Michal Malyska

malyskamichal@gmail.com • michalmalyska.com • github.com/MichalMalyska

Principal ML engineer for AI drug discovery: GPU performance, ML infrastructure, and agentic science workflows

Work Experience

Principal ML Engineer — **Toronto, Canada**
Grove Biopharma (*acquired Fable's ML team*) — *May 2026 – Present*

- Sole engineer in a 7-person ML org, procurement to production: infrastructure, DevOps, MLOps, ML engineering, technical roadmap, and GPU/cloud vendor contracts across ~$3M/yr of on-prem and AWS infrastructure, including the GPU cluster expansion.
- Built human-gated agentic preclinical drug-design workflows on Claude over our tools, open models, and data, used company-wide; wrote skills extending Claude for Life Sciences; set the team's coding-agent practice.
- Architected a multi-region AWS platform for long-running molecular dynamics (CHARMM/Amber) on spot GPUs: fault-tolerant checkpoint/resume, verified cross-region migration of running simulations; ~8× cheaper (~$3.2k→~$400 per sample). Added an EKS/Karpenter cluster for scaling and GPU-selection experiments.
- Built a spot-tolerant batch design pipeline of CPU/GPU stages (MSAs, folding, design, inverse folding, refolding, scoring) over Boltz-2, Protenix, OpenFold/AF2, BioEmu, diffusion models; peaks at ~8,000 AWS GPUs; run by scientists and via Claude skills on Bedrock.
- Sped up Boltz-2 1.12–1.16× on H100 (up to 1.23× on A10G) vs cuEquivariance by integrating Anthropic's FlashPairformer kernels via Boltz Community, with per-GPU dispatch and fallback; extended OpenFold to use multichain templates.
- Built a lab-analysis platform and its research database from scratch: 27 wet-lab scientists submit assay data and get reports back.

MLOps Lead — **Toronto, Canada**
Fable Therapeutics — *January 2024 – May 2026*

- Cut training compute 24× (~1,920→~80 H100 GPU-hours) at matched validation metrics by taking scientists' research PyTorch to production speed: bug fixes, profiling, org-wide CUDA-sync and CPU-op removal, per-config compiled kernels; taught the patterns.
- Cut inference cost >99% (~$1→<$0.01/prediction) on >90% of workloads at equivalent quality: input pruning, compiled kernels, GPU choice.
- Stood up an on-prem GPU cluster (8 H100, 16 A6000) from spec and racking to scientist onboarding: GPU scheduling, isolation, alerting; H100s ~100% occupied, A6000s ~65%.
- Built self-hosted monitoring and alerting across CI, AWS Batch, Slurm, and on-prem: exposed utilization gaps (filled by a background Slurm partition) and flagged jobs underusing GPUs, which justified the on-prem expansion.
- Replaced 27 hand-run notebooks with config-driven AWS Batch orchestration for the production design pipeline; with spot routing and runtime optimizations, cut cost $130–170k→$50–60k/run and operator time ~2 weeks→~8 hours.
- Set ML engineering standards for a 9-person team: CI gates, container builds, model lineage, PR templates; did a third of all reviews.
- Built the research database with scientists (inventory, assays, traceable releases, QC), closing the wet-lab loop: designs out, assay results back.

ML Scientist → ML Lead — **Toronto, Canada**
Semantic Health (*acquired*) — *June 2019 – December 2022*

- First employee: built the ML stack and orchestrated deployment pipelines; bandit tests took model releases from monthly to weekly.
- Led clinical NLP/OCR from R&D to production at multiple clients through US expansion and a $60M valuation; hired and managed 7.
- Designed prediction APIs with fallbacks that removed ML on-call; active learning cut labeling cost >75%.

Sessional Lecturer — **Toronto, Canada**
University of Toronto — *2020 – May 2023*

- Taught Probabilistic Machine Learning (STA414/CSC412); modernized the curriculum with transformers and diffusion models.

Publications & Open Source

- Publications: BioNLP 2022, ACM CHIL 2021, EMNLP ClinicalNLP 2020 (co-first author)
- Open source: Boltz Community, MedSpaCy, SciSpaCy, AllenNLP, MS-BERT

Skills

- **Infra & MLOps**: Python, Bash, SQL, Linux, AWS (EC2, EKS, Batch, S3, RDS, IAM, Bedrock, SageMaker), Claude Code, Agent SDK, Terraform, Slurm, Docker, Kubernetes, GitHub Actions, Prometheus, Grafana, PostgreSQL + pgvector, MLflow, W&B
- **ML & GPU performance**: PyTorch, PyTorch Lightning, Triton, CUDA, cuEquivariance, FlashAttention, DeepSpeed, FSDP/DDP, mixed precision, Nsight, Transformers, inference optimization
- **Methods**: protein folding, molecular dynamics, distributed systems, context engineering, agent architectures, diffusion models, LLMs, RAG, NLP, active learning, Bayesian statistics

Education

- **MSc Statistics** (2022), **BSc Statistics** (2019), University of Toronto