The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

AI Meets Chemistry

Akihiro Kishimoto, Beat Buesser, Adi Botea
IBM Research, Ireland

Abstract

We argue that chemistry should be the next grand challenge for Artificial Intelligence. The AI research community and humanity would benefit tremendously from focusing AI research on chemistry on a *regular basis*, as a benchmark as well as a real-world application domain. To support our position, we review the importance of chemical compound discovery and synthesis planning and discuss the properties of search spaces in a chemistry problem. Knowledge acquired in domains such as two-player board games or single-player puzzles places the AI community in a good position to solve critical problems in the chemistry domain. Yet, we show that searching in chemistry problems poses significant additional challenges that will have to be addressed. Finally, we envision how several AI areas like Natural Language Processing, Machine Learning, planning and search, are relevant for chemistry.

Introduction

Artificial Intelligence research has generated multiple essential technologies, such as Natural Language Processing, Machine Learning, planning, and search, with a wide range of applications. These fields have led to advances where AI is already outperforming humans, for example the Chinook Checkers-playing program (Schaeffer 2009), IBM's Deep Blue in chess (Campbell, Jr., and Hsu 2002) and Watson for Jeopardy! (Ferrucci et al. 2010), and Google DeepMind's Alpha-Go in Go (Silver et al. 2016). However, there is still an uncountable number of unsolved challenges we can think of. One of them is RoboCup (Kitano et al. 1997), where a team of robots attempts to win against human soccer players.

In this paper we promote chemistry as an ideal domain for AI research, encouraging the AI community to use chemistry as a benchmark domain on a regular basis. Chemistry has applications in many industries like pharmaceuticals, food and nutrition, and materials. The commercial impacts and effects on society of these industries are significant. For example, a single chemical compound named atorvastatin, better known as drug Lipitor®, generated annual revenues of over 12 billion US dollars, before its patent expired (Heifets and Jurisica 2012). That way AI research focusing on chemistry as a target translates directly into benefits for society, much faster than research focused on games.

In the past, researchers have attempted to develop expert systems and knowledge-reasoning systems to solve problems in chemistry. Examples include the Dendral Project (Lindsay et al. 1980) and Project Halo (Friedland et al. 2004). However, despite its importance in practice and its difficulty in nature, over the last couple of years a relatively small number of papers, such as (Duvenaud et al. 2015; Jin et al. 2017; Savage et al. 2017), have been published at premier computer science conferences.

At the same time, automation of chemistry has been an essential topic for chemists, e.g., (Dragone et al. 2017; Peplow 2014), and papers particularly about employing Machine Learning approaches have started to appear in chemistry journals (Gómez-Bombarelli et al. 2016; Segler and Waller 2017a; 2017b; Coley et al. 2017).

Due to more powerful hardware resources and recent significant advances in AI technologies including Machine Learning, NLP, search and planning, we are convinced that with the maturity of AI research it is now the right time to address chemistry on a much bigger scale than previously.

Research Motivation

A chemist developing a new drug molecule that cures a certain disease typically needs to design the chemical structure of the target compound, and plan a sequence of chemical reactions, similar to pathways or routes, to synthesize the target compound in an incredibly large combinatorial space of possible chemical reactions. Most importantly, a chemist needs to experimentally validate each step of the process and finally use the gained insights to validate new knowledge and creatively think of new hypotheses about chemistry. Success depends on whether the products of each reaction step are synthesized as predicted in sufficient quantity and on whether additional requirements like efficacy and non-toxicity are fulfilled.

Developing new drugs involves generating and evaluating very large numbers of chemical compounds. Most compounds turn out as negative samples, as they do not exhibit the desired properties, such as being an effective drug. The development of one new drug typically takes over 10 years and costs over one billion US dollars (Dickson and Gagnon

2004; Scannell et al. 2012). Most resources and time are spent during the research and development phase (typically 5-7 years) and clinical trial stages (6 years). Therefore it is most critical to find the best, most promising candidate molecule as quickly as possible, and we are convinced AI can play a critical role here.

For decades, discovering new compounds and planning their chemical syntheses has been a scientific research challenge in organic chemistry and material science. Grand Challenge #2 defined by the US Department of Energy, still unsolved, refers to designing and perfecting atom- and energy-efficient synthesis of revolutionary new forms of matter with tailored properties (Fleming and Ratner 2007). This challenge may be considered to include chemical compound discovery and synthesis planning.

Databases such as ReaxSys[1] and SciFinder[2] allow to search through the literature for reactions discovered in the past. Additionally, analyzing a large scale graph representing existing chemical reactions in organic chemistry (Kowalik et al. 2012; Szymkuć et al. 2016) enables chemists to find useful existing reaction pathways to synthesize existing compounds. However, discovering new compounds and new reaction pathways remains largely a manual process, depending on the experience and the intuition of a human expert.

Chemistry takes humans decades to master and it is exciting to imagine an AI that can reach or exceed expert human performance in this field. Such an AI would be very valuable and significantly advance chemistry-related industries like pharmaceuticals, food, and materials.

Why Is Chemistry Challenging?

At most a few hundred million chemical substances are currently known.[3] On the other hand, in the chemistry research community, there is a consensus that the number of candidate drug compounds which could theoretically exist is estimated to be 10^{60} (Peplow 2014). This is larger than the number of possible positions in chess, estimated to 10^{52} (Allis 1994), but smaller than the number of positions in Go, estimated to 10^{172} (Allis 1994). Upper bounds estimations are also available for puzzle domains used as AI benchmarks. For example, 10^{25} for the 24-puzzle, 10^{19} for the Rubik's Cube, and 10^{98} for Sokoban (Junghanns and Schaeffer 2001).

This size comparison, and the fact that state-of-the-art chess and Go programs outperform the best human players indicate that AI is mature enough for a major shift towards chemistry. At the same time, chemistry poses additional challenges to tackle, besides the size of its state space.

First of all, in games, the root state of a search is either the game start position or the position generated by the opponent's response. That is, the root state is already given to the game-playing program, when it searches for the next move. On the other hand, in chemistry, AI needs to find the target compound to consider from the chemical compound space of 10^{60}. In a sense, this crucial task in chemistry is compared to the task of finding and creating an aesthetic chess problem. This boils down to finding a chess position (the problem) in the large space of all chess positions, estimated to 10^{52} states, as said earlier. This is typically performed by a human chess problem composer. When it comes to AI-based chess problem creation, despite previous attempts, e.g., (Hirose, Matsubara, and Itoh 1997; Schlosser 1988), algorithms still are in an infant stage.

Additionally, games and puzzles, such as those mentioned earlier, can be encoded as a search problem in a perfect way, with no information loss. It is trivially easy to decide whether a given game position is a valid state (e.g., respecting the game rules), or whether it is a goal state. So is enumerating all the valid moves available in a state.

In contrast, in chemistry, even if a candidate of a target compound is designed, deciding whether that candidate is a good target compound (a "goal state") is difficult. Answering this may involve evaluating the toxicity, the effectiveness to the purpose (e.g., can it cure the disease?), and the manufacturing costs. Among the many compounds available, typically a very small number meet the "goal state" criteria. It is often difficult to predict how a compound would behave in practice, unless an experiment is carried out.

Evaluating a candidate sequence of reaction steps meant to synthesize a given compound is equally difficult. In chemical synthesis planning, Szymkuć et al. (2016) discuss that, for a synthesis route with 30 steps, the number of possible pathways to consider is estimated to be 1.2×10^{57}. Among them, there are usually only a very small number of feasible pathways.

As said earlier, game and puzzles have clearly defined rules which enable to easily generate legal transitions from one state to another (moves in games). However, in chemistry, the situation is much more difficult. A reaction rule is a pattern showing how a set of reactants could interact with each other, and what the result of a chemical product would be. Checking whether a reaction rule is applicable or not involves a step that boils down to solving a subgraph isomorphism problem. This is an NP-complete problem, creating a serious bottleneck when generating the successors of a state. For example, the implementation of Heifets and Jurisica (2012) generates at most several tens of compounds per second, as compared to millions of positions per second, which can be achieved in a game like chess.

Compared to typical games and puzzles, there would be a much large number of choices ($> 10,000$) for a one-step reaction in chemistry, depending on the structure of the compounds (Szymkuć et al. 2016).

In chemistry, rules often are ambiguous or even incomplete, as the knowledge available is not always accurate.

AI Technologies for Chemistry

We discuss essential AI technologies for tackling chemistry.

NLP and Chemical Image Processing

The vast literature available, including academic papers and patents, contains many chemical reactions. Automatically

[1] http://www.elsevier.com/online-tools/reaxys

[2] https://scifinder.cas.org

[3] https://www.cas.org/content/chemical-substances

extracting these would lead to an extensive and valuable knowledge base available in a machine-readable format.

Molecule structures as well as reactions are typically illustrated as figures, with a textual description of details such as yields and temperatures. Unlike typical text mining domains, this clearly requires a combination of image processing and NLP. In addition, unlike approaches that attempt to extract only compound structures, reaction extraction is much more difficult, since it needs to detect relations between compounds and reactants.

Lowe (2012) automatically extracts reactions by mining the relevant experimental sections. He extracts millions of chemical reactions from the US patent literature (Lowe 2012) and makes these data publicly available.[4] However, there are still many important reactions whose reactants are incorrectly classified. Additionally, these extracted reactions do not have essential information on the conditions on the reactions, which needs to be obtained by improved text mining technologies.

Machine Learning

Once we have data, Machine Learning is a promising approach to modeling characteristics of chemical compounds. In particular, Deep Learning is a strong candidate due to its success in other domains.

There are a few challenges for Machine Learning in chemistry. First of all, reaction data extracted with NLP and image processing might be noisy and contain erroneous information. Secondly, these data do not have any negative examples. That is, failed reactions are typically not publicly available. Finally, these data do not always best representative reactions. In fact, chemists typically do not employ reactions as described in papers. To increase the yield of the target compound, they perform the reactions in different conditions, such as temperatures and catalysts. These know-hows are rarely shared as public documents. Machine Learning algorithms need to be robust to such circumstances. At the same time, the accuracy of prediction plays a crucial role in chemistry, since a poor prediction could result in days or weeks of wasted time and money.

Machine Learning has started being employed in chemistry-related tasks. For example, while quantum simulations are often used to screen out useless chemical compounds, they have an expensive computational overhead. Gómez-Bombarelli et al. (2016) employ Deep Learning to avoid unnecessary calculations of quantum simulations.

Predicting reactions can be addressed with Deep Learning. Segler and Waller (2017b) generalize existing reactions and one-step reaction rules that hold only essential substructures of products and reactants. Then, with neural networks, Segler and Waller predict which reaction rules can be applied to return a product, given a reactant as input; and to return a reactant, given a product as input. Coley et al. (2017) attempt to model a product as a "true" product if that product is generated by a reaction recorded in the patent literature and as a "false" product otherwise. Their approach performs Deep Learning with features based on the changes of reactants and calculates the score of a generated product. Savage et al. (2017) employ graph-link-prediction-based recommendation algorithms to predict reactants, given a product as input.

In performing prediction tasks, such as predicting reactants, e.g., (Segler and Waller 2017b; Savage et al. 2017), a chemical compound represented as a graph is often transformed into a fingerprint. A fingerprint is a fixed-size bit vector that compactly represents structural information on the compound as chemical features. There are a several approaches to selecting features to include (Morgan 1965; Rogers and Hahn 2010), which do not always reflect the structural characteristics of the compound. The neural graph fingerprints presented by Duvenaud et al. (2015) showed great predictive performance.

Another important task that could be addressed with Machine Learning is the detection of active substructures in a compound. A compound can consist of an active substructure, such as the part that fights a disease, and a supporting part, that completes the compound as a stable structure.

Search and Planning

Given a target compound, chemical synthesis planning is a well-known problem since the 1960s, and it still remains an open problem today (Corey and Wipke 1969; Corey 1967; Corey and Cheng 1995). Chemists typically solve the chemical synthesis planning problem by performing a so-called *retrosynthetic analysis*.

A reaction pathway is a sequence of one-step reaction rules. Retrosynthetic analysis attempts to systematically examine possible combinations of one-step reaction rules and find a reaction pathway that leads to the target compound from a set of commercially available compounds (i.e., start materials). Chemoinformatics researchers have employed a search-and-planning based approach to perform retrosynthetic analysis, e.g., (Law et al. 2009). Challenges in chemical synthesis planning have been overviewed in the previous section. In addition, effective algorithms capable to perform a so-called k-best first search are necessary to return several reasonable solutions, so that an expert choose one or several.

Heifets and Jurisica (2012) have recently modeled retrosynthetic analysis as a procedure of solving a position in two-player games such as chess and checkers. They also make benchmark problems publicly available.[5] We argue that chemical synthesis planning is an ideal domain where game researchers could export their knowledge and technologies. For example, Heifets and Jurisica (2012) employ proof-number search (PNS) (Allis, van der Meulen, and van den Herik 1994), a well-known approach that contributed to solving the game of checkers (Schaeffer et al. 2007). Many algorithms related to PNS are a strong candidate for efficiently discovering pathways. Kishimoto et al. (2012) present a survey of PNS. Additionally, inspired by AlphaGo (Silver et al. 2016), Segler, Preuß, and Waller (2017b; 2017a) apply Monte Carlo Tree Search (MCTS)

[4] https://bitbucket.org/dan2097/patent-reaction-extraction/downloads

[5] http://www.cs.toronto.edu/\~{}aheifets/ChemicalPlanning/

(Kocsis and Szepesvári 2006), combined with Deep Neural Networks that are used to bias Monte Carlo samplings (Segler and Waller 2017b). There are many MCTS techniques that could be useful for synthesis planning (Browne et al. 2012).

When modeling the discovery of new compounds as a search problem, we need an effective representation of the problem. A naive search that would explore the whole search space of compounds (i.e., in the order of 10^{60} states) is not an option. There are constraints and characteristics for the compounds to discover, such as substructures of active parts that determine the functions of the compounds, and chemical core scaffolds that roughly determine the structures of compounds. These constraints and characteristics can be given either by a human expert or by Machine Learning AI. Then, search algorithms should be able to accurately detect the promising portions of the search space.

Conclusions

AI in chemistry will, in its essence, need to master the working principles of modern scientists, the Scientific Method. The seemingly simple, but even for humans challenging, iterative process of scientific discovery consisting of making observations, asking questions, proposing testable hypotheses, designing the right experiments to prove hypotheses, collecting and analyzing data from experiments and drawing conclusions that lead to new theories. Promising developments and advances have been made by researchers of both AI and chemistry in their respective fields. However, a lot of research remains to be done and this paper is outlining how a focused, collaborative effort between AI and chemistry research communities could produce valuable discoveries and contributions to Science for the benefit of humanity.

References

Allis, L. V.; van der Meulen, M.; and van den Herik, H. J. 1994. Proof-number search. *Artificial Intelligence* 66(1):91–124.

Allis, L. V. 1994. *Searching for Solutions in Games and Artificial Intelligence*. Ph.D. Dissertation, University of Limburg.

Browne, C.; Powley, E.; Whitehouse, D.; Lucas, S.; Cowling, P. I.; Rohlfshagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; and Colton, S. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games* 4(1):1–43.

Campbell, M.; Jr., A. J. H.; and Hsu, F. 2002. Deep Blue. *Artificial Intelligence* 134(1–2):57–83.

Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; and Jensen, K. F. 2017. Prediction of organic reaction outcomes using machine learning. *ACS Central Science* 3:434–443.

Corey, E. J., and Cheng, X.-M. 1995. *The Logic of Chemical Synthesis*. Wiley.

Corey, E. J., and Wipke, W. T. 1969. Computer-assisted design of complex organic syntheses. *Science* 166:178–192.

Corey, E. J. 1967. General methods for the construction of complex molecules. *Pure and Applied Chemistry* 14:19–38.

Dickson, M., and Gagnon, J. P. 2004. The cost of new drug discovery and development. *Discovery Medicine* 4(22):172–179.

Dragone, V.; Sans, V.; Henson, A. B.; Granda, J. M.; and Cronin, L. 2017. An autonomous organic reaction search engine for chemical reactivity. *Nature Communications* 8:1–8.

Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, 2224–2232.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; Schlaefer, N.; and Welty, C. 2010. Building watson: An overview of the DeepQA project. *AI Magazine* 31(3):59–79.

Fleming, G., and Ratner, M. 2007. Directing matter and energy: Five challenges for science and the imagination. Technical report, US Department of Energy, Office of Basic Energy Sciences. Available at http://www.sc.doe.gov/bes/reports/abstracts.html#GC.

Friedland, N. S.; Allen, P. G.; Matthews, G.; Witbrock, M.; Baxter, D.; Curtis, J.; Shepard, B.; Miraglia, P.; Angele, J.; Staab, S.; Moench, E.; Oppermann, H.; Wenke, D.; Israel, D.; Chaudhri, V.; Porter, B.; Barker, K.; Fan, J.; Chaw, S. Y.; Yeh, P.; Tecuci, D.; and Clark, P. 2004. Project halo: Towards a digital aristotle. *AI Magazine* 25(4):29–47.

Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; andSunghan Kim, M. N.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; and Aspuru-Guzik, A. 2016. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* 15:1120–1127.

Heifets, A., and Jurisica, I. 2012. Construction of new medicines via game proof search. In *AAAI*, 1564–1570.

Hirose, M.; Matsubara, H.; and Itoh, T. 1997. The composition of tsume-shogi problems. In *Advances in Computer Chess*, volume 8, 299–318.

Jin, W.; Coley, C. W.; Barzilay, R.; and Jaakkola, T. 2017. Predicting organic reaction outcomes with Weisfeiler-Lehman network. In *NIPS*. To appear. The preprint version is available at https://arxiv.org/abs/1709.04555.

Junghanns, A., and Schaeffer, J. 2001. Sokoban: Enhancing general single-agent search methods using domain knowledge. *Artificial Intelligence* 129:219–251.

Kishimoto, A.; Winands, M.; Müller, M.; and Saito, J.-T. 2012. Game-tree search using proof numbers: The first twenty years. *ICGA Journal, Vol. 35, No. 3* 35(3):131–156.

Kitano, H.; Asada, M.; Kuniyoshi, Y.; Noda, I.; and Osawa, E. 1997. RoboCup: The robot world cup initiative. In *Proceedings of the 1st International Conference on Autonomous Agents*, 340–347.

Kocsis, L., and Szepesvári, C. 2006. Bandit based Monte-Carlo planning. In *17th European Conference on Machine Learning (ECML 2006)*, volume 4212 of *Lecture Notes in Computer Science*, 282–293. Springer.

Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; and Bishop, K. J. M. 2012. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angewandte Chemie (International ed. in English)* 51(32):7928–32.

Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; and Ando, H. Y. 2009. Route Designer: A retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *Journal of Chemical Information and Modeling* 49(3):593–602.

Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; and Lederberg, J. 1980. *Application of Artificial Intelligence for Organic Chemistry: The Dendral Project*. McGraw-Hill.

Lowe, D. M. 2012. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. Dissertation, University of Cambridge.

Morgan, H. L. 1965. The generation of a unique machine description for chemical structure. *Journal of Chemical Documentation* 5(2):107–113.

Peplow, M. 2014. Organic synthesis: The robo-chemist. *Nature* 512:20–22.

Rogers, D., and Hahn, M. 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 50(5):742–754.

Savage, J.; Kishimoto, A.; Buesser, B.; Diaz-Aviles, E.; and Alzate, C. 2017. Chemical reactant recommendation using a network of organic chemistry. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys)*, 210–214.

Scannell, J. W.; Blanckley, A.; Boldon, H.; and Warrington, B. 2012. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature Reviews Drug Discovery* 191–200.

Schaeffer, J.; Burch, N.; Björnsson, Y.; Kishimoto, A.; Müller, M.; Lake, R.; Lu, P.; and Sutphen, S. 2007. Checkers is solved. *Science* 317(5844):1518–1522.

Schaeffer, J. 2009. *One Jump Ahead: Computer Perfection at Checkers*. Springer.

Schlosser, M. 1988. Computers and chess-problem composition. *ICCA Journal, Vol. 11, No. 4* 11(4):151–155.

Segler, M. H. S., and Waller, M. P. 2017a. Modelling chemical reasoning to predict and invent reactions. *Chemistry – A European Journal* 1521(3765).

Segler, M. H. S., and Waller, M. P. 2017b. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry – A European Journal* 1521(3765).

Segler, M.; Preuß, M.; and Waller, M. P. 2017a. Learning to plan chemical syntheses. Available at https://arxiv.org/abs/1708.04202.

Segler, M.; Preuß, M.; and Waller, M. P. 2017b. Towards "AlphaChem": Chemical synthesis planning with tree search and deep neural network policies. Available at https://arxiv.org/abs/1702.00020.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; and Hassabis, D. 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529:484–489.

Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; and Grzybowski, B. A. 2016. Computer-assisted synthetic planning: The end of the beginning. *Angewandte Chemie International Edition* 55(20):5904–5937.