Learning Meaningful Representations of Life

At the Thirty-third Conference on Neural Information Processing Systems

Friday, December 13
Vancouver Convention Center
East Ballroom B

Sponsored by

G + NIH

Full abstracts with all authors, references, and figures can be found in the electronic version at `http://lmrl-bio.github.io/`; online videos can be watched after the event at `http://www.broadinstitute.org/mia`

The open-source LaTeX template, AMCOS_booklet, used to generate this booklet is available at `https://github.com/maximelucas/AMCOS_booklet`

About

LRML at NeurIPS 2019 will begin at 8:45am on Friday, December 13th, at the Vancouver Convention Center, East Ballroom B. Over 300 participants have registered from over 40 institutions and plan to be in attendance. Any questions can be directed to the LMRL Organizing Committee. Up-to-date information will be available online at `lmrl-bio.github.io`.

LMRL

The NeurIPS 2019 workshop Learning Meaningful Representations of Life features some of the world's best efforts in biology and machine learning to spur the next generation of data-driven biological problem-solving. An emphasis on interpretable learning of structure and principles will be applied to work on the level of the genome, molecule, cells, and phenotype.

Organizing committee

Elizabeth Wood	Yakir Reshef	Jon Bloom
Ray Jones	Anne Carpenter	Debora Marks
Barbara Englehardt	Chang Liu	Nikolai Slavov
Kresten Lindorff-Larsen	Wouter Boomsma	James Zou
Suchi Saria	Gopal Sarma	Alexander Wiltschko
Casey Greene	Scott Linderman	Jasper Snoek

Sponsors

LMRL was made possible through the generous support of the National Institutes of Health, GV, and the Broad Institute of Harvard and MIT.

Schedule

Friday, 13 of December – Morning

7:30–8:30	**Poster Setup and Registration**	
		Presenters of accepted posters are welcome to join us at the hall at 7:30AM for early snacks, coffee, and a chance to hang their posters. Presenters are asked to have their posters hung by 8:30AM.
8:30–9:00	**Welcome Addresses**	
8:30–8:45	WA	**Francis Collins (via video)** NIH
8:45–9:00	WA	**Krishna Yeshwant** GV
9:00–10:30	**Keynotes**	
9:00–9:30	KN	**Aviv Regev** MIT, Broad Institute, HHMI Investigator
9:30–10:00	KN	**Max Welling** University of Amsterdam, Qualcomm
10:00–10:30	KN	**Daphne Koller in conversation with Barbara Englehardt** insitro
10:30–10:45	**Coffee Break**	
		Coffee and tea.
10:45–12:00	**Molecules and Genomes (Panel)**	
	S	**Barbara Englehardt, Jennifer Wei, David Haussler, David Jones, Michael Keiser, David Duvenaud, Alan Asparu-Guzik** Jointly Identified Challenges
12:00–12:30	**Synthetic Systems (Panel)**	
	S	**Pamela Silver, Debora Marks, Chang Liu, Possu Huang** Synthetic Biology and ML
12:30–1:15	**Lunch & Poster I**	
		A light lunch will be provided. Posters will be viewable on ballroom walls as well as online after the workshop.

Friday, 13 of December – Afternoon

1:15–3:00	**Phenotype (Breakout)**	
	S	**Challenge Presenters: Casey Greene, Dylan Kotliar, Smita Kirshnaswamy Conversation Facilitators: Alex Wiltschko, Aurel Nagy, Brendan Bulik-Sullivan, Casey Greene, David Kelley, Dylan Kotliar, Eli van Allen, Gokcen Eraslan, James Zou, Matt Johnson, Meromit Singer, Nir Hacohen, Samantha Morris, Scott Linderman, Smita Krishnaswamy** Challenges
3:00–3:15	**Coffee Break**	
		Coffee, tea, and light snacks.
3:15–5:00	**Cell (Session)**	
	S	**Anne Carpenter, Hui Ting Grace Yeo, Jian Zhou, Maria Chikina, Alexander Tong, Benjamin Lengerich, Aly O. Abdelkareem, Gokcen Eraslan, Stephen Ra, Daniel Burkhardt, Emanuel Flores Bautista, Frederick Matsen, Alan Moses, Zhenghao Chen, Marzieh Haghighi, Alex Lu, Geoffrey Schau, Jeff Nivala, Luke O'Connor, Miriam Shiffman, Hannes Harbrecht, Shimbi Masengo Wa Umba Papa Levi** Talks and Lightning Presentations
5:00–6:00	**Closing Remarks and Drinks**	
		Drinks.
5:15–5:30	CA	**Chris Sander** Harvard Medical School, Dana-Farber Cancer Institute
5:30–5:45	CA	**Ila Fiete** MIT, HHMI
5:45–6:00	CA	**Dana Pe'er** Sloan Kettering Institute, Columbia

Useful Information

Talks will be held at the **East Ballroom B** of the Vancouver Convention Center. A layout can be seen at `https://www.vancouverconventioncentre.com/facility/floor-plans-and-specs`. East Ballroom B is situated on the first floor (Convention Level) of the East building, with ground-level accessible access directly from the East Plaza off of Canada Place.

Coffee breaks are provided by the NeuRIPS Conference. In addition, LMRL will offer a light lunch and some more informal coffee and snacks where our unconventional formatting has prevented us from breaking at workshop-wide breaks.

The **LMRL poster session** will be held concurrently, with posters mounted directly on the walls of the ballroom. **Supplies for mounting** will be provided. Presenters are asked to have their posters mounted **no later than 8:30AM**.

Wi-Fi will be available during the conference.

An official but informal **after-conference meet-up** will be held at a local restaurant, beginning around 8:30PM. Details will be provided in person at the workshop. Graduate students who submitted posters will receive a USD$50 stipend for food expenses.

How to get to the East Building of the Vancouver Convention Center?

The Vancouver Convention Centre is in Vancouver, British Columbia, Canada. The East Building is located in Canada Place and is located at 999 Canada Pl, Vancouver, BC V6C 3T4, Canada.

- **Directions:** `https://www.vancouverconventioncentre.com/visiting/getting-here`
- **Transit planning:** `https://tripplanning.translink.ca/#/app/nextdepartures`

Participants

Danilo Bzdok	McGill University
Aurel Nagy	Harvard Medical School
Alex Lu	University of Toronto
Sandhya Prabhakaran	Moffitt Cancer Center
Kehinde Owoeye	University College London
Casey Greene	UPenn/ALSF
Rafael Gomez-Bombarelli	MIT DMSE
Baihan Lin	Columbia University
Anne Carpenter	Broad Institute of Harvard and MIT
Joshua Tan	University of Oxford
Kexin Huang	Harvard University
Siddharth Jain	Caltech
Jack Lanchantin	University of Virginia
David Duvenaud	University of Toronto
Chang Liu	UC Irvine
Romain Lopez	UC Berkeley
Daphne Koller	insitro
Rediet Abebe	Harvard University
Guruprasad Raghavan	Caltech
Yanqing Zhang	Georgia State University
Max Welling	University of Amsterdam & Qualcomm
Paul Villoutreix	Aix-Marseille University
Chris Sander	Harvard Medical School & Dana-Farber Cancer Institute
Sameer Antani	National Institutes of Health
Olga Vitek	Northeastern University
Robert Ness	Gamalon
Wouter Meuleman	Altius Institute
Frederick Matsen	Fred Hutchinson Cancer Research Center
Jian Li	Futurewei Technologies
Gokcen Eraslan	Broad Institute of MIT and Harvard
Hannu Rajaniemi	Helix Nanotechnologies Inc
Jonathan Frazer	Harvard Medical School
Dennis Wang	University of Sheffield
Shuangjia Zheng	Sun Yat-sen University

Ge Liu	MIT
Michael Keiser	UCSF
Miriam Shiffman	MIT & Broad Institute
David Ding	Harvard University/MIT
Samantha Morris	Washington University in St. Louis
Jlan Zhou	UT Southwestern
Robin Winter	Bayer AG
Haoyang Zeng	Massachusetts Institute of Technology
Seonwoo Min	Seoul National University
Arpad Vezer	BenevolentAI
Povilas Norvaisas	BenevolentAI
Craig Glastonbury	BenevolentAI
Eirini Arvaniti	BenevolentAI
AnDr.eas Mayr	Johannes Kepler University Linz
Peter DeWeirdt	The Broad Institute
Zhenghao Chen	Calico Life Sciences
AnDr.eas Mayr	Johannes Kepler University Linz
Hannes Bretschneider	University of Toronto
Satpreet H. Singh	University of Washington
Robert ietswaart	Harvard Medical School
Nick Bhattacharya	UC Berkeley
Neil Thomas	UC Berkeley
Yutong Wang	University of Michigan
Mike Dimmick	University of Toronto / Vector Institute
Christopher Aicher	University of Washington
Shimbi Masengo Wa Umba Papa Levi	University of Pretoria
Samuel Friedman	Broad Institute
Nicki Skafte Detlefsen	Technical University of Denmark
Christian Dallago	Technical University of Munich
Michael Heinzinger	Technical University of Munich
Tomasz Blazejewski	Columbia University
Geoffroy Dubourg-Felonneau	Cambridge Cancer Genomics
Polina Kirichenko	New York University
Stefan Groha	Dana Farber Cancer Institute
Vincent Fortuin	ETH Zurich
David Yang	Harvard College
Tien-yu Hsin	CVS Health/Harvard Medical School
Hui Ting Grace Yeo	Massachusetts Institute of Technology
Jung-Eun Shin	Harvard University

Hyunjin Shim	UC Berkeley
Jose Juan Almagro Armenteros	Technical University of Denmark
Iddo Drori	Columbia University and Cornell University
Sebastian Keller	University of Basel
Janina Kueper	Massachusetts General Hospital/Shriners Hospital for Children
Alina Selega	University of Toronto
Noam Bar	Weizmann Institute of Science
Anika Gupta	Broad Institute of MIT and Harvard
Daniel Burkhardt	Yale University
Jonathan Warrell	Yale University
Mirae Parker	MIT
James Brown	Aldevron
Nathan Rollins	Harvard Medical School
Adnan Akbar	Cambridge Cancer Genomics
Zitong Wang	California Institute of Technology
Jeff Nivala	University of Washington
Mafalda Dias	Harvard Medical School
Zhipeng Wang	Apple, Inc.
Alexander Tong	Yale University
Mariano Gabitto	Simons Foundation
Sai Raghaven Maddhuri Venkata Subramaniya	Purdue University
Felix Raimundo	Google Research / Institut Curie
Jingping Qiao	University of Toronto
Abdullah Yonar	Harvard University
Rahul Nadkarni	University of Washington
Somesh Mohapatra	Massachusetts Institute of Technology
Zachary Wu	Caltech
Libby Zhang	Stanford University
Hannes Harbrecht	University of Cambridge
Tsuyoshi Okita	Kyushu Institute of Technology
Justin Kinney	Cold Spring Harbor Laboratory
Aly Abdelkareem	University of Calgary
Timothy Dunn	Duke University
Alexander Sasse	University of Toronto
Ben Lengerich	CMU
Geoffrey Schau	OHSU
Quanzhi Li	Alibaba Group
Weiguang Mao	University of Pittsburgh

Prashant Emani	Yale University
Kadina Johnston	Caltech
Erik Burlingame	Oregon Health & Science University
Salvatore Loguercio	Scripps Research
Mohamed Kane	WL Research
Wei-Cheng Tseng	National Tsing Hua University
Jeffrey Chan	UC Berkeley
Sergey Ovchinnikov	John Harvard Science Fellow
Stephen Ra	Pfizer, Inc.
Emanuel Flores Bautista	Caltech
Lav Varshney	Salesforce Research and University of Illinois at Urbana-Champaign
Kaushal Paneri	Microsoft
Tzu-Yu Liu	Freenome
Payel Das	IBM Research AI
Stephan Eismann	Stanford University
Samson Koelle	University of Washington Department of Statistics
Ada Shaw	Harvard Medical School
Gonzalo Mena	Harvard University
David Haussler	Uc Santa Cruz Genomics Institute
Ila Fiete	MIT
Dana Pe'er	Sloan Kettering Institution
Matt Karikomi	UC Irvine
Samuel Yang	Google Research
Itsik Pe'er	Columbia University
Louis Cammarata	Harvard University
Ellen Zhong	MIT
Mirae Parker	MIT
Elizabeth Wood	Broad Institute of Harvard and MIT & JURA Bio, Inc.
Mor Nitzan	Harvard
Ray Jones	Broad Institute of Harvard and MIT
Yakir Reshef	HMS & Harvard SEAS
Orr Ashenberg	Broad Institute of Harvard and MIT
Debora Marks	Harvard Medical School
Isabelle Chambers	NDSU
Julie Norville	JURA Bio, Inc.

List of Posters

Label scarcity in biomedicine: Data-rich latent factor discovery enhances phenotype prediction

Marc-Andre Schulz, Gael Varoqua, Alexandre Gramfort, Bertrand Thirion, Danilo Bzdok[1]

[1]McGill

Growing data richness opens the door to leveraging machine learning techniques to gain new insights into disease. Latent factor discovery on large-scale brain imaging data may allow going beyond current diagnostic catalogs by revealing and exploiting objectively measurable biomarkers, potentially enabling early diagnosis and individualized treatment and prognosis. While general population datasets provide a wealth of data on healthy subjects, the amount of high-quality descriptions from diseased subjects remains a major limiting factor. Semisupervised machine learning combines unsupervised structure discovery and supervised prediction, attempting to learn a representation of the data with the help of unlabeled samples that is useful for the prediction task at hand (labeled samples) and benchmarks performance improvements of combined unsupervised-supervised learning strategies that discover hidden endo-phenotype structure in high-dimensional biomedical data.

MultiPLIER: learning representations from public data to study rare diseases

Taroni JN[1], Grayson PC[2], Hu Q[3], Eddy S[4], Kretzler M[5], Merkel PA[6], Greene CS[7]

[1] Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA; Childhood Cancer Data Laboratory, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA.
[2] National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD, USA.
[3] Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA.
[4] Division of Nephrology, Department of Internal Medicine, Michigan Medicine, Ann Arbor, MI, USA.
[5] Division of Nephrology, Department of Internal Medicine, Michigan Medicine, Ann Arbor, MI, USA; Department of Computational Medicine and Bioinformatics, Michigan Medicine, Ann Arbor, MI, USA.
[6] Division of Rheumatology and the Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA.
[7] Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA; Childhood Cancer Data Laboratory, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA; Institute of Translational Medicine and Therapeutics, University of Pennsylvania, Philadelphia, PA, USA; Institute of Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA.

Accurate Cancer Detection by Discovering Meaningful Structures in Gene Data and Using Deep Fuzzy Neural Networks

Thosini K. Bamunu Mudiyanselage, Yanqing Zhang, Yi Pan

Georgia State University

Analysis of gene expression levels for diagnosis has received more popularity today. But current challenges include availability of very few samples, high dimensionality, sparseness and noise data which always make the medical decision process difficult and also reduce the performance significantly. Also, most of the conventional gene selection methods are based on an assumption that genes are independent of each other though these biological data have meaningful relations. To achieve higher performance, we first investigate the dependencies in gene data, and then propose a new algorithm to select an optimal gene set which contain meaningful and most informative genes.

Integrating Markov processes with structural causal modeling enables counterfactual inference in complex systems

Robert Ness[1], **Kaushal Paneri**[2], **Olga Vitek**[2], **Kaushal Paneri**[2]

[1] Gamalon
[2] Northeastern University

Modeling causal relationships between components of dynamic systems helps predict the outcomes of interventions on the system. Upon an intervention, many systems reach a new equilibrium state. Once the equilibrium is observed, counterfactual inference predicts ways in which the equilibrium would have differed under another intervention. Counterfactual inference is key for optimal selection of interventions that yield the desired equilibrium state. Moreover, counterfactual inference provides robustness of causal effect under model misspecification, by making use of past interventional or observational data to condition the misspecified model.

Index and biological spectrum of accessible DNA elements in the human genome

Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglass Dunn, Fidencio Neri, Athanasios Teodosiadis, Alex Reynolds, Eric Haugen, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Richard Sandstrom, Jeff Vierstra, Rajinder Kaul, John Stamatoyannopoulos

Altius Institute

A canonical feature of actuated cis-regulatory elements – promoters, enhancers, silencers, chromatin insulators/enhancer blockers, and locus control regions – is focal alteration in chromatin structure resulting in heightened DNA accessibility to nucleases and other protein factors. From their discovery 40 years ago, DNase I hypersensitive sites (DHSs) have provided reliable signposts for high-precision regulatory DNA delineation in the human and other complex genomes. Here we sought to expand a broad set of high-quality DHS maps, and to unify them into a common reference framework that both (i) incorporates precise genomic annotation that reflects observed biological variability in the pattern of DNA accessibility seen at individual elements, and (ii) captures complex cell-selective behaviors in a quantitative fashion.

Abundant opportunity in machine learning of adaptive immune repertoires

Branden Olson[1], Frederick "Erick" Matsen IV[1], Jean Feng[1], Julia Fukuyama[2], Noah Simon[1], Phil Bradley[1], Fred Hutch[1], Thayer Fisher[1], Vladimir Minin[3], Will DeWitt[1]

[1] University of Washington
[2] Indiana University
[3] UC Irvine

The adaptive immune system is how the body learns to recognize and fight off pathogens that make us sick. Massive diversity of B cell receptors (BCRs, aka antibodies) and T cell receptors (TCRs) is generated via random recombination, filtered through a selective step and selected via clonal expansion. It is now possible to sequence these immune receptors in high throughput to obtain what is called an "immune repertoire." The immune repertoire is determined by immune state and history as well as host genetics (which alleles are present to form immune receptors, and the larger immune context). Unsupervised analysis can get us surprisingly far, identifying host genetics and exposure. Logistic regression can classify past CMV infection with high accuracy. Simple VAEs can do density estimation as accurately as highly tuned probabilistic models.

Defining subpopulations of differential drug response to reveal novel target populations

Dennis Wang[1], Nirmal Keshava[2], Tzen S. Toh[1], Haobin Yuan[1], Bingxun Yang[1], Michael P. Menden[3], Dennis Wang[1]

[1] University of Sheffield
[2] Cerevel Therapeutics
[3] Helmholtz Zentrum München

Various anti-cancer therapies have been developed to target the same gene and pathway. The heterogeneity of cancers prevents clinicians from easily selecting the most effective treatment regimens for their patients. Scientists working on drug development also face a conundrum in predicting the efficacy of their drug of interest. Here, we developed an approach, SEABED (SEgmentation And Biomarker Enrichment of Differential treatment response) based on unsupervised machine learning to identify novel cancer subpopulations, their genetic biomarkers, and effective drug combinations.

A quasilinear framework for interpretable exploratory analysis of single-cell omics data

Jian Zhou[1], **Olga Troyanskaya**[2,3]

[1] UT Southwestern
[2] Flatiron Institute
[3] Princeton University

We contribute a general and practical framework for casting a Markov process model of a system at equilibrium as an SCM, thus leveraging the benefits of both approaches. The SCMs are defined in terms of the parameters and the equilibrium dynamics of the Markov process models, and counterfactual inference flows from these settings. The framework alleviates the identifiability drawback of the SCMs, in that the counterfactual inference is consistent with the counterfactual trajectories simulated from the Markov process model. Moreover, counterfactual inference from the derived SCMs is robust to model misspecification.

Protein structure prediction with deep learning representations

Iddo Drori, Darshan Thaker, Arjun Srivatsa, Daniel Jeong, Yueqi Wang, Linyong Nan, Fan Wu, Dimitri Leggas, Jinhao Lei, Weiyi Lu, Weilong Fu, Yuan Gao, Sashank Karri, Anand Kannan, Antonio Khalil Moretti, Chen Keasar, Itsik Pe'er

Columbia University and Cornell University

Protein structure prediction (PSP) from amino acid sequences is a fundamental problem in computational biology. We use embeddings and deep learning models for prediction of backbone atom distance matrices and torsion angles. We recover 3D coordinates of backbone atoms and reconstruct full-atom proteins by optimization. Key contributions:

- Gold standard dataset of around 75k proteins which is easy to use in developing deep learning models for PSP.
- Competitive results with the winning teams on Critical Assessment of Techniques for Protein Structure Prediction (CASP13) and a comparison with AlphaFold (A7D), results mostly superseding winning teams (CASP12).
- Publicly available source code for both protein structure prediction using deep learning models and protein reconstruction.

Massively parallel reporter assays and interpretable neural networks for biophysical studies of individual cis-regulatory elements

Justin B. Kinney

Cold Spring Harbor Laboratory

There is justifiable excitement about the potential for deep learning to shed light on cis-regulatory codes, i.e., rules that govern how DNA and RNA regulatory sequences control gene expression. Much of the work in this direction has used convolutional and residual neural networks to model the genome-wide activities of cis-regulatory elements (CREs) in diverse biological contexts, including transcriptional regulation and alternative mRNA splicing [1]. A common theme in these efforts is that the resulting models aim to describe the activities of all possible input sequences, even if those sequences do not resemble sequences that occur naturally in the genome. I argue that increased effort should be directed toward a related but distinct problem—the quantitative modeling of how random combinations of single nucleotide polymorphisms (SNPs) affect the activities of specific CREs of interest. This proposal is motivated by the capabilities and limitations of massively parallel reporter assays (MPRAs) [2], as well as the fact that the data produced by SNP-CRE MPRAs is well-suited for quantitative modeling using biophysically interpretable neural networks [3].

What do generative models learn from protein sequences?

Dylan Marshall,[1], **Per Greisen**[2], **Haobo Wang**[1], **Peter Koo**[3], **Sergey Ovchinnikov**[4]

[1] Harvard University
[2] enEvolv
[3] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory
[4] JHDSF Program, Harvard University

Homologous proteins organized by families and represented as a multiple sequence alignment (MSA) provides a rich source of evolutionary information. Statistical models informed by this rich data source are frequently used for two different applications: phylogenetics - quantifying the evolutionary relationship between species, and functional inference - characterizing physical aspects of proteins. Whereas phylogenetics is based on patterns between sequences, functional inference is drawn from patterns between the amino acids of these sequences. Functional patterns, which include conservation and coevolution, can be captured by generative models. However, the patterns that describe phylogeny and function are not easily disentangled, because they both exhibit covariance. Determining the origins of the covariance signal in an MSA remains an outstanding problem in the field. Early models, such as Position-Specific Scoring Matrices (PSSMs), only consider conservation per position. As more sequences became available, Multivariate Gaussian (MG) and Markov Random Field (MRF) models were developed to capture covariance due to molecular coevolution. Recently, variational autoencoders (VAEs) have been introduced to capture higher-order interactions. VAEs employ non-linear latent variables which enable it to capture complex representations in the data. In one application, VAEs trained on MSAs have demonstrated improved performance over MRFs at predicting the functional consequence of mutations. However, it remains unclear what the VAE has learned. Elucidating this can provide a deeper understanding of protein biology. Here, we interpret VAEs trained on MSAs by exploring how their latent variables organize sequences, directly probing the VAE with first-order and second-order in silico mutagenesis, and analyzing the statistical properties of the VAE generated sequences. We unexpectedly find that our VAEs learn a mixture of PSSMs, with a weak covariance signal due to structural coevolution. We suspect this occurs because the largest variance in MSAs is dominated by phylogenetic signal. We demonstrate VAEs approximate mixture models of PSSMs, clustering sequences according to their phylogeny. This suggests that a PSSM from a subset of phylogenetically related sequences serves as a better predictor of the functional consequence of mutations compared to a MRF. Indeed we verify this is true. By extending the one-body term in a MRF model to a mixture-one-body term, we further improve upon predicting the functional effects of mutations. Moreover, we find that the two-body term now captures more accurate protein contacts compared to a MRF that employs a single one-body component. Together, this demonstrates the utility of interpreting highly-flexible black box models to provide data-driven insights into modeling biological

data. Moving forward, our mixture-one-body-MRF provides a robust and interpretable approximation of phylogenetic and functional constraints, laying a path towards a unified statistical model of protein evolution.

Towards Learning Human-Interpretable Laws of Neurogenesis from Single-Cell RNA-Seq Data via Information Lattices

Lav R. Varshney[1,2], **Haizi Yu**[1], **Lav R. Varshne**[1,2], **Genevieve Stein-O'Brien**[3]

[1] University of Illinois at Urbana-Champaign
[2] Salesforce Research
[3] Johns Hopkins University

Basic questions and approach: Is it possible to explain the intricate dance of gene expression in pattern formation by learning a human-interpretable representation directly from raw single-cell RNAseq data? How little prior biological domain knowledge needs to be encoded in learning algorithms to make such discoveries? Can such interpretable representations be connected to the orchestration dynamics of gene regulatory networks, and yield experimentally testable hypotheses? Here we aim to build on our recent approach to automatic concept learning via information lattices [Yu, et al., 2017; Yu, et al., 2019] to discover the laws of retinal neurogenesis. We focus on retina neurogenesis as a simplified model system of the temporal complexity and cellular heterogeneity of the rest of the developing central nervous system We use a dataset that has been used to comprehensively characterize changes in gene expression that occur during initiation of neurogenesis, changes in developmental competence, and specification and differentiation of each major retinal cell type [Clark, Stein-O'Brien, et al., 2019]. The aim is to discover invariant representations that succinctly explain gene expression in neurogenesis

Benchmarking the quality and robustness of representation learning from single-cell RNAseq data

Felix Raimundo[1,2], **Celine Vallot**[2], **Jean-Philippe Vert**[1]

[1] Google Research
[2] Institut Curie

- scRNA-seq allows to understand cell functions at a high granularity but is fairly new.
- Explosion both of new datasets and methods in the last 2 years.
- All downstream analysis of this data passes through a dimension reduction phase, that is never benchmarked in itself.
- The task is unsupervised, and the methods can be very sensible to hyperparameters. Benchmarking needs to report both results and influence of parameters.

Manipulating geometry to facilitate information transmission

Zitong Wang, Matt Thomson

California Institute of Technology

Environmental stimuli are non-uniformly distributed across space. Living things must represent these distributions in order to use this spatial information about the stimuli to learn about its environment, and compute an appropriate behavioural response. Biological systems often represent and transmit information through chemical channels where information is represented using diffusible molecules. However, spatial information is challenging to represent and transmit through chemical channels because diffusion promotes spatial information loss. In this work, we introduce an information theoretic framework to identify fundamental limits associated with transmitting spatial information in biological reaction-diffusion systems. More broadly, our work aims to understand how we can use molecules to extract useful representations of a physical environment. To analyze the representation and transmission of spatial information by cellular signal transduction pathways, we construct a minimal reaction-diffusion channel capable of transmitting spatial information across a cell membrane. Our channel model consists of a cell membrane and a membrane-bound, diffusible response molecule. The membrane separates the cytosol from the environment. Given inducer molecules in the environment, our physical channel allows for the spatial distribution of the inducer to be transmitted to the cytosol via the arrangement of response molecules on the inner membrane surface. From this simplified model, we obtain a channel transition function that maps the inducer distribution (input) to a set of possible response molecule arrangement on the membrane, with their associated probabilities (output). The transition function allows us to establish bounds on channel performance, and characterize trade-offs between the rate and accuracy of spatial information transfer, for different signal transduction architectures. We show that receptor clustering is important for information transmission in predictable environments, whereas uniform distributions of receptors are more beneficial in unpredictable environments. Since receptor positioning is one way to change the system's representation of the environment, our result demonstrates that different representation provides different advantages depending on the statistical structure of the environment. In conclusion, we establish a theoretical framework for studying trade-offs of spatial information transmission via a biochemical communication channel. Our results offer insights into the functions of natural biological channels such as transmembrane signal transduction in immune sensing, and the designs of novel nanosensors. In the future, we hope our work will inform the design of adaptive chemical machines, which can change autonomously to maximize the rate with which it learns about its surrounding.

Estimating Mutual Information Content of Biomedical Data Modalities through Self-Supervised Domain Translation in Prostate Cancer

Geoffrey F. Schau, Zeynep Sayar, Erik A. Burlingame, Jenny Eng,, Koei Chin, Ece Eksi, Joe W. Gray, Young Hwan Chang

OHSU

Multiplex imaging and genomic technologies independently measure highly dimensional yet non-orthogonal sets of cellular features of heterogeneous biological samples at single-cell resolution. While often complementary, the application of both imaging and genomic technologies to the same specimen is generally experimentally prohibited, as most biological profiling technologies either irreversibly alter or destroy the sample of interest during the data collection process. Throughout medicine and the biological sciences, the broad absence of paired imaging and genomic data at single-cell resolution profoundly limits our ability to understand and learn relationships between complementary biomedical domains under uncertainty of the mutual information content shared between them. The work presented herein illustrates a self-supervised deep learning-based approach that leverages unpaired data across two domains to learn cross-wise mapping while disentangling mutual information content from distinct profiling modalities, with potential applications to omics-omics and omics-imaging translation at single-cell resolution. We consider a pair of domain-specific autoencoder networks composed of encoders Φ_A and Φ_B and decoders Ψ_A and Ψ_B that independently learn latent representation encodings of data from complementary domains A and B in a joint latent space z. The self-supervised autoencoder system learns to minimize domain-specific autoencoder reconstruction loss while concurrently learning to minimize cross-domain cycle consistency loss following domain translation through the shared latent representation. The system architecture performs cross-domain translation by swapping the decoder networks during evaluation such that $f_{A\to B}(A_i) = \Psi_B(\Phi_A(A_i))$ and $f_{B\to A}(B_i) = \Psi_A(\Phi_B(B_i))$. Mutually informative data features are estimated through the introduction of a novel Programmable Weighted Gate Layer (PWGL) designed to identify mutually informative features between domains that enable cross-domain translation. Composed of a single regularized element-wise multiplication layer, a non-trainable yet programmable binary element-wise multiplication layer, and a hyperbolic tangent activation function, the PWGL identifies domain-specific information by injecting random noise sampled from the respective dataspaces into input data samples and filtering features based on the discriminatory magnitude of learned weights. Further, we introduce a Mutual Encoding Divergence (MED) learning objective that encourages the model to learn similar latent representations of singular inputs by both encoding functions through minimization of the divergence between cyclic encodings. We evaluate the ability of our approach to discern informative data features from non-informative features that facilitate domain translation on well-studied benchmark datasets

and simulated biological data. This work introduces a proof-of-concept approach to potentially uncover relationships between imaging and genomics domains and identify specific feature sets informative for domain translation. We believe this work may provide a compelling avenue towards multimodal integration of biomedical data and contribute to spatially-resolved genomics by understanding biology through learned representations of cellular state at single-cell resolution.

Systematic discovery of mechanisms underlying protein co-evolution to learn predictive representations of proteins beyond current pairwise abstractions

Ding David[1], **Debora Marks**[1], **Michael Laub**[2]

[1] Harvard University
[2] MIT

Proteins function by interacting with other biomolecules, irrespective of their function. What are the next mutational steps that an interacting pair of proteins can? This central question in evolutionary biology also underlies many therapeutic applications such as the design of biologics. Our traditional understanding of protein coevolution views interacting proteins in terms of coevolving pairs of residues which undergo a stepwise process where local and specific mutations directly at the interface compensate for each other. This aligns with recent success of using covariation models, which represent proteins as covarying pairs of residues, in allowing us to determine de novo crystal structures of proteins purely from similar sequences found in nature, or to identify specificity-determining residues in protein interactions. However, are such pairwise abstractions of proteins useful for predicting actual 2-step mutational trajectories? To this end, we have collected a reproducible and systematic dataset, consisting of 100.000s of protein variants, which captures how a mutation in one protein can affect all of the possible future mutational possibilities in its binding partner. This allows us to quantify the probability of specific, local suppressors, which can be captured by current pairwise models of protein evolution, versus long-range and potentially globally suppressing mutations, which are not captured in previous models. We find that current pairwise descriptions of protein evolution only explain a small fraction of compensatory evolution. Rather, there exist significant long-range as well as globally suppressing mutation effects. Such mutations can be found at almost any distance, and allow a protein to tolerate a diverse set of subsequent mutations that would otherwise disrupt binding to its binding partner, hence promoting their evolvability. We also find that seemingly neutral mutations in one binding partner can have restrictive long-range constraints on the possible set of mutations tolerated in its partner. These observations call for a need in alternate abstractions of coevolving proteins that go beyond pairs of residues. With this systematic dataset, we are now able to query representations of proteins which capture higher-order and long-range dependencies, as well as mutational effects on structure and dynamics of interacting proteins for understanding mutational effects. Our hope is to develop meaningful abstractions of protein biophysics for the development of coarse-grained models with predictive abilities, ultimately facilitating the design of novel proteins with therapeutic benefit.

Disentangling Experimental Noise from Fluorescent Microscopy Images with Multi Cell Type

Wei-Cheng Tseng[1,2], **Cheng-Kuan Chen**[1,3], **Tayden Jui-Ting Li**[4], **Min Sun**[1], **Hsiao-Chun Huang**[5]

[1] Equal contribution
[2] National Tsing Hua University
[3] Columbia University
[4] Stanford University
[5] National Taiwan University

Experimental noise is common in biomedical field, which is caused by many factors such as environmental factor or technology used in the experiment. We consider the experiment noise of fluorescent microscopy image taken in different batch (batch effect + plate effect) Challenge:

- Our dataset is large and diverse. contain four cells (HEPG2, HUVEC, U2OS, RPE), 51 batches in total and 1108 siRNA.
- The batch effect is serious that deeply confound the cell information (see Fig 1.) we are interested in

Results

- We test our model on the separate testing batch and achieve accuracy, which means our model can generalize well to disentangle the important cell information in the unseen batch
- We use single model trained by supervised learning without any further performance boosting tricks (e.g., ensemble, pseudo-labelling...etc)

Under Development: Craniofacial Genetics

Janina Kueper[1], **Joshua Tan**[2]

[1] Harvard University
[2] Oxford University

Previous studies of craniofacial genetics have combined genome-wide association studies and 3D morphometry in order to correlate genomic loci with adult facial phenotypes. However, these findings do not address the actual mechanics of craniofacial development and its associated pathologies, which severely restrict their clinical applications (as opposed to forensic applications, which are more straightforward prediction tasks). They also suggest a wrong view of biology: a genome does not specify a face but rather the precursors of a face along with a developmental pathway. In particular, craniofacial defects often occur not because the genome specifies "bad" phenotypes but because something goes wrong in the developmental pathway. The main problem in clinical applications is to characterize these pathological boundary cases, e.g. by seeking out genetic loci active in producing the defect. Here, we propose constructing a deep, generative representation of normal development that can account for physiological variation and then using that representation as a baseline. Drawing inspiration from image representations of the face that, conditioned on an input image, preserve visual identity under variations of pose, expression, and age, we argue that genomic representations of the face, conditioned on an input genome, should preserve developmental identity under variations of phenotype—namely, the visual parameters of the face may change but the basic mechanisms of development should not. As a first step, we propose a deep generative model that produces visual approximations of developmental milestones based on genomic inputs, using a latent vector optimization approach to preserve developmental identity. For example, given a set of gene expression data from a single patient, the model would produce a reconstruction of the patient's face at the 1-week, 3-week, 2-month, post-natal, and adult stages of development. Such a model could then be used to synthetically augment existing data sets and to gauge the possibilities of more focused genomic studies.

Graph Convolutional Networks for Epigenetic State Prediction Using Both Sequence and 3D Genome Data

Jack Lanchantin, Yanjun Qi

University of Virginia

Predicting chromatin state labels such as transcription factor (TF) binding or DNA accessibility are crucial tasks in biology due to the importance of such states in gene regulation. Computational prediction models are essential in order to label previously un-experimented locations in the genome, as to understand the underlying biology. Deep learning models have proven to be effective methods for classifying DNA segments into their respective chromatin state labels. However, current methods consider small segments of DNA independently, ignoring long range dependencies and the 3D shape of DNA which are influential in chromatin state labeling. In this work, we extend previous methods that classify segments independently by utilizing the 3D shape of DNA from Hi-C data and graph neural networks to model the interactions between segments. Our method leads to stronger classification performance, particularly for labels that have a high degree of interactions with other DNA segments, as indicated by the Hi-C graph.

Creative probabilistic programming for biology

Miriam Shiffman

MIT & Broad Institute

The "meaningfulness" of a learned representation in biology can only be measured with respect to a particular biological context or question. Modeling is the structure that provides this context and endows latent representations with meaning. Probabilistic modeling is often the most suitable choice—not only for its decision theoretic properties and coherent handling of measurement noise, but because biology itself is probabilistic. And probabilistic programming languages are one tool missing from widespread adoption in biology, with the potential to more naturally and holistically meld the modeling process with the process of wet lab science. Probabilistic programming languages (PPLs) add random variables to the long list of built-in types that we expect in a language (strings, ints, and the like). Fundamental operations in probability—like sampling, conditioning, and inference—are fundamental (automated) features of a PPL. In other words, writing down the (mathematical) model and coding up the (executable) model are near-verbatim. PPLs make Bayesian methods accessible to non-experts. Better yet, PPLs do for creativity in generative modeling what differentiable languages like TensorFlow and PyTorch have done for neural networks: promote a flowering of experimentation through assembly of complex architectures out of legolike, high-level abstractions. In short: tweak the model but not the algorithm. Inference is generally harder than backpropagation, so efficiency may suffer compared to model-specific algorithms. However, wall time is distinct from user time—and the involved process of deriving and implementing custom inference can follow the valuable experimentation phase. A recent example: dropout (observed abundance of zeros in single-cell RNA sequencing) has been explained as zero inflation since scRNA-seq's inception. Several papers this year [1,2,3] independently contradict this long-held assumption, showing that zero counts in droplet single-cell data closely mirror the expected pattern from count models alone. In a probabilistic program, the model comparison to draw this conclusion (by fitting various flavors of count models, with or without zero inflation, to scRNA-seq datasets) is as simple as changing a few words or lines of code. PPLs are useful for building and extending current workhorses in computational biology, like latent factor models and variational autoencoders. They also enable straightforward implementation of hierarchical models, reaping inferential power (and interpretability) by sharing parameters among genes in a common pathway or single cells from a common individual. Could PPLs be more intimately integrated into the wet lab process, like optimizing experimental protocols? Could probabilistic programs of biological processes be synthesized automatically from experimental data? Could useful structures like Gene Ontology and KEGG pathways be encoded as PPL primitives? Could PPL-enabled uncertainty quantification inform the next gene to perturb or tissue to sequence? A call to action for scientists at the intersection of machine learning,

language design, and biology: we need better support for discrete structures like trees, a common regime in biology. This is a hard problem since existing black-box methods like variational inference and Hamiltonian Monte Carlo require differentiability of the posterior with respect to its parameters (and so exclude uncollapsed discrete variables, like tree topologies). In tandem, to interpret the dense information contained in high-dimensional, multimodal posteriors, we need new methods for intuitive visualization of uncertainty. And until journals accept graphics with interactivity and animation, we would benefit from new publishing venues in the spirit of machine learning's distill.pub, where in-depth, manipulable graphics (often with inventive interfaces) are the centerpiece and conduit for insight. We should be exploring how experimental biology can be restructured around probabilistic modeling—as an ongoing part of data collection and experimental design, beyond post hoc analysis—and how PPLs can be extended to meet the particular challenges of biology and promote model-tinkering in new and creative ways. Bring generative models out of the silo of the lengthy appendix! REFS [1] Townes FW, Hicks SC, Aryee MJ, Irizarry RA (2019) Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. bioRxiv:574574. [2] Svensson V (2019) Droplet scRNA-seq is not zero-inflated. bioRxiv:582064. [3] Silverman JD, Roche K, Mukherjee S, David LA (2018) Naught all zeros in sequence count data are the same. bioRxiv:477794.

Genetic Interaction Maps: Using CRISPR to Characterize Apoptosis

Peter C Deweirdt, Ruth E Hanna, Mudra Hegde, Marissa N Feeley, Annabel K Sangree, Kendall R Sanson, John G Doench

Broad Institute - Genetic Perturbation Platform

Uncovering genetic interactions is an important challenge for therapeutic development. For example, in many cancers these interactions can be used to develop effective combination therapies and tailor treatment for patient populations. Furthermore, genetic interaction networks can provide important context for uncharacterized or novel gene hits. The yeast genetic interaction network took many dollars and years of effort to complete, highlighting the need for more streamlined approaches to characterize genes and pathways of interest. Here we demonstrate three approaches to study genetic interactions in apoptosis. We perform a one by all anchor screen with the anti-apoptotic gene BCL2L1 to determine its genetic interactors. We then test top hits along with known apoptotic genes in a some by some combinatorial screen. We uncover a cluster of four genes - BCL2L1, MCL1, WSB2 and MARCH5 - in which four of the six possible combinations are synthetically lethal. Using coessentiality data we note that MARCH5 and WSB2 are tightly linked with apoptosis. Together, these approaches suggest a path to accelerating the discovery genetic interactions.

Domain adaptation for spatial and dissociated gene expression data integration

Yutong Wang[1], **Joshua Welch**[2], **Laura Balzano**[1], **Clayton Scott**[1]

[1] Department of EECS, University of Michigan
[2] Department of Computational Medicine and Bioinformatics, University ofMichigan Medical School

Gene expression of cells can be measured either in its original spatial context or after dissociation without spatial information. The advantage of dissociated method is that cells can be measured at the single-cell resolution. On the other hand, spatial methods often work on a coarser level, where voxels of hundreds of cells are measured. In this work, we aim to recover the spatial information of dissociated transcriptomics data using a reference spatial dataset.

Find Archetypal Spaces Using Neural Networks

Daniel B. Burkhardt[1], **David van Dijk**[1], **Matthew Amodio**[1], **Alexander Tong**[1], **Guy Wolf**[2], **Smita Krishnaswamy**[1]

[1] Yale University
[2] Université de Montréal

Archetypal analysis (AA) decomposes each observation in a dataset into a convex combination of pure types or archetypes. These archetypes represent extreme combinations of features and thus are extrema of the data space. In biology, this might be an extreme cellular state, combination of gut microbes, or pattern of health biomarkers. This interpretation has several applications for exploratory data analysis. For example, the archetypes can be characterized in the feature space to understand the extrema of a dataset. Additionally, when considering the archetypal space, i.e. the mixture of archetypes for each data point, AA provides a new factor space for data exploration. These applications have led to the application of AA for exploratory data analysis in a number of disciplines including astronomy market research, document analysis, and genomic inference. Identifying archetypes is the primary challenge in AA. Most methods for AA identify archetypes by fitting a simplex to the data space where the vertices are linear combinations of the input data. A limitation of this approach is that if the relationships between features in the dataset are non-linear, as is often the case in biology, then the extrema of the data space may not correspond to the extrema of the data geometry. Take, for example, a triangle projected onto a sphere. Although the vertices of the triangle remain the extrema of the data geometry, they may no longer conform to extrema of the data space. In this case, linear AA methods fail to capture correct archetypes. To overcome these limitations, we introduce the Archetypal Analysis network (AAnet), a neural network framework for learning and generating from a latent archetypal space. AAnet uses an autoencoder with a novel regularization on the latent layer in which the encoder learns the transformation from the data space (input) to the archetypal space (bottleneck layer), and the decoder learns the transformation back to the feature space (reconstruction). Performing AA in this manner also provides powerful generative properties. Single activations of each node in the latent space represent an archetype of the data that the decoder transforms back to the feature space. It is also possible to generate new data with a specific mixture of each archetype. Using AAnet, we demonstrate state-of-the-art recovery of ground-truth archetypes in non-linear data domains with quantitative benchmarks to previously published AA methods. Finally, we show AAnet can generate from data geometry rather than from data density and is capable of identifying biologically relevant archetypes in single-cell RNA-sequencing data of tumor-infiltrating lymphocytes and gut microbiome profiles from thousands of individuals. Code and a tutorial for AAnet is publicly available on GitHub.

Minimal I-MAP MCMC: A Software Package for Fast, User-Friendly Bayesian Inference of Bayesian Networks

Mirae Parker, Raj Agrawal, Tamara Broderick

MIT

Our goal is for this software to be accessible and useful to people with many different research interests. These are the type of general data analysis goals we believe we can help you with:

1. You have 2-100 observed quantities (e.g. the concentration for each pro- tein in a signaling network) – and you want to understand how they affect each other.

2. You wish to be able to express how certain you are that particular rela- tionships exist between observed quantities.

3. You want to automatically generate a concise, intuitive, and interactive visualization of these relationships.

Graphical models (in particular, causal networks) allow us to represent relationship structure (goal #1 above):

- They are relatively easy to interpret and use for inference

- They are modular and can be used to represent complex systems.

To report coherent uncertainties (goal #2), we take a Bayesian approach. Bayesian methods output a distribution that ex- presses our uncertainty over graphs. That distribution may be difficult to access computationally; Markov Chain Monte Carlo (MCMC) provides a useful approximation. Minimal I-MAP MCMC is a particularly fast version of MCMC for graphs. It samples only from a small subset of highly probable graphs.

Modeling Cellular Dynamics with Continuous Normalizing Flows

Alexander Tong[1], **Jiexi Huang**[1], **Guy Wolf**[2], **David van Dijk**[1], **Smita Krishnaswamy**[1]

[1] Yale University
[2] University of Montreal

Single-cell technologies have low time resolution and are destructive. Want to model cells as dynamic continuously evolving entities over gene space. Continuous picture of transition / state change or differentiation from coarse-grained time measurements.

Understanding breast cancer heterogeneity at the multi-omics and imaging levels

Jingping Qiao, Jane Bayani, Theo Cleland, John Bartlett, Martin Yaffe, Lincoln Stein, Quaid Morris

University of Toronto

Recent advances in understanding of tumor clonal evolution and intra-tumoral heterogeneity have identified these features as important to the clinical course of cancer and may soon factor into clinical therapeutic decisions. This project aims to understand breast cancer heterogeneity at the multi-omics and imaging levels, which together will provide a deep perspective on cancer heterogeneity for diagnosis and treatment. The ultimate goal for this study is to model imaging surrogates, which could represent or predict the underlying molecular profiles, by using machine learning algorithms. Currently, we have discovered that a mixture of molecular subtypes can co-exist within the same breast cancer patient, and we also acquired insights on the evolutionary relationships among in situ/invasive carcinomas and different molecular subtypes.

scTranslate: Learning to Translate Between Epigenetic and Transcriptomic Single-Cell Assays

Benjamin Lengerich, Michael Kleyman, Andreas R. Pfenning, Eric P. Xing

CMU

Guiding Questions: Is there a latent space which summarizes both epigenetic and transcriptomic cell state? Can we use this to translate between single cell RNA-seq data and single cell ATAC-seq data? Motivation:

1. Impute missing assay from single assay data Majority of single cell data is scRNA-seq or scATAC-seq. Ability to infer the other assay would provide understanding without needing to perform expensive dual assays.

2. Cell type-specific gene regulatory mechanisms Interpretable translator could link changes in open chromatin events to gene expression events.

3. Learn a highly accurate latent space to define cell state from multiple views.

Machine Learning for Improved Directed Evolution Efficiency and Outcome

Kadina Johnston, Zachary Wu, Frances Arnold

Caltech

Enzymes inexpensively, efficiently, and selectively catalyze useful and challenging reactions under mild conditions. With an increasing drive to adopt greener practices, these biocatalysts are becoming more prevalent in industrial syntheses. Optimal enzymes are developed by searching the protein fitness landscape, the conceptual relationship between protein amino acid sequence and fitness. This landscape is beyond astronomical in size (for reference, there are 20100 unique 100-residue-long proteins and 1080 atoms in the observable universe), and so cannot be exhaustively searched. Rather than randomly searching the fitness landscape, protein engineers improve enzyme activity using directed evolution (DE). Standard DE techniques are burdened by a limited ability to screen protein variants, and often arrive at local fitness optima rather than the global. By integrating machine learning into the directed evolution workflow, we were able to both lower the number of protein variants screened and increase the probability of finding the global optimum. In the future, we will encode substrates in our workflow to specify enzyme productivity with multiple substrates.

Balanced learning of cell state representations

Erik Burlingame, Jennifer Eng, Guillaume Thibault, Geoffrey Schau, Koei Chin, Joe W. Gray, Young Hwan Chang

Oregon Health & Science University

Cell state characterization is essential to patient diagnosis and treatment and can be defined by a cell's morphology or the markers it expresses. We aim to infer cell state from morphology. High-dimensional imaging methods like cyclic multiplexed immunofluorescence (cmIF) enable unprecedented in situ cell state characterization through simultaneous labeling of tens of markers within tissues. Despite the great depth of information it can provide, cmIF is laborious, expensive, and requires specialized reagents. This motivates our search for cell state representations that express the depth of cmIF, but can be encoded by images acquired using a minimal number of low-cost and widely-available reagents like the nuclear stain DAPI. Operating under the hypothesis that morphology reflects features of cell state, here we present a deep curriculum learning framework that leverages the nuclear morphology of a cell as visualized by DAPI staining to progressively infer its state.

Hierarchical, rotation-equivariant neural networks to predict the structure of protein complexes

Stephan Eismann, Raphael Townshend, Nathaniel Thomas, Milind Jagota, Bowen Jing, Ron Dror

Stanford University

Predicting the structure of multi-protein complexes is a grand challenge in biochemistry due to its implications for basic science and drug discovery. Experimental determination of such structures tends to be hard and predicting structures of complexes computationally has proven much more difficult than predicting structures of individual proteins. To date, computational modeling methods have relied upon hand-crafted, local features to score the quality of a hypothesized protein complex structure. Here, we demonstrate that it is possible to learn end-to-end from very large molecular structures such as those of protein complexes. We present a convolutional neural network that learns directly from the 3D positions of atoms and their associated element types what makes for a favorable protein complex. As the method is point-based, it does not require a discretized grid representation of space. Our architecture is exactly rotation, translation, and permutation equivariant. This equivariance removes the need for rotational data augmentation during training. The architecture also preserves these symmetries over multiple layers, including a subsampling operation. The subsampling operation takes advantage of the fact that input and output points in rotation-equivariant convolutions do not have to be the same. We construct this operation to recognize patterns efficiently at different scales of the protein structure and to aggregate information hierarchically—for example, at the level of alpha carbons, which each correspond to one amino acid residue. The combination of built-in symmetries, hierarchical architecture, and local convolutions enables us to learn end-to-end from molecular structures containing more than 10,000 atoms. We specifically consider the problem of protein complex scoring—that is, the task of assigning a quality score to a given structural model that indicates how close it is to an experimentally determined structure. The goal is to rank accurate models at the top. The approach we present improves upon the previous state of the art in this regard. Our architecture is readily applicable to other tasks involving learning on 3D structures of large atomic systems.

Gradient Group Lasso Identifies Sparse Manifold Parameterizations

Samson Koelle, Marina Meila

University of Washington Department of Statistics

A major challenge at the intersection of medicine and machine learning is to leverage the increasing amount of available and parsable data to make clinically-relevant predictions and decisions. Meanwhile, the clinical use of increasingly complex phenotypic assays is further complicating this already difficult problem. This abundance of data and its complex structure has motivated the use of 'black box' machine learning methods such as neural networks. These methods often show optimum empirical results, at the cost of model interpretability. The lack of interpretability of the representations learned by these methods impairs their use. Clinicians and patients want interpretable justifications for health-care decisions, and interpretable justifications are more reliable, since, for example, a sensible justification indicates that a trained model has not overfit, and gives confidence in its generalizability. Efforts to interpret a model are hampered by the non-linearity of the learned feature map, and potential non-trivial data geometry and topology in the latent space. A unifying idea underlying many methods that add interpretability to black-box models is that the learned deformation is smooth, and therefore can be considered on a differential level. For example, the saliency map method leverages this approach by interrogating how individual features impact the latent space representation. However, this is not easily interpretable, since there is no restriction on multifactorality of the gradient of the latent space representation with respect to the input features. Such approaches may also fail due to problems such as vanishing gradients and unsuitable priors in the latent space. Our recent work addresses these problems with Diffusion Map embeddings of molecular dynamics data. We estimate gradients of latent space representations of high-dimensional molecular motion, and compare these gradients with the gradients of a user-defined dictionary of explanatory smooth functions. In our case, these are bond torsions, but generally they can be any smooth function of the features. We employ a novel normalization and group-lasso optimization strategy that encourages selection of a sparse set of functions that parameterize the data manifold: that is, whose gradient bundles form a basis for the tangent bundle of the denoised data manifold, which is estimated using the gradients of the latent space representation. The group lasso approach allows flexibility in whether an interpretation needs to be valid across the whole manifold, or just in some region, and comes with statistical guarantees. Importantly, it will tend to favor covarying dictionary elements that are higher-level descriptors of a data generating process. It would select an appropriate gene set over its subset genes to describe a cellular differentiation trajectory. This method can be seen as finding an interpretable approximation to the saliency map of the latent space. My personal interest in this workshop comes from my background in hematopoietic stem cell biology, in particular transgene barcode labelling. I am interested in extending my current work back to my original field of health sciences. I

am also interested in data analysis for emerging assays that are enabling association of barcode labelling and other clonal information such as TCR with single-cell phenotype.

Learning Sparsely-Coupled Signaling Networks from Data

Matthew Karikomi, Qing Nie

UC Irvine

Populations of cells rely on intercellular communication networks for fate specification, giving rise to spatial patterning, boundaries and other multiscale properties. Juxtacrine communication between two cells requires physical contact. Within a contiguous population, juxtacrine network states arise when pairs of adjacent cells express complementary surface markers. The expression of target genes in the receiving cell provides a noisy signal whose dynamics are observable by several current measurement methods. Such methods always entail a tradeoff between bandwidth and temporal resolution: Destructive methods provide a comprehensive snapshot of cellular state, while real-time imaging provides longitudinal measurements of several state variables (messenger RNA and protein) over the course of hours or days. While comprehensive state measurements are important for exploratory analysis, the lack of longitudinal and spatial information hinders dynamic inference of relevant pathways. We investigate the intersection between phenomenological modeling, high-resolution live-cell imaging and single-cell transcriptomics in a two-step approach: First, we apply sparse, nonlinear variable-selection to simulated populations and [position-masked] live-cell data, to search for the spatial neighborhood each cell. These distributed networks define a Markov-equivalent set of possible global topologies. These global topologies demonstrate the phenomenological relationship between cell-communication and tissue-level pattern formation. Second, we explore spatial inference in high-dimensional snapshot data from single-cell mRNA sequencing experiments. It has been previously shown that cell fate is controlled by spatial dynamics which in turn arise from intercellular communication and cell-autonomous regulation. Spatial dynamics are encoded in multicellular states such as the phase gradient of oscillatory Hes1 expression, which is necessary for the formation of bilateral symmetry in vertebrates. Experiments have shown that the instantaneous phase of known oscillators predict cell location based on this gradient. We therefore explore spatial inference of single cells based on transcriptomic snapshot data as a clustering technique, where cells are clustered based on the instantaneous phase of selected oscillatory markers. Instantaneous phase of an oscillator is defined at the single-cell level by mRNA sequencing. Our method of unsupervised clustering uses phase-neighborhood symmetry to constrain the space of possible network topologies for these cells, since juxtacrine networks arise from the locally asymmetric expression of pathway genes. This work adopts a two-pronged approach to the analysis of intercellular networks. First we utilize experimental measurements of single-cell dynamics to discover population dynamics. Second, we apply dynamic models of cell populations to snapshot single-cell omics data to explore ways in which spatially-constrained regulatory interactions are driven by intercellular communication.

Explainable Substructure Partition Fingerprint for Protein, Drug, and More

Kexin Huang[1], **Cao Xiao**[2], **Lucas Glass**[2], **Jimeng Sun**[3]

[1] Harvard University
[2] IQVIA
[3] Georgia Institute of Technology

What is Explainability? A tractable path from meaningful sub-structures to predictive outcome. Objectives: A moderate-sized discrete partition where each partition is meaningful sub-structures.

Employing Geometry to rescue damaged networks

Guruprasad Raghavan, Matt Thomson

Biology and Biological Engineering,Caltech

Questions: 1. How resilient are state-of-the-art Artificial systems to physical damage? 2. Can we identify efficient repair strategies to rescue damaged artificial systems (ANN's).

Predicting Drug Protein Interaction using Quasi-Visual Question Answering System

Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, Yuedong Yang

Sun Yat-sen University

Identifying novel drug-protein interactions (DPI) is crucial for drug discovery. Many machine learning-based methods have been developed based on drug descriptors and one-dimensional protein sequences. However, protein sequence cannot accurately reflect the interactions in three-dimensional space. On the other hand, a direct input of 3D structure is of low efficiency due to the sparse 3D matrix, and is also prevented by limited number of co-crystal structures available for training. Thus, we propose an end-to-end deep learning framework to predict the interactions by representing proteins with 2D distance map from monomer structures (Image), and drugs with molecular linear notation (String), following the Visual Question Answering mode.

Pre-Training of Deep Bidirectional Protein Sequence Representations with Structural Information

Seonwoo Min, Seunghyun Park[2], Siwon Kim[1], Hyun-Soo Choi[1], Sungroh Yoon[1]

[1] Seoul National University
[2] Clova AI Research

Motivation: A structure of a protein has a direct impact on its properties and functions. However, identification of structural similarity directly from amino acid sequences remains as one of the most challenging problems in computational biology. Instead of a raw amino acid sequence, we need a better representation of a protein to infer and compare structural information for various downstream tasks. Results: This paper presents a novel BERT-wise pre-training scheme for a protein sequence representation model that incorporates:

- PLUS - Protein representations Learned Using Structural information
- Bidirectional RNN model with two BERT-wise pre-training objectives (1) Masked Language Modeling (MLM) (2) Same Family Prediction (SFP)
- One of the first works to use the BERT-wise pre-training on a biological sequence as well as any non-NLP (Natural Language Processing) tasks
- Advances SOTA for protein structural similarity prediction

Enabling naturalistic systems neuroscience: Interpretable representations for analysis of long-term human neural activity and behavior

Satpreet H. Singh, Steven M. Peterson, Rajesh P. N. Rao, Bingni W. Brunton

University of Washington

Much of our understanding in human neuroscience has been informed by data collected in predesigned and well-controlled experimental tasks, where timings of cues, stimuli, and behavioral responses are known precisely. Recent advances in data acquisition and machine learning have enabled us to study longer and increasingly naturalistic brain recordings, where we try to understand neural computations associated with spontaneous behaviors. Analyzing such unstructured, long-term, and multi-modal data remains a substantial challenge, due in part to the lack of a priori experimental design and the difficulty of isolating interpretable behavioral events.

Dirichlet Process Mixture Models for single-cell RNA sequencing

Shimbi Masengo Wa Umba Papa Levi, Adnan Abu-Mafhouz, TD Ramotsoela, GP Hancke

University of Pretoria

Encoding Evolutionary Information through Multi-Task, Alignment Free Protein Prediction

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Burkhard Rost

Technical University of Munich

Correctly predicting features of protein structure and function from amino acid sequence alone remains a supreme challenge for computational biology. For almost three decades, state-of-the-art approaches combined machine learning and evolutionary information from multiple sequence alignments. Here, we introduced a novel approach fusing self-supervised language modeling on an unlabeled big dataset with supervised training on labeled high-quality data in one single end-to-end network. These new results push the boundaries of predictability towards grayer and darker areas of the protein space, allowing to make reliable predictions for proteins which were not accessible by previous methods.

Language modelling for biological sequences - curated datasets and baselines

Jose Juan Almagro Armenteros[1], Alexander Rosenberg Johansen[1], Ole Winther[2], Henrik Nielsen[1]

Technical University of DenmarkAFfUniversity of Copenhagen

In computer science, Natural Language Processing (NLP) is the field that studies how computers are able to understand and process human language. These methods can be also applied in bioinformatics to better understand the language of proteins. We propose a new dataset and benchmarks for language modelling in proteins.

Exploring Data Through Archetypal Representatives

Sebastian Mathias Keller, Fabricio Arend Torres, Maxim Samarin, Mario Wieser, Volker Roth

University of Basel

Archetypal Analysis represents each individual in a data set as a mixture of extreme types or archetypes and at the same time constrains these archetypes to be mixtures of the individuals in the data set. In the original publication a head shape data set was analyzed with the goal to identify archetypal head shapes. Archetypal Analysis (AA) is a form of non-negative matrix factorization where a matrix $X \in \mathbb{R}^{n \times p}$ of n data vectors is approximated as $X \approx ABX = AZ$ with $A \in \mathbb{R}^{n \times k}$, $B \in \mathbb{R}^{k \times n}$. The matrix $Z \in \mathbb{R}^{k \times p}$ contains the k archetypes $\mathbf{z}_1, .., \mathbf{z}_j, .., \mathbf{z}_k$ while A and B are weight matrices with positivity constraints and rows, resp. columns, summing up to one. Conceptually, AA is interesting as it naturally welcomes interpretations with an evolutionary flavour: Extreme types are individuals that have adapted optimally to a given task and now occupy an evolutionary niche which, in an appropriate representation, can be identified with the vertices of the data convex hull. With less emphasis on an interpretation as an evolutionary process in the Darwinian sense, the goal of AA is to describe structured objects as superpositions of purer or in some sense simpler objects. We proposed Deep Archetypal Analysis as an extension which, unlike the original model, is able make use of side information when learning an appropriate representation: It is, for example, very likely that a species optimal at hunting a certain type of prey cease to be optimal at that task if it is changed, e.g. exchanging the task 'primary food source' for the task 'longevity'. In general, a representation is dependent on the task, i.e. the side information provided. The second extension we proposed is to learn archetypes in the latent space of a variational autoencoder (VAE). The benefit is that the VAE learns a linear representation of the data such that linear AA can be used even if non-linear relations between the different dimensions of the data exist. The potential of DeepAA was demonstrated on the QM9 data set, which contains the chemical structures and properties of 134 kilo molecules: Using the heat capacity of each molecule as the side information, an appropriate representation was learnt and archetypes were identified. In summary, DeepAA finds potentially meaningful representations in biomedical data sets in phenotype space or in chemical space. In general, if an appropriate encoding for a data set is available such that a VAE can be trained, deepAA can become a useful exploratory tool and a method helpful in the formation of initial hypotheses.

Learning metabolic representations of health and disease in multi-omic datasets

Noam Bar[1], **Hagai Rossman**[1], **Amir Gavrieli**[1], **Uri Shaham**[2], **Eran Segal**[1]

[1] Weizmann Institute of Science
[2] Yale University

- The human serum metabolome contains a rich set of biomarkers and causative agents, and is a known mediator for progression of metabolic diseases.

- High quality serum metabolomics is still absent from large scale cohorts, and is seldom coupled with other valuable omics data, limiting current potential to provide biological inference.

- We measure the levels of 1251 circulating metabolites in serum samples from a healthy cohort of 10K individuals at baseline, for whom we also obtained a comprehensive profiling consisting of host genetics, gut microbiome, clinical parameters, diet, lifestyle, two weeks of continuous glucose monitoring, anthropometric measurements and medication data, resulting in a large and deeply phenotyped cohort.

- We aim to leverage our unique cohort along with UKBioBank data (N=500K), in order to learn metabolic representations of medical conditions. This can be achieved by devising a joint learning procedure, where a disease risk model is trained under structural constraints, forcing the model to learn a metabolic representation of the input features by introducing interchangeable learning batches from the two datasets.

- Inspired by recently introduced semi-supervised iterative learning methods (Xie 2019) we aim to exploit the large unlabeled (labels=Metabolites) UKBiobank dataset for improving the prediction of blood metabolites from diet and baseline covariates on our smaller labeled dataset.

Emap2sec: Protein Secondary Structure Detection in Intermediate Resolution Cryo-EM Maps Using Deep Learning

Sai Raghavendra Maddhuri Venkata Subramaniya[1], **Genki Terashi**[1], **Daisuke Kihara**[1,2]

[1] Department of Computer Science, Purdue University
[2] Department of Biological Sciences, Purdue University.

- With the advancements of cryo-Electron Microscopy (EM) in the field of structural biology, determination and validation of 3D structures of macro biological molecules such as proteins has improved over the past few years.

- For EM maps of medium range resolution (6 – 10 Å), extracting structure information from a map and building a structure model is a challenge.

- We aim to make use of deep learning to accurately identify protein secondary structures such as α helix, β sheet, coil/turn from maps of medium range resolution.

- The deep learning architecture makes classification predictions, voxel-wise, on 3D EM maps.

A Compressed Sensing Framework for Efficient Dissection of Neural Networks

Abdullah Yonar, Jeffrey B. Lee, Timothy Hallacy, Hannah Shen, Josselin Milloz, Askin Kocabas, Sharad Ramanathan

Harvard University

An important question in neuroscience is how neural circuits modulate behaviors. One approach to addressing this question is to first identify the neurons essential for a behavior and then determine how activity patterns in these essential neurons drive that behavior. Conventional methods for identification the small fraction of neurons in neural circuits that are essential for specific behaviors require searching through each of the subtypes of neurons in the neural circuit, one at a time. Genetic and molecular approaches to implement such a search is challenging due to the lack of unique markers, promoters or specific mutations that affect only one subtype of neurons. Here we argue that the lack of specific genetic tools to address individual neuronal subtypes, which are a challenge for conventional approaches, can in fact be exploited using a compressed sensing framework to find the key neuronal types controlling a behavior much faster. We developed and tested a compressed sensing framework on the small nervous system of C. elegans by identifying the neurons that control its speed of locomotion. Then, we validated our findings and investigated the role of the key neurons controlling speed with a novel stabilization microscope that we developed.

Inferring cell-penetrating peptide design principles

Somesh Mohapatra, Rafael Gómez-Bombarelli

MIT

Background and challenges

- Artificial peptides (in this case, cell-penetrating peptides) used in various medical and chemical applications1
- Combinatorial search-space of possible sequences scaling as nN (n positions, N monomers), order of 10150
- Complex sequence-biological activity relationships
- Representation of monomer/sequence with adequate physio-chemical parameters

Approach

- Synthesis and characterization of library of CPP-drug conjugates, under standard conditions
- Development of graph representation for peptides
- Machine learning to model sequence - activity relationships
- Experimental validation of predicted sequences
- Interpretation of ML Model for peptide design principles

Scalable and accurate optimization methods for maximum pseudo-likelihood estimation of Markov networks

Hannes Harbrecht[1], **José Miguel Hernández-Lobato**[2]

[1] University of Cambridge and MRC Laboratory of Molecular Biology
[2] University of Cambridge

- Markov networks allow us to represent the conditional independence structure of both discrete- and continuous-valued variables.
- Allow causal interpretation of fully-observed systems due to confounding removal.
- Applications are omnipresent in computational biology.

Tracking cell transitions in scRNA-Seq time series data

Aly O. Abdelkareem, **A. Sorana Morrissy**

Department of Biochemistry and Molecular Biology, Cumming School of Medicine, University of Calgary

Single-cell RNA sequencing (scRNA-seq) technologies have increased our knowledge of the molecular landscape of brain tumors as they allow researchers to study the variability between co-existing individual cell transcriptomes. Toward understanding single-cell data, many computational tools are developed to study cell dynamics by grouping them according to their molecular capabilities (e.g. clustering and pseudo-time trajectory building). Various methods are used to reconstruct developmental trajectories by applying dimensionality reduction or probabilistic branching modeling. In our work with scRNA-seq data from 40,000 normal mouse brain cells sampled at 12 stages throughout embryonic development, we have observed that pseudotime trajectories generated at each independent time-point cannot be accurately matched up with subsequent time-points and in many cases are incorrect. For instance, without considering time-point information, two groups of cells that remain transcriptionally distinct through all time-points are consistently placed on the same pseudotime trajectory by standard methods, incorrectly implying that they represent two endpoints of a biological process. However, there is currently a lack of tools in the field that study cell changes in time. Our goal is to explore the relationships of cells within and among time-points. In this work, we aim to understand the cellular origins, cell fates, and differentiation contributing to the development of the murine cerebellum by tracking the cell transcriptions through several time points captured with single-cell resolution. Additionally, this will lead to infers branching trajectories and assign cells to right paths.

Industry-scale Application and Evaluation of Deep Learning for Drug Target Prediction

Noé Sturm[1], Andreas Mayr[2], Thanh Le Van[3], Vladimir Chupakhin[3], Hugo Ceulemans[3], Jörg Wegner[3], Jose-Felipe Golib-Dzib[4], Nina Jeliazkova[5], Yves Vandriessche[6], Stanislav Bohm VSB[7], Vojtech Cima VSB[7], Jan Martinovic VSB[7], Nigel Greene[1], Tom Vander Aa[8], Thomas J. Ashby[8], Sepp Hochreiter[2], Ola Engkvist[1], Günter Klambauer[2], Hongming Chen[1]

[1] AstraZeneca
[2] Johannes Kepler University Linz
[3] Janssen Pharmaceutical
[4] Janssen Cilag SA
[5] Ideaconsult Ltd.
[6] Intel Corporation
[7] Technical University of Ostrava
[8] Imec

Artificial intelligence (AI) is undergoing a revolution thanks to the breakthroughs of machine learning algorithms in computer vision, speech recognition, natural language processing and generative modelling. Recent works on publicly available pharmaceutical data showed that AI methods are highly promising for Drug Target prediction. However the quality of public data might be different than that of industry data due to different labs reporting measurements, different measurement techniques, fewer samples and less diverse and specialized assays. As part of a European funded project, that brought together expertise from pharmaceutical industry, machine learning, and high-performance computing, we investigated how well machine learning models obtained from public data can be transferred to internal pharmaceutical industry data. Our results show that machine learning models trained on public data can indeed maintain their predictive power to a large degree when applied to industry data. Moreover, we observed that deep learning derived machine learning models outperformed comparable models, which were trained by other machine learning algorithms, when applied to internal pharmaceutical company datasets. To our knowledge, this is the first large-scale study evaluating the potential of machine learning and especially deep learning directly at the level of industry-scale settings and moreover investigating the transferability of publicly learned target prediction models towards industrial bioactivity prediction pipelines.

Supervised learning of transcriptional regulatory networks

Emanuel Flores Bautista[1], **Ernesto Pérez-Rueda**[2]

[1] Caltech
[2] UNAM

Transcriptional regulatory networks (TRN) orchestrate gene expression programs in response to environmental stimuli. Despite decades of experimental and theoretical studies of the TRN of E. coli K-12, we still lack knowledge of the complete set of interactions between transcription factors (TF) to their promoters. Specifically, close to one-third of the transcription factors in E. coli are not functionally annotated in knowledge bases like RegulonDB. In this work, we used a neural network model to predict the most likely regulatory program to which each of the unannotated TFs in E. coli belongs. In order to do so, we first define the regulatory modules using a community detection algorithm on the TRN. Using a defined set of modules, we then apply a principal component analysis (PCA) to reduce the variability of a global transcriptomics dataset. We then connect the output of the PCA-reconstructed dataset to a multilabel classifier. We validate our model against random forest, k-Nearest Neighbors and multilayer perceptron classifiers and show that a deep neural net architecture gives better test accuracies. We then apply our pipeline to to other organisms and show that our model is robust to different datasets. Future work could focus on expanding this approach to predict the functional association of different molecules like signaling proteins in higher organisms. Finally, our module predictions could guide the perturbations to confirm our results experimentally.

What About Higher-Order Cellular Complexity? An Inquiry with Topological Simplicial Analysis

Baihan Lin, Raul Rabadan, Nikolaus Kriegeskorte

Columbia University

The lack of a formal link between cell-cell cohabitation and its emergent dynamics into cliques during development has hampered our understanding of how cell populations proliferate, differentiate, and compete, i.e. the cell ecology. With the advancement of single-cell RNA-sequencing (RNA-seq), we have now come closer to describing such a link by taking cell-specific transcriptional programs into account, constructing graphs of a network that reflect the similarity of gene expression, and analyzing these graphs using algebraic topology. We proposed single-cell topological simplicial analysis (scTSA). Applying this approach to single-cell gene expression profiles from local networks of cells in different developmental stages with different outcomes revealed a previously unseen topology of cellular ecology. These networks contain an abundance of cliques of single-cell profiles bound into cavities that guide the emergence of more complicated habitation forms. We visualize these ecological patterns with topological simplicial architectures of these networks, compared with the null models. Benchmarked on single-cell RNA-seq of zebrafish embryogenesis over 25 cell types and 12 time steps, our approach highlights the gastrulation as the most critical stage, consistent with consensus in developmental biology. As a nonlinear, model-independent, and unsupervised framework, our approach can also be applied to tracing multi-scale cell lineage, identifying critical stages, or creating pseudo-time series.

HiCSR: A Hi-C Super-Resolution Framework for Creating Highly Realistic Contact Maps

Michael C. Dimmick[1,2], **Leo J. Lee**[1,2], **Brendan J. Frey**[1,2]

[1] University of Toronto
[2] Vector Institute

The Hi-C method produces a heat map contact matrix where each pixel represents the interaction frequency between two different regions of the genome several kilobases wide [1]. The resolution of the contact matrix is dependent on the number of sequencing read during the sequencing process. In general, a linear increase in resolution a quadratic increase in sequencing reads is required [2]. Deep learning methods provide a way to artificially increase the number of sequencing reads and thereby increase the resolution at which the contact matrix can be generated. We propose a novel Hi-C Super-Resolution (HiCSR) framework capable of accurately recovering the fine details found in high resolution Hi-C contact maps. HiCSR optimizes both an adversarial loss [3] and feature reconstruction loss [4] obtained from the latent representation of a denoising autoencoder trained to reconstruct high resolution Hi-C data. HiCSR was able to produce visually convincing and highly accurate Hi-C matrix enhancements given Hi-C data with 16 times fewer aligned reads.

Disentangling unwanted sources of variation in single-cell RNA-sequencing data under weak supervision

Hui Ting Grace Yeo, **David K Gifford**

Massachusetts Institute of Technology, Computational and Systems Biology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA

Motivation: Removal of unwanted sources of variation presents a major challenge for the analysis of large multiplexed perturbational single-cell RNA sequencing (scRNA-seq) studies Challenge: Removal of these nuisance factors typically requires expert knowledge to identify the factors and tedious curation of factor associated gene sets

Exchangeable Variational Autoencoders for Genomic Data

Jeffrey Chan, Jeffrey P Spence, Yun S Song

UC Berkeley

Define data with exchangeable-structured datapoints as data with each datapoint x(i) whose features are invariant to permutation. Such data is ubiquitous in biology:

- Cryo-EM
- Technical Replicates
- Bootstrap or Posterior Samples
- Population Data

Learning representations of biology with small and homogeneous training datasets

Alex Lu, Alan Moses

University of Toronto

Motivation: Can we still learn generalizable representations even when our training data is small or homogeneous? We compared representations learned by convolutional neural networks trained using supervised classification versus self-supervised inpainting. For the task of classifying protein localization in yeast cells, we trained models using datasets with cells from one (or few) proteins per class, and measured how well learned representations could classify a test dataset with cells from hundreds of proteins per class.

Deep Generative Models for Single-cell Transcriptomics

Romain Lopez[1], **Jeff Regier**[2], **Michael I. Jordan**[1], **Nir Yosef**[1]

[1] UC Berkeley
[2] University of Michigan, Ann Arbor

scVI is a deep generative model with suitable noise assumptions and taking into account technical variations. For each cell $n = 1, ..., N$ and gene $g = 1, ..., G$, the expression entry xng is drawn independently through the following process:

$$z_n \sim N(0, I) l_n \sim LogN(l_\mu, l_{\sigma^2}) \rho_{ng} = f_w(z_n, s_n) \pi_{ng} = f_h(z_n, s_n) x_n g \sim \mathbf{ZINB}(l_n \rho_{ng}, \theta_g, \pi_{ng}) \tag{0.1}$$

Integrating Markov processes with structural causal modeling enables counterfactual inference in complex systems

Robert Osazuwa Ness, Kaushal Paneri, Olga Vitek

Northeastern University

Modeling causal relationships between components of dynamic systems helps predict the outcomes of interventions on the system. Upon an intervention, many systems reach a new equilibrium state. Once the equilibrium is observed, counterfactual inference predicts ways in which the equilibrium would have differed under another intervention. Counterfactual inference is key for optimal selection of interventions that yield the desired equilibrium state. Moreover, counterfactual inference provides robustness of causal effect under model misspecification, by making use of past interventional or observational data to condition the misspecified model. In systems biology, complex dynamic systems are often described with mechanistic models, expressed as systems of ordinary or stochastic differential equations. Mechanistic models of stoichiometric protein signaling networks describe specific biochemical events, such as phosphorylation and degradation. Unfortunately, mechanistic models have no prescribed formula for counterfactual inference. Having observed equilibrium of a system trajectory in time, it is not clear how to re-play that trajectory with all things equal except those affected by the intervention. An alternative representation relies on structural causal models (SCMs). SCMs can represent the system at equilibrium, only require equilibrium data for parameter estimation, and support counterfactual inference. Unfortunately, multiple SCMs can represent the same observational or interventional distributions but provide different counterfactual insights. This drawback limits their practical use. We contribute a general and practical framework for casting a Markov process model of a system at equilibrium as an SCM, thus leveraging the benefits of both approaches. The SCMs are defined in terms of the parameters and the equilibrium dynamics of the Markov process models, and counterfactual inference flows from these settings. The framework alleviates the identifiability drawback of the SCMs, in that the counterfactual inference is consistent with the counterfactual trajectories simulated from the Markov process model. Moreover, counterfactual inference from the derived SCMs is robust to model misspecification. We showcase the benefits of this framework in studies of complex biomolecular systems with nonlinear dynamics, namely the Mitogen-Activated Protein Kinase (MAPK) cascade and the insulin-like growth factor (EGF) signaling system. Importantly, we demonstrate that, in the presence of model misspecification, the proposed approach estimates the outcome of an intervention more accurately than a direct simulation. Note: this work is an extended version of a manuscript that will appear in Proceedings of NeurIPS 2019.

Representation of peptide-MHC complexes for optimized peptide vaccines

Ge Liu[1], **Haoyang Zeng**[1,2], **Siddhartha Jain**[1], **Brandon Carter**[1], **Brooke Huisman**[1], **Michael Birnbaum**[1], **David Gifford**[1]

[1] MIT
[2] Insitro

We discuss the representations of peptides and their MHC presentation in the context of novel peptide vaccine therapeutics for cancer. To formulate peptide vaccines the accurate prediction of which non-self peptides will be presented by tumor cells is essential. We find that including uncertainty metrics as one output of machine learning peptide presentation models permits the computation of presentation likelihoods that result in the superior formulation of peptide vaccines. We further extend the presentation problem to the optimization of peptide anchor residues to improve vaccine efficacy. These machine learning representation and optimization problems are motivated by the important role that the human adaptive immune system plays in cancer therapy. The adaptive immune system is activated by T-cell receptors that recognize the presentation of a non-self peptide by an MHC molecule as part of a peptide-MHC complex (pMHC) complex. There has been considerable interest in engineering peptides that can bind as strongly to the MHC as possible without affecting the pMHC binding to the TCR. The key element to engineering peptides is to have a highly accurate predictor of the peptide-MHC binding affinity. We describe a peptide-MHC binding prediction model (PUFFIN) to score a peptide's affinity to MHC II. Whereas past models have used very shallow neural networks to predict the binding, PUFFIN uses an ensemble of deep residual networks on the peptide and MHC amino acid sequences and outputs the predictive distribution of the affinity of the peptide to the MHC. Each amino acid is encoded as the concatenation of the one-hot encoding and the BLOSUM50 vector corresponding to that amino acid. The full MHC sequence encoding is then concatenated to each element of the peptide sequence encoding to form the input to the residual block. The model achieves state of the art results on MHC II binding prediction on a variety of metrics. We then use the predictions of the model as input to 3 acquisition functions taken from Bayesian optimization. We search for peptides that maximize the acquisition functions and get experimental affinity data for them. To make the search for peptides more efficient, we exploit the fact that while peptide length can vary, there is usually a 9 amino acid "core" that binds to the peptide pocket of the MHC. In particular, one can identify a group of 4 anchor residues (positions 1, 2, 4, 9) that is most responsible for the binding of the peptide to the MHC.

Deep Protein-Ligand Binding Prediction with Unsupervised-Learned Embeddings

Paul Kim[1], **Robin Winter**[1,2], **Djork-Arné Clevert**[1]

[1] Bayer AG
[2] FU Berlin

In-silico protein-ligand binding prediction has been an ongoing area of research in computational chemistry and drug discovery, as it can accelerate the identification of novel pharmaceutical treatments. Recently, machine learning and deep learning approaches to this problem have become more feasible with the increasing amounts of publicly available bioactivity data. Most of this work focuses on learning to predict the binding affinity of ligands (i.e. small molecules) towards a certain biological target (i.e. a protein), using the ligand's structural similarity to ligands in a training set of compounds that have been experimentally measured on this target. Although these so called quantitative-structure-activity-relationship (QSAR) models have demonstrated success in predicting binding affinities in some settings, these approaches fail if the amount of measured molecules for a protein of interest is scarce, or if the compounds that have been measured are too structurally dissimilar from the query molecules. An attractive solution to this problem is to also include information from the biological domain in a so called proteochemometric (PCM) model. In contrast to a QSAR model, a PCM model does not only model structural similarity in the ligand domain but also in the target domain. Thus, in the case of a protein target with no or very few experimental measurements available, a PCM model could incorporate additional bioactivity information from similar proteins, enabling accurate predictions even in settings where a regular QSAR model would fail. The key to a successful PCM models lies in the way ligands and targets are represented, enabling a machine learning model to recover similarities of the biochemical mechanisms. Most previous work relies on predefined feature extraction methods, which encode a small molecule or protein according to a handcrafted protocol, e.g. by counting substructures in molecules or amino acids in proteins. However, recent advances in the field of deep neural networks have shown that it can be beneficial to use a representation of the data generated via unsupervised learning on more basic input features such as raw pixels in images or characters in natural language processing. These embeddings seek to encode the high-dimensional, often discrete raw inputs into continuous, lower-dimensional vectors, while capturing and extracting relevant features more effectively than sophisticated, human-engineered representations. Following this reasoning, we propose two methods to learn meaningful unsupervised representations from large datasets of small molecules and proteins, respectively. We train recurrent autoencoders on SMILES string representations of compounds and single-letter-code string representations of protein amino acid sequences, using techniques based on sequence-to-sequence neural machine translation models. Furthermore, we show that these representations can be combined in a PCM

model to model the interaction (i.e. binding activity) between ligands and targets in silico. Using both public and large in-house benchmark datasets, we compare our models using these unsupervised descriptors against baseline models that use state-of-the-art handcrafted descriptors. We find models based on our proposed unsupervised descriptors exhibit superior performance, which implies that they represent chemical and biological matter in a more meaningful way.

Machine learning optimization of MHC class II presented peptides

Haoyang Zeng, Brandon Carter, Siddhartha Jain, Brooke Huisman, Michael Birnbaum, David Gifford

Massachusetts Institute of Technology

Evaluating Protein Transfer Learning with TAPE

Neil Thomas, Nick Bhattacharya, Roshan Rao

UC Berkeley

Machine learning applied to protein sequences is an increasingly popular area of research in the ML community. Semi-supervised learning for proteins has emerged as a promising paradigm due to the high cost of generating interesting labels for proteins and the availability of large databases of unlabelled protein sequences. Unfortunately, the current literature is fragmented when it comes to datasets and standardized evaluation techniques. To facilitate progress in this field, we introduce the Tasks Assessing Protein Embeddings (TAPE), a set of five biologically relevant semi-supervised learning tasks spread across different domains of protein biology. We curate tasks into specific training, validation, and test splits to ensure that each task tests biologically relevant generalization that transfers to real-life scenarios. We benchmark a range of approaches to semi-supervised protein representation learning, which span recent work as well as standard sequence learning techniques based on alignment or simple one-hot encodings. We find that self-supervised pretraining is helpful for almost all models on all tasks, more than doubling performance in some cases. Of the five models we pretrained, no one model clearly stands out from the rest. This is perhaps surprising because one model we used (the Transformer) has seen pretty uniform success recently in NLP applications of a similar flavor. Despite the overall increase in performance due to pretraining, in several cases features learned by self-supervised pretraining still lag behind features extracted by PSI-BLAST or HMMs by a considerable margin. These results suggest a huge opportunity for innovative architecture design and improved modeling paradigms that better capture the signal in databases of biological sequences. Further, we find that the types of performance increased by pretraining are not always the most interesting for downstream applications. This highlights a key need for more discussion about which properties of embeddings are most relevant for downstream use and benchmarks that clearly evaluate these metrics. Our goal is for TAPE to help the machine learning community focus on scientifically relevant problems; accordingly, all data and code used to run these experiments are available. We plan to continue expanding on TAPE by improving datasets, evaluation metrics, and baselines; LMRL is an excellent venue for feedback and new direction. (NOTE: TAPE is published in the main NeurIPS2019 conference.)

Latent State Modeling of Electrocardiograms Prior to Cardiac Arrest

Christopher Aicher, Jacob E. Sunshine, Emily B. F IS

University of Washington

Goal: Using long recordings of electrocardiogram (ECG) waveforms, train latent variable models that categorize common trends of heart rhythm dynamics prior to onset of cardiac arrest. Motivating Questions:

- What are the patterns of dynamics in the morphology and frequency of cardiac rhythms that precede cardiac arrest?
- Are these patterns generalizable? Can they be shared across different ECG leads and across different patients?

Deep Multiple Instance Learning for Taxonomic Classification of Metagenomic Read Sets

Andreas Georgiou, Vincent Fortuin, Harun Mustafa, Gunnar Rätsch

ETH Zurich

- DNA reads from closely related species in metagenomic read sets tend to occur together.
- Current methods for taxonomic classification rely on classifying each individual read in a read set ignoring this co-occurrence pattern [1, 2].
- We combine permutation invariant multiple instance learning (MIL) pooling layers with existing models to directly predict the distribution over the taxa of whole metagenomic read sets.

Accelerating Protein Design Using Autoregressive Generative Models of Biological Sequences

Adam Riesselman[1], Jung-Eun Shin[2], Aaron Kollasch[2], Conor McMahon[2], Elana Simon[3], Chris Sander[4], Aashish Manglik[5], Andrew Kruse[2], Debora Marks[2]

[1] insitro
[2] Harvard University
[3] Reverie Labs
[4] Dana Farber Cancer Institute
[5] UCSF

Generative models of biological sequences can find organization and structure, predict the effects of mutations, and generate novel sequences in an unsupervised manner. Deep autoregressive generative models are predictive of mutation effects, including insertions and deletions. We apply these models to single domain antibodies, or nanobodies, by fitting to a naïve llama immune repertoire and generate a diverse, optimized synthetic nanobody library.

Systematically learning cellular programs from single-cell transcriptional and chromatin accessibility data

Anika Gupta, Layla Siraj, Thouis R. Jones, Alex Bloemendal, Vidya Subramanian, Eric S. Lander

Broad Institute of MIT and Harvard

Efforts such as the Human Genome Project and the ongoing Human Cell Atlas are catalyzing our understanding of the genomic underpinnings of biology and disease, down to single-cell resolution. An important goal ahead is a comprehensive catalog of all distinct biological pathways that govern cell identity and function, across all cell types. Extracting meaningful signal about these fundamental biological modules in a standardized and unbiased way from increasingly high dimensional data remains challenging. From a functional perspective, distinct sets of transcription factors bind to enhancers to induce transcription of genes, representing activity that can be quantified by gene expression profiles. Further, six different enhancer classes have been previously identified based on transcription factor binding patterns. Defining a 'pathway' as a group of co-activated enhancers, transcription factors, and genes, we pursue complementary approaches to systematically identify cell identity-related pathways.

How to generate novel proteins from models of natural sequences

Trenton Bricken[1], **Nathan Rollins**[2], **Nikki Thadani**[2], **Debora Marks**[2]

[1] Duke University
[2] Harvard Medical School

A Hierarchical State-Space Model with Gaussian Process Dynamics for Functional Connectivity Estimation

Rahul Nadkarni, Nicholas J. Foti, Adrian KC Lee, Emily B. F IS

University of Washington

- We present an approach for learning directed, dynamic functional interactions between brain regions across multiple subjects from magnetoencephalography (MEG) data
- We pool subjects' data to learn a group-level connectivity structure while also modeling interesting subject-specific differences
- Our method incorporates Gaussian processes to model functional connections that are smoothly time-varying while capturing estimation uncertainty

Modelling 3D Body Posture and Dynamics with a von-Mises-Fisher Gaussian Model

Libby Zhang, Scott Linderman

Stanford University

Objectives: Improve behavioral recording to noise, occlusion, and complex environs. Develop posture model & dynamics with interpretable structures.

Eukaryotic-wide reconstruction of RNA binding protein specificities by joint linear embedding of protein and RNA sequence k-mers

Alexander Sasse[1,2,4], **Debasish Ray**[2], **Timothy Hughes**[1,2], **Quaid Morris**[1,2,3,4,5]

[1] Department of Molecular Genetics, University of Toronto
[2] Donnelly Centre for Cellular and Biomolecular Research, Toronto
[3] Department of Computer Science, University of Toronto
[4] Vector institute for Artificial Intelligence, Toronto
[5] Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, USA

Predicting RNA binding specificity from protein sequences. RNA binding proteins (RBPs) bind to specific RNA sequences, also called motifs. RNA binding specificities are measured in vitro by RNAcompete assay. Specificities are represented as enrichmrnts of 7- mers in bound RNA probes (Z-scores).

Non-negative Independent Factor Analysis for single cell RNA-seq

Maria Chikina, Weiguang Mao, Dennis Kostka, Maziyar Baran Pouyan, Maria Chikina

University of Pittsburgh

Single-cell RNA sequencing (scRNA-seq) technologies generate large amounts of data and proper dimensionality reductions techniques which are able to reduce data redundancy while maintaining biological interpretability are an active area of research.In the context of scRNAseq data variants of PCA, ICA and NMF are widely used for dimensionality reduction. Principle component analysis (PCA) seeks directions that maximize variance and provides the best rank-k approximation under least squares loss. Independent Component Analysis (ICA) finds maximally independent or maximally non-Gaussian directions but doesn't minimize reconstruction error and thus doesn't model the likelihood of the data. Non-negative matrix factorization (NMF) imposes non-negativity constraints on both factors and loadings which often leads to interpretable decompositions. We propose Non-negative Independent Factor Analysis (NIFA) that combines properties of ICA, PCA and NMF. Our approach simultaneously models single- and multi-modal factors thus isolating discrete cell-type identity and continuous pathway-level variations into separate components. Furthermore, our model constrains the loading to be positive providing greater biological interpretability.

New Models for Discovering Genotype to Phenotype in Antimicrobial Resistance

Ada Shaw, Mafalda Figueiredo-Dias, Jonathan Fraser, Debora Marks

Harvard Medical School

The evolution of antibiotic resistance is rendering current antibiotics ineffective against bacterial infections. Understanding the relationship between bacterial genotype and phenotype helps us better design antimicrobials and develop longer term strategies for managing the evolution of resistance. Genome-wide association studies (GWAS) are used to find significant genetic variations that confer phenotypes. Single loci GWAS does not account for all the heritability of phenotypes but when interactions between loci are taken into consideration, some of this missing heritability can be accounted for. However, one of the major challenges in studying epistatic interactions is the severe loss of statistical power due to the vast number of interactions to be tested. We model epistasis between SNPs of different genes by extending the maximum entropy (Potts) model that has been applied to finding strongly coupled interactions between protein residues, to find epistatic interactions between strongly coupled loci. In doing so we narrow down our search space to those loci with strong co-evolutionary history and thus gain statistical power. We build an alignment from a pan-genome — allowing us to account for the presence or absence of genes that are included within the accessory genome. This is especially relevant to antibiotic resistance as the horizontal gene transfer of resistant genes accelerates the acquisition of resistance among bacterial populations. We apply the model to studying three different clinically relevant bacteria — E. coli, S. aureus, S. pneumoniae — whose resistance data has been collected by hospitals around the world. Our method disentangles the genetic contingencies driving antibiotic resistance in these three bacterial species, providing genomic insights that could help us better predict and treat antibiotic resistance in bacterial infections.

Augmenting Protein Network Embeddings with Sequence Information

Hassan Kane, Mohamed Coulibali, Pelkins Ajanoh, Ali Abdalla

WL Research

- Proteins play a key role in cellular processes
- Representation learning has been a very helpful paradigm to understand complex data manifolds
- Proteins can be represented using primary, secondary and tertiary structure or nodes in PPI Network

→ How do we develop representations integrating those data sources ?

Knowledge Complex: Toward Learning Representation of Digital Human

Xiaotian Yin[1], Tim Tingqiu Yuan[2], Jian Li[2]

[1] Futurewei Technologies
[2] Huawei Technologies

Towards a Disease-Relevant Benchmark for Co-expression Module Detection

Árpád Vezér, Eirini Arvaniti, Craig Glastonbury, Francesca Mulas, Povilas Norvaišas, Poojitha Ojamies, Aaron Sim, Páidí Creed

BenevolentAI

The detection and analysis of co-expression modules from high-dimensional gene expression data is a common approach to identifying biological processes and their defects in disease. While several module detection methods have been proposed, benchmarking these methods remains challenging [1]. A good benchmark that captures pertinent aspects of bioinformatic analyses would be a useful tool for method development, method selection and hyperparameter tuning, acting as a proxy measure of performance on exploratory analyses for which evaluation is difficult. Here we build on the benchmarking framework developed by [1], discuss some of the difficulties in designing a conclusive module detection benchmark and work towards defining a benchmarking framework that can be focused on specific diseases or processes.

Amyotrophic Lateral Sclerosis endotype detection using Bayesian Biclustering

Craig A. Glastonbury, Povilas Norvaisas, Arpad Vezer, Aaron Sim, Francesca Mulas, Hamish Tomlinson, Poojitha Ojamies, Joanna Holbrook, Paidi Creed

BenevolentAI

After fitting latent variable models to omics data, downstream interpretation is a significant unsolved problem. We have developed a pipeline that exploits both known clinical and biological metadata and utilises a joint model fit with both case (ALS) and control samples. Using this pipeline we identify representations that distinguish ALS cases from controls and latent representations that are free from confounding. Additionally, our pipeline establishes latent representations that are associated to known covariates. This workflow highlighted 5 latent representations that are candidate ALS subtypes, 2 of which have been previously described in the ALS literature.

Towards a Disease-Relevant Benchmark for Co-expression Module Detection

Árpád Vezér, Eirini Arvaniti, Craig Glastonbury, Francesca Mulas, Povilas Norvaišas, Poojitha Ojamies, Aaron Sim, Páidí Creed

BenevolentAI

The detection and analysis of co-expression modules from high-dimensional gene expression data is a common approach to identifying biological processes and their defects in disease. While several module detection methods have been proposed, benchmarking these methods remains challenging [1]. A good benchmark that captures pertinent aspects of bioinformatic analyses would be a useful tool for method development, method selection and hyperparameter tuning, acting as a proxy measure of performance on exploratory analyses for which evaluation is difficult. Here we build on the benchmarking framework developed by [1], discuss some of the difficulties in designing a conclusive module detection benchmark and work towards defining a benchmarking framework that can be focused on specific diseases or processes.

References

[1] Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. Nature communications, 9(1), 1090.

Learning Good Representations of Cell State from Cell Painting

Zhenghao Chen, Calvin Jan, Frank Li, Jun Xu

Calico Life Sciences

A modified ResNet architecture achieves $> 99\%$ accuracy in identifying genetic perturbations using Cell Painting (RxRx1 dataset). Our model allows us to quantify information content of Cell Painting stains and de-multiplex combinations of stains. Our model also generalizes well to new perturbations and (less well) to new cell lines; initial results on small molecule perturbations are promising.

Flatsomatic: A Method to CompressSomatic Mutation Profiles

Geoffroy Dubourg-Felonneau[1], Yasmeen Kussad[1,2], Dominic Kirkham[1], John W Cassidy[1], Nirmesh Patel[1], Harry W Clifford[1]

[1] Cambridge Cancer Genomics
[2] University of Lancaster

With the widespread adoption of next generation sequencing technologies, we are beginning to understand how each cancer tumor is unique on the genetic level, and there has recently been an increasing optimism surrounding the development of personalized cancer treatments. Precision oncology involves the process of identifying genomic features driving an individual tumor and designing a personalized therapeutic strategy in response. This presents a classification problem that is well suited to supervised machine learning algorithms, although due to the high complexity and dimensionality of such genomic data, applying models directly on the raw data can be difficult. A common way of reducing the dimensionality is to select features with known impact (e.g. driver genes, cell signaling pathways, etc). Another way is to use models that compress the data whilst keeping most of the signal. We present Flatsomatic, a Variational Auto Encoder (VAE) optimized to compress somatic mutations that allow for unbiased data compression whilst maintaining the signal. We show that the Flatsomatic representations of 64 dimensions keep the same predictive power that the original 8298 dimensions vector has for drug response prediction.

Ultra-sensitive serum Assay Development toward the Early Detection of Breast Cancer using Single Molecule Arrays (Simoa), an innovative biomarker detection platform

Tien-yu Hsin

CVS Health

Effective Sub-clonal Cancer Representation to Predict Tumor Evolution

Adnan Akbar, Geoffroy Dubourg-Felonneau, Andrey Solovyev, John W Cassidy, Nirmesh Patel, Harry W Clifford

Cambridge Cancer Genomics

- Cancer is an evolutionary process where cellular sub-populations known as sub-clones compete with each other under conditions of Darwinian natural selection (Fig 1).
- Resultant Intra-Tumor Heterogeneity (ITH) has been associated with higher likelihood of relapse and increased resistance to therapy in cancer patients.
- The ability to precisely predict how these sub-clones will evolve over time can help clinicians to develop an effective cancer treatment and reduce failures.
- In this paper, we present a novel data-driven approach to predict cancer evolution for real-world data.
- Our proposed method is based on the intuition that if we can capture the true characteristics of sub-clones within a tumor and represent it in the form of features, a sophisticated machine learning algorithm can be trained to predict its behavior.

Two Novel Unsupervised Machine Learning Algorithms for High-dimensional datasets with applications for Gene Expression Studies

Zhipeng Wang[1,2], **David W. Scott**[1]

[1] Rice University
[2] Apple Inc

In genomics, gene expression datasets such as RNA-seq is widely used for pathway analysis, differential gene expression analysis and gene expression correlation studies. Unsupervised learning algorithms such as clustering are heavily applied to those datasets to extract relevant information in the biological context. Gene expression datasets are usually high-dimensional, with the number of genes (variables) much larger than the number of samples. In order to perform unsupervised learning tasks for gene expression studies, We need to develop unsupervised machine learning algorithms tailored for high-dimensional datasets. In this work, we will introduce two novel unsupervised machine learning algorithms suited for high-dimensional datasets, and explore their applications in genomics, particularly in gene clustering and regulatory pathway analysis. We first introduce the unsupervised variable selection algorithms based on the projection density score (PDS), a statistics metric defined by me and my co-authors to quantify the importance of variables for the unsupervised learning tasks, such as clustering. In addition to PDS, we will present another unsupervised machine learning algorithm called Sub-Matrix Principle Component Analysis (SMPCA). This algorithm automatically picks the relevant set of variables to reconstruct the original signal. In biological context especially in gene expression studies, the gene expression data is a compound signal consisting of many overlaid signals, each of them contains a set of genes. Some of those signals are noise while others are important components. SMPCA is able to identify important components of the original gene expression signal in a completely unsupervised fashion.

Distributed Vector Representation for Medical Natural Language Processing

Quanzhi Li

Alibaba Group

To learn meaningful representation of life, it is unavoidable that we need to deal with different textual data, such as patient information, symptom description, medical and and life science literature, etc. Natural language processing (NLP) and natural language understanding technology is the key to analyze and discover knowledge from various textual data. In the past, researchers used machine learning algorithms with traditional text representation methods, such as bag-of-words (BOW), for NLP tasks. The BOW representation and other traditional representation methods are high dimensional, sparse, and they usually ignore the order and syntactic and semantic information of the words. In contrast, embedding (or distributed vector representation) is a dense, low-dimensional and real-valued vector for a text unit, e.g. word, sentence, paragraph. Text embedding maps the variable length text to dense vector representations, and it can capture both the syntactic structure and semantics of the text. Traditional bag-of-words and bag-of-n-grams hardly capture the semantics of text. Therefore, embedding overcomes the curse of dimensionality and the lack of syntactic and semantic information in representations. Moreover, embeddings are learned in an unsupervised manner, which captures the knowledge in a large unlabeled corpus, and it can be transferred to the downstream tasks with small labeled data sets. An (word, sentence, or other types of text units) text embedding typically consists of dozens or hundreds of dimensions, and each dimension represents a feature. In an embedding, the meaning of a text unit is distributed across dimensions. Text embedding is particularly suitable for deep learning models which consists of multiple layers employing matrix operations to find the high-level representations of text data. Hence, embedding has become an unavoidable and ideal choice for text representation in deep learning era, and it has been used in many NLP tasks in various domains. Even though embeddings have become de facto standard for text representations for deep learning based NLP tasks in both general and medical domains, there is no paper presenting a detailed review of embeddings in the form of classification of embeddings as well as the challenges to be solved in medical or life science domain. Medical embeddings can be classified into ten categories, with each category mapping variable length text (e.g. character, words, phrases, sentences or documents), medical codes, or CUIs to dense vector representations. These ten types of clinical/medical embeddings are: character embedding, word embedding, code embedding, CUI embedding, patient embedding, phrase embedding, sentence embedding, augmented embedding, including augmented word embedding and augmented sentence embedding, document embedding, and document collection embedding. This study aims to present a detailed review of embeddings in medical/clinical natural language processing - how they are generated, their key features, their pros and cons, the challenges,

and where and how to use them.

Representation Learning for Transcriptomics-based Phenotype Prediction

Aaron M. Smith[1], Jonathan R. Walsh[1], John Long[2], Craig B. Davis[3], Peter Henstock[3], Martin R. Hodge[3], Mateusz Maciejewski[3], Xinmeng Jasmine Mu[3], Stephen Ra[3], Shanrong Zhang[3], Daniel Ziemek[3], Charles K. Fisher[1]

[1] Unlearn.AI
[2] Columbia University
[3] Pfizer, Inc.

- We report a comprehensive analysis of phenotype prediction from transcriptomics data spanning prediction tasks from ulcerative colitis, atopic dermatitis, diabetes, to many cancer subtypes for a total of 24 binary and multiclass prediction problems and 26 survival analysis tasks.

- Using the recount2 database, we systematically investigate the influence of gene subsets, normalization methods and prediction algorithms, with a particular emphasis on representation learning.

- We also explore the novel use of deep representation learning methods on large transcriptomics compendia, such GTEx and TCGA, to boost the performance of state-of-the-art methods. The resources and findings in this work should serve as both an up-to-date reference on attainable performance, and as a benchmarking resource for further research.

Integrating Markov process and structural causal model enables counterfactual inference in complex systems

Rober Ness, Kaushal Paneri, Olga Vitek

Microsoft

- This research contributes a framework for casting a Markov process model of a system at equilibrium as a structural causal model, and carrying out counterfactual inference.
- We define the structural causal models in terms of the parameters and the equilibrium dynamics of the Markov process models, and counterfactual inference flows from these settings.
- The proposed approach alleviates the identifiability drawback of the structural causal models. The counterfactual inference is consistent with the counterfactual trajectories simulated from the Markov process model.
- We showcase the benefits of this framework in case studies of complex biomolecular systems with nonlinear dynamics. In presence of Markov process model misspecification, counterfactual inference leverages prior data, and therefore estimates the outcome of an intervention more accurately than a direct simulation.

Latent Representations from Factor Analysis and CNNs of Genomic and Proteomic Data Reveal Immune Pathways in Colorectal Cancer

Tzu-Yu Liu, Francesco Vallania, Michael Dzamba, Mitch Bailey, Charles Roberts, Barbara Engelhardt, C. Jimmy Lin

Freenome

- Plasma proteins and cell-free DNA (cfDNA) are important classes of cancer biomarkers
- Our goal is to identify markers of CRC from this high-dimensional space
- Characterization of their biological functions and interrelationships is ongoing

Generative Models for Target Specific Drug Design

Vijil Chenthamarakshan, Payel Das, Tom Sercu

Inkit Padhi

- Drug discovery is still a cost and time-consuming process with low success rate.
- The aim is to design/optim molecules with a specific set of attributes, e.g. binding affinity (BA) to a specific target protein and drug likeliness (QED).
- We are building an efficient design platform to automatically find optimal molecules that bind to a given target protein, leveraging public molecular and protein-ligand binding affinity data.
- Variational autoencoder (VAE) model defined over SMILES molecular representation is combined with a conditional sampling scheme in latent space to generate novel and valid molecules with high drug-likeliness and high predicted binding affinity with a specific target protein.

A Novel Signature for Cancer Classification from Healthy DNA

Siddharth Jain, Bijan Mazaheri, Netanel Raviv, Jehoshua Bruck

Caltech

Summary:

- New microscope to view the genome.
- Decodes the evolutionary memory of tandem repeat regions to measure the accumulation of mutations.
- Detected the cancer-type signal from the healthy genome.
- Implicitly inferring about a process of acquiring mutations in the blood that is associated with cancer in a tissue-specific way.
- Has potential applications in predicting future cancer risk and early cancer detection.

What is the function of form? A data-driven approach of developmental biology

Paul Villoutreix[1], **Sarah Rubin**[2], **Tomer Stern**[3,4], **Elazar Zelzer**[2], **Joakim Anden**[7], **Bomyi Lim**[8], **Ioannis Kevrekidis**[9], **Amit Singer**[6], **Stanislav Shvartsman**[3,4,5]

[1] Turing Center for Living Systems (IBDM, UMR 7288 & LIS, UMR 7020), Marseille, France
[2] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel
[3] The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[4] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[5] Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA
[6] Program in Applied and Computational Mathematics and Department of Mathematics, Princeton University
[7] Center for Computational Mathematics, Flatiron Institute, New York City, NY, USA
[8] Department of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia, PA 19104;
[9] Department of Chemical and Biomolecular Engineering and Department of Applied Mathematics and Statistics, Johns Hopkins University

Multicellular organisms develop from a single fertilized egg. The sequence of events leading to the precise positioning of individual cells with the required fate within the developing organism is orchestrated by gene regulation within and between cells, cell proliferation and rearrangements as well as global morphological changes, involving coordinated dynamics at multiple scales, from single molecules, to cells, to tissues. Even though various parts of these processes have been elucidated, a generic framework that would capture the interactions between them is still missing. In particular, the question of how morphology affects and is affected by cell differentiation is still open. We propose to describe a developing organism as an evolving graph where each cell correspond to a node, spatial relationships are encoded as the edges and transcriptomic states as labels on the nodes. We will explore how the action of cell proliferation, spatial rearrangement and differentiation dynamics affect the characteristics of these graphs. Dynamical processes in biology are studied using an ever-increasing number of acquisition techniques, each of which brings out unique but partial features of the system. Multiple recent works have shown the utility of the search for an underlying low dimensional manifold such as pseudo-time technique for single data RNASeq data in developing embryos. This approach can be completed by similar search for low dimensional manifolds from microscopy techniques on morphological features of nuclei, cells and tissues. A general assumption is that the parameters governing the processes in the various measurement spaces are the same, hence leading to latent

manifolds of similar dimensions. Therefore, integrating data from various sources amounts to finding mapping between these manifolds. As a first step in this direction, we will show how we used machine learning methods such as multi-view and semi-supervised learning to integrate heterogeneous measurements, and how we developed innovative data visualization tools to extract meaningful patterns in 10000+ cells. We illustrate our approaches using confocal microscopy datasets of studies of pattern formation in Drosophila embryo and light sheet microscopy datasets from studies of bone formation in mouse development.

Towards interpretable single-cell representations using disentangled variational autoencoders

Gokcen Eraslan, Aviv Regev

Broad Institute

A good data-driven representation of a cell should recapitulate known biological processes and should ideally contain different facets of the cell identity such as cell type and state programs. However, the standard single-cell RNA sequencing (scRNA-seq) pipelines mainly operate on clusters and therefore focus mostly on the major sources of biological variation such as the "cell types". Consequently, relatively minor sources of variation that might be shared across multiple cell types such as cell cycle or immune activation cannot be easily identified. Therefore, the representation used and/or produced by the standard pipelines lacks important aspects of cells. As the size and the biological diversity of single-cell datasets grow, better cell representations are needed even more since the shared programs expected to exhibit more profoundly and explain the data better. In a parallel universe, namely the field of machine learning, learning meaningful representations of the data is already an important challenge and the community is making significant progress towards obtaining both expressive, interpretable and data-driven features using unsupervised machine learning techniques. One promising line of research focuses on a concept called "the disentanglement" in variational autoencoders (VAEs) which encourages the VAE to extract more independent features where each feature represents only a single true generative factor (Burgess et al. 2018, Chen et al. 2018, Dupont et al. 2018, Mathieu et al. 2019). VAEs are getting more widespread in single-cell genomics. These models are currently leveraged for removing unwanted effects from data e.g. batch correction (Lopez et al. 2018) and predicting stimulation effects (Lotfollahi et al. 2019). Especially with the latest developments in the machine learning field, VAEs are proposed as promising feature extractors which might then suggest that this can be translated to computational biology and be exploited for obtaining better cell representations that align well with the known biology. Such approaches would then be a good alternative to existing feature extraction techniques such as topic models and non-negative matrix factorization (NMF) which are preferred often by the community today. In this work, we focused on a particular type of VAE variant called beta-TCVAE (Chen et al. 2018) and mainly investigated whether disentanglement can be achieved in scRNA-seq datasets and whether it improves our understanding of biological systems by evaluating how well disentangled VAE dimensions aligns with known biological features. Furthermore, we compared these features with those from topic models, independent component analysis and NMF using heuristics and empirical evaluation. Our results suggest that the selection of hyperparameters (such as the number of dimensions and the disentanglement strength), and the data normalization plays a crucial role in disentanglement and with a proper configuration of hyperparameters, VAEs can infer biologically meaningful representations

and patterns that other methods cannot reveal.

Extracting the Relevant Information from Evolution for Mutation Effect Prediction

Jonathan Frazer, Mafalda Dias, Debora Marks

Harvard Medical School

The use of natural protein sequences for training generative models has already led to a number of successes, in 3-dimensional structure prediction, mutation effect prediction and even more recently, for therapeutic design. State-of-the-art methods (EVCouplings, variational autoencoders, autoregressive models) tend to essentially treat sequences as independent and identically distributed, thereby neglecting the structure of natural sequence variation — the end result of millions of real-life evolutionary experiments. An open question is then how to go beyond current methods by taking advantage of the rich complexity of evolutionary information. For tasks such as mutation effect prediction this challenge can be recast as a search for models that identify and capture the relevant aspects of the dataset, as relevant information is primarily, but not necessarily entirely, contained in a small fraction of the sequence space. We therefore need methods that can account for a spectrum of relevance across the natural protein dataset. Multiple areas of applied statistics face similar challenges. One example is the modeling of complex survey data, where one is concerned with extracting information about a population from non-representative samples. In this case, treating the problem in terms of a Bayesian weighting scheme has been shown to be an effective strategy. Inspired by this success, we take a similar approach to address the issue of extracting relevant information from natural protein sequence data. Focusing on mutation effect prediction, we show that a Bayesian weighting approach has a number of advantages as compared to training on raw sequence data. Firstly, it enables us to ask fundamental questions about where the complexity of modeling mutation effects lies. For instance, does the successful modeling of mutation effects for a given protein require a highly expressive model, or can simple models perform well once the complexity of the data is accounted for by the weights? Preliminary work indicates the latter to be true. We seem to be finding that extremely simple models can perform just as well as, if not better than, deep learning methods, once trained on the weighted data. This leads to a second, more practical advantage, which is that state-of-the-art performance can be achieved with far fewer sequences than used in previous studies, thereby opening the door to studying proteins that were previously thought inaccessible.

Discovering and classifying deep intronic cryptic splice sites with the COSSMO splicing predictor

Hannes Bretschneider

University of Toronto

- We apply COSSMO, a deep learning model to predict alternative splice site usage to pathogenic deep intronic variants and disease-linked variants from the Human Gene Mutation Database.

- COSSMO is an algorithm based on representation learning that predicts alternative splicing patterns from sequence alone with neither expert-derived features nor any information (such as conservation) unavailable to the splicing machinery.

- COSSMO learns relevant representations such as splicing consensus motifs and binding motifs for splicing regulatory elements and core spliceosome components.

- Deep intronic variants cause disease through the activation of cryptic splice sites within introns that compete with constitutive splice sites.

- COSSMO can predict the splicing impact of variants that are thousands of base pairs away from the constitutive splice site.

- Our algorithm is more accurate than MaxEntScan [4] at classifying pathogenic versus common deep intronic variants.

- On the Human Gene Mutation Database, COSSMO predicts that hundreds of missense/nonsense variants have secondary splicing effects.

- Many intronic HGMD variants at up to 1kb distance from a splice site alter splicing.

GeneWalk identifies relevant gene functions for a biological context using network representation learning

Robert ietswaart, Benjamin M. Gyori, John A. Bachman, Peter K. Sorger, L. Stirling Churchman

Harvard Medical School

A general question in high-throughput sequencing data analysis is what biological function a gene of interest has in a particular experimental context, for instance an RNA-seq experiment or CRISPR screen. Gene Ontology (GO) annotation lists all known gene functions, but it is not context-specific. Enrichment analysis methods generally determine what biological processes are enriched across the differential transcriptome, but do not address what function any individual gene could have in an experimental context.

Genome Wide Associations of Learned Low Dimensional Representations of Cardiac MRI

Samuel Friedman, Nathaniel Diamant, James Pirruccello, Puneet Batra, Steven Lubitz, Anthony Philippakis

Broad Institute of MIT and Harvard

Cardiac magnetic resonance imaging (MRI) studies provide high resolution data about the structure, temporal dynamics and health of the human heart. These images are massive and require expert interpretation. Learning a low dimensional representation of these rich images gives insight into their undergirding structure. Furthermore, when the Cardiac MRI subjects have genetic data, as is the case for forty thousand people included in the UK Biobank, the learned encoding provides a target for genetic study. But what low dimensional representation to use, and how to learn its components? We explore unsupervised and semi-supervised methods for learning these representations with 3D Convolutional autoencoders. Our motivation is to discover latent spaces endowed with genomic information, clinical implications, and, ideally, features of preclinical cardiovascular disease. To that end, the quality of the encoding is assessed by its ability to discern known (and discover new) genomic loci with cardiovascular relevance via genome wide association studies (GWAS). Better representations are learned by incorporating supervisory signals from anatomical segmentations of the heart tissue into the myocardium and the left ventricle. Denoising autoencoders, trained to reconstruct the input in the presence of random noise, yield more robust and clinically compelling representations. Variational autoencoders (VAEs) give deeper insight into the latent space. Between the encoder and the decoder of a VAE we sample from learned means and variances. During training, the KL-divergence from the Gaussian distribution is computed and added to the reconstruction loss. Sampling from interpolations between learned encodings reveals the organization of the latent space. For example, in digit classification one axes may correspond with tilt and another with loopiness. An additional loss term penalizes a subset of the components of the latent space according to their distance from the largest principal components of ancestry. This encourages the encoding to be aware of common genetic variation and to localize ancestral information in the latent space. This factorization of the learned representation allows us to investigate genetic and lifestyle components of the encodings separately. In the end, GWAS on learned values were able to uncover loci at genome-wide significance after Bonferroni correction. Many of these loci have been previously identified in the literature as significant for cardiovascular health, confirming that the models were able to extract physiologically relevant information from imaging data. Other loci found to have genome-wide significance have not yet been described in the literature, but belong to genes known to be involved in the cardiovascular system.

What is a meaningfull representation?

Nicki Skafte Detlefsen[1], **Søren Hauberg**[1], **Wouter Boomsma**[2]

[1] Technical University of Denmark
[2] University of Copenhagen

Representation learning is gaining increasing attention as a way to distill abstract and complex data into compact and interpretable vector spaces, and could potentially have a dramatic impact on the life sciences. However, it remains unclear what is required for a representation to be meaningful/useful. Towards such a definition, we here ask five open questions covering different aspects of representation learning, in the hope of starting a discussion about fruitful future directions of research. As the basis for the discussions, we present illustrative examples within the domain of protein sequence modelling.

Interpretable multistate survival analysis using an ODE neural network

Stefan Groha, Alexander Gusev

Dana Farber Cancer Institute

Motivation and Goal:

- Success of immune checkpoint inhibitors (ICI): Significant increase of PFS and OS over chemotherapy
- However: Difference in response, acquired resistance, adverse events
- Adverse events in 20–30% of ICI patients
- BUT: No biomarkers
- Large cohort of ICI patients (∼3000) with germline & somatic mutation data, lab results, diagnoses, charts, etc.

⇒ Prediction of toxicity/death from available data for clincial decision support

Expanding genome editing tools through exploration of new CRISPR-Cas proteins and DNA repair enzymes: A top-down approach using big data and unsupervised machine learning

Hyunjin Shim, Jillian F. Banfield

UC Berkeley

Genome-resolved metagenomics has greatly contributed to the discovery of new CRISPR-Cas systems, including the first reported CRISPR-Cas enzymes in bacteriophages. We now have sequence datasets approaching a billion proteins encoded in genomes of the most extensive microbes to identify novel proteins involved in DNA manipulation. We propose a top-down approach of applying unsupervised machine learning methods such as autoencoders on these big datasets to explore genome editing tools in a systematic, automatized and unbiased manner. We aim to use patterns of co-expression involving the hypothetical proteins of interest to reinforce hypothesized roles in DNA manipulation, with the goal of validating candidate CRISPR-Cas proteins and DNA repair enzymes with functional testing.

Inferring copy number state aberrations from scRNA-seq data

Alina Selega, Quaid Morris

University of Toronto

Copy number aberrations are gains (losses) of large genomic segments arising in cells of an individual during their lifetime. CNAs can span regions from few kilobases to whole chromosomes and are a major factor in cancer development. Tumours are comprised of multiple cell populations with distinct patterns of accumulated point mutations and CNAs (intra-tumour heterogeneity). These populations may differ in their response to treatment, impacting disease progression. Single-cell technologies can profile individual cells in a tumour. Clonal architecture reconstruction and CNA calling is typically performed using bulk or scDNA-seq data. Combining both DNA and RNA readout can map genomic changes to their transcriptional effects for a cancer.

Tandem Data-Driven and Hypothesis-Driven/Bio-Inspired Paradigms

Prashant Emani, Hussein Mohsen, Jonathan Warrell

Yale University

The ascendance of machine learning in the realm of biology, coupled with the eruption of biomedical data, has provided unprecedented opportunities for purely data-driven approaches. Machine learning offers the promise of uncovering representations of underlying biological structure even in the absence of guiding hypotheses. This promise, however, is susceptible to abuse by overreliance on 'blackbox' methods opaque to biological interpretation. On the other hand, proponents of traditional hypothesis-driven research must also acknowledge the burden of historical labels and categories prevalent in scientific research. For example, is the widespread use of putative ancestral groups still justified in the age of affordable genotyping and personalized medicine? Do we continue to see mental disorders in diametric opposition to 'healthy controls', when complex phenotypes may occupy a spectrum in the population? We need to be cognizant of the possible inequities propagated by dogmatically inherited phenomic categorizations. We propose that the respective pitfalls and promises of hypothesis-driven and data-driven approaches can be balanced in collaborations, and accordingly call for the creation of principled guidelines for interactions between biologists and machine learnists: (1) Biologists should pursue studies carefully informed by theoretical models and hypotheses, as well as machine learning methods, using pre-existing knowledge to inform their analyses at all stages. To complement data-driven practices, we call for a higher emphasis on the underlying theory before executing experiments. Using the same data for both hypothesis generation and testing limits the credibility of findings, and hypothesis pre-registration used in other fields such as sociology can help serve as a model. (2) Machine learnists should, in parallel, develop computational frameworks that are strongly informed by experimental assays at all possible cellular, tissue and organ scales, but should learn phenomic relationships de novo. (3) With proper blinding of the two approaches, a comparison of the two sets of results can be made in the end. The hope is that, with a tandem approach that is blinded to avoid confirmation bias, collaborations can yield representations that simultaneously carry biological insight and avoid prejudicial phenomic categorizations.

Meaningful models need diets high in protein: Rationale for developing next-generation proteomics with nanopore technology

Jeff Nivala

While the human genome encodes for $\sim 20,000$ genes, these genes are differentially expressed, spliced, and translated to yield hundreds of thousands of different protein variants. Still further, most proteins undergo multiple post-translational modifications (PTMs) to yield a bewildering level of variation approaching the millions. Recently, mass spectrometry (MS) has made incredible progress in its ability to detect and quantify these variants, particularly at the single-cell level. However, MS technology currently has inherent limitations that prevent it from providing a complete understanding of proteomic complexity. Ultimately, these limitations lead to a missing link in genotype-to-phenotype and constrain "full-stack" molecular models. In fact, we argue that the proteome represents the largest source of hidden variables for models of life at the molecular level. The question, then, is how to develop technologies that can complement MS to find these hidden variables and ultimately capture true proteomic complexity to better inform our models? Here, we hypothesize that nanopore technology is among the most promising technological approaches to generating deeper levels of proteomic data. For example, recent advances in long read nanopore-based sequencing of DNA and RNA are leading to breakthroughs in genomics, transcriptomics and epigenetics, including illumination of genomic 'dark regions,' direct detection of modified bases, and full-length characterization of mRNA splice variants. Current nanopore methods for proteins, on the other hand, have yet to realize the potential that nascent nanopore technology platforms (eg. Oxford Nanopore Technologies' MinION) offer for proteomics, such as the development of single-molecule sequencing of full-length proteins, or single-cell protein analysis. We believe, however, a path to enabling next-generation proteomics analysis with nanopore technology is possible and, as a community, we can work to address the remaining barriers that must be overcome to feed our models with the protein they need to thrive.

Extracting the Relevant Information from Evolution for Mutation Effect Prediction

Mafalda Dias, Jonathan Frazer, Debora Marks

Harvard Medical School

The use of natural protein sequences for training generative models has already led to a number of successes, in 3-dimensional structure prediction, mutation effect prediction and even more recently, for therapeutic design. State-of-the-art methods (EVCouplings, variational autoencoders, autoregressive models) tend to essentially treat sequences as independent and identically distributed, thereby neglecting the structure of natural sequence variation — the end result of millions of real-life evolutionary experiments. An open question is then how to go beyond current methods by taking advantage of the rich complexity of evolutionary information. For tasks such as mutation effect prediction this challenge can be recast as a search for models that identify and capture the relevant aspects of the dataset, as relevant information is primarily, but not necessarily entirely, contained in a small fraction of the sequence space. We therefore need methods that can account for a spectrum of relevance across the natural protein dataset. Multiple areas of applied statistics face similar challenges. One example is the modeling of complex survey data, where one is concerned with extracting information about a population from non-representative samples. In this case, treating the problem in terms of a Bayesian weighting scheme has been shown to be an effective strategy. Inspired by this success, we take a similar approach to address the issue of extracting relevant information from natural protein sequence data. Focusing on mutation effect prediction, we show that a Bayesian weighting approach has a number of advantages as compared to training on raw sequence data. Firstly, it enables us to ask fundamental questions about where the complexity of modeling mutation effects lies. For instance, does the successful modeling of mutation effects for a given protein require a highly expressive model, or can simple models perform well once the complexity of the data is accounted for by the weights? Preliminary work indicates the latter to be true. We seem to be finding that extremely simple models can perform just as well as, if not better than, deep learning methods, once trained on the weighted data. This leads to a second, more practical advantage, which is that state-of-the-art performance can be achieved with far fewer sequences than used in previous studies, thereby opening the door to studying proteins that were previously thought inaccessible.

A Bayesian Nonparametric approach to super resolution localization microscopy

Mariano Gabitto[1], **Herve Marie-Nellie**[2,3], **Ari Pakman**[4], **Andras Patak**[1], **Michael Jorda**[2]

[1] Simons Foundation
[2] UC Berkeley
[3] Insitro
[4] Columbia University

Spatio-Temporal Model for Intracerebral Hemorrhage: Embeddings Methods Solving Violating Assumptions on ML

Tsuyoshi Okita

Kyushu Institute of Technology

Motivation: Rapid Growth of Intracerebral Hemorrhage

- Intracerebral hemorrhage (ICH): rapid growth after the breeding occurs.
- Within 24-72 hours after the injury: an important period whether the hemorrhage grows or not.
- Such decision should lead to the brain surgery.
- Mistake in predicting the rapid progress causes the death of patient.
- The death rate is up to 75%.

Mapping behavioral repertoires by quantifying 3D body and limb kinematics during naturalistic behavior

Timothy W. Dunn[1], **Jesse D. Marshall**[2], **Diego Aldarono**[2], **William Wang**[2], **Kyle Severson**[1], **David Hildebrand**[3], **Fan Wang**[1], **David Carlson**[1], **Bence Olveczky**[1]

[1] Duke University
[2] Harvard University
[3] The Rockefeller University

Tandem Data-Driven and Hypothesis-Driven/Bio-Inspired Paradigms

Prashant S. Emani, Hussein Mohsen, Jonathan Warrell, Mark B. Gerstein

Yale University

One may pause to question what a "meaningful" representation of biology signifies. In an end-goal-directed paradigm, models predicting the behavior of a biological system under restricted conditions are sufficient. We derive ‘meaning’ from the prediction and any ensuing clinical decisions. Machine learning (ML) excels at this, with an arsenal of methods capable of complex interdependencies between variables to build accurate predictive models of biological & clinical phenotypes. The intermediate components may not have direct correspondence with human-interpretable biological models. However, there is often a necessity to consider the intervening levels in terms of causal mechanisms and interactions between biological entities amenable to human interpretation. In this more mechanism-oriented paradigm of thought, ‘meaning’ arises in terms of the interplay of more tangible biological elements, and of ascribing causation. Here, hypothesis-driven approaches have reigned supreme, with biological constraint limiting the parameter space of models every step of the way. Bio-inspired modelling forms an important subcategory of this paradigm, and one which lends itself to easy collaboration with machine-learning methods. Neither approach is complete by itself. We are concerned that hypothesis-based approaches perpetuate historical phenotypic categorizations, which are almost invariably simplifications of true behavior. This is especially true of human behavioral phenotypes with significant social stigma.

Probabilistic Atlases of C.elegans Neurons in NeuroPAL

Erdem Varol, Gonzalo Mena, Amin Nejatbakhsh, Eviatar Yemini, Liam Paninski

Harvard University

- The worm C.elegans is a unique species since its nervous system is stereotypical: neurons (300) and connections remain the same from animal to animal. Therefore, each neuron has a name.

- New technologies enables whole brain imaging to better understand mind (Fig. 1).

- However, before that, a technical problem has to be solved: given colored volumetric images of the worm we need to identify the neurons, that is, assign a canonical label (a name) to each of them.

- One preliminary step is to build an atlas that represents the variability of each neuron in color and position coordinates.

Deep representation learning of biomolecular sequences to improve prediction of protein levels from genome-wide transcriptomics data

Daniel Shak[1], **Chao Wang**[2], **William E. Balch**[2], **Salvatore Loguercio**[2]

[1] UCSD School of Engineering
[2] Scripps Research

Accurate prediction of protein levels from widely available transcriptomics data is crucial for interpreting regulatory genetic variations in personal genomes and in genetic engineering for biotechnological or gene therapy applications. However, the extent to which mRNA levels are predictive of protein abundances remains debated. Current state-of-the-art prediction methods rely on several handcrafted features extracted from protein, RNA and DNA sequences but still suffer from unreliability and lack of precision. Millions of years of evolution have sampled the portions of the biomolecular sequence space that are relevant to life, so large unlabeled datasets of DNA, RNA and protein sequences are expected to contain significant biological information. Advances in natural language processing (NLP) have shown that self-supervised learning is a powerful tool for extracting information from unlabeled sequences, which raises a tantalizing question: can we adapt NLP-based and other contemporary representation learning techniques to extract useful biological information from massive sequence datasets? Herein, we are looking at representation learning of biomolecules as an enabler for studying predictive and generative techniques for biology, with particular regard to the prediction of protein levels from transcripts. Using a comprehensive proteome and transcriptome abundance atlas of 29 human tissues as benchmark, we implement and test several representation learning frameworks in order to automatically extract informative features from DNA, RNA and protein sequences that significantly improve the prediction of protein abundance from transcriptomics data. We found that representation learning of protein sequences is the most developed area thus far. We implemented recent work (i.e. UniRep embeddings, seq2seq autoencoder embeddings of protein sequences – both LSTM-based) finding that self-supervised pretraining is helpful for almost all models on all tasks, significantly improving performance in some cases. Despite this increase, in several cases features learned by self-supervised pretraining still lag behind features extracted by state-of-the-art non-neural techniques. This gap in performance suggests a huge opportunity for innovative architecture design and improved modeling paradigms that better capture the signal in protein sequences. When considering DNA and RNA sequences, the problem of adapting natural language-based models becomes even more complex and challenging. For instance, proteins sequences have defined lengths and coded in a 20-symbol alphabet of aminoacids; for nucleic acids, lengths are much harder to define given multiple (spliced) transcripts for RNA; intron-exon combinations for genes; and a plethora of regulatory regions for both – with all this complexity coded in a 4-symbol, repetitive nucleotide alphabet. Recently, NLP research has seen spectacular advances in the ability to model long-range interdependencies in text, with the introduction

of Transformers and other attention-based neural architectures (e.g. BERT, GPT 2 etc.). However, such abilities still pale in comparison of the long range interactions commonly observed in DNA, and to a lesser extent, RNA and protein sequences. These characteristics of biological sequences pose unique challenges for representation learning, but also hold the promise of a true transformative potential for biology and medicine – ultimately, the construction of an universal language base to fully describe the genotype to phenotype transformation.

Generative models for codon prediction and optimization

Samuel L. Goldman[1], David K Yang[2], Eli Weinstein[3], Debora S. Marks[4,5]

[1] MIT Computational and Systems Biology
[2] Harvard University
[3] Harvard University Department of Biophysics
[4] Harvard Medical School Department of Systems Biology
[5] Broad Institute of Harvard and MIT

Optimizing foreign DNA sequences for maximal protein production in a specied host organism is an important problem for synthetic biology and biomanufacturing. Experimental results have demonstrated that simply interchanging codons, triplets of three DNA bases, with "synonymous" alternatives can in fact amplify protein production several-fold while holding the produced protein constant. Previous methods for codon optimization are frequency based, which cannot consider factors such as RNA secondary structure that contribute to protein expression. Here, we apply a deep learning framework to model the distribution of codons in highly expressed bacterial and human transcripts. We show that our LSTM-Transducer model is able to predict the next codon of a genetic sequence with improved accuracy and lower perplexity on a held out set of transcripts, outperforming the previously state of the art frequency based approach and dening future codon-modeling challenges.

Learning Sparsely-Coupled Signaling Networks from Data

Matthew Karikomi, Qing Nie, UC Irvine

UC Irvine

Case studies on identifying bias and confounding in biomedical datasets

Subhashini Venugopalan[1], **Arunachalam Narayanaswamy**[1], **Samuel Yang**[1], **Anton Gerashcenko**[1], **Scott Lipnick**[2], **Nina Makhortova**[2], **James Hawrot**[3], **Christine Marques**[3], **Joao Pereira**[3], **Michael Brenner**[1,2], **Lee Rubin**[2], **Brian Wainger**[3], **Marc Berndl**[1]

[1] Google
[2] Harvard University
[3] Massachusetts General Hospital

COMET: Combinatorial Marker Detection from Single-Cell Transcriptomic Data

Louis Cammarata[3], **Conor Delaney**[2], **Alexandra Schnell**[1,4,6], **Aaron Yao-Smith**[5], **Aviv Regev**[6,7,8], **Vijay Kuchroo**[1,4,6] **and Meromit Singer**[1,2,6]

[1]Harvard Medical School
[2]Dana-Farber Cancer Institute
[3]Harvard University
[4]Brigham and Women's Hospital
[5]Cornell University
[6]Broad Institute
[7]Massachusetts Institute of Technology
[8]Howard Hughes Medical Institute

Reconstructing continuous distributions of 3D protein structure from cryo-EM images

Ellen Zhong, **Tristan Bepler**, **Joseph Davis**, **Bonnie Berger**

MIT

Partner Institutions and Sponsors

The Learning Meaningful Representations of Life Workshop is part of the Thirty-third Conference on Neural Information Processing Systems, funded by generous contributions by the NIH and GV, with in-kind support by the Broad Institute of Harvard and MIT.

Sponsors