Improving Generalization in Meta-learning via Task Augmentation: Appendix

Huaxiu Yao[†][1] **Long-Kai Huang**[2] **Linjun Zhang**[3] **Ying Wei**[4] **Li Tian**[2]
James Zou[1] **Junzhou Huang**[2] **Zhenhui Li**[5]

A. Validity of Different Task Augmentation Strategies (Detailed Proof)

A.1. Proof of Corollary 1

Proof 1 (Proof of Corollary 1) *We check the validity of MetaMix as a task augmentation algorithm by examining whether the two criteria in Definition 1 in Section 3 are met. First, we check the increase of mutual information between predictions of the query set and the support set.*

$$\begin{aligned}
&I(\hat{\mathbf{Y}}^{mix};(\mathbf{X}^s,\mathbf{Y}^s)|\theta_0,\mathbf{X}^{mix})-I(\hat{\mathbf{Y}}^q;(\mathbf{X}^s,\mathbf{Y}^s)|\theta_0,\mathbf{X}^q)\\
=&H(\hat{\mathbf{Y}}^{mix}|\theta_0,\mathbf{X}^{mix})-H(\hat{\mathbf{Y}}^{mix}|\theta_0,\mathbf{X}^{mix},\mathbf{X}^s,\mathbf{Y}^s)-H(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q)+H(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s)\\
=&\mathbb{E}[-\log(p(\lambda\hat{\mathbf{Y}}^s+(1-\lambda)\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix}))]-\mathbb{E}[-\log(p(\lambda\hat{\mathbf{Y}}^s+(1-\lambda)\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix},\mathbf{X}^s,\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\lambda\hat{\mathbf{Y}}^s+(1-\lambda)\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^{mix},\mathbf{X}^s,\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix},\mathbf{X}^s,\mathbf{Y}^s))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\lambda\hat{\mathbf{Y}}^s+(1-\lambda)\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix},\mathbf{X}^s,\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
\geq&\mathbb{E}[-\log(p(\lambda\hat{\mathbf{Y}}^s+(1-\lambda)\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mix},\mathbf{X}^s,\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^q,\mathbf{X}^s))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^s))]=H(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^s)\geq 0
\end{aligned} \tag{1}$$

Note that the third and the sixth equality holds as the mapping $(\hat{\mathbf{Y}}^s,\hat{\mathbf{Y}}^q|\mathbf{X}^{mix},\mathbf{X}^s)\mapsto(\mathbf{Y}^{mix}|\mathbf{X}^{mix},\mathbf{X}^s)$ *is one-to-one after* λ *is specified. Besides, labels of the support (query) set are independent of features of the query (support) set, leading to the seventh equation. We investigate the capability of MetaMix producing additionally informative tasks in the following.*

$$\begin{aligned}
&I(\theta_0;\mathbf{X}^{mix},\mathbf{Y}^{mix}|\mathbf{X}^q,\mathbf{Y}^q)\\
=&H(\theta_0|\mathbf{X}^q,\mathbf{Y}^q)-H(\theta_0|\mathbf{X}^q,\mathbf{Y}^q,\mathbf{X}^{mix},\mathbf{Y}^{mix})\\
=&\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^q,\mathbf{Y}^q))]-\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^q,\mathbf{Y}^q,\mathbf{X}^{mix},\mathbf{Y}^{mix}))]\\
=&\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^q,\mathbf{Y}^q))]-\mathbb{E}[-\log(\frac{p(\mathbf{X}^{mix},\mathbf{Y}^{mix}|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{X}^q,\mathbf{Y}^q|\theta_0)p(\theta_0)}{p(\mathbf{X}^{mix},\mathbf{Y}^{mix}|\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{X}^q,\mathbf{Y}^q)})]
\end{aligned}$$

[†] Part of the work was done when H.Y. was a student at Penn State University. H.Y. and LK.H. contribute equally. [1] Stanford University, CA, USA [2] Tencent AI Lab, Shenzhen, China [3] Rutgers University, NJ, USA [4] City University of Hong Kong, Hong Kong [5] Pennsylvania State University, PA, USA. Correspondence to: Ying Wei <yingwei@cityu.edu.hk>.

Proceedings of the 38th *International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

$$
\begin{aligned}
&=\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^q,\mathbf{Y}^q))]-\mathbb{E}[-\log(\frac{p(\mathbf{X}^{mix},\mathbf{Y}^{mix}|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)}{p(\mathbf{X}^{mix},\mathbf{Y}^{mix}|\mathbf{X}^q,\mathbf{Y}^q)})]-\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^q,\mathbf{Y}^q))]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^{mix}|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{X}^{mix}|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)}{p(\mathbf{Y}^{mix}|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{X}^{mix}|\mathbf{X}^q,\mathbf{Y}^q)})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^s|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{Y}^q|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{X}^s|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{X}^q|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)}{p(\mathbf{Y}^s|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{Y}^q|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{X}^s|\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{X}^q|\mathbf{X}^q,\mathbf{Y}^q)})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^s|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{X}^s|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)}{p(\mathbf{Y}^s|\mathbf{X}^{mix},\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{X}^s|\mathbf{X}^q,\mathbf{Y}^q)})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^s|\mathbf{X}^s,\mathbf{X}^q,\mathbf{Y}^q,\theta_0)p(\mathbf{X}^s|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)}{p(\mathbf{Y}^s|\mathbf{X}^s,\mathbf{X}^q,\mathbf{Y}^q)p(\mathbf{X}^s|\mathbf{X}^q,\mathbf{Y}^q)})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{X}^s,\mathbf{Y}^s|\mathbf{X}^q,\mathbf{Y}^q,\theta_0)}{p(\mathbf{X}^s,\mathbf{Y}^s|\mathbf{X}^q,\mathbf{Y}^q)})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{X}^s,\mathbf{Y}^s|\theta_0)p(\theta_0)}{p(\mathbf{X}^s,\mathbf{Y}^s)})]+\mathbb{E}[-\log(p(\theta_0))]\\
&=H(\theta_0)-H(\theta_0|\mathbf{X}^s,\mathbf{Y}^s)). \qquad (2)
\end{aligned}
$$

This indicates that MetaMix contributes a novel task as long as the support set of the task being augmented is capable of reducing the uncertainty of the initialization θ_0, which is often the case. Again, we would also note that the sixth equation holds due to the one-to-one mapping mentioned above after λ is specified. The tenth equation holds because the support set is assumed to be sampled independently from the query set.

A.2. Analysis of MetaMix enhanced with channel shuffle

We consider the support and the query set with channel shuffle to be $\mathbf{X}^{s,cf}=\varphi_{cf}(\mathbf{X}^s)$ and $\mathbf{X}^{q,cf}=\varphi_{cf}(\mathbf{X}^q)$, where φ_{cf} is the non-linear function that replaces some channels of one class with the corresponding ones of the other class (refer to Eqn. (6) for the detailed discussion). Building on this, we validate the first criterion as follows.

$$
\begin{aligned}
&I(\hat{\mathbf{Y}}^{mmcf};(\mathbf{X}^{s,cf},\mathbf{Y}^s)|\theta_0,\mathbf{X}^{mmcf})-I(\hat{\mathbf{Y}}^q;(\mathbf{X}^s,\mathbf{Y}^s)|\theta_0,\mathbf{X}^q)\\
=&H(\hat{\mathbf{Y}}^{mmcf}|\theta_0,\mathbf{X}^{mmcf})-H(\hat{\mathbf{Y}}^{mmcf}|\theta_0,\mathbf{X}^{m+cf},\mathbf{X}^{s,cf},\mathbf{Y}^s)-H(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q)+H(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s)\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{mmcf}|\theta_0,\mathbf{X}^{mmcf}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mmcf},\mathbf{X}^{s,cf},\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
\geq&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{mmcf}|\theta_0,\mathbf{X}^{q,cf},\mathbf{X}^{s,cf}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{mmcf},\mathbf{X}^{s,cf},\mathbf{Y}^s))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^{q,cf},\mathbf{X}^{s,cf}))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{q,cf},\mathbf{X}^{s,cf}))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^{q,cf},\mathbf{X}^{s,cf},\mathbf{Y}^s))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]\\
&+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\varphi_{cf}))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\varphi_{cf}))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s,\varphi_{cf}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]\\
&+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\mathbf{X}^s,\mathbf{Y}^s))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^s,\varphi_{cf}))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\varphi_{cf}))]\\
&-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q,\varphi_{cf}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]\\
&+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^q|\theta_0,\mathbf{X}^q))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^s,\varphi_{cf}))]=H(\hat{\mathbf{Y}}^s|\theta_0,\mathbf{X}^s,\varphi_{cf})\geq 0 \qquad (3)
\end{aligned}
$$

In the next, we verify that the channel shuffle as expected produces a task that contributes more knolwedge to the initialization θ_0 compared to using MetaMix only, thereby improving meta-generalization.

$$I(\theta_0;\mathbf{X}^{mmcf},\mathbf{Y}^{mmcf}|\mathbf{X}^q,\mathbf{Y}^q)$$

$$\begin{aligned}
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^{mmcf}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)p(\mathbf{X}^{mmcf}|\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)}{p(\mathbf{Y}^{mmcf}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q})p(\mathbf{X}^{mmcf}|\mathbf{X}^{q},\mathbf{Y}^{q})})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^{s}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)p(\mathbf{Y}^{q}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)p(\varphi_{cf}(\mathbf{X}^{s})|\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)p(\varphi_{cf}(\mathbf{X}^{q})|\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)}{p(\mathbf{Y}^{s}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q})p(\mathbf{Y}^{q}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q})p(\varphi_{cf}(\mathbf{X}^{s})|\mathbf{X}^{q},\mathbf{Y}^{q})p(\varphi_{cf}(\mathbf{X}^{q})|\mathbf{X}^{q},\mathbf{Y}^{q})})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^{s}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)p(\varphi_{cf}(\mathbf{X}^{s})|\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)}{p(\mathbf{Y}^{s}|\mathbf{X}^{mmcf},\mathbf{X}^{q},\mathbf{Y}^{q})p(\varphi_{cf}(\mathbf{X}^{s})|\mathbf{X}^{q},\mathbf{Y}^{q})})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{Y}^{s}|\mathbf{X}^{s},\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0,\varphi_{cf})p(\mathbf{X}^{s},\varphi_{cf}|\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)}{p(\mathbf{Y}^{s}|\mathbf{X}^{s},\mathbf{X}^{q},\mathbf{Y}^{q},\varphi_{cf})p(\mathbf{X}^{s},\varphi_{cf}|\mathbf{X}^{q},\mathbf{Y}^{q})})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{X}^{s},\mathbf{Y}^{s},\varphi_{cf}|\mathbf{X}^{q},\mathbf{Y}^{q},\theta_0)}{p(\mathbf{X}^{s},\mathbf{Y}^{s},\varphi_{cf}|\mathbf{X}^{q},\mathbf{Y}^{q})})]\\
&=-\mathbb{E}[-\log(\frac{p(\mathbf{X}^{s},\mathbf{Y}^{s},\varphi_{cf}|\theta_0)p(\theta_0)}{p(\mathbf{X}^{s},\mathbf{Y}^{s},\varphi_{cf})})]+\mathbb{E}[-\log(p(\theta_0))]\\
&=H(\theta_0)-H(\theta_0|\mathbf{X}^{s},\mathbf{Y}^{s},\varphi_{cf}))\geq H(\theta_0)-H(\theta_0|\mathbf{X}^{s},\mathbf{Y}^{s})). \qquad (4)
\end{aligned}$$

A.3. Analysis of meta-augmentation

Meta-augmentation that augments a task by randomly injecting a noise to the labels of both support and query set is not an effective task augmentation method, as it fails to meet the second criterion in Definition 1. First of all, we check the validity of the first criterion.

$$\begin{aligned}
&I(\hat{\mathbf{Y}}^{q}+\epsilon;(\mathbf{X}^{s},\mathbf{Y}^{s}+\epsilon)|\theta_0,\mathbf{X}^{q})-I(\hat{\mathbf{Y}}^{q};(\mathbf{X}^{s},\mathbf{Y}^{s})|\theta_0,\mathbf{X}^{q})\\
=&H(\hat{\mathbf{Y}}^{q}+\epsilon|\theta_0,\mathbf{X}^{q})-H(\hat{\mathbf{Y}}^{q}+\epsilon|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}+\epsilon)-H(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q})+H(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s})\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}+\epsilon|\theta_0,\mathbf{X}^{q}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}+\epsilon|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}+\epsilon))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q}))]\\
&+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}))]\\
=&\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q}))]+\mathbb{E}[-\log(p(\epsilon|\theta_0,\mathbf{X}^{q}))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}+\epsilon))]\\
&-\mathbb{E}[-\log(p(\epsilon|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}+\epsilon))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q}))]+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}))]\\
=&\mathbb{E}[-\log(p(\epsilon))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}+\epsilon))]-\mathbb{E}[-\log(p(\epsilon|\mathbf{Y}^{s}+\epsilon))]\\
&+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}))]\\
\geq&\mathbb{E}[-\log(p(\epsilon))]-\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}))]-\mathbb{E}[-\log(p(\epsilon|\mathbf{Y}^{s}+\epsilon))]\\
&+\mathbb{E}[-\log(p(\hat{\mathbf{Y}}^{q}|\theta_0,\mathbf{X}^{q},\mathbf{X}^{s},\mathbf{Y}^{s}))]\\
=&H(\epsilon)\geq 0 \qquad (5)
\end{aligned}$$

Note that the third equality holds following the one-to-one assumption from $(\epsilon, x, y) \mapsto (x, y')$ in (Rajendran et al., 2020) and the fact that ϵ is independent of $\hat{\mathbf{Y}}^q$, and the fourth holds as ϵ is independent of $\mathbf{X}^s$, $\mathbf{X}^q$, and θ_0. The first inequality is due to the theorem that conditioning reduces entropy. In the next, we prove that the algorithm of meta-augmentation fails to generate tasks with additional information.

$$\begin{aligned}
&I(\theta_0;\mathbf{X}^{q},\mathbf{Y}^{q}+\epsilon|\mathbf{X}^{q},\mathbf{Y}^{q})\\
&=H(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q})-H(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q},\mathbf{X}^{q},\mathbf{Y}^{q}+\epsilon)\\
=&\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q}))]-\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q},\mathbf{X}^{q},\mathbf{Y}^{q}+\epsilon))]\\
=&\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q}))]-\mathbb{E}[-\log(\frac{p(\mathbf{X}^{q},\mathbf{Y}^{q},\mathbf{Y}^{q}+\epsilon|\theta_0)p(\theta_0)}{p(\mathbf{X}^{q},\mathbf{Y}^{q},\mathbf{Y}^{q}+\epsilon)})]\\
=&\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q}))]-\mathbb{E}[-\log(\frac{p(\mathbf{X}^{q},\mathbf{Y}^{q},\mathbf{Y}^{q}|\theta_0)p(\epsilon|\theta_0)p(\theta_0)}{p(\mathbf{X}^{q},\mathbf{Y}^{q})p(\epsilon)})]\\
=&\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q}))]-\mathbb{E}[-\log(p(\theta_0|\mathbf{X}^{q},\mathbf{Y}^{q}))]=0 \qquad (6)
\end{aligned}$$

B. Additional Information for MetaMix and MMCF

B.1. Figure for the Beta Distribution

For the $\mathrm{Beta}(\alpha, \beta)$ distribution, we illustrate both symmetric ($\alpha = \beta$) and skewed (i.e., $\alpha \neq \beta$) scenarios in Figure 1.

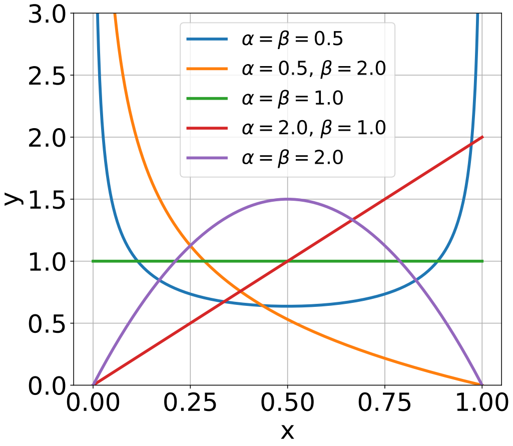


Figure 1. Illustration of the Beta Distribution. Here $\alpha = \beta$ and $\alpha \neq \beta$ represent the symmetric and skewed distributions, respectively.

B.2. Pseudo-codes

Take MAML-MMCF as examples, we show the pseudo-codes for meta-training in Alg. 1, respectively. The meta-testing process of MetaMix and MMCF are same, which is described in Alg. 2.

Algorithm 1 Meta-training Process of MAML-MMCF

```
Require: Task distribution $p(\mathcal{T})$; Learning rate $\mu$, $\eta$; Beta distribution parameters $\alpha$, $\beta$; MetaMix candidate layer set $\mathcal{C}$
Randomly initialize parameter $\theta_0$
while not converge do
    Sample a batch of tasks $\{\mathcal{T}_i\}_{i=1}^n$
    for all $\mathcal{T}_i$ do
        Sample support set $\mathcal{D}_i^s = \{(\mathbf{x}_{i,j}^s, \mathbf{y}_{i,j}^s)\}_{j=1}^{K^s}$ and query set $\mathcal{D}_i^q = \{(\mathbf{x}_{i,j}^q, \mathbf{y}_{i,j}^q)\}_{j=1}^{K^q}$ from $\mathcal{T}_i$
        Sample a mixed layer $l$ from $\mathcal{C}$
        Sample Channel Shuffle parameter $\mathbf{R}_{c,c'}$ for each pair of classes $c$ and $c'$
        Perform Channel Shuffle on the support set as (use a pair of classes as an example) via Eqn. (6) in the original paper: $\mathbf{X}_{i;c}^{s,cf} = \mathbf{R}_{c,c'} f_{\phi_i^l}(\mathbf{X}_{i;c}^s) + (\mathbf{I} - \mathbf{R}_{c,c'}) f_{\phi_i^l}(\mathbf{X}_{i;c'}^s),\ \mathbf{Y}_{i;c}^{s,cf} = \mathbf{Y}_{i;c}^s$.
        Compute the task-specific parameter $\phi_i$ via the inner-loop gradient descent, i.e., $\phi_i = \theta_0 - \mu \nabla_{\theta_0} \mathcal{L}(f_{\theta_0}(\mathbf{X}_i^{s,cf}), \mathbf{Y}_i^{s,cf})$
        Perform Channel Shuffle on the query set via Eqn. (6) in the original paper: $\mathbf{X}_{i;c}^{q,cf} = \mathbf{R}_{c,c'} f_{\phi_i^l}(\mathbf{X}_{i;c}^q) + (\mathbf{I} - \mathbf{R}_{c,c'}) f_{\phi_i^l}(\mathbf{X}_{i;c'}^q),\ \mathbf{Y}_{i;c}^{q,cf} = \mathbf{Y}_{i;c}^q$.
        Sample MetaMix parameter $\boldsymbol{\lambda} \sim \mathrm{Beta}(\alpha, \beta)$
        Forward both support and query sets and mixed them at layer $l$ as: $\mathbf{X}_{i,l}^{mmcf} = \boldsymbol{\lambda} f_{\phi_i^l}(\mathbf{X}_i^{s,cf}) + (\mathbf{I} - \boldsymbol{\lambda}) f_{\phi_i^l}(\mathbf{X}_i^{q,cf})$, $\mathbf{Y}_i^{mmcf} = \boldsymbol{\lambda} \mathbf{Y}_i^{s,cf} + (\mathbf{I} - \boldsymbol{\lambda}) \mathbf{Y}_i^{q,cf}$
        Continual forward $\mathbf{X}_{i,l}^{mix}$ to the rest of layers and compute the loss as $\mathcal{L}(f_{\phi_i^{L-l}}(\mathbf{X}_{i,l}^{mmcf}), \mathbf{Y}_i^{mmcf})$
    end for
    Update $\theta_0 \leftarrow \theta_0 - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\lambda} \sim \mathrm{Beta}(\alpha,\beta)} \mathbb{E}_{l \sim \mathcal{C}} [\mathcal{L}(f_{\phi_i^{L-l}}(\mathbf{X}_{i,l}^{mmcf}), \mathbf{Y}_i^{mmcf})]$
end while
```

Algorithm 2 Meta-testing Process of both MAML-MetaMix or MAML-MMCF

```
Require: Learning rate $\mu$; Optimized parameter $\theta_0^*$ via MMCF or MetaMix
Compute the task-specific parameter $\phi_t$ as $\phi_t = \theta_0^* - \mu \nabla_{\theta_0} \mathcal{L}(f_{\theta_0}(\mathbf{X}_t^s), \mathbf{Y}_t^s)$
Predict $\hat{\mathbf{Y}}_t^q$ on the query set $\mathcal{D}_t^q$
Evaluate the performance via predicted value $\hat{\mathbf{Y}}_t^q$ and actual value $\mathbf{Y}_t^q$
```

C. Detailed Proof of Generalization Analysis

Proof 2 (Proof of Theorem 1) *We first state a standard uniform deviation bound based on Rademacher complexity (c.f. (Bartlett & Mendelson, 2002))*

Lemma 1 *Let the sample $\{z_1, ..., z_N\}$ be drawn i.i.d. from a distribution P over $\mathcal{Z}$ and let $\mathcal{G}$ denote a class of functions on $\mathcal{Z}$ with members mapping from $\mathcal{Z}$ to $[a, b]$. Then for $\delta > 0$, we have that with probability at least $1 - \delta$ over the draw of the sample,*

$$\sup_{g \sim G} \|\mathbb{E}_{\hat{P}} g(z) - \mathbb{E}_P g(z)\| \leq 2R(G; z_1, ..., z_n) + \sqrt{\frac{\log(1/\delta)}{n}}, \tag{7}$$

where $R(G; z_1, ..., z_n)$ denotes the Rademacher complexity of the function class $\mathcal{G}$.

We now write $R(\{\mathbf{Z}_i\}_{i=1}^{n_T}) - R$ as

$$\begin{aligned} R(\{\mathbf{Z}_i\}_{i=1}^{n_T}) - R =& \mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \hat{p}(\mathcal{T}_i)} \mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q) - \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathcal{T}_i} [\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q)] \\ =& \underbrace{\mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \hat{p}(\mathcal{T}_i)} \mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q) - \mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathcal{T}_i} [\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q)]}_{(i)} \\ &+ \underbrace{\mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathcal{T}_i} \mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q) - \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathcal{T}_i} [\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q)]}_{(ii)} \end{aligned}$$

Recall that we consider the function $f_{\phi_i}(\mathbf{X}_i) = \phi_i^\top \sigma(\mathbf{W}\mathbf{X}_i)$ and the function class

$$\mathcal{F}_{\mathcal{T}} = \{\phi^\top \sigma(\mathbf{W}\mathbf{X}) : \phi^\top \Sigma_{\sigma, \mathcal{T}} \phi \leq \gamma\}. \tag{8}$$

For each $\mathcal{T}_i$, let us consider $f_{\phi_i}(\cdot) \in \mathcal{F}_{\mathcal{T}}$. By Theorem A.1 in (Zhang et al., 2021), we have the following result for the Rademacher complexity:

$$\begin{aligned} R(\mathcal{F}_{\mathcal{T}}; z_1, ..., z_n) \leq& 2\sqrt{\frac{\gamma \cdot (rank(\Sigma_{\sigma,\mathcal{T}}) + \|\Sigma_{\sigma,\mathcal{T}}^{\mathbf{W}\dagger/2} \mu_{\sigma,\mathcal{T}}\|)}{K^m}} \\ \leq& 2\sqrt{\frac{\gamma \cdot (r + B)}{K^m}}. \end{aligned} \tag{9}$$

Then the first term (i) can be bounded as below.

$$\begin{aligned} &\mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \hat{p}(\mathcal{T}_i)} \mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q) - \mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathcal{T}_i} [\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q)] \\ \leq& \mathbb{E}_{\mathcal{T}_i \sim \hat{p}(\mathcal{T})} |\mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \hat{p}(\mathcal{T}_i)} \mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q) - \mathbb{E}_{(\mathbf{X}_i, \mathbf{Y}_i) \sim \mathcal{T}_i} [\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q), \mathbf{Y}_i^q)] \\ \leq& C_1 \sqrt{\frac{\gamma \cdot (r + B)}{K^m}} + C_2 \sqrt{\frac{\log(n_T/\delta)}{K^m}}, \end{aligned} \tag{10}$$

where the additional $\log(n_T)$ term in the last inequality above is due to we take union bound on n_T tasks.

Denote function $g : \mathcal{T} \to \mathbb{R}$ such that $g_f(\mathcal{T}) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{T}}(\mathcal{L}(f_\phi(\mathbf{X}), \mathbf{Y}))$. Denote

$$\mathcal{G} = \{g_f(\mathcal{T}) : g_f(\mathcal{T}) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim \mathcal{T}}(\mathcal{L}(f_\phi(\mathbf{X}), \mathbf{Y})), f_\phi \in \mathcal{F}_{\mathcal{T}}\}. \tag{11}$$

The second term (ii) requires computing the Rademacher complexity for the function class over distributions

$$
\begin{aligned}
R(\mathcal{G};\mathcal{T}_1,...,\mathcal{T}_{n_T}) =&\mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n_T}|\sum_{i=1}^{n_T}\sigma_i g(\mathcal{T}_i)| = \mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n_T}|\sum_{i=1}^{n_T}\sigma_i\mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{T}_i}(f_{\phi_i}(\mathbf{X})-\mathbf{Y})^2|\\
\le&\mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n_T}|\sum_{i=1}^{n_T}\sigma_i\mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{T}_i}f_{\phi_i}(\mathbf{X})| + \mathbb{E}\sup_{g\in\mathcal{G}}\frac{1}{n_T}|\sum_{i=1}^{n_T}\sigma_i\mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim\mathcal{T}_i}\mathbf{Y}|\\
\le&\mathbb{E}\sup_{g\in G}\frac{1}{n_T}|\sum_{i=1}^{n_T}\sigma_i(\Sigma^2_{\sigma,\mathcal{T}}\phi_i)^\top\Sigma^{\dagger/2}_{\sigma,\mathcal{T}}\mu_{\sigma,\mathcal{T}}| + \sqrt{\frac{1}{n_T}}\\
\le&\sqrt{\frac{\gamma\cdot B+1}{n_T}}
\end{aligned}
\tag{12}
$$

Then we have the following bound on (ii):

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{T}_i\sim\hat{p}(\mathcal{T})}\mathbb{E}_{(\mathbf{X}_i,\mathbf{Y}_i)\sim\mathcal{T}_i}\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q),\mathbf{Y}_i^q) - \mathbb{E}_{\mathcal{T}_i\sim p(\mathcal{T})}\mathbb{E}_{(\mathbf{X}_i,\mathbf{Y}_i)\sim\mathcal{T}_i}[\mathcal{L}(f_{\phi_i}(\mathbf{X}_i^q),\mathbf{Y}_i^q)]\\
\le& C_3\sqrt{\frac{\gamma\cdot(B+1)}{n_T}} + C_4\sqrt{\frac{\log(1/\delta)}{n_T}}.
\end{aligned}
\tag{13}
$$

Combining the pieces, we obtain the desired result. With probability at least $1-\delta$,

$$
|R(\{\mathbf{Z}_i\}_{i=1}^{n_T}) - R| \le C_1\sqrt{\frac{\gamma\cdot(r+B)}{K^m}} + C_2\sqrt{\frac{\log(n_T/\delta)}{K^m}} + C_3\sqrt{\frac{\gamma\cdot B+1}{n_T}} + C_4\sqrt{\frac{\log(1/\delta)}{n_T}}. \tag{14}
$$

Besides the detailed proof, we also provide the empirical results to show the equivalence between the symmetric version of MAML-MetaMix for generalization analysis (i.e., Mixup($\mathcal{D}^s\oplus\mathcal{D}^q,\mathcal{D}^s\oplus\mathcal{D}^q$)) and the proposed MAML-MetaMix (i.e., Mixup($\mathcal{D}^s,\mathcal{D}^q$)). The experiments are conducted on both omniglot and miniImagenet under the non-exclusive setting. In Table 1, we report the comparison results:

Table 1. Performance comparison between Mixup($\mathcal{D}^s\oplus\mathcal{D}^q,\mathcal{D}^s\oplus\mathcal{D}^q$) and Mixup($\mathcal{D}^s,\mathcal{D}^q$).

Model	Omniglot		MiniImagenet	
	20-way 1-shot	20-way 5-shot	5-way 1-shot	5-way 5-shot
MAML-MetaMix (Mixup($\mathcal{D}^s,\mathcal{D}^q$))	$91.53\pm0.53\%$	$97.63\pm0.15\%$	$38.53\pm1.79\%$	$57.55\pm1.01\%$
MAML-MetaMix (Mixup($\mathcal{D}^s\oplus\mathcal{D}^q,\mathcal{D}^s\oplus\mathcal{D}^q$))	$91.93\pm0.52\%$	$97.17\pm0.16\%$	$38.27\pm1.72\%$	$57.38\pm0.96\%$

Proof 3 (Proof of Theorem 2) *To prove Theorem 2, first, we would like to note that since* $\frac{1}{K^{m_0}}\sum_{j=1}^{K^{m_0}}\sigma(\mathbf{W}\mathbf{x}_{i,j;0}) = \frac{1}{K^{m_1}}\sum_{j=1}^{K^{m_1}}\sigma(\mathbf{W}\mathbf{x}_{i,j;1}) = 0$*, we have*

$$
\mathbb{E}[\mathbf{x}_{i,j;k}^{cf}\mid\mathbf{x}_{i,j;k}] = \sigma(\mathbf{W}\mathbf{x}_{i,j;k}). \tag{15}
$$

Recall that $\mathcal{L}(f_{\phi_i}(\mathbf{X}_i),\mathbf{Y}_i) = \frac{1}{2}(f_{\phi_i}(\mathbf{X}_i)-\mathbf{Y}_i)^2$. *Then let us compute the second-order Taylor expansion on* $\mathcal{L}(\mathbf{Z}_i^{cf}) = \frac{1}{K^m}\sum_{j=1}^{K^m}\mathcal{L}({\phi_i}^\top(\mathbf{x}_{i,j}^{cf}),y_{i,j})$ *with respect to* $(\mathbf{x}_{i,j}^{cf})$ *around* $\mathbb{E}[\mathbf{x}_{i,j;k}^{cf}\mid\mathbf{x}_{i,j;k}] = \sigma(\mathbf{W}\mathbf{x}_{i,j;k})$*, we have that the Taylor expansion of* $\mathbb{E}_\xi\mathcal{L}(\mathbf{Z}_i^{cf})$ *up to the second-order equals to*

$$
\mathcal{L}(\mathbf{Z}_i) + \frac{1}{K^m}\sum_{j=1}^{K^{m_0}} + \frac{1}{K^m}\sum_{j=1}^{K^m}\phi_i^\top Cov(\mathbf{x}_{i,j}^{cf}\mid\mathbf{Z}_i)\phi_i \tag{16}
$$

Let us denote $\sigma(\mathbf{W}\mathbf{x}_{i,j;k})$ *by* $\mathbf{x}_{i,j;k}^\sigma$.

For the quadratic term, we have that given $\mathbf{Z}_i$

$$\begin{aligned} Cov(\mathbf{x}^{cf}_{i,j}) =& \frac{1}{\delta^2} Cov((\mathbf{R}\sigma(\mathbf{W}\mathbf{x}_{i,j;k}) + (\mathbf{I}-\mathbf{R})\sigma(\mathbf{W}\mathbf{x}_{i,j';1-k}))) \\ =& \frac{1}{\delta^2}(Cov\left(\mathbf{R}\mathbf{x}^{\sigma}_{i,j;k}\right) + Cov\left(\mathbf{R}\mathbf{x}^{\sigma}_{i,j;k}, (\mathbf{I}-\mathbf{R})\mathbf{x}^{\sigma}_{i,j';1-k}\right) + Cov\left(\mathbf{I}-\mathbf{R})\mathbf{x}^{\sigma}_{i,j';1-k}\right)) \\ =& \frac{1}{\delta^2}(\delta(1-\delta)diag(\mathbf{x}^{\sigma\circ 2}_{i,1;k}) + 0 + \delta(1-\delta)\frac{1}{K^{m_{1-k}}}\sum_{j=1}^{K^{m_{1-k}}}\mathbf{x}^{\sigma}_{i,j;1-k}\mathbf{x}^{\sigma\top}_{i,j;1-k}) \end{aligned} \tag{17}$$

Plugging into Eq (16)*, we obtain*

$$\begin{aligned} &\mathcal{L}(\mathbf{Z}_i) + \frac{1-\delta}{\delta}\phi_i^{\top}(\frac{1}{K^m}\sum_{j=1}^{K^m} diag(\sigma(\mathbf{W}\mathbf{x}_{i,j})^{\circ 2})\phi_i + \\ &+ \phi_i^{\top}(\frac{1}{K^{m_0}}\sum_{j=1}^{K^{m_0}}\sigma(\mathbf{W}\mathbf{x}_{i,j;0})\sigma(\mathbf{W}\mathbf{x}_{i,j;0})^{\top} + \frac{1}{K^{m_1}}\sum_{j=1}^{K^{m_1}}\sigma(\mathbf{W}\mathbf{x}_{i,j;1})\sigma(\mathbf{W}\mathbf{x}_{i,j;1})^{\top})\phi_i. \end{aligned} \tag{18}$$

D. Detailed Experimental Setup

In this section, we provide more details of the experimental setups of our paper. All experiments are run on a GPU cluster and implemented by Tensorflow (Abadi et al., 2016). In the next, we discuss the setups for all the problems, including drug activity prediction, pose prediction, and image classification.

D.1. Drug Activity Prediction

For drug activity prediction, we use the publicly available dose-response activity assays from ChEMBL[1] and preprocessed in (Martin et al., 2019). All 4,276 assays, as 4,276 tasks, are accessible and downloadable from this site[2]. In each assay, there are a few training drug compounds with biologically tested activities against the target protein in this assay, as well as several testing compounds. The split of training and testing compounds follows the realistic split in (Martin et al., 2019). The number of drug compounds varies from assay to assay, with an median of only 70 drug compounds per assay. To describe each compound, we follow (Martin et al., 2019) to use 1,024 dimensional Morgan fingerprint implemented in RDkit[3]. As mentioned in the experimental section in the main text, we randomly take 100 assays as meta-testing assays, and 76 assays for meta-validation and the rest of 4100 assays for meta-training. Here we report the assay IDs that belong to meta-validation and meta-testing, respectively, for all the four groups. Note that due to space limit we do not report the assay IDs for meta-training, which can be easily obtained by deducting the meta-validation and meta-testing assays from all 4276 assays.

- **Group 1**
  - *Meta-validation:* 972800, 688641, 610565, 1536390, 211079, 1625735, 1641357, 688654, 1641103, 457234, 450707, 195220, 1366808, 49308, 924, 49312, 828065, 737313, 1528100, 596645, 1641767, 1535401, 688427, 969260, 453677, 978479, 1641008, 574385, 911154, 446257, 878513, 1640955, 902584, 1276473, 752567, 306492, 736957, 1640384, 1454018, 2755, 579907, 1527622, 761927, 89542, 809158, 978889, 556876, 478840, 688464, 1330005, 144341, 1528791, 1301597, 1641310, 209245, 608993, 1528801, 89064, 1527913, 4202, 688616, 1513, 510189, 1641197, 1527791, 688495, 89839, 1641201, 1528688, 752371, 688379, 938230, 596087, 835704, 566779, 688767.
  - *Meta-testing:* 752640, 972801, 737284, 954885, 1528837, 1587725, 1527823, 1640977, 157713, 1285138, 1437208, 1349151, 1592870, 93228, 465460, 954934, 84556, 1567308, 1577550, 1285709, 654928, 620647, 864364, 575603, 1280627, 688257, 1443970, 1527947, 737424, 201877, 1457820, 603293, 809120, 883875, 1641128, 1534634, 1641655, 955073, 954571, 736971, 577227, 45264, 455393, 728290, 688357, 1301747, 105205, 865015, 665348, 820998, 759559, 1301769, 609034, 80649, 1641240, 965916, 34078, 1470241, 1348900, 333106, 1527607, 954703, 1641298, 1641300, 727385, 304989, 981861, 212325,

[1] https://www.ebi.ac.uk/chembl

[2] https://pubs.acs.org/doi/10.1021/acs.jcim.9b00375#i21

[3] http://rdkit.org/

756584, 331630, 473976, 63356, 51590, 1640328, 954762, 1642379, 1527698, 1527704, 543133, 954781, 1301405, 619939, 605612, 585134, 1433006, 934321, 1642435, 1637320, 936907, 54735, 70610, 1508820, 1292758, 104407, 992729, 199642, 160234, 1528304, 629753, 931327.

- **Group 2**
    - *Meta-validation:* 7296, 1276546, 87173, 688645, 1350406, 955016, 697223, 1163, 201739, 809231, 1528850, 1528212, 752533, 971798, 954388, 1626011, 1528480, 501795, 1527972, 470053, 1640867, 809128, 737064, 1642538, 954282, 978478, 786095, 29233, 1642418, 737075, 1536179, 1641399, 1527735, 609465, 1640506, 1641659, 307259, 1537597, 769089, 140229, 789189, 860488, 766795, 48587, 1528909, 1451727, 219472, 737105, 955090, 311637, 1528022, 1632983, 727385, 1456602, 1641179, 688347, 67039, 434528, 1564001, 727521, 688483, 595939, 1436004, 736997, 1528160, 1640426, 4202, 102381, 45422, 1641073, 47858, 37363, 1641720, 688889, 1301756, 556797.
    - *Meta-testing:* 835072, 539657, 1641997, 1639955, 1638422, 1639959, 1622038, 637980, 28188, 91168, 954915, 425511, 688685, 155185, 39493, 155208, 1641035, 1288277, 755797, 954462, 812132, 87656, 1536113, 48248, 744057, 210045, 1642144, 50337, 325795, 1527974, 1642150, 814256, 1641143, 438974, 217297, 1641170, 688340, 1641688, 688357, 649964, 930033, 447747, 566532, 1641737, 49425, 562451, 817939, 688403, 817944, 52506, 452895, 984872, 311595, 899888, 646978, 1642307, 664904, 1641802, 1466703, 1466704, 809297, 147797, 1640791, 305497, 209245, 603488, 701282, 752485, 302952, 122731, 563052, 1561972, 1528692, 1642361, 1528698, 737150, 1301374, 51590, 364426, 1642378, 899993, 752538, 1640355, 1446827, 62394, 842684, 1640893, 44489, 688589, 208335, 1642449, 858065, 1640919, 1528791, 1528294, 1520, 1640952, 92156, 63997, 69119.
- **Group 3**
    - *Meta-validation:* 856700, 688641, 448646, 1613063, 1301767, 624014, 559247, 1527953, 1640339, 49558, 737046, 809242, 1642522, 1641371, 1527965, 592925, 954655, 688416, 305569, 538786, 1535011, 208672, 1592863, 688550, 1527974, 1527976, 1642272, 305065, 809259, 1640189, 96941, 688685, 954799, 978480, 934321, 1637168, 29233, 45236, 306221, 1535033, 1640506, 1290683, 158524, 936637, 647615, 422463, 1459648, 1640904, 954953, 1361352, 654923, 1641164, 954959, 1301583, 688210, 1508820, 45272, 688346, 737371, 162397, 775393, 535396, 1301477, 4197, 651627, 75756, 3819, 737391, 1641201, 1528692, 1294964, 456311, 737273, 1301756, 1640573, 42878.
    - *Meta-testing:* 688643, 688645, 159749, 1527820, 539663, 303638, 1638422, 954399, 330271, 70695, 1295917, 1642542, 1527862, 1528890, 200254, 540741, 1365575, 761928, 688201, 37966, 1537108, 336476, 511069, 1301599, 1528416, 206959, 478840, 1503357, 1642117, 1536654, 688274, 1527963, 688288, 688293, 688816, 745138, 438974, 88771, 459971, 714443, 1289425, 1451729, 1284820, 954602, 954604, 208118, 198910, 809216, 610565, 1528071, 453897, 770827, 216843, 828171, 306447, 562451, 1642271, 1528097, 28965, 367910, 1642296, 1528125, 1528145, 1555281, 49489, 493905, 876885, 1290079, 468834, 1528677, 756582, 1338728, 1528170, 845165, 1528696, 617338, 1301888, 102785, 1527682, 940424, 1528724, 809380, 1446827, 864186, 1291714, 642499, 688586, 1513931, 32721, 954834, 1536468, 688598, 1556442, 901084, 954845, 1527780, 1640932, 477677, 829947, 1528829.
- **Group 4**
    - *Meta-validation:* 688391, 1641992, 1641737, 306314, 311816, 813068, 68748, 1536775, 823822, 1639955, 696215, 1469079, 971801, 41884, 637980, 1298461, 76063, 688416, 1637151, 619938, 1642274, 885155, 572966, 510887, 1641000, 1528488, 954411, 1642415, 138287, 1527985, 56498, 27571, 458930, 311855, 809146, 307259, 950588, 736957, 1527742, 809152, 592450, 809156, 1528648, 954957, 209231, 490576, 1528273, 1641170, 45265, 1528917, 1528149, 1292759, 809175, 53367, 737370, 822749, 154333, 67039, 737273, 1640162, 737379, 763492, 809193, 954987, 104172, 510189, 1528695, 208754, 688243, 45044, 954482, 1640307, 1528183, 1642489, 1527674, 688254.
    - *Meta-testing:* 1556484, 688644, 737290, 1301520, 1642001, 1528850, 688661, 971799, 1642520, 46624, 1537067, 1641005, 1641010, 688185, 954938, 443966, 599616, 439367, 1537607, 954956, 688719, 615506, 654934, 1589851, 104542, 457824, 1641573, 1527916, 737391, 1301619, 211078, 52874, 955024, 32404, 158358, 566940, 50848, 1527970, 1349288, 1586856, 860330, 688818, 1536181, 48316, 491718, 1296583, 954567, 1566412, 66255, 66267, 809183, 425699, 467683, 464617, 752377, 508163, 872708, 1640197, 1301765, 809221, 809239, 1528102, 809255, 1536298, 1640747, 688431, 1636657, 213817, 1466703, 1301330, 1545042, 737622, 1452895, 950625, 856937, 493931, 954743, 809346, 558984, 1642378, 591251, 1640852, 305050, 934299, 1640862, 1640351, 1642418, 558515, 1511354, 1330619, 1528770, 1291715, 901575, 1640904, 1439182, 1537998, 737235, 1301469, 37371, 797692.

The base model of drug activity prediction is a two-layer Multilayer Perceptron(MLP) neural network with 500 neurons in each layer. Each fully connected layer is followed by a batch normalization layer and leaky ReLU activation (negative slope is 0.01). In $\mathrm{Beta}(\alpha, \alpha)$, we set $\alpha = 0.5$. We set the candidate layer $\mathcal{C}$ as layer 1 and layer 2. During meta-training, the task batch size, the outer-loop learning rate, the inner-loop learning rate are set to 8, 0.001, and 0.01, respectively. The meta-training process altogether runs for 50 epochs, each of which includes 500 iterations. In either meta-training or meta-testing, the number of inner-loop adaptation steps equals to 5.

D.2. Pose Prediction

In the pose prediction problem, we follow (Yin et al., 2020) to preprocess the pose tasks[4]. The meta-training and meta-testing include 50 and 15 categories, respectively, where each category contains 100 gray images in the size of 128×128.

In pose prediction, following (Yin et al., 2020), the base model is comprised of a fixed encoder with three convolutional blocks and an adapted decoder with four convolutional blocks. Each convolutional block is composed of a convolutional layer, a batch normalization layer and a ReLU activation layer. During the inner-loop optimization, we fix the encoder and only update the parameters in the decoder (i.e., the encoder layers are only meta-updated in the outer-loop optimization). For the hyperparameters in pose prediction, both inner-loop and outer-loop learning rates are set as 0.01. In addition, we set the hyperparameter α in the Beta distribution as 0.5 and the number of adaptation steps in the inner-loop optimization as 5. The candidate set $\mathcal{C}$ for mixup is set to include the input layer (layer 0) as well as all hidden layers (i.e., layer 1, layer 2, and layer 3). All hyperparameters are selected according to the performance on the meta-validation set (10 categories), which are randomly selected from the meta-training categories.

D.3. Image Classification

In image classification, the image sizes of Omniglot, MiniImagenet, Multi-dataset are set to be $28 \times 28 \times 1$, $84 \times 84 \times 3$ and $84 \times 84 \times 3$, respectively. Under the non-mutually exclusive setting, taking 5-way miniImagenet as an example, 64 meta-training classes are split to 5 sets, where 4 sets have 13 classes and the rest one has 12 classes. For each set, a fixed class label is assigned to each class within this set, which remains unchanged across different tasks. During meta-training, we randomly select one class from each set and take all the five selected classes to construct a task, which ensures that each class consistently has one label across tasks. In our experiments, we list the classes within each set as follows.

- **Set 1**: n07584110, n04243546, n03888605, n03017168, n04251144, n02108551, n02795169, n03400231, n03476684, n04435653, n02120079, n01910747, n03062245
- **Set 2**: n03347037, n04509417, n03854065, n02108089, n04067472, n04596742, n01558993, n04612504, n02966193, n07697537, n01843383, n03838899, n02113712
- **Set 3**: n04604644, n02105505, n02108915, n03924679, n01704323, n09246464, n04389033, n03337140, n06794110, n04258138, n02747177, n13054560, n04443257
- **Set 4**: n13133613, n01770081, n02606052, n02687172, n02101006, n03676483, n04296562, n02165456, n04515003, n01749939, n02111277, n02823428, n01532829
- **Set 5**: n02091831, n07747607, n03998194, n02089867, n02074367, n02457408, n04275548, n03220513, n03527444, n03908618, n03207743, n03047690

A similar process is applied to Omniglot, where 1200 meta-training classes are randomly split into 20 sets with 60 classes in each set. In Multi-dataset, each subdataset is split into 5 sets. In the subdatasets Bird, Aircraft and Fungi, we have 4 sets each of which includes 13 classes while the rest one includes 12. In the subdataset Texture, however, each set contains 6 classes.

For all datasets, we utilize the classical convolutional neural network with 4 convolutional blocks as the base model (Finn et al., 2017; Snell et al., 2017). It is worth to mention that (Yin et al., 2020) adopts a deeper network as the base model under the non-mutually exclusive setting. The deeper network includes 3 convolutional layers with a fully connected layer as the encoder and 3 convolutional decoder layers, where the encoder is fixed during inner-loop optimization. In our practice, the shallower network achieves better performance in all meta-learning algorithms, as a result of more serious overfitting issues caused by the deeper network. In Table 2, we illustrate the comparison of pre inner-update accuracy and meta-testing post

[4] code link: https://github.com/google-research/google-research/tree/master/meta_learning_without_memorization/pose_data

inner-update accuracy during meta-training under the Omniglot 20-way, 1-shot setting, where MAML, MR-MAML are included as baselines. The results indicate that the deeper structure is easier to memorize all data samples via the learned initialization; therefore, we adopt the shallow network (i.e., standard 4-block convolutional layers) in this experiment.

Table 2. Comparison between the shallow and deeper base model under the Omniglot 20-way 1-shot setting.

Methods	Meta-training Pre-update		Meta-testing Post-update	
	Shallow	Deep	Shallow	Deep
MAML	$14.38 \pm 0.40\%$	$98.59 \pm 0.05\%$	$87.40 \pm 0.59\%$	$8.82 \pm 0.42\%$
MR-MAML	$5.63 \pm 0.36\%$	$5.12 \pm 0.34\%$	$89.28 \pm 0.59\%$	$83.75 \pm 0.67\%$

For hyperparameter settings, in both MiniImagenet and Multi-dataset, the inner-loop learning rate μ and the outer-loop learning rate η are set as 0.01, 0.001, respectively. In Omniglot, μ and η are set as 0.1, 0.005, respectively. The hyperparameter α of the Beta distribution $\mathrm{Beta}(\alpha, \alpha)$ is set as 2.0 for all datasets. Besides, the candidate layer set $\mathcal{C}$ for both MiniImagenet and Multi-dataset is set as layer (0, 1, 2, 3). In Omniglot, the candidate set $\mathcal{C}$ is set as layer (1, 2, 3). All hyperparameters are determined by the performance on the meta-validation set.

E. Additional Results for Drug Activity Prediction

E.1. Hyperparameter Analysis on Drug Activity Prediction

E.1.1. Analysis of the Candidate Layer Set $\mathcal{C}$

We further analyze the effect of different candidate layer sets $\mathcal{C}$ on drug activity prediction. The results are reported in Table 3. Compared with $\mathcal{C} = 2$, $\mathcal{C} = 1$ leads to higher performances, suggesting that mixing low-level representations with the resulting compactness contributes more to the overall improvement. Furthermore, mixing all layers (i.e., $\mathcal{C} = (1, 2)$) achieves the best performance, indicating the necessity of jointly mixing the representations in all levels.

Table 3. Effect of the candidate layer set $\mathcal{C}$ in MetaMix.

Mixed layers $\mathcal{C}$	Group 1			Group 2			Group 3			Group 4		
	Mean	Med.	>0.3	Mean	Med.	>0.3	Mean	Med.	>0.3	Mean	Med.	>0.3
(1)	0.412	0.362	59	0.326	0.302	50	0.355	0.318	51	0.390	0.349	56
(2)	0.405	0.324	51	0.324	0.256	44	0.354	0.304	50	0.387	0.353	57
(1,2)	**0.413**	**0.393**	59	**0.337**	**0.301**	**51**	**0.381**	**0.362**	**55**	**0.380**	**0.348**	**55**

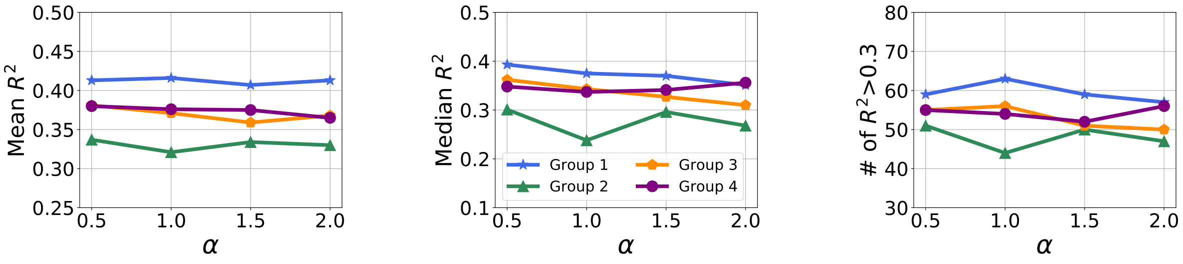
E.1.2. Analysis of the Mixup Ratio

In MetaMix, the mixup ratio for the support and the query sets are controlled by the parameter λ, which is sampled from the Beta distribution $\mathrm{Beta}(\alpha, \alpha)$. Here, we analyze the performance w.r.t. the change of mixup ratio. Specifically, we conduct two experiments: (1) we analyze the performance concerning the change of hyperparameter α; (2) we fix the mixup ratio λ without being sampled from the Beta distribution. The results for the experiments (1) and (2) are shown in Figure 2 and Figure 3, respectively. In the analysis of α, though the overall performance is slightly better when $\alpha = 0.5$, our MetaMix strategy is still robust and not very sensitive to the shape of Beta distribution (i.e., different α). The conclusion is further strengthened by the analysis of fixed λ in Figure 3, where the performance remains relative stable between $\lambda \in [0.4, 0.75]$.

F. Additional Results for Pose Prediction

F.1. Effect of Mixup Strategies on Pose Prediction

In pose prediction, the performance w.r.t. different mixup strategies are reported in Table 4. The superiority of MetaMix over the other strategies further corroborates our analysis that MetaMix is capable of improving the meta-generalization by enhancing the dependence on the support set in the outer-loop optimization.

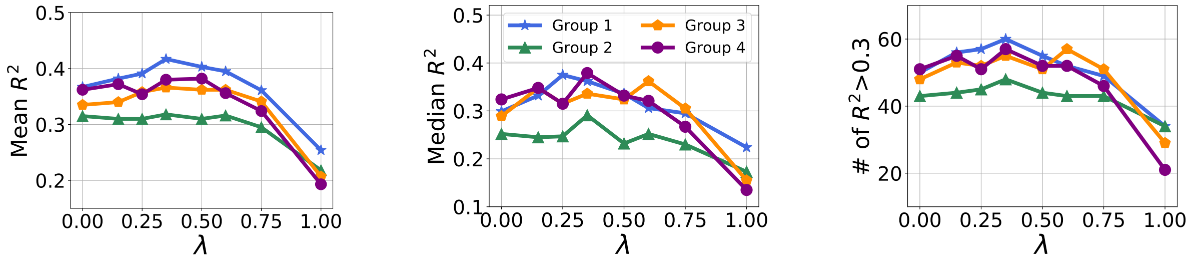


(a)

(b)

(c)

Figure 2. Performance on drug activity prediction w.r.t. the change of α in $\mathrm{Beta}(\alpha, \alpha)$. The three subfigures (a), (b), (c) represent the results under different evaluation metrics.



(a)

(b)

(c)

Figure 3. Performance w.r.t. the fixed λ in MetaMix (i.e., $\boldsymbol{\lambda} f_{\phi_i^l}(\mathbf{X}_i^s) + (\mathbf{I} - \boldsymbol{\lambda}) f_{\phi_i^l}(\mathbf{X}_i^q)$). The three subfigures (a), (b), (c) show the performance under different evaluation metrics (i.e, mean R^2, median R^2, the number of assays with $R^2 > 0.3$).

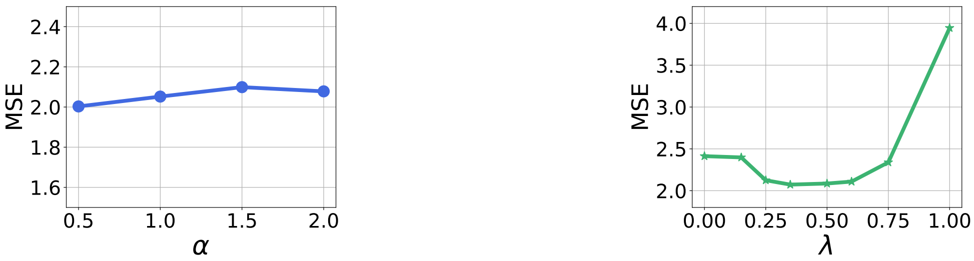
F.2. Hyperparameter Analysis on Pose Prediction

F.2.1. Analysis of the Candidate Layer Set $\mathcal{C}$

Table 5 reports the performance w.r.t. the change of the mixup layer set $\mathcal{C}$. Though including the input layer, i.e., layer 0, slightly hurts the performance in some cases, MetaMix still achieves relatively stable improvements with different mixup layer sets $\mathcal{C}$. Besides, mixing more layers in general enjoys better performance.

F.2.2. Analysis of the Mixup Ratio

In pose prediction, we analyze the effect of the mixup ratio by investigating the performance w.r.t. the changes of two key parameters: (1) α in $\mathrm{Beta}(\alpha, \alpha)$; (2) the mixup ratio λ in $\mathbf{X}_{i,l}^{mix} = \boldsymbol{\lambda} f_{\phi_i^l}(\mathbf{X}_i^s) + (\mathbf{I} - \boldsymbol{\lambda}) f_{\phi_i^l}(\mathbf{X}_i^q)$. We show the results of α and λ in Figure 4a and Figure 4b, respectively. The stability of performance w.r.t. α and the stable region $[0.4, 0.75]$ in the analysis of λ indicates the robustness of MetaMix under different Beta distribution shapes. In addition, the preference of $\alpha = 0.5$ may result from that the regression problem is not linear and the value after mix-up should not deviate too much.



(a)

(b)

Figure 4. Performance w.r.t. (a) α in $\mathrm{Beta}(\alpha, \alpha)$, and (b): the fixed mixup ratio λ under 15-shot pose prediction.

Table 4. Performance (MSE) of pose prediction w.r.t. different mixup strategies. All data augmentation strategies are applied on MAML.

Setting	$\mathcal{D}^q$	Mixup($\mathcal{D}^s$, $\mathcal{D}^s$)	Mixup($\mathcal{D}^q$, $\mathcal{D}^q$)	$\mathcal{D}^{cob}$	**MetaMix**
10-shot	3.098 ± 0.242	4.937 ± 0.210	2.881 ± 0.194	3.112 ± 0.165	$\mathbf{2.438 \pm 0.196}$
15-shot	2.413 ± 0.177	2.701 ± 0.168	2.175 ± 0.153	2.397 ± 0.173	$\mathbf{2.003 \pm 0.147}$

Table 5. Performance (Accuracy) w.r.t. MAML-MetaMix candidate layer set $\mathcal{C}$ under Pose 15-shot setting.

$\lvert\mathcal{C}\rvert = 1$	Performance	$\lvert\mathcal{C}\rvert = 2$	Performance	$\lvert\mathcal{C}\rvert = 3$	Performance	$\lvert\mathcal{C}\rvert = 4$	Performance
(0)	2.117 ± 0.152	(0,1)	2.136 ± 0.165	(0,1,2)	2.032 ± 0.135	(0,1,2,3)	2.003 ± 0.147
(1)	2.120 ± 0.150	(0,2)	2.080 ± 0.156	(0,1,3)	2.090 ± 0.164	-	-
(2)	2.127 ± 0.165	(0,3)	2.053 ± 0.171	(0,2,3)	2.047 ± 0.149	-	-
(3)	2.093 ± 0.154	(1,2)	2.112 ± 0.173	(1,2,3)	2.109 ± 0.162	-	-
-	-	(1,3)	2.094 ± 0.153	-	-	-	-
-	-	(2,3)	2.134 ± 0.167	-	-	-	-

G. Additional Results for Image Classification

G.1. Additional Results on Multi-dataset

In Table 6, we report the results (accuracy with 95% confidence interval) on Muli-datasets.

Table 6. Accuracy with 95% confidence interval on Multi-dataset.

Setting	Model	Bird	Texture	Aircraft	Fungi
5-way 1-shot	MMAML	$40.03 \pm 1.87\%$	$25.43 \pm 1.61\%$	$29.33 \pm 1.69\%$	$31.13 \pm 1.63\%$
	HSML	$40.49 \pm 1.78\%$	$26.40 \pm 1.66\%$	$31.67 \pm 1.68\%$	$30.43 \pm 1.66\%$
	ARML	$40.83 \pm 1.81\%$	$27.03 \pm 1.63\%$	$30.17 \pm 1.67\%$	$30.66 \pm 1.61\%$
	MMAML-MMCF	$51.31 \pm 1.84\%$	$29.62 \pm 1.76\%$	$\mathbf{35.41 \pm 1.75}\%$	$37.67 \pm 1.80\%$
	HSML-MMCF	$51.78 \pm 1.89\%$	$29.51 \pm 1.80\%$	$34.97 \pm 1.74\%$	$38.20 \pm 1.84\%$
	ARML-MMCF	$\mathbf{53.17 \pm 1.86}\%$	$\mathbf{30.08 \pm 1.76}\%$	$35.04 \pm 1.78\%$	$\mathbf{38.70 \pm 1.83}\%$
5-way 5-shot	MMAML	$61.64 \pm 0.96\%$	$34.76 \pm 0.80\%$	$51.89 \pm 0.93\%$	$44.48 \pm 0.96\%$
	HSML	$61.07 \pm 1.04\%$	$35.48 \pm 0.83\%$	$48.07 \pm 0.91\%$	$43.42 \pm 0.94\%$
	ARML	$64.31 \pm 0.99\%$	$36.11 \pm 0.83\%$	$50.76 \pm 0.97\%$	$46.11 \pm 0.95\%$
	MMAML-MMCF	$72.04 \pm 0.93\%$	$40.14 \pm 0.85\%$	$64.59 \pm 0.90\%$	$51.11 \pm 1.00\%$
	HSML-MMCF	$72.53 \pm 0.92\%$	$40.39 \pm 0.83\%$	$64.31 \pm 0.92\%$	$51.04 \pm 1.04\%$
	ARML-MMCF	$\mathbf{73.30 \pm 0.90}\%$	$\mathbf{40.88 \pm 0.83}\%$	$\mathbf{65.18 \pm 0.89}\%$	$\mathbf{51.56 \pm 1.03}\%$

G.2. Results under Mutually-exclusive Setting

In Table 7, we report the results under the standard mutually-exclusive setting on MiniImagenet. Under the mutually-exclusive setting, the mechanism of label shuffling is introduced to construct meta-training tasks, which significantly alleviates the meta-overfitting issue. However, applying the proposed MetaMix and Channel Shuffle on this setting still achieves comparable and even better performance than original MAML, which further demonstrates the effectiveness of our data augmentation strategies to improve meta-generalization.

G.3. Hyperparameter Analysis

G.3.1. Analysis of the Candidate Layer Set $\mathcal{C}$

In Table 8, we analyze the effect of the candidate layer set $\mathcal{C}$ and report the performance of MAML-MMCF under the 5-shot MiniImagenet scenario. Similar to the findings in drug activity prediction, in all scenarios, incorporating MMCF into MAML improves the performance, indicating the robustness of MMCF with different candidate layer sets. Besides, we observe that involving all layers achieves the best performance.

Table 7. Performance (Accuracy) of MiniImagenet under the mutually-exclusive setting.

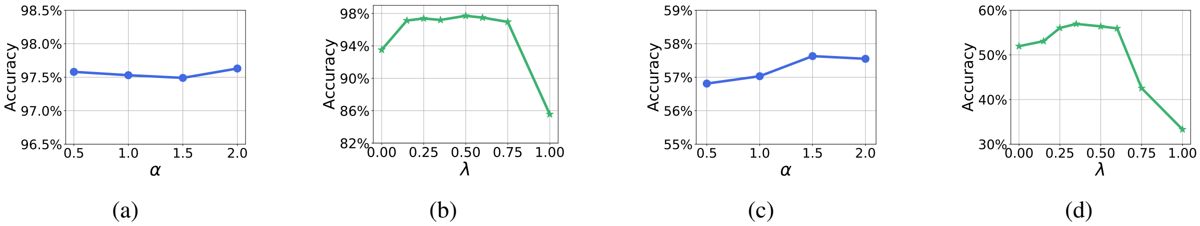
Model	MiniImagenet	
	5-way 1-shot	5-way 5-shot
MAML	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$
MAML-Channel Shuffle	$50.08 \pm 1.86\%$	$64.70 \pm 0.95\%$
MAML-MetaMix	$50.02 \pm 1.83\%$	$64.13 \pm 0.95\%$
MAML-MMCF	$\mathbf{50.35 \pm 1.82\%}$	$\mathbf{64.91 \pm 0.96\%}$

Table 8. Performance (Accuracy) w.r.t. the selected layer set $\mathcal{C}$ under the MiniImagenet 5-shot scenario.

$\|\mathcal{C}\| = 1$	Performance	$\|\mathcal{C}\| = 2$	Performance	$\|\mathcal{C}\| = 3$	Performance
(1)	$57.74 \pm 0.95\%$	(1,2)	$57.88 \pm 0.94\%$	(1,2,3)	$58.96 \pm 0.95\%$
(2)	$57.31 \pm 0.96\%$	(1,3)	$56.91 \pm 0.97\%$	-	-
(3)	$57.19 \pm 0.98\%$	(2,3)	$58.19 \pm 0.93\%$	-	-

G.3.2. Analysis of the Mixup Ratio and the Skewed Beta Distribution

Under the Omniglot and MiniImagenet 5-shot setting, we further investigate the effect of key hyperparameters for MetaMix (i.e., α and λ). The results of MAML-MetaMix on MiniImagenet and Omniglot are shown in Figure 5. Similar to the previous analyses on drug activity prediction and pose prediction, the stability of performance w.r.t. α and the stable region in λ analysis demonstrates the robustness of MetaMix. The conclusion is further supported by the analysis of skewed Beta distribution (i.e., $\alpha \neq \beta$), whose results under the MiniImagenet 5-shot setting are reported in Table 9.



(a) (b) (c) (d)

Figure 5. Performance w.r.t. (a)(c) α in $\mathrm{Beta}(\alpha, \alpha)$ distribution; (b)(d) mixup ratio λ. (a)(b) show the results under the Omniglot 20-way, 5-shot setting; (c)(d) illustrate the performance under the MiniImagenet 5-way, 5-shot scenario.

Table 9. Effect of skewed Beta distribution (i.e., $\lambda \sim \mathrm{Beta}(\alpha, \beta)$ and $\alpha \neq \beta$) under the MiniImagenet 5-shot setting.

Settings	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 2.0$	no MetaMix
$\beta = 0.5$	$55.35 \pm 0.96\%$	$53.82 \pm 0.99\%$	$53.05 \pm 0.93\%$	
$\beta = 1.0$	$53.38 \pm 0.94\%$	$56.12 \pm 1.02\%$	$54.91 \pm 1.01\%$	$51.95 \pm 0.97\%$
$\beta = 2.0$	$50.01 \pm 0.96\%$	$53.69 \pm 0.96\%$	$57.55 \pm 0.97\%$	

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pp. 1126–1135, 2017.

Martin, E. J., Polyakov, V. R., Zhu, X.-W., Tian, L., Mukherjee, P., and Liu, X. All-assay-max2 pqsar: Activity predictions as accurate as four-concentration ic50s for 8558 novartis assays. *Journal of chemical information and modeling*, 59(10): 4450–4459, 2019.

Rajendran, J., Irpan, A., and Jang, E. Meta-learning requires meta-augmentation. *NeurIPS*, 2020.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *NIPS*, pp. 4077–4087, 2017.

Yin, M., Tucker, G., Zhou, M., Levine, S., and Finn, C. Meta-learning without memorization. *ICLR*, 2020.

Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., and Zou, J. How does mixup help with robustness and generalization? In *ICLR*, 2021.