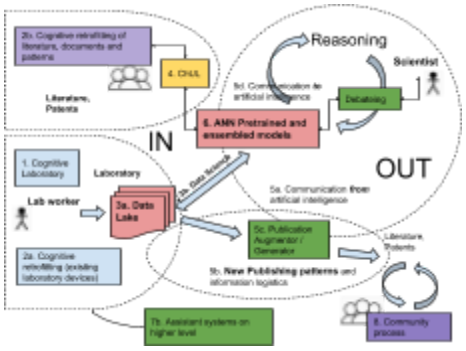
Autochemistry: A new research paradigm based on Artificial Intelligence and Big Data

Thorsten Gressling*
*ARS Computer und Consulting GmbH, Munich, Germany

ABSTRACT: Artificial Intelligence technologies affect every domain and process in industry. Many solutions with different maturity levels have been created or are in development. With this paper we collect initiatives in the domain of chemical science and bring these resources together into a common process model. We define ten building blocks, analyse their role in the architecture and evaluate their impact to the current system. Finally we discuss the changes and the transition that occurs to the lab worker and the chemist. This paper introduces *Autochemistry* as a meme and for further development and discussion. We just can provide a first sketch to this exciting new area of scientific principles changing the anthropocentric fundament of chemistry research to a technocentric one.



"Cognitive technologies will have as much impact on chemistry as quantum mechanics."

INTRODUCTION

In 2017 we introduced a new kind of laboratory environment: integrated by cognitive technologies and driven by artificial intelligence[1] of the 3rd Generation. That was the beginning of rethinking digitalization of the chemists workplace from a different 'cognitive' perspective. We developed the idea further and upon this we extended this to the whole research process in chemistry. In this article we discuss that areas within classical chemistry has to be changed and new memes arise with the introduction of artificial intelligence.

Current research architectures do not handle the impact of modern artificial intelligence[2] or just focus on technical solutions based on classical information technology[3]. Moreover most current solutions in chemistry and pharma research are based at least on the 2nd generation of artificial intelligence[4] or - if neural Networks are used - do not cover the whole research process[5].

BUILDING BLOCKS

All Modules we described in the next sections contain artificial intelligence in several portions. *There is no building block without artificial intelligence.* That is the main reason why we have to think about a new research architecture from the birds eye view.

INGESTION AND LABORATORY

1. Cognitive Laboratory[6,7]**.** We discovered that the idea of changing the perspective of digitalization to a *human centric approach* opens a complete new field of chemometrics. Main aspect was the inspection of manual tasks by the lab worker, guidance of the worker by situation prediction as well as creating a digital twin of the lab.

Since the beginning of science the environment was built around the human ability to read analog scales and further throughout the five traditional Aristotelian senses and their respective sensory organs[8–10]. Moreover as the laboratory worker not only perceive things differently in the same situation or environment, he also apply different meanings to what is perceived. So we introduce neural network ensembles consisting of specific abilities in chemical understanding and reasoning. Cognitive chemometrics bridges the last mile in deep measurement of the laboratory, that up to now was not clearly accessible. Cognitive chemometrics includes visual technologies as well as permanent logging of all data (data lake principal), augmented interpretation of experiment results, remote assistance and training for the worker.

Extending the cognitive space with machine learning. For designing the experiments in the laboratory it is hard to find optimal region for a experimental observable. With methods like Bayesian optimisation in combination using expert system it is possible to identify sample points in the parameter space[11].

Lack of cognitive abilities of the lab worker. But as the density off the experimental observables are limited to the *capabilities of the cognitive performance of the lab worker* also with statistical methods for defining these parameters somewhere there is a natural limit.

2a. Cognitive retrofitting of existing laboratory devices. By now all effort in getting data from instruments were to introduce digital technologies to the instrument. All suggested solutions were *technology-centric*. By now even new equipment is shipped without any digital interfaces. So this problem still persists. But also non scalable, *subjective findings* like observing the meniscus of a liquid, measuring the decay rate of a tablet or the visual shape of a monocrystal can be addressed with our cognitive approach.

Meanwhile the accuracy of image classification is less than 5% wrong[12] and far outbeats the human eye especially in defined environments like the laboratory. As part of our laboratory 4.0 we introduce a small device placed in front of the scale that is now able to read this analogue input, reads the value by image recognition and performs interpretation of the curve by artificial intelligence (EYE). All data then is available via standard representations like Allotrope[13].

2b. Cognitive retrofitting of literature, documents and patterns. Not only in the techno-sphere we have to deal with old equipment but also in the science-sphere we have to deal with old information. Decades of scientific literature must be transformed and brought into understandable digital format. For this retrofitting some services and startups address these process[14].

PERSISTENCE AND SYMBOLIC ANALYSIS

3a. Data Lake. That leads into a new form of persistence where all data is unstructured and start without any preconditions or any structures. This unlimited data Lake is the fundament of a disruptive new form of information logistics[15].

3b. The role of Data Science. Because feature extraction in ANN-based systems is generic is the reason of performing Data Science is shifting. So systems doing just Data Science in a conventional way will just have a short period of relevance. However, today feature extraction is still a important step in engineering[15,16].

Only reasoning may be subject for a longer time as we know examples of “Auto Data Science” emerging[17]. From that point of view we can see data Science as a subset of the communication design to the scientists.

COMMUNICATION

5a. Communication *from* artificial intelligence. Transfer of knowledge between man and machine takes place by Communication. There are a few examples of designing this interface by using current communication patterns[27].

One crucial aspect of the design of Autochemistry is to provide the best information Understanding for the researcher or the scientific community. This means that results discovered by the system has to be communicated in the right way.

Basic assumption of Autochemistry is that is single mind is not capable enough anymore to understand patterns or even projections of reality. Starting with the current communication patterns like chemical formulas the interaction between Man and machine can evolve by suggesting alternative ways of communication and test these systemically by the scientists.

Debateing with the algorithms[28]. That means that we have to design a interface and a language that supports this Deeper immersion which is the next level of science.

5b. New Publishing patterns and information logistics. One classic pattern of information exchange was the scientific publication. Since there was a lack of logistics and material only positive results were published. Now with the new unlimited resources this paradigma to be changed. With the availability of non-positive data we can make use of it[29].

"The permanent collection of all data including subjective findings and Type I and Type II Errors will lead into a observable space with a coverage to powers of ten than the current practice."

5c. Publication Augmentation or Generation. It can be considered as a reverse direction of digesting chemical language into a neural network. Also the duration of the content by a reverse process like a thousand monk's can be designed. By now we are not aware of a implementation.

5d. Communication *to* the artificial intelligence. Also for the communication with the ANN's we have to define the way of communication, for example the representation of molecules[30,31].

"The transformation away from the anthropocentric view of the world also takes place in chemistry."

Lack of cognitive abilities of the scientist. Memory performance of humans are far out beaten by Memory performance of ANN's. We can extend the concept of cognition to more general cognitive functions, such as verbal expression, qualitative problem solving, abstract (usually visual) exploration[32].

Weak and unsolved properties of ANN's:

- *Conceptual mind*
- *Deep Reasoning*
- *Intuition*
- *Creativity*
- *(Consciousness - unclear if necessary)*

"With the combination of ANN's and Big Data we get a new relationship to knowledge."

REASONING AND SUBSYMBOLIC ANALYSIS

4. Modern approach to a chemical understanding language 'ChLU'. We know a few working examples on organic chemistry prediction using artificial neural networks. This is discussed in section 6. By now these Solutions use intermediate natural language as symbolic representation[18–22], for example simplified molecular-input line-entry system, or SMILES. SMILES represents a molecule as a sequence of characters. However, there is research on improvements of integration with deep neural networks with deepSMILES[23] as well as in more general terms[24–26].

In chemistry visual notation expands the human natural language and writing in a way that other features relevant in chemistry can be communicated in a better way. Another example of using a high level abstract location is the Mathematica language.. There are examples of feature extraction with a neural networks that are capable to identify and classify chemical structures in the way a chemist does.

Like quality in language translation dramatically improves with *one shot translation* we also need a specific language understanding for the chemical language. As the domain space size for chemistry is much smaller than the human language space also education in university does not reflect this situation.

Evolution of representation systems. So we see that Solutions built up on natural language understanding can just be a starting point developing the full potential of artificial intelligence further. We predict that even the human approach by using the chemical formula notation is also a representation of the atomic space designed for the *cognitive capabilities of humans*. That means that other Minds can have their own and in terms of complexity with other cognitive parameters.

The roles of the different types of machine learning, artificial intelligence and neural networks in the core process. Artificial intelligence has transformed three times in the last 50 years. In the first generation the basic assumption was that with the Logical calculus all real world problems can be solved. This phase one and it when it was realised that the complexity of the problems are exponential. After a few years in the early 80s with the second generation of a i statistics and semantics were introduced. later in the 90s also this episode ended with the problem of rising complexity. However even today this technologies and design patterns are used and useful[33]. But after the second Ice Age now since 2010's with ANN we expect to handle the complexity problem.

6. ANN Design Patterns: Pretrained and ensembled. One of the big obstacles of artificial intelligence of the first and second generation was the necessary of *feature engineering*. With artificial intelligence of the third-generation feature engineering happens automatically within the neural networks. Several applications for the usage of simple topologies are implemented in organic[21,34–37]. as well as inorganic chemistry[38]. Main application fields are synthesis path prediction and material properties like toxicity[39] or drug design[40].

With the introduction of higher levels of ANN topologies like ensembles[41–44], capsules[45,46] and pre-trained networks we are able to design systems that delivers higher level of understanding within AI of the third generation. Whenever a neural network find a suitable explanation for all circumstances this can be considered as a plateau in the solution space. This stable situation can be considered as *cognitive stationary state (Eigenzustand)* of the neural network like the solution of the Schödinger equation in quantum physics. Depenting on the capacity and the number of obersvables more than on stationery state is possible.

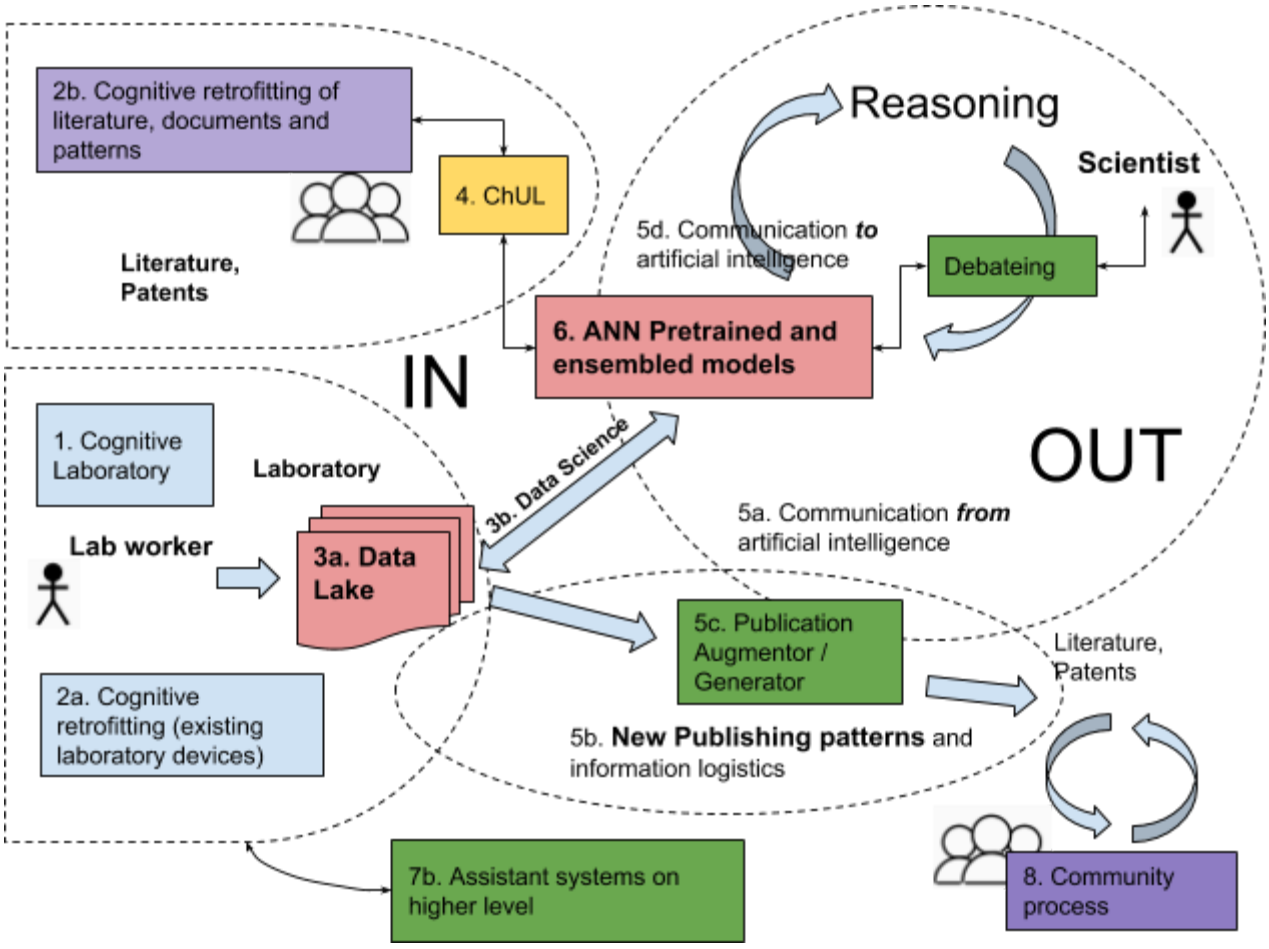


Figure 1. Orchestration of the building blocks

IMPLEMENTATION AND FRAMEWORK

6a. Autochemistry foundation. We suggest to create a open-source community where all AI related models are collected. This structured approach will be the fundament of the ensemble solutions. There are example of such communities[47,48] working well.

7a. Changes in field of expertises. Till the job profile for chemists at we know today will not exist in future anymore. It will undergo a change.

7b. Assistant systems on higher level. As we have described the cognitive laboratory environment and we have discussed the information space that is dramatically broadened in future of course the chemist as a theoretical designer and experimental planner will also have assistance system. Current approaches also use AI of the second generation[49,50] or there is research on designing this systems[51].

8. Community process. The role of the community processes in science is to introduce objectivity. So still there will be peer reviews. But like in politics the group and the delegates that has to decide about quality of results is formed by what we call liquid democracy augmented by artificial intelligence actors. All patterns of delegation and discussion in Liquid Democracy[52] and their adaptation to the scientific validation process will be subject of further research.

AI as an actor. In all structures of the community process also a digital person can act in many roles[53].

The Community process should also cover the infrastructure and not only the scientific objectivity process. analysing the maturity level of the Open Source Project seems that there is currently an ice age which results to the same pattern than nai. So a new level of consolidation to this operational layer[54] will preserve in the next step of objectivity.

9. Implementation and Meta-level. It is obvious that a new type of science administration is created. We do not have a system yet for this new architecture.

10. The role of quantum computing. Both domains - artificial intelligence and quantum computing - are based on the same mathematical topology. So quantum computing is a candidate for improving both areas[55]. Also a intersections the fields are created with applying deep learning on quantum mechanical problems[56].

"Chemists have to accept that deep science can be created by machines. This change of perspective is as fundamental as Galileo Galilei's findings that the Earth rotates around the Sun and is therefore not the center of the universe.

But It is not the expulsion from paradise, it is the discovery of a new tool that will take us to the next level in the history of research."

AUTHOR INFORMATION

Corresponding Author

*thorsten.gressling@ars.de, +49 172 5328003

Notes

The author declares no competing financial interest.

REFERENCES

(1) Gressling, T.; Madl, A. A New Approach to Laboratory 4.0: The Cognitive Laboratory System. In *ResearchGate*; 2017.
(2) AI Meets Chemistry. *Nature* **1988**, *334*, 659.
(3) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. 2018.
(4) Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; Aspuru-Guzik, A. Parallel and Distributed Thompson Sampling for Large-Scale Accelerated Exploration of Chemical Space. *arXiv [stat.ML]*, 2017.
(5) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent Sci* **2017**, *3* (12), 1337–1344.
(6) Thurner, V.; Gressling, T. A Multilayer Architecture for Cognitive Systems -- Supporting Well-Defined Processes That Are Partially Executed Manually in Technical Work Places. **2018**.
(7) smartLAB

http://www.labvolution.de/en/conferences-events/themenschwerpunkte/smartlab/ (accessed Oct 21, 2018).
(8) Cognitive senses find their way to our digital lives https://www.businesstoday.in/moneytoday/technology/cognitive-senses-next-big-thing-in-computing-world-ibm/story/192657.html (accessed Oct 26, 2018).
(9) IBM thinks computers will have senses in five years https://www.businesstoday.in/technology/news/ibm-thinks-computers-will-have-senses-in-five-years/story/190810.html (accessed Oct 26, 2018).
(10) Azarbayjani, M. Technology and the Senses: Multi-Sensory Design in the Digital Age. *Huichawaii.org* **2018**.
(11) Cognitive parameterization http://research.ibm.com/labs/uk/parameterization.html (accessed Oct 26, 2018).
(12) Geirhos, R. Comparing Deep Neural Networks against Humans: Object Recognition When the Signal Gets Weaker.
(13) Data Standard | Allotrope Foundation https://www.allotrope.org/ontologies (accessed Jun 24, 2018).
(14) 1000 Monks – A.I. your documents http://1000monks.com/ (accessed Oct 26, 2018).
(15) Chiang, L.; Lu, B.; Castillo, I. Big Data Analytics in Chemical Engineering. *Annu. Rev. Chem. Biomol. Eng.* **2017**, *8*, 63–85.
(16) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114* (10), 105503.
(17) Lam, H. T.; Thiebaut, J.-M.; Sinn, M.; Chen, B.; Mai, T.; Alkan, O. One Button Machine for Automating Feature Engineering in Relational Databases. *arXiv [cs.DB]*, 2017.
(18) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *arXiv [cs.LG]*, 2017.
(19) IBM RXN for Chemistry https://rxn.res.ibm.com/ (accessed Oct 21, 2018).
(20) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604.
(21) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *arXiv [cs.LG]*, 2017.
(22) Segler, M.; Preuß, M.; Waller, M. P. Towards "AlphaChem": Chemical Synthesis Planning with Tree Search and Deep Neural Network Policies. *arXiv [cs.AI]*, 2017.
(23) O'Boyle, N. M.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures.
(24) Asai, M.; Fukunaga, A. Classical Planning in Deep Latent Space: Bridging the Subsymbolic-Symbolic Boundary. *arXiv [cs.AI]*, 2017.
(25) Steinert, L.; Hoefinghoff, J.; Pauli, J. Online Vision- and Action-Based Object Classification Using Both Symbolic and Subsymbolic Knowledge Representations. *arXiv [cs.AI]*, 2015.
(26) Symbolic and Sub-Symbolic Representations in Computational Models of Human Cognition.
(27) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94* (2), 421–430.
(28) IBM Research Project Debater https://www.research.ibm.com/artificial-intelligence/project-debater/ (accessed Oct 20, 2018).
(29) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76.
(30) Huang, B.; von Lilienfeld, O. A. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145* (16), 161102.
(31) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *arXiv [physics.comp-ph]*, 2012.

(32) Miller, G. A. The Magical Number Seven, plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychol. Rev.* **1956**, *2* (63), 81–97.
(33) Jacob, P.-M.; Lapkin, A. Prediction of Chemical Reactions Using Statistical Models of Chemical Knowledge. 2018.
(34) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent Sci* **2016**, *2* (10), 725–732.
(35) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent Sci* **2017**, *3* (5), 434–443.
(36) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. 2018.
(37) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274.
(38) Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57* (42), 13973–13986.
(39) Duvenaud, D.; Maclaurin, D.; Gomez-Bombarelli, J. A.-I. R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints.
(40) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci Adv* **2018**, *4* (7), eaap7885.
(41) Smolyakov, V. Ensemble Learning to Improve Machine Learning Results https://blog.statsbot.co/ensemble-learning-d1dcd548e936 (accessed Jul 8, 2018).
(42) Yao, X.; Islam, M. M. Evolving Artificial Neural Network Ensembles. *IEEE Comput. Intell. Mag.* **2008**, *3* (1), 31–42.
(43) Zhou, Z.-H.; Wu, J.; Tang, W. Ensembling Neural Networks: Many Could Be Better than All. *Artif. Intell.* **2002**, *137* (1), 239–263.
(44) Opitz, D. Popular Ensemble Methods: An Empirical Study. *Journal of Arti cial Intelligence Researc h* **1999**, *11*, 169–198.
(45) Sabour, S.; Frosst, N.; Hinton, G. E. Dynamic Routing Between Capsules. *arXiv [cs.CV]*, 2017.
(46) Pechyonkin, M. Understanding Hinton's Capsule Networks. Part II: How Capsules Work https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-networks-part-ii-how-capsules-work-153b6ade9f66 (accessed Dec 19, 2017).
(47) O'Boyle, N. M.; Guha, R.; Willighagen, E. L.; Adams, S. E.; Alvarsson, J.; Bradley, J.-C.; Filippov, I. V.; Hanson, R. M.; Hanwell, M. D.; Hutchison, G. R.; et al. Open Data, Open Source and Open Standards in Chemistry: The Blue Obelisk Five Years on. *J. Cheminform.* **2011**, *3* (1), 37.
(48) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46* (3), 991–998.
(49) Goh, G. AI-assisted computational chemistry: Predicting chemical properties with minimal expert knowledge - O'Reilly Artificial Intelligence Conference in New York 2017 https://conferences.oreilly.com/artificial-intelligence/ai-ny-2017/public/schedule/detail/59072 (accessed Oct 21, 2018).
(50) Segler, M. H. S.; Preuss, M.; Waller, M. P. Learning to Plan Chemical Syntheses. *arXiv [cs.AI]*, 2017.
(51) Towards a Cognitive Assistant for Computational Chemistry: Investigating automatable methods to analyse the output of simulations (iCase joint with IBM Research UK) - University of Liverpool https://www.liverpool.ac.uk/study/postgraduate-research/studentships/computational-chemistry/ (accessed Oct 26, 2018).
(52) Liquid Democracy https://wiki.piratenpartei.de/Datei:Liquid_demok.PNG (accessed Oct 21, 2018).
(53) European Civil Law rules in robotics http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf.

(54) LabLayer - A open source operational IT initiative for laboratories http://lablayer.org/index.php?title=Hauptseite (accessed Oct 28, 2018).
(55) Qiskit | Quantum Information Science Kit https://qiskit.org/ (accessed Oct 28, 2018).
(56) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Title: Outsmarting Quantum Chemistry through Transfer Learning.