Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.38850

Artificial Intelligence Assistance with CRISPR to Optimize Drug Development

Huijun Yu

Department of Chemistry, University of California, Los Angeles, USA
huijunyu@g.ucla.edu

Abstract. Drug development remains a costly and time-consuming process since identifying therapeutic targets that are both biologically relevant and clinically effective remains challenging. New technologies including bulk RNA sequencing, single-cell RNA sequencing, CRISPR screens and AI are enabling new target discovery opportunities, but there are still a number of challenges in bringing the heterogeneous data sets together. This paper describes how AI-assisted integration of these three technologies can improve drug development. This paper explains in detail how Bulk RNA sequencing can detect differentially expressed genes, how Single-cell RNA sequencing can uncover cell-specific expression patterns, how CRISPR screening can identify dependencies of genes in disease-related phenotypes, and how AI models can recognize complex molecular patterns and prioritize targets for therapeutic intervention using these complementary datasets. With AI support and current technologies, researchers could decrease false-positive target identification and enhance drug sensitivity forecasting. This paper suggest a way to improve the precision-drug-discovery strategies. But there are still issues with data heterogeneity and with model interpretability. To address the remaining challenges, interpretable AI models that combine multi-omics, functional genomic, and patient-derived data should be developed in the future.

Keywords: Artificial intelligence, CRISPR, Bulk-RNA sequencing, Single cell-RNA sequencing

1. Introduction

Nowadays, the development of novel drugs and therapies remains money- and time-consuming while also having a high failure rate. Approximately 90% of candidate drugs that enter clinical trials ultimately fail due to insufficient efficacy to treat diseases or due to unexpected toxicity that prevents them from passing the safety test [1, 2]. This insurmountable challenge is particularly evident in complex diseases such as AML. Human screening and testing are inefficient, and may require weeks, months or even years to gather and process thousands of data points. Therefore, a method or a tool to ameliorate this problem is urgent needing. As AI is developing fast in contemporary society, it could be promising to apply AI assistance with existing and feasible screening tools and datasets such as CRISPR that can identify oncogenes to improve drug development.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.38850

According to available studies, CRISPR already possesses the ability to examine disease-driving genes and to knock out genes to see if it would affect cell phenotype, DNA repair, cell proliferation, and if it would lead the cell to undergo apoptosis, which further helps us to identify genes that could contribute to diseases. But, there are some drawbacks in the CRISPR technology that make it not as accurate and effective as it could be. For instance, it might categorize some highly expressed genes that are passenger genes instead of disease-driving target genes. Also, the CRISPR technique removes each gene in thousands of genes—this may be very time-consuming. The new developments of CRISPR-based functional genomics allow for a systematic interrogation of gene function on a genome-wide scale thus increasing the rate of processing [3, 4].

However, there are still some challenges to make the CRISPR technique more efficient. Specifically, the genomes-wide CRISPR datasets are created under simplified in vitro conditions, and can't fully recapitulate the complexity of human immune interactions, spatial organization, or tumor microenvironment. On top of that, CRISPR perturbation screens are prevalent and show results that are highly dependent on disparate cancer types, genetic backgrounds, and experiments, which hinders the study from distinguishing clinically relevant targets from experimental artifacts. Thus it is not enough to use the CRISPR method to select drugs that have a high success rate in clinical applications.

The complementary high-throughput genomic datasets should be taken into account to address this problem. For instance, using bulk RNA sequencing and single-cell RNA (scRNA) sequencing with AI support could help to provide a more comprehensive understanding of the genes and processes involved in diseases and the development of treatments. AI, particularly machine learning and deep learning, is showing increasing value under such circumstances for integrating multidimensional biological data. In detail, it is now known that machine learning models can predict gene essentiality, drug sensitivity, and patient prognosis by learning patterns from massive datasets. Besides, it can also identify unrecognized molecular subtypes and cellular populations. What's more. Graph neural networks, transformer-based architectures, and multimodal deep learning models demonstrate the ability to integrate transcriptomic, genomic, proteomic, and functional screening datasets into unified predictive frameworks [5]. More importantly, these works enable simultaneous evaluation of gene expression, pathway activity, protein interactions, mutation status, and functional dependencies with lower time and economic cost, thereby optimizing target development meanwhile improve the discovery efficiency.

Although AI has ameliorated the low efficiency problem in drug development a lot, it is not perfect. Many existing computational frameworks remain limited by incomplete biological interpretability and insufficient external validation. Also, the most highlighted problem is that AI possesses biases from training datasets, which can provide wrong results. Therefore, further studies with more datasets and different AI models ought to be contemplated. By analyzing feasible papers and generating possible future optimization works upon developed techniques and AI models, this paper aims to improve the accuracy of candidate drug results and clinical relevance.

2. Single method's mechanism and limitation in treating cancer

2.1. CRISPR

CRISPR systems stem from a bacterial adaptive immune mechanism and are now artificially repurposed as a gene-editing tool. Take the most prevalent known CRISPR Cas 9 system, as an example, it possesses a single-guide RNA (sgRNA) that contains a complementary sequence to the selected target gene, which then directs the Cas 9 nuclease to the target gene section. According to

DOI: 10.54254/2753-8818/2026.38850

Jinek et al., after binding, Cas9 will utilize its HNH and RuvC nuclease domains to cleave opposite DNA strands near an adjacent protospacer-adjacent motif (PAM), producing a double-strand break that will then be repaired by cellular DNA repair pathways [6]. After that, Error-prone non-homologous end joining (NHEJ) inserts or deletes gene sections to disrupt the original gene's function. At the same time, homology-directed repair can introduce more precise sequence changes when there is an available repair template [6]. As mentioned, CRISPR has the ability to change gene sections, which then will affect the amino acid chain and protein produced by the gene, it provides an efficient method for scientists to determine how specific genetic alterations will influence cellular phenotypes through CRISPR experiments, contributing to drug development for disease associated genes validation and for determine candidate targets by studying the target's association with therapeutically relevant phenotypes [7].

On top of individual gene alteration, CRISPR screening has developed to go over such mechanisms across hundreds or thousands of genes simultaneously, contributing to identifying if a tumor phenotype depends on a particular gene section and to determining the mechanisms of drug sensitivity and resistance. Specifically, a typical pooled CRISPR knockout screen possesses a large sgRNA library that can target genes across the genome. It prevalently introduces the library into a population of Cas9-expressing cells that have low multiplicity of infection to ensure that most cells only receive approximately one genetic perturbation. Being artificially manipulated by CRISPR screening, the resulting cell population will be treated with candidate drugs, experiencing selective pressure. Subsequently, the genomic DNA will be isolated, producing next generation sequencing to quantify the abundance of individual sgRNAs. Accordingly, if the number of sgRNAs becomes depleted or decreased significantly, disruption of their target genes indicates a high probability of reducing cancer cellular fitness or increasing tumor cells' sensitivity to the treatment. On the other hand, if the number of sgRNA increased, it is demonstrated that loss of the corresponding genes provides a survival advantage or promotes drug resistance [7]. Hence, it is relatively easy and fast to test gene essentiality for cancer cell survival through CRISPR screening. According to Gilbert et al., CRISPR interference and CRISPR activation further expand this approach by decreasing or increasing gene expression rather than permanently knocking out genes [8].

Nevertheless, CRISPR encounters several restrictions, as mentioned in the introduction section, that many CRISPR screens are performed in simplified two-dimensional cell-culture systems, which means these experiments cannot fully reproduce tumor heterogeneity, interactions among different cell types, immune responses, or the in vivo microenvironment. Additionally, Loss-of-function screens can also miss important gain-of-function resistance mutations and regulatory effects that cannot be sufficiently modeled through complete gene deletion, according to Gemma et al. [9]. That said, other complementary technologies and data sets are required to ameliorate such limitations.

2.2. Bulk RNA sequencing

Bulk RNA sequencing is one technique that could provide transcriptomic profiles across a large number of cells to identify differentially expressed genes and disease-associated pathways. In general, bulk RNA sequencing will first extract RNA from the sample, converted into complementary DNA for sequencing library preparation, and then scientists will subject the complementary DNA to high-throughput sequencing to get an aligned and quantified read result, determining transcript abundance across the transcriptome [10]. The result usually appears as a count matrix, where each row represents a gene and each column represents a biological sample, normalized by differential-expression tools such as DESeq2 mentioned by Love et al, identifying genes that express in a different quantitate result which could make them be the potential disease

DOI: 10.54254/2753-8818/2026.38850

related gene [11]. Since bulk RNA sequencing possesses such an ability, it contributes to drug development by comparing gene-expression profiles between disease and healthy tissues or between drug-treated and untreated samples. Rosati et al. state that differential gene-expression analysis can be used to find the genes whose expression levels differ greatly from one experimental condition to another. These genes then can be used in further pathway analyses to link them to disease associated biological processes and therapeutic targets [12]. In particular, the genes that are highly expressed in tumor cells compared to normal cells can be selected as potential oncogenes for additional target validation; and comparing tumor and drug-treated cells before and after drug treatment can identify the pathways that are affected by the drug and help study its mechanism of action [13].

Although bulk RNA-seq has these benefits, there are also several drawbacks to its use in the discovery of drug targets. The information it offers is the average expression of the mosaic of cells, and this will obscure the gene expression of the subpopulation of cells that may give rise to therapeutic escape and tumor recurrence more often [14]. For heterogeneous diseases, like cancer, where tumor cells, immune cells, endothelial cells and other cellular populations are mixed together and form a complex microenvironment, this is a particular problem. In detail, this circumstance, which contains more abundant cellular population that the targeted ones, will mask the signals that originate from rare but biologically important cell populations, obscuring subpopulations associated with tumor progression or therapeutic resistance. Moreover, differential expression alone is not convincing proof of direct involvement of a gene in a disease phenotype, drug response or either, since changes to gene expression can be caused by both direct or other effects of the altered phenotype [14]. Therefore, there is insufficient evidence to establish a causal link between any specific targets identified by bulk RNA sequencing and cancer, and other data sets need to be incorporated, like scRNA sequencing.

2.3. ScRNA sequencing

ScRNA sequencing is another high-throughput transcriptomic technology that measures cellular gene expression. It is, however, different from bulk RNA sequencing, which pools the RNA from many cells, and therefore misses out on the characterization of cellular heterogeneity that is lost when averaging RNA from many cells [15, 16]. In general, the scRNA sequencing workflow will first need to dissociate one sample tissue into individual cells. Then, cells are separated and RNA molecules are reverse transcribed into complementary DNA (cDNA), amplified, sequenced and computationally remapped to their cells of origin, using cell-specific barcodes [16]. The gene-expression matrix can then be analyzed for distinct cell populations and cellular states within a heterogeneous sample by normalizing, reducing dimensions, clustering, comparing gene-expression patterns and annotating cells [13]. This capability of the scRNA sequencing is what helps to identify disease-associated genes and pathways in specific cell types rather than in the average signal across the tissue, which can be determined by bulk RNA sequencing [16].

Nevertheless, scRNA sequencing also holds several limitations. It might fail to detect transcripts presented due to low mRNA capture efficiency and cannot capture the complexity of the tumor microenvironment comprehensively due to changing cell interactions promoted by the tissue dissociation process [16, 17]. Accordingly, each dataset is complementary in addressing the mechanisms of tumor, providing a more thorough picture for further drug development, but each method can introduce errors in the result, which means other technologies ought to be utilized to ameliorate existing issues and to improve available technologies' efficiency.

DOI: 10.54254/2753-8818/2026.38850

3. AI assistance with available technology to ameliorate existing issues

Dealing with these existing problems, AI tools can and are already being introduced to assist available technologies such as the three mentioned before. For example, machine learning models, as mentioned, can predict gene essentiality, drug sensitivity, and patient prognosis by learning statistical relationships between molecular features like gene expression, mutations, copy number alterations, and CRISPR dependency scores. In general, AI first receives datasets provided by bulk RNA sequencing, single-cell RNA sequencing, and CRISPR screening. Then, gene features related to diseases will be analyzed and categorized for the model to learn. The trained model then applies such "past experience" to unseen samples, resulting in the prediction of potential targets.

For instance, machine-learning models like DeepDEP developed by Chiu et al. can be trained with CRISPR screening as the ground-truth labels, inputting cancer cell line features [18]. Integrating and using separate neural network encoders to compress gene expression, DNA mutations, copy number alterations, and DNA methylation, while learning how combinations of these molecular features correspond to gene dependencies identified by CRISPR, DeepDEP combines these representative example data with information regarding how the genes are being evaluated. Subsequently, the prediction network will compare the relationship between these features and experimentally measured dependency scores [18]. By doing so, machine-learning models demonstrate how a gene could be essential in tumors with a particular molecular feature combination. On top of that, to make the AI results more thorough, numerous other datasets, including patient-derived organoids, spatial omics, and longitudinal sampling, ought to be introduced to AI for learning [19].

4. Conclusion

Showing how AI models could integrate bulk RNA sequencing, single-cell RNA sequencing, and CRISPR screening to improve target identification in drug development, it is promising to introduce more AI technologies with other various types of datasets like proteomics as original training data. Accordingly, bulk RNA sequencing provides a broad transcriptomic view that enables the identification of genes that are expressed at different levels and in non-regular pathways, while single-cell RNA resolves these expression patterns at single-cell resolution, revealing disease-related cell populations and heterogeneous drug responses. CRISPR technologies, as the main technology which were used prevalently in drug development and target identification, complement these transcriptomic approaches by experimentally perturbing genes to determine whether candidate targets are functionally required for diseases. By integrating datasets generated by these three technologies and other existing or developing technologies, scientists could further connect specific genetic changes with molecular responses at the cellular level. Moreover, AI, which is evolving rapidly in today's era, can then derive insights from these massive and multi-dimensional datasets within a short time frame, gaining the ability to identify intricate connections between molecular characteristics, functional relationships, and therapeutic outcomes, providing an efficient systematic approach to prioritize potential drug targets in drug development.

However, despite the proven ability of AI to effectively incorporate complementary information and discover drug targets, there are still challenges. The source issue is the first and foremost. For example, single-cell RNA-sequencing is sub-optimal for capturing transcripts, Bulk RNA-sequencing can be noisy, and CRISPR screens can be context-dependent. Learning and incorporating these diverse and possibly ill-formed datasets may create significant challenges to computation, and make more precise prediction difficult. Thus, further studies are needed to create

DOI: 10.54254/2753-8818/2026.38850

and validate AI models with more comprehensive and external data sets that encompass transcriptomic, single-cell, functionalgenomic, proteomic, and clinical data. Also, combining these models with patient-derived organoids, spatial omics, longitudinal sampling, and CRISPR validation could further connect the computational target prediction and clinical efficacy, providing a more precise and efficient pathway from molecular data to clinical therapies, improving the efficacy of available drug development technologies.

References

[1] Paul, S. M., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery, 9*, 203–214. https://doi.org/10.1038/nrd3078.

[2] Scannell, J. W., et al. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery, 11*, 191–200. https://doi.org/10.1038/nrd3681.

[3] Shalem, O., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science, 343*, 84–87. https://doi.org/10.1126/science.1247005

[4] Wang, T., et al. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science, 343*, 80–84. https://doi.org/10.1126/science.1246981.

[5] Replogle, J. M., et al. (2022). Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. *Cell, 185*, 2559–2575. https://pubmed.ncbi.nlm.nih.gov/35688146/

[6] Jinek, M., et al. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science, 337*(6096), 816–821.

[7] Fellmann, C., et al. (2017). Cornerstones of CRISPR–Cas in drug discovery and therapy. *Nature Reviews Drug Discovery, 16*(2), 89–100. https://doi.org/10.1038/nrd.2016.238.

[8] Gilbert, L. A., et al. (2014). Genome-scale CRISPR-mediated control of gene repression and activation. *Cell, 159*(3), 647–661. 10.1016/j.cell.2014.09.029.

[9] Llargués-Sistac, G., et al. (2023). HAP1, a new revolutionary cell model for gene editing using CRISPR-Cas9. *Frontiers in Cell and Developmental Biology, 11*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10020200/

[10] Wang, Z., et al. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57–63. 10.1038/nrg2484. https://pmc.ncbi.nlm.nih.gov/articles/PMC2949280/

[11] Love, M. I., et al. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

[12] Rosati, D., et al. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal, 23*, 10.1016/j.csbj.2024.02.018

[13] Li, L., et al. (2021). What are the applications of single-cell RNA sequencing in cancer research: A systematic review. *Journal of Experimental & Clinical Cancer Research: CR, 40*(1), 163. PubMed. https://pubmed.ncbi.nlm.nih.gov/33975628/

[14] Nguyen, H., et al. (2024). Fourteen years of cellular deconvolution: Methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research, 52*(9), 4761–4783. PubMed Central. https://pubmed.ncbi.nlm.nih.gov/38619038/

[15] Heath, J. R., et al. (2015). Single-cell analysis tools for drug discovery and development. *Nature Reviews Drug Discovery, 15*(3), 204–216. https://www.nature.com/articles/nrd.2015.16. https://www.science.org/doi/10.1126/science.1225829.

[16] Van de Sande, B., et al. (2023). Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery, 22*(6), 496–520. https://www.nature.com/articles/s41573-023-00688-4

[17] Ren, X., et al. (2018). Understanding tumor ecosystems by single-cell sequencing: Promises and limitations. *Genome Biology, 19*(1), 10.1186/s13059-018-1593-z. https://pmc.ncbi.nlm.nih.gov/articles/PMC6276232/

[18] Chiu, Y.-C., et al. (2021). Predicting and characterizing a cancer dependency map of tumors with deep learning. *Science Advances, 7*(34), 10.1126/sciadv.abh1275. https://pubmed.ncbi.nlm.nih.gov/34417181/

[19] Chen, H., et al. (2026). Integrating multi-omics and artificial intelligence fuels advanced target identification and drug discovery. *Biotechnology Advances, 87*, 108785. https://www.sciencedirect.com/science/article/pii/S073497502500271X