Max Homilius, PhD

max@homili.us
mxhm.github.io

Professional summary

Machine learning scientist and computational biologist with expertise in **multimodal phenotyping**, **scalable phenotype extraction and integration**, and **biobank-scale genetic studies**. Designed self-supervised and contrastive models for ECG, cytometry, and clinical text; built AWS serverless pipelines and graph databases supporting a phenotype data lake; led GWAS and PRS analyses of perturbation responses and deep learning-derived traits. Focused on turning large, complex human datasets into actionable insights for biological discovery, clinical translation, and therapeutic development.

Technical skills

Machine learning & biomedical applications: self-supervised & contrastive learning, multimodal transformers, classification/regression on text and biomedical signals (ECG, cytometry, imaging); phenotype extraction (PhenoBERT, ClinPhen, small & large language models); biomarker discovery.
Genetics & biobanks: GWAS (plink2, regenie), PRS (PRSice2, LDpred), genetic correlation (HDL), PheWAS, EHR-derived trait analysis.

Data processing: graph-schema design, HPO reasoning, EDW pipelines for notes, labs, medications & diagnoses including longitudinal analysis at patient and cohort levels.
Local & cloud computing: AWS CloudFormation, Lambda & Step Functions, SageMaker, DynamoDB, Neptune (Graph DB), ZFS, Ansible, Docker.
Frameworks & tools: PyTorch, TensorFlow, HuggingFace, spaCy, Python, R, bash, SQL.

Selected projects

Phenotype extraction & data lake architecture.
Contributed to AWS serverless pipelines (CloudFormation, Lambda & Step Functions) that ingest clinical notes, extract HPO phenotypes with rule-based models plus lightweight Small Language Models, and store evidence in an ontology-aware Neptune graph. Improved schema to handle longitudinal data; validated on 7 000 curated subjects; piloting NLP rollout in NICU notes. Co-first author on forthcoming conference submission.

Self-supervised multimodal learning for biomarker discovery.
Built multimodal encoders (ECG, blood cytometry) aligned with structured EHR context via cross-attention; >100 000 samples processed. Pretext task predicting new diagnoses within 6 months uncovered novel ECG/cytometry biomarkers; manuscript in preparation.

Perturbational phenotyping & genetic analysis of blood cells.
Led computational work of large functional genomics study (*Nat Genet* 2023): GWAS of perturbation-response traits in 4 600 subjects, PRS transfer to MGB and UK Biobanks, identified 119 loci/96 genes and links to cardiometabolic and renal disease subsets.

Professional experience

since 08/2025 Principal Data Scientist, Mass General Brigham

09/2024-08/2025 Senior Data Scientist, Office of Data Sciences, Nationwide Children's Hospital
Contributed to production AWS pipelines and graph architecture; developed SLMs for phenotype contextualization and analyzed phenotypic annotations of hospital-wide cohort. Participated in code review and technical deep dives; mentored junior data scientists on modeling strategy and code quality.

04/2019-09/2024 Postdoctoral Research Fellow, MacRae Lab & One Brave Idea, Brigham and Women's Hospital
Led GWAS & PRS of perturbational phenotypes; developed multimodal self-supervised ECG/cytometry models; built NoteContrast contrastive model for ICD-10 coding; integrated proteomics & transcriptomics data. Mentored data scientists on ML modeling, genetics, statistical methods, compute architecture, and scientific writing.

09/2018-02/2019 Postdoctoral Research Associate
Troyanskaya Lab, Princeton University
Developed functional-network methods to prioritise disease genes, animal models and drug targets.

Education

2018 PhD Computer Science, Princeton University
Thesis: "Network-Based Prioritization of Disease Genes, Animal Models, and Drug Targets" (Advisor: O. Troyanskaya)

2014 MA Computer Science, Princeton University

2011 MPhil Computational Biology, University of Cambridge

2009 BSc Bioinformatics, Free University of Berlin

Selected publications

Deep phenotyping, GWAS, PRS. **Homilius M***, Zhu W* *et al.* "Perturbational phenotyping of human blood cells reveals genetically determined latent traits associated with subsets of common diseases." *Nat Genet* 10.1038/s41588-023-01600-x (2024).

Contrastive learning, notes. Kailas P*, **Homilius M*** *et al.* "Contrastive Language-Diagnostic Pretraining for automated adjudication of medical notes." *ML4H / PMLR* (2023).

De-identification, transformers. **Homilius M***, Kailas P*, *et al.* "Robust de-identification of medical notes using transformer architectures, sentence context, and recall thresholding." Model weights downloaded >1M times on HuggingFace (deid_roberta_i2b2). Submitted.

Federated ECG & Echo models. Goto S, Solanki D, John JE, Yagi R, **Homilius M**, *et al.* "Multinational Federated Learning Approach to Train ECG and Echocardiogram Models for Hypertrophic Cardiomyopathy Detection." *Circulation* 146:755–769 (2022).

Risk prediction, cytometry, NLP. Truslow JG, Goto S, **Homilius M**, Mow C, Higgins JM, MacRae CA, Deo RC. "Cardiovascular Risk Assessment Using Artificial Intelligence-Enabled Event Adjudication and Hematologic Predictors." *Circ Cardiovasc Qual Outcomes* 15(6):e008007 (2022).

Deep phenotyping, RNA-Seq. Zhu W, Guo S, **Homilius M**, *et al.* "PIEZO1 mediates a mechanothrombotic pathway in diabetes." *Sci Transl Med* 14(626):eabk1707 (2022).

Honors and awards

2022-23 Drs. Tobia & Morton Mower Fellow

2010-11 German Academic Scholarship Foundation (Study Abroad Stipend)

2008-11 German Academic Scholarship Foundation Fellow

2008 DAAD Travel Award