Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Medicines Discovery Catapult

AI^3 Science Discovery Network+ & Medicines Discovery Catapult: AI in Drug Discovery & Drug Safety Workshop
06/03/2019
Medicines Discovery Catapult
Alderley Park Conference Centre, Nether Alderley, Macclesfield, SK10 4TG

Dr Wendy A. Warr
Wendy Warr & Associates

18/04/2019

AI3SDEventSeries:Report-7

AI^3 Science Discovery Network+ & Medicines Discovery Catapult: AI in Drug Discovery & Drug Safety Workshop
AI3SDEventSeries:Report-7
18/04/2019
DOI: 10.5258/SOTON/P0008
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 Event Details

Title	AI in Drug Discovery & Drug Safety Workshop
Organisers	AI^3 Science Discovery Network+ and Medicines Discovery Catapult
Dates	06/03/2019
Programme	Programme
No. Participants	32
Location	Alderley Park Conference Centre, Nether Alderley, Macclesfield, SK10 4TG
Organisation Committee	Georgina Hett, Medicines Discovery Catapult & Dr Samantha Kanza, AI^3 Science Discovery Network+
Chairs	Professor John Overington, Medicines Discovery Catapult & Professor Jeremy Frey, AI^3 Science Discovery Network+

2 Introduction

2.1 The Network+

The Artificial Intelligence and Augmented Intelligence for Automated Investigations (AI^3) for Scientific Discovery Network+ (AI^{3}SD, http://www.ai3sd.org/) is funded by EPSRC (https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/S000356/10), and hosted by the University of Southampton. The network aims to bring together researchers looking to show how cutting edge artificial and augmented intelligence technologies can be used to push the boundaries of scientific discovery.

The recent success of particular types of machine learning (e.g., deep neutral nets) has excited the interest of the scientific community in delivering insight into the complexity of the real world. This type of approach requires massive amounts of data to be trained. Traditional approaches to scientific discovery work with relatively small amounts of often uncertain data which are distilled by human insight to yield predictions and testable theories which may evolve as new data become available. The impact of "larger data" parallels the reality that almost all science now depends on computational assistance. Nevertheless the quantity of quality data needed to train the new AI systems is not directly available even with recent advances in automation. As a basis for the network the team at the University of Southampton proposes to use "amplification by simulation" as a key element of the cycle of automated experiments, simulation, AI learning, prediction, comparison, design, and further experiments, to create the environment in which leading AI developments can be applied to chemical and materials discovery.

2.2 Medicines Discovery Catapult

Catapults are not-for profit, independent centres which connect businesses with the U.K.'s research and academic communities. The Medicines Discovery Catapult (MDC, https://md.catapult.org.uk) is a national facility connecting the UK community to accelerate innovative drug discovery. It is an independent not-for-profit company, funded by Innovate UK (https://www.gov.uk/government/organisations/innovate-uk), to bring together the large and diverse sector of industry, academia, charities, technologist, finance companies, small and medium enterprises (SMEs), and start-ups.

3 The Event

Drug discovery is a long and long-term scientific investigation involving interdisciplinary research methods coupled with large heterogeneous datasets. The research and data space in this area is vast, and AI^{3}SD and MDC believe that the use of AI and machine learning technologies can help spur on advances in this domain. The current workshop was designed to draw together those with a keen interest in using AI and machine learning technologies in the domain of drug discovery, both to aid future drug discovery, and to help improve drug safety. AI^{3}SD firmly believes that interdisciplinary collaboration is the key to many of these advances. At the workshop, keynote talks were interspersed with general group discussions and working groups around the key topics that arose.

4 Introduction to MDC Informatics

CATAPULT
Medicines Discovery
Accelerating innovative drug discovery
Connecting the UK drug discovery community
Transforming ideas into better medicines for patients, faster
CATAPULT
Medicines Discovery
md.cat

John P. Overington was the first speaker. He is the Chief Informatics Officer, Medicines Discovery Catapult. He leads the development and application of informatics approaches to promote and support innovative, fast-to-patient drug discovery in the United Kingdom through collaborative projects across the applied R&D community. John described the informatics projects as “plumbing” and listed the current ones (not including two Innovate UK grants that have not yet been announced):

- theCollaboratory: inter-organisational data transfer from ELNs
- VESPA: a “multiscale” Bayesian network approach to variant effect prediction
- definitive *a priori* target validation using Mendelian randomisation
- AssayNet: directed graph of bioassays from gene to clinical trial and translational predictive models, annotated with supplier, academic lab, etc.
- CRISPY: MDC’s collaborative intelligence platform, with AI-enabled search across the UK life science sector
- MPO-constrained optimisation using generative adversarial networks, computer vision, etc.
- Deep ADMET: SAR data “on demand” combining Optibrium’s StarDrop and Intellegens’ Alchemite (using deep learning),[1] working with MDC under an Innovate UK grant
- drug combinations to improve efficacy, resistance, and safety
- addressing new target classes (the RNA-world, transporters, channels, etc.)
- cryptic pathogenic infectious agents.

John gave more detail about CRISPY for collaborative intelligence.[2] It is difficult to identify people and organisations with specific skills and experience, and expert curation of resources is slow and expensive. CRISPY builds a live knowledge graph of UK drug discovery assets and capability, using data from Companies House, the Charities Commission, patents, published papers, grant applications, UK universities, the Financial Control Authority, theses, the British Private Equity & Venture Capital Association, angel networks, professional societies, consultant networks, and so on. Natural language processing, named entity recognition, the software word2vec,[3] and faceted searching are used to provide a focused search engine for UK life science.

Search results can be visualised in various types of display. John showed some bar charts, and also hot spots for AI on a map of the United Kingdom. Most of the companies involved are around London and Cambridge but there is also one around Alderley Edge! Search terms are entered in a Google-like graphical user interface. John looked for “surface plasmon resonance”. The system does not need an ontology: word2vec is able to find that “SPR”, for example, is a synonym for “surface plasmon resonance”. CRISPY returns a list of terms that can be selected (or deselected). John ticked five of the boxes for terms related to surface plasmon resonance, and obtained a table of the URLs and titles of 1,000 hits (from 40 different organisations) out of a total of 311,328 records. The search took only 1.49 seconds. More detail can be obtained for each hit. A preview column is also available in the table enabling a preview with marked-up text to be displayed on top of the table for a selected hit.

The system will make it easier to find collaborators, to “fill holes” in a project, and to study competitors. It is part of MDC’s efforts to disrupt traditional skill sets, in a positive way, and, for example, to help chemists make the right compounds. Currently CRISPY is an internal tool for MDC staff, but John mentioned that he would be keen to find collaborators to apply and extend the system’s content and uses.

5 Using Machine Learning to Drive Reaction-based *de Novo* Design

Val Gillet is a Professor of Chemoinformatics at the University of Sheffield where she heads the Chemoinformatics Research Group. She recommended a review[4] on *de novo* design. There was a flurry of attempts at *de novo* design in the 1990s. Three phases are involved in the design: making molecules, searching a vast chemical space, and scoring the molecules. In the 1990s restricted sampling of chemical space and scoring were difficult and most approaches were agnostic of synthesis.

Interest in *de novo* design went into abeyance, but has re-emerged recently. Current approaches are reaction based, using rules, and generative models, using AI. AI methods typically use SMILES and no attention is paid to synthetic accessibility, except in that the system is trained on databases of molecules. Val’s team uses reaction vectors (RVs) in an AI approach. RVs are counts of atom pairs removed from reactants and gained in products. Atom-pair descriptors of reaction components can be applied *in silico* to generate new products. In reality, a more complex descriptor is needed. A reaction centre one step out from the atom-pair is used in a sophisticated forward prediction algorithm, followed by fast structure generation. The reaction vectors are applied to previously unseen starting materials in order to suggest novel molecules for synthesis.[5] Each transformation that is applied has a precedent in the literature, and thus a high degree of confidence is established in the synthetic feasibility of the resulting molecules.

The approach has been implemented in KNIME and was validated by reproducing known reactions. For a wide range of reaction types 95% are well represented. Accuracy is not quite so good for more complex reaction types such as rearrangements: about 90% for 6000 reactions extracted from the *Journal of Medicinal Chemistry.* The literature source of the reactions is stored with the library of potential products.

The reaction based approach initially required datasets that were hand-crafted but a large collection of reactions has now been made publicly available[6,7] by NextMove Software. This database from US patents is referred to here as USPD. Schneider *et al.* have used text mining to extract 1.15 million unique, whole reaction schemes, including reaction roles and yields, from pharmaceutical patents. The reactions were assigned to well-known reaction types such as Wittig olefination or Buchwald-Hartwig amination using an expert system, and analysed to show the evolution of reaction types over time.

The Sheffield team have a process for cleaning the NextMove dataset prior to making reaction vectors. Stoichiometrically balanced reactions are needed, atom mapping is carried out, and catalysts etc. are removed. RVs are then calculated and validation carried out to see if the known product of a reaction can be generated. From the 1.8 million reactions,[6] 92,530 unique and validated RVs were obtained that also had classification labels. The distribution of the number of reactions per RV was highly skewed.

Val showed a flowchart for fully enumerated fragment expansion for a single step reaction from a starting set of 771 DSPL fragments (https://www.diamond.ac.uk/Instruments/Mx/Fragment-Screening/Fragment-Libraries.html), 24,000 Sigma Aldrich reagents and the USPD derived RVs, which produced 6.5 million virtual products. Multistep reactions cause a combinatorial explosion. The Sheffield team sought to mine USPD to find reaction classes more likely to be applied to a specific starting material. Their reaction recommender program does this, and reduces the number of predictions to synthetically accessible products. For a given starting material, the applicable RVs are limited to those that belong to reaction classes suggested by the recommender. The recommender therefore requires that reactions are grouped by class.

A classification procedure was first developed using reactions extracted from the USPD which were represented by RVs encoded as dynamic fingerprints. This approach is similar to that reported[8] by NextMove and Novartis, but Val's team modified the classification scheme to be compatible with the RVs. Reactions were grouped into a four-layer hierarchy. Validation involved an external dataset of 25,000 unseen RVs from USPD. Performance was similar to that reported by the Novartis team but there were 336 classes not 50 and these were combined with a confidence predictor.

The reaction recommender starts with starting materials and reaction class and aims to take account of features outside the reaction centre. The hope is to reduce the size of the chemical space while increasing the number of synthetically accessible molecules. This is now a multi-label classification problem. A starting set of classified reactions is used. Each reaction is represented by only a starting material. Starting materials represented by identical descriptors are merged and reaction labels are appended.

The recommender is trained on 1.1 million cleaned reactions from USPD. Starting materials and reaction class are extracted, duplicates are removed, descriptors are generated, and training and validation are carried out. Extensive experimentation is necessary to make the best decision on reaction class layers, learning methods etc. Two layers appear to be better than three. Val presented a table of the different types of descriptor used.

Retrospective validation was carried out with 26,000 reactions from the *Journal of Medicinal Chemistry* not seen by the model. The reactions were cleaned and classified, and then starting materials were extracted and duplicates removed. The trained recommender was used for each starting material and the reaction suggested was compared with the true class. Val presented some results. In some cases there was no recommendation or a wrong recommendation. The results could be improved if sufficient training data were used. Some wrong recommendations were due to a starting material having more than one reaction centre and the "incorrect" one being suggested. These could be corrected by applying the recommended reaction and making a new recommendation for that product.

In a second validation, behaviours with and without the recommender were compared, starting with fragments from DPSL (https://www.diamond.ac.uk/Instruments/Mx/Fragment-Screening/Fragment-Libraries.html) and reagents from Sigma Aldrich. The results showed that the recommender is successful in reducing the number of products, and in saving time. The synthetic accessibility was measured using the MOE rsynth descriptor (https://www.macinchem.org/reviews/moe-review2.php) and mean SA score.[9] The number of products decreases and the synthetic accessibility scores improve as the classification labelling becomes more fine-grained.

This work is being carried in collaboration with Mike Bodkin and others at Evotec. The team is currently exploring the use of the recommender in augmented *de novo* design and in fully automated design.

6 Re-energising Small Molecule Drug Discovery

Willem van Hoorn (pictured left with the logo modified from Darwin's tree of life) is Chief Decision Scientist at Exscientia (https://www.exscientia.co.uk/). He said that although 90% of drugs are small molecules, and 50% of clinical trials are for small molecules, small molecule drug discovery remains inefficient: it takes five years to get 2500 compounds from idea to drug candidate.[10] Hit to candidate is the most expensive part of drug discovery in terms of cost of capital. Exscientia's vision is to produce a candidate from 500 compounds in just 1.5 years.

Pioneering automated drug design methodologies[11] developed by Professor Andrew Hopkins and other researchers at the University of Dundee led to the spin out of Exscientia in 2012. The company now has an office in Oxford as well as the original one in Dundee. It has become a scale-up, not a start-up company, with clients' molecules heading towards the clinic. The company's AI-driven systems actively learn best practice from vast repositories of discovery data and are further enhanced with knowledge acquired from seasoned drug hunters. The proprietary AI design module uses a genetic algorithm to predict new structures; machine learning did not work as well or as quickly.

Novel compounds prioritised for synthesis by the AI systems simultaneously balance potency, selectivity and pharmacokinetic criteria in order to deliver successful experimental outcomes. By applying a rapid design-make-test cycle, the Exscientia AI platform actively learns from the preceding experimental results and rapidly evolves compounds towards the desired candidate criteria. Exscientia systems learn from both existing data resources and experimental data from each design cycle. The company is developing single-target small molecules as well as compounds with more challenging target product profiles, through a novel bispecific small molecule strategy (compounds with dual pharmacology in an integrated pharmacophore) and phenotypic-driven drug design.

In 2015, the pharmaceutical firm Sunovion asked 10 of its chemists to play a game[12] to see who could discover the best leads for new drugs. The chemists were presented with hundreds of chemical structures, just 10 of which were labelled with information on their biological effects. The experts had to select other molecules that could turn out to be drug candidates. The 11th player was an Exscientia computer algorithm. The chemists took several hours, the computer only milliseconds. Only one chemist out of the 10 beat the machine. By 2017 the machine was as good as the best chemist.

Nevertheless the human is not redundant. In 1996, the first chess match between world chess champion Garry Kasparov and the IBM supercomputer Deep Blue was won by Kasparov. The second, in 1997, was won by Deep Blue. The 1997 match was the first defeat of a reigning world chess champion by a computer under tournament conditions. In 2016, in a five-game Go match between 18-time world champion Lee Sedol and AlphaGo, a computer Go program developed by Google DeepMind, AlphaGo won all but the fourth game. Advanced Chess, sometimes called centaur chess, was introduced by Garry Kasparov in 1998. A centaur chess player is one who plays the game by marrying human intuition, creativity and empathy with a computer's brute-force ability to remember and calculate a staggering number of chess moves. By 2008 it had been shown that, not surprisingly, a centaur beats the solo human, but less surprisingly, a centaur beats the solo computer (under certain time constraints).

In Exscientia's Centaur Chemist technology the machine and the chemist work together. The principle is similar to how a human would learn, but the AI process is far more effective at identifying and assimilating multiple subtle and complex trends to balance potency, selectivity and pharmacokinetic criteria. For example, the decision to focus on, say, hERG demands human strategic thinking, but finding compounds that fulfil that brief is better done by the AI. Willem presented some case studies.

In a collaborative psychiatric drug design project between Exscientia and Sunovion, the research focused on developing novel approaches to compound design by analysing data arising from phenotypic drug discovery. This collaboration builds upon Exscientia's delivery of novel bispecific compounds that combine activities at the GPCR and ion channel target families. In the work reported by Willem, one target had literature and patent SAR, but the other had limited literature SAR. Exscientia's clients had worked previously on the second target (in a single-target approach) and had run a high throughput screen. They shared the screening data so that Exscientia could build a model. In the lead identification stage 20 compounds, on average, were made per two-week cycle. Willem presented a table showing which compounds were taken into lead optimisation. Just over 100 compounds were made in two series. There were 80 further compounds for each prioritised scaffold.

Exscientia and Sumitomo Dainippon researchers are working in partnership to deliver novel multi-target small molecules that have potential to deliver enhanced therapeutic performance in the treatment of psychiatric diseases. The first compound passed over to Sumitomo Dainippon for further internal development is a bispecific, dual-agonist compound that selectively activates two GPCR receptors from two distinct families. The accelerated project delivered the molecule that fulfilled development quality criteria in only 12 months. Starting from a product concept, fewer than 400 compounds needed to be synthesized[13] to identify molecules that matched the required development criteria. Clinical trials will begin for one compound in 2019.

It has been said that AI will not replace chemists but chemists who do not use AI will be replaced by those who do. The same maxim can be applied to companies: those that use AI will out-compete the ones who do not. The positive message is that chemists and companies who do not use AI can become chemists and companies who do.

7 Understanding the holes in the metabolome

Accelerating innovative drug discovery
Connecting the UK drug discovery community
ideas into better medicines for patients, faster
APULT
md.catap

Nicola Richmond is a director in the Artificial Intelligence and Machine Learning team at GlaxoSmithKline (GSK). She described some work carried out by Casey Wojcik, a postgraduate researcher at Stanford University who worked at GSK under the Cambridge Mathematics Placement scheme after finishing Part III of the Mathematical Tripos at the University of Cambridge.

GSK's motivation for the work was to develop objective approaches to analysing metabolomics data. Metabolomics is the scientific study of chemical processes involving metabolites, the small molecule intermediates and products of metabolism. GSK wanted to find out if they could add more insights with topological data analysis (TDA) in an objective way. Nicola showed a typical "hairball" network: the simplicity and beauty of node-link diagrams turns into clutter and confusion when the number of nodes and links gets too high.

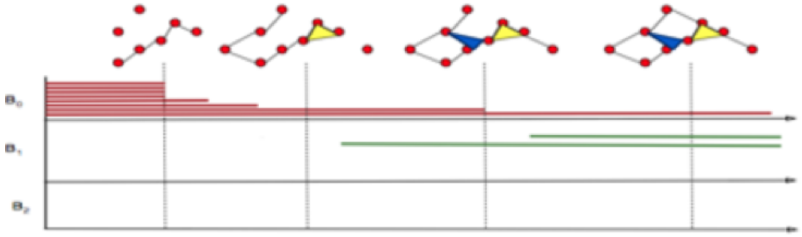
TDA is a mathematical approach to the analysis of datasets using techniques from topology. TDA provides a general framework to analyse datasets that are high-dimensional, incomplete and noisy, in a manner that is robust to noise. TDA has combined algebraic topology and other tools from pure mathematics to allow mathematically rigorous study of shape.

Topology classifies spaces based on their invariant properties. A well-known example is the similarity between a doughnut and a coffee mug: you can deform a doughnut into a coffee mug. The shapes can stretch or deform but not break. Data also have structure and shape. TDA can be used to study the shape of data, in particular, connectedness and voids.

A simplicial complex is a geometrical representation of a topological space which is realised as a union of simplices, such as points (0-simplex), line segments (1-simplex), triangles (2-simplex), tetrahedrons (3-simplex) and other higher dimensional cousins. Simplicial complexes provide a simple combinatorial way to describe certain topological spaces.

You can build a family of simplicial complexes from a point cloud and study the complexes in order to study the point cloud. Persistent homology is a method for computing topological features of a space at different spatial resolutions. More persistent features are detected over a wide range of spatial scales and are deemed more likely to represent true features of the underlying space rather than artefacts of sampling, noise, or particular choice of parameters.

Topological spaces can be characterised by using topological invariants: algebraic objects which are invariant under homeomorphisms. Homology, one of the topological invariants, is a mechanism for counting the number of n-dimensional holes. Persistent homology is the computational implementation of homology, allowing you to describe a simplicial complex in terms of n-dimensional holes. These holes are expressed by Betti numbers, B_i, where B_{0} corresponds to the number of connected components, B_{1} corresponds to the number of planar holes, B_{2} corresponds to the number of voids in solid objects (2-dimensional holes), and so on. Persistent homology is an incremental construction of the final filtered simplicial complex. In the figure below, the Betti numbers are visualised through barcodes. Barcodes are a graphical representation of Betti generators whose horizontal axis corresponds to the filtration parameter and whose vertical axis represents ordering of homology generators. Barcodes could be useful descriptors for input to machine learning approaches.



A weighted network is a network where the ties among nodes have weights assigned to them. In a social network, the weights represent how well people know each other. In a metabolic network the weights relate to reaction rates. Topological features are important. While the null hypothesis (H0) in any experiment or research project is that the connection or conclusion suggested by the experiment is false, the alternative hypothesis (H1) is always the assertion that there is a meaningful connection to be investigated. Open source software called jHOLES (https://www.jholes.eu/home.html) is available for computing persistent homology.[14]

Nicola described Casey's project using TDA and metabolic networks using metabolomics data on Panobinostat, a drug by Novartis for the treatment of multiple myeloma and other cancers. It is a histone deacetylase inhibitor, which promotes cell-cycle arrest and apoptosis of tumour cells *via* multiple pathways.

Casey used the Kyoto Encyclopaedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/) to make a metabolic network and computed correlation coefficients for each metabolite with panobinostat and constructed a network with edges weighted by the average correlation coefficients of the end vertices. Next jHOLES was used to study cycles, with visualisation in Cytoscape (https://cytoscape.org/), an open source software tool for visualising complex networks and integrating these with any type of attribute data. The results were grouped by cycle and coloured by pathway. It could be seen that the first few cycles most strongly correlate with treatment. One pathway was singled out: a tricarboxylic acid (TCA) cycle for cysteine and methionine metabolism. That cycle is known to interact strongly with the drug. The practitioner can understand this. Note that this was a purely data-driven identification.

Nicola outlined the benefits of the methodology. It singles out strongly correlating subnetworks which are small enough to inspect manually. It can detect affected pathways, key metabolites, and cross-talk between pathways. Data imputation through diffusion allowed for unmeasured metabolites.

In summary, persistent homology has a strong theoretical foundation and is useful for understanding changing network structure, especially in metabolism. It is easy to compute and simplifies the task of analysing complex networks. GSK wants to understand systems biology. A lot of the data are present as omics datasets, and images from phenotypic screening could be used to advantage. Pure mathematics has a role to play in systems biology.

8 Discussion

Interspersed with the talks were discussions among six groups of attendees. Professor Jeremey Frey of the University of Southampton and AI^{3}SD introduced the discussions; and Samantha Kanza, the AI^{3}SD Network Coordinator, ably documented the feedback gathered on all the flipcharts. In Jeremy's introduction he said that we have some (mediocre) property prediction algorithms and some budding *de novo* design algorithms, but the number of applications that combine the two is limited. We need to make molecules with the right properties.

As a discipline, chemistry is splitting into biochemistry, molecular biology, nanotechnology, and so on. These sub-disciplines need to be linked so that data can be transferred across them. The subject of the discussions was the way that AI can transform science, and, in particular, applications of AI in drug discovery and use. By the end of the day Jeremy hoped to have a summary of the issues, and an agenda for making progress in addressing them.

Samantha analysed the feedback from each of the six groups and the ideas summarised on various flipcharts. She drew them together into seven themes: skills and training; data quality, access, collation, and reproducibility; interdisciplinary data sharing and interoperability; optimisation of the drug discovery process (reducing attrition, failure and cost); explainable AI and models; making decisions, and trustworthiness of machine learning; and miscellaneous other areas such as automated synthesis and candidate selection.

9 Conclusions

We, the attendees wanted to consolidate a view on the impact that AI will have over the next 5-10 years on the drug discovery process. We need tips for embarking on this journey. What does a good, AI-driven project look like? What are the outputs, constraints and limitations? Are the data reproducibile? Different communities have developed methods for handling "big data" but in drug discovery we do not always have big data. Instead we have complicated data on a few things. There are gaps in the suite of tools we can use. Lists of standards, resources and ontologies are needed. We must focus on the biology as well as the chemistry. AI is not really new, but it now has a lot of momentum, and it has caused some damage. We need a roadmap with realistic timescales. Since we are "punching above our weight" for AI, and we have a great range of talent, why do we not make greater demands on the government to support and retain this capability?

10 Related Events

For those who are interested in getting involved with the AI and drug discovery community and attending related events there are some additional events that cover similar areas of interest.

2nd RSC-BMCS / RSC-CICAG Artificial Intelligence in Chemistry in Fitzwilliam College, Cambridge (2nd-3rd September 2019) - https://www.maggichurchouseevents.co.uk/bmcs/AI-2019.htm

20th SCI/RSC Medicinal Chemistry Symposium in Churchill College, Cambridge (8th-11th September 2019) - http://www.rsc.org/events/detail/35363/20th-sci-rsc-medicinal-chemistry-symposium

Upcoming events of interest can be found on the AI3SD website events page.
http://www.ai3sd.org/events/ai3sd-events
http://www.ai3sd.org/events/events-of-interest

11 References

References

(1) Whitehead, T. M.; Irwin, B. W. J.; Hunt, P. A.; Segall, M. D.; Conduit, G. J. Imputation of Assay Bioactivity Data using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59* (3), 1197–1204.

(2) Markova, D.; McArthur, A., *Collaborative Intelligence. Thinking with People Who Think Differently*; Penguin Random House: New York, NY, 2015.

(3) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013., arXiv.org e-Print archive. https://arxiv.org/abs/1301.3781 (accessed March 18, 2019).

(4) Hartenfeller, M.; Schneider, G. Enabling future drug discovery by de novo design. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (5), 742–759.

(5) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to de Novo Design using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49* (5), 1163–1184.

(6) Lowe, D. M. Chemical reactions from US patents (1976-Sep2016)., https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed March 20, 2019).

(7) Lowe, D. M. Extraction of chemical structures and reactions from the literature., Ph. D. Thesis, University of Cambridge. https://www.repository.cam.ac.uk/bitstream/handle/1810/244727/lowethesis.pdf?sequence=1&isAllowed=y (accessed March 20, 2019).

(8) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **2016**, *59* (9), 4385–4402.

(9) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(10) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203–214.

(11) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X.-P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **2012**, *492* (7428), 215–220.

(12) Mullard, A. The drug-maker's guide to the galaxy. *Nature* **2017**, *549* (7673), 445–447.

(13) Yoshinaga, H.; Uemachi, H.; Ohno, T.; Besnard, J. Preparation of 2,6-disubstituted pyridine compounds for the treatment of symptoms associated with anxiety disorders., WO2018168738A1, 2018.

(14) Binchi, J.; Merelli, E.; Rucco, M.; Petri, G.; Vaccarino, F. jHoles: A Tool for Understanding Biological Complex Networks via Clique Weight Rank Persistent Homology. *Electronic Notes in Theoretical Computer Science* **2014**, *306*, 5–18.