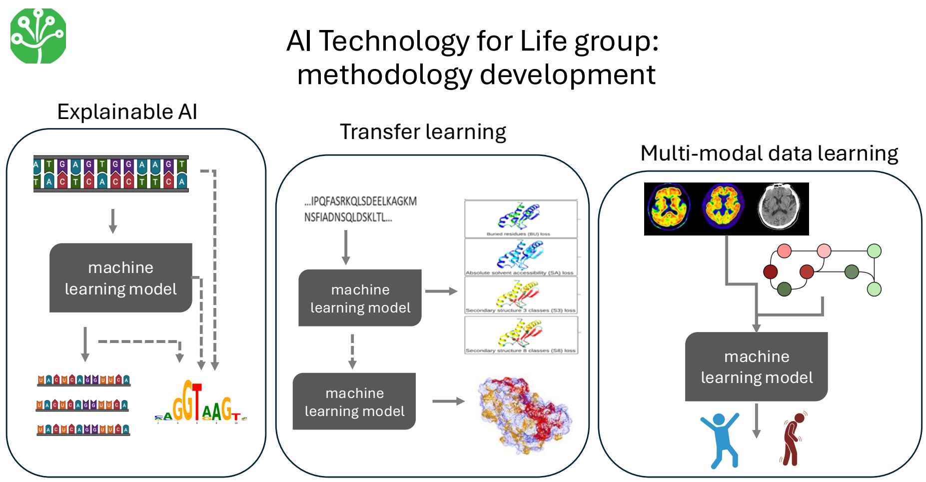
For example, from the omics data from healthy and diseased samples, we would like to gain an understanding of the underlying disease mechanisms. We can only target specific mechanisms in the cell with drugs if we understand what is broken. To develop new drugs, it is essential to understand which biomolecules are responsible for any changes observed.

Possible solutions:

In most of the examples I have presented to you, we did not only make predictions, but we also learned about the systems we were studying. Remember the vesicle prediction task? We learned that post-translational modifications in proteins were associated with being inside or attached to a vesicle.

In addition, we can use machine learning models to question whether it is possible to learn a task. This helped us for example by predicting the importance of genomic alterations.

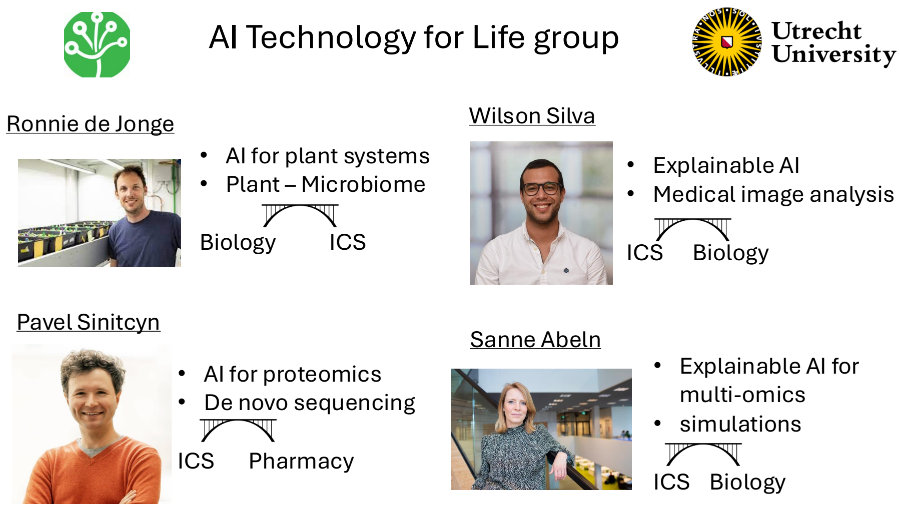
Lastly, to get further in science, collaboration is essential. While AI methods may make nice predictions and give good leads for new ideas, follow-up by laboratory experiments is generally essential.



Here in Utrecht, I am setting up a new research group - AI Technology for Life. The challenges I have just described set a natural focus for the group in terms of method development.

This means that there will be a focus on developing and using explainable AI methods. These are methods that allow you to explain the predictions that a model makes. However, existing techniques need to be improved and adapted for life science data. The second focus is transfer learning or fine-tuning; this means that we want to train large models on public data and fine-tune these models to learn over smaller datasets for a very specific task.

Lastly, we want to develop methods that can handle multi-modal Life Science data. This means learning over different data types: for example different types of omics data, or genomics data and images.



Now, I would also like to introduce the AI Technology for Life group to you. Within our group, we have different areas of expertise, some with a focus on AI methodology development, and others with a focus on specific problems within the life sciences. We will have four principal investigators in our group.

First of all, Ronnie de Jonge, is focussing on AI for plant-microbiome interactions. Hereby, he also forms a bridge between Information and Computing Sciences and the Biology department. In particular, he forms a bridge between our group and the very strong Utrecht plant research community. An example of an AI prediction is if we can predict which microbes are happy around the roots of a specific plant.

Secondly, there is Wilson Silva, who will focus on the development of explainable AI methodology and multi-modal prediction problems, thereby connecting to the Information and Computing Sciences Department. In addition, he has a research focus on medical images. One of the problems we will be focussing on is if we can predict clinical outcomes, from different types of data - for example, genomic data and images - and which parts of the data in this setting can explain most of the prediction.

Thirdly, Pavel Sinitcyn will focus on AI for proteomics. He will form a bridge between the Information and Computing Science department, and the Pharmacy department, in particular with the highly innovative Biomolecular Mass Spectrometry and Proteomics group of Utrecht. Typical problems we may be focussing on are: if we can predict which proteins are more easily detected by mass-spectrometry, and if we can generate artificial mass spectra that could help by identifying proteins for which we do not know the sequence.

I have already given you plenty of examples of research performed under my supervision. Nevertheless, one point that is worth mentioning here is that we will also use simulated data to gain further understanding of data in the biological systems we study and to develop explainable AI methods.

Ethical considerations of AI technology

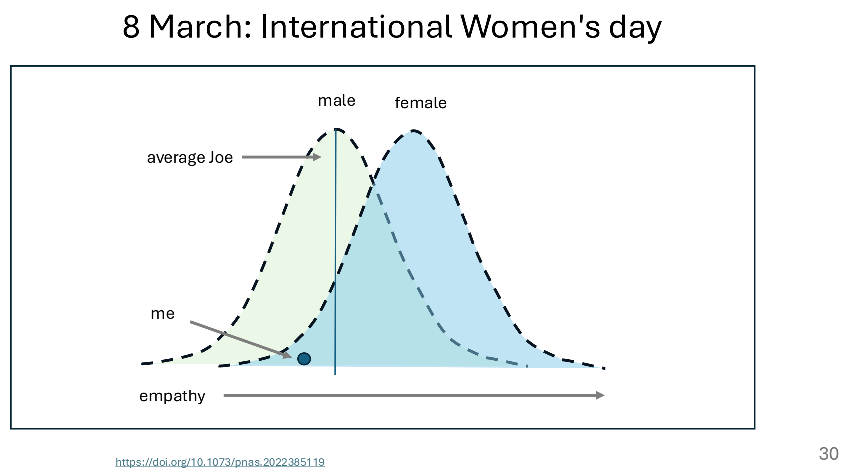
29

Something that needs to be mentioned are the ethical considerations of AI technology development. Many of you may have heard of doom scenarios, whereby artificial general intelligence will outsmart humans.

First of all, I would like to state that the purpose of our group is not to develop artificial general intelligence (AGI). That is artificial intelligence that can perform tasks beyond those it has been trained for. However, it is not unthinkable that our group will develop techniques that can also be applied by groups that do aim to develop AGI. For this reason, I am very glad that measures are being taken - for example by the EU - to regulate the development of AI. These regulations predominantly work on the basis for which purpose the AI techniques are being developed, and what resources they use in terms of compute and data.

Some ethical considerations that are immediately relevant for AI technology in the Life Sciences are those around data usage. As may have become clear from my lecture, data is crucial to train AI models. We need to be careful how we deal with this data. Luckily, tight regulations are in place to ensure the privacy of the data that we work with, gathered from people who could be, for example, very ill with cancer. Here, I would like to note that not sharing data may also be viewed as an ethical issue: some research groups are still not keen to share their data with other researchers, and while of course consent from the patient to reshare their data should be in place, not sharing can also be seen as highly unethical. Not sharing in the field of cancer research means that more people have to get very ill before we have enough data to learn what made them so ill.

Generally, this is really a call to collaborate even more intensively to gather data with care and ensure no data is wasted. Lastly, open science - sharing not only data but also the algorithms to make predictions - is essential to make scientific progress. This prevents other researchers from having to do the same work again, meaning we can get solutions to our questions faster. In addition, it is almost impossible to verify claims about prediction accuracy if the method and data have not been made available. Take GPT for example: the lack of transparency in terms of the machine learning model and data used for training makes any claims about performance and AGI capabilities unverifiable.



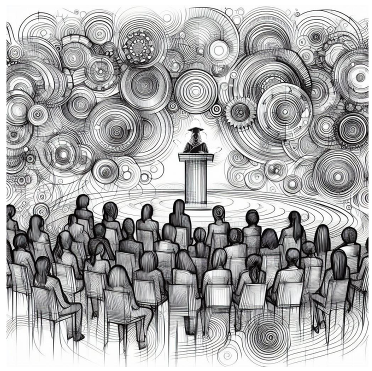
Just before I finish this lecture, I want to briefly draw your attention to today's date - 8 March - International Women's Day. Firstly, I would like to say that I am happy, that steadily - but very slowly - there are more women in Science, and that there is more diversity. For me, this makes everyday work more pleasant. Diversity of people leads to a diversity in motivations to conduct Scientific Research, which leads to a more creative, open and pleasant atmosphere for everyone working in academia.

More generally, I would briefly like to touch upon another point. Personally, I see female and male characteristics more as a spectrum. Let us consider a trait that may be associated more with women, like empathy[**?**]. Now on average, it may be that women are more empathetic than men. However, as I have sketched on the slide, these are highly overlapping distributions. I think strict classification based on this trait could be very harmful.

For example, if we make rules based on being female or male, this could lead to a society where men would no longer be allowed to lead teams - as they may not empathise sufficiently with the people they supervise. Such rules would of course be nonsense - as I, as a female, may well be less empathetic than the average Joe. In other words, no decisions should be made based simply on gender: we need to look at the individual, and this is something we see more and more of today.

However, one of the remaining problems in our society today is that traits that are classically ascribed to women attract much lower economic rewards and are not valued as highly as those ascribed to men. For example, even though it is very clear that a functioning society needs people to raise and teach our children and care for the vulnerable, these tasks do not have the same societal status or rewards as leadership or risk-taking. It is this historic legacy in the way we value services that means that women are economically less independent. I hope the increase of women on the (scientific) work floor will lead to a higher valuation of the characteristics typically associated with women, like collaboration and empathy.

Thank you!



31

Now, I would like to thank the people that have enabled me to give this lecture here today.

The guidelines for this lecture state the acknowledgments - or thank yous - should be short. As most research is nowadays done in larger consortia and teams, we should celebrate that we can work together as scientists, so as you may have noticed, I already sneaked in quite a few names during this lecture - and I would like to thank these young scientists and collaborators for allowing me to present their research here and also for making several of the impressive figures you have seen.

I also would like thank my husband, Peter Crowe. You have been helping me for more than 22 years with my English, my speaking and writing, and are supporting me in every way possible – also by taking on a nonconforming gender role and taking up most of the household tasks.

I would like to thank my close friends and family for their continuing support. Some have travelled all the way from the UK, Hungary, Groningen, Twente and Den Haag to be here today, while others are watching online.

From the Science crowd, I would first like to thank all the PhD students, master students and postdocs that have worked with me over the years. I have been incredibly lucky to work with such bright and talented people, making research wonderful - every day. Thanks to my own supervisors and mentors for leading the way, being a role model and supporting me. Specifically, I would like to thank my PhD supervisor Charlotte Deane, who showed me how you can lead in a powerful way, while preserving a great team spirit.

In addition, I would like to thank Daan Frenkel, who showed me how to lead a research group driven by curiosity. Also, I would like to thank Jaap Heringa, for his support and belief in me, and for being an example of a kind and respectful leader.

Specific thanks also to my close collaborators, and in particular Remond Fijneman, Charlotte Teunissen and Alexander Büll. Without close collaboration, interdisciplinary science would have little value.

In addition, I would like to thank all my colleagues at the Vrije Universiteit Amsterdam and the University of Amsterdam, with whom I have enjoyed teaching for 14 years, during which more than 500 students have graduated with an interdisciplinary master's in Bioinformatics and Systems Biology.

Moreover, I would like to thank everyone in the new AI Technology for Life group, for being enthusiastic and willing to work on difficult but exciting challenges.

Lastly, a special thanks to everyone at Utrecht University, including the board of the university, the faculty of Sciences, the Department of Biology, and the Department of Information and Computing Sciences who made this new group possible: by making plans, by being in interview committees and by sharing their vision of such a new group.

Ik heb gezegd - Dixi