James O’Neill, PhD

AI Scientist — Biomedical NLP & Machine Learning

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

Education

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

- **Dissertation:** “Biomedical Research Analytics through Temporal & Semantic AI”
- Research Focus: Development of NLP and ML methods for biomedical text processing, synonym disambiguation, and temporal network embeddings.

MS in Biological & Medical Informatics, San Diego State University Aug 2019

- Topic: Statistical & machine learning approaches for genomic classification.

BS in Biochemistry & Molecular Biology, University of California, Davis Dec 2015

- Minor in Nutrition

AA in English, Pasadena City College

Research Experience

AI Scientist, FAIR Data Innovations Hub, California Medical Innovations Institute Sept 2025 - Present

- **ENVISION Portal:** Developing an open-source platform for managing and sharing eye imaging data, ensuring datasets are AI-ready and reusable following FAIR principles.
- **Project Lead, posters.science:** Leading development of an AI-powered platform for AI-ready poster metadata extraction.

Head Biomedical AI Scientist / Research Specialist II, Computational Active Matter Mechanics Lab, SDSU
Aug 2025 - Present

- Leading biomedical AI research initiatives in the Mechanical Engineering Department. (CSU Classification: 5681A)

Chief Technology Officer, 42Degrees Journal Dec 2024 - Present

- Leading technical infrastructure and platform development for 42Degrees, an open-access scientific publishing platform.
- Overseeing AI-integrated hybridized semantic network analysis for publication landscape novelty, impact, and methodology metrics.

Head Biomedical AI Scientist, Computational Active Matter Mechanics Lab, SDSU 2020 - 2025

- **CarD-T:** Developed ”CarD-T: Interpreting Carcinomic Lexicon via Transformers,” an NLP framework utilizing transformer-based models for carcinogen identification, achieving high recall (0.857) and F1 score (0.875).
- **PCaR-D:** Created ”Probablistic Carcinogenic Denomination” or ”PCar-D”; a temporal analytical tool with Bayesian hierarchical modeling and Markov Chain Monte Carlo (MCMC) sampling for probabilistic evaluations of carcinogen discourse in literature based on evolving evidence, classifying ˜1,600 new carcinogens yet listed by major databases.
- **Syn-Lustre:** Created hybridized transformer and graph community detection based method for classifying context-specific synonymy levels, ”Synonym-Lustre”, with 88.30% accuracy.
- **Temporal Network Embeddings (TNE):** Built predictive models using TNEs to analyze and predict research trends in biomedical sciences, integrating temporal AI and semantic analysis.
- **PubVerse, Transdisciplinary Collaboration** Created proprietary self-supervised hybridized graph network & language model algorithm for transdisciplinary research collaboration, increasing success in achieving R1 grants by 180%.

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

- **Graph Modeling:** Applied network science to model antibiotic resistance pathways, uncovering novel resistance mechanisms.
- **Statistical Modeling:** Leveraged statistical methods for sequence analysis, accelerating drug resistance detection by 40%.
- **Data Wrangling:** Assembled and analyzed large-scale datasets using Hive and Spark.
- **Collaboration:** Integrated diverse datasets to create a comprehensive antibiotic resistance database used by over 500 researchers globally.

Co-Inventor, PubVerse (Licensed Technology) 2023 - Present

- **Algorithm Development:** Created proprietary self-supervised hybridized graph network & language model algorithm for transdisciplinary research collaboration.
- **Technology Transfer:** Licensed through SDSU Research Foundation Technology Transfer Office (50% royalty share).

Technical Skills

Programming Languages: Python, R, C++, Java, Bash

Machine Learning: PyTorch, KNN, K-Means, DBSCAN, TensorFlow, Hugging Face Transformers, TFX

Deep Learning: Transformer Models, Large Language Models (LLMs), Generative AI, Model Alignment, Reinforcement Learning

Natural Language Processing: Llama 3.0-3.3, ollama, llama.cpp, Named Entity Recognition (NER), sentiment analysis, NLTK, spaCy, Scikit-learn, transformers

Graph and Network Science: NetworkX, DGL, Graph Neural Networks, Temporal Network Embeddings

Statistical Modeling: Bayesian Methods, MCMC Sampling, Probabilistic Modeling

Data Engineering: Apache Spark, Hadoop, Hive, Presto, PySpark, SQL, NoSQL, HDF

Cloud Computing: AWS (EC2 P4, S3), Azure, SLURM, CUDA, PRP/NRP Kubernetes

Data Analysis: Pandas, NumPy, SciPy, Matplotlib, Seaborn

DevOps: Git, Docker, Kubernetes, RESTful APIs, Few-shot learning & finetuning Llama models, parallelized deployment of Llama-405b for personal use

Open Source Models & Datasets

CarD-T Model — CarD-T-NER Dataset

- Bio-ELECTRA-based NER model (335M params) for carcinogen identification from scientific texts.
- Trained on 19,975 annotated PubMed abstracts; achieves F1: 0.875, Precision: 0.894, Recall: 0.857.

BSG CyLlama — Training Dataset

- Biomedical Summary Generation through Cyclical Llama—corpus-level summarization using cyclical embedding averaging with named entity integration. LoRA fine-tuned on Llama-3.2-1B-Instruct.
- Trained on 19,174 clustered scientific abstract corpora for multi-document synthesis.

PseudoGenius

- BERT-based transformer for classifying gene sequences as normal or pseudogene, trained on Mycobacterium leprae. Achieves precision, recall, and F1 of 1.0 on test set.

Key Projects

CarD-T: An Automated Pipeline for Carcinogen Nomination

- Developed an NLP framework combining transformer-based ML with probabilistic analysis for carcinogen identification from scientific texts.
- Achieved superior recall (0.853) compared to GPT-4 (0.757), nominating ˜1,600 potential new carcinogens.
- Applied Bayesian Probabilistic Carcinogen Denomination (PCarD) analysis for temporal evaluation of carcinogenic evidence.
- Published in *Carcinogenesis*: doi: 10.1093/carcin/bgaf074

Synonym-Lustre: Biomedical Jargon Toolkit for Synonym Disambiguation

- Developed a toolkit for processing and disambiguating biomedical texts across diverse subfields.
- Leveraged transformer embeddings and graph community detection to calibrate to user-defined levels of synonymy.
- Validated using datasets like UMLS and IARC, achieving high completeness and homogeneity scores.

Temporal Network Embeddings for Research Trend Prediction

- Built temporal graph networks (TGNs) to model the evolution of biomedical research topics over time.
- Utilized TNEs to predict emerging research trends, aiding in strategic allocation of research funding.
- Validated models using 25 years of data, achieving high precision and accuracy in trend prediction.

Publications

O'Neill, J., Reddy, G.A., Dhillon, N., Tripathi, O., Alexandrov, L., Katira, P. (2025). *CarD-T: An Automated Pipeline for the Nomination and Analysis of Potential Human Carcinogens*. *Carcinogenesis*, bgaf074. doi: 10.1093/carcin/bgaf074

Intellectual Property

PubVerse: Contextual Graph Network for Biomedicine & Global Health Intervention Collaboration

- San Diego State University Research Foundation, Technology Transfer Office
- Co-Inventor/Author (50% royalty share); File: Katira P 1
- Copyright Assignment & Royalty Sharing Agreements executed May 2023
- Available for licensing via SDSURF InPart portfolio

Certifications & Affiliations

Certifications:

- AWS Certified Cloud Practitioner
- Healthcare NLP for Data Scientists (John Snow Labs)
- Linux Cluster Institute: Certification (2022)

Affiliations:

- Associate Member, American Association for Cancer Research (AACR)
- Ethics Reviewer, NeurIPS Conference (2024)

Awards:

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

Additional Experience

SDSU Senior Design: Technical Advisor, SDSU 2022 - 2025

- Advised undergraduate engineering students on advanced technical aspects of their Senior Design Projects.

Volunteer Work, SDSU Upward Bound Summer Academy 2023 - Present

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