🧬 Agentic RNA Biomarker Detection

Problem Statement

Predicting RNA structure and assessing cancer risk from sequence data remains a challenge due to limited structural databases, complex folding, and fragmented tools. Most labs lack integrated, explainable, and privacy-preserving platforms for actionable cancer risk insights.

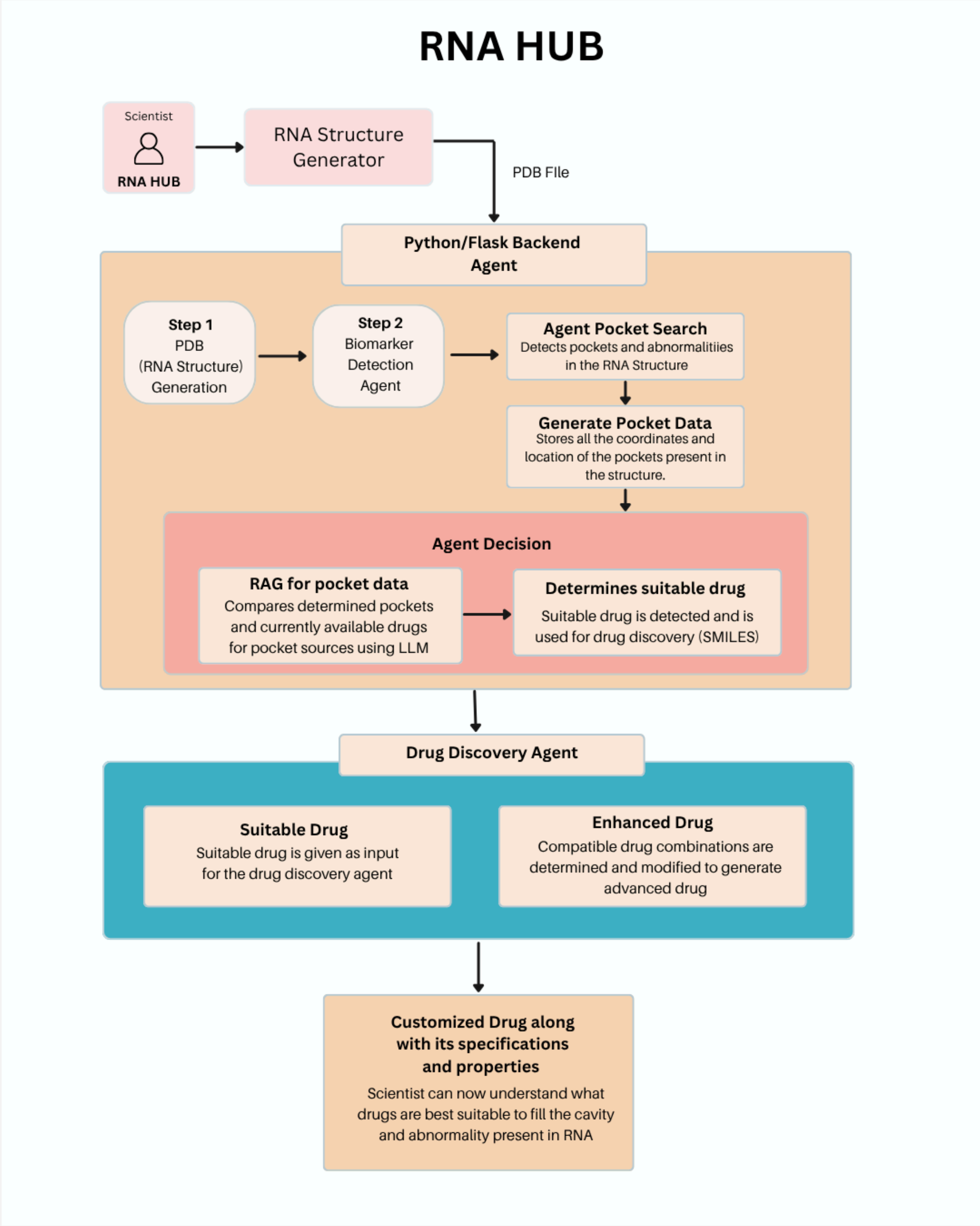
Our project: enables researchers to input RNA sequences, uses local agentic AI for structure prediction and cancer classification, and outputs clear, interpretable results all in one web interface. This accelerates cancer biomarker discovery, empowers clinicians, and ensures compliance, setting a new standard for accessible, explainable, and secure RNA-driven cancer diagnostics with next-generation AI.

Overview

This project implements an Agentic AI pipeline for detecting biomarker regions in RNA 3D structures (`.pdb` files). It integrates structural bioinformatics tools and a locally running BioGPT LLM to automate the workflow — from structure parsing to functional annotation. The project also demonstrates the integration of LangChain for orchestrating AI reasoning and RAG (Retrieval-Augmented Generation) for combining structured RNA data with external knowledge sources.

After detecting biomarker regions, the system can suggest potential drug candidates using molecular SMILES representations and diffusion-based generative models for virtual compound generation. Although this project currently uses BioGPT and the same process for analysis, this step demonstrates the potential for AI-assisted RNA-targeted drug discovery.

🔍 System Architecture



RNA HUB

Scientist

RNA HUB

RNA Structure Generator

PDB FIle

Python/Flask Backend Agent

Step 1
PDB
(RNA Structure)
Generation

Step 2
Biomarker
Detection
Agent

Agent Pocket Search
Detects pockets and abnormalitiies in the RNA Structure

Generate Pocket Data
Stores all the coordinates and location of the pockets present in the structure.

Agent Decision

RAG for pocket data
Compares determined pockets and currently available drugs for pocket sources using LLM

Determines suitable drug
Suitable drug is detected and is used for drug discovery (SMILES)

Drug Discovery Agent

Suitable Drug
Suitable drug is given as input for the drug discovery agent

Enhanced Drug
Compatible drug combinations are determined and modified to generate advanced drug

Customized Drug along with its specifications and properties
Scientist can now understand what drugs are best suitable to fill the cavity and abnormality present in RNA

Agentic AI Orchestrator

- Coordinates multiple tools:
    - Parses RNA `.pdb` files
    - Runs fpocket to detect potential binding pockets
    - Uses LangChain to manage AI reasoning and task planning
    - Utilizes RAG techniques to enhance BioGPT predictions with external databases (e.g., RNAcentral, PubMed)
    - Produces biomarker predictions in JSON format
    - Suggests drug candidates via SMILES-based reasoning and generative models

---

⚙️ Tools Used

1️⃣ PDB Parser (Biopython)

- Extracts atomic coordinates, residue sequences, and secondary structure motifs from `.pdb` files
- Outputs structured JSON summarizing the RNA structure

Example Output:

```
{
  "chains": ["A"],
  "residue_count": 46,
  "motifs_detected": ["stem-loop", "bulge"],
  "sequence": "AUGCUAGU..."
}
```

2️⃣ fpocket

- Detects binding pockets/cavities in RNA structures
- Provides pocket coordinates, volumes, hydrophobicity, and druggability score

Example Output:

```
{
  "pocket_id": 1,
  "score": 52.4,
  "center": [12.4, 5.7, -3.1],
  "residues": ["A12", "U13", "G14"]
}
```

3️⃣ BioGPT (Local LLM)

- Analyzes structure + pocket data to predict:
  - Biomarker regions
  - Structural motifs
  - Function (ligand/metal binding)
  - Disease relevance
- Enhanced with RAG for incorporating external biomedical knowledge
- Can also suggest drug candidates via SMILES and diffusion-based reasoning (demonstrative)

Example Output:

```
{
  "predicted_biomarker_region": "Residues A12–G14",
  "structural_motif": "hairpin loop",
  "function": "metal ion binding",
  "associated_disease": "Lung Adenocarcinoma (LUAD)",
  "potential_drugs": ["CCO", "CCN", "C1=CC=CC=C1"]
}
```

🧩 Input and Output (Agentic AI Workflow)

Input

An RNA Gene Sequence Expression (e.g., `8RBJ`)

Workflow Steps

1. Generates 3D structure of RNA gene
2. Biomarker detection to locate pockets present in the 3D structure
3. BioGPT to find suitable drugs for the detected pockets
4. Enhanced drugs for targeted and personalized therapy

Output Example

```
{
  "predicted_biomarker_region": "Residues A12–G14",
  "structural_motif": "hairpin loop",
  "function": "metal ion binding",
  "associated_disease": "Lung Adenocarcinoma (LUAD)",
  "potential_drugs": ["CCO", "CCN", "C1=CC=CC=C1"]
}
```

Stored in: `outputs/results.json`

```json
{
  "target": "luad_rna",
  "structure_summary": {
    "chains": ["A"],
    "total_residues": 46,
    "motifs_detected": []
  },
  "total_pockets_found": 3,
  "best_biomarker_site": {
    "Pocket_ID": "1",
    "Score": 0.043,
    "Druggability_Score": 0.0,
    "Number_of_Alpha_Spheres": 98.0,
    "Total_SASA": 349.937,
    "Polar_SASA": 281.102,
    "Apolar_SASA": 68.835,
    "Volume": 1620.036,
    "Mean_local_hydrophobic_density": 3.0,
    "Mean_alpha_sphere_radius": 4.284,
    "Mean_alp._sph._solvent_access": 0.557,
    "Apolar_alpha_sphere_proportion": 0.041,
    "Hydrophobicity_score": 4.083,
    "Volume_score": 3.333,
    "Polarity_score": 8.0,
    "Charge_score": 0.0,
    "Proportion_of_polar_atoms": 67.105,
    "Alpha_sphere_density": 7.494,
    "Cent._of_mass___Alpha_Sphere_max_dist": 19.006,
    "Flexibility": 0.0
  },
  "biomarkers": [
    {
      "pocket_id": "1",
      "predicted_region": "A12-A20",
      "motif_type": "stem-loop",
      "function": "Potential ligand-binding site",
      "associated_disease": "Lung carcinoma (hypothetical)",
      "recommended_drug_targets": [
        "Mg2+ pocket stabilizers",
        "RNA-ligand intercalating agents"
      ]
    },
    {
      "pocket_id": "2",
      "predicted_region": "A25-A30",
      "motif_type": "hairpin loop",
      "function": "Metal-binding site",
      "associated_disease": "Renal carcinoma (hypothetical)",
      "recommended_drug_targets": [
        "Metal ion chelators",
        "RNA folding stabilizers"
      ]
    },
    {
      "pocket_id": "3",
      "predicted_region": "A33-A38",
      "motif_type": "tetraloop",
      "function": "Ligand recognition",
      "associated_disease": "Lung carcinoma (hypothetical)",
      "recommended_drug_targets": [
        "RNA-ligand intercalating agents"
      ]
    }
  ]
}
```

🧩 Workflow Diagram

```
PDB Parser → fpocket → LangChain + RAG → BioGPT → Drug Discovery
(SMILES/Generative) → JSON Report
```

🧠 Agent Role Table

Agent Role	Function in RNA Hub	Implementation Strategy (The "How")
Planning & Supervisor Agent	Coordinates the overall workflow and decomposes tasks into biomarker analysis and drug suggestion sub-tasks	Implemented via BioGPT as the central reasoning engine with LangChain orchestration capability
Data Acquisition Agent	Collects and validates RNA structure and sequence data from PDB and RNAcentral	Uses RAG to incorporate external knowledge into BioGPT predictions
Structure & Analysis Agent	Predicts secondary/3D structure, detects pockets, and identifies potential biomarker regions	Combines PDB Parser, fpocket, and BioGPT reasoning to produce structured biomarker outputs
Drug Discovery Agent	Suggests potential small molecules or drug candidates targeting the biomarker regions	Uses SMILES representations and diffusion-based generative reasoning guided by BioGPT

📊 Example Output

```json
{
  "rna_file": "luad_rna.pdb",
  "pockets_detected": 3,
  "best_biomarker_site": {
    "region": "A12-G14",
    "motif": "loop",
    "binding": "metal ion",
    "disease_association": "LUAD",
    "potential_drugs": ["CCO", "CCN", "C1=CC=CC=C1"]
  }
}
```

🚀 Future Enhancements

- Integrate AlphaFold or RoseTTAFold RNA for de novo structure prediction
- Query PubChem/UniProt APIs for biological validation
- Extend to drug docking simulations for target validation
- Implement full agentic loops with LangChain to allow BioGPT to plan, execute, and refine predictions
- Visualize pockets, biomarkers, and suggested compounds in 3D (PyMOL/NGLview)

🧬 References

- Liu et al., 2023. *BioGPT: Generative Pre-trained Transformer for Biomedical Text Mining*
- Le Guilloux et al., 2009. *fpocket: An open-source platform for ligand pocket detection*
- Berman et al., 2000. *The Protein Data Bank: a database of macromolecular structures*
- LangChain documentation: https://www.langchain.com/
- RAG: Lewis et al., 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*