
ML4CO-KIDA: Knowledge Inheritance in Dataset Aggregation

Zixuan Cao
Peking University
Megvii Research
caozixuan.percy@stu.pku.edu.cn

Yang Xu
Peking University
Megvii Research
1800010740@pku.edu.cn

Zhewei Huang
Megvii Research
huangzhewei@megvii.com

Shuchang Zhou
Megvii Research
zsc@megvii.com

Abstract

The Machine Learning for Combinatorial Optimization (ML4CO) NeurIPS 2021 competition aims to improve state-of-the-art combinatorial optimization solvers by replacing key heuristic components with machine learning models. On the dual task, we design models to make branching decisions to promote the dual bound increase faster. We propose a knowledge inheritance method to generalize knowledge of different models from the dataset aggregation process, named KIDA. Our improvement overcomes some defects of the baseline graph-neural-networks-based methods. Further, we won the 1st Place on the dual task. We hope this report can provide useful experience for developers and researchers. The code is available at <https://github.com/megvii-research/NeurIPS2021-ML4CO-KIDA>.

In the dual task of ML4CO NeurIPS 2021 competition, as shown in Table 1, our team (Nuri) ranks 1st, 6th, and 1st in three benchmarks of Balanced Item Placement, Workload Apportionment, and Anonymous Problem separately. Our report is organized as follows: Section 1 introduces the basic information of ML4CO; Section 2 introduces issues we observe when following the baseline method; Section 3 explains in detail the main method we develop to improve the quality of branching decisions further; Section 4 shows the performance of different models in each benchmark; Section 5 discusses the failure of Strong Branching [1] which is regarded as expert knowledge; We conclude our report in Section 6.

Table 1: **Leaderboard of the Dual Task in ML4CO NeurIPS 2021 competition**

Team	Item Placement Cum. Reward \uparrow	Load Balancing Cum. Reward \uparrow	Anonymous Cum. Reward \uparrow	Score \downarrow
Nuri (Ours)	6684.00	630787.18	27810782.42	6 ($1 \times 6 \times 1$)
EI-OROAS	6670.30	631744.31	27158442.74	8 ($2 \times 1 \times 4$)
EFPP	6487.53	631365.02	26340264.47	117 ($3 \times 3 \times 13$)
KAIST_OSI	6196.56	631410.58	26626410.86	126 ($7 \times 2 \times 9$)
qqy	6377.23	630557.31	27221499.03	132 ($6 \times 11 \times 2$)

1 Preliminary

We will introduce the background knowledge needed in the dual task and the official baseline based on graph neural network (GNN) [17, 12].

1.1 Dual Task

For mixed-integer linear programs (MILPs), a well-known solving method is branch-and-bound (B&B) [13]. One of the key issues of B&B is how to select variables for **Branching**. In ML4CO competition, the goal of the dual task is to promote the dual bound to increase faster by selecting proper branching variables. The final metric is the dual integral over time, shown in Figure 1.

1.2 Branching Algorithms

Although making branching decisions has received little theoretical understanding to this day [15], there are many heuristic algorithms. Strong Branching involves testing which of the candidate variable gives the best improvement to the objective function before actually branching on them. Strong Branching is a high-quality but computationally expensive method. In practice, modern B&B solvers rely on hybrid-branching [14] and reliability branching [1], which only use Strong Branching at certain nodes of the search tree and use pseudo-cost branching [3] at other nodes.

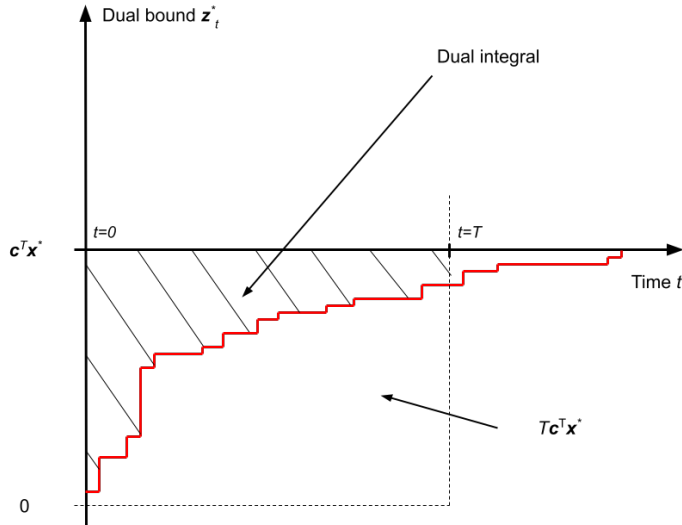


Figure 1: **Evaluation Metric of the Dual Task.** Our goal is to minimize the dual integral over time

1.3 Official Baseline

The baseline method [7] regards the decision of Strong Branching as expert knowledge and train models by imitating the expert knowledge. Specifically, we convert the combinatorial optimization problems to bipartite graphs, illustrated in Figure 2. Then we use the GNN to obtain the embedding of each variable. These embeddings are used to generate branching strategies for the current state. We train this GNN by minimizing the cross-entropy loss between its output and the expert decision.

2 Bad Generalization Issue

Following the method in [7], we summarize two issues that may reduce the final performance.

First, there is a major gap between offline training and the online deployment phase. In offline training, the data is collected beforehand. The GNN model achieves an accuracy of around 0.80 (validation). However, the accuracy shows a sharp decline when deploying a trained model. As shown in Figure 3,

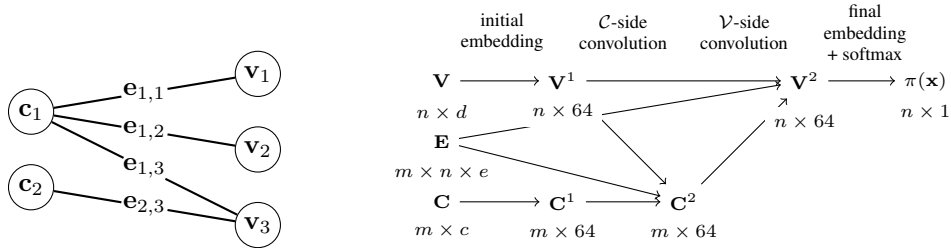


Figure 2: Copy from [7]. Left: the bipartite state representation $s_t = (\mathcal{G}, \mathbf{C}, \mathbf{E}, \mathbf{V})$ with $n = 3$ variables and $m = 2$ constraints. Right: the bipartite GCNN architecture for parametrizing baseline policy $\pi_\theta(\mathbf{a} \mid s_t)$

when making branching decisions for a random online instance, the averaged accuracy of different steps is about 0.1.

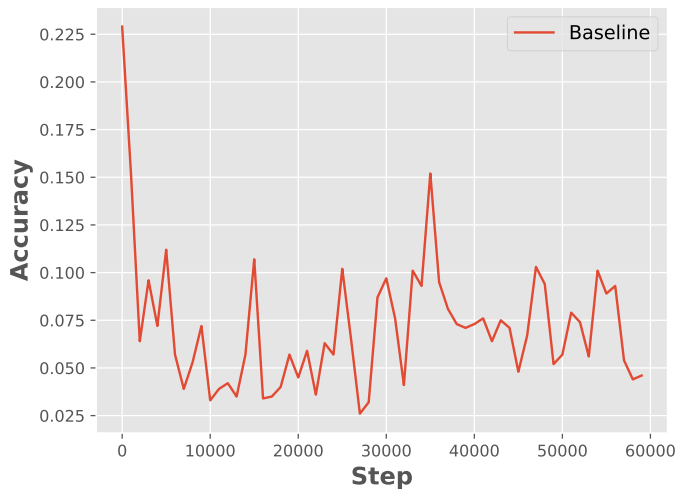


Figure 3: **Accuracy when Interacting with SCIP Solver.** During the deployment phase, the accuracy drops a lot

Second, the higher accuracy does not necessarily lead to a higher reward. As shown in Table 2, comparing models in different epochs, the accuracy is not strongly related to the final reward.

Table 2: **Performance Comparison of Different Epochs.** The imitation accuracy is not directly related to model performance

Epoch	Top 1 Acc.	Top 3 Acc.	Top 5 Acc.	Cum. Reward
1	0.780	0.932	0.971	5202.6
5	0.803	0.948	0.981	5545.7
10	0.808	0.946	0.978	5131.9
20	0.810	0.947	0.979	5038.8

3 Method

3.1 DAgger

The major performance gap between training and deployment can be explained by the defects of Behavior Cloning [2], which forces agents to learn expert behavior. There is a common problem that the distribution of collected data and the data the model encounters in the real scene may be inconsistent. For example, a driving-related dataset is rich in driver’s operation data during normal driving. But when the car is about to hit the wall, there may be a lack of data to deal with the contingency. To make a more completely collected dataset, we need to constantly explore the unseen data and label these data with expert knowledge.

Dataset Aggregation (DAgger) [16] is such a method to reduce the gap between training and deployment. Based on DAgger, we collect data iteratively and train new models constantly. Details are described in Algorithm 1.

Algorithm 1 DAgger in Dual Task

```
1: Initialize a random model  $\pi_0$  and an empty dataset  $D$ 
2: for each  $i \in [1, 50]$  do
3:   Interact with the solver with 0.95 probability of using model  $\pi_{i-1}$  and 0.05 probability of
   using Strong Branching
4:   Collect data obtained by Strong Branching as  $D_i$ 
5:    $D = D \cup D_i$ 
6:   if  $i \bmod 10 == 0$  then
7:     Train  $\pi_i$  with 100 epochs
8:   else
9:     Train  $\pi_i$  with 10 epochs
10:  end if
11: end for
```

3.2 KIDA

As shown in Section 2, in the Item Placement benchmark, higher accuracy does not lead to higher rewards. Experiments show that the performance of different models obtained in DAgger is not stable. Therefore, we use model ensemble [4, 11] to improve the performance of our model further. A noticeable difficulty is that the practicality of the model is time-sensitive. Averaging the output of different models is time-consuming. We consider averaging the weights of different models [19]. Here, we define Knowledge Inheritance in Dataset Aggregation (KIDA) as building a new model π_{avg} by averaging the weights of trained models during the dataset aggregation process. The framework is shown in Figure 3.2. The pipeline of KIDA is similar to Born-Again Neural Networks [5]. In KIDA, the training of the current model depends on the generation using the last model. The final model is the ensemble of trained models.

Formally, for models obtained from dataset aggregation with parameters $(\theta_0, \theta_1, \dots, \theta_{n-1})$, the parameters of π_{avg} are obtained by: $\theta_{avg} = \sum_{i=0}^{n-1} \theta_i/n$.

Figure 3.2 shows the difference between KIDA and the popular epoch weight average method. KIDA focuses on models trained by different data from different iteration rounds of DAgger. Although the training data of models are different, these data are related. Common epoch weight average such as Snapshot ensembles [9] and SWA [10] focuses on models that are obtained in the same training process of different epochs.

4 Experiment

In ML4CO competition, there are three benchmarks, including Balanced Item Placement, Workload Apportionment, and Anonymous Problem ¹ which need to be evaluated separately. In this section, we discuss the performance of different models in each benchmark.

¹Details about each benchmark can refer to <https://www.ecole.ai/2021/ml4co-competition/>

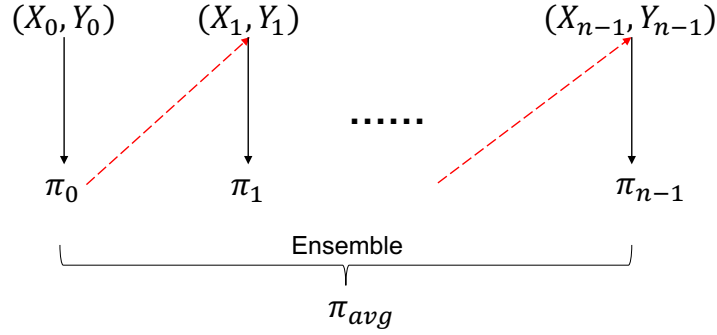


Figure 4: **Framework of KIDA.** We train each model based on the data generation using the last model. Finally, we average the parameter of trained models

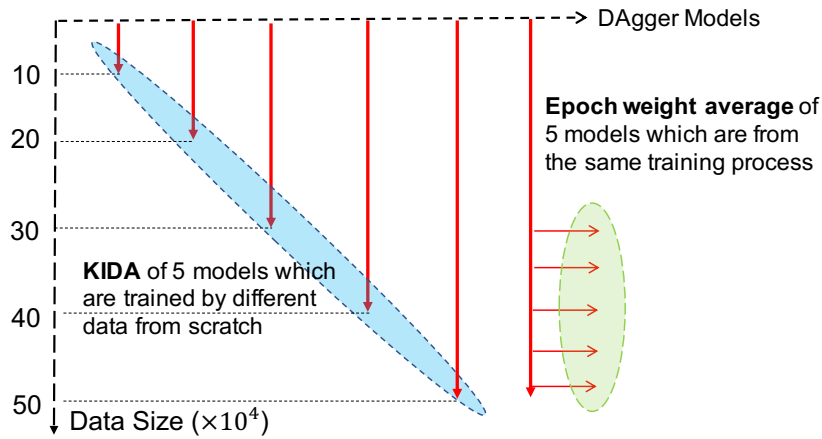


Figure 5: **KIDA vs. Epoch Weight Average**

4.1 Anonymous Problem

We show the performance of models with different settings in the validation set in Figure 6. With the same data size, the model using DAgger performs better. With the increase of data size, the performance increases. The model we finally submitted is a new DAgger model with Dropout [18] layers and a longer solver time when collecting data.

4.2 Balanced Item Placement

Overall Performance. The performance of DAgger at different iteration rounds is shown in Figure 7. Unlike in the Anonymous benchmark, the performance of DAgger is unstable. But when we apply KIDA in this benchmark, the performance has significantly improved.

Detailed Evaluation. To further explore our method, we randomly process a problem from the validation set using Strong Branching and compare the output of different models with Strong Branching labels. The accuracy of different models is shown in Figure 8. DAgger model still has the highest accuracy, and the accuracy of the baseline model is the worst.

Models used in KIDA are the top 3 performance models of DAgger. Table 3 shows the comparison between these models and the KIDA model. Although KIDA has lower accuracy in collected validation data, it has a lower loss value and higher reward.

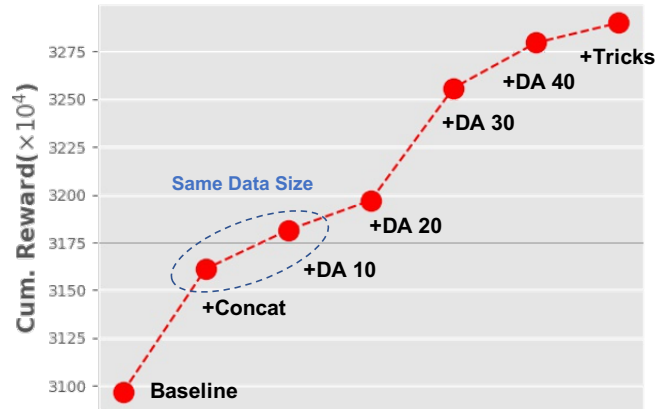


Figure 6: **The Cumulative Reward of Anonymous in Validation Dataset.** The performance of DAgger models is related to the number of iteration rounds

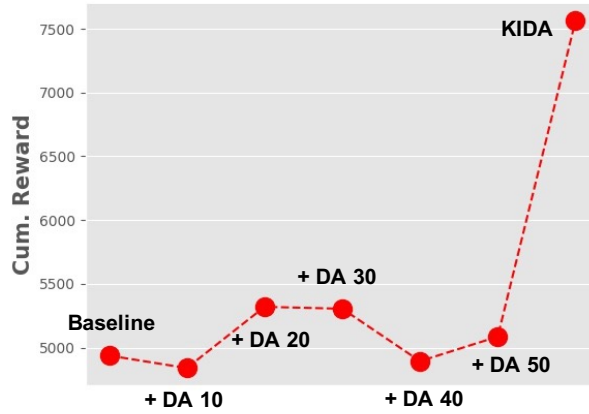


Figure 7: **The Cumulative Reward of Item Placement in Validation set.** KIDA has significant advantages over just applying DAgger

Table 3: **Performance Comparison between DAgger models and the KIDA model.** KIDA suffers from an accuracy drop but has a lower loss

Model	Top 1 Acc.	Top 3 Acc.	Top 5 Acc.	Loss	Cum. Reward
Model 0	0.850	0.957	0.981	7.38	5304.9
Model 1	0.797	0.917	0.966	9.50	5319.8
Model 2	0.795	0.916	0.961	6.18	5237.5
KIDA	0.721	0.822	0.870	2.96	7561.6

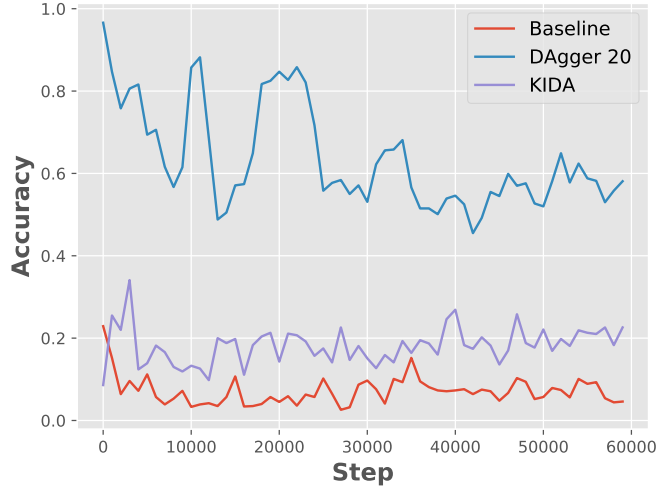


Figure 8: **Accuracy when Interacting with the Solver.** Overall accuracy of DAgger 20, KIDA and the baseline are 0.635, 0.182, 0.071 respectively.

4.3 Workload Apportionment

In Workload Apportionment, baselines are trained and compared with random policy. From the table, we can see that the top 1 accuracy of the baseline model is 45.6%. The accuracy of a classification task with more than one hundred labels shows that the model has learned much expert knowledge. However, random strategies that do not rely on prior knowledge can obtain higher cumulative rewards.

Table 4: **Performance Comparison between Baseline and Random Policy.** The baseline model gets great accuracy but fails to achieve higher rewards

	Top 1 Acc.	Top 3 Acc.	Top 5 Acc.	Cum. Reward
Baseline	0.456	0.729	0.820	624043.6
Random	0.013	0.034	0.052	624928.9

5 Discussion

Overall, we show that Strong Branching can not produce completely reliable labels. To further explore the performance of Strong Branching, we compare the variation of the dual bound using different methods.

As shown in Figure 9, we select a random instance for each benchmark, the variation of the dual bound using different strategies. Even if we ignore the expensive time cost of Strong Branching, the dual bound improvement may still fall behind GNN models. This observation may be explained by dual degeneracy [8]. If there is a high dual degeneracy in LP solution, the product score of Strong Branching that SCIP uses to combine the improvements of the two-child nodes would be close to zero [6]. In that case, Strong Branching may fail, and the expert knowledge needs to be redesigned based on other information accumulated in the problem-solving process.

6 Conclusion

In this paper, we devise a knowledge inheritance method in the dataset aggregation process for the dual task of ML4CO competition. By dataset aggregation, the inconsistency between training and deployment is reduced. By averaging the weights of different models from the dataset aggregation process, learned knowledge is generalized to get better results. Our model gets great improvements on

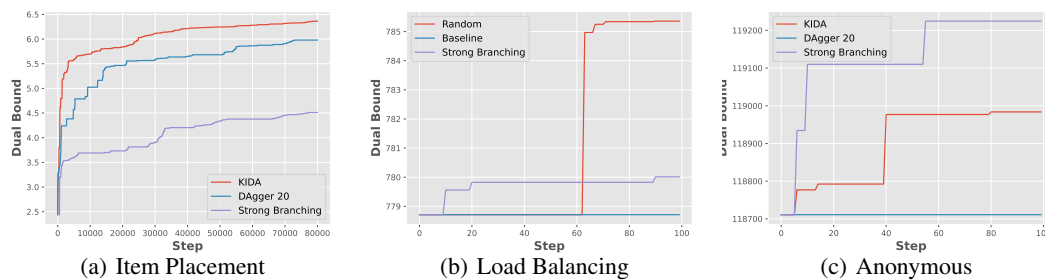


Figure 9: **Dual Bound Improvements with Steps.** Strong Branching may still fall behind GNN-based policy or even random policy

Item Placement and Anonymous. Further, our experiments show models closer to expert knowledge do not necessarily achieve better results, which indicates Strong Branching fails in some cases. More heuristic branching algorithms need to be taken into account to build more reasonable expert knowledge to facilitate learning to branch.

References

- [1] Tobias Achterberg, Thorsten Koch, and Alexander Martin. Branching rules revisited. *Operations Research Letters*, 33(1):42–54, 2005.
- [2] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129, 1995.
- [3] Michel B enichou, Jean-Michel Gauthier, Paul Girodet, Gerard Hentges, Gerard Ribiere, and O Vincent. Experiments in mixed-integer linear programming. *Mathematical Programming*, 1(1):76–94, 1971.
- [4] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.
- [5] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.
- [6] Gerald Gamrath, Daniel Anderson, Ksenia Bestuzheva, Wei-Kun Chen, Leon Eifler, Maxime Gasse, Patrick Gemander, Ambros Gleixner, Leona Gottwald, Katrin Halbig, et al. The scip optimization suite 7.0. 2020.
- [7] Maxime Gasse, Didier Ch etelat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. *arXiv preprint arXiv:1906.01629*, 2019.
- [8] Harvey J Greenberg. An analysis of degeneracy. *Naval Research Logistics Quarterly*, 33(4):635–655, 1986.
- [9] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.
- [10] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.
- [11] Łukasz Kidziński, Sharada Prasanna Mohanty, Carmichael F Ong, Zhewei Huang, Shuchang Zhou, Anton Pechenko, Adam Stelmazczyk, Piotr Jarosik, Mikhail Pavlov, Sergey Kolesnikov, et al. Learning to run challenge solutions: Adapting reinforcement learning methods for neuromusculoskeletal environments. In *The NIPS’17 Competition: Building Intelligent Systems*, pages 121–153. Springer, 2018.
- [12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [13] Ailsa H Land and Alison G Doig. An automatic method for solving discrete programming problems. In *50 Years of Integer Programming 1958-2008*, pages 105–132. Springer, 2010.
- [14] Jeff T Linderoth and Martin WP Savelsbergh. A computational study of search strategies for mixed integer programming. *INFORMS Journal on Computing*, 11(2):173–187, 1999.
- [15] Andrea Lodi and Giulia Zarpellon. On learning and branching: a survey. *Top*, 25(2):207–236, 2017.
- [16] St ephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.
- [17] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- [18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [19] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.