Supplementary Material

Additional results, specifically visualisations of the learned representations and model evaluation via correlations analysis, from the etherum data experiments are presented in Sec. A. Sec. B contains additional results for the synthetic flow datasets. This includes the ablation study results, flow prediction performance results, and visualisation of the flow distributions.

A. Additional Results: Ethereum Dataset

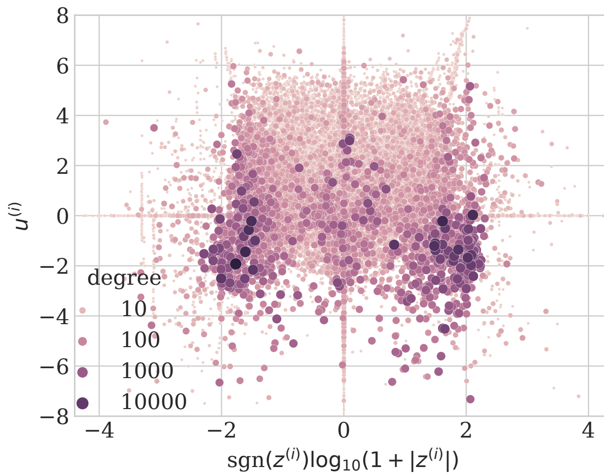
Similar to the transaction amount distributions, the distribution of $\mathbf{z}$ parameters learned by the gated gradient model span multiple orders of magnitude and therefore, a power transform that allows for both zero and negative values is useful for visualisation. We choose a modified version of the Yeo-Johnson transform (Yeo & Johnson, 2000),

$$T(y) = \operatorname{sgn}(y) \log_{10}(1 + |y|). \tag{1}$$

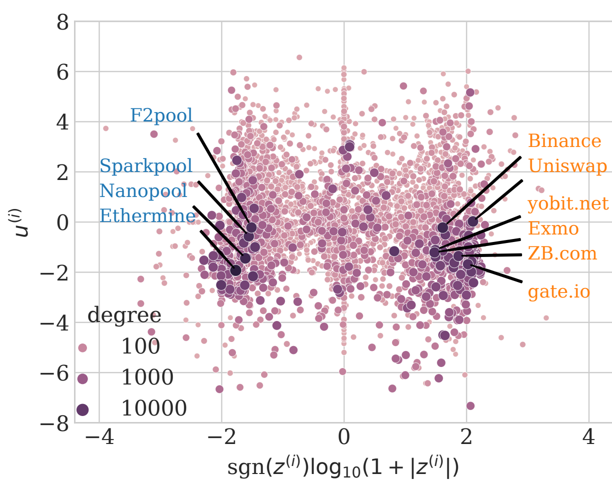
Compared to using $\log_{10}(|y|)$, the cost of preserving the sign of y is the loss of values $|y| \lessapprox 1$.

As stated in Sec. 6 in the main paper, the best performing gated gradient model used $K = 1$, and consequently, the learned representations $\{\mathbf{z}^{(i)}\}$ and gate parameters $\{\mathbf{u}^{(i)}\}$ for each node can be visualised in a 2D scatter plot, see Fig. 1(a). The colour and size of each marker is determined by the nodes total degree in the transaction graph. We note that the nodes of the tail-end of the degree distribution tend to have learned large values for $|z^{(i)}|$, and this observation is confirmed in Fig. 1(b) where only the parameters of nodes with degree > 15 are visualised.

In the case of $K = 1$ and with the chosen sign convention, a large negative values of $z^{(i)}$ indicate that the node functions as a source of ether in the network, while a large positive value correspond to sink behaviour. When investigating the identity of the ten highest degree nodes using https://etherscan.io/, we find that the four nodes with large negative z are mining pools, while the six nodes with large positive values are exchanges. This makes sense since miners generate ether as payment for the the proof of work, while exchanges accept ether in exchange for other currencies, e.g. USD or other tokens on the ethereum blockchain.



(a) All 452862 node parameter pairs.



(b) 4635 node parameter pairs for nodes with degree > 15, and with the top ten nodes by degree labelled.

Figure 1. Parameters, $z^{(i)}$ and $u^{(i)}$, learned by the gated gradient model and coloured by node degree. The power transform in Eq. (1) is used for $z^{(i)}$ to visually capture patterns across multiple orders of magnitude. In (b), the top ten nodes by degree are labelled. Blue font indicates a miner and orange font an exchange.

In addition to the cumulative relative errors distributions and amount histograms, Fig. 2 and 3 in the main paper, the correlation between model predictions and ground truth transactions can be used to evaluated the performance

Table 1. Coefficients of determination of model predictions and ground truth transactions on ethereum data test split. Values are calculated after applying the power transform in Eq. (1). Higher values indicate better performance is better.

MODEL	COEF. OF DETERMINATION
GATED	0.38
GRAD	0.32
F.E.+DNN2	0.24
N2V+DNN2	0.31
KUMAR ET.AL.	0.00

of a regression model. In Fig. 2, correlations between ground truth, x-axis, and model predictions, y-axis, on the ethereum data test split are visualised as 2D histograms. A perfect predictor would produce a histogram tracing the diagonals, highlighted in red. To quantify the correlation, the coefficient of determination, defined as the square of the correlation coefficient between two variables (Everitt & Skrondal, 2002, p.89), can be used. Intuitively, this is the proportion of the variation in the ground truth flow accounted for by the model predictions. As expected, this results in the same ordering of model performance as seen in Fig. 1 in the main paper.

B. Additional Results: Synthetic Flow Data

The result of the ablation study for the gated gradient model is presented in Tbl. 2. Here, the errors are defined as

$$\text{error}^* = \log_{10} \operatorname*{median}_{ij} \delta^{(ij)}, \tag{2}$$

with $\delta^{(ij)}$ being the relative error as defined in Eq. (9) in the main article. The means and standard deviations are calculated over the 10 different flows. The conclusions are the same as in the main paper, repeated here: LSQR+ performs the best out of the three initialisation strategies. LSQR performs similarly to LSQR+ on the multimodal data but worse on the unimodal data, and vice versa for the normal noise initialisation. Overfitting is the bottleneck for the unimodal data and L1 regularisation improves the validation errors slightly for the cora and bitcoin graph, while having a detrimental effect in the multimodal case. We further observe that overfitting is less of an issue for the complete graph and that all models overfit more on the bitcoin graph compared to cora. The reason is believed to relate to the graph sparsity or possibly the clustering coefficient and further analysis is left as future work.

The full results of the flow prediction performance experiments for both validation and training splits are presented as tables, see Tbls. 3 and **??**, and as box plots, see Figs. 5 and 6, with the error defined in (2) as performance measure.

For the gated gradient model, we note large discrepancies between validation and training error for the cora and bitcoin graphs, indicating that overfitting is an issue. We also note that the model using node2vec features performs well on average on the complete graph, but with a large variance. We also note a significant overfitting of the MLP using the node2vec features on the complete graph due to the large embedding dimension used for the node2vec representations.

The distributions of $\mathbf{z}^{(i)}$ used to generate synthetic flow samples in the multimodal case are visualised in Fig. 3 for the three graphs. Also visualised are the parameter values inferred by the gated gradient model. Ellipses are used to highlight the ground truth modes (red edges) and the same nodes of the learned parameters (black edges). For all three graphs we see that the gated gradient model is able to separate the three modes, albeit along the diagonal on which the parameters are initialised.

In Fig. 4, the distribution of flow values are visualised as histograms. The values are collected from the validation edges of one flow sample per graph and parameter distribution. The synthetic ground truth flow is shown in black, the predictions of the gated gradient model in blue and the gradient model in orange. As explained in the main paper, the unimodal setting aims to mimic the distribution of real data, with a single mode spanning multiple orders of magnitude. Conversely, three distinct peaks are observed in the multimodal setting, corresponding to flows within modes, flows between group 0 and group 1, and group 0 and group 2, see Fig. 3 . We see that the gradient model is unable to capture the third peak of the flow distributions in the multimodal case since it lacks the gate parameters.

Finally, in Fig. 7, the cumulative relative errors curves are shown for one flow sample in each setting. The gated gradient model generally performs better than the other models in the multimodal case, as was also observed in the box plots in Fig. 5, since it is able to infer the different modes of the ground truth parameter distribution.

References

Everitt, B. and Skrondal, A. *The Cambridge dictionary of statistics*. Cambridge University Press Cambridge, 2002.

Yeo, I.-K. and Johnson, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

Table 2. Ablation study results for the gated gradient model, showing training and validation error for each graph and flow distribution. The errors are $\log_{10}$ of the median relative errors aggregated over 10 different flow samples, lower is better. Regularisation strengths were $\lambda_z = \lambda_u = 0.5$ for the unimodal case and 0.05 for the multimodal case.

(a) Cora, unimodal

	VAL ERROR*		TRAIN ERROR*	
	MEAN	STD	MEAN	STD
INIT AND REG.				
NORMAL NOISE	-0.04	0.00	-1.23	0.01
LSQR	-0.04	0.00	-0.45	0.01
LSQR+	-0.03	0.00	-1.24	0.01
LSQR+, L1(u)	-0.07	0.01	-0.95	0.01
LSQR+, L1(z)	-0.05	0.00	-0.49	0.01
LSQR+, L1(u), L1(z)	-0.07	0.00	-0.39	0.01

(b) Bitcoin, unimodal

	VAL ERROR*		TRAIN ERROR*	
	MEAN	STD	MEAN	STD
INIT AND REG.				
NORMAL NOISE	-0.04	0.01	-1.84	0.11
LSQR	-0.03	0.01	-0.96	0.05
LSQR+	-0.03	0.01	-2.02	0.13
LSQR+, L1(u)	-0.06	0.01	-1.35	0.06
LSQR+, L1(z)	-0.02	0.01	-1.13	0.08
LSQR+, L1(u), L1(z)	-0.05	0.02	-0.90	0.05

(c) Complete graph, unimodal

	VAL ERROR*		TRAIN ERROR*	
	MEAN	STD	MEAN	STD
INIT AND REG.				
NORMAL NOISE	-0.19	0.05	-0.47	0.06
LSQR	-0.16	0.06	-0.22	0.04
LSQR+	-0.28	0.07	-0.52	0.13
LSQR+, L1(u)	-0.27	0.12	-0.36	0.19
LSQR+, L1(z)	-0.10	0.04	-0.16	0.04
LSQR+, L1(u), L1(z)	-0.10	0.03	-0.13	0.03

(d) Cora, multimodal

	VAL ERROR*		TRAIN ERROR*	
	MEAN	STD	MEAN	STD
INIT AND REG.				
NORMAL NOISE	-0.01	0.00	-0.30	0.01
LSQR	-1.32	0.02	-1.95	0.01
LSQR+	-1.20	0.03	-2.08	0.02
LSQR+, L1(u)	-1.06	0.03	-2.03	0.01
LSQR+, L1(z)	-0.88	0.02	-1.39	0.01
LSQR+, L1(u), L1(z)	-0.91	0.02	-1.41	0.01

(e) Bitcoin, multimodal

	VAL ERROR*		TRAIN ERROR*	
	MEAN	STD	MEAN	STD
INIT AND REG.				
NORMAL NOISE	-0.01	0.00	-0.36	0.16
LSQR	-0.78	0.08	-2.27	0.06
LSQR+	-0.81	0.06	-2.39	0.07
LSQR+, L1(u)	-0.67	0.09	-2.22	0.04
LSQR+, L1(z)	-0.56	0.07	-1.40	0.02
LSQR+, L1(u), L1(z)	-0.62	0.07	-1.39	0.03

(f) Complete graph, multimodal

	VAL ERROR*		TRAIN ERROR*	
	MEAN	STD	MEAN	STD
INIT AND REG.				
NORMAL NOISE	-0.24	0.48	-0.36	0.47
LSQR	-1.93	0.15	-1.98	0.14
LSQR+	-2.15	0.12	-2.23	0.08
LSQR+, L1(u)	-2.22	0.18	-2.30	0.14
LSQR+, L1(z)	-1.34	0.10	-1.42	0.04
LSQR+, L1(u), L1(z)	-1.44	0.09	-1.48	0.06

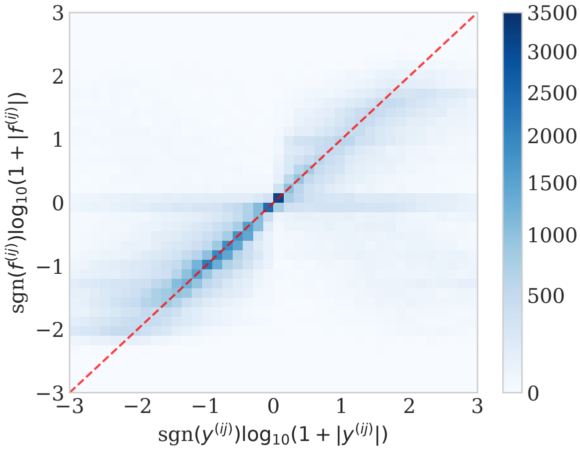
Table 3. Mean and standard deviations of the median $\log_{10}$ relative error for the train and validation splits on the synthetic flow data. For the validation split, lower values are better. The means and standard deviations are calculated over 10 generated flows for each graph.

(a) Validation split

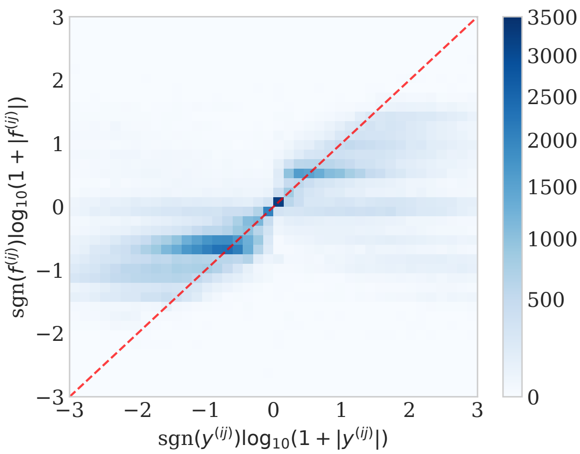
MODELS	UNIMODAL CORA	UNIMODAL BITCOIN	UNIMODAL COMPLETE	MULTIMODAL CORA	MULTIMODAL BITCOIN	MULTIMODAL COMPLETE
GATED	-0.07 ± 0.00	-0.06 ± 0.01	-0.28 ± 0.07	-1.32 ± 0.02	-0.81 ± 0.06	-2.22 ± 0.18
GRAD	-0.07 ± 0.00	-0.07 ± 0.01	-0.14 ± 0.03	-0.57 ± 0.00	-0.56 ± 0.02	-0.62 ± 0.04
F.E.+DNN2	-0.07 ± 0.01	-0.04 ± 0.01	-0.12 ± 0.03	-0.63 ± 0.06	-0.01 ± 0.02	-1.20 ± 0.15
N2V+DNN2	-0.00 ± 0.00	-0.02 ± 0.01	-0.26 ± 0.05	-0.00 ± 0.00	-0.00 ± 0.00	-0.92 ± 0.48
KUMAR ET.AL.	-0.01 ± 0.00	-0.00 ± 0.00	-0.02 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00

(b) Train split

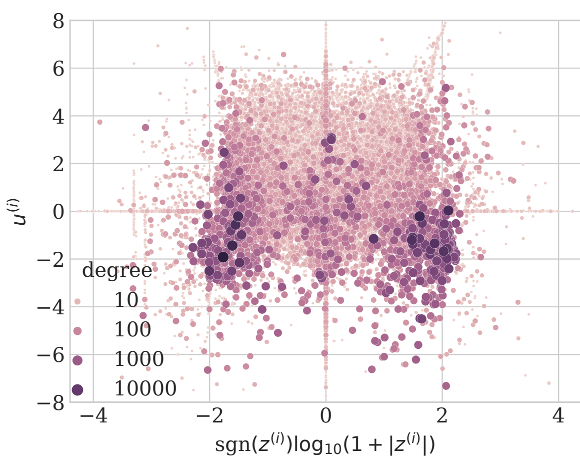
MODELS	UNIMODAL CORA	UNIMODAL BITCOIN	UNIMODAL COMPLETE	MULTIMODAL CORA	MULTIMODAL BITCOIN	MULTIMODAL COMPLETE
GATED	-0.39 ± 0.01	-1.35 ± 0.06	-0.52 ± 0.13	-1.95 ± 0.01	-2.39 ± 0.07	-2.30 ± 0.14
GRAD	-0.24 ± 0.00	-0.29 ± 0.02	-0.15 ± 0.02	-0.69 ± 0.01	-1.05 ± 0.09	-0.62 ± 0.03
F.E.+DNN2	-0.17 ± 0.00	-0.12 ± 0.04	-0.34 ± 0.04	-0.88 ± 0.07	-0.04 ± 0.09	-1.28 ± 0.10
N2V+DNN2	-0.03 ± 0.04	-0.13 ± 0.06	-1.70 ± 0.20	-0.00 ± 0.00	-0.01 ± 0.01	-2.32 ± 1.28
KUMAR ET.AL.	-0.05 ± 0.00	-0.08 ± 0.01	-0.03 ± 0.01	-0.01 ± 0.00	-0.01 ± 0.00	-0.00 ± 0.00



(a) gated

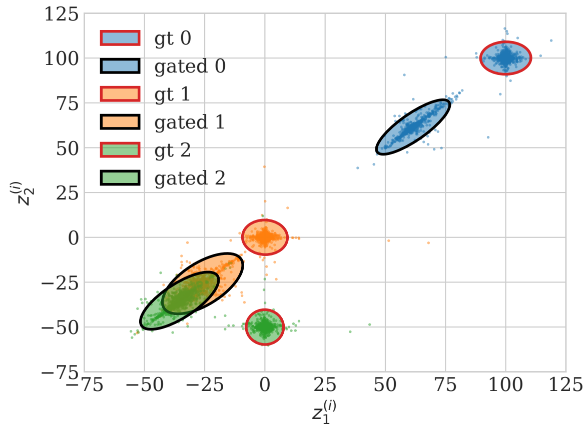


(b) grad

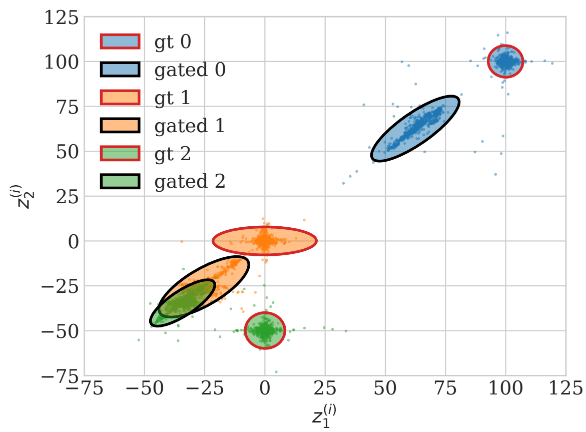


(c) n2v+dnn2

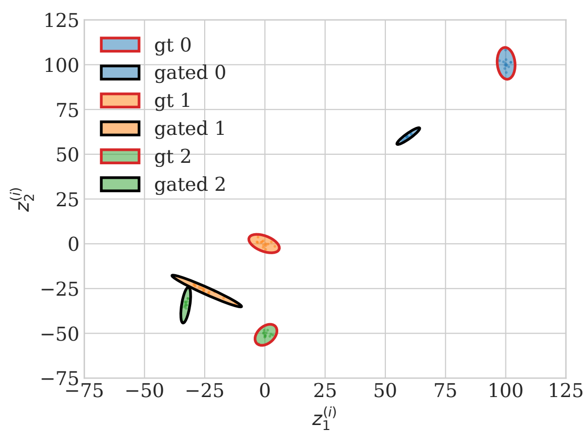
Figure 2. Correlation histograms for the gated gradient model, the gradient model and the two layer MLP model with node2vec features. The x-axes is the ground truth transactions on the ethereum test split and the y-axes are the predictions. Each variable has been transformed using Eq. (1). The result of an ideal model is highlighted by the red diagonals.



(a) Cora graph

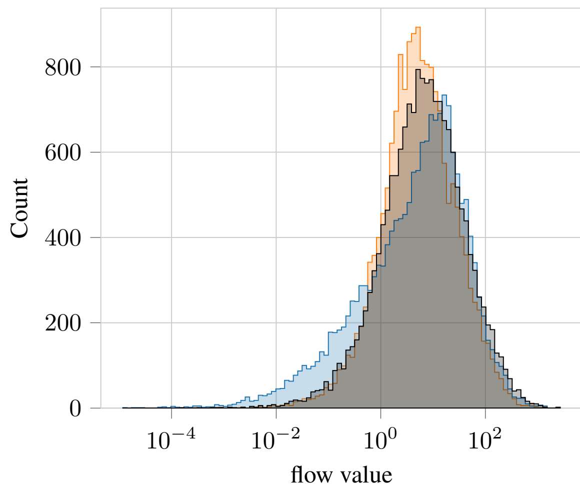


(b) Bitcoin graph

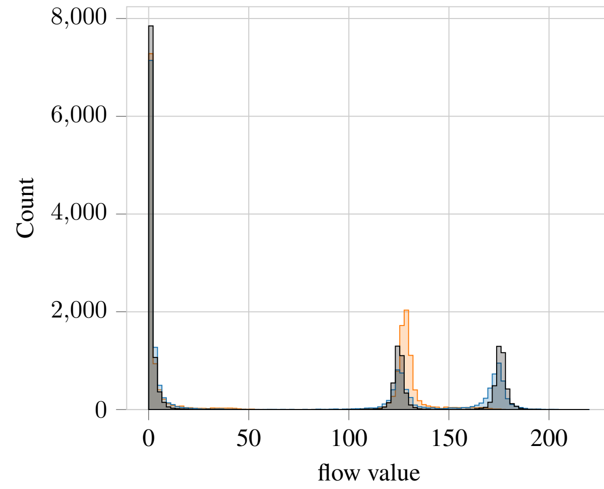


(c) Complete graph

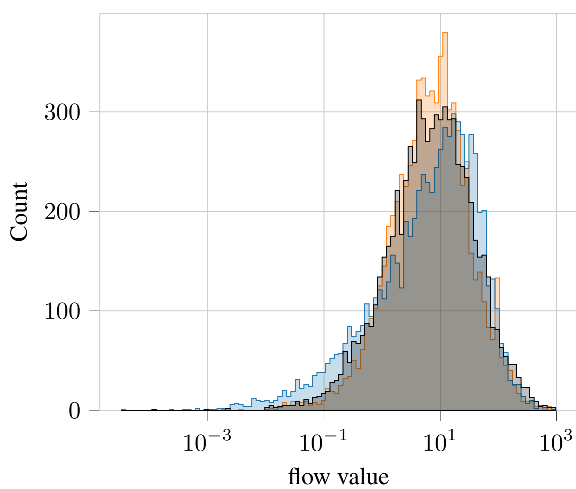
Figure 3. $\{z^{(i)}\}$ samples from the distribution used to generate synthetic flows in the multimodal case, highlighted in red, and the values learned by the gated gradient model, highlighted in black. The ellipses are used as a visual aid and Student-t distributions are used for each mode, as specified in the main article.



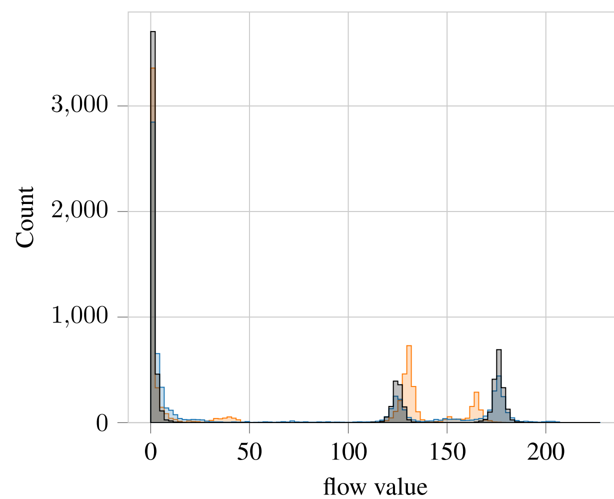
(a) Cora graph, unimodal



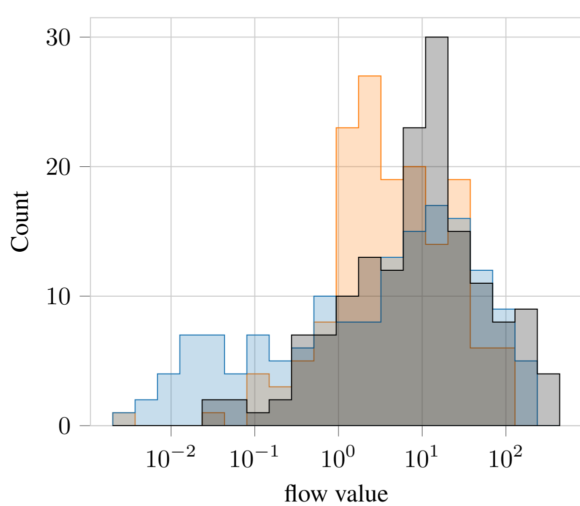
(b) Cora graph, multimodal



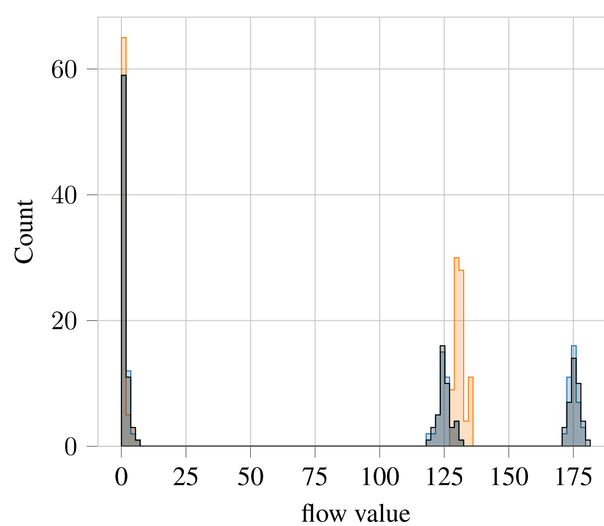
(c) Bitcoin graph, unimodal



(d) Bitcoin graph, multimodal

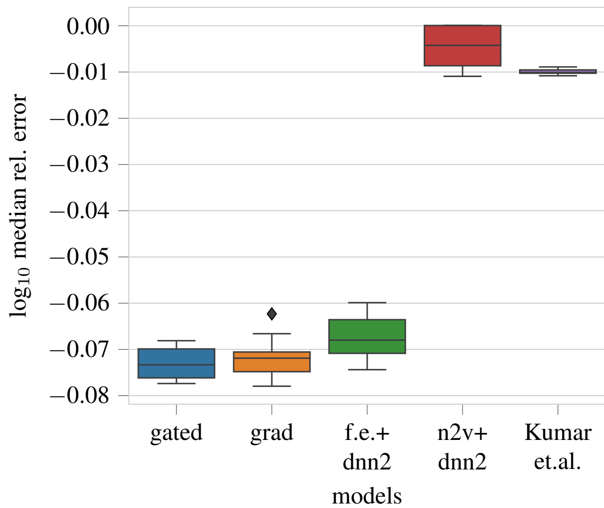


(e) Complete graph, unimodal

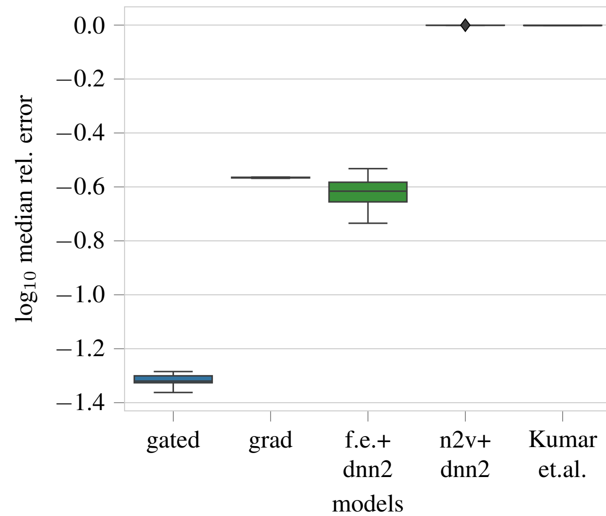


(f) Complete graph, multimodal

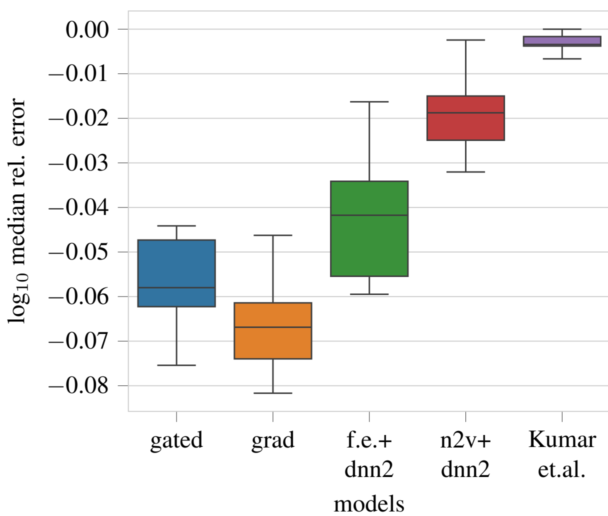
Figure 4. Histograms showing distributions of ground truth validation flows (absolute values) for each of the three graphs and two flow distributions, together with the histograms of predictions for the gated gradient model and the gradient model.



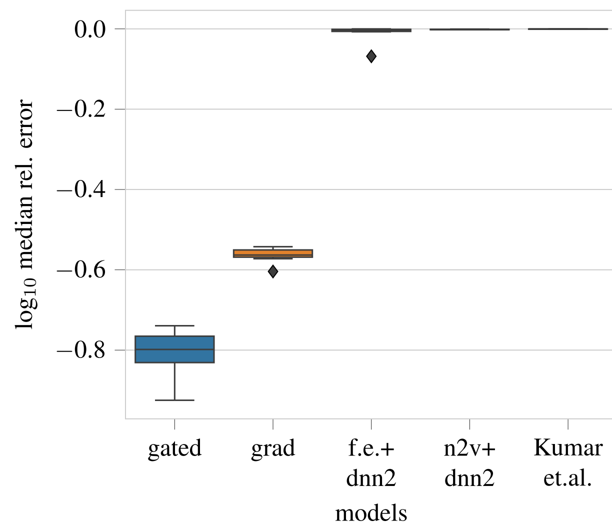
(a) Cora graph, unimodal



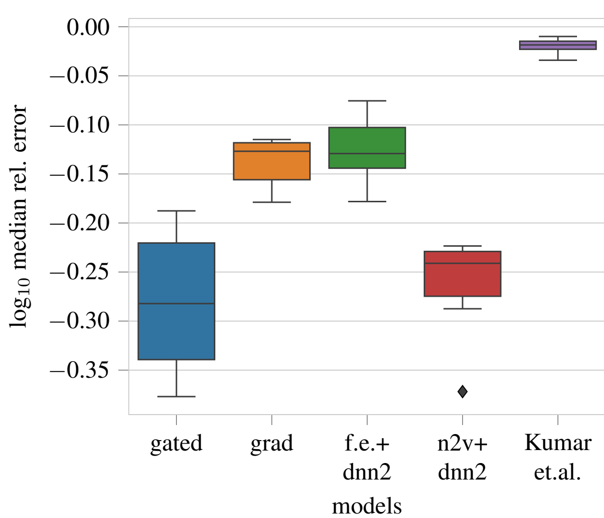
(b) Cora graph, multimodal



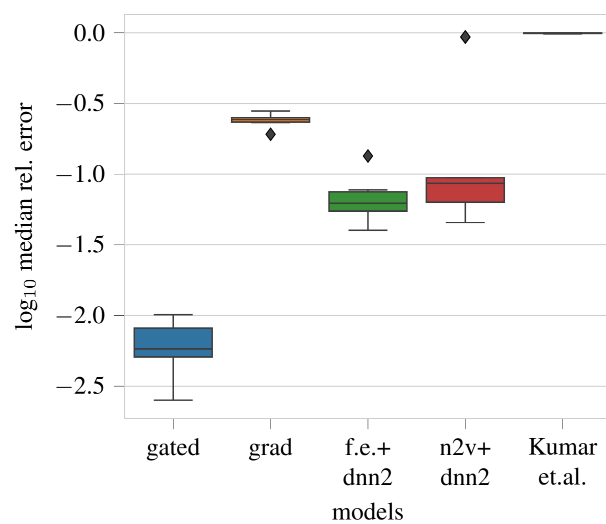
(c) Bitcoin graph, unimodal



(d) Bitcoin graph, multimodal

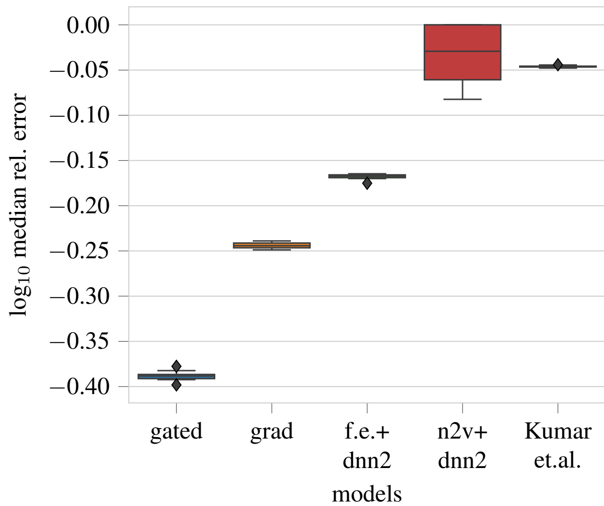


(e) Complete graph, unimodal

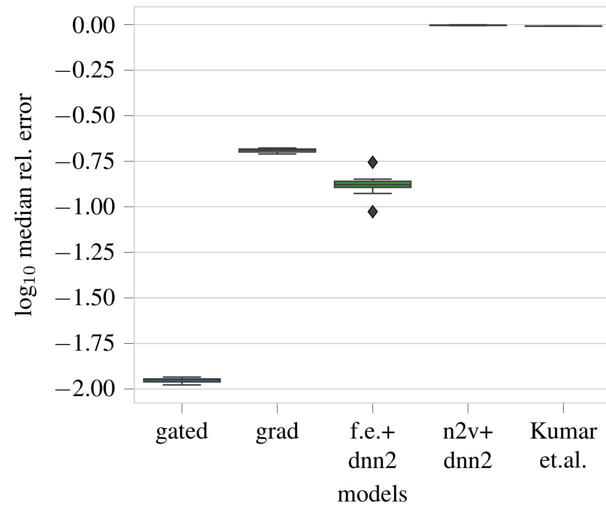


(f) Complete graph, multimodal

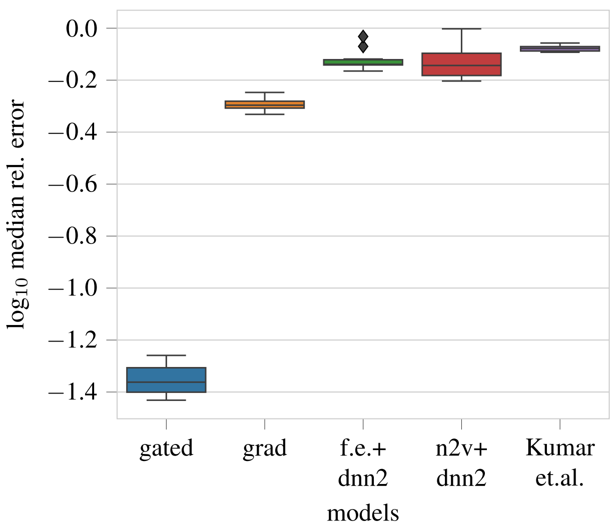
Figure 5. Validation split flow prediction performance results for each of the three graphs and two flow distributions, with each box plot created using 10 different flow samples. The errors are calculated on the validation set not seen during training.



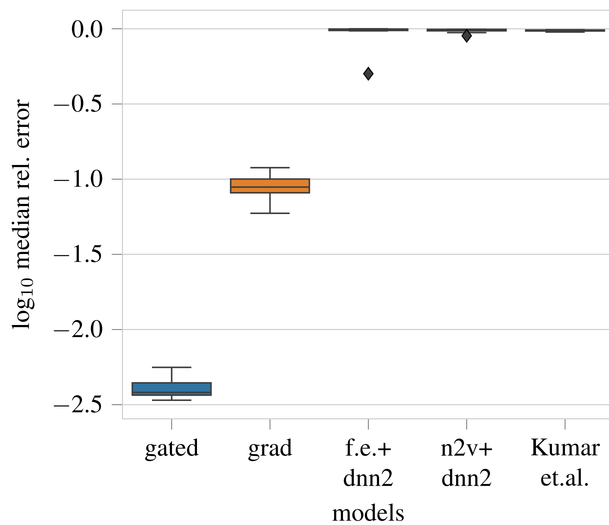
(a) Cora graph, unimodal



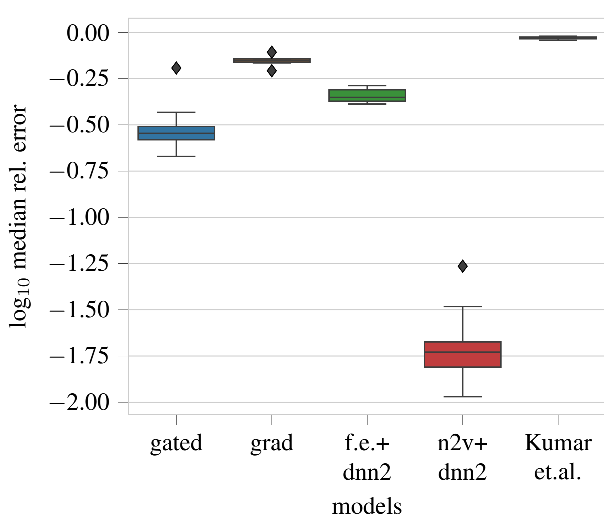
(b) Cora graph, multimodal



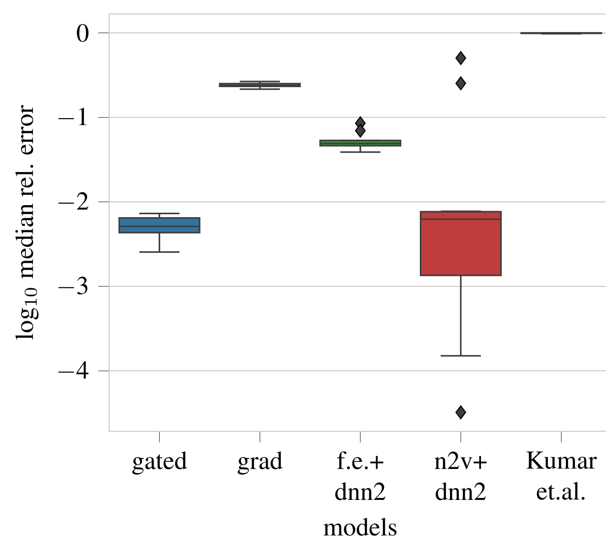
(c) Bitcoin graph, unimodal



(d) Bitcoin graph, multimodal

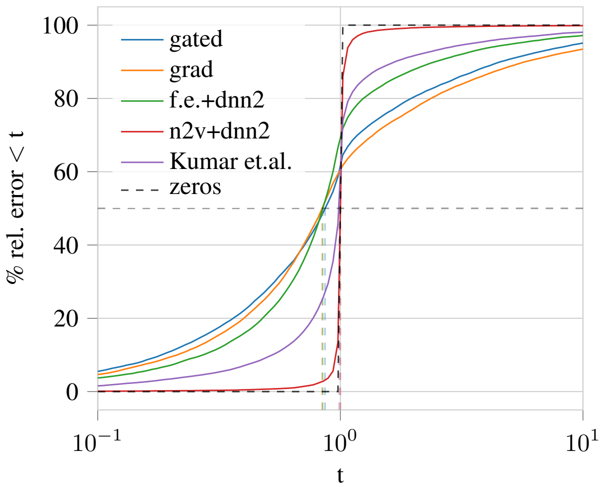


(e) Complete graph, unimodal

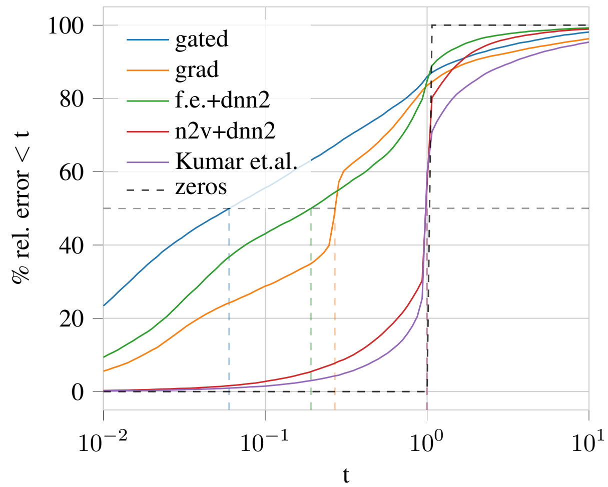


(f) Complete graph, multimodal

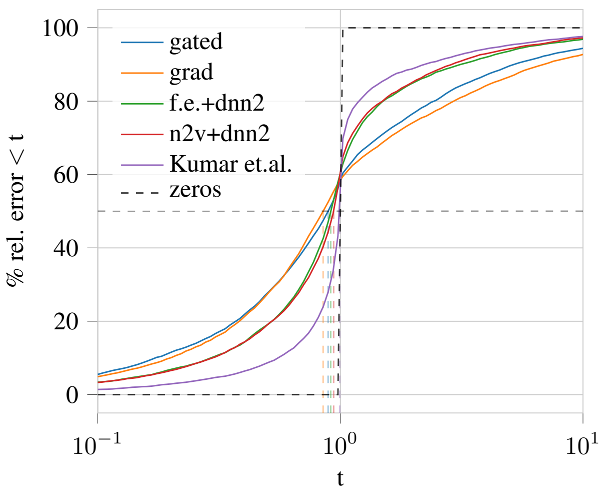
Figure 6. Train split flow prediction performance results for each of the three graphs and two flow distributions, with each box plot created using 10 different flow samples.



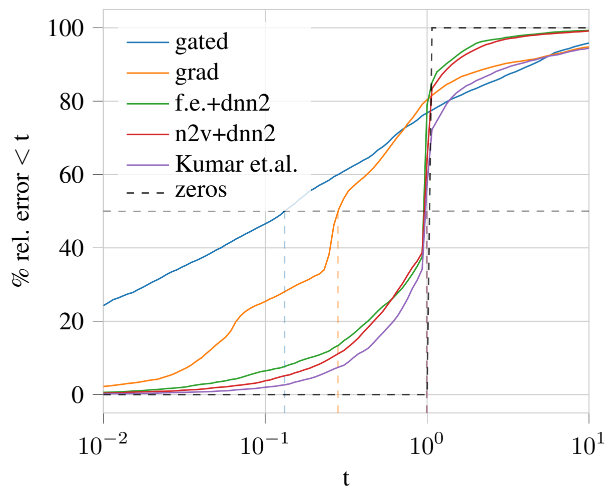
(a) Cora graph, unimodal



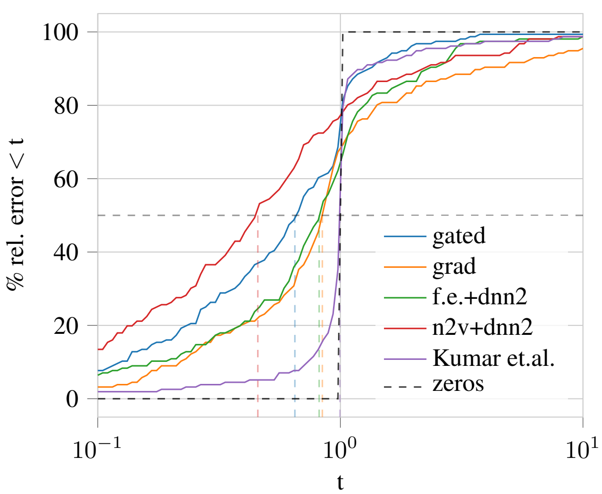
(b) Cora graph, multimodal



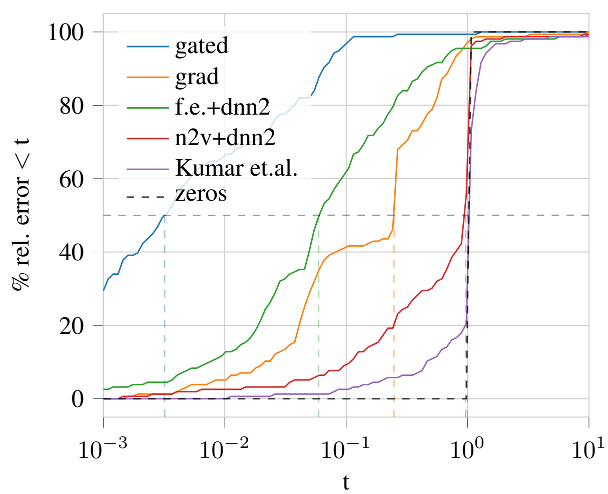
(c) Bitcoin graph, unimodal



(d) Bitcoin graph, multimodal



(e) Complete graph, unimodal



(f) Complete graph, multimodal

Figure 7. Cumulative relative error distributions using one flow samples for each of the three graphs and two flow distributions. The coloured, dashed, vertical lines indicate the median $\log_{10}$ errors for each curve. Note the different scales on x-axes in the multimodal cases.