From Atoms to Bytes: Prediction of Molecular Phenomena Enabled by Chemical Data Mining

by
Jonathan Zheng

Envisioning the Future of Computing Prize
Social and Ethical Responsibilities of Computing
Massachusetts Institute of Technology

Massachusetts Institute of Technology, Envisioning the Future of Computing
Prize: Social and Ethical Responsibilities of Computing

From Atoms to Bytes

Prediction of Molecular Phenomena
Enabled by Chemical Data Mining

February 3, 2024

Executive Summary

Artificial intelligence (AI) models for generating digital media have captured the attention of the world. But such models do not yet exist for physical phenomena. Chemistry, in particular, has a variety of challenges that preclude it from maximally benefiting from AI technologies.

Molecules, as inherently non-digital items, are difficult to model. Chemical deep learning requires a lot of data to work effectively, but molecular data is much more scarce than digital media.

One issue is that chemical data existing in print literature has not yet been digitized, and therefore remains inaccessible. Literature involving chemistry tends to be difficult for computers to parse, as a full understanding of the science requires simultaneous comprehension of language and chemical diagrams. Emergent AI-driven technologies in data mining are promising tools for extracting this information, enabling the compilation of large datasets that can be used to train AI models for molecular property prediction and generative chemistry. Such technology will also lower the financial and time burdens of accessing chemical information, allowing data to be more freely available.

These AI technologies carry risks. Data mining could perpetuate false information and biases, and improperly utilize intellectual property. Generative AI models have the risk for harm, as they could be used by people to create chemicals that hurt others, though technological and policy-based safeguards can temper these risks. But the potential benefits outweigh the potential drawbacks; better AI models in chemistry can help us develop materials that address pressing societal issues.

There is ultimately a risk that AI models for chemistry will never reach the heights of ChatGPT or DALL-E. But we will never know unless we try, and we owe it to ourselves to invest as much effort into building a chemical oracle as we have spent crafting articulate chatbots.

1 Introduction

Digital media is now as accessible as a Google search. Large language models (LLMs) and their derivatives, such as OpenAI's popular ChatGPT, can dream up plausible human-sounding text responses to practically any prompt. Generative artificial intelligence (AI) can transform words into vivid pictures. This is not to mention the many recent developments in generating videos, sound, and 3D models.

But what about AI for the natural world? The possibility of using AI for chemistry is particularly promising. Instead of conjuring up images that match a text prompt, what if AI could generate molecules that match the properties we want? Perhaps this sounds mundane, but the ability to dream up performant materials would revolutionize the world. We as a society are in need of better materials for renewable energies and alternative fuels; new biodegradable polymers to supersede plastic; novel cures for the manifold maladies that afflict life on Earth; and countless other as-of-yet unknown chemicals central to solving society's key issues.

There are many reasons why such advances haven't arrived already. Chemicals are not digital media, and the consequences of that inconvenient truth are plentiful and will be discussed herein. A major issue is that chemical data is not nearly as widespread as text and image data. For instance, GPT-3 (an old version of OpenAI's large language model released in 2020) was trained on approximately 400 billion tokens (units of semantic text). In contrast, chemical datasets typically include just tens of thousands of entries.

Vast numbers of scientific experiments have been conducted - and continue to be conducted daily. However, compiling and standardizing them is difficult. Scientific results are published independently, reported non-uniformly, and embedded in figures, charts, or paragraphs of text. Compilations of data exist, but many are exclusively available in print form and require effort to transform into a useful digital form.

Data mining tools will be critical to furthering our understanding of chemistry. These AI-driven technologies will vastly improve the availability of chemical data and enable bigger, better, and more accessible models to be built. In the short-term, these will allow scientists to accurately predict the properties of materials, which could aid computer-assisted screening and the design of chemical processes. In the long-term, these advances may unlock the ability to construct chemicals for solving societal issues. While the advent of such technologies has its risks, the potential for societal good is tremendous.

2 Why we need data mining in chemistry

2.1 The case for more data

Data is the lifeblood of modern artificial intelligence. Most AI is powered by deep learning, which uses data to update parameters of sequences of computing units. Though these technologies are able to learn arbitrarily complex relationships, they usually require vast amounts of training data to work well.

Chemical data is challenging to work with. Chemicals are physical items that we can only model at varying degrees of complexity: for instance, as 2D graphs or as 3D collections of atoms. Chemical representations also tend to be big. A simple representation might require 512 pieces of information to describe a single molecule. Larger representations can be more descriptive, but are more challenging to train and may require more data to learn meaningful insights.

Then there is the aspect of feasibility. Popular estimates for the number of possible drug-like compounds vary between 10^{23} to 10^{180}, many magnitudes higher than the number of molecules that have ever been studied.[1] The design space for molecules is huge, and we have only explored a comparatively tiny fraction of that cosmos. There is a need for diverse data - especially for generative AI, whose goal is often to find exceptional chemicals.

Recently released property datasets include tens of thousands of datapoints and tend to be manually compiled, or are collections of smaller compilations of data. By themselves, these are often too small and homogeneous to work well. One approach is to augment experimental data with large datasets of properties estimated through molecular simulation. With these advances, property prediction has seen its share of recent success stories. But this approach is are limited by the accuracy and speed of the underlying models, and thereby are currently feasible for only a subset of properties. More experimental data is still very much needed.

Automated data mining serves to increase the quantity and diversity of experimental data. Recent applications in the physical sciences show the potential upsides - and the ongoing challenges.

2.2 The potential for data mining in chemistry

The amount of scientific literature out there is staggering. PubMed, a database that archives research in the biomedical and life sciences, indexes more than 1 million papers annually.[2] It is infeasible for a human to read and parse every article. But what about a computer?

Data mining in chemistry has seen some recent success. In 2018, research by the Cole group in Cambridge employed natural language processing to extract nearly 40,000 datapoints related to magnetic materials.[3] Then, in 2020, the same group released a dataset with almost 300,000 records of data related to batteries using a similar approach.[4] Since then, several similar works have appeared in the literature, often yielding datasets with thousands or tens of thousands of entries.[5,6].

However, such approaches are limited by the accuracy of the data mining. In the battery dataset paper, the authors estimates that about 20% of the scraped data were extracted without errors and that "only about three fifths of the data records were extracted from text."[4]

Tools for data mining have historically not been well-suited for chemistry. Scientific literature is difficult to parse, especially when information is encoded into images and tables. Much of the chemical data literature is decades-old and need to be digitized from images. However, tools for digitizing chemical literature are often not very performant.[7] One challenge is the identification of molecular structures from drawings, which can vary widely in shape, style, and size. Another is the complexity of the language involved in describing chemical information. To address these issues, new computational technologies are being developed.

2.3 Emergent data mining technologies will involve AI

Researchers at MIT have recently developed AI tools tailored for chemical data mining: MolScribe, RxnScribe, and ChemRxnExtractor.

MolScribe translates molecular images into computer-parsable structures. MolScribe leverages modern computer vision architecture, and was pre-trained to classify 14 general million images then fine-tuned to parse diagrams of chemical structures, both synthetic and extracted from patent literature.[8] RxnScribe similarly parses images of chemical reaction sequences.[9] ChemRxnExtractor is a tool for extracting chemical reaction information from article text.[10] Several modules outside of MIT have also been developed.[11,12] Each tool is trained on large datasets of compiled data, and will be used to obtain data that future developers can utilize to build even more capable software - leading to a snowball effect of advancement in AI.

All of the software discussed in this section involved extensive manual annotation. The manual element of compiling data is not going anywhere. Still, there are opportunities in the immediate future for human effort and computer-assisted data extraction to go hand-in-hand. Several digital versions of existing print compilations have been recently published online. These can include tens

of thousands of data, and tend to be reliable due to the manual effort that went into their compilation.[13,14] Such digitizations often involve computer analysis, such as text recognition, followed by manual validation and post-processing to standardize the data for computer usage. Advances in character recognition, language processing, and chemical structure parsing will serve to make the digitization process even faster.

3 Societal benefits

Data mining has the potential to democratize access to information. Much of the scientific literature corpus is privately held by publishers, who charge hefty fees to access articles. Many high-quality datasets, and models trained on such data, are also proprietary, costing in the thousands of dollars to access (and often with restrictions on usage).

But the cost of accessing a data-mined compilation is nearly, if not completely, free, allowing any person with a computer to make models, draw insights, and study trends. Data mining will enable academic and non-profit groups to develop competing software, lowering the barrier to use chemical technology. This may change the pricing and access model for publishers - more journals may publish their articles openly, but charge for programmatic article downloads. The net result, regardless, is that the control of information returns from large, powerful institutions to the rest of the world.

As data mining develops, AI models will also improve. AI in chemistry typically involves either property prediction or generative modeling. The latter is similar to what you expect from a text-to-image model; we might ask the computer to generate the structure of a molecule with desired properties, like a highly-soluble compound that binds to a certain cellular receptor.

Though such technologies wouldn't be directly used by most people, they certainly would still transform society. For one, the pharmaceutical world will look completely different. At present, drug discovery involves synthesizing many thousands of different drug candidates, and then testing them to see which one reacts favorably. This entire process can cost billions of dollars and take many years of work.[15] However, as our chemical AI models become more powerful, this development process will become quicker and cheaper, accelerating the availability of cures for treatable diseases. An analogous application is screening for porous zeolite and metal-organic framework materials to use in gas storage and separation (which could enable usage of environmentally friendly fuels) and as catalysts (used in a large majority of manufacturing processes to speed up reactions). We are already starting to see AI-driven advancements in these realms, and

will only continue to see more as the technology develops.

Many of today's most pressing issues are driven by a need for advanced materials. For instance, we would like to discover biodegradable polymers to replace the current paradigm of using more-or-less permanent plastics. Ongoing research efforts, including those at MIT, investigate whether biodegradability can be predicted from chemical structure.[16] Although the chemistry of biodegradation is less well-understood than those for drug interactions, data mining could help scientists discover heretofore unknown trends.

4 Potential societal risks

No technology is without its risks, and AI in chemistry is no exception.

4.1 Risks of data mining

Scraping data correctly is not easy. There is the risk of perpetuating incorrect information; it is infeasible for humans to double-check the accuracy of every data entry. "Good" datasets tend to be foundational, in the sense that many models will use the same dataset. In this sense, incorrect data can lead to erroneous predictions in *many* chemical predictors and have unintended consequences in the real-world; for instance, in models that are used to make decisions for clinical applications. Or, such data could include biases that unfairly affect different groups of people, such as biomedical data extracted from an unrepresentative subset of the population.

Data mining also currently does not evaluate the legitimacy and accuracy of the information it collects. Scientific theory evolves over time, and sometimes experimental data are discarded or superseded by experiments conducted afterwards. Combined with the fact that data scraping tends not to be as accurate as manual compilation by a human, this problem of poor data is not only a risk, but an ongoing challenge precluding more widespread usage.

Then there is the risk of privacy. Similar to the copyright issues that arise between generative image models and artists, there is an issue of who owns the copyright when involving publishers. In the case of digitizing existing compilations of data from print, copyright policies may prevent the digital formats from being published. These issues are crucial to discuss as data mining accelerates. Many academic publishers today include limitations on data mined from scholarly works, and it is entirely possible to see a world where this becomes the default, vastly limiting the potential for data mining.

Finally, data mining is a costly process. It involves both heavy computation and memory storage. There are energy and carbon emission costs associated with running server architecture. As societal reliance on computation increases, it is important for sustainable energy to be considered in tandem.

These issues are all solvable, and will require ample debate and discussion in society. Technology and policy must come together to address these issues.

4.2 Risks of advanced chemical AI models

Just as in any other AI technology, chemical deep learning is vulnerable to potential misuse. A chemical calculator that's capable of generating a capable curative could be retooled to find a powerful poison.

Another more insidious risk is the possibility of unintended consequences from constraining such technology to prioritize profit while disregarding underemphasizing personal and environmental safety. One can imagine an engineer asking the oracle for the least expensive route to manufacture a chemical, then unwittingly fomenting an environmental disaster. Humans need to thoughtfully consider AI predictions and (at least in the immediate future) will continue to be part of the decision-making process any time a chemical is involved.

But despite the existence of these risks, there are several ways to dampen their impact. Training data can be omitted and safeguards can be placed such that malicious usages of AI are blocked. These safeguards may be supplemented by regulations regarding the allowable usage of generative chemical AI. Also, even if someone is able to use such models to learn how to synthesize a dangerous chemical, they are far more likely to be constrained by the ability to synthesize it than the knowledge of making it. And in the case of governments using such technologies to develop chemical weapons, this is possible today even without AI (and restricted under global disarmament laws). Hence, there is far more potential for chemical AI to help than to harm.

Ultimately, despite its potential, there is even a risk that the technology never works. Chemical data comes with its difficulties and might not be suitable for generative modeling after all. The risk here is that time and financial investments - and the carbon footprint of powering the supercomputers for data mining and model development - sputter out and lead nowhere. But we will never know unless we try.

5 Conclusion

AI in chemistry faces several unique challenges, stemming from the disconnect between physical and digital forms of data. Experimental data is not frequently available in a computerized format, opening up opportunity for data and text mining. The availability of large datasets, facilitated by using AI-driven data mining tools, is likely to result in more powerful and accessible chemical models.

On the other hand, there is no guarantee that generative AI will *ever* reach the heights and usability that models like ChatGPT have in stock. And even if they do, there are serious risks to consider: perpetuation of false information, privacy concerns, and usage of chemical AI with harmful intentions.

Still, the potential benefits for society are astronomical. The sheer potential for good outweighs the risks, which can be mollified through technological and policy-based safeguards. Data mining in chemistry will lower the financial and time barriers to accessing data. In the future, such research might lead to models address issues in medicine, energy, sustainability, manufacturing, and any other industry involving a chemical substance. We owe it to ourselves to invest as much effort into building a chemical oracle as we have spent making articulate chatbots.

References

(1) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. *Journal of computer-aided molecular design* **2013**, *27*, 675–679.

(2) Landhuis, E. *Nature* **2016**, *535*, 457–458.

(3) Court, C. J.; Cole, J. M. *Scientific data* **2018**, *5*, 1–12.

(4) Huang, S.; Cole, J. M. *Scientific Data* **2020**, *7*, 260.

(5) Foppiano, L.; Castro, P. B.; Ortiz Suarez, P.; Terashima, K.; Takano, Y.; Ishii, M. *Science and Technology of Advanced Materials: Methods* **2023**, *3*, 2153633.

(6) Jensen, Z.; Kwon, S.; Schwalbe-Koda, D.; Paris, C.; Gómez-Bombarelli, R.; Román-Leshkov, Y.; Corma, A.; Moliner, M.; Olivetti, E. A. *ACS Central Science* **2021**, *7*, 858–867.

(7) Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. *Iscience* **2021**, *24*.

(8) Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; Barzilay, R. *Journal of Chemical Information and Modeling* **2023**, *63*, 1925–1934.

(9) Qian, Y.; Guo, J.; Tu, Z.; Coley, C. W.; Barzilay, R. *arXiv preprint arXiv:2305.11845* **2023**.

(10) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. *Journal of chemical information and modeling* **2021**, *62*, 2035–2045.

(11) Dong, Q.; Cole, J. M. *Journal of Chemical Information and Modeling* **2023**, *63*, 7045–7055.

(12) Swain, M. C.; Cole, J. M. *Journal of chemical information and modeling* **2016**, *56*, 1894–1904.

(13) Biswas, S.; Chung, Y.; Ramirez, J.; Wu, H.; Green, W. H. *Journal of Chemical Information and Modeling* **2023**, *63*, 4574–4588.

(14) Zheng, J. IUPAC/Dissociation-Constants: v1. 0 (v1-0_initial-release)[Data set]. Zenodo, 2022.

(15) Kim, E.; Yang, J.; Park, S.; Shin, K. *Therapeutic Innovation & Regulatory Science* **2023**, 1–14.

(16) Fransen, K. A.; Av-Ron, S. H.; Buchanan, T. R.; Walsh, D. J.; Rota, D. T.; Van Note, L.; Olsen, B. D. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2220021120.