James ONeill

NLP Scientist — PhD Candidate in Bioengineering

✉ joneilliii@sdsu.edu linkedin.com/in/jimnoneill github.com/jimnoneill huggingface.co/jimnoneill

Work Experience

NLP Scientist, Computational Active Matter Mechanics Lab, SDSU 2020 - Present

- Pioneered CarD-T, a novel NLP framework achieving 20% higher recall than GPT4 in carcinogen classification within scientific literature. Discovered ˜1,600 new carcinogens not in major databases.
- Spearheaded large-scale biomedical literature analysis, processing 10TB+ of data, reducing research time by 60% for oncology teams.
- Engineered scalable ML pipelines on AWS/Azure, enabling real-time processing of 100,000+ scientific papers daily.
- Mentored 20+ junior researchers; 3 MSci researchers, resulting in 5 successful ML/NLP project launches and 3 conference paper acceptances.

Research Associate III, Lab for Pathogenesis of Clinical Drug Resistance, SDSU 2016 - 2020

- Developed ML-powered pipelines for sequence analysis, accelerating drug resistance detection by 40%.
- Led cross-functional collaboration, integrating diverse datasets to create a comprehensive antibiotic resistance database used by 500+ researchers globally.

Founder, PubVerse AI 2023 - Present

- Launching an NLP-driven graph network hybridized, proprietary algorithm projected to reduce literature review time by 70% for biomedical researchers; reduce the time from publication to patent by 60%.
- Developing AI algorithms capable of analyzing 1M+ papers monthly, identifying emerging research trends with 92% accuracy.

Education

PhD, Bioengineering, SDSU-UCSD Joint Doctoral Program Expected Aug 2025

MS, Biological & Medical Informatics, San Diego State University 2020

AA, English, Pasadena City College

Technical Skills

Languages: Python, R, Bash, C++, Javascript

ML/NLP: LLaMa, llama.cpp, PyTorch, TensorFlow, Hugging Face Transformers, NLTK, spaCy, Scikit-learn, training LLMs, Fine-Tuning LLMs, NER, sentiment analysis

Big Data/Cloud/HPC: Local parallelized deployment of Llama-3.1-405B across multiple GPUs, AWS EC2 P4, Azure, SLURM, OpenMP, MPI, CUDA, PRP/NRP Kubernetes, Apache Spark, Hadoop, PySpark, No/SQL, HDF5

Data Analysis: Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly, SQL, dynamic programming, unsupervised learning/unsupervised clustering, graph community detection

Other: Git, LaTeX, Virtual Environments, Obsidian, Unix/Linux Systems, RESTful API, web-scraping

Key Projects

CarD-T: State-of-the-art NLP framework for carcinogen identification (F1 score: 0.875)

PseudoGenius: Deep Learning tool for pseudogene classification in bacterial genomes

Certifications & Affiliations

- AWS Certified Cloud Practitioner

- Healthcare NLP for Data Scientists (John Snow Labs)

- Associate Member, American Association for Cancer Research (AACR)

- Ethics Reviewer, NeurIPS Conference (2024)

- Linux Cluster Institute: Certification (2022)

- μMBA: UCSD, Rady School of Management

- ”1st Prize” and also ”Best Team Work” - Big Data Hackathon of San Diego: Public Health (2017)

Additional Experience

SDSU Senior Deisign: Technical Advisor 2022 - Present

- Advised undergraduate senior engineering students on advanced technical aspects of their Senior Design Projects; resulting in 12 successful teams of 4 students, working alongside corporate sponsors.

Volunteer Work, SDSU Upward Bound Summer Academy 2023 - Present

- Taught high school students with little or no coding background Python, building an environment in VS Code, and Sentiment Analysis fine-tuning within a four week period.

Publications

Conkle-Gutierrez, D., Ramirez-Busby, S. M., Thosar, N., O'Neill, J., Espinoza, M. E., Ahmadi Jeshvaghane, M., & Valafar, F. (2025). Epigenetically mediated gene conversion drives antigenic variation in B-cell epitopes of Mycobacterium tuberculosis PE_PGRS proteins. *Microbiological Research*. Manuscript submitted.

Shaffar, S., Dorr, B., & O'Neill, J. (2025). Lessons from two hundred undergraduate engineering senior design capstone projects including the implications of cross discipline projects, and different types of project sponsors. *International Journal of Engineering Education, 41*(1), 1-25.

Tripathi, O., Parada, H., Sosnoff, C., Matt, G. E., Quintana, P. J. E., Shi, Y., Liles, S., Wang, L., Caron, K. T., O'Neill, J., Nguyen, B., Blount, B. C., & Bellettiere, J. (2025). Exposure to secondhand cannabis smoke among children. *JAMA Network Open, 8*(1), e2455963. https://doi.org/10.1001/jamanetworkopen.2024.55963

O'Neill, J., Reddy, G. A., Dhillon, N., Tripathi, O., Alexandrov, L., & Katira, P. (2024). CarD-T: Interpreting carcinomic lexicon via transformers (Version 2) [Preprint]. *medRxiv*. https://doi.org/10.1101/2024.08.13.24311948