Multi-Omics Data Fusion Through Attention-Based Neural Networks: AI-Enhanced Systems for Integrated Genomic, Proteomic, and Metabolomic Analysis in Drug Discovery

Dr. Paulo Leitão, Professor of Informatics, University of Minho, Portugal

1. Introduction

The integration of genomic, proteomic, and metabolomic data in a systems biology approach is expected to be transformative for drug discovery and development, providing an opportunity to expand the discovery of novel molecules for new targets and, importantly, to determine whether these novel treatments will be effective and safe. This potential is still not fully realized, making it difficult for many researchers to obtain biologically meaningful results from their data. The reasons for this are several: omics data are increasingly large and complex, and include many sources and a diverse range of data types, with a variety of terminologies, measurement units, and assays to measure them. Also, data from R&D and standard medical practice have necessarily been collected at different scales (from the gene to the pathway, organ, and the whole organism/homeostasis).

In this essay, we will provide a brief overview of systems biology, systems medicine, and omics-related background to provide context to technological advances to integrate omic data in drug discovery, with case studies and insights into the methodologies. We will also discuss how artificial intelligence and machine learning, in particular, can streamline the process of interpreting – and especially predicting and building models from – these data, and how its application allows links with previously established scientific and clinical knowledge. We will describe existing applications of AI in interpreting omics data in pharmaceutical drug development, from the use of blood omics to develop blood-based gene signatures for neural disorders, through to the use of AI to interpret a panel of immune proteomic biomarkers to aid informed protocol design for peptide vaccination therapies.

1.1. Background of Omics Data Integration in Drug Discovery

1.1. Background Due to recent advancements in omics technologies, a wide array of molecular measures reflecting the biological processes inside a living organism can now be simultaneously measured. Although the complete understanding of all the complex interactions happening between the different levels of these molecular omics data is an illusion, big data and multi-omics are already offering the field of drug discovery never observed detailed knowledge of biological pathways as well as disease mechanisms. The Human Genome Project marked an inflection point in the field of omics technologies, from which, due to the fast declining costs, large-scale multi-omics data generation efforts were initiated, based on the belief that having access to variation at the genome level would ultimately result in a direct understanding of individual predispositions for a wide array of disease conditions.

In fact, molecular omics data have been widely and successfully employed in target discovery, validation, patient stratification, and pharmacogenomics and are expected to be crucial for decision making in personalized medicine. Merging different biological omics measurements using integration approaches can significantly improve systems-level understanding. A broad array of data is generated on a diverse array of platforms with a wide variety of coverage, and at very different layers of molecular interaction. As a consequence, combining such disparate data sources offers large opportunities for systems-level drug discovery analyses as well as compound-induced effect prediction. Hypotheses extracted from such joint analysis are expected to be greater in quality than the sum of their individual parts. The indicator mentioned in this document is the development of systems using technologies as well as handling diverse contextual information, and are able to handle the heterogeneous, large-scale integrative data sets generated in a systematic way in order to yield more informative hypotheses for drug impact, patient subset stratification as well as chemical compound bio-relevancy.

1.2. Importance of Target Identification in Drug Discovery

Target identification, the process of assessing the biological target of compounds, is one of the main challenges in the discovery of a new drug. Though phenotypic screens are used to discover therapeutic activities of compounds in cells or model organisms on a system-wide basis without a prior hypothesis, a pharmacological and genetic opening confirmed on the desired mechanism of action would be critical in distinguishing drugs

from poison. Both on- and off-target effects to find drugs have similarly proved important. As a consequence, the identification of biological targets and a clear comprehensive knowledge base is important in the process of drug development, which can, accordingly, guide this development more intelligently. This can also contribute to target selection and costly and laborious process optimization in deeper preclinical and clinical medicine. Different forms of techniques can be used, based on the case and other logistical issues, for target identification.

Some of the methods for traditional target validation include both the determination of gene expression modifications induced by the substance in a particular cell type and/or the validation of the target's awareness at the level of the cell type through a solution of genetic material that modulates the pathway or the desired mechanism of action. Similarly, biomarker discovery can be assessed on the target as would be biologically or immunologically consistent, primarily using omics. The molecular entity that modulates the chosen techniques or any gene product or metabolite connected to it would theoretically identify as a biomarker the biological entities of minimal change. Omics techniques are used to provide a company and a well-functioning system for its goals in the discovery and the target. All omics have clean tools that AI is used with basic sample patterns for induction or enhanced target association for aid in molecular target detection. So multidimensional consistent data evaluations can select suitable targets. In a selective manner of the distinct lineages of information, AI has been intelligent. Involvement of proteins and DNA will be significant to unravel the involvement of the proteins in the holder of the information that can be used to perform the molecular assay. Furthermore, AI would be a guided framework for target knowledge development and suggest two distinct compounds.

2. Genomic, Proteomic, and Metabolomic Data in Drug Discovery

Omics data in genomic, proteomic, and metabolomic domains are key to understanding biological systems, pathology development, and host response. Such disease-relevant cascades and networks are likewise strongly recommended as an integrated part in all phases of drug development. At the genetic level, approaches provide details concerning complex diseases and inform which biological system to explore. At the protein expression level, studies elucidate key activation patterns. Metabolomics provide a view of the end phenotype of the integration of biological processes at the

mRNA and protein levels and changes in molecular function and interaction between internal and external forces. Each of these data types not only elucidates different aspects of the biological complexity of human biochemistry and metabolism but also myriad ways in which drugs may intersect and therefore affect overall biology. The genetic architecture levels, i.e., genotype, gene expression, protein, and metabolites, are interconnected and drive human metabolism. Integrated omics analysis can provide valuable insights into systemic responses to therapy and molecular-level interaction. Thus, readings from the multi-omics not only support the mode of action but help to predict the mechanism of drug action, possible drug metabolism, and interaction with different complex biological systems. Additional benefits include the ability to find several new data dimensions/populations about the putative network of pathways accountable for the probable varying factors. AI technologies can be integrated into such systems to provide heuristic as well as statistically based predictions for new drug chemicals based on the derived patterns from these omic studies. It is for these reasons that these omic data points are important and are sought to be part of the predictive models and prognosis panels.

2.1. Overview of Genomic Data

In the topics that follow each specific 'omics level, an overview of the workflow will be provided that details the steps needed to integrate and incorporate data generated at those levels. We will introduce the specific type of data collected at each 'omics level in turn, beginning with genomic data. The genetic influence of drug responses has been investigated for over 50 years, but this field underwent dramatic changes as a result of the genomic revolution. The availability of affordable high-throughput technologies has brought a wealth of genomic data amenable to machine learning analytics. This, in turn, has the potential to revolutionize drug discovery and development.

At the genomic level, genetic variability can be interrogated at various scales, from whole-genome polymorphisms to more targeted gene-specific analyses. The most frequently used genomic information consists of single-nucleotide polymorphisms (SNPs), individual nucleotide bases where the variation can influence gene function. In addition to SNPs, copy number variants (CNVs) are another class of structural variation that can be of importance in drug responses. Also, genes can be sequenced to describe the expression of a specific target. This gene expression profile in the form of RNA levels

can provide valuable information about a potential drug target. This review aims to provide an overview of the relevance of genomic variability in the success and failure of drugs. It begins with an overview of basic topics in genetics at the foundation of the new technologies and the dimensions of variability within human and other populations. While technological advances in sequencing and genotyping have progressed rapidly, successful interpretation has lagged. Integration with other 'omics data is a new means to identify causal molecular drivers of a disease disorder. This will be discussed in the tools and challenges in data integration. As the quantity of generated data is now of big data proportions, machine learning techniques to move past a description of a population to learn novel disease pathways from the data is at hand. This will be reviewed from each 'omics level. AI-automated methods of detection of molecular disease drivers are on the horizon. For personalized medicine, machine learning techniques can make strong predictions of drug side effects.

2.2. Overview of Proteomic Data

In contrast to the central dogma of molecular biology, proteins and their underlying processes are very complex. Proteins can undergo several basic post-translational modifications (PTMs), such as phosphorylation and methylation, added to more than 10 amino acids out of 21, the so-called frequent PTM, and an undefined fraction of up to 595 different PTMs added to any possible amino acid have been discovered so far. This adds a high level of sub-proteomic complexity and severely complicates drug action, particularly concerning in situ selectivity of binding and drug–drug interaction. Proteomic data can provide additional evidence for the validity of genome and transcriptome-based targets and can also reveal off-targets of previously designed compounds. This information is useful, especially in the context of systems biology and drug repositioning.

Over the last decade, a lot of work has been done to improve the quality of mass spectrometry-based data analyses. Innovative methods to improve sensitivity and identification rates have been introduced, such as DIA and PRM for targeted analysis. Therefore, it has become possible to measure systems biology-related proteomics data across 5,000 samples in a time-dependent manner. However, the clinical applicability of the aforementioned technologies and of proteomics data, in particular, is very challenging due to the vast data heterogeneity of measurement methods and the lack of

standards in repetitive compound-related research or clinical studies. Proteomic data has received a lot of attention due to the use of neuroproteomics as part of the first successful clinical proteomic project. The project was able to use AI methods to integrate proteomics data with related information from genomics, metabolomics, and neuroimaging. The project aims to define the typical fluctuation of the blood immune system's response in Alzheimer's patients compared to a healthy population. From the single perturbation reaction, the first step in the physiological reaction cascade, we were able to describe the cellular pathway mechanisms leading to different dynamics in protein levels. Having identified the relevant dynamic pathways, a selection of one or multiple proteins therein can become experimental starting point targets for drug intervention studies. AI deep learning methods can be directly used to acquire relevant dynamic blood proteomics leading to new clinical relevant outcomes.

2.3. Overview of Metabolomic Data

Metabolites are the substrates and products of enzymes, and their chemical levels can be used to analyze biological systems. They can provide comprehensive insights into the metabolic state of cells and tissues. From the standpoint of drug development, metabolomics offers the fastest cut on the overall effect. In particular, phenotyping can reveal the likelihood of long-term drug effects such as drug responses and toxicity. A straightforward means of metabolite determination is provided by nuclear magnetic resonance, and added sensitivity is offered through mass spectrometry. There are a number of refinements under these two major classes of methodologies, for example, gas chromatography mass spectrometry, liquid chromatography–mass spectrometry, direct infusion mass spectrometry, and metabolic profiling based on nuclear magnetic resonance; gel-based approaches; targeted metabolomics and so on have been explored. Besides measuring metabolites, methods in metabolomics also include enzymatic assays of activity or predictions of the availability of metabolically modified drug candidates.

The study of metabolic pathways and drug candidate pharmacology is becoming particularly important for effective scaling up of drug discovery research. Metabolites are one of the major categories of potential drug targets or molecules that can act as biomarkers for disease, and a number of metabolite databases that integrate metabolism with genomics and proteomics data have been developed. Metabolites, in turn, constitute a part of the InChI key format, so integration of metabolomics data with

cheminformatics fingerprints and other types of data used in overall drug development is easy. Metabolomics data are also being used to provide evidence for the mode of action of herbal remedies and traditional medicines. However, metabolomics data have a variety of added complexities which range from the variability in metabolic state composition between individuals resulting from lifestyle and dietary differences, age, and gender to analytical complexities. Its integration with genomic and proteomic data of a population is system biology—one of the current strategies for drug discovery. The increasing application of different purposes and the growing body of data has also started bringing in artificial intelligence technologies with which they are integrated for better data interpretation.

3. Machine Learning Techniques for Omics Data Integration

This section presents various machine learning techniques to facilitate omics data integration in drug discovery. Machine learning techniques possess the capacity to capitalize on predictive modeling for analyzing multiple omics datasets that, due to new high-throughput techniques, have been recently translated from functional genomics to transcriptomics, epigenomics, proteomics, and metabonomics. In addition, they can be deployed to perform exploratory data analyses using multivariate techniques suitable for situations in which the data are too complex to be interpreted by viewing just one variable at a time. These advanced methodologies can integrate multiple data sources for the interpretation of biological datasets. Machine learning presents potential for utilizing the wealth of data present in omics sooner in the part of the drug discovery process. Today, machine learning is used to develop models to classify patient groups or groups of diseases' studies by their profiles of multiple omics datasets.

This machine learning-based systems biology strategy necessitates the use of various predictive modeling techniques that have been developed within the field of cheminformatics and bioinformatics. Many repositories with omics data are available and have been collected over decades. Machine and deep learning procedures are widely used to mine insight from these datasets and to increase information on the efficacy and safety of new medicines. Several machine learning techniques can be used to perform integration of the omics data. There are two broad approaches: supervised and unsupervised machine learning techniques. Depending on the specific biological or clinical question, these machine learning techniques might be deployed to do either

exploratory data analyses, together with other multiple techniques, or to actually build predictive models. These predictive models are powerful approaches available for a number of analyses. Certain predictive approaches to data are powerful methods that fuel this machine learning-based systems biology approach. These methods, whose basic principles will be explained next, are widely used and are currently being applied within the platform. We believe that future advancements in machine learning will likely allow for the extraction of biologically meaningful information by integration and data correction, without the need for data labeling.

3.1. Supervised Learning Methods

Supervised learning is a task of machine learning where one is presented with a number of pre-labeled instances from an input space and their corresponding output to predict new instances in the case of unknown output. The model is trained on labeled datasets to predict new outputs. This type of learning algorithm is also being used in drug discovery applications to make predictions based on omics data. For example, gene expression data can be used for predicting response levels, protein-protein interactions, and pathways as a phenotype, protein structures associated with their functions, proteins and molecules as targets, and biomarkers.

Some of the supervised learning techniques that are used to make predictions using omics data are:

Regression models, e.g., linear regression, ridge regression, elastic net, support vector machines, and decision trees. Here, the output variable is usually continuous. Classification techniques, e.g., logistic regression, random forests, support vector machines, and neural networks. Here, the question could be changed to be a binary or multi-class problem. Dimensionality reduction methods, e.g., principal component regression, partial least squares, and independent component analysis. These methods are used because data is presented in a high-dimensional space, and the number of variables is likely to exceed the number of subjects in the datasets. However, it is challenging to apply these methods directly to omics data because of overfitting and the need for variable selection. Overfitting is a phenomenon in which a model learns noise in the data instead of the underlying pattern, which reduces its ability to fully generalize. Feature selection is the process of identifying and retaining the most significant features.

Several omics-based case studies demonstrate the applicability of supervised learning to target identification, as well as binding affinity and biological response prediction. Additionally, supervised learning can be combined with omics data to enhance these studies. However, to ensure the reliability and generalizability of the models, repeated methods for model learning and validation should be carried out. Some of these methods are cross-validation, external validation, mixed validation, and others.

3.2. Unsupervised Learning Methods

Unsupervised learning is a machine learning technique focused on exploring and uncovering hidden patterns and structures within an unlabeled dataset. Unsupervised learning methods can generate models to describe how these hidden structures are organized, thereby allowing researchers to identify and characterize groups in a dataset that share similar features. The research community has developed a number of clustering algorithms, dimensionality reduction techniques, and data association methods within the realm of unsupervised learning that remain very relevant in the omics data integration toolbox. Each of these methods plays a role in data exploration and hypothesis generation. One of the most common motivations for using unsupervised learning is data-driven hypothesis generation. The ability of unsupervised clustering methods to organize data due to its underlying characteristics is useful for both discovering novel biomarkers and also representing novel relations between entities. One illustrative example is the application of clustering methods to multilayer omics data, where there is prior biological knowledge that protein and mRNA expression measurements are often correlated.

The key advantage of this unsupervised approach is that the information derived still remains relevant despite changing biological knowledge. Similarly, the application of unsupervised dimensionality reduction techniques can identify structured relationships between multiple types of omics data. In another example, a combination of unsupervised dimensionality reduction techniques was applied to integrate gene expression and copy number alteration tumor datasets across cancer types. The goal was to focus on both frequently mutated driver genes and on associated pathways and functions that are regulated by infrequently mutated members of the same protein families. By using these methods, a number of cancer-signaling networks were identified that stratified patients according to poor and better survival. Model interpretability can

be of particular concern with unsupervised hybrid models. Evolutionarily and temporally conserved proteins are often not present in modern drug target vocabularies or in drug target databases and resources, because these proteins are not likely to generate drugs with high specificity. Thus, there is a need for these models to be interpretable, so that these potential side effects can be identified. The application of a model offers a highly interpretable approach because it allows users to understand the human proteins and associated cancers in the prioritized functional contexts.

4. Case Studies and Applications

In this section, we present a series of case studies illustrating the real-world utility of integrated omics data at various pharmaceutical companies. Each case study demonstrates how different omics data types collaborate to address specific pharmaceutical challenges such as better understanding of a disease system, patient stratification, biological understanding and mechanism of action, safety assessment, target identification, and target tractability. Integrated data may be employed for predictive modeling to optimize factors such as chemical structure, model systems, response measurements, pathway modulations, and pharmacokinetics in these descriptions. This enables better assessment of these factors for decision-making along the continuum from discovery to the clinic, improving the development process.

We believe that the best examples we could offer would be data from real settings in drug discovery. As sequencing costs rapidly decreased, the field has witnessed extensive biomarker and targeted therapy trials designed to identify genetic markers associated with drug efficacy or identification of resistance mechanisms. An increasing number of drugs have been FDA-approved as companion diagnostics. All of these have increased demand for integration of specific information across trials to understand both drug mechanisms, the genomics underpinning the disease, and patient response to personalize treatment. Some of our case studies integrate high-quality primary data that was collected following the Good Clinical Practice guidelines enabling regulatory submissions; other data sets use snapshot studies and are used internally for translational research. In this article, we showcase what has been achieved using AI and other data integration methods at these companies and provide key insights from these examples. These case studies can be taken as individual examples of the power of fully integrated approaches, and they illustrate some of the transformative changes in

healthcare by contributing to personalized medicine initiatives. These case study examples represent best practices and proof of concept and have each contributed to some level of decision-making, as well as providing insights for innovative methodologies.

4.1. Drug Target Identification Using Integrated Omics Data

4.1. Drug Target Identification Using Integrated Omics Data

Over the last decades, a vast amount of 'omics' (e.g., genomics, transcriptomics, proteomics) techniques have been developed to capture information on biomolecules in biological systems. Because of rapid technological advances in omics methods, systems biology has emerged as a data-driven approach to system-level biological investigation. As drugs can indirectly affect systems and the development of any interventions needs the understanding of only targeted pathways, thus, over the last decade, the development of new drugs through multi-omics datasets has increased significantly. In addition, the burden on mankind due to increased costs and the very limited number of drugs developed for a few known druggable genes has been substantial. Therefore, in silico systems-based omics integration pathways-based drug discovery is serving as one of the rational and only choices for mankind.

The integration of innovative methods to utilize multi-omics data over recent years has also been enabled by machine learning and artificial intelligence methodologies. Such systematic approaches have been considered in the context of target identification to identify therapeutically relevant targets for drug efficacy and safety. Validation and prioritization of new drug targets remain one of the biggest unmet needs in order to improve the success of drug development and to develop therapeutically relevant mechanism-based drugs in molecular subtyping and precision medicine. Considering the evolutionary parts and functional variations, the development of side effect-free precision drugs is still challenging. A rational and highly accurate quality of prediction parameters is required to develop AI and systems-based in silico approaches. Decisively, improper identification of drug targets leads to the failure of the drugs. The complexity of molecular networks can only be studied with the help of close collaboration between biological expertise, multivariable biostatistics, bioinformatics, and those with expertise in pharmaceutical drug development strategies. With that, the in silico systems-based integration could likely be accomplished. Few examples include: identified targets of

small molecules. The current times are still blinded by such in silico consolidated pathway-based target identification, digitalization, and miniaturization. Decades have passed with high-throughput sequencing and connectivity mapping data accelerating each day; a few other commercial, academic, pharmaceutical companies, and research laboratories are still looking for experimental models, omic levels, integration validation, and in silico model strategy integration to employ affordable times and resources vested at the highest confidence. The development and integration of various omics-based systems to build a series of in silico models to identify altered, coherent biological pathways which may be utilized for therapeutic moieties for the stratification of cancer patients were described in the position manuscript. Furthermore, there is data available about the methodology utilized for the identification of biological targets, their validation, downstream pathways, and for the predictive mechanism-based drug identification that is in discourse.

4.2. Predictive Models for Drug Response

In this context, predictive models have been developed that merge different omics layers to anticipate patients' drug response. In descriptive studies, it was shown that merged data better reflect the drug treatment effects at a molecular level and suggest that dosage and/or combination therapies should be personalized if differences between individuals exist. As most single omics testing or combinations are performed only in one standardized manner, this does not yet reflect the clinical scenario of a systemic approach. In contrast, it is important to show that standard oncology patients indeed differ and personalized treatment strategies would be needed. A straightforward way to validate the predictive power of an integrated model is by comparing it with single molecular changes. A combined model is only useful in application if it provides an advantage over simple screening tests.

In all presented studies, the applied AI algorithm obviously improved the prescription of patients' clinical treatment; otherwise, reporting would not occur. Collection of a huge number of pathophysiological factors can identify apparent robust integrated models. Consequently, if desperately desired, an integrated model can be constructed with any dataset, providing the possibility of having a clinically relevant added value. Having a deep insight into pathophysiology is invaluable for constructing a reasonable integrated model. Potential quality predictors based on genotypic, phenotypic, and molecular data

may already indicate who is more responsive to the new personalized medicines to guide the study. The great challenge is to find robust data of sufficient quality and size that allow conversion into informative predictive and especially integrative models. In the future, the drug dosage and/or combination therapy regimes should be personalized; otherwise, a certain proportion of the patient population does not receive proper treatment, thus increasing treatment costs and mortality. Finally, all the results can only be considered as preliminary. It should be emphasized that model formulation and validation cannot currently be separated based on therapy. It is assumed that it is only the interaction of researchers, omics, and drug science that produces gradually validated robust predictive integrated models suitable for application.

5. Challenges and Future Directions

Despite the multitude of omics data that represent and integrate information at the molecular, tissue, system, and patient levels, there are several challenges that need to be overcome to enable the full use of these technologies in drug discovery. Data heterogeneity and the omics technologies used to acquire system-level data raise a key concern, as each omics discipline records data differently, and comprehensive data integration requires standardization, such as quantifying data in a similar format across different data sources. Aligning technologies across disciplines is very important, given that different research and drug discovery areas, including academic researchers and the pharmaceutical industry, should have a similar understanding of data. Although omics data are essential in drug discovery, an important challenge that needs to be addressed is the existing gap between drug-induced modifications captured by omics data and the traditional in vitro toxicity assays to increase confidence in potential target-drug interactions. The standardization of in vitro assays and how to link data present a fundamental challenge. Moreover, establishing the quality and validity of both in silico and preclinical models that utilize omics data for the ability to predict potential off-target drug interactions remains challenging. Once the drug reaches the stage of registration trials, the potential for side effects emerges from the noise.

Another main limitation of using omics big data is data quality, including processing data, how data are recorded and compiled. Basic descriptive statistics are often undervalued in systems biology, omics, drug discovery, and personalized medicine. Finally, there is an ethical and legal need to handle omics big data, including clinical

trials, biomedical research, clinical reasons, privacy, and protection of data sharing. Data storage should be delineated with patient consent, identification of personally identifiable information, and who has the authority to access the data. The use of omics technologies and big data must respect the legal and ethical agreements of the country. For example, researchers should use de-identified specimens before omics technology and big data analysis begins, without accessing personal identifiers. Following the identification of potential biomarkers resulting from the integration of big data for a personalized intervention, in terms of potential future directions, AI and systems medicine can address these challenges in omics integration to support drug discovery faster than it currently takes. In order to progress further, it is mandatory to have interdisciplinary involvement in standardizing processes in integrating omics big data. This involvement of an interdisciplinary group to facilitate research, clinical data, and the pharmaceutical industry is critically important to ensure there is no bias in agreement. The interdisciplinary group could certainly seek novel methodological approaches and/or tools to overcome the challenges listed. The focus on technologies that will help to address the highlighted challenges in the next part of this discussion is outlined and analyzed.

5.1. Data Integration Challenges

Data Integration Challenges

The application of multi-omics data integration in drug discovery and development encounters a variety of challenges. Firstly, multi-omics data is a heterogeneous type of big data from different omics fields and other sources that have been harmonized in diverse formats and quality. For example, transcriptomics data is obtained through RNA-Seq and hybridization techniques, which can be handled at two different levels of expression.

Data quality in the multi-omics and multi-source input data might be highly variable. In an RNA-Seq-based gene expression measurement service, for example, the libraries are first assessed to ensure the quality of the data. Then, once the library is constructed, it is quality checked and quantified before an equal pool is sequenced to acquire the necessary sequencing reads. This stringent process ensures that high-quality data is generated by standard technologies. Therefore, to avoid data-related biases, the integration of high-throughput data requires proper data management. Prior studies

have highlighted harmonization-related problems resulting from a lack of proper data management. Data discrepancies, such as gene symbols for identical measurements, might negatively affect the outcome of multi-omics analysis.

Besides the inconsistency in the data collected, the increasing level of data generation hampers effective data processing, analysis, and interpretation. Following the growth of data generated, it becomes more challenging to integrate these biological big data. The increase in the rate of data production is obviously outpacing the development of computational storage strategies. This becomes a bottleneck, resulting in a delay in data retrieval, input, processing, and integration. Therefore, there is a clear need to harmonize, integrate, and interpret omics data for the benefit of biological and clinical validation. The integration of multi-omics input data is currently performed manually to create multi-omics features that are then used for modeling purposes, which can be supported by AI technologies, leading to more efficient solutions. However, in the use of AI in multi-omics, the development and investigation of robust AI validation techniques are lacking in the domain. This shortcoming necessitates extensive collaborative work among AI, omics, individual research, and industrial partners to address the issues of omics analytics and data integration.

5.2. Ethical Considerations in Omics Data Usage

Concerns about the use of omics data in discriminatory algorithms contemplate generative adversarial networks and call for an ethical reflection regarding "defensive" AI research. Frameworks to discuss the ethical, legal, and social implications of omics data often highlight, among other principles and considerations, two topics: the need for informed consent and the importance of maintaining the privacy of patients and research participants. The legal position of informed consent is clearly outlined in regulations. While the updated paper still does not clarify the position, a proposed regulation again aimed to enhance penalties.

It is an obvious ethical and value-laden choice in the development of a drug discovery or repurposing initiative to exclude or include disease-related "-omics" data and analyze this data for valuable insights. Informed consent and the protection of patients' privacy should be safeguarded through fair and transparent use of the data donors' consent. In practical steps, including considering consent and privacy preservation aspects or conditions at early stages of the design of the drug discovery or repurposing project,

incorporation of consent and privacy preservation measures at later stages of the decision, versus exclusion of omics data, need to be considered subsequently. Ethical considerations drive the operationalization of integrated omics research, and the multidisciplinary engagement of all stakeholders is vital to achieving this.

6. Future Direction

Continuous advances in high-throughput and high-resolution technologies for data collection, next-generation sequencing, and proteomics analyses will still provide more significant amounts of data. Meanwhile, there is much room for improving the methodologies of collection, storage, and pre-processing of omics data, not only for the omics data but also for phenotypic data and for the metadata. The results of all these parts are also necessary to be stored, analyzed, and queried further with up-to-date methodologies and algorithms for future new drugs to speed up the approvals and the drug-repurposing processes. One possibility would be to include further machine learning methodologies for developing predictive mathematical models from the pre-filtered results determined from the hard data of the omics to lower their number, make robust the predictive power, and reduce the number of tissues, which have adequate predictions that are useful. Furthermore, one possibility could be addressing system-level new methodologies for drug discovery that, in combination with the traditional drug discovery methodologies, could help researchers adapt the data for real-time analysis.

Theranostics and tailored therapies are additional hot topics, which are also candidates for being included in this revision. This minireview also identified that new informatics and molecular biology case studies are ready to be developed to find these multi- or dir-inhibitors that may be toxic only in a particular and 'rare' patient but can also be used efficiently in a personalized medicine approach. Finally, the main system integration drug discovery methodologies are reported in this revision. Some gaps and opportunities are pointed out, and new case studies can be developed addressing non-supervised network-based drug repurposing and new data mining and machine learning methodologies. The ontologies with other sources are also candidates for being explored. Some of these techniques may be preliminary, with a lot of manual experimental testing based on gene expression studies and clinical testing, and dangerous for a high number of people. Therefore, a new interdisciplinary strategy can

be presented to explain to the readers the potential difficulties, criticisms, real future applications, and the importance of tools for drug discovery. In conclusion, a comprehensive system approach from molecular biology data to clinical personalized medicine is also a future possible direction; therefore, the role of system biology and systems for precision pharmacology will increase.

7. Conclusion

Given the remarkable success stories described in the different sections, one can only remain optimistic about the future of the field. It is generally agreed that, beyond offering a wider understanding of disease mechanisms, an integrated approach to drug discovery and the development of candidate compounds with proven efficacies requires the compatibility of genomics, proteomics, and/or metabolomics data. These types of detailed analyses can enable the analysis of data, from understanding the multiscale nature of biological responses at the subcellular molecular level to contextualizing the biological information on the well-being of patients. Moreover, some patient-omics data can directly guide drug repositioning or match patients to existing drugs if there is known information about the disease and drugs or pathways involved.

Nearly two decades of significant consistent increase in the annual drug approvals, together with the weight of the supporting case studies presented, furtively corroborate another important conclusion. Although still progressing, machine learning provides us with valuable tools for analyzing high-dimensional datasets. This can be considered an innovation capable of addressing some of the major limitations of a multi-omics approach in the early 2000s. In conclusion, we dare say that even if there are always challenges lying ahead in the relentless quest for the continuous improvement and ethical use of artificial intelligence systems, some things today unquestionably stand to reason. First, one has to remain creative, thinking of new intelligent ways to combine the different omics modalities in a pioneering manner. Second, data collection and integration will continue to be essential to take full advantage of AI and learn from the available, rich, multi-omics landscape. By joining all these forces and summing all the potential arising from the cutting-edge laboratories around the world, we are now in a much more favorable position to help improve the future health and safety of the whole population, hence opening the way to truly precision medicine.