Poster presentation at **RSGDREAM 2022**, 7-11th November 2022, Las Vegas, NV

DREAM Challenge 2022
Predicting Gene Expression Using a Residual CNN

A description of team Camformers' submission (4th place) to the DREAM 2022 challenge "Predicting gene expression using millions of random promoter sequences".

Fredrik Svensson[1], Maria-Anna Trapotsi[2], Susanne Bornelöv[2]

[1]Alzheimer's Research UK UCL Drug Discovery Institute, University College London, UK
[2]Cancer Research UK Cambridge Institute, University of Cambridge, UK

UNIVERSITY OF CAMBRIDGE

Alzheimer's Research UK | Make breakthroughs possible

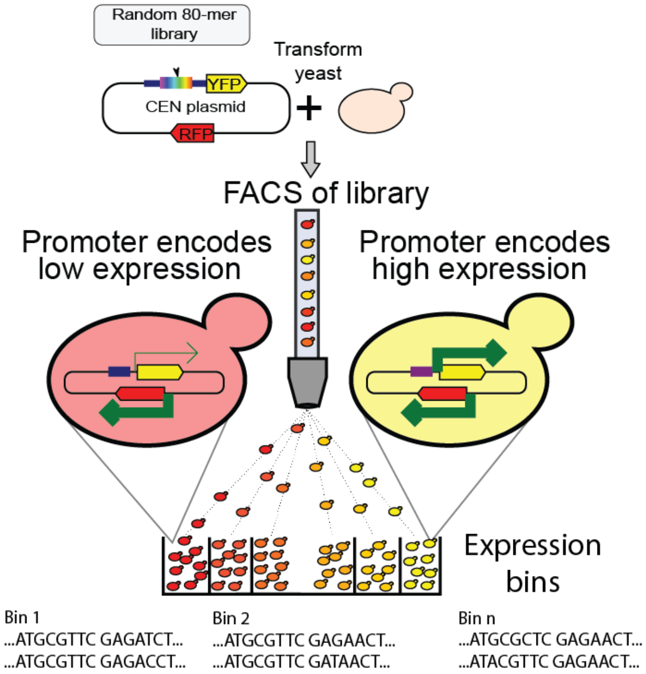
UCL

DRUG DISCOVERY INSTITUTE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

DREAM Challenge 2022

“Predicting gene expression using millions of random promoter sequences”



<u>Task</u>

Train a sequence-to-expression model using 6.7 million random promoter sequences

<u>Data</u>

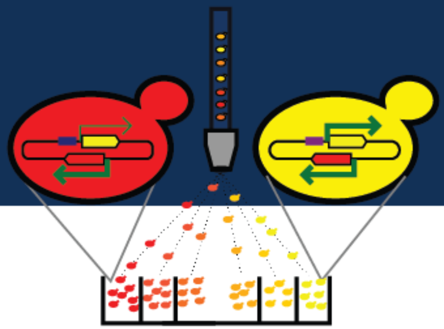
110 nt sequences + expression

<u>Evaluation</u>

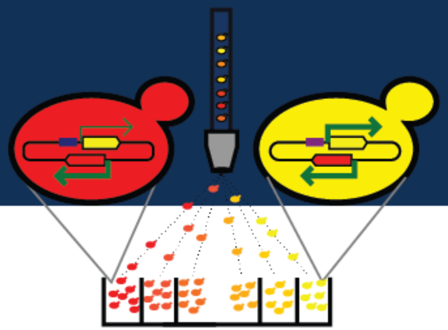
71,103 unseen sequences

Challenge description and image credit: https://www.synapse.org/#!Synapse:syn28469146/wiki/

DREAM challenge results



Predicting gene expression using millions of
random promoter sequences
DREAM Challenge 2022



SBME UBC School of Biomedical Engineering
IBM Research
deep genomics
Google Research TPU Research Cloud
SageBionetworks

Position	Team Name	Mean rank in competition metrics
1	autosome.org	1.01175
2	BHI - dream challenge	1.98825
3	Unlock_DNA	3.6497
4	Camformers	4.5854
5	NAD	5.81105
5	wztr	5.8152
7	High Schoolers Are All You Need	7.21835
8	BioNML	7.93655
9	BUGF	8.5263
10	mt	9.3033

100+ teams participating
27 final submissions

https://www.synapse.org/#!Synapse:syn28469146/wiki/619131

Sequence representation

110 nt sequences and their expression

6,739,258 sequences with known expression
    71,103 sequences to predict

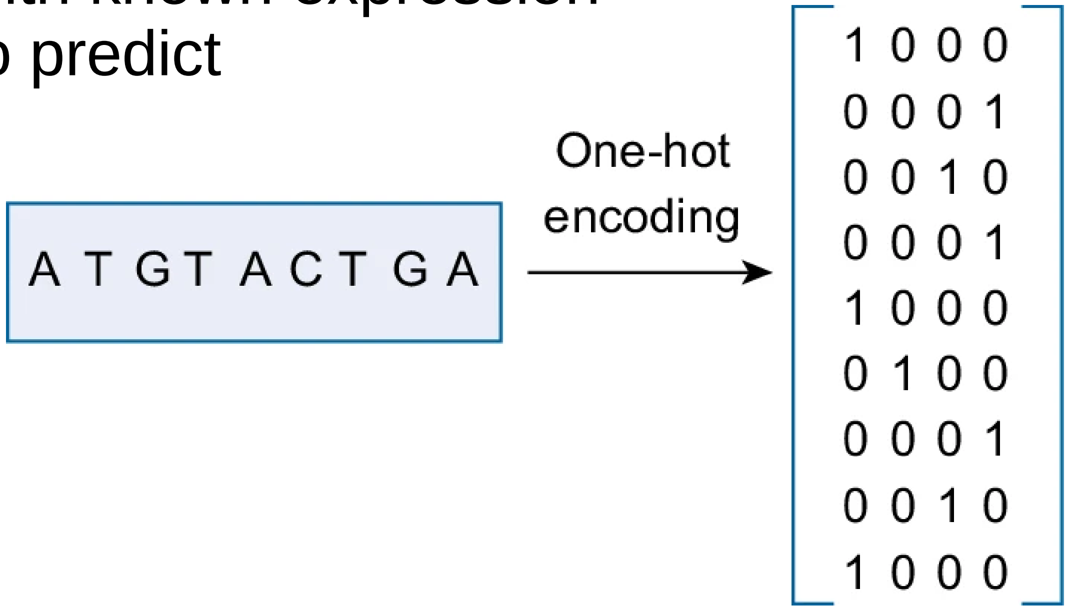
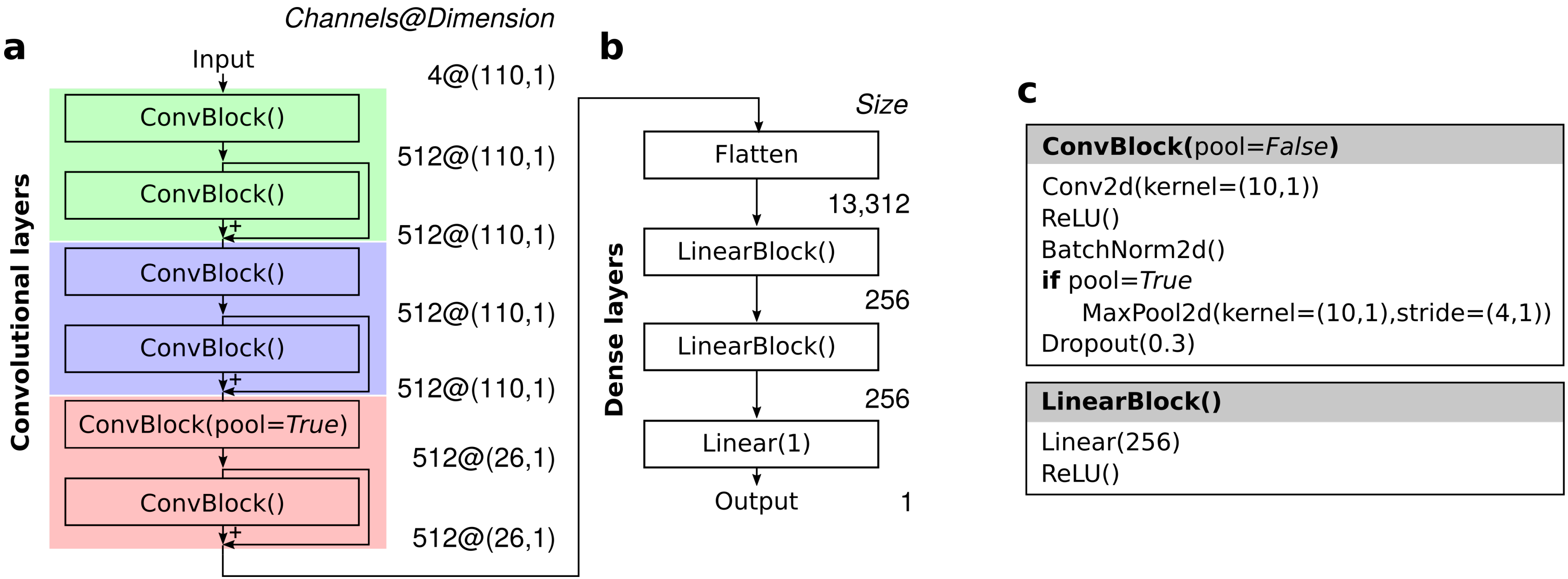


Illustration from
Al-Ajlan & El Allali, 2019

Data processing

- Data inclusion criteria
  - No more than three “N”
  - Length 110 ±3 nt
    - Padding with N and truncation at 110
- Data split
  - Model design and hyperparameter optimisation
    - Training set (72%), validation set (8%), test set (20%)
  - Final submission
    - Training set (90%), validation set (10%)

Model architecture



16,611,073 trainable parameters

<u>Optimisation</u>
AdamW, L1 loss, ReduceLROnPlateau

<u>Early stopping</u>
No improvement in r+ρ for 10 epochs

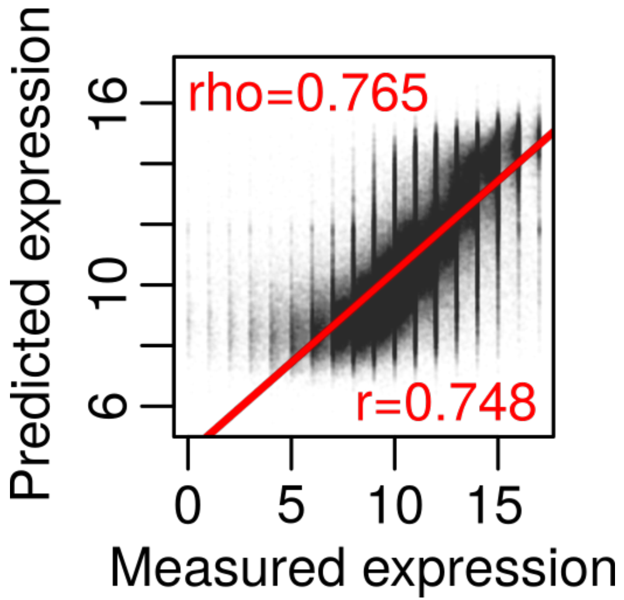
Model training

T, training; V, validation

	T loss	V loss	T r	V r	T rho	V rho	
Epoch 1:	1.3174	1.3255	0.6716	0.7136	0.6909	0.7293	*
Epoch 2:	1.2353	1.1921	0.7153	0.7320	0.7315	0.7503	*
Epoch 3:	1.2092	1.1782	0.7267	0.7382	0.7428	0.7552	*
Epoch 4:	1.1962	1.1802	0.7318	0.7387	0.7480	0.7571	*
Epoch 5:	1.1876	1.1669	0.7350	0.7423	0.7513	0.7594	*
Epoch 6:	1.1806	1.1731	0.7376	0.7436	0.7539	0.7605	*
Epoch 7:	1.1754	1.1589	0.7396	0.7445	0.7559	0.7619	*
Epoch 8:	1.1710	1.1576	0.7412	0.7454	0.7576	0.7627	*
Epoch 9:	1.1671	1.1574	0.7427	0.7452	0.7591	0.7624	
Epoch 10:	1.1634	1.1589	0.7439	0.7455	0.7603	0.7633	*
Epoch 11:	1.1606	1.1650	0.7447	0.7463	0.7613	0.7635	*
Epoch 12:	1.1577	1.1537	0.7457	0.7467	0.7623	0.7641	*
Epoch 13:	1.1551	1.1692	0.7467	0.7463	0.7633	0.7642	
Epoch 14:	1.1528	1.1625	0.7475	0.7463	0.7640	0.7638	
Epoch 15:	1.1507	1.1561	0.7482	0.7469	0.7648	0.7648	*
Epoch 16:	1.1487	1.1576	0.7489	0.7465	0.7655	0.7642	
Epoch 17:	1.1466	1.1590	0.7495	0.7467	0.7661	0.7647	
Epoch 18:	1.1451	1.1608	0.7500	0.7475	0.7667	0.7648	*
Epoch 19:	1.1429	1.1558	0.7508	0.7463	0.7675	0.7640	
Epoch 20:	1.1411	1.1559	0.7513	0.7464	0.7681	0.7639	
Epoch 21:	1.1399	1.1576	0.7517	0.7471	0.7684	0.7647	
Epoch 22:	1.1381	1.1542	0.7524	0.7468	0.7690	0.7644	
Epoch 23:	1.1366	1.1688	0.7528	0.7457	0.7695	0.7637	
Epoch 24:	**1.1200**	**1.1509**	**0.7581**	**0.7481**	**0.7748**	**0.7655**	*
Epoch 25:	1.1161	1.1500	0.7594	0.7480	0.7761	0.7654	
Epoch 26:	1.1138	1.1552	0.7601	0.7479	0.7769	0.7653	
Epoch 27:	1.1123	1.1511	0.7606	0.7478	0.7774	0.7652	
Epoch 28:	1.1110	1.1509	0.7610	0.7478	0.7778	0.7653	
Epoch 29:	1.1096	1.1510	0.7614	0.7479	0.7783	0.7653	
Epoch 30:	1.1085	1.1510	0.7618	0.7479	0.7786	0.7653	
Epoch 31:	1.1074	1.1500	0.7621	0.7481	0.7790	0.7654	
Epoch 32:	1.1064	1.1534	0.7624	0.7476	0.7793	0.7651	
Epoch 33:	1.1058	1.1539	0.7626	0.7470	0.7795	0.7647	
Epoch 34:	1.1050	1.1559	0.7628	0.7474	0.7797	0.7651	

Performance on validation set

(660,559 sequences)



Model fine-tuning

- Optuna
  - Batch size, learning rate, weight decay,
  - Kernel size, number of layers, number of channels, dropout rates
  - Position(s) of max pooling
    - Max pooling operation at the penultimate layer improved generalisation

Tricks that did not work (in our hands...)

- Manipulation of target values
  - Quantile normalization (**no difference**)
  - Add noise (**no difference**)
- Preprocessing
  - Extend/shorten flanking sequence length (**no difference**)
- Data augmentation
  - Upsample distribution tails (**no difference** or **reduced performance**)
  - Extend and cut sequences (**no difference**)
  - Reverse-complement sequences (**reduced performance**)

Caveat: Evaluation was done on a model that differs from the final submission.

Performance breakdown

Category	# Sequences	PearsonR	Rank	Spearman	Rank	Weight
All	71103	0.956	6th	0.961	5th	1.00
High expression	968	0.557	8th	0.575	6th	0.30
Low expression	997	0.622	6th	0.611	3rd	0.30
Native	578	0.825	5th	0.818	6th	0.30
Random	6349	0.968	6th	0.972	5th	0.30
Challenging	1953	0.941	4th	0.940	3rd	0.50
SNVs	44340	0.821	5th	0.674	7th	1.25
TFBS perturbation	3287	0.975	4th	0.959	8th	0.30
Motif tiling	2624	0.899	16th	0.912	9th	0.40

Based on all data with no subsampling

100+ teams participating
27 final submissions

ScorePearsonR²	0.753	5th
ScoreSpearman	0.821	4th
Overall	**0.787**	**4th**

https://www.synapse.org/#!Synapse:syn28469146/wiki/619131

Contact

Fredrik Svensson, Senior Research Associate
Alzheimer's Research UK UCL Drug Discovery Institute, University College London
The Cruciform Building, Gower Street, London, WC1E 6BT, United Kingdom

Maria-Anna Trapotsi, Bioinformatician
Cancer Research UK Cambridge Institute, University of Cambridge
Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, United Kingdom

Susanne Bornelöv, Senior Research Associate
Cancer Research UK Cambridge Institute, University of Cambridge
Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, United Kingdom
susanne.bornelov@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Alzheimer's Research UK | Make breakthroughs possible
UCL
DRUG DISCOVERY INSTITUTE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE