# Jianyu Yang

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

## Skills

| | |
|---|---|
| **Programming** | Python, JAVA, R, bash |
| **Machine learning & Deep learning** | Bayesian models, Training + evaluating large genomic transformer and CNN models |
| **Software Packages** | Tensorflow, DALI, Pytorch, Pytorch-lightning, Huggingface„ CuML, Ray distributed training |
| **NGS analysis** | ChIP-seq, ATAC-seq, RNA-seq, MNase-seq, BS/TAB-seq analysis, Snakemake |

## Experience

### Developed interpretation method for Genomics Deep Learning Models via Concept Attribution

*Penn State*

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

- Developed TPCAV, an generalizable and input-agnostic method for deep learning model interpretation.
- Extended model interpretability beyond sequence motifs to additional feature such as chromatin accessibility and epigenomic features.
- Demonstrated applicability to interpreting tokenized foundation models.

### Identified cell type specific transcription factor FOXA1 cofactor through deep learning model training and interpratation

*Penn State*

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

- Designed and applied a bimodal convolutional neural network integrating DNA sequence and chromatin features to dissect the determinants of FOXA1 binding across multiple cell lines.
- Identified cell type specific TF cofactor AP-1 and quantified the chromatin contribution to FOXA1 binding.

### Multimodal Deep Learning for induced Fox Factor Binding Prediction

*Penn State*

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

- Designed and trained a multimodal CNN/Transformer framework integrating DNA sequence and chromatin features to predict induced binding of Fox family transcription factors in mouse embryonic stem cells (mESCs).
- Performed systematic model interpretation to quantify factor-specific reliance on pre-existing chromatin states.
- Enabled systematic analysis of how epigenetic context modulates transcription factor binding and pioneer activity.

### Nucleosome Calling with Bayesian Gaussian Mixture Models

*SMU & Penn State*

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

- Designed and implemented a hierarchical Bayesian Gaussian mixture model (Size-based Expectation Maximization; SEM) in Java to classify nucleosome structural subtypes from large-scale MNase-seq data.
- Applied SEM to mESC MNase-seq datasets, enabling the discovery and characterization of a previously unrecognized nucleosome subtype.
- Provided new insights into nucleosome structural heterogeneity through probabilistic modeling of fragment size distributions.

### Regulatory Network Analysis on RUNX1

*SMU*

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

- Developed Snakemake workflows for integrated analysis of differential gene expression and transcription factor binding in leukemia models.
- Performed joint analysis of ChIP-seq and RNA-seq data from leukemia and wild-type cells to infer RUNX1 regulatory networks.
- Identified CENPE as a RUNX1-regulated downstream target associated with cell proliferation.

## 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* |