CO problem type	MaxCut		MIS		MDS		MaxCl
Dataset	BA-small	BA-large	RB-small	RB-large	BA-small	BA-large	RB-small
n_{Query}	482	197	8.596	48.216	348.830	31.525	1.752.059

Table 5. Table that shows the number of n_{Query}, where $t_{\text{train}} + n_{\text{Query}} \cdot t_{\text{eval}} < t_{\text{Gurobi}}$. Here, t_{train} is the training time of DiffUCO, t_{eval} is DiffUCO's evaluation time per graph and t_{Gurobi} is the evaluation time per graph of Gurobi.

CO problem type	MaxCut|*Size* ↑		MIS | Size ↑		MDS | Size ↓		MaxCl | Size ↑
Dataset	BA-small	BA-large	RB-small	RB-large	BA-small	BA-large	RB-small
Train set	722.96 ± 0.88	2932.56 ± 1.45	$\mathbf{18.59 \pm 0.02}$	36.7 ± 0.07	28.36 ± 0.05	107.56 ± 0.10	14.17 ± 0.2
Validation set	$\mathbf{721.07 \pm 1.13}$	$\mathbf{2936.06 \pm 8.02}$	18.58 ± 0.03	$\mathbf{36.79 \pm 0.06}$	$\mathbf{28.286 \pm 0.10}$	$\mathbf{107.54 \pm 0.42}$	$\mathbf{14.56 \pm 0.1}$
Test set	$\mathbf{723.34 \pm 2.32}$	2928.53 ± 0.83	$\mathbf{18.52 \pm 0.04}$	$\mathbf{36.81 \pm 0.07}$	28.44 ± 0.10	108.02 ± 0.33	14.36 ± 0.1

Table 6. Generalisation properties of DiffUCO on various datasets. The average solution quality without CE is shown for the train, val, and test datasets. In most cases, the performance on the validation set is the best, while the performance on the test set is the worst. In contrast to the results shown in Tab. 3 and Tab. 2 the number of diffusion steps is not increased during the evaluation on the train, val and test datasets.

Method	$\|V\| = 800$	$\|V\| = 1K$	$\|V\| = 2K$	$\|V\| \geq 3K$
Greedy (r)	411.44	359.11	737.00	774.25
SDP (r)	245.44	229.22	N/A	N/A
RUNCSP (r)	185.89	156.56	357.33	401.00
ECO-DQN (r)	65.11	54.67	157.00	428.25
ECORD (r)	8.67	8.78	39.22	187.75
ANYCSP (r)	**1.22**	**2.44**	**13.11**	**51.63**
DiffUCO: CE-ST$_8$	4.11	7.17	31.67	119

Table 7. Comparison to MaxCut results form (Tönshoff et al., 2023). Models are evaluated according to the average deviation from the best-known cut size on the Gset dataset. The graphs are grouped according to their node size. (r) indicates that the results are taken from (Tönshoff et al., 2023).

Method	BA-small	BA-large	BA-huge	BA-giant
	MDS Size ↓			
Gurobi - $t_{lim} = 300s$	**27.91**	**105.05**	**153.89**	**253.53**
DiffUCO: CE-ST$_8$	28.31	106.08	157.46	259.02
	MaxCut Size ↑			
Gurobi - $t_{lim} = 300s$	**735.67**	2945.43	4414.29	7319.47
DiffUCO: CE-ST$_8$	731.72	**2949.02**	**4444.47**	**7390.99**

Table 8. Out-of-distribution (OOD) results on MDS and MaxCut. DiffUCO is trained on BA-small and then evaluated on three larger OOD datasets BA-large, BA-huge with an average of 2000 nodes, and BA-giant with an average of 3000 nodes. The best method is marked as bold.

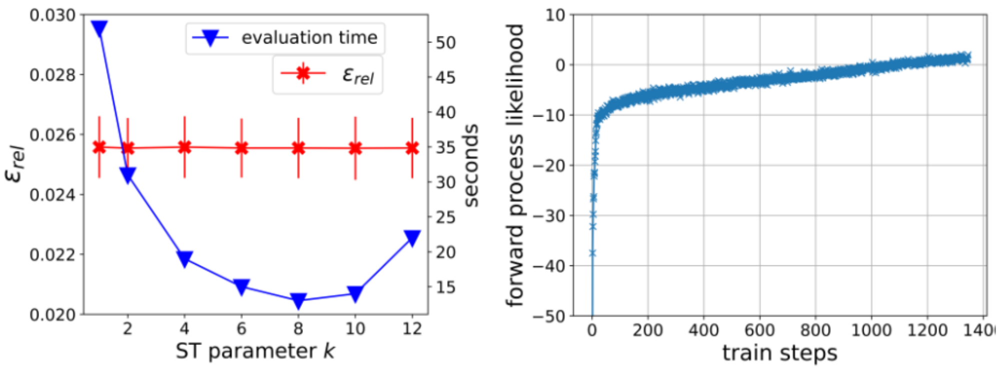

Figure 2. Left: Evaluation of the solution quality of CE-ST over k on the RB-small MIS dataset. The plot shows that the relative error does not change when k is increased, whereas the evaluation time decreases up to $k = 8$. As some operations scale exponentially with k we observe that the evaluation time increases at $k > 8$. Right: The unnormalized forward process likelihood $\mathbb{E}_{X_{0:T} \sim q_\theta(X_{0:T})}[\log p(X_{0:T})]$ over the number of training steps on the RB-100 MIS dataset. The model is trained at a constant temperature. The plot shows that the model learns to generate samples that become more likely under the forward diffusion process.