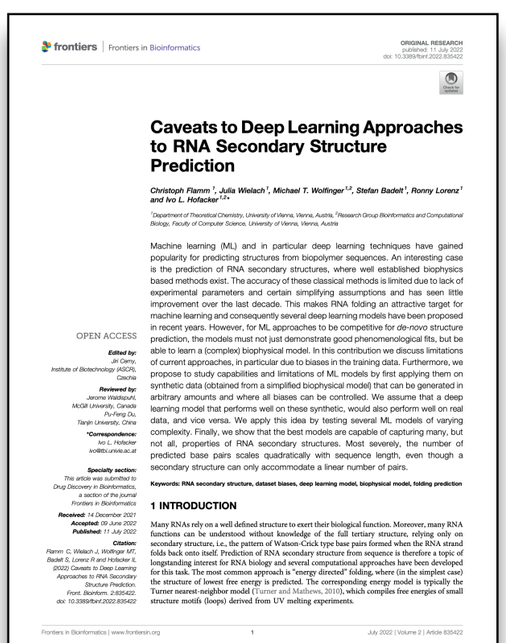
Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction

Christoph Flamm, Julia Wielach, Michael T. Wolfinger, Stefan Badelt, Ronny Lorenz, Ivo L. Hofacker

Figures and title from :

Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction
Flamm, C., Wielach, J., Wolfinger, M.T., Badelt, S., Lorenz, R., and Hofacker, I.L.
Front. Bioinform. 2:835422 (2022)
doi: 10.3389/fbinf.2022.835422



frontiers | Frontiers in Bioinformatics

ORIGINAL RESEARCH
published: 11 July 2022
doi: 10.3389/fbinf.2022.835422

Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction

*Christoph Flamm [1], Julia Wielach [1], Michael T. Wolfinger [1,2], Stefan Badelt [1], Ronny Lorenz [1] and Ivo L. Hofacker [1,2]**

[1]Department of Theoretical Chemistry, University of Vienna, Vienna, Austria, [2]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

Machine learning (ML) and in particular deep learning techniques have gained popularity for predicting structures from biopolymer sequences. An interesting case is the prediction of RNA secondary structures, where well established biophysics based methods exist. The accuracy of these classical methods is limited due to lack of experimental parameters and certain simplifying assumptions and has seen little improvement over the last decade. This makes RNA folding an attractive target for machine learning and consequently several deep learning models have been proposed in recent years. However, for ML approaches to be competitive for *de-novo* structure prediction, the models must not just demonstrate good phenomenological fits, but be able to learn a (complex) biophysical model. In this contribution we discuss limitations of current approaches, in particular due to biases in the training data. Furthermore, we propose to study capabilities and limitations of ML models by first applying them on synthetic data (obtained from a simplified biophysical model) that can be generated in arbitrary amounts and where all biases can be controlled. We assume that a deep learning model that performs well on these synthetic, would also perform well on real data, and vice versa. We apply this idea by testing several ML models of varying complexity. Finally, we show that the best models are capable of capturing many, but not all, properties of RNA secondary structures. Most severely, the number of predicted base pairs scales quadratically with sequence length, even though a secondary structure can only accommodate a linear number of pairs.

Keywords: RNA secondary structure, dataset biases, deep learning model, biophysical model, folding prediction

OPEN ACCESS

Edited by:
Jiří Černý,
Institute of Biotechnology (ASCR), Czechia

Reviewed by:
Jerome Waldispühl,
McGill University, Canada
Pu-Feng Du,
Tianjin University, China

***Correspondence:**
Ivo L. Hofacker
ivo@tbi.univie.ac.at

Specialty section:
This article was submitted to Drug Discovery in Bioinformatics, a section of the journal Frontiers in Bioinformatics

Received: 14 December 2021
Accepted: 09 June 2022
Published: 11 July 2022

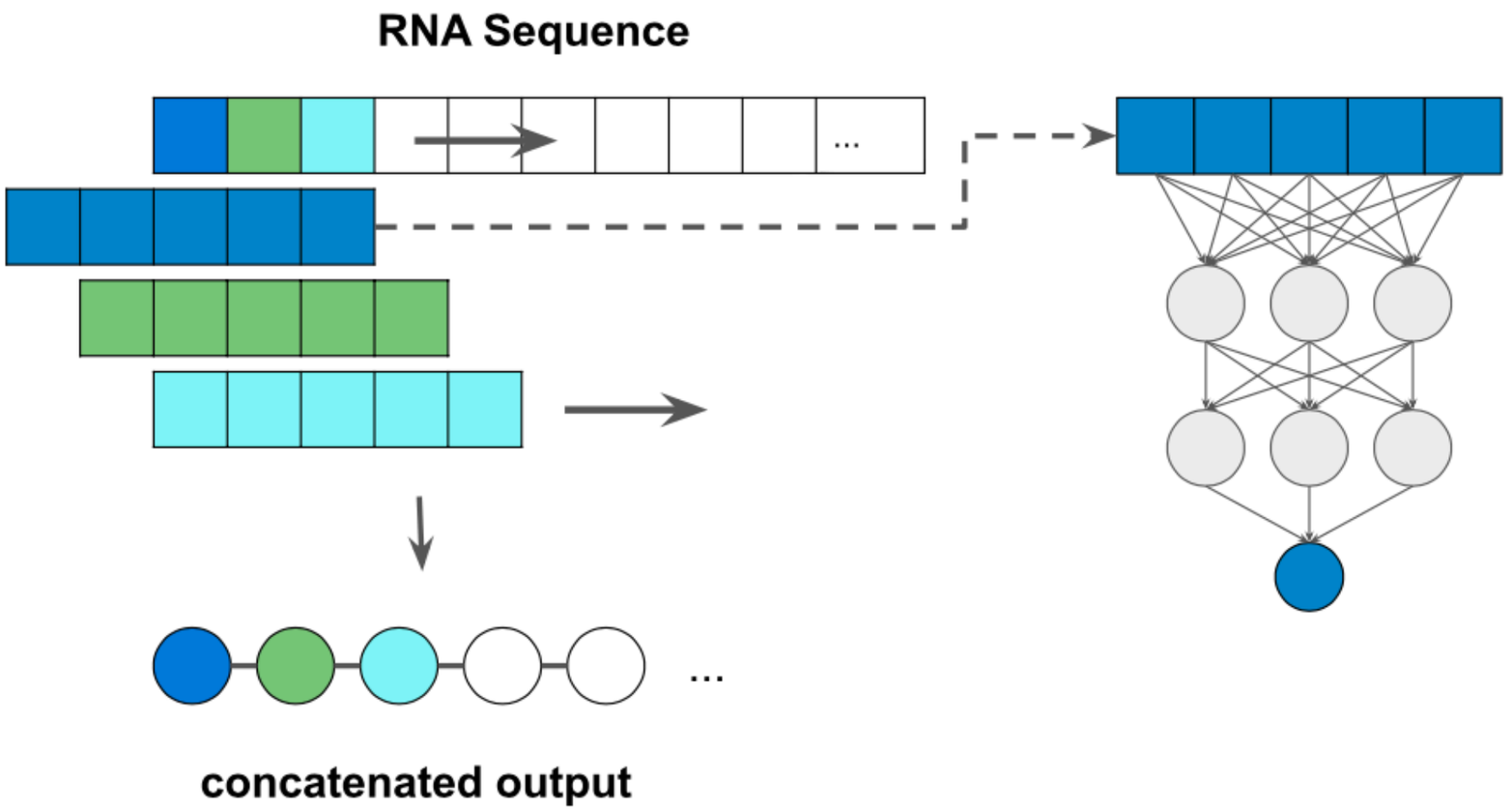
Citation:
Flamm C, Wielach J, Wolfinger MT, Badelt S, Lorenz R and Hofacker IL (2022) Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction. Front. Bioinform. 2:835422. doi: 10.3389/fbinf.2022.835422

1 INTRODUCTION

Many RNAs rely on a well defined structure to exert their biological function. Moreover, many RNA functions can be understood without knowledge of the full tertiary structure, relying only on secondary structure, i.e., the pattern of Watson-Crick type base pairs formed when the RNA strand folds back onto itself. Prediction of RNA secondary structure from sequence is therefore a topic of longstanding interest for RNA biology and several computational approaches have been developed for this task. The most common approach is "energy directed" folding, where (in the simplest case) the structure of lowest free energy is predicted. The corresponding energy model is typically the Turner nearest-neighbor model (Turner and Mathews, 2010), which compiles free energies of small structure motifs (loops) derived from UV melting experiments.

Frontiers in Bioinformatics | www.frontiersin.org 1 July 2022 | Volume 2 | Article 835422

Published under a Creative Commons Attribution CC BY License
http://creativecommons.org/licenses/by/4.0/



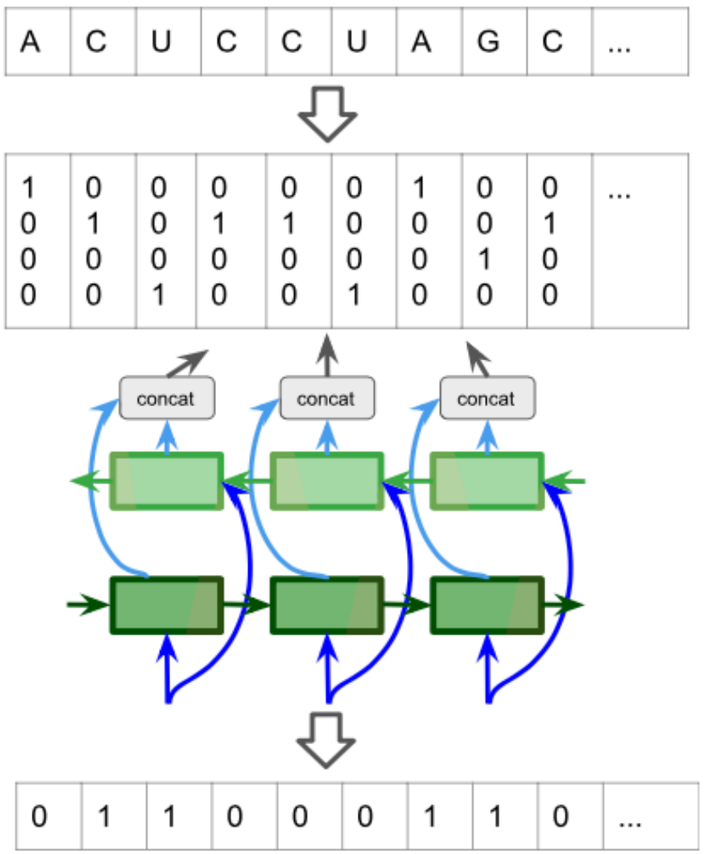
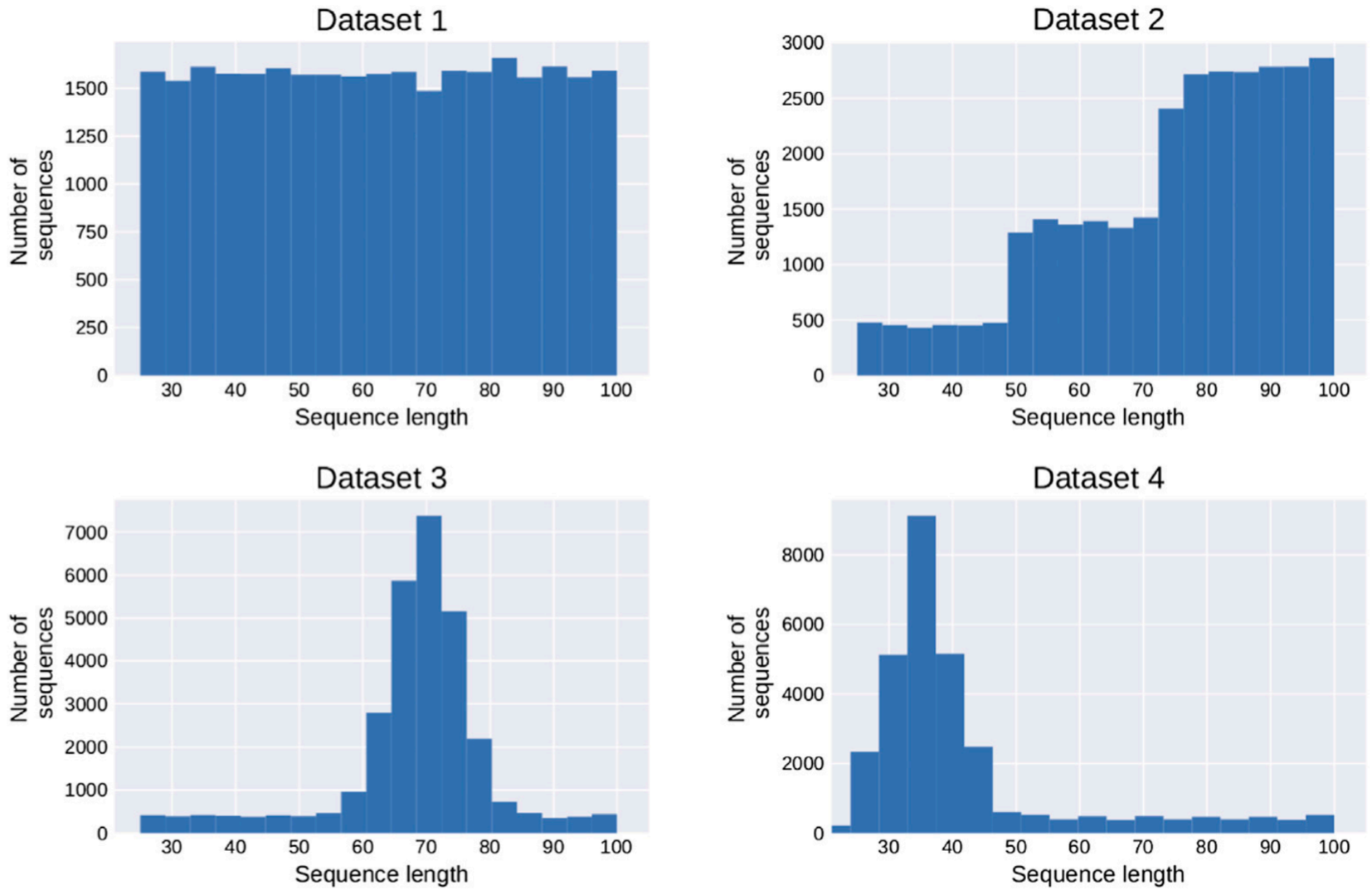


Figure from doi:10.3389/fbinf.2022.835422



michaelwolfinger.com

Figure from doi:10.3389/fbinf.2022.835422

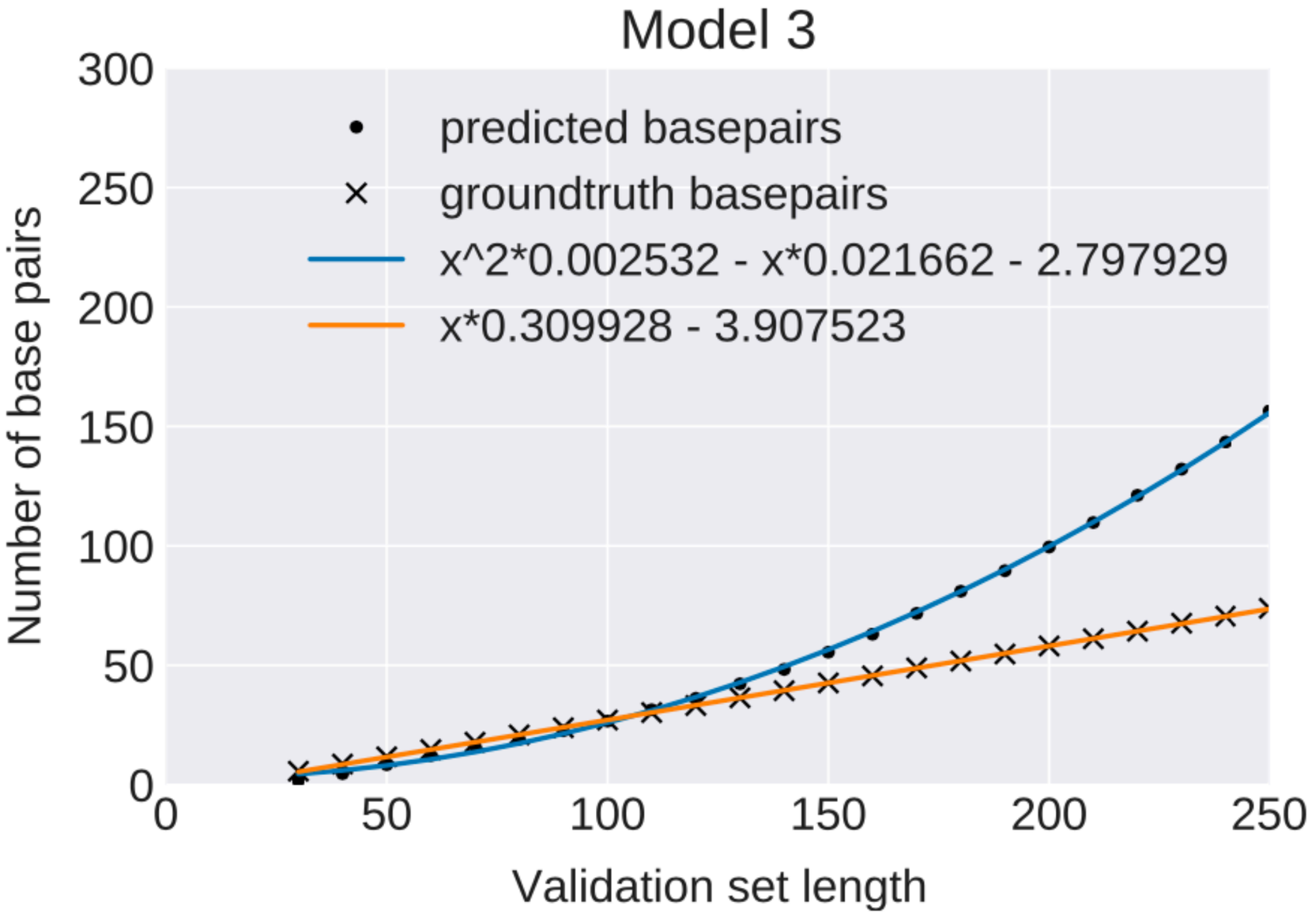
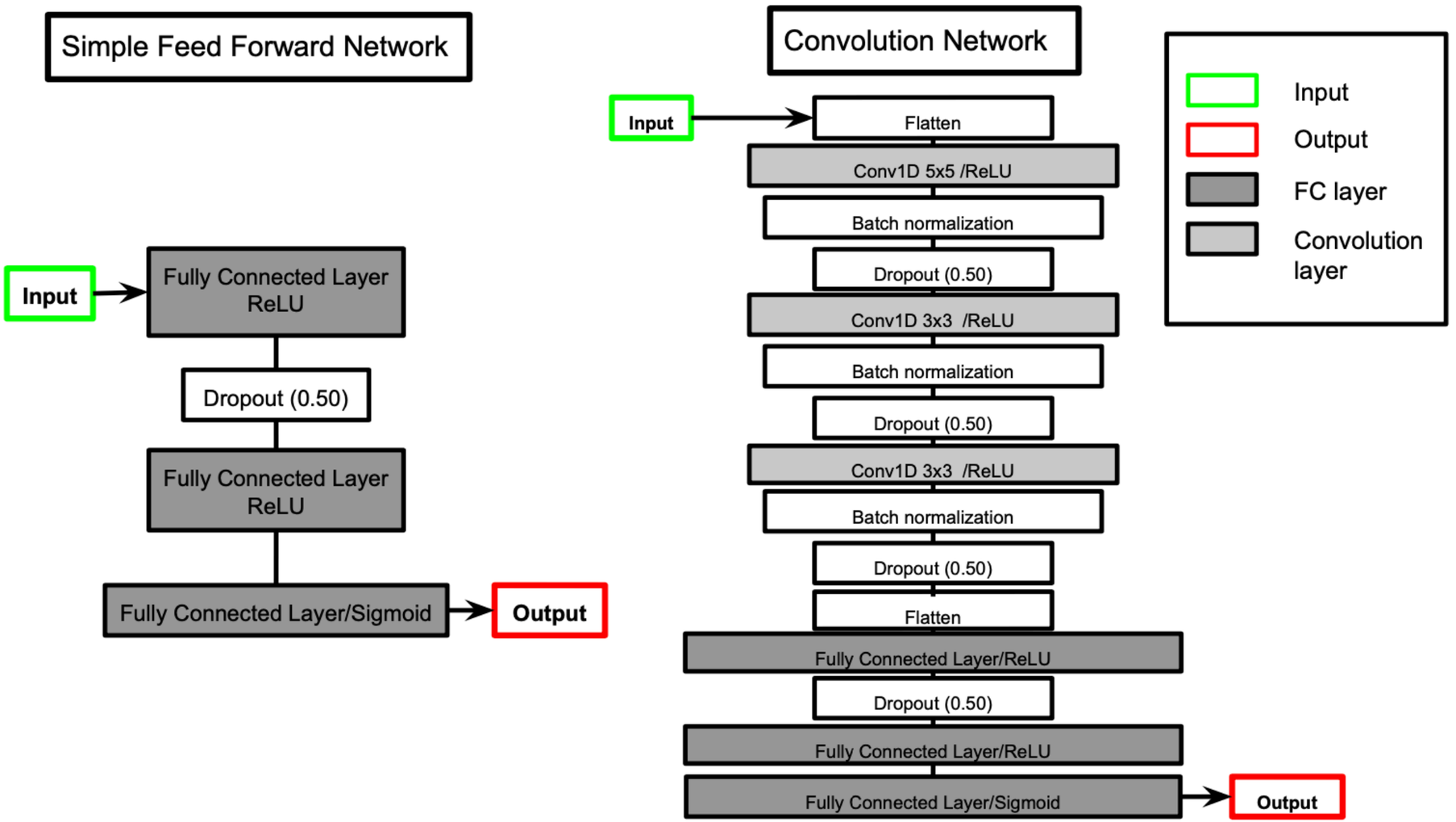


Figure from doi:10.3389/fbinf.2022.835422

Modeltype	Parameters	Epochs	Accuracy	F1	Loss	MCC
BLSTM	1 Layer, 40 Neurons	43	0.667	0.594	0.609	0.166
	1 Layer, 80 Neurons	27	0.664	0.589	0.612	0.168
	3 Layers, 40 Neurons	38	0.676	0.609	0.604	0.207
Sliding Window	Window 15	89	0.654	0.559	0.623	0.120
	Window 35	94	0.659	0.559	0.620	0.118
	Window 71	59	0.661	0.569	0.618	0.118
CNN Sliding Window	Window 15	67	0.660	0.588	0.616	0.156
	Window 35	65	0.666	0.586	0.609	0.166
	Window 71	30	0.668	0.580	0.608	0.170

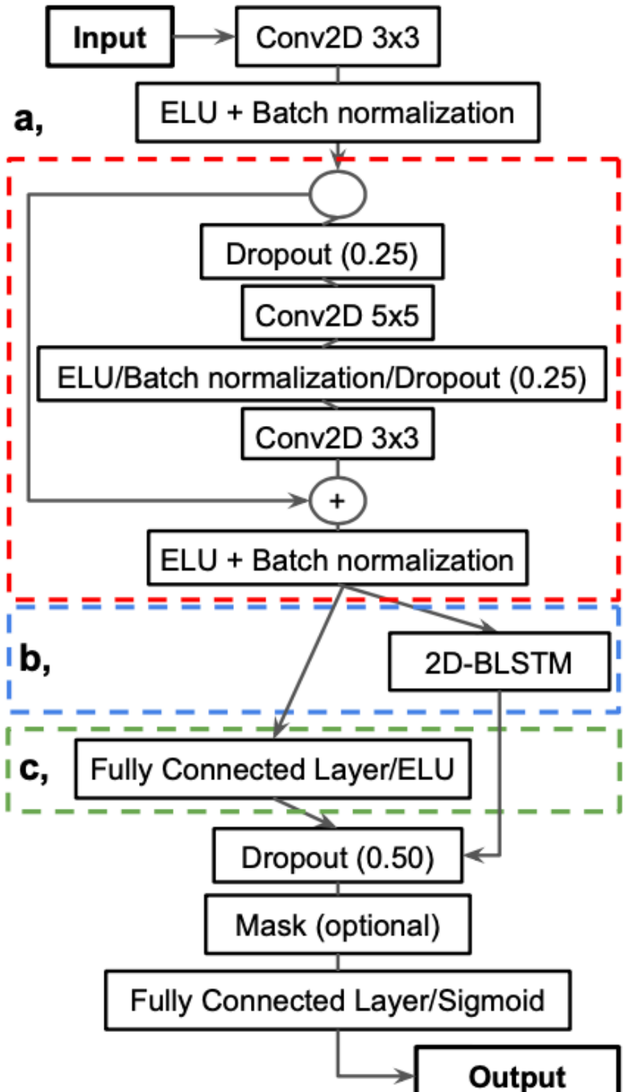
michaelwolfinger.com

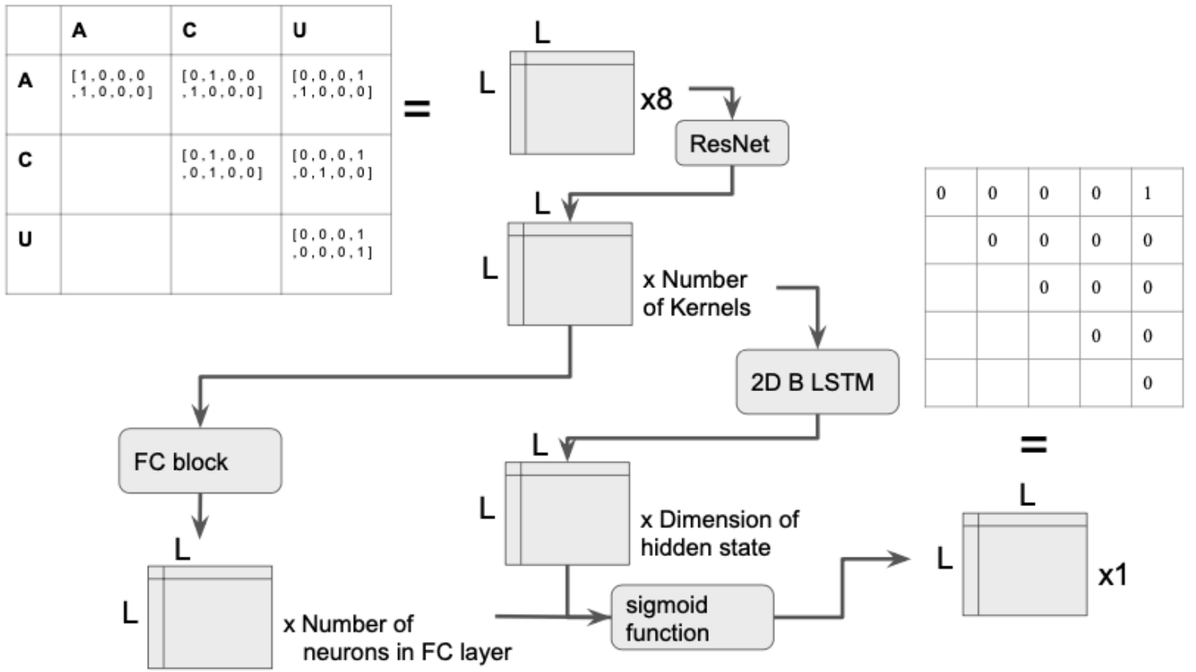
Figure from doi:10.3389/fbinf.2022.835422



michaelwolfinger.com

Figure from doi:10.3389/fbinf.2022.835422





Model	Filters	ResNet blocks	LSTM blocks	LSTM neurons	FC blocks	FC neurons
0	64	20	0	0	2	512
1	48	16	0	0	2	512
3	64	30	1	200	0	0

michaelwolfinger.com

Figure from doi:10.3389/fbinf.2022.835422

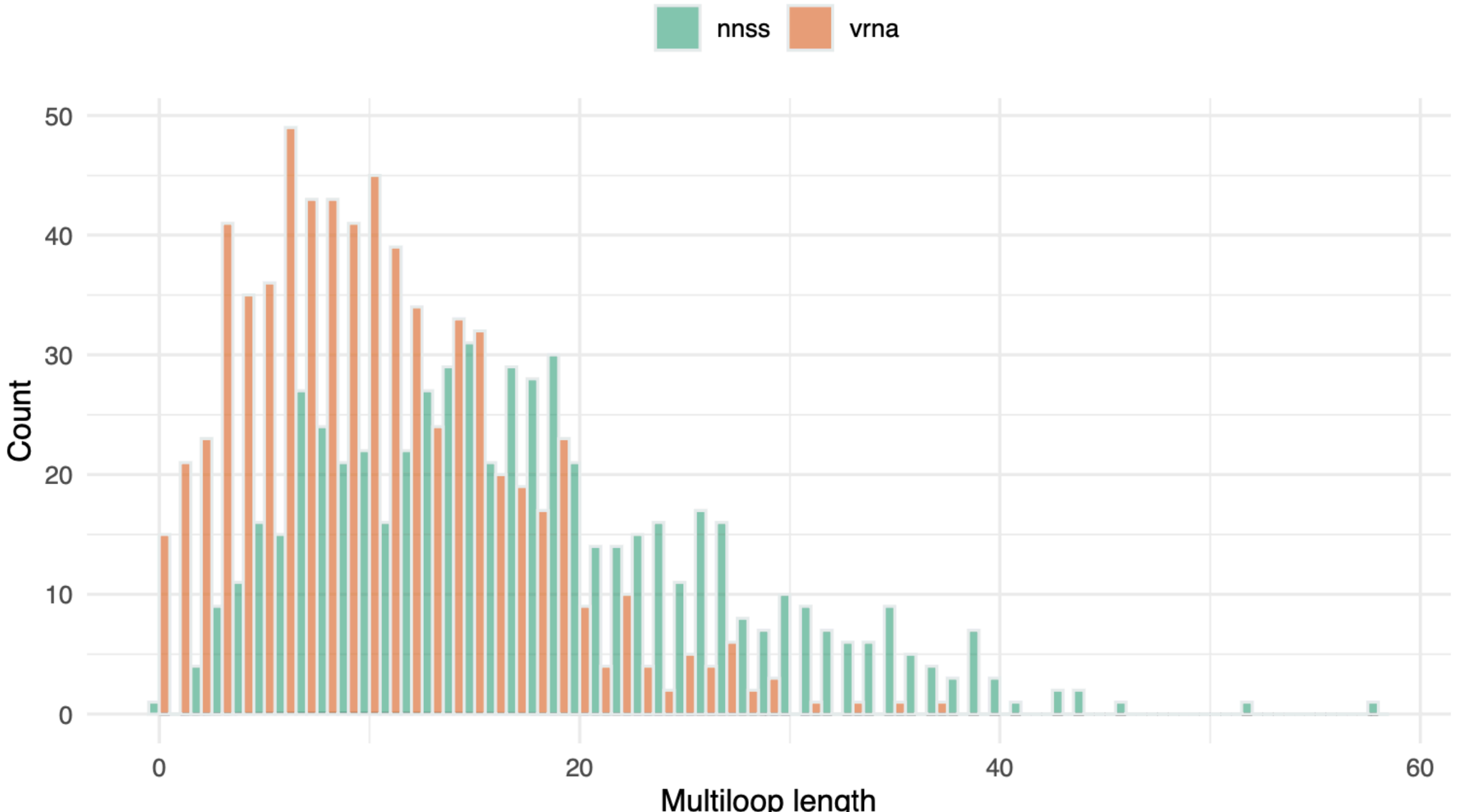


Figure from doi:10.3389/fbinf.2022.835422