We would like to thank the reviewers for their feedback and comments, which we shall address below.

To Reviewer1

> Missing References and Comparisons: As you commented, and as written at the last sentence in the conclusion section, attention-based approaches can be used in our framework. Our contribution is not to develop permutation invariant networks, but to develop a few-shot learning method for heterogeneous attribute spaces using permutation invariant networks. Since Prototypical nets cannot handle heterogeneous attribute spaces, we did not compare with them. The computational time (hours) were: Ours: 7.5, DS:3.5, DS+FT:10.0, DS+MAML:34.2, NP:7.2, NP+FT:22.3, NP+MAML:101.0. We will include the missing references and computational complexity of baselines.

> Results on more realistic data benchmarks: Meta-Dataset is image data, and Hetro-lingual text classification dataset is text data. Their attribute sizes might be different, but the modality is shared. On the other hand, OpenML data contains datasets with different modality. Since our aim is to develop a model that can be learned from any datasets, we believe that OpenML is more suitable.

> quickly adapt to various tasks with heterogeneous attribute spaces was difficult with MAML. Why?: MAML learns good initial parameters that achieve good performance when finetuned. Good initial parameters would be different across various tasks with different attributes.

To Reviewer 2

> How would one train the task specific parameters on the unseen test tasks? By taking the support set as input, we can obtain the task specific parameters on the unseen test task using the neural networks that are shared across all tasks. Some meta-learning methods (e.g., matching networks and conditional neural processes) also use shared neural networks to obtain the task specific parameters.

> Do the inference network and prediction network together form a good prediction model?: The inference network infers the task specific parameters given the support set, which can be seen as training on regular supervised learning, where the training procedure is approximated by the neural networks. The prediction network predicts a response of an instance using the task specific parameters.

> compare to standard meta-learning methods on standard meta-learning datasets: We compared with standard meta-learning methods, MAML and NP, with heterogeneous datasets as written in our experiments. The standard meta-learning methods on standard meta-learning datasets are not fair since the standard meta-learning methods know that their attribute spaces are the same, but the proposed method does not know.

> $[\bar{\mathbf{v}}_i, x_{ni}]$ and $[\bar{\mathbf{c}}_j, y_{nj}]$, which are not even in the same feature space and does not even have the same feature dimension: x_{ni} and y_{ni} in Eq(2) are scalar values. $\bar{\mathbf{v}}_i$ and $\bar{\mathbf{c}}_j$ are the outputs of neural netowks g in Eq(1), and their dimensions are the same by using neural networks with the same output unit size. Therefore, $[\bar{\mathbf{v}}_i, x_{ni}]$ and $[\bar{\mathbf{c}}_j, y_{nj}]$ have the same dimension. We can use different functions f_u for $[\bar{\mathbf{v}}_i, x_{ni}]$ and $[\bar{\mathbf{c}}_j, y_{nj}]$, but used the same function for simplicity. We used Eq(2) to calculate the instance representation using all attributes and all responses.

To Reviewer3

> The artificial construction of the regression and classification: We admit that the classification task in our task is a bit artificial. But, we included the classification experiments to demonstrate that the proposed method is applicable to classification tasks. The regression experiments with OpenML demonstrates that our method can learn from various datasets.

> if prior knowledge existed about which subset of attributes were shared among pairs of tasks: Yes. We think the proposed method can be improved by sharing attribute representations for shared attributes.

To Reviewer 4

> it has no real-world motivating problem: We want to develop a model that can be learned from any datasets, and that can be used for an unseen task. For example, consider anomaly detection for various machines in various factories. Attributes (e.g., sensors) are different across machines. But, there are related machines. We want to detect anomalies for a new machine in a new factory with only a few labeled data, by utilizing data of existing machines. We include real-world motivating examples.

> complex approach: Although the difference of the performance between the proposed method and kNN was not large in our classification experiment, the performance of the proposed method was statistically better than that of kNN. We believe that our work is an important step for learning from a wide variety of datasets. Since there are no existing methods for solving this problem, we used the baselines that were not designed to solve the problem.