Jianyu Yang

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

Education

Bioinformatics and Genomics Program (Ph.D. candidate) — *State College, US*

Pennsylvania State University — *Aug. 2020 - May. 2026 (Planned)*

Developmental Biology (Master's Degree) — *Guangzhou, China*

Southern Medical University — *Sep. 2017 - Jun. 2020*

Preclinical Medicine (Bachelor's Degree) — *Guangzhou, China*

Southern Medical University — *Sep. 2012 - Jun. 2017*

Publications

Link to Google Scholar

TPCAV: Interpreting deep learning genomics models via concept attribution — *Biorxiv*

10.64898/2026.01.20.700723 — *2026*

Jianyu Yang, Shaun Mahony

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*

Cheng Xu, Holly Kleinschmidt, ***Jianyu Yang***, Erik Leith, Jenna Johnson, Song Tan, Shaun Mahony, Lu Bai

SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes — *Genome Research*

10.1101/2023.10.17.562727 — *2024*

Jianyu Yang, Kuangyu Yen, Shaun Mahony

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*

Sonny Arora*, ***Jianyu Yang****, Tomohiko Akiyama, Daniela Q James, Alexis Morrissey, Thomas R Blanda, Nitika Badjatia, William KM Lai, Minoru SH Ko, B Franklin Pugh, Shaun Mahony

Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification — *NAR Genomics and Bioinformatics*

10.1093/nargab/lqab094 — *2021*

Houyu Zhang, Ting Lu, Shan Liu, ***Jianyu Yang***, Guohuan Sun, Tao Cheng, Jin Xu, Fangyao Chen, Kuangyu Yen

RUNX1 upregulates CENPE to promote leukemic cell proliferation — *Frontiers in Molecular Biosciences*

10.3389/fmolb.2021.692880 — *2021*

Shan Liu, ***Jianyu Yang***, Guohuan Sun, Yawen Zhang, Cong Cheng, Jin Xu, Kuangyu Yen, Ting Lu

Research Experience

Interpret Genomics Deep Learning Models via Concept Attribution

Skills: Pytorch, Captum/DeepLiftSHAP, TF-Modisco, Scikit-Learn — *Oct. 2024 - Now*

- Developed an input-agnostic concept attribution framework, Testing with PCA-transformed Concept Activation Vectors (TPCAV), to overcome limitations of existing genomics interpretation methods that are largely restricted to one-hot DNA motif analysis. TPCAV enables consistent and reliable global interpretation of deep learning models for transcription factor binding, including CNNs, Transformer-based and tokenized foundation models, and extends interpretability beyond sequence motifs to chromatin accessibility and other genomic features. This framework provides a flexible and generalizable complement to existing model interpretation approaches.

Nucleosome Calling with Bayesian Gaussian Mixture Models (SEM Algorithm)

Related skills: Java, Bayesian Methods, Cell Culture, Molecular Biology, CRISPR-Cas9 — *Sep. 2017 - April. 2024*

- Designed and implemented a hierarchical Gaussian mixture model in Java, termed Size-based Expectation Maximization (SEM), to classify structural subtypes of nucleosomes from large-scale MNase-seq data. Application of SEM to mouse embryonic stem cells (mESCs) enabled the discovery and characterization of a previously unrecognized nucleosome subtype, providing new insight into nucleosome structural heterogeneity.

Training and interpreting deep learning model for FOXA1 binding partner in A549

Related skills: Pytorch, Pytorch-Lightning, Captum/DeepLiftSHAP, TF-Modisco *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. Using model interpretation methods (DeepLIFT-SHAP and TF-MoDISco), I identified AP-1 as a FOXA1 cofactor uniquely in A549 cells, revealing cell-type-specific cofactor dependence. This regulatory interaction was subsequently supported by downstream experimental validation.

Multimodal Deep Learning for induced Fox Factor Binding Prediction

Related skills: Tensorflow, Pytorch, Pytorch-Lightning, Captum/DeepLiftSHAP, TF-Modisco *Jan. 2022 - Sep. 2023*

- Developed a multimodal CNN/Transformer-based neural network integrating DNA sequence and chromatin features to predict induced binding of Fox family transcription factors in mouse embryonic stem cells (mESCs). Model interpretation revealed differential dependence on pre-existing chromatin states across Fox factors, with FoxG1 and FoxP3 showing stronger chromatin reliance than FoxA1 and FoxL2. This work establishes a general analytical framework for dissecting how epigenetic context modulates transcription factor binding specificity and pioneer activity.

Regulatory Network Analysis on RUNX1 in leukemia cell

Related skills: R, DESeq, Snakemake *Sep. 2020 - April. 2021*

- Developed automated Snakemake workflows for integrated analysis of differential gene expression and transcription factor binding in leukemia models. By jointly analyzing ChIP-seq and RNA-seq data from leukemia and wild-type cells, I identified CENPE as a RUNX1-regulated downstream target associated with cell proliferation, with supporting validation from wet-lab experiments.

Presentations

2025	**Poster,** Explaining genomics deep learning models via concept attribution	*MLCB*
2024	**Proceeding talk (selected for Genome Research),** SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes	*RECOMB*
2024	**Talk,** Jointly characterizing the sequence and chromatin binding preferences of transcription factors using neural networks	*GLBIO*
2023	**Poster,** SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes	*Penn State Summer Symposium*
2023	**Poster,** SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes	*GLBIO*
2023	**Poster,** SEM: size-based expectation maximization for characterizing nucleosome positions and subtypes	*Keyston Symposium*
2022	**Poster,** Size-based expectation maximization for characterizing nucleosome positions and subtypes	*RSGDREAM*

Honors & Awards

2025	**Rising Researcher Collaborations Funding Opportunity,** Penn State ICDS	*State College, US*
2024	**Travel Fellowship,** 16th Great Lakes Bioinformatics conference	*Pittsburgh, US*
2020	**Graham Endowment Fellowship,** Bioinformatics and Genomics Program Recruitment	*State College, US*
2018	**1st Prize,** Excellent Graduate Student of Southern Medical University	*Guangzhou, China*
2018	**1st Prize,** National Scholarship for Graduate students	*Guangzhou, China*

Teaching

Tutoring Junior Lab Members

State College, PA

Senior Student Mentor *2021-Now*

- Mentor junior lab members on project design, experimental/technical planning, and troubleshooting.
- Provide guidance on scientific communication and long-term career development.

Summer Internship Project

State College, PA

Undergraduate Research Mentor *Aug. 2025*

- Supervised an undergraduate intern working on adapting the SEM algorithm for TF peak calling in CUT&Tag data.
- Held weekly meetings to resolve technical issues, review progress, and provide computational guidance.

Summer Internship Project

State College, PA

Undergraduate Research Mentor *Aug. 2024*

- Mentored an undergraduate intern on applying deep learning models to predict TF ChIP-seq peaks in unmappable genomic regions.
- Conducted weekly check-ins and supported problem-solving and model development.

Penn State Bioinformatics & Genomics Data Reproducibiliy Bootcamp

State College, PA

Co-Host, *Aug. 2022*

- Co-organized the BG program's data reproducibility bootcamp.
- Led the Snakemake workshop, delivering lectures on workflow design and presenting hands-on reproducibility examples.
- Workshop repository: *github.com/biostars/bootcamp-central*

Snakemake Reproducibility Workshop

State College, PA

Workshop Instructor *2022*

- Designed and delivered a workshop for lab members on using Snakemake to enhance computational workflow reproducibility.

Personal 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.

General Snakemake Pipelines for NGS Data

Skills: R, Python, Snakemake, Common packages used in 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.

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