Come with the Sequence, Stay for the Structure: RNA Conformational Learning to Predict Protein Expression

Anonymous Authors[1]

Abstract

Developing efficient representations of genomic sequences is crucial to unlock the full potential of machine learning for bioengineering applications. One of the key challenges is predicting protein expression directly from DNA sequence, and has been traditionally tackled by sequence-focused biophysical or deep learning models. Here, we show that these models are prone to covariate shift, severely limiting their accuracy. To overcome this, we propose the use of a graph attention network that integrates additional RNA structural information that can be inferred from a DNA sequence. We find the generalization of our model outperforms purely sequence-based models on several data sets, bringing us closer to reliable general-purpose prediction of protein expression.

1. Introduction

RNAs play several roles in living cells, acting as intermediate information carriers and as regulators of biological processes through their sequence and structural conformations (Mathews et al., 2010). One process that RNA strongly regulates is the translation of protein. This is strongly dependant on sequences at and near the point where protein translation initiates called the 5' Untranslated Region (5'UTR) of an RNA. Varying the sequence of the 5'UTR is commonly used to fine-tune the protein translation rate of an RNA. Control of protein synthesis rate is vital in virtually all biological engineering applications, enabling the optimisation of cellular processes, such as the production of biologics. However, the relationship between sequence and the rate of protein expression is complex and still unclear, hampering the development of models to unlock the promise of engineered biology.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

To overcome this challenge, data-centric models have gained traction over recent years, leveraging new high-throughput experiments that are able to generate vast corpuses of sequence to protein expression level mappings. Sequence-based deep learning models can often outperform thermodynamics models of protein expression prediction. This success can in part be attributed to the ability of these models to infer important sequence motifs or biological process that may yet to be discovered. However, sequence-based models typically fail to generalize outside their training set, severely limiting wider use (Gilliot & Gorochowski, 2023a).

A possible reason for this lack of generalisation is the challenge of integrating data beyond sequence alone. For example, the structure of RNAs is known to play a key role in regulating translation initiation. While purely structural modelling approaches have failed to achieve comparable predictive performance to sequence-based models (Angenent-Mari et al., 2020), it remains unclear whether sequence-based models can effectively recover RNA structure. Compounding this challenge is the fact that most existing data sets only vary the 5'UTR or the protein coding sequence and not both simultaneously, which introduces bias in the data due to the non-additive effects between these regions (Goodman et al., 2013).

A recent high-throughput experiment (Höllerer & Jeschek, 2023) collected data where the CDS and 5'UTR were simultaneously mutated, which could support the development of less biased models. In this paper, we propose the use of Graph Neural Networks (GNNs) to leverage this corpus and model the mapping between sequence and protein abundance. GNNs enable us to integrate both sequence and structural information, which we show provides improved generalization performance across sequence contexts. To the best of our knowledge, this work constitutes the first attempt at using a GNN framework to predict protein expression from RNA sequence and structure.

2. Related work

2.1. Predicting RNA secondary structure

Messenger RNAs (mRNAs) in a cell are single stranded nucleic acid polymers which are able to form secondary

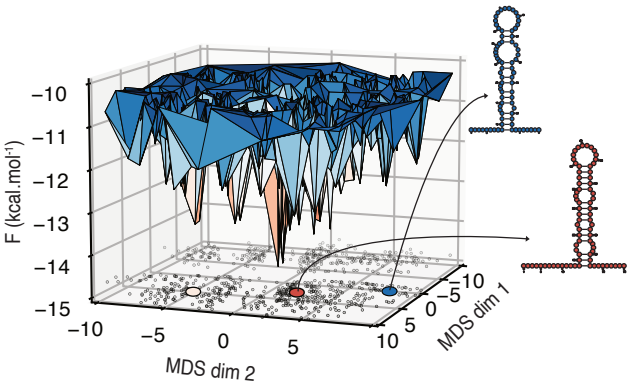

Figure 1. Conformational landscape of secondary structures for a single RNA sequence computed using MDS. Red point corresponds to the structure with the minimum free energy (MFE), while blue point corresponds to a structure with a lower probability.

structures through nucleotide base pairing. Due to thermal fluctuations and differences in base pair affinities, their structures are highly dynamic (Ganser et al., 2019) with some conformations being more probable than others (Figure 1). The energetics of RNA secondary structures can be calculated using physics-based models (Zuker & Stiegler, 1981), with parameters either derived from optical melting experiments (Xia et al., 1998) or learnt from examples (Do et al., 2006). Using these biophysical models, efficient dynamic programming algorithms can be used to predict the structure of the minimum free energy (MFE) conformation, i.e., the most likely observed structure (see Figure 1).

The dynamic nature of RNA structure means that the MFE structure may occur with a low probability (Mathews et al., 2010). However, it is often the only structure considered when incorporating structure into bioinformatic analyses. Accounting for the fluctuations in RNA secondary structure generally leads to better predictions and can be done by computing the base pairing probabilities between all pairs of nucleotides (McCaskill, 1990). Conditioned on a sequence x, the probability $p(i, j|x)$ that nucleotide i and j form a base pair is obtained through a Boltzmann equilibrium distribution:

$$p(i, j|x) = \sum_{\sigma \in S(i,j,x)} p(\sigma|x), \text{ with } p(\sigma|x) = \frac{e^{-\frac{E(\sigma|x)}{RT}}}{Z(x)},$$

where $p(\sigma|x)$ is the probability of observing the structure σ, $S(i, j, x)$ is the ensemble of secondary structures with a base pair between nucleotide i and j, $Z(x)$ is the partition function and $E(\sigma|x)$ is the free energy associated with the structure σ, and R and T are two physical constants. For an RNA sequence of length n, the set of all base pair probabilities can be computed in $\mathcal{O}(n^3)$ time using dynamic programming (McCaskill, 1990).

2.2. Predicting protein abundance

There have been numerous attempts to predict translation efficiency from RNA sequence alone using thermodynamic models (Terai & Asai, 2020). However, these types of models typically lack sufficient accuracy for many bioengineering tasks (Angenent-Mari et al., 2020). Improvements have been seen when using deep learning models based on a one-hot encoding of input sequence and convolutional (Höllerer et al., 2020), recurrent (Angenent-Mari et al., 2020) and hybrid convolutional-recurrent neural network architectures (Valeri et al., 2020; Gilliot & Gorochowski, 2023b). These models often display improved accuracy on test data sets, but struggle to generalise to highly diverse sequences and require transfer learning approaches to reach good accuracy (Gilliot & Gorochowski, 2023b).

The information content of RNA structure has been explored by directly using the output from RNA secondary structure prediction models, including properties such as: 1. a sliding window average of the base pairing probabilities (Goodman et al., 2013), 2. the accessibility around the start codon (Terai & Asai, 2020) and 3. the MFE (Kudla et al., 2009; Angenent-Mari et al., 2020). However, while offering promising directions, the limited processing of the structural measures often hampers performance.

To improve accuracy, we use a GNN to process the ensemble of RNA secondary structures. The most similar approach is by Yan et al. (2020) who proposed RPI-Net, a convolution-based graph neural network (GCN) to predict protein-RNA interaction based on the RNA secondary structure. However, GCNs are limited in their ability to effectively handle multidimensional edge information (Schlichtkrull et al., 2017). In the task considered here, it is important to distinguish between covalent and hydrogen bonds, and the varying probabilities of base pairing which range over 10^4. Therefore, we adopt a variant of the Graph Attention Network (GAT) (Veličković et al., 2018) with a more advanced and dynamic architecture that can better process important edge features.

3. Approach

3.1. Datasets

To maximize the diversity of RNA secondary structures our model was exposed to, we selected data sets that exhibited global sequence diversity (i.e., did not vary only a small region of the RNA). For model training, we use the data set from Höllerer & Jeschek (2023) (Hollerer_23), which characterized the strength of 1.2 million generic constructs with varying 5'UTR and protein coding sequences. The original data set was subsampled to obtain a balanced distribution of library types and protein expression levels. This resulted in a training/validation/test data sets of 318,550/30,000/30,000 samples, respectively. To simplify comparisons, protein

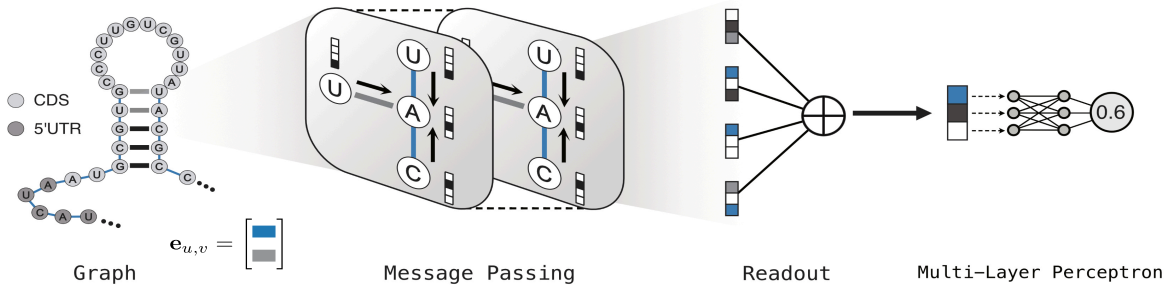

Figure 2. Processing RNA graph structure data. The graph is initially processed through multiple rounds of message passing. Subsequently, a readout layer aggregates the maximum value across all features, focusing on a window of 40 nt surrounding the AUG start codon. Finally, this information is processed by a standard Multilayer Perceptron (MLP) to predict the protein abundance.

expression levels were normalised between 0 (min) and 1 (max).

Two other data sets were used to evaluate the generalization of our models. The first from Höllerer et al. (2020) (Hollerer_20) was collected using the same experimental protocol by the same laboratory, but only the 5'UTR region was varied. This test set included 30,000 examples with balanced protein levels. We also used data collected using a different experimental protocol by Cambray et al. (2018), which measured protein expression for a library of variants where only the CDS was varied. The mean log-fluorescence was used as proxy for protein expression for the 30,000 examples in the test set and was inferred using FORECAST (Gilliot & Gorochowski, 2023a), followed by a normalization by the maximum mean log fluorescence value. To account for differences in units, we only compared the slope's magnitude when linearly regressing the predictions as a function of the ground truth.

3.2. Sequence-based models

To act as a baseline for the generalisation performance of the graph-based models, we used a hybrid Convolutional-LSTM (Conv-LSTM) model which has proven effective in similar tasks (Gilliot & Gorochowski, 2023a). After hyper-parameter optimisation (Table 2), the successful architecture directly processes the one-hot encoded sequence through a first stage of 1D convolutional layers, followed by a BiLSTM layer, and finally a multi-linear perceptron (MLP), to return the protein expression level.

3.3. RNA graph representation

Two graphs representations of RNA structure were investigated (Appendix A). In the first, a graph was used to represent the MFE RNA structure, where each node represents a nucleotide with a feature including a one-hot encoding of the nucleotide type. Nucleotides linked by a covalent bond were connected in the graph through an edge with

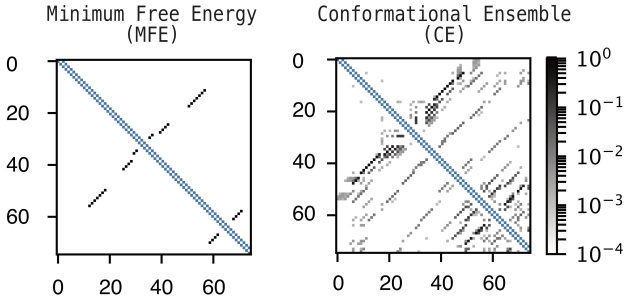

Figure 3. Edge features associated with a single RNA sequence for both MFE and CE graphs. Each matrix value at index (u, v) denotes the corresponding edge feature $e_{u,v}$ (blue for the first edge component indicating covalent bonds, level of gray for the second edge component indicating the base pair probability).

feature [1, 0], while nucleotides forming base pairs were connected with an edge feature of [0, 1]. For the second representation, the graph captured the RNA Conformational Ensemble (CE). This was generated in a similar way as for the MFE graph, but with the edge feature for a base pair now capturing the probability of that base pair occurring in the CE. The CE graph is always more connected than the MFE graph and contains more information regarding the possible secondary structures (Figure 3)

3.4. Neural network architecture

Our proposed model to integrate both sequence and structural information (GATv2_40nt) consists of four stages. First, a graph representation is generated by using the RNA secondary structure predictions. Then, we use a modified variant of the GAT (Veličković et al., 2018) to refine the initial node embeddings. This architecture updates the embedding h_u associated with each node u using the most relevant nodes in its neighbourhood $\mathcal{N}(u)$ through an attention-based weighted average (Brody et al., 2022). The propagation rule for the node representation h'_u in the subsequent

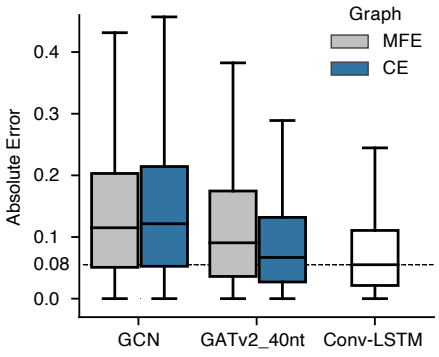

Figure 4. Distribution of absolute errors on the Hollerer_23 test set for various models trained on the Hollerer_23 train set.

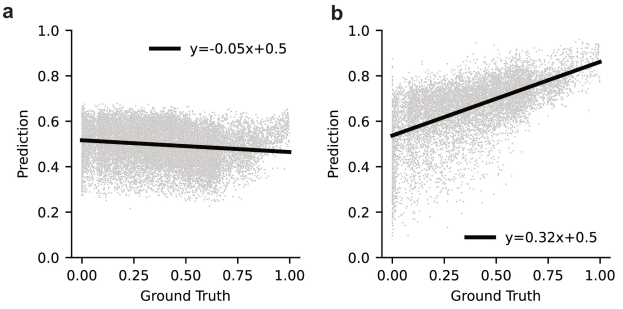

Figure 5. Evaluating the generalization performance. Predictions of the Conv-BiLSTM model (a) and the GAT_40nt model (b) on the Hollerer_20 test set. A linear regression was fitted on each set of points (black line).

layer is given by:

$$h'_u = \sigma\left(\sum\nolimits_{v\in\mathcal{N}(u)\cup\{u\}} \alpha_{u,v} W h_v\right), \text{ where}$$

$$\alpha_{u,v} = \frac{a^\intercal \text{LeakyReLu}(\Theta[h_u \parallel h_v \parallel e_{u,v}])}{\sum_{w\in\mathcal{N}(u)} a^\intercal \text{LeakyReLu}(\Theta[h_u \parallel h_w \parallel e_{u,w}])}.$$

Here, $\parallel$ indicates vector concatenation, the matrices $\Theta \in \mathbb{R}^{d'\times(2d+2)}$, $W \in \mathbb{R}^{d'\times d}$ and the vector $a \in \mathbb{R}^{2d'}$ are learnt; d and d' are the dimensions of the node embeddings before and after each graph layer. Following the message passing phase, the readout layer consists of taking the feature-wise maximum over the node embeddings using a window of 40 nt around the start codon. This restriction is motivated by local window analysis, which identified the pairing patterns for the nucleotides around the start codon to be the most influential (Terai & Asai, 2020). Finally, the vector is processed through a standard MLP to predict the protein expression level.

4. Experiments

When evaluating the GATv2_40nt model trained on the Hollerer_23 dataset, we found that using the CE graph led to better predictive performance compared to the MFE graph. The improvement was significant, with a difference of 0.03 in mean absolute error (Figure 4). On the other hand, a simple graph convolutional network (GCN) failed to leverage the additional information from the CE, highlighting the need for an architecture handling of edge features more effectively.

While the sequence-based Conv-LSTM model achieved the highest performance on the Hollerer_23 test dataset, it did not demonstrate generalization. Specifically, its predictions were nearly constant on the Cambray and Hollerer_20 datasets (Figure 5a, Figure 6h). In contrast, the GATv2_40nt model exhibited significantly better generalization, as reflected by a larger slope modelling the linear trend between the predictions and the ground truth (Table 1, Figure 5b).

Table 1. Slope of the best linear fit between the predictions versus ground truth. Higher value is better (given equal intercepts).

Dataset	Conv-LSTM	GATv2_40nt
Hollerer_23	**0.809 ± 0.003**	0.740 ± 0.004
Hollerer_20	−0.052 ± 0.002	**0.325 ± 0.004**
Cambray	0.009 ± 0.003	**0.198 ± 0.003**

Furthermore, we found that restricting the readout layer to a 40 nucleotide window yielded significant benefits compared to utilizing the entire graph (Figure 6). This outcome can be attributed to the likelihood that the crucial nucleotides are located within this window. By focusing on this specific sequence, the readout layer is better able to capture and learn the key features essential for accurate prediction.

5. Discussion

In this paper, we presented a pipeline for processing RNA secondary structure derived from computational simulations to aid prediction of protein expression. We found that a GAT is well-suited to extracting information from the CE of RNA secondary structures. Although there is room for improvement in the training of the GAT, we find it is more generalizable compared to models based on sequence alone, which were sensitive to covariate shift.

Despite the ever growing availability of sequence to protein expression data sets, our results suggest that sequence alone is insufficient for building generalizable and robust models. Incorporating structural information is a promising way to regularize models beyond specific sequence contexts and enable the more rational engineering of diverse genetic systems.

References

Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G., and Collins, J. J. A deep learning approach to programmable RNA switches. *Nature Communications*, 11(1):5057, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18677-1.

Brody, S., Alon, U., and Yahav, E. How Attentive are Graph Attention Networks?, January 2022.

Cambray, G., Guimaraes, J. C., and Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nature Biotechnology*, 36(10):1005–1015, October 2018. ISSN 1546-1696. doi: 10.1038/nbt.4238.

Do, C. B., Woods, D. A., and Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, July 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl246.

Ganser, L. R., Kelly, M. L., Herschlag, D., and Al-Hashimi, H. M. The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology*, 20(8):474–489, August 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0136-0.

Gilliot, P.-A. and Gorochowski, T. E. Effective design and inference for cell sorting and sequencing based massively parallel reporter assays. *Bioinformatics*, 04 2023a. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad277. btad277.

Gilliot, P.-A. and Gorochowski, T. E. Transfer learning for cross-context prediction of protein expression from 5'utr sequence. *bioRxiv*, 2023b. doi: 10.1101/2023.03.31.535140.

Goodman, D. B., Church, G. M., and Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science*, October 2013. doi: 10.1126/science.1241934.

Höllerer, S. and Jeschek, M. Ultradeep characterisation of translational sequence determinants refutes rare-codon hypothesis and unveils quadruplet base pairing of initiator tRNA and transcript. *Nucleic Acids Research*, 51(5):2377–2396, March 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad040.

Höllerer, S., Papaxanthos, L., Gumpinger, A. C., Fischer, K., Beisel, C., Borgwardt, K., Benenson, Y., and Jeschek, M. Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nature Communications*, 11(1):3551, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17222-4.

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science*, 324(5924):255–258, April 2009. doi: 10.1126/science.1170160.

Mathews, D. H., Moss, W. N., and Turner, D. H. Folding and Finding RNA Secondary Structure. *Cold Spring Harbor Perspectives in Biology*, 2(12):a003665, January 2010. ISSN , 1943-0264. doi: 10.1101/cshperspect.a003665.

McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990. ISSN 1097-0282. doi: 10.1002/bip.360290621.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. Modeling Relational Data with Graph Convolutional Networks, October 2017.

Terai, G. and Asai, K. Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. *Nucleic Acids Research*, 48(14):e81, August 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa481.

Valeri, J. A., Collins, K. M., Ramesh, P., Alcantar, M. A., Lepe, B. A., Lu, T. K., and Camacho, D. M. Sequence-to-function deep learning frameworks for engineered riboregulators. *Nature Communications*, 11(1):5058, December 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18676-2.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph Attention Networks, February 2018.

Xia, T., SantaLucia, J. J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochemistry*, 37(42):14719–14735, October 1998. ISSN 0006-2960. doi: 10.1021/bi9809425.

Yan, Z., Hamilton, W. L., and Blanchette, M. Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. *Bioinformatics*, 36 (Supplement_1):i276–i284, July 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa456.

Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, January 1981. ISSN 0305-1048. doi: 10.1093/nar/9.1.133.

A. RNA structure prediction and processing

A.1. Prediction of RNA secondary structures

ViennaFold was used to compute the MFE structures using the RNAfold function. To generate the base pairing probabilities, we used the ContraFold model (Do et al., 2006) with the Eternafold parameters. Sequences in all datasets were restricted to the first 75 nt of the 5'UTR and the protein coding sequence, which is typically sufficient to account for the potential influence of the protein coding sequence on expression rate (Höllerer & Jeschek, 2023).

A.2. Projection of the RNA conformational space

Conformations within 4 kcal/mol of the MFE structure were sampled using the RNAsubopt function from the Vienna package. Pairwise distances between all resulting 772 dot-braket files were computed using the Hamming distance. Two-dimensional coordinates for each structure were extracted by using the 'manifold.MDS' function implemented in the scikit learn package with parameters max_iter = 3000, eps = 10^{-9}, random_state = 28, normalized_stress = 'auto'. Two-dimensional interpolation was then performed using the surface triangulation function 'create_trisurf' implemented in matplotlib.

A.3. Model training

We found that GNNs are more challenging to train than sequence-based models as they tend to exhibit a higher degree of loss fluctuation. To cope with this instability, we performed a systematic model hyperparameter optimization to find appropriate ranges of values that best enabled learning. Learning rates and architecture parameters for each candidate model (sequence- and graph-based) were optimised to minimise the MAE loss after training for 15 epochs on the validation data set. 100 hyperparameter combinations were then dynamically selected using the Tree-structure Parzen estimator implemented in Optuna and evaluated for each model.

B. Model evaluation

The Conv-LSTM, GAT_20nt and the GAT_all models were evaluated across three data sets: 1. Hollerer_23 (Höllerer & Jeschek, 2023), which was also used for training; 2. Hollerer_20 (Höllerer et al., 2020); and 3. Cambray (Cambray et al., 2018). Predictions were assessed against the ground truth values and a linear regression performed to assess accuracy. Results are shown in Figure 6.

Table 2. Hyperparameter optimisation of the sequence-based models for the Hollerer_2023 dataset. Range of all hyperparameters considered when optimising the Conv-LSTM and CNN models are shown. The bold values indicate the final optimise parameters.

Hyperparameter	**Parameter search space**
CNN Channels	[64, 128, 256,**512**]
CNN Layers	[1, **2**, 3, 4,5]
CNN Kernel	[4, 6, 8, **10**]
CNN Pool	[1,**2**]
LSTM Hidden Size	[5, 10, 25, **50**, 100, 200, 500]
BILSTM	[**True**, False]
Activation	[ReLU, ELU, **LeakyReLU**]
MLP Hidden Size	[64, 128, 256, 512, **1024**]
MLP Layers	[**1**, 2, 3]
MLP dropout Rate	[0..**0.06**..0.5]
Learning Rate	[5e-5..**1e-4**..5e-3]
Trainable parameters	**5,490,597**

Table 3. Hyperparameter optimisation of the graph-based models for the Hollerer_2023 dataset. Range of all hyperparameters considered when optimising the GATv2_40nt and GCN model processing the CE graph structure are shown. The bold values indicate the final optimised parameters. The 'set2set' model corresponds to the 'seq2seq' model and 'all' represent the aggregation of the max, mean, min and STD pooling layers.

Hyperparameter	**Parameter search space**	
	Gatv2_40nt	GCN
GNN node embeddings size	[64, **128**, 256]	[64, 128, 256,**512**]
GNN number of layers	[2,3,4, **5**,6 , 7]	[1, **2**, 3]
GNN number of heads	[2,3,4, **5**,6 , 7]	∅
GNN activation	[**ReLU**, LeakyReLU]	[**ReLU**,LeakyReLU]
GNN norm	[**True**, False]	[**True**, False]
GNN readout layer	[set2set, all, **max**]	[set2set, **all**, max]
MLP Hidden Size	[64, **128**, 256, 512]	[64, **128**, 256, 512]
MLP Layers	[1,**2**, 2, 3]	[1, 2, **3**]
Dropout Rate	[0..**0.17**..0.5]	[0..**0.33**..0.5]
Learning Rate	[5e-4..**2.39e-4**..5e-3]	[5e-4..**3e-4**..5e-3]
Number of parameters	3,385,217	547,201

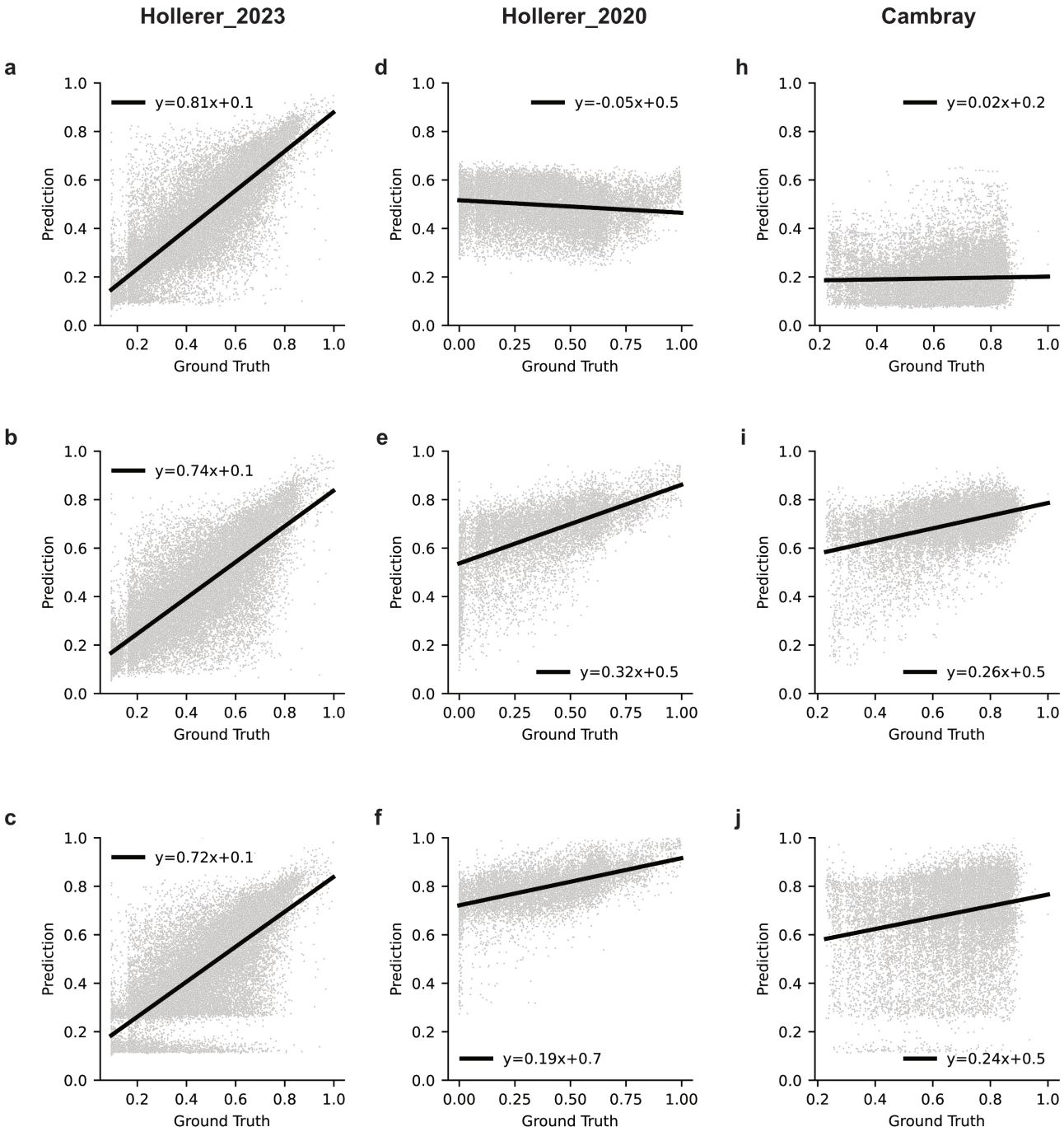

Figure 6. Plotting the predictions of the Conv-LSTM model (a, d, h), the GAT_40nt (b, e, i) or the GAT_all (c, f, j) for the Hollerer_23 data set (first column), Hollerer_20 data set (second column) and Cambray data set (third column). A linear regression was fitted on each set of points (black line).