

Yanjie Huang

+86-13122396686 | huangyanjie@sjtu.edu.cn | <https://github.com/kevinhyj>

SKILLS

- **Model Architecture & Systems:** MoE Transformer (Megablocks); Unified diffusion model (Transfusion, Bagel); Mechanistic Interpretability (SAE); Agentic Systems (LangChain; Qwen-Agent).
- **Training & Optimization:** Pretraining (CLM+GLM), Mid-training (SFT) (via MosaicML); Finetuning (LoRA, Full-parameter tuning); Post-training (DPO, OPSD); vLLM
- **Data Engineering:** Large-scale data processing (>150TB), Parallel data acquisition, Streaming QC pipelines, SELEX/NGS preprocessing, Metagenomic mining, Structure-based data mining.
- **Scientific Visualization:** Conceptual schematics, data viz, and molecular rendering (ChimeraX, pymol).
- **Molecule Design:** CRISPR-Cas; RNA aptamer; CircRNA; Phase-separating proteins; TCR; Binder; Antibody
- **Programming:** Python, C++, Shell Scripting, LaTeX

EDUCATION

Shanghai Jiao Tong University (SJTU) | *B.E. Biomedical Engineering (Minor: Computer Science)* 09.2023~06.2027

- **Comprehensive Evaluation Score: 92.4/100 | Ranked 2nd of 22 students** in the cohort; evaluation based on academic performance (50%) and research, competitions, leadership, and service (50%)
- **GPA: 3.7/4.3 | Selected as the sole student** from SJTU's Global Institute of Future Technology and **one of 30 among 18,584 SJTU undergrads** for Bolè Plan (Elite Scientific Talent Program), recognizing exceptional research potential

PUBLICATIONS

- [1] [EVA: a Long-Context Generative Foundation Model Deciphers RNA Design Principles](#). Yanjie Huang*, Guang Lv*, ..., Yi Yang†, Shuangjia Zheng† (2026). *Nature Machine Intelligence* (Under Revision)
- [2] [PhaseFlow: A Unified Model for Multi-scale Biomolecular Condensates Understanding and Design](#). Yi Zhu*, Yanjie Huang*, ..., Shuangjia Zheng†, Yi Lin† (2026). *Nature* (Under Review)
- [3] [ProteoCraft: Proposal-Verification Search for Protein Design with Diffusion Structure Predictors](#). F. Wu*, H. Lin*, B. Duan*, Yanjie Huang*, Y. Li*, ..., Jure Leskovec†, Yejin Choi† (2026). *Neurips 2026* (Submitted)
- [4] [What Matters in RL for Diffusion Models? The Dominant Role of Noise](#). Da Xing, Haokai Zhao, Yanchao Li, Hanqun Cao, Yanjie Huang, ..., Yejin Choi†, Fang Wu† (2026). *Neurips 2026* (Submitted)
- [5] [Capturing Natural Evolution in RNA Design for in Vivo Functionality via Genomic Foundational Models](#). Yanjie Huang*, B. Zhang*, ..., Yi Yang†, Shuangjia Zheng† (2025). *Communications Biology* (Under Review)
- [6] [Energy-Adaptive Elastic Equivariant Transformer-SSM for Noise-Robust Protein Structure Representation](#). Zhongyue Zhang, Runze Ma, Yanjie Huang, Shuangjia Zheng† (2025). *Neurips 2026* (Submitted)

RESEARCH EXPERIENCES

Stanford University | Research Assistant (Advisors: Prof. Yejin Choi & Dr. Fang Wu)

Developing training code and conducting ablation study

Self-evolving diffusion model via Test-time On-policy Self Distillation

05.2026~Present

○ [Source Code](#) | Co-first author | Self-evolving | OPSD | Unified diffusion model

- Built the first version of distributed On-policy Self Distillation (OPSD) pipeline for unified diffusion, covering on-policy sampling, self-distillation, LoRA finetuning and evaluation.
- Implemented ablations over noise, prompts, seeds, denoising steps, and teacher dynamics to diagnose quality/stability and speed iteration, achieving 2.6x training efficiency and 20% performance improvement.

Optimizing model architecture and benchmarking

ProteoCraft: Proposal-Verification Search for Protein Design with Diffusion Structure Predictors^[3] ★ 03.2026~05.2026

○ [Source Code](#) | Co-first author | Hallucination | Inference-time search | Binder/TCR design | Neurips (Submitted)

- Built reproducible benchmarks for binder, TCR CDR redesign across search strategies and diffusion-based & hallucination-based baselines. (Outperformed BindCraft, PXDesign-d and RFDiffusion3 on success rate).
- Drove model architecture design, including template design, domain redesign, and Monte Carlo Tree Search and reduced memory usage by 40%.

Training model for atom-level reasoning ability & Synthesizing multi-scale reasoning data

Proteo-R2: Multi-scale Reasoning Foundation Models for Biomolecule Design

05.2026~Present

○ [Source Data](#) | Reasoning model | Alignment | Multi-modal

- Built programmatic pipelines solely to generate, filter, and standardize reasoning traces for alignment (aligning text and molecule embedding), mid-training (learning structural and spatial information) and joint training.
- Trained model for residue-level and atom-level reasoning ability and translated *reasoning ability* into controllable Protein/RNA/DNA/Ligand design.

Columbia University | Research Team Leader (Advisors: Prof. Erran Li & Dr. Fang Wu)

Synthesizing reasoning data

Proteo-R1: Reasoning Foundation Models for De Novo Protein Design

03.2026~04.2026

★ denotes work of exceptional impact and significance.

- [Source Code](#) | Reasoning model | Alignment | Antibody design | ICML (*Accepted*)
- Synthesized reasoning data for alignment, midtraining (most notably improving the index ability by 40%) and joint training and translated understanding into controllable antibody design.

Tsinghua University | Research Team Leader (Advisor: Prof. Yi Lin)

Developing and training model & Processing data & Benchmarking

PhaseFlow: A Unified Model for Multi-scale Biomolecular Condensates Understanding and Design^[2] ★ 05.2025~07.2026

- [Source Code](#) | Unified diffusion model | The first high-throughput phase-separating protein dataset
- Built data/QC pipelines for the first high-throughput phase-separation dataset, producing model-ready supervision and performed analysis to uncover driving features of protein phase separation.
- Built PhaseFlow, a unified multi-modal model solely (transformer for sequence generation and flow matching for phase diagram prediction), achieving better performance than DDPM and transformer-only models.
- Realized Inference-time scaling for controllable and high-quality protein design (improving 30% self-consistency).

Shanghai Jiao Tong University | Research Team Leader (Advisor: Prof. Shuangjia Zheng)

Developing and training model & Active Learning & Protein Design

De Novo Protein Design Incorporating Non-canonical Amino Acids 05.2025~07.2026

- [Source Code](#) | 2 Successful Design Cases & 14-Fold Activity Enhancement & Solved Crystal Structures
- Implemented DPO-based few-shot learning on experimental data to optimize protein design strategies with 30% improvement in high fitness protein enrichment.
- Constructed generator-predictor pipelines leveraging BoltzGen, Boltz2, and AlphaFold, yielding 2 successful designs and 14x activity improvement.

Developing and training model & Curating data & Benchmarking & Interpretability analysis & Molecule Design

EVA: A Long-context Generative Foundation Model Deciphers RNA Design Principles^[1] ★ 07.2025~03.2026

- [Source Code](#) | 10x Structural Modeling Accuracy | **First RNA Scaling Law**
- Developed MoE architectures via MegaBlocks and the first version of training code, including pretraining (incorporating CLM and GLM) and midtraining (for Lineage conditioning), which achieved 30% lower perplexity than the traditional dense model.
- Curated and opened the largest and high-quality RNA sequence data, [OpenRNA v1](#), covering more than 15 types, 1.14 millions of full-length RNAs.
- Conducted benchmark on zero-shot ncRNA/mRNA/Protein fitness prediction and gene essentiality prediction, with EVA achieving SOTA performance on ncRNA/mRNA and generalizing to protein and gene.
- Conducted Mechanistic Interpretability analysis leveraging Sparse Autoencoders (SAE), identifying latent RNA structure/function features (Spearman>0.9) and steering these features to controllable design.
- Realized De novo design and optimization of tRNA, mRNA, circRNA, RNA aptamers, and CRISPR-Cas systems. Our best designed aptamers improve both fluorescence and binding affinity by 2x over the positive control.

Developing model & Benchmarking & RNA Design

Capturing Natural Evolution in RNA Design via Genomic Foundational Models^[5] 07.2024~03.2025

- [Source Code](#) | 2.6x *in vivo* Performance Improvement | Multi-Modal Model
- Optimized fluorescent RNA aptamers using RNA inverse folding models and RNA language models, achieving 2.6-fold enhancement in fluorescence intensity for multiple variants.
- Benchmarked existing RNA inverse folding models and RNA language models on zero-shot RNA fitness prediction, and found the multi-modal model to achieve the highest spearman on fitness prediction.

Massachusetts Institute of Technology | Research Assistant (Advisor: Dr. Jack Xiaoyu Chen)

Building Agentic System

SynthArchitect: Agent for Plasmid Design and Laboratory Automation 08.2025~11.2025

- [Source Code](#) | Automated Lab | LIMS | Plasmid construction
- Built agent integrating 20+ tools including Golden Gate, DNA Cauldron, and other bioinformatics tools for automated plasmid engineering.

WORK EXPERIENCES

Lingang Laboratory | Research Intern, AI Research Division | Shanghai

Developing software to facilitate the design and optimization of MCP

[Origene: A Self-Evolving Virtual Disease Biologist Automating Therapeutic Target Discovery](#) 08.2025~11.2025

- [Source Code](#) | Contributor | Self-Evolving Agent | Multi-Agent System
- Developed software to facilitate the design and optimization of MCP with structured tool-call feedback, schema-based tool-use rules, and modular interfaces for collaborative development.

COMPETITION AWARDS

2022 Chinese Mathematical Olympiad (CMO) | Bronze Medal 12.2022

2021 The 7th International Mathematical Modeling Challenge (IMMC) | First Prize (International Round) 07.2021