




SIDDHARTH BETALA

+91-8619197880 | betalas5@gmail.com

 Siddharth Betala |  [sid-betalol](#) |  Siddharth Betala
Bengaluru, India

EXPERIENCE

• Entalpic

Machine Learning Research Engineer

January 2025 -

Paris, France (Remote)

- **LeMat-Synth[P.1]**: Designed and deployed integrated LLM- and OCR-driven pipelines leveraging frameworks like **DSPy** to extract chemical synthesis procedures from academic PDFs, combined with automated evaluation systems employing **LLM-as-a-judge** methodologies for benchmarking extraction quality and accuracy.
- Leading the development of **LeMat-GenBench[P.2]**, a benchmarking suite for generative materials models, introducing metrics for novelty, uniqueness, diversity, and stability to rigorously assess model performance in unconditional and property-guided materials discovery; soon to be integrated as a **leaderboard on HuggingFace**.
- Architecting an agentic retrieval pipeline combining **RAG** and **LangGraph** to systematically extract structured technological hierarchies from patent corpora and identifying technological gaps, significantly enhancing precision and efficiency for downstream analytics and decision-making tasks.

• Piramal Capital and Housing Finance Limited

Data Scientist, Risk Analytics, Business Intelligence Unit (BIU)

July - December 2024

Mumbai, India

- Pioneered a feature selection pipeline that leveraged **feature-engine** and **BorutaPy**, now **adopted as the standard across the division**, post-presentation at the monthly BIU Townhall in August 2024.
- Developed a home loan acquisition model, driving **Rs. 4.8 Billion** in additional business by reducing false negative-induced rejection rates, using a meta-learning ensemble of XGBoost and LightGBM.
- Built a bank statement narration tagging tool using **named entity recognition** and fine-tuned **LLaMA2**.
- Designed a two-track data science hackathon to identify top talent from 3 premier institutions and mentored the top 10 participants for the finale.

• LMSE, University of Toronto

MITACS Globalink Research Intern

May - August 2023

Toronto, Canada

- Implemented novel **convolution** and **attention**-based architectures for protein function prediction.
- Executed parameter-efficient fine-tuning of language models using **LoRA** for enzyme activity prediction.

• Baker Lab, Institute of Protein Design, University of Washington

Summer Research Fellow under Nobel Laureate, Dr. David Baker

June - September 2022

Seattle, USA

- Explored GNNs and transformers such as **ProteinMPNN** and **MIF-ST** for protein inverse folding.
- Demonstrated improved performance by **augmenting** the pre-training set with **AlphaFold-predicted** structures and training with noisy backbone coordinates.

EDUCATION

• Indian Institute of Technology Madras

Dual Degree: B.Tech in Biological Engineering and M.Tech in Data Science

July 2019 - July 2024

Chennai, India

- CGPA: 9.02/10.00 (Rank 3)
- Thesis: Enhancing Protein Fitness with Deep Learning: Sequence-Structure Fusion using Language Models and Graph Neural Networks for Function Prediction and Generative Sequence Design via Conditional Diffusion (**Nominated for the Best Thesis in Data Science Award**)
- Highlights:
MITACS Globalink Research Fellowship, University of Washington's IPD Summer Research Fellowship, Indian Academy of Sciences Research Fellowship, Amazon ML Summer School, HTTA Scholarship, 2x Inter-IIT Tech Contingent Member
- Relevant Coursework:
Deep Learning for Imaging, Pattern Recognition & Machine Learning, Database Management Systems, Introduction to Data Analytics, Data Analytics Laboratory, Big Data Laboratory, Mathematical Foundations of Data Science, Linear Algebra, Probability & Statistics, Data Structures & Algorithms, Numerical Methods, Computational Neuroscience, Mathematical Modelling for Industry, Analysis and Interpretation of Biological Data

PUBLICATIONS

C=CONFERENCE, J=JOURNAL, S=IN SUBMISSION, P=POSTER, *=FIRST AUTHOR

- [P.1] **LeMat-Synth: a multi-modal toolbox to curate broad synthesis procedure databases from scientific literature.** [AI for Accelerated Materials Discovery \(AI4Mat\) Workshop NeurIPS 2025.](#)
- [P.2*] **LeMat-GenBench: Bridging the gap between crystal generation and materials discovery.** [AI for Accelerated Materials Discovery \(AI4Mat\) Workshop NeurIPS 2025.](#)
- [S.1*] **Out-of-Distribution performance as a proxy for explanation quality in graph neural networks.**
- [C.1] **De-Identification of sensitive personal data in datasets derived from IIT-CDIP.** [Empirical Methods in Natural Language Processing Main Conference \(EMNLP Main\) 2024](#)
- [C.2*] **Leveraging LLM-generated contextual conversations for cross-lingual image captioning.** [Ninth Conference on Machine Translation \(WMT\) at Empirical Methods in Natural Language Processing \(EMNLP\) 2024](#)
- [P.3*] **Out-of-Distribution performance as a proxy metric for graph neural network explainers in the absence of ground-truth explanations.** [WiML Workshop at Neural Information Processing Systems \(NeurIPS\) 2024](#)
- [P.4*] **Screening protein sequences generated via conditional diffusion for enhanced fitness using a GNN-based function predictor.** [Machine Learning for Computational Biology 2024](#)
- [J.1] **Advances in generative modeling methods and datasets to design novel enzymes for renewable chemicals and fuels.** [Current Opinion in Biotechnology 2023](#)

STUDENT CONFERENCES

*=FIRST AUTHOR

- [1*] **Sequence-informed structured GNNs to screen diffusion-generated proteins for enhanced function.**
Poster Session at [WSAI Annual Research Showcase](#), Wadhvani School for Data Science and AI, IIT Madras, 2024.
- [2*] **Utilizing Whey Water to Produce Bacterial Cellulose-based Insulin Patch.**
Unique Idea Award at Bioinnovate Competition, National Bioengineering Competition 2022.
- [3] **Team GEnoM: Utilizing Whey Water to Produce Bacterial Cellulose-based Insulin Patch.**
Gold Medal with an award for Best Computational and Overall Project at [Global Open Genetic Engineering Conference 2021.](#)

ACADEMIC SERVICE

- **Reviewer, AI for Accelerated Materials Discovery (AI4Mat) Workshop** 2025
Neural Information Processing Systems (NeurIPS)
- **Reviewer, Conference on Machine Translation (WMT)** 2025
Empirical Methods in Natural Language Processing (EMNLP)
- **Reviewer, XAI4Science Workshop** 2025, 2026
International Conference on Learning Representations (ICLR), AAAI Conference on Artificial Intelligence
- **Reviewer, Machine Learning for Structural Biology (MLSB) Workshop** 2024, 2025
Neural Information Processing Systems (NeurIPS); San Diego and EurIPS
- **Reviewer, Machine Learning for Life and Material Science (ML4LMS) Workshop** 2024
International Conference on Machine Learning (ICML)

TECH AND RESEARCH COMMUNITY INVOLVEMENT

- **Cohere Labs Open Science Initiative** Since 2025
Community Member
- **ML Collective** Since 2023
Member; published work in EMNLP Main 2024

PROJECTS

- **Graph Active Learning for 3D Molecular Property Prediction** July 2024
Tools: Python, PyTorch, PyTorch Geometric, NumPy, scikit-learn [G]
 - Designed a GNN surrogate model from scratch using gated equivariant blocks to process scalar & vector features.
 - Integrated pre-trained models as labelers with true labels in active learning loops for property prediction.
 - Implemented diverse acquisition functions, including Monte Carlo Dropout-based uncertainty, Expected Improvement, [BADGE](#), and Model Disagreement.
- **Fixing Label Errors and Train-Test Overlap in RVL-CDIP** February - September 2024
Tools: Python, PyTorch, HuggingFace Transformers, OpenAI CLIP [G]
 - Implemented CLIP embeddings for identifying and rectifying label errors by isolating outliers based on their distance from class centroids.

- Employed the [SuperGlue](#) pre-trained model for feature-based similarity assessment of document pairs, facilitating the identification of train-test duplicates.
 - Used minimum hashing and locally sensitive hashing to efficiently identify groups of similar documents and further refined these groups using DBSCAN to enable accurate deduplication.
 - Designed scripts for the training and eventual evaluation of advanced models, including DiT, Donut, and LayoutLM, on the cleaned dataset.
- **An Attempt at Optimized Implementation of GPT-2** November-December 2023
Tools: Python, PyTorch, HuggingFace Accelerate, HuggingFace Transformers, wandb [🔗]
- Implemented GPT2-small, integrating its advanced features like positional encodings, multi-head attention, and position-wise feedforward networks in transformer layers.
 - Upgraded context capture by integrating rotary positional embeddings, Group Query Attention, and Sliding Window Attention mechanisms with implementation from scratch.
- **Hierarchical Loss Functions and Soft Labels for Mitigating Neural Collapse** July-November 2023
Course Project: Deep Learning for Imaging under Dr. Kaushik Mitra [🔗]
- Demonstrated the prevalence of Neural collapse across loss functions (MSE, Cross Entropy), various image datasets (CIFAR-10, ImageNet), and model architectures (VGG, ResNet).
 - Investigated the integration of hierarchical representation to align similar classes more closely in the learned feature space, countering the typical NC attributes.
 - Constructed a hierarchy tree for the CIFAR-10 dataset and incorporated soft labels to enforce class hierarchy.
 - Designed a custom loss function imposing higher penalties for misclassifications farther apart in the class hierarchy.
- **Robotic Personal Assistant, Project Lead with Robotics Club at IIT Madras** 2020-2021
Tools: Python, PyTorch, Tensorflow, OpenCV, SLAM, RViz, Gazebo, NVIDIA Jetson Nano, Arduino
- Real-time face and emotion detection via OpenCV using Haar feature-based cascade classifiers.
 - Object detection and identification enabled using YOLO. Used anchor boxes to find the best fit object, non-maximum suppression to filter out the best predictions and intersection over union for image segmentation.
 - Employed ORB-SLAM2 with a probabilistic DL model for mapping dynamic environments.
 - Integrated an intent-based chatbot and a product recommendation system by scraping data from Amazon.
 - Used SLAM and RViz to make the bot capable of mapping multiple unknown environments and localizing itself with respect to the environment on Gazebo.

HACKATHONS

- **Temenos AI for Green Finance Hackathon** January 2024
Tools: Python, Amazon Web Services (AWS) Bedrock, PyTorch, RAG, Streamlit
- **Placed 3rd out of 340 teams** across India and won prizes worth Rs. 600,000.
 - Engineered a comprehensive platform empowering users to evaluate the investment potential of green finance projects using project design documents (PDDs).
 - Used Titan Multimodal embeddings to assimilate multimodal information from PDDs. Upgraded to Donut embeddings to ensure OCR-free parsing for rapid computation and later facilitate semantic search during RAG.
 - Deployed a user interface MVP with RAG and Claude on AWS Bedrock via Streamlit, allowing users to review projections of carbon credit metrics and prices for various projects.
- **Piramal Finance Data Science Hackathon** March 2024
Tools: Python, NumPy, Pandas, scikit-learn, feature-engine, XGBoost, LightGBM, TabTransformer, CatBoost
- Developed an ensemble of classical ML models and an attention-based architecture for a home loan acquisition scorecard, achieving an AUC-ROC of 0.88.
 - Secured **1st place among 1000+ participants**, winning prizes worth 100,000 rupees and a full-time job offer.
- **American Express Campus Challenge: Credit Card Default Prediction** November 2022
Tools: Python, NumPy, Pandas, scikit-learn, statsmodels, XGBoost, LightGBM, CatBoost [🔗]
- Achieved **2nd position** and received a summer internship offer for the data scientist role.
- **SENAI Lab Data Contest: Auto-grader Tool** May 2022
Tools: Python, NumPy, Pandas, scikit-learn, nltk, spaCy, HuggingFace Transformers, Flask [🔗]
- Built an NLP auto-grader that evaluates student answers using semantic similarity to model solutions.
 - Pre-processed a [dataset](#) of 2500 questions using lemmatization and tokenization.
 - Evaluated and compared the performance of Doc2Vec, BERT and RoBERTa on the data set, achieved 82% accuracy, and **placed 3rd** in the contest.