Jianyu Yang

☎ +1 8149968462 | ✉ yztxwd@gmail.com | ⌂ yztxwd.github.io | yztxwd | Jianyu

Skills

Programming	Python, JAVA, R, bash
Machine Learning & AI	Deep learning (CNNs, Transformers), large-scale model training & evaluation, Bayesian modeling
Frameworks & Tools	PyTorch, PyTorch Lightning, TensorFlow, Hugging Face, NVIDIA DALI, cuML, Ray (distributed training)
Genomics & Bioinformatics	ChIP-seq, ATAC-seq, RNA-seq, MNase-seq, BSTAB-seq analysis
Data Engineering & Pipelines	Snakemake, HPCSlurm workflows, large-scale data processing

Experience

Interpretable AI for Genomics and Biomarker Discovery

Penn State

Biorxiv, doi.org/10.64898/2026.01.20.700723 — *Oct. 2024 - Now*

- Developed TPCAV, a scalable and input-agnostic model interpretation framework to enable biologically meaningful feature discovery from deep learning models.
- Enabled identification of regulatory signals beyond sequence motifs, including chromatin accessibility and epigenomic features, supporting biomarker discovery and hypothesis generation.
- Demonstrated applicability across CNNs, transformer-based models, and tokenized foundation models, improving interpretability of modern genomics AI systems.

Discovery of Cell Type–Specific TF Cofactors via Interpretable Deep Learning

Penn State

Molecular cell, doi.org/10.1016/j.molcel.2024.06.022 — *Jan. 2022 - Oct. 2023*

- Designed and deployed a multimodal deep learning model integrating DNA sequence and chromatin features to predict transcription factor binding across cell types.
- Identified cell type–specific cofactors (AP-1) and quantified the contribution of chromatin context, enabling mechanistic insight and candidate regulatory biomarker discovery.

Multimodal Deep Learning for Predicting Induced FOX Factor Binding

Penn State

Biorxiv, doi.org/10.1016/j.molcel.2024.06.022 — *Jan. 2022 - Sep. 2023*

- Developed a multimodal deep learning framework (CNN/Transformer) integrating DNA sequence and chromatin features to predict induced transcription factor binding in mESCs.
- Quantified factor-specific dependence on pre-existing chromatin states using model interpretability analyses.
- Enabled systematic characterization of how epigenetic context modulates transcription factor binding and pioneer activity, supporting regulatory mechanism discovery.

Bayesian Gaussian Mixture Modeling for Nucleosome Positioning and Subtype Discovery

SMU & Penn State

Genome Research, doi.org/10.1101/gr.279138.124 — *Sep. 2019 - April. 2024*

- Developed a hierarchical Bayesian Gaussian mixture model (SEM) to classify nucleosome subtypes from large-scale MNase-seq data.
- Discovered and characterized a previously unrecognized nucleosome subtype in mESCs.
- Enabled quantitative analysis of nucleosome structural heterogeneity through probabilistic modeling of fragment size distributions.

Construction of RUNX1 Regulatory Networks for Target and Biomarker Discovery

SMU

Frontiers in Molecular Biosciences, doi.org/10.3389/fmolb.2021.692880 — *Sep. 2020 - April. 2021*

- Built Snakemake pipelines for integrated ChIP-seq and RNA-seq analysis to reconstruct RUNX1 regulatory networks in leukemia.
- Identified CENPE as a proliferation-associated downstream target of RUNX1, providing a candidate therapeutic target and biomarker for leukemic progression.

SOFTWARE & TOOLING PROJECTS

Seqchromloader (Training Data Toolkit for Genomic DL)

Skills: Python, Pytorch, Webdataset — *github.com/seqcode/seqchromloader*

- Built a production-ready toolkit to construct training datasets for sequence/chromatin DL models. Optimized for high-throughput, distributed dataset streaming. Has been widely adopted by lab members.

HDF5-Backed Genome Coverage & Heatmap Engine

Related skills: Python, HDF5 — *github.com/yztxwd/chiptoolkit*

- Developed a Deeptools-like plotting engine using HDF5 to pre-store genome-wide tracks, enabling extremely fast data retrieving and heatmap/composite plot generation for thousands of regions.

SEM (Nucleosome caller)

Skills: Java, HDF5, Bayesian modeling — *github.com/YenLab/SEM*

- Greatly improved SAM/BAM file loading speed by integrating multi-processing and HDF data format.
- Built a nucleosome caller that is able to predict nucleosome locations and types across mammalian genomes.

General Snakemake Pipelines for NGS Data

Skills: R, Python, Snakemake, NGS analysis, Slurm, HPC — *github.com/yztxwd/snakemake-pipeline-general*

- Built modular snakemake pipelines for ATAC-seq, ChIP-seq, RNA-seq, BS-seq, and MNase-seq. Designed for reproducibility, portability, and HPC batch environments (Slurm). Adapted by lab members for routine preprocessing workflows.

Presentations

2024 **Talk** Jointly characterizing the sequence and chromatin binding preferences of transcription factors using neural networks *16th Great Lakes Bioinformatics conference*

2024 **Proceeding talk (selected for Genome Research)** SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes *RECOMB 2024*

Publications

Link to Google Scholar(scholar.google.com/citations?user=r7sRhzoAAAAJ)

TPCAV: Interpreting deep learning genomics models via concept attribution *Biorxiv*
10.64898/2026.01.20.700723 *2026*

Systematic Dissection of Sequence Features Affecting the Binding Specificity of a Pioneer Factor Reveals Binding Synergy Between FOXA1 and AP1 *Molecular Cell*
10.1016/J.MOLCEL.2024.06.022 *2024*

SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes *Genome Research*
10.1101/2023.10.17.562727 *2024*

Joint sequence & chromatin neural networks characterize the differential abilities of Forkhead transcription factors to engage inaccessible chromatin *Biorxiv*
10.1101/2023.10.06.56122 *2023*

Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification *NAR Genomics and Bioinformatics*
10.1093/NARGAB/LQAB094 *2021*

RUNX1 upregulates CENPE to promote leukemic cell proliferation *Frontiers in Molecular Biosciences*
10.3389/FMOLB.2021.692880 *2021*

Education

Ph.D., Bioinformatics and Genomics Program, Pennsylvania State University *Aug. 2020 - May. 2026 (Expected)*
M.S., Developmental Biology, Southern Medical University *Sep. 2017 - Jun. 2020*
B.S., Preclinical Medicine, Southern Medical University *Sep. 2012 - Jun. 2017*