THE WHITE HOUSE
WASHINGTON

AI ASPIRATIONS

An AI Aspiration for Health

Wade Shen

ARPA-H

IMAGINE IF…

… new medicines for seemingly intractable diseases could be developed and approved in months rather than decades.

Today

The vast majority of over 10 thousand known diseases have no effective medication today. With fewer than 40 new drugs becoming available each year, many millions of people and their families are suffering and waiting for decades. The process of drug development is long and costly, taking 10-15 years to complete, and costing $1-2 billon per drug. Most drug candidates fail, but it can take years to screen and test before reaching that failure point. Even for promising candidates, it takes 8-10 years to test.

AI opens the door

In the last few years, development of large-scale AI models has made previously hard problems more solvable. These advances have been driven by vast amounts of data and compute capabilities in conjunction with advanced modeling techniques to expand existing models to massive scales. Large-scale AI models are being employed for chemical and biological purposes, such as protein fold structure prediction, chemical language models to predict function, and structure and function prediction for small molecules.

To harness AI to change drug development will require even more data—and in particular clinical data about how successful and unsuccessful medicines have interacted with patients in trials. As individual companies have invested billions of dollars over many years in proprietary drug development, they have each amassed separate and closely held collections of this kind of data. The result is that the development of AI systems for drug development is currently blocked by lack of access to the quantities of data needed to reach effective scale.

If the data access issue can be overcome, AI holds tremendous promise for drug development. AI models for pharmacokinetics and toxicity could speed the process of screening new drug candidates to enter clinical trials. The trials themselves could be designed in far more optimized ways to achieve statistically meaningful results more quickly and with fewer subjects and their results may become partially predictable with AI. AI models could help quickly repurpose existing drugs for new conditions. And ultimately, AI models could generate powerful new candidates that are able to demonstrate safety and efficacy in highly robust but far faster regulatory processes.

The work ahead

To achieve rapid and effective drug development, a number of key R&D and implementation problems have to be addressed.

June 2024

Secure and private data access for AI models: If AI models can be trained without sharing proprietary drug development data, that would overcome a key limitation. New privacy-preserving technologies such as federated and confidential machine learning could make this possible, but significant engineering is required to make this viable for AI drug development. These methods are computationally expensive and they require bespoke engineering for particular applications.

Building in safety: The scale of data that could become available for R&D also comes with safety risks. AI-enabled biological design tools will carry increased safety risk with every advance in capability. Technological and administrative means will need to be established to contain risks of bio-error and bio-terror.

Development and performance characterization of AI models for pharmacokinetics and toxicity, trial design, repurposing, and generative drug design: To date, no one has built large-scale models with aggregate data, so the potential of these models is unknown. To build these models and understand their performance characteristics requires extensive data engineering to deal with disparate data types and formats used across industry, academic research, and government. Significant work is required to generate the empirical evidence that will be needed to validate the use of these models in place of existing biochemical and animal gold standards.

Value assignment for generative AI drug models: Generative AI models can produce novel compounds and protein candidates. At least in part, these candidates are derived from training data. A core problem in AI research is the assignment of value to a specific data element that was used during training for any given candidate an AI model might generate. Solving this problem for drug discovery would enable proportionate and fair licensing models for candidates created by AI models that ultimately succeed. If these methods are successful, standards organizations and FDA could encourage implementation of these methods in a marketplace built on the proposed infrastructure.

A first step: We will establish a rapid AI innovation platform to aggregate experimental data about candidate therapies from across public, private, and non-profit drug developers, making them available with strong intellectual property protections, secure access provisions, and legal agreements that enable training of AI models while maintaining IP security and IP ownership. If successful, these models may have significant capabilities to predict toxicology and ADME characteristics for new compounds, optimize and automate trial designs for new compounds, predict repurposing uses for existing compounds, and, ultimately, generate new candidate compounds given disease targets and populations of interest.

This platform will enable safe, auditable, and large-scale experimentation for predictive toxicology and safety modeling via the largest set of drug discovery data ever assembled. In conjunction with a concerted set of research activities by both government and industry, this platform would be used to validate the most advanced AI absorption, distribution, metabolism, and excretion (ADME) and toxicology models and help to set the standards for their use.

Ultimate capability: If these challenges can be overcome with focused effort, developers will be able to use an array of predictive AI models that have been certified and validated by FDA and partner standards agencies to:

- *Predictively assess candidate compounds for efficacy, ADME characteristics, and safety at near-zero cost and with potentially higher predictive power than in-vitro or animal models.* Candidates with favorable assessments by certified models could be fast tracked by the FDA for human trial with limited or no animal study needed.
- *Design optimal trials for clinical validation.* Candidate drugs characterized by AI models in terms of their safety can also receive automated trial designs for follow-on animal and/or human trials that minimize trial size, cost, and time-to-execute. Trials proposed by certified models could also receive fast-track approval from FDA.

- *Find alternative indications and target populations for existing therapies.* Existing drugs and compounds could be evaluated using AI repurposing models for new indications and new treatment populations. These models would predict the efficacy and safety of existing compounds either in isolation or as part of combined regimes. New uses with high predicted potential as indicated by certified models could be fast tracked during the approval process and could receive minimized study designs by the FDA.
- *Generate de novo therapeutic compounds and proteins given forms and/or functions of interest.* In addition to the generation of these candidates, AI models could provide corresponding ADME and safety characterizations along with fast-tracked trial design and AI-predicted efficacy for clinical validation, to dramatically shorten the entire development and approval lifecycle.

Major hurdles and societal risks

This Aspiration is critically dependent on access to vast amounts of high-quality data. Data exists today from failed and/or abandoned development of therapies in the pharmaceutical industry. This data could be used to drive AI development, but they are typically protected by IP. Data also exists in the form of clinical evidence and patient treatment outcomes, all of which live in the medical system, scattered across numerous provider networks, IT systems, and insurers. Making these data available to AI modeling is a major challenge in the U.S. healthcare context. Finally, all this data may not be enough, and more data may be needed both the characterize molecules and to understand their real-world effects.

A second major hurdle will be to create the significant evidence needed to validate that AI models trained with these data resources can generate results that consistently predict real-world safety and efficacy. This bar is high and will have to be surpassed before we can revolutionize existing development and regulatory processes.

This undertaking comes with some societal risks. Unless implemented carefully and with purpose, this drug-development advance can exacerbate inequities and bias in two ways. One is in the data used to train models. Clinical data and real-world evidence that are used to build these models are often derived (in part) from clinical trials and datasets that are unrepresentative of underserved populations. These data must be curated and managed carefully to make them useful and appropriate for AI training. The other is in the diseases and populations that are addressed with an accelerated drug development process. While the promise of AI to reduce the cost of development could incentivize the develop of therapies for rarer diseases, data on these populations is more limited and AI modeling abilities will likely be correspondingly worse. A continued investment in rare disease data is needed to ensure that the advances of AI are equitable distributed to diseases affecting smaller populations. Finally, it will be important to make sure that all patients are able to access new treatments affordably.

This advance in biological design tools will also reduce the knowledge, expertise, and cost required to create dangerous molecules and pathogens and could lead to dangerous but unintended consequences. Safeguards, standards, and careful oversight will be required to manage this risk.

A transformative national capability

Employing AI in all phases of the drug development process can ultimately reduce the time it takes to bring new medicines to market from decades to months. AI models can unlock significant capabilities to predict toxicology and pharmacokinetics characteristics for new compounds, optimize and automate trial designs for new compounds, predict repurposing uses for existing compounds, and, ultimately, generate new candidate compounds given disease targets and populations of interest. Achieving this AI aspiration will transform the lives of millions of people whose diseases have no cure today.