ParslFest 2023

MoStream: Enabling ML-Guided Adaptive Molecular Simulations on Real-Time Stateful Stream Processing Systems

Jianshu Liu

Ph.D. candidate

Division of Computer Science and Engineering

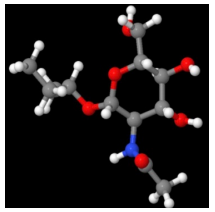
Louisiana State University

October 19, 2023

LSU THE UNIVERSITY OF CHICAGO Argonne NATIONAL LABORATORY

ML-guided Molecular Simulation

- Conventional molecular simulations are expensive

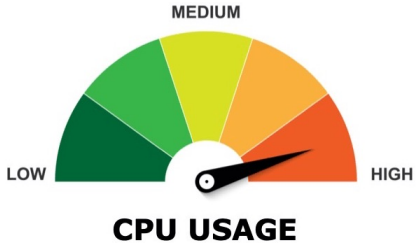


Molecular Simulation

Ionization potential (IP)

a key property for organic electrolytes design

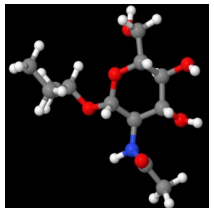
1) High computing resource usage



2) Time-consuming expert efforts

ML-guided Molecular Simulation

- Conventional molecular simulations are expensive



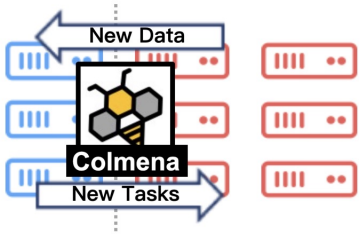
Molecular Simulation → **Ionization potential (IP)**

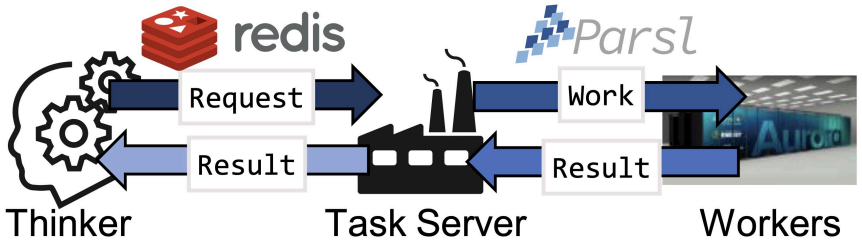
a key property for organic electrolytes design

AI/ML-based models → **Predicted IP** → Interesting Molecules

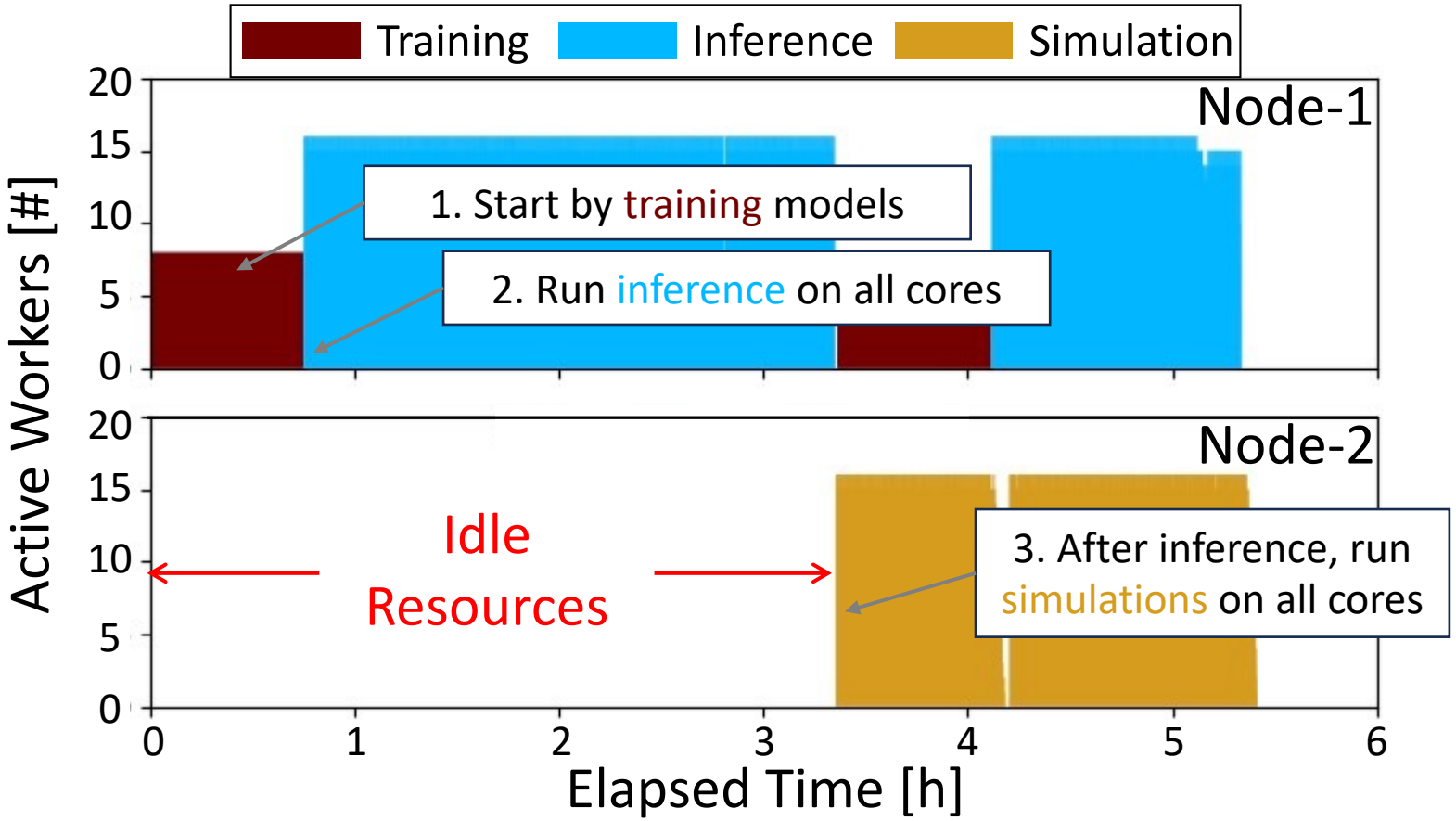
How can we organize ML-guided ensemble computations to maximize resource efficiency and timeliness of ML guidance?

Colmena: Steering Ensemble Simulations in a Sequential Manner



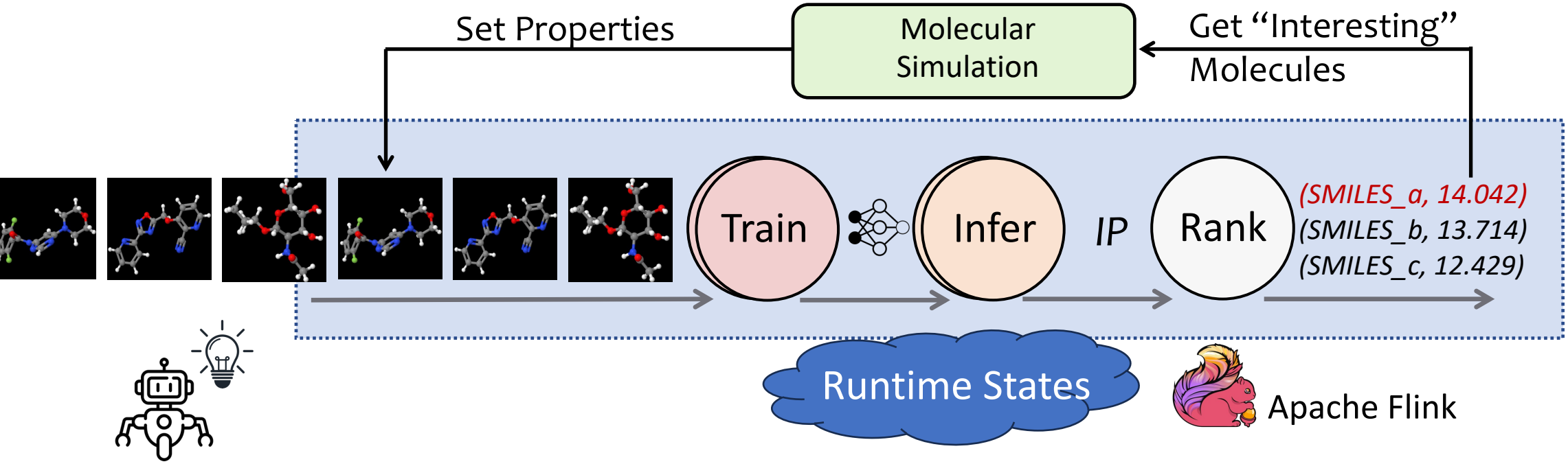


Colmena: Steering Ensemble Simulations in a Sequential Manner



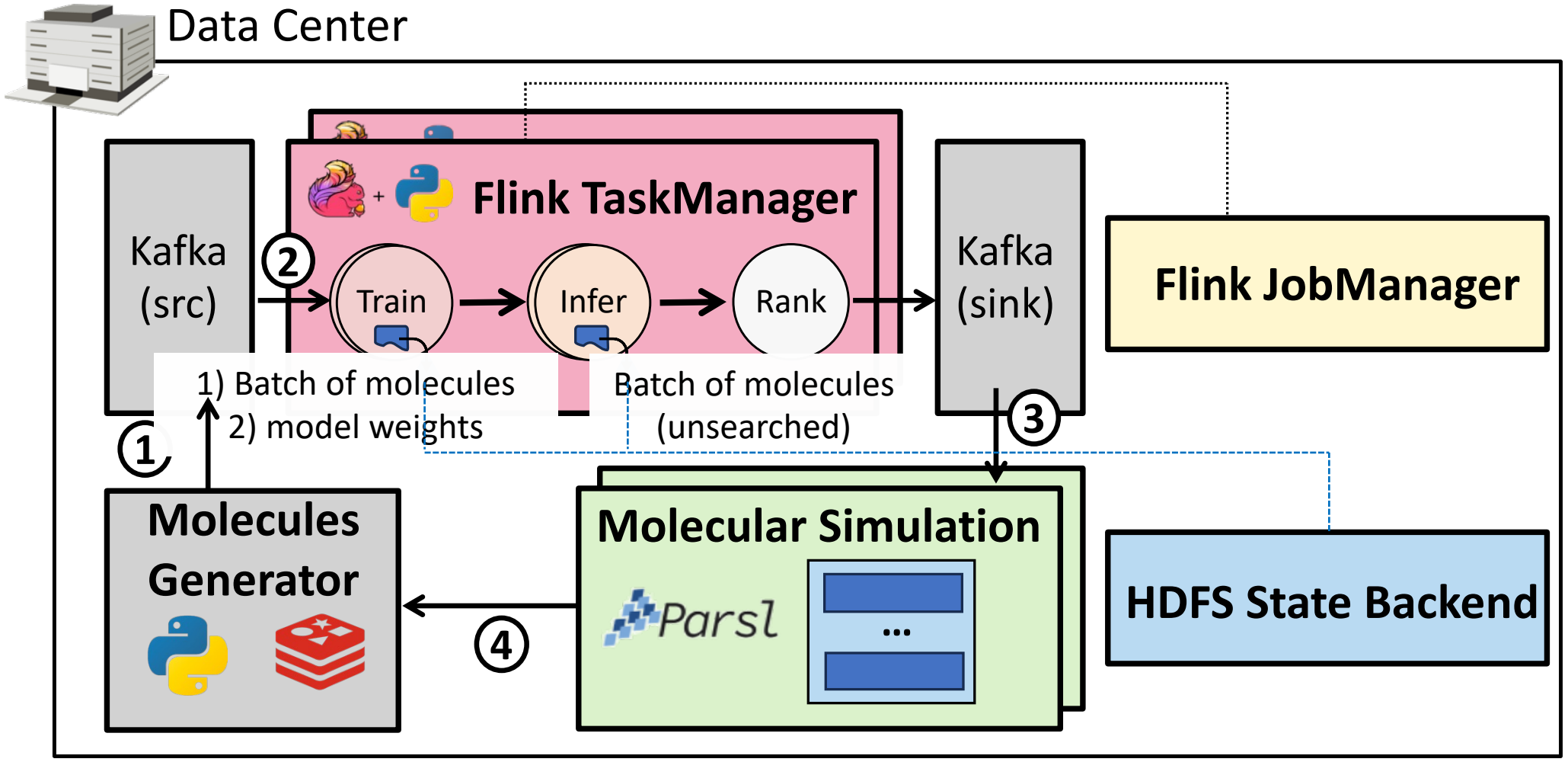
How about running these tasks in a streaming manner?

Applying Stream Learning using Apache Flink



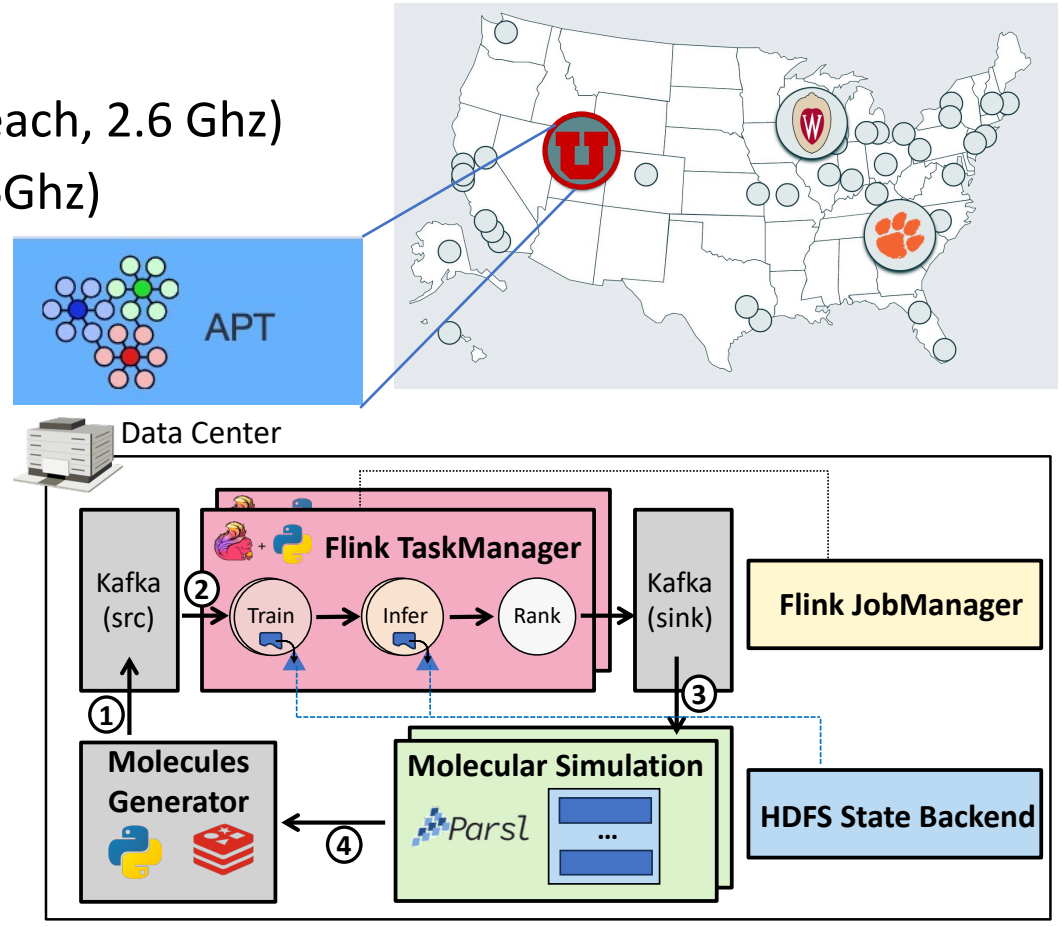
- **Iterative online ML training** based on continuous data stream (not static dataset)
- **Dynamic ML inferences** saves exploration time within the whole dataset
- **Timely recommendations** can guide molecular simulations

MoStream Framework



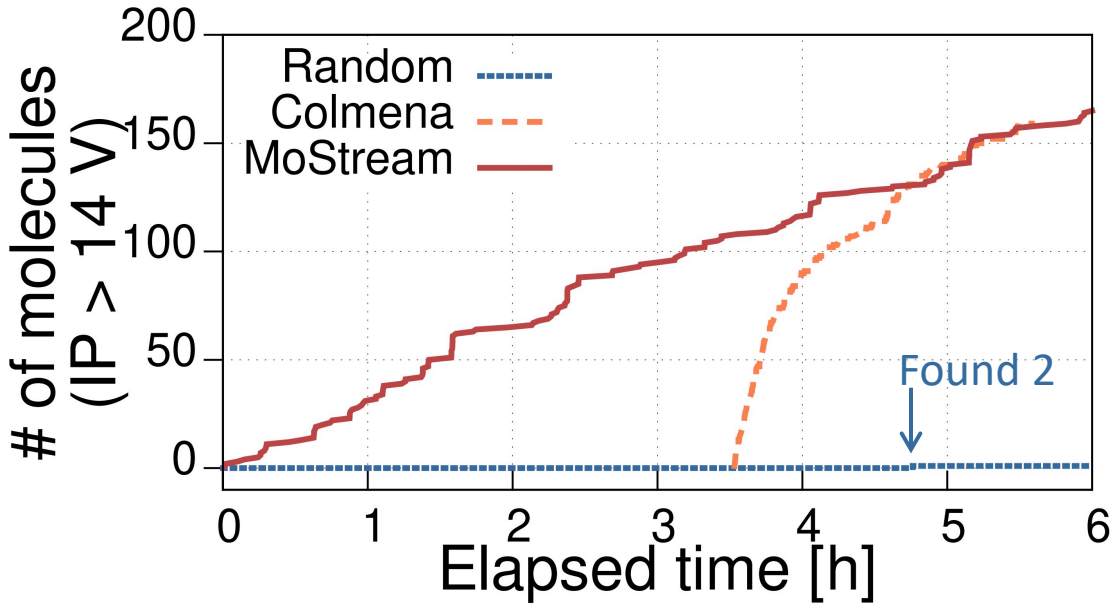
Deployment of MoStream on CloudLab

- Hardware Specification: (6 x c6220 nodes)
  - 16 vCPU, 2 x Xeon E5-2650v2 processors (8 cores each, 2.6 Ghz)
  - RAM 64GB Memory (8 x 8GB DDR-3 RDIMMs, 1.86Ghz)
- Software Stack
  - Message Queue: Kafka_2.11-1.1.1
  - Streaming Engine: PyFlink-1.17.1
  - Molecular Simulation: Parsl-2023.3.27



Effectiveness Evaluation

- Measure how many molecules with IP > 14 V were found over 6 hours

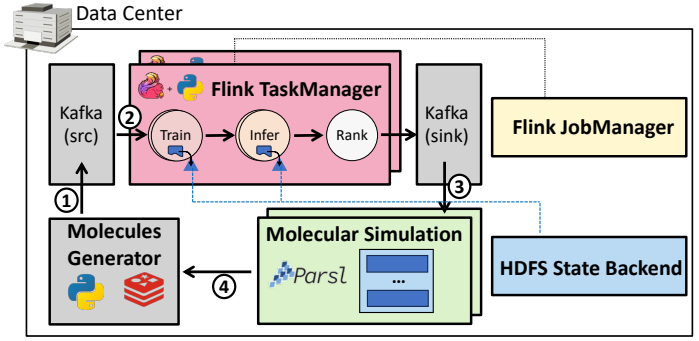


- "Random" identifies only 2 target molecules after 4.8 hours, success rate of 0.05%
- "Colmena" finds 159 molecules with resource wastage over the initial 3.5 hours
- "MoStream" finds 165 molecules with improved resource utilization

Conclusion and On-going works

What did we cover today?

- MoStream enables ML-guided molecular simulation on **stateful stream processing** system

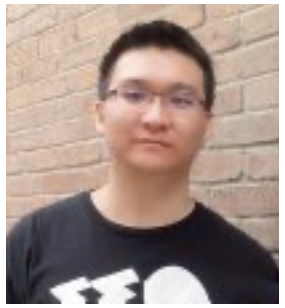


- Applying stream learning on Flink to support large(r) AI/ML models training

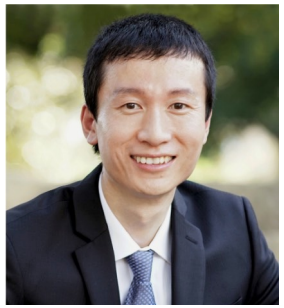
What do we plan to do?

- Studying the novelty and practice of MoStream in scientific research
- Optimizing performance and resource management of MoStream
- Research on impact of runtime states on online ML training

Acknowledgement



Jianshu Liu
LSU



Qingyang Wang
LSU

Kyle Chard
UChicago

Ian Foster
UChicago

Logan Ward
ANL

Any Questions?

jliu96@lsu.edu
https://jianshuliu1721.github.io/