Barbara E Engelhardt, PhD

Contact Information

Computer Science 322
Princeton University
35 Olden Street
Princeton, NJ 08540 USA

Voice: (609) 258-0933
Cell: (510) 333-8823
Email: bee@princeton.edu
http://www.cs.princeton.edu/∼bee/

Research Interests

Machine learning, nonparametric Bayesian statistics, statistical genetics, latent factor models, hypothesis testing, approximate Bayesian inference, functional genomics, medical data, missing data.

Academic experience

Princeton University, Princeton, New Jersey USA

Assistant Professor **September 2014 - Present**
Computer Science Department and Center for Statistics and Machine Learning. Affiliate in the Lewis-Sigler Institute for Integrative Genomics.

Duke University, Durham, North Carolina USA

Assistant Professor **September 2012 - August 2014**
Biostatistics & Bioinformatics Department, Institute for Genome Sciences & Policy, Department of Statistical Science, and Computer Science Department. Member of Computational Biology & Bioinformatics Faculty and University Program in Genetics and Genomics.

Duke University, Durham, North Carolina USA

Visiting Research Scientist **September 2011 - August 2012**
Independent researcher studying statistical models applied to complex phenotype association studies, differential gene expression analysis, and RNA-seq data to understand mechanism of human disease.

University of Chicago, Chicago, Illinois USA

Postdoctoral Researcher **August 2008 - August 2012**
Advised by Prof. Matthew Stephens, performed research on new statistical approaches for association mapping for complex phenotypes and identifying population structure in genotype data.

Education

University of California, Berkeley, Berkeley, California USA

Ph.D., Computer Science, December, 2007
- Dissertation Title: "Predicting protein molecular function"
- Advisor: Michael I. Jordan
- Designated Emphases: Computational Biology; Communication, Computation and Statistics

Stanford University, Stanford, California USA

M.S., Computer Science (Theory), June, 1999
B.S., Symbolic Systems, June, 1999

Honors and Awards

NIH NHLBI R01 (2017-2021)
Princeton Innovation Helen Shipley Hunt Fund Award (2017-2019)
Sloan Faculty Fellowship (2016-2018)
E. Lawrence Keys, Jr. Emerson Electric Co. Faculty Advancement Award in Engineering (2016)
Princeton Innovation J. Insley Blair Pyne Fund Award (2015)
NIH NIMH R01 GTEx Supplement (2014-2015)
Duke iiD Research Incubator Award (2013)
NIH NIMH GTEx R01 Award (2013-2016)
NIH NHGRI K99/R00 Award (2011-2015)
Bioinformatics Research Development Fund Postdoctoral Fellowship (2008-2011)
Google Anita Borg Scholarship (2005-2006)
Walter M. Fitch Prize from Society for Molecular Biology and Evolution (2004)
National Science Foundation Graduate Research Fellowship (2001)

Teaching Experience

Princeton University, Princeton, NJ USA

Lecturer for Fundamentals of Machine Learning (COS424) **February - June 2015 - 2017**
Designed and taught upper level undergraduate course introducing machine learning with three projects, e.g., music genre classification, imputing DNA methylation, predicting BitCoin transactions.

Lecturer for Data-driven Statistical Genomics (COS597D) **September 2014 - January 2015**
Designed and taught graduate level course on data-driven statistical models for genomic analyses.

Duke University, Durham, NC USA

Lecturer for Probabilistic Machine Learning (STA561/CS571) **August 2013 - December 2013**
Designed and taught graduate level course on machine learning from the probabilistic perspective.

Lecturer for Statistical Methods in Computational Biology (STA613) **January 2013 - May 2013**
Graduate level, case-based course on statistical models used for biology and population genetics.

University of California, Berkeley, Berkeley, California USA

Instructor **January 2005 - May 2005**
Graduate student instructor for undergraduate course CS174 Discrete Probability and Combinatorics (taught by Prof. Karp). Responsible for two sections, solution sets, and administrative issues.

Instructor **August 2004 - December 2004**
Graduate student instructor for graduate course CS281A/Stat241A Statistical Learning Theory (taught by Prof. Jordan). Responsible for recitation section, office hours, and administrative issues.

Stanford University, Stanford, California USA

Instructor **September 1995 - June 1997**
Teaching assistant for Stanford undergraduate programming courses (106 series). Responsible for section, exams, homework assignments, and grading.

Professional Experience

23andMe, Mountain View, California USA

Scientist, Research Group **June 2007 - August 2008**
Under Dr. Joanna Mountain, developed statistical methods for calling SNPs.

Google, Mountain View, California USA

Summer Intern, Research Group **May 2005 - August 2005**
Under Dr. David Pablo Cohn, developed nonparametric statistical models for population structure.

Jet Propulsion Laboratory, NASA, Pasadena, California USA

Member, Technical Staff **June 1999 - August 2001**
Research focused on continuous-time and -resource system models, planning algorithms, and statistical analyses of algorithm robustness in support of autonomous spacecraft.

Preprints in review

V Feinberg, L-F Cheng, K Li, **BE Engelhardt**. Large linear multi-output Gaussian process learning for time series. *(in review)* arXiv:1705.10813.

IC McDowell, D Manandhar, CM Vockley, A Schmid, TE Reddy, **BE Engelhardt**. Clustering gene expression time series data using an infinite Gaussian process mixture model. *(in review)* bioRXiv:131151.

L-F Cheng, G Darnell, C Chivers, ME Draugelis, K Li, **BE Engelhardt**. Sparse multi-output Gaussian processes for medical time series prediction. *(in review)* arXiv:1703.09112.

D Aguiar, L-F Cheng, B Dumitrascu, F Mordelet, AA Pai, **BE Engelhardt**. BIISQ: Bayesian nonparametric discovery of Isoforms and Individual Specific Quantification. *(in review)* arXiv:1703.08260.

ME Basbug, **BE Engelhardt**. Coupled compound Poisson factorization. *(in review)* arXiv:1701.02058.

A Saha, Y Kim, ADH Gewirtz, B Jo, C Gao, IC McDowell, GTEx Consortium, **BE Engelhardt***, A Battle*. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *(in review)* bioRXiv:078741.

B Jo, Y He, BJ Strober, P Parsana, F Aguet, AA Brown, SE Castel, ER Gamazon, A Gewirtz, G Gliner, B Han, AZ He, EY Kang, IC McDowell, X Li, P Mohammadi, CB Peterson, G Quon, A Saha, AV Segré, JH Sul, TJ Sullivan, KG Ardlie, CD Brown, DF Conrad, NJ Cox, ET Dermitzakis, E Eskin, M Kellis, T Lappalainen, C Sabatti, GTEx Consortium, **BE Engelhardt***, A Battle*. Distant regulatory effects of genetic variation in multiple human tissue. *(in review)* bioRXiv:074419.

G Sabnis, D Pati, **BE Engelhardt**, N Pillai. A divide and conquer strategy for high dimensional Bayesian factor models. *(in review)* arXiv:1612.02875.

JT Ash, **BE Engelhardt**, Robert E Schapire. Unsupervised domain adaptation using approximate label matching. *(in review)* arXiv:1602.04889.

IC McDowell, AA Pai, C Guo, CM Vockley, CD Brown, TE Reddy*, **BE Engelhardt***. Many long intergenic non-coding RNAs distally regulate mRNA gene expression levels. *(in review)* bioRXiv:044719.

A Valente, G Ginsburg, **BE Engelhardt**. Nonparametric reduced-rank regression for multi-SNP, multi-trait association mapping. *(in review)* arXiv:1512.02306.

B Dumitrascu, G Darnell, J Ayroles, **BE Engelhardt**. A Bayesian test to identify variance effects. *(in review)* arXiv:1512.01616.

ME Basbug, **BE Engelhardt**. Clustering with beta divergences. *(in review)* arXiv:1510.05491.

BE Engelhardt, RP Adams. Bayesian structured sparsity from Gaussian fields. *(in review)* arXiv:1407.2235.

C Gao, CD Brown, **BE Engelhardt**. A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. *(in review)* arXiv:1310.4792.

Publications

N Prasad, L-F Cheng, C Chivers, M Draugelis, **BE Engelhardt**. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *(in press) Proceedings of Uncertainty in Artificial Intelligence (UAI).* arXiv:1704.06300.

S Zhao, **BE Engelhardt**, S Mukherjee, DB Dunson. Fast moment estimation for generalized latent Dirichlet models. *(in press) Journal of the American Statistical Association (JASA).* arXiv:1603.05324.

S Srivastava, **BE Engelhardt**, DB Dunson. Expandable factor analysis. *(in press) Biometrika.* arXiv:1407.1158.

G Darnell, S Georgiev, S Mukherjee, **BE Engelhardt**. Adaptive randomized dimension reduction on massive data. *(in press) Journal of Machine Learning Research (JMLR).* arXiv:1504.03183.

G Jerfel, ME Basbug, **BE Engelhardt** (2017). Dynamic collaborative filtering with compound Poisson factorization. *Proceedings of Artificial Intelligence and Statistics (AISTATS) Conference* 54:738-747.

JD Cohen, N Daw, **BE Engelhardt**, U Hasson, K Li, Y Niv, KA Norman, J Pillow, PJ Ramadge, NB Turk-Browne, TL Willke (2017). Computational approaches to fMRI analysis. *Nature Neuroscience* 20(3):304313.

PD Tonner, CD Darnell, **BE Engelhardt***, A Schmid* (2016). Detecting differential growth of microbial populations with Gaussian process regression. *Genome Research* 27:320-333.

S Zhao, C Gao, S Mukherjee, and **BE Engelhardt** (2016). Bayesian group latent factor analysis with structured sparsity. *Journal of Machine Learning Research (JMLR)* 17:1–47.

ME Basbug, **BE Engelhardt** (2016). Hierarchical compound Poisson factorization. *International*

Conference on Machine Learning (ICML).

C Gao, S Zhao, IC McDowell, CD Brown, and **BE Engelhardt** (2016). Differential gene co-expression networks via Bayesian biclustering models. *PLOS Computational Biology* 12(7):e1004791.

D Mimno, DM Blei, **BE Engelhardt** (2015). Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences* 112(26):E334150.

W Zhang, TD Spector, P Deloukas, JT Bell, **BE Engelhardt** (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology* 16(1):14.

Genetics of Personality Consortium (2015). Meta-analysis of Genome-wide Association Studies for Neuroticism, and the Polygenic Association With Major Depressive Disorder, *JAMA Psychiatry* 72(7):642–650.

BE Engelhardt & CD Brown (2015). Diving deeper to predict noncoding sequence function, *Nature Methods* 12(10):925–926. *Nature News & Views article; not peer reviewed*

AB Hart, ER Gamazon, **BE Engelhardt**, P Sklar, AK Kähler, CM Hultman, PF Sullivan, BM Neale, SV Faraone, Psychiatric Genomics Consortium: ADHD Subgroup, H de Wit, NJ Cox, A Palmer (2014). Genetic variation associated with euphorigenic effects of *d*-amphetamine is associated with diminished risk for schizophrenia and ADHD. *Proceedings of the National Academy of Sciences* 111(16):5968-5973.

LM Mangravite*, **BE Engelhardt***, MW Medina, JD Smith, CD Brown, DI Casman, BH Mecham, B Howie, H Shim, D Naidoo, QP Feng, MJ Rieder, YD Chen, JI Rotter, PM Ridker, JC Hopewell, S Parish, J Armitage, R Collins, RA Wilke, DA Nickerson, M Stephens*, RM Krauss* (2013). A statin-dependent QTL for *GATM* expression is associated with statin-induced myopathy. *Nature* 502(7471):377-380.

CD Brown, LM Mangravite, **BE Engelhardt** (2013). Integrative modeling of eQTLs and cis-regulatory elements suggest mechanisms underlying cell type specificity of eQTLs. *PLoS Genetics* 9(8):e1003649.

KE Muratore, **BE Engelhardt**, JR Srouji, MI Jordan, JF Kirsch (2013). Molecular function prediction for a family exhibiting evolutionary tendencies towards substrate specificity swapping: Recurrence of tyrosine aminotransferase activity in the 1α subfamily *Proteins* 81(9):1593–1609.

F Mordelet, J Horton, A Hartemink, **BE Engelhardt** and R Gordan (2013). Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* 29(13):i117–i125.

AB Hart*, **BE Engelhardt***, MC Wardel, G Sokoloff, M Stephens, H de Wit, AA Palmer (2012). Genome-wide association study of *d*-amphetamine response in healthy human volunteers identifies putative associations, including cadherin 13 (*CDH13*). *PLoS ONE* 7(8):e42646.

BE Engelhardt, MI Jordan, JR Srouji, and SE Brenner (2011). Genome-scale phylogenetic function annotation of large and diverse protein families. *Genome Research* 12:1969–1980.

BE Engelhardt, M Stephens (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* 6(9):e1001117.

JK Pickrell, JC Marioni, AA Pai, JF Degner, **BE Engelhardt**, E Nkadori, JB Veyrieras, M Stephens, Y Gilad, JK Pritchard (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.

BE Engelhardt (2007). "Predicting protein molecular function." PhD Thesis, Electrical Engineering and Computer Science Department, University of California, Berkeley.

BE Engelhardt, MI Jordan, and SE Brenner (2006). A Graphical Model for Predicting Protein Molecular Function. *International Conference on Machine Learning* (ICML).

BE Engelhardt, MI Jordan, KE Muratore, and SE Brenner (2005). Protein Molecular Function Prediction by Bayesian Phylogenomics. *PLoS Computational Biology* 1(5):e45.

E Amir and **BE Engelhardt** (2003). Factored Planning. *International Joint Conference on Artificial Intelligence (IJCAI).*

BE Engelhardt, S Chien (2001). Stochastic Optimization for Adapting Behaviors of Exploration Agents. *International Symposium on Artificial Intelligence, Robotics, and Automation for Space (i-SAIRAS 2001).*

BE Engelhardt, S Chien, D Mutz (2000). Hypothesis Generation Strategies for Adaptive Problem Solving. *IEEE Aerospace Conference.* Big Sky, MT.

* indicates equal authorship

Talks

Bayesian Nonparametrics Conference, Keynote (2017). "Bayesian nonparametrics in the wild: Opportunities and challenges in genomic analysis."

Oxford University, Big Data Institute Seminar (2017). "Intersecting pathology images and gene expression data to understand drivers of complex phenotypes."

Cold Spring Harbor Laboratory, Biology of Genomes (2017). "Intersecting pathology images and gene expression data to understand drivers of complex phenotypes."

Banff Statistical Genetics and Genomics Workshop (2017). "Intersecting pathology images and gene expression data to understand drivers of complex phenotypes."

California Institute of Technology, CMS Colloquium (2017). "Structured factor models to find interpretable signal in genomic data."

NIPS Advances in Approximate Bayesian Inference Workshop Invited Speaker (2016). "Structured latent factor models to recover interpretable networks from transcriptomic data."

University of Toronto Department of Statistics Seminar (2016). "Structured latent factor models to recover interpretable structure from transcriptomic data."

UC Berkeley Center for Computational Biology Seminar (2016). "Structured latent factor models to recover interpretable networks from transcriptomic data."

University of Pennsylvania Mathematical Biology Seminar (2016). "Structured latent factor models to recover interpretable networks from transcriptomic data."

Probabilistic Modeling and Genomics (Probgen 2016). "Statistical analysis of transcriptomic time series data."

Wellcome Trust Sanger Institute Invited Speaker (2016). "Distant regulatory effects of genetic variation in multiple human tissues in GTEx."

Interational Conference on Machine Learning (ICML) Computational Biology Workshop Invited Speaker (2016) "Unsupervised estimation of context-specific gene networks using a Bayesian biclustering model."

Cold Spring Harbor Systems Biology (2016). "Using gene co-expression networks to study the genetic basis of complex disease."

Dana Farber Cancer Institute (2016). "Latent factor models for the study of complex traits."

UCLA Computational Biology Seminar (2016). "Latent factor models for the study of complex

traits."

Human Genetics in New York City Colloquium (2016). "Using gene co-expression networks to study the genetic basis of complex disease."

Memorial Sloan Kettering Computational Biology Colloquium (2015). "Recovering usable hidden structure using exploratory data analyses on genomic data."

Neural Information Processing Systems (NIPS) Machine Learning Methods in Computional Biology Workshop Keynote (2015). "Bayesian structured sparsity using Gaussian fields."

CSHL Probabilistic Modeling in Genomics (2015). "Heteroskedastic linear models: Going beyond mean effects in genomics."

American Society for Human Genetics (2015). "Sex-specific gene co-expression networks."

Microsoft Research New England Seminar Series (2015). "Recovering usable hidden structure using exploratory data analyses on genomic data."

Banff Statistical Genetics and Genomics Workshop (2015). "Heteroskedastic linear models for functional genomics."

Gotham Seminar on Genomics and Statistics, Columbia University (2015). "Bayesian latent factor models: moving beyond exploration in genomic data."

ENAR (2015). "Allele specific expression to identify causal functional QTLs."

Johns Hopkins University, Biostatistics Seminar (2014). "Bayesian latent factor models recover gene networks and expression QTLs."

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology Seminar (2014). "Bayesian latent factor models recover gene networks and expression QTLs."

Columbia University, Statistics Seminar (2014). "Bayesian latent factor models recover gene networks and expression QTLs."

Joint Statistical Meeting (2014). "Bayesian structured sparsity to uncover eQTLs."

Biology of Genomes (2014). "Identification of long intergenic non-coding RNA QTLs in four tissue types reveals association with metabolic phenotypes."

University of Washington, Statistics Seminar (2014). "Bayesian structured sparsity for genetic association mapping."

Princeton University, Computer Science Department (2013). "Bayesian models of structured sparsity for discovery of regulatory genetic variants."

GTEx Community Meeting (2013). "Replication of cis- and trans-eQTLs across cell types."

Biological Sequence Analysis and Probabilistic Models, Janelia Farm (2013). "Efficient Bayesian structured sparsity models to discover multiple regulatory genetic variants."

American Society of Human Genetics (2012). "Comparative eQTL analyses within and between seven tissue types suggest mechanisms underlying cell type specificity of eQTLs."

Research Triangle Statistical Genetics Conference (2012). "Statistical approaches for analysis of complex phenotypes in genome-wide association studies"

Duke University, University Program on Genetics and Genomics (2012). "Interpreting genome-wide association studies: mechanisms underlying eQTL cell type specificity"

University of North Carolina, Chapel Hill, Genetics Department (2012). "Analysis of complex phenotypes for genome-wide association studies"

Columbia University, Computer Science Department (2011). "Analysis of complex phenotypes in two genome-wide association studies."

Princeton University, Computer Science Department (2011). "Genome-wide associations studies

with complex phenotypes: How statistics can help."

Duke University, Biostatistics and Bioinformatics Department (2011). "Sparse factor analysis for genomic problems."

Mount Holyoke College, Department of Computer Science (2011). "Graphical models and HMMs for gene finding."

University of Massachusetts, Amherst, Department of Computer Science (2011). "Sparse factor analysis for genomic problems."

Duke University, Department of Statistical Science (2010). "Sparse factor analysis in diverse genomic systems."

SuSTain Sparsity Workshop (2010). "Sparse factor analysis applied to biological problems."

Oxford University, Wellcome Trust Centre for Human Genetics (2010). "Sparse factor analysis applied to biological problems."

American Society of Human Genetics (2009). "Modeling population structure using sparse factor analysis."

Automated Function Prediction (2006). "SIFTER: a model of molecular function evolution to predict protein function."

International Conference on Phylogenomics (2006). "A statistical model of molecular function evolution to predict protein function."

Princeton University, Department of Computer Science (2006). "A statistical model of molecular function evolution to predict protein function."

International Conference on Machine Learning (2006). "A graphical model for predicting protein molecular function."

Society for Molecular Biology and Evolution (2004). "Protein function prediction using a Bayesian model of molecular function evolution." (*Walter M. Fitch Prize*)

Service

Committee on Athletics and Campus Recreation (Appointed), Princeton University (2016-2017).
Co-Organizer, NIPS Machine Learning and Computational Biology Workshop (2016).
Associate Editor, Annals of Applied Statistics (2016-present).
Program Committee, RECOMB (2016).
Co-Organizer, RECOMB Genomics Satellite Meeting (2016).
Princeton's Big Data Requirements Group (2015-present).
Princeton's Women's Water Polo Faculty Liason (2014-present).
Co-Organizer, Center for Statistics & Machine Learning Reading Group (2014-2016).
Co-Organizer, Center for Statistics & Machine Learning Seminar Series (2015-2016).
Lewis-Sigler Affiliated Faculty (2015-present).
PICSciE's Associated Faculty (2015-present).
Graduate Admissions Committee, Computer Science, Princeton University (2015-2016).
Faculty Search Committee, Ecology & Evolutionary Biology and Lewis-Sigler Institute, Princeton University (2015-2016).
Faculty Search Committee, Center for Statistics & Machine Learning, Princeton University (2015-2016).
Faculty Search Committee, Computer Science, Princeton University (2015-2016).
University College Life Committee (Elected), Princeton University (2015-2017).
Organizer, Probabilistic Models for Genetics and Genomics (ProbGen) (2015).
Co-Organizer, New York Area Population Genomics Workshop (2015-2017).
Program Committee, International Conference on Machine Learning (ICML) (2015).
Faculty Search Committee, Department of Computer Science (2015).
Women in Machine Learning (WiML) Board of Directors (2014-present).
Co-Organizer, Duke University Machine Learning Seminar Series (2013-2014).
Faculty Search Committee, Electrical and Computer Engineering, Duke University (2013).
Academic Editor, PeerJ (2012-present).

Program Committee, Artificial Intelligence and Statistics (AISTATS) (2013).
Scientific Program Committee, Translational Bioinformatics (TBI) (2013).
Graduate Student Admissions Committee, Graduate Program in Computational Biology and Bioinformatics, Duke University (2012).
Faculty Search Committee, Center for Computational Biology, UC Berkeley (2006).
Graduate Student Admissions Committee, Electrical Engineering and Computer Science Department, UC Berkeley (2005).

Journal/Conference Reviewer:

Science (2016,2017), *Nature* (2015), *Nature Genetics* (2014,2015,2016,2017), *Nature Methods* (2015,2016), *Pacific Symposium on Biocomputing* (2013), *Public Library of Science (PLoS) Computational Biology* (2013), *Annals of Applied Statistics* (2013), *Public Library of Science (PLOS) Genetics* (2012, 2014, 2015), *Journal of the American Statistical Association* (2012, 2014, 2016), *Molecular Biology and Evolution* (2012), *Genome Research* (2012, 2013, 2014), *Conference on Artificial Intelligence and Statistics (AISTATS)* (2011), *Genetics* (2011), *American Journal of Human Genetics* (2011, 2015), *Bioinformatics* (2011, 2012, 2014, 2015, 2016), *Briefings in Bioinformatics* (2011), *Journal of Machine Learning Research (JMLR)* (2010, 2012, 2013, 2015), *Proceedings of the National Academy of Sciences (PNAS)* (2009,2016), *BMC Evolutionary Biology* (2008), *Conference on Research in Computational Molecular Biology* (2016), *Conference on Intelligent Systems for Molecular Biology* (2016), *The International Conference on Machine Learning* (2007, 2008, 2009, 2010, 2012, 2016), *Automated Function Prediction* (2007), *Advances in Neural Information Processing Systems* (2006, 2007, 2009, 2010)