Mohit Pandey CV

CONTACT INFORMATION
mpandey@bu.edu LinkedIn profile GitHub profile Google Scholar
Menlo Park, CA 94025 +1 (626) 213-4065

EDUCATION

PhD in Physics *Boston University, USA* Aug 2015- June 2021
BS-MS Dual Degree *IISER Kolkata, India* July 2010- May 2015

SUMMARY

I'm a Machine Learning Scientist with over five years of experience in deep learning, optimization algorithms, and software engineering. I'm passionate about leveraging these skills to analyze high-throughput biological datasets. My background includes a Ph.D. in Computational Statistical Physics and Biophysics, along with biotech industry experience in i) peptide and protein modeling, and CRISPR-Cas gene-editing and epigenetic-editing medicine. I'm eager to collaborate on projects that use AI to accelerate drug discovery and improve patient outcomes

Protein Language Models | Generative AI | Graph Neural Network | Cheminformatics | 3D structure-based drug discovery | Peptide and protein modeling | CRISPR-Cas gene editing | Monte Carlo Sampling | Uncertainty ML quantification

TECHNICAL SKILLS

Programming: Python, R, and Bash
Molecular modeling: RDKit, Molsoft ICM, Rosetta Commons / PyRosetta, PyMOL, Jalview, QSAR modeling
Machine Learning: Deep neural networks (e.g. CNN, LSTM, GNN and Transformers), generative AI (e.g. VAE, GAN and Diffusion), protein language models (e.g. ESM-2 and ProtTrans) LASSO, ridge regression, XGBoost, clustering algorithms (e.g., PCA, UMAP and t-SNE)
Packages: PyTorch, PyTorch Lightning, PyTorch Geometric, Huggingface transformers, Huggingface evaluate, Keras, TensorFlow, DeepChem, Rosetta, pandas, numpy, scikit-learn, Jupyter lab, Streamlit, Requests (for REST API)
Computing: Amazon Web Services (AWS) EC2, AWS S3, AWS Lambda, AWS Step function, AWS EventBridge, Docker container, Sun Grid Engine, Computing Cluster job scheduling, Linux environment
MLOps: MLflow, AWS SageMaker and Domino Data Labs

WORK EXPERIENCE

Sr. Data Scientist *Scribe Therapeutics, Inc.* June 2024 - Present

- Hit optimization: Achieved a prospective success rate of 43% in designing novel and potent enzyme variants using a finetuned protein language model (ESM-2) on proprietary CRISPR-Cas gene editing data.
- Hit identification: Analyzed high-throughput CRISPR-Cas gene editing mammalian screening data using 3D protein structure features that contributed to the success rate of 74% in hit selection.
- LLM pretraining: Explored and implemented pre-training of a protein language model on CRISPR-Cas sequence landscape.
- ML training on multiple GPU cores: Finetuned a protein language model (ESM-2) with 35M parameters on 8 NVIDIA GPU cores using Huggingface transformers for 8 hours.
- Epigenetics study: Improved organization's understanding of novel epigenetic repressor domains by analyzing high-throughput CRISPR-Cas epigenetic screening data using protein clustering, sequence alignment, and physicochemical properties correlation studies.
- ETL pipeline: Managed a Data Engineer who implemented cloud ETL data pipelines for organization's Data Foundation using AWS Glue, AWS RDS, AWS Redshift, AWS Lambda and Streamlit.
- Presented ML modeling results related to hit optimization to an external partner.

Machine Learning Engineer *Fog Pharmaceuticals/ Parabilis Medicines, Inc.* Oct 2022 - May 2024

- Developed novel custom-built graph neural network VAE (Generative AI) to generate peptide sequences with an alpha-helical constraint, which are predicted to have high cell permeability.
- Designed peptide sequences using generative AI that i) achieved single-digit nanomolar binding with 1.5x improvement in cell permeability over parent sequence, and ii) improved biological cellular activity by 500 times over the parent sequence.
- Developed a novel metric for quantifying applicability and confidence measure of ML classifier models
- Achieved prospective ROC AUC of more than 0.7 for predicting cell permeability with 4000 peptide sequence

- Communicated insights to drug designers for hit identification and hit-to-lead optimization obtained from machine learning modeling and data analysis

Quantum Computing, Senior Scientist *Menten AI, Inc.* Aug 2021- Aug 2022

- Experienced with peptide and protein modeling, and computer-aided drug discovery using Rosetta
- Tackled protein design and multibody molecular docking using simulated annealing under Rosetta score
- Contributed to Platform by building a protein design pipeline using Python, Bash and Rosetta scripts.
- Worked on a client-facing project to increase thermostability of a target protein using a quantum optimization algorithm
- Extensively used Gitlab for version control, tracking issues and collaborating with colleagues while coding drug discovery related algorithms

Data Scientist Intern *Amazon WorkForce Staffing* June - Aug 2020

- Drove future excellence by tackling the problem of hiring more than a million people a year from a technical perspective
- Built a CNN-LSTM model in tensorflow to improve time-series forecasting of proprietary data with goal of informing business decisions
- Used Agile-like project management to work with stakeholders to understand a rapidly evolving and ambiguous problem space
- Communicated highly technical ML information to non-technical stakeholders and increased understanding of time-series forecasting and neural networks in organization

Ph.D. Research Experience

Machine Learning – *CNN-based Image Classifier using Transfer Learning* Nov - Dec 2018

- Web - scraped around 120K images from wikiart.org using BeautifulSoup
- Built a model in Keras using pre-trained VGG-16 followed by a few fully-connected layers
- Trained the CNN-based image classifier on around 96K images using GPU computing cluster nodes
- Achieved accuracy of 49% with 59 genre classes (e.g. Portrait, Symbolic)
- Visualized output of Deep Neural Network using clustering algorithms (PCA and t-SNE)

Theoretical and Computational Statistical Physics – *PhD research* May 2016 - June 2021

- Published manuscripts in peer-reviewed journals in the field of Statistical Physics
- Expertise in interacting with computing clusters (Sun Grid Engine) and job scheduling through bash scripts
- Coded linear algebra operations on large matrices ($d = 2^{16} \times 2^{16}$) including diagonalization and inner products in Python
- Increased speed of code by 40x using parallel computing (OMP and MKL) for high-dimensional data ($d = 4^{12} \sim 16$ million)

Publications

- ☐ **Mohit Pandey***, Tristan Zaborniak*, Hans Melo, Alexey Galda and Vikram K. Mulligan. (2022) *Multibody molecular docking on a quantum annealer.* arXiv:2210.11401 [q-bio.BM]. (* Co-primary authors)
- ☐ **Mohit Pandey**, Pieter W. Claeys, David K. Campbell, Anatoli Polkovnikov, & Dries Sels (2020). *Adiabatic eigenstate deformations as a sensitive probe for quantum chaos.* Physics Review X, 10, 041017
- ☐ Eric Boyers*, **Mohit Pandey***, David K. Campbell, Anatoli Polkovnikov, Dries Sels & A. O. Sushkov (2019). *Floquet-engineered quantum state manipulation in a noisy qubit.* Physics Review A, 100, 012341 (* Co-primary authors)
- ☐ Pieter. W. Claeys, **Mohit Pandey**, Dries Sels, & Anatoli Polkovnikov (2019). *Floquet-engineering counter-diabatic protocols in quantum many-body systems.* Physics Review Letter, 123, 090602 [**Editors' Suggestion**]
- ☐ Tamiro Villazon,, Pieter W. Claeys, **Mohit Pandey**, Anatoli Polkovnikov, & Anushya Chandran (2020). *Persistent dark states in anisotropic central spin models.* Nature Scientific Reports, 10, 16080
- ☐ Sayak Ray, **Mohit Pandey**, Anandmohan Ghosh, & Subhasis Sinha (2015). *Localization of weakly interacting Bose gas in quasiperiodic potential.* New Journal of Physics, 18(1), 013013.

Manuscripts in Preparation

- ☐ Vikram K. Mulligan, **Mohit Pandey**, Haley I. Merritt, Michael R. Sawaya, Stewart Slocum, Brian D. Weitzner, Andrew M. Watkins, P. Douglas Renfrew, Craig Pelissier, Todd O. Yeates, Julia K. Leman, Paramjit S. Arora, Richard Bonneau, and Hans Melo. (2022). *A peptide designed on a quantum computer*, In preparation. Incorporates work from this earlier preprint

Mentorship and Supervisory Experience

- ☐ Led a Data Science team of N=2 whose mission was to accelerate data-powered discovery of genetic therapeutics through ML modeling and cutting-edge Data Engineering. – *Scribe Therapeutics Inc, June 2024 - Present*
- ☐ Supervised an intern successfully to finish a project on multibody docking – *Menten AI, June - Aug 2022*