Appendix 1

Researcher Designed Scenarios for Five Standardised Tests to Evaluate AI-Driven No-Code Applications in Drug Discovery

Scenario 1: Target Identification

(Used to test the functionality of apps such as Insilico Medicine's Pharma.AI, Exscientia & BenevolentAI)

Objective: Evaluate the application's ability to identify potential drug targets based on disease-related proteins.

Description:

Disease Selection:

• Alzheimer's Disease: A neurodegenerative disorder characterised by the accumulation of amyloid-beta plaques and tau tangles in the brain.

• Breast Cancer: A common cancer type where mutations in certain genes, such as BRCA1 and HER2, play a critical role in the disease's progression.

Dataset Preparation:

Genomic and Proteomic Data:

For Alzheimer's Disease:

• Genomic Data: Whole-genome sequencing data from Alzheimer's patients, including known mutations in genes such as APP, PSEN1, and APOE.

• Proteomic Data: Protein expression profiles from Alzheimer's brain tissues, focusing on proteins like amyloid-beta (Aβ) and tau.

For Breast Cancer:

• Genomic Data: Whole-exome sequencing data from breast cancer patients, including mutations in *BRCA1*, *BRCA2*, and *TP53*.

• Proteomic Data: Protein expression profiles from tumor samples, with an emphasis on proteins like HER2 (ERBB2) and estrogen receptor (ER).

Task:

The application is to be tasked with analysing these genomic and proteomic datasets to identify novel drug targets:

• For Alzheimer's Disease: The application might be supposed to identify proteins involved in the amyloid-beta pathway or tau phosphorylation as potential drug targets.

• For Breast Cancer: The application could be expected to pinpoint mutations in PIK3CA or overexpression of HER2 as potential targets for therapeutic intervention.

Validation:

• Known Targets: The application's identified targets is to be compared against known targets, such as those listed in databases like DrugBank, OMIM, or the Therapeutic Target Database (TTD).

• Novel Targets: For novel targets, additional validation is to be performed using literature searches and experimental data to determine their relevance and potential for drug development.

Evaluation Metrics:

Accuracy: The proportion of correctly identified targets relative to known targets.

Novelty: The application's ability to identify previously unknown targets.

Speed: The time taken to process the datasets and generate results.

Scenario 2: Virtual Screening

(Used to test the functionality of apps such as Insilico Medicine's Pharma.AI & Atomwise)

Objective: Assess the application's capability to screen large chemical libraries for compounds with potential binding affinity to a target protein.

Description:

Target Protein Selection:

• EGFR (Epidermal Growth Factor Receptor): A well-known target in non-small cell lung cancer (NSCLC). Mutations in EGFR can lead to uncontrolled cell proliferation, making it a prime target for cancer therapies.

• BACE1 (Beta-Secretase 1): A target for Alzheimer's disease, involved in the production of amyloid-beta peptides, which aggregate to form plaques in the brain.

• SARS-CoV-2 Main Protease (Mpro): A critical enzyme in the life cycle of the SARS-CoV-2 virus, making it a target for COVID-19 antiviral drug development.

Chemical Library Selection:

• ZINC15 Database: A publicly available chemical library containing over 230 million purchasable compounds. For this scenario, a subset of 1,000,000 diverse compounds will be used for screening.

• ChEMBL: A database of bioactive drug-like small molecules with over 2 million compounds, widely used in drug discovery research.

• Enamine REAL Database: A chemical library containing over 3 billion synthetically accessible compounds, suitable for large-scale virtual screening projects.

Task:

• The application is to be tasked with virtually screening a library of 1,000,000 chemical compounds against a specific target protein.

• The virtual screening process involves docking the compounds into the binding site of the target protein and predicting their binding affinities.

• The output will be a ranked list of compounds, with those predicted to have the highest binding affinity at the top.

Validation:

• The top-ranked compounds are to be cross-referenced with known inhibitors or binders of the target protein from databases such as DrugBank, PubChem, and BindingDB.

• Experimentally validated binding affinities are to be compared to the predicted affinities to assess the accuracy of the virtual screening.

Evaluation Metrics:

Accuracy: The percentage of top-ranked compounds that exhibit experimentally validated binding affinity.

Efficiency: The computational resources and time required to complete the virtual screening.

Scalability: The application’s ability to handle larger libraries without significant performance degradation.