Shrey Goel

503-709-0813 | shrey.goel@duke.edu | LinkedIn | GitHub | Google Scholar

Education

Duke University May 2027
Bachelor of Science, Computer Science & Mathematics (GPA: 3.83) *Durham, NC*

Related Coursework: Foundations of Generative Models, Deep Learning, Natural Language Processing, Generative AI in Protein Design, Linear Algebra, Advanced Probability, Stochastic Processes, Data Structures & Algorithms, Databases

Dr. Bart Kamen Memorial Scholar: $40,000 merit scholarship awarded to students with high research output

Volunteer Services: Workshop paper reviewer for the 2025 Neural Information Processing Systems Conference

Peer-Reviewed Articles

1. Vincoff S., **Goel S.**, Kholina K., Pulugurta R., Vure P., & Chatterjee P. (2025). FusOn-pLM: a fusion oncoprotein-specific language model via adjusted rate masking. Nature Communications, 16(1), 1436.
2. Smela M.P., Kramme C.C., Fortuna R.J.P., Wolf B., **Goel S.**, Adams J., Ma C., Velychko S., Widocki U., Kavirayuni V.S., Chen T., Vincoff S., Dong E., Kohman R.E., Kobayashi M., Shioda T., Church G.M., Chatterjee P. (2025). Rapid Human Oogonia-like Cell Specification via Combinatorial Transcription Factor-Directed Differentiation. EMBO Reports, 1-30.
3. Bhat S., Palepu K., ..., **Goel S.**, ... & Chatterjee, P. (2025). De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. Science Advances, 11(4), eadr8638.
4. Chen T., Dumas M., Watson R., Vincoff S., Peng C., Zhao L., Hong L., Pertsemlidis S., Shaepers-Cheu M., Wang T., Srijay D., Monticello C., Vure P., Pulugurta R., Kholina K., **Goel S.**,... & Chatterjee, P. (2024). Target sequence-conditioned design of peptide binders using masked language modeling. Nature Biotechnology, 1-9.

Preprints

1. **Goel S.**, Schray P., Zhang Y., Vincoff S., Kratochvil H., Chatterjee P. (2025). Token-Level Guided Discrete Diffusion for Membrane Protein Design. arXiv preprint arXiv:2410.16735.
2. Hong L., Ye T., Wang T., Srijay D., Zhao L., Watson R., Vincoff S., Chen T., Kholina K., **Goel S.**,... & Chatterjee P. (2024). Programmable Protein Stabilization with Language Model-Derived Peptide Guides. Research Square, rs-3.
3. Ye T., Alamgir A., Robertus C., Colina D., Monticello C., Donahue T.C., Hong L., Vincoff S., **Goel S.**,... & DeLisa MP. (2024). Programmable protein degraders enable selective knockdown of pathogenic β-catenin subpopulations in vitro and in vivo. bioRxiv, 2024-11.

Proceedings

1. **Goel S.**, Schray P., Zhang Y., Vincoff S., Kratochvil H., Chatterjee P. (2026). Token-Level Guided Discrete Diffusion for Membrane Protein Design. International Conference on Learning Representations – Main Conference. Under Review.
2. **Goel S.**, Schray P., Zhang Y., Vincoff S., Kratochvil H., Chatterjee P. (2025). Token-Level Guided Discrete Diffusion for Membrane Protein Design. Neural Information Processing Systems – AI4Science Workshop **(Oral Presentation)**.
3. **Goel S.**, Thoutam V., Marroquin E. M., Gokaslan A., Firouzbakht A., Vincoff S., ... & Chatterjee P. (2025). MeMDLM: De Novo Membrane Protein Design with Property-Guided Discrete Diffusion. International Conference on Learning Representations – Generative Perspectives for Biology Workshop.

Research Experience

Chatterjee Lab April 2023 – Present
Machine Learning Researcher *University of Pennsylvania*

- Developed and trained discrete diffusion model for membrane protein sequence generation on High-Performance Computing cluster using PyTorch Lightning, Wandb, and HuggingFace.
- Generated protein sequences achieved wet-lab performance equivalent to naturally existing controls and a 44% decrease in perplexity over state-of-the-art autoregressive models.
- Designed and implemented novel classifier-guided sampling algorithm combining attention scores and classifier gradients to selectively edit specific sequence tokens during inference.
- **First-author manuscript** published at **NeurIPS 2025** workshops and selected for **oral presentation** .

Machine Learning Engineer

- Fine-tuned ESM-2 protein language model on a novel masked language modeling objective that leverages dynamic masking rates to engineer cancer protein-specific sequence embeddings.
- Trained over 40 model variants to reach a lower perplexity compared to standard MLM fine-tuning, conducting extensive ablation studies and hyperparameter tuning.
- Evaluated embedding quality by training Scikit-learn classifiers on downstream tasks, achieving a 25% improvement in AUROC and F1 score over pretrained embeddings.
- **Second-author manuscript** published at **ICML 2024** workshops and published in *Nature Communications* .

Gameto April 2023 – Feb 2025

Bioinformatics Researcher *Duke University*

- Reduced experimental analysis time from 4 months to 3 days by training language models for cell type classification.
- Created automated statistical analysis pipeline in R and Python to plot PCAs, volcano plots, and heat maps.
- Co-author manuscript published in *EMBO Reports*

Professional Experience

Latus Bio September 2025 – Present

Machine Learning Engineer (Part-Time) *Remote*

Qualcomm Technologies May 2025 – August 2025

Machine Learning Engineer Intern *San Diego, CA*

- Diagnosed performance bottlenecks in Meta's Llama LLM using PyTorch Profiler, identifying redundant computations over padded tokens in KV-cache system.
- Engineered optimized self-attention tensor computations that bypass pad tokens, reducing inference latency by 14x in quantized Llama models deployed on edge-devices.

Technical Skills

Languages: Python, Java, R
ML Technologies: PyTorch, Hugging Face, Scikit-learn, Parameter Efficient Fine Tuning, HPC, Pandas, NumPy
Developer Tools: Git, Jira, Docker, Jupyter, Figma