jacob**schreiber**

https://jmschrei.github.io/

whoami

jmschreiber91@gmail.com
@jmschreiber91
jmschreiber91
@jmschrei

awards

Stanford Dean's Award Fellowship
Ruth L. Kirschsten Fellowship
NSF IGERT Big Data Fellowship
ACM-BCB Best Paper 2020
ISMB RegSys Best Presentation 2023

education

2020-Now	**Postdoctoral Fellow, Genetics** Anshul Kundaje	Stanford University
2020-2020	**Postdoctoral Fellow, Genome Science** William Stafford Noble	University of Washington
2016-2020	**Ph.D.** Comp. Sci. and Eng. & Advanced Data Science William Stafford Noble	University of Washington
2014-2016	M.S. Computer Science and Engineering	University of Washington
2009–2013	B.S. Cum Laude Biomolecular Engineering	University of California, Santa Cruz

experience

Current — **Editorial Roles**
I am on the editorial board of reviewers for the Journal of Machine Learning Research (JMLR, `https://jmlr.csail.mit.edu/`) and the editorial board for the Journal of Open Source Software (JOSS, `https://joss.theoj.org/`) and the Stanford AI Lab Blog (SAIL, `https://ai.stanford.edu/blog/`).

2/24/17-4/4/19 — **Core Developer, scikit-learn**
Reference: Gael Varoquaux <gael.varoquaux@inria.fr>
I served as a core developer on the scikit-learn team with a focus on the tree code-base (e.g., decision trees, random forests, gradient boosting) but also reviewed issues and PRs related to probabilistic models.

2017, Summer — **Research Intern, Autopilot Maps, Tesla**
Reference: Nathan Jones <najones@tesla.com>
This internship focused on exploring new ways that machine learning can improve Tesla AutoPilot. The projects involved processing terabytes of fleet data, doing exploratory data analysis, and building working machine learning prototypes.

2016, Summer — **Research Intern, Aspen Technology**
Reference: Mike Noskov <Mike.Noskov@aspentech.com>
This internship focused developing a machine learning implementation that could be deployed in-house to analyze internal data and make structured predictions.

2015, Summer — **Software Engineering Intern, Neurospin, INRIA**
Reference: Olivier Grisel <olivier.grisel@inria.fr>
This internship focused on speeding up the gradient boosting implementation in scikit-learn and resulting in speedups for most tree-based models.

publications

12 first author journal articles + **2 first author perspectives** + **9 non-first journal articles** + **1 first author highlight** + **1 first author workshop paper** + **2 first author preprints** + **1 non-first author preprint**

[28] S. Nair, M. Ameen, L. Sundaram, A. Pampari, A. Balsubramani, **J. Schreiber**, Y.X. Wang, D. Burns, I. Karakikes, H. Blau, K.C. Wang, A. Kundaje. *bioRxiv* (2023). "Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency" **Contribution: I performed supplemental analyses and computationally confirmed findings.**

[27] B. Baur, J. Shin, **J. Schreiber**, S. Zhang, Y. Zhang, M. Manjunath, J.S. Song, W.S. Noble, S. Roy. "Leveraging epigenomes and three-dimensional genome organization for interpreting regulatory variation" *PLOS Computational Biology* (10). **Contribution: I trained an imputation model on the data, and helped edit the paper.**

[26] **J. Schreiber**, C. Boix, H. Li, Y. Guan, C. Chang, J. Chang, A. Hawkins-Hooker, B. Schölkopf, G. Schweikert, M.R. Carulla, A. Canakoglu, F. Guzzo, L. Nanni, M. Masseroli, M.J. Carman, P. Pinoli, C. Hong, K.Y. Yip, J.P. Spence, S.S. Batra, Y.S. Song, S. Mahony, Z. Zhang, W. Tan, Y. Shen, Y. Sun, M. Shi, J. Adrian, R. Sandstrom, N. Farrell, J. Halow, K. Lee, L. Jiang, X. Yang, C. Epstein, J.S. Strattan, B. Bernstein, M. Snyder, M. Kellis, W. Stafford, A. Kundaje. "The ENCODE Imputation Challenge: a critical assessment of methods for cross-cell type imputation of epigenomic profiles" *Genome Biology* (2023). **Contribution: I oversaw the challenge, analyzed the submitted models, and wrote the paper.**

[25] A. Frankish, S. Carbonell-Sala, M. Diekhans, I. Jungreis, J.E. Loveland, J.M. Mudge, C. Sisu, J.C. Wright, C. Arnan, I. Barnes, A. Banerjee, R. Bennett, A. Berry, A. Bignell, C. Boix, F. Calvet, D. Cerdán-Vélez, F. Cunningham, C. Davidson, S. Donaldson, C. Dursun, R. Fatima, S. Giorgetti, C.G. Giron, J.M. Gonzalez, M. Hardy, P.W. Harrison, T. Hourlier, Z. Hollis, T. Hunt, B. James, Y. Jiang, R. Johnson, M. Kay, J. Lagarde, F.J. Martin, L.M. Gómez, S. Nair, P. Ni, F. Pozo, V. Ramalingam, M. Ruffier, B.M. Schmitt, **J.M. Schreiber**, E. Steed, M. Suner, D. Sumathipala, I. Sycheva, B. Uszczynska-Ratajczak, E. Wass, Y.T. Yang, A. Yates, Z. Zafrulla, J.S. Choudhary, M. Gerstein, R. Guigo, T.J.P. Hubbard, M. Kellis, A. Kundaje, B. Paten, M.L. Tress, P. Flicek. "GENCODE: reference annotation for the human and mouse genomes in 2023 " *Nucleic Acids Research* (2022). **Contribution: I participated in consortium activities.**

[24] **J. Schreiber**, S. Nair, A. Balsubramani, A. Kundaje. "Accelerating in-silico saturation mutagenesis using compressed sensing" *Bioinformatics* (2022). **Contribution: I conceived of the method, performed analyses, and was involved in writing the paper.**

[23] S. Nair, A. Shrikumar, **J. Schreiber**, A. Kundaje. "fastISM: performant in silico saturation mutagenesis for convolutional neural networks" *Bioinformatics* (2022). **Contribution: I performed supplementary analyses of the method.**

[22] S. Whalen*, **J. Schreiber***, W.S. Noble, K. Pollard. "Navigating the pitfalls of applying machine learning in genomics" *Nature Reviews Genetics* (2022). **Contribution: I performed analyses and helped with the organization and writing of the paper. I was co-first author, with ordering determined by a random number generator.**

[21] **J. Schreiber**, R. Singh. "Machine learning for profile prediction in genomics" *Current Opinion in Chemical Biology* (2021). **Contribution: I co-wrote this invited article.**

[20] J. Rozowsky, J. Gao, B. Borsari, Y.T. Yang, T. Galeev, G. Gürsoy, C.B. Epstein, K. Xiong, J. Xu, T. Li, J. Liu, K. Yu, A. Berthel, Z. Chen, F. Navarro, M.S. Sun, J. Wright, J. Chang, C.J. Cameron, N. Shoresh, E. Gaskell, J. Drenkow, J. Adrian, S. Aganezov, F. Aguet, G. Balderrama-Gutierrez, S. Banskota, G.B. Corona, S. Chee, S.B. Chhetri, G.C.C. Martins, C. Danyko, C.A. Davis, D. Farid, N.P. Farrell, I. Gabdank, Y. Gofin, D.U. Gorkin, M. Gu, V. Hecht, B.C. Hitz, R. Issner, Y. Jiang, M. Kirsche, X. Kong, B.R. Lam, S. Li, B. Li, X. Li, K.Z. Lin, R. Luo, M. Mackiewicz, R. Meng, J.E. Moore, J. Mudge, N. Nelson, C. Nusbaum, I. Popov, H.E. Pratt, Y. Qiu, S. Ramakrishnan, J. Raymond, L. Salichos, A. Scavelli, **J.M. Schreiber**, F.J. Sedlazeck, L.H. See, R.M. Sherman, X. Shi, M. Shi, C.A. Sloan, J.S. Strattan, Z. Tan, F.Y. Tanaka, A. Vlasova, J. Wang, J. Werner, B. Williams, M. Xu, C. Yan, L. Yu, C. Zaleski, J. Zhang, K. Ardlie, J.M. Cherry, E.M. Mendenhall, W.S. Noble, Z. Weng, M.E. Levine, A. Dobin, B. Wold, A. Mortazavi, B. Ren, J. Gillis, R.M. Myers, M.P. Snyder, J. Choudhary, A. Milosavljevic, M.C. Schatz, B.E. Bernstein, R. Guigó, T.R. Gingeras, M. Gerstein. "The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models" *Cell* (2021). **Contribution: I ran our imputation method on the data tensor and analyzed the results.**

[19] **J. Schreiber**, J. Bilmes, W.S. Noble. "Prioritizing transcriptomic and epigenomic experiments using an optimization strategy that leverages imputed data" *Bioinformatics* (2021). **Contribution: I conceived of the method, implemented it, performed analyses, and was involved in writing the paper.**

[18] **J. Schreiber**, R. Singh, J. Bilmes, W.S. Noble. "A pitfall for machine learning methods aiming to predict across cell types" *Genome Biology* (2020). **Contribution: I conceived of the issue, designed the experiments, and was involved in writing the paper.**

[17] **J. Schreiber**, T. Durham, W.S. Noble, J. Bilmes. "Avocado: Deep tensor factorization characterizes the human epigenome via imputation of tens of thousands of functional experiments" *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (2020). **Contribution: I wrote the highlight.**

[16] **J. Schreiber**, J. Bilmes, W.S. Noble. "apricot: Submodular selection for data summarization in Python" *J. Mach. Learn. Res.* (2020). **Contribution: I conceived of and designed the package, wrote all the code, performed all the analysis, and wrote the paper.**

[15] A. Erijman, L. Kozlowski, S. Sohrabi-Jahromi, J. Fishburn, L. Warfield, **J. Schreiber**, W.S. Noble, J. Söding, S. Hahn. "A High-Throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning" *Molecular Cell* (2020). **Contribution: I advised on the design of method and the analyses.**

[14] **J. Schreiber**, T. Durham, J. Bilmes, W.S. Noble. "Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome" *Genome Biology* (2020). **Contribution: I implemented the method, performed analyses, and was involved in writing the paper.**

[13] **J. Schreiber**, J. Bilmes, W.S. Noble. "Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples" *Genome Biology* (2020). **Contribution: I conceived of the method, implemented it, performed analyses, and was involved in writing the paper.**

[12] **J. Schreiber**, Y.Y. Lu, W.S. Noble. "Ledidi: Designing genome edits that induce functional activity" *Proceedings of the ICML Workshop on Computational Biology* (2020). **Contribution: I conceived of the method, performed analyses, and was involved in writing the paper.**

[11] **J. Schreiber**, D. Hedge, W.S. Noble. "Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics" *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (2020). **Contribution: I conceived of the method, implemented it, performed analyses, and was involved in writing the paper.**

[10] W. Chen, A. McKenna, **J. Schreiber**, M. Haeussler, Y. Yin, V. Agarwal, W.S. Noble, J. Shendure. "Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair" *Nucleic Acids Research* (2019). **Contribution: I advised on the design of computational experiments and performed some of them.**

[9] M. Gasperini, A.J. Hill, J.L. McFaline-Figueroa, B. Martin, S. Kim, M.D. Zhang, D. Jackson, A. Leith, **J. Schreiber**, W.S. Noble, C. Trapnell, N. Ahituv, J. Shendure. "A genome-wide framework for mapping gene regulation via cellular genetic screens" *Cell* (2019). **Contribution: I designed the submodular optimization strategy for selecting loci and ran it.**

[8] **J. Schreiber**. "pomegranate: Fast and Flexible Probabilistic Modeling in Python" *J. Mach. Learn. Res.* (18). **Contribution: I conceived of and designed the package, wrote all the code, performed all the analysis, and wrote the paper.**

[7] **J. Schreiber**, W.S. Noble. "Finding the optimal Bayesian network given a constraint graph" *PeerJ Computer Science* (2017). **Contribution: I conceived of the method, implemented it, performed analyses, and was involved in writing the paper.**

[6] **J. Schreiber**, M. Libbrecht, J. Bilmes, W.S. Noble. "Nucleotide sequence and DnaseI sensitivty are predictive of 3D chromatin architecture" *biorXiv* (2017). **Contribution: I conceived of the method, implemented it, performed analyses, and was involved in writing the paper. After years of work we found a bug in the preprocessing of the data, which was done before I joined the lab, and after correction many of our results no longer held. At that point, I made the personal choice to move on to another project rather than continue. We made a note in the preprint of this issue.**

[5] **J. Schreiber**, K. Karplus. "Analysis of nanopore data using hidden Markov models" *Bioinformatics* (2015). **Contribution: I conceived of the method, implemented it, performed analyses, and was involved in writing the paper.**

[4] **J. Schreiber**, K. Karplus. "Segmentation of Noisy Signals Generated By a Nanopore" *bioRxiv* (2015). **Contribution: I implemented the computational method, performed analyses, and was involved in writing the paper. This paper was submitted to Bioinformatics alongside ``Analysis of nanopore data..." above. Both papers came back with major revisions requested but, because I had just started graduate school, I only had enough time to finish one of them.**

[3] J. Nivala, L. Mulroney, G. Li, **J. Schreiber**, M. Akeson. "Discrimination among protein variants using an unfoldase-coupled nanopore" *ACS Nano* (2014). **Contribution: I designed and performed computational analyses and proof-read the paper.**

[2] Z. Wescoe, **J. Schreiber**, M. Akeson. "Nanopores discriminate among five C5-cytosine variants in DNA" *Journal of the American Chemical Society* (2014). **Contribution: I designed and performed computational analyses and proof-read the paper.**

[1] **J. Schreiber**, Z.L. Wescoe, R. Abu-Shumays, J.T. Vivian, B. Baatar, K. Karplus, M. Akeson. "Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands" *Proceedings of the National Academy of Science* (2013). **Contribution: I designed and performed wetlab experiments and computational analyses and was involved in writing the paper.**

in-progress publications

[4] **J. Schreiber**, S. Nair, A. Kundaje. "DragoNNFruit: Dissecting *cis*- and *trans*-regulation of chromatin accessibility at single-cell resolution"

Abstract: Cellular processes are governed by continuous interactions between cis- and trans-regulatory factors; however, most modeling strategies consider only one of these axes at a time. We introduce DragoNNFruit, a method that explicitly

models the cis- and trans-regulatory factors that drive chromatin accessibility. At a high level, DragoNNFruit is a convolutional neural network whose parameters are dynamically generated by a second network that operates on cell state, effectively conditioning the processing of nucleotide sequences on the cell that predictions are being made for. We validate DragoNNFruit on a reprogramming timecourse and show that it accurately makes base-pair resolution predictions of chromatin accessibility without the need for peak calling or cell clustering. Further, by explicitly accounting for Tn5 bias, we show that DragoNNFruit's denoised predictions can identify dynamic footprints, predict variant effects, and precisely timestamp when regulatory elements turn on across continuous trajectories. Interpreting DragoNNFruit yields explanations for the protein binding sites driving accessibility at individual loci and how this accessibility changes across trajectories. Overall, DragoNNFruit offers a powerful new paradigm for analyzing single-cell experiments.

[3] K. Cochran, **J. Schreiber**, M. Yin, A. Mantripragada, A. Kundaje. "Dissecting the *cis*-regulatory syntax of transcription initiation with deep learning"

Abstract: While many aspects of mammalian Pol II transcription initiation have been extensively characterized, our understanding of the DNA sequence determinants of initiation remains incomplete. A third of human promoters contain no known initiation motifs, and in promoters with known motifs, how those sequence features modulate rates of ranscription and TSS positioning is poorly characterized. We understand even less about transcription initiation at enhancers. To address these knowledge gaps, we trained a deep learning model to learn the mapping between DNA sequence and transcription initiation, measured genome-wide at base resolution by PRO-cap experiments. The model accurately predicts both exact locations of TSSs and overall initiation activity. Next, we applied a model interpretation framework to obtain a high-sensitivity collection of motifs predictive of transcription initiation that includes both core promoter sequence features (TATA box, Inr) and several other TF motifs. To dissect the association between these identified sequence features and initiation, we performed in silico mutational experiments using our model. Systematically, we quantified the nuanced contributions of all motifs to both TSS positioning and the rate of initiation. Our simulations suggest that the sequence features driving initiation are highly epistatic, with initiation outcomes determined by combinations of motifs, but that different motifs play different roles. Results recapitulated known spacing constraints between promoter features and also identified motifs that can perform a dual function as direct initiation sites. Finally, we characterized the sequence determinants of initiation between promoters and enhancers; our findings support a unified model of transcription initiation across both.

[2] A. Pampari, A. Shcherbina, S. Nair, **J. Schreiber**, A. Wang, S. Kundu, A. Shrikumar, A. Kundaje. "Bias factorized, base-resolution deep learning models of chromatin accessibility reveal *cis*-regulatory sequence syntax, transcription factor footprints and regulatory variants"

Abstract: The high resolution morphology of chromatin accessibility profiles measured by DNASE-seq and ATAC-seq experiments is a result of cooperative binding of transcription factors (TFs) to complex sequence syntax encoded in regulatory DNA. However, the DNase-I and Tn5 enzymes have strong and distinct sequence preferences that confound discovery of TF footprints, underlying causal cis-regulatory sequence syntax and regulatory genetic variants from sparse base-resolution chromatin accessibility profiles. Here we introduce ChromBPNet as an integrated solution to all three problems. ChromBPNet is an optimized convolutional neural network architecture that models the influence of genomic sequence context on base-resolution chromatin accessibility profiles. ChromBPNet trained on five ENCODE canonical cell lines achieved superior predictive performance in held-out chromosomes, while automatically learning and optimally regressing out DNase-I/Tn5 enzyme bias. The models are highly performant over a range of sequencing depths, while de-noising and de-sparsifying low coverage signal profiles and dramatically improving concordance between DNase-seq and ATAC-seq experiments. We improved interpretation methods for de-novo inference of predictive TF binding motif models, driver motif instances and associated bias-free TF footprints in accessible elements as well as quantitative effects of higher-order motif syntax on base-resolution profiles. Finally, we develop new in-silico mutagenesis scores to predict the impact of non-coding variants on the strength and shape of base-resolution chromatin profiles. ChromBPNet models of bulk DNase-seq and ATAC-seq expts. outperform deltaSVM and Enformer models in predicting dsQTLs in LCLs and MPRA effects from the CAGI prediction challenge. ChromBPNet also accurately predicts directional effect sizes of SPI1 binding QTLs and caQTLs in LCLs as well as caQTLs in smooth muscle cells and microglia using models trained on cell-type resolved pseudobulk scATAC-seq profiles.

[1] A. Tseng, V. Ramalingam, A. Banerjee, Z. Zafrulla, **J. Schreiber**, A. Shrikumar, A. Kundaje. "Deciphering DNA sequence motifs and motif syntax from neural-network models of in vivo transcription-factor binding profiles"

Background: Much of cellular biology is driven by transcription factors (TFs) binding to DNA. Experimental assays like TF ChIP-seq provide functional readouts of genome-wide binding of a specific TF. A major computational challenge is to use these experiments to reveal the short sequence preferences (i.e. motifs) that elicit binding of a TF, and also the syntax and grammars between motifs that induce binding. Although deep neural networks trained on these assays achieve state-of-the-art predictive performance, and tools now exist to reliably extract local importance scores from these models, it has remained challenging to recover the global decision rules that govern TF binding.

Results: We present a novel framework for extracting motifs, as well as their syntax and grammars, from local importance scores. Our framework interprets a trained neural network to identify the most important subsequences that drive binding. Our framework then extracts higher-order binding rules from the neural network, by examining instances of these discovered

motifs across the genome, and analyzing model predictions on both natural and synthetic sequences. We demonstrate our method's ability to discover novel cis-regulatory biology on a wide variety of ChIP-seq experiments.

Conclusions: Our proposed method distills a set of human-interpretable global decision rules from the model. This allows for computational de novo motif and motif-grammar discovery, all from a single biological experiment. This work constitutes a major step in leveraging neural networks to discover protein–DNA binding rules from data.

talks

25 invited talks + **30 submitted talks** + **2 invited class lectures** + **7 workshops** + **6 lightning/spotlight taks**

Talus Bioscience (10/2/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
Young Investgiator Symposium, IMP/IMBA Vienna (9/25/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
Stowers Institute (9/13/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
ImmunoVec (8/24/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
Insitro (8/16/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
Genentech (6/26/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
Royal Bank of Canada (3/31/2023). apricot: Submodular Selection for Data Summarization
Royal Bank of Canada (8/12/2022). Navigating the pitfalls of applying machine learning in genomics
Retro Bio (6/29/2022). Navigating the pitfalls of applying machine learning in genomics
UK Dementia Institute (6/28/2022). Navigating the pitfalls of applying machine learning in genomics
Talus Bioscience (4/28/2022). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
NVIDIA (1/10/2022). Navigating the pitfalls of applying machine learning in genomics
SFU (10/8/2021). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
23andMe (3/4/2021). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
Valo Health (11/10/2020). Ledidi: designing edits for biological sequences
ODSC Europe 2020 (9/18/2020). pomegranate: fast and flexible probabilistic modeling in Python
Freenome (8/11/2020). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
Lawrence Livermoore National Laboratory (10/29/2019). pomegranate: fast and flexible probabilistic modeling in Python
HudsonAlpha (4/19/2019). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
Nordstrom (7/7/2018). pomegranate: fast and flexible probabilistic modeling in Python
Stanford SCGPM (6/11/2018). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
Jump Trading (2/2/2018). pomegranate: fast and flexible probabilistic modeling in Python
Metis Seattle (10/16/2017). pomegranate: fast and flexible probabilistic modeling in Python
Stanford SCGPM (9/19/2017). Rambutan: predicting three-dimensional genome structure
Tesla Autopilot (7/20/2017). pomegranate: fast and flexible probabilistic modeling in Python
ISMB 2023 (7/25/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
scipy2023 (7/12/2023). tfmodisco-lite: Attribution-guided discovery of biological motifs
CSHL Biology of Genomes (5/11/2023). DragoNNFruit: Learning cis- and trans-regulatory drivers of chromatin accessibility at single-base and single-cell resolution
ODSC West 2022 (11/3/2022). Navigating the pitfalls of applying machine learning in genomics
ISMB 2022 (7/14/2022). Accelerating in-silico saturation mutagenesis using compressed sensing
RECOMB 2022 (5/22/2022). Navigating the pitfalls of applying machine learning in genomics
ENCODE Consortium (3/22/2022). A critical assessment of functional imputation methods
SGTP (2/24/2022). Navigating the pitfalls of applying machine learning in genomics
PyData Global 2021 (10/29/2021). apricot: Submodular Selection for Data Summarization
RSG Dream (11/17/2020). Ledidi: designing edits for biological sequences
CSHL BoG 2020 (11/6/2020). Ledidi: designing edits for biological sequences
ACM-BCB 2020 (9/23/2020). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
ACM-BCB 2020 (9/23/2020). Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics
scipy 2020 (7/9/2020). Avocado: Deep tensor factorization learns a latent representation of the human epigenome

Moore-Sloan Data Science Summit (11/6/2019). apricot: Submodular Selection for Data Summarization
ASHG 2019 (10/17/2019). Deep tensor factorization characterizes the human epigenome through imputation of thousands of genome-wide epigenomics and transcriptomics experiment
ISMB 2019 (7/25/2019). A pitfall for machine learning methods aiming to predict across cell types
scipy 2019 (7/11/2019). apricot: Submodular Selection for Data Summarization
CSHL BoG 2018 (11/8/2018). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
Moore-Sloan Data Science Summit (10/12/2018). pomegranate: fast and flexible probabilistic modeling in Python
ISMB 2018 (7/9/2018). Avocado: Deep tensor factorization learns a latent representation of the human epigenome
PyData NYC (11/30/2017). pomegranate: fast and flexible probabilistic modeling in Python
ODSC West 2017 (11/5/2017). pomegranate: fast and flexible probabilistic modeling in Python
Strata New York (9/27/2017). pomegranate: fast and flexible probabilistic modeling in Python
scipy 2017 (7/18/2017). pomegranate: fast and flexible probabilistic modeling in Python
scipy 2017 (7/14/2017). Rambutan: predicting three-dimensional genome structure
Data Intelligence (7/5/2017). pomegranate: fast and flexible probabilistic modeling in Python
Great Lakes Bioinformatics (5/16/2017). Rambutan: predicting three-dimensional genome structure
Seattle Data Analytics / Machine Learning Meetup (4/25/2017). pomegranate: fast and flexible probabilistic modeling in Python
PyData Chicago (8/26/2016). pomegranate: fast and flexible probabilistic modeling in Python
UW GENOM 560 (5/31/2022). Navigating the pitfalls of applying machine learning in genomics
UC Berkeley CS 294-172 (11/12/2020). Ledidi: designing edits for biological sequences
ODSC West 2021 (11/18/2021). apricot: Taming redundancy in massive data sets using submodular optimization
ODSC West 2019 (10/31/2019). pomegranate: fast and flexible probabilistic modeling in Python
ODSC East 2019 (5/3/2019). pomegranate: fast and flexible probabilistic modeling in Python
ODSC West 2018 (11/3/2018). pomegranate: fast and flexible probabilistic modeling in Python
ODSC West 2017 (11/2/2017). Cython and Multithreading
ODSC East 2017 (5/3/2017). pomegranate: fast and flexible probabilistic modeling in Python
Moore-Sloan Data Science Summit (10/24/2016). pomegranate: fast and flexible probabilistic modeling in Python
ENCODE Consortium (12/10/2019). A critical assessment of functional imputation methods
Moore-Sloan Data Science Summit (11/6/2017). Probabilistic Modeling in the Wild
Moore-Sloan Data Science Summit (10/24/2016). Parallelized Out-of-Core Mixture Modeling in pomegranate
MLCB 2022 (11/21/2022). The ENCODE Imputation Challenge
MLCB 2021 (11/22/2021). Accelerating in-silico saturation mutagenesis using compressed sensing
MLCB 2019 (11/23/2019). Zero-shot imputations across species are enabled through joint modeling of human and mouse epigenomics

software

pomegranate (3,157 stars, 570 forks, used by 941 repos, >2.5M downloads as of 8/16/2023)
pomegranate[8] is a Python package for probabilistic modeling that is a NumFOCUS Affiliated Project (`https://numfocus.org/sponsored-projects/affiliated-projects`). It extends scikit-learn by offering a more flexible API for building and training complex probabilistic models, such as Bayesian networks, hidden Markov models, and mixture models. Users can build models with the many pre-defined distributions or easily implement their own custom ones. `https://github.com/jmschrei/pomegranate`

apricot (480 stars, 45 forks, >117k downloads as of 8/16/2023)
apricot[16] is a Python package that implements submodular optimization for the purpose of summarizing massive data sets into non-redundant subsets that still represent the space of the full data. The package follow the format of scikit-learn so that selection can be done easily and without background knowledge and dropped into existing pipelines. `https://github.com/jmschrei/apricot`

Avocado (112 stars, 20 forks, >50k downloads as of 8/16/2023)
Avocado[13-14, 17] is a Python package that implements deep tensor factorization for the purpose of modeling large, but incomplete, compendia of epigenomic data. The model both learns a low-dimensional representation that is broadly useful and can be used to impute the missing values in the tensor. `https://github.com/jmschrei/avocado`

ledidi (40 stars, 2 forks, >10k downloads as of 8/16/2023)
ledidi[12] turns any machine learning model into a biological sequence editor by swapping the normal training procedure to update data rather than model weights. `https://github.com/jmschrei/ledidi`

tfmodisco-lite (30 stars, 9 forks, >13k downloads as of 8/16/2023)
tfmodisco-lite is a rewrite of the tfmodisco package to take significantly less memory and be faster. tfmodisco-lite extracts

repeated biological motifs based on attribution scores from machine learning models. https://github.com/jmschrei/tfmodisco-lite

PyPore (26 stars, 7 forks)
PyPore is a package for analyzing data generated by nanopore devices, including event detection and event segmentation. https://github.com/kundajelab/yuzu

bpnet-lite (9 stars, 1 forks, >17k downloads as of 8/16/2023)
bpnet-lite is a PyTorch implementation of the BPNet and ChromBPNet models as well as common analysis functions involving these models, such as attribution scores and marginalization experiments. https://github.com/jmschrei/bpnet-lite

yuzu (6 stars, 1 forks, >2k downloads as of 8/16/2023)
yuzu[24] uses compressed sensing to speed up in-silico saturated mutagenesis (ISM) when performed on models with convolution layers. https://github.com/kundajelab/yuzu

scikit-learn (>55.4k stars, >24.6k forks, used by >568k repos, >1.2B downloads as of 5/5/20)
scikit-learn is a Python package that implements classic supervised and unsupervised machine learning algorithms as well as many components of the machine learning ecosystem, such as model evaluation, hyperparamer selection, and data preprocessing steps. I contributed for several years and was a core contributor for around a year, focusing on the tree-based methods (specifically gradient boosting) and probabilistic models. I am now an emeritus core developer. https://github.com/scikit-learn/scikit-learn