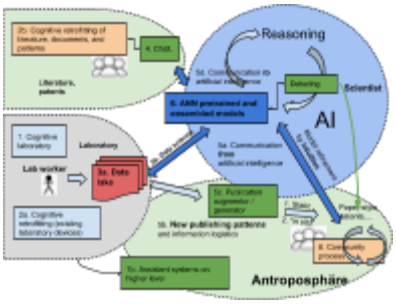
Autochemistry: A Research Paradigm Based on Artificial Intelligence and Big Data

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

Version 2, 2018-Nov-10

ABSTRACT: Artificial intelligence (AI) technologies affect every domain and process within industry. Many solutions with different maturity levels have been created or are in development. With this paper we collect the initiatives within the domain of chemical science, add missing items, and bring these resources together into a common process model, based on artificial neural networks (ANN). We define ten building blocks, analyze their role within the architecture, and evaluate their impact to the current system. Finally we discuss the changes and the transition that is experienced by the lab worker and the chemist. This paper introduces *autochemistry* as a meme describing a self-improving research approach. With this we begin development and discussion of an exciting new area of scientific principles, that is changing the anthropocentric fundament of chemistry research into a technocentric one.



KEYWORDS: Deep learning, artificial intelligence, chemistry, research, neural network

"Cognitive technologies will have as much of an 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] (AI) of the third-generation. That was the beginning of rethinking digitalization of the chemist's workplace from a different *cognitive* perspective. We developed the idea further and thus extended it to the whole of research processes in chemistry. In this article we discuss areas within classical chemistry that have to be changed and new memes that will arise with the introduction of artificial intelligence.

Current research architecture cannot handle the impact of modern artificial intelligence[2], or it merely focuses on technical solutions based on classical information technology[3]. Moreover, most current solutions in chemistry and pharmaceutical research are based on at least the second-generation of artificial intelligence[4] or – *if* artificial neural networks (ANN) are used – do not cover the whole research process[5].

BUILDING BLOCKS

All of the modules we will describe in the next sections contain AI technology in their implementation. *There are no building blocks without artificial intelligence.* This is the reason why we have to consider a new research approach from a bird's-eye view. And with this, a disruptive new moment may be introduced: the self-modification of chemical knowledge.

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 completely new field of chemometrics. The central aspects were the inspection of manual tasks by the lab worker, guidance of the worker utilized by situation prediction, and the creation of a digital twin of the lab.

Since the beginning of science, the lab environment was built around the human ability to read analog scales, a result of the utilization of the five traditional Aristotelian senses and their respective sensory organs[8–10]. Moreover, as the laboratory worker not only perceives things differently within the same situation or environment, they may also apply different meanings to what is perceived. Therefore, we have introduced neural network ensembles

consisting of specific abilities in environmental understanding and reasoning. Cognitive chemometrics bridges the last gap in deep measurement of the laboratory, which up until now has not been measurable to a degree of detail that is necessary to get invariably reproducible results and cover an observable space of a sufficient quality. Cognitive chemometrics include visual technologies as well as the permanent logging of all data (the data lake principal), augmented interpretation of experimental results, remote assistance, and training of the worker.

Extending the observable space with machine learning. Within the process of designing experiments in the laboratory it is difficult to find the optimal scope of application for a required observable. With methods like Bayesian optimization in combination with the use of an expert system, it is possible to identify sample points in the parameter space[11].

Lack of cognitive abilities of the lab worker. The determination of an experimental observable is 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. Up until now, any effort made to recover data from scientific instruments was made by introducing digital technologies to the instrument; i.e., all suggested solutions were *technology-centric*. Even now, most new equipment is shipped without any form of digital interface; clearly the problem persists. That said, non-scalable, *subjective findings* like observing the meniscus of a liquid or measuring the decay rate of a tablet or the visual shape of a monocrystal, can now be addressed with our new cognitive approach.

Meanwhile the accuracy of image classification is less than 5% error[12] and far outweighs the capabilities of the human eye, especially in defined environments like the laboratory. As a part of our laboratory 4.0, where 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. It is not only in the technosphere that we have to deal with outmoded equipment but within the world of science we are also faced with outmoded information. Decades of scientific literature must be transformed and brought forth into legible digital format. For this retrofitting, some services and startups already address these processes[14].

PERSISTENCE AND SYMBOLIC ANALYSIS

3a. Data lake. This leads to a new form of persistence, wherein all data is unstructured and begins without any preconditions or structures. This unlimited data lake is the fundament of a disruptive new form of information logistics[15].

3b. The role of data science. Performing data science is shifting because feature extraction in ANN based systems is generic. Systems which merely calculate data science in a conventional way will have a short period of relevance. That said, today's feature extraction is still an important step in engineering[15,16].

Reasoning technologies may be relevant for a longer period as we have continued examples of the emergence of "automated data science." From that perspective we can see data science as a subset of the communication design of the scientists.

COMMUNICATION

5a. Communication *from* artificial intelligence. The transfer of knowledge between human and machine takes place via communication. Herein, examples are given for designing this interface using current communication patterns[17].

One crucial aspect of the design of autochemistry is to provide the optimal means of understanding the information given to the researcher or the scientific community. This means that results discovered by the system have to be communicated in the right way.

A basic presumption about autochemistry is that a single mind is not capable of understanding patterns or even projections of reality. Starting with the current communication patterns, e.g., chemical formulas, the interactions

between human and machine may evolve by suggesting alternative means of communication and testing these means systemically.

Debating with the algorithms[18]. This means that we have to design an interface and a language that supports this deeper immersion which is the next level of science.

5b. New publishing patterns and information logistics. One classical pattern of information exchange is the scientific publication. While there has been a general lack of logistics and material, traditionally only positive results have been published. Now with new, unlimited resources this paradigm has to be changed. With the availability of non-positive data we can make use of its efficacy[19].

"The permanent collection of all data, including subjective findings and type I and type II errors, will lead to an observable space with a coverage to powers of ten greater 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. For now we are not aware of a means of implementation.

5d. Communication *to* the artificial intelligence. For the communication with the ANNs we have to define the means of communication, e.g., the representation of molecules[20,21].

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

Lack of cognitive abilities of the scientist. The memory performance of humans is far outmatched by the memory performance of ANNs. We can extend the concept of cognition to more general cognitive functions, such as verbal expression, qualitative problem solving, and abstract (usually visual) exploration[22].

Weak and unsolved properties of ANNs:

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

"With the combination of ANNs and big data we get a new relationship to knowledge."

REASONING AND SUBSYMBOLIC ANALYSIS

4. Modern approach to a chemical understanding language, "ChLU." There exist a few well known working examples of organic chemistry prediction using artificial neural networks (these will be discussed further in section 6). By now these solutions use intermediate natural language as a means of symbolic representation[23–27], e.g., simplified molecular-input line-entry system, or SMILES. SMILES represents a molecule as a sequence of characters. However, there exists research on improvements of integration within deep neural networks, with deep SMILES[28], as well as in more general terms[29–31].

In chemistry, visual notation expands upon natural human language and writing in a way that other features, which are more relevant to chemistry, are better communicated. Another example of using a high level abstract location is the Mathematica language. There are examples of feature extraction within a neural network that is capable of identifying and classifying chemical structures much the way a chemist would.

Like quality in language translation dramatically improves with *one shot translation;* however, 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 within systems. We see that solutions built upon natural language understanding may only be a starting point with which to further develop the full potential of artificial intelligence. We predict that even the human approach of utilizing a chemical formula of notation is also a representation of the atomic space designed for the *cognitive capabilities of*

humans. This suggests that other minds may have their own means of representation in terms of complexity with regard to other cognitive parameters.

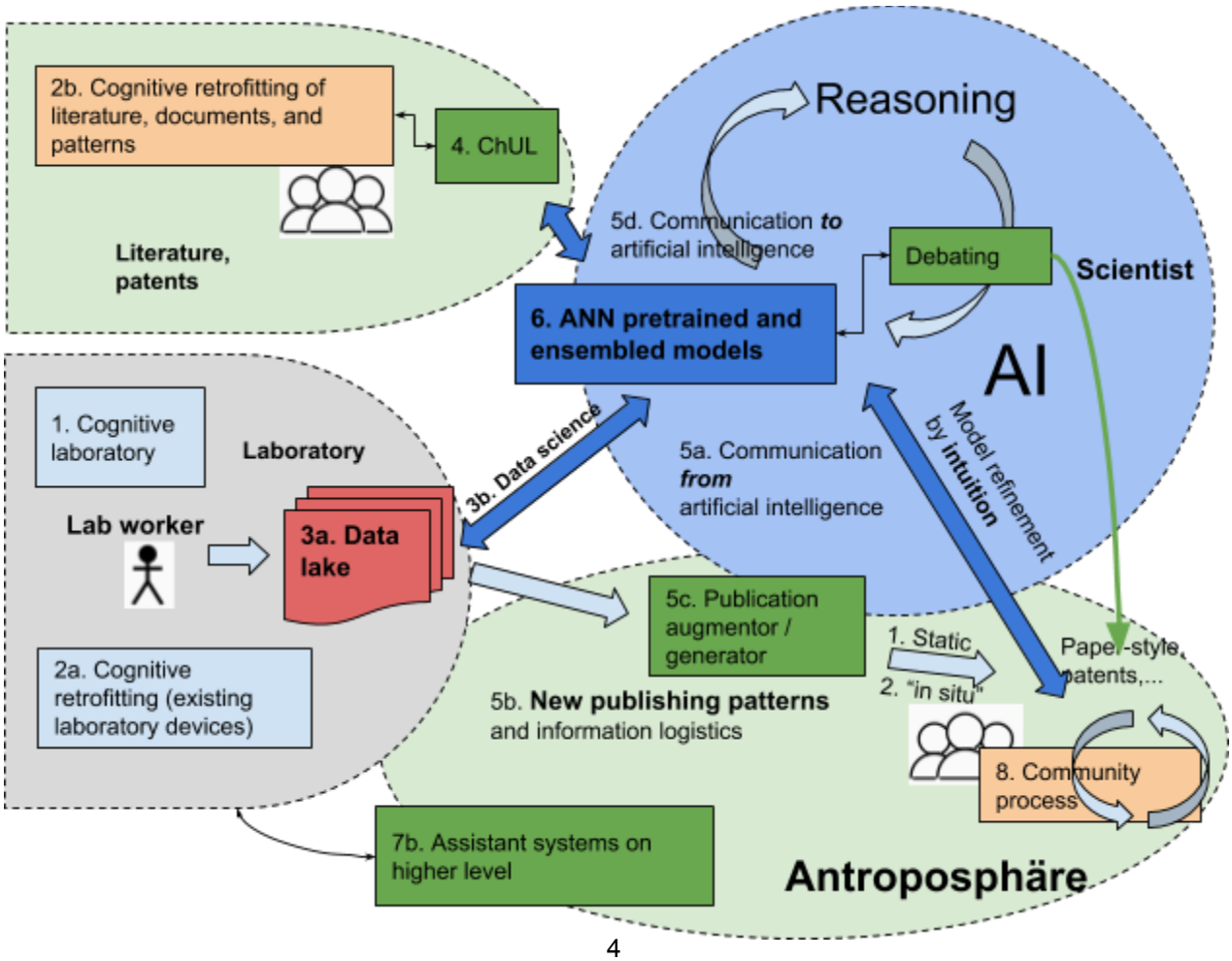
The roles of the different types of machine learning, artificial intelligence, and neural networks in the core process. Artificial intelligence has transformed threefold in the last fifty years. In the first generation, the basic assumption was that with logical calculus all real world problems could be solved. This was phase one and when it was realized that the complexity of the problems are exponential. In the early 80s, with the arrival of the second generation of AI, statistics and semantics were introduced. Later, in the 90s, this episode ended with a problem of rising complexity. However, even today these technologies and design patterns are used and are considered to be sufficiently useful[32]. 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 major obstacles of artificial intelligence of the first and second generations was the necessity of *feature engineering*. With artificial intelligence of the third-generation, feature engineering happens automatically within neural networks. Several applications for the

Figure 1. Orchestration of the building blocks.

usage of simple topologies are implemented in organic[26,33–36] as well as inorganic chemistry[37]. The main application fields are synthesis path prediction and material properties like toxicity[38] or drug design[39].

With the introduction of higher levels of ANN topologies like ensembles[40–43], capsules[44,45], and pretrained networks we are able to design systems that deliver a higher level of



understanding within AI of the third generation. Whenever a neural network finds a suitable explanation for all circumstances, this may be considered a plateau within the solution space.

This stable situation may be considered to be a *cognitive stationary state (Eigenzustand)* of the neural network, similar to the Schödinger equation in quantum physics. Depending on the capacity and the number of observables, more than one stationery state is possible.

IMPLEMENTATION AND FRAMEWORK

6a. Autochemistry foundation. We suggest the creation and implementation of an open-source community wherein all AI related models are collected. This structured approach will be the fundament of the ensemble solutions. There exist examples of other such communities[46,47] which are working well.

7a. Changes in fields of expertise. The job profile for chemists, as we know of them today, will no longer exist in the future. It will undergo a considerable change.

7b. Assistant systems at higher levels. As we have described, the cognitive laboratory environment and 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 second-generation AI[48,49] or there is research on designing this systems[50].

8. Community processes. The role of community processes in science is to introduce objectivity; thus the use of peer review as a methodology will continue. However, as with politics, the group and the delegates that has to determine the quality of results is formed by what we refer to as *liquid democracy augmented by artificial intelligence actors*. All patterns of delegation and discussion in delegative democracy[51] and their adaptation to the scientific validation process will be subjected to further research.

AI as actor. Within all of the structures of the community process, a digital person may act within various roles[52].

The community process should also cover infrastructural elements and not only processes of scientific objectivity. Upon analyzing the maturity levels of the open source project, it seems that there is currently an ice age which results in the same pattern as that of nai. So a new level of consolidation ithin this operational layer[53] will preserve the next steps of objectivity.

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

10. The role of quantum computing. Both domains — artificial intelligence and quantum computing — are based on the same mathematical topology; thus quantum computing is a candidate for improving both areas[54]. Intersections within the fields are also created by applying deep learning to quantum mechanical problems[55].

"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 apex of the universe.

However, this is not an 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) 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.
(18) IBM Research Project Debater https://www.research.ibm.com/artificial-intelligence/project-debater/ (accessed Oct 20, 2018).
(19) 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.
(20) 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.
(21) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *arXiv [physics.comp-ph]*, 2012.
(22) 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.
(23) 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.
(24) IBM RXN for Chemistry https://rxn.res.ibm.com/ (accessed Oct 21, 2018).
(25) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604.
(26) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *arXiv [cs.LG]*, 2017.
(27) Segler, M.; Preuß, M.; Waller, M. P. Towards “AlphaChem”: Chemical Synthesis Planning with Tree Search and Deep Neural Network Policies. *arXiv [cs.AI]*, 2017.
(28) O’Boyle, N. M.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures.

(29) Asai, M.; Fukunaga, A. Classical Planning in Deep Latent Space: Bridging the Subsymbolic-Symbolic Boundary. *arXiv [cs.AI]*, 2017.
(30) Steinert, L.; Hoefinghoff, J.; Pauli, J. Online Vision- and Action-Based Object Classification Using Both Symbolic and Subsymbolic Knowledge Representations. *arXiv [cs.AI]*, 2015.
(31) Symbolic and Sub-Symbolic Representations in Computational Models of Human Cognition.
(32) Jacob, P.-M.; Lapkin, A. Prediction of Chemical Reactions Using Statistical Models of Chemical Knowledge. 2018.
(33) 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.
(34) 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.
(35) 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.
(36) 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.
(37) 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.
(38) 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.
(39) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci Adv* **2018**, *4* (7), eaap7885.
(40) Smolyakov, V. Ensemble Learning to Improve Machine Learning Results https://blog.statsbot.co/ensemble-learning-d1dcd548e936 (accessed Jul 8, 2018).
(41) Yao, X.; Islam, M. M. Evolving Artificial Neural Network Ensembles. *IEEE Comput. Intell. Mag.* **2008**, *3* (1), 31–42.
(42) Zhou, Z.-H.; Wu, J.; Tang, W. Ensembling Neural Networks: Many Could Be Better than All. *Artif. Intell.* **2002**, *137* (1), 239–263.
(43) Opitz, D. Popular Ensemble Methods: An Empirical Study. *Journal of Arti cial Intelligence Researc h* **1999**, *11*, 169–198.
(44) Sabour, S.; Frosst, N.; Hinton, G. E. Dynamic Routing Between Capsules. *arXiv [cs.CV]*, 2017.
(45) 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).
(46) 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.
(47) 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.
(48) 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).
(49) Segler, M. H. S.; Preuss, M.; Waller, M. P. Learning to Plan Chemical Syntheses. *arXiv [cs.AI]*, 2017.
(50) 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).
(51) Liquid Democracy

https://wiki.piratenpartei.de/Datei:Liquid_demok.PNG (accessed Oct 21, 2018).
(52) European Civil Law rules in robotics http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf.
(53) LabLayer - A open source operational IT initiative for laboratories http://lablayer.org/index.php?title=Hauptseite (accessed Oct 28, 2018).
(54) Qiskit | Quantum Information Science Kit https://qiskit.org/ (accessed Oct 28, 2018).
(55) 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.