Journal of Translational Medicine

REVIEW

Open Access

AI-driven CRISPR screening: optimizing gene editing through automation and intelligent decision support

Yuxi Zhao[1], Xingrui Li[1*] and Yaying Du[1*]

Abstract

Background CRISPR-based genetic screening has become a central methodology in functional genomics, enabling systematic interrogation of gene function, genetic interactions and context-dependent vulnerabilities at scale. However, the rapid expansion of screening modalities—including multi-condition designs, combinatorial perturbations, in vivo applications and single-cell readouts—has exposed fundamental limitations of heuristic-driven experimental design and post hoc statistical analysis.

Main body This Review synthesizes how artificial intelligence is reshaping CRISPR screening by introducing predictive, adaptive and system-level intelligence across the experimental lifecycle. We organize recent advances into two tightly coupled modules. First, machine learning and deep learning (ML/DL) methods optimize experimental design by learning context-dependent perturbation behavior, anticipating confounding effects and enabling iterative, information-efficient screening strategies. Second, large language model–agent (LLM–agent) systems complement these advances by externalizing scientific reasoning, integrating biological knowledge at scale and coordinating analysis and decision-making in human-in-the-loop workflows.

Conclusions Together, ML/DL and LLM–agent approaches reframe CRISPR screening from a static analytical pipeline into an intelligent experimental system, with important implications for robustness, scalability and biological discovery.

Keywords CRISPR screening, Machine learning, Deep learning, Large language models, Intelligent agents, Genomic editing, Functional genomics, Experimental design and automation, Human-in-the-loop system

*Correspondence:
Xingrui Li
lixingrui@tjh.tjmu.edu.cn
Yaying Du
yayingdu@hust.edu.cn
[1]Department of Thyroid and Breast Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, People's Republic of China

Introduction

Genome-scale CRISPR screening has transformed functional genetics by enabling systematic perturbation of gene function across diverse cellular contexts. Early pooled CRISPR knockout (CRISPR–KO) screening established the feasibility of genome-wide loss-of-function interrogation and provided a foundation for identifying essential genes and genetic vulnerabilities [1–3]. Large-scale efforts such as the Cancer Dependency Map further demonstrated that aggregating CRISPR screening data across cell lines can uncover both conserved and context-specific dependencies with translational relevance [4, 5]. These early successes relied on experimental and analytical paradigms that emphasized standardized library design, empirical guide scoring and downstream statistical aggregation, which proved effective for relatively homogeneous screening designs.

Over the past decade, however, the scope of CRISPR screening has expanded substantially. Perturbation modalities now include CRISPR interference (CRISPRi) and activation (CRISPRa), which enable regulatory interrogation without inducing double-strand breaks [6–8], as well as RNA-targeting systems such as Cas13 that extend screening to the transcriptome [9, 10]. In parallel, in vivo pooled screening enables functional interrogation within physiologically relevant tissue environments [11]. Most notably, single-cell CRISPR screening platforms now link genetic perturbations to high-dimensional transcriptomic phenotypes, enabling systematic mapping of gene programs and cellular states at unprecedented resolution [12–14]. Together, these advances have transformed CRISPR screening from a comparatively uniform assay into a heterogeneous family of experimental systems with distinct sources of variability and analytical demands.

Traditional CRISPR screening workflows reflect an implicit separation between design, analysis and interpretation. Experimental design is typically guided by static heuristics, while downstream analysis relies on count-based statistical frameworks such as MAGeCK [15, 16] or BAGEL [17] to identify depleted or enriched perturbations, followed by manual biological interpretation. Although this division of labour has proven robust for simple viability screens, it is increasingly strained by the complexity of modern designs. Large-scale datasets have revealed that systematic confounders—including copy-number–associated cutting toxicity, heterogeneous guide efficacy and selection dynamics—can dominate observed signals and generate false positives that cannot be eliminated by post hoc statistical filtering alone [18, 19].

Mechanistic inference models such as CERES [20] and, more recently, Chronos [21] address these limitations by explicitly modeling key confounding processes rather than treating them as residual noise. Chronos in particular demonstrated that accurate inference of gene essentiality requires integrating guide efficacy, temporal dynamics and experimental noise within a unified population-level framework [21]. These developments point to a broader conceptual shift: as CRISPR screening becomes increasingly complex and context-dependent, analytical success depends less on downstream correction and more on designing experiments that are informative by construction.

This realization motivates an AI-centred rethinking of CRISPR screening paradigms. In this Review, we frame this transition through a dual-module perspective (Fig. 1). Machine learning and deep learning (ML/DL) approaches address how CRISPR experiments can be designed to succeed by learning context-dependent perturbation behavior and optimizing experimental parameters prior to data generation. LLM–agent systems address the complementary challenge of interpreting results and guiding subsequent actions by integrating biological knowledge, orchestrating analytical tools and supporting transparent, human-in-the-loop decision-making. Together, these complementary AI layers recast CRISPR screening from a linear pipeline into an adaptive experimental system capable of iterative refinement.

ML/DL-enabled optimization of CRISPR screening design

Context-dependent gRNA design as preventive confound control

Guide RNA (gRNA) design represents one of the earliest and most visible intersections between artificial intelligence and CRISPR screening. Empirical design rules and sequence-based scoring schemes improved single guide RNA (sgRNA) efficiency and reduced off-target activity, establishing a practical baseline for genome-scale screens [22, 23]. However, these approaches implicitly assume that guide performance is largely sequence-intrinsic and broadly transferable across contexts. Accumulating evidence indicates that sgRNA efficacy is strongly context-dependent, shaped by chromatin accessibility, transcriptional activity, delivery modality and perturbation mechanism, such that identical guides can exhibit markedly different behaviors across screening platforms and biological settings.

Recent ML/DL approaches therefore recast guide selection as a conditional prediction problem rather than a static ranking task, enabling gRNA libraries to be optimized for specific experimental regimes and anticipated sources of variability. A notable example is GuideScan 1/2 [24, 25], which integrates genome-wide specificity analysis with empirical screening data to show that low-specificity guides introduce systematic confounding across CRISPR–KO, CRISPRi and CRISPRa screens, and that specificity-aware library design can reduce

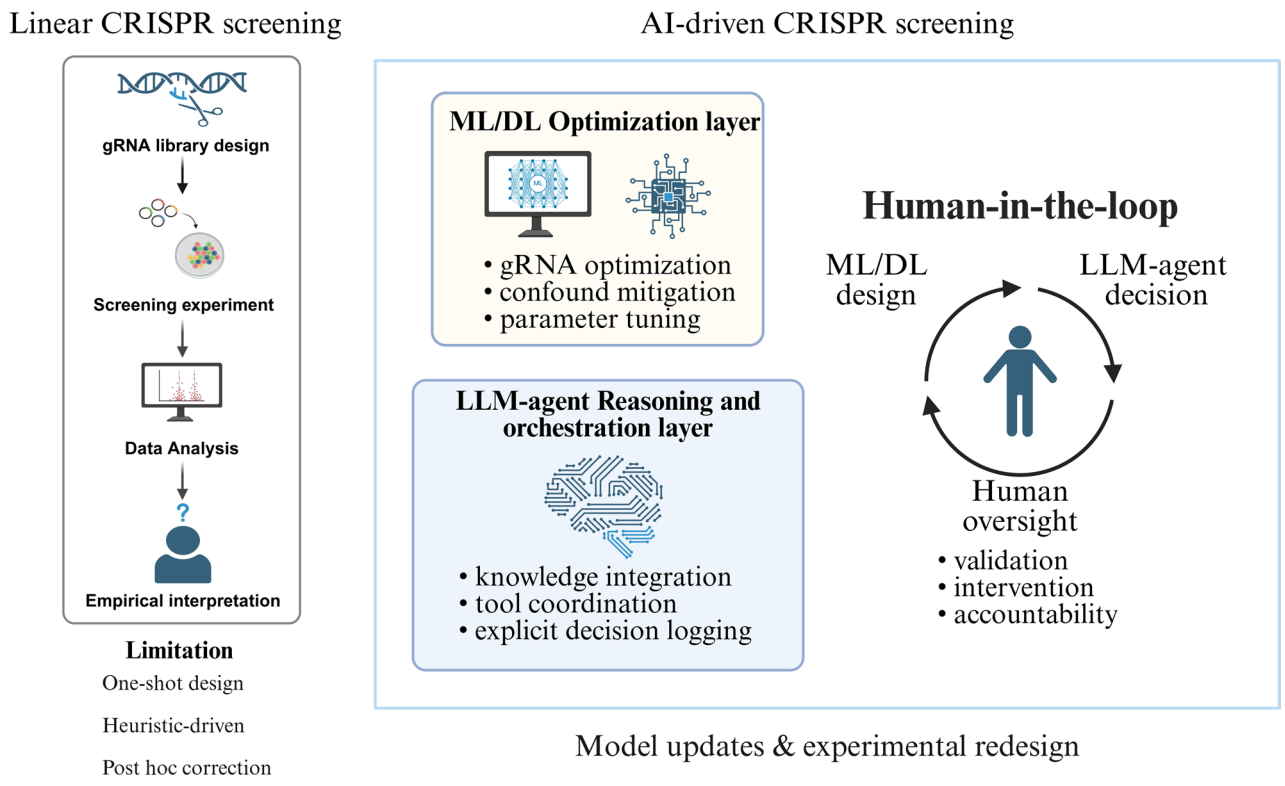

Fig. 1 Conceptual comparison between traditional and AI-driven CRISPR screening workflows. Traditional CRISPR screening typically follows a linear, one-shot pipeline, in which gRNA library design, pooled screening, data analysis and biological interpretation are conducted as largely independent stages, with confounding effects addressed primarily through post hoc statistical correction. In contrast, AI-driven CRISPR screening introduces a modular architecture composed of an ML/DL optimization layer and an LLM–agent reasoning layer. ML/DL models support data-driven optimization of gRNA design, confound mitigation and parameter tuning, while LLM–agent systems integrate functional genomics knowledge, coordinate analytical tools and log decision rationales. Human oversight is embedded throughout the workflow to validate proposed actions, intervene when assumptions fail and ensure accountability. Together, these components transform CRISPR screening from a static analytical pipeline into an adaptive, decision-aware and human-in-the-loop experimental system

artifactual dependencies at the experiment level, shifting bias management upstream rather than relying on downstream correction. Similar predictive models of gRNA targeting activity [26, 27] can inform rational library design by prioritizing perturbations with improved efficacy and reduced confounding, moving beyond ad hoc guide exclusion and post hoc adjustment that treat confounding as an unavoidable cost of scale.

Over the past five years, the accumulation of large and diverse screening datasets has enabled ML-driven approaches that treat specificity, efficacy and toxicity as explicit design constraints learned from data. RNA-targeting CRISPR screening with Cas13 provides a clear illustration: TIGER [10], trained on large-scale Cas13d screening data, uses deep learning to predict guide efficacy and mismatch tolerance directly from sequence, enabling rational construction of RNA-targeting libraries with improved performance and fewer unintended effects. By learning from screening outcomes aligned with the intended use case, TIGER enables design decisions to be made before experiments are performed, rather than inferred retrospectively.

A parallel design-first logic has also emerged for DNA-targeting CRISPR screening. Deep learning models trained on large-scale Cas9 datasets learn non-linear sequence determinants of cleavage efficiency and support principled prioritization of highly active guides at genome scale; for example, DeepSpCas9 [28] improves prediction of SpCas9 activity beyond rule-based scoring. Related work extending efficiency prediction across Cas13d orthologs further suggests that learned representations can generalize across effectors and biological contexts [29], reinforcing the broader conclusion that ML/DL transforms screening data into a predictive design engine that mitigates predictable confounding at the source rather than correcting it after the fact.

Revealing weak and non-linear genetic effects from bulk to single-cell CRISPR screening

Many biologically important genetic effects are subtle, conditional or non-linear, emerging only under specific perturbations or environmental contexts. In bulk pooled CRISPR screening, such effects are often obscured when guide-level measurements are aggregated into gene-level

statistics, particularly in the presence of heterogeneous guide efficacy and complex noise structures. ML/DL-based modeling strategies address these limitations by explicitly representing guide activity, experimental dynamics and confounding factors, thereby redistributing statistical power toward informative perturbations. Mechanistic population-level models such as Chronos [21] demonstrate how incorporating guide efficacy and temporal dynamics improves inference of gene fitness effects across heterogeneous datasets, enabling more reliable detection of weak but reproducible dependencies. Complementary approaches, including copy-number–aware correction models and hierarchical Bayesian frameworks that jointly infer guide- and gene-level effects, further illustrate how structured modeling can rescue subtle biological signals that would otherwise be lost in conventional analyses [20, 30]. Together, these methods indicate that ML/DL reshapes not only statistical estimation but also the effective resolution of bulk CRISPR screening, determining which classes of genetic effects become observable in practice (Fig. 2).

This challenge becomes even more pronounced in single-cell CRISPR screening, where increased phenotypic resolution is accompanied by orders-of-magnitude higher dimensionality. By linking perturbations to transcriptome-wide changes, single-cell CRISPR screening enables systematic discovery of gene programs and state-dependent effects that are inaccessible to bulk readouts [12–14]. However, sparsity, confounding and model misspecification can dominate signal structure, particularly in low-multiplicity designs, inflating false-positive rates under naïve differential expression testing. A recent comprehensive benchmark highlighted these failure modes and underscored the need for rigorous, model-based inference tailored to the data-generating process of single-cell CRISPR screening [31]. Recent AI-enabled

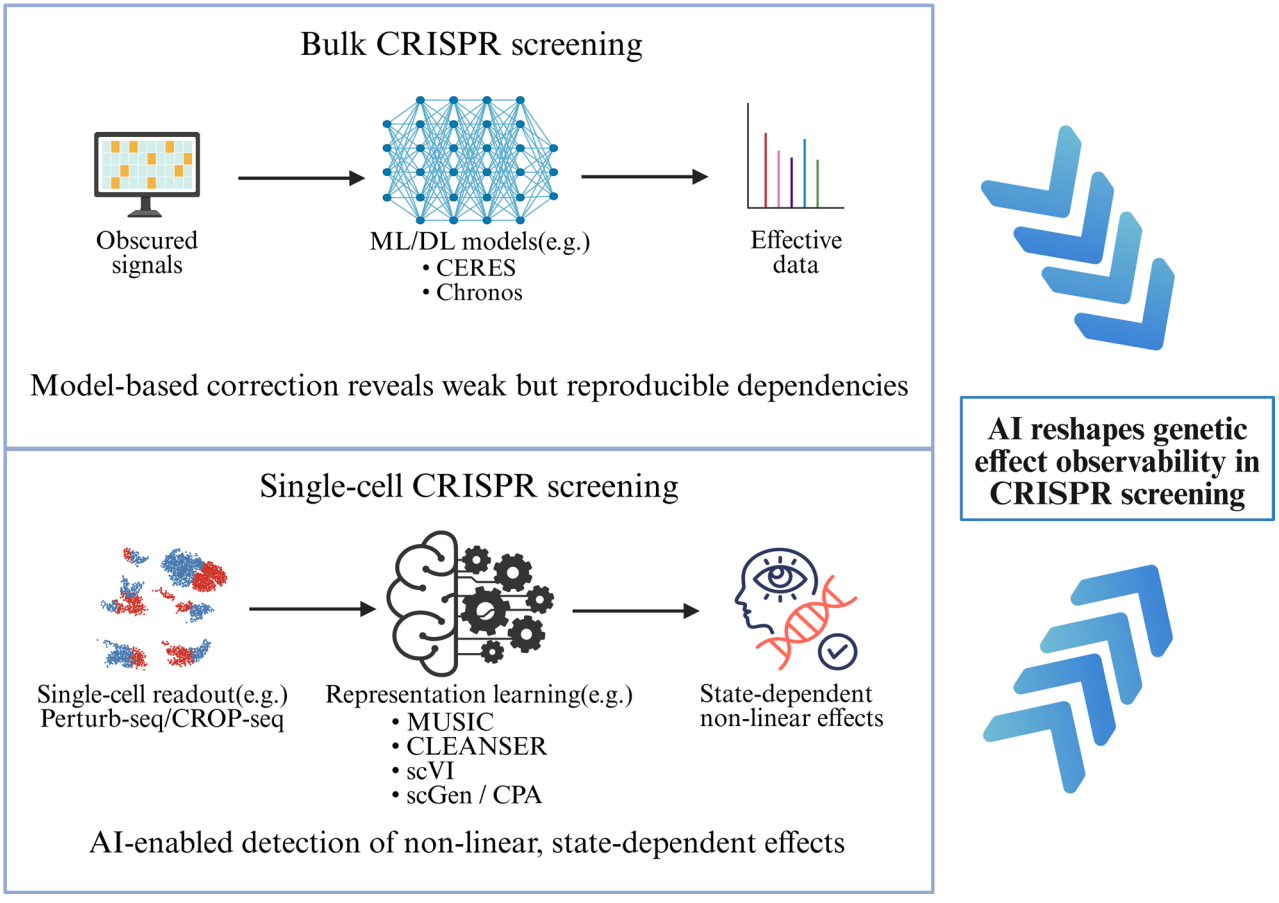

Fig. 2 AI-enabled reshaping of genetic effect observability across CRISPR screening scales. In bulk pooled CRISPR screening, heterogeneous gRNA efficacy, copy-number–associated toxicity and population dynamics obscure weak but reproducible genetic dependencies when guide-level measurements are aggregated. ML/DL-based inference models, such as CERES and Chronos, explicitly model these confounding factors and redistribute statistical power toward informative perturbations, enabling recovery of subtle fitness effects. In single-cell CRISPR screening, increased phenotypic resolution is accompanied by high dimensionality, sparsity and technical noise. Representation learning and perturbation-aware models—including scVI, MUSIC, CLEANSER, scGen and compositional perturbation autoencoders—denoise high-dimensional transcriptomic data, model latent perturbation structure and reveal non-linear, state-dependent genetic effects that are inaccessible to bulk readouts. Across scales, AI methods reshape which classes of genetic effects become statistically detectable in CRISPR screening

methods address these challenges by modeling latent perturbation structure and technical noise directly. For example, the topic-modeling framework MUSIC [32] aggregates weak but coherent transcriptional responses across cells by learning shared perturbation-associated programs, substantially improving sensitivity to effects diluted at the single-gene level. In parallel, mixture-model–based approaches such as CLEANSER [33] explicitly correct for ambient guide contamination and misassignment, identifying and filtering a substantial fraction of cells as non-informative or misassigned (dataset- and threshold-dependent; ~40–45% reported in the original evaluation across multiple Perturb-seq datasets), thereby improving effect-size estimation and false-discovery control.

ML/DL approaches further extend this logic through representation learning frameworks that capture structured variation in high-dimensional single-cell data. Deep generative models such as scVI [34] enable denoising and latent structure learning, while perturbation-aware models, including scGen [35] and compositional perturbation autoencoders (CPA), explicitly model how genetic or chemical perturbations transform cellular states in latent space [36]. Noise-aware deep learning models such as CrisprDNT [37] incorporate denoising objectives directly into training, enabling robust recovery of perturbation effects under low signal-to-noise conditions and yielding consistent gains in classification accuracy and hit reproducibility relative to conventional pipelines. Because perturbation assignment uncertainty and technical noise directly shape the effective sample size and variance structure, denoising alters not only signal clarity but the statistical regime under which effects are tested. These results highlight that AI-driven denoising functions not merely as preprocessing but as a quantitative determinant of which weak, non-linear and context-dependent genetic effects can be detected in single-cell CRISPR screening.

When integrated with single-cell CRISPR screening, these approaches support robust detection of subtle perturbation effects and enable information-driven experimental planning, in which follow-up perturbations are selected based on expected information gain rather than heuristic prioritization. Viewed collectively, these developments establish a coherent trajectory from bulk to single-cell CRISPR screening in which ML/DL enables weak, context-dependent genetic effects to become discoverable across increasingly complex phenotypic landscapes (Fig. 2). A defining advantage of ML/DL is therefore its capacity to shift CRISPR screening from a predominantly one-shot assay toward an iterative, model-updating experimental process, in which each screening round refines predictive models of guide behavior, noise and uncertainty, improving the design and expected informativeness of subsequent experiments. This closed-loop logic aligns CRISPR screening with "self-driving laboratory" paradigms in adjacent fields, where model-guided planning and increasingly automated execution enable efficient navigation of large experimental spaces [38–41]. As combinatorial perturbation modalities expand the design space while practical constraints favour targeted refinement over exhaustive enumeration, iterative strategies coupling screening outcomes to adaptive redesign are likely to become a central organizing principle for next-generation CRISPR screening.

Knowledge-integrated and LLM–agent-orchestrated CRISPR screening

From manual interpretation to knowledge-integrated reasoning

Even optimally designed CRISPR screens ultimately generate outputs that must be translated into mechanistic insight, a step that has traditionally relied on manual literature review, pathway annotation and expert intuition. As CRISPR screening datasets expand in scale, perturbation modality and phenotypic complexity, this human-centred interpretive layer increasingly becomes a bottleneck, limiting both scalability and reproducibility. Domain-adapted language models such as SciBERT, BioBERT and PubMedBERT [42–44] provided an essential foundation for addressing this challenge by demonstrating that pretraining on scientific and biomedical corpora substantially improves literature understanding, entity recognition and knowledge extraction relevant to functional genomics and multi-omics data interpretation. In the context of CRISPR, RNAi and base editor screening, these models enable more systematic annotation of screening hits, automated aggregation of prior evidence across genes, pathways and disease contexts, and structured synthesis of background knowledge at a scale that would be impractical to curate manually [45].

Generative language models further extend these capabilities from information extraction toward integrative reasoning. Work summarizing the broad application of LLMs in bioinformatics highlights their ability to produce biologically relevant summaries, synthesize insights across genomic domains, and support downstream interpretive tasks such as functional annotation and candidate prioritization [46]. Similarly, systematic benchmarking of LLMs on genomic knowledge tasks demonstrates that large language models possess substantive latent genomic understanding and can be combined with domain-specific tools to improve genomic intelligence, suggesting utility in contexts such as interpreting perturbation screens [47]. Such generative reasoning is particularly valuable for interpreting subtle, context-dependent effects commonly observed in genetic perturbation screens, where individual hits may appear

weak or ambiguous in isolation but gain biological meaning when contextualized through integrated knowledge and explicit evidence synthesis.

Agentic extensions of LLM reasoning further enhance interpretive capacity by explicitly structuring reasoning over diverse knowledge sources. Emerging frameworks in transcriptomic analysis illustrate how interpretable, evidence-grounded LLM-based agents can produce transparent and reproducible functional explanations anchored to literature and pathway resources [48]. These developments point toward systems that integrate screening outputs with pathway databases, genetic interaction maps and disease annotations to produce citation-backed rationales and auditable decision traces that externalize analytical judgement typically embedded implicitly in expert interpretation. Collectively, such knowledge-integrated LLM reasoning frameworks signal a transition from manual, ad hoc interpretation toward scalable, structured reasoning systems that augment human expertise while preserving interpretability, positioning LLM-based reasoning as a critical complement to ML/DL-driven experimental design in genetic perturbation screening.

LLM-agent-driven automated analysis and experimental planning

Importantly, in CRISPR screening, interpretive reasoning is inseparable from experimental decision-making, as biological interpretation directly informs model selection, prioritization criteria and the design of subsequent perturbations. Accordingly, integrating LLM-powered agents into gene-editing workflows promises not merely incremental efficiency gains but a reconfiguration of experimental orchestration in functional genomics. Static pipelines struggle with the branching decisions inherent to CRISPR screening—quality control, model choice, confound diagnostics, prioritization criteria and corrective strategies—yet agentic systems can re-plan workflows in response to emerging evidence, maintain explicit decision trails, and propose corrective actions in an auditable form. Foundational perspectives have articulated how AI agents configured through tool use, role assignment and feedback-driven refinement can take on scientific roles—planning, tool invocation, execution and interpretation—positioning agents as collaborators rather than mere interpretive aids in experimental science [49]. Complementary work defines the emerging paradigm of agentic bioinformatics, in which autonomous yet auditable systems tackle complex biological data workflows through dynamic planning-and-execution loops, enabling adaptive decision logic and continual hypothesis refinement [50].

A prominent example is CRISPR-GPT [51], an LLM–agent system that integrates structured reasoning, task decomposition, curated domain knowledge and programmatic tool interaction to automate end-to-end CRISPR experiment design and analysis. Rather than functioning as a passive assistant, CRISPR-GPT operates as a goal-directed planning agent, translating high-level objectives into executable workflows spanning modality selection (including knockout, CRISPRi/a, base editing and prime editing), gRNA design and prioritization, delivery strategy choice and downstream data interpretation. By coupling LLM planning to deterministic bioinformatics tools and curated databases, the system externalizes intermediate decision points, improves interpretability and limits failure propagation across multi-step workflows. Empirical validation across diverse human gene-editing tasks demonstrated that CRISPR-GPT demonstrated performance comparable, on the evaluated task suite and metrics, to expert-designed workflows on the evaluated task suite and metrics in both design quality and analytical robustness, while maintaining transparency through structured reasoning traces [51]. These results support the view that LLM planners can function as co-pilots for complex gene-editing workflows under human supervision, rather than as purely conversational interfaces.

Beyond gene-editing–specific tasks, multi-agent scientific systems such as SciToolAgent [52] leverage knowledge graphs and retrieval-augmented generation to orchestrate large tool suites across complex workflows, illustrating architectures in which agents mediate between domain tools and high-level scientific objectives. Similarly, frameworks such as GenoMAS [53] show how teams of LLM-based agents can maintain logical coherence across multi-step genomic analyses while adapting to dataset idiosyncrasies, with interpretability preserved through explicit, inspectable reasoning artifacts. In the broader bioinformatics landscape, systems such as BioAgents [54] propose multi-agent workflows tailored to autonomy and personalization, reporting performance on conceptual genomics tasks comparable to human experts and further underscoring the feasibility of agentic orchestration in analysis pipelines.

Collectively, these developments point to adaptive, auditable orchestration layers that couple "reasoning" (LLM planning) with "acting" (tool execution), enabling scientific workflows that are resilient to assumption drift and responsive to experimental feedback. This architectural shift supports not only gene editing but also scalable bioinformatics and discovery workflows that can dynamically restructure analysis strategies, propose corrective interventions and maintain structured reasoning histories that remain accessible to human researchers and regulators [55, 56].

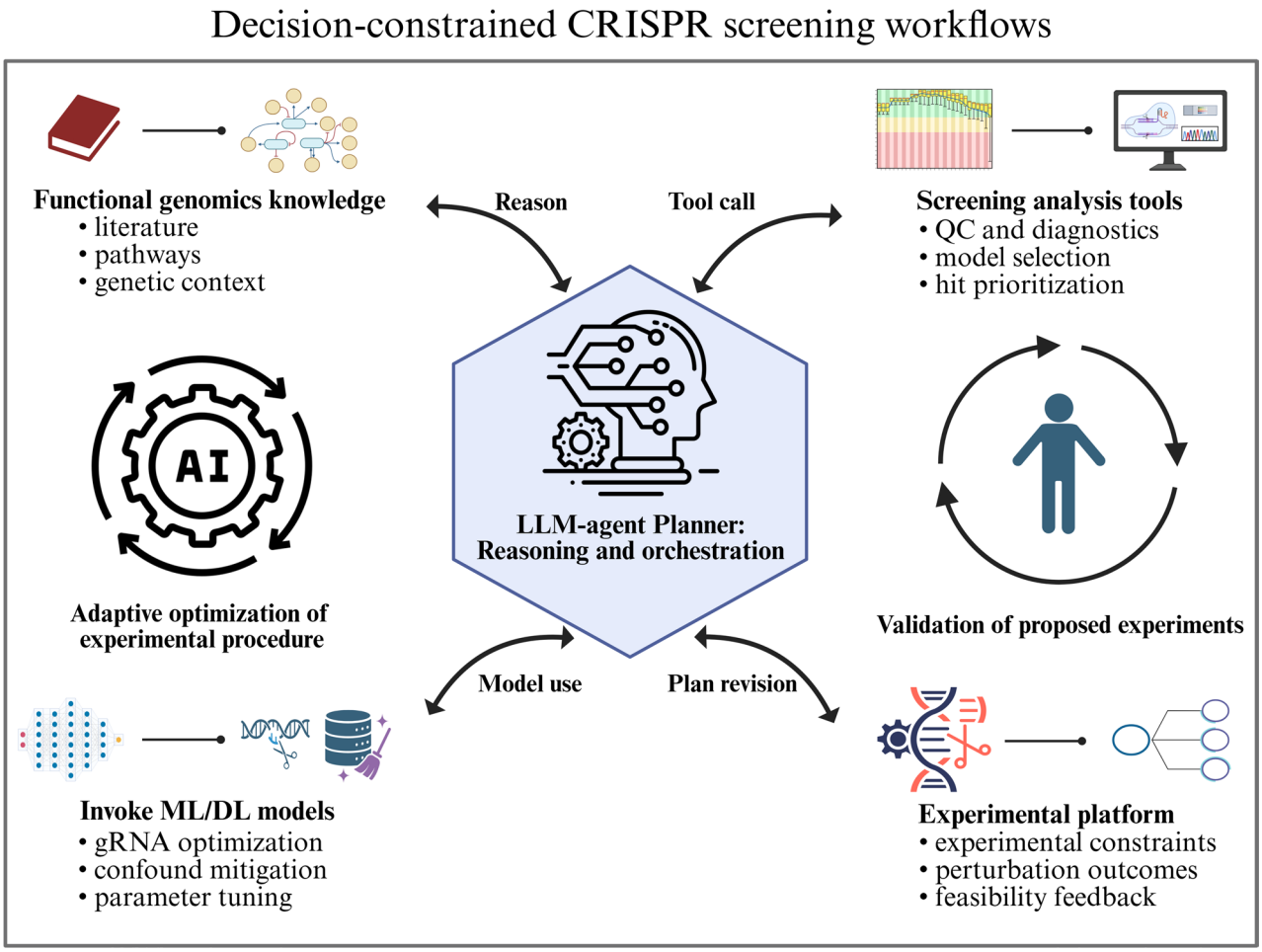

Fig. 3 Decision-constrained CRISPR screening enabled by LLM–agent orchestration and ML/DL optimization. An LLM–agent planner serves as a central orchestration layer that integrates functional genomics knowledge, screening analysis tools and experimental constraints to support adaptive decision-making in CRISPR screening. The agent reasons over prior biological knowledge and screening outputs, invokes ML/DL models for gRNA optimization, confound mitigation and parameter tuning, and updates experimental plans based on feasibility feedback from experimental platforms. Human experts remain embedded in the loop to validate proposed experiments, intervene when necessary and ensure accountability. This architecture enables closed-loop CRISPR screening workflows in which analytical reasoning, model-driven optimization and experimental execution are iteratively coupled under explicit decision constraints

Human-in-the-loop closed-loop screening systems

The convergence of ML/DL design models with LLM–agent reasoning is enabling human-in-the-loop, closed-loop CRISPR screening systems, in which experimental outcomes iteratively update predictive models and directly inform subsequent experimental choices (Fig. 3). In this setting, CRISPR screening is no longer treated as a terminal analytical exercise but as a sequential decision process, where each experimental round refines both biological hypotheses and the computational models that guide future perturbations. Prior demonstrations of agentic gene-editing systems, including CRISPR-GPT, illustrate how LLM-based planners can coordinate domain knowledge with programmatic tool use to support iterative experiment design and interpretation under human supervision, providing a concrete instantiation of evidence-driven iteration rather than a fixed analytical endpoint [51].

Critically, the value of agentic systems in CRISPR screening is not limited to automation or throughput gains; it lies in externalizing and structuring scientific decision-making in ways that are explicit, inspectable and reproducible (Fig. 3). Conceptual syntheses argue that biomedical AI agents should be constructed as collaborative systems that integrate predictive models, analytical tools and experimental platforms with feedback-driven refinement, thereby supporting long-horizon scientific workflows that remain aligned with human judgment [49]. Crucially, decisions are made under explicit experimental, biological and operational constraints—including perturbation feasibility, platform limitations and safety considerations—rather than unconstrained optimization objectives. Consistent with the emerging paradigm

of agentic bioinformatics, such systems emphasize autonomy bounded by auditability—treating provenance (what was done), rationale (why it was done) and intervention points (where a human can override or redirect the workflow) as first-class design principles rather than afterthoughts [50]. Notably, closely related requirements have been formalized in medical AI evaluation and reporting frameworks, which treat system versioning, input data acquisition, human–AI interaction context and error analysis as essential metadata for transparency and reproducibility in real-world deployment—principles that map naturally onto agent-managed, tool-using screening pipelines [57, 58].

Methodologically, this supervision-friendly behavior is underpinned by frameworks that interleave reasoning traces with tool-mediated actions, enabling agents to articulate hypotheses, justify diagnostic steps, execute analyses via external tools and revise plans in response to returned evidence. The ReAct paradigm and related reasoning–acting frameworks exemplify this coupling of explanation and execution, ensuring that intermediate decisions remain visible and contestable rather than implicit within opaque pipelines [55].

In practice, such architectures support closed-loop CRISPR screening workflows in which anomalous signals or violated assumptions trigger targeted diagnostics, alternative models or revised prioritization criteria, preventing error propagation and enabling adaptive correction under human oversight rather than silent failure within static workflows. This "bounded autonomy" mirrors clinical expectations for early-stage, live evaluation of AI decision-support systems, where human factors, safety monitoring and iterative refinement are treated as core evaluation targets rather than post hoc considerations. More broadly, the convergent view in high-performance medicine is that the goal is not maximal automation but a calibrated division of labor between humans and AI—an orientation that strengthens the case for agentic screening systems designed around auditability, override points and continual recalibration [59].

Viewed through this lens, the evolution of CRISPR screening workflows reflects a broader shift in computational biology: from pipelines optimized for efficiency under fixed assumptions toward reasoning-centric systems designed to operate under uncertainty. Knowledge-integrated LLM reasoning addresses the interpretive bottleneck inherent to large-scale perturbation data, while agentic orchestration externalizes the decision logic that governs how analytical methods and experimental designs are selected, combined and revised over time. When embedded within human-in-the-loop, closed-loop architectures, these capabilities enable adaptive, data-driven experimentation without relinquishing scientific accountability. Rather than replacing established statistical models or experimental expertise, LLM-agent systems reconfigure how such components are coordinated, positioning structured reasoning and transparent oversight as central pillars of next-generation CRISPR screening.

Challenges and future directions

Despite rapid progress, the deployment of ML/DL- and LLM–agent-enabled CRISPR screening systems raises a set of intertwined challenges spanning model generalization, interpretability, evaluation and experimental integration. A persistent limitation of ML/DL design and analysis models lies in their ability to generalize beyond the conditions represented in their training data. Models trained on large-scale CRISPR compendia often perform well in distribution but can degrade substantially when transferred across cell types, perturbation modalities, laboratories or screening conditions, reflecting sensitivity to domain shift and partially unmodelled confounders [20, 21, 60]. In pooled CRISPR knockout screens, for example, copy-number–associated cutting toxicity and population dynamics can dominate apparent gene dependency if not explicitly modelled, necessitating mechanistic correction frameworks such as CERES and Chronos [20, 21]. As screening designs increasingly incorporate diverse perturbation modalities, time-resolved readouts and single-cell phenotypes, batch effects and context-specific biases further complicate cross-study transfer and highlight the need for uncertainty-aware, mechanism-informed and domain-robust modeling approaches [60, 61]. Collectively, these challenges suggest that the limiting factor for next-generation CRISPR screening is no longer data volume, but the design of trustworthy decision systems that can operate under uncertainty without obscuring responsibility.

LLM–agent-based systems introduce complementary opportunities and risks. While agentic layers can integrate heterogeneous evidence and orchestrate adaptive workflows, their reasoning processes remain vulnerable to hallucination, implicit assumption drift and overconfident synthesis—failure modes that are particularly consequential when systems propose experimental actions rather than post hoc interpretations [62]. Retrieval-augmented generation mitigates some risks by grounding responses in external sources, yet errors can still arise from incomplete retrieval, biased corpora or faulty synthesis [63]. Tool-using paradigms such as ReAct and Toolformer further improve reliability by coupling reasoning to explicit actions and external checks, but they also shift the burden of trust toward tool provenance, decision-trace completeness and the correctness of intermediate arguments and tool outputs [21, 55]. Consequently, a central future challenge is to move from narrative-style explanations toward verifiable, structured

reasoning artifacts—including citation-backed claims, explicit uncertainty estimates and auditable decision logs—that can be inspected and overridden by human experts.

These concerns intersect directly with the problem of evaluation and reproducibility. Although established statistical methods such as MAGeCK provide robust benchmarks for hit calling, there is no widely accepted framework for evaluating end-to-end agentic screening workflows, which encompass quality control, confound diagnosis, model selection, prioritization and iterative decision-making under uncertainty [15]. General agent benchmarks demonstrate the feasibility of testing planning and tool use in interactive environments, but biology-specific benchmarks must additionally encode experimental constraints, control behavior and modality-specific artifacts that are intrinsic to perturbation screens [64]. Reproducibility further depends on standardized data and metadata practices: without machine-readable provenance, parameter tracking and workflow versioning, the same agentic system may yield irreproducible outcomes across laboratories or software updates. The FAIR principles provide a foundational framework for addressing these issues, but their systematic adoption remains uneven in functional genomics [65].

Finally, the realization of truly closed-loop CRISPR screening depends on deep integration with automated experimental platforms. Experience from self-driving laboratories and biofoundries suggests that, beyond modeling, execution reliability, fault handling and real-time data integration become dominant bottlenecks as autonomy increases [66–68]. Robotic platforms capable of long-duration biological experimentation highlight the importance of standardized, machine-interpretable protocols and feedback channels that allow computational agents to adapt plans safely in response to experimental outcomes [67, 68]. In this context, LLM–agent systems will need to interface seamlessly with laboratory automation while operating under explicit human oversight, in line with emerging governance frameworks that emphasize transparency, accountability and controllability for high-impact AI systems [69–71]. Progress in this area will require coordinated efforts across algorithm development, experimental standardization and community-driven benchmarking. Meeting these requirements in concert—rather than addressing them piecemeal—will be essential for translating agent-enabled CRISPR screening from promising prototypes into robust, trustworthy components of routine functional genomics.

Conclusion

CRISPR screening has evolved from a largely standardized, one-shot assay into a diverse and increasingly complex experimental landscape, encompassing multiple perturbation modalities, high-dimensional phenotypic readouts and context-dependent biological questions. This evolution exposes fundamental limitations of heuristic-driven design, static analysis pipelines and manual interpretation, while simultaneously creating an opportunity for artificial intelligence to reshape how functional genomics experiments are conceived, executed and interpreted.

Across this Review, we have outlined a unifying framework in which ML/DL methods and LLM–agent systems play complementary roles. ML/DL models transform accumulated screening data into data-driven optimization layers that inform guide selection, bias mitigation and experimental parameter tuning that anticipate guide behavior, confounding effects and experimental uncertainty, enabling more informative experiments by construction. LLM–agent systems, in turn, externalize scientific reasoning by integrating biological knowledge, coordinating analytical tools and supporting transparent, human-in-the-loop decision-making across the experimental lifecycle. Notably, this division of labour does not imply a hierarchy between ML/DL and LLM–agent systems, but rather reflects complementary forms of intelligence operating at different levels of abstraction. Together, these components recast CRISPR screening as an adaptive, closed-loop system rather than a linear analytical pipeline.

Looking forward, the impact of AI on CRISPR screening will depend not only on algorithmic advances, but also on progress in evaluation standards, interpretability, reproducibility and integration with automated experimental platforms. Addressing these challenges will be essential to ensure that increasingly autonomous systems remain scientifically trustworthy and biologically grounded. If successfully navigated, the convergence of predictive modeling, agentic reasoning and experimental automation has the potential to elevate CRISPR screening from a high-throughput discovery tool to a continuously learning experimental system, capable of navigating complex genotype–phenotype landscapes with both efficiency and accountability.

Author contributions

All the authors read and approved the final manuscript.

Funding

This work was supported by the Bethune Charitable Foundation (No. Z04J2024E107-B-12).

Data availability

All of the material is owned by the authors, and is available upon reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable.

Conflict of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the work reported.

Received: 28 November 2025 / Accepted: 5 February 2026
Published online: 19 February 2026

References

1. Shalem O, Sanjana NE, Hartenian E, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343(6166):84–7. https://doi.org/10.1126/science.1247005.
2. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343(6166):80–4. https://doi.org/10.1126/science.1246981.
3. Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32(3):267–73. https://doi.org/10.1038/nbt.2800.
4. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a cancer dependency map. Cell. 2017;170(3):564–e57616. https://doi.org/10.1016/j.cell.2017.06.010.
5. Behan FM, Iorio F, Picco G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature. 2019;568(7753):511–6. https://doi.org/10.1038/s41586-019-1103-9.
6. Gilbert LA, Larson MH, Morsut L, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell. 2013;154(2):442–51. https://doi.org/10.1016/j.cell.2013.06.044.
7. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014;159(3):647–61. https://doi.org/10.1016/j.cell.2014.09.029.
8. Horlbeck MA, Xu A, Wang M, et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife. 2016;5:e19760. https://doi.org/10.7554/eLife.19760.
9. Abudayyeh OO, Gootenberg JS, Konermann S, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science. 2016;353(6299):aaf5573. https://doi.org/10.1126/science.aaf5573.
10. Wessels HH, Méndez-Mancilla A, Guo X, et al. Prediction of on-target and off-target activity of CRISPR–Cas13d guide RNAs using deep learning. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-023-01830-8.
11. Santinha AJ, Strano A, Platt RJ. Methods and applications of in vivo CRISPR screening. Nat Rev Genet. 2025;26(10):702–18. https://doi.org/10.1038/s41576-025-00873-8.
12. Dixit A, Parnas O, Li B, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell. 2016;167(7):1853–e186617. https://doi.org/10.1016/j.cell.2016.11.038.
13. Datlinger P, Rendeiro AF, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14(3):297–301. https://doi.org/10.1038/nmeth.4177.
14. Replogle JM, Norman TM, Xu A, et al. Mapping information-rich genotype–phenotype landscapes with genome-scale Perturb-seq. Cell. 2022;185(14):2559–e257528. https://doi.org/10.1016/j.cell.2022.05.013.
15. Li W, Köster J, Xu H, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15:554. https://doi.org/10.1186/s13059-014-0554-4.
16. Li W, Xu H, Xiao T, et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome Biol. 2015;16:281. https://doi.org/10.1186/s13059-015-0843-6.
17. Kim E, Hart T. BAGEL2: A fast and accurate method for analysis of genome-wide CRISPR knockout screens. Genome Med. 2021;13(1):2. https://doi.org/10.1186/s13073-020-00811-0.
18. Aguirre AJ, Meyers RM, Weir BA, et al. Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. Cancer Discov. 2016;6(8):914–29. https://doi.org/10.1158/2159-8290.CD-16-0154.
19. Munoz DM, Cassiani PJ, Li L, et al. CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. Cancer Discov. 2016;6(8):900–13. https://doi.org/10.1158/2159-8290.CD-16-0230.
20. Meyers RM, Bryan JG, McFarland JM, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49(12):1779–84. https://doi.org/10.1038/ng.3984.
21. Dempster JM, Pacini C, Pantel S, et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol. 2021;22:343. https://doi.org/10.1186/s13059-021-02540-7.
22. Doench JG, Hartenian E, Graham DB, et al. Rational design of highly active SgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat Biotechnol. 2014;32(12):1262–7. https://doi.org/10.1038/nbt.3026.
23. Doench JG, Fusi N, Sullender M, et al. Optimized SgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34(2):184–91. https://doi.org/10.1038/nbt.3437.
24. Perez AR, Pritykin Y, Vidigal JA, et al. GuideScan software for improved single and paired CRISPR guide RNA design. Nat Biotechnol. 2017;35(4):347–9. https://doi.org/10.1038/nbt.3804.
25. Schmidt H, Zhang M, Chakarov D, et al. Genome-wide CRISPR guide RNA design and specificity analysis with GuideScan2. Genome Biol. 2025;26(1):41. https://doi.org/10.1186/s13059-025-03488-8. Published 2025 Feb 26.
26. Yu Y, Gawlitt S, de Andrade E, Sousa LB, et al. Improved prediction of bacterial CRISPRi guide efficiency from depletion screens through mixed-effect machine learning and data integration. Genome Biol. 2024;25(1):13. https://doi.org/10.1186/s13059-023-03153-y. Published 2024 Jan 11.
27. Guo J, Wang T, Guan C, et al. Improved SgRNA design in bacteria via genome-wide activity profiling. Nucleic Acids Res. 2018;46(14):7052–69. https://doi.org/10.1093/nar/gky572.
28. Kim HK, Kim Y, Lee S, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. Sci Adv. 2019;5(11):eaax9249. https://doi.org/10.1126/sciadv.aax9249. Published 2019 Nov 6.
29. Rauch BJ, Silvis MR, Hultquist JF, et al. Cas13d orthologs for robust RNA targeting with improved guide design principles. Cell Syst. 2023;14(2):153–e1677. https://doi.org/10.1016/j.cels.2023.01.003.
30. Allen F, Behan F, Khodak A, et al. JACKS: joint analysis of CRISPR/Cas9 knockout screens. Genome Res. 2019;29(3):464–71. https://doi.org/10.1101/gr.238923.118.
31. Barry T, Mason K, Roeder K, Katsevich E. Robust differential expression testing for single-cell CRISPR screens at low multiplicity of infection. Genome Biol. 2024;25(1):124. https://doi.org/10.1186/s13059-024-03254-2. Published 2024 May 17.
32. Duan B, Zhou C, Zhu C, et al. Model-based Understanding of single-cell CRISPR screening. Nat Commun. 2019;10(1):2233. https://doi.org/10.1038/s41467-019-10216-x. Published 2019 May 20.
33. Liu S, Hamilton MC, Cowart T, et al. Characterization and bioinformatic filtering of ambient gRNAs in single-cell CRISPR screens using CLEANSER. Cell Genom. 2025;5(2):100766. https://doi.org/10.1016/j.xgen.2025.100766.
34. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2.
35. Lotfollahi M, Wolf FA, Theis FJ. ScGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715–21. https://doi.org/10.1038/s41592-019-0494-8.
36. Lotfollahi M, Naghipourfar M, Theis FJ, Wolf FA. Compositional perturbation autoencoder for single-cell response modeling. Nat Methods. 2021;18(10):1200–6. https://doi.org/10.1038/s41592-021-01267-1.
37. Guan Z, Jiang Z. Transformer-based anti-noise models for CRISPR-Cas9 off-target activities prediction. Brief Bioinform. 2023;24(3):bbad127. https://doi.org/10.1093/bib/bbad127.
38. Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282–91. https://doi.org/10.1016/j.trechm.2019.02.007.
39. Bédard AC, Adamo A, Aroh KC, et al. Reconfigurable system for automated optimization of diverse chemical reactions. Science. 2018;361(6408):1220–5. https://doi.org/10.1126/science.aat0650.
40. MacLeod BP, Parlane FGL, Morrissey TD, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6(20):eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.
41. Nikolaev P, Hooper D, Webber F, et al. Autonomy in materials research: a case study in carbon nanotube growth. Npj Comput Mater. 2016;2:16031. https://doi.org/10.1038/npjcompumats.2016.31.
42. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained Language model for scientific text. arXiv. 2019. https://doi.org/10.48550/arXiv.1903.10676.
43. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical Language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.

44. Gu Y, Tinn R, Cheng H, et al. Domain-specific Language model pretraining for biomedical natural Language processing (PubMedBERT). arXiv. 2020. https://doi.org/10.48550/arXiv.2007.15779.
45. Sarumi OA, Heider D. Large Language models and their applications in bioinformatics. Comput Struct Biotechnol J. 2024;23:3498–505. https://doi.org/10.1016/j.csbj.2024.09.031. Published 2024 Oct 5.
46. Lin A, Ye J, Qi C, et al. Bridging artificial intelligence and biological sciences: a comprehensive review of large Language models in bioinformatics. Brief Bioinform. 2025;26(4):bbaf357. https://doi.org/10.1093/bib/bbaf357.
47. Shang X, Liao X, Ji Z, Hou W. Benchmarking large Language models for genomic knowledge with geneturing. Brief Bioinform. 2025;26(5):bbaf492. https://doi.org/10.1093/bib/bbaf492.
48. Hossain E, Shoeibi M, Garibay I, Yousefi N. Interpretable RNA-Seq clustering with an LLM-based agentic evidence-grounded framework. arXiv; 2025. abs/2510.16082.
49. Gao S, Fang A, Huang Y, et al. Empowering biomedical discovery with AI agents. Cell. 2024;187(22):6125–51. https://doi.org/10.1016/j.cell.2024.09.022.
50. Zhou J, Jiang J, Han Z, Wang Z, Gao X. Streamline automated biomedical discoveries with agentic bioinformatics. Brief Bioinform. 2025;26(5):bbaf505. https://doi.org/10.1093/bib/bbaf505.
51. Qu Y, Huang K, Yin M, et al. CRISPR-GPT for agentic automation of gene-editing experiments. Nat Biomed Eng. 2025;30. https://doi.org/10.1038/s41551-025-01463-z.
52. Ding K, Yu J, Huang J, Yang Y, Zhang Q, Chen H. SciToolAgent: a knowledge-graph-driven scientific agent for multitool integration. Nat Comput Sci. 2025;5(10):962–72. https://doi.org/10.1038/s43588-025-00849-y.
53. Liu H, Li Y, Wang H. GenoMAS: a multi-agent framework for scientific discovery via code-driven gene expression analysis. arXiv; 2025, abs/2507.21035.
54. Mehandru N, Hall AK, Melnichenko O, Dubinina Y, Tsirulnikov D, Bamman D, Alaa A, Saponas S, Malladi VS. BioAgents: democratizing bioinformatics analysis with multi-agent systems. arXiv; 2025, abs/2501.06314.
55. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, Cao Y. ReAct: synergizing reasoning and acting in Language models. arXiv. 2022. https://doi.org/10.48550/arXiv.2210.03629.
56. Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Zettlemoyer L, Cancedda N, Scialom T. Toolformer: Language models can teach themselves to use tools. arXiv. 2023. https://doi.org/10.48550/arXiv.2302.04761.
57. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364–74. https://doi.org/10.1038/s41591-020-1034-x.
58. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health. 2020;2(10):e549–60. https://doi.org/10.1016/S2589-7500(20)30219-3.
59. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.
60. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–203. https://doi.org/10.1038/s41592-021-01252-x.
61. Yu Y, Mai Y, Zheng Y, Shi L. Assessing and mitigating batch effects in large-scale omics studies. Genome Biol. 2024;25(1):254. https://doi.org/10.1186/s13059-024-03401-9. Published 2024 Oct 3.
62. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A survey on hallucination in large Language models: Principles, Taxonomy, Challenges, and open questions. ACM Trans Inform Syst. 2023;43:1–55.
63. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kuttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv; 2020, abs/2005.11401.
64. Liu X, Yu H, Zhang H, Xu Y, Lei X, Lai H, Gu Y, Gu Y, Ding H, Men K, Yang K, Zhang S, Deng X, Zeng A, Du Z, Zhang C, Shen S, Zhang T, Shen S, Su Y, Sun H, Huang M, Dong Y, Tang J. AgentBench: evaluating LLMs as agents. arXiv; 2023, abs/2308.03688.
65. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18. Published 2016 Mar 15.
66. Florian Häse LoïcM, Roch. Alán Aspuru-Guzik, next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282–291. https://doi.org/10.1016/j.trechm.2019.02.007. ISSN 2589–5974.
67. Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. Nature. 2020;583(7815):237–41. https://doi.org/10.1038/s41586-020-2442-2.
68. Lee DH, Kim H, Sung BH, et al. Biofoundries: bridging automation and biomanufacturing in synthetic biology. Biotechnol Bioproc E. 2023;28:892–904. https://doi.org/10.1007/s12257-023-0226-x.
69. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST Trustworthy and Responsible AI. Gaithersburg, MD: National Institute of Standards and Technology, Gaithersburg, MD; 2023. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225. Accessed Jan 25, 2026.
70. World Health Organization. Ethics and governance of artificial intelligence for health. (2021).
71. European Union. Regulation (EU) 2024/1689 (Artificial intelligence Act). Official J Eur Union (2024).

Publisher's note