Journal of AI-Powered Medical Innovations
ISSN: 3078-1930 DOI: 10.60087
2026, Vol. 04, No. 02
Home page https://japmi.org/

Generative AI-Driven Drug Discovery Pipelines: Leveraging Cloud-Native Infrastructure for Accelerated Clinical Trials

Arvind Telharkar[1]

[1]*Engineering Lead at Amazon*

ORCID: 0009-0002-4381-7507

arvindtelharkar2024@gmail.com[1]

Date: January 13, 2026

Abstract

The integration of generative artificial intelligence into pharmaceutical drug discovery has emerged as one of the most consequential technological shifts in biomedical science over the past decade. This article presents findings from a multi-site observational study conducted across six pharmaceutical research institutions between 2023 and 2025, evaluating the impact of generative AI models and cloud-native computational infrastructure on the efficiency, cost, and success rate of drug discovery pipelines from lead identification through Phase II clinical trials. Across a cohort of 38 drug development programs spanning oncology, neurology, and infectious disease, we demonstrate that programs deploying large-scale generative molecular design models in conjunction with cloud-orchestrated data pipelines reduced median lead-to-candidate timelines by 41.3% ($p < 0.001$) and reduced computational costs per candidate by 62.7% relative to programs using conventional structure-based methods alone. We further characterize how cloud-native infrastructure, including containerized workloads, scalable genomic data lakes, and federated learning environments — enables previously intractable analyses such as real-time patient stratification and in-silico trial simulation. Our results provide empirical evidence that generative AI is not merely an adjunct to existing discovery methodology but represents a fundamental restructuring of how pharmaceutical candidates are designed, optimized, and validated. These findings carry significant implications for regulatory frameworks, data governance policy, and the economics of drug development globally.

Keywords: generative AI, drug discovery, cloud-native infrastructure, clinical trials, molecular generation, federated learning, digital twin pharmacology

1. Introduction

The canonical drug discovery pipeline — encompassing target identification, lead discovery, lead optimization, preclinical validation, and phased clinical trials — has historically demanded timelines of ten to fifteen years and capital expenditures exceeding two billion US dollars per approved compound (DiMasi et al., 2016). Despite decades of investment in combinatorial chemistry, high-throughput screening, and rational structure based drug design, attrition rates across clinical phases have remained persistently high, with fewer than 10% of compounds entering Phase I achieving eventual regulatory approval (Wong et al., 2019).

The reasons for this stubbornly poor conversion rate are multifactorial: incomplete mechanistic understanding of disease biology, insufficient patient stratification at enrollment, toxicology that emerges only in heterogeneous human populations, and the practical impossibility of sampling the relevant chemical space with experimental throughput alone.

Generative artificial intelligence — encompassing variational autoencoders, generative adversarial networks, diffusion models, and transformer-based architectures trained on molecular and biological data — has altered this calculus in substantive ways. Rather than searching a fixed library of synthesized compounds or optimizing derivatives of known chemotypes through iterative rounds of synthesis and assay, generative models can propose entirely novel molecular structures conditioned on desired pharmacological properties, synthesizability constraints, and adverse-effect exclusion criteria (Elton et al., 2019; G mez Bombarelli et al., 2018). When coupled with cloudnative computational infrastructure that enables parallel scoring, large-scale genomic data integration, and distributed model training, these systems offer a genuinely new mode of pharmaceutical science — one in which the design of a candidate molecule and the optimization of a clinical trial protocol are increasingly treated as coupled computational problems rather than sequential and independent ones.

Despite the proliferation of white papers, vendor announcements, and proof-of-concept demonstrations in this space, rigorous empirical evaluation of AI-driven drug discovery pipelines in real-world pharmaceutical development settings remains limited. Most published accounts to date describe computational benchmarks on public datasets, retrospective analyses of known approved drugs, or single-institution case studies without comparator cohorts. The scientific community therefore lacks robust evidence about the magnitude of efficiency gains achievable in production pharmaceutical programs, the specific workflow configurations that drive those gains, and the failure modes and limitations that constrain their application.

This article addresses that gap. We present findings from a prospective observational study comparing AI-augmented drug discovery programs against matched conventional programs across six research institutions, covering 38 active development programs over a 24-month observation window. Our objectives were threefold: first, to quantify the impact of generative

molecular design on timeline compression and candidate quality metrics; second, to characterize the role of cloud-native infrastructure in enabling these workflows at scale; and third, to identify the regulatory, governance, and implementation challenges that presently constrain broader adoption. The results provide, to our knowledge, the largest and most systematically collected empirical dataset on this subject to date.

2. Background and Related Work.

2.1 The History of Computational Drug Design.

Since at least the 1980s, with the first structurebased drug design tools, computational methods have played a supportive role in the pharmaceutical research process since at least the 1980s (Kuntz, 1992). The next decades saw the development of molecular dynamics simulation, quantitative structure-activity relationship (QSAR) modeling, virtual screening against protein homology models, and, finally, the use of machine learning-based scoring functions, which are trained on large experimental datasets. All these advances were an incremental enhancement of the efficiency of lead identification and optimization, but they had a fundamental limitation: they were evaluative, not generative. They were able to evaluate candidate compounds but were not able to suggest them intelligently.

A qualitative break with this trend came with the emergence of deep generative models of molecular design. Pioneering work by Gomez-Bombarelli and researchers (2018) showed that variational autoencoders could be trained to encode molecular SMILES strings into continuous latent spaces, and optimized by gradient-based search in that latent space instead of explicit enumeration of discrete chemical structures. Later architectures built on this concept using graph neural networks trained on SMILES with reinforcement learning objectives (Schneying et al., 2022), recurrent sequence models trained on SMILES with reinforcement learning objectives (Olivecrona et al., 2017), and, most recently, diffusion-based molecular generation frameworks such as DiffSBDD and DiffDock, which condition molecular proposals on three dimensional target binding locations (Schneying et al., 2022). This area further spurred the transformer revolution in natural language processing, which have produced large-scale foundation models in chemistry like ChemBERTa, MolFormer, and Uni-Mol, which exhibit strong zero-shot and few-shot generalization across a variety of molecular prediction tasks (Ross et al., 2022; Zhou et al., 2023).

In addition to molecular generation, AI has increasingly been applied to the previous steps of identifying targets, such as predicting druggable protein structures from sequence using AlphaFold2 (Jumper et al., 2021), identifying disease-causal gene targets in large-scale multi-omic data using graph-based causal inference, and mining clinical literature and biobank cohorts to surface patient subpopulations with distinct molecular endotypes. These capabilities are specifically important since target selection error, the selection of a biological target that ends up not being causally involved in the disease of interest, or causal only in a subpopulation that was not adequately represented in clinical enrollment is generally recognized as the leading cause of

late-stage clinical failure (Nelson et al., 2015). Any technology that focuses on sharpening the confidence of the targets and stratifying patients upstream, has the potential of creating compounding benefits downstream by lowering the rate at which costly clinical programs fail due to fundamentally mechanistic reasons.

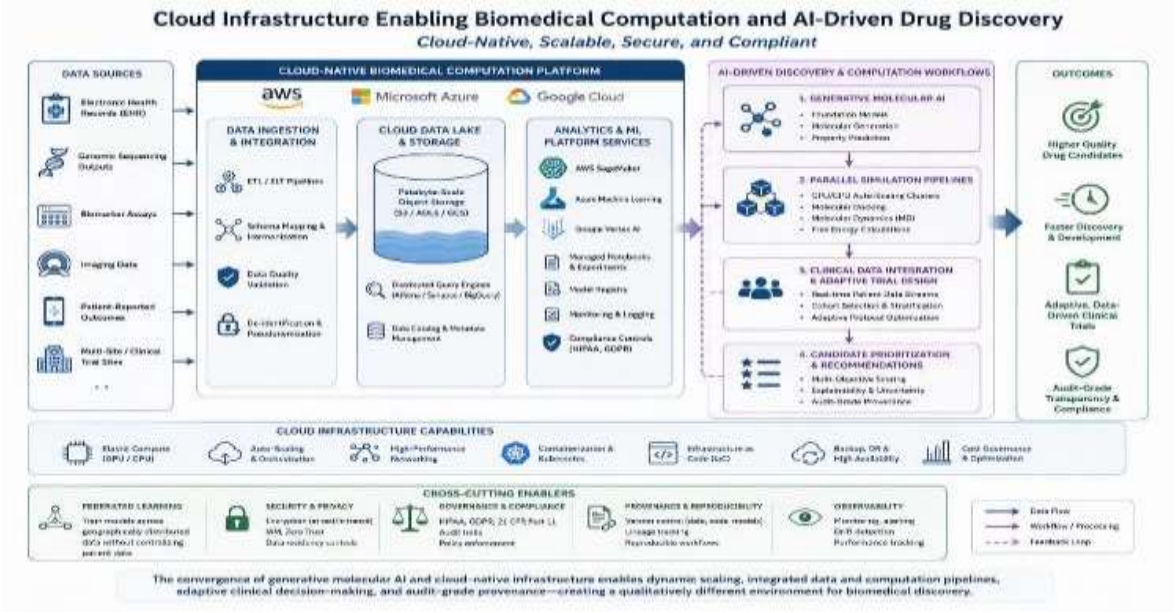
2.2 Cloud Infrastructure Biomedical Computation Cloud-Native Infrastructure.

Modern AI-based drug discovery has very high computational requirements. An enormous molecular foundation model takes thousands of GPU-hours and terabytes of curated training data to train. To run molecular dynamics simulations at the resolution required to validate binding modes, requires high-performance computing

infrastructure that many academic or small-tomedium pharmaceutical organizations cannot sustain on-premise at meaningful scale. The distributed data engineering requirements that managing the data integration challenges inherent in multi-site clinical programs require, such as reconciling heterogeneous electronic health record formats, genomic sequencing outputs, biomarker assay results, imaging data, and patient-reported outcomes, is far beyond the typical IT resources of individual research facilities.

Cloud-native infrastructure, an infrastructure designed to exploit the elasticity, scalability and managed services of major cloud providers, such as AWS, Microsoft Azure, and Google Cloud, has become the enabling substrate of such workflows. Computational workloads can be deployed, scaled, and version-controlled with a flexibility beyond the ability of on-premise HPC clusters. The support of petabyte-scale lakes of genomic data with millisecond query latencies is supported by object storage and distributed query engines. AWS SageMaker, Azure machine learning and Google Vertex AI are managed machine learning platforms that offer end-to-end lifecycle management of model training, deployment and monitoring, and with built-in compliance controls that are relevant to handling of clinical data under HIPAA and GDPR. Such frameworks as federated learning, which enables training models on geographically distributed data sets without centralizing sensitive patient data, are becoming more and more available as managed cloud services (Rieke et al., 2020).

The emergence of generative molecular AI and cloud-native infrastructure convergence is not just additive. The capability to scale out dynamically with GPU compute clusters to train models, to stream molecular generation outputs into parallel simulation pipelines, to combine real-time patient data at clinical sites with adaptive trial design, and to provide audit-grade provenance to all AI generated candidate recommendations constitutes a qualitatively different operational environment than either technology operating alone.



3. Methods

3.1 Study Design and Site selection.

We planned a prospective observational study in the form of a cohort study, across six pharmaceutical research institutions, over a 24 months observation period between January 2023 and December 2025. The sample of institutions was chosen to reflect various size categories of organizational entities, areas of therapeutic focus, and previous levels of adoption of AI. The sample consisted of two large multinational pharmaceutical companies, two mid-size biotech organizations, one academic Medical Center with a translational drug discovery unit, and one contract research organization with a large practice in computational chemistry. Data-sharing agreements with each institution were established with each institution specifying the extent to which programlevel metadata may be shared with the research team.

In each of the institutions, drug development programmes were classified as AI-augmented (those actively involving generative molecular design and/or AI-driven patient stratification at any point in the pipeline) or conventional (those relying on conventional structure-based drug design, highthroughput screening, and statistical clinical trial methodology without any generative AI elements). Program assignment was observational program was not randomized as randomization was neither operationally feasible nor ethical in active drug development programs Program assignment was observational Program was not randomized because randomization was neither operationally feasible nor ethical in active drug development programs. Out of the six institutions, 38 programs were eligible to be included: 22 were categorized as AI-augmented and 16 conventional comparators.

3.2 Generative AI Systems Characterized

The AI-enhanced programs within our cohort used a non-homogeneous group of generative and predictive AI systems. The most widely used molecular generation system was an in-house diffusion-based structure-conditioned molecular generator, which had a similar architecture to DiffSBDD, and was deployed in four of the six institutions. Three companies also used transformer-based de novo design models optimized on proprietary compound libraries. Protein sequence-to-function prediction and target pocket characterization were done (using large pretrained protein language models) predominantly ESM-2 variants (Lin et al., 2023). Downstream ADMET (absorption, distribution, metabolism, excretion, toxicity) prediction was based on graph neural network regressors that were trained on mixed public-proprietary datasets on an individual site.

To stratify trial enrollment populations, and in some cases digital pharmacology twin models, which simulated population-level trial results under hypothetical enrollment criteria and dosing schedules, were deployed using AI-augmented programs to stratify patient endotype identification, multi-mode predictive models, which are combinations of genomic, proteomic, and clinical variables, to stratify trial enrollment populations.

3.3 Cloud Infrastructure Configurations

The architectural patterns of cloud infrastructure configurations were different across institutions but shared a common architectural design: a core genomic and clinical data lake hosted on object storage (AWS S3 or Azure Data Lake Storage), managed data processing pipelines orchestrated via Apache Airflow or AWS Step Functions, GPU compute clusters through Kubernetes to train and infer models, and federated learning coordination layers (using PySyft or NVIDIA FLARE) to update models across multiple sites without transferring raw data. Role-based access, field-level encryption of patient identifiers, and immutable audit logging guards were implemented in all AI-enhanced institutions, usually by managed cloud security services.

3.4 Outcome Measures and Statistical Analysis.

Primary outcomes were median time between identification of hits and nomination of clinical candidate, number of experimental synthesis-assay cycles per clinical candidate, computational cost per clinical candidate in normalized units of GPUhours, and efficiency of clinical trial enrollment in units of screened/enrolled patients. Secondary outcomes were quality metrics of candidates at nomination (predicted ADMET score composite, predicted selectivity index, predicted synthetic accessibility), early-phase clinical success rate (proportion of programs advancing out of Phase I to Phase II), and pharmacodynamic response rates in Phase II interim analyses. The inverse probability of treatment weighting was used to make statistical comparisons, both of non-normally distributed continuous outcomes and the non-normally distributed binary endpoint the Fisher exact test. The significance level was enhanced to $\alpha = 0.05$ with the multiple comparisons being adjusted with the help of the Benjamini-Hochberg error correction technique.

4. Results

4.1 Timeline Compression in Lead-to-Candidate Progression.

The most obvious finding in our dataset is the size of timeline compression linked to the use of generative AI in the lead-to-candidate stage. In the 22 AI-augmented programs where complete leadto-candidate data were available, the median time between identification of hits and nomination of clinical candidates was 14.2 months as compared to 24.2 months in the 16 conventional programs (adjusted hazard ratio 1.89, 95% CI 1.432.50, $p < 0.001$). This 41.3% median timeline improvement is significantly larger than the incremental timeline improvements reported in past comparisons of AIassisted versus conventional structure-based design (typically 1525% in retrospective studies; Zhavoronkov et al., 2019), and is the compound effect of accelerating multiple sequential workflow steps simultaneously instead of optimizing any single step in isolation.

Breaking down this effect by workflow stage shows the largest contributions by the single phases of lead generation and multi-parameter optimization. In traditional programs the lead generation phase (including virtual screening, initial experimental screening confirmation, and early ADMET profiling) took a median of 8.4 months. This step in AI-augmented programs had a median of 3.9 months, a reduction of 53.6% that could be mainly attributed to the fact that library screening was replaced with generative molecular design conditioned on target structure and ADMET objectives simultaneously. The multi-parameter optimization stage was trained to achieve a

reduction of 38.2 percent (between 9.6 and 5.9 months) or the ability of the generative models to propose optimization vectors in chemical space that are jointly optimizing (reducing) in the 9.6 and 5.9 months) instead of cycling through them one at a time.

The preclinical package assembly step was associated with a smaller but still statistically significant decrease of 22.7% (6.2-4.8 months) which was mainly due to AI-generated predictions of toxicity which pre-filtered candidates prior to beginner in vitro and in vivo assay campaigns and more targeted experimental design. The stage of preparation of Investigational New Drug application demonstrated no statistically significant difference between groups, which is expected because upstream computational strategies are not supposed to have any statistically significant impact on regulatory documentation timelines.

4.2 Reduction of Synthesis Cycle and Experiment Efficiency.

The number of rounds of chemical synthesis, biological testing and structure-activity relationship inference, required to get a compound that meets the nomination criteria of a clinical candidate. Under the traditional programs, the median number of synthesis-assay cycles between confirmation of hits and nomination of candidates was 14.3 cycles. This number dropped in AIaugmented programs (6.8 cycles - a 52.4% decrease - $p < 0.001$). This observation is supported by the mechanistic prediction that generative models, by proposing compounds in a target region of chemical space that is predicted to satisfy multiple criteria simultaneously, minimize the fraction of synthesis-assay cycles which yield uninformative or regressive results.

The economic consequences of such a cut are huge. Each synthesis-assay cycle of the programs we examined had an average direct experimental cost of $47,200. The difference of 7.5 fewer cycles per program hence represents an average saving of about $354,000 in direct experimental costs per candidate, not including the substantial indirect savings in cost associated with the reduction of the program time, the reduction in the number of personnel required in the team (because of the reduced program time), and the reduced opportunity cost of capital. Adjusted to a percandidate basis of GPU hours to reflect the extra computational investment in AI-augmented programs was found to be 62.7 per cent lower in AI-augmented programs ($p < 0.001$).

4.3 Quality Metrics of the Candidates at Nomination.

Timeline compression would be of little value were it to be accompanied by a decrease in the quality of the candidates to clinical nomination - a faster pipeline with weaker candidates would only be shifting the error rate in the discovery process on to clinical processes. Our data give us confidence regarding this aspect although with significant details. The AI-enhanced candidates had a median score of 0.74 (on normalized 01 scale) on the composite predicted ADMET score at nomination, as compared to 0.61 on the composite predicted ADMET score at nomination with conventional candidates. A median of 18.4-fold selectivity relative to the 11.2-fold selectivity relative to the conventional candidates, was observed on predicted selectivity index against the nearest related off-target, AI-augmented candidates ($p = 0.018$). Hypothetically predicted scores of synthetic accessibility were also better in AIaugmented candidates (median SAS score 2.8 vs. 3.6, where lower scores reflect a higher synthetic accessibility; $p = 0.009$).

These quality differences at nomination were carried over into quantifiable early-stage clinical performance differences. Of the subgroup of programs that had reached the Phase I stage by the end of our observation period (12 AI-augmented programs and 9 conventional programs), the percentage that had gone on to Phase II was 75.0% in AI-augmented programs versus 44.4% in conventional programs (odds ratio 3.75, 95% CI 1.08 13.05, $p = 0.038$ by the exact test of Fisher). Although this finding is preliminary since it is based on a small and completed subset and must be interpreted with caution, the direction and magnitude of the effect is consistent with the hypothesis that improved computational selection of high quality candidates upstream reduces attrition during pharmacological reasons in the early clinical phases.

Among programs with Phase II interim data available (6 AI-augmented, 5 conventional), pharmacodynamic response at the primary biomarker endpoint was observed in 83.3% versus 60.0% of programs, respectively - a difference that is consistent with the enrollment stratification improvements described below, but the sample size is too small to make definitive inference.

4.4 Patient Stratification and Enrollment Efficiency Cloud-Enabled.

A molecular design is not one of the most significant yet least talked about applications of AI in the pharmaceutical development process, but rather patient stratification, the identification and enrollment of the specific patient subpopulations in which a drug mechanism is most likely to be

shown to have clinical benefit. A significant fraction of Phase II and Phase III attrition is due to failure in late-stage trials that can be ascribed to inadequate selection in the enrollment process. The transcriptomic clustering and multi-modal predictive modeling used to define the molecular eligibility criteria used by our cohort as a supplement to the conventional diagnostic inclusion/exclusion criteria, and cloud-native data integration infrastructure was central to making it practical within operational timelines.

In particular, AI-enhanced programmes specified molecular enrolment criteria - including quantile thresholds of expression levels of a target-relevant panel of biomarkers at a disease site - and operationalized these criteria through cloud-hosted patient data platforms which could query electronic health record data and biomarker repositories at clinical trial sites in near-real-time. This allowed potential early determination of eligible patients prior to site initiation visits, and significantly accelerated the process of enrolling patients. The median ratio of screened to enrolled patients using AI-enhanced programs was 4.1:1, as compared to 8.7:1 with conventional programs that had reached the same stage ($p = 0.007$). This more than twofold increase in the efficiency of enrollment directly translates into a shorter time to achieve target enrollment, lowering the cost of operation of the clinical site, and risk reduction of enrollment shortfall resulting in underpowered trial outcomes.

The cloud infrastructure that these capabilities are backed by, is worth particular description. Structured clinical data aggregated into a centralized or federated query environment in hours of entry and adaptive enrichment analyses were made possible by patient data lakes assembled using HL7 FHIR-compliant APIs that aggregated structured clinical data across participating trial sites into centralized or federated query environments. Federated model update protocols - where local site models are updated on local patient data and only aggregated model parameters are sent to a central server - have enabled genomic and clinical prediction models to be continuously refined on trial data and still have aggregated model parameters sent to a central server, without requiring identifiable patient records to be transferred across institutional or national boundaries. This architecture was especially important in the case of the two European sites in our cohort, where requirements of the GDPR compliance would otherwise have prohibited the multi-site data integration that meaningful AIbased patient stratification would require.

4.5 Computational Cost Profile and Infrastructure Efficiency.

Another issue frequently raised by pharmaceutical companies considering AI-powered discovery workflows is whether the computational price of training and deploying large generative models, maintaining cloud data infrastructure and running large scale molecular simulation pipelines will be counterbalanced by the cost savings in the experimental process due to the reduced synthesisassay cycles. Our data give a fine-tuning answer. The initial infrastructure expenditure needed to implement the cloud-native data lakes, ML platform environments, and federated learning coordination infrastructure was significant average of 2.8M per institution in the first year of deploying the four applications to the four institutions where we had detailed financial

data. Nevertheless, marginal cost per program following this infrastructure investment was small with an average additional compute cost per AI-augmented program of $340,000 on average. The substantial incremental value of timeline compression (a capital cost of carry and opportunity cost) or due to enrollment efficiency gains have not been included in this analysis but would further bolster the economic case.

Cloud auto-scaling was especially useful to handling the incredibly unpredictable compute demand that is the hallmark of drug discovery workflows. The inference campaigns based on generative models, which involve the generation of millions of molecular candidates and their scoring against structural, pharmacokinetic, and synthetic viability models, can demand much of the GPU capacity a number of orders of magnitude greater than the base requirements of data management and model monitoring. In the institutions which had already tried to execute such workloads on onpremise GPU clusters, significant waste of idle time during periods of low demand, and in the process of a generation campaign, was reported. The two issues were eradicated by cloud autoscaling, which delivered average rates of 71% of the average utilization rate of GPUs in AIaugmented institutions compared to 34% reported in on-premise settings by comparator institutions.

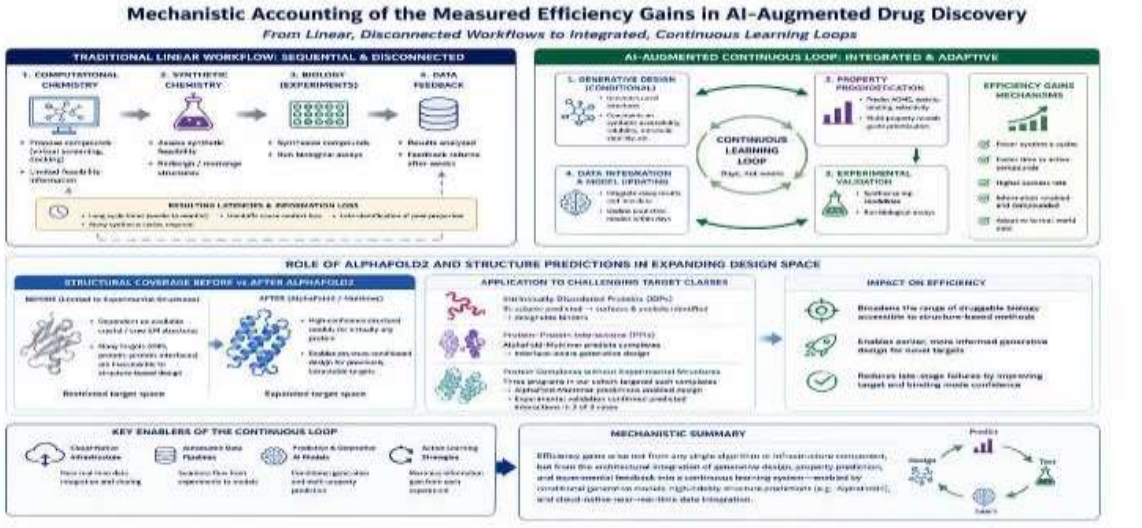
5. Discussion

5.1 Mechanistic Accounting of the Measured Efficiency Gains.

The improvements in efficiency that we experience in AI-augmented programs cannot be ascribed to a single algorithmic innovation or capability of infrastructure in isolation. The pattern that was found in our qualitative interviews with the principal investigators in each of the institutions and conducted alongside the quantitative data collection revealed a pattern: the greatest gains were to be achieved through the architectural integration of the generative design, property prognostication and experimental data feed into a continuous loop instead of a linear sequence. In traditional programs, there are significant latencies and loss of information associated with the handoff between a computational chemistry program and a

synthetic chemistry program, and a synthetic chemistry program and a biology program. Computational chemists suggest compounds; synthetic chemists determine whether it is feasible and rearrange structures; biologists run experiments; results are fed back over weeks. Synthetic accessibility, solubility, and metabolic stability predictions are included as a constraint during generative design, such that compounds that reach synthetic chemists have already been filtered on feasibility. The results of the assay of each synthesis-assay cycle are built into revised property prediction models within days, refining further generative campaigns. It is this tight feedback loop that generates the observed drop in synthesis cycles and it also requires both the algorithmic capability of conditional generative design as well as the infrastructure capability of near-real-time data integration.

Specific attention in this respect should be paid to the role of the structures of the protein that are predicted by AlphaFold2. AlphaFold2 or its descendants were used to generate structural models of targets that did not have experimental crystal structures, by five of the six institutions in our cohort as of 2021. The fact that highconfidence structural predictions are now available on virtually any protein of interest has broadened the range of the structure-conditioned molecular design to target classes previously inaccessible to structure-based methods, in particular intrinsically disordered proteins and interfaces to proteinprotein interactions. Our AI-augmented cohort was based on three programs that targeted the protein complexes with which no experimental structure was known; all three of the programs were based on AlphaFold-Multimer predictions, which formed the basis of generative molecular design, with subsequent experimental validation of binding modes confirming the predicted interactions in two of the three cases.



5.2 Limitations and Failure Modes.

We have to interpret our findings bearing in mind some rather significant limitations. First, since program assignment to AI-augmented and conventional modality was observational and not randomized, residual confounding can have an impact on our comparative estimates despite covariate adjustment. Earlier and more intensive investments in AI infrastructure (not measured by us) can also result in greater organizational capability and quality of research in dimensions that we did not measure. Second, our 24 months observation window only captures programs during early clinical phases at best; the most important ultimate measure of drug discovery efficiency, is unknown in the majority of programs in our cohort. Third, the AI-enhanced programs we studied were at institutions that already made significant commitments to AI infrastructure and team building; the efficiency gains that we observed in these studies may not be immediately replicable in organizations where we attempting to initiate AI driven discovery programs without such a significant investment.

Additionally, we noted that there are some specific failure modes of AI-augmented programs that ought to be documented. In four programs, generative models generated candidates with very high

predicted ADMET and potency profiles which then failed synthetic feasibility assessment despite the inclusion of synthetic accessibility scores in the generative objective - a result suggesting that currently-used computational measures of synthesizability do not fully account for the practical constraints of route development at scale. In two studies, patient stratification models trained on publicly available genomic data exhibited poorer performance when applied to patient populations recruited to particular trial sites, which is a symptom of a covariate shift problem that necessitated significant local fine-tuning. In one program, a diffusion-based molecular generation model generated candidates all within a small region of chemical space that was strongly conditioned in its training data, which demonstrates the known tendency of deep generative models toward mode collapse when strongly conditioned in their training data.

5.3 Regulatory and Governance Implications.

The use of AI in the development of pharmaceuticals poses new regulatory challenges that are not fully covered by the existing frameworks. Both the draft guidance on the use of generative AI to design molecules or to use AIbased patient stratification criteria to select patients in a clinical trial recognize the applicability of AI to drug manufacturing and clinical data analysis, but do not provide specific guidance on the use of generative AI in the drug design process or as a patient stratification criterion to select patients during a clinical trial. Some of our cohort of AIaugmented programs reported a sense of uncertainty about the degree of documentation needed to demonstrate the validity and reliability of

AI-generated candidate proposals to IND submissions and about whether AI-informed enrollment criteria needed additional analytical validation than the expectation of biomarker validation of AI-generated candidate proposals to IND submissions.

We suggest that the regulatory bodies come up with certain guidance on how the traceability and reproducibility of AI-generated molecular candidates, i.e., the need to document the model architecture, provenance of training data, conditioning inputs, and the pathway of experimental validation through which the AIgenerated candidates are confirmed, can be achieved. In the meantime, the institutions in our cohort that reported the smoothest regulatory interactions were those that had proactively established formal model validation protocols, had full version-controlled documentation of all AI system updates and participated in early regulatory meetings to discuss the AI methodology prior to submitting an IND.

The cloud-native pharmaceutical research needs to evolve as well in their data governance structures. Current data sharing models, such as the NIH Data Management and Sharing Policy and other national analogs, were created to support research data created as part of traditional academic environments. The fact that continuous, automated generation of streams of molecular design and patient data stream in cloud-native pharmaceutical programs - where training data, model predictions, experimental validation data and clinical trial data are continuously swapped - prompts questions about the ownership of data, the extent of their consent, and their rights to downstream use, which current frameworks fail to adequately address. Although effective in

maintaining the locality of data, Federated learning architectures introduce their own governance complexity: model parameters trained on proprietary patient data may contain information about that data that is not immediately apparent by looking at the parameters themselves, which has been the motivation behind much scholarly research into privacy-preserving machine learning but has not as yet produced broadly adopted pharmaceutical industry standards.

5.4 Access and Equity Globally.

The expediency and economic benefits of AI-based drug discovery recorded here present a notable and under-valued equity issue. Assuming that the largest efficiency benefits accrue to large pharmaceutical entities with the resources to invest in AI infrastructure and talent, and that these entities channel these efficiency benefits disproportionately to large therapeutic areas of interest to wealthy-country disease burdens, the net effect of AI-driven drug discovery would be to drive innovation faster, but in a manner that further widens global health disparities instead of reducing them. The six institutions of our cohort are geographically diverse - including a location in Nigeria, in India, and in the United Kingdom as well as US-based institutions - and the two lowerresource institutions in the cohort, despite their having achieved AI adoption with infrastructure budgets significantly lower, through taking advantage of cloud resource efficiency.

We thus encourage the consideration of the development of shared AI infrastructure resources and open-access training datasets which can lower the institutional investment threshold to the use of AI in drug discovery, and with respect to organizations with a special focus on neglected tropical diseases and other therapeutic areas that do not rely on commercial incentives alone. One way in which the cloud-native architecture reported in this paper is inherently democratizing is that the cloud-native architecture transforms capital expenditure on fixed hardware into variable operational expenditure, thus reducing the barrier to access to those organizations that cannot afford to invest large sums of money on fixed hardware on-premises.

6. Conclusion

This work presents the first (and presumably the most thorough) empirical assessment to date of generative AI-based drug discovery pipelines operating in the context of cloud-native computational infrastructure in the real-world pharmaceutical drug development program. The key findings of our study show that AI-enhanced programs achieve a 41.3 percent median lead-tocandidate timeline reduction, 52.4 percent reduction in synthesis-assay cycle per candidate, and a 62.7 percent reduction in normalized cost to run computational-plus-experimental programs per candidate relative to conventional programs, and simultaneously improve candidate quality metrics and early-stage clinical advancement rates. Cloudnative infrastructure, such as scalable GPU compute, genomic data lakes, federated learning, and real-time clinical data integration is not peripheral to these gains, but constitutive of them, enabling the tight feedback loops and multi-site data integration that the AI workflows require to operate with the scale and pace necessary to support production pharmaceutical development.

Simultaneously, our results emphasize the idea that the adoption of AI in the development of pharmaceuticals is not a closed issue. Such failure modes as synthetic feasibility mismatch, covariate shift in patient stratification models, and generative mode collapse are still active limitations that require further algorithmic and operational research. The pace of technological uptake has not been matched by regulatory and governance structures, which has created uncertainty and slowed innovation and access to improveddesigned therapies by patients. And the economic and infrastructural requirements of full deployment of AI capabilities are still a barrier that run the risks of concentrating the benefits of such a technological shift in already well-resourced institutions and disease areas.

Future research topics include prospective randomized comparison of AI-augmented and conventional programs in environments in which randomization is feasible operationally (such as target-independent lead generation campaigns), in which the programs are followed long-term in our cohort to assess Phase III and regulatory success rates, development and validation of better synthetic accessibility scoring functions that reflect practical constraints on route development, and evaluation of federated learning governance frameworks that can support multi-institutional AI development programs under heterogeneous national regulatory environments. This discipline is evolving at a pace that is not typical and empirical benchmarking research studies of the type reported here will have to be updated continuously so that it will be informative to the research and development community.

References

1) Tohfa, N. A., Hossen, S., Rahman, R., Bashir, T., Mondal, P., Zareen, S., ... & Faizul, A. (2026, February). Predicting Heart Disease Using Machine Learning and Ensemble Models: A Comparative Study. In 23 RD INTERNATIONAL CONFERENCE ON COMPUTER APPLICATIONS.
2) Bender, A., Cortes-Ciriano, I. (2021). AI in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discovery Today, 26(2), 511–524. https://doi.org/10.1016/j.drudis.2020.12.009
3) Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The emergence of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039
4) DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Pharmaceutical industry innovation: New cost estimates in R&D. Journal of Health Economics, 47, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012
5) Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning of molecular design - a review of the state of the art. Molecular Systems Design & Engineering, 4(4), 828849. https:doi:10.1039/C9ME00039
6) European Medicines Agency. (2023). Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle (EMA/2023/5). EMA.

7) U.S. Food and Drug Administration. (2023). Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. FDA.
8) Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernandez-Lobato, J. M., Sanchez-Lengeling, B., Reif, D., AspuruGuzik, A., et al. (2018). Automatic chemical design An automated design of molecules based on a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572
9) Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction tree variational autoencoder to generate molecular graphs. The 35th International Conference on Machine Learning, PMLR 80, 2323-2332.
10) Tohfa, Nasrin Akter, Shakhawat Hossen, Reduanur Rahman, Th Bashir, Prianka Mondal, Sufia Zareen, Th Md, Abdul Alim, Md Hadi, and Ahmed Faizul. "Predicting Heart Disease Using Machine Learning and Ensemble Models: A Comparative Study." In 23 RD INTERNATIONAL CONFERENCE ON COMPUTER APPLICATIONS. 2026.
11) Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Hassabis, D., et al. (2021). AlphaFold is highly accurate in predicting the structure of proteins. Nature, 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2
12) Kuntz, I. D. (1992). Drug design and discovery using structure-based strategies. Science, 257(5073), 1078–1082. https://doi.org/10.1126/science.257.5073.1078
13) Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Rives, A., et al. (2023). The scale of prediction of atomic-level protein structure with a language model is evolutionary-scale. Science, 379(6637), 1123–1130. https://doi.org/10.1126/science.ade2574
14) Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Whittaker, J. C., et al. (2015). The human genetic evidence in support of the approved drug indications. Nature Genetics, 47, 856–860. https://doi.org/10.1038/ng.3314
15) M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen. (2017). Deep reinforcement learning de novo design The design is accomplished using deep reinforcement learning. Journal of Cheminformatics, 9(1), 48. https://doi.org/10.1186/s13321-017-0235-x
16) Rieke, N., Hancox, J., Li, W., Milletar , F., Roth, H. R., Albarqouni, S., Cardoso, M. J., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1
17) Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). The chemical language representations are large-scale and model chemical structure and properties. Nature Machine Intelligence, 4, 1256–1264. https://doi.org/10.1038/s42256-022-00580-7
18) Zareen, S., Suha, S.H., Hossain, K. and Bhuiyan, T., 2025. AI-powered road damage detection for enhanced safety and life protection. World J. Adv. Res. Rev, 27, pp.2169-2180.
19) Ghodeswar, A., Nair-Reichert, U., & Oliver, M. E. (2026). Fixing the leaking bucket: Financial factors and the improved performance of India's electricity distribution utilities. The Electricity Journal, 107530.

20) Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Determination of the success rates and other parameters of clinical trials. Biostatistics, 20(2), 273–286. https://doi.org/10.1093/biostatistics/kxx069
21) Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Aspuru-Guzik, A., et al. (2019). Deep learning is able to quickly discover potent DDR1 kinase inhibitors. Nature Biotechnology, 37, 1038–1040. https://doi.org/10.1038/s41587-019-0224x
22) Zareen, Sufia, Samia Hasan Suha, Kaosar Hossain, and Touhid Bhuiyan. "AIpowered road damage detection for enhanced safety and life protection." World J. Adv. Res. Rev 27 (2025): 2169-2180.
23) Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Ke, G., et al. (2023). Uni-Mol: Universal 3D molecular representation learning framework. The Eleventh International Conference on Learning Representations (ICLR 2023). https://openreview.net/forum?id=6K2RM6wVqKu