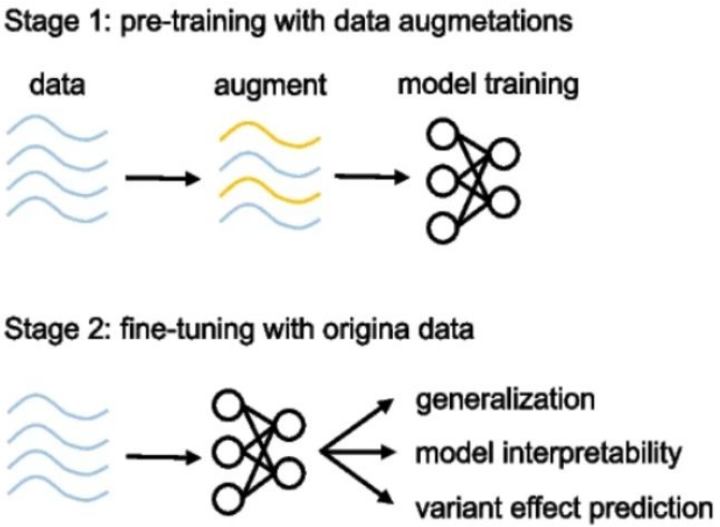
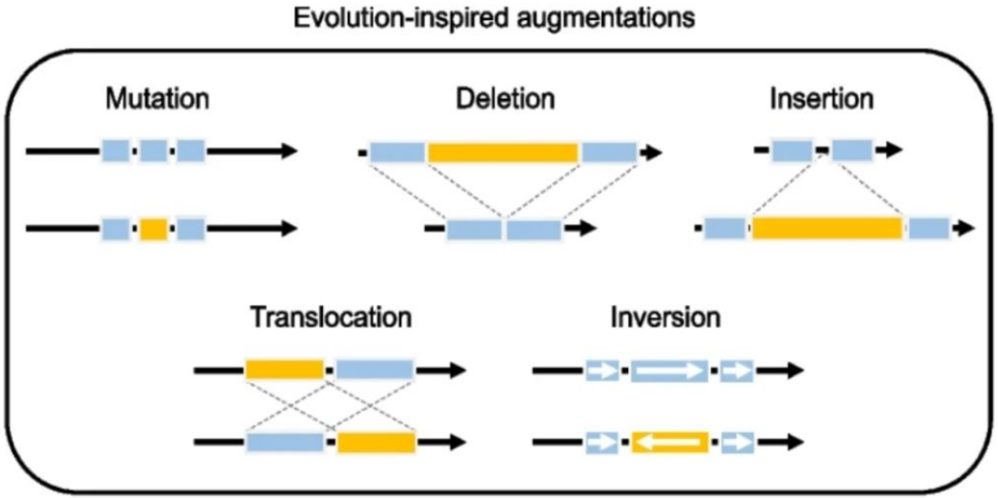
Application of evolutionary augmentations to work with genomic sequences

Penzin N.[1], Penzar D.[1,2]
[1] Lomonosov Moscow State University, Moscow, Russia
[2] Institute of Protein Research, RAS, Pushchino, Russia

Framework EvoAug substantially improves the generalization and interpretability of established DNNs across prominent regulatory genomics prediction tasks.

The purpose of this work is to test the effectiveness and test the framework for various tasks of regulatory genomics.



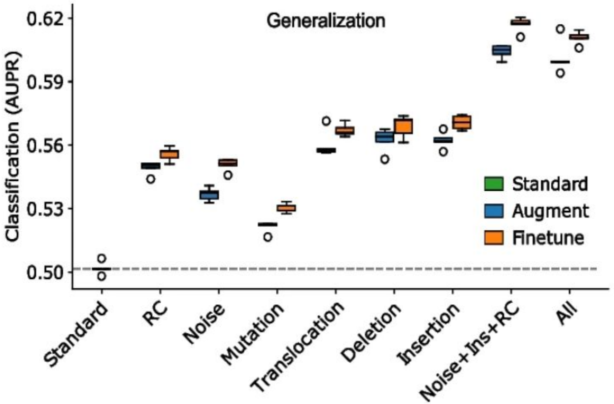


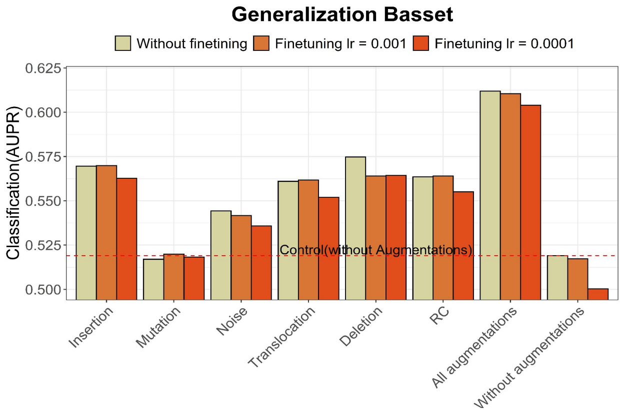
Lee, N.K., Tang, Z., Toneyan, S. *et al.* EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biol* 24, 105 (2023). https://doi.org/10.1186/s13059-023-02941-w

Results reproducibility

The Basset dataset consists of a multi-task binary classification of chromatin accessibility sites across 161 cell types/tissues.

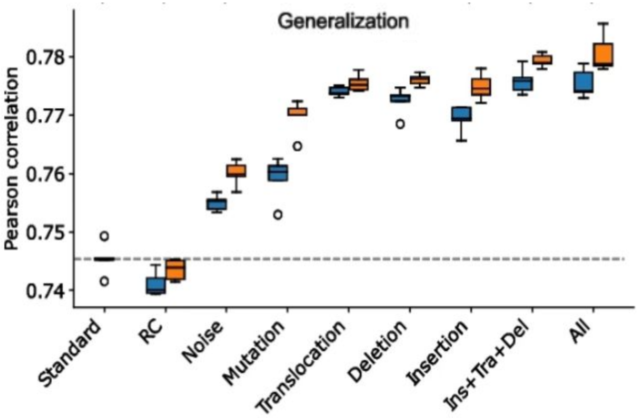
Top left graphic - article Basset results
Top right graphic - reproduced Basset results

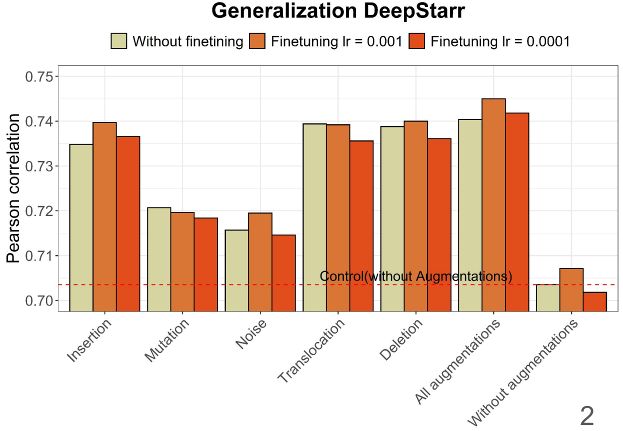




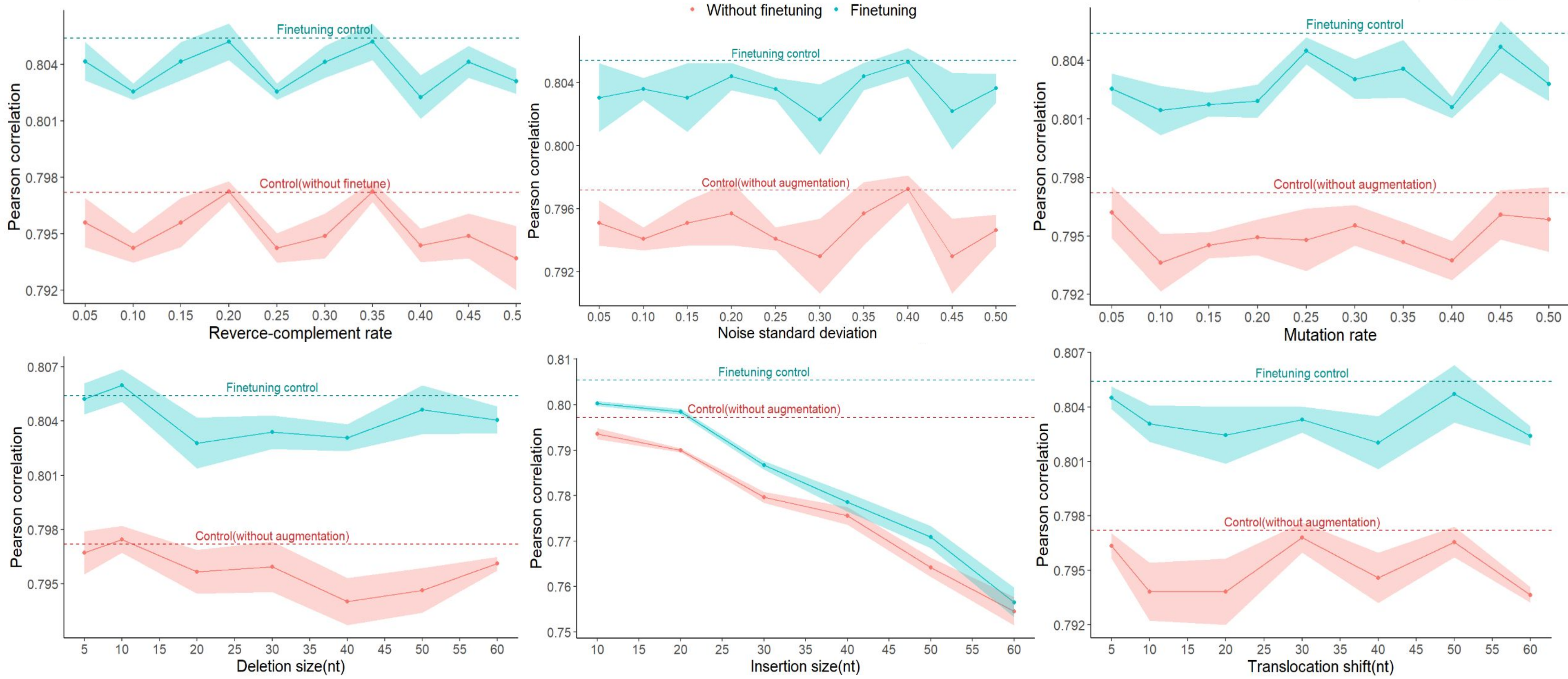
The DeepSTARR dataset consists of a multi-task regression of enhancer activity promoters in D. melanogaster S2 cells.

Bottom left graphic - article DeepStarr results
Bottom right graphic - reproduced DeepStarr results





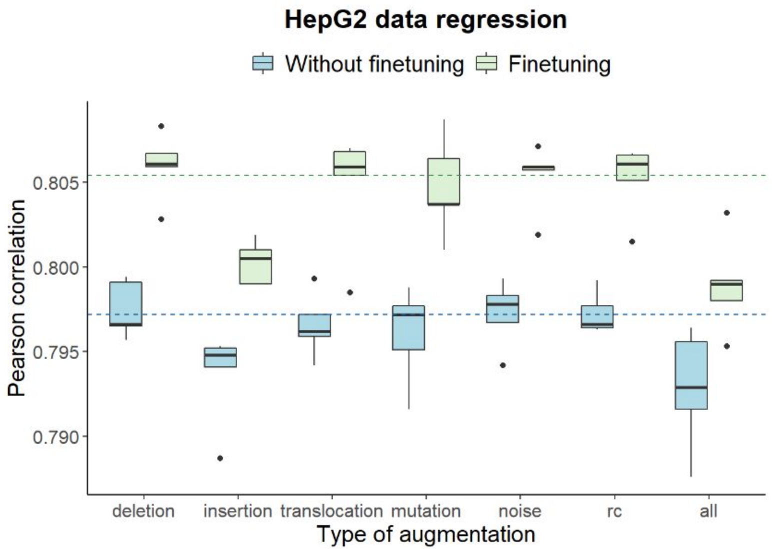
Prediction of human MPRA gene expression



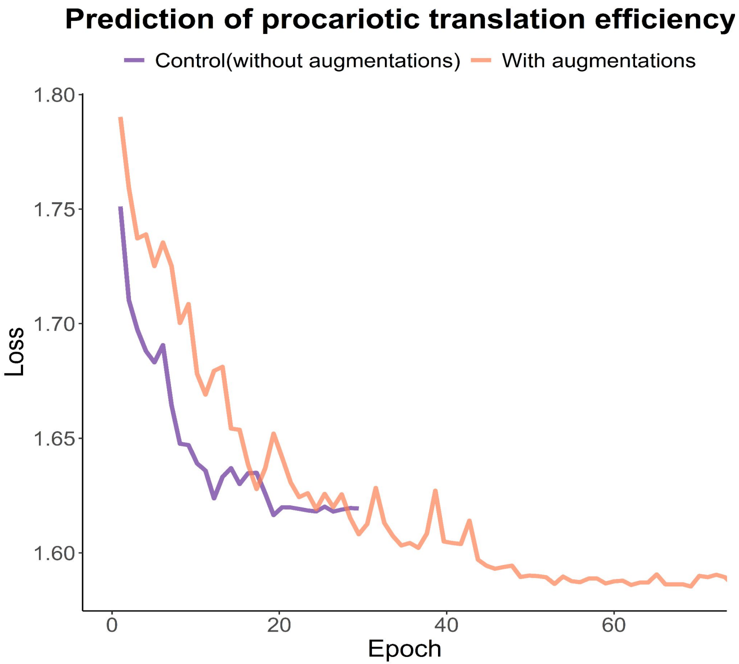
Hyperparameter sweep of each augmentation method for HepG2 data using LegNet.

Prediction of human MPRA and prokaryotic gene expression

HepG2 - human hepatocyte MPRA data.



Dashed blue and green lines represent the performance without augmentations (standard) without fine-tuning and with fine-tuning respectively.



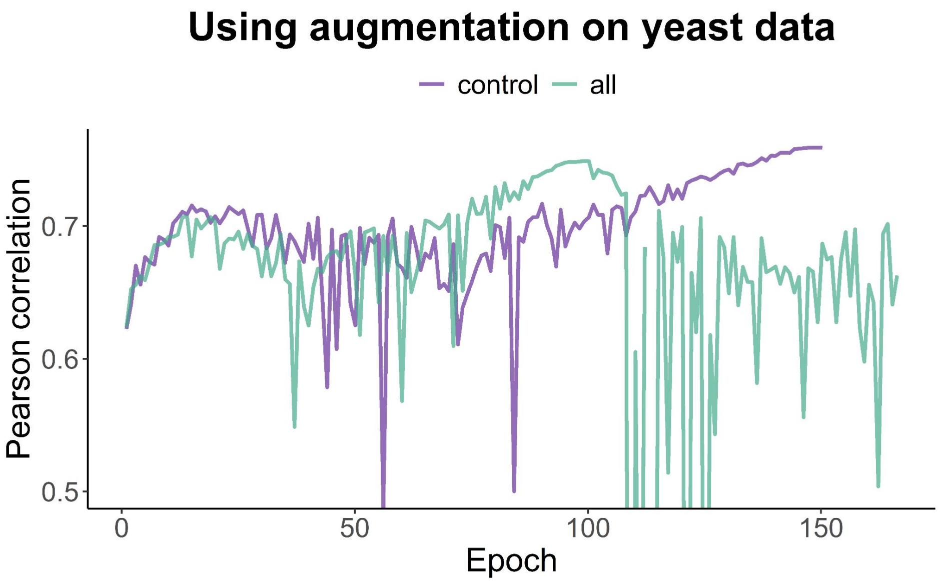
For human MPRA data EvoAug augmentations may provide improvements in individual cases, but do not improve the quality of predictions using all or several augmentations simultaneously.

EvoAug augmentations leads to improving generalisation DNN capable of predicting the level of prokaryotic translation efficiency.

Prediction of yeast GPRA gene expression

	Mean Random pearson ± STD	**Mean SNV's pearson ± STD**
With augmentation	0,9576 ± 0,0007	0,7398 ± 0,0107
Fine-tuning after augmentations	0,9552 ± 0,0005	0,7352 ± 0,0019
Standard	0,9614 ± 0,0006	0,7884 ± 0,0015

Conclusion: For some regulatory genomic problems the EvoAug framework actually allows one to obtain a model with better quality than the original ones. However, for some datasets it may not lead to an improvement in the quality of prediction at all or even result in suboptimal results. Moreover, further model fine-tuning suggested by authors may not allow it to get rid of bias in the data introduced by augmentations and in general doesn't result in performance gains.



EvoAug checking runs were performed for yeast GPRA data so LegNet's prediction quality being lower than training without EvoAug. Using augmentations did not improve prediction of yeast GPRA gene expression.