Chenghan Sun

(+1)530-551-4025 sunchenghan1995@gmail.com, chesun@ucdavis.edu Chenghan Sun Chenghan-Sun

Education

University of California, Davis **Sep. 2017 – Dec. 2023**
Doctor of Philosophy (Ph.D.) in Chemical Engineering *Davis, CA*

University of California, Davis **Sep. 2019 – Jun. 2021**
Master of Science in Statistics (Data Science Track) *Davis, CA*

Iowa State University **Aug. 2014 – May. 2017**
Bachelor of Science in Chemical Engineering (Magna Cum Laude Academic Honor) *Ames, IA*

Work Experience

Meta Platforms, Inc. **June. 2022 – Sep. 2022**
Ph.D. Data Scientist Intern, Ads Problem DS Team, Business Integrity *Seattle, WA*

- Quantified, optimized, and implemented end-to-end data-driven infrastructure modifications during the political ads enforcement by suppressing component-level classification models for high parent-ads components; This solution impacted and implemented by the core infrastructure team before the intern ends with an improved precision of **7.5%**.
- Applied different strategies (data-specific, fine-tuning models, and multi-task learning) and build multiple country-level NLP models which span various embeddings and encoders to improve performance of the legacy model; The final model applied for body-text classification task improved average recall at same (80%) precision by at least **5%+** for several targeting countries.
- Built and optimized ETL data pipelines to boost up query efficiency by **65%** with more than 100TB total processed data; Implemented a bootstrapping method in Presto SQL engine to construct confidence intervals for political ads recall dashboard.

Research Projects

Dissertation: Applications of Machine Learning for Multiscale Atomistic Modeling in Catalysis Informatics

- Generated, processed, and analyzed high-throughput catalytic material datasets calculated from quantum simulation methods by constructing data pipelines through multiple High-Performance Computing (HPC) servers and GPUs for scalable computations.
- Designed and applied hybrid structural representations of molecules graphs by leveraging both geometric and physical information on atomistic trajectories for generalized AI4Science catalysis informatics tasks (e.g., adsorption energies predictions, partial atomistic charge predictions for pharmaceutical molecules and proteins, etc.).
- Applied LASSO, Kernel-based methods, ensemble methods (XGBoost, LightGBM) and deep neural networks (DeepMDkit, GPUMD, etc.), combining with active learning, reinforcement learning, and transfer learning to develop specific machine-learned force-fields and publish pre-trained models for general novel catalysts discoveries and acceleration of quantum simulations.

Selected First-author Publications

- **Chenghan Sun**, Sonti Siddharth, Zekun Chen, Davide Donadio, Surl-Hee Ahn, Ambarish R. Kulkarni. "Elucidating the Fluxionality and Dynamics of Zeolite-confined Au Nanoclusters using Machine Learning Potentials" (*preprint*). Accepted by ACS 2023 Conference, AIChE 2023 Annual Meeting. Submitted to *Journal of the American Chemical Society*.
- **Chenghan Sun**, Rajat Goel, Ambarish R. Kulkarni. "Developing Cheap but Useful Machine Learning based Models for Investigating High-Entropy Alloy Catalysts" (*Paper link*). Published by *Langmuir*
- Wang-Yeuk Kong, **Chenghan Sun (co-first author)**, Zekun Chen, Ambarish R. Kulkarni, Davide Donadio, Dean J. Tantillo. "Efficient Prediction of Partial Charges with a Size Extensive Multi-objective Deep Neural Network". Pending submission to *Journal of Chemical Information and Modeling*.

Selected Data Science Projects

Deep Learning for Drugs Multi-Classification Based on Biological Activities **Github Repo Link**

- Implemented a TensorFlow-based, state-of-the-art deep learning architecture named Neural Oblivious Decision Ensemble (NODE) to achieve an improved log-loss score of 0.017 in Kaggle competition.

Extensive Yelp Data Analysis and Data Interface Visualizations **Github Repo Link**

- Implemented a Scrapy-based Yelp web crawler module to collect raw restaurants data spanning various cities in California.
- Performed a comprehensive workflow of exploratory data analysis, data filtering & feature selection; Implemented several machine learning algorithms (KNN, SVM, RF, XGBoost) to deliver data insights through graphical dashboard.

Technical Skills

Programming Languages: Expert in Python, SQL, R, Linux/Bash, MATLAB; Experience with C++, Java, Julia.
Frameworks: Scikit-Learn, XGBoost, lightGBM, TensorFlow/Keras, Pytorch, Pandas, NumPy, SciPy, Visualization Libraries (Matplotlib, Seaborn, Plotline), Hyperparameter optimization frameworks (Optuna, Hyperopt).
Knowledge: Presto SQL engine, Recommender Systems, A/B testing, Docker/Kubernetes, Git, Statistical Machine Learning, Deep Learning, High Performance Computing, Probability Theory, Convex Optimization, Product Sense.