Meghana Kshirsagar

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

Work & Research Experience

AI for Good research, Microsoft Aug 2019–current

Senior Applied Research Scientist

Genomics, proteomics, public health, wildlife conservation and social good

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

Research Scholar

Building machine learning models to combine knowledge from diverse and large scale DNA sequencing data arising from studies of various cellular phenomena.

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 analyzed data coming 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*

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*

Vasavi Eng. College, Osmania University, Hyderabad 2000–2004

Bachelor of Engineering, Computer Science Dept., CPI: 8.45/10

Awards & Achievements

- **Media coverage**: (1) The Wall Street Journal, The Economist: Article on work done with an NGO working with Syria, Benetech, on weapons detection from audio data using deep learning. (2) Medical news websites: COVID-19 breakthroughs work covered.
- 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 topping the Computer Science Dept, 2001, 2002

- Selected for the meritorious Pratibha scholarship by the Govt. of Andhra Pradesh (India) for academic excellence in higher secondary education, 2000

Internships

- Microsoft Research, Redmond and IBM Research Labs, Delhi.

Ongoing projects

- **Estimation of pocket volume in protein structures**: I am leading a collaboration with Folding@Home and Washington Univ-St. Louis, to understand cryptic pockets, that are crevasses on the protein surface, where a drug compound can bind. Due to the transient nature of cryptic pockets, we use the Molecular Dynamics simulations of proteins, that were generated by Folding@home, to study them. Given a particular state, we model the probabilities of individual residues ending up in a state where the residue is part of a cryptic pocket, encoded by computing the solvent accessible surface area (SASA) or root mean square fluctuation (RMSF). We build 3D deep learning models of the protein structure using Graph Neural Networks and 3D CNNs and evaluate them on structures obtained from AlphaGo.
- **Late effects of childhood cancers**: In this collaboration with St Jude Children's hospital, I am leading the effort towards building models to analyze genomic markers (SNPs) in conjunction with clinical features to understand the various late effects of early childhood chemotherapy, such as cardiovascular dysfunction. We are building a feature selection approach using ideas from the causal modeling to address issues such as confounding variables, data drift.
- **Decoding epigenetic regulation**: With the broad goal of decoding what drives cell fate, cell identity and regulation of genes, I have built unsupervised deep learning models, in particular Variational Auto Encoders (VAEs), on DNA sequence data coming from chromatin accessibility experiments to learn general representations of DNA sequence. We use Next Generation Sequencing based *in vivo* assays such as ATAC-seq, and *in vitro* DNA sequencing data to learn representations for DNA-binding proteins (called transcription factors) across several cell types and over similar members of a transcription factor family.
- **Observational studies**: In the following projects, I have lead efforts with the goal of understanding the impact of COVID-19 on public health, where we use statistical methods from epidemiology. **Long COVID and the impact of social determinants of health (SDOH)**: We identify clinical conditions that are potential long-term sequelae of SARS-CoV-2 infection and the role of SDOH features such as race, gender and comorbidities from medical claims data. **Vaccines and breakthroughs**: We study the impact of SDOH, on the time to get a breakthrough from the point of vaccination.
- **Fairness and privacy trade-offs**: We study synthetic data generation from the perspective of fairness – i.e what happens to various fairness metrics such as demographic parity etc, in the context of various differentially private synthetic data generation approaches, as we vary ϵ to change the level of privacy guaranteed by the model.
- **Ethics and fairness concerns underlying organ donation**: We are collaborating with UNOS, the organ transplantation system of the U.S., on developing a deeper understanding of the dynamic, organ to donor matching process. My goal is to understand the implications of deploying a machine learning model that computes organ acceptance rates, from an ethical and fairness perspective and develop an approach to mitigate such concerns.

Publications in submission

Meghana Kshirsagar, Han Yuan, Christina Leslie, and Juan-Lavista Ferres. Dirichlet variational auto encoders for *de novo* motif discovery from atac-seq data. `https://www.biorxiv.org/content/10.1101/2021.09.23.461564v2`, 2021. Under review, Genome Biology.

Artur Meller, Michael Ward, Meghana Kshirsagar, Felipe Oviedo, Juan-Lavista Ferres, and Gregory Bowman. Predicting cryptic pocket opening from protein structures using graph neural networks. *In*

preparation, 2021.

Meghana Kshirsagar, Sumit Mukherjee, Md Nasir, Nicholas Becker, Juan-Lavista Ferres, and Barbra Richardson. Risk of hospitalization and mortality after breakthrough sars-cov-2 infection by vaccine type and previous sars-cov-2 infection utilizing medical claims data. `https://medrxiv.org/cgi/content/short/2021.12.08.21267483v1`, 2021. Under review, Journal of Infectious Diseases.

Meghana Kshirsagar*, Sumit Mukherjee*, Yixi Xu, Nicholas Becker, Juan-Lavista Ferres, and Michael Jackson. Identifying long-term effects of sars-cov-2 and their association with social determinants of health using a large medical claims database. `https://assets.researchsquare.com/files/rs-1032897/v1/ae6d3cff-1bfa-4cf5-887f-0766e5be7a5f.pdf?c=1636646537`, 2021. Under review.

Gabriele Ciceri, Hyunwhoo Cho, Meghana Kshirsagar, Arianna Baggiolini, Kelly Aromolaran, Ryan M. Walsh, Peter Goldstein, Richard, Christina Leslie, and Lorenz Studer. An epigenetic barrier in neural progenitor cells and early neurons set the timing of human neuronal maturation. *Under review, Science*, 2021.

Journal Publications

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

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

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

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

Meghana Kshirsagar, 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, Meghana Kshirsagar, 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

Meghana Kshirsagar*, 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.

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

Meghana Kshirsagar, 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.

Meghana Kshirsagar, 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.

Meghana Kshirsagar, 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.

Meghana Kshirsagar, 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.

Meghana Kshirsagar, 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*, Meghana Kshirsagar*, 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

- Predicting cryptic pocket opening from protein structures using graph neural networks. M. Ward, A. Meller, M. Kshirsagar, F. Oviedo, J. L. Ferres, G. Bowman. *Workshop on Structural Biology, Neural Information Processing Systems (NeurIPS) 2021*
- 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

Patents

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

Mentoring

- I have a lot of experience mentoring interns, rotation students, graduate students and team members as part of collaborations and projects. I want to mention students involved in the most significant of these projects: Gaurav Gupta (PhD student, USC), Zhongqi Miao (PhD student, UCBerkeley), Han Yuan (PhD student, MSKCC), Cassandra Burdzyak (PhD student, MSKCC), Jeffrey Law (PhD student, UVA), Michael Ward (PhD student, WU-STL), Artur Meller (PhD student, WU-STL), Jonathan Borowsky (PhD student, WU-STL). Recently, I have also lead an effort to define AI4Good in my team by recruiting several team members, laying out the foundation to aggregate our team's experiences. This has culminated in our AIES 2021 paper.

Other professional activities

- **Consortium activities**: I attend the ORCHARDS meetings (by invitation), which is a multi-institutional collaboration on Coral Bioinformatics. ENCODE consortium meetings (2017-2019).
- **Reviewing AI for health grants**: I help review grant proposals from clinical/ biological research organizations, universities and hospitals, that apply to Microsoft AI-for-health grants.
- **Organizational**: Co-organizer of ICML Workshop for Computational Biology (WCB 2017-2018)
- **Program Committee**: NeurIPS 2016-2021, ICML 2017-2021, ICLR 2018-2021, Bioinformatics, PLoS Computational Biology, MSJAR (Microsoft Journal for Applied Research), Neural Computation, BMC Genomics 2013, IJCAI 2016, WWW Posters 2017-2018, Workshop for ML in Comp Bio 2016-2018, Biotechnology Journal 2017
- **Invited talks**: Invited Panel on Structural Biology at Pacific Symposium of Biocomputing (2021), Panel on healthcare research at Boston University (2020), Fred Hutch immunology lab, Machine Learning seminars (CMU), Pro-active Learning and applications to Computational Biology, University of Pittsburgh (2013)
- **Reading groups**: Organized the matrix factorization reading group at CMU, Machine Learning reading group at IBM Research, Deep Learning reading group at MSKCC
- **White papers**: Wrapper Induction for automatic extraction, TechPulse 2008; Site-Specifc Conditional Random Fields, TechPulse 2008; Web-Scale Information Extraction, TechPulse 2009
- **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

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

Jaime Carbonell	Judith Klein-Seetharaman	Chris Dyer	Christina Leslie
Carnegie Mellon University	Arizona State University	Google DeepMind	Memorial Sloan Kettering Cancer Center

- References available upon request