AI for Genomics: from CNNs and LSTMs to Transformers

Nikolay Oskolkov, Group Leader (PI) at LIOS, Riga, Latvia

Physalia course, 09.09.2025

@NikolayOskolkov

@oskolkov.bsky.social

Personal homepage:
https://nikolay-oskolkov.com

Image generated at www.canva.com

Session content

Topics we'll cover in this session:

1) Introduction to the analysis of high-dimensional data

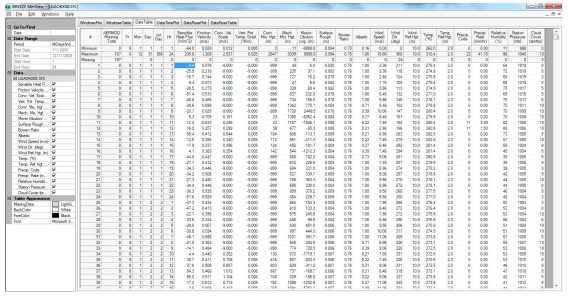
2) Potential Big Data in biology and biomedicine

3) Examples of applications of neural networks for Life Sciences

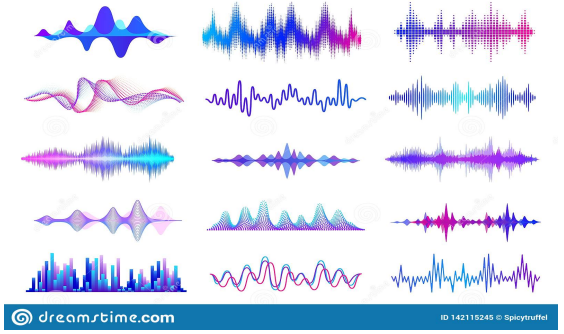
Introduction to high-dimensional data analysis: potential Big Data in Life Sciences

Various types of data around us

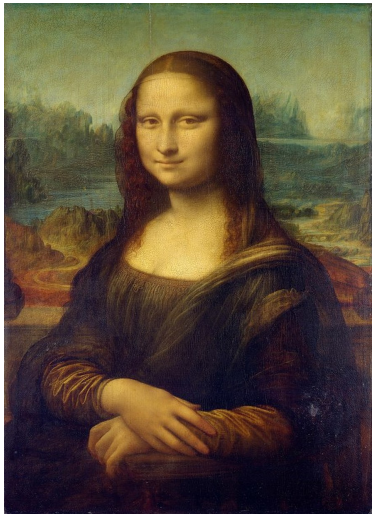
Tabular



Sound



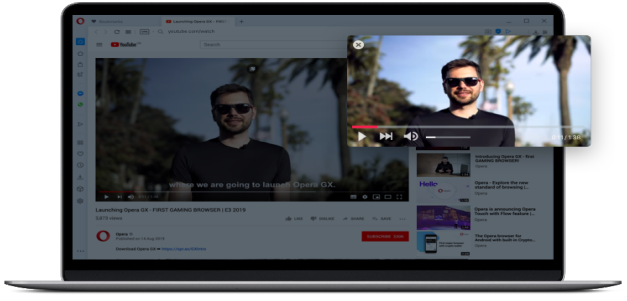
Image



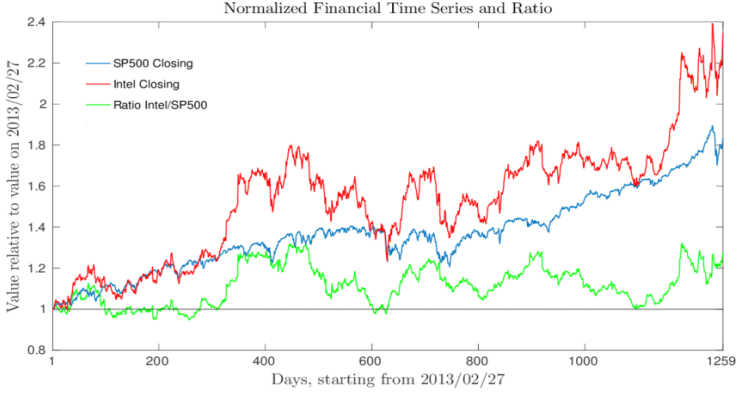
Text

DATA

Video



Time Series



Biological data are high dimensional

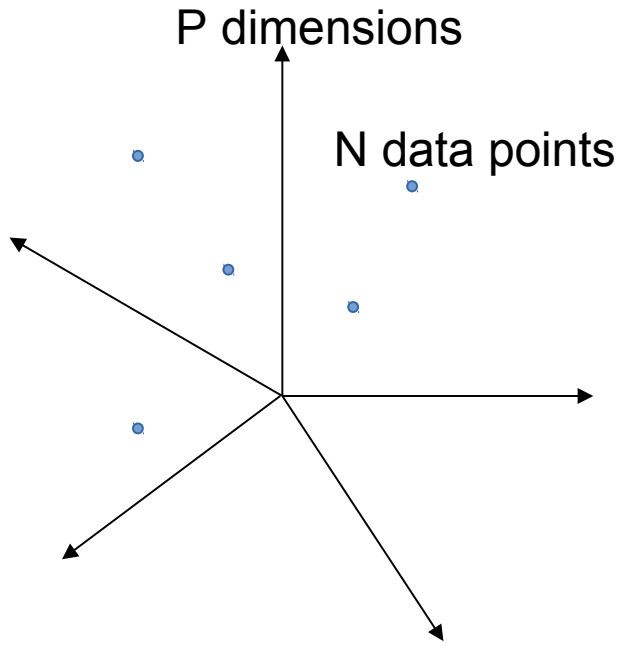
Statistical observations:
e.g. samples, cells etc.

N

P

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

Features: genes, proteins, microbes, metabolites etc.



High Dimensional Data:
P >> N

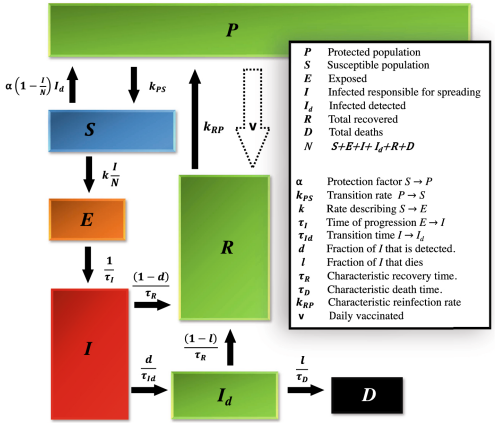
For a robust statistical analysis, one should properly “sample” the P-dimensional space, hence large sample size is required, N >> P

Some types of data analysis in Life Sciences

P is the number of features (genes, proteins, genetic variants etc.)
N is the number of observations (samples, cells, nucleotides etc.)

Biology / Biomedicine

Mathematical modeling



N ⟶ 0

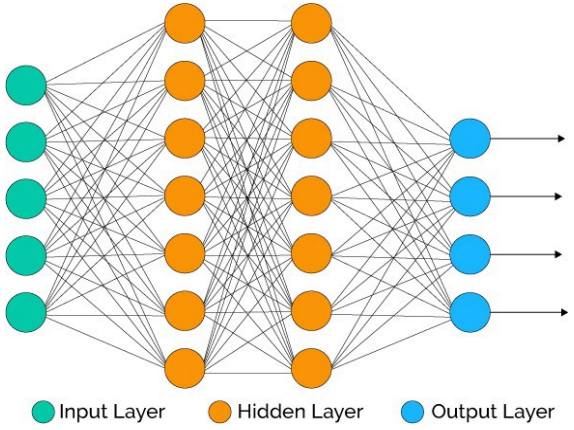
Bayesianism

N << P

Frequentism

N ≈ P

Machine Learning



N >> P

Hypothesis-driven ⟶ **Data-driven**

Amount of Data

The Curse of Dimensionality

Ex.1

$$Y = \alpha + \beta X$$

$$\beta = (X^T X)^{-1} X^T Y$$

$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \cdots \to \infty, \quad n << p$$

Ex.2 $$E[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2$$

Biased ML variance estimator in HD-space

Literature on the Curse of Dimensionality

POINTS OF SIGNIFICANCE

The curse(s) of dimensionality

There is such a thing as too much of a good thing.

Naomi Altman and Martin Krzywinski

We generally think that more information is better than less. However, in the 'big data' era, the sheer number of variables that can be collected from a single sample can be problematic. This embarrassment of riches is called the 'curse of dimensionality'[1] (CoD) and manifests itself in a variety of ways. This month, we discuss four important problems of dimensionality as it applies to data sparsity[1,2], multicollinearity[3], multiple testing[4] and overfitting[5]. These effects are amplified by poor data quality, which may increase with the number of variables.

Throughout, we use n to indicate the sample size from the population of interest and p to indicate the number of observed variables, some of which may have missing values for some samples. For example, we may have $n = 1{,}000$ subjects and $p = 200{,}000$ single-nucleotide polymorphisms (SNPs).

First, as the dimensionality p increases, the 'volume' that the samples may occupy grows rapidly. We can think of each of the n

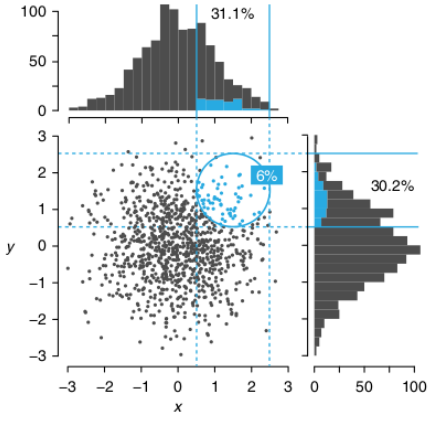


Fig. 1 | Data tend to be sparse in higher dimensions. Among 1,000 (x, y) points in which both x and y are normally distributed with a mean of 0 and s.d. $\sigma = 1$, only 6% fall within σ of (x, y) = (1.5, 1.5) (blue circle). However, when the data are projected into a lower dimension—shown by histograms—about 30% of the points (all bins within blue solid lines) are within σ of 1.5. Blue

A and 100 to have the minor allele a. If we tabulate on two SNPs, A and B, we will expect only ten samples to exhibit both minor alleles with genotype ab. With SNPs A, B and C, we expect only one sample to have genotype abc, and with four or more SNPs, we expect empty cells in our table. We need a much larger sample size to observe samples with all the possible genotypes. As p increases, we may quickly find that there are no samples with similar values of a predictor.

Even with just five SNPs, our ability to predict and classify the samples is impeded because of the small number of subjects that have similar genotypes. In situations where there are many gene variants, this effect is exacerbated, and it may be very difficult to find affected subjects with similar genotypes and hence to predict or classify on the basis of genetic similarity.

If we treat the distance between points (e.g., Euclidian distance) as a measure of similarity, then we interpret greater distance as greater dissimilarity. As p increases, this

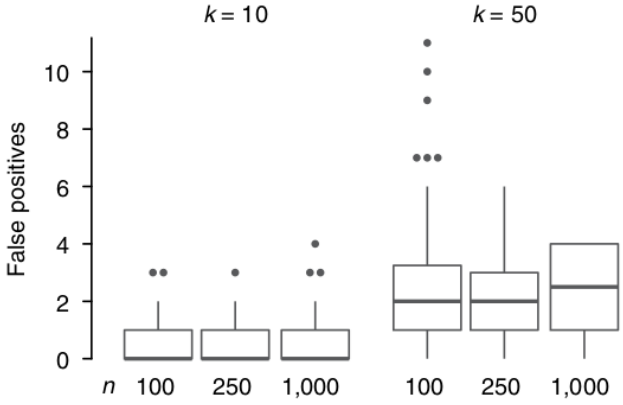
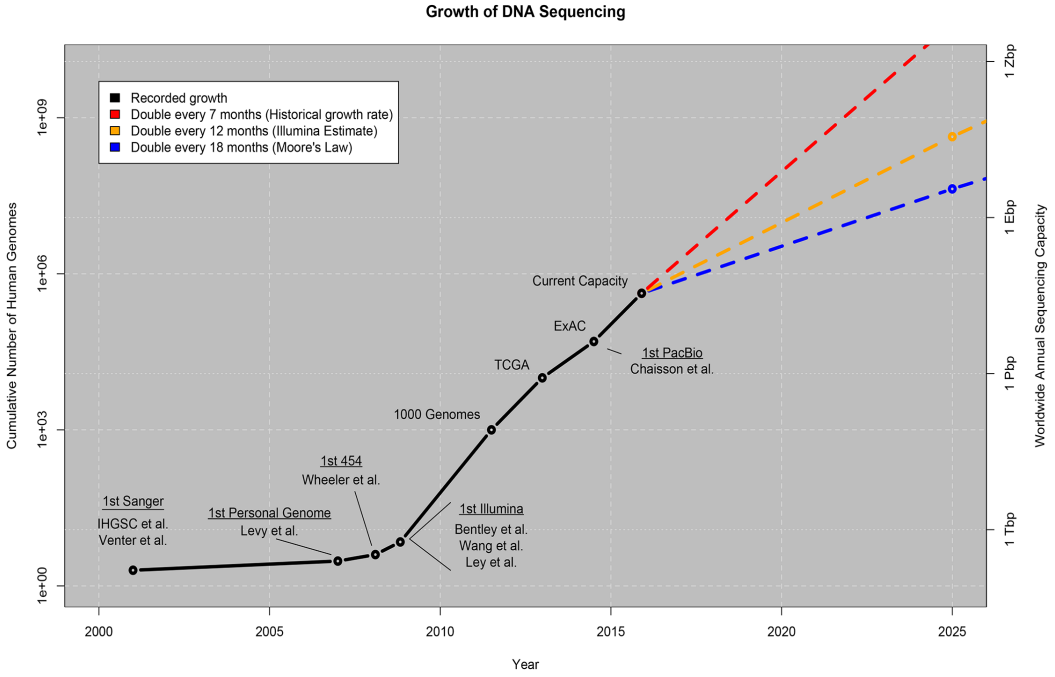


Fig. 3 | The number of false positives increases with each additional predictor. The box plots show the number of false positive regression-fit P values (tested at $\alpha = 0.05$) of 100 simulated multiple regression fits on various numbers of samples ($n = 100$, 250 and 1,000) in the presence of one true predictor and $k = 10$ and 50 extraneous uncorrelated predictors. Box plots show means (black center lines), 25th and 75th percentiles (box edges), and minimum and maximum values (whiskers). Outliers (dots) are jittered.

Correcting for multiple testing does not solve the problem of too many false-positive hits

Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018 Jun;15(6):399-400. doi: 10.1038/s41592-018-0019-x. PMID: 29855577.

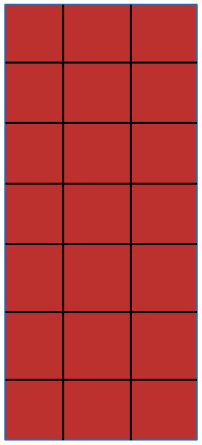
Do we have Big Data in Life Sciences?



Stephens et al., (2015). Big Data: Astronomical or Genomical? PLoS Biology 13(7)

Genomics / WGS: **Little Data**

$N_1 \sim 10^3$



$P_1 \sim 10^6$

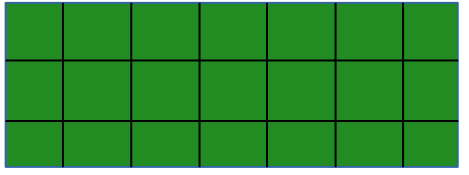
$N_1 * P_1 = N_2 * P_2 = 10^9$

Potential Big Data in Life Sciences:

- Microscopy imaging (well known, AI widely used)
- Single cell Omics (novel type of data for AI)
- Metagenomics (possibly, not high-dimensional)
- Genomics (possibly, sequence is an observation)
- Epidemiology (population level data)

scRNAseq: **Big Data**

$N_2 \sim 10^6$



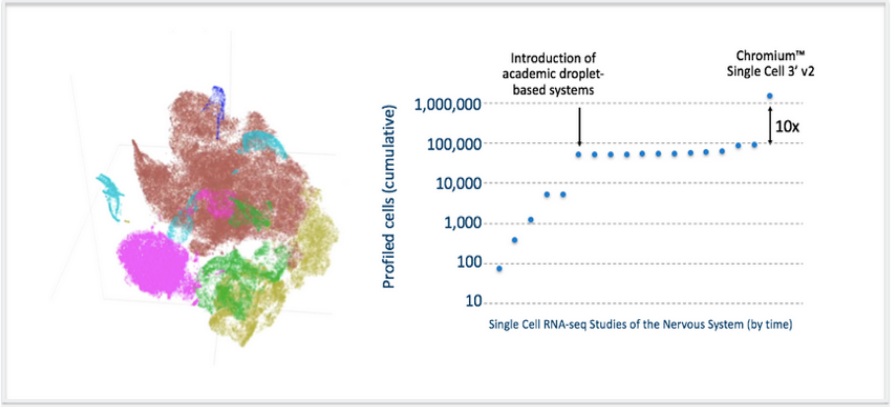
$P_2 \sim 10^3$

Big Data in single cell genomics

CAREERS BLOG 10X UNIVERSITY

10X GENOMICS SOLUTIONS & PRODUCTS RESEARCH & APPLICATIONS EDUCATION & RESOURC

« Back to Blog ‹ Newer Article Older Article ›



Our 1.3 million single cell dataset is ready to download

0 KUDOS

POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic mice brains. Read more in our application note Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Single cells make big data: New challenges and opportunities in transcriptomics

Philipp Angerer[1], Lukas Simon[1], Sophie Tritschler[1], F. Alexander Wolf[1], David Fischer[1] and Fabian J. Theis[1,2]

Abstract

Recent technological advances have enabled unprecedented insight into transcriptomics at the level of single cells. Single cell transcriptomics enables the measurement of transcriptomic information of thousands of single cells in a single experiment. The volume and complexity of resulting data make it a paradigm of big data. Consequently, the field is presented with new scientific and, in particular, analytical challenges where currently no scalable solutions exist. At the same time, exciting opportunities arise from increased resolution of single-cell RNA sequencing data and improved statistical power of ever growing datasets. Big single cell RNA sequencing data promises valuable insights into cellular heterogeneity which may significantly improve our understanding of biology and human disease. This review focuses on single cell transcriptomics and highlights the inherent opportunities and challenges in the context of big data analytics.

Addresses

[1] Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany
[2] Department of Mathematics, Technical University of Munich, Garching, Germany

Corresponding author: Theis, Fabian J. Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany. (fabian.theis@helmholtz-muenchen.de)

Current Opinion in Systems Biology 2017, 4:85–91

This review comes from a themed issue on **Big data acquisition and**

volume — the amount of data, velocity — the required processing speed, veracity — trustworthiness and availability, and variety — necessary model complexity [2]. The traditional scientific big data field is astronomy because of the huge *volume* of image data produced by telescopes with a high daily *velocity* [3]. Big data has also reached biology, mainly driven through the advent of next generation sequencing technology. For biologists, assessing *veracity* through statistical means is nothing new.

Recent technological advances now allow the profiling of single cells at a *variety* of omic layers (genomes, epigenomes, transcriptomes and proteomes) at an unprecedented level of resolution [4]. Single cell transcriptomics (SCT) entails the profiling of all messenger RNAs present in a single cell and constitutes the most widely-used sc profiling technology [4]. Unlike bulk RNA-seq profiling where sequencing libraries are generated from thousands of cells, scRNA-seq technologies isolate single cells and generate cell-specific sequencing libraries (e.g. Fluidigm [5]) mark RNA content with a cell-specific molecular barcode [6–9]. Both approaches generate gene expression estimates at the single cell level [10]. SCTenables, for the first time, the measurement of the transcriptomic information of thousands, and up to millions of single cells, in a single experiment [7]. The complexity of SCT data coupled with the massive volume inherent to next generation sequencing data makes it a paradigm of big data.

AI applications for single cell biology

nature methods
ARTICLES
https://doi.org/10.1038/s41592-018-0229-2

Deep generative modeling for single-cell transcriptomics

Romain Lopez[1], Jeffrey Regier[1], Michael B. Cole[2], Michael I. Jordan[1,3] and Nir Yosef[1,4,5*]

Single-cell transcriptome measurements can reveal unexplored biological diversity, but they suffer from technical noise and bias that must be modeled to account for the resulting uncertainty in downstream analyses. Here we introduce single-cell variational inference (scVI), a ready-to-use scalable framework for the probabilistic representation and analysis of gene expression in single cells (https://github.com/YosefLab/scVI). scVI uses stochastic optimization and deep neural networks to aggregate information across similar cells and genes and to approximate the distributions that underlie observed expression values, while accounting for batch effects and limited sensitivity. We used scVI for a range of fundamental analysis tasks including batch correction, visualization, clustering, and differential expression, and achieved high accuracy for each task.

nature methods

Article
https://doi.org/10.1038/s41592-024-02201-0

scGPT: toward building a foundation model for single-cell multi-omics using generative AI

Received: 12 July 2023
Accepted: 30 January 2024
Published online: 26 February 2024
Check for updates

Haotian Cui[1,2,3,8], Chloe Wang[1,2,3,8], Hassaan Maan[1,3,4], Kuan Pang[2,3], Fengning Luo[2,3], Nan Duan[5] & Bo Wang[1,2,3,4,6,7]

Generative pretrained models have achieved remarkable success in various domains such as language and computer vision. Specifically, the

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04368-5 OPEN

Interpretable dimensionality reduction of single cell transcriptome data with deep generative models

Jiarui Ding[1,2,3,4], Anne Condon[1] & Sohrab P. Shah[1,2,3,5]

Huang et al. Genome Biology (2023) 24:259
https://doi.org/10.1186/s13059-023-03100-x

Genome Biology

RESEARCH Open Access

Evaluation of deep learning-based feature selection for single-cell RNA sequencing data analysis

Hao Huang[1,2,3], Chunlei Liu[1,3], Manoj M. Wagle[1,2,3] and Pengyi Yang[1,2,3,4*]

Biomedicine & Pharmacotherapy 165 (2023) 115077

Contents lists available at ScienceDirect
Biomedicine & Pharmacotherapy
journal homepage: www.elsevier.com/locate/biopha

Review

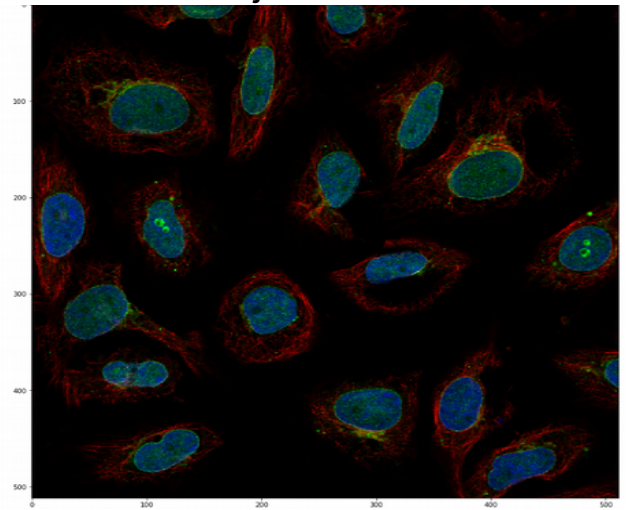
Deep learning applications in single-cell genomics and transcriptomics data analysis

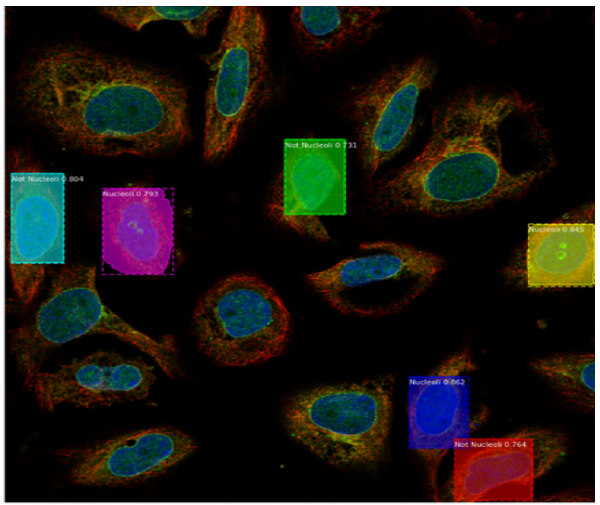
Nafiseh Erfanian[a], A. Ali Heydari[b,c], Adib Miraki Feriz[a], Pablo Iañez[d], Afshin Derakhshani[e], Mohammad Ghasemigol[f], Mohsen Farahpour[g], Seyyed Mohammad Razavi[g], Saeed Nasseri[h], Hossein Safarpour[h,*], Amirhossein Sahebkar[i,j,k,**]

A few examples of applications of artificial neural networks in Life Sciences

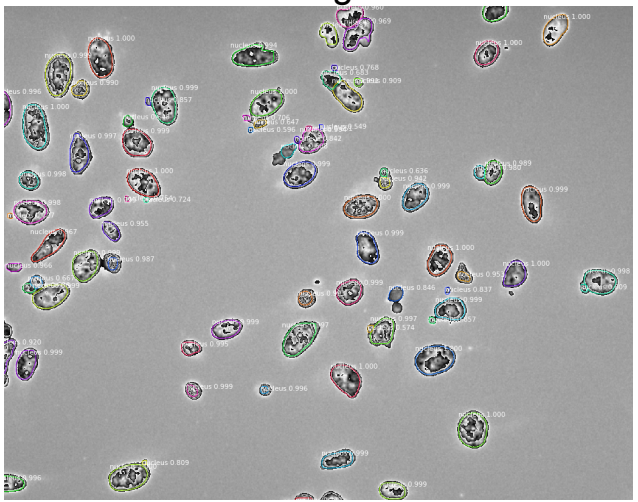
Deep Learning for Microscopy Imaging

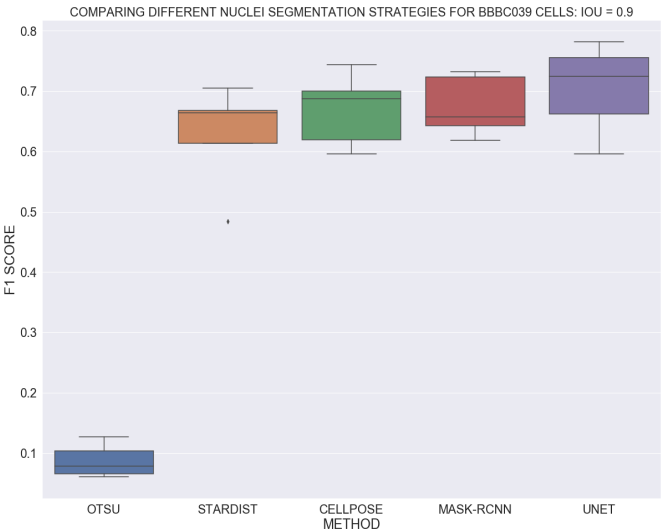
Object detection



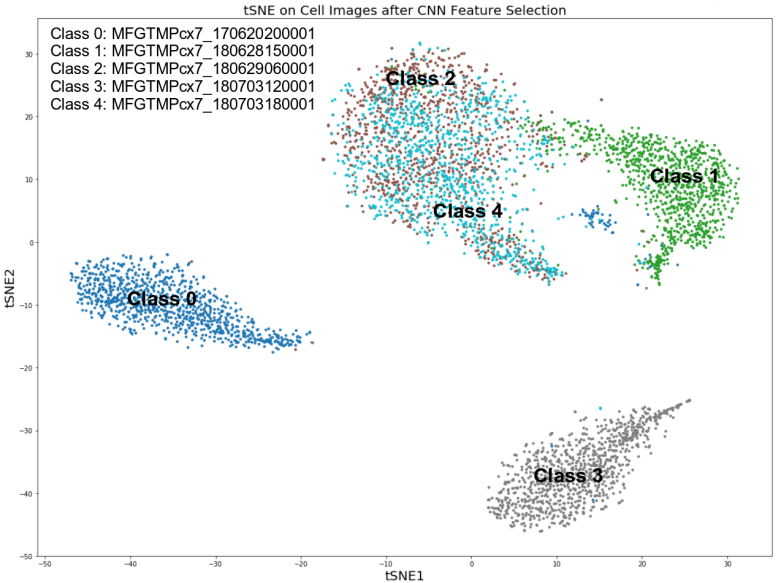


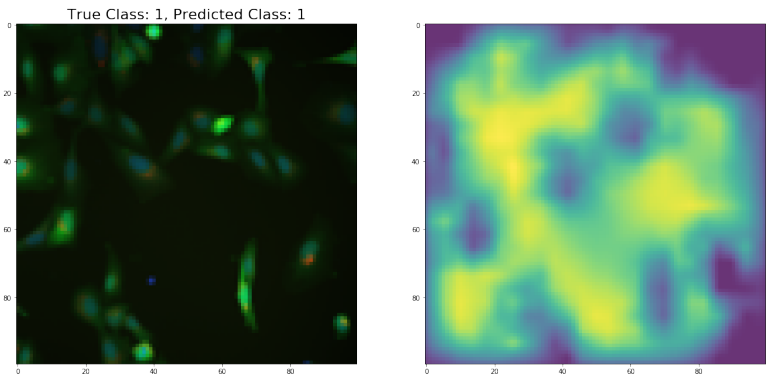
Instance segmentation





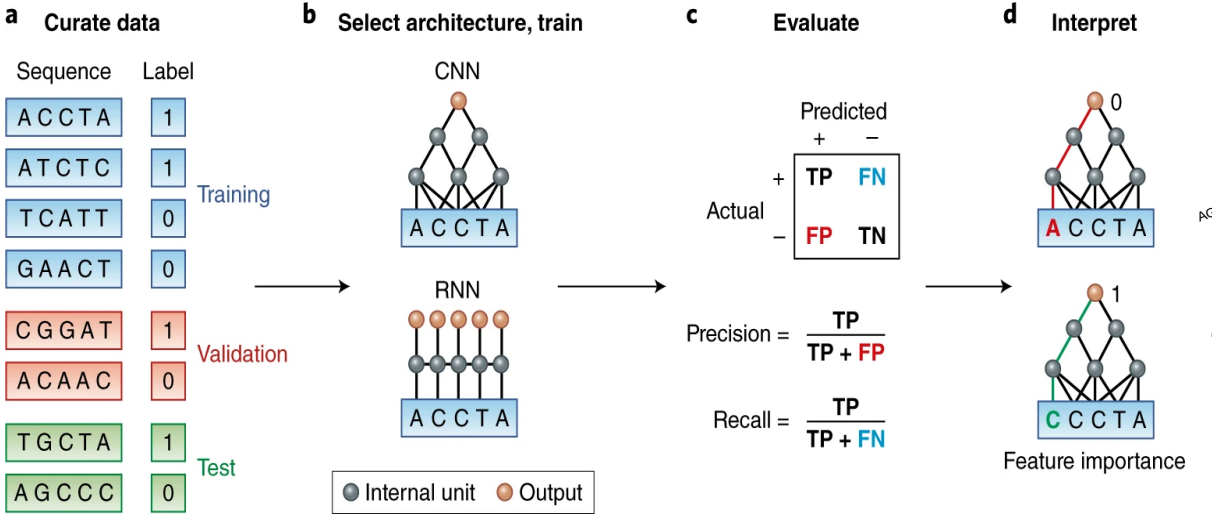
Grad-CAM image cluster interpretation



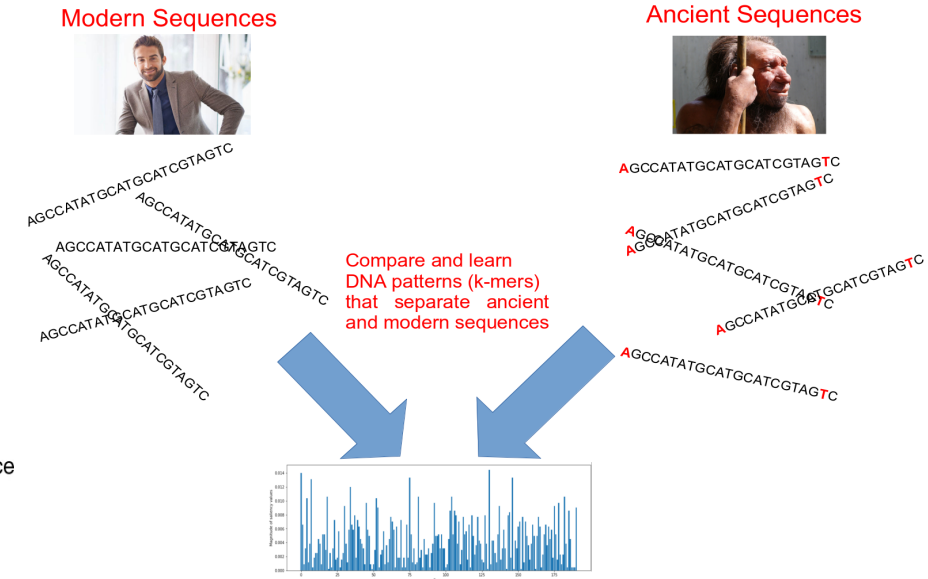


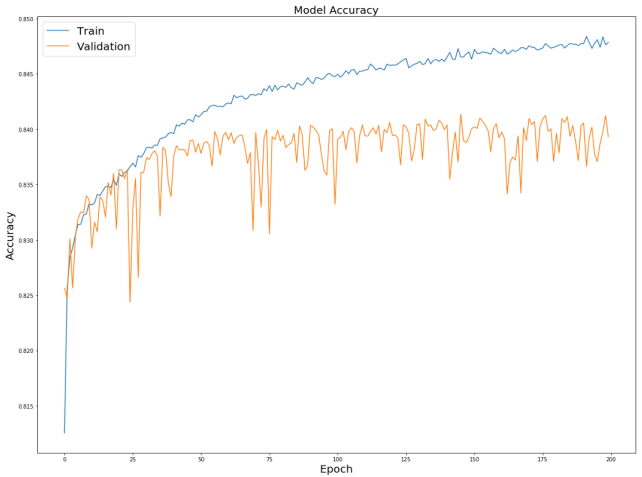
Oskolkov et al., unpublished

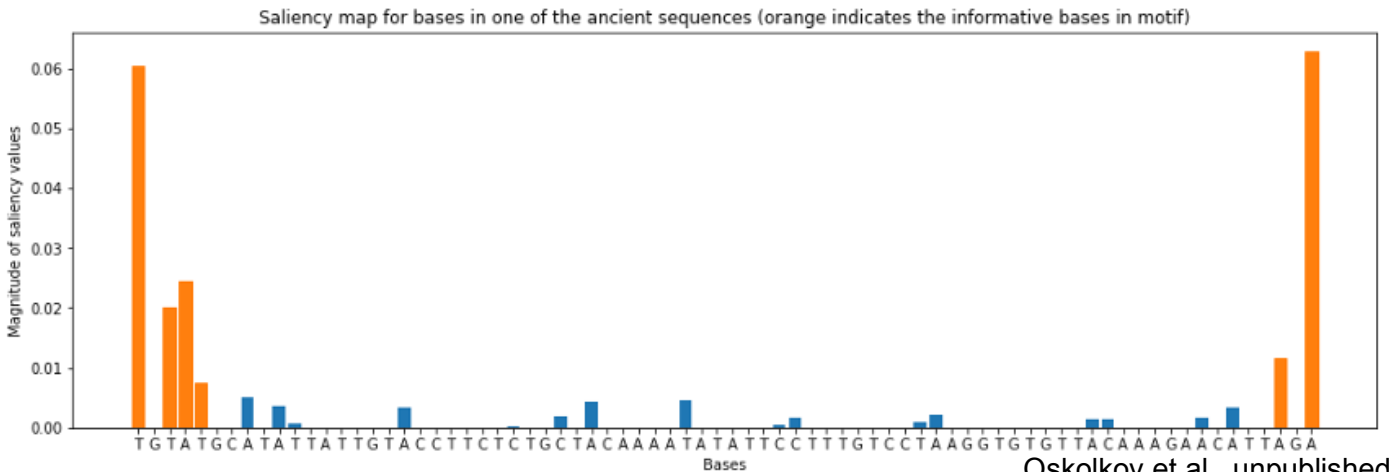
Deep Learning for Genomics



Zou et al., 2018, Nature Genetics 51, 12-18

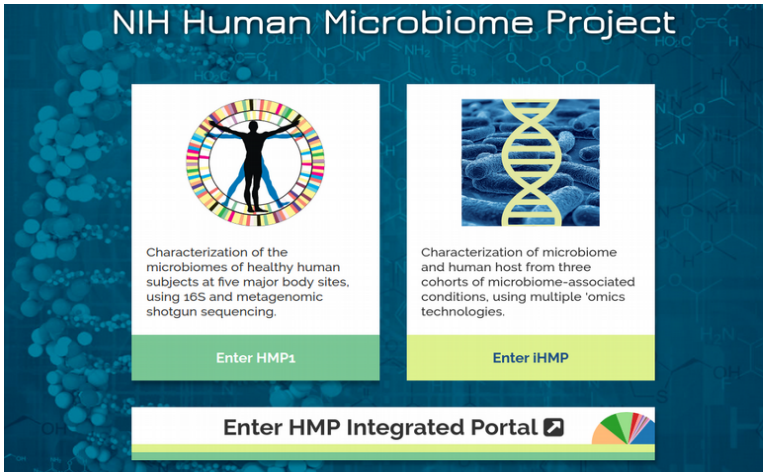


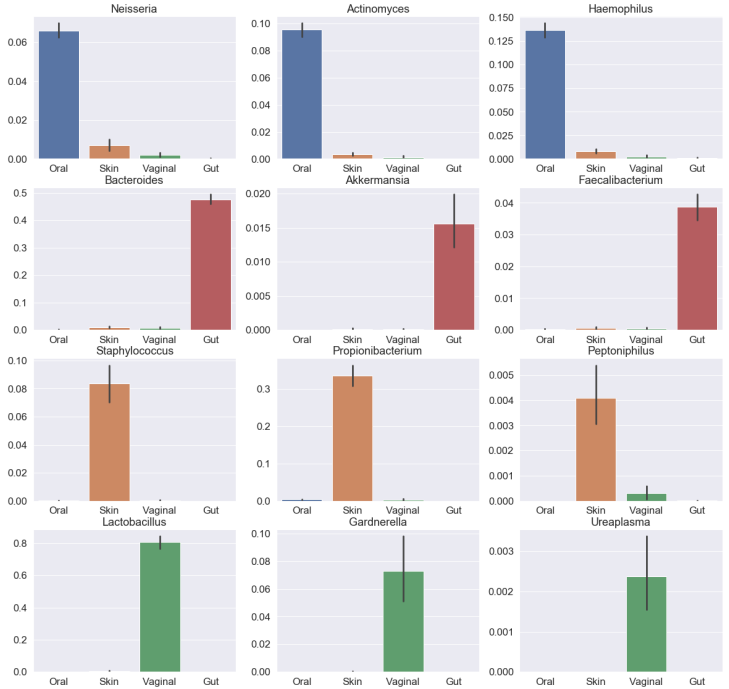


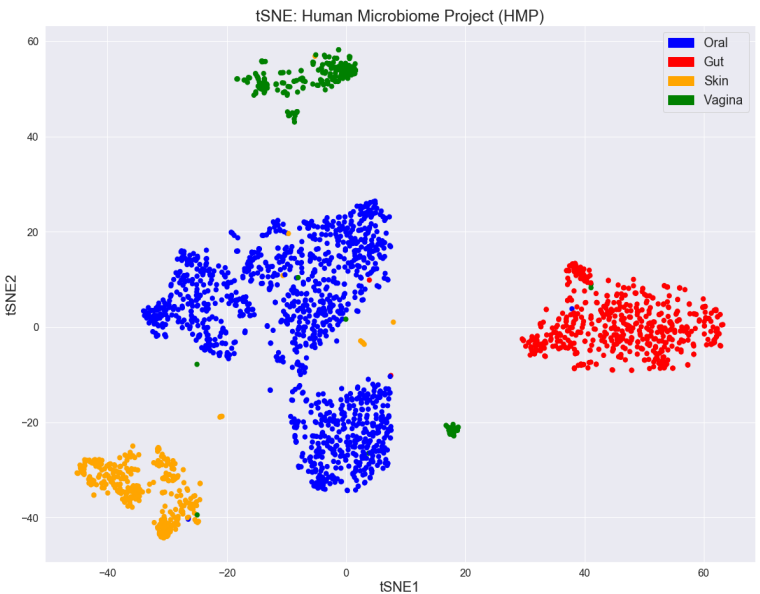


Oskolkov et al., unpublished

Deep Learning for Metagenomics

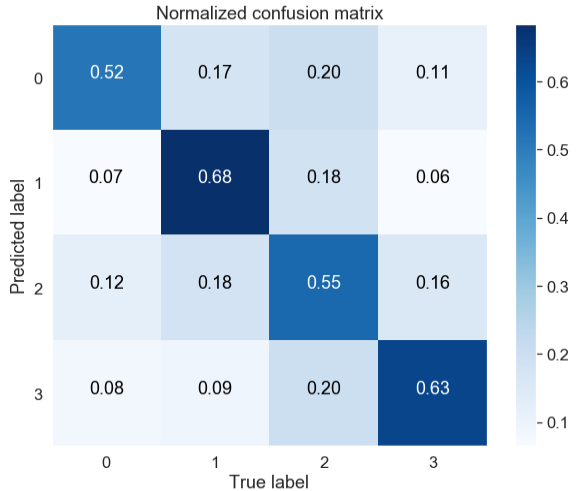


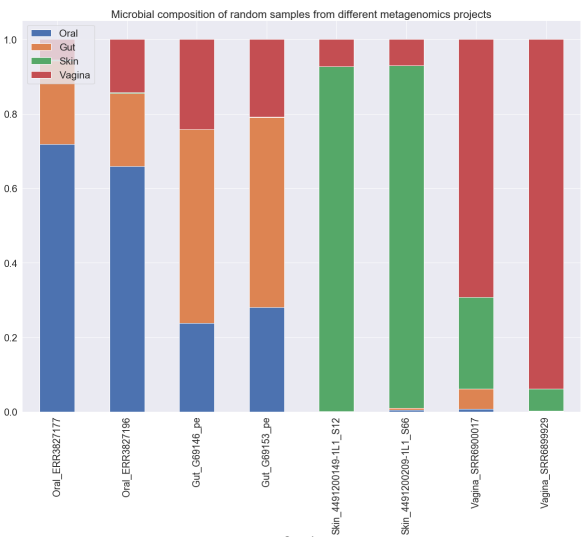




	SEQ	LAB
0	GCACGACATTTACAGCTTACCGCCACGTTTTATCCGCAATTTGAGTGCGATGGTCAGGCAGGACGATGTTTTATAGTGGA	0
1	CAGCCTGAAAAATTCACATACCTTGCAAGCAGGAAAACGCATTCGGATTAAAGCAAATAACCTTGATAATGCAGTACAAG	0
2	GCCAAATACCCGTCCCCTTACCATTCCCGAAATATCCGTTCAAACTTGGAGCAGCGTTACGGCAAAGAAAACATCACCTC	0
3	GGCGGGGGGATGCCACGCCGGCAGCGCCGTTGTCGGCGGGCCATTTGCCGTTATTCAGGTAATACCCGGTAACGGCTGATT	0
4	AATCCCAACGTACCGGCTGATATTCTTCGCGCTTGGGGTGTTTTGGTTCAGGTCTATCAAGTAAGCCAGGCCACCCGCGT	0
...	...	...
12971523	CCATGGTTGCCCCGCCCCCGCCAACGCCAATCACTCAAGTCCAACAGGGCGGTGACGGTGGCGCCGGGGGCACCGGATCC	3
12971524	GAGCGGAGCACATGACACGATACGACTCGCTGTTGCAGGCCTTGGGCAACACGCCGCTGGTTGGCCTGCAGCGATTGTCG	3
12971525	TCGACGGGGGTGGCGTTCGCGTTCGGGCCGGGCGTCACTGTCGAAGGCATGCTGTTCGACATCATCCGACGGTGACCGCC	3
12971526	GCGAGGTCCGGAGCGGGGCGGAAATTTCATTGAACAGCCGTAGAAGTTCAGCCAGGACCGGAACGGATCCAGCGGCAAGC	3
12971527	AACTGGGCACGGCGCTTGACGCACCGGTGCTGATGCATCCCGGCGACGCCGTGCTGTGGCGAATGACTCACCCGGACAAA	3

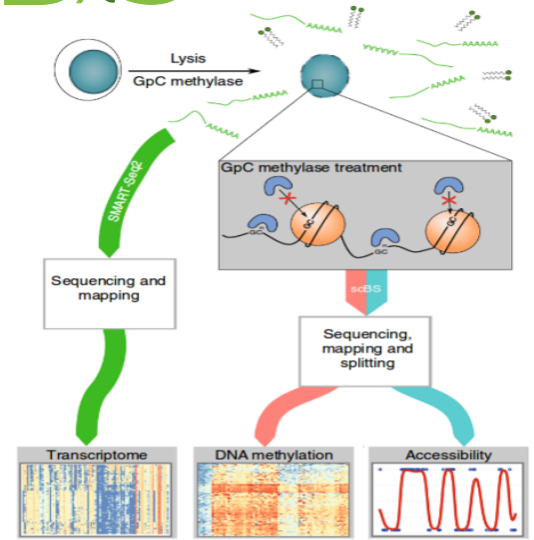
12971528 rows × 2 columns



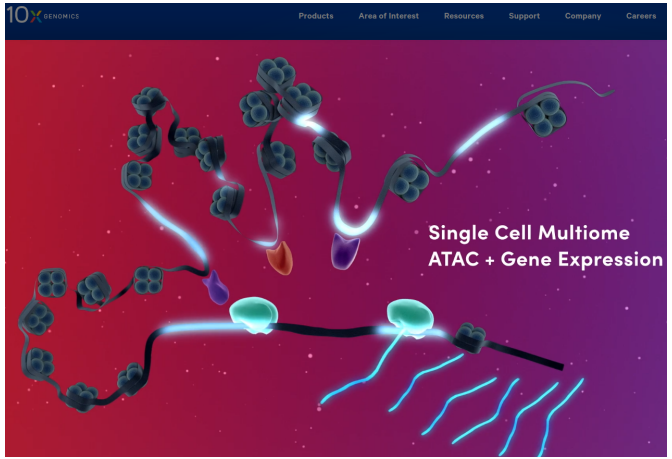


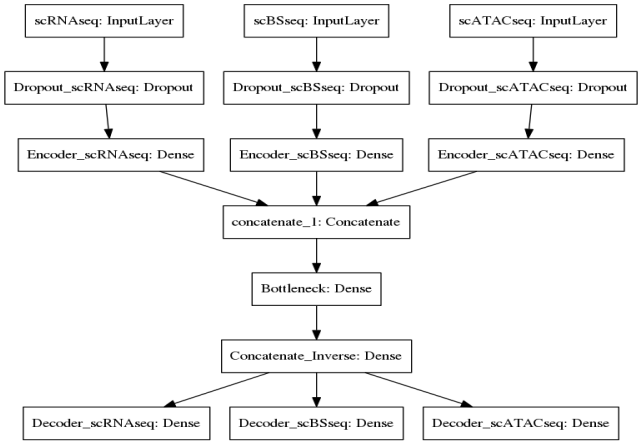
Oskolkov et al., unpublished

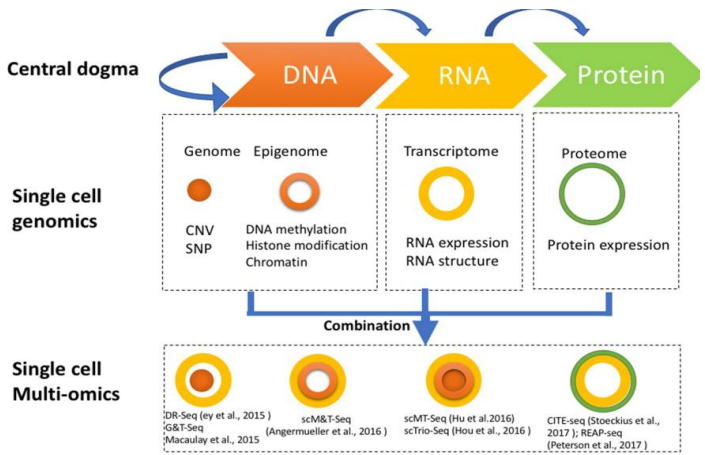
Deep Learning for Data Integration



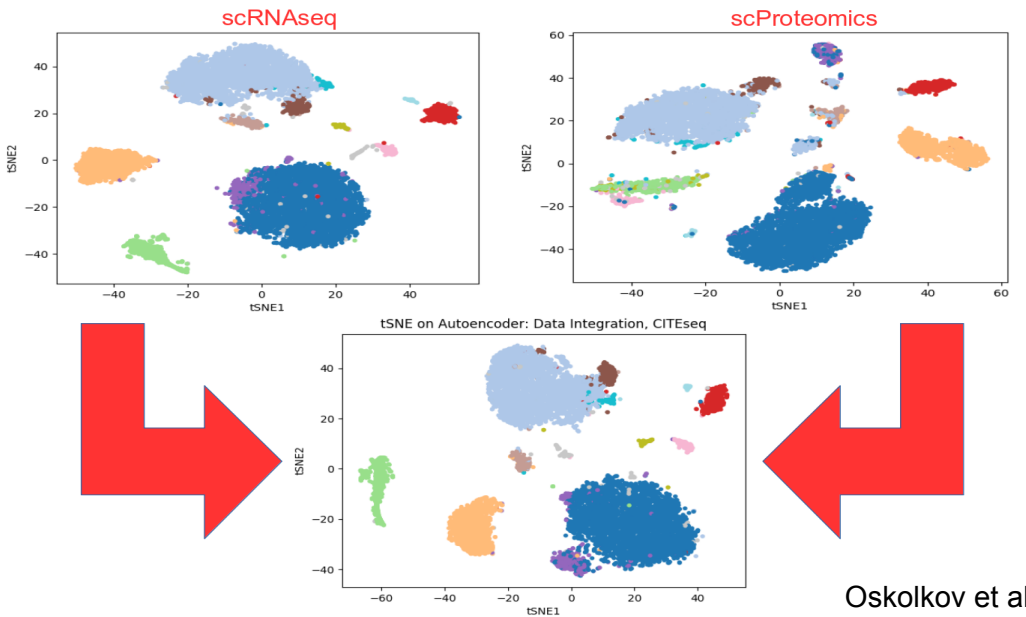
Clark et al., Nature Communications 2018





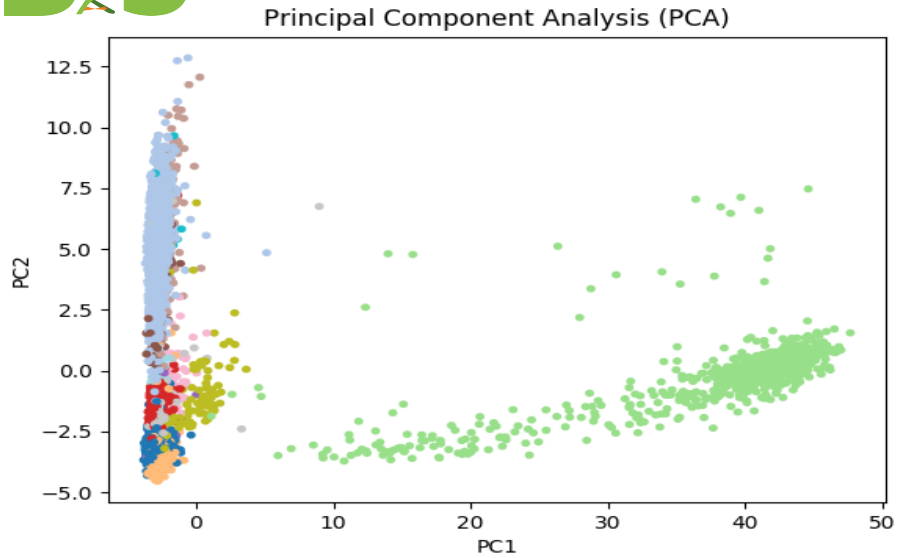


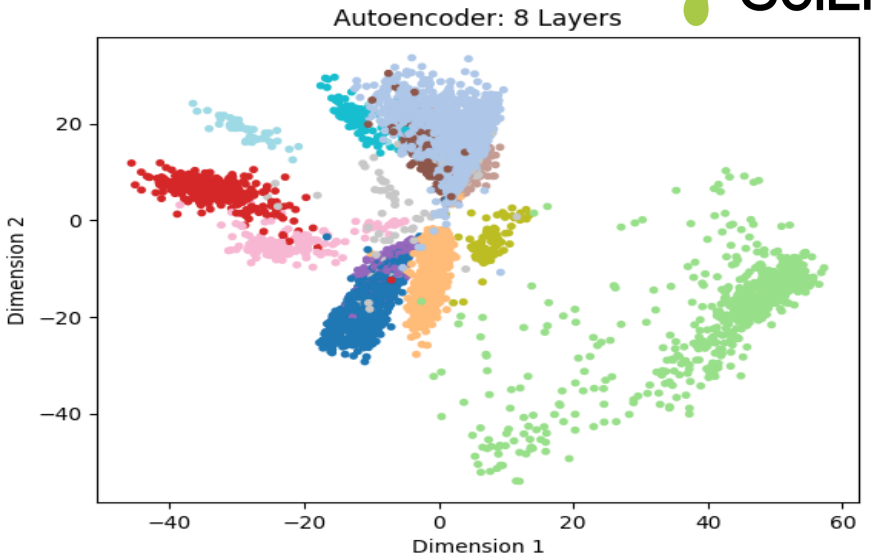
Hu et al., Frontier in Cell and Developmental Biology 2018

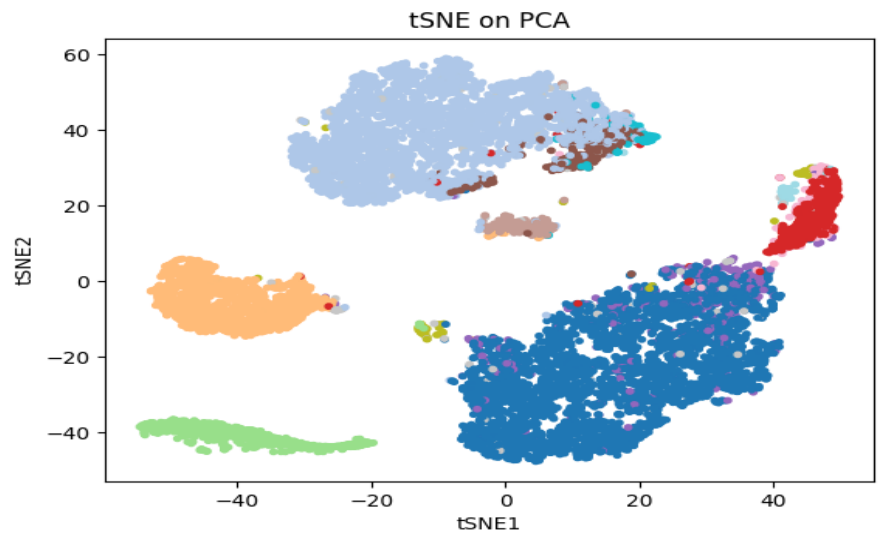


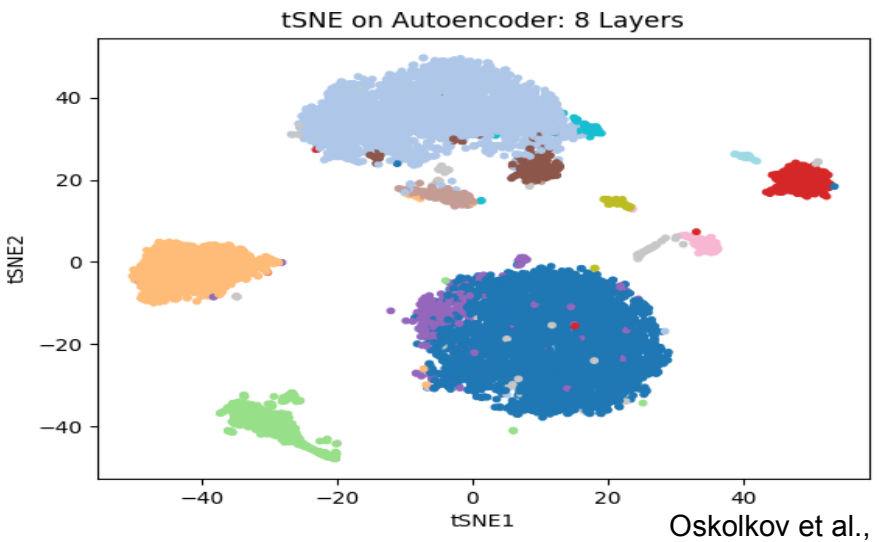
Oskolkov et al., unpublished

Deep Learning for Single Cell Biology









Oskolkov et al., unpublished

Popular summary of AI for Life Sciences

towards data science

LATEST | EDITORS' PICKS | DEEP DIVES | ABOUT | CONTRIBUTE | NEWSLETTER

TAGGED IN

DI For Life Sciences

tds Towards Data Science

Your home for data science. A Medium publication sharing concepts, ideas and codes.

More information

FOLLOWERS

688K

ELSEWHERE

MORE, ON MEDIUM

DI For Life Sciences

Nikolay Oskolkov in Towards Data Science
Aug 18, 2021 · 13 min read ★

DEEP LEARNING FOR LIFE SCIENCES

Deep Learning on Human Microbiome

Infer microbial...

Read more...

133 4 responses

Nikolay Oskolkov in Towards Data Science
May 20, 2019 · 7 min read ★

Deep Learning for Life Sciences

Deep Learning for Clinical Diagnostics

Read more...

141

Nikolay Oskolkov in Towards Data Science
Aug 6, 2019 · 6 min read ★

Deep Learning for Life Sciences

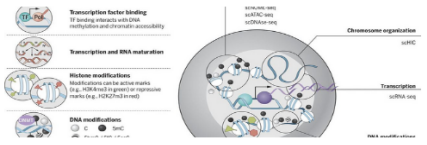
Why Biology is Sceptic Towards AI

And why Precision...

Read more...

159 1 response

Nikolay Oskolkov in Towards Data Science
May 12, 2019 · 7 min read ★



Deep Learning for Life Sciences

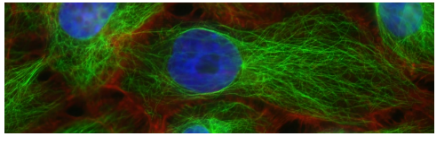
Deep Learning for Data Integration

Synergistic effects...

Read more...

126 5 responses

Nikolay Oskolkov in Towards Data Science
Jul 7, 2019 · 9 min read ★



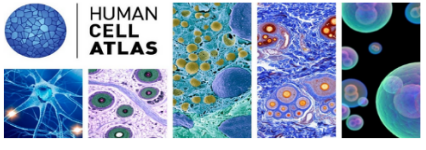
Deep Learning for Life Sciences

Deep Learning on Microscopy Imaging

Read more...

207 1 response

Nikolay Oskolkov in Towards Data Science
May 5, 2019 · 9 min read ★



Deep Learning for Life Sciences

Deep Learning for Single Cell Biology

Read more...

406 7 responses

Nikolay Oskolkov in Towards Data Science
Apr 28, 2019 · 6 min read ★

Deep Learning for Life Sciences

Deep Learning on Ancient DNA

Reconstructing the Human...

Read more...

381 4 responses

Session summary

Take home messages of the session:

1) Challenge of high-dimensional data: Curse of Dimensionality

2) Balance between numbers of samples and features

3) Singe cell genomics represent promising big data in Biology

4) Deep Learning sequence classification is common in Genomics

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet