Shawn T. O'Neil, PhD

shawn@tislab.org
http://shawntoneil.com
T 541 250 6755

Professional Experience

University of North Carolina Chapel Hill, 2024 – Present
Assistant Professor, Department of Genetics. Developing machine-learning and statistical approaches for health informatics. Observational health research. Teaching and curriculum design in real-world data analytics.

University of Colorado, Anschutz Medical Campus, 2020 – 2024
Assistant Professor, Department of Biomedical Informatics, Center for Health AI. Developing machine-learning and statistical approaches for health informatics. Observational health research. Teaching and curriculum design in real-world data analytics.

Oregon State U., Center for Genome Research and Biocomputing, 2012 – 2020
Senior Faculty Research Assistant; Advanced Cyberinfrastructure Teaching Facility Manager; Lead Bioinformatics Trainer. Bioinformatics teaching and research. Developing curricula in data analysis and programming, developing HPC and cloud (grid engine, kubernetes) teaching infrastructure, analysis and software development for research projects, project management.

Education

2012 **Ph.D., University of Notre Dame, Computer Science and Engineering**
Non-Model Transcriptomics: Applications, Assessments, and Algorithms
Advisors: Dr. Scott J. Emrich (Comp. Sci.), Dr. Jessica J. Hellmann (Bio. Sci.)

2009 **M.S., University of Notre Dame, Computer Science and Engineering**
Expert Advice and the Newsvendor Problem
Advisor: Dr. Amitabh Chaudhary (Comp. Sci. and Eng.)

2005 **B.S., Northern Michigan University, Computer Science**
Minor Mathematics, Summa Cum Laude

Publications

O'Neil ST, Schilder BM, Schaper K, Cox C, Korn D, Gehrke S, ... Haendel MA. 2025. **monarchr: an R package for querying biomedical knowledge graphs.** Bioinformatics. **2025**;41(10), btaf549. doi:10.1093/bioinformatics/btaf549.

O'Neil ST, Madlock-Brown C, Wilkins KJ, McGrath BM, Davis HE, Assaf GS, Wei H, Zareie P, French ET, Loomba J, McMurry JA, Zhou A, Chute CG, Moffitt RA, Pfaff ER, Yoo YJ, Leese P, Chew RF, Lieberman M, Melissa AH, and the N3C and RECOVER Consortia. **Finding Long-COVID: Temporal topic modeling of electronic health records from the N3C and RECOVER programs**. Nature Digital Methods. **2024**;7:296. doi:10.1038/s41746-024-01286-3.

Hurwitz E, Butzin-Dozier Z, Master H, O'Neil ST, Walden A, Holko M, Patel RC, Haendel MA. **Harnessing consumer wearable digital biomarkers for individualized recognition of postpartum depression using the All of Us Research Program data set: cross-sectional study**. JMIR Mhealth Uhealth. **2024**; 12:e54622. doi:10.2196/54622.

O'Neil ST, Schaper K, Elsarboukh G, Reese JT, Moxon SAT, Harris NL, Munoz-Torres MC, Robinson PN, Haendel MA, Mungall CJ. **Phenomics Assistant: An interface for LLM-based biomedical knowledge graph exploration**. bioRxiv. **2024**. doi:10.1101/2024.01.31.578275.

Putman TE, Schaper K, Matentzoglu N, Rubinetti VP, Alquaddoomi FS, Cox C, Caufield JH, Elsarboukh G, Gehrke S, Hegde H, Reese JT, Braun I, Bruskiewich RM, Cappelletti L, Carbon S, Caron AR, Chan LE, Chute CG, Cortes KG, De Souza V, Fontana T, Harris NL, Hartley EL, Hurwitz E, Jacobsen JOB, Krishnamurthy M, Laraway BJ, McLaughlin JA, McMurry JA, Moxon SAT, Mullen KR, O'Neil ST, Shefchek KA, Stefancsik R, Toro S, Vasilevsky NA, Walls RL, Whetzel PL, Osumi-Sutherland D, Smedley D, Robinson PN, Mungall CJ, Haendel MA, Munoz-Torres MC. **The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species**. Nucleic Acids Research **2024**;gkad1082. doi:10.1093/nar/gkad1082.

Toro S, Anagnostopoulos AV, Bello S, Blumberg K, Cameron R, Carmody L, Diehl AD, Dooley D, Duncan W, Fey P, Gaudet P, Harris NL, Joachimiak M, Kiani L, Lubiana T, Munoz-Torres MC, O'Neil S, Osumi-Sutherland D, Puig A, Reese JP, Reiser L, Robb S, Ruemping T, Seager J, Sid E, Stefancsik R, Weber M, Wood V, Haendel MA, Mungall CJ. **Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI)**. *J Biomed Semant* **2023**;15:19. doi:10.1186/s13326-024-00320-3.

Salah HM, Fudim M, O'Neil ST, Manna A, Chute CG, Caughey MC. **Post-recovery COVID-19 and incident heart failure in the National COVID Cohort Collaborative (N3C) study**. Nature Communications **2022**;13(1):4117. doi:10.1038/s41467-022-31801-3.

Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, Payne PRO, Pfaff ER, Robinson PN, Saltz JH, Spratt H, Suver C, Wilbanks J, Wilcox AB, Williams AE,

Wu C, Blacketer C, Bradford RL, Cimino JJ, Clark M, Colmenares EW, Francis PA, Gabriel D, Graves A, Hemadri R, Hong SS, Hripscak G, Jiao D, Klann JG, Kostka K, Lee AM, Lehmann HP, Lingrey L, Miller RT, Morris M, Murphy SN, Natarajan K, Palchuk MB, Sheikh U, Solbrig H, Visweswaran S, Walden A, Walters KM, Weber GM, Zhang XT, Zhu RL, Amor B, Girvin AT, Manna A, Qureshi N, Kurilla MG, Michael SG, Portilla LM, Rutter JL, Austin CP, Gersing KR, and the N3C Consortium. **The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment**. Journal of the American Medical Informatics Association **2021**;28(3):427–443. doi:10.1093/jamia/ocaa196. [Named consortial author]

O'Neil ST. **TidyTensor: Utilities for multidimensional arrays as named hierarchical structures**. Journal of Open Source Software. **2021**;6(66):3543. doi:10.21105/joss.03543.

Annalora AJ, O'Neil S, Bushman JD, Summerton JE, Marcus CB, Iversen PL. **A k-mer based transcriptomics approach for antisense drug discovery targeting the Ewing's family of tumors**. Oncotarget. **2018**;9(55):30568. doi:10.18632/oncotarget.25763.

Ryan SF, Fontaine MC, Scriber JM, Pfrender ME, O'Neil ST, Hellmann JJ. **Patterns of divergence across the geographic and genomic landscape of a butterfly hybrid zone associated with a climatic gradient**. Molecular Ecology **2017**;26(18):4725–4742. doi:10.1111/mec.14256.

O'Neil S, Zhao X, Sun D, Wei JC. **Newsvendor problems with demand shocks and unknown demand distributions**. Decision Sciences **2016**;47(1):125–156. doi:10.1111/deci.12171.

Fister AS, O'Neil ST, Shi Z, Zhang Y, Tyler BM, Guiltinan MJ, Maximova SN. **Two Theobroma cacao genotypes with contrasting pathogen tolerance show aberrant transcriptional and ROS responses after salicylic acid treatment**. Journal of Experimental Botany **2015**;66(20):6245–6258. doi:10.1093/jxb/erv345.

O'Neil ST. **Implementing Persistent O(1) Stacks and Queues in R**. The R Journal **2015**;7(1):118–126. doi:10.32614/RJ-2015-011.

Gouthu S, O'Neil ST, Di Y, Ansarolia M, Megraw M, Deluc LG. **A comparative study of ripening among berries of the grape cluster reveals an altered transcriptional programme and enhanced ripening rate in delayed berries**. Journal of Experimental Botany **2014**;65(20):5889–5902. doi:10.1093/jxb/eru329.

O'Neil ST, Dzurisin JDK, Williams CM, Lobo NF, Higgins JK, Deines JM, Carmichael RD, Zeng E, Tan JC, Wu GC. **Gene expression in closely related species mirrors local adaptation: consequences for responses to a warming world**. Molecular Ecology **2014**;23(11):2686–2698. doi:10.1111/mec.12765.

Abrudan J, Ramalho-Ortigão M, O'Neil S, Stayback G, Wadsworth M, Bernard M, Shoue D, Emrich S, Lawyer P, Kamhawi S. **The characterization of the Phlebotomus papatasi transcriptome**. Insect Molecular Biology **2013**;22(2):211–232. doi:10.1111/imb.12017.

O'Neil ST, Chaudhary A, Chen DZ, Wang H. **The topology aware file distribution**

problem. Journal of Combinatorial Optimization **2013**;26(4):621–635. doi:10.1007/s10878-012-9471-5.

O'Neil ST, Emrich SJ. **Assessing De Novo transcriptome assembly metrics for consistency and utility**. BMC Genomics **2013**;14:1–12. doi:10.1186/1471-2164-14-465.

O'Neil ST, Emrich SJ. **Haplotype and minimum-chimerism consensus determination using short sequence data**. BMC Genomics **2012**;13:1–12. doi:10.1186/1471-2164-13-15.

O'Neil ST, Emrich SJ. **Robust haplotype reconstruction of eukaryotic read data with Hapler**. Proceedings of the IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) **2011**;141–146. doi:10.1109/ICCABS.2011.5729887.

O'Neil ST, Dzurisin JDK, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ. **Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon**. BMC Genomics **2010**;11:1–15. doi:10.1186/1471-2164-11-310.

O'Neil S, Chaudhary A. **Comparing online learning algorithms to stochastic approaches for the multi-period newsvendor problem**. Proceedings of the Algorithm Engineering and Experiments (ALENEX) Conference **2008**;49–63. doi:10.1137/1.9781611972893.5.

Books

Researcher's Guide to N3C: A National Resource for Analyzing Real-World Health Data O'Neil ST, Beasley W, Loomba J, Patrick S, Wilkins KJ, Crowley KM. (Eds.) DOI: 10.5281/zenodo.7749367 (**2023**). An edited volume with researcher resources for the National COVID Cohort Collaborative, 35+ contributors. Lead of editorial team.

Bio/Recursion: Exploring CS and Bioinformatics in R O'Neil ST, Self Published (**2018**). Introduces foundational computer science topics via examples in bioinformatics and the R programming language. Self-published, available in digital and print.

A Primer for Computational Biology O'Neil ST, OSU Press, ISBN 978-0-87071-926-4 (**2017**). An open-access textbook covering skills needed for success in computational biology, in 3 parts: Unix/Linux, Python, and R.

Awards and Honors

University of Notre Dame:

Eck Institute for Global Health Bioinformatics Fellow, 2010.

Kaneb Center Outstanding Graduate Student Teacher, 2008.

Arthur J. Schmitt Fellow, 2006.

Educational Infrastructure

UNC Chapel Hill:

2020–Present N3C Educational Resources. Contributed to the ingestion, harmonization, and curation of notional data in the National Clinical/COVID Collaborative infrastructure, including schema-aligned OMOP common data model datasets and documentation.

Oregon State University:

2018–2020 Data Science @ OSU. IT Division. Co-lead steering committee consisting of campus-wide Associate Deans for Education, a technical advisory subcommittee, and a faculty advisory subcommittee to identify a next-generation platform for data science instruction across colleges at OSU. Performed campus-wide needs assessment with steering committee help, peer institution interviews, and faculty listening groups. After settling on a suitable platform (the Zero to JupyterHub kubernetes-based deployment developed at UC Berkeley and elsewhere, supporting Jupyter, RStudio, and other tools), implemented and deployed on autoscaling AWS infrastructure, with a number of technical improvements including class-shared data storage, role-based Canvas integration, a quota system for cost control, and other features. As of 2024, the system has been used by over a dozen courses and supported thousands of students.

2015–2019 Advanced Cyberinfrastructure Teaching Facility. Center for Genome Research and Biocomputing. Managed class access, software deployments, and systems administration tasks of 8-node HPC cluster developed for multi-departmental bioinformatics workshop and class use.

Conference Presentations

O'Neil ST. **Data, Knowledge, and Intelligence: What's Next for Health Informatics?** American Association of Medical Colleges, Group on Information Resources (AAMC-GIR). *Invited plenary session*, **2023**.

O'Neil ST. **The Monarch Assistant**. AMIA Knowledge Representation and Semantics Working Group Pre-Symposium. *Invited workshop presentation*, **2023**.

O'Neil S, Walden A. **From Health Equity to Long COVID: Exploring big questions with electronic health records. ScienceWriters conference**. *Accepted abstract presentation*, **2022**.

O'Neil ST, Madlock-Brown C, Wilkins K, Zareie P, McGrath B. **Finding Long COVID: Topic modeling of post-infection trends in patient EHR profiles.** ROCKY Bioinformatics conference. *Accepted abstract presentation*, **2022**.

O'Neil ST. **Teaching and Using R: A CS Perspective.** Cascadia R Conference. *Accepted abstract presentation*, **2017**.

O'Neil ST. **Assessing De-Novo Transcriptome Assemblies.** Association of Biomolecular Research Facilities (ABRF). *Accepted abstract presentation*, **2017**.

Hellman J, O'Neil S, Emrich S, Dzurisin J, Williams C. **Related insects show differing amounts of population differentiation and localization of transcribed genes in response to climate.** Arthropod Genomics Symposium. *Accepted abstract presentation*, **2013**.

O'Neil ST, Chaudhary A, Chen D, Wang H. **The topology-aware file distribution problem. Computing and Combinatorics Conference** (COCOON). *Presentation on accepted conference paper*. **2011**.

O'Neil ST, Chaudhary A. **Comparing online learning algorithms to stochastic approaches for the multi-period newsvendor problem.** Algorithm Engineering and Experiments (ALENEX). *Presentation on accepted conference paper*, **2008**.

Other Presentations

O'Neil ST. **LLM Evaluation: Do You Trust Your AI?** UNC Translational and Clinical Sciences Institute, Data Science Lab, *Invited presentation*, **2025**.

O'Neil ST. **Data, Knowledge, and Intelligence: What's Next for Health Informatics?** University of Colorado Anschutz Psychology Dept. Grand Rounds, *Invited presentation*, **2023**.

O'Neil ST. **Experiences in Bioinformatics Education**. University of Oregon, *Invited presentation*, **2014**.

Posters

O'Neil ST, Hurwitz E, Presskreicher R, Roeder C, Varun D, Magesh S, Alhassan S, Lowe D, Haendel MA. **Data Driven Prescribing for Anxiety & Depression: Better, Faster Care with Generative AI & Interpretable ML**. Poster, Wellcome Trust Accelerator, **2025**.

Patrick S, Anzalone J, O'Neil S, Beasley W, Hong S, Zhou A, Wilkins K, Loomba J, Hodder SL. **Developing a Shared Education and Training Curriculum for Real World Data Science in the National COVID Cohort Collaborative (N3C).** National IDeA Symposium of Biomedical Research Excellence, **2024**.

O'Neil ST, Madlock-Brown CR, Wilkins KJ, Zareie P, McGrath BM. Finding Long COVID: **Topic modeling of post-infection trends in patient EHR profiles.** Rocky Mountain Bioinformatics Conference, **2022**.

O'Neil ST. **Fast Reduction of Non-Redundant (Sub-)Interval Graphs with Application to Haplotyping**. Oregon State University Center for Genome Research and Biocomputing Annual Conference, **2016**.

Brenberg T, Colaco A, O'Neil S, McLachlan J. **New Genetic Tools for Estimating Long Term Changes in Forest Composition**. Ecological Society of America, **2011**.

Reviewing

2022, 2026 DIAI Data Access, Integration, and Analysis Grant. Genome British Columbia. Scientific panel reviewer for program emphasizing multi-modal data integration projects.

Organizing

2025 UNC Genetics Department Colloquium Series. Speaker recruitment and co-host for weekly departmental seminar.

2024-2025 Centers for Translational Sciences Award (CTSA), RWD Workforce Development Working Group. Co-lead for national working group focused on understanding the training landscape for real-world data use and research. Lead inventory subgroup.

2024-2025 Real-World Analytics with Harmonized Multi-Site EHR Data *½ day conference workshop*. American Medical Informatics Association (AMIA) Summit. Co-lead organization and presented at hands-on workshop exploring data and tools for real-world health data analysis.

2024 Strategies for Successful Real World Data Research and Workforce Development *½ day workshop*. Addition to Association for Clinical and Translational Science (ACTS). Co-lead organization and presented at community-building workshop focusing real-world data training and development needs.

2023 Exploring AI Seminar *Local 1-day workshop*. CU CRIO Office. Lead organization and presented at workshop on new AI technologies, informing and generating discussion ranging from basic information, use cases, and limitations, to applications in health, practice, research, and education, as well policy considerations.

2023 From Health Equity to Long COVID: Exploring big questions with electronic health records. *½ day conference workshop*. ScienceWriters 2023. Co-lead organization and presented at workshop, focusing on developing relationships between scientists and science journalists and enhancing journalistic data literacy.

2022–2024 Education and Training Domain Team. *Bi-weekly working group*. N3C Consortium. Co-lead a team of researchers with interest in teaching or training using real-world electronic health data.

2021–2022 Enclave Users' Group. *Monthly presentation series*. N3C Consortium. Organized speakers and presented at discussion-oriented presentation series, focusing on N3C data, tools, and methods.

2018 Deep Learning in the Life Sciences. *Conference workshop*. International Conference on Biological Ontology (ICBO) 2018. Organized speakers and logistical planning for IBM-presented deep learning workshop.

2013–2019 CGRB Annual Conference. *1-day conference*. Center for Genome Research and Biocomputing. Contributed to workshop planning and operations, technical, logistical for speakers and poster session.

2013–2018 Bioinformatics Users' Group. *Monthly presentation series*. Center for Genome Research and Biocomputing. Organized speakers and presented at discussion-oriented presentation series, focusing on bioinformatics tools and techniques; over 100 speaker sessions hosted.

2008-2009 Michiana Science Cafe. *Monthly community presentation series*. Society of Schmitt Fellows. Founded and co-organized presentation series, inviting university scientists to share their work with the local community.

Teaching (Credit)

UNC Chapel Hill:

2025, GNET 742 Introduction to Unix/Linux and Python for Biomedical Researchers. Genetics and Molecular Biology (GMB) Program. 1 Credit. Re-designed and taught course on high performance computing and Python programming.

CU Anschutz:

2023–2024, PMED 6410 Methods and Challenges in Observational Health Data Analysis. Personalized Medicine Program. *3 Credit*. Designed and taught course on electronic health record data analysis, focusing on OMOP and introducing vocabularies and mapping, SQL and R, cohort definition, logistic regression, cohort matching, and introductory ML techniques. Taught remote, synchronous.

Oregon State University:

2018–2019, OSU-14-1011 Introduction to R and RStudio. Center for Genome Research and Biocomputing, Professional and Continuing Education (PACE) Program. *2 Credits*. Designed and taught course introducing programming and data analysis with R in RStudio. Fully online, asynchronous and graded for continuing education credit.

2015–2017, MCB599 Simulating Natural Systems. Center for Genome Research and Biocomputing, Dept. of Molecular and Cellular Biology. *1 Credit*. Designed and taught simulation course, covering forces and vectors, agent-based systems, flocking, cellular automata, genetic algorithms, simulated annealing, neural networks, and other techniques.

2014–2016, MCB599 Recursion & Dynamic Programming for Sequence Analysis.. Center for Genome Research and Biocomputing, Dept. of Molecular and Cellular Biology. *1 Credit*. Designed and taught course introducing fundamental computer science concepts used in various bioinformatics algorithms.

2014–2019, ST499/599 Data Programming in R. Center for Genome Research and Biocomputing, Dept. of Statistics. *2 Credits*. Designed and taught programming and data analysis techniques for the R programming language. Includes data types, control structures, functional and object-oriented programming.

2014–2019, MCB599 Introduction to Python I and II. Center for Genome Research and Biocomputing, Dept. of Molecular and Cellular Biology. *1 Credit*. Designed and taught this matched pair of courses to introduce students to bioinformatics programming and CS concepts ranging from mutability to class constructors, APIs, and packages.

2013–2016, MCB525 Techniques in Molecular and Cellular Biology. Dept. of Molecular and Cellular Biology. *3 Credits*. Co-lead bioinformatics instruction for lab-based bench+bioinformatics course. Designed and ran exercises, lectures.

2013–2015, MCB599 Command-Line Data Analysis. Center for Genome Research and Biocomputing, Dept. of Molecular and Cellular Biology. *1 Credit*. Developed and taught companion course to Introduction to Unix/Linux, similarly attended. Focuses on GNU utilities, awk, sed, grep, and others, bash scripting.

2013–2015, MCB599 Introduction to Unix/Linux. Center for Genome Research and Biocomputing, Dept. of Molecular and Cellular Biology. *1 Credit*. Developed and taught a biocomputing-focused course covering bash, paths, permissions, executables, grid engine, and more. A highly popular course, taken by over 60 students per year during my tenure and taught through at least 2019 by others.

2013, MB668 Bioinformatics and Genomics. Dept. of Molecular Biology. *4 Credits*. 33% responsibility for introduction to bioinformatics and genomics course.

2013, MCB599 Bioinformatics Programming. Dept. of Molecular and Cellular Biology. *2 Credits*. 33% responsibility for general introduction to bioinformatics scripting course.

University of Notre Dame:

2010–2011, CSE60132 Basic Computing for Bioinformatics. Computer Science and Engineering. *3 Credits*. Developed and taught a service course aimed at biology graduate students.

Teaching (Non-Credit)

Not included are guest lectures, presentations, or workshops with a primarily organizing role.

UNC Chapel Hill:

2024, 2025 Research at UNC With AllOfUs Data. *Non-credit 3-session workshop*. Partnership with UNC PPMH program and other Genetics department labs. 20% contributor. Provided slides, lectures, and individual help for researchers interested in using AllOfUs data.

CU Anschutz:

2024 AIM-AHEAD and NCATS Training Program, Introduction to Infrastructure. *Non-credit 10 week course*. AIM-AHEAD Consortium. 25% contributor. Provided several lectures and consulting for other content as part of a larger AIM-AHEAD developed training program.

2022, 2024 Introduction to Analyzing Real-World Data Using the National COVID Cohort Collaborative (N3C). *Non-credit 4 or 6 week course*. N3C Consortium. 25% contributor. Introduces learners to N3C access & policy, data, and tools, including introductory OMOP and analyses.

Oregon State University:

2019 Deep Learning for Life Scientists. *Non-credit 6 week course*. Center for Genome Research and Biocomputing. Designed and taught using Keras and Tensorflow in R, introducing students to data loading and augmentation, tensors, autograd, stochastic gradient descent, ML evaluation metrics, activation functions, CNNs, RNNs, autoencoders, and more, with examples drawn from biological sciences.

2019 Data Science for the Public Good. *USDA NIFA Coordinated Innovation Network Program*. Center for Genome Research and Biocomputing. Originally developed at Virginia Tech, the DSPG summer program recruits teams of undergraduate and graduate students, faculty advisors, community stakeholders, and extension faculty to apply data science techniques to issues of community need. In 2019 I co-organized OSU's arm of the 5-university collaborative effort, and developed and taught OSU's associated 2-week training program.

2018 Deep Learning Day. *1-day workshop for graduate program*. Center for Genome Research and Bioinformatics. Designed and taught a full-day workshop for the UO Bioinformatics program in deep learning in R + Keras, covering data loading, tensors, loss functions, batched training, CNNs and RNNs.

2017 Comparing and contrasting R and Python. *½-day workshop for graduate program*. Center for Genome Research and Bioinformatics. Designed and taught a half-day workshop for the UO Bioinformatics program, investigating how these two languages handle classes and objects, variable passing, functional features, and other under-the-hood considerations.

2017 3D Modeling and Printing. *Non-credit 5-week course*. Oregon State U. Craft Center. Designed and taught an evening class introducing 3D model design using TinkerCAD and similar tools, printing student models at the university library.

2016 Artistic Programming. *Non-credit 5-week course*. Oregon State U. Craft Center. Designed and taught an evening class introducing graphical programming with the javascript-based P5.js library.

2012 Introduction to Programming Perl. *Non-credit 2 week course*. Center for Genome Research and Biocomputing. Designed and taught introduction to bioinformatics scripting course with Perl.

Service

CTSA Real World Data Workforce Development Across the Translational Spectrum Working Group. Co-lead a multi-organizational working group focused on understanding the landscape of educational efforts in the use of real-world clinical data. 06/2024–Present.

UNC Chapel Hill Genetics Department Faculty Advisory Committee. Member of a departmental committee discussing and disseminating policy changes and updates in the department 09/2025–Present.

CU Anschutz Department of Biomedical Informatics Education Committee. Member of a departmental committee tasked with defining educational criteria for faculty evaluation and promotion.. 2020.

Oregon State University Search Advocate. Certified training and service program promoting inclusive and meritorious hiring practices at OSU. Invited to serve as Advocate on several search committees. 2017–2020.

Benton County Search and Rescue. Search and rescue organization directed by the Benton Co, OR Sheriff's office. Co-lead of the GIS working group 2021–2023, Member 2017–2024.

Research Statement

With a background in theoretical and applied Computer Science, my research has touched on several inter-related fields. Common themes in my work include 1) development of novel computational methods, 2) application of new and existing methods to questions of scientific interest, and 3) development of tools and resources to support the larger research community.

My earliest work focused on development of machine learning and online algorithms for problems in operations management, a field dominated by parametric statistical approaches. We demonstrated that not only can non-parametric, learning-based methods be applied to such problems, but they can simultaneously provide strong theoretical guarantees and competitive performance.

Most of my research, however, has been in biological sciences, specifically bioinformatics and applied computational biology. In these areas my work has emphasized the study of non-model organisms: those without well-curated genetic or genomic resources, but that are nevertheless useful for understanding our natural environment and relationships to it. Included here are the development of new genomic/genetic resources such as de-novo transcriptomes for important species, and their use in understanding ecological effects of climate change. Also included are novel methods for working with population-level genomic data, new metrics for evaluating transcriptomic resources, and evaluations of commonly used existing metrics. These are well-cited (approximately 150 citations each), highlighting the impact of resource and guidance development for the broader community.

While some of the tools I have developed implement niche novel methods, such as Hapler for haplotype phasing of pooled short-read data, others are of broad interest. These include published R packages: *rstackdeque*, implementing foundational data structures well-integrated with R, *tidytensor*, for manipulating tensor data in deep learning applications, and *monarchr*, for querying and manipulating knowledge graph data, amongst others. *Rstackdeque* has a significant continuing impact, as a dependency for other packages and downloaded 49k+ times.

More recently, my research has focused on clinical informatics, especially via my involvement in the National COVID Cohort Collaborative (N3C), the nation's largest database of de-identified electronic health records, accessed by thousands of researchers producing hundreds of COVID-19 related publications. My core role contributed to infrastructure development, logistical support, and researcher training, but I also contributed data analysis to a variety of interdisciplinary teams. My own research has developed novel methods for EHR analysis, integrating unsupervised machine learning and statistical techniques to better understand Long-COVID subtypes across patient demographics. Relatedly, my supervisory work with the Center for Linkage and Acquisition of Data contributed to educational materials development for new data modalities linked to the NIH AllOfUs database.

At UNC, I plan to engage in similar work in collaboration with other groups integrating recent advances in ML and AI to clinical research. I collaborate on the Cancer Identification and Precision Oncology Center with the NC TraCS institute, RENCI, Lineberger Cancer Center, and others, to more comprehensively identify cancer patient metadata with Large Language Models for clinical trial recruitment. More recently I was awarded pilot funds from the Wellcome Trust for the development of generative AI and ML tools for the measurement and treatment of anxiety and depression, with new colleagues in Genetics, Psychiatry, and Neurology. These activities aim to improve the lives of North Carolinians especially via data-driven translational research.

Overall, my broad research experience makes me a quality collaborator and researcher, and I look forward to contributing further to the scientific enterprise at UNC and beyond.

Teaching Statement

Teaching has been a personal passion since my time as a Computer Science graduate student studying bioinformatics. I recognized a vital need for my collaborators in the Biology department to develop independent computational skills to further their own research. I thus developed and taught a service course, Basic Computing for Bioinformatics, introducing concepts in programming and data analysis in an accessible way. The impact was immediate, rapidly advancing their research and improving collaboration via new shared vocabulary.

Since then, teaching foundational techniques to those most in need has been a consistent theme. While at Oregon State University's Center for Genome Research and Biocomputing (CGRB), I developed and offered short (1 to 2 credit) courses

through the departments of Molecular and Cellular Biology (MCB) and Statistics; these were also available and popular for auditing by faculty, staff, and post-docs. Principally these covered 1) usage and analysis within a Unix/Linux command-line environment, 2) programming for data analysis and software development in Python, and 3) programming for data analysis in R. Together, these served as a curricular foundation that grew to support three other instructors and a half-dozen other courses offered through the CGRB and MCB, as well as a basis for my book A Primer for Computational Biology published by OSU press. Beyond these basics, I also developed and taught more advanced courses in bioinformatics algorithms, simulation, and deep learning. Over 500 learners passed through my classroom at OSU, with consistently high evaluations and more than half taking more than one course.

Since joining the Translational and Integrative Sciences Lab as remote faculty (until recently), my teaching portfolio has shifted to more informal, virtual, and grant-focused efforts. From 2020 to 2024 I served as the Training Coordinator for the National COVID Cohort Collaborative (N3C), hosting the nation's largest de-identified EHR dataset accessed by thousands of researchers. As many researchers were unfamiliar with the relevant data, tools, and regulations, I led a significant effort to rapidly onboard them via various channels. These included live orientations and seminars, recorded videos (with thousands of views), example analyses and notional data, regular discussion groups, and an open-access volume The Researchers' Guide to N3C, where I served as lead editor and author along with 36 colleagues. The Guide has seen subsequent use as instructional materials for the AIM-AHEAD program developed by NCATS. Rounding out my teaching in real-world clinical data analytics, I developed and taught "Methods and Challenges for Observational Health Data Analysis" for the Precision Medicine program at CU Anschutz. Most recently I inherited a course in the Genetics and Molecular Biology (GMB) program at UNC, GNET 742 Introduction to Unix/Linux and Python for Biomedical Researchers. With an updated syllabus UNC students learned fundamental techniques in computing applicable to their own research.

Beyond teaching, I also have an interest in enabling data science education via technology access. Many beginners are unable to set up analysis environments on their personal devices, and availability of computer labs cannot keep up with modern demand. At the CGRB I founded and managed the Advanced Cyberinfrastructure Teaching Facility, a high-performance computing cluster serving bioinformatics education needs across several departments. Later, I co-chaired a steering committee consisting of IT personnel, faculty, and associate deans for education from across campus to identify more scalable and general teaching infrastructure. I designed and deployed the resulting cloud-hosted, auto-scaling platform known as DataScience@OSU, still serving students across campus. Collaborating with

leadership in the Genetics department at UNC, I recently developed and distributed a survey to catalogue the department's educational activities and highlight the contributions of faculty, staff, and students in support of our research mission.

Finally, I will say a few words about my teaching philosophy, especially as it relates to introductory courses. First, I am a strong believer in individual hands-on-keyboards homeworks: as I tell my students, knowledge is absorbed through fingers as much as eyes and ears. Second, exercises should be challenging: the satisfaction of overcoming a difficult task is an incredibly powerful motivator. This requires that I be consistently available for students, via office hours, hallway chats, and asynchronous communication. Whenever possible I allow students to correct mistakes through feedback and earn credit for doing so; very little in professional programming or data analysis is perfect on the first iteration. While I have taught remotely, I prefer in-person classrooms where I can better develop a story and connect with students.

While I have not yet had a chance to teach for UNC, I am looking forward to meeting learnings of all stages, continuing my own education, and advancing the educational mission of the institution.

Service and Engagement Statement

Many of my service and engagement activities take the form of organizing discussion groups, events, and workshops. While at OSU I founded and organized the Bioinformatics Users' Group, a bi-weekly discussion group hosting 120+ topics and presenters over 6 years with 20 to 40 regular attendees. With N3C I organized the similar Enclave Users' Group discussing real-world-data analysis tools and techniques, for 70+ topics and presenters over 3 years, and co-founded the Education and Training Domain Team, a working group dedicated to training and education activities in N3C and beyond. I have helped organize conferences and workshops, including the annual CGRB research conference at OSU, computational ½-day workshops at the University of Oregon, a deep-learning event at CU Anschutz, and clinical data workshops at major conferences AMIA and ACTS.

Beyond academia, I co-founded the Michiana Science Cafe as a graduate student, bringing academic expertise to the local community, taught workshops in 3D printing and Creative Coding at the OSU community arts center, and volunteered as an Oregon state-certified search and rescue member. I've participated in a number of hiring committees at OSU, CU Anschutz, and UNC, and while at OSU was a certified Search Advocate–a volunteer role on search committees ensuring fairness and best practices during hiring.

At UNC I am a member of the Genetics Department Faculty Advisory Committee, served as co-host for the Genetics department's colloquium series, and worked with departmental leadership to survey educational efforts. I am also recently co-lead for a Centers for Translational Sciences Award (CTSA) working group on workforce development for real-world data analysis, leading a component inventorying educational resources in this area nationwide.

I look forward to continuing service and outreach at UNC and in the local community.

—

Last revision: 01/2026