New UK Collaborative Uses AI to Predict Missing Data Points in Compound Data

WRITTEN BY: Frederick Dawson

A new UK collaboration focuses on taking sparse data – data where a significant amount of points are missing from the complete sets – or "noisy" data – data where a significant amount of variables could contribute to issues and changes in results – and making predictive models that fill in missing points with degrees of certainty and without having to undergo costly experimentation.

A new UK collaborative start-up is looking to use AI to predict missing data points in compound data. The collaboration is between Intellegens, a process modeling company spun out of Cambridge University, Optibrium, a provider of software solutions to the pharmaceutical industry, and the UK Medicines Discovery Catapult, a non-profit created by the UK government to support the creation of businesses in targeted sectors such as medicine.

Together the three entities have created a technology that will be able to predict missing values in sparse data sets to provide better direction in drug discovery and cut down on the amount of costly testing companies need to undertake. This has led to the collaboration being awarded £1m in funding from the UK government.

"There's a lot of hype about AI in drug discovery and it's building to a crescendo at the moment," says Matt Segall, chief executive officer and company director of Optibrium. "We've been looking at what tech will make a real difference. Lots of people are doing the same old things in the same old processes with shiny new toys. And there are also lots of same old toys being rebadged as new." Each member in the collaboration brings an essential tool to the project.

Intellegens developed the actual software used in making the predictions. Optibrium has experience in drug discovery and has developed a software setup called StarDrop that helps with the complete process for analysis and visualization of data.

And the Medicines Discovery Catapult organizes the collaboration while also undertaking database facilitation, benchmarking and access.

Intellegens originally developed its software in the material sciences field. It thought there could be other applications for it and wondered if drug discovery would be a candidate. After discussions with Optibrium, it was found to be a near perfect match.

The model focuses on taking sparse data – data where a significant amount of points are missing from the complete sets – or "noisy" data – data where a significant amount of variables could contribute to issues and changes in results – and making predictive models that fill in missing points with degrees of certainty and without having to undergo costly experimentation.

"The work translates beautifully to biology," says

Segall. "It's very noisy in biology. You can do the same test five times and get five different results. It's also very sparse. A big pharmaceutical company may have some data on a couple million compounds – all of which can have thousands of assays run on them such as different physical chemical properties or solubility. But for all of those different experimental data points, they'll only have measured only a handful per compound."

No compound will have had all assays run on it. And no assay will have been run on all compounds. Meanwhile biological experiments can produce different results from something as simple as how cells are handled, he adds.

The companies hope the new development can go some way towards solving this without involving significant amounts of expensive testing of compounds. A proof of concept study for the collaboration's system has been undertaken on an industry database. Results were favorable and have been submitted for peer review in the Journal of Chemical Information and Modeling.

Optibrium will be responsible for commercializing the technology once fully proven and developed. But that is still some degree of time away, according to Segall.

It is about a two year development program before fully going out to market, though there is a web app that customers in pharma as well as other verticals are paying to access, adds Gareth Conduit, chief technology officer and co-founder of Intellegens.

The full idea is that a company would be able to take its proprietary information on what happens to proteins when a certain drug is injected and combine that with other data such as publicly available sets to train a model and predict missing values.

This then means a company avoids having to conduct an experiment to check that particular process. "It can predict what value it would be," says Conduit. "You can also say: 'We want to activate this protein and deactivate that one. What would do that?' And the modeling should be able to propose a brand new chemical that could satisfy those targets."

Further uses in commercial situations could be for double-checking information. The model could go in and predict all values in a data set then compare that to what has been identified in a set. Those values the furthest away from what is in the sheet could point to potentially incorrect values – for example numbers that have been mistyped or lost in transcription – both common enough occurrences in drug discovery, he adds.

And the model could be used to identify which assay pairs could be of interest for further experiments. "You can say: 'There are ones in which if we perform one additional experiment here, it would give us a lot of information about that local chemical space to really help us extrapolate into that new domain and understand what is going on,'" says Conduit. "We can do experiment design to recommend what is most important to do to gain the most additional information out of each of the experiments the clients perform."

Overall the collaboration has a commercial advantage over others working in similar areas when it comes to data sets with sparse points. "That's where the company has the competitive edge," Conduit adds.

The Medicines Discovery Catapult recognized this. As part of its mandate it provides capabilities that small companies would otherwise be unable to access due to prohibitive costs and effort, according to John Overington, chief informatics officer at the Medicines Discovery Catapult.

In the case of Intellegens and Optibrium, the Catapult provides access to large databases and puts them in a format capable of being used for machine learning. "[We're] good at finding data and putting it in a form that feeds machine learning right away," he says. "For a company to do that in-house is a huge overhead. It would require a librarian type organization or mindset. Leveraging outside help is a good way to do that instead."

In the field of drug discovery there is not easy table of data for assays and compounds. The information may exist but it will be published as part of an in-depth study. Medicines Discovery Catapult has found a way to draw this data out and model it in a way machine learning algorithms can process.

"There are not these nice tables of data you can extract and put into a database. They tend to be published as one compound studied in-depth in a specific journal in that field. It's very hard to extract that data from the literature to understand and model it."

Medicines Discovery Catapult is only a couple months into the two year period of the £1m grant. During that time it will work with the companies before moving on to help other partners. By that point Optibrium and Intellegens shall have hopefully proven that their model is significantly more accurate than anything else out there. If it proves to be more than 10% more accurate than current models, it could create a quantum effect on productivity, Overington says.

"In olden days, which turns out to be the 1990s, people used to make compounds and then test them and that make and test cycle was very expensive. But it turns out that for a lot of tasks you're interested in something like solubility. So because solubility is a common feature, there's a lot of data out there for it. So people began to investigate predictive models that replace the experiment in many cases," he adds.

"The field has gone quite a long way in the development of predictive models but there's a limit to how accurate they are and they need to be more accurate to have a quantum effect [which is hopefully what Optibrium and Intellegens have done]."