ENPICOM

AI in biologics discovery and engineering: A practical guide to driving adoption

White paper

Table of contents

Executive summary

AI and machine learning (ML) are revolutionizing the development of biotherapeutics. However, despite its potential, AI adoption in this field faces significant challenges. The ENPICOM Platform provides a solution to integrate AI into biologics research, bridging the gap between lab science and Machine Learning Operations (MLOps). This whitepaper explores the roadblocks to AI-driven biologics innovation and presents ENPICOM's blueprint for overcoming them through scalable, automated, and efficient MLOps workflows.

The promise of AI in biologics discovery

Bringing a new drug to market is a lengthy and costly endeavor, taking an estimated 12 to 15 years and costing between $879 million [1] and $2.87 billion [2] when factoring in failed candidates and regulatory setbacks. Just 35% of investigational drugs reach clinical trials, and only 9–14% progress from Phase I to approval [3]. The early discovery phase is especially resource-intensive, requiring teams to screen vast libraries of candidates, identify promising hits, and optimize leads, all before clinical testing begins.

To accelerate this process and reduce costs, pharma companies are turning to AI. In an AI-augmented pipeline, early discovery processes like hit identification, screening, and lead optimization can be accelerated by predictive models and automation. Morgan Stanley estimates a 20% to 40% reduction in costs for preclinical development, that would represent as much as a 15% increase of approved therapies [5]. This promise is already reshaping investment priorities. Two-thirds of pharma companies are increasing IT and AI investments in 2024 [6].

Despite AI's growing promise in biopharma, large organizations struggle to fully integrate it into their research workflows.

Challenges to AI adoption

As AI adoption grows, the industry is facing shared challenges that often lead to **underutilized AI investments**. From infrastructure limitations to AI integration into daily practices, companies face technical and operational barriers that slow AI adoption.

Here's a look at the top five challenges:

1. Fragmented and inaccessible data
High-quality data is essential for model training, yet it is often fragmented, unstructured, siloed, and stored in inconsistent formats, disconnected from LIMS and other environments, making retrieval and integration slow and inefficient.

2. Complexity of integrating AI into research workflows
Biologics research teams struggle to incorporate AI into existing workflows, as it requires major adjustments in data preparation, analysis, and model deployment. Without integration, AI remains disconnected from daily research.

3. Maintenance burden
Maintaining AI and data infrastructure requires ongoing updates, system integrations, and resource allocation. Without dedicated support, in-house solutions become outdated, inefficient, and costly to sustain, slowing down biologics discovery instead of accelerating it.

4. Versioning and documentation challenges
AI models and bioinformatics workflows constantly evolve, making version tracking and documentation critical for reproducibility and compliance. Without structured tracking, teams risk using outdated models, losing performance history, and facing validation or regulatory challenges.

5. Computational power and scalability
AI models require significant computational power, but many biologics research teams lack the infrastructure to support large-scale data processing, model training, and deployment. Without scalable resources, training is slow, costs rise, and AI struggles to move beyond proof-of-concept.

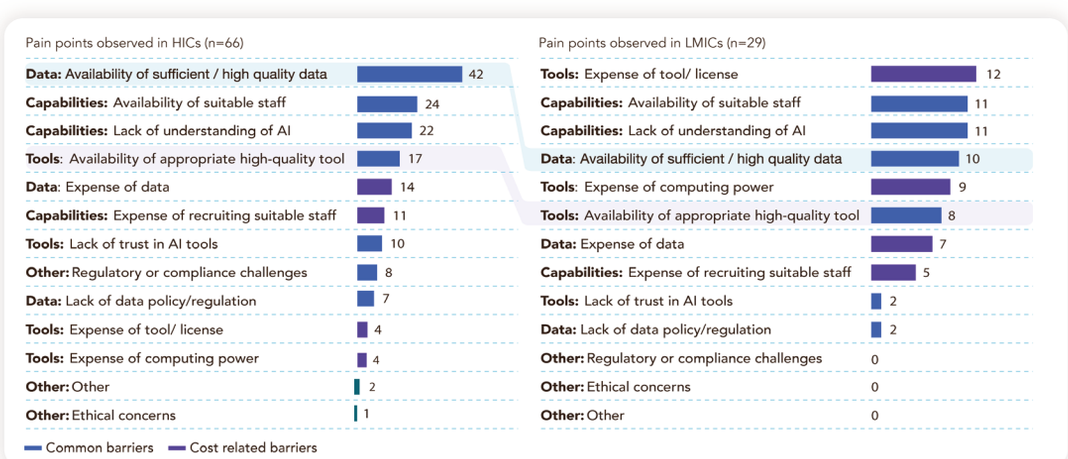


Figure 1. Ranked pain points for high-income (HICs) and low- and middle-income countries (LIMCs) for adopting AI-enabled drug development workflows [3].

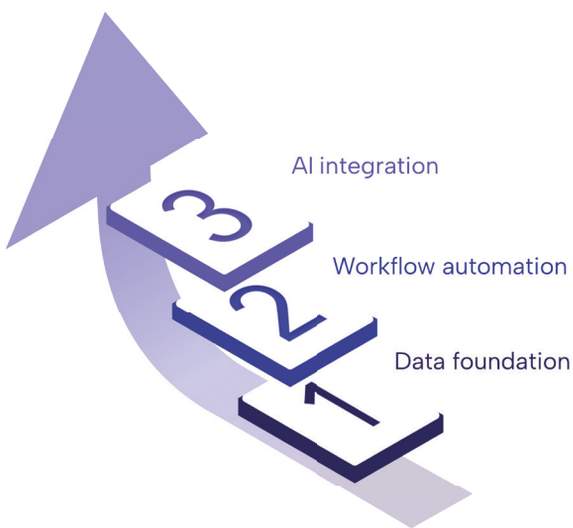
ENPICOM's blueprint to AI adoption

ENPICOM developed a three-step approach that provides a clear path to successfully integrating AI into antibody development workflows. With a strong data foundation, streamlined bioinformatics workflows, and seamless AI integration, organizations can fully leverage AI to accelerate lead selection, optimize candidates, and shorten development timelines.

1. **Data Foundation** – A centralized, structured, and scalable data architecture for well-labeled, accessible, and harmonized datasets for efficient AI model training and research.

2. **Workflow Automation** – Standardized and automated data analysis workflows within a unified ecosystem, consolidating tools and structuring data processing.

3. **AI Integration** – A streamlined approach to train and version track models, along with an automated system for deploying models and embedding them into workflows.



Introduction to the ENPICOM Platform

To enable seamless AI adoption in biologics discovery, we have developed the ENPICOM Platform, a comprehensive solution that integrates data management, workflow automation, ML model training, and deployment of production-ready models.

The ENPICOM Platform provides:

- **A scalable and integrated data architecture** that centralizes, structures, and harmonizes research data.
- **Lab scientists' interface** to analyze data, select and optimize lead candidates, and build and run automated analysis workflows.
- **ML scientists' interface** to streamline MLOps by tracking model performance and registering models.
- **Model deployment and consumption** through a CloudOps layer for AI deployment, enabling seamless integration into research workflows and decision-making processes.

In the next sections, we go into detail on how the ENPICOM Platform, and our expertise help partners implement each step of the AI adoption blueprint, addressing key challenges and enabling successful AI adoption in discovery workflows.

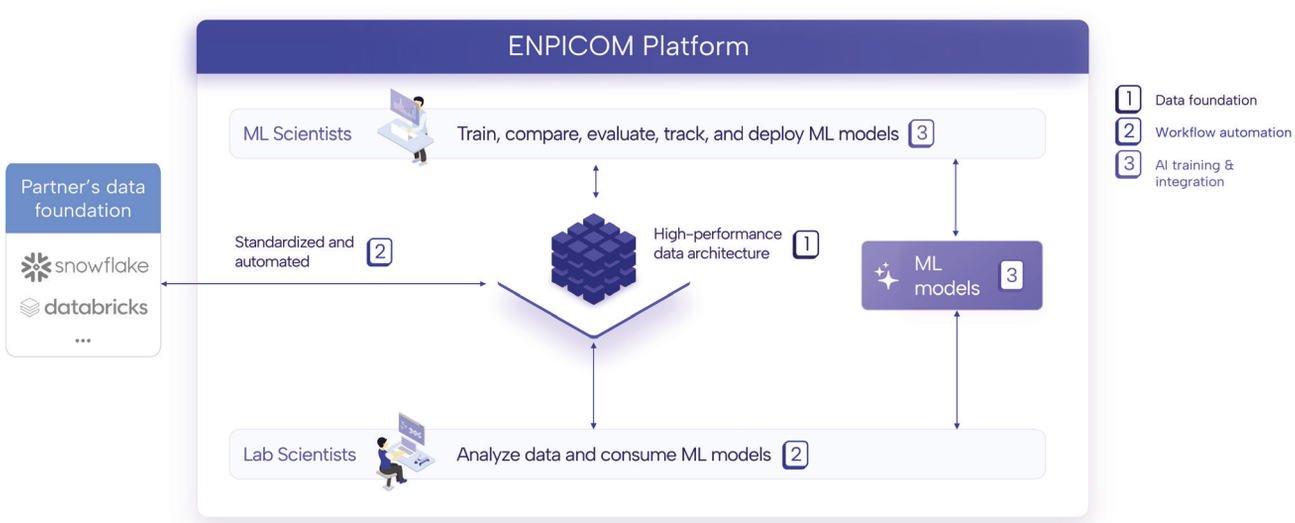


Figure 2: ENPICOM Platform is a comprehensive cloud-based software platform for lab, informatics, and AI teams.

STEP 1

Data foundation

The backbone of AI in biologics research

AI is only as good as the data it learns from, yet pharma organizations struggle with fragmented, inconsistent, and inaccessible data, making AI adoption and scaling difficult. Legacy data architecture creates data silos, resulting in hidden costs, resource duplication, collaboration obstacles, and reduced productivity [2, 3].

To fully leverage AI, annotated data must be stored in high-performance databases that enable rapid retrieval and seamless access to large-scale datasets. Findable, Accessible, Interoperable, and Reusable (FAIR) data drives scientific innovation by making information searchable, ready for advanced analytics, and able to flow freely across systems, tools, and teams.

This involves:

- **Structured data**: Well-structured, labeled data that is easily accessible to be consumed by ML models
- **Scalability**: An infrastructure that supports growing datasets without performance bottlenecks.
- **Centralized storage and integration**: A single source of truth, accessible across departments and compatible with existing systems.

Robust data foundation by ENPICOM

Designed for biologics, the ENPICOM Platform integrates seamlessly into any ecosystem or federated data warehouse, eliminating silos. Powered by a proprietary high-performance architecture, it delivers instant access to all your data, accelerating discovery and AI-driven insights.

Built for biologics

The ENPICOM Platform is engineered to process large-scale biologics data with speed and accuracy. State-of-the-art algorithms and tools help you make full use of your sequencing and assay data to identify high-quality, developable candidates with greater efficiency.

Structured data

ENPICOM Platform ensures that all data is well-labeled, structured, and easily accessible through the UI or API. With well-organized data, you can easily augment and correlate experimental results within and across projects. Uniquely, our platform instantly incorporates new data into all previously conducted analyses, no manual reprocessing needed. This real-time data integration capability keeps your teams working with the latest insights, accelerating decision-making and improving scientific outcomes compared to traditional approaches that create isolated data snapshots.

STEP 1

High-performance infrastructure

Effective data management is crucial to discovery analyses including diverse experimental metadata and/or large sequencing datasets, and the ENPICOM Platform excels in this domain. All annotated data is stored in a rapid, scalable database that ensures effortless access and extraction for analyses. This database houses sequencing data and enables seamless association of any assay or in-silico generated data with individual clones, allowing you to integrate these in your analysis to find the best leads.

At the core of this system is Amazon's Redshift Massive Parallel Processing (MPP) architecture, combined with our proprietary high-throughput SQL engine. This powerful combination allows for blazing-fast queries, enabling you to search receptor motifs among hundreds of millions in mere seconds. The database is built to scale to **petabyte-sized datasets,** ensuring smooth data access and retrieval. Querying massive datasets remains seamless, with no drop in performance from **10M to 500M clones.**

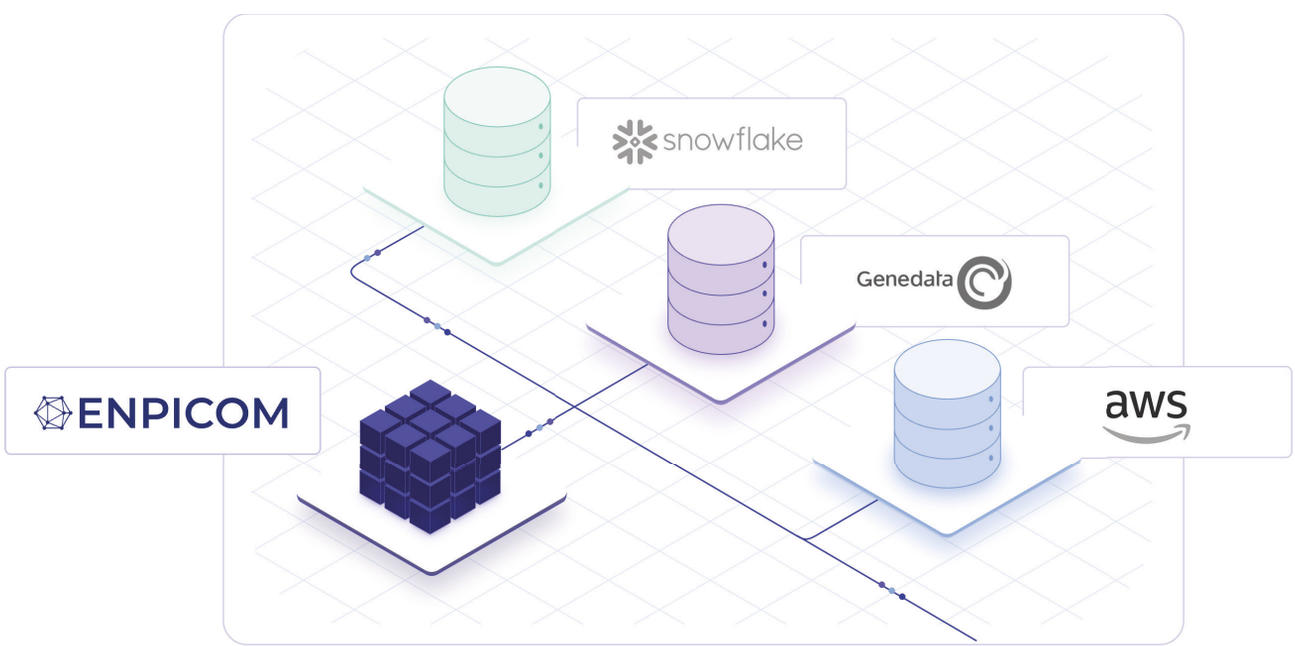
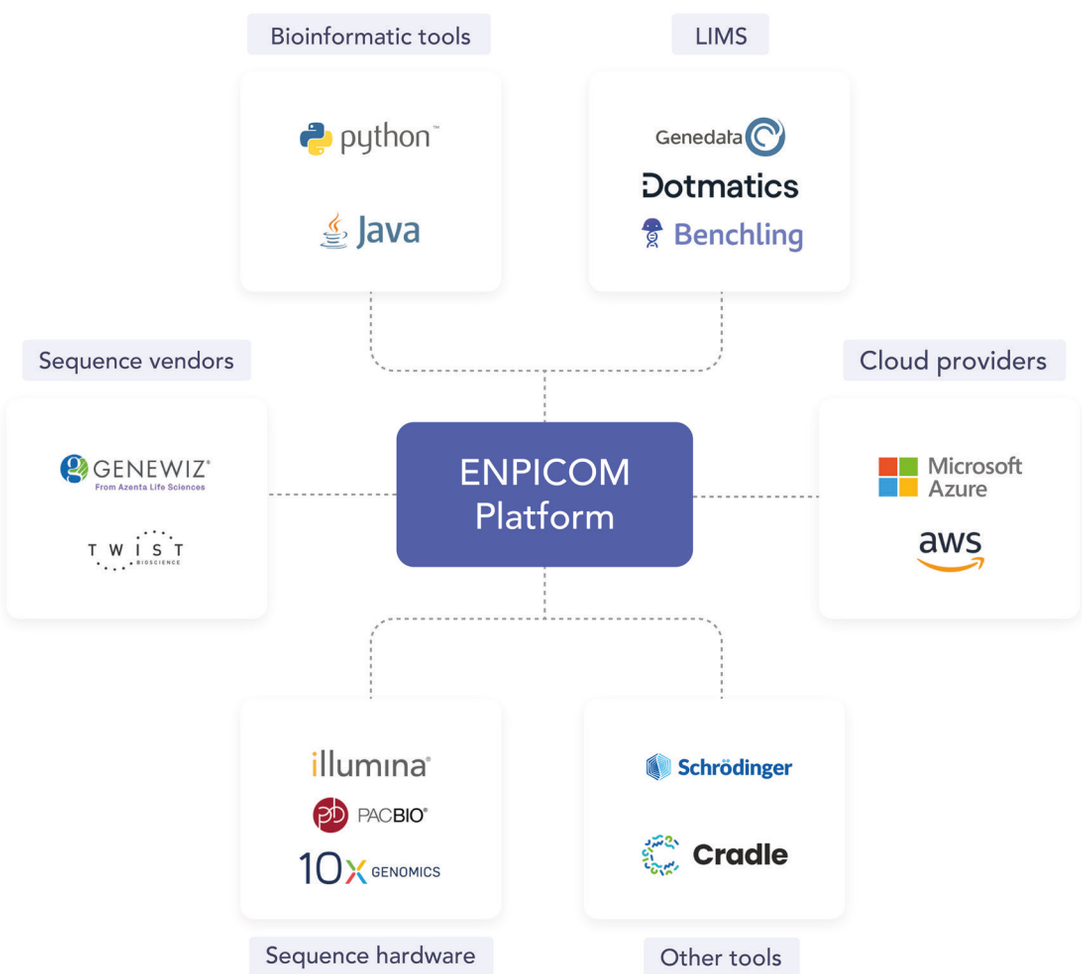


Figure 3: ENPICOM's data foundation integrates seamlessly into any existing ecosystem or federated data warehouse, eleminating the risk of data silos.

STEP 1

Centralized storage and integration

Eliminate data silos and fragmented systems. Our data foundation integrates seamlessly with existing systems and federated data warehouses, creating a single source of truth, FAIR by design, accessible across teams, and optimized for advanced analytics and AI applications.With a comprehensive API, ENPICOM enables seamless connectivity with Laboratory Information Management Systems (LIMS), Electronic Laboratory Notebooks (ELNs), and various cloud storage solutions (like S3, Google Cloud, Azure) for automatic synchronization of experimental data and metadata. For developers, our user-friendly Python SDK simplifies the creation of custom workflows and integrations, empowering teams to tailor the platform to their specific project needs.



STEP 2

Workflow automation and consolidation

Workflow automation: Unlocking speed and precision in data analysis

Biologics discovery relies on large, complex datasets that must be processed, analyzed, and structured efficiently for reproducibility and to be consumed by models. However, research teams often face fragmented workflows, disconnected tools, and inconsistent data formats, leading to data silos and unstructured datasets that introduce inconsistencies, bias, and reduced model reliability.

When lab scientists store data locally or in isolated systems, it further increases overhead and hinders data access across teams. To ensure smooth data, scientists need user-friendly analysis tools that integrate effortlessly into their workflows without causing delays. Any gaps or inefficiencies in data processing can lead to fragmentation, jeopardizing the success of AI initiatives.

Well-defined data analysis protocols and adaptable tools that unify and streamline data consolidation are critical to enable cross-functional collaboration, eliminate data silos, and ensure effective AI adoption. This requires scalable, reliable workflow orchestration capable of managing various data streams and processes at scale.

Automation plays a vital role in keeping pace with the rapid growth of data in modern labs. By minimizing human error and reducing manual intervention, automated systems ensure higher data quality and consistency. They also reduce user fatigue and free scientists from repetitive tasks, allowing them to focus on high-value work.
Finally, automation is the only way to achieve the scalability and efficiency needed for AI to truly transform drug discovery.

Workflow automation and consolidation by ENPICOM

At ENPICOM, we work closely with companies to automate their workflows, combining our expertise in biology, bioinformatics, and software engineering. We translate the data analysis processes into automated workflows, ensuring efficiency, reproducibility, and scalability.

State-of-the-art tools

Bioinformatic scripts and legacy tools often fall short when handling today's sequencing and screening technologies. They are also not built for integration into complex data ecosystems. Our biologics discovery platform was purposefully designed to address these challenges. It combines state-of-the-art algorithms with a comprehensive API to ensure high performance and seamless integration.

Cross-domain expertise

Pipeline automation without scientific alignment can lead to flawed processes and results that don't reflect your scientific priorities. With deep expertise in B-cell immunology, bioinformatics, and machine learning, we work closely with your team to develop automation strategies that reflect your discovery platform and scientific goals. This collaborative approach ensures accurate and reproducible data capture.

STEP 2

Unmatched implementation speed

Historical software solutions are often developed at a static, slow-evolving pace. Traditional pipeline rollouts and customizations can take years, delaying discovery and progress. Our team has the expertise to continuously advance our solutions, cut through complexity to deliver fully operational pipelines in just a few months, without compromising flexibility or performance.

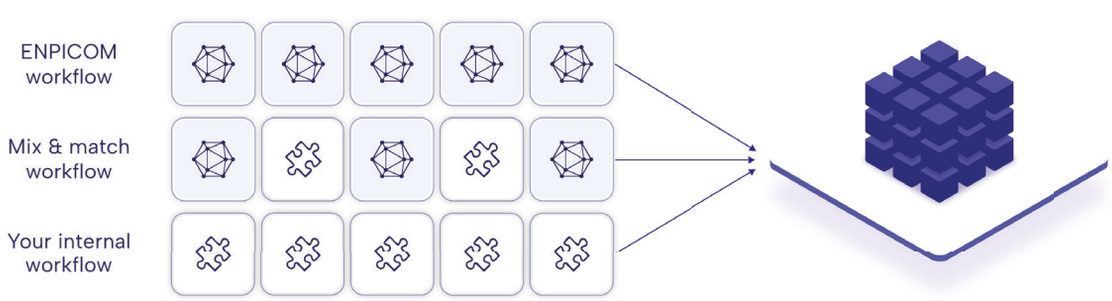
Our platform enables researchers to automate pipelines and standardize data processing, and maintain flexibility for project-specific needs by providing:

- **Automation** – Automated pipelines that execute analysis steps efficiently, generating structured, harmonized, and well-labeled data.
- **Standardization** – Implementing automated pipelines across all users to ensure consistent data processing, making all generated data AI-ready and eliminating variability.
- **Integration** – A data infrastructure that consolidates research data from multiple sources, enabling streamlined analysis and AI adoption.
- **Expertise** – ENPICOM's experts collaborate with partners to optimize automation, ensuring alignment with scientific goals and accurate data capture.

Flexible options for workflow automation

Together with your team, we define the most effective approach to workflow automation, ensuring seamless integration and full adaptability as your needs evolve.

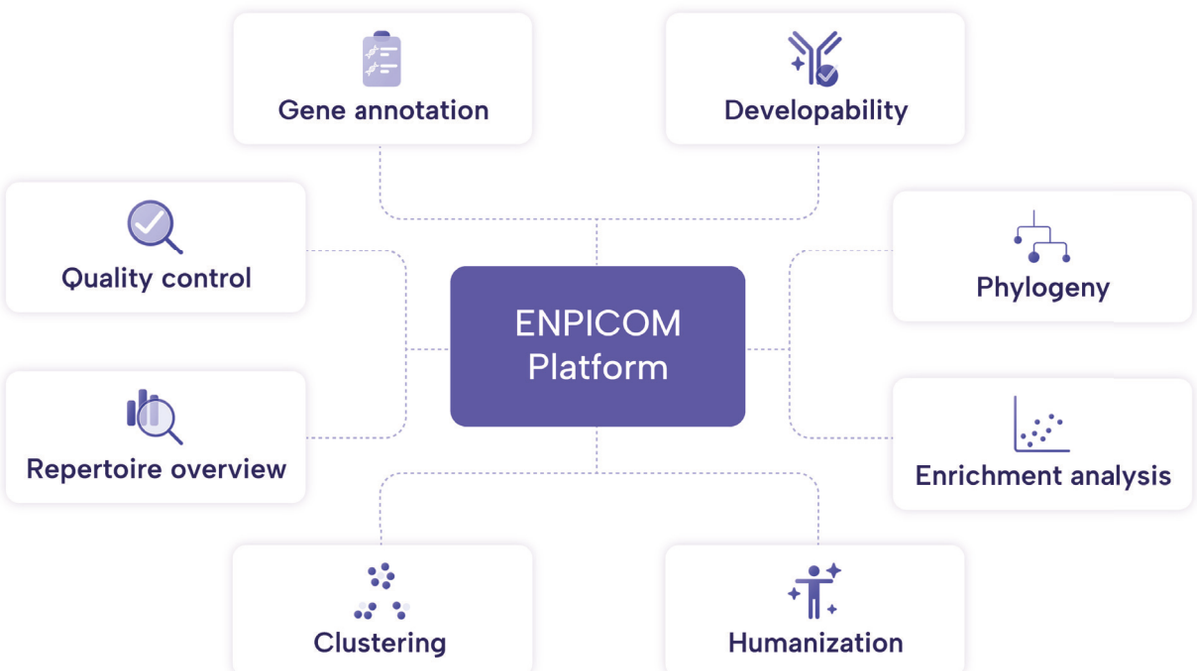
- **ENPICOM workflow** – A fully optimized bioinformatics workflow built with ENPICOM's cutting-edge tools and models.
- **Mix-and-match workflow** – A hybrid approach integrating your tools with ENPICOM's bioinformatics solutions.
- **Your workflow** – A workflow where your bioinformatics tools are automated and optimized.



STEP 2

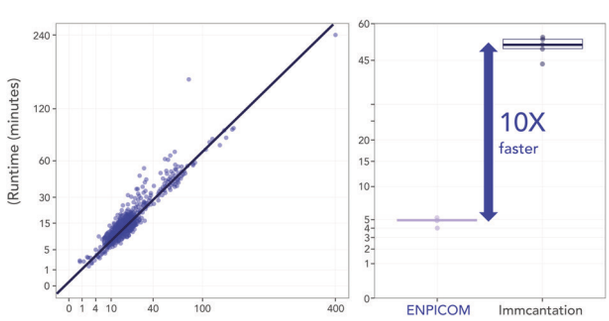
Tools & Models by ENPICOM

Whether you choose to use our tools for workflow automation or manual data processing through the lab scientist interface, integrating ENPICOM's solutions gives you access to comprehensive tools and models that are tailored to support efficient analysis, selection, and optimization of lead candidates.



Clone annotation

Transforms raw sequencing data into high-quality, annotated clones with accuracy and scalability. Supports B and T cell annotation across species and sequencing technologies, including plate- and microfluidics-based protocols, with UMI error correction and paired-chain linking.
ENPICOM's tool for processing raw sequencing data (germline alignment, gene annotation) is highly scalable and outperforms industry standards.



Quality control

Interactive visualizations for fastq file quality control, including Read Length Distribution and Base/Sample Quality. Users can inspect read fate and detect sequencing biases.

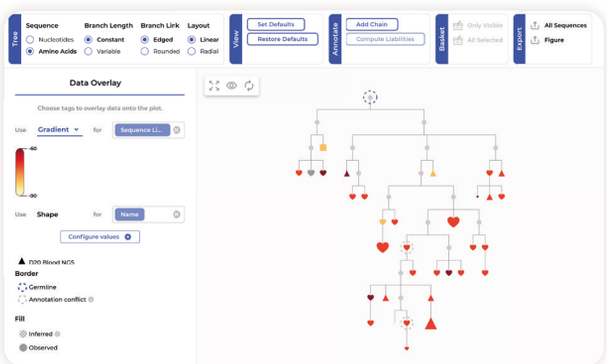
Repertoire overview

Analyzes high-level characteristics of immune system responses, including V/J gene usage, CDR3 length distribution, and diversity measures.

STEP 2

Clustering

Cluster clones based on user-defined parameters like gene usage and CDR3 length, and any combination of sequence regions, with the flexibility to group across samples and sequencing technologies.

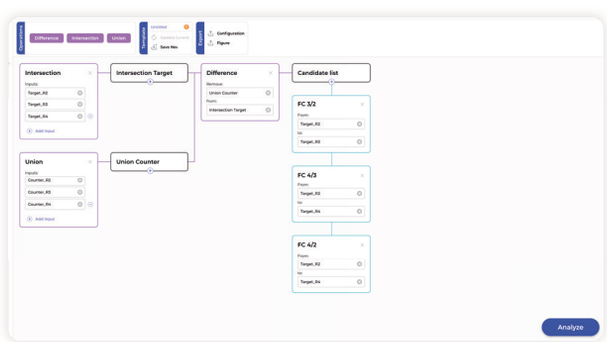


Phylogeny

Generate interactive phylogenetic tree visualizations with the ability to overlay metadata such as assay results, liability predictions, and clone counts to identify trends and select optimal candidates.

Enrichment

Select highly enriched candidates from display screening using dataset joining, intersection, and subtraction features. Track specific sequences across multiple datasets and analyze enrichment dynamics over selection rounds.



ML-based developability predictions

High-throughput model that generates structural models and predicts exposed liabilities, enabling early risk identification. Also assesses developability by comparing with antibodies that reached the clinic.

ML-based humanization

Humanizes antibody candidates while maintaining high germline identity and superior binding affinity, outperforming industry standards. It supports humanization from any species, **requiring only 3–5 variants** to achieve clinically relevant humanness levels (>85% and >90% germline identity) while retaining binding affinity.

STEP 3

AI Integration

AI integration: Standardized, reproducible machine learning

Integrating AI into biologics research comes with challenges in establishing reproducible and scalable ML model development, and in making production-ready models readily available to lab scientists.

In biologics discovery, large datasets can be hard to obtain and models are often generated in an iterative way while the data improves incrementally. Meanwhile, lab and data science teams often work in isolation, leading to hurdles for data scientists in obtaining the most up-to-date data, and hurdles for lab scientists in consuming the ML models in their analyses.

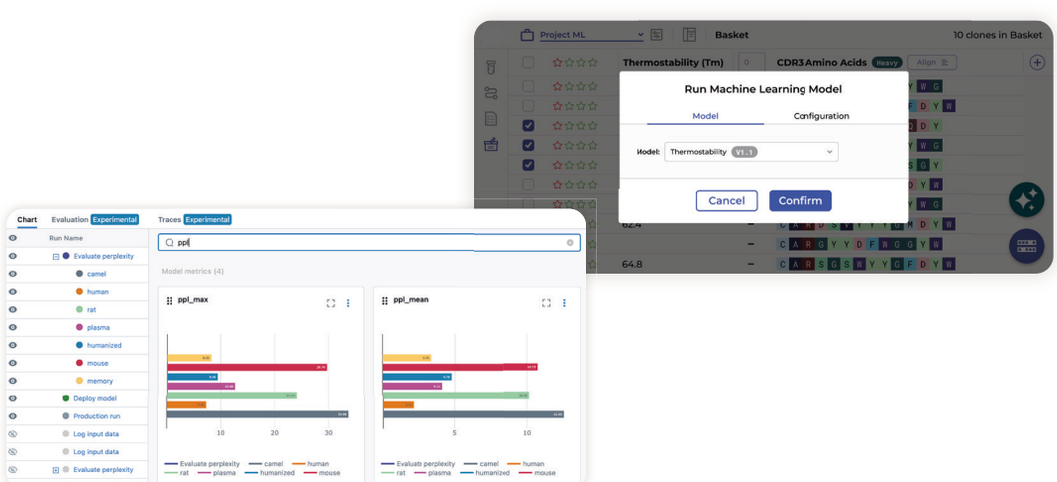
To fully unlock the power of AI in biologics discovery, ML scientists need a structured environment that enables them to develop, manage, and deploy models efficiently. A well-designed infrastructure allows them to track model performance, ensure reproducibility, and maintain consistency, leading to more reliable predictions, faster iteration cycles, and seamless integration into research workflows.

Streamlining MLOps with the ENPICOM platform

Train and track ML models

The ENPICOM Platform provides an integrated AI environment, providing ML scientists with:

- **Model performance tracking.** Compare models parameters, metrics, and track their performance over time to confidently select the best-performing model for production.
- **Model registry**, ensuring proper versioning, accessibility, and compliance with best practices through an intuitive model registry. Maintain a clear history of versions, datasets, and results, enabling scientists to trace data origins, track model evolution, and ensure reproducibility without disrupting workflows.
- **Easy model deployment**, allowing ML scientists to control which ML models are made available to lab scientists. Once a model is deployed, it becomes immediately available in the lab scientists' analysis workflow, ensuring full integration between environments for seamless adoption and immediate impact.



STEP 3

Deploy and consume models

The ENPICOM Platform enables optimized AI deployment, allowing ML scientists to make models accessible in the lab scientist UI through a well-defined deployment and consumption architecture. This ensures AI-driven insights are easily integrated into daily research workflows. The next section dives deeper into how models are deployed using our platform.

1. Select production-ready model

ML scientists select a model from their ML environment to deploy. The model will be pushed and hosted using a CloudOps layer developed by ENPICOM, a cloud-based infrastructure that manages deployment, scalability, and accessibility. Models will be scaled according to demand and are configured to process many invocations in parallel, ensuring timely results.

2. Ensuring seamless workflow integration

Enabling one-click consumption of the models is achieved by defining where the models get their input and by defining their intents. This works as follows.

First, ML scientists determine which data the model requires and configure it to automatically pull the necessary inputs from the database. For example, a liability prediction model may need antibody sequence data, while a clustering model might require CDR3 sequences and similarity metrics.

Then, the ML scientists clarify the intent of the model by defining the type of data it will generate, ensuring it appears in the right place in the UI for lab scientists to use, or that it is integrated into the correct stage of automated workflows, allowing scientists to access and apply machine learning insights where they matter most.

Model intent examples include:

- **Metadata** (e.g., liability scores, confidence levels)
- **File outputs** (e.g., reports, structured datasets)
- **Sequences** (e.g., optimized antibody variants)
- **Clustering results** (e.g., grouping of similar candidates for selection)
- **Liability assessments** (e.g., identification of developability risks)

The resulting experience is one of one-click insights, where lab scientists only need to select their clones and click a model. There is no additional configuration needed. The system automatically supplies the necessary data to the model and integrates the results it generates directly into the scientist's workflow. For even greater efficiency, models can also be embedded into automated workflows creating a fully integrated and uninterrupted process.

3. Provenance

To ensure full traceability and reproducibility of what models are used by the lab scientist UI, ENPICOM also takes care of model registry, logging which model version was used, who used it, and what data was the input. This allows teams to track predictions, validate results, and maintain compliance effortlessly.

Applications of ML adoption in biologics discovery

Adopting machine learning in biologics discovery goes beyond just implementing AI models. It transforms how researchers analyze data, identify lead candidates, optimize designs, and predict key properties. By integrating AI into daily workflows, scientists can accelerate decision-making, improve efficiency, and enhance the success rate of therapeutic discovery.

In this section, we explore practical applications of machine learning in biologics, demonstrating how AI-driven insights can streamline discovery, reduce timelines, and unlock new possibilities.

Use case: Quick register and deploy new models for candidate optimization

In biologics discovery, new AI models are constantly being developed to enhance lead optimization, antibody engineering, and candidate selection.
In this example, we are looking at a candidate optimization model described in this study [15], that utilizes machine learning algorithms to predict and generate antibody sequences with high binding affinity and diversity.

By learning from existing antibody-antigen interaction data, the model can suggest novel antibody candidates that are more likely to exhibit strong binding properties, accelerating the development of effective therapeutics. While this model is publicly available on GitHub, to be effectively used in biologics discovery, it needs to be trained with target-specific data, validated, and deployed in a structured way so that lab scientists can easily access and apply its predictions.

From research paper to implementation in a day:

1. Retrieve the model from external sources

- ✓ If the model is publicly available on Hugging Face or GitHub, ML scientists download it to their local or cloud environment.

2. Upload and train the model in MLFlow

- ✓ ML scientists retrieve relevant dataset(s) from the database and **train the model for** specific targets.
- ✓ The trained model is then **uploaded into the integrated MLFlow environment** within the ENPICOM Platform.
- ✓ Multiple datasets are tested to **compare model quality and evaluate different training strategies.**

3. Track model performance and select the best version

- ML scientists track each trained model's **performance metrics, input data, and results** within MLFlow.
- By comparing different versions, they can **identify the best-performing model** for deployment.

4. Deploy to CloudOps for automated execution

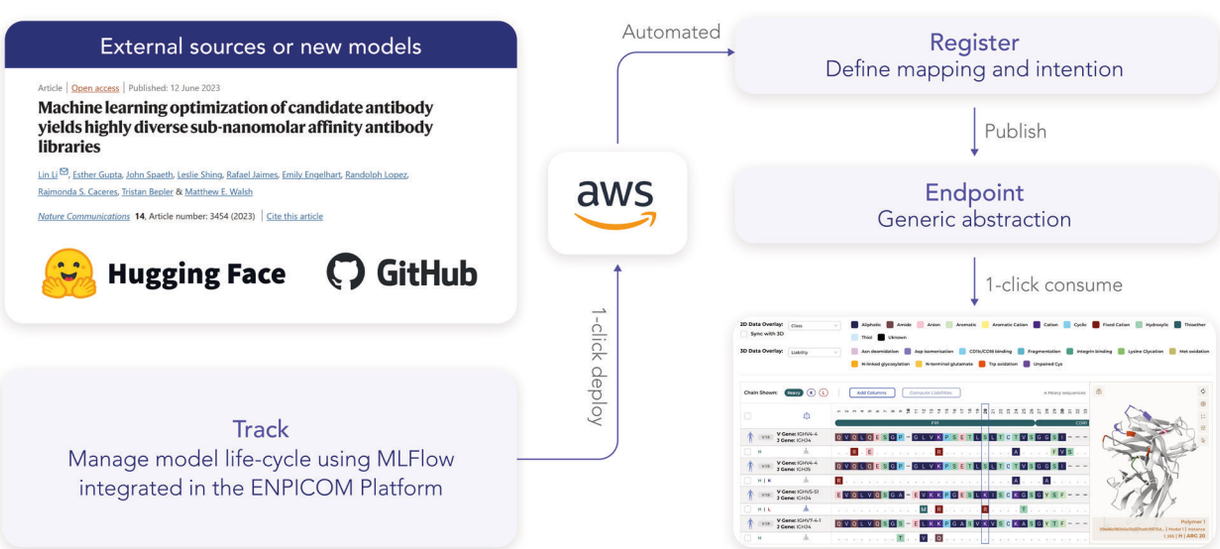
- Once a final model is selected, it is **deployed to CloudOps**, ensuring it can be automatically executed without manual intervention.
- CloudOps manages **scalability, resource allocation, and uptime**, ensuring the model is always available when needed.

5. Register and define model intent

- The **automated model registry** logs which version is being deployed, ensuring full traceability.
- The model's **intent is determined**, including **what type of input data it requires and where it can be accessed** in the interface for lab scientists.
- Inputs are mapped for **automated data extraction**, ensuring lab scientists can use the model **without technical setup.**

6. Model becomes available in the right interface

- The model is now **accessible to lab scientists** in the appropriate section of the ENPICOM Platform.
- In this case, the model becomes available in the "Basket" environment, where scientists can generate optimized variants of initially selected candidates.



References

1. Solomon, L. (July 2, 2024.) Mean cost of bringing new drug to U.S. market is $879.3 million. Drugs.com. Accessed September 26, 2024.
2. DiMasi, JA, et al. (May 2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. Accessed October 10, 2024.
3. Understanding the potential of AI in drug discovery: Status, barriers and future opportunities. (June 25, 2023.) Wellcome Trust.
4. Fueling the AI transformation: Four key actions powering widespread value from AI, right now. (October 2022). Deloitte's State of AI in the Enterprise, 5th Edition report.
5. Why Artificial Intelligence Could Speed Drug Discovery. (September 9, 2022.) Morgan Stanley.
6. Buntz, B. (December 14, 2023). Two-thirds of pharma companies plan to up IT investments in 2024, survey finds. Drug Discovery & Development.
7. Gen AI: A game changer for biopharma operations. (January 28, 2025) McKinsey&Company.
8. Scaling gen AI in the life sciences industry. (January 10, 2025) McKinsey&Company.
9. Chadha, S et al. (2023.) Technology Vision 2023 for Biopharma: Merging atoms and bits. Accenture. Accessed October 24, 2024.
10. 2024 Global Life Sciences Sector Outlook. deloitte.com. Accessed November 14, 2024.
11. AI in the pharmaceutical industry – statistics & facts. statista.com. Accessed November 14, 2024.
12. What's the future of generative AI? An early view in 15 charts. (2023). McKinsey & Company. https://www.mckinsey.com/ featured-insights/mckinsey-explainers/whats-the-future-of-generative-ai-an-early-view-in-15-charts
13. Oldfield, J. (2023, May 4). Researchers use generative AI to design novel proteins. Phys. https://phys.org/news/2023-05- generative-ai-proteins.html
14. Melanie Senior, "Fresh from the biotech pipeline: fewer approvals, but biologics gain share," Nature Biotechnology 41 (2023): 174–182, https://doi.org/10.1038/s41587-022-01630-6.
15. Li L, Gupta E, Spaeth J, Shing L, Jaimes R, Engelhart E, Lopez R, Caceres RS, Bepler T, Walsh ME. Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. Nat Commun. 2023 Jun 12;14(1):3454. doi: 10.1038/s41467-023-39022-2. PMID: 37308471; PMCID: PMC10258481.

ENPICOM

info@enpicom.com | +31852500575 | enpicom.com
Sint Janssingel 88 | 5211 DA ‘s-Hertogenbosch | The Netherlands