Rohan Yadav

Education

2020–Present **Ph.D. in Computer Science**, *Stanford University*, Stanford, CA.
Advised by Alex Aiken and Fredrik Kjolstad

2015–2019 **BS in Computer Science**, *Carnegie Mellon University*, Pittsburgh, PA.
Advised by Umut Acar
Dean's List, University and SCS College Honors

Experience

2023-Present **Part-time Researcher**, *NVIDIA*, Santa Clara, CA.
- Working on parallel programming systems.

2024 **Research Intern**, *NVIDIA*, Santa Clara, CA.
- Researching techniques to effectively program emerging GPU architectures.

2023 **Research Intern**, *NVIDIA*, Santa Clara, CA.
- Researching compilation-based techniques to compose parallel programs in the Legate framework.

2022 **Research Intern**, *NVIDIA*, Santa Clara, CA.
- Developed `legate.sparse`, a distributed and accelerated drop-in replacement for `scipy.sparse`.

2019-2020 **Software Engineer**, *Cockroach Labs*, New York, NY.
- Improved stability and performance of CockroachDB's distributed SQL engine and schema management infrastructure.
- Contributed to development of a variety of large features in CockroachDB including ENUM types, User Defined Schemas, and Online Primary Key Changes.

2018 **Software Engineering Intern**, *Uber Advanced Technologies Group*, San Francisco, CA.
- Developed infrastructure for a migration from an internal data center to AWS.
- Implemented a file access system within AWS for integration with existing data center services.
- Dramatically enhanced scalability of batch compute jobs processing internal data.

2017 **Software Engineering Intern**, *Facebook*, Menlo Park, CA.
- Developed system to perform disruptive upgrades on network switches.
- Added packet subscription service for network switch agent debugging and maintenance.

Selected Research Projects

2025 **Overheads in Task-Based Runtime Systems** *with Michael Bauer, Joseph Guman, Michael Garland, Alex Aiken, Fredrik Kjolstad*
Developed techniques to enable task-based runtime systems to support fine-grained heterogenous workloads with performance competitive to low-level systems like MPI through a theoretical connection to actor-based programming models.

2024 **Programming Languages for Tensor Core GPUs** *with Michael Bauer, Alex Aiken, Michael Garland*
Developing programming language techniques to manage the asynchrony and hierarchy in modern GPUs that contain accelerators within the SM, such as the Tensor Core and TMA within the Hopper GPU architecture.

2024 **Automatic Tracing in Task-Based Runtime Systems** *with Michael Bauer, David Broman, Michael Garland, Alex Aiken, Fredrik Kjolstad*
Developed dynamic program analyses to automatically apply the tracing optimization in task-based runtime systems, enabling significantly reduced runtime overhead at scale in complex distributed applications.

2023 **Composing Distributed Computations Through Task and Kernel Fusion** *with Michael Bauer, Shiv Sundram, Wonchan Lee, Michael Garland, Alex Aiken, Fredrik Kjolstad*
Developed dynamic program analysis techniques to fuse computations across library boundaries on distributed machines, enabling applications built through the composition of high-level libraries to approach the performance of hand-written code.

2022 **Legate Sparse** *with Michael Bauer, Wonchan Lee, Manolis Papadakis, Melih Elibol, Michael Garland*
Developing `legate.sparse` a distributed and GPU-accelerated drop-in replacement for `scipy.sparse`, enabling supercomputer scale performance from high-level Python code.

2021-2022 **Compiling Tensor Computations to Supercomputers** *with Fred Kjolstad, Alex Aiken*
Developed DISTAL, a compiler for sparse and dense tensor algebra that targets distributed systems.

2020 **Automated Mapping of Computation and Data** *with Alexandra Henzinger, Thiago Teixeira, Alex Aiken*
Developed system to automatically discover strategies for mapping computation and data onto different processors and memories in a heterogenous system.

rohany@cs.stanford.edu • *rohany.github.io*

2018-2019 **Disentanglement** *with Sam Westrick, Umut Acar*
Designed efficient memory management systems for the memory access patterns of fork-join parallel programs.

Publications

PLDI 2025 **Task-Based Tensor Computations on Modern GPUs** *Rohan Yadav, Michael Garland, Alex Aiken, Michael Bauer*

ASPLOS 2025 **Automatic Tracing in Task-Based Runtime Systems** *Rohan Yadav, Michael Bauer, David Broman, Michael Garland, Alex Aiken, Fredrik Kjolstad*

ASPLOS 2025 **Composing Distributed Computations Through Task and Kernel Fusion** *Rohan Yadav, Shiv Sundram, Wonchan Lee, Michael Garland, Michael Bauer, Alex Aiken, Fredrik Kjolstad*

SC 2023 **Legate Sparse: Distributed Sparse Computing in Python** *Rohan Yadav, Wonchan Lee, Melih Elibol, Manolis Papadakis, Taylor Lee-Patti, Michael Garland, Alex Aiken, Fredrik Kjolstad, Michael Bauer*

SC 2023 **Automated Mapping of Task-Based Programs onto Distributed and Heterogenous Machines** *Thiago S. F. X. Teixeira, Alexandra Henzinger, Rohan Yadav, Alex Aiken*

SC 2022 **SpDISTAL: Compiling Sparse Distributed Tensor Computations** *Rohan Yadav, Alex Aiken, Fredrik Kjolstad*

PLDI 2022 **DISTAL: The Distributed Tensor Algebra Compiler** *Rohan Yadav, Alex Aiken, Fredrik Kjolstad*

OOPLSA 2021 **Compilation of Sparse Array Programming Models** *Rawn Henry, Olivia Hsu, Rohan Yadav, Stephen Chou, Kunle Olukotun, Saman Amarasinghe, Fredrik Kjolstad*

POPL 2020 **Disentanglement in Race-Free Nested Parallel Programs** *Sam Westrick, Rohan Yadav, Matthew Fluet, Umut A. Acar*

Undergraduate Thesis **Disentanglement, Theory and Practice** *Rohan Yadav*

SPAA 2019 **Brief Announcement: A Parallel Algorithm for Subgraph Isomorphism** *Rohan Yadav, Umut A. Acar*

Submitted for Publication (Under Review)

On The Duality of Task And Actor Programming Models *Rohan Yadav, Joseph Guman, Sean Treichler, Michael Garland, Alex Aiken, Fredrik Kjolstad, Michael Bauer*

Mentoring

2025- **Ahmad Zafar** *Stanford BS*
Integrating METIS-based partitioning methods into the Legion programming ecosystem.

2025- **Rohan Chanani** *Stanford BS*
GPU accelerating partitioning computations within the Legion runtime system.

2024 **Joseph Guman** *Stanford BS and MS, 2024. Now: NVIDIA*
Static specialization techniques to reduce overhead in distributed runtime systems.

Awards

2024 Jane Street Graduate Research Fellowship (Finalist)
2023 NVIDIA Graduate Research Fellowship
2020 NSF Graduate Research Fellowship
2019 CRA Outstanding Undergraduate Researcher Nominee
2019 Carnegie Mellon Senior Leadership Recognition
2015 Presidential Scholar Semifinalist

Talks

Task Based Tensor Computations on Modern GPUs
- Stanford Portal Affiliates Meeting, June 2025
- PLDI 2025, June 2025

Automatic Tracing in Task-Based Runtime Systems
- ASPLOS 2025, April 2025

Computing Distributed Computations Through Task and Kernel Fusion
- ASPLOS 2025, April 2025
- Charm++ Workshop 2024, April 2024

Legate Sparse: Distributed and Accelerated Sparse Computing in Python

- SIAM Parallel Processing, March 2024
- UW PLSE Seminar, December 2023
- SC 2023, November 2023
- UIUC Compilers Seminar, October 2023
- MIT Fast Code Seminar, October 2023
- CMU Catalyst Group Meeting, October 2023
- Berkeley Programming Systems Seminar, September 2023
- Stanford HPC-AI Advisory Council, February 2023

SpDISTAL: Compiling Sparse Distributed Tensor Computations

- Legion Retreat, December 2022
- AHA Affiliates Retreat, December 2022
- SC 2022, November 2022
- Stanford Software Research Lunch, April 2022

DISTAL: The Distributed Tensor Algebra Compiler

- Google Research, November 2022 (Invited)
- PLDI 2022, June 2022
- Vienna University of Technology, April 2022 (Invited)
- Stanford Agile Hardware Project Group Meeting, Jan 2022
- Cerebras Systems, Dec 2021 (Invited)
- Oxford Tensor Computations Seminar, Nov 2021
- Stanford Software Research Lunch, Nov 2021

On the Automated Mapping of Computation and Data Onto Heterogenous Machines

- Stanford Software Research Lunch, Feb 2021
- Legion Developer Meeting, Jan 2021

A Parallel Algorithm for Subgraph Isomorphism

- SPAA 2019, Jun 2019

Disentanglement, Theory and Practice

- CMU Meeting of the Minds, May 2019

Service

2025 **JDPC External Reviewer**

2025 **SPAA'25 Program Committee**

2024- **Stanford Software Research Lunch Organizer**

2024 **Stanford Faculty Hiring Committee**

Teaching

2025 **Teaching Assistant** *Stanford CS242* Programming Languages

2023 **Teaching Assistant** *Stanford CS143* Compilers

2021-2022 **Teaching Assistant** *Stanford CS242* Programming Languages

2017-2018 **Head Teaching Assistant** *CMU 15210* Parallel Algorithms and Data Structures

2016 **Teaching Assistant** *CMU 15150* Functional Programming

2018-2020 **Diderot**
Developed and maintained a new course management platform, now used by 1500 students daily at CMU.