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, Data Structures & Algorithms, Linear Algebra, Advanced Probability

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

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. Chaklai A., O'Neil A., **Goel S.**, Margolies N., Krenik D., Perez R., ... & Raber J. (2024). Effects of Paraquat, Dextran Sulfate Sodium, and Irradiation on Behavioral and Cognitive Performance and the Gut Microbiome in A53T and A53T-L444P Mice. Genes, 15(3), 282.
3. 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.
4. 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.
5. 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). PepMLM: Target Sequence-Conditioned Generation of Therapeutic Peptide Binders via Span Masked Language Modeling. Nature Biotechnology, (In press).

Preprints

1. **Goel S.**, Thoutam V., Marroquin E. M., Gokaslan A., Firouzbakht A., Vincoff S., ... & Chatterjee P. (2024). MeMDLM: De Novo Membrane Protein Design with Masked Discrete Diffusion Protein Language Models. 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.**, 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 and Experimental Perspectives for Biomolecular Design Workshop.
2. **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 – Learning Meaningful Representations of Life Workshop.
3. **Goel S.**, Thoutam V., Marroquin E. M., Gokaslan A., Firouzbakht A., Vincoff S., ... & Chatterjee P. (2024). MeMDLM: De Novo Membrane Protein Design with Masked Discrete Diffusion Protein Language Models. Neural Information Processing Systems – AIDrugX Workshop.
4. **Goel S.**, Thoutam V., Marroquin E. M., Gokaslan A., Firouzbakht A., Vincoff S., ... & Chatterjee P. (2024). MeMDLM: De Novo Membrane Protein Design with Masked Discrete Diffusion Protein Language Models. Molecular Machine Learning Conference.
5. Vincoff S., **Goel S.**, Kholina K., Pulugurta R., Vure P., & Chatterjee P. (2024). FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Focused Probabilistic Masking. Neural Information Processing Systems – Machine Learning for Structural Biology Workshop.
6. Vincoff S., **Goel S.**, Kholina K., Pulugurta R., Vure P., & Chatterjee P. (2024). FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Focused Probabilistic Masking. International Conference on Machine Learning – Accessible and Efficient Foundation Models for Biological Discovery Workshop.
7. Vincoff S., **Goel S.**, Kholina K., Pulugurta R., Vure P., & Chatterjee P. (2024). FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Focused Probabilistic Masking. Molecular Machine Learning Conference.
8. Vincoff S., **Goel S.**, Kholina K., Pulugurta R., Vure P., & Chatterjee P. (2024). FusOn-pLM: A Fusion Oncoprotein-Specific Language Model via Focused Probabilistic Masking. Duke University AI Day Conference.

Experience

Latus Bio September 2025 – Present
Incoming Machine Learning Engineer (Part-Time) *Remote*

Qualcomm AI Research 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.

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

- Developed and trained masked discrete diffusion model for membrane protein sequence generation on multi-node GPU cluster using PyTorch Lightning, Wandb, and HuggingFace.
- Generated proteins 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 presented at leading machine conference workshops (**NeurIPS 2024**, **ICLR 2025**)

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 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 presented at leading machine learning conference workshops **(NeurIPS 2024, ICML 2024)** and published in *Nature Communications*

Gameto April 2023 – Feb 2025
Bioinformatics Scientist *Durham, NC*

- Reduced experimental analysis time from 4 months to 3 days by building bioinformatics pipeline in R and Python
- Created novel platform to validate cell engineering research by training language models for cell type classification
- Co-author manuscript published in *EMBO Reports*

Projects

CHOFormer | *Python, PyTorch, Wandb, HuggingFace, Pandas, Git* Sept 2024 - Oct 2024

- Designed and trained autoregressive Transformer model that converts low-expression protein sequences into highly optimized DNA sequences, achieving a 25% increase in protein expression.

PPI-CLIP | *Python, PyTorch, HuggingFace, Pandas, NumPy, Git* Aug 2024

- Created a computational pipeline to identify PPIs by adopting OpenAI's Contrastive Language-Image Pre-training (CLIP) architecture for protein sequence data

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