He (Jason) Sun

EM: he.sun@wustl.edu | PN: 314 6828433 | Git: github.com/wushanyun64 | LI: linkedin.com/in/he-jason-sun-574350136

SKILLS

Programming Languages: Python (Pandas, NumPy, PyTorch, Scikit-learn, Matplotlib, Fastai), SQL (PostgreSQL)

Tools: Jupyter Notebook, DFT calculation (VASP, CASTEP), Github, Langchain, Huggingface, LLamaIndex, Docker, Kubernets, Linux, Microsoft Excel, Microsoft Word, Google Cloud, Azure

EXPERIENCE

Corteva, Inc. | Data Scientist II | Nov 2022 - Now

- Spearheaded the development of "GenAE," an innovative adaptive experimental design platform, which significantly accelerated the formulation process for plant protection products. This initiative notably reduced dependency on external vendors, resulting in annual savings exceeding $1 million. Responsibilities included overseeing the platform's comprehensive architecture, managing data ingestion and featurization pipelines, developing surrogate models, and refining sampling/optimization mechanisms, search parameters, candidate generation processes, and continuous integration/continuous deployment (CI/CD) pipelines.
- Leader of the development of an automated regulatory document generation platform leveraging a large language model (LLM), enhancing efficiency by 50x and supporting over 100 countries and 20 document types, the project is highlighted in Corteva's stake holder day presentation.
- Pioneered the creation of "ChatPDF," an advanced internal tool leveraging LLM, computer vision models and retrieval augmented generation (RAG) technology for efficient extraction of information from complex PDF documents such as academic papers, regulatory documents, technical documents such as patents, etc..
- Initiated and led an internal LLM-focused journal club, aimed at enhancing the team's expertise and skills in the fields of Natural Language Processing (NLP) and Large Language Models.
- Played a pivotal role in formulating and standardizing technical guidelines within the Data Science department. Contributed valuable insights and recommendations that were integral to achieving consensus on technical standards, thereby improving the reliability, reusability, and transparency of data science initiatives.

Washington University in St. Louis | Grad/Ph.D. Researcher | Aug 2016 - Aug 2022

- Responsible for research, development and validation of novel computational methods (machine learning/quantum chemistry simulations) for the calculation of materials physio-chemical properties such as NMR spectrum parameters.
- Worked independently to perform research on generating and applying new feature engineering methods which suits the characteristic of solid materials based on domain knowledge.
- Collaborated with a team of computational chemist, developers and experimentalists to systematically benchmark the performance of quantum chemistry simulation methods. Constructed a database of material's properties using the validated method.
- Streamlined machine learning pipelines to test the performance of different combinations of features and model algorithms. The best model achieved (R2=0.98) significantly outperformed previous state-of-the-

art (SOTA) model (R2=0.93).
- Developed a module for statistical analysis of the posterior distribution of NMR parameters based on Markov Chain Monte Carlo.
- Performed density functional theory (DFT) calculation over thousands of materials and contribute to mainstream material database the Materials Project.

PROJECTS

Adaptive formulation design with Bayesian optimization | Project Lead | Aug 2022 - Now
- Developed an end-to-end Bayesian optimization model for new formulation development, the model is capable of handling very sparse and unbalanced data sets and is capable of suggesting highly desirable formulation candidates.
- Utilizing tree-based models (random forest/LGBM) instead of the Gaussian process to account for high dimensional and categorical feature space.
- Constructed feature generation and data processing pipeline from scratch including a molecular structure featurizer and multiple imputation strategies to improve the quality of the data set.

Machine learning prediction of NMR spectrum for solid materials | Research Project Lead | Aug 2018 - Jan 2022
- Performed data wrangling/ETL and exploratory data analysis (EDA) using open-source API and relevant Python libraries to parse and store research data.
- Created novel engineered features utilizing domain knowledge and building into modeling pipeline, increasing model performance from R2 = 0.93 to R2=0.98.
- Leveraging the data pipeline and engineered features, developed and tested machine learning models (random forest/XGboost) that outperformed published state-of-the-art (SOTA) models (R2~0.91) for predicting spectroscopic properties.
- Ensured the quality of training data with rigorous benchmarking study between quantum chemistry calculations and experimental data for over 3000 chemical compounds.
- Project Link: github.com/wushanyun64/27Al_CQ_prediction

Paddy Disease Classification Kaggle Competition | Kaggle | Jul 2022 - Aug 2022
- Built image recognition machine for fast detection of 9 different types of paddy disease using transfer learning techniques.
- Improved accuracy from the baseline's 0.87 (resnet) to 0.97 by iterating through CNN architectures using sampled datasets.
- Applied test time augmentation and ensemble learning, combining the top performance models in addition to managing RAM usage with gradient accumulation, resulting in an accuracy of 0.98 [top 6% on the leaderboard].

UNIFESP X-ray Body Part Classifier Kaggle Competition | Kaggle | Jul 2022 - Aug 2022
- Trained a convolution neural network (CNN) model using Pytorch pretrained architectures for X-ray picture recognition with totally 25 different body parts classes. Modified the network for multilabel classification, significantly reduced overfitting with customized early stop function.
- Identify the best architectures for medical images (densenet) with literature research. Explored the performance differences between finetuning and feature extracting. Reached the final F1-score of 0.86 for a test set of 750 X-ray pictures.

Reliability analysis of spectrum fitting model with Markov Chain Monte Carlo | Team Member | Jan 2020 - Mar 2022

- Developed a statistical module to assess the reliability of model parameters for computational simulation vs experimental spectrum.
- Collaborated with the development team to integrate the Monte Carlo module with an open-source spectrum fitting program Mrsimulator using Github Actions CI/CD pipelines.
- Project link: github.com/wushanyun64/monte_carlo_sampling_NMR_params

Enabling materials informatics for 29Si solid-state NMR of crystalline materials | Research Project Lead | Sep 2017 - May 2021

- Leveraged quantum chemistry approaches to construct a data infrastructure for silicon materials and confirm the validity of computational methods. (https://contribs.materialsproject.org/).
- Determined the optimal way of expressing spectroscopic data with statistical error analysis.
- Systematically corrected the machine-generated data via extensive benchmarking against experimental results. Proved a conceptual error in the popular quantum chemistry package VASP and improved the information reliability on the VASP wiki. (https://www.vasp.at/wiki/index.php/LCHIMAG).

PUBLICATIONS

- 2023 Sun, H., West, M., Dwaraknath, S., Ling, H. et al. 27Al NMR benchmarking with DFT and fast prediction of quadrupolar coupling constant from simple local geometry. (Under review with Nature Communication)
- 2021 Sun, H; Hammann, B; Brady, A; Singh, G; Housel, L; Takeuchi, E; Takeuchi, K; Marschilok, A; Hayes, S; Szczepura, L. Structural Investigation of Silver Vanadium Phosphorus Oxide (Ag2VO2PO4) and its Reduction Products. Chem Mater (2021). doi: 10.1021/acs.chemmater.1c00446
- 2021 Cendejas A J, Sun H, Hayes S E, et al. Predicting Plasma Conditions Necessary for Synthesis of γ-Al2O3 Nanocrystals[J]. Nanoscale, 2021. doi: 10.1039/D1NR02488D
- 2020 Sun, H., Dwaraknath, S., Ling, H. et al. Enabling materials informatics for 29Si solid-state NMR of crystalline materials. npj Comput Mater 6, 53 (2020). doi: 10.1038/s41524-020-0328-3
- 2020 Zahan, M., Sun, H., Hayes, S. E., Krautscheid, H., Haase, J., & Bertmer, M. (2020). Influence of Alkali Metal Cations on the Photodimerization of Bromo Cinnamates Studied by Solid-State NMR. The Journal of Physical Chemistry C. doi: 10.1021/acs.jpcc.0c09826
- 2020 Malone, M. W., Espy, M. A., He, S., Janicke, M. T., & Williams, R. F. (2020). The 1H T1 dispersion curve of fentanyl citrate to identify NQR parameters. Solid State Nuclear Magnetic Resonance, 110, 101697. doi: 10.1016/j.ssnmr.2020.101697

EDUCATION & CERTIFICATIONS

Ph.D. in Computational Chemistry | Washington University in St. Louis | Aug 2016 – Jul 2022

B.S. in Physical Chemistry | University of Science and Technology of China | Sep 2012 – Jun 2016

PRESENTATIONS

- 08/02/2022 He Sun, 27Al NMR chemical shielding and quadrupolar tensors benchmarking with DFT: machine learning prediction of quadrupolar coupling constants (CQ) from simple local geometry and ele-

mental properties. RMC conference 2022. (international conference, oral)
- 04/25/2022 He Sun, 27Al NMR quadrupolar and chemical shielding tensors benchmarking with DFT: prediction of quadrupolar coupling constants (CQ) from simple local geometry and elemental properties. ENC conference 2022. (international conference, oral)

AWARDS

- 2024 Corteva, Corteva R&D Hackathon 2024, Accelerate Performance & Execution, Category Winner
- 2014 MIT, International Genetically Engineered Machine Competition, Gold Medal

TEACHING & MENTORING

Washington University in St. Louis | Teaching Assistant | Aug 2016 - Aug 2018
- Gave over 50 lectures to more than 150 students on inorganic chemistry and general chemistry experiments.
- Prepared and graded homework and exam problem sets for inorganic chemistry lectures, answered questions from students during one-on-one sessions.
- Guided and helped students on their general chemistry experiments to ensure safe and correct lab operations. Received overall rating of 4.88/5 from students.