nature reviews drug discovery

nature > nature reviews drug discovery > from the analyst's couch > article

FROM THE ANALYST'S COUCH | 29 August 2023

The company landscape for artificial intelligence in large-molecule drug discovery

Navraj S. Nagra ✉, Lieven van der Veken, Erika Stanzl, David Champagne, Alex Devereson & Matej Macak

Artificial intelligence (AI) techniques such as machine learning are transforming drug research and development (R&D), enabled by ever-increasing amounts of data and computational power. Historically, small molecules have been at the forefront of AI applications in drug discovery, including modelling small-molecule–target interactions, lead candidate optimization and safety prediction. However, AI tools are increasingly being applied to large-molecule modalities, including antibodies, gene therapies and RNA-based therapies. Such therapies represent an important share of the biopharma industry's current portfolio — around 40% of new molecules approved in 2022 — and of its future commercial potential. For example, in oncology, large molecules are forecast to represent ~50% of the market by revenue in 2030, of which more than 80% is expected to be derived from antibodies.

In this article, we outline how AI-based approaches are being applied in large-molecule drug discovery, analyse the landscape of companies developing these approaches and

their pipelines, and provide a perspective on what is required for the biopharma industry to implement these approaches successfully.

AI in large-molecule drug discovery

Machine learning tools are being applied in many areas of drug R&D. Here, we focus on three overlapping aspects of large-molecule drug discovery — structural predictions, functional predictions and new candidate generation — where approaches are maturing quickly. Applications of machine learning in drug development, such as tools to predict responsive patient populations, or to derisk or accelerate trials, are not covered.

Tools to predict large-molecule structure. Prediction of protein structures is valuable for large-molecule drug discovery in areas ranging from target identification (such as predicting antigen structure) through to lead identification and optimization. The success of AlphaFold2, a machine learning model, in predicting three-dimensional protein structure from amino acid sequences alone was a landmark advance in 2020. Many companies in the field are now using AlphaFold2 or other protein structure prediction models with similar accuracy, such as RoseTTAFold. Ongoing developments are improving aspects such as ease-of-use, scalability, performance on orphan proteins and re-trainability (for example, ColabFold, FastFold, OmegaFold and OpenFold), and increasing generalization and speed using different architectures that are similar to large language models such as GPT-4 (for example, ESMFold).

Tools to predict large-molecule functions. AI tools have been developed to support the prediction of the functions of large-molecule therapeutic candidates, including antigen–antibody or RNA–protein binding, as well as aspects relevant to their developability, such as pharmacokinetic clearance. These predictions can be made using machine learning models such as gradient-boosted trees or computational models such as molecular dynamics simulations. More recently, deep learning methods have been used, including graph-based models, convolutional neural networks, recurrent neural networks or 'large-molecule language models', to predict key therapeutic properties such as antibody affinity. Various representations of large

molecules can be used by these methods; for example, three-dimensional coordinates of antibody–antigen amino acids, or sequences of amino acids or nucleotides for convolutional neural network or large-molecule language model architectures.

Generating large-molecule therapeutic candidates. Rapidly growing data availability is supporting the development of algorithms that can generate proteins, antibodies or mRNAs at scale as part of lead generation or optimization; for example, based on diffusion, variational autoencoder models or through employing large language models similar to GPT-4 trained on data specific for the modality, such as protein sequences.

Specific examples where such algorithms have been implemented include development of novel antigen structures, determining optimal mRNA structure for stability and immunogenicity, and novel protein and antibody design. Designed molecules are often subsequently assessed in high-throughput systems to experimentally confirm functional properties and further reinforce and improve candidate generation.

The tools described above are versatile; for example, RFdiffusion or the ESM family of models have been used for novel protein generation as well as structural and functional predictions. These tools are now complementing or replacing traditional computational approaches. Examples of the application of AI tools in the antibody discovery pipeline are highlighted in Supplementary Fig. 1.

Emerging company landscape

We analysed the landscape for AI-driven biotech companies engaged in large-molecule drug design and identified 82 companies active in this field (Fig. 1a; Supplementary Table 1). More than 60% of these companies were founded in the past 5 years, indicating a nascent industry driven by recent technological step changes, such as the advent of AlphaFold. There is also some emerging evidence of consolidation amongst these companies, such as iBio's acquisition of RubrYc Therapeutics in 2022.

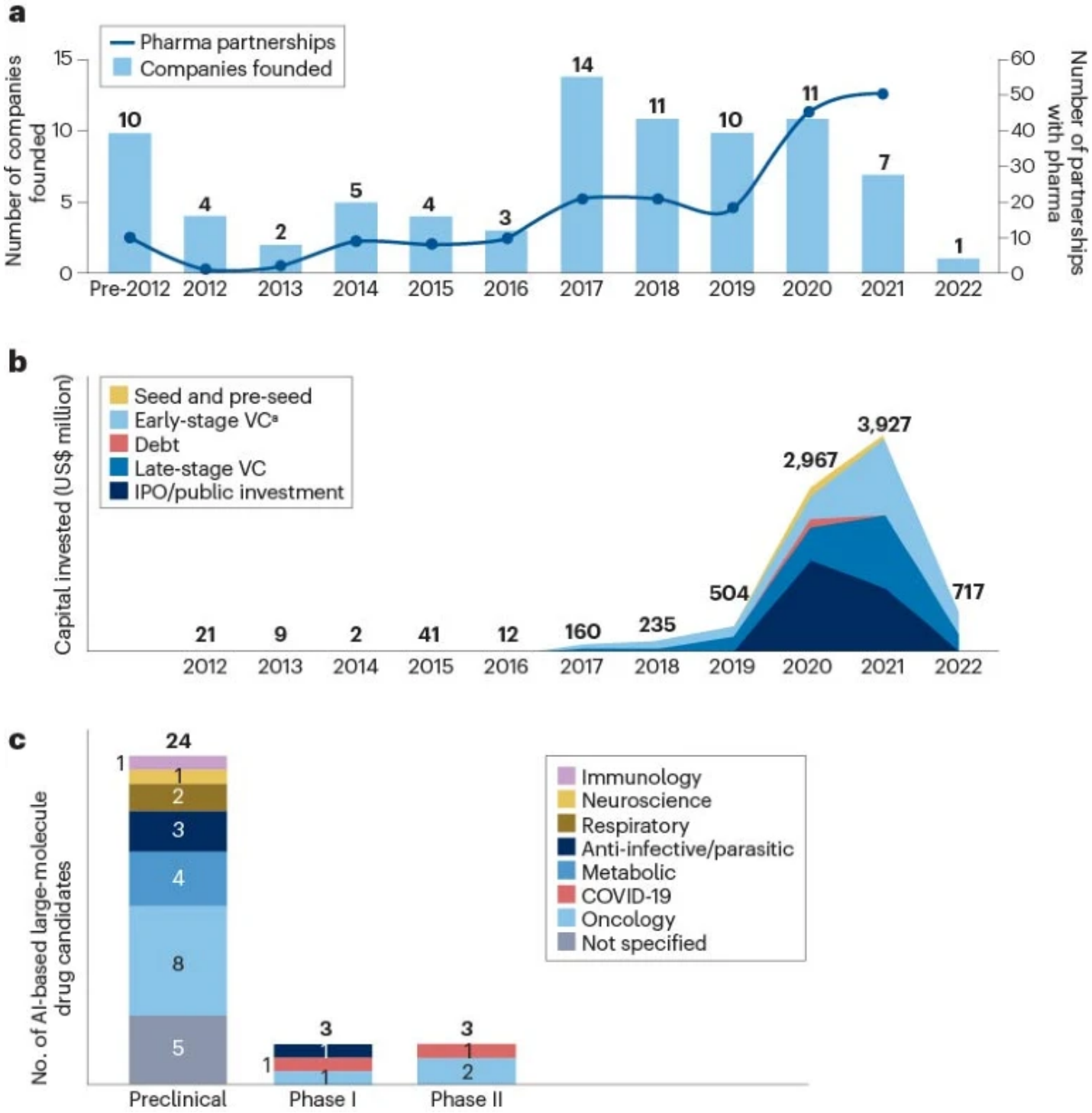

Fig. 1 | Trends in the landscape of biotech companies applying AI in large-molecule drug discovery. a, Number of artificial intelligence (AI)-driven biotech companies focusing on large-molecule drug discovery founded per year, and the number of partnerships formed by these companies with top-20 biopharma companies per year. **b**, Total capital invested in these companies, by investment type. **c**, Number of pipeline assets for these companies by development stage and indication. IPO, initial public offering; VC, venture capital. For details, see Supplementary information. [a]Includes series A and series B funding rounds.

The companies analysed in this space raised US$3.9 billion in 2021, with $2.7 billion of this being raised by venture capital firms alone (Fig. 1b). Total investment declined significantly in 2022, however, to $0.7 billion (Fig. 1a). Notable activity includes AbCellera and Absci (both focused on antibody discovery) raising $555 million and $200 million, respectively, in initial public offerings in 2020 and 2021, and Generate Biomedicines (focused on protein drug discovery) raising $370 million in series B funding in 2021.

Established biopharma companies are investing in building AI capabilities for large-molecule drug discovery both internally and through acquisitions, such as Genentech's 2021 acquisition of Prescient Design, an AI-driven biotech applying machine learning to antibody discovery. Established large biopharma companies have also partnered with AI-driven biotech companies, with 51 partnerships identified in 2021, up from 10 partnerships in 2016 (Fig. 1a). Examples of deals include partnerships between BigHat Biosciences (focused on antibody discovery) with Amgen, AbCellera with AbbVie, and MAbSilico (focused on antibody discovery) with OSE Immunotherapeutics.

The pipelines of AI-driven biotechs are currently at an early stage (Fig. 1c). Three assets were identified in phase II: Evaxion is developing a peptide-based personalized cancer immunotherapy for metastatic melanoma, ZielBio is developing a monoclonal antibody against plectin for solid tumours and PharmCADD has an mRNA vaccine candidate for SARS-CoV-2.

There are also three assets in phase I: Peptilogics' peptide antibiotic for periprosthetic joint infection, SparX Therapeutics' monoclonal antibody targeting claudin 18.2 for gastric carcinoma and another mRNA vaccine against SARS-CoV-2 from PharmCADD.

Companies developing these molecules have reported leveraging AI-based target identification, functional (binding) prediction and antibody generation (including the use of generative AI) as part of candidate development. Details of the molecules in clinical development are available in Supplementary Table 2.

In the preclinical pipeline, the largest group of assets under development by AI-driven biotech companies is in the oncology area, with eight molecules. With regard to modality, there are more RNA therapeutics and peptides in preclinical development (~50% of all molecules) than antibody therapeutics. This is potentially owing to the higher complexity of antibody design and lack of functional data with which to train machine learning tools.

Outlook

Our analysis indicates that the application of AI in large-molecule drug discovery is increasing rapidly. However, while the potential value of these tools has been convincingly demonstrated in academic settings, deployment at scale has so far proved challenging.

Several elements need to be addressed to realize the potential of AI in the field. First, AI models must be fully integrated into research processes, with appropriate capability building of research scientists. By doing so, companies can rapidly train and validate machine learning algorithms, whilst also overcoming potential 'siloing' of AI efforts. For example, when using a large language model for high-throughput prediction of antibody affinity, timely validation in vitro through integrated research systems will further train and enhance the performance of in silico models. Second, technical environments must be established, such as complex data engineering pipelines (integrating and capable of automating labelling of public and internal data), suitable computational infrastructure, and integration of source system modelling environments. This enables companies to train and refine AI models at a rate that can inform and improve the next experiment. Finally, AI technologies need to be combined across the R&D process beyond drug discovery, into areas such as trial design and identification of patient subpopulations to further improve trial efficiency and probability of success.

doi: https://doi.org/10.1038/d41573-023-00139-0

Acknowledgements

The authors wish to thank Jeffrey Algazy, Joachim Bleys, Sam White, Ester Friedlaenderova, Thomas Devenyns, Rachel Moss, Michael Steinmann and Chris Anagnostopoulos for their contributions to this article.

SUPPLEMENTARY INFORMATION

1. Supplementary information

COMPETING INTERESTS

The authors of this article are employees of McKinsey & Company, a management consultancy that works with the world's leading biopharmaceutical and biotechnology companies. The research for this specific article was funded by McKinsey's Life Sciences practice.

Nature Reviews Drug Discovery (*Nat Rev Drug Discov*) | ISSN 1474-1784 (online) | ISSN 1474-1776 (print)