Pharmaceutical Medicine
https://doi.org/10.1007/s40290-023-00504-9

LEADING ARTICLE

Potential of Artificial Intelligence to Accelerate Drug Development for Rare Diseases

Giulio Napolitano[1] · Canan Has[1] · Anne Schwerk[2] · Jui-Hung Yuan[2] · Carsten Ullrich[2]

Accepted: 16 October 2023

Abstract

The growth in breadth and depth of artificial intelligence (AI) applications has been fast, running hand in hand with the increasing amount of digital data available. Here, we comment on the application of AI in the field of drug development, with a strong focus on the specific achievements and challenges posed by rare diseases. Data paucity and high costs make drug development for rare diseases especially hard. AI can enable otherwise inaccessible approaches based on the large-scale integration of heterogeneous datasets and knowledge bases, guided by expert biological understanding. Obstacles still exist for the routine use of AI in the usually conservative pharmaceutical domain, which can easily become disillusioned. It is crucial to acknowledge that AI is a powerful, supportive tool that can assist but not replace human expertise in the various phases and aspects of drug discovery and development.

Key Points

- Artificial intelligence has witnessed an explosion of research and applications in the biomedical domain, but special challenges affect drug development for rare diseases.
- These challenges are mainly driven by data paucity and high costs, and AI can be effectively employed to address them.
- Experts should regard AI as a powerful tool to support, not replace them in their drug development endeavours.

1 Introduction

The term artificial intelligence is not well defined, and it encompasses systems and technologies which may be viewed as ‘intelligent’ today but not tomorrow, by some people but not others. It is currently an umbrella term which, at a high level, includes systems that seem to behave as humans in quality (e.g. display learning or reasoning capabilities) but exceed them in performance in some respects (e.g. in the speed of solving a given problem or in the amount of data they are able to process). In the biomedical domain, artificial intelligence applications are those which employ machine learning techniques, especially deep learning methods, as well as those combining advanced statistical techniques with other AI approaches such as classifiers, reasoners or knowledge representation [1]. Machine learning encompasses computational algorithms and statistical models that enable computers to autonomously learn patterns from ‘training’ data, without explicit programming. Deep learning is a subfield of machine learning inspired by the human brain structure that utilises artificial neural networks, multiple interconnected layers of ‘neurons’ to comprehend intricate patterns within data. Deep learning and machine learning have enabled remarkable scientific advancements, as complex patterns could be learned just by using big amounts of training data, enabling complex models to be built. Yet, a huge challenge of the models, especially in the medical domain is their black box character, as the knowledge-driven approach to modelling (deductive reasoning) is replaced with data-driven modelling (inductive reasoning).

AI can empower medical technologies towards applicable solutions in clinical practice, with current AI-based medical applications ranging from the prediction of disease risk based on electronic patient records [2] to target identification [3], computer-assisted surgery [4] and pain management [5]. In the domain of drug development,[1] deep learning as

Extended author information available on the last page of the article

[1] By ‘drug development’ we mean the full process leading to the marketing of a new medication for a condition, including the drug discovery steps.

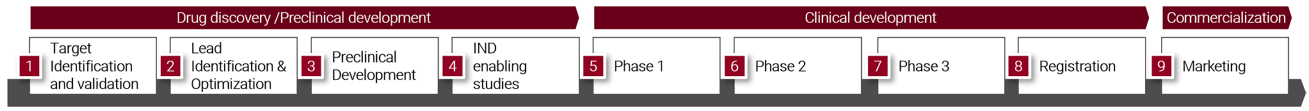

Fig. 1 The drug development process

well as traditional machine learning algorithms have also led to a range of applications. These include the prediction of compound properties, the design of new drug candidates and the in silico monitoring of drug efficacy [6].

1.1 The Drug Development Process

The drug development process is a multi-step process that involves the discovery, pre-clinical research, clinical trials and regulatory approval of a new pharmaceutical product (see Fig. 1).

The discovery stage (Fig. 1 (1)) is focused on identifying new potential drug candidates. In the target-based drug discovery process, on which we focus here, the initial phases are target discovery and validation: researchers identify potential biological targets, such as enzymes or receptors, that are involved in a disease process. This often includes the analysis of large amounts of genomic, transcriptomic and proteomic data to identify genes, proteins or pathways that are dysregulated in a diseased state. It also includes the analysis of protein–protein interactions to identify groups of proteins that are highly interconnected and are, therefore, potential drug targets. These targets are then validated through various experimental techniques, such as genetic, biochemical and pharmacological studies, to ensure that they are relevant to the disease and amenable to drug development.

After validation, the next stage of drug development involves a sequence of hit discovery, hit-to-lead and lead optimization (Fig. 1 (2)). In this phase, potential small molecules, peptides or antibodies that can bind to a validated target are identified, characterised and optimised to improve their therapeutic properties.

Subsequently, the pre-clinical stage (Fig. 1 (3 and 4)) involves evaluating the safety, efficacy and pharmacokinetics of the lead compound in in vitro and in vivo models. Then, the clinical stage concerns the evaluation of the safety and efficacy of the drug candidate in human subjects through a multi-stage process involving phase I, phase II and phase III clinical trials (Fig. 1 (5–7)). Finally, the regulatory stage concludes the process with the submission of a new drug application to the regulatory agency for review and approval (Fig. 1 (8)), until final marketing (Fig. 1 (9)).

In this article we mainly focus on the employment of AI to support target identification, validation and lead identification. For more details on clinical trial support, we refer the reader to [7]. Here, we also focus on rare diseases, which are underrepresented in basic and clinical research.

After a brief introduction to the requirements for the successful application of AI in medicine, and drug development in particular, we focus more on the specific challenges posed by rare diseases. We then point to the obstacles that AI has been facing in the biomedical domain, before listing the most successful industry actors and the drug development programmes currently active. We conclude with a summary of the central points we want to make, with an indication of what may be the most fruitful AI approach in the near future, as well as a short description of the contribution that generative AI has provided so far in this context.

2 Artificial Intelligence for Drug Development

Several application fields for AI thus far have shown promise in improving the accuracy of diagnostic systems, identifying patterns in medical data to create personalised treatments, automating tedious administrative tasks, enabling personalised preventative care and helping to detect disease outbreaks [8–10]. Significantly, all of these applications have required large volumes or diverse datasets to allow breakthroughs by AI and only few disciplines may implement solutions yet. Diagnostics is one of those fields that has benefitted from large digital datasets, especially in the domains of genomics and imaging. Many companies, such as Siemens Healthineers, pioneered the German market with radiology-based AI tools in 2019 to enable the detection of diseases in radiology images, thanks to the large amounts of data available [11].

Most efforts in the application of AI methods to drug target identification have been devoted to the exploration of novel approaches to the detection of signals originating from the various ‘omics’ layers, including genomics, transcriptomics, proteomics and metabolomics as the main focus. The complexity of interactions of entities (from genes to metabolites) within and across layers has stimulated two main approaches, not necessarily disjointed [3]. One approach is based on the analysis of the networks, in which the entities can be represented in their biological relations, to identify patterns and crucial relationships that may indicate potential drug targets. Another is based on machine learning, which is able to discriminate the relevant dataset features that contain the relevant signals from the potential targets. Both approaches have pros and cons, but all techniques rely on the integration of datasets which can be conceived as 'interconnected', by their representing interlinked information of a single biological entity (a cell, an individual or a population). In this holistic view, the hope is that the latent properties of a druggable target that affects the pathology of interest can be discovered by reinforcing their signals.

After a promising target is identified, AI applications have been developed in the subsequent drug development phases of hit discovery and lead compound optimisation. Based on purely the molecular descriptors/fingerprints of the ligands (ligand based) or the complete structure of the ligand-target binding site (structure based), a wide range of machine learning (ML) methods have been applied for quantitative structure–activity relationship modelling and prediction of binding affinities. These can be used as scoring functions to accelerate virtual screening [12]. With the advancement of deep reinforcement learning (RL), in which the algorithm is 'rewarded' when it is pursuing promising avenues, various neural network architectures have been combined for de novo drug design [13], which aims to support selection of novel compounds with optimised property profiles. These structure-based approaches can also be employed for drug repurposing. Another more knowledge-driven approach to achieve drug repositioning is to represent the biological knowledge as a heterogeneous network and to apply graph-based algorithms to discover novel drug–target interactions [14].

In recent years, a significant number of biotech companies have also been making AI part of their core R&D strategy for drug development and the anticipated rise of AI applications in drug discovery processes in the pharma industry is USD 5 billion [10]. Drug development processes are being advanced by AI as we obtain increasing volumes of data from molecular structures, high-throughput screening, compound databases and cell culture experiments, which are usually not as stringently restricted by data protection regulations as most other types of medical datasets. One example of such advancements is provided by the results obtained by Insilico Medicine. With the help of generative AI, supporting all the steps of the pre-clinical drug discovery process, a drug candidate has recently entered phase 2 clinical trials to treat idiopathic pulmonary fibrosis [15]. While rare diseases are, by definition, characterised by the paucity of patients, each patient may be described by huge amounts of omics data. Can this imbalance be compensated by the integration of datasets and knowledge bases, with the injection of expert-generated biological understanding?

3 Challenges in Drug Development for Rare Diseases

A rare disease (RD) is a disease which is seriously life-quality limiting or life threatening and affects a small portion of the population (e.g. with prevalence of less than 1 in 2000 people, although the definition varies [16]). Around half of the RDs affect children, a third of them will not see their fifth birthday [16]. Between 5000 and 8000 RDs have been described worldwide [16], among those only 5% have a treatment [17]. Most RDs are genetic in origin and one of the biggest challenges in RD starts with diagnosis, as almost 5 years elapse in average before an accurate diagnosis is received (when it is received) [17]: so far, most AI-based applications have been used to accelerate diagnosis of RDs. Although ~400 million people are affected by rare genetic diseases worldwide, each rare disease and corresponding phenotype can only be found in relatively few patients [17]. This causes severe challenges to the collection of sufficient amounts of data for research and for designing adequate clinical trials [18]. Hence, demonstrating effectiveness of potential treatments is difficult. To date, even though more than 3000 genes have been identified against over 4000 monogenic diseases [19], the number of approved therapeutics remains extremely limited. This is probably due to a lack of research and funding that hinder the application of modern gene therapies.

3.1 Small Populations

The limited size of rare disease patient populations available for research and clinical trials makes it difficult to gather enough data to identify potential targets and develop effective treatments. Moreover, the small size of these populations often makes the data inherently identifiable, especially when geographical information is included, making open data approaches and the sharing of data complicated. In addition, developing drugs for rare diseases can be very costly, as the small patient population means that there is little financial incentive for pharmaceutical companies to invest in research and development. AI can help to address some of these challenges by analysing multi-dimensional and complex datasets, i.e. genomic or proteomic data, and combining these into so-called multi-omics datasets. Such 'deep' and comprehensive datasets allow for more individual and specific target discoveries as, despite their genetic components, rare diseases can be highly heterogeneous in the expression [20, 21]. Many genetic targets are easily missed due to the failure to detect them based on solely genetic information [20]: via the integration of multi-faceted data by AI, precise subgroups and targets can be highlighted and efficiently explored. Another approach is the use of transfer learning, which has already been successfully employed in other situations where there is great disparity in size among datasets. Transfer learning is a specific machine learning technique: the training is performed with a large dataset, which is available for a domain close to the one of interest, then the resulting model is refined by further training on the smaller dataset available for the target domain. This technique has already been used to characterise disease patterns of rare diseases from transcriptomics, by transfer learning the general patterns that had already been modelled for more common conditions [22].

3.2 Limited Disease Understanding

Next to the challenge of small patient populations, there is also a limited understanding of rare disease mechanisms, as the underlying pathomechanisms are complex, diseases are heterogeneous and research is limited [23].

In some tasks, the AI methodology is applicable for both common diseases and rare diseases, such as the prediction of pharmacokinetic properties based on known compound structures, or structure-based drug design when the target candidate is certain for the disease of interest [24]. However, the pathogenesis of rare diseases tends to not be well understood and can be associated with multiple genetic variants. Often, there are distinct genetic, molecular and clinical characteristics even for a single disease, which poses challenges in developing targeted therapies that can be effective across different subtypes of a rare disease. Knowledge-driven approaches that extract information from various clinical and experimental data sources, and integrate them as a connected context, would allow better reasoning and enable advancements in drug repurposing and polypharmacology for rare diseases [25].

As a result of both the scarcity of clinical trial data and the heterogeneity in response patterns, obtaining regulatory approval for rare disease drugs can be particularly challenging.

3.3 High Costs

Finally, when a drug reaches the stage of market approval, the associated costs can be extremely high due to the massive development expenses: manufacturers face obstacles in pricing and recouping investments in rare disease drugs due to the limited patient population that will benefit from them. This is also accompanied by ethical concerns, as the society may have to agree for the public reimbursement of extremely expensive drugs, to ensure affordable access to these treatments for a relatively small number of patients.

The cost challenge for rare disease development efforts may be ameliorated by the repurposing of drugs, that is the identification of new uses for existing drugs [9, 10]: the exploration of already approved drugs may bring immense time and cost savings. AI can analyse data on the activity of existing drugs to identify new potential uses for them in the treatment of rare diseases. Often, such repurposing approaches benefit from data enriched with different levels of omics information, to gain additional insights into the mechanistic actions on the targets and to ensure safety and efficiency of candidates [21]. Another enrichment approach was shown by a Dutch team [26] that combined data with expert knowledge to generate powerful repurposed drug candidates. The team developed a repurposing platform that includes a repository of 245,000 DNA samples combined with detailed phenotype information, to enable data-driven queries on the association of genetic variation with symptom data. This approach allows one to find variants in drug-target genes and then look for the phenotypic patterns associated with those variants, using the results to extrapolate the findings to other indications that might benefit from the same drug. The expert knowledge that is integrated comes from the description of the clinical pathophysiology. Through this approach, 13 drug–disease pairings have already been developed, 6 of which have been confirmed. Lee et al. proposed a ML-based approach which integrates genetic- and molecular mechanism-related information about complex disorders as well as rare diseases [27]. The approach predicted cisplatin and resveratrol as potential repurposed drugs for anaemia with excess blasts (ORPHA: 86839) and sideroblastic anaemia (ORPHA:1047) which need to be validated experimentally.

3.4 Obstacles to the Adoption of AI in Drug Development

A general challenge, that also applies to common diseases, is the growing need for explainable AI methods. Most AI models behave like 'black boxes', producing results without any indication of the reasoning behind them or any suggestion of what portions of input data or information support those results. Using black-box AI models in the development of new drugs may prevent the proper understanding of the exact properties relevant for treating the specific (rare) disease. This would be particularly relevant for the early phases in the drug development process, such as target identification and pre-clinical target validation, as the understanding of the biological pathways that contribute to target selection allows researchers to prioritise potential drug candidates more effectively. Explainability methods include sophisticated approaches such as Shapley values, which can account even for interacting features but come at the expense of high computational loads [28], or more common approaches such as layer-wise relevance propagation (LRP) [29]. The specific type of sought explainability depends on the use case and should be examined carefully to serve the underlying goal at hand. In some cases, it is important to unravel the biological relevance of candidate targets, in others, it is the mechanisms of action of developed drugs, for instance when there is a need to enable further refinement or repurposing of compounds.

Additional challenges arise from the attitude towards *AI* in biomedical circles and are exacerbated in the case of rare diseases. On the one hand, unrealistic expectations of the abilities of AI turn quickly into disillusionment, when the limitations of AI applications become clearer: AI systems can hardly operate in the presence of limited data, and operate badly with low-quality data. This is often the case with rare diseases. On the other hand, professionals in the usually conservative domain of biomedical and pharma industry are well aware that, despite the enormous, growing interest and investment [30] and isolated success stories, these have not yet translated into widespread use and deployments of AI solutions [31].

4 Main Actors and Success Stories

Despite these difficulties, a few new companies that focus on using AI for drug development are also active in the rare disease space. See the infobox for the most prominent examples.

Infobox: Examples of companies using AI for drug discovery

HealX, founded in 2014, is a company that is using AI to develop therapeutics for a dozen rare diseases, currently in the pre-clinical stage. Their approach is based on a knowledge graph that integrates knowledge extracted from scientific literature, commercial information and bio databases, as well as the analysis of curated biological data. Hypotheses inferred from the knowledge graph are then reviewed by human experts.

- HealX has around 18 candidates in its pipeline. Their candidates for neurofibromatosis type 1 (cutaneous neurofibroma and plexiform neurofibroma) and Fragile X syndrome are the most advanced along the drug discovery process and are now in the clinical planning phase.

Insilico Medicine, founded in 2014, uses AI to develop therapeutics for rare diseases such as cystinosis and idiopathic pulmonary fibrosis. They are currently in the pre-clinical stage for cystinosis and in the clinical phase 1 for idiopathic pulmonary fibrosis.

- A cooperation with CENTOGENE focused on identifying, ranking and annotating proteins to identify novel therapeutic targets for Niemann-Pick disease type C. Again, the AI plays a supporting role. Insilico Medicine's target discovery AI platform was used on CENTOGENE's multi-omics patient data to identify differential metabolites, pathways and genetic modifiers from the data. Experts then analysed the identified targets and validated them using CENTOGENE's cellular models.
- Insilico Medicine utilises various AI methods for different purposes: generative adversarial networks to generate novel molecular structures; reinforcement learning to optimise molecule properties; deep learning to analyse large datasets to identify patterns that inform how molecules can be designed or improved; and various other machine learning methods for tasks like predicting toxicity, protein–ligand interactions and more.
- Insilico Medicine has 29 targets under development and developed the candidate most advanced in the drug discovery process: Their candidate for idiopathic pulmonary fibrosis has advanced to a phase 2 clinical trial to evaluate effectiveness and safety.

Infobox: Examples of companies using AI for drug discovery

Insitro, founded in 2018, aims to overcome the sparsity of data available for machine learning by setting up cell-based disease models that generate data specifically aiming at being used by machine learning. Insitro has not published any details on potential drug candidates nor on the AI methods they are using.

BenevolentAI, founded in 2013, unites AI with expertise of wet-lab scientists to discover new medicines for complex diseases. The BenevolentAI computational pipeline covers all steps of drug discovery and has currently been applied on over 20 diseases.

- BenevolentAI has not published details on their drug candidates. Their most advanced asset is for ulcerative colitis, with a phase I trial to start in the UK in 2023.
- BenevolentAI combines a wide array of AI technologies: knowledge graphs to integrate and analyse biomedical data relationships, natural language processing to extract insights from scientific literature, machine learning for predicting properties and activities, neural networks to identify novel targets and generate molecular structures, and causal inference and reasoning to determine mechanisms of action.

Table 1 lists the AI-based drug development programmes currently in the clinical trial phases, along with the lead indications of the potential first-in-class compounds [32]. Among them, familial adenomatous polyposis and neurofibromatosis type 2 are rare diseases [33, 34].

An area where AI has already proven its value is the diagnosis of rare diseases based on very different types of input, ranging from genetics to phenotypes. Using next-generation sequencing data, machine learning algorithms are able to identify disease-causing genetic variants with near-perfect specificity and sensitivity [35]. Such algorithms are used today by several companies for diagnosis. On the other end of the spectrum, deep convolutional neural networks (machine learning algorithms specifically well suited to analysis of sensory data such as images) are able to recognize a rare disease based on portrait photos [36], again with near-perfect accuracy. We are still far away from such figures in AI for drug development. However, with an increasing number of high-quality datasets and a better understanding of how to model the biological complexity of a disease, we expect radical changes in 5–10 years: AI will be incorporated in the standard drug development pipelines and significantly improve efficiency in its various steps.

5 Conclusive Remarks and Outlook

AI has been increasingly used for drug development, but its application is still in its early phase. The drug development process is a highly complex domain, and AI has yet to fully demonstrate its capabilities in this field.

One of the main challenges of using AI for drug development is the complexity of biological systems. The size and feedback loops of pathways in living organisms are highly intricate

Table 1 List of diseases and corresponding AI-developed drugs, currently in clinical trials

Drug ID	Company	Description	Clinical phase	Disease
REC-2282	Recursion	Small-molecule pan-HDAC inhibitor	2/3	Neurofibromatosis type 2
REC-994	Recursion	Small-molecule superoxide scavenger	2	Cerebral cavernous malformation
REC-4881	Recursion	Small-molecule inhibitor of MEK1 and MEK2	2	Familial adenomatous polyposis
INS018_055	InSilico Medicine	Small-molecule inhibitor	2	Idiopathic pulmonary fibrosis
BEN-2293	BenevolentAI	Topical pan-tyrosine kinase inhibitor	2a	Atopic dermatitis
EXS-21546	Exscientia and Evotec	A2A receptor antagonist	1b/2	Solid tumours carrying high adenosine signatures
RLY-4008	Relay Therapeutics	Inhibitor of FGFR2	1/2	FGFR2-altered cholangiocarcinoma
EXS-4318	Exscientia	PKC-θ inhibitor	1/2	Inflammatory and autoimmune conditions
BEN-8744	BenevolentAI	Small molecule PDE10 inhibitor	1	Ulcerative colitis
Undisclosed	Recursion	Small molecular inhibitor of RBM39, a CDK12-associated protein	Pre-clinical	HRD-negative ovarian cancer

The companies which developed the drug candidates are also given

and not fully understood. Additionally, the limited knowledge about these systems and the limited amount of high-quality data available make it difficult to apply machine learning or to use knowledge-based approaches, such as semantic networks, to manually build reliable models of pathways.

These challenges are even more pronounced when it comes to the development of therapeutics for rare diseases. Due to the small patient populations, there is a limited amount of data and knowledge available for AI models to work with. Furthermore, the small market for rare disease drugs requires careful financial considerations.

To answer our earlier question, we do believe that, in the near future, the most robust advancements will be brought by the principled integration of tabular datasets and biological knowledge. The interactive nature of biological entities suggests that network representation of this knowledge will be central. Ideally, this integration will also be guided by expert-crafted rules, e.g. logical rules that reflect how an expert would reason in choosing or discarding candidate drugs for repositioning. This has already been demonstrated with, again, the employment of reinforcement learning [37]. In particular, the problem of finding a new indication for a drug has been designed as a link prediction problem. This is a typical approach when using knowledge represented in a graph format: a (candidate) discovery is achieved by the prediction of a missing link between entities in the graph, in this case a link between a compound and a disease. In the search for new links, reinforcement learning is employed to reward those paths in the graph, which already connect drugs to their known indications. To achieve higher generalisation and increase the probability to find longer distance dependencies, logical rules are added to the training process to guide the search over longer but meaningful paths. One rule, for instance, rewards those paths that link a compound to a disease, where the compound is in the same pharmacological class of another compound already known to treat the disease. Such approaches will prove essential for rare disease, where the injection of context and expert knowledge could compensate for the lack of data.

A recent new development in artificial intelligence is generative AI. ChatGPT [38] is a well-known example of generative AI: after a text input, the so-called prompt, it generates new text—in contrast to search engines such as Google, which search for and display existing texts in which the input occurs. But generative AI is not limited to the medium of text. DALL-E2 generates unprecedented images to given prompts; other AIs generate text describing a given image. Multi-modal generative AI, for example, combines image and sound to generate video. In the life sciences, practical use of generative AI is still in its infancy. Some areas where progress has been made, relevant to the pharmacological domain, are in the Infobox.

Infobox: Generative AI

- Protein property prediction: Meta's ESM-1 and ESM-2 are generative AIs trained on millions of proteins to learn how amino acid sequences affect protein structure and function [39].
- Small molecule creation: MegaMolBart [40], a generative chemistry AI, is being used to create and optimise new small molecules that are being tested for their ability to bind to other molecules.
- Protein creation: ProtGPT2 can create new protein sequences that can have unprecedented shapes and functions [41].
- Molecular docking: DiffDock can be used to predict how a molecule will bind to a protein [42].

Probably the best-known example of generative AI in drug discovery is AlphaFold2, which predicts the 3D structure of a protein based on its amino acid sequence with high accuracy [43]. In the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition in 2020, AlphaFold2 achieved top scores. Indeed, AlphaFold2 has revolutionised protein structure prediction and practically solved the problem, which will drive the development of safer and more effective drugs.

It is, however, important to recognize that AI is not a 'golden bullet' solution and cannot replace the expertise of human researchers. Instead, AI can be used as a supportive tool to assist human experts in various aspects of drug development. It provides them with capabilities to analyse large amounts of data, extract information from scientific literature, and build and analyse networks. Furthermore, AI can build predictive models and identify new leads for further research. In this way, AI helps researchers to gain new insights and accelerate the discovery of therapeutics, also for rare diseases.

Declarations

Funding No external funding was provided for this article.

Conflict of Interest GN and CH are employees, AS, JHY and CU are former employees of Centogene GmbH.

Availability of Data and Material Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Code Availability Not applicable.

Author Contributions GN and CU conceived the initial idea. All authors contributed to most sections. All authors have read and approved the final version of the manuscript and agree to be accountable for the work.

References

1. Vijayan RSK, Kihlberg J, Cross JB, Poongavanam V. Enhancing preclinical drug discovery with artificial intelligence. Drug Discovery Today. 2022;27(4):967–84. https://doi.org/10.1016/j.drudis.2021.11.023.
2. Mortazavi BJ, et al. Analysis of machine learning techniques for heart failure readmissions. Circulation. 2016;9(6):629–40. https://doi.org/10.1161/CIRCOUTCOMES.116.003039.
3. You Y, Lai X, Pan Y, Zheng H, Vera J, Liu S, et al. Artificial intelligence in cancer target identification and drug discovery. Signal Transduct Target Ther. 2022;7(1):156.
4. Overley SC, et al. Navigation and Robotics in Spinal Surgery: Where Are We Now? Neurosurgery. 2017;80(3):S86–99. https://doi.org/10.1093/neuros/nyw077.
5. Malloy KM, Milling LS. The effectiveness of virtual reality distraction for pain reduction: a systematic review. Clin Psychol Rev. 2010;30(8):1011–8. https://doi.org/10.1016/j.cpr.2010.07.001.
6. Jiménez-Luna J, et al. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discovery. 2021;16(9):949–59. https://doi.org/10.1080/17460441.2021.1909567.
7. Woo M. An AI boost for clinical trials. Nature. 2019;573(7775):S100–2. https://doi.org/10.1038/d41586-019-02871-3.
8. Pandya S, Thakur A, Saxena S, Jassal N, Patel C, Modi K, Mehta P, Gonge S, Kadam K, Kadam P. A study of the recent trends of immunology: key challenges, domains, applications, datasets, and future directions. Sensors. 2021. https://doi.org/10.3390/s21237786.
9. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discovovery. 2019;18:463–77.
10. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discovery Today. 2021;26(1):80–93. https://doi.org/10.1016/j.drudis.2020.10.010. (**Epub 2020 Oct 21**).
11. Alexander A, Jiang A, Ferreira C, Zurkiya D. An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. J Am Coll Radiol. 2020;17(1):165–70.
12. Carpenter KA, Huang X. Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: a review. Curr Pharm Des. 2018;24(28):3347–58. https://doi.org/10.2174/1381612824666180607124038.
13. Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, et al. Advances in de novo drug design: from conventional to machine learning methods. Int J Mol Sci. 2021;22(4):1676.
14. Zhang Z, Chen L, Zhong F, Wang D, Jiang J, Zhang S, et al. Graph neural network approaches for drug-target interactions. Curr Opin Struct Biol. 2022;73: 102327.
15. "First drug discovered and designed with generative AI enters Phase II trials, with first patients dosed", Business Announcement by EurekAlert!, 2023 https://www.eurekalert.org/news-releases/993844. Accessed 20 Jul 2023.
16. Adachi T, El-Hattab AW, Jain R, Nogales Crespo KA, Quirland Lazo CI, Scarpa M, Summar M, Wattanasirichaigoon D. Enhancing equitable access to rare disease diagnosis and treatment around the world: a review of evidence, policies, and challenges. Int J Environ Res Public Health. 2023;20(6):4732. https://doi.org/10.3390/ijerph20064732.
17. "RARE Disease Facts". Information page by Global Genes, 2023 https://globalgenes.org/learn/rare-disease-facts/. Accessed 22 Aug 2023.
18. Shah KK, Kogut S, Slitt A. Challenges in evaluating safety and efficacy in drug development for rare diseases: a review for pharmacists. J Pharm Pract. 2021;34(3):472–9. https://doi.org/10.1177/0897190020930972.
19. Ehrhart F, Willighagen EL, Kutmon M, van Hoften M, Curfs LM, Evelo CT. A resource to explore the discovery of rare diseases and their causative genes. Scientific Data. 2021;8(1):124.
20. Bottini S, Emmert-Streib F, Franco L. Editorial: AI and multi-omics for rare diseases: challenges, advances and perspectives. Front Mol Biosci. 2021;30(8): 719978. https://doi.org/10.3389/fmolb.2021.719978.
21. Kerr K, McAneney H, Smyth LJ, et al. A scoping review and proposed workflow for multi-omic rare disease research. Orphanet J Rare Dis. 2020;15:107. https://doi.org/10.1186/s13023-020-01376-x.
22. Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, Greene CS. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. Cell Syst. 2019;8(5):380–94. https://doi.org/10.1016/j.cels.2019.04.003.
23. Fonseca DA, Amaral I, Pinto AC, Cotrim MD. Orphan drugs: major development challenges at the clinical stage. Drug Discovery Today. 2019;24(3):867–72. https://doi.org/10.1016/j.drudis.2019.01.005.

24. Govindaraj RG, Naderi M, Singha M, Lemoine J, Brylinski M. Large-scale computational drug repositioning to find treatments for rare diseases. NPJ Syst Biol Appl. 2018;4(1):13.
25. Alves VM, Korn D, Pervitsky V, Thieme A, Capuzzi SJ, Baker N, et al. Knowledge-based approaches to drug discovery for rare diseases. Drug Discovery Today. 2022;27(2):490–502.
26. Challa AP, Zaleski NM, Jerome RN, Lavieri RR, Shirey-Rice JK, Barnado A, et al. Human and machine intelligence together drive drug repurposing in rare diseases. Front Genet. 2021;12: 707836.
27. Lee YS, Krishnan A, Oughtred R, et al. A computational framework for genome-wide characterization of the human disease landscape. Cell Syst. 2019;8(2):152-162.e6. https://doi.org/10.1016/j.cels.2018.12.010.
28. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.); 2022. christophm.github.io/interpretable-ml-book/
29. Yang KY, Tresp V, Wunderle M and Fasching PA. Explaining therapy predictions with layer-wise relevance propagation in neural networks. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 2018, p. 152–62. https://doi.org/10.1109/ICHI.2018.00025.
30. "AI continues to gain momentum in the biopharmaceutical industry in 2023" in Pharmaceutical Technology, 2023. https://www.pharmaceutical-technology.com/comment/ai-momentum-biopharmaceutical/. Accessed 29 Jan 2023.
31. Zhang J, Budhdeo S, William W, et al. Moving towards vertically integrated artificial intelligence development. NPJ Digital Med. 2022;5:143. https://doi.org/10.1038/s41746-022-00690-x.
32. Arnold C. Inside the nascent industry of AI-designed drugs. Nat Med. 2023;29:1292–5. https://doi.org/10.1038/s41591-023-02361-0.
33. Mitchem JB, Hall JF. Adenomatous polyposis syndromes: diagnosis and management. Clin Colon Rectal Surg. 2016;29(4):321–9. https://doi.org/10.1055/s-0036-1584089.
34. Tamura R. Current understanding of neurofibromatosis type 1, 2, and Schwannomatosis. Int J Mol Sci. 2021;22(11):5850. https://doi.org/10.3390/ijms22115850.
35. Bauer P, Kandaswamy KK, Weiss MER, et al. Development of an evidence-based algorithm that optimizes sensitivity and specificity in ES-based diagnostics of a clinically heterogeneous patient population. Genet Med. 2019;21:53–61. https://doi.org/10.1038/s41436-018-0016-6.
36. Hsieh TC, Bar-Haim A, Moosa S, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nat Genet. 2022;54:349–57. https://doi.org/10.1038/s41588-021-01010-x.
37. Liu Y, Hildebrandt M, Joblin M, Ringsquandl M, Raissouni R, Tresp V. Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. In: et al. The Semantic Web. ESWC 2021. Lecture Notes in Computer Science, vol 12731. 2021; Springer, Cham. https://doi.org/10.1007/978-3-030-77385-4_22
38. Benoit JR. ChatGPT for Clinical Vignette Generation. Revision, and Evaluation. 2023. https://doi.org/10.1101/2023.02.04.23285478.
39. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123–30. https://doi.org/10.1126/science.ade2574.
40. "A deep learning model for small molecule drug discovery and cheminformatics based on SMILE", in GitHub, 2023, https://github.com/NVIDIA/MegaMolBART, Accessed 29 Aug 2023
41. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022. https://doi.org/10.1038/s41467-022-32007-7.
42. Corso G, Stärk H, Jing B, Barzilay R & Jaakkola T. DiffDock: diffusion steps, twists, and turns for molecular docking. arXiv. 2022. https://doi.org/10.48550/ARXIV.2210.01776
43. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.

Authors and Affiliations

Giulio Napolitano[1] · Canan Has[1] · Anne Schwerk[2] · Jui-Hung Yuan[2] · Carsten Ullrich[2]

✉ Giulio Napolitano
gnapolitano01@qub.ac.uk

Canan Has
canan.has@centogene.com

Anne Schwerk
anne.schwerk@charite.de

Jui-Hung Yuan
juihung.yuan@gmail.com

Carsten Ullrich
ullrich@carstenullrich.net

1 Centogene GmbH, Alboinstraße 36-42, 12103 Berlin, Germany

2 Beriln Institute of Health Center for Regenerative Therapies (BCRT), Berlin Institute of Health, Charité - Universitätsmedizin Berlin, Berlin, Germany