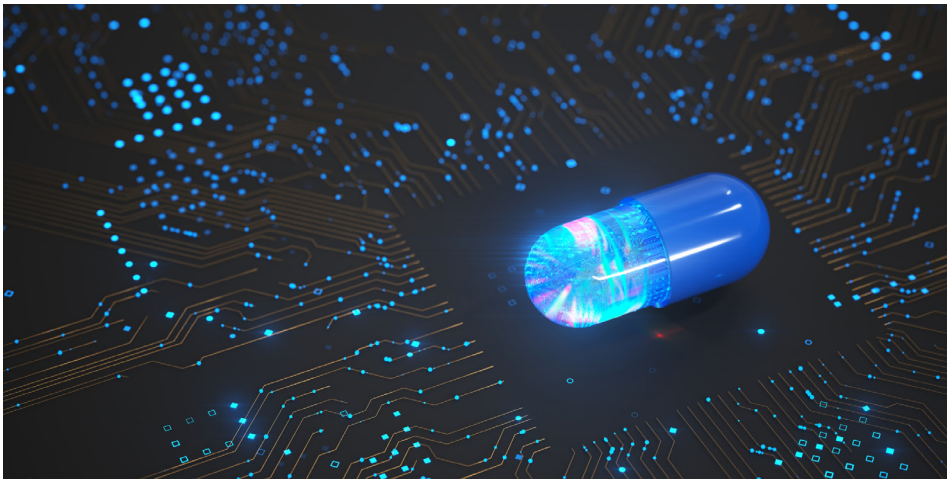
BUILDING A NEXT-GENERATION TOOLBOX FOR AI-POWERED DRUG DISCOVERY

In late 2023, numerous media outlets touted landmark progress in bringing the first truly artificial intelligence–designed drug into clinical trials: researchers at Hong Kong–based Insilico Medicine had used their Pharma.AI workflow to dream up entirely new molecules and had developed a promising candidate for treating lung fibrosis.[1,2] The molecule, INS018_055, succeeded in a Phase 1 safety trial and is now undergoing efficacy testing in a Phase 2 study. If things go well, this drug would represent an important milestone in AI-facilitated drug discovery.

But at times, tremendous fanfare in this field has resulted in questionable claims about "first AI-generated" drugs.[3,4] Some promising early contenders have already fallen by the wayside, including an eczema program shut down by BenevolentAI last spring after a disappointing trial.[5] And some experts caution against overinflating the role of algorithms versus that of the scientists using them. "I think there have been AI contributions," says Pat Walters, chief data officer at Relay Therapeutics. "But I don't buy the claim that there's just a purely AI-invented drug."



AI-generated drugs are still over the horizon, but machine learning and deep learning are already creating opportunities to accelerate and improve the discovery process.

Credit: Just_Super/Getty Images

There's no question that AI is now embedded in the drug discovery process at a growing number of companies, however. "For 70% of our small-molecule drug discovery projects, we apply AI methods," says Ola Engkvist, head of molecular AI at AstraZeneca. As more drug assets are successfully developed with AI methods, this trend is poised to become the industry standard.

Many AI-driven start-ups that initially focused on software have matured into full-service drug companies, with both their own internal discovery programs and partnerships with large pharmaceutical firms, says Ashwini Ghogare, head of AI and automation in drug discovery at MilliporeSigma. "Now we see billions of dollars that are being invested for these strategic partnerships." Use cases for AI in drug discovery are still evolving, however, and AI-powered drug ideation must overcome many hurdles before it becomes commonplace and reliable. Here's the latest on how AI supports drug discovery.

STREAMLINING SCREENING

Most AI methods are based on machine learning (ML), in which an algorithm is fed a selected cohort of training data to create models for the classification and interpretation of novel inputs. Conventional ML approaches such as support vector machines and random forest classifiers have been used for decades, but interest in a subset of ML methods known as deep learning (DL) has surged in the past decade. DL relies on sophisticated neural network computational architectures to extract richer insights from datasets. It powers many mainstream AI tools, including ChatGPT and the image-generating algorithm Midjourney. It's not always the best choice for chemists, however. "If datasets are very small, classical ML methods are still very valuable and extremely hard to beat," says Olexandr Isayev, a computational chemist at Carnegie Mellon University. But he adds that DL methods are highly scalable and demonstrate advantages over ML when they are unleashed on training datasets with many thousands or millions of data points.

Such data-crunching capabilities could be especially powerful at the earliest stages of drug discovery, when the goal is to pluck candidates with promising pharmacological properties out of vast libraries of chemical compounds. This can be done experimentally, with automated high-throughput screening of many compounds against a given protein target in cultured cells or other assays. Computational virtual screening methods can also narrow the field before wet-lab work begins, but it has become burdensome as chemical libraries grow increasingly larger. Consider the Enamine REAL collection—a database of more than 6 billion compounds, any of which can be ordered on demand.[6] Brute-force screening at this scale is impractical, and some virtual libraries are even larger. "MilliporeSigma has one of the largest chemical spaces, which is 10^{20} compounds," Ghogare says. "In order to search that chemical space, it used to take us a minimum of a month's time with the compute that we had at hand." Computational and generative methods can help sift these daunting chemical collections to a manageable scale.

Further advances in AI-assisted drug development will require more high-quality data and robust benchmarking methods to assess algorithmic performance.

Credit: KTSDESIGN/SciencePhotoLibrary

One approach to virtual screening entails ligand-based screening, which assesses the suitability of chemical library candidates according to their similarity or difference to other molecules in terms of physicochemical properties. Pedro Ballester, a data scientist at Imperial College London, says ML-based methods for assessing quantitative structure-activity relationships (QSAR) have been around for more than 30 years. In that time, the field has got a major boost from the rapid growth of public chemistry data resources like ChEMBL and PubChem, which have enabled the use of more powerful, DL-based approaches. "[Deep learning] is extremely simple to use but also extremely simple to misuse," Ballester says. Researchers must be careful in how they train their models with such heterogeneous data repositories, he adds.

A complementary structure-based virtual screening approach focuses on predicting how well different library molecules can physically dock with a target protein of interest. The method has already benefited directly from the rise of DL, which has enabled the development of structural biology tools like AlphaFold2. This software has predicted structures for millions of proteins that have yet to be assessed via experimental techniques like X-ray crystallography, with confidence metrics that depict how robust those predictions are. Such predictions can complement experimental structural data and help researchers find and map promising binding sites for docking. "I think there is a lot of utility," Walters says, "but it's not like it's going to put structural biologists out of business."

The effectiveness of docking experiments depends heavily on the choice of scoring function—a mathematical procedure for computing whether a compound could be an effective drug candidate based on its 3D interactions with the binding pocket. In principle, ML-derived scoring functions can handily outperform classical scoring functions, such as those based on force fields, in predicting the binding affinity of protein-ligand complexes based on their 3D structures. But Ballester points out that it is critically important to select an appropriate scoring function for a given target,[7] in the context of virtual screening, in which the vast majority of of the billions of library compounds will be nontarget binders. In the absence of sufficient training data for a tailored scoring function, generic scoring functions can deliver solid results. Ballester also sees exciting opportunities for transfer learning approaches, which can extrapolate likely target-binding behaviors according to insights gleaned from experiments with similar drug or target candidates.

THE NEXT GENERATION OF AI

Other emerging algorithmic approaches could further accelerate drug development. For example, many groups are now exploring active learning, which couples the analytical throughput of ML or DL methods with the scientific rigor of physics-based modeling methods like free energy perturbation calculations, a well-established molecular dynamics technique for assessing ligand-target binding. “Free energy perturbation is a relatively slow technique—it takes between 4 and 8 hours to calculate one molecule,” Walters says. This method would scale poorly to libraries with thousands of molecules. Teams can dramatically reduce computational requirements by performing iterative cycles of FEP and ML, however. In a 2022 paper by scientists from Relay Therapeutics and Google, the authors identified more than half of the 100 highest-scoring molecules by evaluating only 3% of a 10,000-molecule library.[8] “There’s a lot of interesting work going on now to integrate physics-based methods and machine learning, and I think the power really comes in that combination,” Walters says.

In parallel, generative AI is giving chemists more shots on goal than would otherwise be possible even with today’s vast chemical libraries. These models use DL to design new molecules according to patterns identified in real-world compounds and on other external parameters, like chemical composition or solubility, supplied by the algorithm’s user. The AstraZeneca team works extensively with generative AI and has developed an extensive toolbox for this purpose.[9,10] Engkvist typically applies it slightly later in the process, after initial hits have been identified via experimental or virtual screening. “We iteratively optimize our hits and increase the chance of evolving these compounds into clinical trial candidates,” he says. The group combines generative AI methods with docking as a scoring function, and then in postprocessing, uses more expensive methods, like free energy perturbation methods, for a detailed evaluation of the interaction binding strength, he adds.

This technology is far from mature, however. Isayev at Carnegie Mellon warns that expert oversight is essential and that many algorithms still spit out molecules that are unrealistic or impractical to produce. “If I would take some of those molecules to my experimental colleagues, they would stop talking to me because of how atrocious they are,” he says. It helps to impose rules that limit the algorithm’s “imagination.”

MilliporeSigma has developed a new, AI-powered platform for drug discovery. The software, called AIDDISON™, integrates generative de novo design, molecular docking, and synthetic accessibility scoring. Ghogare says this manufacturability in drug design is steered by retrosynthesis rules codified in the company’s SYNTHIA™ retrosynthetic-planning software. “SYNTHIA™ encompasses more than 100,000 retrosynthesis rules, which took over 15 years to build,” she says. These rules, which are manually coded, can be used inside ML models that strongly favor the generation of bona fide drug candidates that can be synthesized and also meet criteria such as affordability or compatibility with green synthetic methods.

AIDDISON™ takes a holistic drug design approach to accelerate hit-to-lead by incorporating predictive ADMET (absorption, distribution, metabolism, excretion, and toxicity) data at early stages of the design process. Traditionally, this has been a challenge due to the limited availability of large, well-annotated preclinical and clinical datasets with which to train robust ML models. MilliporeSigma has an advantage with access to 30 years of experimentally validated preclinical data from its parent company, Merck KGaA, Darmstadt, Germany. The company also benefits from a partnership with Excelra, the creator of GOSTAR, a repository of curated SAR and ADMET data for nearly 10 million compounds. Ghogare believes that models derived from these experimentally validated data could save researchers a great deal of trouble by increasing the likelihood that hits predicted to show excellent target binding and potency will prove suitable for use in patients. That raises the probability of success for each shot on goal, she says.

This ‘glowing twin’ model produced by AIDDISON™ shows the atomic-level contribution to the ADMET property of CACO-2 permeability. The atoms highlighted in green show positive contribution while the atoms in red show negative contribution and make CACO-2 permeability weaker.

Credit: MilliporeSigma

BUILDING A BETTER ECOSYSTEM

As AI becomes more fundamental to the drug discovery process, however, the limitations of the current toolbox are also coming into stark relief.

Many experts cite the need for better benchmarking of performance and reliability for this generation of ML and DL methods, particularly in emerging domains like generative chemistry. "Our benchmarks cannot anticipate how well a method is going to work in a particular target, and there's a lot to be done," Ballester says. Without these capabilities, scientists cannot objectively weigh the strengths and weaknesses of different algorithms or workflows. But promising steps have been taken. Those include the Critical Assessment of Computational Hit-finding Experiments (CACHE) challenges, in which teams test their pipelines in real-world demonstrations that span the process from hit identification to lead optimization. "I think this is going to be the best kind of validation, not just a synthetic benchmark," says Isayev, whose group recently claimed first place in the intial CACHE challenge.

More and higher-quality data will also be essential, particularly for DL algorithms. Some of the richest data collections are siloed in the servers of individual companies. This is understandable, given that these proprietary data are often more important than the algorithms themselves in determining success or failure of an AI-driven experiment. But it leads to replication of effort and slower progress in the field. This issue remains difficult to resolve. In a recently concluded European Union initiative called Machine Learning Ledger Orchestration for Drug Discovery (MELLODDY), 10 pharmaceutical companies fed internal data related to drug discovery into a federated learning framework that let participants collectively train an ML model while protecting proprietary data. The MELLODDY consortium was able to apply this approach successfully, but not all participants benefited equally: larger companies generally saw small gains relative to those with limited internal data.[11] Engkvist, who coordinated AstraZeneca's involvement in MELLODDY, is positive about the experience. "I think we can see a way forward with privacy-preserving machine learning in the future," he says. But Walters is skeptical and believes that these efforts will remain hamstrung by the conflicting goals of openness and intellectual property protection.

There are also major limitations to publicly available repositories, and much of the data used to train algorithms about SAR and other chemical properties may be inconsistently or incorrectly labeled or obtained from highly heterogeneous sources. This latter aspect could even be a matter of subtle differences in experimental processes between labs. "There is substantial error there, and when you put that into the model without knowing it, you can fool yourself into believing that your model generalizes," Ballester says.

Advances in automated experimentation could make a huge difference by enabling more consistent and reproducible experimental design while collecting the resulting data in appropriately structured, ML-friendly formats. By incorporating these systems into active learning processes, we could ultimately

envision closed-loop robotic laboratories where AI algorithms use incoming data to design and execute future experiments. “That’s my dream, and probably I can retire if the system runs,” says Isayev, whose team developed and operates the fully automated Carnegie Mellon University Cloud Lab. But he acknowledges that the dream will require considerable evolution of software, hardware, and human expertise to become a reality. “There are a lot of things that need to be ironed out before we can replace chemists,” he jokes.

REFERENCES

1. Hayden Field, “The First Fully A.I.-Generated Drug Enters Clinical Trials in Human Patients,” CNBC website, June 29, 2023, https://www.cnbc.com/2023/06/29/ai-generated-drug-begins-clinical-trials-in-human-patients.html.
2. Jamie Smyth, “Biotech Begins Human Trials of Drug Designed by Artificial Intelligence,” *Financial Times* website, June 26, 2023, https://www.ft.com/content/82071cf2-f0da-432b-b815-606d602871fc.
3. Derek Lowe, “Another AI-Generated Drug?,” *In the Pipeline* (blog), Science, Jan. 31, 2020, https://www.science.org/content/blog-post/another-ai-generated-drug.
4. Derek Lowe, “AI-Generated Clinical Candidates, So Far,” *In the Pipeline* (blog), Science, Nov. 8, 2021, https://www.science.org/content/blog-post/ai-generated-clinical-candidates-so-far.
5. Nick Paul Taylor, “BenevolentAI, Cruel R&D: AI-Enabled Drug Flunks Midphase Eczema Trial to Dent Deal Plans,” Fierce Biotech, April 5, 2023, https://www.fiercebiotech.com/biotech/benevolentai-cruel-rd-ai-enabled-drug-flunks-midphase-eczema-trial-dent-deal-plans.
6. REAL Database, Enamine, accessed February 2024, https://enamine.net/compound-collections/real-compounds/real-database.
7. Viet-Khoa Tran-Nguyen et al., “A Practical Guide to Machine-Learning Scoring for Structure-Based Virtual Screening,” *Nat. Protoc.* 18 (Nov. 2023): 3460–511, https://doi.org/10.1038/s41596-023-00885-w.
8. James Thompson et al., “Optimizing Active Learning for Free Energy Calculations,” *Artif. Intell. Life Sci.* 2 (Dec. 2022): 100050, https://doi.org/10.1016/j.ailsci.2022.100050.
9. Thomas Blaschke et al., “REINVENT 2.0: An AI tool for De Novo Drug Design,” *J. Chem. Inf. Model*. 60, no. 12 (Dec. 28, 2020): 5918–22, https://doi.org/10.1021/acs.jcim.0c00915.
10. Jeff Guo et al., “Link-INVENT: Generative Linker Design with Reinforcement Learning,” *Digital Discovery 2* (April 1, 2023): 392–408, https://doi.org/10.1039/D2DD00115B.
11. Wouter Heyndrickx et al., “MELLODDY: Cross-Pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR Without Compromising Proprietary Information,” *J. Chem. Inf. Model*., published online Aug. 29 2023, https://doi.org/10.1021/acs.jcim.3c00799.

AMERICAN CHEMICAL SOCIETY
1155 Sixteenth Street, NW,
Washington, DC 20036