RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry

Friday, 15th June 2018 - Royal Society of Chemistry at Burlington House, London, UK

Post-event Report on Speaker Presentations, written by Bursary Awardees

Introduction from the Organising Committee

Our first conference on aspects of Artificial Intelligence in Chemistry was held on Friday 15th June at The Royal Society of Chemistry in the Library of Burlington House. The conference was a year in the planning and surprised us by being sold out fully two months before the event with 115 delegates. The day itself was great for the quality of scientific talks and poster presentations, and also the chance to network with those interested in this highly exciting and fast-moving field. However, as highlighted in the first talk from Marwin Segler (BenevolentAI), artificial intelligence has been a part of chemistry going back more than half a century.

The success of our first Artificial Intelligence in Chemistry conference has led us to begin work on a second conference to take place in 2019, which we hope will be even bigger and better than our event this year. Watch this space!

The event had an extensive following on Twitter with the #RSC_AIChem hashtag being used 327 times.

Organising Committee: Nathan Brown, Phil Jones, Chris Swain

Presentation Summaries from Bursary Awardees

Speaker: **Marwin Segler, BenevolentAI**
Presentation: **Self-driving synthesis planning**
Written by: **Po-Chang Shih, University College London**

The great talk given by Dr Marwin Segler clearly gives a broad view of current progress of AI technology in the field of retrosynthesis. In the beginning of the talk, a story about the challenge of developing AI in retrosynthesis was introduced – encoding ~20,000 synthetic chemistry rules took 10 years. This is absolutely startling, as there are apparently more than such number of synthetic rules. To begin his journey of AI application in retrosynthesis to a wonderful land, he and his team picked up few factors that could make it work, such factors as focusing on more promising reactions and filtering infeasible reactions, although getting the rules to the machine and having efficient search initially challenged them. Their current findings are that applying deep neural networks and Monte Carlo Tree Search to the machine is able to create synthetic route(s) to a molecule, of which route(s) are not inferior to literature's routes. The success of machine-generated retrosynthesis sheds a light on AI's potential in drug discovery, although the wrap-ups of the talk lists current challenges that AI cannot do so far – no conditions of reactions, no yield prediction, cannot work well for natural products and cannot invent new reactions.

Speaker: **Nadine Schneider, Novartis**
Presentation: **Chemical topic modelling - an unsupervised approach to organize and explore chemical information**
Written by: **Robert Shaw, University of Sheffield**

Topic modelling is a useful AI technique for determining the co-occurrence of words, and from this building a probability distribution of words in topics. Schneider's talk demonstrated how this approach could be used instead for chemicals, classifying molecules by the common (or rare) fragments they contain. In an age where the amount of data, especially in chemical space, is unmanageably large, this approach provides a robust, unsupervised way of identifying chemically "interesting" compounds. Schneider demonstrated the efficacy of chemical topic modelling by showing how a 100-topic model could be built in around an hour for a database of 1.6 million molecules. This led to sensible and interpretable distinctions, highlighted by intuitive visualization tools.

Speaker: **Bob Sheridan, formerly Merck**
Presentation: **What I learned about machine learning - revisited**
Written by: **Shuzhe Wang, ETH Zurich**
Bob gave an insightful review of his experience in QSAR for the past 30 years. In that time, he made important contributions to the field, such as the promotion of random forest models and replacement of random train-test set split by split based on time. He showcased the performance comparison of random forest against gradient boosting tree and deep neural nets, based on their computational cost, accuracy and model size and concluded gradient boosting tree is the most cost-efficient approach to date. Following domain of applicability analysis aiming to assign error bars to predictions, he moved on to some remaining challenges on the interpretability of model outputs. These include assigning input descriptors importance and more so, atoms importance - mapping of activity readout back to subset of atoms in the input molecule. I really enjoyed Bob's talk, especially his emphasis on the importance of data quality over the complexity of model architecture.

Speaker: **Bob Sheridan, formerly Merck**
Presentation: **What I learned about machine learning - revisited**
Written by: **Miha Skalic, University Pompeu Fabra**
Dr. Sheridan provided insight into data science and predictive model development at Merck Sharp & Dohme. The talk was focused on the domain of QSAR and the presenter pointed out that the methods are between worthless and kinda working and more important than anything in the pipeline is the data. If one has insufficient data, no model optimization will help you. The three key points of the talk were:

1. Model development
2. Estimation of model reliability (assigning error bars)
3. Interpretability of the models

In terms of feature engineering they use features extracted from substructure (e. g. ECFP4/MACCS fingerprints), molecular properties or properties such as atom pairs descriptions. As predictive models they usually use one, either random forests, XGB-boost or neural networks, which if they are trained as multitask networks perform better. For the task determining confidence they rely on DA metrics - metrics that evaluate differences between observations and predictions. Finally, the presenter described two methods used for interpretability of models: (i) Universal descriptor importance generator (UDIG) and (ii) coloring by atoms. UDIG uses permutations of descriptors while coloring methods hide atoms in computation of substructure descriptions.

Speaker: **Ola Engkvist, AstraZeneca**
Presentation: **Molecular *de novo* design through deep learning**
Written by: **Fergus Imrie, University of Oxford**
Ola’s talk provided a fantastic summary of the current capabilities of molecular *de novo* design, whilst also hinting at what might come next. Ola began his talk by motivating *de novo* design and discussing its potential impact on the industry. It was made clear that attempts to enumerate chemical space would not succeed and thus developing methods to sample or search this space is of crucial importance. The speed with which the field was advancing towards these goals was made apparent through a brief outline of some of the recent literature, a recurring theme throughout the day. After a primer on neural networks, Ola described how methods from natural language processing could intuitively be applied to molecular structure generation by using SMILES strings. A compelling demonstration of this technique showed exactly how a trained model produced a new molecule. Ideally the generated molecules should have desirable properties, in particular ones that the user can specify. Through the use of reinforcement learning, Ola described how a generative model can be incentivized to produce molecules with specific properties. Ola then addressed some of the criticisms of generative models, two majors ones being that molecules are not diverse and are not synthetically feasible. Ola showed that the generated molecules followed the properties of the dataset used for training in both cases, but conceded that evaluating generative algorithms was

challenging and needed thought. Finally, Ola presented his thoughts for the future of chemistry, and how this would be influenced by automation and AI. A very engaging talk and exciting times lay ahead!

Speaker: **Willem van Hoorn, *Exscientia***
Presentation: **Scaling *de novo* design, from single target to disease portfolio**
Written by: **Lee Steinberg, University of Southampton**
Willem Van Hoorn spoke about the technology that *Exscientia* applies in the automatic design of patentable compounds. In the words of Willem, “You would not go looking for mountains in The Netherlands”. Equivalently, when looking for potential candidates, we should first search the regions of chemical space that we would expect to be most fruitful (maybe we should try looking in Switzerland for our mountains). Furthermore, Willem introduced his idea of the “Centaur Chemist”, where humans design strategy and assess progress, but computers do the running. This leads to an environment where the scientist and computer work in tandem, and this methodology was shown to reduce the number of molecules assessed in a normal target search from 500 to 250, with the time taken reduced by more than half. Lastly, Willem ended with a variant of a quotation that will likely remain with the conference attendees: “Artificial intelligence will not replace chemists – but chemists who don’t use AI will be replaced by those who do”.

Speaker: **Ella Gale, Bristol University**
Presentation: **Investigating clusters in solvent data using K-means**
Written by: **José Jiménez, University Pompeu Fabra**
The need for investigating solvent data arises from several factors. Historically, solvent choice was based on history or previous best practice in the lab bench. However, changes may be needed in order either to save money, be more environmentally friendly or safer. In this study, the speaker focused in the application of Organic Electrolyte Solutions (OES), that is, the best co-solvent choice for the dissolution of cellulose in paper and plastic processing. In order to tackle the problem, there is a clear need to map the solvent landscape, and previous attempts have tried to do so by Principal Component Analysis means, which is the currently used most method. There is always a discrepancy between the problems unique to chemistry when applying machine-learning algorithms, namely few labeled and incomplete examples. The speaker then focused on the chemical properties that are best predictors of solvent properties, and for so, used a solvent, a chemical group and a structure database, with different parameters each. The technique used to tackle the problem was a derivative of the well known k-means++ method and det-K, a way for determining the best value of clusters. In particular, they use a hierarchical approach where clusters are recursively found until eventually they are pure. She found that her data was best described by 2 clusters according to the det-K method, and identified items in the first like alkanes and nitriles, while in the second benzene and molecules with double rings were found. The physical properties that cause most of the separation were molecular weight, melting point, boiling point and molar volume. Another clustering with k=2 was made and found the most important properties to be dipole moment, dielectric constant and molar volume. Hierarchical k-cluster, on the other hand, was shown to separate solvent data into functional groups, allowing for novel re-arrangement. Further work encompasses repeating analysis in order to compare with the previous PCA method, and with solvents that feature incomplete data.

Speaker: **Colin Batchelor, Royal Society of Chemistry**
Presentation: **Deep learning and chemical data**
Written by: **Caroline Bushdid, Institut de Chimie de Nice**
Colin Batchelor’s talk entitled “Deep learning and chemical data” presented the results of using deep artificial neural networks (DNN) to analyse spectra and unstructured text related to chemistry. In particular, he presented this technique applied to three topics: NMR spectra analysis, Chemical named-entity recognition and protein/molecule relationship extraction.

In the case of NMR, the spectra were given as an input in the form of a vector, and the output was the presence of certain functional groups. He found that DNN outperformed classical Random Forest models and the model was capable of predicting accurately functional groups such as Iodine, Nitro groups and even aliphatic groups, but natural products identification remained challenging. Overall, deep learning did not perform as good as humans in this task, likely due to the need for more data.

In contrast, when DL was used to recognise chemical names and patterns, the numerical model was found to perform as well as humans. The nature of the input data (i.e. unstructured text) made the task more difficult than expected. The limit in the performance of the model was found to be mainly due to Humans disagreeing in the annotations. He explored using a word-base system as compared to a character-based system and found that by using an ensemble of these techniques, the best results were obtained.

Lastly, he used DL to extract relationships on how molecules affect proteins. The input consisted of text (for example from CHEMPROT) which resulted in the output of a vector which held information such as “no relation”, “up regulation”, “down regulation”, “agonist”, “antagonist”, etc… The model performed in this case worse than humans, underlying that, like in the case of NMR spectra, the task is much harder for DL.

Speaker: **Darren Green, GlaxoSmithKline**
Presentation: **Automation, analytics and AI**
Written by: **Sam Munday, University of Southampton**
Dr Darren Greens talk on automation, analytics and AI highlighted the role that humans will continue to have alongside an increasingly powerful AI in drug discovery. In comparison to other speakers, he introduced his belief that the chemical community has yet to engage with true AI systems, and that we are currently still applying automation techniques, rather than intelligent algorithms.

Darren Green introduced with an overview of his experience within the field of molecular design at GSK UK. He describes the changing role of the medicinal chemist, intuition borne from experience gives way to a data driven systematic search for possible lead compounds. He highlights how this will lead to an increase in productivity as the chemist will use models built from GSK’s vast database to refine and direct their search, leading to less unsuccessful paths being followed.

Dr Green’s view on the future of AI in chemistry drew from his experiences over the past 10 years, including the rise of “evolutionary algorithms” up to deep neural network techniques. His talk was funny and engaging and a great way to round off what was an already fantastic event.