

Incorporating biological structure into machine learning models in biomedicine

This manuscript ([permalink](#)) was automatically generated from [greenelab/biopriors-review@702ea42](#) on September 30, 2019.

Authors

- **Jake Crawford**

 [0000-0001-6207-0782](#) ·  [jjc2718](#) ·  [jjc2718](#)

Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA · Funded by Grant XXXXXXXX

- **Jane Roe**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [janeroe](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

When applying machine learning techniques to biomedical datasets, it can be challenging to distinguish signal from noise, particularly in the presence of limited amounts of data. Biological knowledge can take many forms, including genomic sequences, pathway databases, gene interaction networks, and knowledge hierarchies such as the Gene Ontology [1]. Incorporating these resources in machine learning models can be helpful in identifying patterns in noisy data [2] and in interpreting model predictions [3]. However, there is often no canonical way to encode these structures as real-valued predictors. This means modelers must be creative when deciding how to encode biological knowledge that they expect will be relevant to the task in models.

Biomedical datasets often contain more input predictors than data samples [4,5]. For example, a genetic study may genotype millions of single nucleotide polymorphisms (SNPs) in hundreds of patients, or a gene expression study may profile the expression of thousands of genes in only a handful of samples. Thus, it can be useful to include prior information describing the relationships between the predictors to inform the representation learned by the model. This stands in contrast to non-biological applications of machine learning, where one might fit a model on millions of images [6] or tens of thousands of documents [7], making inclusion of prior information unnecessary.

In this review, we survey approaches to learning models from biomedical data that incorporate external information about the structure of desirable solutions. One class of commonly used approaches involves using raw sequence data to learn a representation that considers the context of each base pair. For models that operate on gene input, such as gene expression data or genetic variants, it can be useful to incorporate networks or pathways describing relationships between genes. We also consider other examples in this review, such as neural network architectures that are constrained based on biological knowledge.

Sequence models

Early neural network models primarily used hand-engineered sequence features as input to a fully connected neural network [8,9]. As convolutional neural network (CNN) approaches matured for image processing and computer vision, researchers were able to use similar ideas to leverage biological sequence proximity in modeling. CNNs are a neural network variant in which input data are grouped by spatial context to extract features for prediction. The definition of “spatial context” is specific to the input. For example, for images pixels that are nearby in 2D space might be grouped, or for genomic sequences base pairs that are nearby in the linear genome might be grouped.

These approaches work by first encoding input into a numeric matrix (for DNA “one-hot” encoding is often used: A=[1,0,0,0], C=[0,1,0,0], G=[0,0,1,0], T=[0,0,0,1]). They then apply spatial filters to the encoded input, which are adjusted based on the data during model training. In this way, CNNs are able to consider context without making strong assumptions about exactly how much context is needed or how it should be encoded; the data informs the encoding. A detailed description of how CNNs are applied to sequence data can be found in [10].

Applications in regulatory biology

Many of the first applications of deep learning to biological sequence data were in regulatory biology. Example early applications of CNNs on sequence data include prediction of binding protein sequence

specificity from DNA or RNA sequence [11], prediction of variant effects from noncoding DNA sequence [12], and prediction of chromatin accessibility from DNA sequence [13].

Recent sequence models take advantage of hardware advances and methodological innovation to incorporate more sequence context and rely on fewer modeling assumptions. BPNet, a CNN used to predict transcription factor binding profiles from raw DNA sequences, was able to accurately map known locations of binding motifs in mouse embryonic stem cells [14]. The BPNet model considers 1000 base pairs of context around each position when predicting binding probabilities, using a technique called dilated convolutions [15] to make a large input field feasible. This context is particularly important because motif spacing and periodicity can influence protein function. cDeepbind [16] combines RNA sequences with information about secondary structure to predict RNA binding protein affinities. Its convolutional model acts on a feature vector combining sequence and structural information, simultaneously using context for both to inform predictions. APARENT [17] is a CNN used to predict alternative polyadenylation (APA) from a training set of over 3 million synthetic APA reporter sequences. These diverse applications underscore the power of modern deep learning models to synthesize large sequence datasets.

Models that consider sequence context have also been applied to impute and make predictions from epigenetic data. DeepSignal [18] is a CNN that uses contextual electrical signals from Oxford Nanopore single-molecule sequencing data to predict 5mC or 6mA DNA methylation status. MRCNN [19] uses sequences of length 400, centered at CpG sites, to predict 5mC methylation status. Deep learning models have also been used to predict gene expression from histone modifications [20,21]. Here, a neural network model consisting of long short-term memory (LSTM) units was used to encode the long-distance interactions of histone marks in both the 3' and 5' genomic directions. In each of these cases, proximity in the linear genome proved useful for modeling the complex interactions between DNA sequence and epigenome.

Applications in variant calling and mutation detection

Identification of genetic variants can also benefit from models that take into account sequence context. DeepVariant [22] applies a CNN to images of sequence read pileups, using read data around each candidate variant to accurately distinguish true variants from sequencing errors. CNNs have also been applied to single molecule (PacBio and Oxford Nanopore) sequencing data [23], using a different sequence encoding that results in better performance than DeepVariant on single molecule data. However, many variant calling models still use hand-engineered sequence features as input to a classifier, including current state-of-the-art approaches to insertion/deletion calling [24,25]. Detection of somatic mutations is a distinct but related challenge to detection of germline variants, and has also recently benefitted from use of CNN models [26].

Applications in CRISPR guide selection

With the recent rise in popularity of CRISPR gene editing technology, sequence models have proven useful in improving design of single-guide RNAs (sgRNAs). To select the best sgRNA from multiple possibilities, one might be interested in balancing on-target efficiency with likelihood of off-target effects, both of which depend on the sgRNA sequence and its genomic context. Early models for prediction of on-target cleavage efficiency [27] and off-target effects [28] used hand-engineered sequence features, in addition to other RNA-specific information such as thermodynamic sequence properties. More recently, CNN-based models have demonstrated improved performance using data-derived sequence features. CNNs have been successfully applied to CRISPR-Cas9 on-target efficiency prediction [29,30], CRISPR-Cas12a (Cpf1) on-target efficiency prediction [31], and CRISPR-Cas9 off-target effect prediction [29,32]. In each case, the authors show that using a CNN to operate on raw sequence data improves sgRNA design relative to models that use hand-engineered sequence features as input.

Network- and pathway-based models

Rather than operating on raw DNA sequence, many machine learning models in biomedicine operate on inputs without an intrinsic order. For instance, models may make use of gene expression data matrices (e.g. from RNA sequencing or microarray experiments), where each row represents a sample and each column a gene. When modeling data that is indexed by gene, one might incorporate knowledge that describes the relationships or correlations between genes, in an effort to take these relationships into account when making predictions or generating a low-dimensional representation of the data. This is comparable to the manner in which sequence context encourages models to consider nearby base pairs similarly.

Applications in transcriptomics

Models that take gene expression data as input can benefit from incorporating gene-level relationships. One form that this knowledge commonly takes is a database of gene sets, which may represent biological pathways or gene signatures related to a biological state of interest. PLIER [\[33\]](#) uses gene set information from MSigDB [\[34\]](#) and cell type markers to extract a representation of whole blood gene expression data that corresponds to biological processes and reduces technical noise. The resulting gene set-aligned representation was used to perform accurate cell type mixture decomposition. MultiPLIER [\[35\]](#) applied the PLIER framework to the recount2 gene expression compendium [\[36\]](#) to develop a model that shares information across multiple tissues and diseases, including rare diseases with limited sample sizes. PASNet [\[37\]](#) uses data from MSigDB to inform the structure of a neural network for predicting patient outcomes in glioblastoma multiforme (GBM) from gene expression data. This approach has the added benefit of straightforward interpretation, as pathway nodes in the network having high weights can be inferred to correspond to important pathways in GBM outcome prediction.

Alternatively, gene-level relationships can take the form of a network. Nodes in these networks typically represent genes, and real-valued edges in these networks may represent interactions or correlations between genes, often in a tissue or cell type context of interest. netNMF-sc [\[38\]](#) incorporates coexpression networks [\[39\]](#) as a smoothing term to perform dimension reduction and impute dropouts in single-cell gene expression data. The authors show that using a coexpression network to extract a low-dimensional representation increases performance for cell type identification and identification of cell cycle marker genes, both as compared to using raw gene expression data and as compared to other single-cell dimension reduction methods. Combining gene expression data with a network-derived smoothing term has also been shown to improve performance at predicting patient drug response in acute myeloid leukemia [\[40\]](#) and at detecting mutated cancer genes [\[41\]](#). PIMKL [\[42\]](#) combines network and pathway data to predict disease-free survival from breast cancer cohorts. This method takes as input both RNA-seq gene expression data and copy number alteration data, but can be applied to gene expression data alone as well.

Gene regulatory networks can also augment models for gene expression data. These networks describe how the expression of genes is modulated by biological regulators such as transcription factors, microRNAs, or small molecules. creNET [\[43\]](#) integrates a gene regulatory network, derived from STRING [\[44\]](#), with a sparse logistic regression model to predict phenotypic response in clinical trials for ulcerative colitis and acute kidney rejection based on gene expression data. The gene regulatory information allows the model to identify the biological regulators that are associated with the response, potentially giving mechanistic insight into differential clinical trial response. GRRANN [\[45\]](#) uses a gene regulatory network to inform the structure of a neural network, applying it to the same clinical trial data as creNET. Several other methods [\[46,47\]](#) have also used gene regulatory network structure to constrain the structure of a neural network, reducing the number of parameters to be fit by the network and facilitating interpretation of network predictions.

Applications in genetics

Approaches to incorporating gene set or network structure into genetic studies have a long history (see, e.g. [48,49]). Recent applications of these methods include expression quantitative trait loci (eQTL) mapping studies, which aim to identify associations between genetic variants and gene expression. netReg [50] implements the graph-regularized dual LASSO algorithm for eQTL mapping described in [51] in a publicly available R package, based on an efficient C++ backend. This model smooths regression coefficients simultaneously based on networks describing associations between genes (target variables in the eQTL regression model) and between variants (predictors in the eQTL regression model). eQTL information can also be used in conjunction with genetic variant information to predict phenotypes, in an approach known as Mendelian randomization (MR). In [52], a smoothing term derived from a gene regulatory network is used as a component in an MR model. The model with the network smoothing term, applied to a human liver data set, more robustly identifies genes that influence enzyme activity than an MR model that does not consider network interactions. As genetic datasets become larger, efficient methods for gene and genetic variant selection will become even more important, and it is likely that researchers will continue to develop models that leverage gene set and network databases.

Other models incorporating biological structure

Knowledge about biological entities is often organized in an ontology, which is a directed graph that encodes the relationships between entities. The Gene Ontology (GO) [1] describes the relationships between cellular subsystems and other attributes describing proteins or genes. DCell [53] uses GO to inform the connectivity of a neural network predicting the effects of gene deletions on yeast growth. The authors show that DCell performs comparably to an unconstrained neural network for this task. Additionally, it has the advantage of interpretability: a cellular subsystem with high neuron output values under a particular gene deletion can be inferred to be strongly affected by the gene deletion, providing a putative genotype-phenotype association. DeepGO [54] uses a similar approach to predict protein function from amino acid sequence, using a neural network constrained by the dependencies of GO. However, a follow-up paper by the same authors [55] showed that this hierarchy-aware approach can be outperformed by a hierarchy-naïve CNN, which uses only raw amino acid sequence and similarity to labeled proteins in the training set. This suggests a potential tradeoff between interpretability and predictive accuracy in the context of protein function prediction.

Phylogenetic trees, or hierarchies describing the evolutionary relationships between species, can be useful for a similar purpose. glmmTree [56] uses a phylogenetic tree describing the relationship between microorganisms to improve predictions of age based on gut microbiome sequencing data. The same authors combine a similar phylogeny smoothing strategy with a sparse regression model in [57], showing its utility to model caffeine intake and smoking status based on microbiome data. Phylogenetic trees can also be useful in describing the relationships between subclones of a tumor, which are fundamental to understanding cancer evolution and development. In [58], the authors use a tumor phylogeny inferred from copy number aberration (CNA) sequencing data as a smoothing term in an algorithm for deconvolving tumor subclones, providing more robust predictions than a phylogeny-free model. The tree structure of the phylogeny and the subclone mixture model are fit jointly to the CNA data.

Depending on the application, incorporating other forms of structure or prior knowledge can help to inform predictions and interpretation of the model's output. CYCLOPS [59] uses a circular node autoencoder [60] to order periodic gene expression data and measure circadian rhythms. The authors validated the method by correctly ordering samples without temporal labels and identifying genes with known circadian expression. They then applied it to compare gene expression in normal and cancerous liver biopsies, identifying drug targets with circadian expression as candidates for chronotherapy. NetBiTE [61] uses drug-gene interaction information from GDSC [62], in addition to

protein interaction data, to build a tree ensemble model with splits that are biased toward high-confidence drug-gene interactions. The model is used to predict sensitivity to drugs that inhibit critical signaling pathways in cancer, showing improved predictive performance compared to random forests, another commonly used tree ensemble model.

Conclusions

1. What is outside of the scope of this review? (this can also go in introduction?)
 - a. Biological “constraints” vs. feature selection or feature extraction from heterogeneous biological data (e.g. network embedding approaches)
 - Example: one could use a network-based feature extraction method (e.g. Node2Vec) to convert each gene in a PPI network into a set of real-valued features, then use those + gene expression as input to a model
 - For purposes of keeping this review short enough, I’m trying to stay away from papers like [^], but still unclear to me where exactly the line should be drawn. Almost any ML model that operates on sequence data can be viewed as having a feature extraction component, for example.
 - b. Could kind-of consider many single-cell dimension reduction methods biologically constrained (e.g. dropout/zero inflation modeling approaches, etc), but this is way too broad for this review - maybe refer the reader to other recent reviews of these methods.
 - c. Could also consider omics integration methods (combining, for example, gene expression and epigenetic data) to be biologically constrained, but we refer the reader to [10.1016/j.inffus.2018.09.012](https://doi.org/10.1016/j.inffus.2018.09.012) for further detail on these methods.

References

1. **The Gene Ontology Resource: 20 years and still GOing strong***Nucleic Acids Research* (2018-11-05) <https://doi.org/gf63mb>
DOI: [10.1093/nar/gky1055](https://doi.org/10.1093/nar/gky1055) · PMID: [30395331](https://pubmed.ncbi.nlm.nih.gov/30395331/) · PMCID: [PMC6323945](https://pubmed.ncbi.nlm.nih.gov/PMC6323945/)
2. **Network propagation: a universal amplifier of genetic associations**
Lenore Cowen, Trey Ideker, Benjamin J. Raphael, Roded Sharan
Nature Reviews Genetics (2017-06-12) <https://doi.org/gbhkwn>
DOI: [10.1038/nrg.2017.38](https://doi.org/10.1038/nrg.2017.38) · PMID: [28607512](https://pubmed.ncbi.nlm.nih.gov/28607512/)
3. **Visible Machine Learning for Biomedicine**
Michael K. Yu, Jianzhu Ma, Jasmin Fisher, Jason F. Kreisberg, Benjamin J. Raphael, Trey Ideker
Cell (2018-06) <https://doi.org/gdqcd8>
DOI: [10.1016/j.cell.2018.05.056](https://doi.org/10.1016/j.cell.2018.05.056) · PMID: [29906441](https://pubmed.ncbi.nlm.nih.gov/29906441/) · PMCID: [PMC6483071](https://pubmed.ncbi.nlm.nih.gov/PMC6483071/)
4. **Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets**
Michael K. K. Leung, Andrew Delong, Babak Alipanahi, Brendan J. Frey
Proceedings of the IEEE (2016-01) <https://doi.org/f75grb>
DOI: [10.1109/jproc.2015.2494198](https://doi.org/10.1109/jproc.2015.2494198)
5. **Diet Networks: Thin Parameters for Fat Genomics**
Adriana Romero, Pierre Luc Carrier, Akram Erraqabi, Tristan Sylvain, Alex Auvolat, Etienne Dejoie, Marc-André Legault, Marie-Pierre Dubé, Julie G. Hussin, Yoshua Bengio
arXiv (2016-11-28) <https://arxiv.org/abs/1611.09340v3>
6. **ImageNet: A large-scale hierarchical image database**
Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei
2009 IEEE Conference on Computer Vision and Pattern Recognition (2009-06) <https://doi.org/cvc7xp>
DOI: [10.1109/cvpr.2009.5206848](https://doi.org/10.1109/cvpr.2009.5206848)
7. **NewsWeeder: Learning to Filter Netnews**
Ken Lang
Machine Learning Proceedings 1995 (1995) <https://doi.org/gf63mh>
DOI: [10.1016/b978-1-55860-377-6.50048-7](https://doi.org/10.1016/b978-1-55860-377-6.50048-7)
8. **DEEP: a general computational framework for predicting enhancers**
Dimitrios Kleftogiannis, Panos Kalnis, Vladimir B. Bajic
Nucleic Acids Research (2014-11-05) <https://doi.org/gcggk83>
DOI: [10.1093/nar/gku1058](https://doi.org/10.1093/nar/gku1058) · PMID: [25378307](https://pubmed.ncbi.nlm.nih.gov/25378307/) · PMCID: [PMC4288148](https://pubmed.ncbi.nlm.nih.gov/PMC4288148/)
9. **The human splicing code reveals new insights into the genetic determinants of disease**
H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, ... B. J. Frey
Science (2014-12-18) <https://doi.org/f6wzj2>
DOI: [10.1126/science.1254806](https://doi.org/10.1126/science.1254806) · PMID: [25525159](https://pubmed.ncbi.nlm.nih.gov/25525159/) · PMCID: [PMC4362528](https://pubmed.ncbi.nlm.nih.gov/PMC4362528/)
10. **Deep learning for computational biology**
Christof Angermueller, Tanel Pärnamaa, Leopold Parts, Oliver Stegle
Molecular Systems Biology (2016-07) <https://doi.org/f8xtvh>
DOI: [10.15252/msb.20156651](https://doi.org/10.15252/msb.20156651) · PMID: [27474269](https://pubmed.ncbi.nlm.nih.gov/27474269/) · PMCID: [PMC4965871](https://pubmed.ncbi.nlm.nih.gov/PMC4965871/)

11. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey

Nature Biotechnology (2015-07-27) <https://doi.org/f7mkrd>

DOI: [10.1038/nbt.3300](https://doi.org/10.1038/nbt.3300) · PMID: [26213851](https://pubmed.ncbi.nlm.nih.gov/26213851/)

12. Predicting effects of noncoding variants with deep learning-based sequence model

Jian Zhou, Olga G Troyanskaya

Nature Methods (2015-08-24) <https://doi.org/gcggk8g>

DOI: [10.1038/nmeth.3547](https://doi.org/10.1038/nmeth.3547) · PMID: [26301843](https://pubmed.ncbi.nlm.nih.gov/26301843/) · PMCID: [PMC4768299](https://pubmed.ncbi.nlm.nih.gov/PMC4768299/)

13. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

David R. Kelley, Jasper Snoek, John L. Rinn

Genome Research (2016-05-03) <https://doi.org/f8sw35>

DOI: [10.1101/gr.200535.115](https://doi.org/10.1101/gr.200535.115) · PMID: [27197224](https://pubmed.ncbi.nlm.nih.gov/27197224/) · PMCID: [PMC4937568](https://pubmed.ncbi.nlm.nih.gov/PMC4937568/)

14. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Amr Alexandari, Sabrina Krueger, Khyati Dalal, Robin Froepf, Charles McAnany, Julien Gagneur, Anshul Kundaje, Julia Zeitlinger

Cold Spring Harbor Laboratory (2019-08-21) <https://doi.org/gf64fc>

DOI: [10.1101/737981](https://doi.org/10.1101/737981)

15. Multi-Scale Context Aggregation by Dilated Convolutions

Fisher Yu, Vladlen Koltun

arXiv (2015-11-23) <https://arxiv.org/abs/1511.07122v3>

16. cDeepbind: A context sensitive deep learning model of RNA-protein binding

Shreshth Gandhi, Leo J. Lee, Andrew Delong, David Duvenaud, Brendan J. Frey

Cold Spring Harbor Laboratory (2018-06-12) <https://doi.org/gf68nk>

DOI: [10.1101/345140](https://doi.org/10.1101/345140)

17. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation

Nicholas Bogard, Johannes Linder, Alexander B. Rosenberg, Georg Seelig

Cell (2019-06) <https://doi.org/gf66nc>

DOI: [10.1016/j.cell.2019.04.046](https://doi.org/10.1016/j.cell.2019.04.046) · PMID: [31178116](https://pubmed.ncbi.nlm.nih.gov/31178116/) · PMCID: [PMC6599575](https://pubmed.ncbi.nlm.nih.gov/PMC6599575/)

18. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning

Peng Ni, Neng Huang, Zhi Zhang, De-Peng Wang, Fan Liang, Yu Miao, Chuan-Le Xiao, Feng Luo, Jianxin Wang

Bioinformatics (2019-04-17) <https://doi.org/gf66qw>

DOI: [10.1093/bioinformatics/btz276](https://doi.org/10.1093/bioinformatics/btz276) · PMID: [30994904](https://pubmed.ncbi.nlm.nih.gov/30994904/)

19. MRCNN: a deep learning model for regression of genome-wide DNA methylation

Qi Tian, Jianxiao Zou, Jianxiong Tang, Yuan Fang, Zhongli Yu, Shicai Fan

BMC Genomics (2019-04) <https://doi.org/gf48g6>

DOI: [10.1186/s12864-019-5488-5](https://doi.org/10.1186/s12864-019-5488-5) · PMID: [30967120](https://pubmed.ncbi.nlm.nih.gov/30967120/) · PMCID: [PMC6457069](https://pubmed.ncbi.nlm.nih.gov/PMC6457069/)

20. Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin

Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi

Cold Spring Harbor Laboratory (2018-05-25) <https://doi.org/gf66qz>

DOI: [10.1101/329334](https://doi.org/10.1101/329334)

21. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications

Arshdeep Sekhon, Ritambhara Singh, Yanjun Qi

Bioinformatics (2018-09-01) <https://doi.org/gd9mk4>

DOI: [10.1093/bioinformatics/bty612](https://doi.org/10.1093/bioinformatics/bty612) · PMID: [30423076](https://pubmed.ncbi.nlm.nih.gov/30423076/)

22. A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, ... Mark A DePristo

Nature Biotechnology (2018-09-24) <https://doi.org/gd8gkf>

DOI: [10.1038/nbt.4235](https://doi.org/10.1038/nbt.4235) · PMID: [30247488](https://pubmed.ncbi.nlm.nih.gov/30247488/)

23. A multi-task convolutional deep neural network for variant calling in single molecule sequencing

Ruibang Luo, Fritz J. Sedlazeck, Tak-Wah Lam, Michael C. Schatz

Nature Communications (2019-03-01) <https://doi.org/gf4c37>

DOI: [10.1038/s41467-019-09025-z](https://doi.org/10.1038/s41467-019-09025-z) · PMID: [30824707](https://pubmed.ncbi.nlm.nih.gov/30824707/) · PMCID: [PMC6397153](https://pubmed.ncbi.nlm.nih.gov/PMC6397153/)

24. SICaRiO: Short Indel Call filteriNg with bOosting

Md Shariful Islam Bhuyan, Itsik Pe'er, M. Sohel Rahman

Cold Spring Harbor Laboratory (2019-04-07) <https://doi.org/gf68d6>

DOI: [10.1101/601450](https://doi.org/10.1101/601450)

25. Machine learning-based detection of insertions and deletions in the human genome

Charles Curnin, Rachel L. Goldfeder, Shruti Marwaha, Devon Bonner, Daryl Waggott, Matthew T. Wheeler, Euan A. Ashley,

Cold Spring Harbor Laboratory (2019-05-05) <https://doi.org/gf68d7>

DOI: [10.1101/628222](https://doi.org/10.1101/628222)

26. Deep convolutional neural networks for accurate somatic mutation detection

Sayed Mohammad Ebrahim Sahraeian, Ruolin Liu, Bayo Lau, Karl Podesta, Marghoob Mohiyuddin, Hugo Y. K. Lam

Nature Communications (2019-03-04) <https://doi.org/gf68f8>

DOI: [10.1038/s41467-019-09027-x](https://doi.org/10.1038/s41467-019-09027-x) · PMID: [30833567](https://pubmed.ncbi.nlm.nih.gov/30833567/) · PMCID: [PMC6399298](https://pubmed.ncbi.nlm.nih.gov/PMC6399298/)

27. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9

John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, ... David E Root

Nature Biotechnology (2016-01-18) <https://doi.org/f79twf>

DOI: [10.1038/nbt.3437](https://doi.org/10.1038/nbt.3437) · PMID: [26780180](https://pubmed.ncbi.nlm.nih.gov/26780180/) · PMCID: [PMC4744125](https://pubmed.ncbi.nlm.nih.gov/PMC4744125/)

28. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

Jennifer Listgarten, Michael Weinstein, Benjamin P. Kleinstiver, Alexander A. Sousa, J. Keith Joung, Jake Crawford, Kevin Gao, Luong Hoang, Melih Elibol, John G. Doench, Nicolo Fusi

Nature Biomedical Engineering (2018-01) <https://doi.org/gf7hmk>

DOI: [10.1038/s41551-017-0178-6](https://doi.org/10.1038/s41551-017-0178-6) · PMID: [29998038](https://pubmed.ncbi.nlm.nih.gov/29998038/) · PMCID: [PMC6037314](https://pubmed.ncbi.nlm.nih.gov/PMC6037314/)

29. DeepCRISPR: optimized CRISPR guide RNA design by deep learning

Guohui Chuai, Hanhui Ma, Jifang Yan, Ming Chen, Nanfang Hong, Dongyu Xue, Chi Zhou, Chenyu Zhu, Ke Chen, Bin Duan, ... Qi Liu

Genome Biology (2018-06-26) <https://doi.org/gdshcw>

DOI: [10.1186/s13059-018-1459-4](https://doi.org/10.1186/s13059-018-1459-4) · PMID: [29945655](https://pubmed.ncbi.nlm.nih.gov/29945655/) · PMCID: [PMC6020378](https://pubmed.ncbi.nlm.nih.gov/PMC6020378/)

30. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with unparalleled generalization performance

Hui Kwon Kim, Younggwang Kim, Sungtae Lee, Seonwoo Min, Jung Yoon Bae, Jae Woo Choi, Jinman Park, Dongmin Jung, Sungroh Yoon, Hyongbum Henry Kim
Cold Spring Harbor Laboratory (2019-05-15) <https://doi.org/gf7hnp>
DOI: [10.1101/636472](https://doi.org/10.1101/636472)

31. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity

Hui Kwon Kim, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim, Sangeun Lee, Sungroh Yoon, Hyongbum Kim
Nature Biotechnology (2018-01-29) <https://doi.org/gc6xvx>
DOI: [10.1038/nbt.4061](https://doi.org/10.1038/nbt.4061) · PMID: [29431740](https://pubmed.ncbi.nlm.nih.gov/29431740/)

32. Off-target predictions in CRISPR-Cas9 gene editing using deep learning

Jiecong Lin, Ka-Chun Wong
Bioinformatics (2018-09-01) <https://doi.org/gd9m4f>
DOI: [10.1093/bioinformatics/bty554](https://doi.org/10.1093/bioinformatics/bty554) · PMID: [30423072](https://pubmed.ncbi.nlm.nih.gov/30423072/) · PMCID: [PMC6129261](https://pubmed.ncbi.nlm.nih.gov/PMC6129261/)

33. Pathway-level information extractor (PLIER) for gene expression data

Weiguang Mao, Elena Zaslavsky, Boris M. Hartmann, Stuart C. Sealfon, Maria Chikina
Nature Methods (2019-06-27) <https://doi.org/gf75g6>
DOI: [10.1038/s41592-019-0456-1](https://doi.org/10.1038/s41592-019-0456-1) · PMID: [31249421](https://pubmed.ncbi.nlm.nih.gov/31249421/)

34. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov
Proceedings of the National Academy of Sciences (2005-09-30) <https://doi.org/d4qbh8>
DOI: [10.1073/pnas.0506580102](https://doi.org/10.1073/pnas.0506580102) · PMID: [16199517](https://pubmed.ncbi.nlm.nih.gov/16199517/) · PMCID: [PMC1239896](https://pubmed.ncbi.nlm.nih.gov/PMC1239896/)

35. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease

Jaclyn N. Taroni, Peter C. Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A. Merkel, Casey S. Greene
Cell Systems (2019-05) <https://doi.org/gf75g5>
DOI: [10.1016/j.cels.2019.04.003](https://doi.org/10.1016/j.cels.2019.04.003) · PMID: [31121115](https://pubmed.ncbi.nlm.nih.gov/31121115/) · PMCID: [PMC6538307](https://pubmed.ncbi.nlm.nih.gov/PMC6538307/)

36. Reproducible RNA-seq analysis using recount2

Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, Jeffrey T Leek
Nature Biotechnology (2017-04) <https://doi.org/gf75hp>
DOI: [10.1038/nbt.3838](https://doi.org/10.1038/nbt.3838) · PMID: [28398307](https://pubmed.ncbi.nlm.nih.gov/28398307/) · PMCID: [PMC6742427](https://pubmed.ncbi.nlm.nih.gov/PMC6742427/)

37. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data

Jie Hao, Youngsoon Kim, Tae-Kyung Kim, Mingon Kang
BMC Bioinformatics (2018-12) <https://doi.org/gf75g9>
DOI: [10.1186/s12859-018-2500-z](https://doi.org/10.1186/s12859-018-2500-z) · PMID: [30558539](https://pubmed.ncbi.nlm.nih.gov/30558539/) · PMCID: [PMC6296065](https://pubmed.ncbi.nlm.nih.gov/PMC6296065/)

38. netNMF-sc: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis

Rebecca Elyanow, Bianca Dumitrascu, Barbara E. Engelhardt, Benjamin J. Raphael

Cold Spring Harbor Laboratory (2019-02-08) <https://doi.org/gf386x>
DOI: [10.1101/544346](https://doi.org/10.1101/544346)

39. COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH)

Sunmo Yang, Chan Yeong Kim, Sohyun Hwang, Eiru Kim, Hyojin Kim, Hongseok Shim, Insuk Lee
Nucleic Acids Research (2016-09-26) <https://doi.org/f9v38k>
DOI: [10.1093/nar/gkw868](https://doi.org/10.1093/nar/gkw868) · PMID: [27679477](https://pubmed.ncbi.nlm.nih.gov/27679477/) · PMCID: [PMC5210615](https://pubmed.ncbi.nlm.nih.gov/PMC5210615/)

40. Learning Explainable Models Using Attribution Priors

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, Su-In Lee
arXiv (2019-06-25) <https://arxiv.org/abs/1906.10670v1>

41. A novel network regularized matrix decomposition method to detect mutated cancer genes in tumour samples with inter-patient heterogeneity

Jianing Xi, Ao Li, Minghui Wang
Scientific Reports (2017-06-06) <https://doi.org/gcq9jZ>
DOI: [10.1038/s41598-017-03141-w](https://doi.org/10.1038/s41598-017-03141-w) · PMID: [28588243](https://pubmed.ncbi.nlm.nih.gov/28588243/) · PMCID: [PMC5460199](https://pubmed.ncbi.nlm.nih.gov/PMC5460199/)

42. PIMKL: Pathway-Induced Multiple Kernel Learning

Matteo Manica, Joris Cadow, Roland Mathis, María Rodríguez Martínez
npj Systems Biology and Applications (2019-03-05) <https://doi.org/gf8ck6>
DOI: [10.1038/s41540-019-0086-3](https://doi.org/10.1038/s41540-019-0086-3) · PMID: [30854223](https://pubmed.ncbi.nlm.nih.gov/30854223/) · PMCID: [PMC6401099](https://pubmed.ncbi.nlm.nih.gov/PMC6401099/)

43. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes

Kourosh Zarringhalam, David Degras, Christoph Brockel, Daniel Ziemek
Scientific Reports (2018-01-19) <https://doi.org/gcwzdn>
DOI: [10.1038/s41598-018-19635-0](https://doi.org/10.1038/s41598-018-19635-0) · PMID: [29352257](https://pubmed.ncbi.nlm.nih.gov/29352257/) · PMCID: [PMC5775343](https://pubmed.ncbi.nlm.nih.gov/PMC5775343/)

44. STRING v10: protein-protein interaction networks, integrated over the tree of life

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, ... Christian von Mering
Nucleic Acids Research (2014-10-28) <https://doi.org/f64rfn>
DOI: [10.1093/nar/gku1003](https://doi.org/10.1093/nar/gku1003) · PMID: [25352553](https://pubmed.ncbi.nlm.nih.gov/25352553/) · PMCID: [PMC4383874](https://pubmed.ncbi.nlm.nih.gov/PMC4383874/)

45. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data

Tianyu Kang, Wei Ding, Luoyan Zhang, Daniel Ziemek, Kourosh Zarringhalam
BMC Bioinformatics (2017-12) <https://doi.org/gf8cm6>
DOI: [10.1186/s12859-017-1984-2](https://doi.org/10.1186/s12859-017-1984-2) · PMID: [29258445](https://pubmed.ncbi.nlm.nih.gov/29258445/) · PMCID: [PMC5735940](https://pubmed.ncbi.nlm.nih.gov/PMC5735940/)

46. Using neural networks for reducing the dimensions of single-cell RNA-Seq data

Chieh Lin, Siddhartha Jain, Hannah Kim, Ziv Bar-Joseph
Nucleic Acids Research (2017-07-31) <https://doi.org/gcnzb7>
DOI: [10.1093/nar/gkx681](https://doi.org/10.1093/nar/gkx681) · PMID: [28973464](https://pubmed.ncbi.nlm.nih.gov/28973464/) · PMCID: [PMC5737331](https://pubmed.ncbi.nlm.nih.gov/PMC5737331/)

47. Genetic Neural Networks: an artificial neural network architecture for capturing gene expression relationships

Ameen Eetemadi, Ilias Tagkopoulos
Bioinformatics (2018-11-19) <https://doi.org/gfnks6>
DOI: [10.1093/bioinformatics/bty945](https://doi.org/10.1093/bioinformatics/bty945) · PMID: [30452523](https://pubmed.ncbi.nlm.nih.gov/30452523/)

48. Nonparametric pathway-based regression models for analysis of genomic data

Z. Wei, H. Li

Biostatistics (2006-06-13) <https://doi.org/fdmggm>

DOI: [10.1093/biostatistics/kxl007](https://doi.org/10.1093/biostatistics/kxl007) · PMID: [16772399](https://pubmed.ncbi.nlm.nih.gov/16772399/)

49. Network-constrained regularization and variable selection for analysis of genomic data

C. Li, H. Li

Bioinformatics (2008-03-01) <https://doi.org/fk8n4b>

DOI: [10.1093/bioinformatics/btn081](https://doi.org/10.1093/bioinformatics/btn081) · PMID: [18310618](https://pubmed.ncbi.nlm.nih.gov/18310618/)

50. netReg: network-regularized linear models for biological association studies

Simon Dirmeyer, Christiane Fuchs, Nikola S Mueller, Fabian J Theis

Bioinformatics (2017-10-25) <https://doi.org/gcg9xq>

DOI: [10.1093/bioinformatics/btx677](https://doi.org/10.1093/bioinformatics/btx677) · PMID: [29077797](https://pubmed.ncbi.nlm.nih.gov/29077797/) · PMCID: [PMC6030897](https://pubmed.ncbi.nlm.nih.gov/PMC6030897/)

51. Graph-regularized dual Lasso for robust eQTL mapping

Wei Cheng, Xiang Zhang, Zhishan Guo, Yu Shi, Wei Wang

Bioinformatics (2014-06-11) <https://doi.org/f58j6m>

DOI: [10.1093/bioinformatics/btu293](https://doi.org/10.1093/bioinformatics/btu293) · PMID: [24931977](https://pubmed.ncbi.nlm.nih.gov/24931977/) · PMCID: [PMC4058913](https://pubmed.ncbi.nlm.nih.gov/PMC4058913/)

52. Integrative analysis of genetical genomics data incorporating network structures

Bin Gao, Xu Liu, Hongzhe Li, Yuehua Cui

Biometrics (2019-04-29) <https://doi.org/gf8f9q>

DOI: [10.1111/biom.13072](https://doi.org/10.1111/biom.13072) · PMID: [31009063](https://pubmed.ncbi.nlm.nih.gov/31009063/)

53. Using deep learning to model the hierarchical structure and function of a cell

Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, Trey Ideker

Nature Methods (2018-03-05) <https://doi.org/gc46jp>

DOI: [10.1038/nmeth.4627](https://doi.org/10.1038/nmeth.4627) · PMID: [29505029](https://pubmed.ncbi.nlm.nih.gov/29505029/) · PMCID: [PMC5882547](https://pubmed.ncbi.nlm.nih.gov/PMC5882547/)

54. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier

Maxat Kulmanov, Mohammed Asif Khan, Robert Hoehndorf

Bioinformatics (2017-10-03) <https://doi.org/gc3nb8>

DOI: [10.1093/bioinformatics/btx624](https://doi.org/10.1093/bioinformatics/btx624) · PMID: [29028931](https://pubmed.ncbi.nlm.nih.gov/29028931/) · PMCID: [PMC5860606](https://pubmed.ncbi.nlm.nih.gov/PMC5860606/)

55. DeepGOPlus: improved protein function prediction from sequence

Maxat Kulmanov, Robert Hoehndorf

Bioinformatics (2019-07-27) <https://doi.org/gf84d8>

DOI: [10.1093/bioinformatics/btz595](https://doi.org/10.1093/bioinformatics/btz595) · PMID: [31350877](https://pubmed.ncbi.nlm.nih.gov/31350877/)

56. Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model

Jian Xiao, Li Chen, Stephen Johnson, Yue Yu, Xianyang Zhang, Jun Chen

Frontiers in Microbiology (2018-06-27) <https://doi.org/gdtz4z>

DOI: [10.3389/fmicb.2018.01391](https://doi.org/10.3389/fmicb.2018.01391) · PMID: [29997602](https://pubmed.ncbi.nlm.nih.gov/29997602/) · PMCID: [PMC6030386](https://pubmed.ncbi.nlm.nih.gov/PMC6030386/)

57. A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data

Jian Xiao, Li Chen, Yue Yu, Xianyang Zhang, Jun Chen

Frontiers in Microbiology (2018-12-19) <https://doi.org/gf8qcc>

DOI: [10.3389/fmicb.2018.03112](https://doi.org/10.3389/fmicb.2018.03112) · PMID: [30619188](https://pubmed.ncbi.nlm.nih.gov/30619188/) · PMCID: [PMC6305753](https://pubmed.ncbi.nlm.nih.gov/PMC6305753/)

58. Tumor Copy Number Deconvolution Integrating Bulk and Single-Cell Sequencing Data

Haoyun Lei, Bochuan Lyu, E. Michael Gertz, Alejandro A. Schäffer, Xulian Shi, Kui Wu, Guibo Li, Liqin Xu, Yong Hou, Michael Dean, Russell Schwartz

Lecture Notes in Computer Science (2019) <https://doi.org/gf8qck>

DOI: [10.1007/978-3-030-17083-7_11](https://doi.org/10.1007/978-3-030-17083-7_11)

59. CYCLOPS reveals human transcriptional rhythms in health and disease

Ron C. Anafi, Lauren J. Francey, John B. Hogenesch, Junhyong Kim

Proceedings of the National Academy of Sciences (2017-04-24) <https://doi.org/f9796k>

DOI: [10.1073/pnas.1619320114](https://doi.org/10.1073/pnas.1619320114) · PMID: [28439010](https://pubmed.ncbi.nlm.nih.gov/28439010/) · PMCID: [PMC5441789](https://pubmed.ncbi.nlm.nih.gov/PMC5441789/)

60. Circular Nodes in Neural Networks

Michael J. Kirby, Rick Miranda

Neural Computation (1996-02-15) <https://doi.org/ffcww8>

DOI: [10.1162/neco.1996.8.2.390](https://doi.org/10.1162/neco.1996.8.2.390)

61. Network-based Biased Tree Ensembles (NetBiTE) for Drug Sensitivity Prediction and Drug Sensitivity Biomarker Identification in Cancer

Ali Oskooei, Matteo Manica, Roland Mathis, Maria Rodriguez Martinez

arXiv (2018-08-18) <https://arxiv.org/abs/1808.06603v2>

62. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, ... Mathew J. Garnett

Nucleic Acids Research (2012-11-22) <https://doi.org/f4jrdn>

DOI: [10.1093/nar/gks1111](https://doi.org/10.1093/nar/gks1111) · PMID: [23180760](https://pubmed.ncbi.nlm.nih.gov/23180760/) · PMCID: [PMC3531057](https://pubmed.ncbi.nlm.nih.gov/PMC3531057/)