wellcome

BCG

Unlocking the potential of AI in Drug Discovery

Current status, barriers and future opportunities

Disclaimer

This document has been prepared in good faith on the basis of information available at the date of publication without any independent verification. No party guarantees or makes any representation or warranty as to the accuracy, reliability, completeness, or currency of the information in this document nor its usefulness in achieving any purpose. Readers are responsible for assessing the relevance and accuracy of the content of this document. It is unreasonable for any party to rely on this document for any purpose and no party will be liable for any loss, damage, cost, or expense incurred or arising by reason of any person using or relying on information in this document. To the fullest extent permitted by law (and except to the extent otherwise agreed in a signed writing by a party), no party shall have any liability whatsoever to any other party, and any person using this document hereby waives any rights and claims it may have at any time with regard to the document. Receipt and review of this document shall be deemed agreement with and consideration for the foregoing.

All parties are responsible for obtaining independent advice concerning legal, accounting or tax matters. This advice may affect the guidance in the document. Furthermore, no party has made any undertaking to update the document after the date hereof, notwithstanding that such information may become outdated or inaccurate. Any financial evaluations, projected market and financial information and conclusions contained in this document are based upon standard valuation methodologies, are not definitive forecasts, and are not guaranteed by any party. No party has independently verified the data and assumptions from these sources used in these analyses. Changes in the underlying data or operating assumptions will clearly impact the analyses and conclusions. This document is not intended to make or influence any recommendation and should not be construed as such by the reader or any other entity.

The content included herein stems from an engagement to write a commissioned report whereby BCG was compensated by Wellcome.

This document does not purport to represent the views of the companies mentioned in the document. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favouring by any of the parties involved in compiling the document.

Other than the logos or other marks and similar of BCG and Wellcome, the contents of this document may be reproduced, distributed, or circulated provided that the source is acknowledged with the following acknowledgement: “This report and its findings were produced and co-authored by Wellcome and BCG.”

Table of contents

1. Executive Summary

Context and Objectives of this report

The discovery of new medicines is critical to improving human health. As dramatically highlighted by the race to find therapeutics and vaccines for COVID-19, innovation can drive enormous health impact. However, drug discovery is an increasingly expensive, risky, and time-consuming proposition – estimated to cost approximately $2.5 Bn to bring a new drug to market when accounting for the cost of failures [1], [2]. Scientific and technical challenges mean the probability of discovering a new drug and progressing it to clinical trials is in the range of 35% and the probability of successfully taking a candidate from Phase 1 trials to regulatory approval only 9-14% [3], [4]. Overall, the process takes on average 12-15 years [5].

This combination of significant cost, risk, and time acts as a major barrier to innovation, with market forces often channelling R&D efforts into areas with large commercial returns such as Oncology and Immunological diseases. In contrast, therapeutic areas (TAs) not typically supporting high price points, such as many infectious diseases, are often struggling to secure funding.

Over the past decade, the field of artificial intelligence (AI) has progressed enormously, with major advances in machine learning, neural networks, deep learning, generative AI and other areas. The potential to apply AI techniques to accelerate and improve drug discovery has garnered growing interest from the pharmaceutical industry, tech companies, investors, and funders of biomedical research. Public and policy maker awareness is also growing through high-profile successes, such as the AlphaFold2 algorithm successfully predicting the 3D structure of human proteins in 2020 [6].

AI has the potential to materially alter the economics of innovation, allowing new medicines to be discovered for a much wider set of conditions and patient segments, and by a wider membership of the research community than is possible today. If this potential can be realised, the impact on human health and health equity could be profound.

This report takes stock of a rapidly evolving field and aims to (i) identify the key use cases and applications of AI in drug discovery (ii) assess the maturity of these use cases (iii) baseline adoption and determine current barriers and (iv) identify solutions to overcome barriers to unlock the potential of AI in drug discovery. This report is targeted towards the broad drug discovery community and specifically considers how funders could play a role to shape this field.

Key findings

The current status and future potential of AI in drug discovery was analysed through review of published literature, exploration of patent, funding and investment data, expert interviews, and surveying members of the drug discovery community. From this analysis, a mixed picture emerges – we see some areas where AI delivers value today in drug discovery, but also many areas where the promise is yet to be fulfilled.

Five major AI use case families in drug discovery were identified:

- *Understanding disease use cases* deploy AI to identify and validate new targets for drug discovery efforts.
- *Small molecule design and optimisation use cases* deploy AI for identifying hit-like or lead-like small molecule compounds and optimising the identified hits for favourable properties.
- *Vaccines design and optimisation use cases* encompass AI applications specific to the discovery and design of vaccines, with a primary focus on mRNA-based vaccines.
- *Antibody design and optimisation use cases* deploy AI to identify and optimise antibody structures and formats, binding and other properties.
- *Safety and toxicity use cases* deploy AI to evaluate the safety profile of a therapeutic or vaccine of interest.

The field is maturing rapidly, though unevenly:

- Publications and patents related to AI-enabled drug discovery have grown by 34% and 17% respectively year-on-year over the last five years.
- Efforts however are skewed towards a small number of use cases with over 80% of publications in the last five years focused on applying AI to understanding disease, target discovery and small molecule optimisation.
- Private funding is skewed towards the most commercially tractable therapeutic areas with ~70% of AI-related investments in the last five years being made in oncology, neurology, and COVID-19.
- Private sector funding for AI-related drug discovery efforts is almost exclusively flowing to high-income countries (HICs) and China.

Adoption of AI tools varies significantly:

- Despite increasing investment and research activity in developing AI tools for understanding diseases and small molecule optimisation, adoption lags with less than a third of survey respondents across industry segments using AI tools routinely today.

- Overall, industry drug discovery efforts are more likely to be systematically deploying AI approaches today versus academia where adoption varies widely and is typically focused on open-source tools. Even within industry, there is a wide spectrum with adoption led by 'AI-first' biotechs who have built their R&D workflow and value proposition around AI tools, and some pharmaceutical companies who are pioneering AI in drug discovery.
- Adoption is much higher in HICs than in low- and middle-income countries (LMICs) – 42% of HIC survey respondents stated regular use of AI approaches versus 19% of LMIC respondents.
- Despite varying adoption rates, there is broad consensus on the future potential of this technology with 84% of current AI users and 70% of current non-users stating that they expect AI to drive significant impact in drug discovery over the next five years.

Early proof points are emerging, but a key test will come in the clinic:

- Modelling based on extrapolation of publicly available data from early AI programmes suggests AI-driven R&D efforts from discovery up to preclinical could deliver time and cost savings of at least 25-50% (see Appendix section 10.5).
- Drugs developed through AI approaches are now entering the clinic, which will be a critical test of whether these drugs can also deliver on a key postulated benefit – improving on current standards of care and/or having a higher probability of clinical success.
- Whilst time and cost savings are helpful, modelling shows it will be improvements in probability of success in the clinic that delivers the biggest impact from AI in drug discovery and a step change in the economics of R&D.

Barriers must be addressed to unlock the full potential of AI:

- Trust in AI is a major barrier in many of the settings explored in this report. Although sentiment ranges widely, common themes include the perceived lack of value proofs in drug discovery and overall uncertainty about AI in general, and what rapid advances in AI could mean for science and wider society.
- Lack of high-quality data sets, access to mature tools, and relevant AI and drug discovery capabilities constrains the value being delivered from AI today.
- The challenges are particularly acute in applying AI to commercially less attractive therapeutic areas and for LMIC researchers looking to harness AI. For example, longitudinal population datasets that can be mined to understand diseases and identify new targets may be scarce, of lower quality or absent in LMIC settings.
- Lack of commercial potential can limit data-generation and the applicability of key tools – with the potential for AI to amplify disparities in health equity. For example, whilst commercially attractive therapeutic areas such as oncology are well served, there is no equivalent in infectious diseases despite the much greater health-burden in many geographies.
- Even when high-quality data and mature tools are available, access to inter-disciplinary capabilities such as computational chemistry and bioinformatics has emerged as a key barrier in many settings.

Initiatives are emerging to tackle these barriers:

- Efforts such as the World Economic Forum and University of Oxford's AI Governance Research group are working to improve understanding and trust in AI across a range of settings including medical research.
- Initiatives are being established to create or enrich data and enable greater access. For example, the Wellcome-Sanger African Genome Variation Project is laying the foundations for generating high-quality genomic datasets in Africa. The US NIH is also funding grantees to clean and standardise existing datasets to improve their applicability to machine learning techniques – which is particularly critical in data-poor therapeutic areas.
- Powerful AI tools available to commercial R&D organisations are also in some cases being opened for use in less commercially attractive areas, such as, the Moderna Access and Atomnet platforms being deployed for historically under-invested infectious diseases.
- Capability gaps, particularly in LMIC settings are being explored and solutions trialled, such as, H3D-Ersilia's collaboration that provides a fully funded 4-day course to researchers in Africa on the use of AI in discovering drugs for locally relevant infectious diseases.

However, on the current trajectory, existing initiatives will be insufficient to unlock the potential of AI in Drug Discovery in ways that can equitably address urgent health needs.

Recommendations and call to action for funders

Unlocking the potential of AI in Drug Discovery will require a combination of ecosystem-wide actions and a broad portfolio of point solutions advancing specific use cases and expanding the applicability of AI in new therapeutic areas. This report has identified potential actions to be taken across four key areas:

1. Increasing **trust in** and **understanding of** AI in drug discovery
2. Developing high-quality **datasets**
3. Developing AI **tools**
4. Building **capabilities**

The report outlines actions that can be taken to increase **maturity**, expand **access** and support **standardisation** (see Figure 18). For example, actions to develop high-quality datasets may include enriching existing datasets with more entries or fields (increasing maturity), making currently proprietary data widely available to researchers (expanding access) or supporting the adoption of standard data structures to enable AI models to parse more available data (standardisation).

Priorities for funders will be determined by each organisation's strategic goals and capabilities in this space, with this report suggesting a call to action for funders in six key areas:

1. **Find value from AI today** by ensuring current grantees and partners are leveraging AI where use cases are mature e.g., small molecule design and optimisation or target identification in data-rich TAs such as oncology or immunology.
2. **Take no-regret moves to maximise future value** from research efforts generating data that might have utility in training AI models e.g., by mandating that data is published in open-access repositories, is machine readable and contains the requisite meta-data to support future interpretation.
3. **Build coalitions to shape the 'rules of the road'** with funders acting in concert globally to build norms in this rapidly developing field e.g., on data access, maintenance of open-source tools, transparent benchmarking of tool performance and expansion of training and development programs to data scientists and related AI disciplines.
4. **Invest where AI intersects with drug discovery goals** to see value today from mature use cases or to critically assess where AI most closely intersects with drug discovery goals and where intervention may be needed to accelerate progress e.g., in data-poor TAs such as infectious diseases where new foundational datasets may need to be generated before AI can deliver value, or where capabilities in LMICs may need to be upgraded to support local discovery efforts.
5. **Contribute to the public debate** as AI rapidly emerges as an 'all-of-society' topic and as new regulatory and legal instruments are being developed that will impact funders e.g., through transparency on AI-related activities, outcomes (positive and negative) and advocacy for AI-enablers such as equitable data access and standardisation.
6. **Build the organisational capabilities to deliver** a range of AI activities including critically identifying AI opportunities, advising current grantees on application of AI to their research, and appraising future grant applications e.g., by funders determining how AI fits into their funding strategy and programme of calls and then ensuring the organisation has access to the right capabilities (internally or externally) at the right level of capacity to deliver.

AI in drug discovery is at an inflection point. A number of mature use cases are delivering value today and provide immediate opportunities to help researchers discover new medicines to improve human health. At the same time, barriers risk concentrating the benefits of AI to already data-rich and commercially attractive TAs with limited opportunity for researchers in other areas to engage. Concerted action is needed today to shape this emerging field and set the 'rules of the road' that will allow equitable benefit from the transformational opportunities of AI in drug discovery.

2. About this report

2.1. Scope

Report objectives

Applications of AI in drug discovery are advancing rapidly and generating significant attention from industry and academia, as well as funders and policy makers. Interest amongst the public has also grown significantly in recent years as advances in AI are reported by the popular media.

However, many stakeholders have struggled to separate fact from hyperbole. Assessing the current status and future potential of AI in drug discovery typically requires knowledge of both data science and drug discovery. As a result, the ability to evaluate the topic is often limited to a small number of well-funded companies and institutions, and there are few publicly available resources that describe the field.

This report, intended as a global public good, aims to provide a fact-base for stakeholders looking to understand the current status and future potential of AI in drug discovery. It is also intended for funders considering how to engage on this critical topic (see Figure 1 for more detail on report objectives).

Given the broad interest in the field, this report is designed to be informative for stakeholders directly involved in the drug discovery process (e.g., academics, industry researchers) as well as stakeholders that influence the field (e.g., funders, policy makers and non-profits).

Technological scope

In this report, AI is defined as an umbrella term for a range of advanced computational and modelling techniques that analyse and learn from often large and complex data sources and can generate insights or perform tasks that would typically require human-level intelligence, at a scale and speed beyond human capability. Techniques classified as AI for the purposes of this report includes:

- *Machine learning (ML)* – a subfield of AI that focuses on developing algorithms and statistical models that enable computers to learn from and make new predictions or decisions based on data. ML techniques include, for example, the random forest algorithm, or the Naïve Bayes classifier.
- *Deep learning* – a subset of ML which uses artificial neural networks to learn increasingly complex representations of input data, such as unstructured data and image recognition. Model types include, for example, convolutional neural networks (CNNs) and autoencoders.

It is important to point out that Large Language generative Models (LLMs), are a subfield of deep learning, and its impact on drug discovery is an emerging topic that is currently receiving much attention. However, it is not covered in this report due to the paucity of reliable research published to date.

Figure 1 – Objectives of this report

Identify the key **use cases** and **applications** of AI in drug discovery

Assess the **maturity of AI in Drug Discovery** across modalities, stages of the drug discovery value chain, and therapeutic areas

Baseline the **adoption of AI in drug discovery** and determine current **barriers** limiting its use

Identify **solutions** to overcome theses barriers to unlock the **potential of AI** in drug discovery

Scope of R&D activities

This report focuses on the application of AI on the discovery portion of drug research and development, encompassing all steps from target identification up to and including pre-clinical development (see Figure 2 for definitions of each phase). The analysis covers both academic and industrial research, across high-income countries (HICs) as wells as low-and middle-income countries (LMICs).

When considering the application of AI within industry, this report will focus on two types of organisations: pharmaceutical companies and 'AI-first' biotech companies. The latter has been defined as biotechnology companies for which AI is central to their discovery activities.

Whilst there is great value of applying AI to other aspects of drugs and vaccines – such as in clinical trials (e.g., protocol design), chemistry manufacturing and controls (e.g., raw material forecasting), lab efficiency (e.g., reagent identification), diagnostics (e.g., image analysis) – these are beyond the immediate scope of this report owing to their limited activity in academia and resource-constrained settings, thus focusing on a different stakeholder group to that of drug discovery.

Modality scope

This report focuses on the use of AI to develop new small molecules and biologics – specifically monoclonal antibodies (mAbs) and vaccines, focussing primarily on the mRNA modality for the latter. These modalities have been selected due to their potential to serve a significant portion of global health needs. Small molecules, mAbs and vaccines are already widely deployed globally and represent the major focus of drug discovery efforts. In contrast, highly personalised therapies, such as cell therapies and gene therapies, are currently only narrowly adopted and remain challenging to manufacture and administer at scale. These latter therapies are beyond the scope of this report.

Figure 2 – Description of each phase in drug discovery

	Target Identification and Validation		Molecule Optimisation			Preclinical
	Target Identification and Validation	Target Characterisation	Hit Identification	Hit to Lead	Lead Optimisation	Preclinical
Description of Phase	Identifying and validating a target through desktop research, demonstrated *in vitro* evidence	Understanding target and identifying function-modifying sites, designing antigens, etc.	Screening molecules/mAbs/ antigens for desired response, often using high-throughput screening	Confirming basic desirable properties of hit and short-listing best hits for next step	Optimising lead towards desired properties, including absorption, solubility and developability	Selecting candidates for Investigational New Drug application and clinical studies based on studies in animals

2.2. Methodology

This report combines three analytical approaches to develop a perspective of AI in drug discovery:

1. **Desk research** including literature review, patent analysis, private sector funding analysis and review of pharmaceutical and 'AI-first' biotech R&D pipelines.
2. **Stakeholder survey** to gain perspectives from a broad mix of individuals across academia and industry, and across a range of geographies.
3. **Expert interviews** to pressure test findings and supplement analysis via expert input.

These analyses were then synthesised to answer four research questions:

1. What is the current state of AI in drug discovery?
2. How are AI use cases currently being adopted across different drug discovery settings?
3. What are the barriers to further adoption of AI in drug discovery?
4. What initiatives could support further adoption of AI in drug discovery?

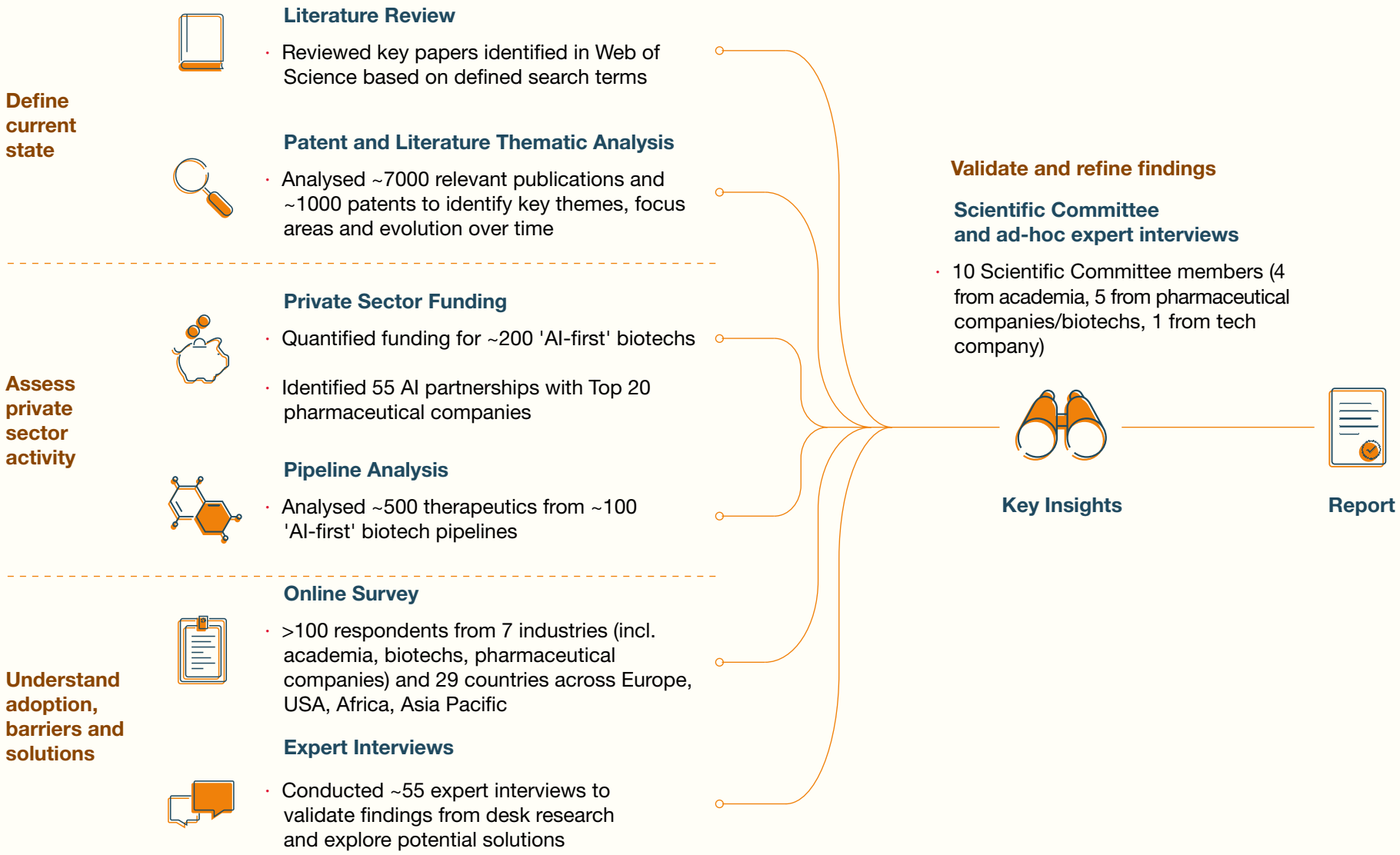
Additional methodology used for specific analyses is described below:

- **Literature review** was performed based on a keyword search for papers appearing in Web of Science (a research aggregation platform) in the last five years (2018-2022). To support interpretation, data was then clustered and visualised using Quid – a tool that deploys AI to semantically cluster data.
- **Patent analysis** was conducted based on LexisNexis PatentSight entries in the last five years (2018-2022), filtered based on relevant keywords.
- **Private sector funding** for approximately 200 'AI-first' biotechnology companies and 20 large pharmaceutical companies was assessed using Pitchbook, Biomedtracker and press releases from the last five years (2018-2022). See section 10.4 for the list of 'AI-first' biotechs.
- **Pipeline review** was limited to 96 of the approximately 200 'AI-first' biotechs, where information was publicly available through Citeline PharmaProjects and company press releases over the last five years (2018-2022).
- **Stakeholder survey** received online responses from 102 individuals across academia, industry, and funders from both HIC and LMIC settings. Topics covered were current and planned future adoption of AI, perceived current and future value of AI, barriers to adoption of AI and potential solutions.
- **Expert interviews** canvassed in-depth perspectives from 55 experts across academia, industry, and funders from both HIC and LMIC settings.

Findings were additionally calibrated and refined with a Scientific Committee consisting of academic and industrial experts in the field. The Committee met three times over the course of producing this report. See acknowledgements for the composition of the Committee (Section 9.2).

Collectively, these data sources and analyses enabled both a quantitative and qualitative assessment of the maturity, adoption, and perceived value of AI techniques within drug discovery today, and barriers that would be most important to overcome to unlock the true potential of AI within this field going forward.

Figure 3 – Data inputs and process used to inform report findings



3. Promise of AI in drug discovery

Significant advances have been made in the field of AI over the past two decades. AI is now routinely applied in many industries, and current AI applications cover a broad range of activities, including image recognition, mining of large and unstructured datasets, personalised learning, and many others. In the last 6-12 months, breakthroughs in large language models, such as GPT-4, are redefining human-computer interactions.

Drug discovery is no exception to these broader trends seen in AI. Whilst computational methods in drug discovery gained widespread adoption through the 1990s and early 2000s, the adoption of AI is more recent. In the 1990s, researchers increasingly employed computer-aided drug design (CADD) to support both data generation and data analysis to build more accurate hypotheses, which consequentially, led to an exponential increase in the quantity of data related to drug discovery. Whilst this digitalisation came with newfound challenges, namely how to best use and analyse such large volumes of data to aid the solving of complex clinical problems, it also brought about a better understanding of algorithmic principles – and thus the dawn of the use of AI in drug discovery [7].

> *AI usage will continue to rapidly ascend as tools improve. AI will play a huge role in processing large amounts of data which is vital for drug and vaccine discovery.*
>
> **AI Lead, Health Research Institute**

Over the past five years, the field has seen numerous breakthroughs. Most high-profile, perhaps, was the development of AlphaFold, a deep learning algorithm which can predict protein structures with an accuracy that approaches experimental methods [6]. This has fuelled research on a range of protein targets which had no experimentally characterised structure. Other key milestones in recent years has been the initiation of Phase I clinical trials for numerous AI-discovered drugs and vaccines, such as Relay Therapeutic's selective inhibitor of FGFR2 [8].

AI has the potential to create significant value in drug discovery, primarily through three main drivers: (i) time and cost savings, (ii) increased probability of success, and (iii) novelty of both the molecular target and optimised therapeutic agent [9].

With regards to ***time and cost***, AI has the potential to impact traditional drug discovery in several ways:

- AI could reduce the reliance on lengthy and expensive experiments and direct experimental efforts to areas with greatest impact (e.g., focusing target validation efforts on predicted disease-relevant targets), or in other cases replacing experiments entirely (e.g., virtual compound screening). These approaches can help researchers to "fail faster" and evaluate a broader range of targets or therapeutic compounds before progressing to experimental testing.
- AI could also adjust the drug discovery workflow. Traditionally, different teams have been responsible for each of the steps across the value chain, such as library design, compound screening, and synthesis. AI provides the opportunity to amalgamate these steps into one, streamlined – and potentially fully automated – process, reducing the need for in-process decision making. AI, therefore, augments the work of experienced human scientists and allows them to concentrate efforts at the end of AI-powered workflows. Additionally, rather than a linear progression from lead optimisation to ADME evaluation, predictive models also enable optimisation for multiple properties in parallel, thereby compressing discovery timelines.
- We see some indications of this where 'AI-first' biotechs have developed large preclinical portfolios in relatively short timelines (43 preclinical and 22 clinical assets from public 'AI-first' biotechs founded in the last 10 years alone)

> *Early-stage drug and vaccine discovery entails several rounds of optimisation and experiments which are costly and not easily accessible in resource-limited settings. AI tools can assist in prioritising the most promising targets and cut the number of experiments and costs in this area.*
>
> **Infectious Disease Lecturer, LMIC**

> *AI has the potential to identify more targets and candidates quicker. There will need to be a significant shift in the industry model to cater for this explosion of opportunity.*
>
> **Senior Executive, AI Software Company**

AI could also improve the ***likelihood of discovering a successful therapeutic***. Firstly, when used on appropriate datasets, AI may develop additional or improved hypotheses than traditional methods alone by synthesising vast datasets. Secondly, AI may be able to better select therapeutics with desired properties than traditional experimental methods. This could take the form of properties such as

efficacy, pharmacokinetic and metabolic properties, or those which result in more globally accessible therapeutics – such as thermostability, which could reduce reliance on cold chain supply and storage. Additionally, AI may be more capable at subsequently optimising these therapeutics for properties such as toxicity and immunogenicity to improve the likelihood of success during clinical phases. Proof points are emerging here, with 73 clinical assets from 'AI-first' biotechs catalogued as part of this work, with read outs expected in the next few years (see Appendix Section 10.6).

> *There are very few things that can increase the chances of luck in finding a good hit and AI is one of them.*
>
> **Head of Vaccines, Pharmaceutical Company**

Finally, the ***novelty of both discovered therapeutics, and the targets they select for***, could be enhanced by AI. Models may be able to generate novel therapeutic structures, such as bi or tri- specific antibodies, owing to their ability to explore a much vaster chemical/ biological space than humans can alone [10]. These novel structures may possess properties which enable the targeting of previously undruggable targets. For example, Absci's zero-shot generative AI model designs novel *de novo* antibodies that bind to specific targets without using any training data of antibodies [11]. AI may also be able to elucidate novel disease-driving pathways by establishing links between disparate biological signals – and thus identify new targets within these pathways that could be modulated therapeutically to treat the underlying disease. Recent partnerships highlight the perceived value of these approaches, for example, notable target discovery partnerships between AstraZeneca and BenevolentAI, and Insilico Medicine and Sanofi [12], [13].

> *AI will create new opportunities and drive innovation. It will help us fully explore the chemical space and hence find novel drugs.*
>
> **Head of R&D, Pharmaceutical Company**

However, whilst there is promise in the value of AI in drug discovery, this potential has yet to be demonstrated at scale, across populations and disease areas. For AI to truly realise its full potential in addressing global health challenges at scale, there is a need for better understanding its current applications and limitations, and the barriers that face the industry today.

> *Given the rapid progress of AI, it is almost inconceivable that AI won't have a role in drug and vaccine discovery. It will work – It's just a matter of when and how.*
>
> **Senior Associate, Life Science Venture Capital**

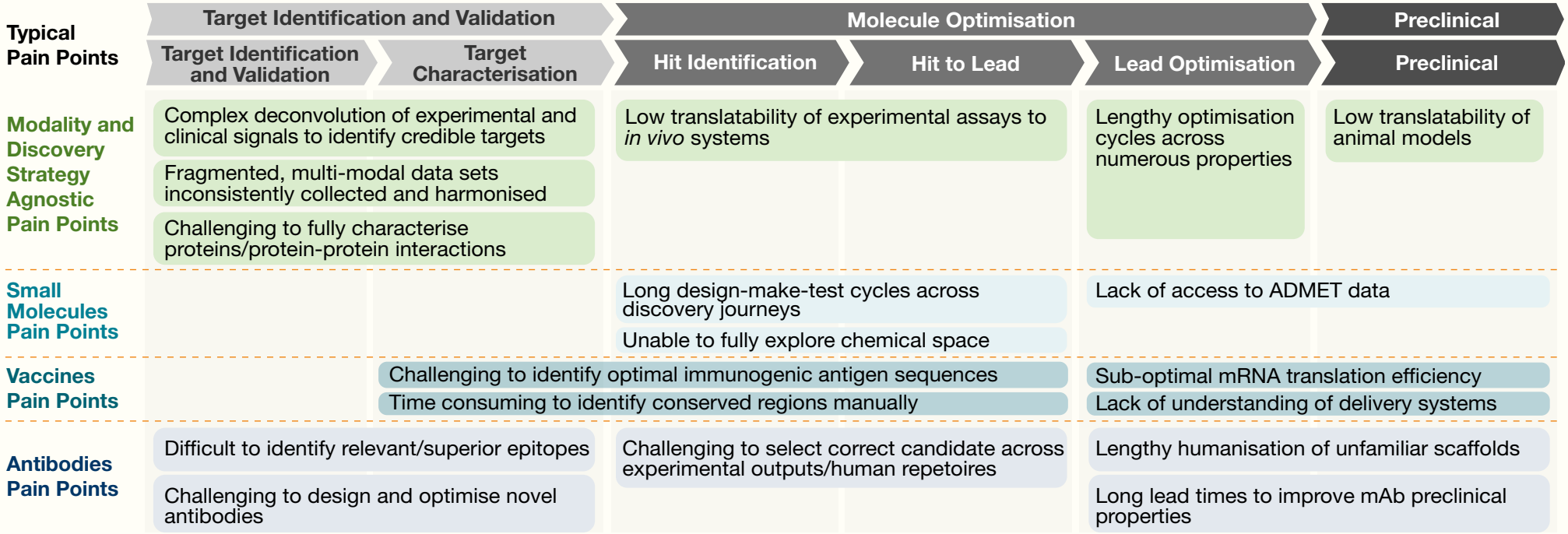
4. AI applications in drug discovery

4.1. Current challenges along the drug discovery workflow

As a starting point for assessing AI applications in drug discovery, the current drug discovery process was examined, and general pain points identified (see Figure 4). For example, within the molecule optimisation phase, there is often a need to complete multiple lengthy design-make-test cycles to define a therapeutic with desired properties – and this not only increases the duration of the discovery process, but also adds significant cost.

Other pain points are specific to a modality or discovery strategy. For example, in vaccine discovery it is often challenging to identify optimal antigenic sequences as experimental methods normally employed to explore the binding affinities of antigens to proteins of the human immune system are often complex and laborious to conduct [14].

Figure 4 – Key pain points along the drug and vaccines discovery process today



4.2. AI use cases along the drug discovery workflow

The analysis of literature, patents as well as input from drug discovery experts via interviews and survey, suggests five major families of use cases for AI in drug discovery (outlined in Figure 5):

1. ***Understanding disease*** use cases deploy AI to identify and validate new targets for drug discovery efforts. These typically include deploying AI to automate image analysis from phenotypic screens; -omics mining (e.g., proteomics, genomics) to better understand how a given target modulates disease progression; protein dynamics modelling to understand how a target interacts with a disease pathway; and biomarker identification to better segment patient populations for drug research.

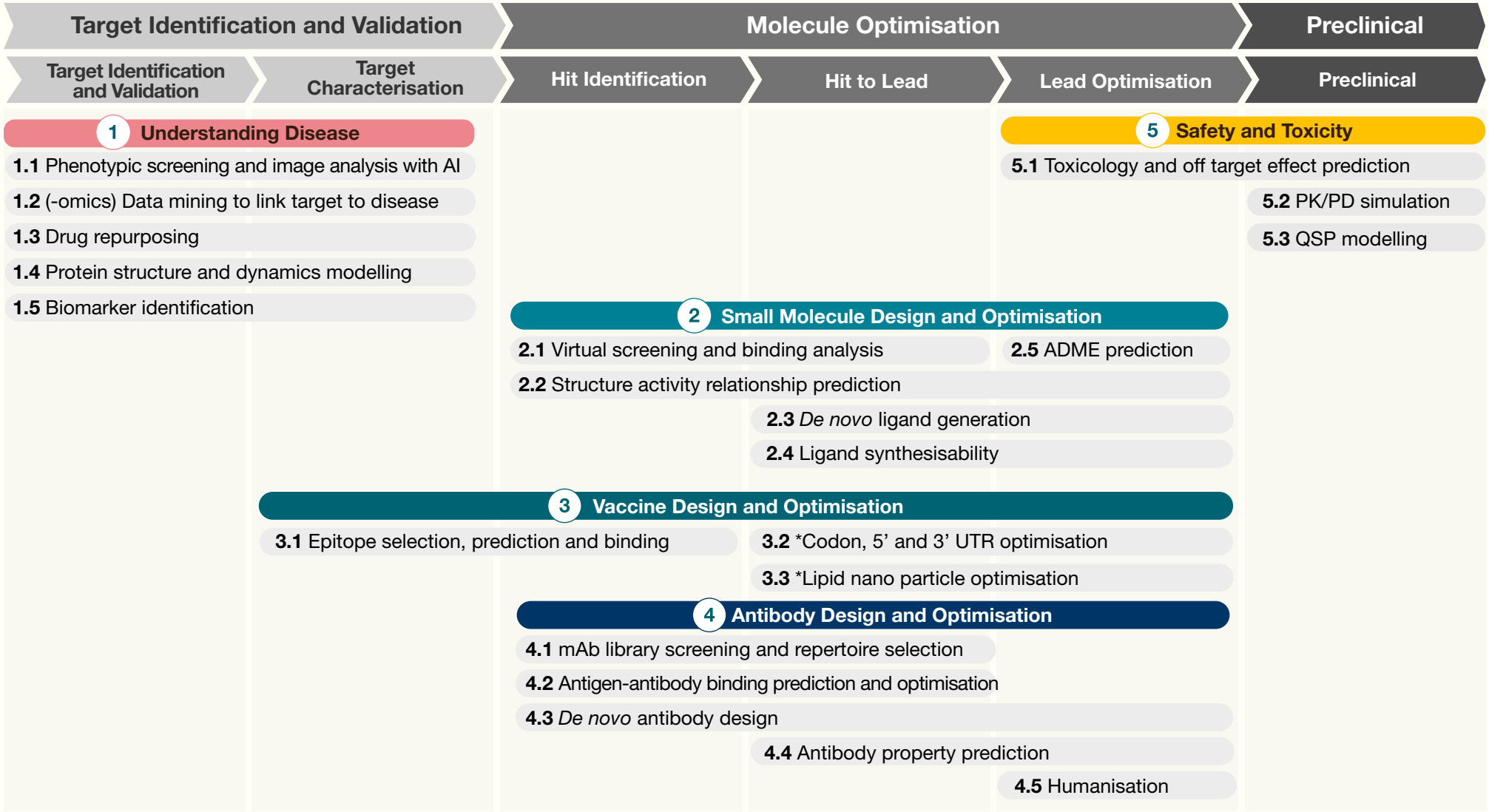
> *"There is an abundance of opportunities to use AI to better understand diseases. There's already a vast amount of existing data that AI can synthesise to establish disease pathways.*
>
> **Head of R&D AI, Pharmaceutical company**

2. ***Small molecule design and optimisation*** use cases deploy AI for two types of activities: identifying hit-like or lead-like small molecule compounds; and optimising the identified hits for favourable properties such as binding affinity, toxicity, and synthesis. AI can be utilised for identifying compounds both via screening of existing chemical libraries and via generative *de novo* design.
3. ***Vaccines design and optimisation*** use cases encompass AI applications specific to the discovery and design of vaccines, with a primary focus on mRNA-based vaccines, but in some instances use cases can be applied to other vaccine modalities. For example, AI use cases on epitope selection, prediction and binding are applicable for all vaccine designs, irrespective of modality. On the other hand, use cases pertaining to codon and delivery system optimisation – to ensure heightened protein production per dose with minimal toxicity – are specific to the mRNA vaccine modality.
4. ***Antibodies design and optimisation*** use cases deploy AI on a wide range of applications focused on identifying and optimising antibody structures and formats, binding and other properties. For identification of molecules, two core AI applications exist – the screening of pre-existing libraries and the more nascent *de novo* design capabilities. There are also emerging examples of where AI has played a role in the subsequent optimisation of these molecules, for example, for binding physicochemical and humanisation properties to ensure high target specificity, affinity and potency.
5. ***Safety and toxicity*** use cases focus on using AI to evaluate the safety profile of a therapeutic or vaccine of interest. Given the highly specific nature of established toxicity approaches, there are relatively fewer applications of AI within this space, compared to the other use cases. Also, it is often challenging to build generalisable AI models for safety ad toxicity which can be applied to a broad range of settings. However, some point solutions do exist which mostly fall into one of three use cases – predicting off-target impacts; simulating pharmacokinetics and dynamics; and modelling the interactions between the molecule of interest and a biological system via quantitative systems pharmacology (QSP).

Irrespective of use case family, experts highlighted that AI would not replace the role of experienced drug discovery scientists, but rather enhance it by allowing scientists to focus on higher-value and more varied tasks.

> *"At nearly every stage of the drug and vaccine discovery process there is an opportunity for AI. But it is not a simple solution where you 'press a button' and a chemical comes out. We need medicinal chemists to be able to interpret the output of the AI models as ultimately that is where the value is.*
>
> **Head of Vaccines, Pharmaceutical company**

Figure 5 – Over of AI use cases based on analysis of literature, patents, as well as input from drug discovery experts



* Applicable for mRNA vaccines only

5. Current state of AI in drug discovery

5.1. Evolution of AI tools in drug discovery

This section assesses the current state of AI tool development in drug discovery across sectors (academia, pharmaceutical and biotech companies), modalities (small molecules, biologics), therapeutic areas (oncology, infectious diseases, neurology etc) and geographic settings (HICs and LMICs). Leading indicators of tool development are then assessed within academic settings and commercial advancements, through the progress of 'AI-first' biotechs across our use case map.

Use cases and applications of AI in drug discovery

Figure 6 shows an analysis of publications on AI in drug discovery by use case family. In the period from 2018 to 2022, AI was most frequently applied to use cases pertaining to understanding diseases and small molecules design and optimisation, suggesting a greater maturity of these use cases. While design and optimisation of vaccines and antibodies historically lagged, they are now accelerating rapidly, driven in part by research efforts in the context of the COVID-19 pandemic. Biologics AI use cases are also growing, driven by increasing sophistication of AI technology and algorithms, growing computing power, increasing availability of data, and evolving discovery workflows [15]. Finally, safety and toxicity use cases are also relatively nascent but growing, and literature suggests that this is driven by the lack of publicly available data from which AI models can be trained [16].

> *"Biological problems are not only more complex to understand (than chemical ones) but also much harder to solve computationally as there is less data available.*
>
> **Microbiology researcher, LMIC**

> *"Safety and toxicity is one of the pockets of pharmaceuticals where AI is neglected, as it is hard to automate in vivo models.*
>
> **Senior Director, Pharmaceutical company**

In our expert interviews and survey, we see that many drug discovery organisations view target identification and validation, as a key competitive differentiator. This attention to improving ***understanding of diseases*** has driven some therapeutic areas to reach a tipping point in terms of the quality and quantity of data available to train AI models against. Not only are new data sources becoming available (such as imaging, -omics, clinical data, data gathered from wearable devices), but the advent of new experimental and patient-derived models have also created richer and more translation-relevant datasets, alongside algorithms that help combine, contextualise, and draw insight from sparse, fragmented, structured & unstructured datasets (from publications to in vitro data). A more detailed analysis (Appendix Figure 20) shows that, (-omics) data mining to link target to disease and drug repurposing were the most frequently published sub-use cases deployed for understanding disease. However, it should be noted that the true extent of -omics publications is likely under-represented in this analysis given the long-established precedence for AI approaches to pattern recognition within this type of data.

> *"AI has large innovation potential especially within Understanding Disease. The biology we are trying to understand cannot reasonably be understood by people – we need the AI models.*
>
> **Senior Executive, Pharmaceutical company**

> *"-omics mining was big a decade ago. Tools are so standardised that no one mentions AI in omics anymore, it's just assumed.*
>
> **Lead Developer, Scientific AI Software**

AI activity in ***small molecule design and optimisation*** is widespread across the drug discovery community, driven by the availability of well-validated tools, such as those for AI-driven screening and design, which are widely available and being applied especially in industry. There are, however, a number of important nuances. For example, there is relatively limited literature about the application of AI to ligand synthesisability. In contrast, in virtual screening, there are more publications and a wide variety of open-source AI tools available (e.g., VirtualFlow, PyRMD) [17], [18]. Small molecule AI approaches have seen substantial attention from the pharmaceutical industry, where many companies are deploying their own solutions, although less information on this progress is publicly available [19].

For ***vaccines design and optimisation*** use cases we see a limited number of publications. Many academics participating in our research mention a lack of clinical serology data for tool development as a primary hindrance, especially for models identifying optimal antigen sequences and predicting immunogenicity. One area where the use of clinical serology data is more commonplace for identifying immunogenic antigens is within the influenza field which has led to the development of models, such as FluLeap, which can accurately categorise novel influenza viruses as either avian or human [20]. Efforts are also underway to explore immunogenicity in more detail – the Human Immunome Project, for example, aims to better understand the molecular basis of immunity through the combination of AI and systems biology techniques (see Section 8 for further detail) [21]. This could be an inflection point for the development of epitope selection, prediction and binding tools, resulting in significant acceleration of the discovery of new vaccines and other therapeutics. Many academics also highlighted the need for, and current lack of, clinical feedback loops to test and validate AI approaches – although experts have speculated this need might change with the advent of rapid mRNA sequence generation.

When it comes to ***antibody design and optimisation*** use cases, a large proportion of AI efforts have been taking place in pharmaceutical companies and 'AI-first' biotechs compared to academia, which is reflected in fewer publication data on the same. However, publication output of the last few years suggests an acceleration of AI, with a focus on sequence-structure determination (e.g., by building on the capabilities within Alphafold), in vitro library screening and antigen-antibody binding prediction. Other use cases further down the antibody value chain focus on the multiparameter optimisation of the antibody ahead of preclinical studies. These efforts require extensive datasets, which historically have resided in pharmaceutical companies and contract research organisations (CROs) and not always available publicly, which perhaps explains the more limited literature in this space.

> *We are starting to see a bigger AI impact on antibody design. AI makes it easier to make and screen 10k antibody molecules.*
>
> **Senior Vice President, 'AI-first' biotech**

For ***safety and toxicity***, the story is somewhat different. Activity in academia is currently limited by the lack of data in the public domain on which to train models, and efforts to model broad toxicity impacts are few and far between. However, some concentrated areas of activity exist, such as those related to pharmacokinetics and pharmacodynamics simulation, or image analysis from patient biopsy samples. There are also increasing efforts in quantitative systems pharmacology (QSP) that may further increase momentum going forward, spurred by the recent FDA Modernisation Act 2.0 in the United States, which includes a provision to allow pre-clinical approval without the need for an animal model evidence base [16].

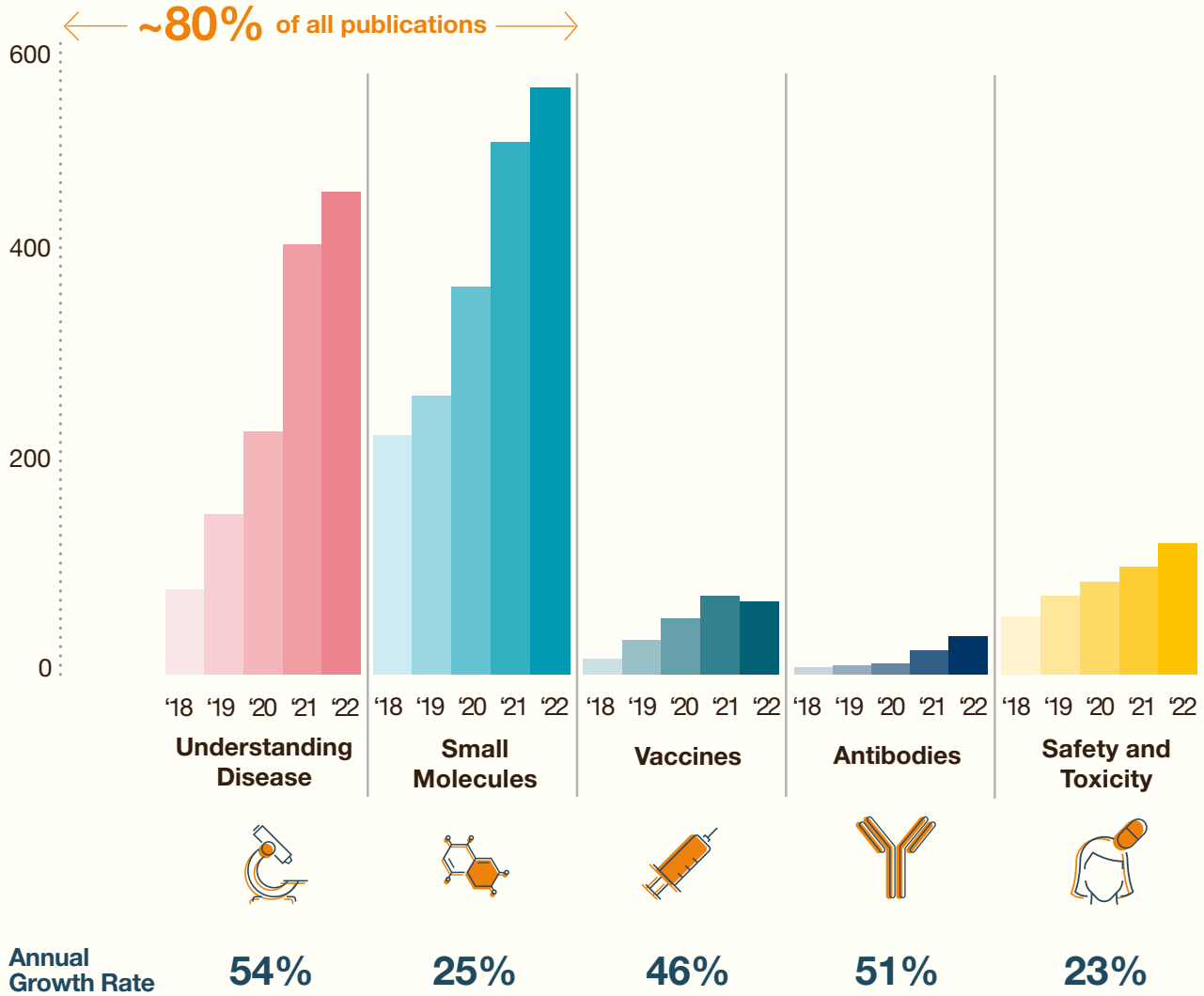
> *It is very difficult to build a generalisable model for ADME / Toxicity because of the idiosyncratic nature of the underlying data. And where generalisable models do exist, teams need to be well versed in evaluating the outputs of predictive tools to ensure we are not over interpreting results.*
>
> **CEO, 'AI-first' biotech**

> *Toxicity models broadly fall into one of two categories; either they predict the impact of drugs on specific toxicity pathways (e.g., CYP340 for liver toxicity) or they help identify idiosyncratic toxicity. The former requires highly specific models and training data, the latter may simply not be detectable with the model capabilities we have today.*
>
> **Lead Developer, Scientific AI Software**

Applications of AI across different Therapeutic Areas

When considering the use of AI across different therapeutic areas we see a similar trend, with most publications focusing on therapeutic areas that are both data-rich and commercially attractive, such as

Figure 6 – Publications on AI in drug discovery, by use case family and by year



oncology (37%) and COVID-19 (23%). Infectious disease (12%) is also well-represented in literature but concentrated in diseases where significant commercial incentives exist or where philanthropic funding is available. As a result, the majority of publications in infectious disease cover malaria, tuberculosis, and HIV, and less than 1% of literature focuses on other infectious diseases including neglected tropical diseases (NTDs).

Many experts we interviewed highlighted that research culture and incentives may further concentrate efforts in existing areas as it is easier to document progress in improving an AI tool where one already exists as a proof point. This ultimately limits the development of tools to within a small sub-set of therapeutic areas where efforts already exist today.

> *We focus our algorithms on the same therapeutic areas because we are forced to compare new methodologies with a technique from a seminal paper a decade ago. And this is worse in the field of AI – it's hard to tell the quality of papers when they are being published at such a fast rate, so people anchor even more heavily on historic ones.*
>
> **Bioinformatics Professor, Academia**

In contrast, other disease areas such as mental health have been comparatively under-served by AI. These disease areas often lack high-quality data (e.g., consistent coding/patient phenotyping) with sufficient dimensionality and depth from which to train models, and in some cases, also lack an understanding of the biological mechanisms that cause the underlying disease.

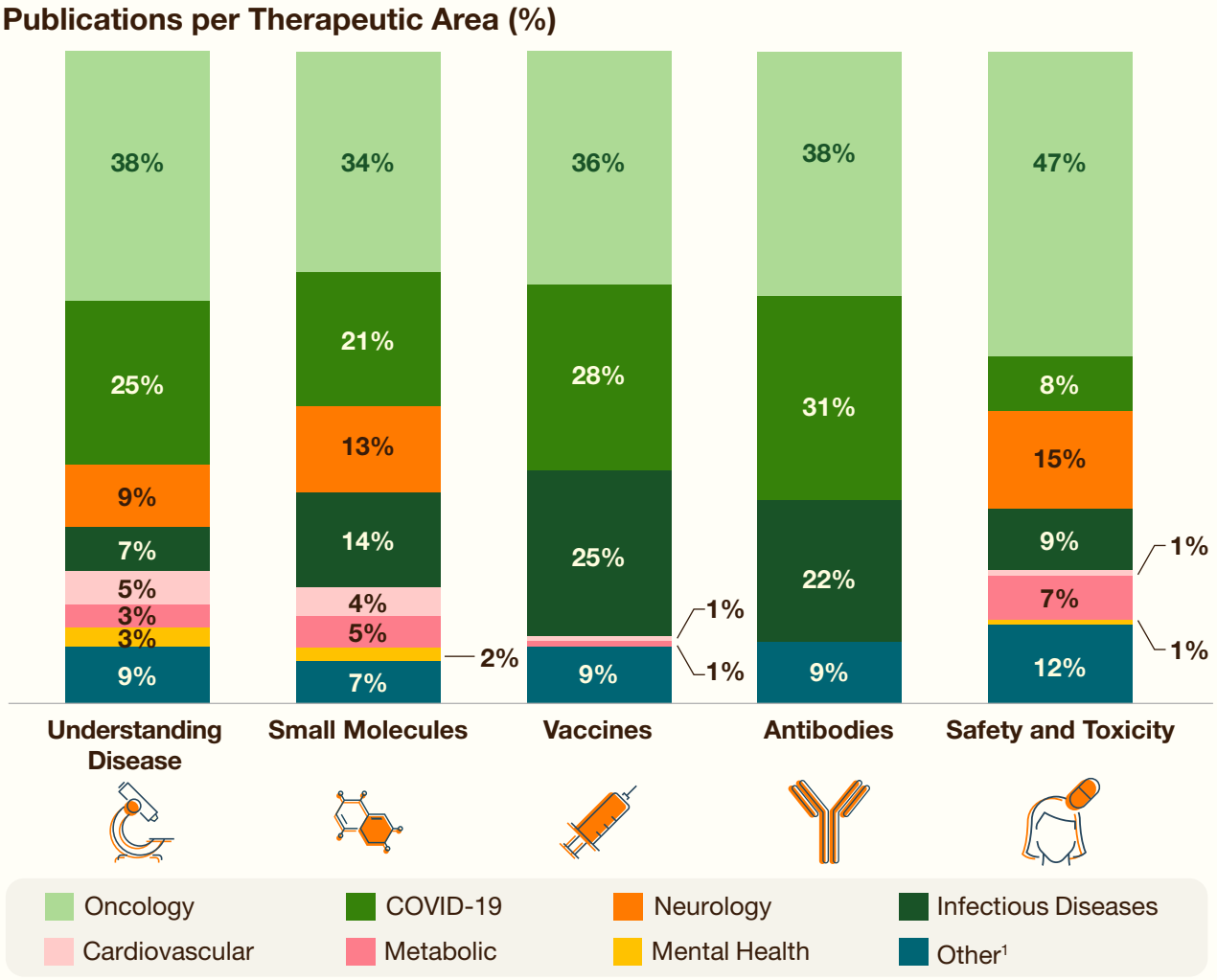
> *A breadth and depth of metadata is required to better understand disease drivers. Within depression, for example, not only is it pertinent to understand genetic predispositions, but factors such as alcohol consumption, sleep schedules and frequency of social interactions should also be considered.*
>
> **Senior Vice President Business Strategy, 'AI-first' biotech**

An analysis of the pipelines of 'AI-first' biotech companies shows a similar pattern. The vast majority of assets in these pipelines are that of small molecules, although vaccines and antibodies have been growing strongly (see Figure 8). Furthermore, their assets are focused on data-rich, commercially tractable therapeutic areas, such as oncology and neurology.

Our results also show that therapeutic area focus is often influenced by the need to manage the pipeline and balance risk, especially in the case of 'AI-first' biotechs. In the early stages of their evolution, when resources are often limited, many 'AI-first' biotechs focus on commercially attractive therapeutic areas and indications. Later in their evolution, when companies are better established and funded, many 'AI-first' biotechs expand to other therapeutic areas.

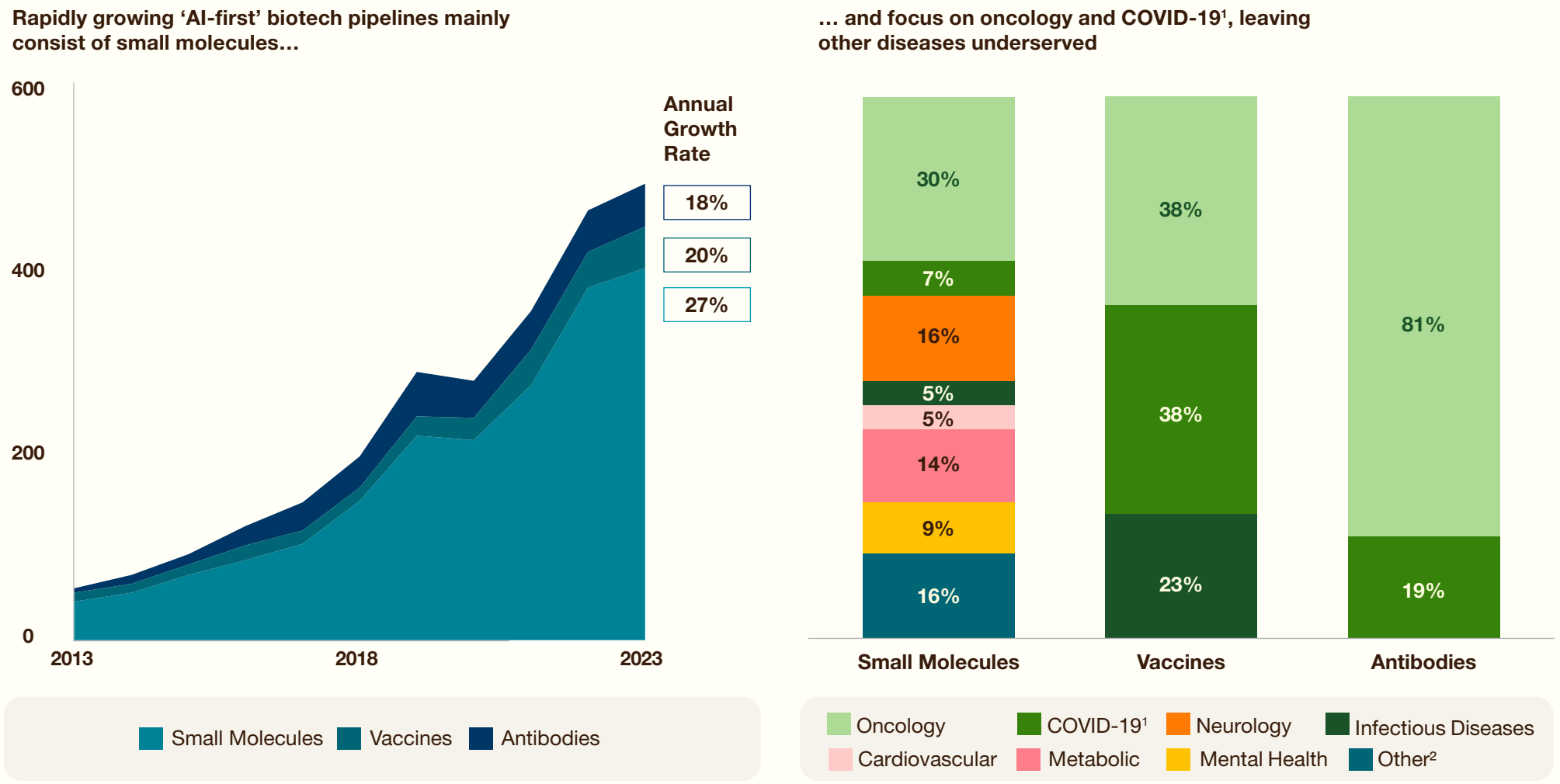
> *Using our platform to improve global health would be a fantastic end result, but it's just not feasible from an economic perspective in a cash-poor, early-stage start-up.*
>
> **Senior Executive, 'AI-first' biotech**

Figure 7 – Publications on AI in drug discovery, by therapeutic area



1. Gastrointestinal, Immunology, Respiratory

Figure 8 – Pipeline and therapeutic area focus of 'AI-first' biotechs



1. Numbers are one-off and not an ongoing trend
2. Gastrointestinal, Immunology, Respiratory

Investment in AI in drug discovery

The investment pattern from private funders follows a similar trend as seen above. Across the approximately 200 'AI-first' biotechs identified (see appendix section 10.4), analysis showed a total investment of over $18Bn over the past ~10 years. However, this has been highly focussed, with 60% of total investment concentrated within the top 20-funded 'AI-first' biotechs. Within this top group, 80% focus on understanding disease and small molecule use case families, with limited signs of expansion into vaccines or antibodies so far.

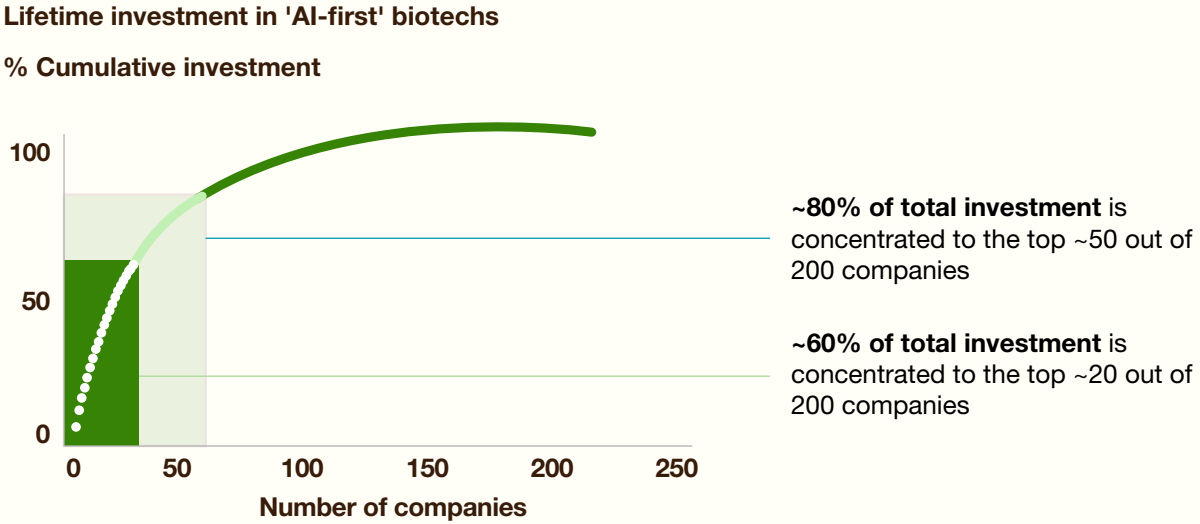
Please not that with regards to investment, it is not easy to distinguish cause and effect. Investment may be going into areas where investors see the greatest transformative and commercial potential for AI. Alternatively, the influx of investment itself may catalyse the development of technologies so that well-funded areas mature more rapidly.

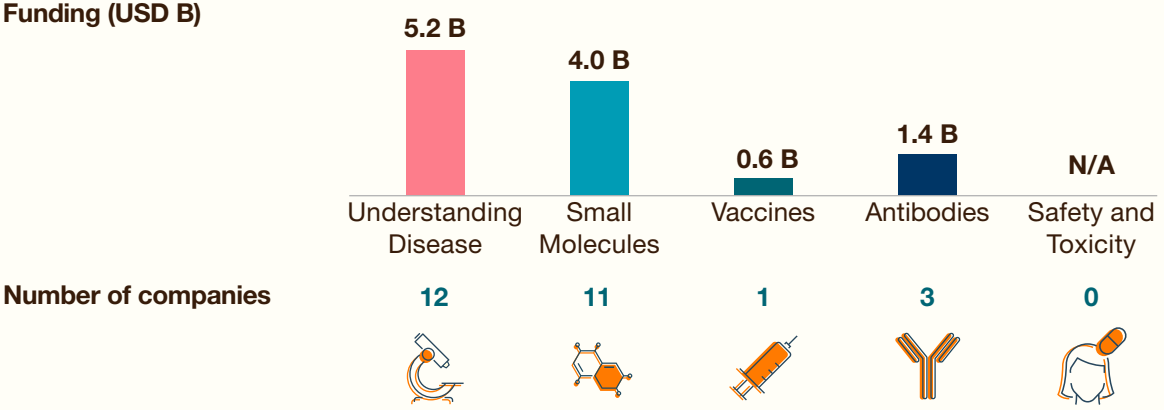
> *"Funding is what limits 'AI-first' biotechs to a specific corner. Funders want to see return on their investment and some use cases are more promising than others.*
>
> **Head of R&D, Pharmaceutical company**

Summary: Current state of AI in drug discovery

Whilst there is no doubt that the innovation and investment activity within the AI drug discovery space has been growing rapidly over the last five years, this growth has been uneven across use cases. Progress has been rapid in areas where data is abundant and publicly available, and industry efforts are often skewed towards commercially tractable, well validated, or data-rich therapeutic areas. In contrast, there has been much less progress in areas where data availability is limited, or areas which are commercially less attractive and therefore less likely to capture investment.

Figure 9 – Investment in 'AI-first' biotech companies





1. Includes IPO funding where relevant

Note: Where companies operate across multiple use cases, funding has been split equally across the use cases and the companies have been double counted

5.2. Adoption of AI in DD

Through expert interviews and an online survey, we assessed the adoption of AI technologies in the drug discovery space today. Despite the promise of AI within drug discovery, our analysis shows that, overall, AI is not routinely adopted across use case families and sectors.

Adoption of AI is highest within the ***Understanding disease*** and ***Small Molecule*** use case families, likely driven by the historic precedence of activity in these fields (as described in section 5.1), and owing to better availability of established tools, open source or otherwise. However, it should be noted that even within these use cases, adoption is uneven and there is much more that could be done to further the application of technologies in the field.

> *"There are quick wins in Small Molecules; we are setting AI on a problem that is easiest to solve initially.*
>
> **Vice President, 'AI-first' biotech**

Vaccines use cases have the lowest adoption amongst users in our survey. It should be noted, however, that in some vaccine discovery organisations, AI has seen much greater adoption, with some experts calling out the COVID-19 pandemic as a turning point for the use of AI across the field. This may reflect a distinct evolution of the field, in which some pharmaceutical companies have taken a leading role in driving the use of AI. Recent efforts, such as vaccine pandemic preparedness initiatives, may help address some of these disparities in roles between industry and the rest of the field [22], [23].

The current adoption of AI is primarily taking place within 'AI-first' biotech players, with academia and pharmaceutical companies showing lower adoption. Overall, HICs and China show higher adoption rates compared to LMICs. This is perhaps not surprising; 'AI-first' biotech players structure their organisations to embed AI across their workflows to maximise its benefits, and this has led to the development of extensive pipelines of therapeutics and vaccines, discovered in an AI-enabled manner. 'AI-first' biotechs are also increasingly building scale through mergers and acquisition activity, for example Schrodinger's acquisition of XTAL Biostructures to create internal laboratory capabilities to generate data to feed models; and Valo Health's acquisition of the protein therapeutics player Courier to enable expansion across the drug discovery value chain (away from their historic focus on *understanding disease* and *small molecules* use cases).

> *"AI has propagated into every area of our research and development.*
>
> **CEO, 'AI-first' biotech**

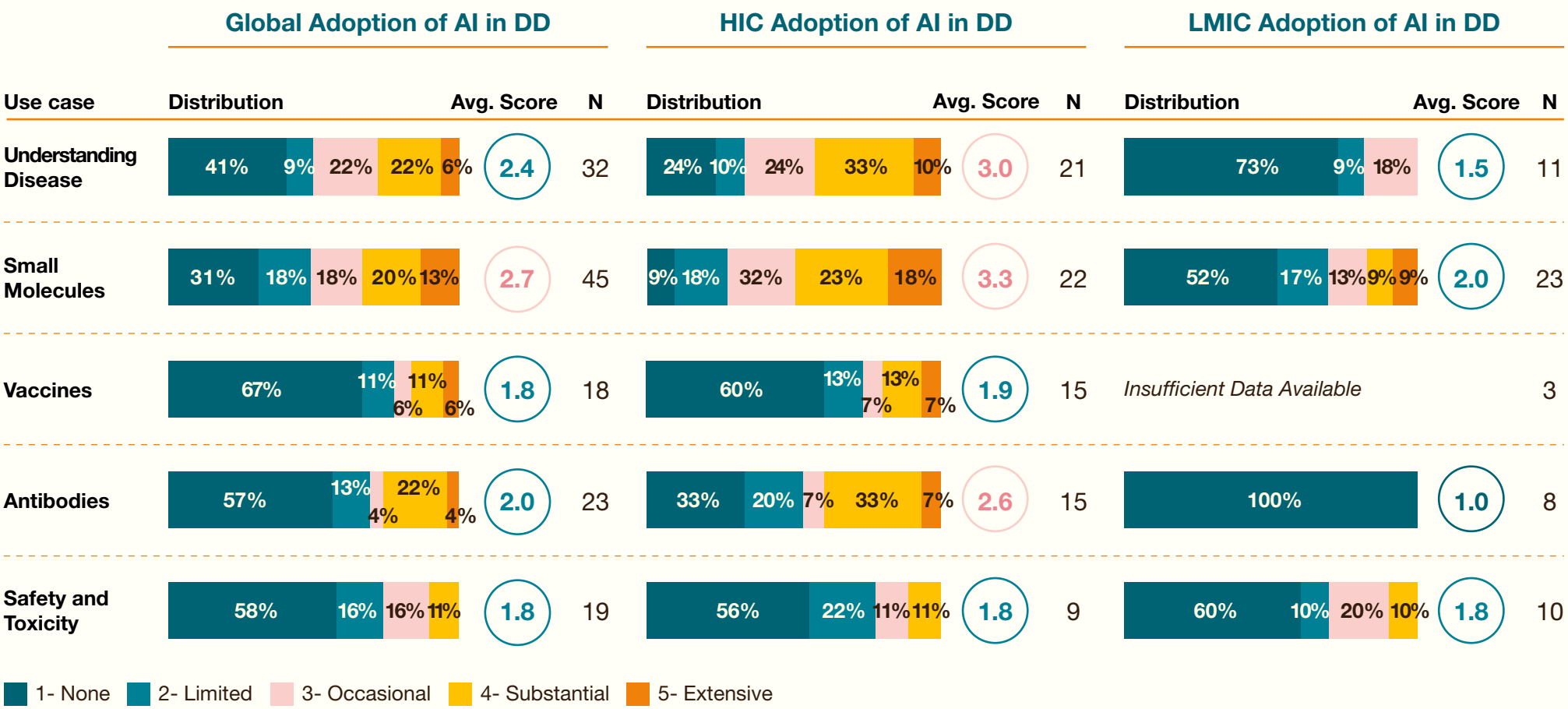
> *"Do we really want to spend our time and money to develop algorithms when we can be a fast follower instead? The latter is our approach to adopting AI.*
>
> **Head of R&D AI, Pharmaceutical company**

In contrast, in more established organisations, ingrained processes, working practices, and functional silos can limit adoption of AI. Our analysis indicates that embedding AI solutions in these organisations requires extensive change management, and cross-functional collaboration to establish new, AI-powered working practices.

> *"Pharmaceutical companies already have trained skill base in one area, so embedding AI tools within their standard workflow would require a huge amount of retraining and rehiring. So, they are sceptical – they see the value of AI, but they are dipping their toes in before they commit fully.*
>
> **Scientist, 'AI-first' biotech**

Whilst these barriers present challenges for the deployment of AI in established R&D organisations, such as large pharmaceutical companies, many are beginning to experiment with AI through partnerships with 'AI-first' biotechs. Other pharmaceutical companies are investing significantly to develop and deploy AI tools internally [24]. If successful, these initial experiments can then be scaled up – as exemplified by the long-term collaboration between AstraZeneca and Benevolent AI, which has resulted in the identification of five novel targets that have entered the AstraZeneca portfolio since 2019 [12].

Figure 10 – Adoption of AI in drug discovery across use cases, HICs and LMICs

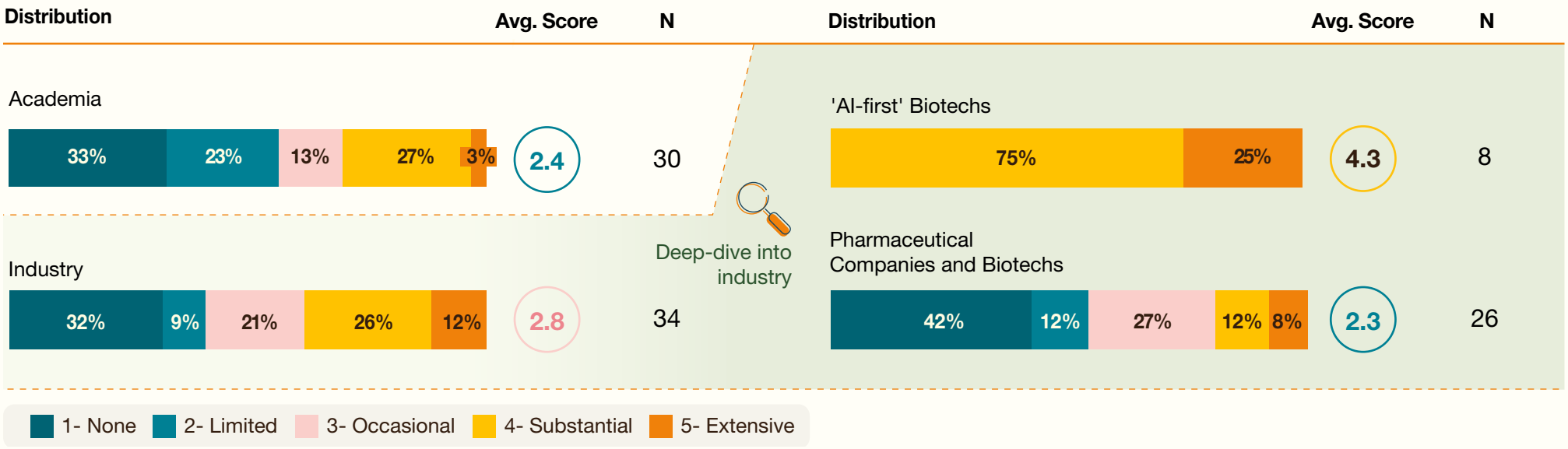


Avg. Score refers to average adoption score. N refers to number of respondents. Survey Question: "How would you characterise your organisation's usage of AI tools to support the use case family?" Survey Options: 1. Don't use AI 2. Some limited experimentation is occurring, but AI tools aren't routinely used. 3. AI tools used on an ad-hoc basis for specific processes where it is valuable, AI tools aren't a part of workflow. 4. AI tools used commonly for many of the critical processes and are part of workflow. 5. AI tools used as a standard for most of the critical processes and are a standard part of workflow. Note: Respondents may overlap between use cases and can only choose one option per use case

> *Amongst the big players, everyone is trying to jump on board quickly one way or another. We are using external partnerships to extend our capabilities.*
>
> **Scientist, Vaccines, Pharmaceutical company**

Across the pharmaceutical industry, large numbers of partnerships can be seen. In the last 5 years, the top 20 pharmaceutical companies have engaged in at least 55 relevant partnerships which have, to date, disclosed over $770M in upfront payments, with a potential future value of at least $38B. In both the 'AI-first' biotech and large pharmaceutical settings, the formation of these partnerships, and in such numbers, underscores the private sector appetite for deploying these technologies at scale. Whilst industry-to-industry partnerships may dominate in HIC settings, one can also envision other types and networks of partnerships, for example between research and industry, that could bring the value of AI to other settings, especially those which are more resource-constrained.

Figure 11 – Adoption of AI in drug discovery, by type of organisation



Avg. Score refers to average adoption score. N refers to number of respondents. Survey Question: "How would you characterise your organisation's usage of AI tools to support the use case family?" Survey Options: 1. Don't use AI 2. Some limited experimentation is occurring, but AI tools aren't routinely used. 3. AI tools used on an ad-hoc basis for specific processes where it is valuable, AI tools aren't a part of workflow. 4. AI tools used commonly for many of the critical processes and are part of workflow. 5. AI tools used as a standard for most of the critical processes and are a standard part of workflow. Note: Respondents may overlap across use cases and can only choose one option per use case, average use case option per respondent was used to determine usage by sector

5.3. Emerging value proofs

While our survey results highlighted the differences in adoption of AI use cases across the industry today, it also highlighted the broad agreement in the potential for AI to transform drug discovery in the next 3-5 years.

As discussed earlier, to determine the potential impact of deploying AI in drug discovery, a high-level analysis was conducted which highlighted three primary value drivers (i) time and cost savings, (ii) increased probability of success, and (iii) novelty of both the molecular target and optimised therapeutic agent [9].

Despite the existence of numerous, publicly reported partnerships, projects and increasingly pipeline assets (see Figure 12 for selected examples), many of the experts we interviewed and surveyed argue that the value proofs of AI in drug discovery so far are point successes. The large number of on-going programs will mature in the coming years to add data points to evaluate the true impact of the technology.

> *"There is undoubtedly a lot of promise for AI in drug and vaccine discovery, but there's also an awful lot of hype – AI has been talked about for the last decade but the challenges facing R&D haven't changed that much yet.*
>
> **Head of AI, Pharmaceutical company**

> *"We need examples to show that AI in drug discovery is broadly applicable, and not just within one particular model or therapeutic.*
>
> **Professor of Bioinformatics, Academia**

On time and cost; many players across the ecosystem are claiming substantially faster timelines and reduced costs (e.g., driven by fewer compounds synthesised), but sceptics have yet to see the externally verifiable impacts [25], [26], although competitions like CACHE will help address this [27]. Early analysis has shown that select AI-partnerships/AI derived assets took on average 4 years to reach the clinic versus 5-7 yrs in benchmarked timelines [28], [29]. COVID-19 efforts in repurposing for small molecule drugs and for antibody discovery also showed dramatic acceleration where AI played a part, albeit in a pandemic-accelerated research and regulatory environment (Figure 12).

On novelty and probability of success, an increasing maturation of proof points in the coming years is expected from the 73 AI-derived clinical pipeline assets identified today across Small Molecules, Antibodies and Vaccines portfolios.

COVID-19 treatments aside, initial indications of the clinical performance of assets are also emerging – from RLY-4008 (selective-FGFR2 reversible inhibitor with a novel mechanism of action) and EVX-01 (peptide-based neoantigen) which both showed superior ORR versus controls in Phase 1/2a studies; to NDI-034858 (selective TYK2 allosteric inhibitor) which read out positive Phase 2b results by Takeda (acquired for $4B from Nimbus Therapeutics) [8], [30], [31] .

These are singular point examples of early clinical successes where we would also expect several failures across these large portfolios. A more thorough analysis in coming years will be needed to truly evaluate the impact of AI on discovery.

84%

Of **current users** expect AI to drive **significant/ transformative** impact in drug discovery over the next 5 years

70%

Of **current non-users** expect AI to drive **significant/ transformative** impact in drug discovery over the next 5 years

Survey Question: "How valuable do you expect the future application of AI to support use case family to be? The term future refers to the next 5 years" Survey Options: 1. None, 2. Limited - AI only makes modest improvements 3. Some – AI gives small incremental impact 4. Significant – AI provides a large positive impact, 5. Transformative – AI helps achieve outcomes that were not possible before. Note: Respondents can only select one option for each use case.

Figure 12 – Selected value-proofs for AI-enabled drug discovery

Structure Activity Relationship Prediction - RLAY-4008 for cholangiocarcinoma

- **Sep '16** — **Research to identify a FGFR2 inhibitor started**
  Relay Therapeutics develops novel FGFR2-specific inhibitor RLY-4008 by analysing the dynamic balance of protein conformations using its AI platform
- *48 months*
- **Sep '20** — **Clinical trials started**
  RLAY-4008 starts Phase I/2 study
- *16 months*
- **Jan '22** — **FDA granted Orphan Drug Designation**
  FDA grants RLY-4008 orphan drug designation for the treatment of cholangiocarcinoma
- *8 months*
- **Sep '22** — **Interim clinical data published**
  88% ORR[1] at phase 2 dose level

Value Drivers

Time: 4 years from Discovery to Clinical Trial
Cost: Not publicly available
PoS: 45-65% higher ORR compared to peers
Novelty: New mechanism of action

AI-enabled Selectivity Improvement- NDI-034858 for moderate-to-severe psoriasis

- **Dec '16** — **Research to identify a candidate for the JH2 domain of TYK2 started**
  Nimbus Therapeutics leverages Schrodinger's FEP+ physics-based model to evaluate 13k molecules and design superior selectivity across the majority of the kinome
- *56 months*
- **Aug '21** — **Clinical trials entered**
  NDI-034858 enters Phase 2b clinical trials; a 75% improvement in skin lesions is observed compared to placebo at 12 weeks
- *16 months*
- **Dec '22** — **Takeda acquires NDI-034858**
  Takeda enters agreement to acquire NDI-034858 from Nimbus Therapeutics for ~$6B

Time: 4.5 years from Discovery to Clinical Trial
Cost: 300-500 molecules synthesised compared to 5-10k in traditional workflows
PoS: 75% improvement in skin lesions and expected superior selectivity
Novelty: Novel allosteric inhibitor of TYK2

Neoantigen prediction- EVX-01 for metastatic melanoma

- **Dec '17** — **Research to identify a neoantigen immunotherapy**
  Evaxion deploys their proprietary AI-based PIONEER platform to identify neoantigens most likely to elicit strong T-cell responses and anti-tumour effects
- *13 months*
- **Jan' 19** — **Clinical trials started**
  EVX-01 starts Phase I study
- *30 months*
- **Jul '21** — **Phase 1 interim data announced**
  EVX-01 demonstrates ORR of 67% in conjunction with anti-PD-1 treatment vs. historical data of 40% with anti-PD-1 treatment alone
- *3 months*
- **Oct '21** — **Merck collaboration announced**
  Enters clinical trial collaboration and supply agreement with Merck to evaluate EV-01 with Merck's anti-PD-1 therapy

Time: 1 year from Discovery to Clinical Trial
Cost: Not publicly available
PoS: ~30% higher ORR compared to peers
Novelty: Personalised medicine approach

1, ORR=Overall response rate
Sources: [32, 33, 8, 34, 35, 30, 36]

Figure 12 – Selected value-proofs for AI-enabled drug discovery (Cont'd)

Drug Repurposing – Baricitinib for severe COVID-19

Feb '20 – Research on existing and approved drugs initiated
Baricitinib, an existing Eli Lilly drug, is hypothesised to be a leading therapeutic for COVID-19 by Benevolent AI's knowledge graph technology

3 months

May '20 – Clinical trials entered
Baricitinib enters control trial with US NIAID

6 months

Nov '20 – FDA EUA granted
FDA grants Baricitinib Emergency Use Authorization (EUA)

5 months

Apr '21 – Trial results obtained
Phase III trial results from Eli Lilly show 38% reduction in mortality, the most significant clinical effect to date

13 months

May '22 – FDA full approval granted
Baricitinib approved for the treatment of COVID-19

Value Drivers

Time: 2 months from Discovery to Clinical Trial
Cost: Not publicly available
PoS: n/a[1]
Novelty: New antiviral mechanism

Antibody Design and Optimisation – Bamlanivimab for mild/moderate COVID-19

Mar '20 – Research to identify a neutralising human monoclonal antibody started
Eli Lilly, in partnership with AbCellera, begins mAb development using convalescent plasma from a COVID-19 patient. Computational methods narrows down ~2000 potential candidates to a final 24, from which bamlanivimab is selected for its 10x-greater neutralisation capacity

3 months

Jun '20 – Clinical trials entered
Bamlanivimab enters Phase II trials

5 months

Nov '20 – FDA EUA granted
Bamlanivimab becomes the first mAb to receive FDA EUA for mild COVID-19

Value Drivers

Time: 3 months from Discovery to Clinical Trial
Cost: Not publicly available
PoS: n/a[1]
Novelty: n/a

> *"It took less than 3 months for Baricitinib to enter clinical trials. There's no way this would've been possible in such a short time frame without AI."*
>
> **CEO, 'AI-first' biotech**

> *"The discovery of bamlanivimab was cruical during the pandemic. I think in the future, AI will play a large role in curing diseases."*
>
> **Vice President, 'AI-first' biotech**

> *"Preliminary clinical data suggests that RLY-4008 is more effective than alternative treatments currently available. AI is helping us innovate and find better solutions."*
>
> **Senior Vice President, 'AI-first' biotech**

1. PoS for COVID examples are not comparable to peers
Sources: [37,40]

BOX 1: Value modelling

To catalogue the potential impact of AI in drug discovery, a value model was built to assess the potential implications from adopting AI technologies, especially for resource-limited settings and/or commercially less attractive diseases.

For simplicity, the model focusses on the time and cost impact of AI within small molecules, where interviews, publications and emerging proof points were used to triangulate potential impacts across the discovery value chain against an adjusted published baseline [29].

Given that drug discovery programs vary enormously in scale and complexity and there is no "one size fits all", 3 scenarios were assessed at a high level:

- Scenario 1: New molecule for a difficult or poorly understood target
- Scenario 2: Molecule from existing chemical series for well understood target
- Scenario 3: Repurposing of existing molecule for a target

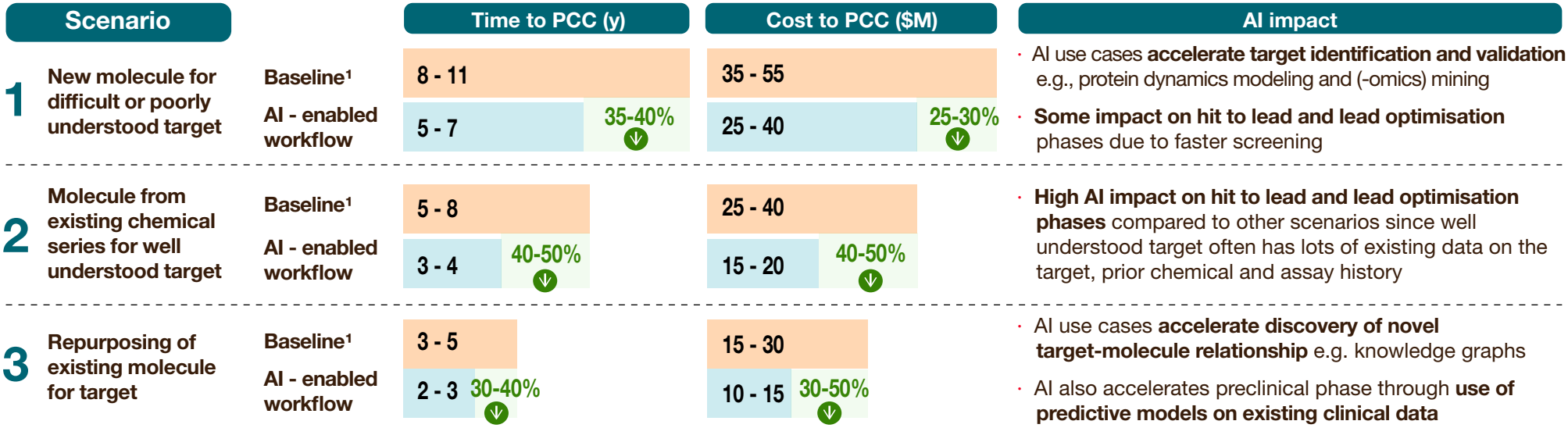
To evaluate the potential of AI across these scenarios impacts were triangulated from available value proofs, publications and expert interviews. See appendix section 10.5 for details of how these scenarios were modelled and the high-level assumptions used.

Across all scenarios, the model suggests that AI has the potential to materially reduce timelines from discovery to the preclinical candidate stage (from 2-3 years in repurposing or with well validated targets), and significantly reduce costs, dropping as low as $10-15M in Scenario 3.

This is driven by faster and better hypothesis generation in target identification and validation for Scenarios 1 and 3, as well as improved molecule optimisation through fewer design-make-test cycles. This latter impact is particularly improved in Scenario 2, when there is greater information on the target, prior chemical and assay history (e.g., kinase inhibitors) that can act as starting points for quicker lead optimisation efforts.

It remains to be seen to what extent these benefits can be realised in drug discovery programmes. However, even if only some of these benefits materialise, it could represent a fundamental reshaping of the economics of discovery. This could allow industry to take more "shots on goal" and thereby increase success rates of discovery programmes. Also, it could enable LMIC discovery teams to pursue preclinical programmes against diseases which are currently economically not feasible.

Figure 13 – Potential impact of AI on time and cost of drug discovery



1. Adjusted based on expert input Source: Publication [29]

6. Key barriers to adoption

Alongside emerging value proofs and increasing appetite to use AI in the future, our interviews and survey identified several barriers that may be limiting current adoption of AI in drug discovery. These barriers can be grouped into 4 dimensions, defined in **Box 2:**

- Trust
- Data
- Tools
- Capabilities.

BOX 2: Barrier definitions

Trust: Trust barriers typically stem from scepticism or a lack of understanding of the potential and maturity of AI tools in drug discovery. This can inhibit investment and adoption of these technologies by practitioners and leaders across the industry. On a smaller scale, it also refers to the lack of trust users feel with AI-derived outputs. Specifically, some AI algorithms are perceived as "black boxes" without a full understanding of how the algorithms work, including their strengths and limitations.

Data: Data barriers refers to the challenges of accessing or using raw datasets required to train and run AI models for specific tasks. Many of the experts we interviewed and surveyed highlighted the need for deep, broad datasets (often at the disease area level depending on the use case). Similarly, being able to access data for model training is critical for developing AI solutions and can vary substantially across data types and industry players. Lastly, some experts also called out a need for common data standards or formats to enable greater interoperability of data sets to allow AI models to run at scale (and not just for bespoke analysis).

Tools: In the context of AI in drug discovery, tool barriers refer to the lack of maturity of tools to derive accurate predictions, and to the accessibility and usability of tools (e.g., open source/licensed or simple interfaces). This also includes ways of deploying AI tools, either through integrations with existing workflow solutions, or through standardised deployment across research settings (e.g., protein families, disease types etc).

Capabilities: Capability barriers refer to the skillsets and infrastructure needed to conduct AI in drug discovery. Typically, this work requires multidisciplinary skillsets which combine deep drug discovery expertise as well as data science and AI capabilities. This can be achieved through cross-functional teaming within institutions, but also through partnerships between different organisations. In terms of infrastructure, access to large-scale computing power, a stable cloud connection, and access to wet labs are usually required. Lastly, education and training for drug discovery practitioners are important for growing AI capabilities.

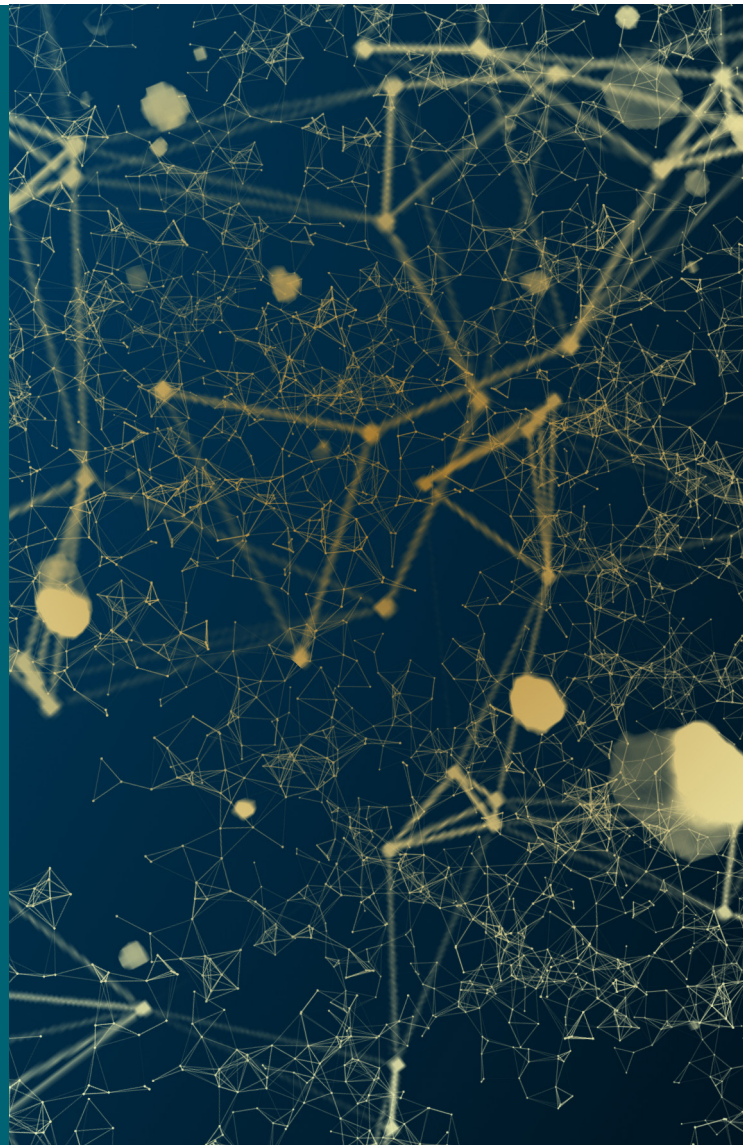


Figure 14 – Overview of barriers limiting the adoption of AI in drug discovery

AI Enablers	Maturity	Access	Standardisation
Data	**Insufficient breadth and depth of data** for specific therapeutic areas, use cases, and population sets	**Fragmented datasets in public domain;** limited access to privately-held proprietary data	**Data collected in inconsistent formats and levels of granularity** across geographies and industries for specific use cases and therapeutic areas
Tools	**Lack of suitable tools** (e.g., user-friendly, energy-efficient) for specific use cases and therapeutic areas	**Lack of access to best-in-class tools,** even in licensed models	**Algorithms not readily interoperable** with existing workflows / other toolkits or **comparable** across diseases or use cases
Capabilities	**Insufficient drug discovery resources** (e.g., labs, computing power) **and relevant AI expertise** in specific geographies and sectors	**Silos of resources and knowledge** within sectors, functions and geographies	**Lack of standardised skillsets** (e.g., training curriculum) **and approaches** used across geographies and sectors

Barriers relating to the lack of trust in AI drug discovery

Our survey and interviews highlighted a wide spectrum of trust and understanding of AI in drug discovery.

Whilst many drug discovery experts believe in the potential of AI in the future, some opinion leaders – including policy makers, funders, and senior academic and industrial leaders – remain sceptical about the near-term impact of AI in drug discovery.

This scepticism can present challenges, especially in LMICs, where there is less familiarity of the field, compounded by the lack of awareness of relevant tools and industry proof points from which to challenge the perception of policy makers.

> *AI is the future – it will help us explore areas that have never been explored before. One day AI will help us understand biology so deeply that we can form new scientific laws and drug design principles.*
>
> **Head of Data and Platform Strategy, 'AI-first' biotech**

> *AI is somewhat valuable. In our work, AI has helped make a lot of molecules synthesisable faster & cheaper.*
>
> **Translational Scientist, Academia**

> *It's too early to recognize the true impact of AI – we will only be able to see the true impact once we can see the productivity over time.*
>
> **Deputy Director, Global Health, Non-profit Organisation**

> *AI is currently only used for solving simple problems. The InSilico screen would only have had a 4% failure rate, even without AI.*
>
> **Computational Biologist, Research Institute**

> *AI is a new hype – investors buy into the desire to be hip, cash is raised, Pharma cos do deals to be in the news. There's lots of noise in this field but it has not been proven yet.*
>
> **Chief Executive, Data Consortium**

Believers ← → **Sceptics**

Barriers relating to data required for AI in drug discovery

Data related barriers were those most frequently cited by experts across all industries and therapeutic areas. These barriers can be broadly categorised into three themes - (i) lack of suitable datasets to feed AI models, (ii) lack of access to proprietary databases, and (iii) limited interoperability of existing datasets.

> *Data needs to be treated and managed like a strategic asset. We have a lot of data in-house but it is not managed well. Our data needs to be cleaned before it can be used to feed AI models.*
>
> **Senior Director, Pharmaceutical company**

i. Lack of suitable datasets to feed AI models

Where open-source datasets exist today, many experts we interviewed highlighted that depth, dimensionality, and scale are often too limited for the application of AI to better characterise diseases (e.g., missing metadata on cell culture conditions; or assay conditions beyond just experimental outcomes). There are increasing efforts to build out these multi-modal datasets. Ochre Bio, for example, are leveraging a deep phenotyping approach for liver disease using multi-omics, imaging and novel translational models using livers unsuitable for transplant. However, this type of approach to understanding disease is highly specific, often requiring patient samples or novel experimental approaches to develop and test hypotheses – and this can create challenges when considering the expansion of these approaches across disease types, or the creation of scalable open-source algorithms.

> *It's important to understand the limitations of AI. In silico predictions are only as good as the in vivo/vitro data they are trained on.*
>
> **Manager, Gene therapy biotech**

ii. Lack of access to proprietary databases

Proprietary datasets, on the other hand, are often sufficiently rich but are typically inaccessible to the broader field. In our interviews and survey, several experts stated that proprietary datasets often contain high-quality data for a given use case (e.g., use cases pertaining to large-scale small molecule synthesis or safety and toxicity use cases), but lack of access to these datasets can significantly hinder the development of tools within academia. As a result, tool developed in an academic setting, using publicly available data, are sometimes less accurate than they could otherwise be. This challenge can be addressed in several ways – Proprietary datasets can be replicated in the public domain; however, this can be costly and time consuming. Alternatively, federated learning approaches can be used on proprietary data. For example, the MELLODDY platform in the small molecules space powered by Owkin's Substra, and LHASA's Effiris model have been deployed to enable learning across proprietary datasets with the end goal of building a more accurate algorithm, though these types of initiatives are currently relatively rare [41], [42].

iii. Limited interoperability of existing datasets

Making multiple datasets interoperable (i.e., being able to use these datasets on different systems, or in combination with other data) is a complex and costly task, particularly for less-experienced teams deploying open-source solutions. Inconsistencies in data structure, metadata and normalisation exist in many databases, and precludes both the easy application of AI techniques, and the amalgamation of data to drive better target-disease hypotheses. For example, the Cancer Genome Atlas Project (TCGA) and the Catalogue of Somatic Mutations in Cancer (COSMIC) databases could provide significant value when used simultaneously, but the lack of consistent data standards renders their integration complex, especially for academics [43].

> *When you think about Infectious Diseases – the strain of the virus, the cellular growth model, the time of data collection – all have a massive impact on the data and these are often values that aren't captured.*
>
> **Vice President, 'AI-first' biotech**

In our interviews and survey, some AI experts mentioned initiatives that are beginning to tackle this need for more standardised and interoperable data through advocating for consistent data standards (e.g., FAIR) or supporting minimum metadata requirements for the publishing of data within a database (e.g., PRIDE). These standards have yet to be universally agreed, implemented, or adopted, but point to a strong appetite from industry players to support greater collaboration on data.

> *Academics are very good at generating data (proteomics, transcriptomics etc.) but it is not yet done consistently in a way that can be used by AI models.*
>
> **Research Scientist, Biotech**

Figure 15 – Principles of FAIR data

Findable

(Meta)data is referenced with unique and consistent identifiers (e.g., DOIs), and **can be discovered by both humans and computers** (e.g., by exposing key words to search engines)

Accessible

Whilst the data does not have to be openly available to everyone, **information on how the data could be retrieved must be available**

Interoperable

Data can be exchanged and used across different applications, systems and geographies using metadata standards, standard ontologies and standard structures

Reusable

(Meta)data is well described such that **it can be replicated or combined with other datasets;** encourages collaboration and avoids duplication of efforts by allowing assessment and validation of results

Source: Website [44]

Barriers relating to AI tools in drug discovery

Limitations of existing AI tools was cited as another hinderance to adoption, with barriers relating to the (i) lack of mature tools (ii) limited tool usability, and (iii) lack of access to relevant tools most frequently mentioned.

i. Lack of mature tools

For tool maturity, experts noted that the development of novel AI tools can often be hindered by an absence of underlying datasets required to train the algorithms. This is particularly pertinent in the infectious disease space where robust tools for predicting immunogenicity of pathogenic proteins is often lacking. For example, several experts cited the need for algorithms to determine the impact of bacterial protein glycosylation on major histocompatibility complex (MHC) binding, but limited large-scale data exists today from which this can be achieved.

> *Data you feed into a model is the most important determinant of outcome and neglected tropical diseases has always been behind other therapeutic areas for data availability. It's frustrating!*
>
> **Infectious Disease Investigator, LMIC**

ii. Limited tool usability

The usability of AI tools in drug discovery, especially open-source tools, was also cited as a major barrier. Interviews highlighted that open-source tools are often not developed with wide-spread use in mind, and in some cases not regularly maintained once developed (often due to cost). As a result, some AI tools lack simple user interfaces, cannot easily integrate into existing workflows, and are rarely updated with new features or bug fixes.

> *We are massively limiting ourselves if these tools can only be used by academics or AI-specialists – students, doctors, industry professionals should all be able to play a part – we can't expect everyone to have a PhD in AI.*
>
> **Head of AI Platforms, 'AI-First' biotech**

> *There is no incentive for academics to make usable open-source tools that are also maintained.*
>
> **Professor of Bioinformatics, Academia**

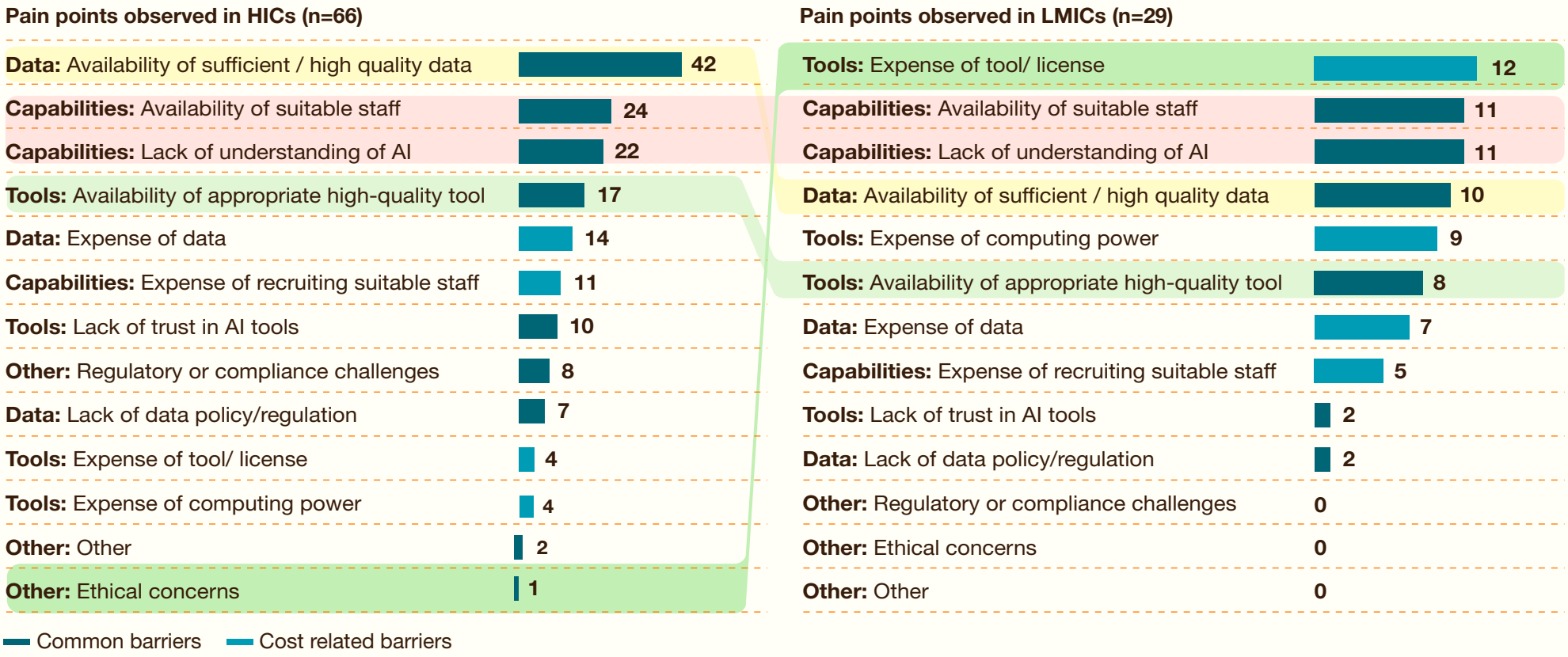
> *Lots of 'AI-first' biotechs are producing tools but they are not accessible; they are like a black box – we put data in, they give us data back – but without understanding what the tool is doing, it's difficult to convince anyone to trust the findings.*
>
> **Head of Drug and Vaccine Discovery, Academia**

Whilst industry players can address the challenges of open-source tools – by either purchasing commercial tools or deploying internal teams to build upon existing open-source tools – academics, particularly those from LMIC settings, often struggle to overcome these challenges. In our interviews, many LMIC researchers highlighted a particular need for small molecule drug discovery programs in

infectious disease or natural product identification. Open-source tools for both these areas exist, albeit at different levels of maturity [45], [46]. Commercial tools also exist, however, the high costs of licensing these tools often present a major barrier. This was substantiated from our survey analysis, where over 40% of LMIC survey respondents cited cost of tools as their main barrier to adoption (see figure 16).

> *"Licenses deemed "good value" in the western world, even for commonplace tools such as the Microsoft Office suite, are often completely unaffordable to us.*
>
> **Head of AntiMicrobial Agents Unit, LMIC**

Figure 16 – Specific barriers to the adoption of AI, observed by HICs and LMICs



Survey Question: "What are the current barriers to increasing the usage of AI tools in drug discovery?"
Survey Options: Shown above
Note: Respondents can select up to 3 options

iii. Lack of access to relevant tools

Finally, our analysis shows that challenges in accessing tools can also hinder adoption.

To date, much of AI in drug discovery has been driven by the private sector, with investors and pharmaceutical companies funding 'AI-first' biotechs who apply AI most extensively. As a result, most cutting-edge AI algorithms, tools and databases are patented or otherwise protected, and sometimes unavailable to a broad community of drug discovery researchers.

In our survey and interviews, some experts expressed a concern that extensive patenting and IP protections could lead to "intellectual property lock up" of the AI drug discovery space. This is being compounded by changes in the business models of many 'AI-first' biotechs, away from software or fee-for-service model towards proprietary pipeline development (See figure 17). This shift in turn heightens the need to patent or otherwise protect intellectual property, and disincentivises provision and support for widely used external software solutions.

As with other barriers, "IP lock up" disproportionately affects research academia and LMICs who often struggle to pay licensing or access fees for proprietary tools (as mentioned above) or have the capability networks to deploy/develop open-source tools.

In section 7 of this report, we discuss how cross-industry initiatives could help address this challenge.

> *"Much of what works and goes on in pharmaceutical companies is behind walls. Knowledge, data, and tools are locked up.*
>
> **Senior Vice President, 'AI-first' biotech**

Figure 17 – Spectrum of business models observed in 'AI-first' biotech companies

Providing AI platform solutions as fee-for-service

Whilst some software players intend to play only in this space, some **'AI-first' biotech companies begin as fee-for-service to allow validation of models** and begin proprietary data generation

Employing a mixed fee-for-service plus pipeline model

Once models are validated, 'AI-first' biotechs employ a mixed model where they in-licence assets, to **begin significant revenue generation**

Using AI solutions to create own pipeline

As capabilities and models improve, 'AI-first' biotechs mature to create their own pipeline, which is further **driven by investors valuing pipeline assets over proprietary tech platforms**

Barriers relating to capabilities

Capability-specific barriers are mostly related to the (i) lack of expertise needed to develop and deploy AI tools (ii) lack of required training to use existing tools and (iii) lack of infrastructure necessary to support AI in drug discovery efforts. The severity of these barriers varies by geographies and sectors, with the last two most applicable to LMIC settings.

i. Lack of required expertise to develop and deploy AI tools

Both the development and deployment of AI tools in drug discovery requires a combination of technical drug discovery experience (e.g., structural biology, medicinal chemistry etc.), and data science/ data engineering expertise. Experts repeatedly highlighted the need to establish these multi-disciplinary teams internally or through collaborations to truly embed AI techniques into drug discovery workflows.

Both academics and industry leaders reported a lack of relevant subject matter expertise. They also highlighted significant challenges in the formation of multidisciplinary teams, such as experts being unable to "speak the same language", and difficulties working across organisational siloes to foster collaboration between teams. Furthermore, they noted that as large organisations adjust processes and workflows to embed AI, these challenges will likely be exacerbated. Thus, increased training and skill building will be required across both AI and drug discovery experts, not only to support the capability gap today but also to support interdisciplinary teaming in the future [47].

> *Hiring talent at the intersection of biology and AI is very challenging. There is a chronic shortage of people with the right skillsets for using AI in drug discovery. And we struggle to keep those with the right skillset – we are competing not only with biotechs but also the likes of Google. We can't compete with better paid industries.*
>
> **Head of Drug and Vaccine Discovery, Academia**

> *There's a lack of communication and understanding between biologists and computer scientists – people think in silos.*
>
> **Organic Chemistry Professor, LMIC**

ii. Lack of required training to use existing tools

At present, drug discovery and computational science research within some LMICs is not yet fully established. As a result, it is often challenging to find talent with relevant experience [48], [49], [50]. Furthermore, in our interviews, many researchers in LMICs highlighted the lack of structured trainings on the use of AI tools, and the lack of formal university training programs on AI fundamentals, let alone on its application to drug discovery [51], which further hinders the adoption of AI in LMICs

> *Pharmaceutical companies and 'AI-first' biotechs will use the skillsets they have at their disposal to accelerate, but LMICs still have a long way to go to catch up.*
>
> **Head of Drug Discovery, Non-profit Organisation**

iii. Lack of required infrastructure required to support AI in drug discovery efforts

In addition to a strong talent pool, computational infrastructure and experimental set-ups are also required to support models and validate outputs from AI. Within LMICs, this infrastructure is often missing, ranging from basic technology infrastructure, such as lack of stable cloud access and compute power to build and run models smoothly, to more drug discovery specific capabilities such as the lack of supporting wet lab capabilities to test model hypotheses.

> *Building infrastructure for AI in drug discovery is more than just providing computers – in some of these places even accessing reagents is a struggle.*
>
> **Senior Program Officer, Global Health Non-profit Organisation**

Conclusion: barriers limiting adoption of AI in drug discovery

Overall, AI in drug discovery has seen uneven rates of progress across use cases, therapeutic areas, and settings often driven by a patchwork of available data, tools, and capabilities. Significant opportunities remain to support the development and adoption of AI technologies to discover new medicines, particularly for under-served diseases. The next section explores some of the emerging solutions across the use cases in focus.

7. Potential solutions to drive adoption

Given the exciting potential of AI in drug discovery, but also the barriers and challenges outlined above, we see a need for broad initiatives to drive adoption of AI in drug discovery across sectors, geographies and therapeutic areas.

These initiatives could help to build trust, enrich datasets, and foster the development of new tools and capabilities. Figure 18 gives an overview of these potential initiatives.

Figure 18 – Potential initiatives for addressing barriers limiting the adoption of AI in drug discovery

Non-exhaustive

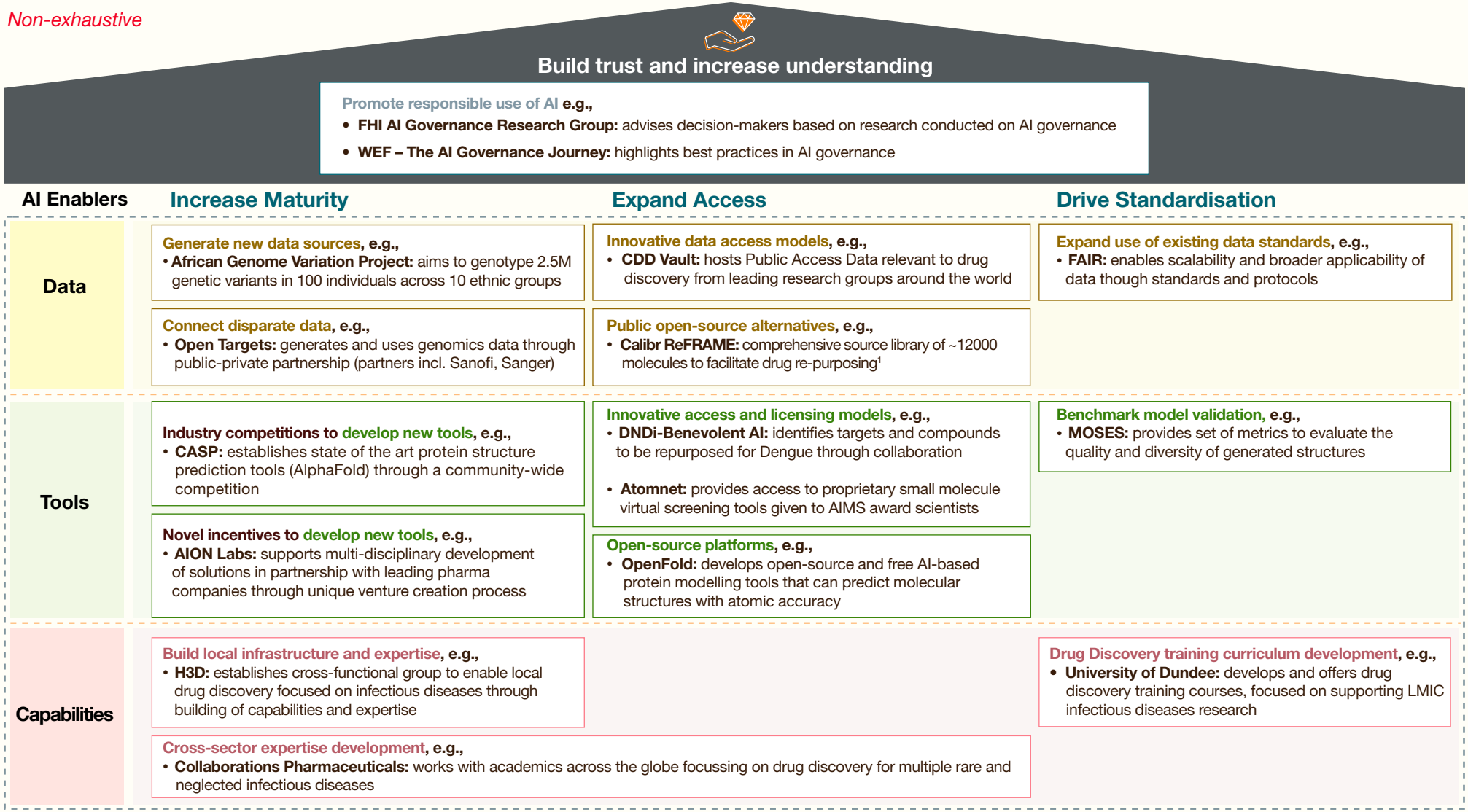
Build trust and increase understanding

Catalogue successes and failures of derived assets to **build objective value proofs**	**Demonstrate performance and breadth of AI-tools** in the field of drug discovery	**Transparently communicate the limitations** of tools that are newly developed or available	**Promote responsible use** of AI in drug discovery

AI Enablers	Increase Maturity	Expand Access	Drive Standardisation
Data	**Enrich depth and breath of existing data** sets (e.g., expand quantity, add metadata) **Generate new data sources** to fill high priority gaps (e.g., for disease understanding)	Broaden **innovative data access models** for private data including federated models, public access initiatives Generate **public open-source alternatives** to proprietary data, e.g., libraries, repertoires	**Expand use of existing data standards** and protocols for AI data quality and interoperability (e.g., FAIR) **Encourage development of advanced data standards** for specific use-cases or modalities
Tools	**Develop new tools** (e.g., generative design) to address frontier problems via industry competitions, novel incentive structures etc. **Improve tool quality and applicability** for existing tools across the drug discovery ecosystem	Propagate and maintain **open-source platforms** that are user-friendly and low-code Support **innovative access and licensing models** for proprietary tools and platforms	Establish best practices for **benchmark model validation** and quality assurance of tools Align on **interoperability standards** for tools
Capabilities	**Build local research capacity and infrastructure**, i.e. labs, materials and computing resources Foster **expertise development**, knowledge sharing and talent retention	Encourage **collaboration models** for sharing capabilities across regions **Create peer-to-peer networks** of researchers across global community	Develop and implement standard **drug discovery training curriculum** Expand standardised **knowledge-sharing forums** for wider community exchange

Figure 19 – Examples of solutions already underway today to address barriers to adopting AI in drug discovery

Non-exhaustive



1. Library created by combining three databases (Clarivate Integrity, GVK Excelra GoStar, and Citeline Pharmaprojects)

Several initiatives to address these barriers are already underway, such as those creating novel training data or tools, expanding application of existing technology to under-served diseases or building capabilities in LMIC settings (see Figure 19).

However there remains an opportunity to scale or broaden initiatives across all disease areas.

BOX 3: Examples of existing initiatives

Human Immunome Project

Human Immunome Project

A global non-profit aiming to create the first AI model of the human immune system to enable faster, cheaper, and more effective development of vaccines and treatments

Description

The consortium **brings together leaders across industry, academia, governments and non-profits to compile the biggest dataset of biomedicine at a population scale, and fund, develop and advance key scientific goals.** Collaborators include GSK, Moderna, Illumina, National Institutes of Health and Harvard School of Public Health.

A set of key initiatives and partnerships have been established to **focus on those most at risk of disease, and the hardest to protect** – such as the elderly, those most susceptible to antimicrobial resistance, and those living in developing countries – as well as **setting themselves up to tackle pandemic preparedness in the future.**

Open Molecular Software Foundation

Open Molecular Software Foundation

A non-profit organisation focused on building open source software and communities for molecular sciences.

Description

Develop cutting-edge **open-source AI-based protein modelling tools** that can predict molecular structures with atomic accuracy.

Software released under **permissive license** enables both academic and industry researchers globally to use, validate, and improve the tools.

Complete training, inference stack and training datasets are also shared under the permissive license.

Additionally, they develop packages that **maintain and increase interoperability of existing free energy method tools.**

CACHE CHALLENGE

Critical Assessment of Computational Hit-finding Experiments (CACHE)

A series of competitions targeted towards computational chemists and scientists from the drug discovery field aiming to define the cutting edge of molecular design of small molecule therapeutics, similar to the Critical Assessment of protein Structure Prediction (CASP) competition

Description

Sponsored competitions that focus on specific protein targets of biological or pharmaceutical relevance e.g., predicting hits for a specific domain within a Parkinson's disease target molecule, or identifying ligands targeting components of SARS-CoV2.

Participants **use computational algorithms to predict hits; algorithms are then tested experimentally by CACHE**, and **all data and chemical structure information are released publicly.**

CACHE allows **for comparison of tools, and creation of additional data.**

Source: Website [21], [27], [52]

To truly unlock the potential of AI in drug discovery, our analysis identified a number of opportunities across the areas, as described in Figure 18.

Whilst in most cases, solutions will differ by use case, there is also an overarching need to build broader trust in AI, and the value it could deliver to the drug discovery field.

There are many paths by which this can be attained, with transparency being fundamental to help cut through the hype in the field today. Initiatives could include cataloguing the successes and failures of AI-derived assets, demonstrating tool performance and breadth (the CASP and CACHE competitions are already doing this), and transparently communicating the limitations of tools that are newly developed or available [27], [53]. Examples of the latter can be seen in OpenAI's publication of the outcomes of red-team testing of GPT-4, or AlphaFold's publication of a per-residue confidence score in its protein structure predictions [54], [55].

> *Serious uptake requires fully transparent real-world examples where AI has successfully delivered. For AI efforts on fundamental aspects of R&D, quality has to be 100% unquestionable (e.g., no one running a R&D project is going to take toxicity guidance from AI unless AI can provide hundreds of examples of being right and zero examples of being wrong)*
>
> **Director, Life Science Venture Capital**

> *Companies need to be transparent about what they actually can do, only then we can see the value; but for now, we can't be believers*
>
> **Chief Executive, Data Consortium**

> *There is not going to be a silver bullet to build trust; a myriad of approaches will be needed*
>
> **Professor of Bioinformatics, Academia**

Together, initiatives such as these helps clarify the potential value at stake, as well as help users better understand the relative strengths of tools, and how to best deploy them.

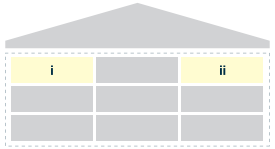
Understanding Disease: Solutions to drive adoption of AI

Whilst understanding disease use cases have seen considerable research efforts to date (as described in section 5.1), furthering our understanding requires the availability of large, interoperable datasets from which AI techniques can triangulate datapoints to identify and validate molecular drivers of disease.

Thus, to advance the field and broaden the applicability of AI solutions to underserved diseases, there are two key opportunities to improve data access and integration:

i. Enriching or combining datasets to better understand molecular drivers of disease

ii. Enabling data standardisation and facilitating sharing



> *The value AI will bring in understanding disease is in the linking of datasets to build hypotheses based on far more data than any human could even fathom synthesising*
>
> **COO, 'AI-first' biotech**

i. Enriching or combining datasets to better understand molecular drivers of disease

AI algorithms used for understanding disease typically look for connections between clinical data, molecular mechanisms and biological pathways, either for repurposing efforts or for finding new targets. This usually involves combining different types of disease specific data (multi-omics data, patient data, treatment history etc). Examples include precision medicine approaches for oncology and immunology across both academia and the private sector (e.g. Tempus, Immunai, OWKIN, TCGA) [56]–[59]. The main challenges in building these data sets are fragmentation of datasets, especially real-world data, and inconsistent data availability, especially diagnostic data.

An example of where data presents a critical challenge today is Mental Health. Inconsistent coding of mental health diagnoses, and a lack of rich molecular diagnostic data hinders the application of AI in drug discovery efforts.

To address these challenges, targeted efforts are required to improve data maturity through the generation of new datasets, or via enrichment of datasets that might already exist. For example, longitudinal population study data can be enriched with Mental Health-specific data, such as diagnoses and symptomatology, to help identify patient phenotypes. For some diseases, these efforts have already begun to draw a closer link to molecular drivers for disease precision medicine. For example, real-world-data players such as Tempus are expanding to pharmacogenetic testing of neurology and mental health patients, and models such as DRIAD are aiming to quantify the association between early stage Alzheimer's and biological processes that can be defined by a set of genes [60], [61].

> *In Mental Health, there is a lot that needs to happen first at the disease understanding level outside of AI. For example, translatable models, better defined clinical end points and better molecular diagnostics*
>
> **Vice President, 'AI-first' biotech**

> *Animal models for mental health are terrible; cellular models are limited. The omnigenic nature and confounders of mental health make it very difficult to have model systems*
>
> **CEO, 'AI-first' biotech**

ii. Enabling data standardisation and facilitating sharing

The ability to share data is critical for understanding diseases with AI, including in vitro assay data (e.g., experimental conditions, cell lines), clinical trial data, and real-world evidence. To achieve this, many experts we interviewed and surveyed have highlighted the need for greater data standardisation.

For some diseases, this is already beginning to happen. For example, in infectious diseases, the Poolbeg-CytoReason partnership standardises and analyses clinical data from their influenza and RSV challenge trials to identify drug targets for the treatment of these diseases [62]. Examples such as this highlight the potential for AI applications in the infectious disease space should greater data sharing become possible within clinical settings.

Whilst approaches such as FAIR data (as mentioned above) can lay out the principles for data sharing and interoperability, driving adoption of concrete and specific data standards is likely to require considerably more effort – namely, buy-in from leading databanks, as well as mandates from academic institutions, publishers and funders to encourage broad adoption of data standards.

> *To tackle the problems we are facing as a field today, we need to first sort the deeply unfashionable areas such as data structure and standardisation – we need to walk before we can run*
>
> **Senior Vice President, 'AI-first' biotech**

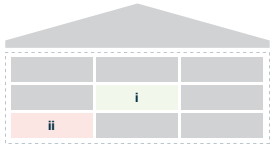
Small Molecules: Solutions to drive adoption of AI

For small molecule discovery, the use of AI is more mature, with relatively accessible data, and open source and proprietary tools available.

Our expert interviews and survey have identified two opportunities:

i. Improving access of established tools and platforms to underserved diseases

ii. Supporting the deployment of existing open-source tools across the ecosystem, particularly in LMICs



i. Improving access to established AI tools and platforms, for small molecule discovery in currently underserved diseases

Expanding access to open-source and proprietary tools (i.e., through licensing and partnership) could change the economics of drug discovery, as described in Section 5.3 Box 1.

Regarding licensing, many companies are already offering their platforms in a fee-for-service model. To expand access, some companies have started providing licenses to their platforms for free – most often to academic groups looking for solutions for an unaddressed health need, or to those operating in underserved disease areas. Atomwise, for example, have opened their virtual screening platform to scientists involved in their AIMS award programme [63].

> *We wouldn't ordinarily be looking at diseases such as Malaria (due to commercial viability for an early-stage start-up), but under the umbrella of a targeted grant, we are of course more than happy to use our platform for these types of diseases*
>
> **Vice President 'AI-first' biotech**

In addition, there are several major partnerships between pharmaceutical companies, Global Health and academic centres. In these partnerships, access to proprietary AI platforms is granted in a more targeted manner. Within Global Health particularly, this often takes the form of funders providing specific investment to enable the deployment of a proprietary tool within a disease area of interest to them, such as the Gates-Exscientia collaboration in Malaria and anti-viral discovery efforts [64], [65].

However, our expert interviews highlighted that opening access to AI tools in itself is often not sufficient. To truly drive adoption, developers of open-source tools should focus on tool usability via simple user interfaces and easy maintenance, especially for use in settings where technical skills are less readily available.

ii. Supporting the deployment of open-source tools across the ecosystem, particularly in LMICs

Many experts we interviewed mentioned that open source tools theoretically could be deployed in LMIC settings. However, this often requires substantial education, training and experience-sharing across centres of excellence [66].

Regarding training and education, we see a global need for improved education on the AI tools available today. Some of this is already happening e.g., webinars on the REINVENT platform open sourced by AstraZeneca. Further efforts are likely required, with Global Health funders, industry players, and local trailblazing institutions driving maximum impact here.

Regarding experience-sharing and networking, we see opportunities to foster more partnerships across LMICs – to train and upskill researchers, share capabilities and infrastructure (including cloud access, computing infrastructure) or collaborate in multi-disciplinary teams [67].

> *Currently you need a certain level of expertise to use open-source tools because a lot of the tools out there are not user friendly – there is no incentive to maintain user friendly platforms*
>
> **CEO, AI Software Non-profit Organisation**

Notable examples include the H3D-Ersilia collaboration, which provided a fully sponsored four-day training in AI drug discovery to African-based researchers working in infectious disease; and Zindi, an online platform and community of data scientists which provides online training in AI, runs competitions to develop and validate models, and encourages users to connect to openly discuss models, share feedback and ask questions [68]–[70].

> *We have established collaboration to implement fully functional AI tools in our institution. We anticipate that extensive future use will lead to success stories to tell and share with other groups in Africa*
>
> **Natural Products Professor, LMIC**

> *We must increase hands-on training workshops in LMIC on the use of AI tools, especially for underexplored research areas such as natural products*
>
> **Infectious Diseases PhD student, LMIC**

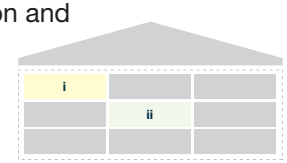
Vaccines: Solutions to drive adoption of AI

As discussed above, the use of AI in vaccines discovery is more nascent than in other areas, and the AI technology (e.g., for mRNA vaccines) typically exists in small number of industrial companies.

We therefore see two main solutions that could drive greater adoption of AI in vaccines.

i. Supporting data collection and sharing for AI in vaccine discovery

ii. Fostering improved access and deployment of AI technologies to underserved diseases



i. Supporting data collection and sharing for AI in vaccine discovery

As discussed above, one of the great challenges of vaccine discovery today is the often limited understanding of how pathogens interact with the human immune system. AI is well-placed to help address these challenges but requires large datasets to train algorithms e.g., data on pathogen structures or the impact of post-translational modifications on immune response. Our expert interviews suggest that systematic data collection and data sharing could boost the use of AI in vaccines discovery.

> *Neglected disease is always behind other therapeutic areas for data availability and, in turn, we also have far fewer AI tools available at our disposal*
>
> **Organic Chemistry Professor, LMIC**

Whilst a number of examples of AI-based open-source immunogenicity prediction models do exist (e.g., TRAP, a deep learning platform that predicts CD8+ T-cell recognition of MHC-I presented pathogenic peptides), the use of these models is not yet fully established [71]. Our interviews highlighted the importance of building relevant datasets to support model training through data sharing and other global initiatives to improve our current understanding of immunogenicity. For molecular data, generation will likely need to be driven by researchers in the field (e.g., bacterial protein characterisation); whilst for patient data, a combined effort between academia and industry may be most effective [72].

Private endeavours aimed at building internal data sets and developing novel AI approaches are also underway (e.g. Evaxion's PIONEER and RAVEN models based on patient genomics data from samples of healthy and tumour tissues) [73]. In addition, there are public-private partnerships to bring different stakeholders together to address these challenges (e.g., Human Immunome Project, Box 3).

For mRNA vaccines specifically, more data is required to improve the tools for designing mRNA constructs and optimising delivery. As discussed above, mRNA research has so far been driven by the private sector through COVID-19. For greater applications in academia, access to some of these data and tools would be helpful in further accelerating the field.

ii. Fostering improved access and deployment of AI technologies to underserved diseases

Given that most of the AI-based tools for vaccine discovery reside in industrial companies, collaboration to ensure broader access is critical.

Some vaccine companies with established platforms are already developing extensive pipelines against global health diseases (E.g. Moderna have two assets in the clinic for Zika and Nipah [74]). Efforts are already underway to improve access to these platforms for neglected diseases.

To date, the most notable of these is the Moderna Access program which allows partners to rapidly test parallel antigen design for priority pathogens in preclinical testing [75]. Other industry players and funders alike could take this as inspiration when considering how they could look to support drug discovery for further underserved disease areas.

Antibodies: Solutions to drive adoption of AI

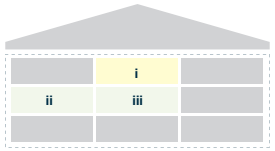
Antibody AI discovery efforts are increasingly gaining traction in industry and through public-private solutions. Whilst there are more and more AI-derived antibodies, our analysis suggests opportunities to improve the maturity of tools (especially open-source solutions) today and increase the breadth of applications beyond oncology and COVID-19.

Key opportunities include:

i. Improving data access to support new AI tools for antibody discovery

ii. Improving usability, validation, and deployment of existing AI tools for antibody discovery

iii. Broadening applicability of AI tools, particularly for infectious diseases



i. Improving data access to support tool development

Antibody discovery efforts have historically been driven by 'AI-first' biotech and pharmaceutical companies, with much of the relevant data e.g., sequence-to-structure or sequence-to-property relationships not widely accessible. This is particularly true for tools that support *de novo* antibody design or multi-property prediction and optimisation (e.g., across solubility, aggregation, immunogenicity etc.) although some point examples of open-source tools do exist, such as Rosetta [76].

Public efforts have started to catalogue relevant data, resulting in tools that support humanisation prediction (e.g., BioPhi) and structure prediction (e.g., ABlooper) for example. However, greater coordination between different organisations could be beneficial particularly where internal data alone could be limiting (e.g. developability parameters) [77], [78].

> *"Traditional drug discovery processes, particularly with antibodies, don't gather data in a way that is useful for AI, so datasets have to be built from scratch.*
>
> **C-Suite, 'AI-first' biotech**

ii. Improving usability, validation, and deployment of existing AI tools for antibody discovery

Open-source AI tools do exist today, most notably in antibody structure prediction (e.g., ABlooper, IGfold) or binding use cases (e.g., Alphafold Multimer) [77], [79], [80].

However, in our interviews many practitioners highlighted that these tools can be slow (>20 minutes to run) which limits their application in early discovery campaign, where large libraries of sequences typically need to be analysed. Thus, initiatives such as those to improve running times of tools, employ non-code user interfaces, and reduce required compute power to deploy the tool could be helpful to increase adoption. For example, the Baker Lab at the Institute for Protein Design aims to predict protein–protein docking and design new protein structures through their Rosetta@Home initiative which allows volunteers to donate CPU capacity by running the software on their computers [81].

Additionally, as with small molecules, better external comparison and validation of open-source tools would help users identify optimal tools to use.

iii. Broadening applicability of AI tools, particularly for infectious diseases

So far, the most advanced examples of using AI in antibody discovery come from industry, especially 'AI first' biotech companies. To increase adoption, we see the opportunity for greater public-private partnerships in AI-antibody discovery to apply this technology more broadly to underserved diseases.

One such example is Abcellera's partnerships with the Bill & Melinda Gates Foundation in 2017 and 2019, which supported the discovery of antibodies for diagnostic testing in Mycobacterium tuberculosis infection, and more recently supported researchers in HIV, malaria, and TB [82], [83].

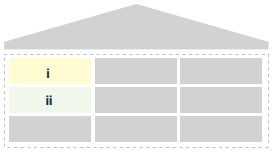
Safety and Toxicity: Solutions to drive adoption of AI

While Safety and Toxicity continues to be a key area of opportunity for AI, academic and private efforts to develop algorithms and tools have been limited to date.

Our analysis suggested two solutions which could unlock progress within this field:

i. Sharing proprietary preclinical and clinical datasets to better understand and model safety signals

ii. Experimenting with innovative preclinical approaches (experimental and AI-derived)



i. Sharing proprietary preclinical and clinical datasets to better understand and model safety signals

In our interviews, experts highlighted the challenges of predicting safety and toxicity based on experimental data with insufficient supporting clinical validation. Combining different data sources and collaborating is therefore critical.

A number of organisations are already trying to address these challenges. One such example is OMEC.AI, founded by AION labs as a result of a challenge set to identify an AI-based system to predict safety risks. This collaborative project aims to amalgamate historical pre-clinical data on a drug of interest and deliver potential safety liabilities that may have been initially overlooked [84].
For diseases where existing data can be challenging to standardise and access, our analysis indicates that greater data sharing would be beneficial.

For example, networks such as Datacelerate are providing a platform on which partner companies can upload and merge deidentified research and development data types, including preclinical toxicology [85]. This network currently has 4,500 collaborating experts across 20 member companies and presents an interesting model for data sharing that could provide the bedrock for advances in this field. Going forward, these efforts could include data sets for a broad range of disease areas, particularly neglected diseases.

ii. Experimenting with innovative preclinical approaches (experimental and AI-derived)

The regulatory framework for preclinical testing is evolving, with recent changes such as the FDA Modernisation Act 2.0 encouraging a shift away from animal models and towards in silico and novel experimental predictive models [16].

As a result, we see increasing interest from different organisations to explore novel mechanisms of predicting safety and toxicity, moving away from in vivo models to novel experimental methods (e.g., single-cell experiments, organoids etc) or in silico predictive models (e.g., quantitative systems pharmcology, QSP). Over time, this is likely to increase adoption of AI in the safety and toxicity space.

> "*AI in predictive toxicology may work well in the development sphere, but international regulations (e.g., WHO and ICH) will still require appropriate in vivo modelling. Cell-based toxicity may be an accepted alternative in some cases. AI may provide a rationale for the selection of species or what to look for, but at this time, there would be insufficient regulatory acceptance of safety and/or toxicity based solely on AI prediction*
>
> **Scientist, Global Health Non-profit**

Conclusion: Driving adoption of AI in drug discovery

AI in drug discovery is at an exciting inflection point. Not only is there a huge amount of dynamism across solutions – with initiatives already spanning data generation, tool optimisation, pan-industry, and pan-geography collaborations – but there are also a range of further tangible opportunities for funders and key players within the broader drug discovery ecosystem to make a marked impact on a host of human health challenges.

8. Call to Action for Funders

The findings of this report suggest that all funders of health research – from basic science to translation and product development – stand to benefit from unlocking the potential of AI in drug discovery. Some research funders, especially those focused on product development, may see immediate and clear links to their strategy. Other funders, especially those more upstream in basic science are likely to increasingly see AI as a key tool in the pathway to impact from the research they fund. Funders collectively will have enormous influence in how this field develops over the coming years, and if funder efforts help overcome access restrictions, how and when AI delivers on the potential outlined in this report.

Our analysis suggests six key actions for funders to take now.

1. Find value from AI today

Based on the landscaping described in this report, drug discovery efforts focused on small molecule discovery and optimisation could immediately benefit from applying AI to accelerate current efforts and programmes. This could be through deployment of open-source tools, such as those for virtual screening or *de novo* drug design (e.g., VirtualFlow, REINVENT 2.0), or through seeking partnerships with 'AI-first' biotechs with well-known small molecule platforms such as Benevolent AI, Exscientia, or Recursion. Funders supporting efforts focused on antibody and vaccine discovery may be able to find value from AI use cases today, but these are more likely to require public-private partnerships given the paucity of open-source tools for these modalities.

Funders may also find value today from applying AI to target identification in data-rich therapeutic areas such as oncology and immunology.

Funders can review their current portfolio and latest funding calls to identify efforts that could find value from AI today and engage pro-actively with investigators to review and re-tool discovery efforts as appropriate.

2. Take no-regret moves to maximise future value

Almost all research efforts will generate data that can have future value in training AI models to support drug discovery. Interviews and survey responses in this landscaping have highlighted how factors such as variable access to research data, data quality, structure and presence of metadata can have an enormous impact on how valuable a research effort is to training AI models. Trends described in this report suggest that ability to incorporate research data into AI models will be an increasingly important part of how that research can deliver impact in future drug discovery efforts.

Funders can help future-proof investments by specifying requirements in grants for the need to (i) publish data in open-access repositories that support APIs for future linkage to AI tools (ii) optimise machine-readability of published data (iii) publish relevant metadata to support interpretation. These actions will have the most impact when taken in concert with other funders (see below)

Funders can take stocks of their current data access and publication policies and make updates to ensure research data is maximally useful for emerging AI applications.

3. Build coalitions to shape the 'rules of the road'

Scale is critical for delivering value from AI – funders acting in concert will be critical to realise the potential described in this report, especially in less data-rich therapeutic areas and less mature use cases. Where standards exist and are well adopted e.g., in areas such as protein structures, genomics and medicinal chemistry, AI has been able to rapidly deliver value. Beyond these areas, landscaping has identified lack of standardisation as a critical barrier to scale and impact.

Funders have a critical role in building the 'rules of the road' in data, tools and capabilities described in this report. An initial set of actions could include (i) agreeing a common minimum standard for publishing research data to support future AI applications (as above), similar to IDDO's creation of a data platform for collation and standardisation of individual patient data from clinical studies in Chagas disease [86] (ii) requiring and supporting grantees to ensure ongoing maintenance of any open-source AI tools (iii) expanding access to existing training and development programmes to data scientists and related AI disciplines.

New measures to assess and compare the quality of new AI tools will be critical for funders to navigate this space. Funders can look to support efforts to transparently benchmark AI tools within specific use cases to build greater understanding of their capabilities and limitations and also measure the impact of their funding. This type of benchmarking could be achieved through competitions akin to CASP and CACHE.

Funders can identify and convene potential partners across the public, private and philanthropic space to build coalitions to support a common goal of maximising the value of AI in drug discovery.

4. Invest where AI intersects with drug discovery goals

Funders focused on drug discovery may see value from AI today (as above) and material investments in AI approaches could be appropriate immediately to support their drug discovery goals. Funders less focused on drug discovery, or in therapeutic areas and modalities where AI is less mature, will need to critically assess where AI most closely intersects with their goals and what their role should be to develop data

and use cases to the point of being able to deliver value from AI.

For example, funders in data-poor therapeutic areas, such as infectious diseases and mental health, may see value in supporting the foundational datasets that are essential for training AI models. These foundational efforts may typically suit large-scale funders able to make multi-year investments and comfortable with a time horizon of 5-10+ years to impact. An example of this type of foundational data set build is efforts to digitally map the entire Drosophila brain 'the fly connectome'- funded by Howard Hughes Medical Institute, Wellcome and the UK Medical Research Council amongst others.

Funders focused on topics where large-scale data generation isn't feasible for financial, practical, or ethical reasons may see value in targeted investments to improve the AI-readiness of historical research data. This could apply to research into neglected tropical diseases, rare diseases or pathogens with high pandemic potential. An example of funder support for this type of activity is the US National Institute of Health where grant recipients can apply for additional funding to support a variety of activities such as, but not limited to; cleaning and filtering data, preparing data for multi-modal multi-scale AI applications, developing, and sharing documentation to highlight recommended uses for the data [25]. Funders should also be aware that the boundaries of feasibility are changing rapidly with new AI tools being developed to work with sparse or lower quality data.

Funders supporting drug discovery efforts in LMICs may see value from building the local capabilities needed to deploy AI techniques which may be under-developed currently. These include computational chemistry, bioinformatics and access to computing power and re-agents for confirmatory testing.

Identifying the most appropriate and impactful investments will also require funders to maintain strong awareness of where the private sector and academia is likely to make progress, and also areas where progress is slower than ideal. This report highlights the rapid pace but also the heterogeneity of progress - requiring funders to maintain close links across the ecosystem and to tailor their approach over time.

Funders can start identifying where AI can help them deliver their drug discovery goals and consider targeted investments where they see a pathway to impact

5. Contribute to the public debate

AI has rapidly emerged as an 'all-of-society' topic with national and international bodies responding with regulatory and legal instruments as they grapple with the potential implications. An example of this is the EU's Regulatory Framework and Coordinated Plan on AI that could enter force later in 2023.

Interviews and surveys as part of this landscaping suggest that funders looking to harness the potential of AI in drug discovery are likely to rapidly feel the implications of these debates in diverse areas, such as, perceptions of their governing bodies and peer reviewers to AI investments, and ability to move or access research data across borders.

Contributing to, and helping shape, the public debate on AI will be critical for funders and may include (i) transparency on AI-related activities and their outcomes (positive and negative) akin to the Royal Society conference on AI in drug discovery (ii) advocating for the use of AI and enablers to AI (such as equitable data sharing) where clear value proof exists.

Like all AI models, those in drug discovery are likely to contain biases based on the makeup of training data and how this is, or isn't, representative of the wider population. Funders will play a key role to build trust in the field, through better understanding and tackling of these biases.

Funders can be a key voice to make the case for AI as part of the wider public debate and can make positive contributions to public understanding and trust

6. Build the organisational capabilities to deliver

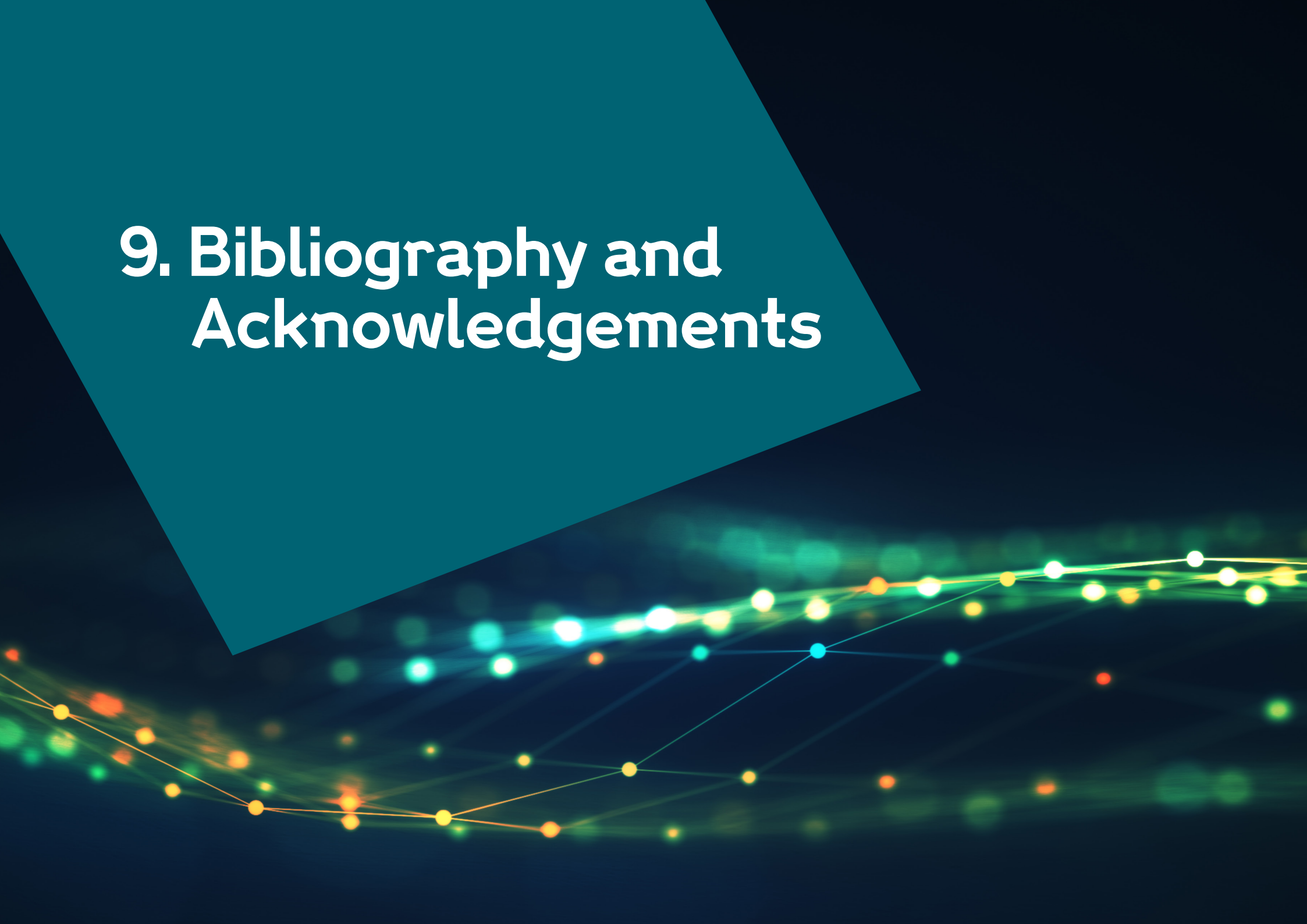
Many funders may currently lack the capabilities to understand the AI space, critically identify opportunities, advise current grantees on opportunities, and appraise future grant applications. The breadth of the field, limited talent pools and competition from tech, 'AI-first' biotech and pharmaceutical companies are likely to make it challenging for most funders to build deep internal expertise on these topics.

Funders may also face choices about how to best incorporate AI as part of their activities. For example, will a funder have an 'AI strategy' or expect this to be a core part of any strategic approach to drug discovery? Will a funder consider AI-focused funding applications alongside more traditional approaches or issue a specific call for AI-focused proposals? These choices have implications on the depth of AI expertise a funder may require and how extensively this is required across the organisation.

Funders can review their current capability mix and build a plan to fill any gaps as their engagement with AI topics grows over time

> *"If we truly want to drive any marked impact on human health with AI, funders, academics and industry need to work together. One single company can't solve this alone*
>
> **Vice President, 'AI-first' biotech**

9. Bibliography and Acknowledgements



9.1. Bibliography

[1] N. Fleming, “How artificial intelligence is changing drug discovery,” Nature, vol. 557, no. 7707, pp. S55–S57, May 2018, doi: 10.1038/d41586-018-05267-x.

[2] D. H. Freedman, “Hunting for New Drugs with AI,” Nature, vol. 576, no. 7787, pp. S49–S53, Dec. 2019, doi: 10.1038/d41586-019-03846-0.

[3] O. J. Wouters, M. McKee, and J. Luyten, “Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018,” JAMA, vol. 323, no. 9, pp. 844–853, Mar. 2020, doi: 10.1001/jama.2020.1166.

[4] R. C. Mohs and N. H. Greig, “Drug discovery and development: Role of basic biological research,” Alzheimers Dement (N Y), vol. 3, no. 4, pp. 651–657, Nov. 2017, doi: 10.1016/j.trci.2017.10.005.

[5] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott, “Principles of early drug discovery,” Br J Pharmacol, vol. 162, no. 6, pp. 1239–1249, Mar. 2011, doi: 10.1111/j.1476-5381.2010.01127.x.

[6] J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, no. 7873, Art. no. 7873, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[7] A. N. Ramesh, C. Kambhampati, J. R. T. Monson, and P. J. Drew, “Artificial intelligence in medicine.,” Ann R Coll Surg Engl, vol. 86, no. 5, pp. 334–338, Sep. 2004, doi: 10.1308/147870804290.

[8] L. Goyal et al., “First Results of RLY-4008, a Potent and Highly Selective FGFR2 Inhibitor in a First-in-Human Study in Patients with FGFR2-Altered Cholangiocarcinoma and Multiple Solid Tumors”.

[9] “Adopting AI in Drug Discovery,” BCG Global, Mar. 23, 2022. https://www.bcg.com/publications/2022/adopting-ai-in-pharmaceutical-discovery

[10] X. Li, Y. Xu, H. Yao, and K. Lin, “Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors,” Journal of Cheminformatics, vol. 12, no. 1, p. 42, Jun. 2020, doi: 10.1186/s13321-020-00446-3.

[11] “Absci First to Create and Validate De Novo Antibodies with Zero-Shot Generative AI | Absci Corp.” https://investors.absci.com/news-releases/news-release-details/absci-first-create-and-validate-de-novo-antibodies-zero-shot

[12] “BenevolentAI Achieves Further Milestones In AI-Enabled Target Identification Collaboration With AstraZeneca,” BenevolentAI (AMS: BAI). https://www.benevolent.com/news-and-media/press-releases-and-in-media/benevolentai-achieves-further-milestones-ai-enabled-target-identification-collaboration-astrazeneca/

[13] G. MassonNov 8 and 2022 09:00am, “Amid ‘biotech winter,’ Insilico turns up the heat with Sanofi deal worth $1.2B in biobucks,” Fierce Biotech, Nov. 08, 2022. https://www.fiercebiotech.com/biotech/amid-biotech-winter-insilico-turns-heat-sanofi-deal-worth-12b-biobucks

[14] A. Patronov and I. Doytchinova, “T-cell epitope vaccine design by immunoinformatics,” Open Biol, vol. 3, no. 1, p. 120139, Jan. 2013, doi: 10.1098/rsob.120139.

[15] “AI In Biologics Discovery: An Emerging Frontier,” In Vivo, Oct. 11, 2022. https://invivo.pharmaintelligence.informa.com/IV146716/AI-In-Biologics-Discovery-An-Emerging-Frontier

[16] J. J. Han, “FDA Modernization Act 2.0 allows for alternatives to animal testing,” Artif Organs, vol. 47, no. 3, pp. 449–450, Mar. 2023, doi: 10.1111/aor.14503.

[17] C. Gorgulla et al., “An open-source drug discovery platform enables ultra-large virtual screens,” Nature, vol. 580, no. 7805, Art. no. 7805, Apr. 2020, doi: 10.1038/s41586-020-2117-z.

[18] G. Amendola and S. Cosconati, “PyRMD: A New Fully Automated AI-Powered Ligand-Based Virtual Screening Tool,” J. Chem. Inf. Model., vol. 61, no. 8, pp. 3835–3845, Aug. 2021, doi: 10.1021/acs.jcim.1c00653.

[19] “Automating drug discovery | Nature Reviews Drug Discovery.” https://www.nature.com/articles/nrd.2017.232?draft=collection

[20] S. Makin, “Could an algorithm predict the next pandemic?,” Nature, vol. 610, no. 7933, pp. S42–S44, Oct. 2022, doi: 10.1038/d41586-022-03358-4.

[21] “Human Immunome Project.” https://www.humanimmunomeproject.org/

[22] “CEPI’s 100 Days Mission,” CEPI. https://100days.cepi.net/

[23] “Preparing for pandemics.” https://www.who.int/westernpacific/activities/preparing-for-pandemics

[24] D. V. S. Green et al., “BRADSHAW: a system for automated molecular design,” J Comput Aided Mol Des, vol. 34, no. 7, pp. 747–765, Jul. 2020, doi: 10.1007/s10822-019-00234-8.

[25] “Exscientia-March-2023-presentation.pdf.” AAvailable: https://s28.q4cdn.com/460399462/files/doc_presentations/2023/03/Exscientia-March-2023-presentation.pdf

[26] “a15bfdeb-c705-45e6-b3a9-a4939471e117.pdf.”. Available: https://ir.recursion.com/static-files/a15bfdeb-c705-45e6-b3a9-a4939471e117

[27] “Critical assessment of computational Hit-Finding experiments | CACHE.” https://cache-challenge.org/

[28] M. K. P. Jayatunga, W. Xie, L. Ruder, U. Schulze, and C. Meier, “AI in small-molecule drug discovery: a coming wave?,” Nature Reviews Drug Discovery, vol. 21, no. 3, pp. 175–176, Feb. 2022, doi: 10.1038/d41573-022-00025-1.

[29] S. M. Paul et al., “How to improve R&D productivity: the pharmaceutical industry’s grand challenge,” Nat Rev Drug Discov, vol. 9, no. 3, pp. 203–214, Mar. 2010, doi: 10.1038/nrd3078.

[30] “Evaxion Biotech Reports Data from Phase 1/2a Trials of EVX-01 and EVX-02 | Evaxion Biotech.” https://investors.evaxion-biotech.com/news-releases/news-release-details/evaxion-biotech-reports-data-phase-12a-trials-evx-01-and-evx-02/

[31] “Takeda Announces Positive Phase 2b Psoriasis Results for Oral TYK2 Inhibitor.” https://www.takeda.com/newsroom/newsreleases/2023/takeda-announces-positive-results-in-phase-2b-study-of-investigational-tak-279

[32] “Relay Therapeutics Reports Fourth Quarter and Full Year 2022 Financial Results and Corporate Highlights | Relay Therapeutics.” https://ir.relaytx.com/news-releases/news-release-details/relay-therapeutics-reports-fourth-quarter-and-full-year-2022/

[33] J. O. Park et al., “76MO Efficacy of RLY-4008, a highly selective FGFR2 inhibitor in patients (pts) with an FGFR2-fusion or rearrangement (f/r), FGFR inhibitor (FGFRi)-naïve cholangiocarcinoma (CCA): ReFocus trial,” Annals of Oncology, vol. 33, pp. S1461–S1462, Nov. 2022, doi: 10.1016/j.annonc.2022.10.112.

[34] “Nimbus Therapeutics Announces Positive Topline Results for Phase 2b Clinical Trial of Allosteric TYK2 Inhibitor in Psoriasis – Nimbus.” https://www.nimbustx.com/2022/11/30/nimbus-therapeutics-announces-positive-topline-results-for-phase-2b-clinical-trial-of-allosteric-tyk2-inhibitor-in-psoriasis/

[35] “Takeda Completes Acquistion of Nimbus Therapeutics’ TYK2 Program Subsidiary” https://www.takeda.com/newsroom/newsreleases/2023/takeda-completes-acquisition-of-nimbus-therapeutics-tyk2-program-subsidiary/

[36] I. M. Svane, “A Pilot Study of the Safety, Tolerability, Feasibility and Efficacy of Anti-PD-1 or Anti-PD-L1 in Combination With a Personalized Neo-antigen Vaccine in Advanced Solid Tumors (NeoPepVac),” clinicaltrials.gov, Clinical trial registration NCT03715985, Jan. 2022.. Available: https://clinicaltrials.gov/ct2/show/NCT03715985

[37] P. J. Richardson, B. W. S. Robinson, D. P. Smith, and J. Stebbing, “The AI-Assisted Identification and Clinical Efficacy of Baricitinib in the Treatment of COVID-19,” Vaccines (Basel), vol. 10, no. 6, p. 951, Jun. 2022, doi: 10.3390/vaccines10060951.

[38] “FDA Converts Emergency Approval Of Baricitinib — First Identified As A COVID Treatment By BenevolentAI — To A Full Approval,” BenevolentAI (AMS: BAI). https://www.benevolent.com/news-and-media/blog-and-videos/fda-converts-emergency-approval-baricitinib-first-identified-covid-treatment-benevolentai-full-approval/

[39] M. Tuccori et al., “An overview of the preclinical discovery and development of bamlanivimab for the treatment of novel coronavirus infection (COVID-19): reasons for limited clinical use and lessons for the future,” Expert Opin Drug Discov, pp. 1–12, doi: 10.1080/17460441.2021.1960819.

[40] “Lilly’s neutralizing antibody bamlanivimab (LY-CoV555) receives FDA emergency use authorization for the treatment of recently diagnosed COVID-19 | Eli Lilly and Company.” https://investor.lilly.com/news-releases/news-release-details/lillys-neutralizing-antibody-bamlanivimab-ly-cov555-receives-fda

[41] T. Hanser, “Federated learning for molecular discovery,” Current Opinion in Structural Biology, vol. 79, p. 102545, Apr. 2023, doi: 10.1016/j.sbi.2023.102545.

[42] “Substra - powering federated learning research,” OWKIN. https://owkin.com/substra/

[43] Jha, A., Khan, Y., Mehdi, M. et al. Towards precision medicine: discovering novel gynecological cancer biomarkers and pathways using linked data. J Biomed Semant 8, 40 (2017). https://doi.org/10.1186/s13326-017-0146-9

[44] “FAIR Principles,” GO FAIR. https://www.go-fair.org/fair-principles/

[45] D. Vemula, P. Jayasurya, V. Sushmitha, Y. N. Kumar, and V. Bhandari, “CADD, AI and ML in drug discovery: A comprehensive review,” European Journal of Pharmaceutical Sciences, vol. 181, p. 106324, Feb. 2023, doi: 10.1016/j.ejps.2022.106324.

[46] F. I. Saldívar-González, V. D. Aldas-Bulos, J. L. Medina-Franco, and F. Plisson, “Natural product drug discovery in the artificial intelligence era,” Chemical Science, vol. 13, no. 6, pp. 1526–1546, 2022, doi: 10.1039/D1SC04471K.

[47] “Sanofi, Pfizer and More Use Upskilling to Solve the Life Science Talent Shortage,” BioSpace. https://www.biospace.com/article/upskilling-a-solution-to-the-life-science-talent-shortage/

[48] T. Ciecierski-Holmes, R. Singh, M. Axt, S. Brenner, and S. Barteit, “Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review,” npj Digit. Med., vol. 5, no. 1, Art. no. 1, Oct. 2022, doi: 10.1038/s41746-022-00700-y.

[49] S. R. P. Franzen, C. Chandler, and T. Lang, “Health research capacity development in low and middle income countries: reality or rhetoric? A systematic meta-narrative review of the qualitative literature,” BMJ Open, vol. 7, no. 1, p. e012332, Jan. 2017, doi: 10.1136/bmjopen-2016-012332.

[50] A. Hosny and H. J. W. L. Aerts, “Artificial intelligence for global health,” Science, vol. 366, no. 6468, pp. 955–956, Nov. 2019, doi: 10.1126/science.aay5189.

[51] H. Ejaz, H. McGrath, B. L. Wong, A. Guise, T. Vercauteren, and J. Shapey, “Artificial intelligence and medical education: A global mixed-methods study of medical students’ perspectives,” Digit Health, vol. 8, p. 20552076221089100, May 2022, doi: 10.1177/20552076221089099.

[52] “OMSF Projects,” May 30, 2018. https://omsf.io/projects/project-list/

[53] “Home - CASP15.” https://predictioncenter.org/casp15/index.cgi

[54] OpenAI, “GPT-4 Technical Report.” arXiv, Mar. 27, 2023. Available: http://arxiv.org/abs/2303.08774

[55] M. Varadi et al., “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models,” Nucleic Acids Res, vol. 50, no. D1, pp. D439–D444, Nov. 2021, doi: 10.1093/nar/gkab1061.

[56] Y. Zhao et al., “PO2RDF: representation of real-world data for precision oncology using resource description framework,” BMC Medical Genomics, vol. 15, no. 1, p. 167, Jul. 2022, doi: 10.1186/s12920-022-01314-9.

[57] H. Landi, “Precision medicine company Tempus inks 3rd major pharma deal, securing nearly $1B revenue boost,” Fierce Healthcare, Mar. 02, 2023. https://www.fiercehealthcare.com/health-tech/precision-medicine-company-tempus-inks-3rd-major-pharma-deal-securing-nearly-1b-revenue

[58] “Immunai Raises $215 Million to Accelerate Development of Its Immune-First Drug Actuary Platform,” WebWire. https://www.webwire.com/ViewPressRel.asp?aId=280863

[59] “Integrating multimodal data to meet clinical challenges,” OWKIN. https://owkin.com/publications-and-news/blogs/integrating-multimodal-data-to-meet-clinical-challenges/

[60] E. Carron, “Tempus Launches Psychiatric Real-World Data Program to Advance Personalized Medicine,” Tempus, May 20, 2022. https://www.tempus.com/news/pr/tempus-launches-psychiatric-real-world-data-program-to-advance-personalized-medicine/

[61] S. Rodriguez et al., “Machine learning identifies candidates for drug repurposing in Alzheimer’s disease,” Nat Commun, vol. 12, no. 1, Art. no. 1, Feb. 2021, doi: 10.1038/s41467-021-21330-0.

[62] R. L. S. Exchange, “Poolbeg Pharma PLC Announces Influenza AI model build completed,” ACCESSWIRE News Room, Nov. 29, 2022. https://www.accesswire.com/729168/Poolbeg-Pharma-PLC-Announces-Influenza-AI-model-build-completed

[63] “Atomwise Opens Applications for Historic AI Drug Discovery Awards,” Atomwise, Apr. 09, 2017. https://www.atomwise.com/2017/04/09/atomwise-opens-applications-for-historic-ai-drug-discovery-awards/

[64] “Exscientia applies Genome scale AI-Drug Discovery to critical global health challenges - Company receives $4.2M grant from Bill & Melinda Gates Foundation to identify new targets and leads for malaria, tuberculosis, and non-hormonal contraception.” https://investors.exscientia.ai/press-releases/press-release-details/2020/Exscientia-applies-Genome-scale-AI-Drug-Discovery-to-critical-global-health-challenges---Company-receives-4.2M-grant-from-Bill--Melinda-Gates-Foundation-to-identify-new-targets-and-leads-for-malaria-tuberculosis-and-non-hormonal-contraception/default.aspx

[65] "Exscientia enters $70M collaboration to develop anti-viral therapeutics against Coronavirus and other viruses with pandemic potential." https://investors.exscientia.ai/press-releases/press-release-details/2021/exscientia-enters-70m-collaboration-to-develop-anti-viral-therapeutics-against-coronavirus-and-other-viruses-with-pandemic-potential/Default.aspx

[66] "AI Driven De Novo Design with REINVENT • BioSolveIT," BioSolveIT. https://www.biosolveit.de/webinar/ai-driven-de-novo-design-with-reinvent/

[67] S. Winks, J. G. Woodland, G. 'Colin' Pillai, and K. Chibale, "Fostering drug discovery and development in Africa," Nat Med, vol. 28, no. 8, Art. no. 8, Aug. 2022, doi: 10.1038/s41591-022-01885-1.

[68] H. Foundation, "AI/ML Workshop Opens New Doors for Young Scientists | H3D Foundation | Pioneering World-Class Drug Discovery in Africa," H3D Foundation, Nov. 22, 2022. https://h3dfoundation.org/aiml-workshop-opens-new-doors-for-young-scientists/

[69] "Indaba Grand Challenge: Curing Leishmaniasis," Zindi. https://zindi.africa/competitions/indaba-grand-challenge-curing-leishmaniasis

[70] J. Smith, H. Xu, X. Li, L. Yang, and J. M. Gutierrez, "Compound Screening with Deep Learning for Neglected Diseases: Leishmaniasis." bioRxiv, p. 2021.10.02.462874, Oct. 02, 2021. doi: 10.1101/2021.10.02.462874.

[71] C. H. Lee et al., "A robust deep learning platform to predict CD8+ T-cell epitopes." bioRxiv, p. 2022.12.29.522182, Dec. 29, 2022. doi: 10.1101/2022.12.29.522182.

[72] C. Soto et al., "High frequency of shared clonotypes in human B cell receptor repertoires," Nature, vol. 566, no. 7744, Art. no. 7744, Feb. 2019, doi: 10.1038/s41586-019-0934-8.

[73] "evaxion-corpdeck-aug-2022_pdf.pdf." Available: https://www.evaxion-biotech.com/media/nnmcnfd3/evaxion-corpdeck-aug-2022_pdf.pdf

[74] "Moderna Announces First Participant Dosed in a Phase 1 Trial of its Nipah Virus mRNA Vaccine, mRNA-1215." https://investors.modernatx.com/news/news-details/2022/Moderna-Announces-First-Participant-Dosed-in-a-Phase-1-Trial-of-its-Nipah-Virus-mRNA-Vaccine-mRNA-1215/default.aspx

[75] "Moderna Announces Its Global Public Health Strategy." https://investors.modernatx.com/news/news-details/2022/Moderna-Announces-Its-Global-Public-Health-Strategy/default.aspx

[76] T. M. Chidyausiku et al., "De novo design of immunoglobulin-like domains," Nat Commun, vol. 13, no. 1, Art. no. 1, Oct. 2022, doi: 10.1038/s41467-022-33004-6.

[77] "ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation | Bioinformatics | Oxford Academic." https://academic.oup.com/bioinformatics/article/38/7/1877/6517780

[78] D. Prihoda, "BioPhi Antibody design platform." https://biophi.dichlab.org/

[79] J. A. Ruffolo, L.-S. Chu, S. P. Mahajan, and J. J. Gray, "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies." bioRxiv, p. 2022.04.20.488972, Apr. 21, 2022. doi: 10.1101/2022.04.20.488972.

[80] R. Evans et al., "Protein complex prediction with AlphaFold-Multimer." bioRxiv, p. 2021.10.04.463034, Mar. 10, 2022. doi: 10.1101/2021.10.04.463034.

[81] "What is Rosetta@home?" https://boinc.bakerlab.org/rosetta/rah/rah_about.php

[82] "AbCellera Receives Grant to Help Fight Tuberculosis." https://investors.abcellera.com/news/news-releases/2017/AbCellera-Receives-Grant-to-Help-Fight-Tuberculosis/default.aspx

[83] "AbCellera Signs Agreement with Global Health Foundation to Fight Infectious Disease." https://investors.abcellera.com/news/news-releases/2019/AbCellera-Signs-Agreement-with-Global-Health-Foundation-to-Fight-Infectious-Disease/default.aspx

[84] C. Hale, "Pfizer, AstraZeneca, Merck KGaA-backed Israeli AI incubator launches first biopharma startup," Fierce Biotech, Sep. 28, 2022. https://www.fiercebiotech.com/medtech/israeli-ai-incubator-launches-drug-discovery-startup-backed-pfizer-astrazeneca-merck-kgaa

[85] "TransCelerate and BioCelerate Launch New Technology Platform to Enable R&D Data Sharing," Bloomberg.com, Jul. 31, 2018. Available: https://www.bloomberg.com/press-releases/2018-07-31/transcelerate-and-biocelerate-launch-new-technology-platform-to-enable-r-d-data-sharing

[86] "Chagas Disease | Infectious Diseases Data Observatory." https://www.iddo.org/research-themes/chagas-disease

9.2. Acknowledgements

This report was commissioned by the Wellcome Trust and authored by Boston Consulting Group (BCG), drawing on research and analysis conducted by BCG. Input and oversight from the Wellcome Trust was led by Colleen Loynachan, Harriet Unsworth, Kim Donoghue, Raphael Sonabend and Sabrina Lamour-Julien. The BCG team was led by Andrew Rodriguez, Christoph Meier, Emily Serazin, John Gooch, Priyanka Aggarwal, Asher Steene, Madura Jayatunga, Shruti Nayak, and Will Randall with contributions from Lotte Bruens, Priyanka Harley, Maria Antunica, Methuna Kailanathan and Nana Balser.

The authors would like to thank the members of the Expert Scientific Advisory Committee for their support, insights, and guidance to contextualise and amplify this work. Their contributions were vital in validating and refining findings presented in this report. The committee consisted of:

- **Alain Bouckenooghe** – CSO, Hilleman Laboratories Singapore
- **Charlotte Deane** – Chief Scientist, Exscientica/ University of Oxford
- **Chris Rackauckas** – Research Affiliate, MIT; Lead Developer, SciML and Director of Scientific Research, Pumas-AI.
- **Dorcas Osei-Safo** – Associate Professor, University of Ghana
- **Emna Harigua** – Research team leader, Institut Pasteur de Tunis
- **Ivan Griffin** – COO, Benevolent AI
- **Joseph Lehar** – SVP Business Strategy, Owkin
- **Jing Li** – CEO, VelaVigo
- **Kelly Chibale** – Professor, University of Cape Town
- **Ziv Bar-Joseph** – R&D Data & Comp. Sci, Sanofi

In addition, the project team extends sincere thanks to the 55 experts who provided their time and insights through expert interviews, as well as the 102 experts who participated in the online survey. These experts were industry professionals and academics from multiple geographies and across varying tenures that provided an objective perspective on the topic and were instrumental in informing the insights presented in this report.

The views and opinions expressed in this report represent those of the joint Wellcome Trust / The Boston Consulting Group project team, and do not necessarily reflect those of the any specific individual or organisation mentioned above.

10. Appendix

10.1. Glossary

Use cases definitions

1.1: High throughput, unbiased screens run for drug candidates that modulate disease-relevant phenotypes to enable the identification of (novel) molecules that act on disease pathways

1.2: Using AI to analyse large datasets and disparate sources to uncover indirect associations between disease and cellular target(s), as well as interactions between drugs

1.3: Knowledge graphs organise data from multiple sources to represent a network of real-world entities (e.g., objects, situations, concepts) and illustrate the relationship between them

1.4: Prediction of 3D protein structure based on amino acid sequence and subsequent dynamic protein interaction modelling to accurately model protein binding pockets, protein-protein binding and protein-ligand binding

1.5: Identification of diagnostic and prognostic biomarkers (molecules by which a particular disease etc. can be recognised) to support drug development and treatments

2.1: Computational techniques used to search libraries of small molecules to identify structures most likely to bind to a drug target; using computer simulations to analyse the physical movements of each atom and molecule

2.2: AI analysis of molecule structure and experimental data to predict the biological activities of small molecules

2.3: The design of novel chemical entities that fit a set of constraints using computational growth algorithms; often used to generate lead-like small molecules after analysis of prospective protein targets

2.4: The AI analysis, and ranking, of all possible synthetic routes for a given compound according to various metrics of synthetic accessibility

2.5: Prediction and multiparameter optimisation of small molecule pharmacokinetic properties using AI

3.1: Use of AI to predict potential (conserved) antigenic sites on a protein, simulate molecular docking using antibody sequence and structure to predict binding surfaces or Ag-Ab binding

3.2: Optimisation of coding and non-coding sequences to improve mRNA stability and translation efficiency and, therefore, protein production

3.3: Optimisation of mRNA delivery systems to enhance transportation into the cell and reduce toxicity (e.g., through improving lipid composition, molar ratios, and structure)

4.1: High-throughput screening of natural antibody repertoires for identification and selection of optimal therapeutic antibodies, and to extrapolate to further evolutionary repertoires that have not yet been observed

4.2: Simulation of molecular docking using antibody sequence and structure to predict Ab-Ag binding

4.3: Creation of optimised antibody sequences beyond natural repertoires (incl. assessing impact of mutations and PTMs), and *in silico* library design

4.4: Prediction of protein properties based on their sequence (e.g., solubility, aggregation)

4.5: Modification of antibody protein sequences from non-human species to increase their similarity to antibody variants produced naturally in humans

5.1: Employment of AI for the pre-emptive flagging of drug candidates for predicted toxic and/or off-target effects

5.2: The modelling of pharmacokinetics and pharmacodynamics to predict the time course of effect intensity in response to the administration of a drug dose

5.3: The modelling of dynamic interactions between biological systems and drugs; allows prediction of efficacy, and the reasons behind it

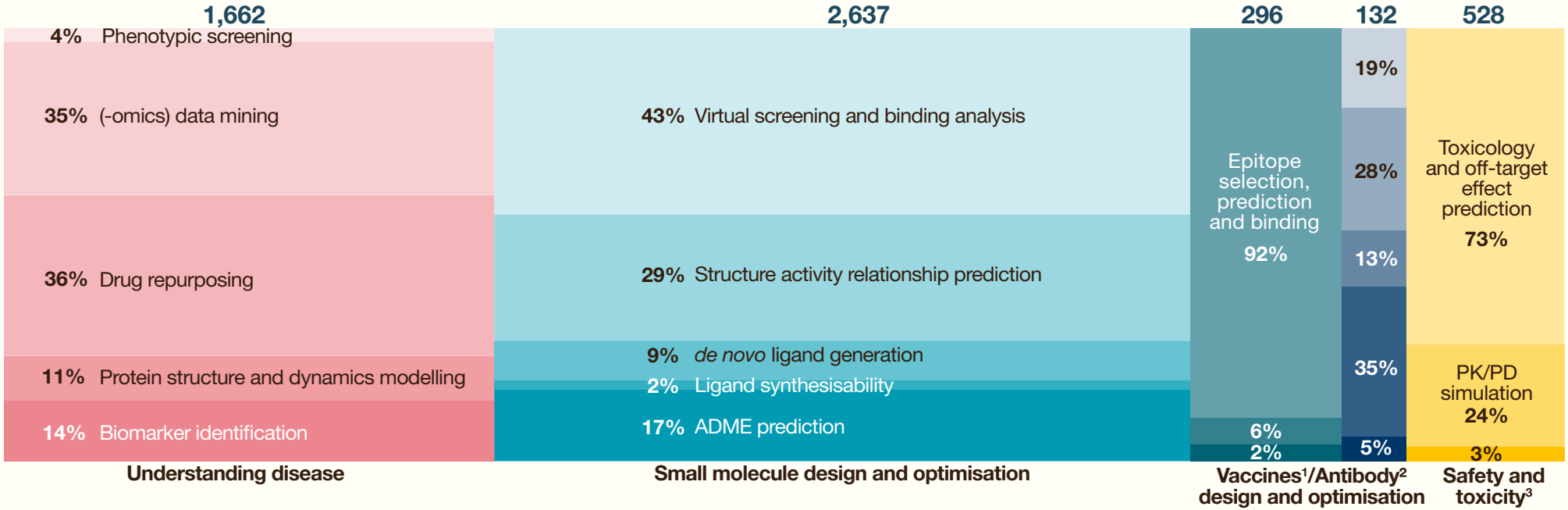
10.2. Abbreviations

ADME(T): Absorption, distribution, metabolism, excretion, toxicity
AI: Artificial Intelligence
CACHE: Critical Assessment of Computational Hit-finding Experiments
CASP: Critical Assessment of Protein Structure Prediction
DD: Drug Discovery
FAIR: Findable, Accessible, Interoperable, Usable
FDA: Food and Drug Administration
GANs: Generative Adversarial Networks
GPT: Generative Pre-trained Transformer
EUA: Emergency Use Authorisation
HICs: High-income countries
HIV: Human Immunodeficiency Virus
LMICs: Low-to-middle income countries
LLMs: Large Language Models
LPS: Longitudinal population studies
ML: Machine Learning
MH: Mental Health
MHC: Major Histocompatibility Complex
NTDs: Neglected tropical diseases
OCD: Obsessive-compulsive disorder
PK/PD: Pharmacokinetic/pharmacodynamic
QSP: Quantitative Systems Pharmacology
R&D: Research and Development
SMILEs: Simplified Molecular-Input Line-Entry Systems
TA: Therapeutic Area

10.3. Sub-use case Analysis

Figure 20 shows an analysis of publications on AI in drug discovery published in the last 5 years by sub-use case. The most frequently published sub-use cases within understanding disease were *(-omics) data mining* to link target to disease and *drug repurposing* due to the high quality and quantity of data available to train AI models. Within small molecule design and optimisation, AI was most frequently applied to sub-use cases pertaining to *virtual screening and binding analysis* and *structure activity relationship prediction*, which is likely driven by the availability of well-validated tools. For vaccines design and optimisation use cases, *epitope selection, prediction and binding* sub-use case dominates due to the focus on better understanding the molecular basis of immunity. AI activity within antibody design and optimisation is focused on *antibody property prediction*. A large proportion of AI work in safety and toxicity is concentrated on *toxicology and off-target effect prediction*.

Figure 20 – Publications on AI in drug discovery published in the last five years, by sub-use case



1. Epitope selection, prediction and binding; Codon, 5' and 3' UTR optimisation; LNP optimisation. 2. mAb library screening and repertoire prediction; Ag-Ab binding prediction and optimisation; de novo antibody design ; Antibody property prediction; Humanisation 3. QSP modelling

10.4. 'AI-first' Biotechs

The 'AI-first' biotechs analysed within this landscaping report are listed in the table below. Those marked with (*) were included in the pipeline analysis based on the source data used. Those marked with (#) are public 'AI-first' biotechs founded in the last 10 years.

1910 Genetics	**Athos Therapeutics, Inc. ***	**Cloud Pharmaceuticals ***	Entos Pharmaceuticals
3T Biosciences *	**Atomwise ***	Clover Therapeutics	Entos, Inc.
A2A Pharmaceuticals *	Auransa	Collaborations Pharmaceuticals	**Enveda Biosciences ***
AbCellera Biologics *#	**BenevolentAI *#**	**Compugen Ltd. ***	**Envisagenics ***
AbSci	Berg	Computational Medicine Beijing Co., Ltd.	EpiVax
Accutar Biotechnology *	BigHat Biosciences	Creyon Bio, Inc.	EpiVax Therapeutics
Acelot Inc *	**BioAge Labs ***	Cyclica	**e-therapeutics ***
AcuraStem	**BioMap ***	CytoReason	Evaxion Biotech
Adagene *	Biomatter Designs	Data2Discovery	**Exscientia *#**
Adapsyn Bioscience	Biorelate	Deargen	**Frontier Medicines ***
Adimab *	BioSymetrics	Deep Intelligent Pharma	**Gain Therapeutics *#**
AI Therapeutics	biotx.ai	Deepcell	Galapagos
Ai-biopharma	Biovista	DeepCure	Galixier
Aimble	**BioXcel Therapeutics *#**	DeepLife	Galixir
Ainnocence LLC	**Black Diamond Therapeutics *#**	DeepMatter	Gandeeva Therapeutics Inc.
Anagenex, Inc.	C4X Discovery	DeepTrait	GATC Health
Anima Biotech	CardiaTec Biosciences LTD	Delta 4	Gatehouse Bio
Animol Discovery, Inc.	Causaly	Denovicon Therapeutics	Generate Biomedicines
Antiverse	**Celeris Therapeutics ***	Differentiated Therapeutics, Inc.	Genesis Therapeutics
Aqemia	Cellarity	Eleven Therapeutics Ltd	Genialis
Arctoris *	**Celsius Therapeutics ***	Elucidata	Gero
Aria Pharmaceuticals (Formerly: TwoXAR) *	Charm Therapeutics	Empiric Logic	GigaCeuticals
Arpeggio Biosciences, Inc. *	CHARM Therapeutics Inc.	**Empirico ***	Glympse Bio
Artivila Therapeutics *	ChemAlive SA	**Engine Biosciences ***	GNS Healthcare
Asimov	ChemPass	ENSEM Therapeutics Inc.	**Gritstone Bio (formerly -- Gritstone Oncology) *#**

GT Apeiron Therapeutics *	Micar21	**Pharos I&BT Co., Ltd ***	Shanghai GV20 Biotechnology Co., Ltd.
Harmonic Discovery Inc.	Micrographia Bio	**Pharos iBio ***	Shanghai Matwings Technology Co., Ltd.
Healx *	**MindRank AI ***	**Pharos iBT ***	Shenzhen NeoCura Biotechnology
HelixNano	**Model Medicines ***	Phenomic AI	Shuimu BioSciences
HemoShear Therapeutics, Inc. *	**Modulus Discovery ***	Polaris Quantum Biotech	Silexon AI Technology
Herophilus, Inc. *	Molecule.one	PostEra	Silicon Therapeutics
HiFiBiO Therapeutics *	**Molomics ***	Pragma Biosciences Inc.	**Sinopia Biosciences ***
Hotspot Therapeutics *	Nabla Bio	Profluent Bio Inc.	**Soley Therapeutics, Inc. ***
Huashen Zhiyao Technology (Beijing) Co., Ltd	Nanjing Suikun Intelligent Technology Co., Ltd	Protai Bio	**SOM Innovation Biotech S.L. ***
Hummingbird Bioscience	NeoCura	ProteinQure	Spring Discovery
Immunai	**neoX Biotech ***	PsychoGenics	Standigm
IMMUNITOAI PRIVATE LIMITED	Neumora Therapeutics	Purposeful	StoneWise
Insilico Medicine *	**Neuron23 ***	Pythia Labs	Strateos
Insitro	New Equilibrium Biosciences, Inc.	**Q-State Biosciences, Inc. ***	**Syntekabio ***
InterAx Biotech AG	Nimbus Therapeutics	Qubit Pharmaceuticals	Systems Oncology
Interprotein	**Nobias Therapeutics, Inc. ***	Quris Technologies LTD.	TandemAI
InveniAI	NonExomics	RECEPTOR.AI	Ten63 Therapeutics, Inc.
InVivo AI	Novoheart	**Recursion Pharmaceuticals *#**	Terray Therapeutics
Juvena Therapeutics *	NuMedii	**Relation Therapeutics Ltd. ***	Totient
Keen Eye Technologies	OccamzRazor	**Relay Therapeutics ***	**Totus Medicines ***
Kuano	**Ochre Bio ***	**Resonant Therapeutics ***	Turbine AI
LabGenius	**Octant, Inc. ***	Reverie Labs	Turbine Ltd.
Landos Biopharma *#	OmniAb Technologies	**ReviveMed ***	**Valo Health ***
Lassogen	OncXerna Therapeutics	Rezo Therapeutics, Inc.	Variational AI
LifeMine Therapeutics, Inc.	OneThree Biotech	**RNAimmune, Inc. ***	**Verge Genomics ***
Lodo Therapeutics	Optina Diagnostics	**RubrYc Therapeutics ***	VERISIM Life
Lunit	Ordaos, Inc.	**Schrödinger ***	Vesalius Therapeutics Inc.
Matchpoint Therapeutics, Inc	OWKIN	Scipher Medicine	Vevo Therapeutics, Inc.
Meliora Therapeutics, Inc.	PACT Pharma, Inc.	Seismic Bio	WhiteLab Genomics
Menten AI	Pepticom	Seismic Therapeutic, Inc.	**X37 ***
Metanovas Inc.	Peptone	SEngine Precision Medicine	XtalPi
Metis Pharmaceuticals *	PharmCADD	Serimmune Inc.	**ZebiAI Therapeutics ***

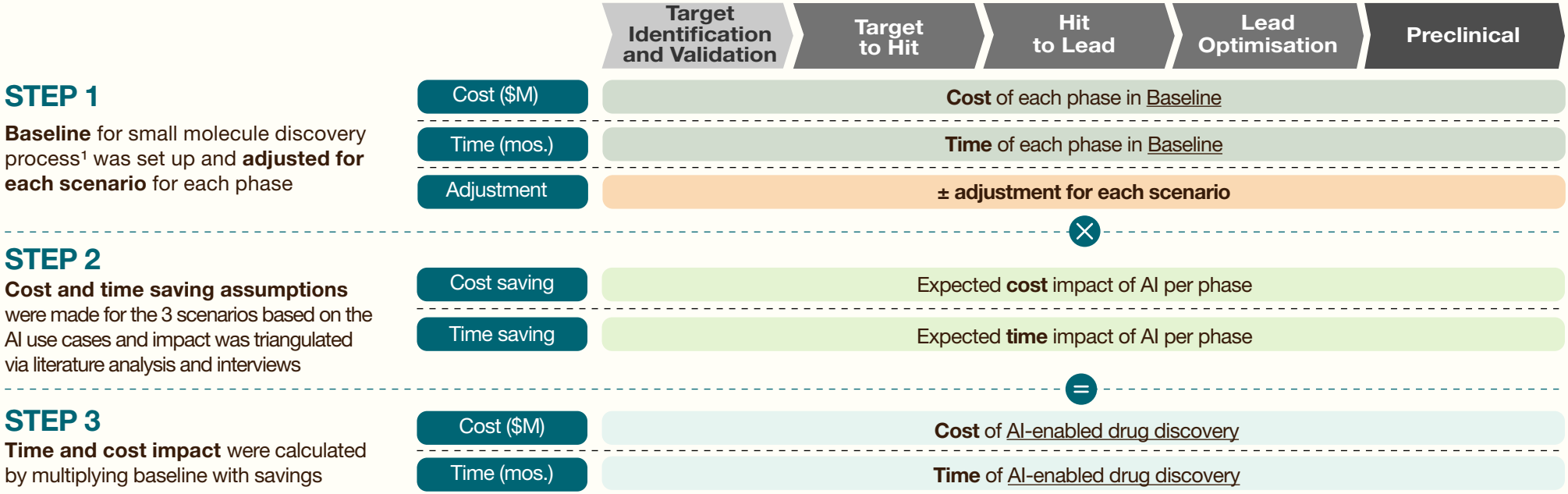
10.5. Value modelling

In our high-level value model, three different discovery scenarios were assessed. All three scenarios were within the small molecule discovery space. The scenarios were defined starting from baseline of typical experimental drug discovery, as it is practised today, and adjusting that baseline for each phase of the discovery value chain and testing potential impact ranges of AI on time & cost.

This was done in three steps (Figure 21).

- The original baseline for each phase was informed from prior literature and adjusted for inflation and advancement. For each phase, the baseline was adjusted to reflect the example scenarios (assuming no AI use cases are deployed). The adjusted baselines were validated by experts within the field.
- The potential time and cost impact of AI for each phase and scenario was determined by triangulating interviews, publications, and emerging proof points [8, 26, 28, 30-38].
- The adjusted baseline was then multiplied with potential time and cost impact to find the time and cost of AI-enabled drug discovery.

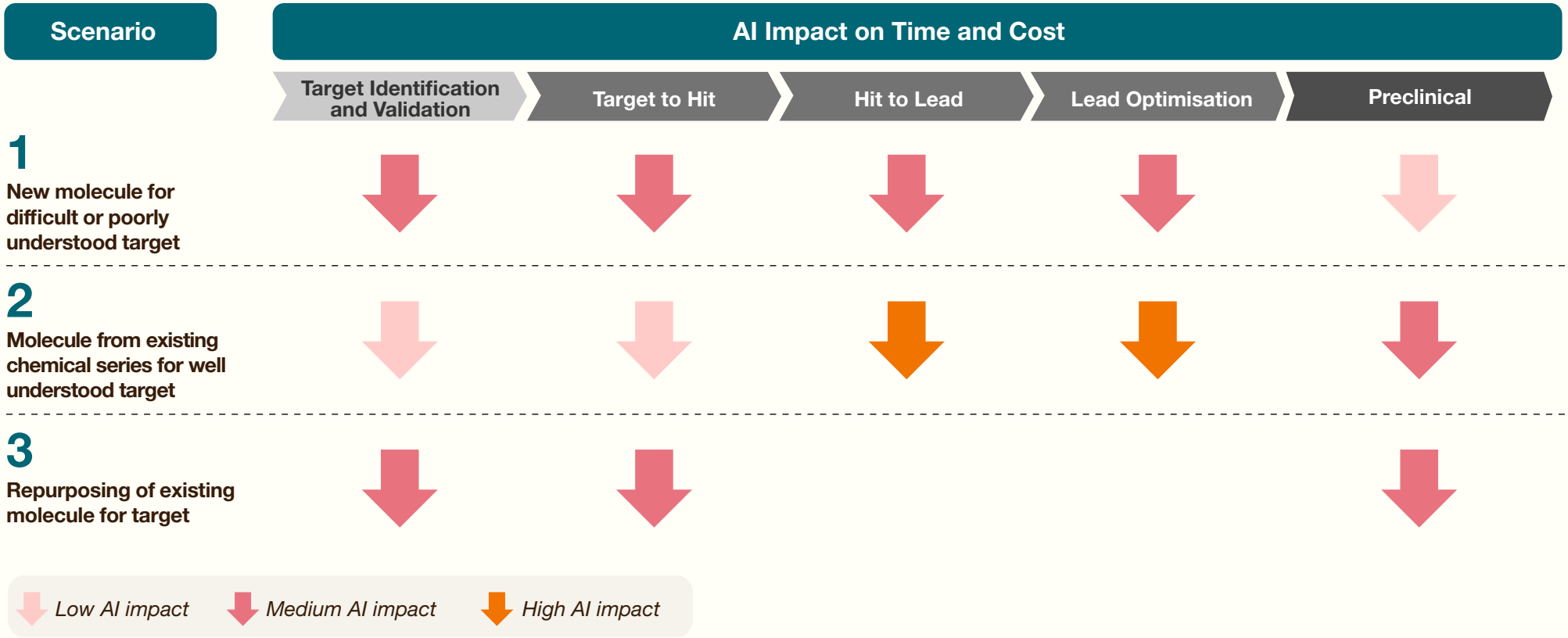
Figure 21 – Methodology used for value modelling



1. Baseline informed by [24], updated to adjust for inflation Note: The impact of AI on PoS is not modeled as this is difficult to quantify given long timelines

Figure 22 shows how the impact of AI varies for each phase and scenario.

Figure 22 – AI impact on time and cost across the discovery value chain



In Scenario One, AI has the largest impact on the time and cost of the following phases:

- Target identification and validation: AI use cases such as *(-omics) data mining* to link target to disease and *protein structure and dynamics modelling* can help understand the structure-function relationship more quickly, resulting in quicker and better hypothesis generation. For a difficult target, extensive biology and validation is required; AI can help prioritise targets systematically.
- Target to hit: Large DNA-encoded chemical libraries enable large, tailored libraries (which include de *novo* structures) to be generated and screened. This enables a larger chemical space to be explored at moderate costs and therefore increases the chance of discovering targets of interest and good hits for a difficult target.
- Hit to lead and lead optimisation: AI can significantly reduce the number of compounds and experiments required to find and optimise leads in various ways. For example, activity prediction enables experiments to be more targeted and predictive analytics can help forecast compound properties. However, the impact is limited by the lack of existing clinical data for a new molecule.

In Scenario Two, AI has the largest impact on the time and cost of the following phases:

- Hit to lead and lead optimisation: As mentioned in Scenario Two, AI use cases reduce the number of compounds and experiments required to find and optimise leads. Particularly for a well-understood target which has a wealth of existing data target (e.g., prior chemical & assay history) as this significantly reduces the number of design-make-test cycles required.

In Scenario Three, AI has the largest impact on the time and cost of the following phases:

- Target identification and validation: AI has the potential to accelerate the discovery of a novel disease-target relationship. For example, (-omics) mining and patient data (e.g., from historic trials or real-world evidence) can be used to draw novel links between diseases.
- Target to hit: AI can speed up the discovery of a novel target-molecule relationship using knowledge graphs or screening of licensed libraries. However, it is worth noting that the AI impact on cost is limited as screening of licensed libraries often results in additional costs.
- Preclinical: AI use cases such as predictive analytics on toxicity and PK/PD can enable compounds for testing to be prioritised which can save both time and costs. For a repurposed drug, the large availability of existing data (e.g., pharmacological, and functional impact data) and models (e.g., target-based toxicity prediction models) can be leveraged.

10.6. AI-derived clinical assets

The 'AI-derived assets' catalogued as part of this project are listed in the table below. The list was curated by extracting current clinical pipelines of 'AI-first' biotech companies from Citeline's Pharmaprojects, a global drug development database. All discontinued programmes and programmes regarding cell therapies were excluded to form the list. We recognise that pipelines are challenging to track and update given dynamic changes across clinical and preclinical portfolios. As such, we have used an external source from Citeline to assess the evolution of pipelines over time, but recognise that this source does not have full coverage across all 'AI-first' biotechs identified. This approach is not exhaustive but provides an indication of the size and growth of AI-derived portfolios.

Company	Generic Drug Name	Global Status	Modality	Therapeutic area
A2A Pharmaceuticals	AO-001	Phase II Clinical Trial	Small molecule	Oncology
AbCellera Biologics	bamlanivimab	Launched	Antibody	Covid-19
AbCellera Biologics	bebtelovimab	Phase II Clinical Trial	Antibody	Covid-19
Accutar Biotechnology	AC-682	Phase I Clinical Trial	Small molecule	Oncology
Accutar Biotechnology	AC-0176	Phase I Clinical Trial	Small molecule	Oncology
Accutar Biotechnology	AC-699	Phase I Clinical Trial	Small molecule	Oncology
Adagene	ADG-106	Phase II Clinical Trial	Antibody	Oncology
Adagene	ADG-104	Phase II Clinical Trial	Antibody	Oncology
Adagene	ADG-116	Phase II Clinical Trial	Antibody	Oncology
Adagene	ADG-126	Phase II Clinical Trial	Antibody	Oncology
Adagene	BC-006	Phase I Clinical Trial	Antibody	Oncology
Adimab	PM-1022	Phase I Clinical Trial	Antibody	Oncology
AI Therapeutics	sirolimus, LAM Therapeutics	Phase I Clinical Trial	Small molecule	Immunology
AI Therapeutics	apilimod dimesylate	Phase II Clinical Trial	Small molecule	Covid-19
AI Therapeutics	AIT-101	Phase II Clinical Trial	Small molecule	Neurology
BenevolentAI	BEN-2293	Phase II Clinical Trial	Small molecule	Other
Berg	ubidecarenone, BERG Pharma	Phase II Clinical Trial	Small molecule	Oncology
BioAge Labs	asapiprant	Phase II Clinical Trial	Small molecule	Covid-19
BioAge Labs	BGE-105	Phase I Clinical Trial	Small molecule	Musculoskeletal

Company	Generic Drug Name	Global Status	Modality	Therapeutic area
BioXcel Therapeutics	dexmedetomidine, BioXcel	Launched	Small molecule	Mental health
Black Diamond Therapeutics	BDTX-1535	Phase I Clinical Trial	Small molecule	Oncology
C4X Discovery	INDV-2000	Phase I Clinical Trial	Small molecule	Mental health
Compugen	bapotulimab	Phase I Clinical Trial	Antibody	Oncology
Compugen	COM-701	Phase II Clinical Trial	Antibody	Oncology
Compugen	COM-902	Phase I Clinical Trial	Antibody	Oncology
Entos Pharmaceuticals	COVID-19 vaccine, Entos Pharmaceuticals	Phase II Clinical Trial	Vaccine	Covid-19
EpiVax	influenza vaccine, H7N9, Epivax	Phase I Clinical Trial	Vaccine	Infectious disease
EpiVax	influenza vaccine, H7N9, Protein Sciences	Phase I Clinical Trial	Vaccine	Infectious disease
Evaxion Biotech	EVAX-01	Phase II Clinical Trial	Vaccine	Oncology
Evaxion Biotech	EVX-02	Phase II Clinical Trial	Vaccine	Oncology
Exscientia	EVOEXS-21546	Phase I Clinical Trial	Small molecule	Oncology
Exscientia	DSP-0038	Phase I Clinical Trial	Small molecule	Mental health
Gritstone Bio	GRANITE-001	Phase III Clinical Trial	Vaccine	Oncology
Gritstone Bio	SLATE-001	Phase II Clinical Trial	Vaccine	Oncology
Gritstone Bio	COVID-19 vaccine, Gritstone Oncology	Phase I Clinical Trial	Vaccine	Covid-19
Gritstone Bio	HIV vaccine, Gilead Sciences	Phase I Clinical Trial	Vaccine	Infectious disease
Gritstone Bio	SLATE v2	Phase II Clinical Trial	Vaccine	Oncology
Gritstone Bio	COVID-19 vaccine, Gritstone Bio	Phase I Clinical Trial	Vaccine	Covid-19
Gritstone Bio	COVID-19 vaccine, Gritstone Bio-1	Phase I Clinical Trial	Vaccine	Covid-19
Gritstone Bio	COVID-19 vaccine, Gritstone Bio-2	Phase I Clinical Trial	Vaccine	Covid-19
Healx	sulindac, Healx	Phase II Clinical Trial	Small molecule	Neurology
HemoShear Therapeutics	HST-5040	Phase II Clinical Trial	Small molecule	Metabolic
HiFiBiO Therapeutics	HFB-301001	Phase I Clinical Trial	Antibody	Oncology
HiFiBiO Therapeutics	HFB-200301	Phase I Clinical Trial	Antibody	Oncology
HiFiBiO Therapeutics	HFB-30132A	Phase I Clinical Trial	Antibody	Covid-19
InSilico Medicine	INS018-055	Phase I Clinical Trial	Small molecule	Respiratory
Landos Biopharma	omilancor	Phase II Clinical Trial	Small molecule	Metabolic
Landos Biopharma	NX-13	Phase I Clinical Trial	Small molecule	Anti-inflammatory
Landos Biopharma	LABP-104	Phase I Clinical Trial	Small molecule	Immunology

Company	Generic Drug Name	Global Status	Modality	Therapeutic area
METiS Pharmaceuticals	central nervous system disease therapy, METiS Pharmaceuticals	Phase I Clinical Trial	Small molecule	Neurology
Neumora Therapeutics	NMRA-140	Phase II Clinical Trial	Small molecule	Mental health
Neumora Therapeutics	NMRA-511	Phase I Clinical Trial	Small molecule	Neurology
Nimbus Therapeutics	firsocostat	Phase II Clinical Trial	Small molecule	Oncology
Nimbus Therapeutics	NDI-034858	Phase II Clinical Trial	Small molecule	Oncology
Nimbus Therapeutics	NDI-101150	Phase II Clinical Trial	Small molecule	Oncology
Nobias Therapeutics	fasoracetam, Nobias Therapeutics	Phase II Clinical Trial	Small molecule	Neurology
Pharos iBio	PHI-101	Phase I Clinical Trial	Small molecule	Oncology
Recursion Pharmaceuticals	ruboxistaurin mesylate	Phase I Clinical Trial	Small molecule	Neurology
Recursion Pharmaceuticals	REC-2282	Phase III Clinical Trial	Small molecule	Oncology
Recursion Pharmaceuticals	REC-4881	Phase II Clinical Trial	Small molecule	Oncology
Recursion Pharmaceuticals	REC-994	Phase II Clinical Trial	Small molecule	Neurology
Recursion Pharmaceuticals	REC-3964	Phase I Clinical Trial	Small molecule	Infectious disease
Relay Therapeutics	RLY-2608	Phase I Clinical Trial	Small molecule	Oncology
Relay Therapeutics	RLY-4008	Phase I Clinical Trial	Small molecule	Oncology
Schrödinger	SGR-1505	Phase I Clinical Trial	Small molecule	Oncology
SOM Biotech	tolcapone, SOM Biotech	Phase II Clinical Trial	Small molecule	Neurology
SOM Biotech	bevantolol, SOM Biotech	Phase II Clinical Trial	Small molecule	Neurology
SOM Biotech	SOM-1311	Phase I Clinical Trial	Small molecule	Metabolic
SOM Biotech	prexasertib, SOM Biotech	Phase I Clinical Trial	Small molecule	Covid-19
Valo Health	SAR-407899	Phase II Clinical Trial	Small molecule	Analgesic
Valo Health	OPL-0301	Phase II Clinical Trial	Small molecule	Cardiovascular
Verge Genomics	VRG-50635	Phase I Clinical Trial	Small molecule	Covid-19

Wellcome supports science to solve the urgent health challenges facing everyone. We support discovery research into life, health and wellbeing, and we're taking on three worldwide health challenges: mental health, infectious disease, and climate and health.

Wellcome Trust, 215 Euston Road, London NW1 2BE, United Kingdom
T +44 (0)20 7611 8888, E contact@wellcome.org, wellcome.org