Meghana Kshirsagar

B meghana.ksagar@gmail.com • www.meghanak.net

Work & Research Experience

AI for Good research, Microsoft Aug 2019–current

Senior Research Scientist

In my current role, I lead AI-for-science collaborations and contribute to AI-for-health projects, working closely with top academic labs and healthcare organizations. I help review grant proposals and contribute to internal reponsible-AI efforts.

Memorial Sloan Kettering Cancer Center, NY Jun 2016–Jan 2019

Research Scholar (post-doctoral researcher, Christina Leslie's lab)

Understanding epigenetic regulation by building models to learn representations of transcription factors and cell states that combine knowledge from diverse and large scale DNA sequencing data. Working with data from chromatin accessibility (ATAC-seq), transcriptomic (RNA-seq), spatial transcriptomics and single-cell studies.

IBM T.J Watson Research, Yorktown Heights Sept 2015–Apr 2016

Postdoctoral researcher, Machine Learning

As part of an ARPA-e funded grant for crop science (TERRA project), we trained predictive models on data from plant genomic information and hyperspectral images of fields.

Yahoo! Labs, Bangalore 2007–2009

Research Engineer, Search Relevance & Information Extraction

Applied and extended algorithms and machine learning techniques for large scale classification and information extraction from the Web.

Education

School of Computer Science, Carnegie Mellon University, Pittsburgh 2010–2015

PhD from Language Technologies Institute (LTI), GPA: 3.98

Advisors: *Jaime Carbonell and Judith Klein-Seetharaman*

Thesis: *Combine and Conquer: Methods for Multitask Learning in Biology and Language*. My thesis work focused on building new multi-task learning models for host-pathogen protein-protein interaction prediction.

Indian Institute of Technology, Bombay 2004–2007

Master of Technology, Computer Science Dept., CPI: 9.3/10

Advisor: *S. Sudarshan*

Thesis: *Graph Algorithms for Keyword Search on External Memory Data Graphs*

Ongoing projects

- **Generation of accurate structures for higher-order protein homo-oligomers (joint with the Institute of Protein Design, UW)**: We characterize protein inter-chain contacts as primary / secondary. Current models perform very poorly on secondary contacts in general and primary contacts in higher-order proteins involving several chains. The goal is to train a model that is good at generating inter-chain contacts.
- **Protein-DNA binding (joint with the Institute of Protein Design, UW)**: We are working on extensions of a diffusion-based protein foundation model (like AlphaFold3) with the goal of incorporating knowledge from DNA binding studies from epigenetic databases like TRANSFAC and ENCODE to improve the accuracy and generalizability of these models. We are exploring distillation and reinforcement learning approaches.
- **Understanding the vulnerabilities in deepfake detection (internal)**: I have been red-teaming an in-house deepfake detector to understand vulnerabilities and propose solutions to mitigate the risks.

– **Peptide design for Cardiac Myosin protein (mentoring students at Univ. of Pennsylvania in Gregory Bowman's lab)**: The goal of this project is apply protein generative models to generate diverse protein sequences that will bind the Cardiac Myosin protein and modify the protein's transition from super-relaxed state (SRX) to disordered relaxed state (DRX).

Publications in submission

Preethi K Periyakoil, Melanie H Smith, **Kshirsagar, Meghana**, Daniel Ramirez, Edward F DiCarlo, Susan M Goodman, Alexander Rudensky, Laura Donlin, and Christina S Leslie. Deep topic modeling of spatial transcriptomics in the rheumatoid arthritis synovium identifies distinct classes of ectopic lymphoid structures. *Under review at Nature Communications*, pages 2025–01, 2025.

Carlos Martinez, Md Nasir, **Kshirsagar, Meghana**, Cass Mccharen, Rae Shean, Andrew Placona, Juan Lavista Ferres, Rahul Dodhia, and Bill Weeks. Predictive models for kidney offer acceptance: Challenges and strategies. *Under review at Transplantation Direct*, 2025.

Journal Publications

Kshirsagar, Meghana, Artur Meller, Ian Humphreys, Samuel Sledzieski, Yixi Xu, Rahul Dodhia, Eric Horvitz, Bonnie Berger, Gregory Bowman, Juan Lavista Ferres, David Baker, and Minkyung Baek. Rapid and accurate prediction of protein homo-oligomer symmetry with seq2symm. *Nature Communications*, 2025.

Shahrzad Gholami, Lea Scheppke, **Kshirsagar, Meghana**, Yue Wu, Rahul Dodhia, Roberto Bonelli, Irene Leung, Ferenc B Sallo, Alyson Muldrew, Catherine Jamison, et al. Enhanced macular telangiectasia type 2 detection: Leveraging self-supervised learning and ensemble models. *Ophthalmology Science*, page 100710, 2025.

Samuel Sledzieski, **Kshirsagar, Meghana**, Minkyung Baek, Rahul Dodhia, Juan Lavista Ferres, and Bonnie Berger. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences*, 121(26):e2405840121, 2024.

Shahrzad Gholami, Lea Scheppke, **Kshirsagar, Meghana**, Yue Wu, Rahul Dodhia, Roberto Bonelli, Irene Leung, Ferenc B Sallo, Alyson Muldrew, Catherine Jamison, et al. Self-supervised learning for improved optical coherence tomography detection of macular telangiectasia type 2. *JAMA ophthalmology*, 142(3):226–233, 2024.

Gabriele Ciceri, Arianna Baggiolini, Hyein S Cho, **Kshirsagar, Meghana**, Silvia Benito-Kwiecinski, Ryan M Walsh, Kelly A Aromolaran, Alberto J Gonzalez-Hernandez, Hermany Munguba, So Yeon Koo, Nan Xu, Kaylin J. Sevilla, Peter A. Goldstein, Joshua Levitz, Christina S. Leslie, Richard P. Koche, and Loren Studer. An epigenetic barrier sets the timing of human neuronal maturation. *Nature*, 626(8000):881–890, 2024.

Mayana Pereira, **Kshirsagar, Meghana**, Sumit Mukherjee, Rahul Dodhia, Juan Lavista Ferres, and Rafael de Sousa. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *Plos one*, 19(2):e0297271, 2024.

Artur Meller, Michael D Ward, Jonathan H Borowsky, Jeffrey M Lotthammer, **Kshirsagar, Meghana**, Felipe Oviedo, Juan Lavista Ferres, and Gregory Bowman. Predicting the locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Biophysical journal*, 122(3):445a, 2023.

Ming Zhong, Tamara Glazer, **Kshirsagar, Meghana**, Richard Johnston, Rahul Dodhia, Allen Kim, Divya Michael, Santiago Salcido, Sameer Nair-Desai, Thomas C Tsai, et al. Estimating vaccine-preventable covid-19 deaths among adults under counterfactual vaccination scenarios in the united

states: a modeling study using observational data. *Journal of Pharmacy and Pharmacology Research*, 7(3):163–167, 2023.

Sumit Mukherjee, **Kshirsagar, Meghana**, Nicholas Becker, Yixi Xu, William B Weeks, Shwetak Patel, Juan Lavista Ferres, and Michael L Jackson. Identifying long-term effects of sars-cov-2 and their association with social determinants of health in a cohort of over one million covid-19 survivors. *BMC Public Health*, 22(1):2394, 2022.

Kshirsagar, Meghana, Md Nasir, Sumit Mukherjee, Nicholas Becker, Rahul Dodhia, William B Weeks, Juan Lavista Ferres, Barbra Richardson, et al. The risk of hospitalization and mortality after breakthrough sars-cov-2 infection by vaccine type: observational study of medical claims data. *JMIR Public Health and Surveillance*, 8(11):e38898, 2022.

Kshirsagar, Meghana, Han Yuan, Juan Lavista Ferres, and Christina Leslie. Bindvae: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin. *Genome biology*, 23(1):174, 2022.

Jeffrey N Law, Kyle Akers, Nure Tasnina, Catherine M Della Santina, Shay Deutsch, **Kshirsagar, Meghana**, Judith Klein-Seetharaman, Mark Crovella, Padmavathy Rajagopalan, Simon Kasif, et al. Interpretable network propagation with application to expanding the repertoire of human proteins that interact with sars-cov-2. *GigaScience*, 10(12):giab082, 2021.

Gaurav Gupta, **Kshirsagar, Meghana**, Ming Zhong, Shahrzad Gholami, and Juan Lavista Ferres. Recurrent convolutional neural networks for large scale bird species classification. *Scientific Reports*, 2021.

Kshirsagar, Meghana, Nure Tasnina, Michael D Ward, Jeffrey N Law, TM Murali, Juan M Lavista Ferres, Gregory R Bowman, and Judith Klein-Seetharaman. Protein sequence models for prediction and comparative analysis of the sars-cov-2—human interactome. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 154–165. World Scientific, 2020.

Han Yuan, **Kshirsagar, Meghana**, Lee Zamparo, Yuheng Lu, and Christina Leslie. Bindspace: decoding transcription factor binding signals by large-scale joint embedding. *Nature Methods*, 2019.

Sylvia Schleker, **Kshirsagar, Meghana**, and Judith Klein-Seetharaman. Comparing human–salmonella with plant–salmonella protein–protein interaction predictions. *Frontiers in Microbiology*, 6(36), 2015.

Kshirsagar, Meghana, Sylvia Schleker, Jaime Carbonell, and Judith Klein-Seetharaman. Techniques for transferring host-pathogen protein interactions knowledge to new tasks. *Frontiers in Microbiology*, 6, 2015.

Zhongming Zhao, Junfeng Xia, Oznur Tastan, Irtisha Singh, **Kshirsagar, Meghana**, Jaime Carbonell, and Judith Klein-Seetharaman. Virus interactions with human signal transduction pathways. *International journal of computational biology and drug design*, 4(1):83–105, 2011.

Peer-reviewed Conference Publications

Kshirsagar*, Meghana, Caleb Robinson*, Siyu Yang*, Shahrzad Gholami, Ivan Klyuzhin, Sumit Mukherjee, Md Nasir, Anthony Ortiz, Felipe Oviedo, Darren Tanner, et al. Becoming good at ai for good. *Artificial Intelligence, Ethics and Society, AIES*, 2021.

Kshirsagar, Meghana, Eunho Yang, and Aurélie Lozano. Learning task structure via sparsity grouped multitask learning. *European Conference on Machine Learning (ECML 2017)*, 2017.

Kshirsagar, Meghana, Jaime Carbonell, Judith Klein-Seetharaman, and Keerthiram Murugesan. Multitask matrix completion for learning protein interactions across diseases. In *International Conference on Research in Computational Molecular Biology (RECOMB 2016), Journal of Computational Biology (2017 issue)*, pages 53–64, 2016.

Kshirsagar, Meghana, Jaime Carbonell, and Judith Klein-Seetharaman. Multitask learning for host–pathogen protein interactions. *In Intelligent Systems for Molecular Biology (ISMB 2013) and Bioinformatics*, 29(13):i217–i226, 2013.

Kshirsagar, Meghana, Jaime Carbonell, and Judith Klein-Seetharaman. Techniques to cope with missing data in host–pathogen protein interaction prediction. *In European Conference for Computational Biology (ECCB 2012) and Bioinformatics*, 28(18):i466–i472, 2012.

Kshirsagar, Meghana, Sam Thomson, Nathan Schneider, Jaime Carbonell, Noah A Smith, and Chris Dyer. Frame-semantic role labeling with heterogeneous annotations. In *Association for Computational Linguistics (ACL)*, 2015.

Kshirsagar, Meghana, Rajeev Rastogi, Sandeep Satpal, Sengamedu Srinivasan, and Venu Satuluri. High-precision web extraction using site knowledge. *Proceedings of the Conference on Management of Data (COMAD)*, 2010 **(Best Paper Award)**.

Bhavana Bharat Dalvi*, **Kshirsagar*, Meghana**, and S Sudarshan. Keyword search on external memory data graphs. *Proceedings of the Very Large Data Bases (VLDB)*, 1(1):1189–1204, 2008.

Workshop papers

- An Analysis of the Deployment of Models Trained on Private Tabular Synthetic Data: Unexpected Surprises, M. Pereira, M. Kshirsagar, S. Mukherjee, R. Dodhia, J. L. Ferres *Workshop on Automated Creation, Privacy and Bias, International Conference on Machine Learning (ICML) 2021*
- Inferring transcription factor binding profiles jointly from SELEX and ATAC-seq. M. Kshirsagar, H. Yuan, C. Leslie *Cold Spring Harbor Labs (CSHL) workshop for Quantitative Biology, 2017*
- Iteratively Regrouped Lasso: learning group structures in genome wide studies of crops. M. Kshirsagar, E. Yang and A. C. Lozano, *Data Science for Food, Energy and Water at Conference on Knowledge Discovery and Data Mining* (KDD) 2016
- Automated Sorghum Phenotyping and Trait Development Platform. M. Tuiinstra, C. Weil, A. Thompson, C. Boomsma, M. Crawford, A. Habib, E. Delp, K. Cherkauer, M. Kshirsagar, E. Yang, P. Olsen, K. Natesan and A. C. Lozano, *Data Science for Food, Energy and Water at Conference on Knowledge Discovery and Data Mining* (KDD) 2016
- Leveraging Heterogeneous Data Sources for Relational Semantic Parsing. M. Kshirsagar, N. Schneider and C. Dyer, *Assoc. for Computational Linguistics* (ACL) workshop on Semantic Parsing 2014
- Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks, M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman, *NIPS Workshop on Machine Learning for Computational Biology* 2013
- Confident prediction of Salmonella-human protein-protein interactions. S. Schleker, I. Nouretdinov, M. Kshirsagar, J. Klein-Seetharaman, A Gammerman et al., *European Conf. Computational Biology* 2012
- Transfer learning based methods for new hosts: discovering host-pathogen protein-protein interactions. M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman, *Intelligent Systems for Molecular Biology* (ISMB) 2012

Awards & Achievements

- My work on cluster bomb munition detection was featured on prominent news websites covering "AI in monitoring civil conflict"
- Richard King Mellon Presidential Fellow of Life Sciences, Carnegie Mellon University, 2011-2014

- Ray Ozzie Fellowship awarded by Computer Science Dept at University of Illinois, Urbana Champaign, 2009
- Best Paper award at the Conference on Management of Data, 2010
- Best Poster prize at the CMU Student Research Symposium, 2013
- Won the Carnegie Mellon University Social Innovations Challenge, 2011
- Awards for securing first rank in the Computer Science Dept, 2001, 2002 (undergraduate)
- Selected for the meritorious Pratibha scholarship by the Govt. of Andhra Pradesh (India) for academic excellence in higher secondary education, 2000

Patents

- Three patents on information extraction techniques
  (USPTO Publication # 20100223214, 20100257440, 20090216739)

Other professional activities

- **Book**: Contributed chapters and helped review a book published by my team at Microsoft on "AI for Good: Applications in Sustainability, Humanitarian Action and Health"
- **Organizational**: Co-organizer of ICML workshop on Accessible foundation models for Biology (AccML 2024-2025), Co-organizer of ICML Workshop for Computational Biology (WCB 2017-2018)
- **Program Committee**: NeurIPS 2016-2021, ICML 2017-2021, ICLR 2018-2021, PLoS Computational Biology, Neural Computation, BMC Genomics 2013, IJCAI 2016, WWW Posters 2017-2018, Workshop for ML in Comp Bio 2016-2018, AccML 2024-2025, Biotechnology Journal 2017, Nature Communications
- **Reading groups**: Organized the matrix factorization reading group at CMU, Machine Learning reading group at IBM Research, Deep Learning reading group at MSKCC
- **Mentoring:** graduate students on rotation and undergrad interns at MSKCC, interns at Microsoft
- **White papers**: Internal report on vulnerabilities of deepfake detection, Wrapper Induction for automatic extraction, TechPulse 2008; Site-Specifc Conditional Random Fields, TechPulse 2008; Web-Scale Information Extraction, TechPulse 2009
- **Invited talks**: Machine Learning seminar, Carnegie Mellon University (2015), Pro-active Learning and applications to Computational Biology, University of Pittsburgh (2013)
- **Posters**: Poster at Grad Expo 2010 at Univ. of Illinois, LTI Student Research Symposia 2012, 2013
- **Teaching Assistantship**: Machine Learning, Data-Mining and Information Retrieval
- **Others**: LTI Student committee: helping organize LTI colloquium, allocating student funds, organizing zero-waste events. At IIT Bombay: Elected as Cultural Secretary & Publications Coordinator, Alumni Secretary, Systems Admin for Hostel-11 and Mechanical Eng. Dept., Publicity coordinator for IGSA@CMU

Outreach

- Co-founded LaptopRehab, a campaign to donate phased out computers at Carnegie Mellon, and personal laptops to schools http://sites.google.com/site/cmulaptoprehab
- Taught sessions on CS and Machine Learning at Technights, a women@SCS workshop for school girls organized by Carnegie Mellon
- Organized Roadshows on Computer Science and Machine Learning at Pittsburgh schools

PhD students and interns mentored

- Mentored Han Yuan (PhD student, MSKCC), Gaurav Gupta (USC, intern), Samuel Sledzieski (MIT, intern), Preethi Periyakoil (Weill Cornell, M.D. PhD student), Artur Meller (WU-STL, M.D. PhD student), Loc Trinh (USC, intern), Shahlo Solieva (Univ. of Penn., PhD student)

Programming skills

- Python, R, matlab, C++, Java, Perl, Shell/awk scripting, running cluster jobs, PyTorch

– Code from some papers: https://github.com/meghana-kshirsagar/

References

– References available upon request