WeChat

Heng Yang PhD

+86-19842750173 | +44-7878711663 | yangheng2021@gmail.com
Homepage | LinkedIn | GitHub | Hugging Face | Google Scholar
Department of Computer Science, University of Exeter, EX4 4RN, United Kingdom
Timezone: UTC+1 (e.g., 12:00 PM in the UK corresponds to 7:00 PM in China).

Latest Version

Research Summary

- **Large Language Models (LLMs):** Pioneered **InstOptima**, the first multi-objective optimization framework using LLMs as operators to synergistically enhance prompt diversity, performance (up to **16.8%** gain), complexity (perplexity reduced by **0.07**), and conciseness (average reduction of **200 tokens**). Proposed a GRPO-based framework for RNA sequence design. Proficient in the LLM technology stack.
- **Pre-trained Language Models (PLMs):** Six years of experience in PLM architecture and fine-tuning. Led the development of DeBERTa-v3-base-ABSA-v1.1 (**>1.45 million downloads**) and multiple other pre-trained models. Expertise spans pre-training, data augmentation/synthesis (BoostAug), Code PLMs, textual adversarial attacks/defenses, and ensuring adversarial robustness and fairness.
- **Genome Foundation Models (GFMs):** Led the development of OmniGenome-186M / 2B and OmniGenBench. Increased the success rate of EternaV2 RNA design from 3% to **84%**. The framework supports the vast majority of Transformer and MoE models out-of-the-box and integrates LoRA and FlashAttention for efficient fine-tuning and application.
- **Aspect-Based Sentiment Analysis (ABSA):** Creator and maintainer of **PyABSA** (**1k+ GitHub stars**), a framework enabling one-click access to **31+** ABSA models and **30+** datasets. The proposed LSA model has achieved state-of-the-art performance for three consecutive years. Widely recommended by AI assistants like ChatGPT and DeepSeek as the top choice for a beginner-friendly open-source framework (validated via zero-shot queries).

Education & Honors

- **University of Exeter, PhD in Computer Science** — ***Funded PhD Scholarship; PhD Research Grant***
  Pre-trained Models, Large Language Models, and Genome Models — Exeter, UK; Sep 2021 – Oct 2025 (Expected)
- **South China Normal University, M.Eng. in Computer Science** — ***National Scholarship for Graduate Students (Top 1%)***
  Pre-trained Models, Sentiment Analysis — Guangzhou, China; Sep 2018 – Jun 2021
- **Yangtze University, B.Eng. in Computer Science** — ***Outstanding Undergraduate Thesis (Top 10%)***
  Natural Language Processing, Spatio-temporal Trajectory Clustering — Jingzhou, China; Sep 2014 – Jun 2018

Open Source Leadership

Published 18 GitHub Repositories · 500k+ LOC · 1.6k+ Stars · 194 Followers · Maintained 8 PyPI Packages · >1M Downloads

- **RNADesign-GRPO**
  The first open-source framework to integrate **Reinforcement Learning** with **Genome Foundation Models** for guiding RNA sequence design. Based on the **GRPO** algorithm, it trains a **Seq2Seq** agent for progressive sequence generation and optimization. It introduces a composite reward function that combines a **semantic consistency scorer** (LMScorer), based on a genome sequence-structure alignment model, with metrics for predictive validity and similarity. The framework features **vectorization** and **mixed-precision training** across its environment, model, and training components.
- **OmniGenBench** (***53k+ pip installs · 351 stars***)
  The first unified platform for **Genome Foundation Model** benchmarking and application. Enables training and evaluation on **123** downstream tasks with a single line of code. Provides modular templates for models, data, and metrics, all easily extensible without modifying the source code. Natively supports **LoRA, MoE, and FlashAttention**. It is the first framework to encapsulate common genomic tasks into pipelines, allowing users to apply GFMs to practical tasks like degradation rate prediction, translation efficiency forecasting, RNA structure prediction, and sequence design optimization with one-click, no expert knowledge required.
- **PyABSA** (***>500k pip installs · 1,041 stars***)
  A widely adopted open-source toolkit for Aspect-Based Sentiment Analysis. Supports **31+** models and **30+** open-source ABSA benchmark datasets, enabling one-click fine-tuning, deployment, and inference. Reduces the time for prototyping and validating sentiment analysis systems by an estimated **>80%**. Its modular design, unified encapsulation, and easy-to-use API, along with built-in features like automatic data augmentation (self-developed BoostAug), make it highly accessible. PyABSA is recommended by leading AI assistants like ChatGPT and DeepSeek (in zero-shot queries) as a user-friendly framework for newcomers.

Hugging Face Portfolio

Released 9 Open-Source Models · Single Model with >1.45M Downloads · 10 HF Spaces · >150 Likes · 23 Followers

- **yangheng/deberta-v3-base-absa-v1.1** — *Mar 2022*
  Top-1 ABSA Model on Hugging Face. Featured in the Stanford University 2022 AI Index Report (p.83). — **1.45M+ Downloads**
- **yangheng/OmniGenome-186M** — *Apr 2024*
  *The first RNA sequence-structure alignment model; boosting RNA design success rate from 3% to **84%**.* — **67k+ Downloads**
- **Gradio-Blocks/PyABSA Space** — *May 2022*
  A comprehensive, interactive demo hub for ABSA. Featured as an official demo by Gradio-Blocks. — **142k+ Visits**

Competition Organization

- Organized the **ECML-PKDD Challenge on Genome Foundation Model Exploration (2025)**. Served as a Competition Chair in collaboration with ECML-PKDD 2025, coordinating co-chairs, managing the review process, and securing sponsorships. Led the creation of the first multi-species, multi-modal GFM benchmark, the **OmniGenomic Benchmark (OGB)**, covering 7 tasks (e.g., RNA sequence prediction, structure modeling, function inference, and DNA function prediction). Developed the codebase and designed the fully automated evaluation pipeline and CodaBench leaderboard. The competition attracted 20 participants with 227 valid submissions.

Representative Publications

Authored 12 First-Author Papers · Published 10 Papers · Citations: 625+ · h-index: 8

- **Bridging Sequence-Structure Alignment in RNA Foundation Models** — *CCF-A*
  ***Heng Yang**, Ke Li* — **AAAI 2025**
- **MPRNA: Unleashing Multi-species RNA Foundation Model via Calibrated Secondary Structure Prediction** — *CCF-B*
  ***Heng Yang**, Ke Li* — **EMNLP 2024**
- **The Best Defense is Attack: Repairing Semantics in Textual Adversarial Examples** — *CCF-B*
  ***Heng Yang**, Ke Li* — **EMNLP 2024**
- **InstOptima: Evolutionary Multi-objective Instruction Optimization via LLM-based Instruction Operators** — *CCF-B*
  ***Heng Yang**, Ke Li* — **EMNLP 2023**
- **BoostAug: Boosting Text Augmentation via Hybrid Instance Filtering Framework** — *CCF-A*
  ***Heng Yang**, Ke Li* — **ACL 2023**
- **An Interpretable RNA Foundation Model for Exploration Functional RNA Motifs in Plants** — *Nature Journal*
  *Haopeng Yu#, **Heng Yang#**, et al. (Co-first author)* — **Nat. Mach. Intell.**
- **DaNuoYi: Evolutionary Multi-Task Injection Testing on Web Application Firewalls** — *CCF-A*
  *Ke Li#, **Heng Yang#**, Willem Visser (Technical Lead)* — **IEEE TSE**
- **Modeling Aspect Sentiment Coherency via Local Sentiment Aggregation** — *CORE A*
  ***Heng Yang**, Ke Li* — **EACL 2024**
- **PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis** — *CCF-B*
  ***Heng Yang**, Chen Zhang, Ke Li* — **CIKM 2023**
- **A Multi-task Learning Model for Chinese-oriented Aspect Polarity Classification and Term Extraction** — *CCF-C*
  ***Heng Yang**, Biqing Zeng* — **Neurocomputing**
- **OmniGenBench: A Modular Platform for Reproducible Genomic Foundation Models Benchmarking** — *Preprint*
  ***Heng Yang**, Jack Cole, Ke Li* — arXiv Preprint
- **Robustness Meets Fairness: Investigating Adversarial Attack Effects on Alleviating Model Bias** — *Preprint*
  ***Heng Yang**, Geyong Min, Ke Li* — OpenReview Preprint
- **Tokenization or Featurization? Leveraging Language Models for Code Defect Prediction** — *Preprint*
  ***Heng Yang**, Ke Li* — Under Review

Research Experience — Spanning **Pre-trained NLP/Genome/Code Models**, **Sentiment Analysis**, **Text Augmentation**, and **Framework Development**

- **Large Language Models (LLMs).** Pioneered the application of **Genetic Algorithms** for **multi-objective instruction evolution** in LLMs (**InstOptima**), proposing novel **instructional crossover and mutation** operators. This work is among the first to systematically address the optimization of competing objectives such as instruction clarity, conciseness, and model performance. It achieved significant gains on downstream tasks, including performance (up to **16.8%** improvement), complexity (average prompt perplexity reduced by **0.07**), and conciseness (average prompt length reduced by **200 tokens**). Future work includes refining genetic operators (e.g., using another LLM for context-sensitive crossover/mutation), enhancing scalability (reducing fitness evaluation costs), and expanding objective dimensions (e.g., reward, fairness, creativity, and adversarial robustness). Developed the first GRPO-based framework for RNA sequence design, introducing a composite reward function that combines a semantic consistency scorer (LMScorer) with predictive validity and similarity. The framework features end-to-end vectorization and mixed-precision training. Accepted by **EMNLP 2023**.

- **Genome Foundation Models (GFMs).** Led pre-training research in **AI4Science**, focusing on **Genome Foundation Models and genetic code representation**. Addressed challenges like **sequence sparsity** and **single nucleotide polymorphisms (SNPs/SNVs)** by proposing a base-level semantic modeling approach. Conducted pre-training on RNA structure and sequence-structure alignment, significantly improving performance on tasks like **RNA structure prediction** (>20%) and **sequence design** (>80%). Independently managed the end-to-end pipeline, from **data collection and processing** to model architecture design and the construction of **training scripts and benchmarks**. Produced a series of papers, with three published in Nature Machine Intelligence, AAAI 2025, and EMNLP 2024. Developed the first comprehensive open-source framework, **OmniGenBench**, based on proprietary models and code. Created a modular benchmarking ecosystem by abstracting models, tokenizers, datasets, and metrics, complete with APIs, tutorials, and documentation to facilitate the rapid application of GFMs to genomic sequence processing.

- **Aspect-Based Sentiment Analysis (ABSA).** Independently conducted research on **fine-grained, aspect-based sentiment analysis**, covering tasks such as **aspect-sentiment-opinion triplet/quadruplet extraction**. Released several open-source models that gained significant traction on Hugging Face, becoming one of the platform's most widely used ABSA models. All work has been integrated into the **PyABSA** toolkit, which is widely recommended by AI assistants such as ChatGPT, Gemini, Claude, and DeepSeek. Published in CIKM 2023 and EACL 2024.

- **Text Data Augmentation.** Independently developed a novel **text augmentation technique** (**BoostAug**) for various text modeling tasks. The method filters synthetic samples using **global feature distribution (skewness)** combined with metrics like **perplexity, confidence, and hard labels**. Released as a **standalone PyPI tool**, it has garnered **180k+** installations. Achieved a **1%–2%** performance gain across all experiments on **8** public datasets, demonstrating superior stability and effectiveness compared to methods like EDA and NLPAug. Published in ACL 2023.

- **Text Adversarial Attack & Defense.** Systematically investigated **textual adversarial attack and defense strategies** for pre-trained models lacking **adversarial training**. Proposed a state-of-the-art **defense technique** (**Rapid**) that integrates **adversarial example detection and text restoration**, based on common adversarial patterns. Published in EMNLP 2024.

- **Software Defect Prediction.** Proposed the **LMDP** framework, the first to apply **pre-trained language models** to cross-project and within-project defect prediction. Introduced multi-task objectives like **”Corrupted Code Detection”** to mitigate feature sparsity and improve localization granularity. On **10 CPDP** and **13 WPDP** benchmarks, achieved an average **AUC increase of 4.6%** and **F1 score increase of 5.3%**, supporting **line-level defect localization** and significantly outperforming AST/GNN-based methods. The framework is AST-free, plug-and-play, and compatible with various PLMs like CodeT5 and CodeBERT.

- **Synergizing Robustness & Fairness.** Conducted the first systematic exploration of the impact of **fairness-agnostic adversarial attacks** on PLM bias (**AdvFairness**). Established a **three-stage evaluation pipeline (Benign→Attack→Defend)** and validated on three real-world datasets that while attacks decrease accuracy, they consistently **reduce group disparities (SPD drop of 0.07–0.14)**. Further **adversarial training** can **recover up to 40% of accuracy** while continuing to improve fairness, revealing a potential **”Pareto frontier”** between robustness and fairness.

- **Web Application Firewall (WAF) Injection Security Testing.** Proposed **DaNuoYi**, the first automated WAF testing framework supporting **multi-task, multi-lingual injection**. It utilizes **15 pairs of injection translation models** for shared semantics and a **multi-task evolutionary algorithm** (with a shared mating pool and 6 semantic-preserving mutation operators) to cooperatively generate test cases for six injection types: **SQLi, XMLi, PHPi, OSi, XSSi, and HTMLi**. On three real-world WAFs (**ModSecurity, Ngx-Lua-WAF, Lua-Resty-WAF**), it increased the number of bypass cases by **3.8×–5.78×** on average, significantly outperforming SQLMap and single-task baselines. Published in IEEE TSE 2024.

- **Open-Source Framework Development & Project Maintenance.** Possess comprehensive practical experience in open-source project management and scientific software architecture, having led the full lifecycle of multiple projects from architectural design and software implementation to long-term maintenance and community management.
  - **Framework Design:** Adhere to a design philosophy of modularity, extensibility, and interface abstraction. For example, in **OmniGenBench**, core functions are modularized and extensions are plugin-based, allowing researchers to seamlessly integrate custom models, data, and evaluation metrics without altering the core code. In **PyABSA**, complex calls are simplified to single-line commands via highly encapsulated APIs, significantly lowering the barrier to entry.
  - **Implementation & Project Management:** Demonstrated ability to drive complex projects from concept to completion. As **Chair of the ECML-PKDD 2025 Genome Foundation Model Challenge**, I oversaw competition planning and management, cross-party collaboration, design of the automated evaluation pipeline, and sponsor outreach. All projects emphasize continuous integration, comprehensive documentation (e.g., ReadTheDocs), and active community support to ensure long-term usability and robustness.

Invited Talks

- Google Health, Genomics Team: **OmniGenome** — Sep 2025
- Mila - Quebec AI Institute, Multi-omics Team: **OmniGenBench** — Dec 2024

Appendix: Original Open Source Projects & Models

The following content is optional and may be disregarded.

This appendix systematically catalogues the **original open-source projects, publicly released pre-trained models, and associated software tools** that I have independently developed or led during my research. The content spans application framework design, algorithm implementation, and interactive demonstrations, with corresponding publication statuses or current states indicated. **Some content in this appendix may overlap with the main body of the CV; please disregard any redundancies.**

GitHub Open Source Projects (https://github.com/yangheng95)

- **PyABSA**: Flagship framework for fine-grained sentiment analysis, supporting one-click use of 31+ models and 30+ datasets. **[CIKM 2023, PyPI Index]**
- **ABSADatasets**: A public repository of datasets for ABSA and text classification. **[CIKM 2023]**
- **OmniGenBench**: An automated benchmark and evaluation pipeline for RNA/DNA foundation models. **[arXiv 2024, PyPI Index]**
- **AdvFairness**: A comprehensive evaluation framework for adversarial robustness and model fairness. **[OpenReview 2024]**
- **OmniGenome-Demo**: Online demos and tutorial scripts for OmniGenome. **[AAAI 2025]**
- **Rapid**: A collection of experiments on textual adversarial attacks and defenses. **[EMNLP 2024]**
- **PlantRNA-FM**: Codebase for the Plant RNA Foundation Model. **[Nat. Mach. Intell. 2024]**
- **MP-RNA**: Training scripts for the multi-species RNA foundation model. **[EMNLP 2024]**
- **InstOptima**: Implementation of the multi-objective evolutionary algorithm for instruction tuning. **[EMNLP 2023]**
- **BoostAug**: Implementation of the BoostAug text augmentation method and data generation package. **[ACL 2023, PyPI Index]**
- **LCF-ATEPC**: A multi-task learning model for Chinese-oriented ABSA. **[EACL 2024]**
- **CodeT5DefectDetection**: Data and model experiments for code defect detection. **[TSE 2025, in press]**
- **DaNuoYi**: A multi-task evolutionary injection testing tool for WAFs. **[IEEE TSE, PyPI Index]**
- **EMOO**: A foundational codebase for evolutionary multi-objective optimization. **[PyPI Index]**
- **findfile**: A fast, keyword-level file/directory search utility. **[PyPI Index]**
- **autocuda**: A script for automatic selection and management of CUDA devices. **[PyPI Index]**
- **metric-visualizer**: A unified interface for research metric visualization and logging. **[PyPI Index]**
- **RNADesign-GRPO**: A GRPO-based method for RNA sequence design. **[Pending]**

Hugging Face Models (https://huggingface.co/yangheng)

- **deberta-v3-base-absa-v1.1**: 186M parameters; a lightweight model for fine-grained sentiment analysis. **[ACL 2023, EACL 2024]**
- **deberta-v3-large-absa-v1.1**: 418M parameters; a large-scale ABSA classifier. **[ACL 2023, EACL 2024]**
- **OmniGenome-186M**: 186M parameters; a foundation model for RNA secondary structure prediction and design. **[AAAI 2025]**
- **OmniGenome-52M**: 52M parameters; a fast-inference version of the genome language model. **[AAAI 2025]**
- **OmniGenome-v1.5**: 186M parameters; universal genome foundation model v1.5. **[Pending]**
- **MoEOmniGenomeV2**: 577M/2B parameters; a MoE-architected large genome model. **[Pending]**
- **MP-RNA**: 186M parameters; a multi-species RNA prediction model. **[EMNLP 2024]**
- **PlantRNA-FM**: 35M parameters; an RNA foundation model for mining functional motifs in plants. **[Nat. Mach. Intell. 2024]**

Hugging Face Spaces

- **PyABSA Space**: An interactive demo hub for ABSA models, invited to be an official Gradio-Blocks featured project. **[CIKM 2023]**
- **OmniGenBench**: The official online leaderboard, allowing researchers to easily trial the benchmark framework. **[arXiv 2024]**
- **Text-Adversarial-Attack-Defense (Rapid)**: An interactive demo for textual adversarial attack and defense. **[EMNLP 2024]**

PyPI Packages Published & Maintained (https://pypi.org/user/yangheng)

- **omnigenbench**: An automated benchmark and leaderboard for multi-task, multi-species GFMs. **[arXiv 2024]**
- **pyabsa**: A framework for fine-grained sentiment analysis (31 models × 30 datasets). **[CIKM 2023]**
- **boostaug**: A text augmentation toolkit based on instance filtering. **[ACL 2023]**
- **DaNuoYi**: A multi-task evolutionary injection testing framework supporting 6 types of WAF injection. **[TSE 2024]**
- (Four other PyPI projects mentioned previously are omitted here for brevity.)

Data as of August 5, 2025

— 拾级而上 —