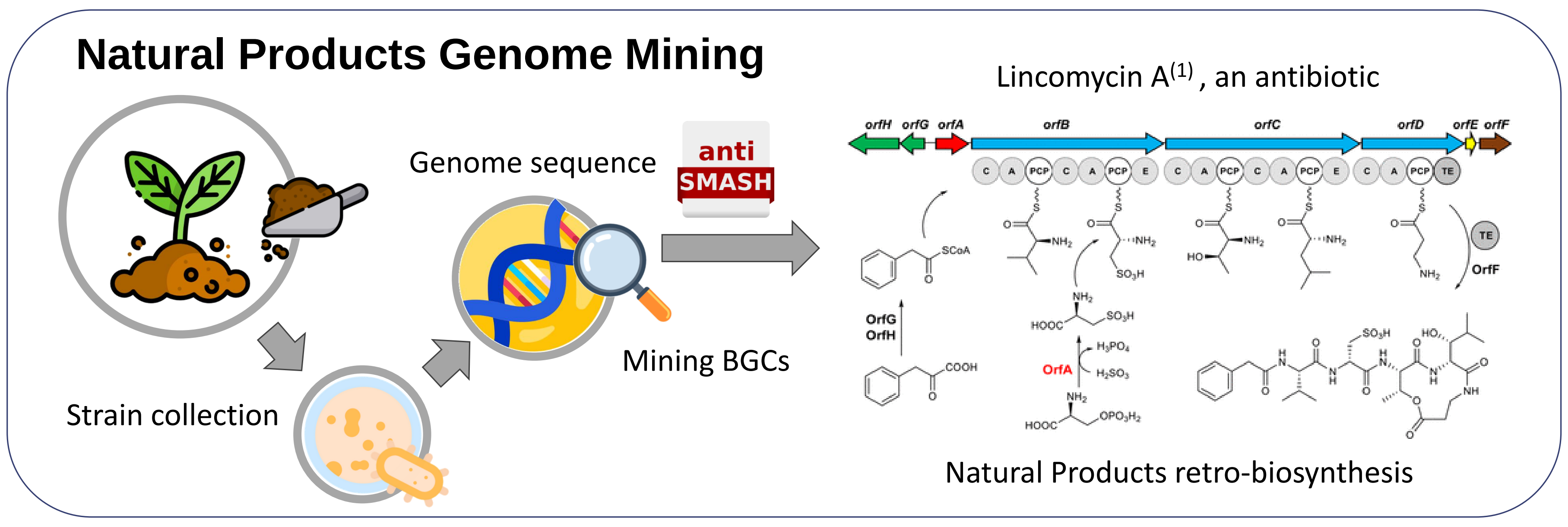
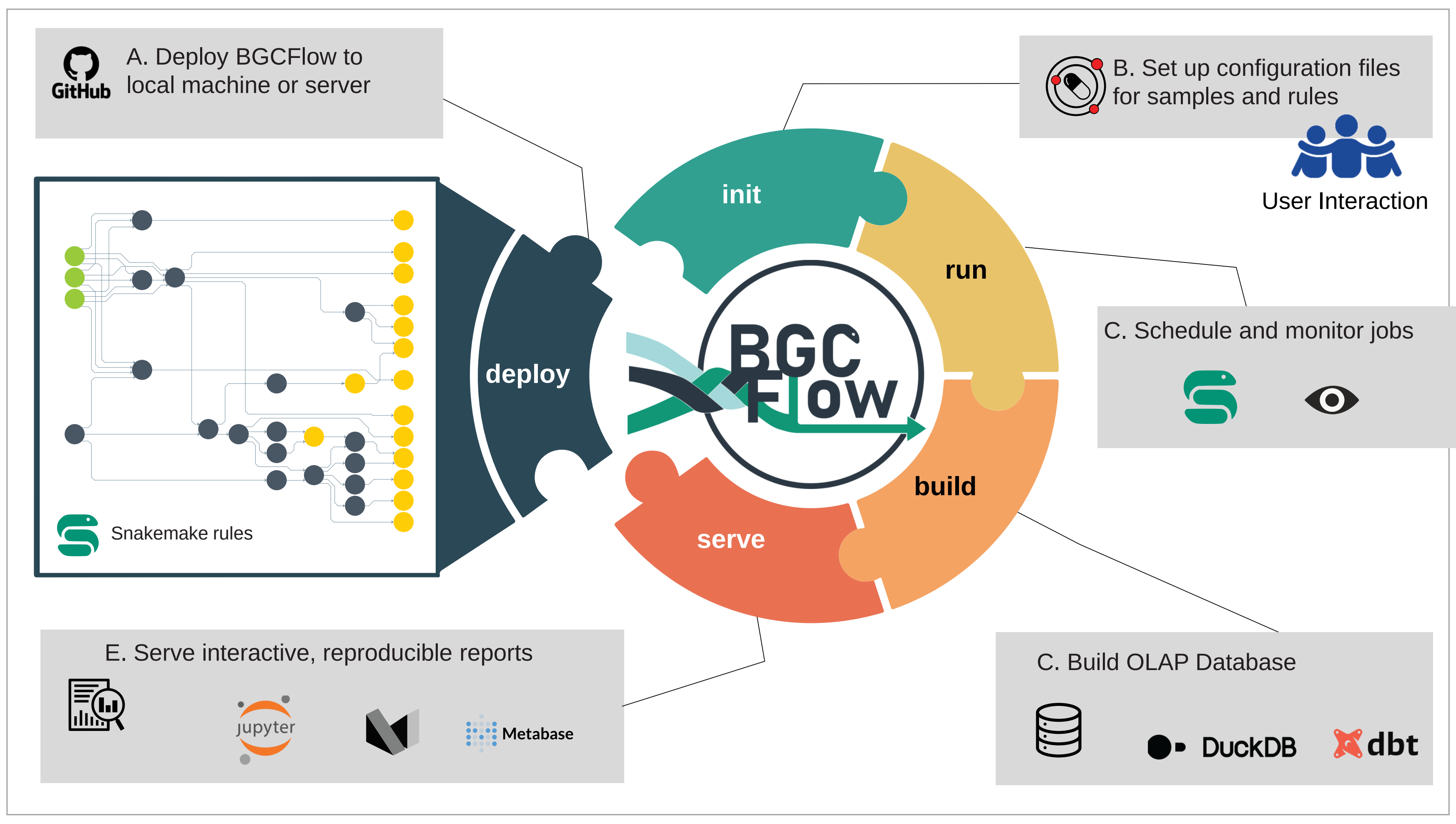
Leveraging Modern Data Stack in a Box for Natural Product Genome Mining in Small-Scale and Private Strain Collection

Matin Nuhamunada[1,2], Omkar S. Mohite[4], Patrick V. Phaneuf[1], Bernhard O. Palsson[1,3], and Tilmann Weber[1,2]

[1] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark,
[2] Center for Microbial Secondary Metabolites, Technical University of Denmark
[3] Department of Bioengineering, University of California San Diego
[4] Bactobio Ltd, London, UK

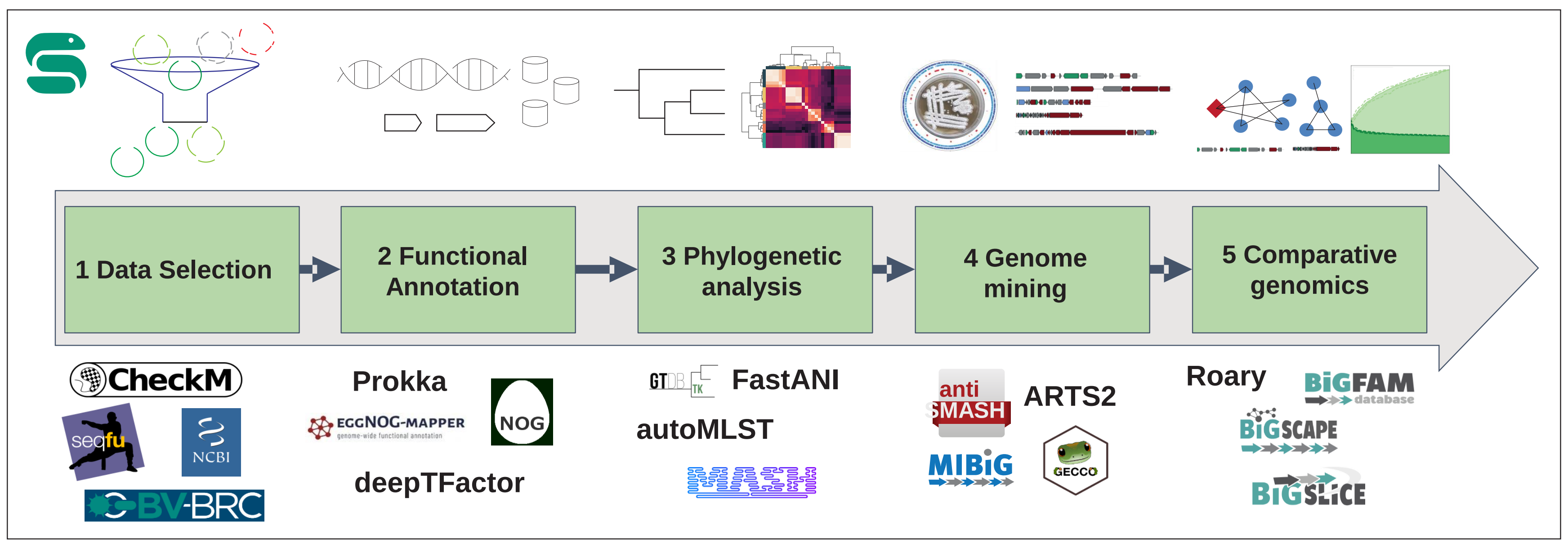

Natural Products Genome Mining
Strain collection
Genome sequence
antiSMASH
Mining BGCs
Lincomycin A(1), an antibiotic
Natural Products retro-biosynthesis

A. Workflow structure enabling iterative genome mining analysis


A. Deploy BGCFlow to local machine or server
GitHub
B. Set up configuration files for samples and rules
User Interaction
C. Schedule and monitor jobs
C. Build OLAP Database
DuckDB
dbt
E. Serve interactive, reproducible reports
jupyter
Metabase
Snakemake rules
deploy
init
run
build
serve
BGCFlow

https://github.com/NBChub/bgcflow_wrapper

B. Snakemake pipelines for end-to-end genome mining



https://github.com/NBChub/bgcflow

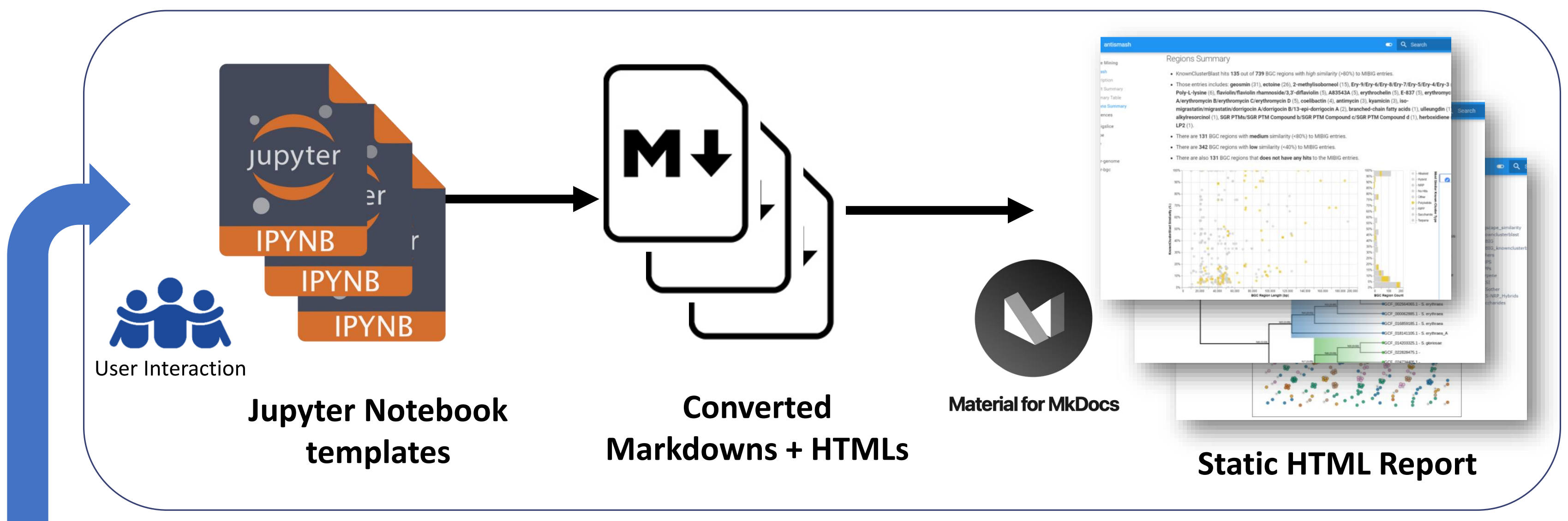
A comprehensive genome mining workflow for the analysis of bacterial pangenomes

- We present **BGCFlow** [2], a comprehensive genome mining workflow for the analysis of bacterial pangenomes.
- BGCFlow integrates a "modern data stack (MDS) in a box" leveraging tools such as dbt, DuckDB, and Metabase to offer **streamlined data engineering pipeline** and efficient platform for the **exploration and management of private strain collections**.
- Each tool is selected for its unique capabilities:
  - dbt for **transforming data** with simplicity and reproducibility,
  - DuckDB for its lightweight, in-process SQL database that facilitates **fast analytical queries**,
  - Metabase for its **user-friendly interface** allowing both data scientists and lab researchers to **visualize and interact with their data**
  - RAG for more **natural, conversational investigation** of the data in the strain collection
- By doing so, we aim to bridge the gap between the potential of genome mining and the practicalities of conducting such research **at a scale** that is both **manageable** and **accessible** to a broader scientific community.

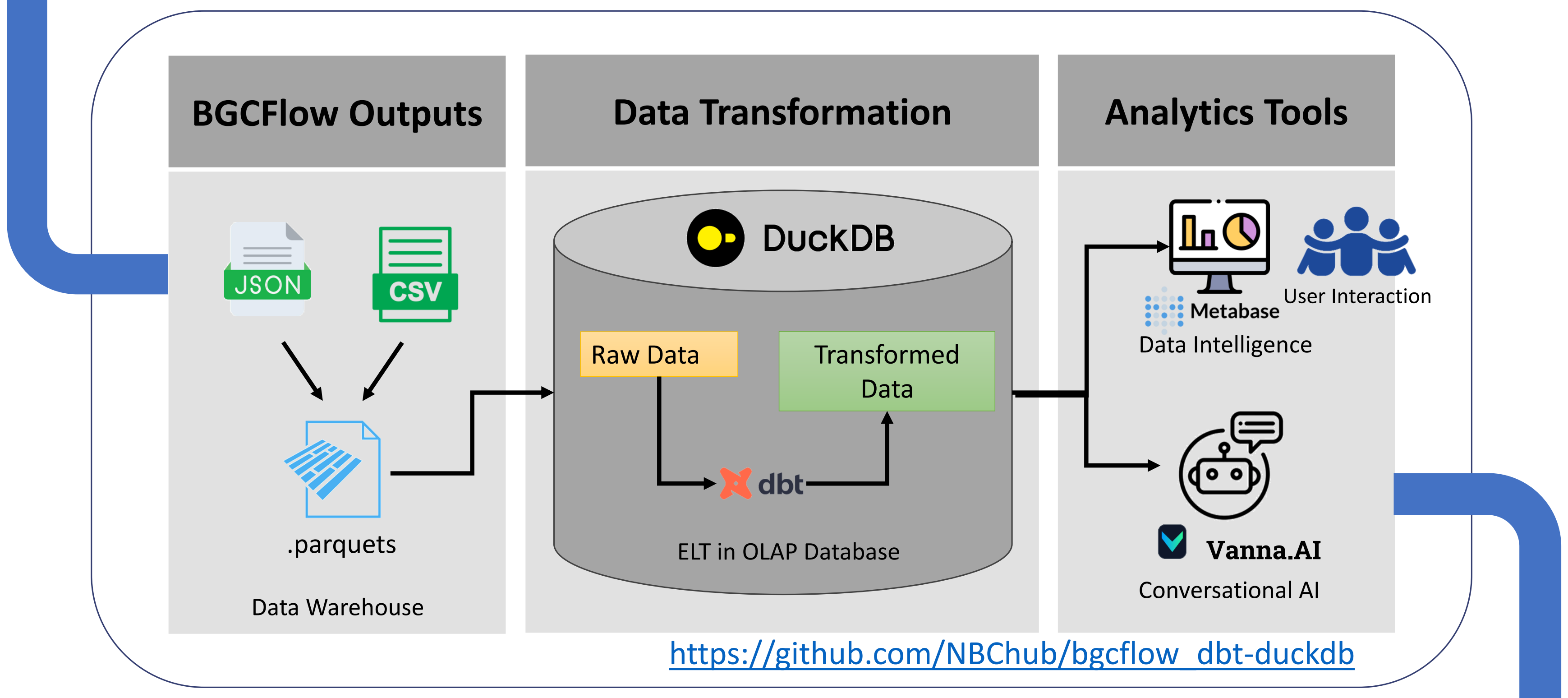
Background

- The advent of third-generation **sequencing** technologies enables individual researchers and small laboratories to **affordably create** and manage **private microbial strain collections**.
- This shift promises to accelerate **natural product discovery** by facilitating the **mining** of biosynthetic gene clusters (**BGCs**) from **genomic sequences**, an important step in unlocking **novel pharmaceuticals**, **agrochemicals**, and other industrially relevant compounds.
- As researchers embark on building and analyzing their own private collections, the **challenge** extends beyond managing large-scale public genomic datasets but also in **providing solutions that cater to the analysis of smaller, more focused collections**.

C. Reproducible reports with Jupyter Notebooks and MkDocs

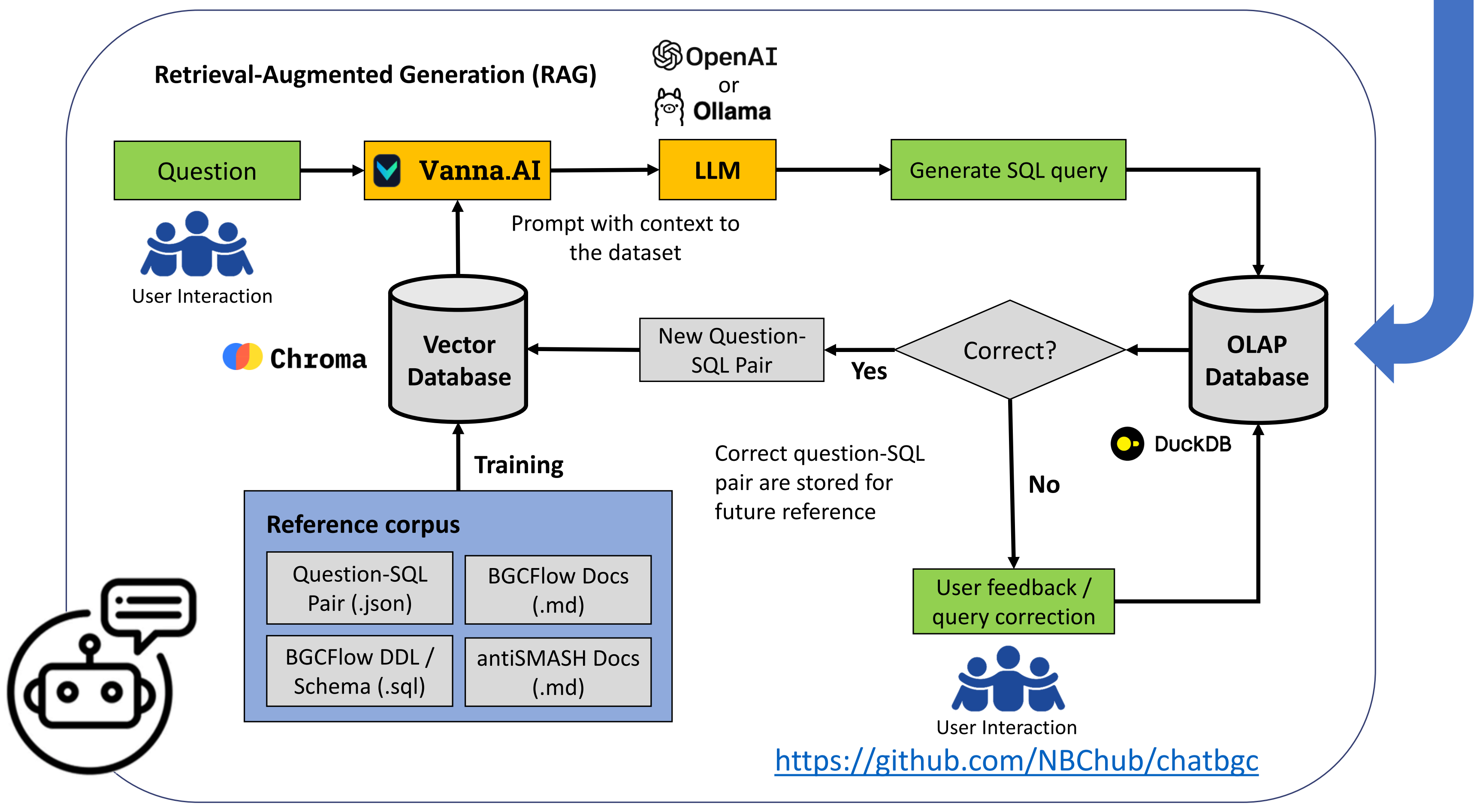


D. Data processing and visualization with MDS in a box



https://github.com/NBChub/bgcflow_dbt-duckdb

E. Chat with the data using LLM-RAG



https://github.com/NBChub/chatbgc

Contact:

✉ matinnu@biosustain.dtu.dk

References

(1) Wang, M., D. Chen, Q. Zhao, W. Liu. 2018. The Journal of Organic Chemistry. DOI: 10.1021/acs.joc.8b00044
(2) Nuhamunada, M., O.S. Mohite, P.V. Phaneuf, B.O. Palsson, T. Weber. 2024. *Nucleic Acids Research*, gkae314, https://doi.org/10.1093/nar/gkae314

Acknowledgements

This work was supported by grants of the Novo Nordisk Foundation (NNF20CC0035580, NNF16OC0021746). MN was supported by the NNF Copenhagen Bioscience PhD program (NNF20SA0035588).