Deep Learning Benchmarks on Gene Expression Data

Matthew B.A. McDermott*, Jennifer Wang†, Wen-Ning Zhao‡, Steven D. Sheridan†, Peter Szolovits*, Isaac Kohane⋆, Stephen J. Haggarty‡, Roy H Perlis†

*CSAIL, MIT; †CEDD & ‡CGM, MGH; ⋆DBMI, Harvard

Get in Touch
mmd@mit.edu
https://github.com/mmcdermott/LINCS_Deep_Learning_Benchmarks

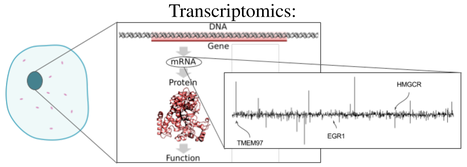
Abstract

We establish and profile three benchmark tasks over two curated views of the large, publicly available LINCS Pilot Phase I gene expression corpus (GSE 92742) and one privately produced gene expression dataset. We test a variety of methods, including graph convolutional neural networks (GCNNs) which incorporate prior biological knowledge via regulatory networks. GCNNs can be very performant, but require large datasets, whereas FF-ANNs consistently perform well. Non-neural classifiers across all tasks and datasets are dominated by linear models and KNN classifiers.

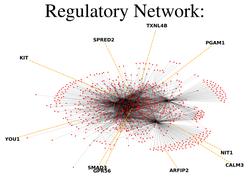
Gene Expression Data

- Transcriptomic gene expression data tracks mRNA prior to its translation into proteins.
- The L1000 technique is a low-cost, high-throughput transcriptomics technology that only directly measures 978 "landmark genes" rather than the full transcriptome. L1000 data is often used both per-sample (Level 4) and per-aggregated-triplicate (Level 5).
- Gene expression self-regulation is commonly visualized by regulatory graphs, where nodes are genes and edges represent regulatory relationships.

Transcriptomics:



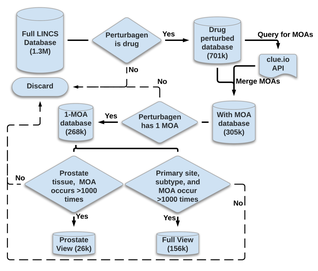
Regulatory Network:



Data

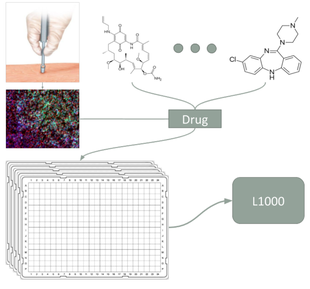
LINCS GSE92742

We downloaded the LINCS GSE92742 L1000 corpus from GEO, and induced two curated views of this dataset suitable for three benchmark tasks: primary site, subtype, and drug mechanism of action (MOA) prediction.



MGH NeuroBank Corpus

The MGH NeuroBank corpus is formed from patient-derived neural progenitor cells (NPCs) of two healthy control subjects, two schizophrenic subjects, and one subject with bipolar disorder.

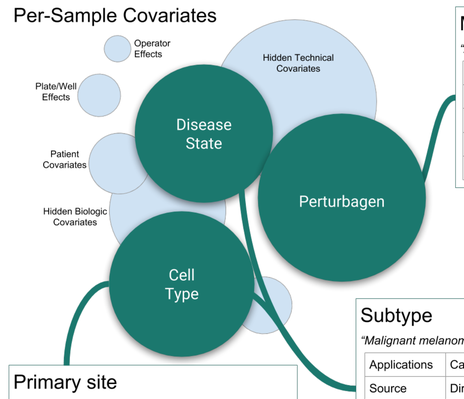


Dataset Statistics

Dataset	# Samples	# Cell Lines	Most Frequent Line	Least Frequent Line
LINCS (Full)	156,461	36	MCF7 (26,546)	NCIH716 (8)
LINCS (Prostate)	25,565	2	PC3 (13,625)	VCAP (11,940)
MGH NeuroBank (Lvl 4)	5602	5	N/A (1133)	N/A (1109)
MGH NeuroBank (Lvl 5)	1894	5	N/A (380)	N/A (377)

Classification Benchmark Tasks

Our tasks spanned three central sources of per-sample variation: perturbagen mechanism of action (MOA), cell primary site, and cell diagnostic subtype. All performance measures are in per-sample accuracy across a fixed set of 10 cross-validation folds. On the MGH corpus, we predicted drug identity, to avoid sample loss. We *do not* predict per-perturbagen or per-cell line accuracies, and our results should be considered only as methodological comparison points.



Mechanism of Action (MOA)

"ATPase inhibitor," "Sodium channel blocker"

Applications	Drug Repurposing, Discovery
Source	https://clue.io/api#perts
# Classes	49
Most Freq.	DMSO (~26k)
Least Freq.	IKK Inhibitor (~840)

Primary site

"Breast tissue," "Large-intestine"

Applications	Cell Differentiation, QC Metrics
Source	Directly Encoded in LINCS Data
# Classes	12
Most Freq.	Prostate (~44k)
Least Freq.	Ovary (~420)

Subtype

"Malignant melanoma," "myoblast"

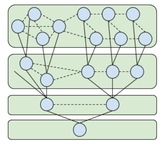
Applications	Case Control Analysis
Source	Directly encoded in LINCS Data
# Classes	14
Most Freq.	Adenocarcinoma (~53k)
Least Freq.	Embryonal Kidney (~1.4k)

Methods

We tested methods spanning non-neural methods, feed-forward neural networks, and graph convolutional networks over fixed, tissue-independent regulatory networks.

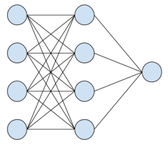
Graph Convolutional Neural Networks (GCNNs)

- Structural
- Weight Sharing
- Theory vs. Practice



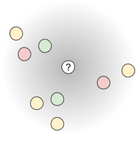
Feed-Forward Artificial Neural Networks (FF-ANNs)

- Unstructured
- No Parameter Sharing
- Inefficient Learners



***K*-Nearest Neighbor Classifiers (KNNs)**

- Theoretically Optimal
- Require Density
- Distance-Metric



Linear Classifiers

- Simple
- Interpretable
- Limited Capacity

In addition, we also tested random forests and decision trees, but they performed worse and are not reported here.

Results

LINCS Corpora

- GCNNs outperform all other methods on the full corpus, and place 2nd on the prostate corpus.
- Shallow FF-ANNs consistently perform well.
- KNNs using the Canberra distance dominate non-neural methods.

Task	Classifier Name	Accuracy	Macro F1
Primary Site	GCNN	$\mathbf{93.9 \pm 0.28}$	$\mathbf{90.5 \pm 0.82}$
	FF-ANN	90.6 ± 0.44	85.6 ± 0.97
	KNNs	89.6 ± 0.30	87.2 ± 0.61
	Linear Classifier	60.9 ± 0.50	47.6 ± 0.63
	Majority Class	27.9 ± 0.16	3.63 ± 0.02
Subtype	GCNN	$\mathbf{93.5 \pm 0.34}$	$\mathbf{91.7 \pm 2.1}$
	FF-ANN	90.5 ± 0.30	88.5 ± 0.54
	KNNs	89.8 ± 0.13	90.2 ± 0.27
	Linear Classifier	62.6 ± 0.62	56.3 ± 1.06
	Majority Class	34.0 ± 0.21	3.62 ± 0.02
MOA	GCNN	$\mathbf{46.4 \pm 0.35}$	$\mathbf{31.6 \pm 0.65}$
	FF-ANN	45.9 ± 0.43	29.6 ± 0.60
	KNNs	43.5 ± 0.50	29.5 ± 0.58
	Linear Classifier	39.1 ± 0.29	20.6 ± 0.39
	Majority Class	16.4 ± 0.16	0.57 ± 0.005
Prostate MOA	GCNN	67.7 ± 0.76	46.0 ± 0.42
	FF-ANN	$\mathbf{68.3 \pm 0.60}$	$\mathbf{50.4 \pm 0.71}$
	KNNs	66.5 ± 0.71	46.2 ± 0.89
	Linear Classifier	63.8 ± 0.52	42.6 ± 1.03
	Majority Class	34.54 ± 0.05	5.71 ± 0.01

All differences between classifiers were statistically significant at $p \leq 0.05$.

MGH NeuroBank Corpus

- FF-ANNs perform best here, followed by linear models, KNNs, and GCNNs.
- Level 5 and level 4 results were statistically insignificantly different save for GCNNs.
- All differences between per-sample and per-patient performance were statistically significant.

Classifier Name	Per-Sample Level 5		Per-Sample Level 4		Per-Patient Level 4	
	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
GCNN	46.0 ± 9.90	44.0 ± 10.8	54.6 ± 3.94	56.4 ± 3.94	47.7 ± 6.78	48.9 ± 7.40
FF-ANN	$\mathbf{63.2 \pm 10.3}$	$\mathbf{62.7 \pm 10.8}$	$\mathbf{57.3 \pm 4.12}$	$\mathbf{58.9 \pm 4.00}$	$\mathbf{48.7 \pm 7.85}$	$\mathbf{50.1 \pm 8.34}$
KNNs	46.9 ± 8.13	44.7 ± 9.15	44.9 ± 3.74	45.7 ± 3.61	37.9 ± 5.39	39.0 ± 6.68
Linear Classifier	52.3 ± 9.61	51.4 ± 10.0	49.1 ± 3.98	50.2 ± 3.63	44.1 ± 4.03	44.7 ± 4.21
Majority Class	7.56 ± 2.37	0.23 ± 0.07	6.88 ± 0.77	0.21 ± 0.02	N/A	N/A

Conclusion

1. Gene expression data is an important modality. These benchmarks will enable more efficient deep learning method development.
2. Biologically structured models such as GCNNs can offer performance benefits, but require large amounts of data.
3. Per-patient generalizability remains an important, unsolved problem.

This research was funded in part by grants from the National Institutes of Health (NIH): National Institute of Mental Health (NIMH) grant P50-MH106933, National Human Genome Research Institute (NHGRI) grant U54-HG007963. We would additionally like to thank the organizers of BioKDD '18 for organizing this workshop.