Datasets Required

1. Viral Genomic Datasets

- **Purpose:** For mutation prediction, evolutionary modeling, and genomic surveillance.
- **Data Needed:**
  - Viral genomes (DNA/RNA sequences).
  - Variant lineages (e.g., Alpha, Delta, Omicron for SARS-CoV-2).
  - Metadata (geographic location, collection date, patient background).
- **Formats:** FASTA, GenBank, JSON metadata.
- **Sources:**
  - INSACOG (India's genomic consortium, 10+ sequencing labs).
  - GISAID (global SARS-CoV-2 genomes).
  - NCBI Virus, EMBL-EBI Virus Pathogen DB.

2. Protein Structural Datasets

- **Purpose:** For protein folding, molecular dynamics, and docking.
- **Data Needed:**
  - Crystal and cryo-EM protein structures.
  - Mutant protein structures for known variants.
  - Receptor–ligand complexes (e.g., Spike–ACE2).
- **Formats:** PDB, mmCIF.
- **Sources:**
  - RCSB Protein Data Bank (PDB).
  - Indian Institute of Science (IISc) structural biology repository.
  - AlphaFold DB / OpenFold outputs.

3. Drug & Ligand Datasets

- **Purpose:** For docking, chemical modification, and generative drug design.
- **Data Needed:**
  - Small molecule inhibitors, antivirals.
  - Peptide libraries and antibody fragments.
  - Chemical properties, SMILES, 3D conformers.
  - ADMET profiles (Absorption, Distribution, Metabolism, Excretion, Toxicity).
- **Formats:** SMILES, MOL2, SDF.
- **Sources:**
  - ChEMBL (bioactive molecules).
  - ZINC database (commercial compounds).
  - PubChem (chemical structures + bioassays).

- Indian Pharmacopeia / ICMR drug trial data.

4. Multi-Omics Datasets

- **Purpose:** To model virus–host interactions, organ-level effects, and symptom mapping.
- **Data Needed:**
  - Transcriptomics (RNA-seq from infected tissues).
  - Proteomics (protein expression changes after infection).
  - Metabolomics (metabolic pathways altered by infection).
- **Formats:** FASTQ (raw sequencing), TSV/CSV (expression matrices).
- **Sources:**
  - GEO (Gene Expression Omnibus).
  - PRIDE (Proteomics database).
  - India-specific DBT-funded omics projects.-

5. Clinical Trial & Toxicity Datasets

- **Purpose:** For in-silico clinical trial simulation and drug safety prediction.
- **Data Needed:**
  - Phase I–IV clinical trial data.
  - Patient pharmacokinetic (PK) & pharmacodynamic (PD) profiles.
  - Reported side effects & toxicity pathways.
- **Formats:** CSV, JSON, HL7/FHIR (health data standards).
- **Sources:**
  - CTRI (Clinical Trials Registry of India).
  - ICMR trial datasets.
  - FDA/EMA open-access clinical trial datasets (for benchmarking).

6. Epidemiological & Outbreak Datasets

- **Purpose:** For SEIR-based outbreak forecasting and Deadliness Score validation.
- **Data Needed:**
  - Case incidence data (district/state level).
  - Reproduction number (R_0) estimates.
  - Contact tracing data.
  - Mortality & recovery statistics.
- **Formats:** CSV, JSON, API feeds.
- **Sources:**
  - MoHFW (India COVID-19 dashboard, epidemic bulletins).
  - WHO Athina API.
  - Covid19india.org (archival datasets).
  - National Health Mission reports.

7. Biomedical Ontologies & Knowledge Graph Datasets

- **Purpose:** For symptom mapping, preventive measure prediction, and translational insights.
- **Data Needed:**
  - Clinical ontologies (disease–symptom relationships).
  - Preventive guidelines mapping (WHO, ICMR).
- **Formats:** RDF, OWL, JSON-LD.
- **Sources:**
  - UMLS (Unified Medical Language System).
  - SNOMED CT India Edition.
  - MeSH ontology.