Multi-Target Oncology Lead Identification Through Graph Neural Networks: AI-Driven Platforms for Cancer Drug Discovery

Dr. Wai-Keung Wong, Professor of Computer Science, The Chinese University of Hong Kong (CUHK)

1. Introduction

Cancer is one of the most devastating diseases worldwide. The traditional avenue of treating patients has been through chemotherapy, but a recent innovation has transformed the growth of early 21st-century biotech companies. Drugs that act on molecular targets in tumors have revolutionized the treatment of cancer because each different drug can reduce the size of tumors at a different site. As oncology is the area most advanced in personalized medicine therapy development, the combination of a patient's cancer and novel drug therapies allows long-term cancer disease management with greater tolerability. Other areas of unmet patient needs and poor molecular diversity further necessitate new approaches to quickly identify novel mechanisms of action and their corresponding targets within the approximately 20,000 protein-encoding human genes.

Artificial Intelligence and advanced machine learning techniques are uncovering novel biological and clinical datasets at speeds once thought impossible. This has become particularly true in drug discovery, providing ways to address success odds plagued by former failed clinical programs, soaring drug development costs, insufficient pipelines, and the probability of targeting the undulating complexities of cancer. Medicine inherently involves clinical trial improvement and participation collaboration, ensuring symptom and quality data in deriving molecular phenotype outcomes congruent with an evolving view of the cancer disease state. This chapter describes how advanced machine learning techniques leverage large-scale molecular, genomic, imaging, and clinical data to directly improve drug therapy development and clinical research across the oncology landscape. Progress in oncology will be showcased, from molecular target

identification to drug discovery. In each section, a vivid case study will illustrate the promise of artificial intelligence in the translation of data into clinical insights that could have a significant impact on drug development, regulatory, and post-marketing decision-making.

1.1. Background of Drug Discovery in Oncology

Drug Discovery in Oncology: Introduction

In the last 20 years, many innovative therapies have been developed for cancer treatment. They have led to some spectacular results, notably in immunotherapy, where long-term responses have been observed. However, the drug discovery process is still long and expensive, and many phase 3 trials have failed to reach their primary endpoints. The biology of cancer is very complex, and every tumor is different between two patients, which is called intra-tumor heterogeneity. Traditional, not completely artificial intelligence-driven platforms may be used in drug discovery, but most of them are "experimental analytics" platforms. They use some features of artificial intelligence in data analyses, but the main principles remain the same as those of traditional platforms. These techniques are very useful and may help identify some potentially novel drugs, but require very large sets of pre-existing data for comprehensive predictions, similar to other experimental and computational systems already existing in this field.

Cancer is a complex, multifactorial disease characterized by abnormal cell growth leading to the invasion of new tissues. There are too many types of cancer (more than 100), and outcomes may differ significantly between two histologically equal tumors. Drug development, especially in oncology, is very time-consuming. The canons of toxicity caused the pharmaceutical industry to move toward innovative drug discovery and development processes characterized by an increasing preclinical candidate selection (from a single molecule identification only a few years back). Moreover, specialized preclinical toxicity studies are also needed to provide risk assessment in the appropriate species, evaluate mechanisms of toxicities, and potential reversibility. The drugs currently on the market have been discovered over 10 years ago. The traditional methods of preclinical and clinical development have become outdated since they do not allow for the rapid and adequate prediction of the behavior of new potential medicines. In addition, the traditional approach neglects the individualization of therapy and the possibility of its cost-effective implementation. In order to discover and

develop a small molecule, it takes 13 to 16 years and costs about 1.5 to 2.5 billion dollars. Moreover, only 1 in 10 drug candidates becomes a marketed product after preclinical development. All those mentioned above is primarily due to limited and hard-to-predict views of human responses to new drug candidates, complex biology of humans today required for understanding "next generation functional omics", high variability of diseased-damaged human tissues, and individual patients. Heterogeneity of oncology targets decreases efficacy and increases side effects by targeting the main subclonal population rather than the rare tumor cell population with drug tolerance. However, the study of drugs tested with predictive platforms is 79% correctly matched. Of these, 90% correctly predicted the cause of reinstitution, and 54% correctly distinguished the true cause of the recondition from other causes with a higher impact on efficiency and/or the sequence that provides treatment tolerability. From the above, it can be concluded that it is necessary to use effective tools that will help speed up the drug development processes. The objectives of the case study in this review were to analyze factors, processes, etc., based on the use of predictive platforms for cancer target definition.

1.2. Role of AI in Drug Discovery

Artificial intelligence (AI) stands as a potential game changer for revolutionizing how drug discovery is conducted, particularly for developing much-needed oncology therapeutics to improve patient outcomes. Multiple AI techniques, including machine learning and deep learning, are increasingly enabling researchers to analyze exceedingly large data sets more efficiently and, in the process, uncover patterns that provide insights that were previously unattainable. AI systems are being designed to handle the discovery of new therapeutics by optimizing decision paths, standardizing procedures, and predicting outcomes in a way that surpasses traditional methods. As a result, the use of an AI-based drug development pipeline is expected to significantly reduce the chance of failure in the preclinical stage by ensuring that a drug candidate reaches the clinical development phase only when it has molecular characteristics indicating potential efficacy. The predictive prowess of machine learning, in particular, has already demonstrated its ability to identify therapeutic targets and drugs with a substantially higher likelihood of success than traditional target identification methods. In the coming years, there is anticipated to be further growth in the use of AI systems to design personalized treatment plans that optimize time, cost, and patient outcomes when

combined with biomarkers that can determine which patients are most likely to respond to a particular treatment.

AI possesses even greater potential as a driver of new innovation to transform the drug discovery landscape by combining bioinformatic analyses with large data sets of clinical and molecular profiles to elucidate the underlying biology of a condition and the mechanism by which existing clinical-stage compounds exert their effects. New platforms bringing together the capabilities of AI and real-world big data increasingly promise to provide a wealth of rich commercial opportunities progressing beyond the traditional incentive-based drug discovery model. Integrating a variety of machine learning-based platforms can provide a much richer data set by combining patient data from a wide variety of sources including wearable health technology, electronic health records, and available molecular profiling. This consolidated data set can then be analyzed using such platforms to find novel relationships that might not be evident just by analysis of a fragment of the overall patient journey. AI can be an enabler to draw insight from data when pressures such as regulations make the anonymization and use of patient data for global insights more difficult.

2. Machine Learning Techniques in Drug Discovery

Machine Learning Techniques in Drug Discovery Pharmacology is a burgeoning field. Rapid increases in digital data, computational capacities, and AI have led to significant advances in drug development. Essentially, AI leverages machine learning (ML) techniques to learn from massive data sets of compounds, genes, and diseases. ML techniques are capable of prioritizing particular molecules or targets for the creation of drugs in early development stages. In this paper, we will discuss the most frequently used supervised and unsupervised ML methods in drug discovery.

Most of the techniques used are closely related to some ML methods previously widely discussed and used in artificial intelligence, including neural networks, support vector machines, Bayesian inference, random forests, decision trees, k-means, hierarchical clustering, and k-nearest neighbors. In oncology, ML methods are commonly applied to: 1) identify cancer molecular subgroups, 2) support the prediction of gene function, or 3) predict the outcome of treatments by clinicians. Supervised learning consists of the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It has this meaning only in the context of training data: the

'labeled' data set for the construction of the predictive model. Practically, supervised learning consists in modeling the dependence of various types of output variables (i.e., response, target, dependent variable(s)) on a specific combination of input features that are used as the model's 'covariates' (i.e., explanatory or independent variables).

Unsupervised learning is a branch of machine learning that requires the training of algorithms in order to model and understand the data structure or distribution. In unsupervised learning, the training data contains only features and no labeled response. The goals of unsupervised learning techniques are quite different from those in supervised learning. With unsupervised learning, the goal is to learn the underlying structure of a high-dimensional dataset without having the output. A typical task is the 'clustering' of samples based on their similarity or dissimilarity within the input feature space, in order to identify subgroups of samples characterized by specific patterns or properties. Among unsupervised learning methodologies, clustering and dimensionality reduction (i.e., extraction of 'features' from raw data) are two of the most commonly applied techniques in drug discovery. The notion of 'clustering microarray' gene expression data as different as cancers has been around for over a decade. Many efforts have been dedicated to clustering such data, revealing gene signatures that lead to the identification of 'molecular cancer subtypes', based on a specific gene signature that somehow recapitulates 'cancer type'-specific features. In unsupervised learning, the quality of any analysis depends entirely on the method's ability to identify reliably different clusters or hidden groups and, therefore, a proper, systematic validation is mandatory.

There are numerous algorithms and software tools available to obtain highly reliable clustering results, along with dynamic cut perturbation clustering, consensus clustering, single sample predictors, cluster stability, or silhouette analysis functionalities. The choice of suitable clustering algorithms for cancer biology applications predominantly pertains to the volume and nature of the input data, including the type of biological phenomenon to consider (i.e., the biological aspect on which emphasis should be placed). Model selection is a key element while defining the best performing algorithm, together with the data feature extraction, input and output scoring, and visualization of the most relevant clusters. Data feature extraction, as well as careful and systematic performance evaluation, quality validation, and statistical interpretation are critical.

2.1. Supervised Learning

Supervised learning refers to the category of machine learning tasks that involve predicting or classifying according to a specified outcome, namely when algorithms are trained on labeled input and output data using various computational procedures. The potential of being able to predict drug sensitivity or resistance makes such techniques appealing in contemporary drug discovery, particularly in the field of personalized medicine. Some of these supervised learning models for drug development that have been developed and utilized in the scientific community include random forest classifier, support vector machine, decision trees, and neural networks. By training data in these models, it is possible to predict efficacy in terms of drug sensitivity or resistance or drug sensitivity associations.

However, the performance of these models is highly dependent on the quantity, quality, and diversity of the training data. The majority of labels in the data result from high-throughput experiments, typically obtained in cancer cell lines. An increasing proportion of the efficacy labels come from the realm of patient cancer samples or patients as cell-sample products or organoids. Data from such an amalgamation can be used to facilitate the formulation of quantitative systems pharmacology models, which can be employed for mechanistic and translational analyses. Supervised learning models have been showcased to predict both drug responses in human cancer cell lines, primary cells, and xenografts, as well as drug efficacy and response biomarkers across a wide variety of cancer types, from pediatric cancers to solid tumors. A number of clinical case studies also act as justifications for using methods in drug development. However, there are drawbacks related to the use of these computational models. Most prominently, overfitting may be a major downside of such generalized approaches. Care should be taken in considering validation options, the use of robust computational procedures, ongoing curation of the models based on new data releases and outputs to ensure prediction accuracy. Furthermore, bias in the data, derived from cell line in vitro models, must be counteracted. Overall, supervised learning approaches may play a decisive role in the development of a new class of effective therapies targeting the well-known targets in cancer biology.

2.2. Unsupervised Learning

Unsupervised learning algorithms are particularly relevant in the context of biology and drug discovery. Indeed, due to their versatility, unsupervised learning approaches represent an alternative when omics features are not labeled. Under these circumstances, researchers might not be aware of the true structure of the data, so they instead try to detect hidden patterns or structures intrinsic to the data. Unsupervised approaches have two major applications in the field of drug discovery. Clustering is used to identify and subclassify patients based on the expression profile of certain genes or proteins. Dimensionality reduction is instead used for the visual inspection of the data where something is known about the training disease and to screen the layers or maps of the data to outline potential new therapeutic avenues or markers.

In terms of drug discovery in oncology, unsupervised methods help to gain insights into the complexity of the diseases, to identify new targets, and to reduce the heterogeneity that exists within tumors. Therefore, clustering approaches are of great relevance in cancer research, where they are commonly used to deconvolute the cancer hallmarks and identify new subgroups of patients that might not have been captured by the clinical and pathological data, groups of cancer genes with similar activities or functions, and clusters of patients sharing mutated genes or copy number alterations. As a result, unsupervised techniques can improve our understanding of the disease by identifying novel actionable targets that could not have been otherwise captured. A major challenge affecting unsupervised learning models, particularly in integrative medicine and drug development, is the follow-up interpretation of the discovered hidden structures. As the essential configuration of these clusters becomes more intricate, the interpretation of the samples can become perplexing because the discovery does not necessarily relate to some variable maximizing the clustering effect. More work is needed to develop interpretability methods addressing this aspect.

3. Applications of AI in Identifying Novel Cancer Therapeutics

Over the past decade, digital tools exploiting artificial intelligence (AI) have made direct impacts on drug discovery. In this section, we focus on the two interrelated applications of AI in the search for new cancer therapeutics.

First, we discuss virtual screening, which uses computational approaches to assess millions of compounds held in libraries to seek compounds with potential activity

against cancer targets. This approach has revolutionized the medicine design field by enabling a rapid and economical search for novel therapies. In particular, we review ways that modern machine learning-based AI algorithms are optimizing and enhancing this approach, focusing on improving the ability to screen candidate oncology drugs to increase the accuracy of hit identification and to more accurately predict how a molecule might interact with its target.

Next, we address de novo drug design—the development of AI algorithms that use knowledge of what a drug needs to do to interact with its molecular target to propose entirely new molecules that might be effective in cancer while having minimal effects on other cells. We provide a case study, as well as some of the lessons learned from these applications.

Overall, these applications demonstrate how AI can be used to help identify new targets for cancer therapies and design novel agents without necessarily requiring trials in humans—potentially opening up innovative rather than therapies that need to show clear clinical benefit to enter the market.

Since its origins, AI has shown increasing relevance for the field of drug discovery, offering prediction capabilities that fundamentally impact the discovery and development processes of new therapeutic agents. In particular, oncology has reaped considerable benefits from AI-driven platforms that are now used across efficient hit generation and lead optimization processes. With their focus on hit identification and hit optimization, these platforms offer new methods to rationally search for largely diverse therapeutic proteins.

3.1. Virtual Screening

Virtual screening is one of the focal points of AI applications in drug discovery and, in particular, in the context of oncology. It is an in-silico approach for the examination of a much larger number of chemical compounds and/or drug surrogates, often as part of a chemical library, or for the synthesis of a new ligand. Additionally, it incorporates approaches that impinge upon specific pharmaceutical targets. Various molecular docking algorithms and approaches may be implemented to predict the binding affinity of chemical compounds to a target of interest with substantial speed and cost reductions in various virtual screening processes. Some molecular docking approaches consider

only the compound and the target, without considering the larger biological context in which the target of interest appears.

Incorporating biological data and chemical knowledge into a molecular docking-based method is consistent with four critical elements of the drug discovery process. For example, it became apparent in the mid-1900s, through refinements to computer-aided drug design methods for the generation of efficacious compounds, that kinetics greatly impacts the biological efficacy of small-molecule drugs as well as the toxicity of these compounds. Predictive kinetics approaches such as ADME, which are well outside the pipeline of computer-aided drug design efforts but clearly impact efficacy and toxicity, have been incorporated into more accurate modeling methods. As an additional example, early computer-aided drug design efforts that focused solely on molecular docking to the active site of a single target yielded important linkage to biological effect, but these methods were modified and iteratively improved to allow the design and selection of lead compounds that operate on multiple targets and pathways to produce maximal clinical effects. In high-impact research projects, some of the development costs are often diverted towards the thoughtful computational design of the iterative process, thereby constructing disease-fighting strategies that are several steps removed from chemical structure. However, it should be noted that the iterative selection of lead series of compounds that are then selected based on in vitro screening against relevant biological targets continues to be a staple process in lead-finding efforts. Consequently, virtual screening has thus emerged as a pivotal application, and these opening statements about the role of iterative design and selection imply a drug discovery team that intentionally incorporates biological and chemical knowledge into its iterative processes around lead compound selection.

3.2. De Novo Drug Design

De novo drug design is an exciting new era in computer-aided drug design, particularly as it uses the power of artificial intelligence to revolutionize the de novo design of new cancer chemicals. Driven by the success of generative models in this era of AI, de novo drug design has the potential to revolutionize the discovery of new therapeutics in oncology, chronic disease, and beyond. De novo drug design covers the realm of recruitment and collaboration of AI, computer-aided drug design, and medicinal chemists to create new chemical structures from scratch. The rationale behind the

exploration of de novo drug design is to design novel chemical entities for the treatment of diseases. The new entities would be specifically directed to the target against which drugs are developed, such as proteins in the human body, while minimizing drug side effects. The attraction for AI data scientists and engineers in drug discovery, especially in oncology, is the aim to create molecules that are able to provoke a treatment response.

De novo drug design is exciting in AI, exploring several techniques such as creating new molecules with machine learning and statistical approaches, including generative models, reinforcement learning, and deep learning generative models. The ability to generate chemical structures without scaffolds effectively demonstrates this. The value of using deep reinforcement learning to assist in the design of novel irreversible covalent inhibitors of the protein tyrosine phosphatase is highlighted. Starting from a set of commercially available analogs, generative models were used in an iterative manner to generate novel hit series, with the trained generative model suggesting compounds for synthesis and testing between training cycles. As the compounds were synthesized iteratively, the generative model was retrained to account for newly synthesized and tested compounds. Highlighting the potential of such a design approach, the diverse set of compounds generated using generative models and the iterative approach led to a promising new class of inhibitors. In this small example, we can begin to see the change in mindset, moving from generating data sets towards generating new molecular hypotheses to test. This de novo approach to both hit optimization and lead generation will possibly make the design and test cycle more efficient and applicable to a range of problems, which in turn brings us into the future of modern AI in oncology.

4. Predicting Tumor Responses with AI

Selecting the most promising candidates for therapies and the first application of human samples is a crucial step in patient treatment. Specificity in drug response enables the implementation of targeted therapies to increase the chances of tumor response. We know that every individual patient carries a unique tumor that has evolved from a healthy cell that accumulated genetic mutations over time. Until recently, specific genetic variations in a patient were out of reach in a clinical setting. However, the application of predictive AI has recently offered new insights into the unique biological makeup of individual patients and to what degree a cancer therapy might work for patients.

AI techniques like statistical modeling and machine learning have shown great promise in learning from past treatments and patient outcome data jointly gathered from patients' clinical data, demographics, drug responses, genomic, and tumor molecular characteristics. This section describes the variety of predictive approaches and their link to patient stratification. This section also elaborates on a specific type of patient data used in predictive AI, personas. Pharmacogenomics comes as a subtopic since the presence of specific variations in the DNA can alter the way individuals actually metabolize and respond to drugs. Overall, using predictive analytics that merges preclinical, clinical, and genetic traits creates unprecedented opportunities for developing and offering tailored therapies to individual patients.

4.1. Patient Stratification

4.1. Subsection: Patient Stratification The stratification of patients into testable or observable populations is one of the key challenges in drug discovery and development. This stratification process ideally reveals patients who are more likely to respond or not respond to a new therapy against a comparator, for instance, standard of care, due to a biological subphenotype relevant for treatment outcome. In an ideal situation, a maximum number of patients can be stratified into a control or investigational regimen, leading to a higher signal-to-noise detection of the effect of an investigational molecule. It is equally important to identify a subpopulation of patients, for instance, due to a genetic propensity, that could be at an increased risk of not responding to the investigational therapy as well as to prevent treatment-related side effects. Patient stratification increases the likelihood of success of a new treatment through a more efficient use of resources and revealing high effect sizes during testing.

Patient stratification on the basis of molecular data is an integral aspect of precision medicine and a principal application of AI in clinical oncology. This subgroup of AI algorithms leverages, for example, unsupervised learning to reveal different patient groups based on molecular profiles, supervised learning in the form of deep learning or knowledge graph algorithms due to their ability to integrate biological knowledge bases, identify deregulated pathways, and reveal multiple correlated factors behind resistance in one patient subgroup, among others. One of the most classical approaches connects genomic changes to drug response outcomes using, among others, decision trees, multiple regression methods, or random forests. AI models can also encompass data

beyond genetics, including demographic factors such as race and sex, or clinical parameters. AI in this context not only maximizes numeracy and hidden information extraction on the side of generated treatment groups but can also greatly aid in predicting the safety of cancer immunotherapeutics related to the increased inflammation in the patient subgroup. The arrival of patient stratification tools in the clinic represents a momentous shift in the personalized medicine world with dramatic game-changing potential.

4.2. Pharmacogenomics

Pharmacogenomics, defined as the study of variations of DNA and RNA traits related to drug response, is essential for explaining individual variability in the outcome of anticancer therapies. Typically, all cells in a tumor, as well as germ cells, carry the mutations that serve as predictive biomarkers. Therefore, predicting the treatment outcome, whether it be efficacy or severe toxicity, is generally more accurate and clinically relevant when addressed as an integrated corollary intrinsic to the host component. Personalized cancer medicine is not just about drugging patients based solely on their tumors; it ideally depends on various patient factors. Fundamental information such as systematically measured parameters from biometric or omic sources has proven predictive in classifying actual responders from nonresponders using AI. Various studies have reported anticipating the probability of both treatment outcomes in multiple artificial neural networks and machine learning models. Key steps in oncological decision-making, such as assessment of metastatic spread as well as prognosis and treatment response, could be advanced by fractal or AI mathematics.

AI, therefore, could serve as a component of a self-organizing strategy to spur an academic immunochemotherapy era in cancer. For example, in pharmacogenomics, AI might be pivotal in establishing predictive genomic models dictating which patients could develop unexpected severe life-threatening side effects or, in the case of alternative therapeutics, prospective idealized negatively predictive patients who will not respond to the treatment. These types of research are particularly welcomed in our healthcare institutions, where difficult patients refuse to undergo traditional therapeutic guidelines. Thus, both the regulatory ethics committees and the healthcare providers are eager to embrace the advances of niche omics. Larger clinical trials are often unable to obtain oncological predictive surrogate endpoints pertinent enough for

pharmacogenomic purposes. AI can help greatly in this regard. These subsequent challenges may affect the excitement for potential pharmacogenomic AI platforms. These ethical reservations are an example of how omics and the potential to develop unique scoring functions accounting for combination stratification computing, unique reactions, or new therapies can often aggravate treating oncologists. Certainly, a thorough intimate relationship with the specific disease under study is likely to aid in avoiding many of these challenges. Solutions incorporating privacy, respect for the right of consent, and the rights to have personal data removed from such predictions may well address some of these concerns. The controversies about giving the public access to pharmacogenomic profiles and personalized medicine recommendations do, however, continue to cast a shadow over this discovery. While controversy surrounding this topic of public use will also revolve around physician resistance to interpreting complex pharmacogenomic network or medication risk ratio predictions, the take-home message for research is clear. Significant potential exists for AI to transform the efficacy of oncology prediction based on direct genomic-based predictivity or pharmacogenomics. In fact, one study found that AI could improve the correlation of a genome data signature in terms of prediction. This suggests a lead role for pharmacogenomics five years from now.

5. Challenges and Future Directions

Data quality and data quantity: AI systems highly depend on the quality and quantity of the data as input. They are mainly used to predict the outcome of research and clinical trials; if the data are flawed or not representative, the model may not be efficient. The datasets used need to be of high quality and ideally available to everyone in order to eliminate bias, including diverse representation. Interpretability and trust: another challenge is that AI allows making predictions without explaining the rationale behind it. It requires deep analysis to understand how and why the model reached a certain conclusion. In this context, holding back the method is undesirable, since learning how AI arrived at a given conclusion influences accuracy and trust. It is a clinical priority to deploy not only accurate predictive models but also models with robustness, reliability, interpretability, and a proper understanding of the uncertainty on new data. Regulatory considerations—pathway to approval: there are very few approved AI-based treatments yet. The US FDA published a discussion paper which states that all participants in the AI workflow may play a role in the development of an AI/machine learning system. If

AI-assisted diagnosing products are outsourced or potential data sources are not clear, the FDA must watch the manufacturers and get an assertion on data rights. The manufacturer must, in the submission to the agency, present all existing information showing product effectiveness. This guidance reflects the challenges inherent to the administration of AI-based products and ensures the development of robust clinical studies, data sources, and algorithmic development. To complement this guidance, the FDA moderated public discussions and released a public docket to get feedback on the generation of real-world data in the development of AI. The European Medicines Agency has also published a concept paper in collaboration with the European Commission on the use of AI in medicine, including AI-based drug discovery and development. Solutions and research directions: one way to work around these limitations is to develop and use universal biomedical language representation approaches. If you embed hundreds or thousands of studies into one language space, you can understand the relative meaning more easily. A key research direction trend should be to integrate AI techniques with more comprehensive datasets that should include overlapping of varying histologies if possible and studies from diversified populations. Since no technologies are mainly machine learning or AI, and machine learning experimental designs are needed to bring out safe and effective drug development, it is critical to include data science technology throughout the entire drug development cycle. AI is one type of data science technology and requires a team approach with traditional biostatistics as a necessary partner both in design and execution.

5.1. Data Quality and Quantity

Shortcomings in data quality and quantity are among the greatest barriers to effective AI application in drug discovery projects. High-quality datasets in cancer biology need to be large; however, researchers often deal with small and highly curated datasets, which do not capture the full complexity and variability of cancers. Expanding this dimension by including a diverse patient population and capturing the pan-cancer transcriptome introduces containment inconsistency and can easily generate unreliable AI models simply because the knowledge gaps and omitted warnings introduce noise into the datasets. Limiting coverage to certain technology platforms or analysis pipelines, which would restrict the ability to integrate the generated omics and non-omics, can be regarded as a strategy to circumvent issues related to data inconsistency and quality.

Here, the two levels of data inconsistency manifest themselves in the onset of unmodeled heterogeneity as a hyperparameter, or the lack of assumption-aware models to consolidate these sources of heterogeneity.

It is a commonly accepted principle that twenty-first-century medical research must be built upon the hybrid foundation of multiple research facilities, thus requiring close cooperation between academia, the pharmaceutical and biotech industries, and AI operators. This is particularly true in the realm of AI and oncology research, where progress is inextricably linked to the acquisition of large structured databases that exhibit critical molecular and clinical homogeneity throughout the study. Given both the paramount need for improving patient care and the strategic importance of the underlying data resources for the advancement of drug development, we envisage that relevant data collection efforts may come to be seen not only in close collaboration with the pharmaceutical industry, but also as maintaining pre-competitive value. Efforts to improve the collection of data, as well as the curation and integration of this multi-omics data, are a revolutionary undertaking that will require a combination of advanced computational, automation, and administrative expertise.

5.2. Interpretability and Trustworthiness

Interpretability and Trustworthiness: AI technologies, in general, show higher prediction performance than current chemoinformatics methods, especially for complex biological processes such as anticancer drug sensitivity prediction. However, it is important to be able to explain to health professionals, regulatory agencies, researchers, and even patients how AI systems yield their predictions and recommendations, as well as to estimate the trustworthiness of these systems. Legally, regulatory, and professional reasons make interpretability a pivotal point in the adoption of new technologies, including AI, in clinical routine. In the context of AI, recent works have discussed that an opaque AI system may not only generate suboptimal decisions but also may lead to adverse consequences, bias, and ethical issues. Current interpretable models do not satisfy, at the same time, the requirements of different stakeholders. Several AI fields are converging today in broader, more comprehensive models, named “explainable AI,” whose prediction process is transparent, interpretable, and understandable, providing information about the model performance and the reliability of the prediction results. At the same time, several case studies demonstrate the potential to use the explainable AI

approach to support diagnoses, prognosis, and therapy choices in different types of cancer and pathology. The explainable AI approach to build trust among users must take these aspects into account. Going deeper into the AI models and providing a rationale is fundamentally important for users' confidence, controlling the interaction between the expert and the algorithm.

6. Future Direction

Future direction Advances in machine learning, such as deep learning and transfer learning, may further drive the drug discovery pipelines. AI can provide improved predictions on the toxicity of drugs and suggest alternative combination treatments as part of a faster and more pervasive approach for multimodal pharmacological analyses. Extensive use of AI in preclinical and clinical data is the future, with in vitro human data from organs on a chip that could predict organ toxicity and response. Increased integration of AI in the clinic is also expected, aiding the management and analysis of patient data, lifestyle, and daily activity. Interdisciplinary studies involving experts in engineering, physical sciences, biology, enabling technologies, corporate and clinical development programs, and academic hospitals are needed to accelerate the application of new computer-aided drug discovery and development.

Key trends In the present review, we discussed different strategies of applying AI to predict response to drugs, either as monotherapy or combination therapy, in a clinical setting using real-world evidence. Innovations leading to better model calibration, based on updated patient data, are discussed in the context of robust digital infrastructure for further development requiring special considerations. These include data standardization and model interpretability for ethical and practical implementation of a data-driven, adaptive framework of drug scheduling. This computational framework shifts from a fixed-dose, empirically determined treatment schedule to an adaptive strategy that encompasses a patient-centered, precision medicine use of drugs, where a schedule of dosing and drug class could be defined based on the latest available data. Future developments include the use of reinforcement learning mechanisms to achieve therapy personalization, especially where treatments from different drug classes can lead to diverging long-term outcomes based on dynamically evolving patient data. The development of any AI-based tool in healthcare must be implemented within a robust regulatory and ethical framework, including data and process transparency, together

with the setting of clinical research protocols and the initiation of informed consent for innovative adaptive clinical trials. Publications in this area are relatively new, raising interest but not saturating the potential of AI and real-world data in the coming future. In oncology, AI can enable faster identification of new targets and the selection of the best-suited drug, unlocking precision medicine and fostering drug safety. An active area of research will involve the development of advanced AI models integrating a mix of in vitro, in vivo, and in silico preclinical data and, in parallel, real-world evidence on the effect of drug treatments. Ethical considerations will need to be discussed to prevent their misuse or exploitation, to ensure safe and patient-empowering real-world evidence data, as well discussed in this area. An effective patient-care hybrid approach to AI should improve and enable drug treatment decisions while keeping patients at the center of this development.

7. Conclusion

In conclusion, AI technologies have the potential to revolutionize oncology drug discovery and significantly impact cancer patients' lives through more efficient and effective treatments. We have presented recent and forthcoming applications across the impact landscape and emphasized the predicted outcome of these developments in terms of both the predicted impact on the efficiency of our software and the acceleration of new drug discoveries. AI technologies enable the deconvolution of complex biology into insightful, predictive models. We have emphasized our ongoing close relationships with hospitals, collaborations, and bench-based experimental evaluation programs to ensure that our AI-based predictions are aligned with hospital-specific requirements and procedures, are clinically relevant, and would result in improved patient outcomes. The challenging task of making new treatments work is central to the conference track and the related healthcare congress. Additionally, we believe that Cancer Frontiers is an appropriate platform for AI-based approaches, as recent examples of successful applications were driven by AI technologies for predicting novel therapies for different tumor types prior to their clinical validation. The pursuit of algorithmic developments has been complemented by a clear striving towards translational and clinical impact. While AI technology continues to advance in our efforts, there are, of course, numerous hurdles yet to overcome in an oncological context, not least data quality, a lack of consistent patient truth data, regulatory reform, and clinician buy-in, along with closer formal integration of AI technologies in clinical trial procedures. Moreover, while some

AI technologies can become a new set piece of patient management, AI solutions need to avoid 'black box' adoption and always focus on patients' safety, security, and outcomes using standard clinical grades of evidence, especially in the context of any 'right first time' solutions. While the research and innovation landscape for ethical AI leverages citizen and stakeholder engagement and surveys to help shape patterns of funding in academia and healthcare, we need to institutionalize these methods on a larger scale to ensure appropriate societal acceptance and governance of ethical AI in healthcare. In summary, AI has already enabled new drugs, new actions, and thereby new hope for many cancer patients, and hopefully will continue to do so throughout the future of oncology drug discovery, using the input from this prestigious conference and state-of-the-art clinic today. AI has the potential to revolutionize our care: let us begin.