Organiczna praca

1. Title: (End of writting): Descriptive, catchy sentence that will trigger interest.

PROTO-NOOS: Orchestrating Open-Access Bioinformatics for Seamless Drug Discovery

2. Abstract:(single paragraph of about 200 words maximum). The less the better The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main.

Computational antibiotic discovery requires combining chemical generation, structural modelling, physical validation, biological interpretation, and synthetic feasibility into a reproducible workflow. This study presents the GUB pipeline, an in silico pipeline for prioritising small-molecule candidates in an *E. coli* DHFR context. The workflow used REINVENT4 for de novo molecule generation, RDKit-based physicochemical and Gram-negative entry filtering, Boltz2 for protein-ligand complex and affinity prediction, GROMACS for molecular dynamics stability checks, BioTransformer3 and COBRApy/iML1515 for metabolite and target-perturbation analysis, and AiZynthFinder for retrosynthetic accessibility. The pipeline produced structured intermediate outputs across stages, allowing candidates to be ranked by multiple weak signals rather than a single affinity score. The highest-priority systems generated by this pipeline are intended to be transferred into a downstream GUB_biocomplex_analysis workflow, where selected protein-ligand-cofactor complexes are examined in greater structural detail. In this companion analysis, the pipeline-derived candidates are not treated as confirmed hits, but as prioritized systems for deeper pocket-focused MD, contact, MM-GBSA, PLUMED metadynamics, and QM/DFT inspection. The main finding is that this integrated workflow can connect chemical, structural, MD, metabolic, target-level, and retrosynthetic evidence in one reproducible screening process. However, several outputs, especially short-MD kinetic proxies and systems-biology labels, should be interpreted as computational ranking features rather than confirmed biological activity. The study demonstrates a practical framework for hypothesis generation and dataset construction, but it does not include synthesis, antibacterial assays, toxicity testing, or wet-lab validation. Therefore, the conclusions are limited to in silico prioritisation and workflow feasibility.

2.1 Background: place the question to be addressed in a broad context and highlight the purpose of the study

Computational antibiotic discovery requires workflows that combine molecular generation, structural modelling, physical checks, biological interpretation, and synthetic feasibility. This study asks whether existing open or accessible tools can be connected into a reproducible in silico pipeline for prioritising small molecules in an *E. coli* DHFR context.

2.2 Do not overstate or overgeneralize your contribution.

The contribution is a practical integration workflow, not a confirmed antibiotic discovery. The study does not claim experimental activity, clinical relevance, or validated toxicity/safety.

2.3 Methods: describe 5 briefly the main methods or treatments applied

The pipeline used REINVENT4 for de novo generation, RDKit-based filtering for physicochemical and Gram-negative entry features, Boltz2 for protein-ligand complex and affinity prediction, GROMACS for molecular dynamics stability checks, BioTransformer3/COBRApy/iML1515 for metabolism and target-perturbation analysis, and AiZynthFinder for retrosynthetic accessibility.

2.4 Results: summarize the article's main 6 findings

The workflow produced structured outputs across chemical, structural, MD, metabolic, target-level, and retrosynthetic stages. It enabled multi-signal ranking rather than relying on affinity alone, and exposed where some signals were weak or should be interpreted cautiously.

2.5 Conclusions: indicate the main conclusions or interpretations.

The GUB pipeline is useful for in silico hypothesis generation, candidate prioritisation, and synthetic dataset construction. Its outputs remain computational weak labels until validated by synthesis and wet-lab antibacterial assays.

3. Introduction

3.1 Universe definition (Be very careful to define it sharp enough).

The universe of this study is the in silico search for small-molecule candidates that could be useful against an *E. coli* DHFR-related antibacterial target. More specifically, we focus on SMILES-represented compounds that can be generated de novo or taken from existing molecular datasets, then ranked by a multi-stage computational pipeline.

This universe is narrower than "all antibiotic discovery". We do not study clinical efficacy, broad-spectrum activity, toxicity in humans, or confirmed MIC values. We study the computational prioritisation problem: how to search chemical space in a more informed way by combining chemical validity, drug-like properties, Gram-negative entry heuristics, receptor binding,

molecular dynamics stability, metabolic interpretation, target perturbation, and synthetic accessibility.

The practical goal is to make this process reproducible and scalable through an API-like pipeline. Instead of using each tool separately, the study connects existing tools into one workflow that can process many compounds and return structured outputs useful for ranking, analysis, and future machine-learning datasets.

3.2 Most relevant achievements (in this very specif topic).

The main achievement of this work is the integration of compound search with biological interpretation. The pipeline connects de novo generation, physicochemical filtering, retention prediction, Boltz2 complex and affinity prediction, GROMACS molecular dynamics, BioTransformer metabolite prediction, COBRApy metabolic modelling, and AiZynthFinder retrosynthesis.

This makes it possible to inspect candidates from several perspectives at once. A compound is not ranked only by a single affinity score. It can also be checked for predicted stability in a protein-ligand complex, predicted bacterial entry, possible metabolic products, target-level growth effects, and synthetic accessibility.

The systems-biology part is especially important. We used metabolite analysis and target perturbation logic, including gene knockout-style simulations in two variants. This adds biological context to what would otherwise be only a structural or chemical screen. The result is a richer set of synthetic labels that can later support machine-learning models trained on pipeline-derived evidence.

3.3 What other people did, and why they fail (Be diplomatic)

Many existing studies and tools solve important parts of this problem. Generative molecular design tools can create new compounds. Structure-prediction tools such as Boltz2 can estimate protein-ligand complexes and affinity-related signals. Molecular dynamics packages such as GROMACS can test physical stability. Retrosynthesis tools can estimate whether a compound may be synthetically accessible. Systems-biology tools can simulate metabolic effects.

The limitation is that these tools are often used separately. A molecule may look promising by one metric, for example affinity, but fail because it is unstable, difficult to synthesise, unlikely to enter Gram-negative bacteria, metabolically irrelevant, or biologically bypassed by the cell. This does not mean the existing tools fail; rather, single-tool workflows can miss conflicts between chemical, physical, and biological constraints.

Commercial and larger integrated platforms address some of these issues, but they are often expensive, closed, or difficult to reproduce. Open tools exist, but combining them into a robust, inspectable, high-throughput pipeline still requires substantial engineering. This study

contributes in that direction, although it is not a full systematic benchmark of all available platforms.

3.4 What we did in this study (From a conceptual point of view).

Conceptually, we built a modular in silico discovery pipeline for candidate prioritisation. We used existing tools as specialised components and connected them through structured intermediate files and stage contracts.

The pipeline was used to explore compounds against an *E. coli* DHFR context, with 6XG5 as the receptor structure and folA as the biological target gene. The same design allows the receptor and target to be changed, making the workflow adaptable to other biological targets.

The study also investigates tool bias and signal agreement. By collecting outputs from chemistry, structure prediction, MD, metabolism, target perturbation, and retrosynthesis, the pipeline allows us to see where different methods agree, where they conflict, and where a score should be treated only as a weak label. A second purpose of the workflow is to create inputs for a deeper downstream analysis layer. The broad GUB pipeline performs candidate generation, filtering, ranking, and first-pass physical and biological annotation. The most promising systems from this stage are then passed to the GUB_biocomplex_analysis workflow, where individual protein-ligand-cofactor complexes can be analysed with higher structural resolution. This downstream analysis focuses on whether the predicted binding mode remains geometrically plausible, whether key pocket contacts are persistent, whether the ligand interacts productively with the DHFR/NADPH context, and whether additional energetic or QM descriptors support the original ranking.

The pipeline separates structural affinity prediction from cellular target accessibility by inserting a mechanistic ODE-based Cell Target Engagement layer, which models compound entry, efflux, and binding dynamics inside the bacterial cell rather than treating KD alone as a proxy for biological effect.

3.5 What we did not do.

We did not perform chemical synthesis, MIC testing, target-engagement assays, cytotoxicity assays, or any wet-lab validation. Therefore, the results should not be interpreted as confirmed antibacterial activity.

We also did not close the full design-make-test-analyse loop. In an ideal setting, selected de novo candidates would be synthesised or purchased, tested experimentally, and then used to retrain or calibrate the pipeline. That was outside the scope of this work.

All current claims are therefore computational and hypothesis-generating.

3.6 Why this is imporatant, for us and other people.

This work is important because it makes a complex drug-discovery-style workflow more accessible. Groups without expensive commercial platforms can still combine open or accessible tools into a structured pipeline for early-stage candidate prioritisation.

It is also useful as a step toward future API-driven and agent-assisted scientific workflows. As more tools expose programmable interfaces, pipelines like this can become reusable research infrastructure rather than one-off scripts.

Finally, the project includes supporting modules for error correction and analysis. These make the pipeline more practical: errors can be detected earlier, outputs can be checked more consistently, and results can be converted faster into interpretable tables, rankings, and figures.

4. Method

4.1 Describe fully, all the tech. details that will possible to others replicate your own experiments.

The experiments were performed with the GUB pipeline, an in silico multi-stage workflow for prioritising potential antibacterial small molecules against an *E. coli* DHFR context. The target protein/receptor used in the pipeline was PDB 6XG5, and the biological target gene used in the systems-biology stages was folA (b0048 in the *E. coli* iML1515 model).

The de novo branch started with REINVENT4. REINVENT4 was used to generate new SMILES candidates in a receptor-focused mode, using the receptor context rather than a wet-lab seed series. Generated molecules were post-processed to remove invalid SMILES, canonicalise structures, deduplicate records, and keep only valid molecules for downstream stages. In the representative de novo run, 100 compounds were passed through the downstream pipeline.

After generation, molecules were processed with an RDKit-based analysis stage. This stage calculated physicochemical descriptors including molecular weight, cLogP, hydrogen-bond donors and acceptors, TPSA, rotatable bonds, ring count, fraction sp3, heavy atom count, and heteroatom count. The same stage applied drug-likeness and bacterial-entry filters, including Lipinski rules, Veber rules, Gram-negative entry heuristics, PAINS/custom structural alerts, scaffold analysis, and Morgan fingerprint similarity logic. The configured Lipinski limits were MW <= 500, logP <= 5, HBD <= 5, HBA <= 10, with at most one violation. Veber filtering used TPSA <= 140 and rotatable bonds <= 10.

GROMACS was used as the molecular dynamics engine to check whether the Boltz2-predicted complexes remained physically plausible over short MD simulations. The GROMACS workflow prepared protein-ligand systems, generated ligand structures from SMILES where needed, parameterised ligands using OpenBabel/ACPYPE/AmberTools, and ran energy minimisation, NVT equilibration, NPT equilibration, and production MD. The production configuration available in the pipeline defines a 10 ns MD protocol at 300 K and 1 bar, with 2 fs timestep, PME

electrostatics, Verlet cut-off scheme, LINCS constraints on hydrogen bonds, V-rescale temperature coupling, and Parrinello-Rahman pressure coupling. For high-throughput screening runs, the fast MD mode used 0.1 ns production MD; these outputs were treated only as weak stability proxies, not as definitive binding kinetics.

The MD stage extracted stability-related metrics such as ligand RMSD, energy summaries, hydrogen-bond occupancy, stability score, residence-time proxy, and koff_proxy. Because many screening runs used fast MD, MD-derived kinetic proxies were interpreted conservatively and were given low confidence.

The systems-biology stages used COBRApy with the genome-scale *E. coli* model iML1515. Stage 6A used BioTransformer3 to predict possible xenobiotic/metabolite transformations, mapped predicted metabolites through MetaNetX/BiGG identifiers into the metabolic model, and compared baseline and perturbed FBA behaviour. Stage 6B simulated target-level inhibition/knockout logic against the folA/DHFR context and evaluated growth-related outputs, including inhibited growth ratio. The configured Stage 6B concentration scenarios were 1.0 uM and 50.0 uM.

A final scoring stage combined available evidence into a ranking rather than a binary activity claim. The score included structural affinity, MD stability, physicochemical/entry features, and systems-level signals. Where data were missing or low confidence, the pipeline recorded those conditions explicitly rather than treating them as confirmed negative or positive biological results.

For de novo candidates that reached the final ranking, AiZynthFinder was used as a retrosynthetic accessibility screen. AiZynthFinder was configured with a USPTO policy, a maximum route depth of 5, and a 300 s time limit per compound. The output recorded whether a route was solved, the number of retrosynthetic steps, route score, and precursor list. This step was used as a practical synthetic feasibility filter, not as proof that the proposed route is optimal. AiZynthFinder is known to be limited by stock choice, policy data, and template coverage; therefore, future work should compare it with more complete computer-aided synthesis planning systems such as ASKCOS, which integrates multiple one-step, multi-step, and feasibility models and can use richer precursor databases.

Stage contracts are additionally protected by an execution integrity layer that validates outputs at each boundary, enforces yield thresholds, and halts the pipeline on critical artefact inconsistencies rather than allowing corrupted intermediate data to propagate silently to downstream scoring stages.

4.2 Cell target engagement modelling

The pipeline includes an ODE-based Cell Target Engagement (CellTE) module that models drug–target binding dynamics inside a Gram-negative bacterial cell. The model tracks two state

variables, free intracellular drug concentration (C_free) and target-bound drug concentration (C_bound), governed by four coupled processes: passive entry driven by the external concentration C_out, efflux removal, target association, and target dissociation. The system is:

```
dCfree/dt  = Pin·Cout − Kefflux·Cfree − kon·Cfree·(Bmax − Cbound) +
koff_int·Cbound
dCbound/dt = kon·Cfree·(Bmax − Cbound) − koff_int·Cbound
```

Physicochemical descriptors are mapped onto ODE parameters via param_maps.py: entry_score determines Pin through an exponential scaling, retention_proxy determines Kefflux through a power-law retention term, and KD_pred (post-Boltz2) determines koff_int via the relationship koff_int = KD_pred × kon.

The module operates in two modes. In the pre-Boltz2 pass, KD_pred is unavailable; kon and koff_int are set to biological priors (1×10^6 M^-1 s^-1 and 1×10^-3 s^-1 respectively), and ranking reflects primarily entry and efflux properties. In the post-Boltz2 rescore pass, Boltz2-derived KD_pred replaces the prior, so the ODE integrates predicted affinity with accumulation dynamics. The module outputs Cout50 (external concentration required for 50% occupancy), AUCbound (integral of C_bound over time), apparent koff_app from washout simulation, rebinding_penalty (ratio koff_app/koff_int), and a composite score_cellTE used for ranking.

This layer is important because KD alone does not determine intracellular target occupancy. A compound with favourable predicted affinity but poor membrane entry or high efflux susceptibility may achieve low C_bound, while a compound with moderate affinity but strong accumulation may outperform it cellularly.

4.3 What software, hardware, strains (biology), chemicals, etc. What constrains you used. uzylismy reinvent4, boltz2, gromacs, biotransformer

The main software components were REINVENT4 for de novo generation, RDKit for chemical standardisation/descriptors/filtering, Boltz2 for complex and affinity prediction, GROMACS for molecular dynamics, BioTransformer3 for metabolism prediction, COBRApy for FBA on iML1515, and AiZynthFinder for retrosynthetic planning. Supporting tools included OpenBabel, ACPYPE, AmberTools, OpenMM, PDBFixer, pandas, NumPy, SciPy, and Python stage runners with CSV-based contracts between pipeline stages.

The computational environment was organised through separate Conda environments for major modules: reinvent4, boltz2, gromacs_gpu_tools, biosys, cellte, molecule, analiza, and aizynthfinder. The Boltz2 environment used Python 3.11; the systems-biology environment used Python 3.10 with COBRApy, RDKit, pandas, NumPy, SciPy, matplotlib, Java/OpenJDK 17,

and BioTransformer3. The GROMACS GPU environment included Python 3.11, OpenBabel, ACPYPE, AmberTools, RDKit, pandas, and GROMACS GPU tooling.

Computations were designed for the PUT Slurm cluster. GPU work was assigned mainly to ngpu or hgx partitions. The ngpu nodes provide RTX 4080/4090 GPUs, while hgx nodes provide NVIDIA A100 80 GB GPUs. CPU-heavy systems-biology work was assigned to the obl partition. The pipeline used /raid scratch storage for heavy intermediate files and /home for code, configuration, and final outputs.

No wet-lab bacterial strain was cultured in this work. The biological system was represented computationally by the *E. coli* K-12 MG1655 iML1515 genome-scale metabolic model. No physical compounds were synthesised or purchased as part of this experiment. The chemical inputs were generated de novo or read from public/curated molecular datasets, and the experimental branch used existing activity records rather than new measurements.

The main constraint was the lack of laboratory validation. Therefore, the pipeline could prioritise and de-risk candidates computationally, but it could not confirm antibacterial activity, toxicity, MIC values, target engagement, or real synthetic success. With access to wet-lab testing, the de novo candidates could be synthesised or purchased, tested against *E. coli*, and used to create a larger project-specific dataset for retraining and calibrating the scoring model.

The pipeline includes a defensive execution integrity layer designed to prevent silent error propagation across stages. This layer performs argument validation before execution, enforces explicit Conda environment isolation per stage, applies retry with exponential backoff for fragile external tools (particularly Boltz-2), and generates a run-level success or failure report with per-stage metrics. At Stage 1, molecular yield is checked against a minimum threshold before downstream stages proceed. At Stage 2, the retention model artefact is resolved canonically rather than selected heuristically. At Stage 4, a resume mechanism detects previously completed predictions, skips recomputation for finished compounds, prunes stale results from prior runs, deduplicates per compound_id, and enforces hard coverage checks for KD_pred and CIF file existence. If any critical check fails, the pipeline halts with a non-zero exit code rather than propagating an incomplete result. This design follows the principle that a loud execution failure is scientifically preferable to a silent semantic error that produces plausible-looking but corrupted output.

4.4 Downstream biocomplex analysis of prioritized systems

The output of the GUB pipeline was designed to serve as the input layer for a downstream GUB_biocomplex_analysis workflow. In this second workflow, only selected high-priority systems are analysed, rather than the full generated population. The selection is based on the pipeline ranking, with preference given to compounds that combine favourable predicted affinity, acceptable physicochemical properties, plausible MD behaviour, systems-level relevance, and synthetic accessibility.

For each selected system, the downstream workflow uses the prepared protein-ligand-cofactor complex as a structural object for deeper analysis. The current implementation is pocket-focused: a reference system and reference trajectory frame define a frozen binding-pocket description, including strict and extended pocket residues, NADPH-associated features, and key distances relevant to the DHFR catalytic context. This definition is then reused across systems to make comparisons less dependent on ad hoc residue selection.

The downstream analysis includes preflight validation, trajectory quality control, ligand and pocket RMSD, NADPH-fragment stability, pocket flexibility, contact persistence, hydrogen bonding, hydrophobic and aromatic contacts, sigma-hole interactions, water bridges, and unsatisfied buried polar atoms. It also includes MM-GBSA/MM-PBSA-style binding-energy estimates, short PLUMED-based metadynamics as a local probe of pocket escape or bound-state persistence, and CP2K DFT single-point calculations on selected pocket regions. These calculations are used as comparative descriptors, not as definitive binding free energies or experimental validation.

This separation gives the project a two-resolution design. The GUB pipeline performs broad triage across many compounds, while GUB_biocomplex_analysis performs deeper mechanistic inspection of the most promising systems. Therefore, data generated by the pipeline are reused here as structured starting points for more expensive methods, allowing computational effort to scale with candidate priority.

5. Results

5.1 Present plots, and tables. With a unifiying narrative why you did that.

The overarching logic of the Results section is to follow the compound from its statistical/chemical origin (Stage 1) through increasing layers of biophysical and systems-level filtering (Stages 2–7), and finally through synthetic accessibility assessment (Stage 8). Each subsection motivates its visualisation before showing it.

5.1.1 De Novo Generation — Chemical Space and Scaffold Diversity (Stage 1)

Why this plot? REINVENT generates molecules by sampling a learned chemical prior. Before trusting any downstream metric we need to verify that the generated set actually covers diverse chemical space and does not collapse onto a single chemotype.

Planned visualisations:

- **UMAP / t-SNE** on Morgan fingerprints (radius 2, 1024 bits) of the 100 generated compounds overlaid on the training set (ChEMBL, BindingDB, Stokes antibiotics set). This shows whether REINVENT sampled inside or outside the known chemical space — a prerequisite for out-of-distribution (OOD) analysis.
- **Scaffold diversity bar chart** using Bemis-Murcko scaffolds: number of unique scaffolds per 100 compounds. For reference, a fully diverse set of 100 compounds would have 100 distinct scaffolds; a collapsed set would have 1–5.
- **~5 fingerprint types comparison** (Morgan, MACCS, RDKit topological, ECFP4, atom-pair): compute pairwise Tanimoto distance matrices and compare their internal cluster structure with a heatmap or violin plot. This checks whether diversity conclusions are fingerprint-dependent.
- **Wasserstein distance** between the QED, molecular weight, and logP distributions of generated vs. training compounds. All 100 generated compounds pass Lipinski and Veber rules (confirmed in Stage 2a), but Wasserstein distance quantifies how far the property distributions are displaced from the training distribution.

What the data already shows (Stage 1 output): QED of generated compounds spans 0.35–0.95, with selection methods mixing `highest_score` and `score_rank`. All compounds were labelled `compound_origin = reinvent`, confirming no experimental scaffolds leaked in. QED is the only Stage 1 scoring metric available, confirming the model was rewarded for drug-likeness during generation.

Source	n compounds	Lipinski pass	Veber pass
REINVENT (de novo)	100	100 (100%)	100 (100%)

5.1.2 Structural Filtering — ADMET Alerts and Drug-likeness (Stage 2a/2b)

Why this plot? Alert-based filters are a fast, rule-based sanity check. Visualising the filter funnel shows the attrition rate before any expensive computation is run.

Planned visualisations:

- **Funnel plot**: Stage 1 (100) → Stage 2a pass (100) → Stage 2b pass (100) → Stage 3 (100). In this run, no compound was rejected. A bar chart of individual alert types (PAINS, Ames, Lipinski, Veber, Brenk) — even if all pass — tells reviewers what risk categories were checked.
- **QED vs. molecular weight scatter plot** coloured by Lipinski pass/fail. Since all 100 pass, the plot illustrates that REINVENT has internalised the Lipinski constraint.

Current data: 0 compounds rejected at Stage 2a. All carry the flag `lipinski_pass=True; veber_pass=True`. This warrants a brief Discussion note (see §6).

5.1.3 WL and WL-OA Graph Kernel Analysis

Why this plot? Morgan fingerprints are molecular graph hashes; graph kernels (Weisfeiler-Lehman, WL-OA) operate directly on the molecular graph and capture higher-order structural patterns, including ring topology, that circular fingerprints can miss.

Planned visualisations:

- **WL kernel Gram matrix heatmap**: a 100×100 similarity matrix of the generated set. Clusters in this matrix correspond to chemically related scaffolds.
- **WL-OA kernel vs. Morgan fingerprint rank correlation**: for each pair of compounds, plot WL similarity vs. Tanimoto similarity. Divergence reveals cases where graph topology and atom-environment fingerprints disagree — these are structurally unusual compounds worth examining.
- **Wasserstein distance** between WL kernel distributions of generated vs. training set: quantifies structural novelty at the graph-topology level.

5.1.3b Cell Target Engagement Scoring

- Rozkład score_cellTE pre-Boltz2 dla wszystkich 100 związków (histogram)
- Porównanie rankingu pre-Boltz2 vs. post-Boltz2 (scatter rank vs. rank)
- Najważniejsza tabela: top 10 związków z Cout50 i AUCbound z post-Boltz2 rescore
- Jedno zdanie interpretacyjne: ile związków zmienia pozycję po recore'owaniu KD_pred z Boltz2

5.1.4 Docking Score Distribution (Stage 4 — Boltz2)

Why this plot? Docking is the first physics-based filter. Its score distribution tells us whether the generative model has learned to place atoms in geometries consistent with the target binding site.

Planned visualisations:

- **Histogram of docking scores** (`d_docking_score` in final output, normalised 0–1). The range observed is ~0.62–0.99, with most compounds clustering near 0.75–0.98 (see final_output.csv, top 30 rows).

- **Docking score vs. QED scatter**: does the generative model trade drug-likeness for docking affinity?
- **Top-10 compounds 3D binding pose visualisation** (if Boltz2 pose outputs are available in work directory).

Statistic	Value
Stage 4 runtime	2684 s (~45 min)
% of total pipeline runtime	33.0%
All 100 compounds docked	Yes

5.1.5 Molecular Dynamics — Binding Stability (Stage 5 — GROMACS)

Why this plot? MD provides kinetic proxies (residence time, k_off, H-bond occupancy, RMSD) that docking alone cannot. This is the most computationally expensive stage (46.6% of total runtime, 3795 s).

Key findings from data:

- **98/100 compounds flagged as** `unstable`, 2 as `converged` (REINVENT_EC569C45, based on `md_quality:converged` flag).
- All 100 compounds received `md_confidence = low`, because the MD was run at 100 ps (`gate_only` mode). The pipeline explicitly notes: *"MD-derived kinetic proxies for md_confidence=low should be treated as indicative only (md_weight=0.15)"*.
- `residence_time` uniformly 0.1 (lower bound clamp for short MD). `koff_proxy` values are in the range 1.1–1.2×10^{-3}.
- `hbond_occupancy = 0` for 98/100 compounds; 2 compounds show partial H-bond occupancy (~0.82).

Planned visualisations:

- **RMSD violin plot** (backbone first-half vs. second-half mean, and ligand RMSD) for all 100 compounds. Distributions of `rmsd_backbone_half_mean_delta` (range: ~0.018–0.045 nm) and `ligand_rmsd_mean` (range: 0.48–6.45 Å — high variance indicating unstable poses).
- **Energy trace boxplot**: `energy_potential_mean` vs. compound ID (sorted by GUB_rank). Are top-ranked compounds energetically more favourable?
- **KD_pred distribution histogram** (range observed: ~1.3×10^{-6} to ~1.2×10^{-4} M). The top compound (REINVENT_DDA288DA) shows KD_pred ~1.3 μM, suggesting moderate predicted affinity.

MD Quality	Count	%
Converged	2	2%
Unstable	98	98%
md_confidence = low (100 ps)	100	100%

5.1.6 Metabolic Network Analysis — FBA with COBRApy (Stage 6a)

Why this plot? Flux Balance Analysis (FBA) maps compound metabolites onto the *E. coli* metabolic network and tests whether the compound perturbs bacterial growth flux. This is where the biological mechanism of action is predicted.

Key findings:

Reason Code	Count	%
`no_parent_or_product_mapping`	66	66%
`integrated_but_no_flux_change`	25	25%
`mapped_but_not_integrated`	8	8%
`discriminative`	1	1%

Only **1/100 compounds** was truly discriminative in FBA — meaning only one compound produced a detectable change in growth-relevant fluxes. 66 compounds had no metabolite mapping at all, meaning the metabolic network database (MetaNetX / MNX) does not contain their parent or any metabolic product.

Planned visualisations:

- **Stacked bar** of Stage 6a reason codes.
- **FBA flux difference magnitude** (`fba_flux_difference_magnitude`) for the 34 compounds with partial integration: scatter vs. docking score and GUB_score.
- **Top perturbed reactions** (`top_perturbed_reactions`) for the 1 discriminative compound: visualisation as a metabolic pathway diagram.
- **Dynamic FBA biomass curves** (`dfba_final_biomass`) for compounds that integrated into the model — comparing the 1 discriminative compound vs. `integrated_but_no_flux_change` compounds.

5.1.7 Growth Inhibition Ratio — Synthetic Lethality and varB (Stage 6b)

Why this plot? The `inhibited_growth_ratio` (IGR) from FBA-based synthetic lethality predicts the degree to which targeting a given gene/reaction kills the bacterium.

Key findings:

- **197/200 compound-target pairs were discriminative** (`discriminative_true`).
- IGR values are predominantly high (median ~0.93), indicating strong predicted growth inhibition for most compound-target combinations.
- The `varB` and `synleth` modes produce **identical IGR values**, suggesting convergence of the two synthetic lethality scoring approaches for this target.
- Exceptions: 3 compounds with `discriminative=False` all show IGR > 0.99 (REINVENT_E42C9F26 IGR=0.991, REINVENT_O=C(O)C1CC... IGR=0.992, REINVENT_O=C(O)C1CSC... IGR=0.993) — these are classified non-discriminative despite high IGR, likely due to a threshold artefact.

Planned visualisations:

- **IGR distribution histogram** across all 100 compounds, with a vertical line at the discrimination threshold.
- **IGR vs. GUB_score scatter**: do high-IGR compounds also rank highly by the composite score?
- **Heatmap of top 20 compounds × scoring dimensions** (KD_pred, docking, md_score, penetration_PC1, qed, systems_score): to visualise multi-criteria trade-offs.

5.1.8 Multi-Criteria GUB Score and Pareto Ranking (Stage 7)

Why this plot? The GUB score is a desirability-weighted composite of all evidence blocks. Visualising Pareto fronts and rank stability reveals which compounds dominate across all criteria vs. which are specialist performers.

Key findings (top 10 compounds):

Rank	Compound ID	GUB Score	Pareto Rank	Classification	Penetration	d_dockir
1	REINVENT_DDA288DA	0.387	7	bacteriostatic	0.300	0.9999
2	REINVENT_C9FA3838	0.347	1	partial	1.000	0.9999
3	REINVENT_00410F32	0.337	2	partial	0.864	0.9999

Rank	Compound ID	GUB Score	Pareto Rank	Classification	Penetration	d_dockir
4	REINVENT_6881EED3	0.334	4	partial	0.782	0.9999
5	REINVENT_2F198768	0.329	4	partial	0.745	0.9999

- GUB scores drop sharply after rank ~27; from rank 28 onward, all compounds score 0.0 (classified as `resistant`). This is driven by the `d_growth_ratio` component collapsing to 0 for resistant compounds — the pipeline's way of hard-filtering compounds with no predicted bactericidal effect.
- `rank_stability_score = 0.996` for all compounds: the ranking is essentially stable across weight perturbations (low rank_std for most top compounds).
- `evidence_completeness_score = 0.842` (systems block missing) for 88 compounds; `1.000` for 12 compounds that also have systems-level data.

Planned visualisations:

- **Pareto front scatter plot** (GUB_score vs. penetration_score, coloured by classification).
- **Desirability weight bar chart** from `desirability_params.json` showing the contribution of each evidence block.
- **PCA biplot** of scoring dimensions from `pca_params.json`, with compounds projected onto PC1/PC2.
- **Sensitivity analysis heatmap** from `sensitivity_summary.json`: how rank changes under weight perturbation.

5.1.9 Retrosynthetic Feasibility — AiZynthFinder (Stage 8)

Why this plot? A computationally optimal compound is of limited value if it cannot be synthesised. Stage 8 assesses each compound's synthetic accessibility using AiZynthFinder's MCTS retrosynthesis.

Key findings (batch_001, top 10 compounds):

Compound	is_solved	retro_steps	route_score	stock_coverage	wall_
REINVENT_DDA288DA (Rank 1)	False	5	0.726	0.304	8.7
REINVENT_C9FA3838 (Rank 2)	False	3	0.670	0.333	18.9

Compound	is_solved	retro_steps	route_score	stock_coverage	wall_
REINVENT_00410F32 (Rank 3)	**True**	3	0.987	0.800	9.1
REINVENT_6881EED3 (Rank 4)	**True**	4	0.975	0.800	11.9
REINVENT_1C07AD84 (Rank 8)	**True**	3	0.987	0.778	14.8
REINVENT_6FF45398 (Rank 9)	**True**	1	0.998	1.000	2.9

- The top-ranked compound (GUB Rank 1) has `is_solved = False` and low stock coverage (30%), meaning its building blocks are not commercially available. This represents a direct conflict between activity prediction and synthetic accessibility.
- Rank 3 and Rank 4 are solved with high route scores (0.987, 0.975) and good stock coverage (80%), making them the most actionable leads.
- REINVENT_6FF45398 (Rank 9) has the simplest synthesis: 1 retrosynthetic step, 100% stock coverage, solved in 2.9 s.

Planned visualisations:

- **Bar chart of `is_solved` rate across all batches** (batches 001–055+ present in the data directory, representing multiple seeds and runs). This gives a population-level synthesis feasibility estimate.
- **route_score vs. GUB_score scatter** (the key actionability plot): compounds in the top-right quadrant are both biologically promising and synthetically accessible.
- **retro_steps distribution histogram**.
- **stock_coverage violin plot** by classification (bacteriostatic / partial / resistant).

5.1.10 Transfer of prioritized systems to GUB_biocomplex_analysis

The ranked GUB pipeline output defines which systems should enter the deeper biocomplex-analysis layer. This is important because methods such as MM-GBSA, PLUMED metadynamics, and CP2K single-point calculations are too expensive to apply blindly to every generated molecule. The pipeline therefore acts as a prioritisation engine: it reduces the chemical search space to a smaller set of protein-ligand systems that justify more detailed structural and energetic analysis.

In the downstream GUB_biocomplex_analysis workflow, the selected systems are compared using pocket-level descriptors rather than only compound-level scores. The analysis records ligand RMSD, pocket RMSD, NADPH-fragment stability, persistent contacts, hydrogen bonds, hydrophobic and aromatic interactions, sigma-hole contacts, MM-GBSA estimates, PLUMED-

derived local free-energy descriptors, and QM-derived charge or bond-order features. This makes it possible to ask whether a high pipeline rank is supported by a coherent binding-mode explanation.

The expected outcome is not a new claim of antibacterial activity, but a more interpretable computational triage. A compound that scores well in the broad pipeline but fails pocket stability, contact persistence, or synthesis feasibility can be deprioritised. Conversely, a compound with moderate global score but strong, reproducible pocket-level evidence may be selected for longer simulation or experimental follow-up.

5.1.11 Pipeline Runtime and Scalability

Stage	Runtime (s)	% of Total	s / compound
Stage 5 (MD/GROMACS)	3795	46.6%	37.95
Stage 4 (Docking/Boltz2)	2684	33.0%	26.84
Stage 6a (FBA/COBRApy)	1133	13.9%	11.33
Stage 2a (ADMET)	338	4.2%	3.38
Stage 6b (IGR)	109	1.3%	1.09
Stage 1 (Generation)	62	0.8%	0.62
Others (2b, 3×)	19	0.2%	<0.2
Total	**8140**	**100%**	**81.4**

Linear scaling projections:

- 100 compounds: **~2.3 hours** (observed)
- 1000 compounds: **~22.6 hours**

Planned visualisations:

- **Horizontal stacked bar** of per-stage runtime contribution.
- **Log-scale scaling projection plot** (10 → 10,000 compounds).

5.1.12 Additional Proposed Analyses

The following three contexts strengthen the biological relevance of the pipeline but were not executed in the current run:

Cell phenotype guiding: Morphological cell profiling data (e.g., Cell Painting assays) can be used to score generated compounds against a desired phenotypic signature. Compounds whose predicted features match the phenotypic profile of known antibiotics (e.g., membrane-disrupting agents) could serve as a complementary filter at Stage 2 or Stage 7. This is especially valuable because it is mechanism-agnostic: compounds can be selected even when the molecular target is unknown.

Transcriptomics guideline: Gene expression signatures (e.g., from GEO/LINCS L1000 database) of *E. coli* treated with known antibiotics can be compared against the FBA-predicted affected pathway sets from Stage 6a. A compound whose FBA perturbation signature resembles the transcriptomic signature of a known effective antibiotic is a stronger mechanistic candidate. Practically, this would add a cosine similarity score between the `top_perturbed_reactions` vector and pre-computed antibiotic transcriptomics signatures.

Docking guideline (structure-based context): The Boltz2 docking in Stage 4 currently produces a single scalar score. A docking guideline would establish reference binding poses from crystal structures (e.g., from PDB) for the target, and compare the generated compound's predicted binding mode to this reference using RMSD of key interacting residues. This would distinguish compounds that dock to the canonical binding site from those that score well by binding elsewhere.

6. Discussion

6.1 Address first, the most obvious relevant data.

The composite GUB score identifies a small set of actionable leads from a large generated set. Of 100 de novo generated compounds, 27 receive a non-zero GUB score (ranked 1–27), all classified as `partial` or `bacteriostatic` inhibitors. The remaining 73 are classified as `resistant` and are assigned a GUB score of 0, meaning the multi-criteria filter is decisive. The top compound (REINVENT_DDA288DA, GUB=0.387) shows predicted KD ~1.3 μM with a moderate docking score, moderate membrane penetration, and bacteriostatic classification.

The CellTE layer adds mechanistic resolution that KD alone cannot provide. Two compounds in the top 10 share nearly identical KD_pred values (~1.3 μM) but differ in penetration score (1.000 vs. 0.300), which is reflected in their Cout50 and AUCbound after the post-Boltz2 rescore. This illustrates the intended function of the ODE: to distinguish compounds that fail because of poor entry or efflux, rather than because of weak binding. Under current pipeline conditions, the CellTE parameters are estimated from descriptor-derived proxies and remain uncalibrated against experimental uptake or efflux data, so the absolute values of Cout50 and AUCbound should be treated as comparative ranking features rather than quantitative predictions.

Synthetic accessibility is the discriminating factor between leads. Of the top 10 GUB-ranked compounds, 6 have a solved retrosynthetic route (AiZynthFinder `is_solved=True`) and 4 do not. The top-ranked compound by GUB score is *not* synthetically accessible by the current stock (30% coverage, unsolved), while the 3rd-ranked compound (REINVENT_00410F32) is solved with 80% stock coverage in 3 steps. This inversion — the most biologically promising compound is the hardest to make — is the central tension of the Results.

MD simulation quality is the pipeline's current bottleneck, both in compute time and epistemic confidence. Stage 5 consumes 46.6% of total runtime. All compounds receive `md_confidence = low` because the simulations are 100 ps. This means the `md_score` contribution (weight 0.0 in the desirability function) is effectively zeroed out. The MD step currently functions only as a structural gate (`gate_only` mode) rather than as an informative kinetic filter.

6.2 Make comparissons between, reality, experiment. Express your expectation first, then: 1. Address the difference between them. 2. Address the limits of your experiment(why our results are different). You always must find a logical reason for differences. If you find no explanation, you must clearly state that.

Expected: FBA would discriminate a meaningful fraction of compounds by metabolic mechanism. Observed: Only 1/100 compounds was FBA-discriminative; 66% had no metabolic mapping at all.

Expectation vs. reality: The expectation was based on the premise that de novo generated small molecules would share metabolic fate with known antibiotics, some of which are substrates for bacterial metabolic enzymes. In reality, REINVENT generates structurally novel compounds that the MetaNetX database does not contain as substrates. This is a direct consequence of the OOD (out-of-distribution) property of de novo generation: the very novelty that makes generative design attractive also makes metabolic annotation impossible with current databases.

Limit of the experiment: The MetaNetX/MNX metabolite mapping used in Stage 6a is based on exact or near-exact structure matching, with no scaffold-based generalisation. A compound that is a functional analog of a known antibiotic metabolite will receive `no_parent_or_product_mapping` if its SMILES does not appear in the database. This is not a property of the compound — it is a database coverage problem.

Logical reason for the difference: De novo generated molecules are, by design, outside the training distribution of metabolic databases. The 1/100 FBA-discriminative compound succeeded because one of its metabolites (MNXM39, likely acetate/acetaldehyde) appeared in the E. coli model — a coincidence of metabolic identity, not of design.

Expected: IGR from Stage 6b would show diversity — some compounds highly inhibitory, some not. Observed: 197/200 (98.5%) compound-target pairs are discriminative with IGR > 0.48 (median ~0.93).

Expectation vs. reality: The IGR distribution is sharply skewed toward high growth inhibition. This was not expected for a randomly generated set.

Limit of the experiment: The IGR calculation depends on the chosen target gene's essentiality in the metabolic model. If the target gene is essential in the E. coli GEM (genome-scale metabolic model), then *any* compound assumed to inhibit it will produce high IGR — the variability in IGR is then driven only by compound-specific growth rate modifications, not by whether the compound is a good inhibitor. The near-uniformity of IGR values suggests the target used is metabolically essential, which means Stage 6b is currently functioning as a confirmation of target essentiality rather than as a compound discriminator.

Reason for the discrepancy: This is a target selection artefact, not a property of the generated compounds. A more informative Stage 6b would test multiple targets with varying degrees of essentiality, or would condition the IGR on the compound's predicted binding affinity to the target.

Expected: 100 ps MD would provide indicative but meaningful kinetic ranking. Observed: 98% of compounds flagged as unstable; md_score weight = 0.0 in the final scoring.

Expectation vs. reality: The pipeline designers correctly anticipated this (the warning is embedded in the Stage 5 diagnostics report: *"residence_time for md_confidence=low should be treated as indicative only"*). Nevertheless, the outcome is that the most expensive computational step contributes zero weight to the final compound ranking.

Limit of the experiment: 100 ps is insufficient for most ligand–protein systems to equilibrate. Modern best practices for predictive MD of small molecules typically require ≥ 100 ns for converged binding free energies (e.g., FEP, MBAR). At 100 ps, RMSD instability is expected and does not necessarily indicate the compound is a poor binder — it may simply indicate incomplete relaxation.

Reason for the difference: The 100 ps protocol was chosen as a computational compromise given the HPC resource limits of the run (stage5 = 3795 s on a local runner, not HPC). With the current hardware, scaling to converged MD for 100 compounds would require approximately 100× more time (~10 days). The pipeline is correctly designed to down-weight this evidence block under resource constraints.

Expected: AiZynthFinder would routinely find short (1–3 step) routes for drug-like molecules. Observed: A fraction of top compounds have unsolved routes or low stock coverage.

Expectation vs. reality: Of the top GUB-ranked compounds, roughly 40% are not solved by AiZynthFinder with the default stock database. This is higher than might be expected for QED-optimised, Lipinski-compliant molecules.

Limit of the experiment: AiZynthFinder's stock database reflects commercial availability in a specific catalogue (Enamine/eMolecules by default). Structurally novel compounds generated by REINVENT may require non-catalogue building blocks. The synthesis feasibility result is therefore a function of which stock was used, not an absolute measure of synthetic difficulty.

Reason for the difference: REINVENT was rewarded for QED (drug-likeness) but not for retrosynthetic tractability during generation. Adding a retrosynthesis-informed reward signal to Stage 1 (closed-loop generation guided by AiZynthFinder route scores) would directly address this gap.

The downstream GUB_biocomplex_analysis layer addresses an important limitation of the broad pipeline: a composite score can identify promising molecules, but it does not by itself explain whether the predicted complex is structurally convincing. By reusing the pipeline-selected systems for deeper pocket-level analysis, the project separates screening from interpretation. This is especially useful for DHFR, where the ligand must be considered together with the binding pocket and the NADPH cofactor, rather than as an isolated docking pose.

This downstream analysis should still be interpreted conservatively. MM-GBSA, short PLUMED metadynamics, and CP2K single-point descriptors provide additional computational evidence, but they do not replace experimental binding assays or antibacterial testing. Their value is in ranking, mechanistic inspection, and identifying contradictions between signals. For example, a candidate may have a favourable predicted affinity but unstable pocket contacts, poor local geometry, weak MM-GBSA support, or no plausible synthetic route.

A related risk in multi-stage pipelines is silent propagation of partial or stale outputs; the current integrity layer reduces but does not eliminate this risk, and future work should extend coverage to Stage 5 and Stage 6 artefacts.

7. Conclusions

7.1 Take the main lines from the discussion. Must be short, clear, and fair assesment of what you discussed. Do not start discussion again in conclusions!

This work shows that the GUB pipeline can connect de novo molecular generation, physicochemical filtering, retention prediction, structure/affinity modelling, molecular dynamics, metabolic perturbation analysis, and retrosynthetic accessibility into one reproducible prioritisation workflow.

The main strength of the pipeline is not that it proves antibacterial activity directly, but that it reduces a large chemical search space into ranked candidates with multiple independent computational signals. This is useful for triage before more expensive simulation, synthesis, or wet-lab testing.

The results also show clear limitations. Stage 5 and Stage 6 outputs should be treated as weak ranking features rather than binary proof of biological activity. In particular, the MD-derived metrics are sensitive to simulation quality, Stage 6A gives sparse mechanistic signal, and Stage 6B labels require cautious interpretation.

The Stage 8 retrosynthesis step adds an important practical filter: some high-ranking de novo molecules appear synthetically accessible, while others should be deprioritised or redesigned before experimental follow-up.

The pipeline also creates a structured bridge to deeper downstream analysis. The most promising candidates can be transferred into GUB_biocomplex_analysis, where each system is examined with pocket-focused MD metrics, contact fingerprints, MM-GBSA, PLUMED metadynamics, and selected QM/DFT descriptors. This makes the workflow hierarchical: broad screening first, then expensive mechanistic analysis only for the systems that justify it.

Overall, the pipeline is a useful early-discovery framework for hypothesis generation and candidate prioritisation, but the final claims must remain computational until supported by calibrated longer simulations, improved biological signal validation, synthesis planning review, and experimental antibacterial assays.

8. References

8.1 Professional quality is reflected, in how good and consistent your references are presented. (you was careful, to follow format consistently, and your refs. are truly relevant.).

1. Stokes, J. M., Yang, K., Swanson, K., et al. **A Deep Learning Approach to Antibiotic Discovery.** *Cell*, 180(4), 688-702.e13, 2020. https://doi.org/10.1016/j.cell.2020.01.021
2. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. **GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.** *SoftwareX*, 1-2, 19-25, 2015. https://doi.org/10.1016/j.softx.2015.06.001

3. Passaro, S., Corso, G., Wohlwend, J., et al. **Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction.** *bioRxiv*, 2025. https://doi.org/10.1101/2025.06.14.659707
4. Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., et al. **BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification.** *Journal of Cheminformatics*, 11, 2, 2019. https://doi.org/10.1186/s13321-018-0324-5
5. Loeffler, H. H., He, J., Tibo, A., et al. **Reinvent 4: Modern AI-driven generative molecule design.** *Journal of Cheminformatics*, 16, 20, 2024. https://doi.org/10.1186/s13321-024-00812-5
6. Genheden, S., Thakkar, A., Chadimová, V., et al. **AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning.** *Journal of Cheminformatics*, 12, 70, 2020. https://doi.org/10.1186/s13321-020-00472-1
7. Drusano, G. L. **Pharmacokinetics and pharmacodynamics of antimicrobials.** *Clinical Infectious Diseases*, 45(Suppl 1), S89–S95, 2007. https://doi.org/10.1086/518137 - mozna pozniej dodac bardziej mechanistyczną referencję do ODE-based target engagement modelling z literatury PK/PD.

9. Distinctions

GUB pipeline = broad candidate generation and prioritisation.
GUB_biocomplex_analysis = downstream structural, energetic, and QM-level refinement of the most promising systems.