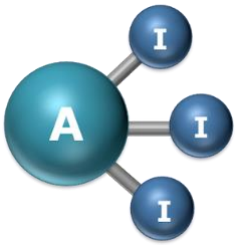

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

ISTCP 2019: 10th Triennial Congress of the International Society for Theoretical Chemical Physics
11-17 July 2019
Tromsø, Norway

Dr J. Grant Hill
University of Sheffield

Publication Date: 07/08/2019

AI3SD-Event-Series:Report-12

ISTCP2019
AI3SD-Event-Series:Report-12
Publication Date: 07/08/2019
DOI: 10.5258/SOTON/P0016
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*

Table of Contents

1. Event Details

Title	ISTCP 2019
Organisers	International Society for Theoretical Chemical Physics
Dates	11-17 July 2019
Programme	Programme
No Participants	550
Location	Clarion Hotel The Edge, Tromsø, Norway
Local Organiser	Prof. Kenneth Ruud / Hylleraas Centre of Quantum Molecular Sciences UiT The Arctic University of Norway
Sponsors	Research Council of Norway, University of Tromso, Hylleraas Centre, International Journal of Quantum Chemistry, ACS OMEGA, Elsevier, Journal of Chemical Theory and Computation, The Journal of Physical Chemistry Letters, The Journal of Physical Chemistry A/B/C, Physical Chemistry Chemical Physics.

2. Event Summary and Format

This seven-day conference mainly consisted of a mixture of plenary lectures and parallel presentation sessions, with two poster sessions and a session on European Research Council funding. Each day had two coffee breaks and time for lunch, which, combined with the poster sessions, allowed time for networking. The conference was broad in its overall scope – covering most aspects of theoretical chemical physics / physical chemistry, but three sessions were dedicated to machine learning / AI. A number of talks within other sessions also contained significant AI for chemical discovery content too.

3. Event Background

Congresses of the International Society for Theoretical Chemical Physics are held every three years, with the most recent marking the tenth congress. The Society was founded by Janos Ladik with the aim of showcasing the achievement and advances of all areas of theoretical chemical physics. An emphasis is placed on the interaction between experimental and theoretical methods. The conference website lists the current directors and national representatives of the Society.

4. Talks

The talks of interest to the Network were mostly clustered into sessions with a stated focus of Machine Learning, although some talks within the Physical Organic Chemistry and Catalysis sessions were also relevant. There was a large variety of other talks detailing the development and application of computational chemistry methods, but the following focusses on machine learning/AI specific elements.

Machine Learning Session 1

Pavlo Dral (Xiamen University) began the session with an overview of how his group have been using machine learning (ML) methods with a goal of reaching accuracy similar to ab initio methods, but with the speed of a molecular mechanics approach.

He discussed how they have found that kernel ridge regression methods tend to give greater accuracy than neural networks for the relatively small amounts of data available from ab initio computation. He briefly presented the MLatom program package [1] [2], which has been designed for carrying out simulation of atomistic systems with ML algorithms. The Δ-ML approach was also briefly discussed, where inexpensive semiempirical quantum chemical calculations are "corrected" using a machine learning model [3]. Once trained, this was able to reproduce density functional theory results in a small fraction of the time. Finally, efforts to reach spectroscopic (wavenumber) accuracy for vibrational spectra were presented in the context of computational astrochemistry [4]. A potentially important advance here is the introduction of structure-based sampling, which produces better results than the random sampling typically employed in ML frameworks.

Cecilia Clementi (Rice University) presented her group's work on *ML models for biomolecular dynamics*. In biophysics there is a problem of different time and size scales, and the question under consideration in this talk was how to bridge atomic molecular dynamics with coarse-graining. There are multiple ways to carry out this dimensionality reduction, with the traditional approach usually based on intuition and detailed knowledge of a system. A neural network was used to produce an energy function (and its parameters) that minimised the loss between an atomistic approach and the coarse-grained model [5]. Prof. Clementi discussed a major challenge for the future, which is to design a transferable coarse-grained model.

Marivi Fernández-Serra (Stony Brook University) gave a talk on *how ML can be used to produce a highly accurate exchange and correlation functional of the electronic density*. The aim here is not to bypass density functional theory (DFT), but to improve its accuracy using ML. This was a second example of a "delta-approach", in this case a standard DFT result is augmented by a machine learned correction to reproduce the exact electron density [6]. The resulting code, named NeuralXC, can be interfaced to a number of popular electronic structure codes [7]. It was noted during the talk that a model trained on ethanol also works quite well for water, displaying some indication of transferability.

The final talk of the session was delivered by Koji Tsuda (University of Tokyo) on the topic of *automated metamaterials discovery via quantum annealing*. The focus here was to find out if there is a class of materials that ML can design that a human would not, with the intended application of sky radiators. These devices take heat from the Earth and radiate it away, and are complex from a materials design and engineering perspective. Prof. Tsuda has used quantum computing as a way to accelerate the design process, more specifically quantum annealing on a D-Wave 2000Q. This machine has an annealing time (time taken to solve the optimisation problem) of 170 μs, and the group performed 50 annealing runs to take the consensus [8]. Although this is incredibly fast, the difficulty was in rewriting the optimisation problem in a way that can be solved by quantum annealing. The bottleneck in the process is now the first-principles simulation of the properties of the materials selected by the quantum annealing.

Machine Learning Session 2

The second session was started with a presentation from Olexandr Isayev (University of North Carolina, moving to Carnegie Mellon University) on the topic of *neural networks learning quantum chemistry*. A major emphasis was on developing new potentials for force fields using a supervised machine learning framework. The ANI-1, ANI-1x and ANI-1cc results were presented [9], where accuracy is 1-2 kcal/mol relative to DFT calculations. Other properties, including charges, can also be predicted accurately [10]. A sampling method for chemical space based upon active learning was also presented [11], which mitigates human-biases in selecting training data. A recent development is the prediction of chemical properties with an atoms-in-molecules neural network that gives some insights into the "black box" of neural networks [12]. Post-presentation discussion revolved around whether the accuracy was sufficient for calculating rates of reactions etc., and whether there is scope for more work to be done there. A desirability for building in uncertainty estimations was also raised.

Johannes Hachmann (State University of New York, Buffalo) presented work on a *ML shortcut to physics-based modelling and simulations*. The focus was on replacing quantum mechanical calculations that rigorously map a structure to a property using machine learning, with an example application of the discovery of high-refractive-index polymers. There was an interesting discussion on hyperparameter optimisation and finding the best ML model for the system of interest, with genetic algorithm optimisation of hyperparameters representing an attractive opportunity. Ideas to reduce the data generation bottleneck were also presented, with on-the-fly assessment of learning rate combined with just-in-time termination of data generation leading to two orders of magnitude reduction in the amount of training data required. The ChemML machine learning suite was also presented [13], which aims to make ML as user-friendly as quantum chemical programs.

Alexandre Tkatchenko (University of Luxembourg) gave a talk on progress towards *exact molecular dynamics simulations with quantum chemistry and machine learning*. The context of the talk was the Cambridge Blind Test of crystal polymorphism, which stresses the importance of non-covalent interactions. Good results were possible with DFT methods including many-body corrections, but resulted from 20 million CPU hours. Prof. Tkatchenko is working towards the inclusion of ML-derived covalent force fields in this workflow, which would also allow path-integral molecular dynamics to be carried out. The approach taken in his group is known as gradient-domain ML [14], where the main focus is on forces rather than energies. It was demonstrated that this can result in lower errors and requires fewer training data. The talk finished with a perspective that ML might not be the way to go for describing non-covalent interactions, but that it is good for local (covalent) interactions.

The second session concluded with a presentation from Volker Deringer (University of Cambridge) on the transition *from machine learning interatomic potentials to atomic-scale materials science*. The aim of this work is finding the unknown function that links different materials of the same element based upon data alone – a form of

potential energy function for materials. Work was presented on a ML potential that is more expensive to evaluate than existing empirical potentials but is significantly more accurate [15]. As with many other talks during the meeting there was some focus on how to obtain training data, with the use of self-guided learning to generate a database and then a potential discussed [16]. Dr Deringer also presented a method for obtaining some chemical understanding from ML [17], which is due to the representation used.

Machine Learning Session 3

The final ML specific session was opened by Şule Atahan-Evrenk (TOBB University of Economics and Technology, Ankara) with a presentation on the *prediction of intramolecular reorganization energy using ML*. The application of interest here was organic semiconductors and it was noted that a lot chemical space has not been considered for this type of material, suggesting that ML should be able to help. The reorganization energy (due to structural change on charge transfer) is a preliminary screening metric for these materials, helping to narrow down the number of candidate molecules before more expensive calculations are carried out. The deep learning model constructed by Dr Atahan-Evrenk works well for data sets larger than 1500 molecules [18], but it was noted that care is needed in the selection of molecules within the training, validation and test sets.

Thomas Heine (TU Dresden) gave a talk on *challenges for automated materials discovery*, with a focus on 2D materials and predicting properties of metal organic frameworks (MOFs). It was made clear from the start that energies don't always help with materials discovery, there is greater interest in properties such as band gaps etc. The problem of finding high quality data was mentioned, especially in terms of including older literature data, with the CoRE MOF database presented [19]. In terms of ML and predicting properties of MOFs one issue discussed was how to encode a MOF (SMILES etc. cannot be used)? Structure factors (related to X-ray diffraction data) would be a convenient choice, but the problem is that common ML frameworks don't work with complex numbers. The solution presented was to use an autoencoder to reduce the dimensionality of the structure factor data with minimal loss of accuracy. A member of the audience made an intriguing suggestion that perhaps properties could be encoded alongside structural data.

The final talk of the ML sessions was delivered by Alán Aspuru-Guzik (University of Toronto) on *the materials for tomorrow, today*. This engaging presentation was focused on "closing the loop" in materials design; predicting a structure, making it, and feeding back the outcomes to the prediction model. A major barrier to translating theoretically-derived materials into physical materials is in persuading people to make them. The route Prof. Aspuru-Guzik has chosen to pursue is the use of robotics combined with high-throughput and high-performance computing. Videos shown during the presentation showed one of these robots in action. The software platform used in the closed-loop approach, known as ChemOS [20], is available for free download in an effort to democratize autonomous discovery [21]. It is possible to interact with ChemOS via the Slack messaging platform and Prof. Aspuru-Guzik

made a suggestion that Slack could also be the way we "talk" to quantum chemistry programs. His group is working on several ChemOS robotic chemistry algorithms and have created Olympus, a playground for benchmarking and testing such algorithms. The audience were invited to upload their own algorithms and test them against the best existing ones. Questions from the audience prompted a perspective on challenges for the field, which included the best way to represent molecules, how to get the best out of autoencoders for molecules, and how to represent chemical graphs in neural networks.

Thomas Miller Plenary

As part of his plenary lecture, Thomas Miller (Cal Tech) presented his group's work on MOB-ML [22], where the ML model predicts correlation energies using molecular orbital based descriptors. An example usage is carrying out a Hartree-Fock calculation and inferring the coupled cluster correlation energy. Relative to conventional atom-based descriptors the advantage is better transferability between molecules.

Physical Organic and Catalysis Session

Although not entirely obvious from the title, this session included many aspects of theoretical chemistry that can be considered as artificial intelligence approaches to molecular discovery.

The first speaker was Satoshi Maeda (Hokkiado University) who gave a presentation on *systematic generation and analysis of reaction path networks by the artificial force induced reaction (AFIR) method*. While a potential energy surface can predict reaction paths by systematic exploration, this process doesn't scale well with the number of atoms and one must look beyond brute force approaches. In particular, we generally do not want to sample regions far from reaction pathways. The AFIR method was presented as a possible solution as it pushes reactants together to create a slice of a potential energy surface that also includes reaction barriers [23]. This is systematically repeated to build up a reaction path network. To solve the problem of this process generating too many minima and transition states, rate equations are generated for the full network and then simplified based upon user defined parameters (such as temperature or desired reaction time) [24]. This has been applied to systems including Rh catalysed isomerisation [25], and is available in the GRRM17 program package.

Natalie Fey (University of Bristol) delivered a talk on *CatLab – putting calculation before experiment in organometallic catalysis*. This centred on the idea of using ligands to tune complex properties, which is a well-established idea reflected in the Ligand Knowledge Base, but it is now supplemented with computational data [26]. Dr Fey presented a methodology of using the same ligands in different environments, then using a principle component analysis to reduce the number of variables and produce a map of that slice of chemical space. Recent work has involved collaborating with a statistician to take an extra look at the data, with

preliminary results indicating that the assumption that catalysis follows a normal distribution is not a good one. The Bristol Reactivity in Catalysis Knowledgebase (BRiCK) was also presented as a computational way of finding new applications for known catalysts, with experimental testing.

Markus Reiher (ETH Zürich) presented work on *quantum chemical exploration of catalytic reaction networks*. This was a second talk on the theme of automating the exploration of possible reaction pathways, this time with a focus on complex chemical processes [27]. One of the problems is that navigating the networks manually is too difficult, so Prof. Reiher's group have developed the Chemoton software to automate exploration and visualisation of chemical reaction networks [28]. A new library known as Molassembler (manuscript in preparation) was also presented, which is similar to the existing RDKit cheminformatics software but also covers transition metal complexes. Prof. Reiher's group have also investigated error assessment in ML models, which was presented as a good way to identify when extra data is required [29]. A question from the audience asked about storing data for large reaction networks, which led to a perspective of collecting data from the user community to prevent "wasted" calculations that have already been done.

The final talk of this session was delivered by Heather Kulik (MIT) on *transition metal catalyst discovery with high-throughput screening and machine learning*. The motivation for this work is that while there is a lot of good chemical discovery practice in organic chemistry, there are numerous problems when moving to inorganic chemistry, including oxidation states, spin states, smaller databases and increased complexity. Prof. Kulik presented the molSimplify software [30], which allows for automated structure generation and optimisation for transition metal complexes. The question of representation of molecules in ML was discussed, with revised autocorrelations suggested for prediction of ground states and splitting [31]. The question of how molecular properties depend on different atoms was also explored, leading to an orthogonal design process [32]; the redox behaviour can be modified without affecting the spin splitting. The talk concluded with highlighting that a current problem for the field is in describing non-covalent interactions with ligands, which may be related to an insufficient amount of data, and highlighting an opinion article on ML in the discovery of transition metal complexes [33].

5. Posters

Many of the posters featured preliminary results, hence details are not included here. However, it was clear that ML and AI techniques are being incorporated into many aspects of theoretical chemistry, from prediction of NMR relaxation times through to non-adiabatic coupling in quantum molecular dynamics. There was ample time for discussion and networking around the two poster sessions.

One noteworthy poster that features published results was presented by J.P. Janet from the Kulik group at MIT. This work developed an uncertainty metric for neural networks used in chemical discovery [34], which holds the promise of predicting when the neural network is navigating into new areas of chemical space where calculations may fail.

6. Participants

There were approximately 550 attendees at the congress, with the majority holding academic affiliations. As the theoretical chemical physics community has traditionally been male dominated, it was particularly pleasing to see a significant number of female speakers (roughly one in three) and the organisers should be commended for working hard to improve gender balance. It was also notable that bursaries were available for primary caregivers, and appropriate facilities were made available at the conference site.

7. Conclusions

The event clearly demonstrated that machine learning and artificial intelligence are influencing many areas of theoretical chemical physics, both in terms of improving the methods used in computational investigations and in allowing exploration and discovery in areas of chemical/materials space that might not otherwise be possible. Although this report is focused on the talks and posters relevant to the AI3SD network, it was impressive to see how the AI content was balanced and interspersed with what could be described as more traditional theory. Two interesting themes that emerged on the discovery front were the use of chemical reaction networks to predict reaction outcomes (including kinetic considerations) and that there is rapid improvement in the software tools used to investigate transition metal complexes.

Bibliography

[1] P. Dral, "MLatom," [Online]. Available: http://mlatom.com. [Accessed 7 August 2019].

[2] P. Dral, *J. Comput. Chem.,* DOI:10.1002/jcc.26004, 2019.

[3] R. Ramakrishnan, P. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.,* vol. 11, p. 2087, 2015.

[4] P. Dral, A. Owens, S. N. Yurchenko and W. Thiel, *J. Chem. Phys.,* vol. 146, p. 244108, 2017.

[5] J. Wang et al., *ACS Cent. Sci.,* vol. 5, p. 775, 2019.

[6] S. Dick and M. Fernandez-Serra, *arXiv:1812.06572.*

[7] S. Dick, "NeuralXC," [Online]. Available: https://github.com/semodi/neuralxc. [Accessed 7 August 2019].

[8] K. Kitai, J. Guo, S. Ju, K. Tsuda, J. Shiomi and R. Tamura, *arXiv:1902.06573.*

[9] J. S. Smith, O. Isayev and A. E. Roitberg, *Scientific Data,* vol. 4, p. 170193, 2017.

[10] A. E. Sifain et al., *J. Phys. Chem. Lett.,* vol. 9, p. 4495, 2018.

[11] J. S. Smith et al., *J. Chem. Phys.,* vol. 148, p. 241733, 2018.

[12] R. Zubatyuk et al., *ChemRxiv:7151435.*

[13] G. Vishwakarma et al., *ChemRxiv:8323271.*

[14] A. Tkatchenko, "sGDML: symmetric gradient domain machine learning," [Online]. Available: http://quantum-machine.org/gdml/. [Accessed 7 August 2019].

[15] M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila and G. Csányi, *Phys. Rev. Lett.,* vol. 120, p. 166101, 2018.

[16] V. L. Deringer, C. J. Pickard and G. Csányi, *Phys. Rev. Lett.,* vol. 120, p. 156001, 2018.

[17] N. Bernstein et al., *Angew. Chemie. Int. Ed.,* vol. 58, p. 7057, 2019.

[18] S. Atahan-Evrenk and F. B. Atalay, *J. Phys. Chem. A,* DOI:10.1021/acs.jpca.9b02733.

[19] Y. G. Chung et al., *Chem. Mater.,* vol. 26, p. 6185, 2014.

[20] L. M. Roch et al., *ChemRxiv:5953606.*

[21] "ChemOS," [Online]. Available: https://github.com/aspuru-guzik-group/ChemOS. [Accessed 7 August 2019].

[22] L. Cheng, M. Welborn, A. S. Christensen and T. F. Miller, *J. Chem. Phys.,* vol. 150, p. 131103, 2019.

[23] S. Maeda and K. Morokuma, *J. Chem. Phys.,* vol. 132, p. 241102, 2010.

[24] Y. Sumiya, T. Taketsugu and S. Maeda, *J. Comput. Chem.,* vol. 38, p. 101, 2017.

[25] T. Yoshimura et al., *Chem. Sci.,* vol. 8, p. 4475, 2017.

[26] D. J. Durand and N. Fey, *Chem. Rev.,* vol. 119, p. 6561, 2019.

[27] G. N. Simm, A. C. Vaucher and M. Reiher, *J. Phys. Chem. A,* vol. 123, p. 385, 2019.

[28] "SCINE," [Online]. Available: https://scine.ethz.ch/download/. [Accessed 7 August 2019].

[29] G. N. Simm and M. Reiher, *J. Chem. Theory Comput.,* vol. 14, p. 5238, 2018.

[30] E. I. Ioannidis, T. Z. H. Gani and H. J. Kulik, *J. Comput. Chem.,* vol. 37, p. 2106, 2016.

[31] J. P. Janet and H. J. Kulik, *Chem. Sci.,* vol. 8, p. 5137, 2017.

[32] J. P. Janet et al., *Inorg. Chem.,* DOI:10.1021/acs.inorgchem.9b00109.

[33] H. J. Kulik, *WIREs Comp. Mol. Sci.,* p. e1439, 2019.

[34] J. P. Janet et al., *Chem. Sci.,* DOI:10.1039/c9sc02298h.