Insights Brief

Machine Learning Applications for Therapeutic Tasks with Genomic Data

KEXIN HUANG, Research Student, IQVIA and PhD Student, Stanford University
CAO XIAO, Senior Director, Amplitude
LUCAS GLASS, Vice President, Analytics Center of Excellence, IQVIA
CATHY CRITCHLOW, Vice President, Amgen
GREG GIBSON, Professor, Georgia Institute of Technology
JIMENG SUN, Chief AI Scientist, IQVIA and Professor, University of Illinois Champaign Urbana

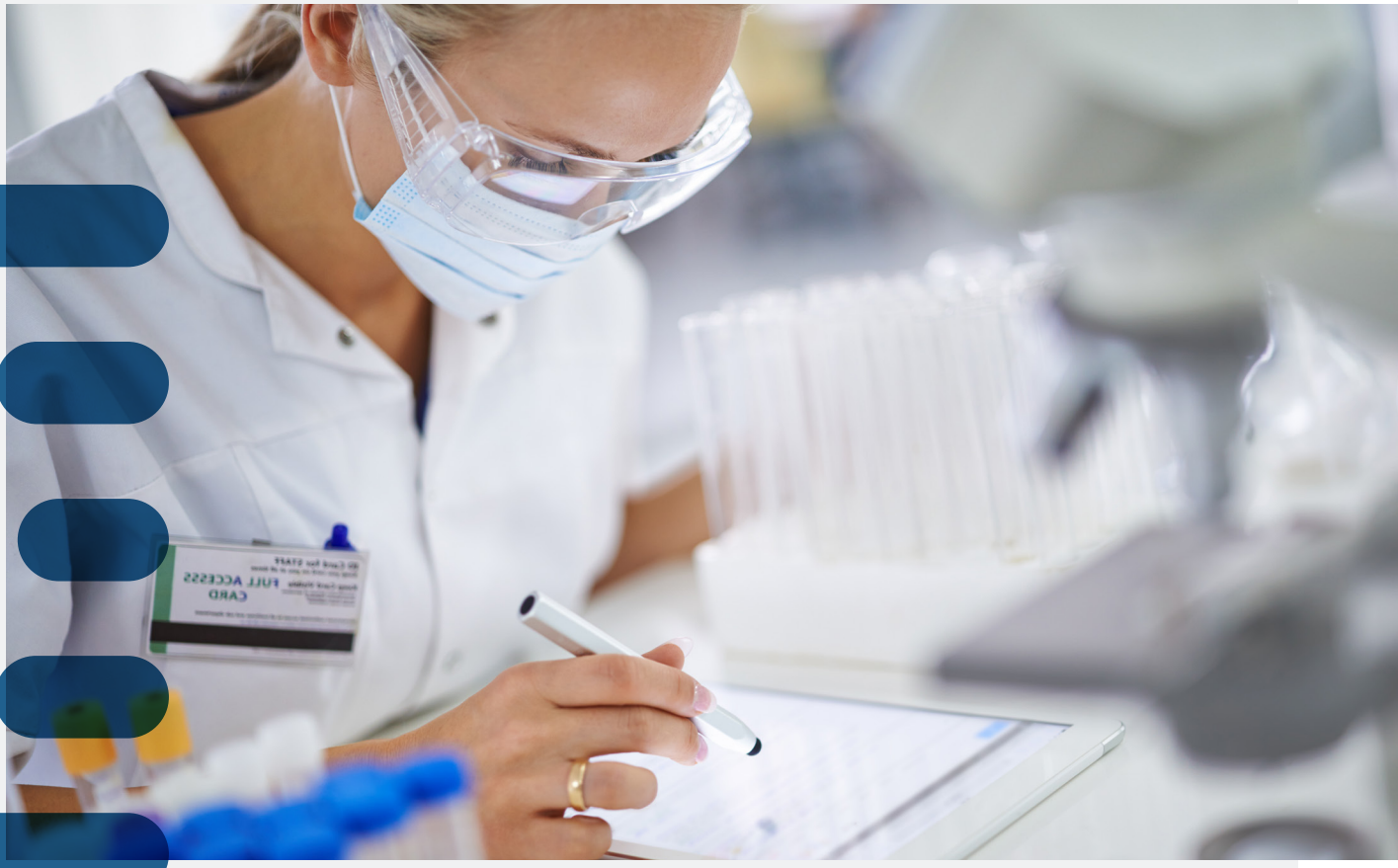


Table of contents

Introduction

The future of medicine is personalized, so understanding the therapeutic tasks with machine learning methods on genomics data is the key to leading ultimate breakthroughs in drug discovery and development.

Thanks to increasing availability of genomics and other biomedical data, many machine learning algorithms have been proposed for a wide range of therapeutic discovery and development tasks. The genome contains instructions for building the function and structure of organisms. Treatments will inevitably be tailored to a patient's particular genomic makeup.

With advances in high-throughput technologies and data management systems, we now have vast and heterogeneous datasets in the field of biomedicine. We're able to generate massive amounts of genomics data, but there are some roadblocks to turning genomic data into tangible therapeutics. Genomics data alone are insufficient for therapeutic development. How genomics data interact with other types of data, such as compounds, proteins, electronic health records, images, texts, etc., needs to be investigated.

Machine learning (ML) techniques can be used to identify patterns and extract insights from these complicated data. We surveyed a wide range of genomics applications of machine learning that can enable faster and more efficacious therapeutic development (Figure 1). Challenges remain, including technical issues, for example learning under different contexts given low resource constraints, and practical issues, such as mistrust of models, privacy, and fairness.

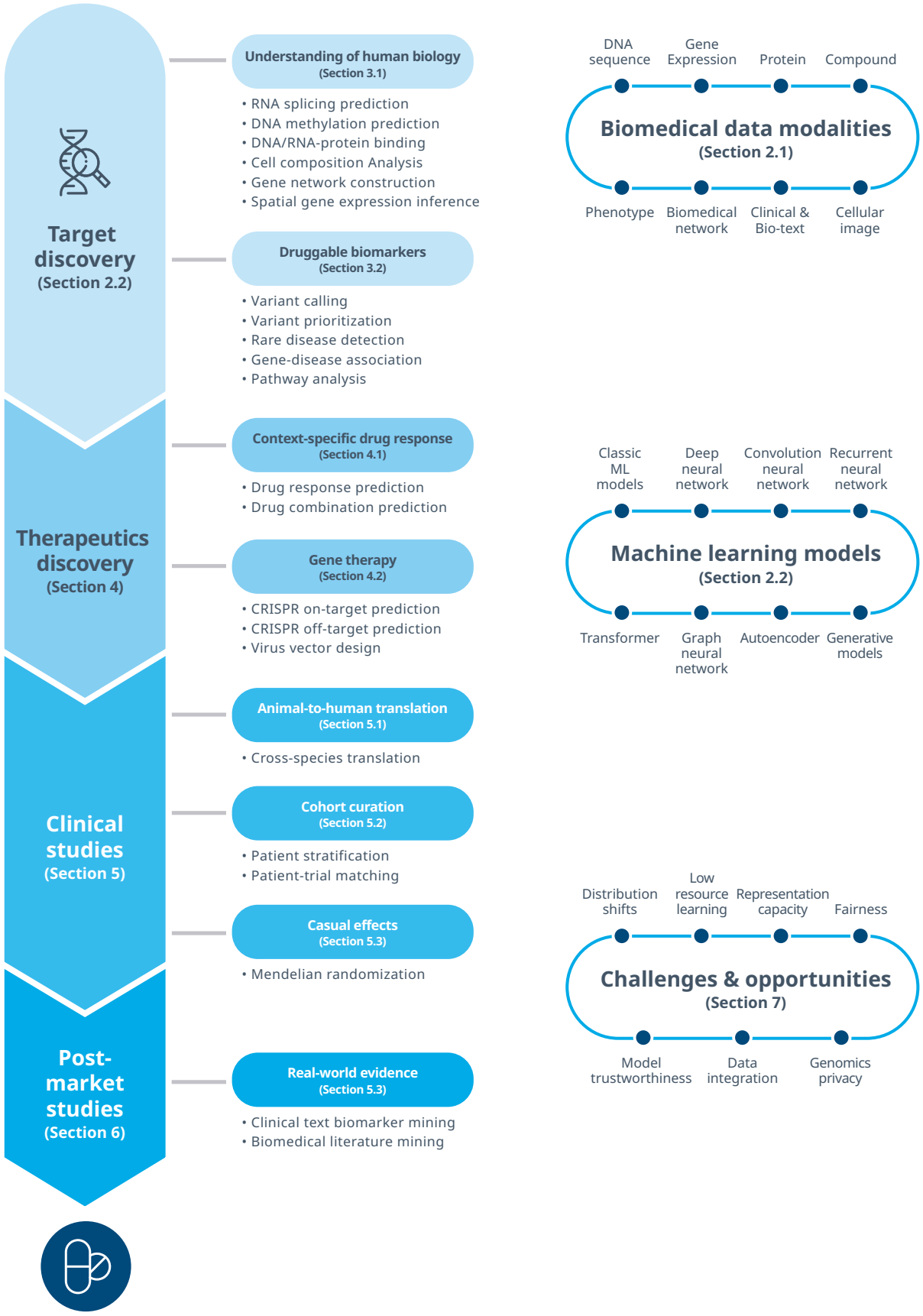
We investigated the interplay among genomics, compounds, proteins, electronic health records (EHR), cellular images and clinical texts. Twenty-two machine learning in genomics applications that span the entire therapeutics pipeline, from discovering novel targets, personalizing medicine, developing gene-editing tools, all the way to facilitating clinical trials and post-market studies, were identified.

Genomics studies the function, structure, evolution, mapping and editing of genomes and allows us to understand biological phenomena, including the roles the genome plays in diseases. This deep understanding of genomics has led to a vast array of successful therapeutics to cure a wide range of diseases. It also allows for more precise treatments or more effective therapeutics strategies, such as genome editing.

There's a lot of research to be done at the intersection of machine learning, genomics and therapeutic development.

Recent advances in high-throughput technologies have led to an outpouring of large-scale genomics data. However, the bottlenecks along the path of transforming genomics data into tangible therapeutics are innumerable. For instance, diseases are driven by multifaceted mechanisms, so to pinpoint the right disease target requires knowledge about the entire suite of biological processes. Personalized treatment requires accurate characterization of disease sub-types and the compound's sensitivity to various genomics profiles. It's our belief genomics data alone are insufficient to ensure clinical implementation, but the integration of a diverse set of data types, from compounds, proteins, cellular image and electronic health records (EHR) to

Figure 1: Organization and coverage of the survey



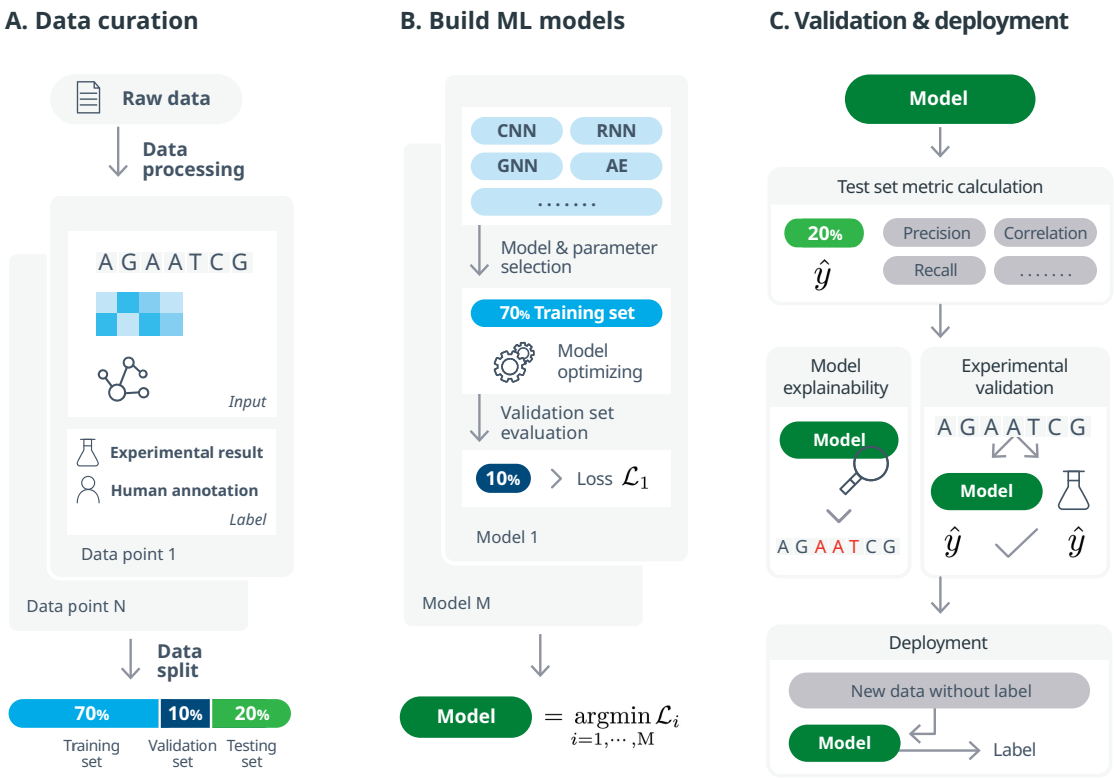
scientific literature is required. This heterogeneity and scale of data enable the application of sophisticated computational methods, such as machine learning.

ML can learn useful and novel patterns from data, often not found by experts, to improve prediction performance on various tasks (Figure 2). This ability is much needed in genomics and therapeutics. Uncovering these patterns can also lead to the discovery of novel biological insights. Therapeutic discovery often consists of large-scale resource-intensive experiments, which limit the scope of experiments, and many potent candidates are therefore missed. Using accurate prediction by ML can drastically scale up and facilitate the experiments, catching or generating novel therapeutics candidates.

Twenty-two "ML for therapeutics" tasks with genomics data were identified, ranging across the entire therapeutic pipeline. We go beyond DNA sequences and study a wide range of interactions among DNA sequences, compounds, proteins, multiomics and EHR data. ML applications have been organized into four therapeutic pipelines:

1. **Target discovery:** basic biomedical research to discover novel disease targets to enable therapeutics
2. **Therapeutic discovery:** large-scale screening designed to identify potent and safe therapeutics
3. **Clinical study:** evaluating the efficacy and safety of the therapeutics in vitro, in vivo and through clinical trials
4. **Post-market study:** monitoring the safety and efficacy of marketed therapeutics and identifying novel indications.

Figure 2: Machine learning for genomics workflow



Machine learning for genomics in:

TARGET DISCOVERY

A therapeutic target is a molecule (e.g., a protein) that plays a role in the disease's biological process. The molecule could be targeted by a drug to produce a therapeutic effect, such as inhibition, thereby blocking the disease process. Much of target discovery relies on fundamental biological research in depicting a full picture of human biology and, based on this knowledge, to identify target biomarkers. By identifying these biomarkers, we can design therapeutics to break the disease pathway and cure the disease. Machine learning can help identify these biomarkers by mining through large-scale biomedical data to predict genotype-phenotype associations accurately.

ML models are good at identifying patterns from complex patient data. For example, rare disease detection can be formulated as a classification task, similar to phenotype prediction. It aims to identify if the patient has a rare disease from the patient's genomic sequence and other information, such as her EHR data. If sufficient data from patients with a rare disease and suitable controls exist, ML models can be trained to detect rare diseases.[1]

THERAPEUTICS DISCOVERY

After a drug target is identified, a campaign to design potent therapeutic agents to modulate the target and block the disease pathway is initiated. These therapeutics can be a small molecule, an antibody, gene therapy and so on. The discovery consists of numerous phases and sub-tasks to ensure the efficacy and safety of the therapeutics.

A machine learning model can be used to predict a drug's response in a diverse set of cell lines in silico. An accurate machine learning model can greatly narrow down the drug screening space and reduce experimental costs and resources.

Drug combination therapy, also called cocktails, can expand the use of existing drugs, improve outcomes and reduce side effects. Drug cocktails can modulate multiple targets to provide a novel mechanism of action in cancer

treatments. Also, by reducing dosages for each drug, it may be possible to reduce adverse effects. Screening the entire space of possible drug combinations is not feasible experimentally. ML models that can predict synergistic responses given the drug combination and the genomic profile for a cell line can prove valuable.[2]

CLINICAL STUDIES

After a therapeutic is shown to have efficacy in the wet lab, it's further evaluated in animals and then humans in full-scale clinical trials. ML can facilitate this process using genomics data.

To study the efficacy of therapeutics in the intended or target patient groups, a clinical trial requires a precise and accurate patient population in each arm. However, due to the heterogeneity of patients, it may be difficult to recruit and enroll appropriate patients. ML can help characterize important factors for the primary endpoints and quickly identify them in patients by predicting patient molecular profiles.

Clinical trials suffer from difficulties in recruiting a sufficient number of patients. Many factors can prevent successful enrollment, including ineffective methods to identify eligible patients in the traditional manual matching system. Automated patient-trial matching using ML models could be desirable to increase enrollment by taking account into heterogeneous patient data and trial eligibility criteria (Figure 3).[3]

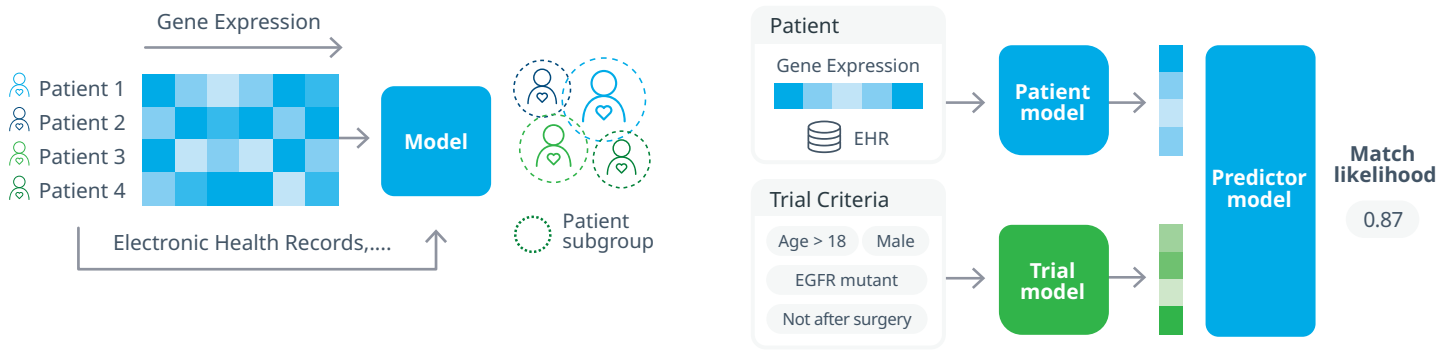
POST-MARKET STUDIES

After a therapeutic is evaluated in clinical trials and approved for marketing, numerous studies monitor its efficacy and safety when used in clinical practice. These studies contain important and often unknown information about therapeutics that was not evident before regulatory approval. ML can mine through a large corpus of texts and identify useful signals for post-market surveillance.

After therapeutics are approved and used to treat patients, voluminous documentation is generated in the EHR system, insurance billing system and scientific literature. These are called real-world data (RWD). The analyses of these data are called real-world evidence (RWE). They contain important insights about therapeutics, such as patients' drug responses given different patient characteristics.

The structured EHR data does not cover the entire picture of a patient. The majority of important variables can only be found in the clinical notes.[4] Using machine learning to automatically process clinical notes can facilitate this process.

Figure 3: Machine learning applications for therapeutic tasks with genomics data



Challenges and opportunities

Machine learning has the potential to revolutionize the use of genomics in therapeutics development. ML models work well when the training and deployment data follow the same data distribution. However, distribution shifts have been a longstanding challenge in ML, and a large body of work in model robustness and domain adaptation could be applied to genomics to improve generalizability. Another challenge is there are usually only a few drug response data points for new therapeutics. Discovering how to make an ML model learn given only a few examples is crucial.

ML models can manifest the bias in the training data. It has been shown that ML models don't work equally well on all subpopulations. ML models that perform well on the discovery population generally have much lower accuracy and are worse predictors in other populations. Since most discovery is performed with European-ancestry cohorts, predictive models may exacerbate health disparities since they will not be available for or have lower utility in African and Hispanic ancestry populations. These imbalances against minorities require specialized ML techniques. The fairness in ML is defined to make the prediction independent of protected variables such as race, gender and sexual orientation. Recent works have been proposed to ensure this criterion in the clinical ML domain.

Abundant genomics data and annotations are generated every day. Aggregation of these data and annotations can tremendously benefit ML models. However, these are usually considered private assets for individuals and contain sensitive private information and are not shareable directly. Techniques to anonymize, de-identify these data using differential privacy can potentially enable genomics data sharing. Recent advances in federated learning techniques allow ML model training on aggregated data without sharing data.

There's an interdisciplinary domain between ML and genomics and huge potential when there's collaboration across these two communities.

Conclusion

We have conducted a comprehensive review of the literature on ML applications for genomics in therapeutics development. We systematically identify diverse ML applications in genomics and provide pointers to the latest methods and resources. For ML researchers, we show that most of these applications have problems that remain unsolved, thus providing many technical challenges for ML method innovations. We also provide concise ML problem formulation to help ML researchers to approach these tasks. For biomedical researchers, we pinpoint a large set of diverse use cases of ML applications, which they can extend to novel use cases. We also introduce the popular ML models and their corresponding use cases in genomics data.

In conclusion, this survey provides an in-depth research summary of the intersection of ML, genomics, and therapeutic developments. We hope that this survey can lead to a deeper understanding of this interdisciplinary domain between ML and genomics and broaden the collaboration across these two communities. As a common belief that the future of medicine is personalized, understanding the therapeutic tasks with ML methods on genomics data is the key that will lead to ultimate breakthroughs in drug discovery and development. We hope that this survey will help to bridge the gap between genomics and ML domains.

For a more in-depth view on machine learning applications for genomics through the lens of therapeutic development, read the complete "Machine learning applications for therapeutic tasks with genomics data" paper featured in the open access journal *Patterns*.

References

1. Bojian Yin, Marleen Balvert, Rick AA van der Spek, Bas E Dutilh, Sander Bohté, Jan Veldink, and Alexander Schönhuth. Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. Bioinformatics, 35(14):i538–i547, 2019.

2. Fangfang Xia, Maulik Shukla, Thomas Brettin, Cristina Garcia-Cardona, Judith Cohn, Jonathan E Allen, Sergei Maslov, Susan L Holbeck, James H Doroshow, Yvonne A Evrard, et al. Predicting tumor cell line response to drug pairs with deep learning. BMC Bioinformatics, 19(18):71–79, 2018.

3. Jessica J Tao, Michael H Eubank, Alison M Schram, Nicholas Cangemi, Erika Pamer, Ezra Y Rosen, Nikolaus Schultz, Debyani Chakravarty, John Philip, Jaclyn F Hechtman, et al. Real-world outcomes of an automated physician support system for genome-driven oncology. JCO Precision Oncology, 3:1–13, 2019.

4. Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? unpacking predictive value in clinical note representations. AMIA Summits on Translational Science Proceedings, 2018:26, 2018.

Additional references can be found in the complete "Machine learning applications for therapeutic tasks with genomics data" paper featured in the open access journal *Patterns*.

CONTACT US

iqvia.com