AI-Driven Drug Discovery and Development

EURON

Use Case: AI-Driven Drug Discovery and Development

Objective

Leverage AI/ML models and big data processing to accelerate drug discovery, predict molecular interactions, optimize clinical trials, and enhance decision-making in pharmaceutical research.

---

1. Functional Architecture

The functional architecture describes the end-to-end workflow of the AI-driven drug discovery system.

Actors

1. **Pharma Researchers & Scientists** – Analyze results and provide domain expertise.
2. **Data Scientists & ML Engineers** – Build and train AI models for drug discovery.
3. **Regulatory & Compliance Team** – Ensure adherence to FDA/EMA regulations.
4. **IT & DevOps Teams** – Maintain infrastructure and monitor system health.

Key Functional Components

1. **Data Ingestion Layer**
   - Sources: Genomics databases, chemical libraries, past clinical trials, patient records, scientific literature (PubMed, FDA, research papers).
   - Data Pipeline: Uses **Apache NiFi or Confluent Kafka** for real-time and batch data ingestion.
   - Data Types: Structured (clinical data), semi-structured (research articles), unstructured (molecular images).
2. **Data Storage & Management**
   - **Molecular & Genomic Data** – Stored in NoSQL databases like **MongoDB** or **AWS DynamoDB**.
   - **Clinical Trial Data** – Stored in **PostgreSQL/MySQL**.
   - **Scientific Literature & Research Papers** – Indexed using **Elasticsearch** for NLP-based searches.
   - **Image Data** (MRI scans, histopathology) – Stored in **AWS S3, Google Cloud Storage** with metadata in **Neo4j** (graph database).
3. **AI/ML Processing & Analytics**
   - **Molecular Structure Analysis**
     - Deep learning models (Graph Neural Networks, Transformers) process molecular structures to predict potential drug interactions.
   - **Clinical Trial Optimization**
     - ML models predict patient eligibility, dropout rates, and trial outcomes.
   - **Drug Target Interaction (DTI) Prediction**

        - Uses **Graph Neural Networks (GNNs)** and **Transformer-based BioBERT models**.
    - **Adverse Drug Reaction (ADR) Prediction**
        - NLP models analyze patient records, social media, and medical reports.

4. **Orchestration & Processing Layer**
    - **Apache Spark (Databricks on AWS/GCP)** – Distributed computing for large-scale analytics.
    - **Ray or Dask** – Parallel computation for AI workloads.
    - **Kubernetes** – Manages AI workloads efficiently.
5. **Model Training & Deployment**
    - **MLFlow** – Model tracking and lifecycle management.
    - **TensorFlow/PyTorch** – Training deep learning models.
    - **Databricks or SageMaker** – Model training, hyperparameter tuning, and deployment.
    - **Inference Engine** – Exposes AI models via **REST/GraphQL APIs** for easy integration.
6. **Visualization & Reporting**
    - **Power BI/Tableau** – Dashboards for real-time insights.
    - **Streamlit** – Web-based interactive AI model result visualization.
7. **Security & Compliance**
    - **Data Anonymization & Encryption** – Ensures compliance with **HIPAA, GDPR, FDA 21 CFR Part 11**.
    - **Audit Logging & Monitoring** – Using **ELK Stack, Prometheus, Grafana**.

---

2. Technical Architecture

Below is a high-level technical architecture diagram outlining major components and interactions.

Technology Stack

Layer	Technology Choices
Data Ingestion	Apache NiFi, Confluent Kafka, AWS Glue
Data Storage	MongoDB, PostgreSQL, Neo4j, Elasticsearch, AWS S3
AI/ML Processing	PyTorch, TensorFlow, Hugging Face, BioBERT, Graph Neural Networks
Big Data Processing	Apache Spark (Databricks), Dask, Ray
Model Deployment	MLFlow, SageMaker, Kubernetes, FastAPI
Visualization	Tableau, Power BI, Streamlit
Security & Compliance	AWS IAM, HashiCorp Vault, Prometheus, Grafana

Technical Flow

1. **Data Ingestion**
   - Apache NiFi ingests genomics, molecular, and clinical trial data.
   - Confluent Kafka streams real-time updates from research papers and adverse drug reports.
2. **Data Storage & Processing**
   - MongoDB stores molecular data.
   - PostgreSQL stores structured clinical data.
   - Elasticsearch indexes research papers.
   - Graph-based representation (Neo4j) links molecular interactions.
3. **AI/ML Training & Inference**
   - Spark on Databricks processes large-scale molecular data.
   - TensorFlow/PyTorch-based models predict drug interactions.
   - SageMaker/MLFlow manages model training and deployment.
4. **Real-Time Monitoring & Analytics**
   - ELK Stack provides logs and auditing.
   - Prometheus/Grafana monitors system performance.
   - Power BI/Tableau visualize AI results for researchers.