Autonomous Molecular Discovery and Chemicals for Peace

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

Autonomous Molecular Discovery and Chemicals for Peace

February 1, 2025

Executive Summary

We humans want chemicals to *do things*, but we don't know what molecules will induce those responses. For instance, we want disease-curing drugs, efficient solar cell materials, carbon-capturing compounds, biodegrading polymers, and so on, but what do those chemicals look like? And even if we could imagine *what* to make, *how* do we make them?

In the early 2020s, researchers at MIT took a leap forward to solving those problems: a platform for semi-autonomous molecular discovery. This technology proposes molecules, using property prediction models to assess each chemical candidate. These candidates are then synthesized using a robotic lab, and tested using onboard analytical chemical tools. These measurements are then used to update the property prediction models. This process is repeated iteratively, exploring new areas of chemical space and becoming more effective with each run. After several cycles, the tool has learned from its exploratory runs, and is used to *generate* compounds that are highly stable and red-absorbing. The human involvement was small, limited to just "setting and adjusting objectives, providing requested materials, and occasionally fixing unrecoverable errors" [1]. In other words, this platform could almost autonomously invent molecules that *do useful things.*

Although that technology is real, its full potential has not yet been actualized. In the fictional history presented in this story, society catches up. Autonomous molecular discovery platforms are used for the first time to autonomously propose and synthesize an effective cancer-treating drug. In addition to working well for biochemical targets, the technology shows promise that this can be applied to other domains, like for developing sustainable materials and renewable energies.

But flip this formula on its head, and something tragic happens. Predictive models work by finding optima - *optimally stable, optimally efficacious, optimally good.* But what if you wanted the opposite? *Optimally toxic, optimally explosive, optimally harmful.* Is this a real threat? Do the downsides outweigh the positives? These are questions we must consider as a society as we advance toward the future of chemical automation.

1 The Hierophant

They had finally done it. Or so they had said, so many times, for so many years; but it seemed like it was really true this time. *A chemical breakthrough. An innovation in AI.* Familiar words ring from the stage. I've heard these words so many times throughout my career as a researcher - usually well-intentioned words, but overstatements nonetheless. At this press conference, something in the air tells me that the situation's quite different.

A team of young men and women in lab coats stands by a news podium. A white-haired woman with sunken eyes and a creased smile is gesturing vivaciously toward a crowd of press reporters. She speaks of automation, innovation, revolution, determinations: all sorts of -ations, before landing on the word "verdenir." A new drug. She points to a bar chart on the screen behind her, and explains that the small molecule had just been approved by the FDA, having passed through an extremely successful round of clinical trials in which the drug eliminated cancerous tumors caused by a certain genetic mutation. Nobody is in doubt when she says that verdenir is poised to save countless lives and curb immeasurable suffering.

A new cancer cure, especially one specific to a single form of the disease, wasn't revolutionary in itself, she goes on to explain. After all, over a decade prior, in 2018, larotrectinib had been approved by the FDA for essentially the same application - tumor-agnostic treatment caused by mutations in NTRK genes [2]. The patient had to have that specific type of cancer for it to work, but if they did, then larotrectinib was potentially life-saving. This was followed up by repotrectinib, which was approved by the FDA in 2023 for targeting cancer caused by NTRK gene fusions [3]. A smattering of other similar drugs followed, the most recent of which was just approved in 2030.

It wasn't the mechanism of this new molecule verdenir that made the headlines, but rather, how the drug itself was discovered. It was found using a fully-autonomous discovery platform, the woman explains. And this platform had the potential to revolutionize the entire way we develop medicine - and beyond this, how we discover novel materials.

There is an air of tense excitement as she makes this last declaration. She pauses, and then explains:

The process of discovering new drugs is a very expensive endeavor: from lab to market, it can cost billions of dollars and take about a decade of work [4]. Traditional drug discovery involves identifying the cause of a disease, proposing numerous drug candidates, and then making and

testing them to see if they "hit" the biochemical target that is linked to the disease. Successful candidates at this stage are then exhaustively studied to understand their mode of action and physicochemical properties, formulated, analyzed, and potentially tested on a small number of animals. This preclinical stage is expensive, often costing many millions of dollars [5]. But the most expensive and risky stage is when the drugs are tested on people in clinical trials. Indeed, about 90% of drugs fail clinical trials [6], about half of which is due to lack of efficacy [7]. At any point, the drug can be deemed unfit, undoing the entire preceding chain of events. Having a good list of drug candidates to start with would save huge amounts of time, vitality, and money across all of these stages.

The goal of the autonomous drug discovery platform is to find the top few candidate drugs computationally, reducing the expenses in pursuing suboptimal drug candidates while maximizing the chance of drug feasibility. To do so, it used a so-called design-make-test-analyze (DMTA) cycle that employed numerous different applications of artificial intelligence.

First, the platform generates candidate compounds, using algorithms that estimate pharmacological and chemical properties deemed relevant to a drug, such as solubility, toxicity, and synthesizability. By using chemical property predictors, which are data and physics-driven models that predict the relevant properties of each compound, the platform identifies the chemicals that it can predict with the *least* confidence. This subset of compounds is then synthesized and analyzed in a robot lab, and tested against the biomarkers. These tests are used to update the AI models, improving both the property predictors as well as the platform's understanding of what makes a good drug candidate. This means that the model becomes increasingly accurate with each subsequent cycle, exploring new areas of chemical space. After numerous calibration steps, the platform is used to generate a list of molecules with *optimal* predicted drug properties. This, apparently, had led to a batch of molecules of which several passed to human trials, including verdenir.

It was odd that the technology had worked so well. The property predictors were refined only in an artificial setting, using a robot lab which could only measure *proxies* for toxicity and drug distribution in the body (which were impossible to ethically study in a robot lab setup, and so an artificial approximation had to suffice). These may improve the pre-clinical process, but there was still no guarantee that the platform would improve the drug development process holistically. But the technology, through many cycles of training, had still apparently learned higher-order information about biochemistry that led to higher success rates in the downstream clinical trials.

I sit back and think - this wasn't anything especially new. When I was a grad student at MIT, my coworkers had used this very same process to iteratively create dye molecules [1]. *Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back.* Their study didn't lead to the discovery of a revolutionary new material, but it did demonstrate that even as of 2023 that this self-improving cyclical procedure could one day be useful.

In fact, all of those individual components were actively studied in the early 2020s at MIT. Methods for selecting synthesis routes based on tailorable objectives (synthesizability, safety, cost) were developed, such as through the ASKCOS project [8]. Numerous models for predicting chemical properties were also developed or spearheaded at MIT, including the message-passing neural network architecture [9]. Similar story, too, with the autonomous chemistry lab [10, 11]. But years of refinement and innovation had finally made autonomous labs real.

Graduate school was many years in the past now. Though a practitioner of computational chemistry, I had been deeply skeptical of the hype of AI for chemistry for years. I had known that issues in data quality and availability severely hampered the ability of models to learn patterns in chemistry, with huge challenges in extrapolating to data-poor regions of chemical space, which unfortunately tended to harbor the most interesting kinds of molecules. Making matters worse, chemical literature contained pervasive issues in data curation often from combining datasets measured under different conditions or standards, leading to inconsistent data and thereby inconsistent predictions [12–14].

It was not at all guaranteed that generative AI would work for chemistry. One big issue is that we simply don't have enough scientific knowledge about the world of "chemical space". Even with high-throughput data generation, we will never come close to the amount of sheer data available for models in natural language processing and computer vision. State-of-the-art transformer models in the early 2020s were pre-trained on hundreds of billions of tokens ($\approx 10^{11}$), whereas the number of available experimental data is typically no more than 10^6 and often is much smaller ($\approx 10^4$), and many datapoints might even be redundant due to high molecular similarity. As another point of comparison, the number of possible drug-like compounds is popularly estimated to lie between 10^{23} to 10^{180}. At the same time, chemical representation is not trivial either. Compounds are often represented as graphs, which for machine learning purposes can be encoded in vectors. But because chemical space is huge, with variations on the elements, formal charge, bond orders, connections, aromaticity, stereochemistry, and varying permutations of such properties, these vectors typically need to be quite big, and hence quite reliant on large quantities of data to be useful.

During graduate school, I had thought that perhaps data mining would be able to extract more information from "locked" compilations of chemical data, perhaps lending just enough additional data for us to make useful chemical models. Although those advances did lead to modest model improvements, they didn't lead to any revolutionary advancements that really impacted people. It would take something more clever.

Now, as I watch this woman vividly describe her successful chemical discovery platform, I can't help but finally feel some optimism at the future of AI in chemistry. An autonomous discovery platform would allow itself to explore chemical space, gather data, and iteratively improve. It addresses the fundamental problems of data-driven methods. Previously, models were trained only on compounds whose experimental data were available, usually because those compounds were already deemed interesting for some purpose, or were easy to synthesize or readily available to buy. This biases and limits the chemical space to chemicals we *already know about*, rather than the remarkable chemicals that we're trying to find. The ability of the robot platform to synthesize compounds, analyze them, and update its predictors combats this issue of poor extrapolation. Also, because the measurements are systematic (i.e., from the same device), they avoid the issues of data mismatch that are so pervasive.

Generalizable to any task, an automated chemical discovery workflow could be applied not just to pharmaceuticals, but also to advancements in nanotechnology, environmental science, renewable energy, sustainable materials, biochemistry, and so much more.

But at the same time, it could open Pandora's box.

2 The Chariot

It isn't even a week after this news briefing that I start seeing the commentary in the news.

I strap on my Conec-Set and check my feed: a scowling senator stares eagle-eyed into the camera. "How long before they use this technology to make a deadly neurotoxin? Or a nuclear bomb?" On another TV channel, a commentator says that we need to invest more into the technology before we fall behind to foreign competitors. Another commentator at his table agrees: apparently we've *already* fallen behind. A CEO of a biotech company is arguing that the autonomous lab is an existential threat to our future and to our jobs. Another CEO says it's the most important invention in recent years. The headset knows I want to log off before I even tell it to, and it shuts off automatically. I move to turn the TV off but it turns off by itself as well. Technology's advanced so much recently. Maybe too much! I think about those electrons coursing through a little

slab of silicon, churning in some fusion-powered server megafarm, representing bits of zeroes and ones, which then assemble into larger-scale computer outputs and finally, information. That this process will continue is the only thing guaranteed; the details are left to us. *What* information, exactly? Information for how to create? Or to destroy? Or just nonsense - raw manifestations of entropy? I think about the recent generative AI wave: the natural language revolution of the early 20s, the realistic fake videos in the mid-late 20s. Many good things came about, but many people were also misled, scammed, and hurt from these as well. Some lost their lives [15]. But technology carried on.

Now, here we are crafting AI tools for chemistry. This isn't funny videos and fake screenshots anymore. Anything that involves chemicals is real, substantive, and *will* affect humans in some way. I realize at this moment that AI-driven chemical technologies, in the wrong hands, could actually make the world *worse*. If a model can find something that *cures the most*, couldn't you invert the problem and predict what *cures the least* (finding the minimum of some reward function rather than the maximum)? The least safe molecule - that's the world's most dangerous chemical. So, then, couldn't someone use this same process to make a bomb? Or a deadly neurotoxin?

People have been working on making such threats for centuries. In fact, though I'm not an expert whatsoever in dangerous weapons, I realize that there are already many, many types of dangerous weapons that people could have already built even without computer-aided assistance, probably far more accessible than we would like to think.

Learnings from autonomous labs could extend this, as they could be used to teach someone how to make chemical weapons from easily-accessible consumer goods, which are harder to track and thwart ahead of time. If someone is so determined to harm others as to go to these lengths, couldn't such models push the needle? Something being just a little easier to do might just make it more likely to happen. It gives me pause.

More than ever, I realize we (*collective* we, meaning the whole world) need to discuss the risks, or at least start thinking of safeguards and international guidelines to prevent dangerous misuse. Realistically, such tools would likely be used by state entities to produce applications for warfare, in which case any regulations are really hard to enforce. Could policy alone solve anything? Thus, we would need to develop new technology to *combat* new chemical weapons discovered by AI, not just policy decisions - and Pandora's box is thereby opened. The nuclear bomb nearly brought the world to extinction during the Cold War. Something just as dangerous could happen once more.

I keep these thoughts in mind - the benefits, the risks, and the unrealized solutions. I'm not

sure what to think until a few years later.

3 The Hermit

The verdenir team has won a Greene Prize for AI Advancements in Health. I happen to be tuned into my Connec-Set so I step into the stream and watch: it's the same team, now a few years older, a bit more tired, but still confident, still proud.

The woman from the first press conference - I learn her name is Dr. Jones - starts to speak.

"On behalf of the team, I'm honored to accept the Greene Prize, and even more honored to see the legacy that our work has left behind. Despite its risks, the generalizability of a truly autonomous molecular discovery platform has realized its promise. Many of the most pressing societal issues are being addressed thanks to advanced materials developed through automation. New photovoltaic cells are in development, as are CO_2 capture materials. Some recent advancements were quieter: advanced nanomaterials, leading to advanced semiconductors; 7G communications; longer-lasting roads; small-but-significant uses for medical tools; and so on. And of course, new drugs are being discovered, and more lives will saved."

"Although I acknowledge the many ways that problems have been solved and lives have been saved by chemical innovations, *we need to also acknowledge the risks.* The truth of every technology we make is that we have no idea how it will affect the world. The internet connected the world, but also transformed it in harmful ways. Cars opened up freedom and accessibility to millions, but also contributed to climate change and caused many thousands of accidental deaths per year. Now, similarly, I could not say beyond a doubt that autonomous labs *won't* be used by threat actors, or *couldn't* be used to make some truly abhorrent chemical-based weapons, or some other thing that we can't even imagine. And we can't rely on financial incentives alone to prevent these threats; sure, there is push to produce things that consumers want, like renewable energy, sustainable materials, life-improving drugs. But there's massive government funding in weaponry as well."

"On the whole, I do believe that people in general *want* to use technology for good. Governments also don't *want* people to use dangerous technologies, and will try to prevent technological misuse too. Although we live in a chaotic and unknowable world, I believe that people will want to fight against misuse and band together to face the risks."

"As we are in a society that is driven by technological development, and as such technologies now exist and will continue to be developed, it is our duty to *choose* to be the ones that use these

tools for good."

"Finally, from a pragmatic sense, it's also much harder to use these tools for harm, because they by nature are destructive in their test cycle - explosives will detonate during test cycles. Biochemical weapons can't be tested as easily as curatives. It is far easier to build safeguards *against* those weapons than to build those weapons explicitly."

Dr. Jones pauses, then continues.

"Before this speech, many of you might never have thought before about what AI looks like in chemistry. Maybe you've just thought about text or image or video generation or self-driving cars. My hope to *you* is that when you step away today, you'll start thinking about the staggering untapped potential of chemistry. What else could we develop? Where else can AI go? Could we broadly cure neurodegenerative diseases and eliminate cancer, reverse aging, or perhaps prevent death itself? What should we work towards as a civilization, with all of our progress and technology? I don't know the answers, but I am hopeful that our best shot is through combining our collective human knowledge via computation."

"We must work together as humankind and devote ourselves to the betterment of life. Together, let us make chemicals for peace."

References

(1) Koscher, B. A.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B. et al. *Science* **2023**, *382*, eadi1407.

(2) Scott, L. J. *Drugs* **2019**, *79*, 201–206.

(3) Dhillon, S. *Drugs* **2024**, *84*, 239–246.

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

(5) Sertkaya, A.; Beleche, T.; Jessup, A.; Sommers, B. D. *JAMA Network Open* **2024**, *7*, e2415445–e2415445.

(6) Mullard, A. *Nature Reviews Drug Discovery* **2016**, *15*, 447–448.

(7) Harrison, R. K. *Nat Rev Drug Discov* **2016**, *15*, 817–818.

(8) Tu, Z.; Choure, S. J.; Fong, M. H.; Roh, J.; Levin, I.; Yu, K.; Joung, J. F.; Morgan, N.; Li, S.-C.; Sun, X. et al. *arXiv preprint arXiv:2501.01835* **2025**.

(9) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. *Journal of Chemical Information and Modeling* **2023**, *64*, 9–17.

(10) Eyke, N. S.; Schneider, T. N.; Jin, B.; Hart, T.; Monfette, S.; Hawkins, J. M.; Morse, P. D.; Howard, R. M.; Pfisterer, D. M.; Nandiwale, K. Y. et al. *Chemical Science* **2023**, *14*, 8798–8809.

(11) Canty, R. B.; Koscher, B. A.; McDonald, M. A.; Jensen, K. F. *Digital Discovery* **2023**, *2*, 1259–1268.

(12) Llompart, P.; Minoletti, C.; Baybekov, S.; Horvath, D.; Marcou, G.; Varnek, A. *Scientific Data* **2024**, *11*, 303.

(13) Landrum, G. A.; Riniker, S. *Journal of Chemical Information and Modeling* **2024**, *64*, 1560–1567.

(14) Zheng, J. W.; Leito, I.; Green, W. H. *Journal of Chemical Information and Modeling* **2024**, *64*, 8838–8847.

(15) Payne, K. *AP News* **2024**.