Few-shot Conformal Prediction with Auxiliary Tasks

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay

Few-shot learning with confidence

- Few-shot tasks are learning problems with severely limited training data.
- Making accurate predictions is challenging (or impossible).
- Predictions with well-calibrated probabilities are thus critical for many domains.

Our goal: quantify the uncertainty in few-shot predictions.

Few-shot learning with confidence

- Few-shot tasks are learning problems with severely limited training data.
- Making accurate predictions is challenging (or impossible).
- Predictions with well-calibrated probabilities are thus critical for many domains.

Our goal: quantify the uncertainty in few-shot predictions.

Distribution-free

Model-agnostic

Any sample size

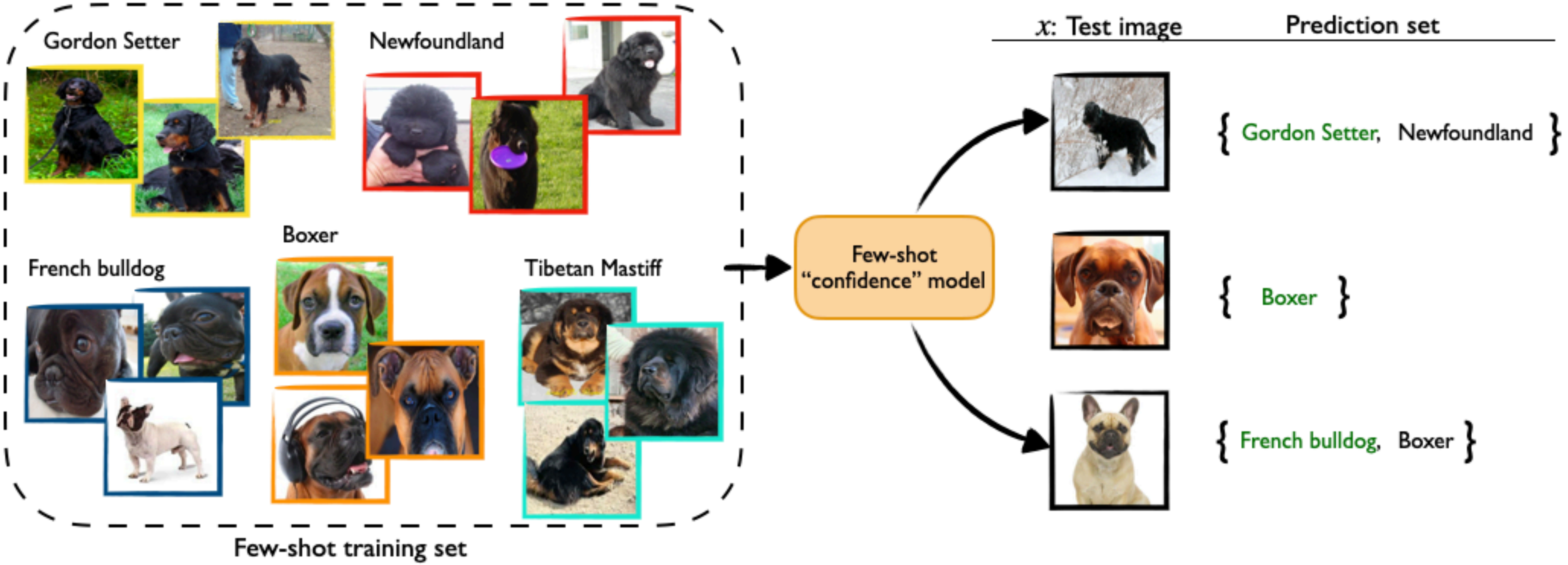
Calibrated *set*-valued predictions

- Ensuring calibrated probabilities for each possible outcome is hard.
- It can be more feasible and ultimately as useful to instead output a **small set of plausible answers**—one of which is likely to be correct.

  e.g., a *confidence* interval.

- Formally, we seek a *prediction set* $C(X)$ such that $\mathbb{P}(Y \in C(X)) \geq 1 - \epsilon$, where the user is able to specify ϵ (i.e., conformal inference).

An example (*mini*ImageNet)



Conformal prediction framework

- Given n exchangeable examples $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ and a desired significance level ϵ, for a new input X_{n+1}, return a **set of predictions** $C_n(X_{n+1}) \subseteq \mathcal{Y}$.

- A predictor is **valid** if $C_\epsilon(X_{n+1})$ covers the correct label Y_{n+1} w.p. at least $1 - \epsilon$:

$$\mathbb{P}\left(Y_{n+1} \in C_\epsilon(X_{n+1})\right) \geq 1 - \epsilon$$

- An **efficient** predictor should satisfy:

$$\mathbb{E}\left[\,|C_\epsilon(X_{n+1})|\,\right] \ll |\mathcal{Y}|$$

Nonconformity measures

- Conformal prediction uses "nonconformity" scores to measure surprise.
- Basic idea: if I assign a possible label to a given input, how strange does it look relative to other examples from my dataset that I know are correct?
- If it is relatively strange, it is considered to be *nonconforming* to the dataset.

(to be defined)

Can be any f: *known pairs* $\times$ *new pair* $\rightarrow \mathbb{R}$

Constructing conformal sets

- For each candidate label y, we compute a **nonconformity score** to quantify how "surprising" the pairing $(X_{n+1} = x_{n+1}, Y_{n+1} = y)$ would be.

- For each candidate $y \in \mathcal{Y}$, we accept or reject it based on its nonconformity score, $V_{n+1}^{(x,y)}$, compared to the $1 - \epsilon$ **quantile** of exchangeable calibration scores, $V_{1:n}^{(x,y)}$:

$$C_\epsilon(x) := \left\{ y \in \mathcal{Y} : V_{n+1}^{(x,y)} \leq \text{Quantile}(1 - \epsilon; V_{1:n}^{(x,y)} \cup \{\infty\} \right\}$$

- Thm (Vovk et. al.): the true Y_{n+1} is covered at least $(1 - \epsilon)$-fraction of the time.

Challenges of few-shot conformal prediction

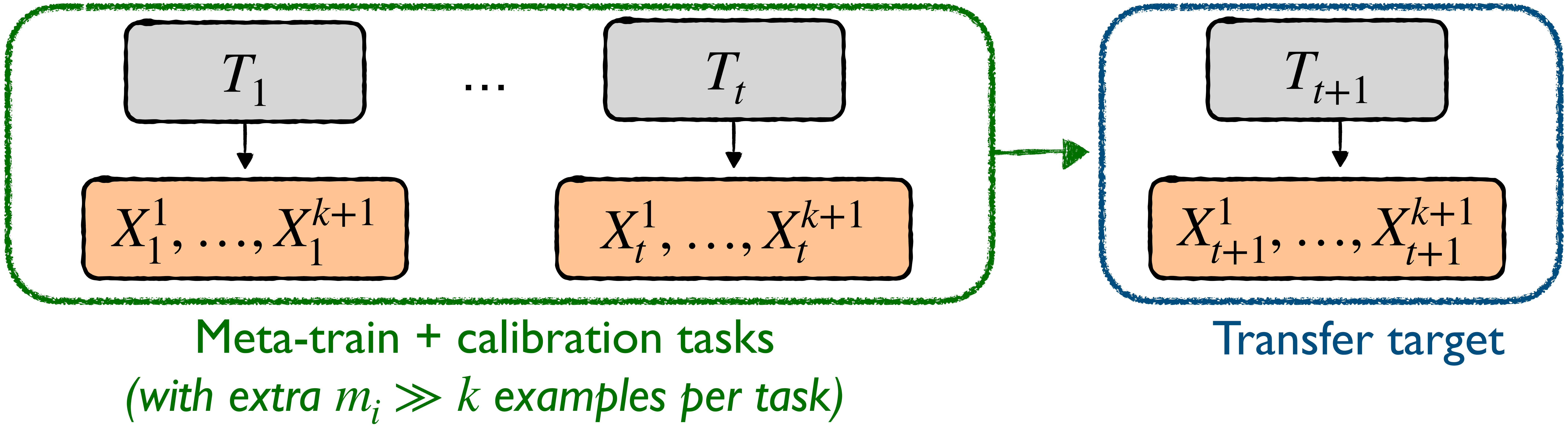
- Good nonconformity models are hard to train with few examples.
- Empirical quantiles with few points can be conservative (large step sizes).
- Leads to **uninformative** prediction sets with **poor statistical efficiency**.

Appealing to auxiliary tasks

- A popular approach to few-shot learning is **meta-learning** using auxiliary tasks.
- By being exposed to a set of similar tasks, a model can **learn to learn quickly** on a target task with much less in-domain data.
- We cast conformal prediction as a meta-learning paradigm over **exchangeable collections of tasks** to obtain *tight* prediction sets with *few* examples.

Meta-learning: two levels of exchangeability

- Assume that we do not have that much data for a target task $t + 1$ (k examples).
- But, we have data for t auxiliary tasks (other classes, regression targets…).
- Assume that *tasks* are exchangeable (i.e., $\mathbb{P}(T_1, \ldots, T_{t+1}) = \mathbb{P}(T_{\sigma(1)}, \ldots, T_{\sigma(t+1)})$).
- Assume that in-task *examples* are exchangeable (i.e., $\mathbb{P}(X_i^1, \ldots, X_i^{k+1}) = \mathbb{P}(X_i^{\sigma(1)}, \ldots, X_i^{\sigma(k+1)})$).



Conformal prediction over exchangeable tasks

- Let task T_{t+1} be the target task with a desired prediction on $X_{t+1}^{\text{test}} := X_{t+1}^{k+1}$.
- A relaxed view of validity: conformal predictor $\mathscr{M}_{\epsilon}\left(X_{t+1}^{\text{test}}\right)$ is valid *across tasks* if

$$\mathbb{P}\left(Y_{t+1}^{\text{test}} \in \mathscr{M}_{\epsilon}(X_{t+1}^{\text{test}})\right) \geq 1 - \epsilon\,.$$

Randomness is over task and task examples.

This work: create a conformal predictor that is valid (on average) on task T_{t+1}.

Few-shot meta conformal prediction

- **Step 1:** **meta-learn and meta-calibrate** a meta nonconformity measure and meta quantile predictor over a set of auxiliary tasks.

- **Step 2:** **adapt** the meta nonconformity measure to the new task using the few-shot in-domain data and meta-learning algorithm.

- **Step 3:** **predict** the $1 - \epsilon$ quantile of the new task's meta nonconformity scores using the meta quantile predictor, given the few-shot in-domain data.

- **Step 4:** keep all labels $y \in \mathcal{Y}$ whose meta nonconformity scores for input $x \in \mathcal{X}$ are below the predicted (and adjusted) quantile, $\hat{Q}_{t+1} + \Lambda(1 - \epsilon, I_{\text{cal}})$.

Few-shot meta conformal prediction

- **Step 1:** **meta-learn and meta-calibrate** a meta nonconformity measure and meta quantile predictor over a set of auxiliary tasks.

- **Step 2:** **adapt** the meta nonconformity measure to the new task using the few-shot in-domain data and meta-learning algorithm.

- **Step 3:** **predict** the $1-\epsilon$ quantile of the new task's meta nonconformity scores using the meta quantile predictor, given the few-shot in-domain data.

- **Step 4:** keep all labels $y \in \mathcal{Y}$ whose meta nonconformity scores for input $x \in \mathcal{X}$ are below the predicted (and adjusted) quantile, $\hat{Q}_{t+1} + \Lambda(1-\epsilon, I_{\text{cal}})$.

Few-shot meta conformal prediction

- **Step 1:** **meta-learn and meta-calibrate** a meta nonconformity measure and meta quantile predictor over a set of auxiliary tasks.

- **Step 2:** **adapt** the meta nonconformity measure to the new task using the few-shot in-domain data and meta-learning algorithm.

- **Step 3:** **predict** the $1 - \epsilon$ quantile of the new task's meta nonconformity scores using the meta quantile predictor, given the few-shot in-domain data.

- **Step 4:** keep all labels $y \in \mathcal{Y}$ whose meta nonconformity scores for input $x \in \mathcal{X}$ are below the predicted (and adjusted) quantile, $\hat{Q}_{t+1} + \Lambda(1 - \epsilon, I_{\text{cal}})$.

Few-shot meta conformal prediction

- **Step 1:** **meta-learn and meta-calibrate** a meta nonconformity measure and meta quantile predictor over a set of auxiliary tasks.

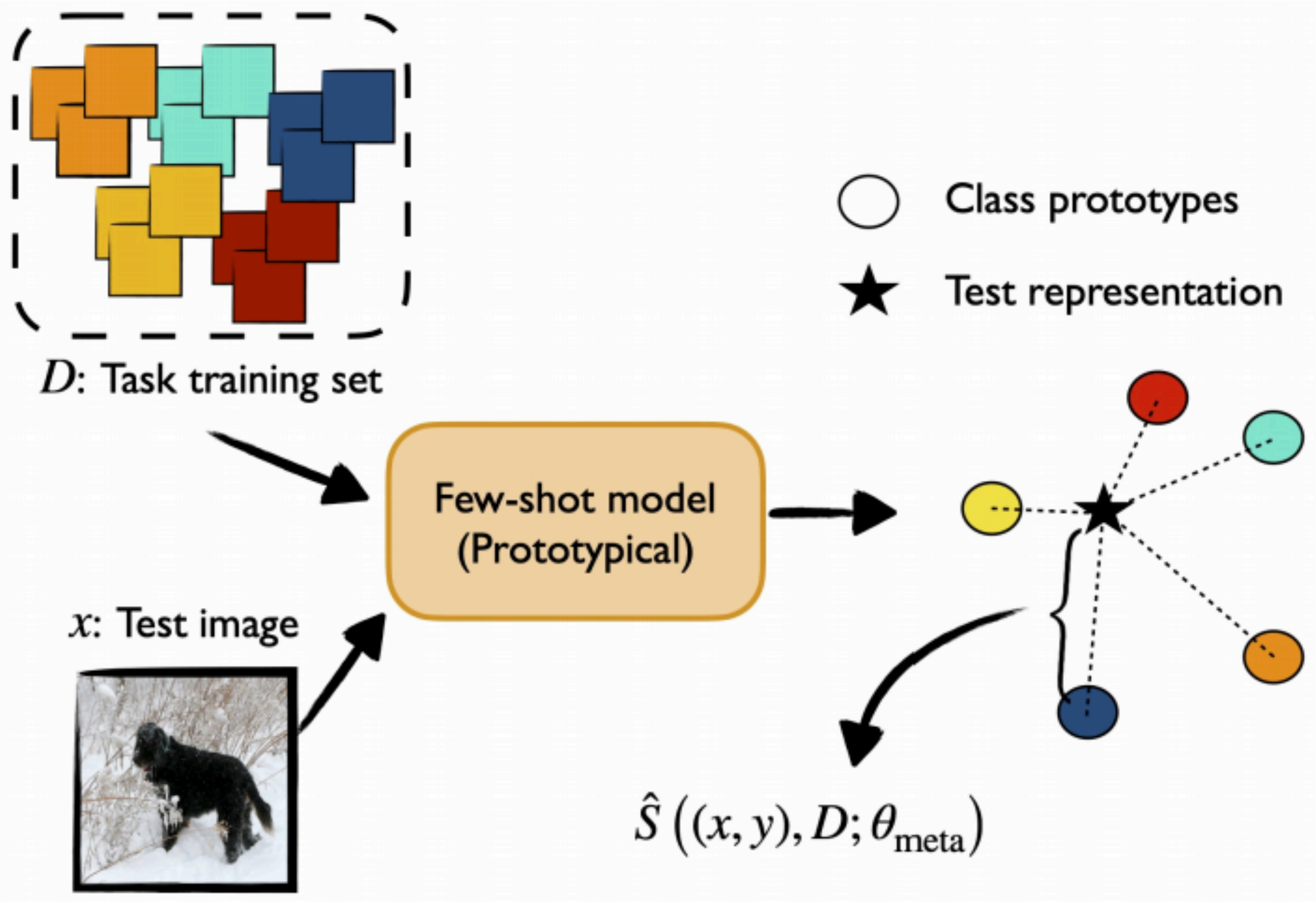
- **Step 2:** **adapt** the meta nonconformity measure to the new task using the few-shot in-domain data and meta-learning algorithm.

- **Step 3:** **predict** the $1-\epsilon$ quantile of the new task's meta nonconformity scores using the meta quantile predictor, given the few-shot in-domain data.

- **Step 4:** keep all labels $y \in \mathcal{Y}$ whose meta nonconformity scores for input $x \in \mathcal{X}$ are below the predicted (and adjusted) quantile, $\hat{Q}_{t+1} + \Lambda(1-\epsilon, I_{\text{cal}})$.

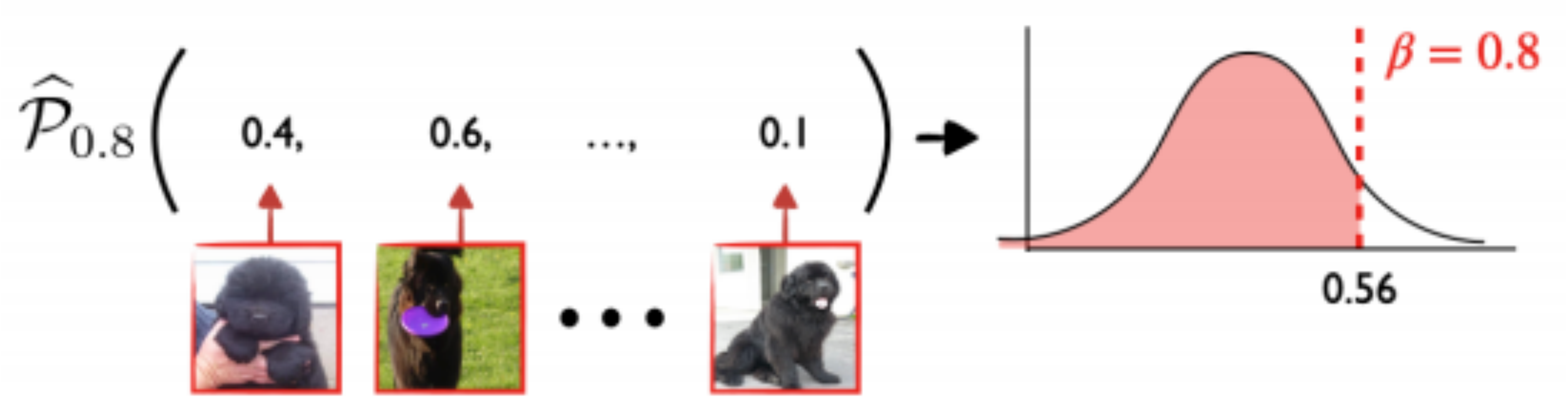
Meta-learning a nonconformity measure

- Generalizes to **any meta-learning framework** (MAML, R2D2, …).
- A set of meta parameters, θ_{meta}, are learned over auxiliary training tasks $\mathcal{I}_{\text{train}}$. θ_{meta} can be fixed or adapted symmetrically, as long as it preserves exchangeability.



Meta-learning a quantile predictor

- We want to know the $1 - \epsilon$ quantile of the new task's nonconformity scores, but we don't have enough data to directly estimate it empirically.
- Auxiliary tasks can help us learn a prior and model to **predict it directly**.
- Wrong? No problem! We **calibrate** the predictor to account for error margins.



Meta calibration (sketch)

- Let F_i be the true distribution function of task T_i's nonconformity scores. Assume F_i is known for calibration tasks I_{cal} only (we relax this to work with $\hat{F}_{m_i}$).
- A valid β-quantile prediction, $\hat{Q}_i$, should satisfy $F_i(\hat{Q}_i) \geq \beta$.
- We account for any error in the predicted quantile via a **calibration term**:

$$\Lambda(\beta, I_{\text{cal}}) = \inf \left\{ \lambda : \frac{1}{|I_{\text{cal}}| + 1} \sum_{i \in I_{\text{cal}}} F_i(\hat{Q} + \lambda) \geq \beta \right\}$$

- … and use the calibrated prediction $\hat{Q}_{t+1} + \Lambda(1 - \epsilon, I_{\text{cal}})$ for the target task.

Contributions

- We prove in our paper that our algorithm provides **valid conformal predictions** (on average) across tasks.
- Given a *consistent* quantile predictor, we further prove **asymptotic conditional validity** for any particular target task, $T_{t+1} = t_{t+1}$.
- We prove additional performance bounds when some uncertainties due to calibration task data sampling need to be accounted for.
- See paper for strong empirical results on few-shot **image classification**, **natural language processing**, and **computational chemistry** tasks.

Conclusion

- Providing precise performance guarantees and confidence-aware predictions is a critical element for many real-world machine learning applications.
- Conformal prediction can afford remarkable theoretical guarantees, but suffers in practice when data is limited (as in few-shot problems).
- We provide a **novel and theoretically grounded** approach to meta-learning conformal prediction, and show **consistent improvements** across **multiple, diverse domains and applications**.

Thank you!

Checkout our other work on principled & practical DF-UQ at the poster sessions:

- "Efficient Conformal Prediction via Cascaded Inference with Expanded Admission"
  - Building $C_\epsilon(X_{n+1})$ can be slow for large label spaces $\mathcal{Y}$ using expensive nonconf. measures.
  - In open-ended problems with large output spaces, the target Y_{n+1} can be nonunique.
  - **Solution:** prediction cascades (simple→complex models) with a calibration twist.

- "Consistent Accelerated Inference via Confident Adaptive Transformers"
  - Multi-layered models are slow; predictions can often be made at intermediate layers with "early exit".
  - How to ensure that the predictions are consistent, i.e., $\mathbb{P}(f_{\text{early}}(X_{n+1}) = f_{\text{full}}(X_{n+1})) \geq 1 - \epsilon$?
  - **Solution:** use conformal inference to identify a conservative set of consistent layers + pick the first.