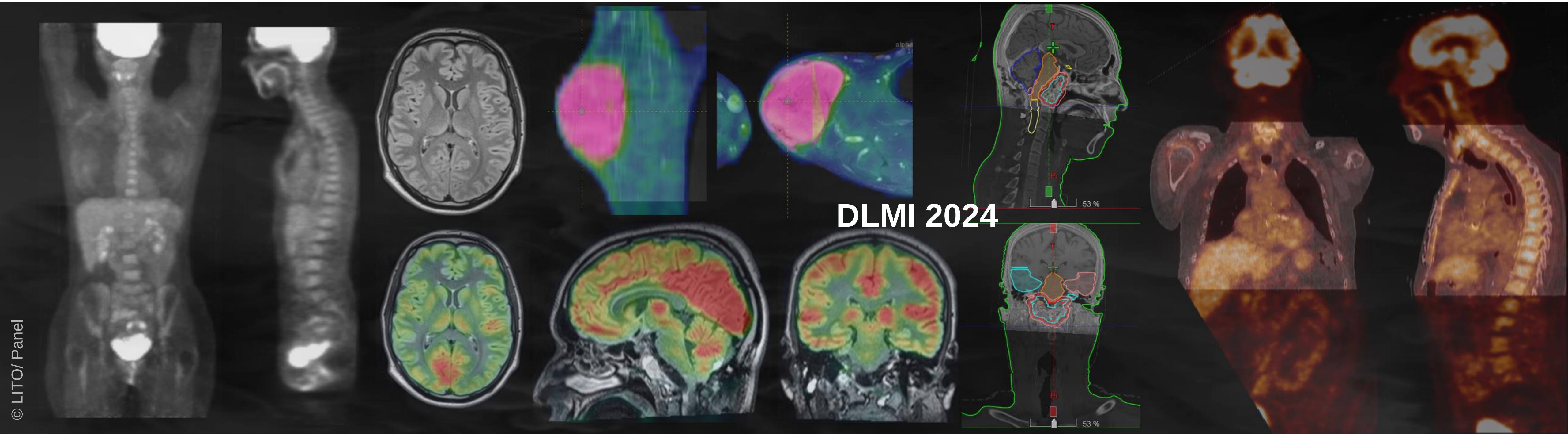
# LITO

institut Curie

Inserm

La science pour la santé
From science to health



DLMI 2024

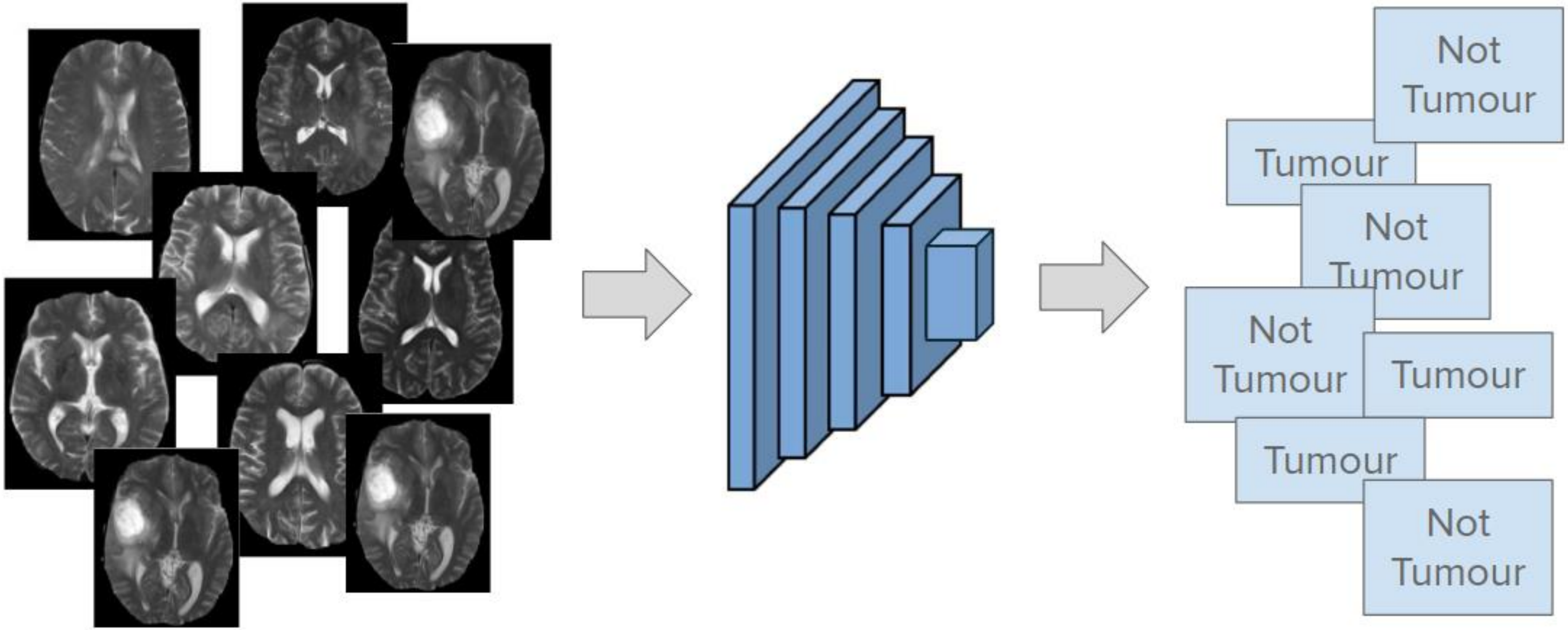
© LITO/ Panel

# DLMI 2024 – Deep Learning for Medical Imaging – Summer School

- Summer school on deep learning methods in the context of medical imaging (5th edition)
- 1 week with conferences and hands-on sessions at ETS Montreal
- Topics:
  - Basics of DL
  - CNN
  - GAN – Diffusion models
  - RNN – Transformers
  - Uncertainty Quantification
  - Medical images typical issues
  - Self-supervised – Weakly supervised learning
  - Foundation models
  - MONAI

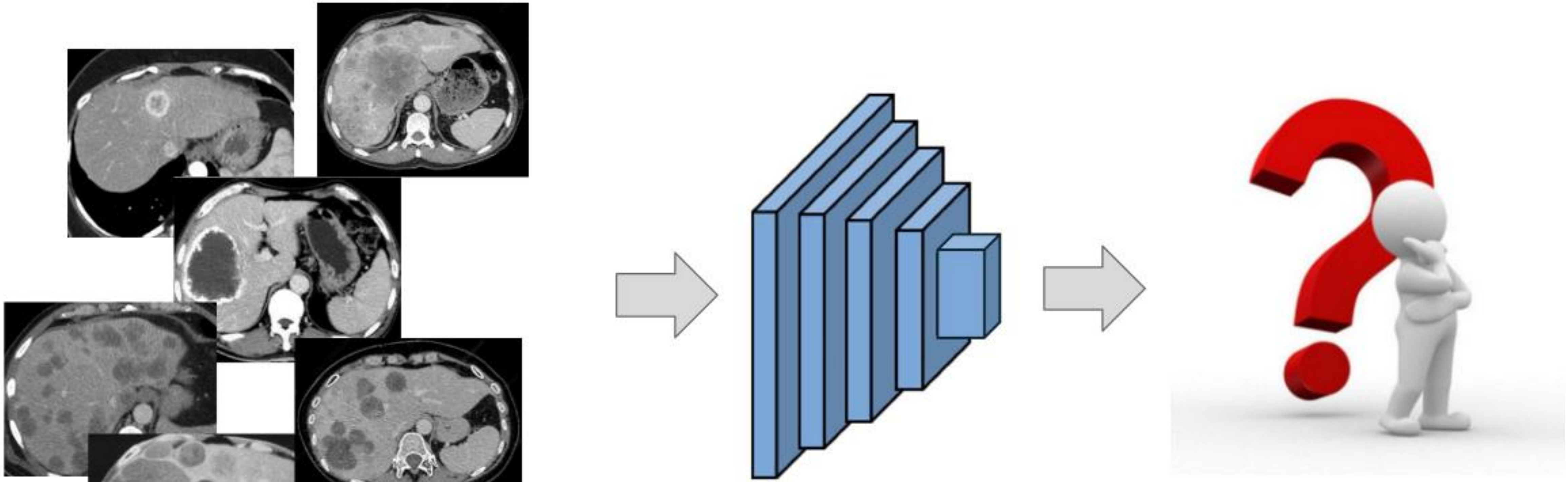
# Foundation Models

## Traditional (task-specific) learning



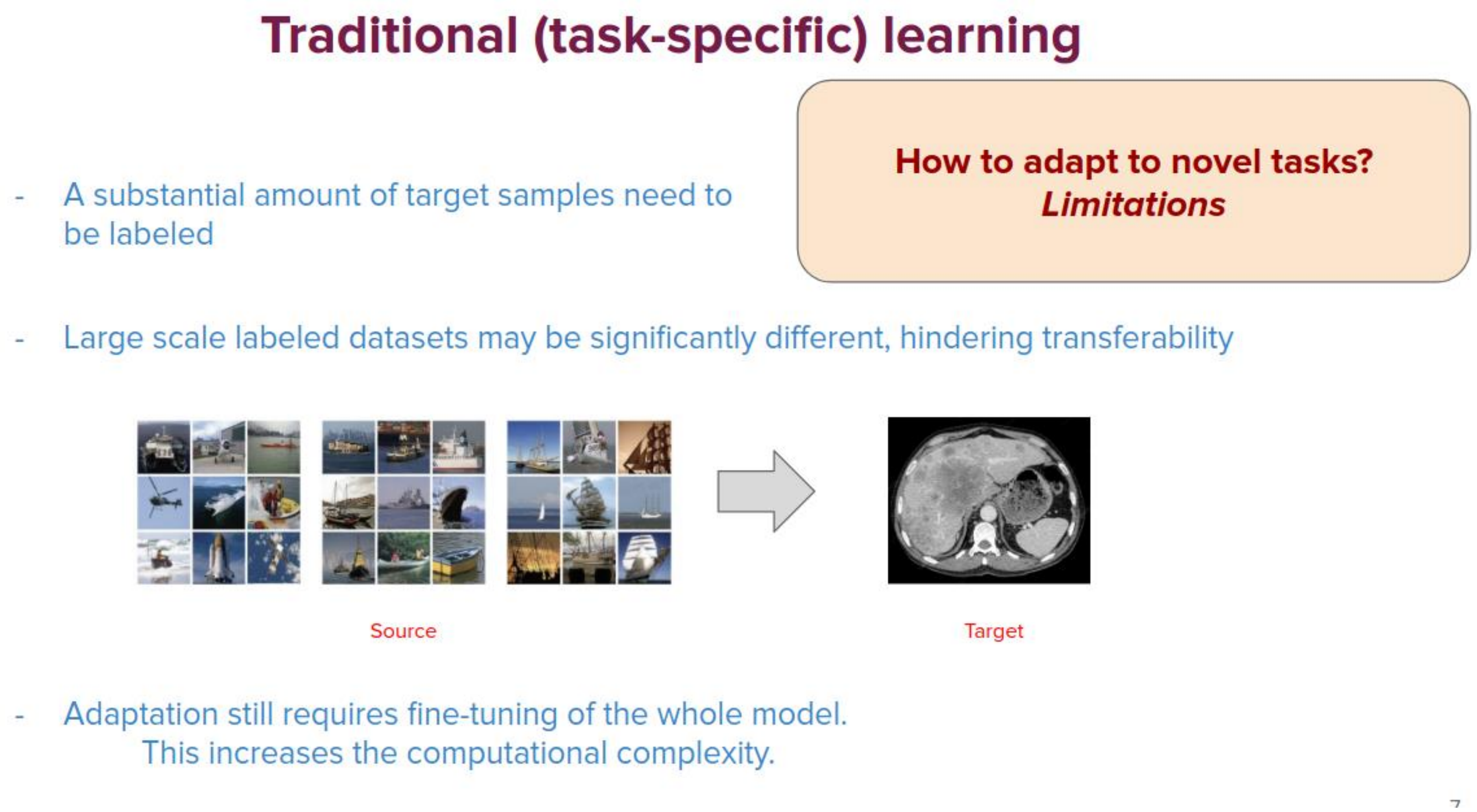
# Foundation Models

## Traditional (task-specific) learning



How to adapt to novel tasks?

# Foundation Models



## Traditional (task-specific) learning

**How to adapt to novel tasks?**
***Limitations***

- A substantial amount of target samples need to be labeled

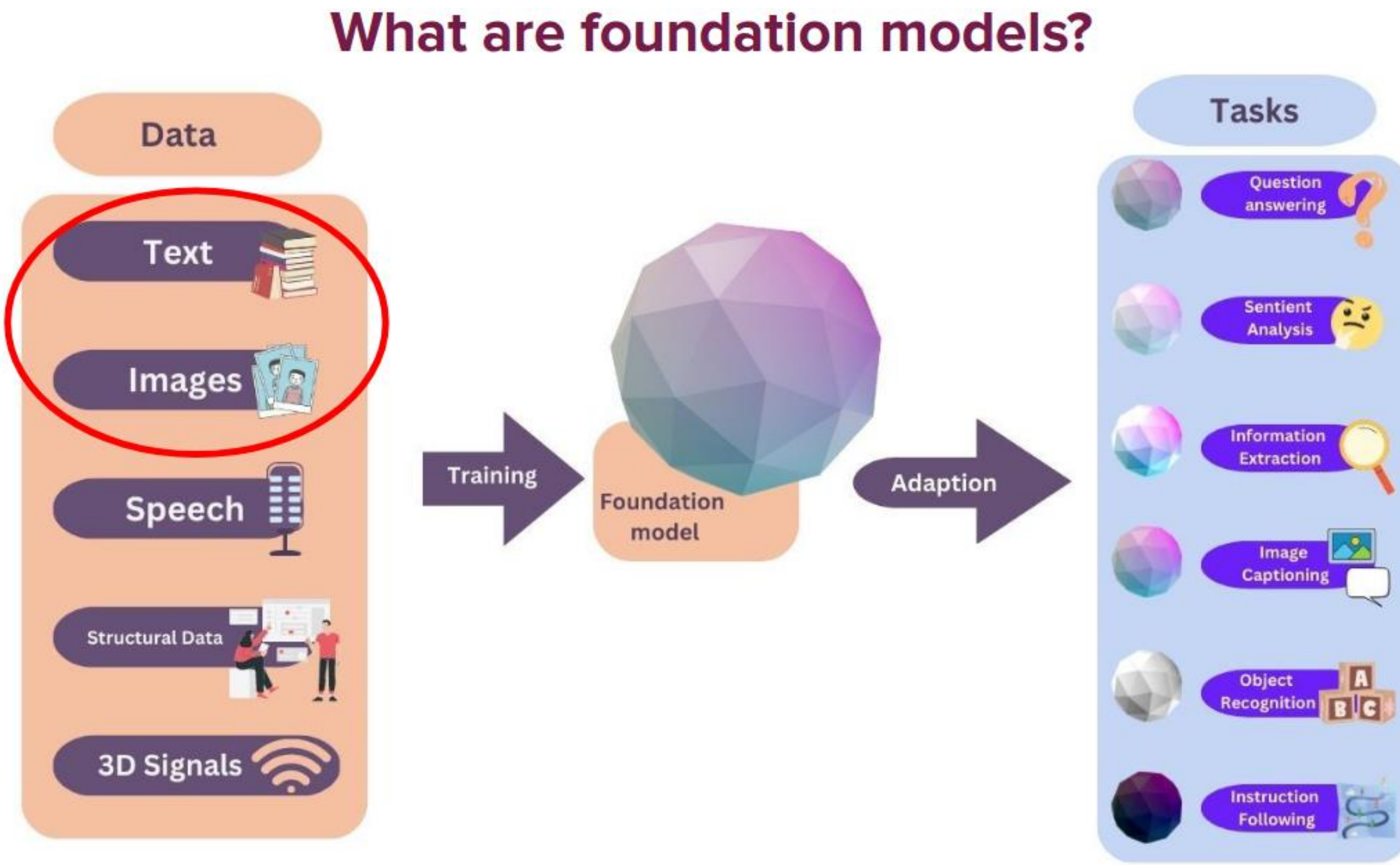
- Large scale labeled datasets may be significantly different, hindering transferability

Source
Target

- Adaptation still requires fine-tuning of the whole model. This increases the computational complexity.

7

# Foundation Models



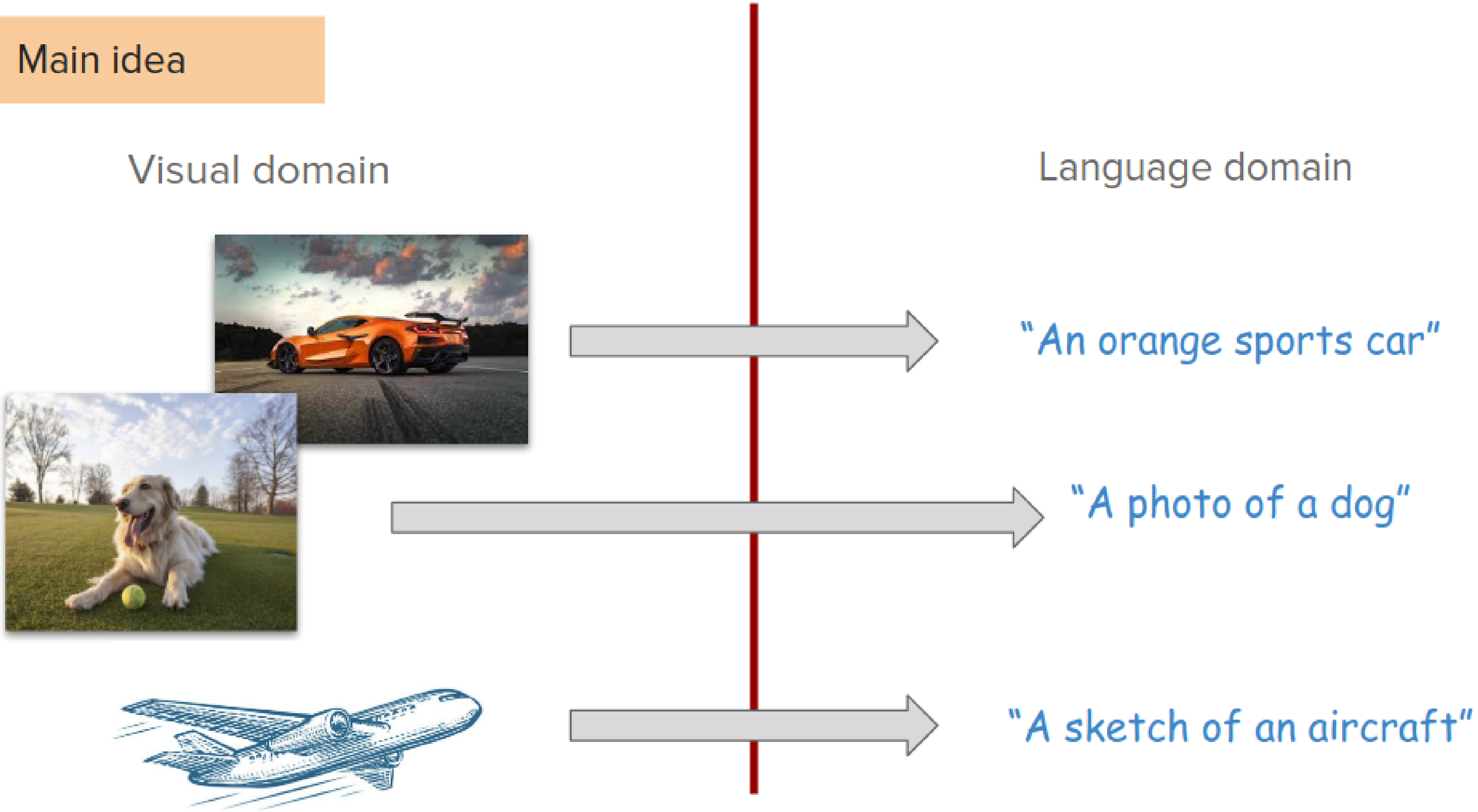
What are foundation models?

# Foundation Models

## Revisit language-vision models

(the main topic of this talk)

Main idea



Visual domain

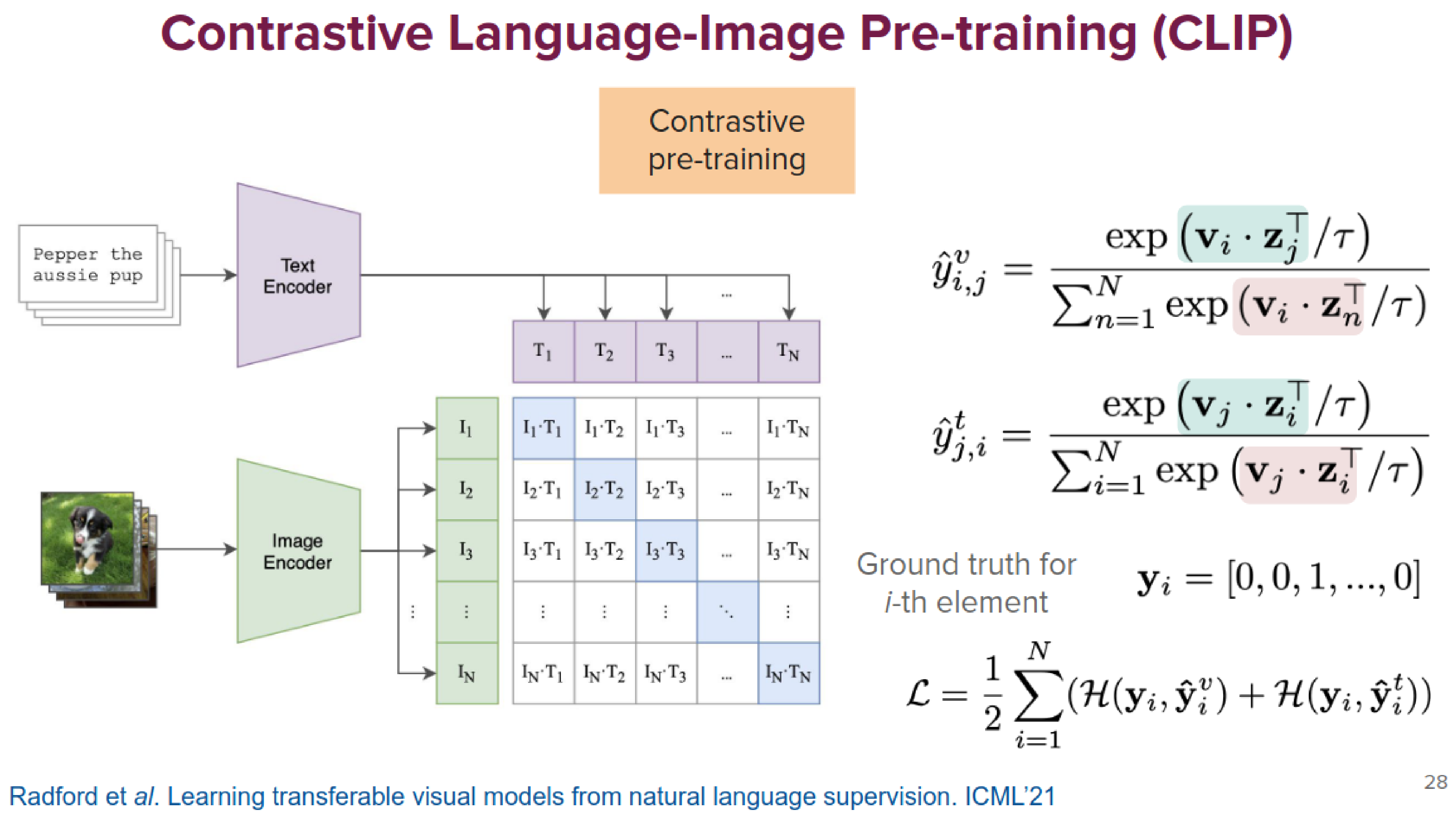
Language domain

"An orange sports car"

"A photo of a dog"

"A sketch of an aircraft"

# Foundation Models



## Contrastive Language-Image Pre-training (CLIP)

|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

$$\hat{y}^{v}_{i,j} = \frac{\exp\left(\mathbf{v}_i \cdot \mathbf{z}_j^{\top} / \tau\right)}{\sum_{n=1}^{N} \exp\left(\mathbf{v}_i \cdot \mathbf{z}_n^{\top} / \tau\right)}$$

$$\hat{y}^{t}_{j,i} = \frac{\exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^{\top} / \tau\right)}{\sum_{i=1}^{N} \exp\left(\mathbf{v}_j \cdot \mathbf{z}_i^{\top} / \tau\right)}$$

Ground truth for *i*-th element $\mathbf{y}_i = [0, 0, 1, ..., 0]$

$$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{N} \left(\mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i^{v}) + \mathcal{H}(\mathbf{y}_i, \hat{\mathbf{y}}_i^{t})\right)$$

Radford et *al*. Learning transferable visual models from natural language supervision. ICML'21

28

# Medical applications: MedCLIP (Chest XRays)

## MedCLIP: Contrastive Learning from Unpaired Medical Images and Text

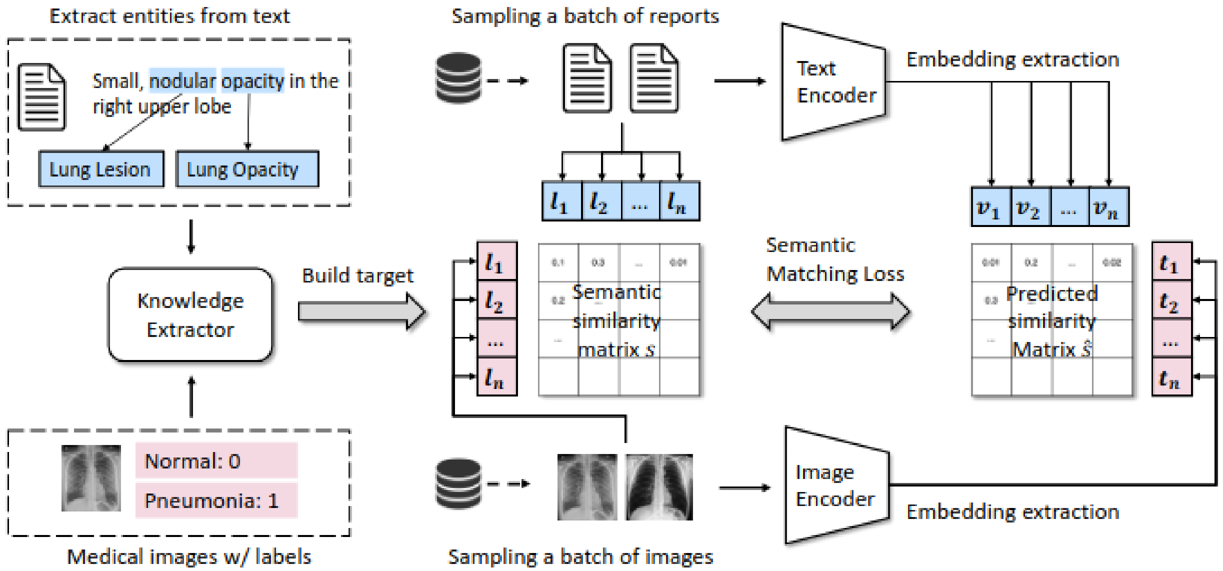
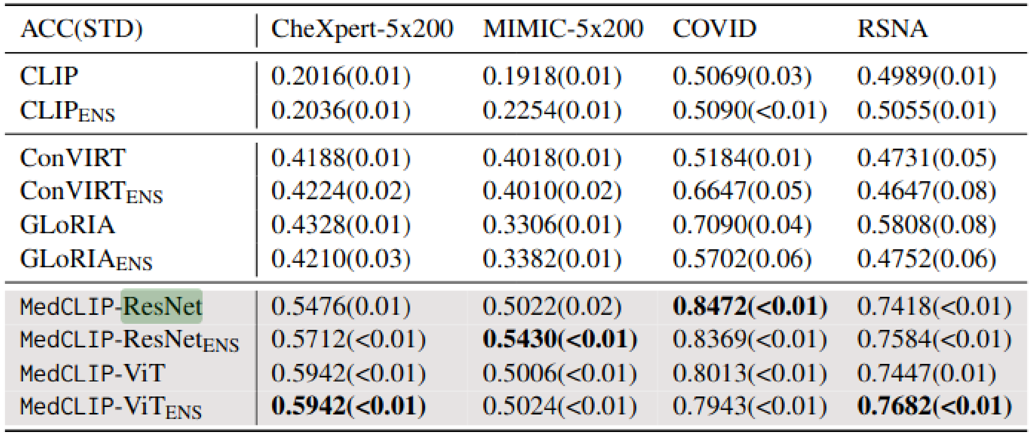


Figure 3: The workflow of MedCLIP. The knowledge extraction module extracts medical entities from raw medical reports. Then, a semantic similarity matrix is built by comparing medical entities (from text) and raw labels (from images), which enables pairing arbitrary two separately sampled images and texts. The extracted image and text embeddings are paired to match the semantic similarity matrix.

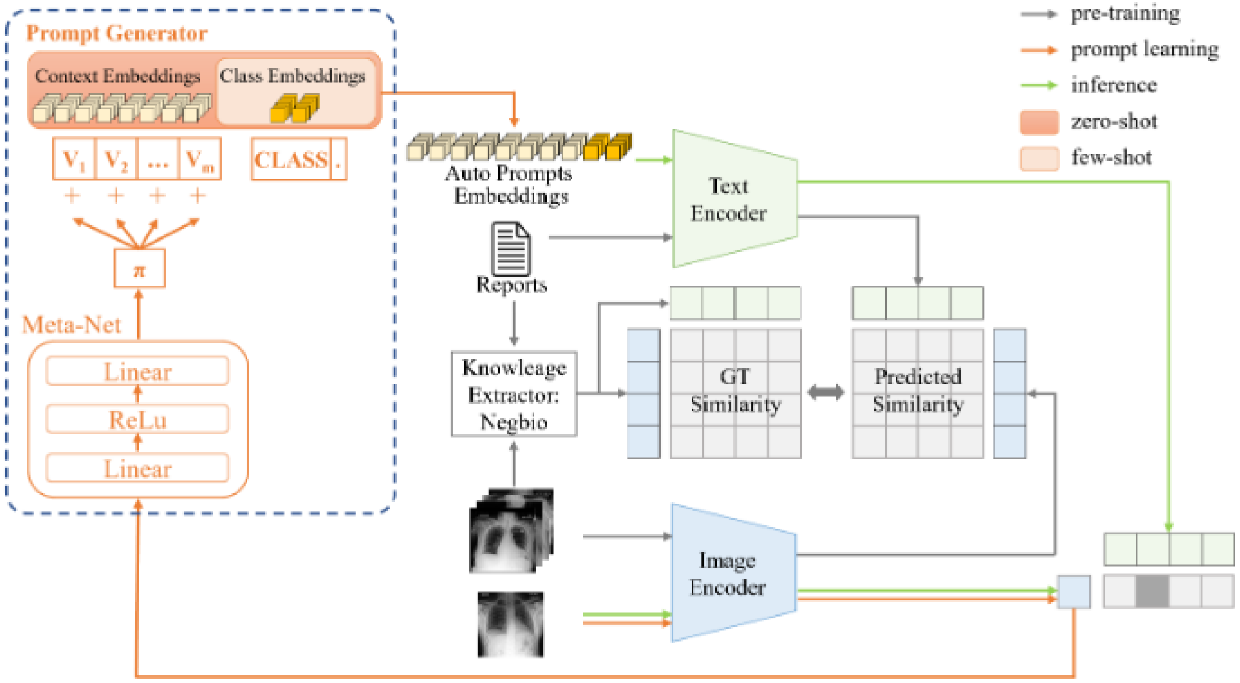


| ACC(STD) | CheXpert-5x200 | MIMIC-5x200 | COVID | RSNA |
|---|---|---|---|---|
| CLIP | 0.2016(0.01) | 0.1918(0.01) | 0.5069(0.03) | 0.4989(0.01) |
| CLIP$_{\text{ENS}}$ | 0.2036(0.01) | 0.2254(0.01) | 0.5090(<0.01) | 0.5055(0.01) |
| ConVIRT | 0.4188(0.01) | 0.4018(0.01) | 0.5184(0.01) | 0.4731(0.05) |
| ConVIRT$_{\text{ENS}}$ | 0.4224(0.02) | 0.4010(0.02) | 0.6647(0.05) | 0.4647(0.08) |
| GLoRIA | 0.4328(0.01) | 0.3306(0.01) | 0.7090(0.04) | 0.5808(0.08) |
| GLoRIA$_{\text{ENS}}$ | 0.4210(0.03) | 0.3382(0.01) | 0.5702(0.06) | 0.4752(0.06) |
| MedCLIP-ResNet | 0.5476(0.01) | 0.5022(0.02) | **0.8472(<0.01)** | 0.7418(<0.01) |
| MedCLIP-ResNet$_{\text{ENS}}$ | 0.5712(<0.01) | **0.5430(<0.01)** | 0.8369(<0.01) | 0.7584(<0.01) |
| MedCLIP-ViT | 0.5942(<0.01) | 0.5006(<0.01) | 0.8013(<0.01) | 0.7447(0.01) |
| MedCLIP-ViT$_{\text{ENS}}$ | **0.5942(<0.01)** | 0.5024(<0.01) | 0.7943(<0.01) | **0.7682(<0.01)** |

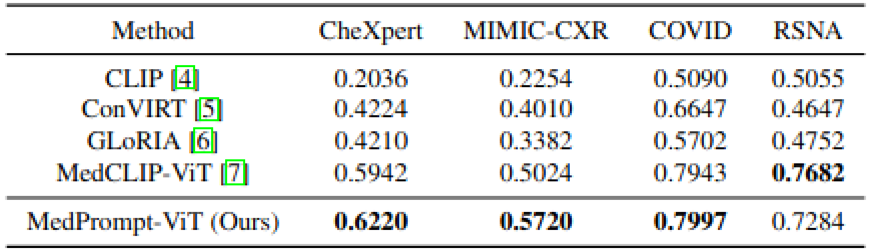
Zero-shot classification task

# Medical applications: MedPrompt (Chest XRays)

Exploring Low-Resource Medical Image Classification with Weakly Supervised Prompt Learning

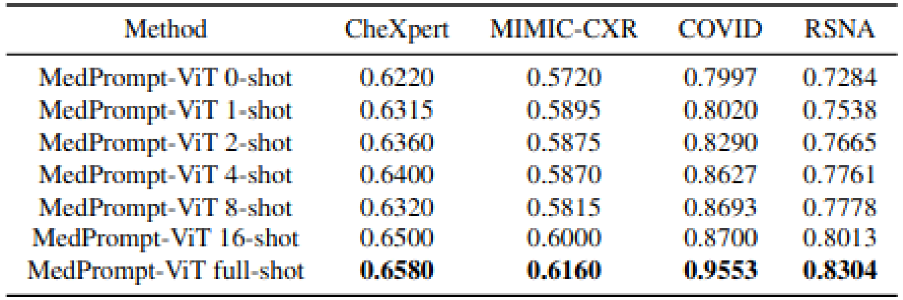


MedPrompt architecture: MedClip + Prompt learning module



| Method | CheXpert | MIMIC-CXR | COVID | RSNA |
|---|---|---|---|---|
| CLIP [4] | 0.2036 | 0.2254 | 0.5090 | 0.5055 |
| ConVIRT [5] | 0.4224 | 0.4010 | 0.6647 | 0.4647 |
| GLoRIA [6] | 0.4210 | 0.3382 | 0.5702 | 0.4752 |
| MedCLIP-ViT [7] | 0.5942 | 0.5024 | 0.7943 | **0.7682** |
| MedPrompt-ViT (Ours) | **0.6220** | **0.5720** | **0.7997** | 0.7284 |

Table 1: Performance of zero-shot image classification on four datasets. For models with manually designed prompts, we only report the results with prompt ensemble. Best performance are in bold.



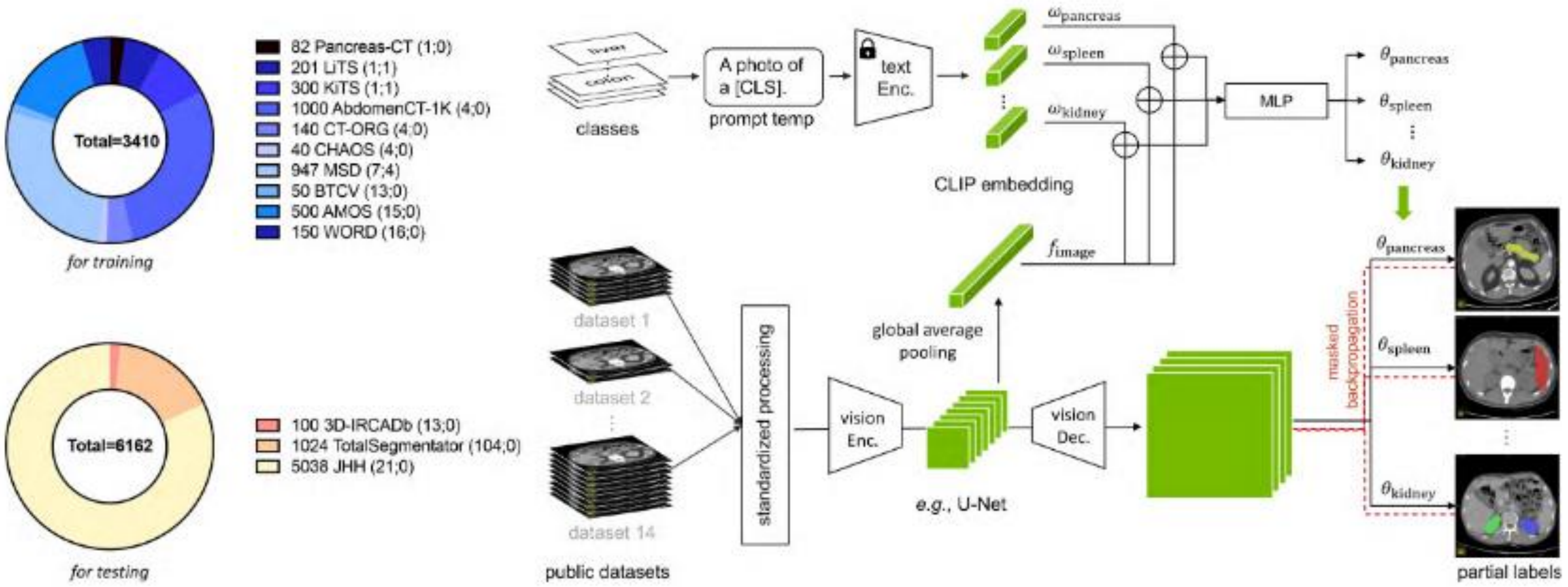
| Method | CheXpert | MIMIC-CXR | COVID | RSNA |
|---|---|---|---|---|
| MedPrompt-ViT 0-shot | 0.6220 | 0.5720 | 0.7997 | 0.7284 |
| MedPrompt-ViT 1-shot | 0.6315 | 0.5895 | 0.8020 | 0.7538 |
| MedPrompt-ViT 2-shot | 0.6360 | 0.5875 | 0.8290 | 0.7665 |
| MedPrompt-ViT 4-shot | 0.6400 | 0.5870 | 0.8627 | 0.7761 |
| MedPrompt-ViT 8-shot | 0.6320 | 0.5815 | 0.8693 | 0.7778 |
| MedPrompt-ViT 16-shot | 0.6500 | 0.6000 | 0.8700 | 0.8013 |
| MedPrompt-ViT full-shot | **0.6580** | **0.6160** | **0.9553** | **0.8304** |

Table 3: Performance of few-shot learning of our model on four datasets. Best performance are in bold.

Results for zero-shot and few-shots image classification

# Medical applications in segmentation: CLIP Driven

## CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection



# Medical applications in segmentation: CLIP Driven

Main idea

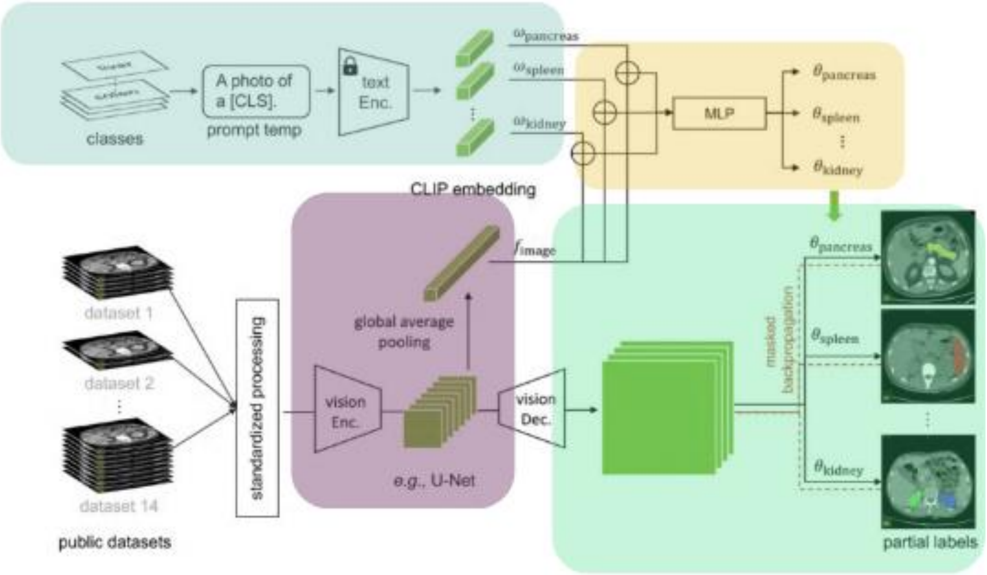
Text branch
(generates text embedding for class k) $\mathbf{w}_k$

Visual branch-encoder
(generates visual embedding for image x) $\mathbf{f}$

Text-based controller MLP
(generates class parameters)

$$\boldsymbol{\theta_k} = MLP(\mathbf{w}_k \oplus \mathbf{f})$$

$$\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$$



Visual branch-decoder
(generates visual embedding for image x)

$$\mathbf{P}_k = \text{sigmoid}(((\mathbf{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3})$$

It represents foreground class k *vs* background

Training loss

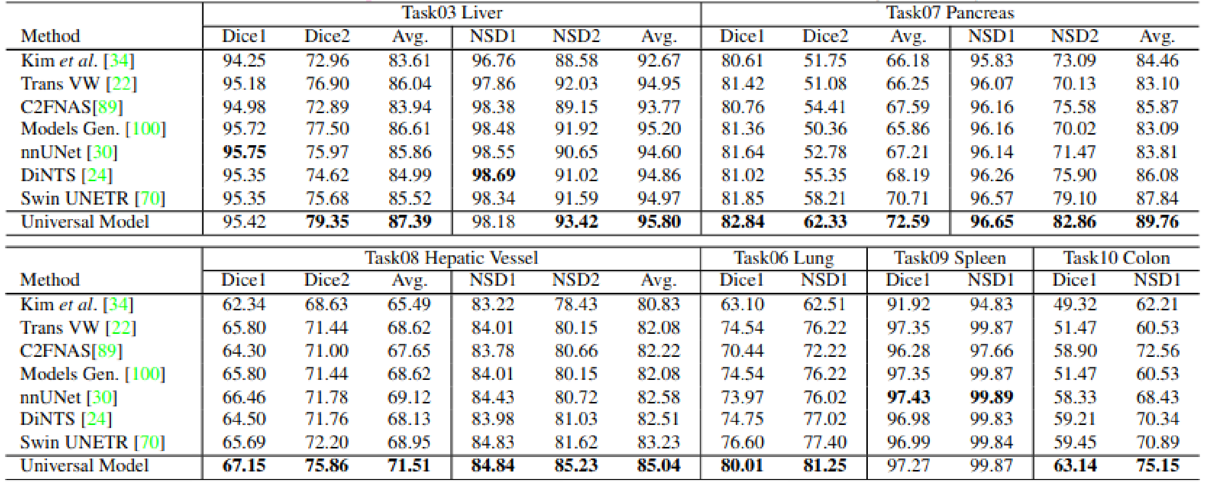
Binary cross-entropy per class (and terms masked for those classes not present)

$$\mathcal{L} = \sum_{k=1}^{K} \mathbf{1}_{\{k \in y\}} \cdot \text{BCE}_k$$

Liu *et al.* CLIP Driven universal model for organ segmentation and tumor detection. ICCV'23

# Medical applications in segmentation: CLIP Driven

Table 2. **Leaderboard performance on MSD.** The results are evaluated in the server on the MSD competition test dataset. All Dice and NSD metrics are obtained from the MSD public leaderboard. The results of MRI-related tasks were generated by Swin UNETR [70].



| | Task03 Liver | | | | | | Task07 Pancreas | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Dice1 | Dice2 | Avg. | NSD1 | NSD2 | Avg. | Dice1 | Dice2 | Avg. | NSD1 | NSD2 | Avg. |
| Kim *et al.* [34] | 94.25 | 72.96 | 83.61 | 96.76 | 88.58 | 92.67 | 80.61 | 51.75 | 66.18 | 95.83 | 73.09 | 84.46 |
| Trans VW [22] | 95.18 | 76.90 | 86.04 | 97.86 | 92.03 | 94.95 | 81.42 | 51.08 | 66.25 | 96.07 | 70.13 | 83.10 |
| C2FNAS[89] | 94.98 | 72.89 | 83.94 | 98.38 | 89.15 | 93.77 | 80.76 | 54.41 | 67.59 | 96.16 | 75.58 | 85.87 |
| Models Gen. [100] | 95.72 | 77.50 | 86.61 | 98.48 | 91.92 | 95.20 | 81.36 | 50.36 | 65.86 | 96.16 | 70.02 | 83.09 |
| nnUNet [30] | **95.75** | 75.97 | 85.86 | 98.55 | 90.65 | 94.60 | 81.64 | 52.78 | 67.21 | 96.14 | 71.47 | 83.81 |
| DiNTS [24] | 95.35 | 74.62 | 84.99 | **98.69** | 91.02 | 94.86 | 81.02 | 55.35 | 68.19 | 96.26 | 75.90 | 86.08 |
| Swin UNETR [70] | 95.35 | 75.68 | 85.52 | 98.34 | 91.59 | 94.97 | 81.85 | 58.21 | 70.71 | 96.57 | 79.10 | 87.84 |
| Universal Model | 95.42 | **79.35** | **87.39** | 98.18 | **93.42** | **95.80** | **82.84** | **62.33** | **72.59** | **96.65** | **82.86** | **89.76** |

| | Task08 Hepatic Vessel | | | | | | Task06 Lung | | Task09 Spleen | | Task10 Colon | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Dice1 | Dice2 | Avg. | NSD1 | NSD2 | Avg. | Dice1 | NSD1 | Dice1 | NSD1 | Dice1 | NSD1 |
| Kim *et al.* [34] | 62.34 | 68.63 | 65.49 | 83.22 | 78.43 | 80.83 | 63.10 | 62.51 | 91.92 | 94.83 | 49.32 | 62.21 |
| Trans VW [22] | 65.80 | 71.44 | 68.62 | 84.01 | 80.15 | 82.08 | 74.54 | 76.22 | 97.35 | 99.87 | 51.47 | 60.53 |
| C2FNAS[89] | 64.30 | 71.00 | 67.65 | 83.78 | 80.66 | 82.22 | 70.44 | 72.22 | 96.28 | 97.66 | 58.90 | 72.56 |
| Models Gen. [100] | 65.80 | 71.44 | 68.62 | 84.01 | 80.15 | 82.08 | 74.54 | 76.22 | 97.35 | 99.87 | 51.47 | 60.53 |
| nnUNet [30] | 66.46 | 71.78 | 69.12 | 84.43 | 80.72 | 82.58 | 73.97 | 76.02 | **97.43** | **99.89** | 58.33 | 68.43 |
| DiNTS [24] | 64.50 | 71.76 | 68.13 | 83.98 | 81.03 | 82.51 | 74.75 | 77.02 | 96.98 | 99.83 | 59.21 | 70.34 |
| Swin UNETR [70] | 65.69 | 72.20 | 68.95 | 84.83 | 81.62 | 83.23 | 76.60 | 77.40 | 96.99 | 99.84 | 59.45 | 70.89 |
| Universal Model | **67.15** | **75.86** | **71.51** | **84.84** | **85.23** | **85.04** | **80.01** | **81.25** | 97.27 | 99.87 | **63.14** | **75.15** |

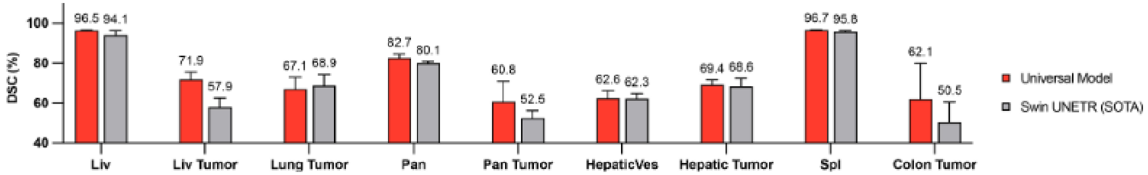
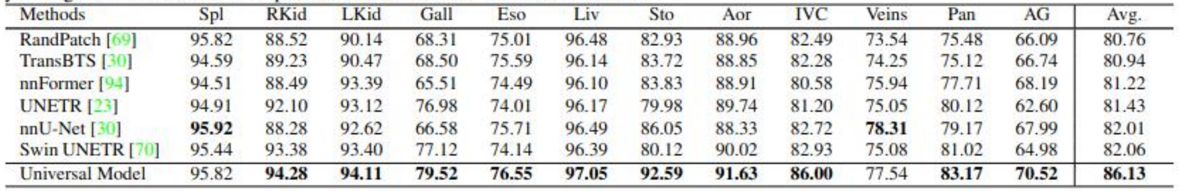


Figure 3. **Benchmark on MSD validation dataset.** We compare Universal Model with Swin UNETR [70] (previously ranked first on the MSD leaderboard) on 5-fold cross-validation of the MSD dataset. Universal Model achieves overall better segmentation performance and offers *substantial* improvement in the tasks of segmenting liver tumors (+14%), pancreatic tumors (+8%), and colon tumors (+11%).

Medical Segmentation Decathlon results comparison

Table 3. **5-fold cross-validation results on BTCV.** For a fair comparison, we did not use model ensemble during the evaluation. All experiments are under the same data splits, computing resources, and testing conditions. Universal Model achieves the overall best performance, yielding at least +3.9% DSC improvement over the state-of-the-art method.



| Methods | Spl | RKid | LKid | Gall | Eso | Liv | Sto | Aor | IVC | Veins | Pan | AG | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RandPatch [69] | 95.82 | 88.52 | 90.14 | 68.31 | 75.01 | 96.48 | 82.93 | 88.96 | 82.49 | 73.54 | 75.48 | 66.09 | 80.76 |
| TransBTS [30] | 94.59 | 89.23 | 90.47 | 68.50 | 75.59 | 96.14 | 83.72 | 88.85 | 82.28 | 74.25 | 75.12 | 66.74 | 80.94 |
| nnFormer [94] | 94.51 | 88.49 | 93.39 | 65.51 | 74.49 | 96.10 | 83.83 | 88.91 | 80.58 | 75.94 | 77.71 | 68.19 | 81.22 |
| UNETR [23] | 94.91 | 92.10 | 93.12 | 76.98 | 74.01 | 96.17 | 79.98 | 89.74 | 81.20 | 75.05 | 80.12 | 62.60 | 81.43 |
| nnU-Net [30] | **95.92** | 88.28 | 92.62 | 66.58 | 75.71 | 96.49 | 86.05 | 88.33 | 82.72 | **78.31** | 79.17 | 67.99 | 82.01 |
| Swin UNETR [70] | 95.44 | 93.38 | 93.40 | 77.12 | 74.14 | 96.39 | 80.12 | 90.02 | 82.93 | 75.08 | 81.02 | 64.98 | 82.06 |
| Universal Model | 95.82 | **94.28** | **94.11** | **79.52** | **76.55** | **97.05** | **92.59** | **91.63** | **86.00** | 77.54 | **83.17** | **70.52** | **86.13** |

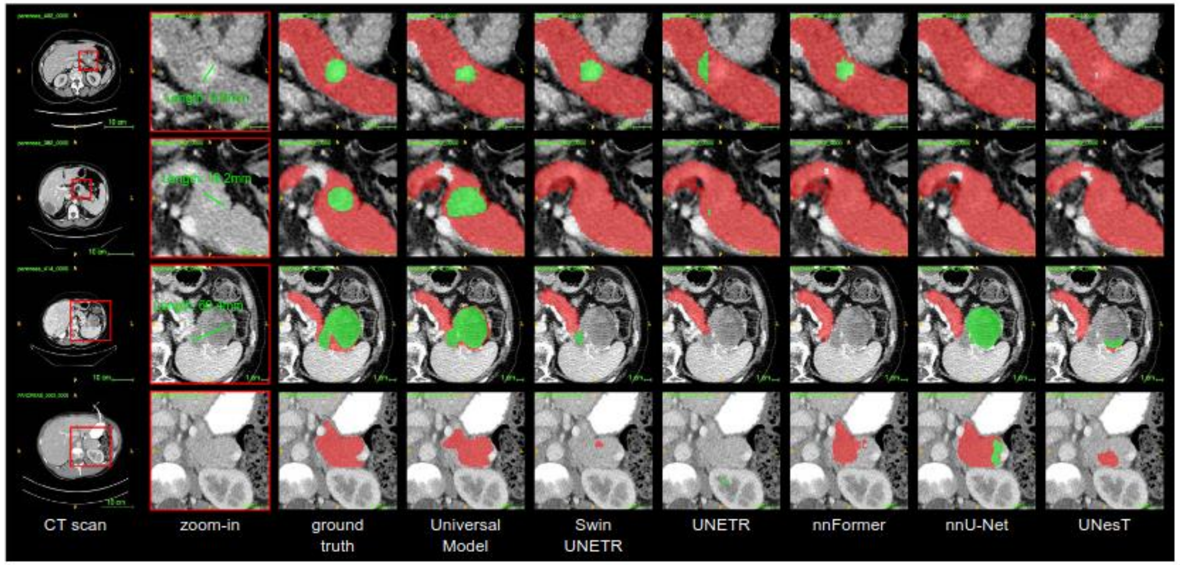


Figure 5. **Pancreatic tumor detection.** Qualitative visualizations of the proposed Universal Model and five competitive baseline methods. We review the detection results of tumors from smaller to larger sizes (Rows 1–3). When it comes to a CT scan without tumor from other hospitals, the Universal Model generalize well in organ segmentation and does not generate many false positives of tumors (Row 4; §4.2). The visualization of tumor detection in other organs (*e.g.*, liver tumors and kidney tumors) can be found in Appendix Figures 10–11.

Beyond the Cranial Vault (BTCV) Segmentation Challenge results comparison

LITO

**Paul Steinmetz - 02/10/2024**

**Développement de méthodes pour la création de modèles d’IA performants et robustes en imagerie médicale**

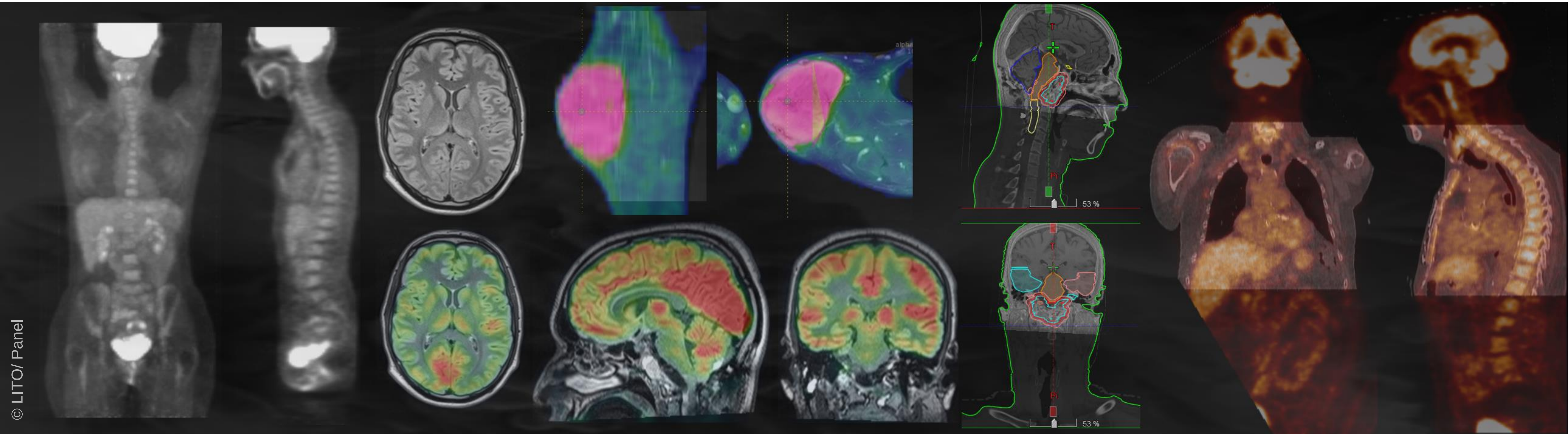
**Thèse débutée le 08/01/2024 – dans le cadre du programme AIDReAM**

INVESTIR L'AVENIR

institut Curie

Inserm

La science pour la santé
From science to health



© LITO/ Panel

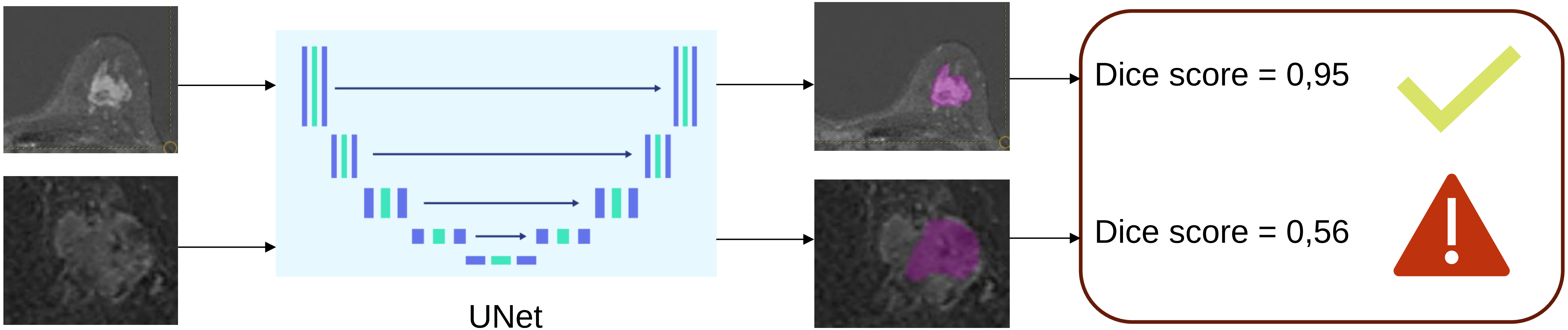
# Plan

- Context
- Use case: BI-RADS classification:
  - Data pre-processing
  - Training strategy and model
  - Training results
  - External evaluation results
- Uncertainty Quantification:
  - Results
  - Next steps

# Context

- Lack of robustness and adaptability of models to domain shift + inability to identify failure cases → main bottleneck for adoption in clinical practice

- Example: segmentation of breast MRI tumors for prognostic evaluation. High performance models but in case of failure → incorrect prognosis, may lead to inappropriate patient management



UNet

Raw images

Segmentations

Dice score = 0,95 ✓

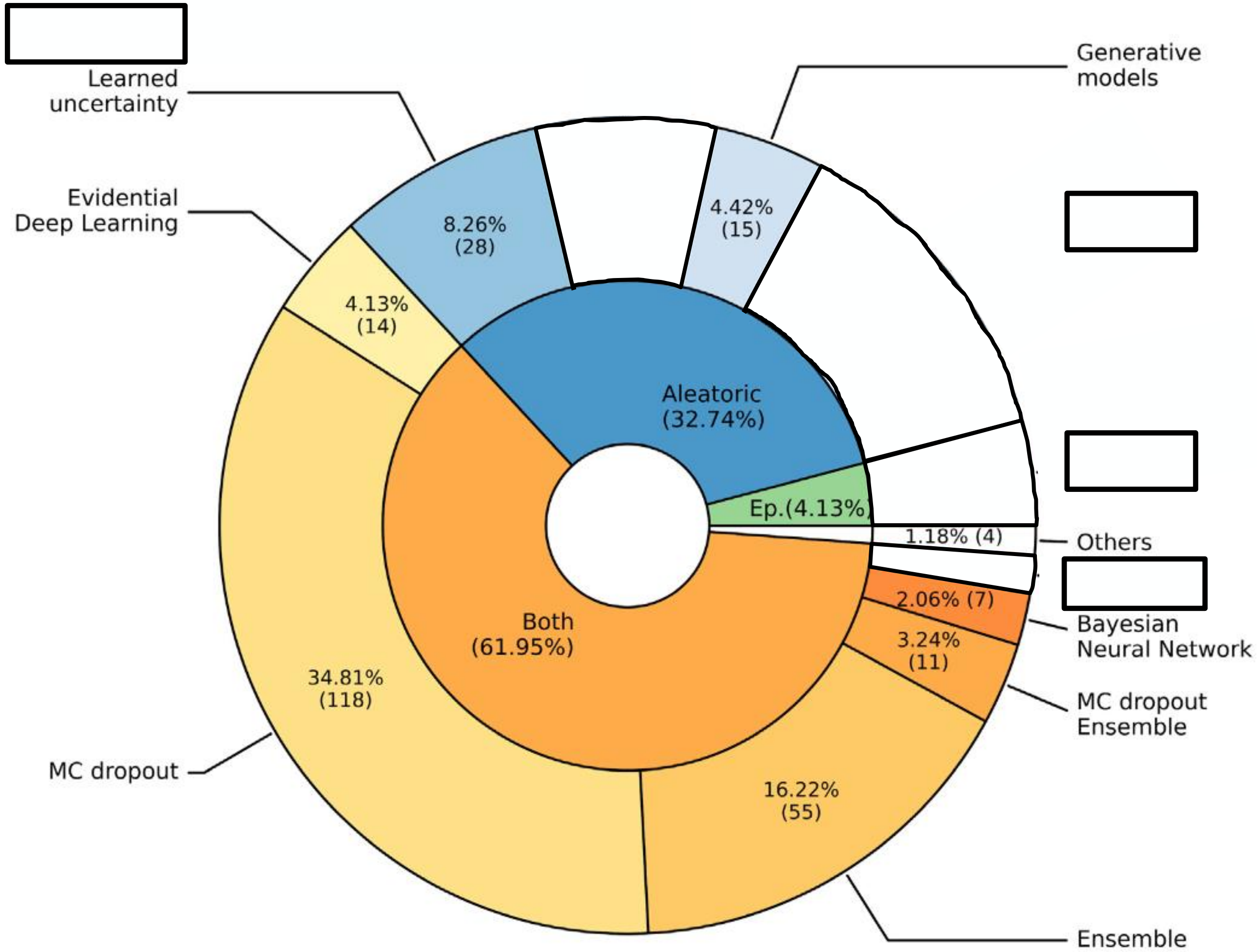
Dice score = 0,56 ⚠

- Uncertainty quantification can help identify cases at risk of poor performance, that need to be reviewed

# Uncertainty Quantification (UQ)

- Multiple approaches described in the literature[1]
- Often directly integrated in architecture / during training (intrinsic methods) → **not easily generalizable**
- Post-hoc methods: applied after model training, with no knowledge on model architecture or weights

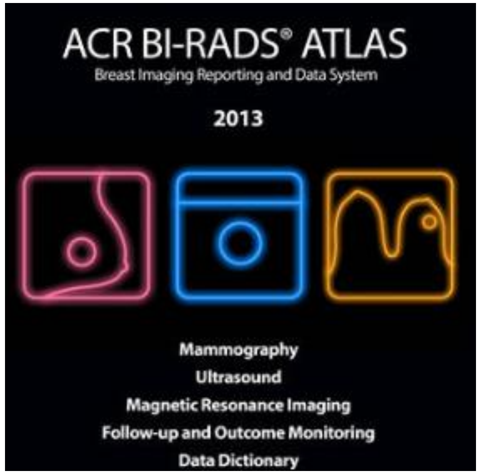
- **Goal of thesis**: develop post-hoc UQ methods as a python library and test it on multiple scenarios (classification, segmentation, regression, with/without access to training data…).

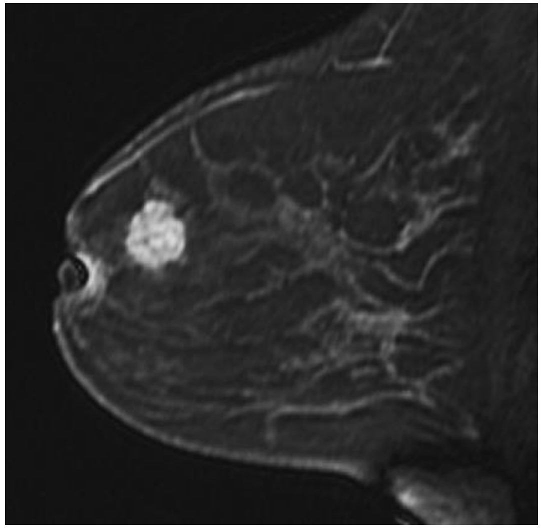


[1] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*, 102830.

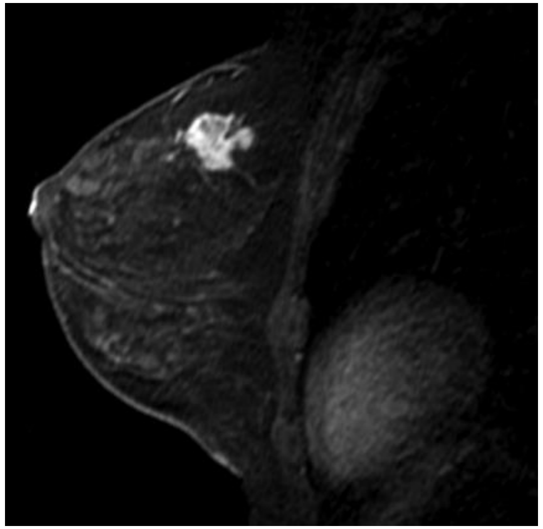
# Use case model – BIRADS classification

- Breast Imaging Reporting and Data System (BI-RADS) : Atlas standardizing breast imaging terminology + structure assessment (shape, margin) and classification system in X-Rays, US and MRI
- Shape criteria shown to be associated with complete response to neoadjuvant chemotherapy[1]
- 103 MRI sequences for training/evaluation + 31 external test cases
- CNN to predict shape of tumors


ACR BI-RADS® ATLAS
Breast Imaging Reporting and Data System
2013
Mammography
Ultrasound
Magnetic Resonance Imaging
Follow-up and Outcome Monitoring
Data Dictionary



Round shape

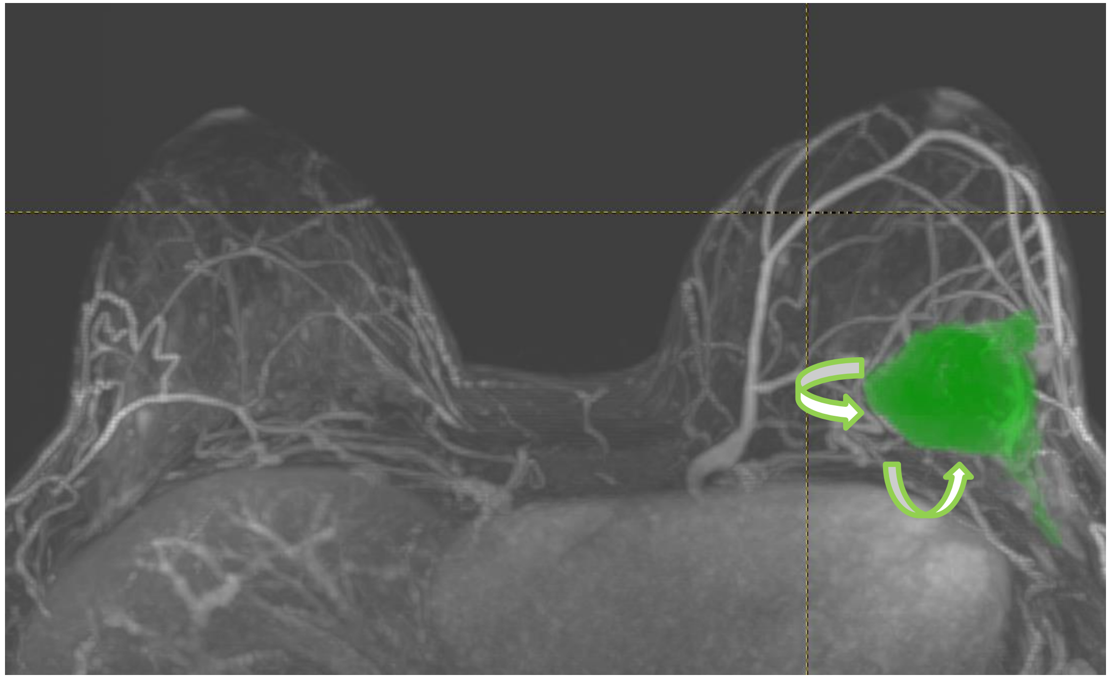


Irregular shape

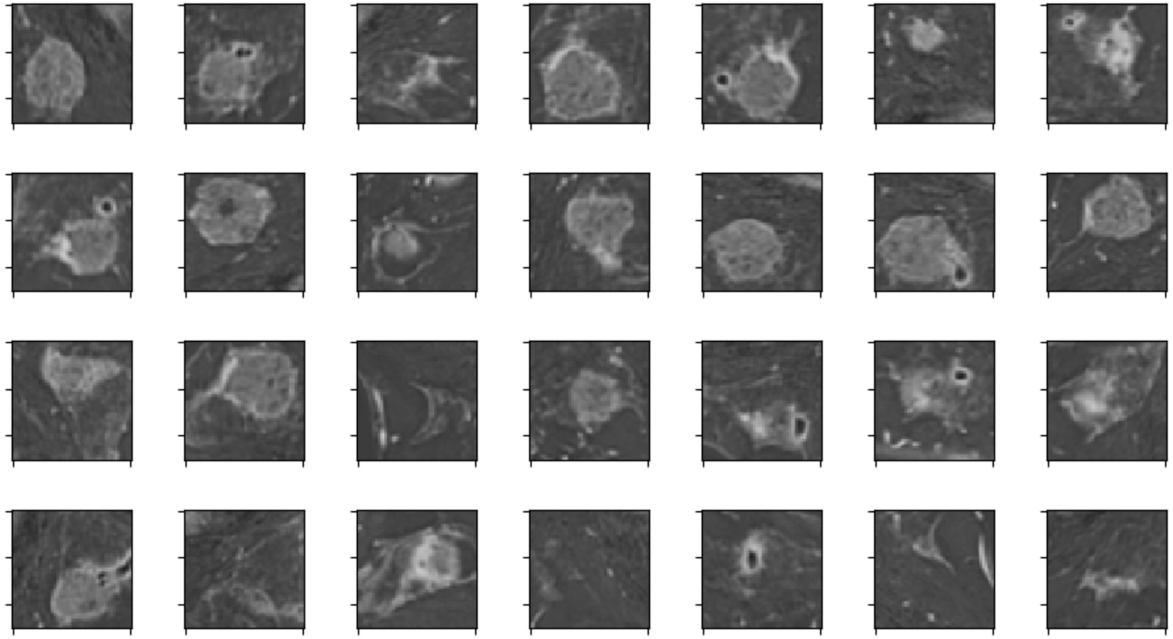
Malhaire, C., Selhane, F., Saint-Martin, M. J., Cockenpot, V., Akl, P., Laas, E., ... & Frouin, F. (2023). Exploring the added value of pretherapeutic MR descriptors in predicting breast cancer pathologic complete response to neoadjuvant chemotherapy. *European Radiology*, *33*(11), 8142-8154.

# Training with natural data augmentation

- Problem: small number of patients (initial DB n=103) + imbalanced classes (1:4) --> data augmentation needed
- Proposal: from 3D to 2D --> use of MRI slices + 3D random rotations of the tumors

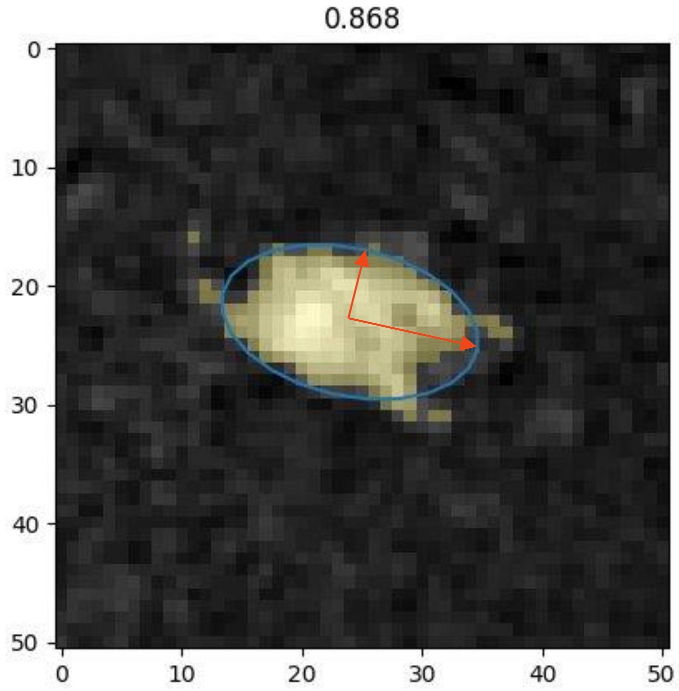


Axial MIP with segmented tumor

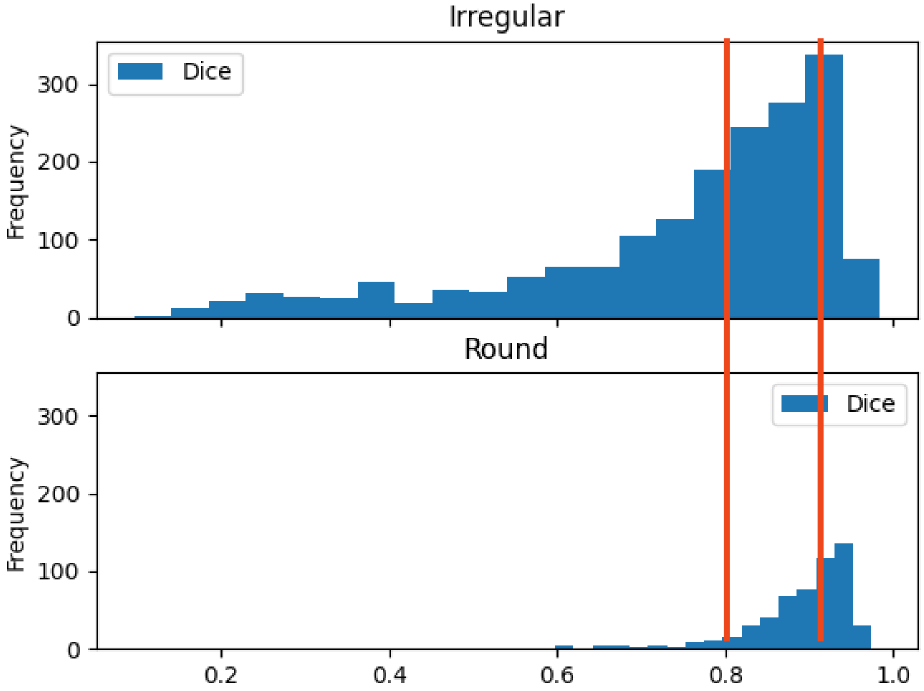


Train: from 103 to 4312 images

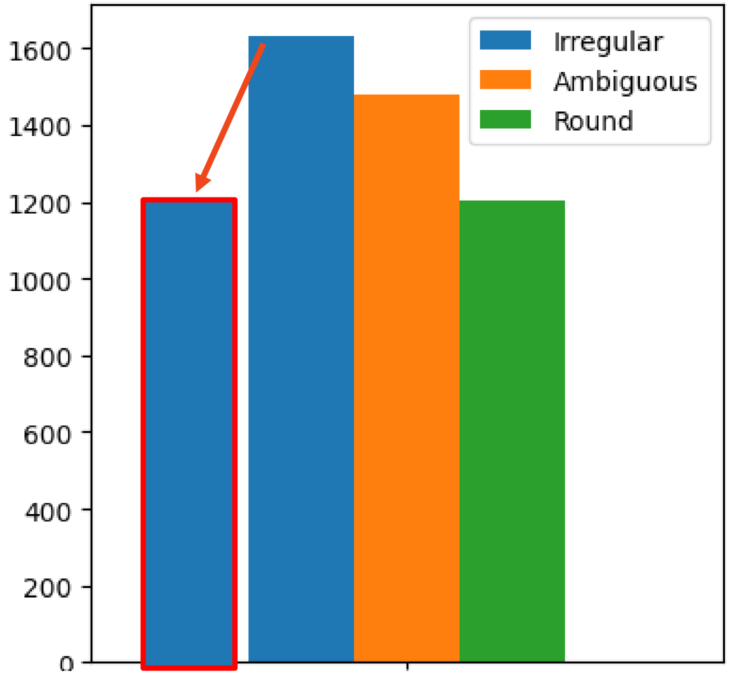
# 2D slices auto-labelling



Dice computation between tumor 2D mask & ROI-enclosing ellipsoid (from PCA of mask)



Slices dice scores distributions for « Round »-labelled tumors and for « Irregular »-labelled tumors



Final repartition (downsampling irregular class to 1200 to balance)

Problem: “round”(“irregular”)-labelled tumors ≠ all slices round (irregular)

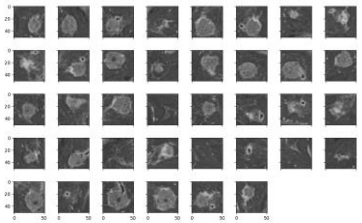
Proposal: 3-class distribution, dice-based: “Irregular” (dice<0.8) | “Ambiguous” (0.8<=dice<0.9) | “Round” (dice>0.9)

**External evaluation DB (domain shift evaluation): from 31 to 1349 images**

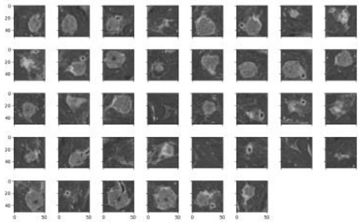
# Training: Data & CNN Architecture

103 patients
5X cross-validation

Train set (n=82)



Test set (n=21)

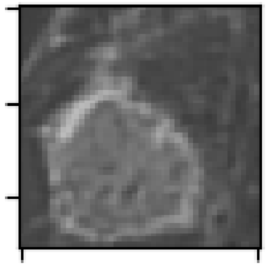


| Folds | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Round | 957 | 1041 | 978 | 988 | 848 |
| Irregular | 970 | 889 | 953 | 979 | 1009 |

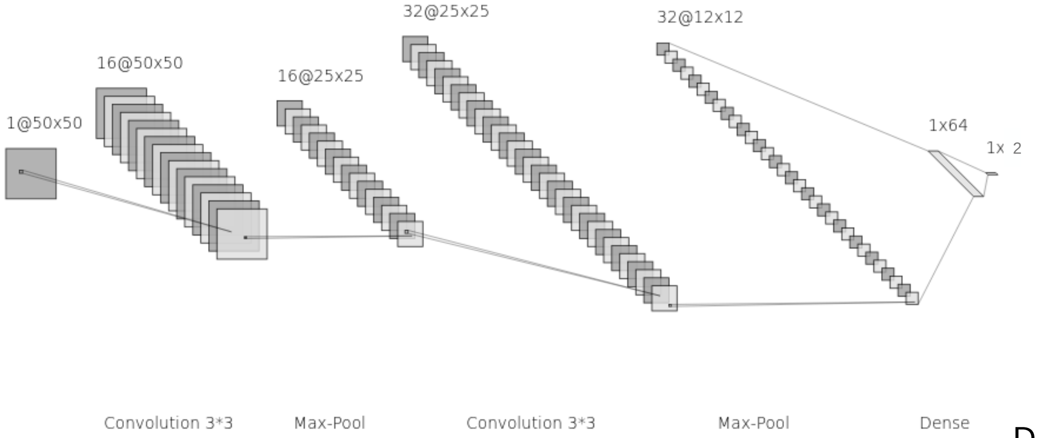
| | | | | | |
|---|---|---|---|---|---|
| Round | 246 | 162 | 225 | 215 | 191 |
| Irregular | 230 | 162 | 247 | 221 | 191 |

Downsampling when imbalance > 2:3

Training with and without random data augmentation



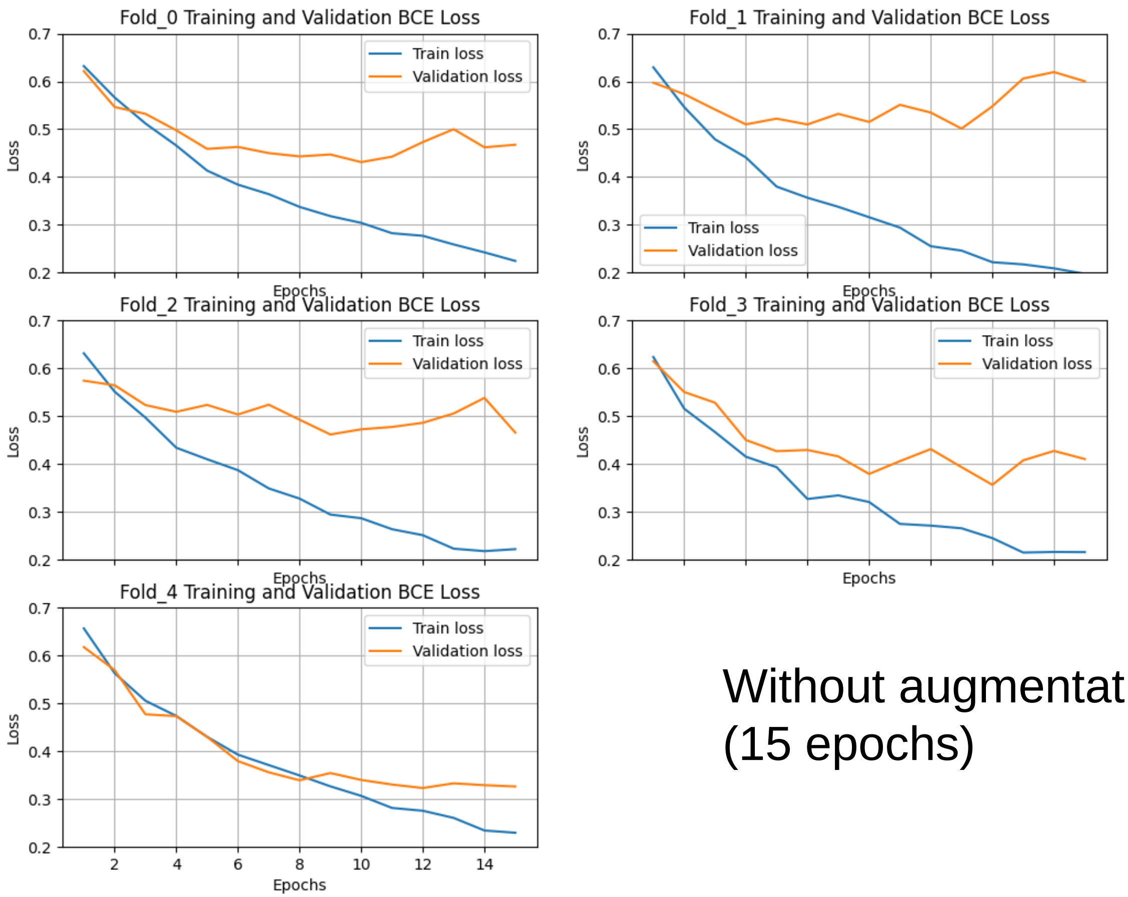
Test slice example



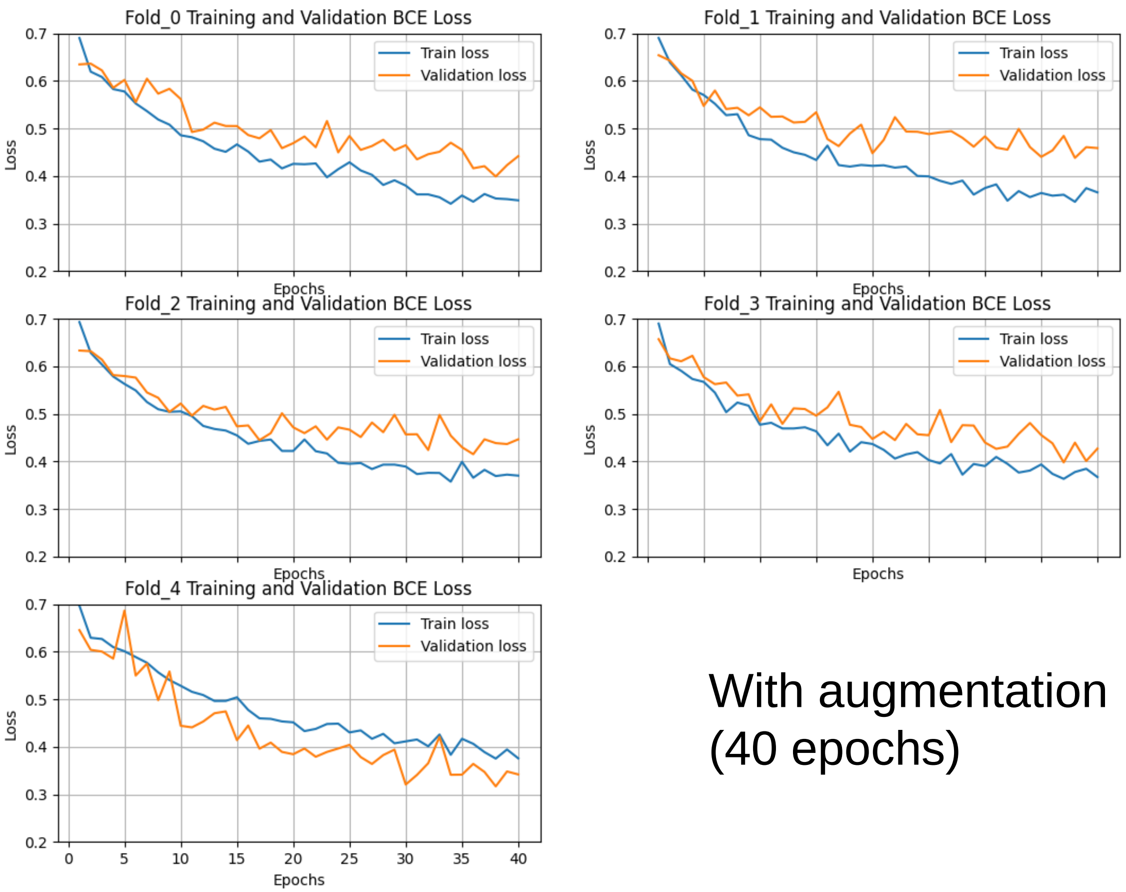
If < 0.5: Round slice
Else: Irregular slice

Dropout + weight decay for regularization

# Training results from cross validation



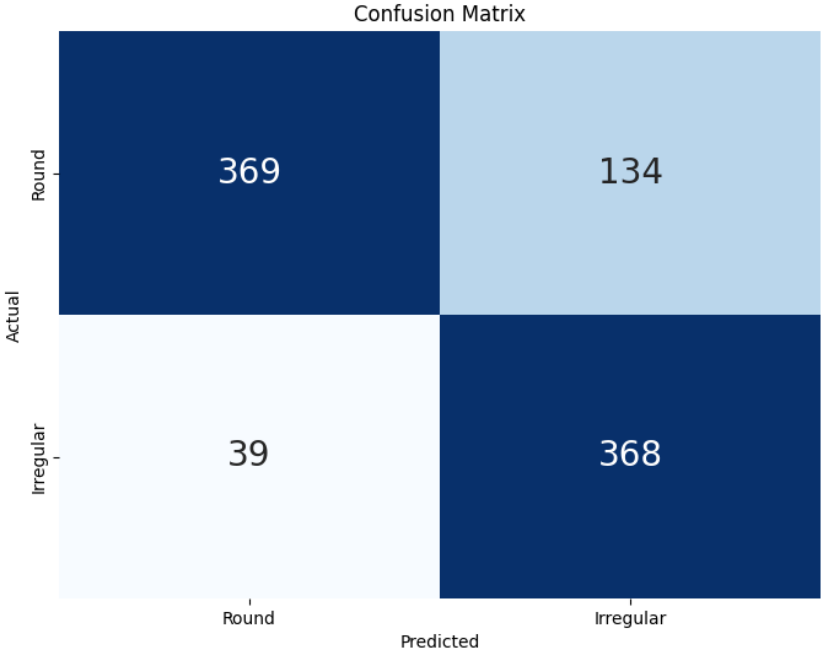
Without augmentation
(15 epochs)

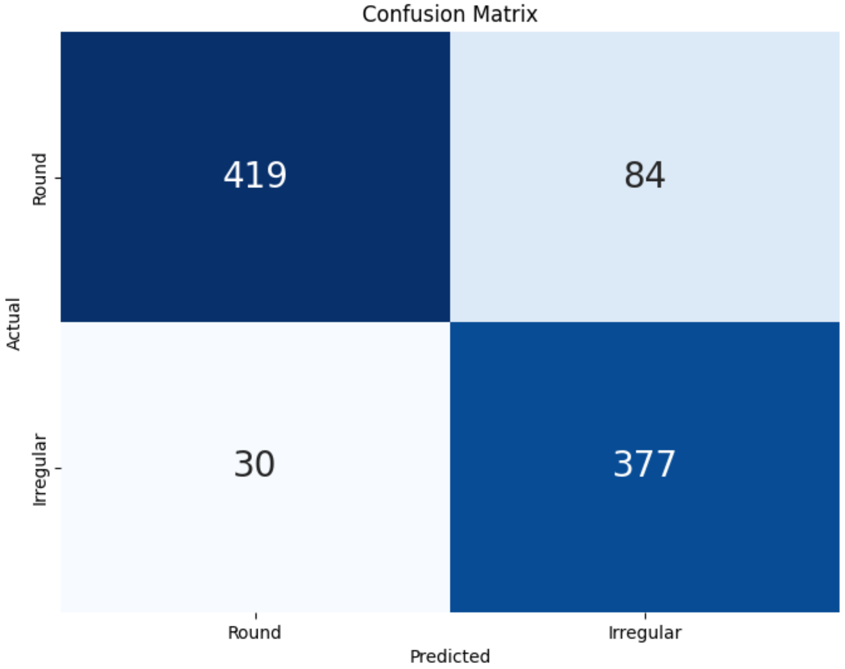


With augmentation
(40 epochs)

- Overall good performance
- No important differences with and without data augmentation at training:
  - w/o data augmentation: mean (SD) **Roc AUC = 0.89 (0.03)**
  - w data augmentation: mean (SD) **Roc AUC = 0.88 (0.02)**

# External evaluation DB





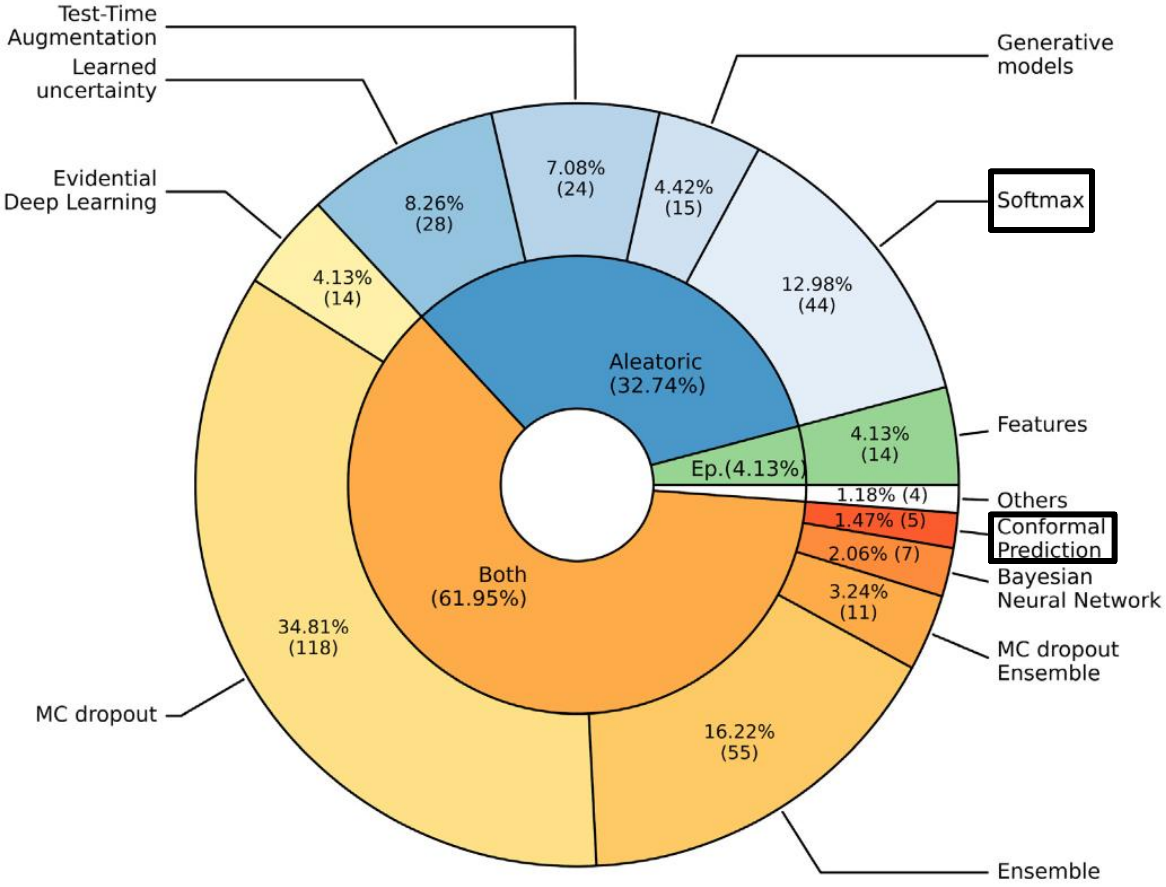
| | Without data augmentation | With data augmentation |
|---|---|---|
| Accuracy | 0.81 | **0.87** |
| Roc AUC | 0.86 | **0.93** |
| Sensitivity | 0.90 | **0.93** |
| Specificity | 0.73 | **0.83** |

Data augmentation increases robustness and performance on new unseen data

# Methods for Uncertainty Quantification

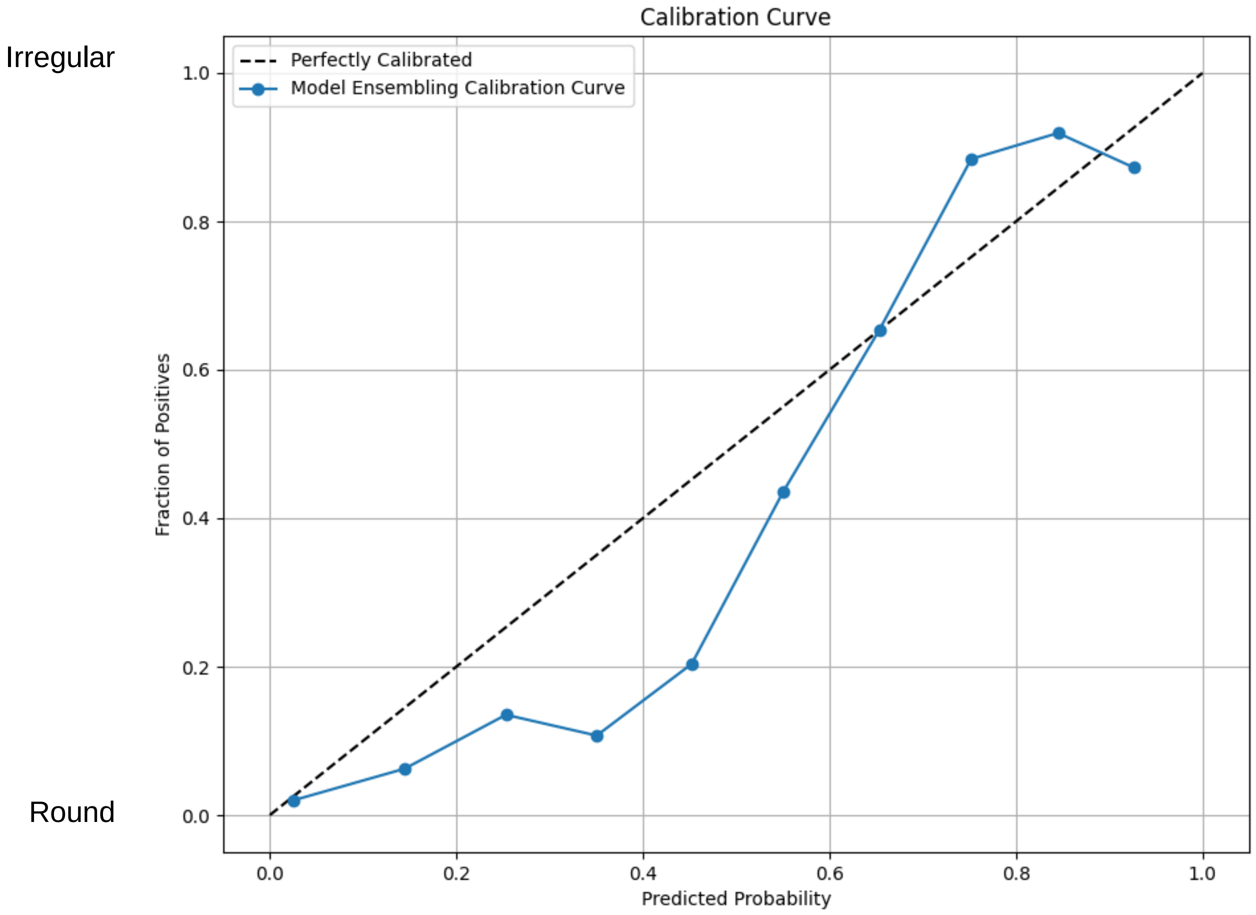
Needed:

- Trained model (w/wo data augmentation)
- Annotated test set



# Uncertainty Quantification: Model calibration – Softmax

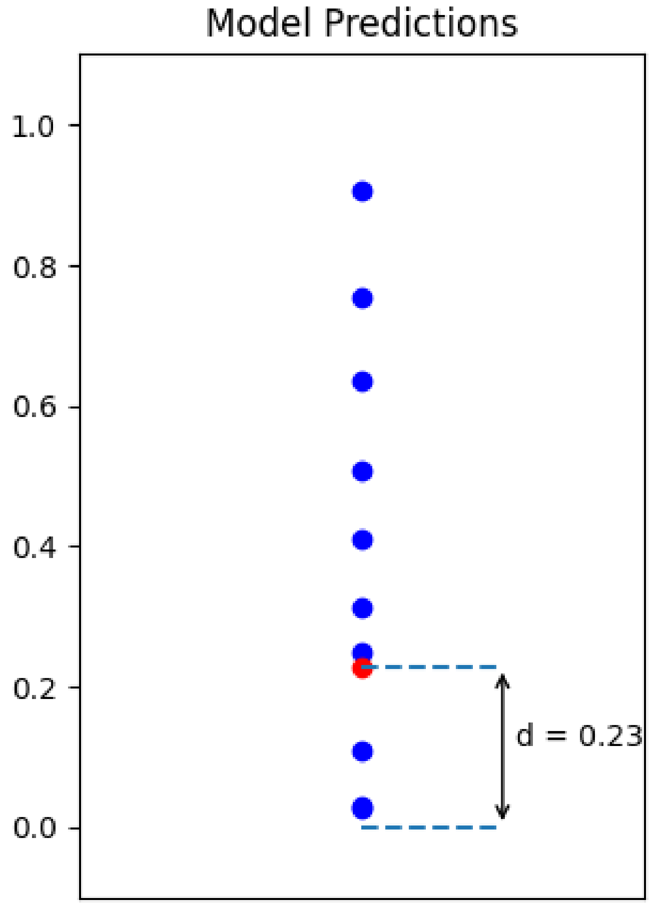
- Is the output probability correlated with the frequency of correct predictions ?

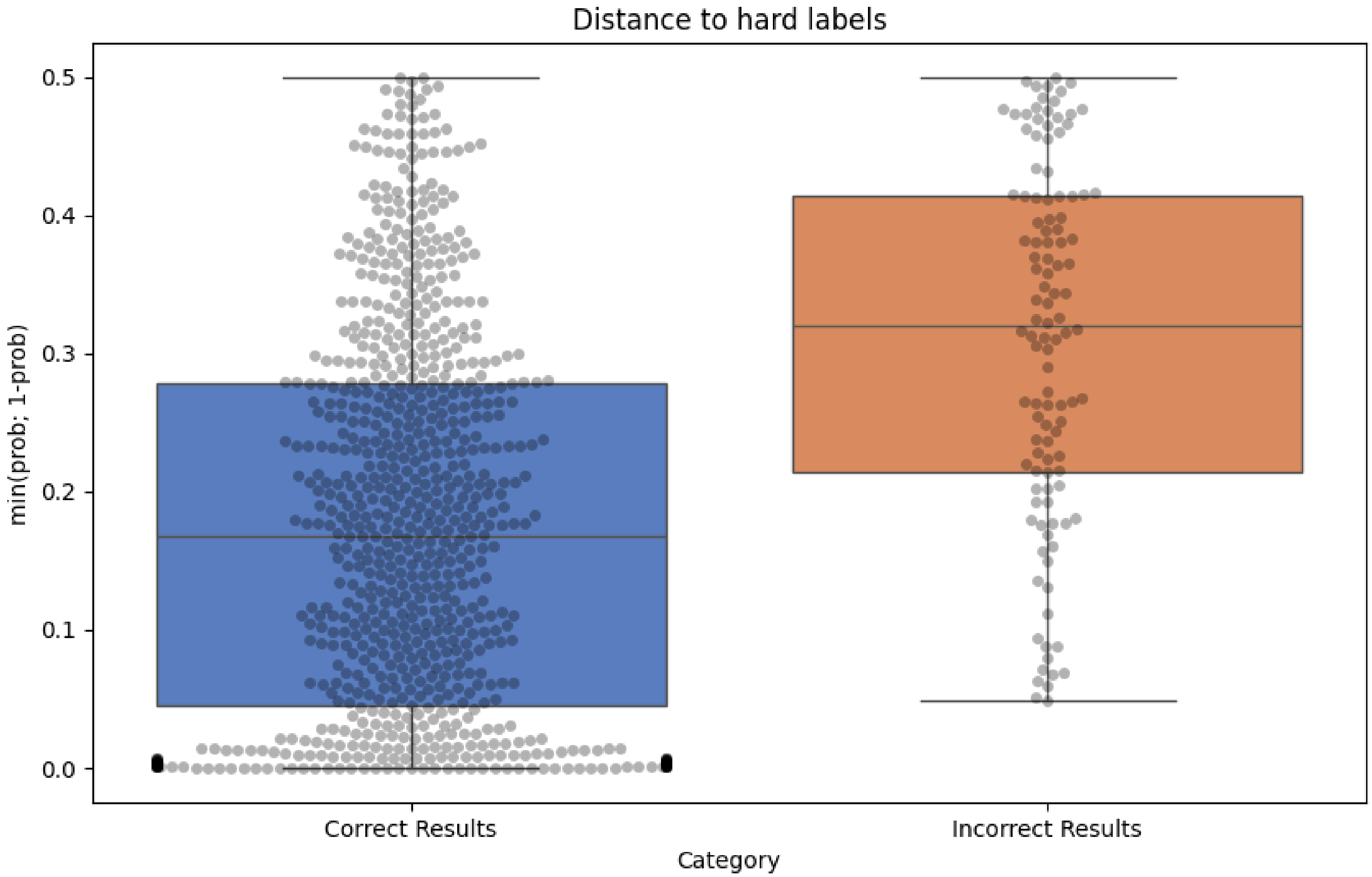


For predictions < 0.5: below x=y curve -->underconfident | above --> overconfident
> 0.5: above x=y curve -->underconfident | below --> overconfident

# Uncertainty Quantification: Model calibration – Softmax

- If model correctly calibrated: predicted probability distance to hard labels (round: 0, irregular: 1) as a proxy for UQ

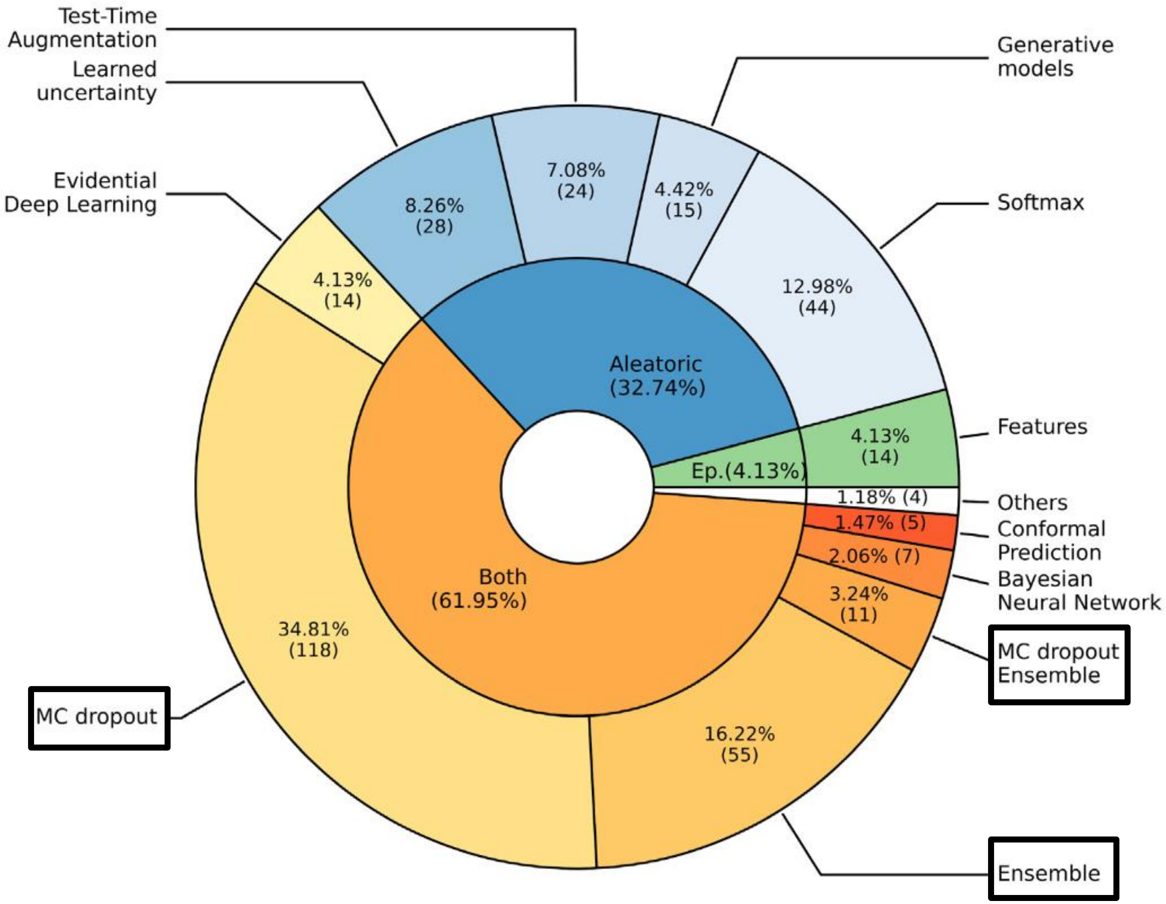




# Methods for Uncertainty Quantification

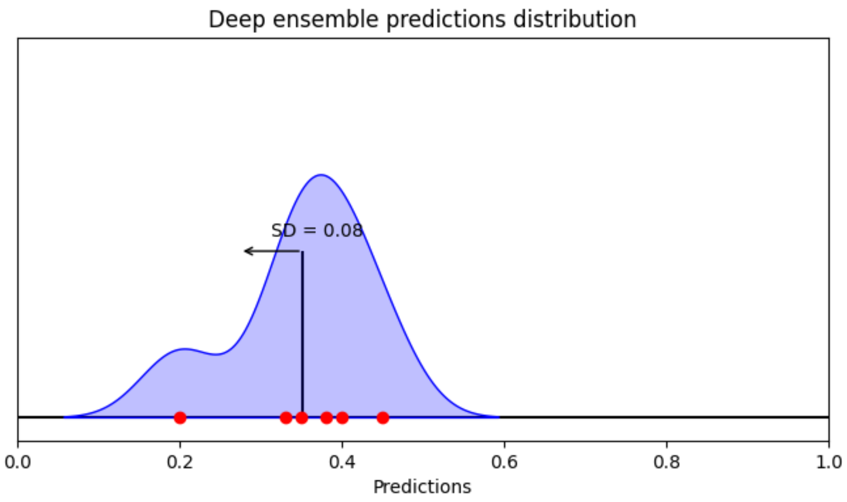
Needed:

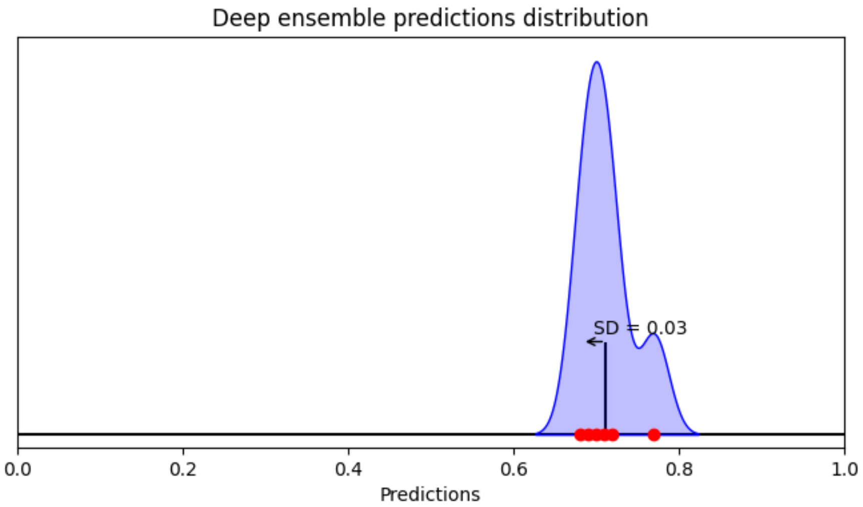
- Trained model (w/wo data augmentation)
- Sub-ensembles results if ensembling approaches used in training

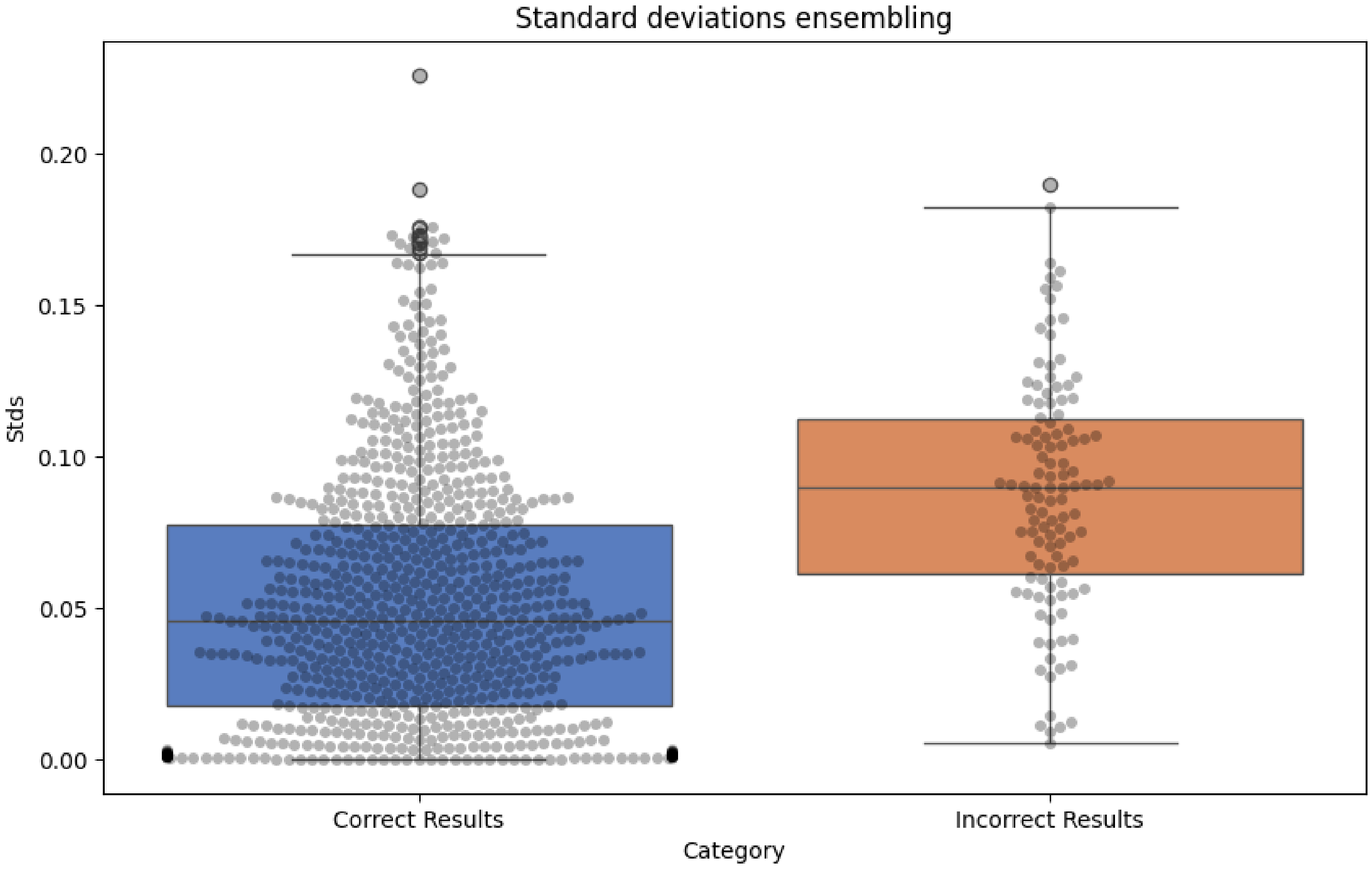


# Uncertainty Quantification: Variability of deep ensemble results

- Variability of ensembling prediction to quantify uncertainty



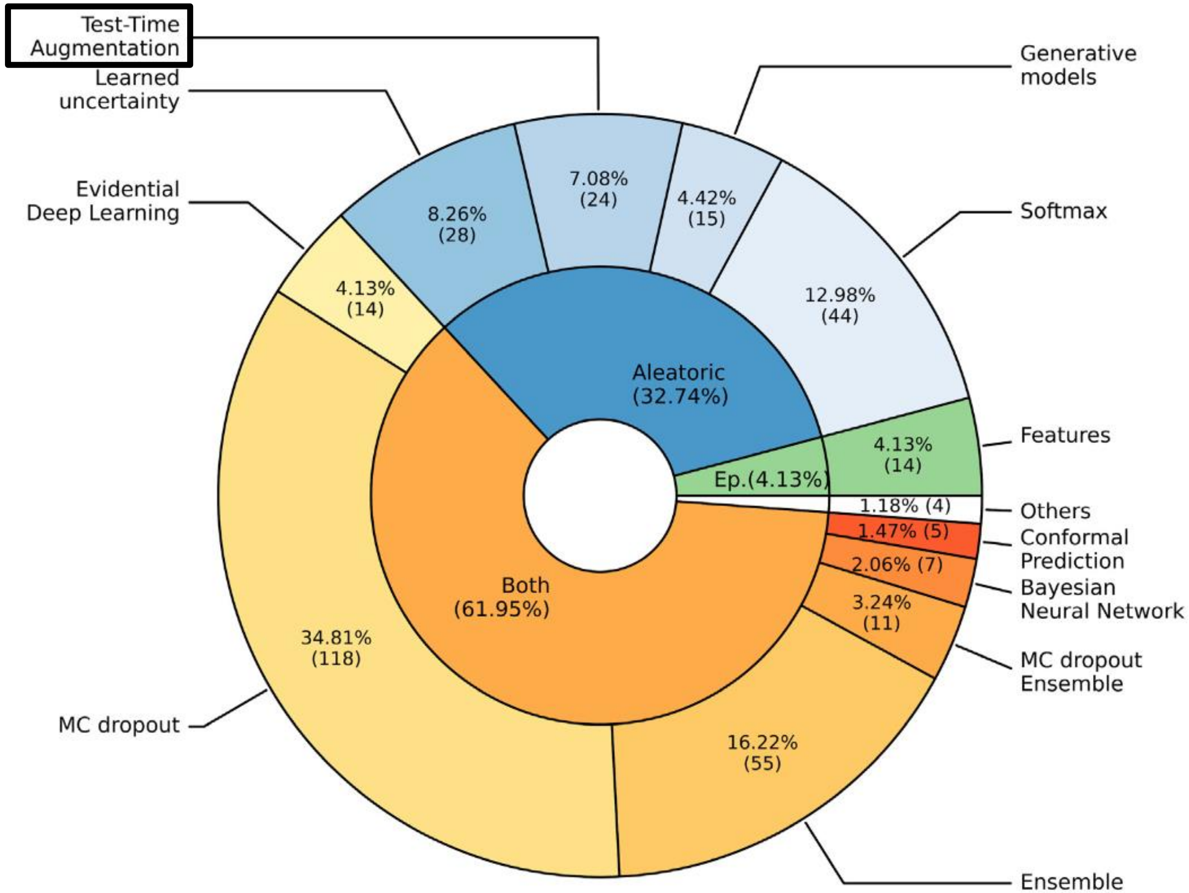




# Methods for Uncertainty Quantification

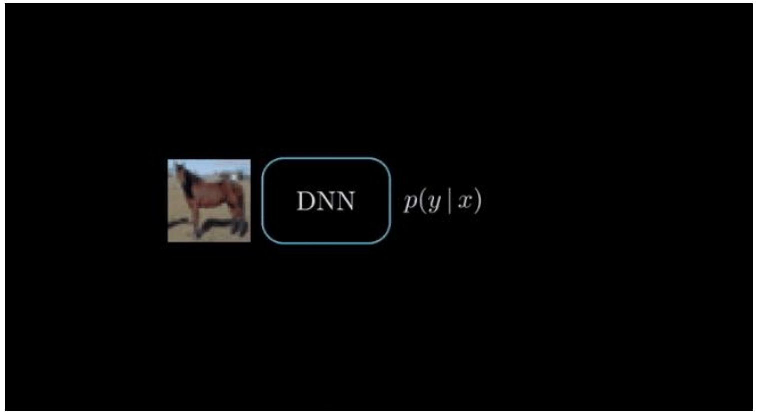
Needed:

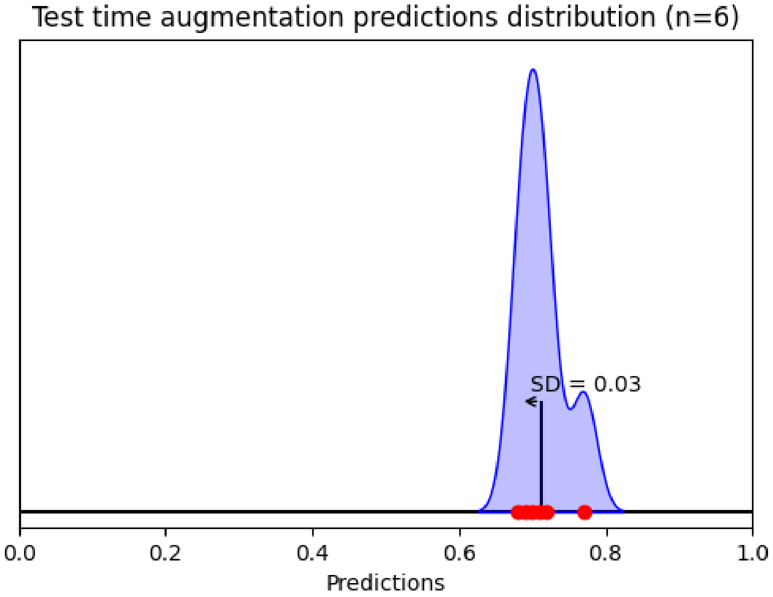
- Trained model (w/wo data augmentation)
- Annotated test set (if augmentation policies search)

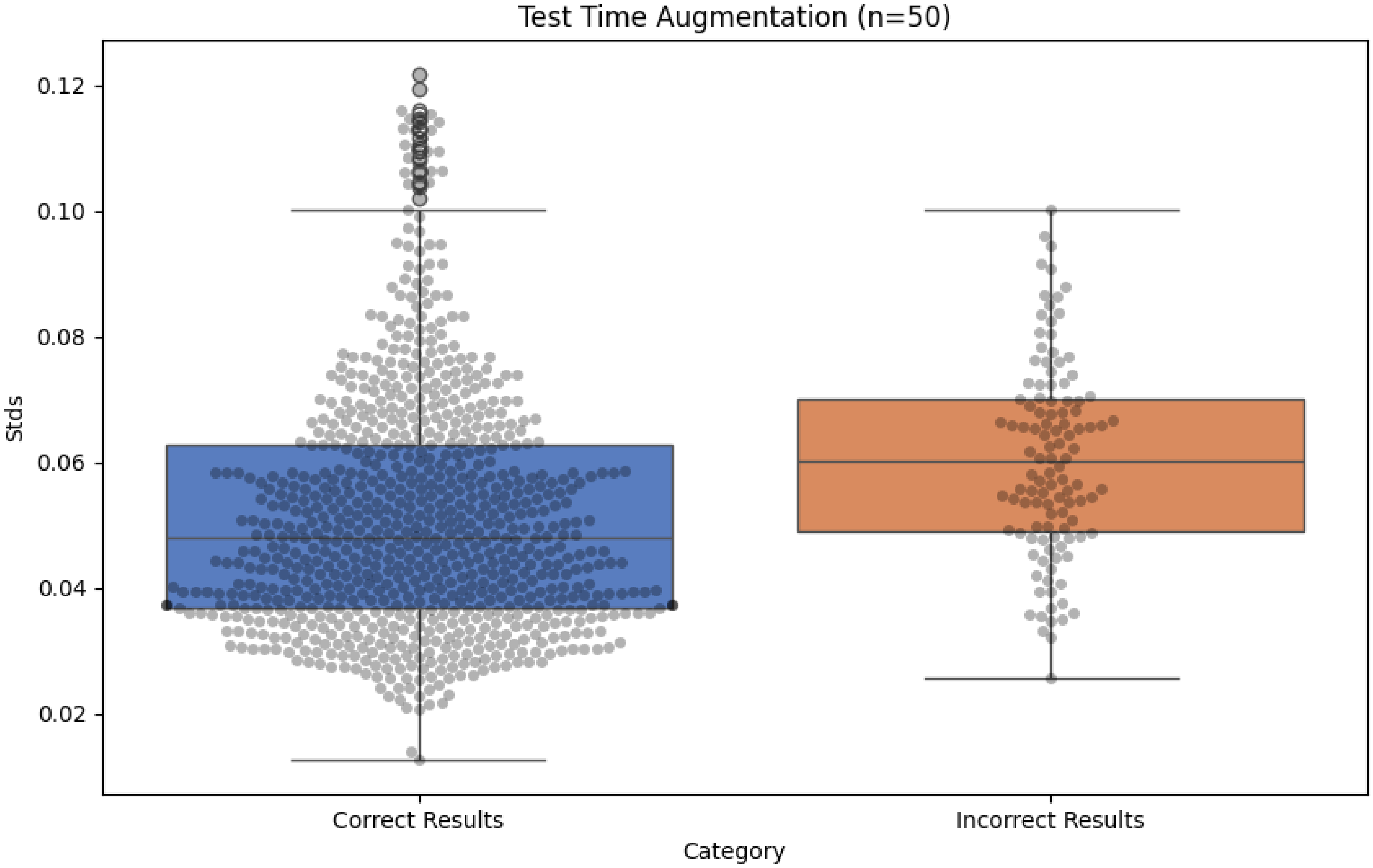


# Uncertainty Quantification: Test-time augmentation

- Variability after augmentations at prediction to quantify uncertainty

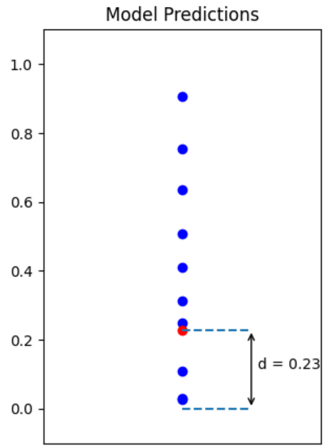


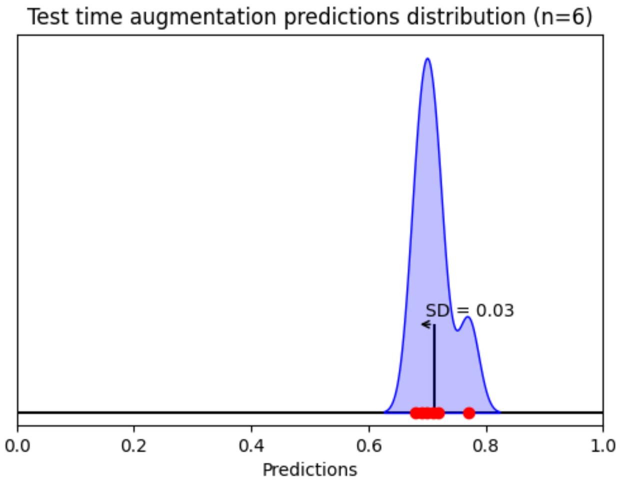




# Uncertainty Quantification: Methods combination

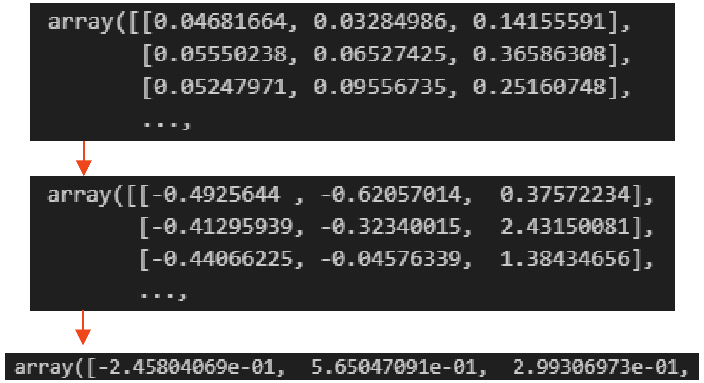
- How to combine results (different metrics and scales) ?
- Standardization + mean result across methods





$$z = \frac{X - \mu}{\sigma}$$

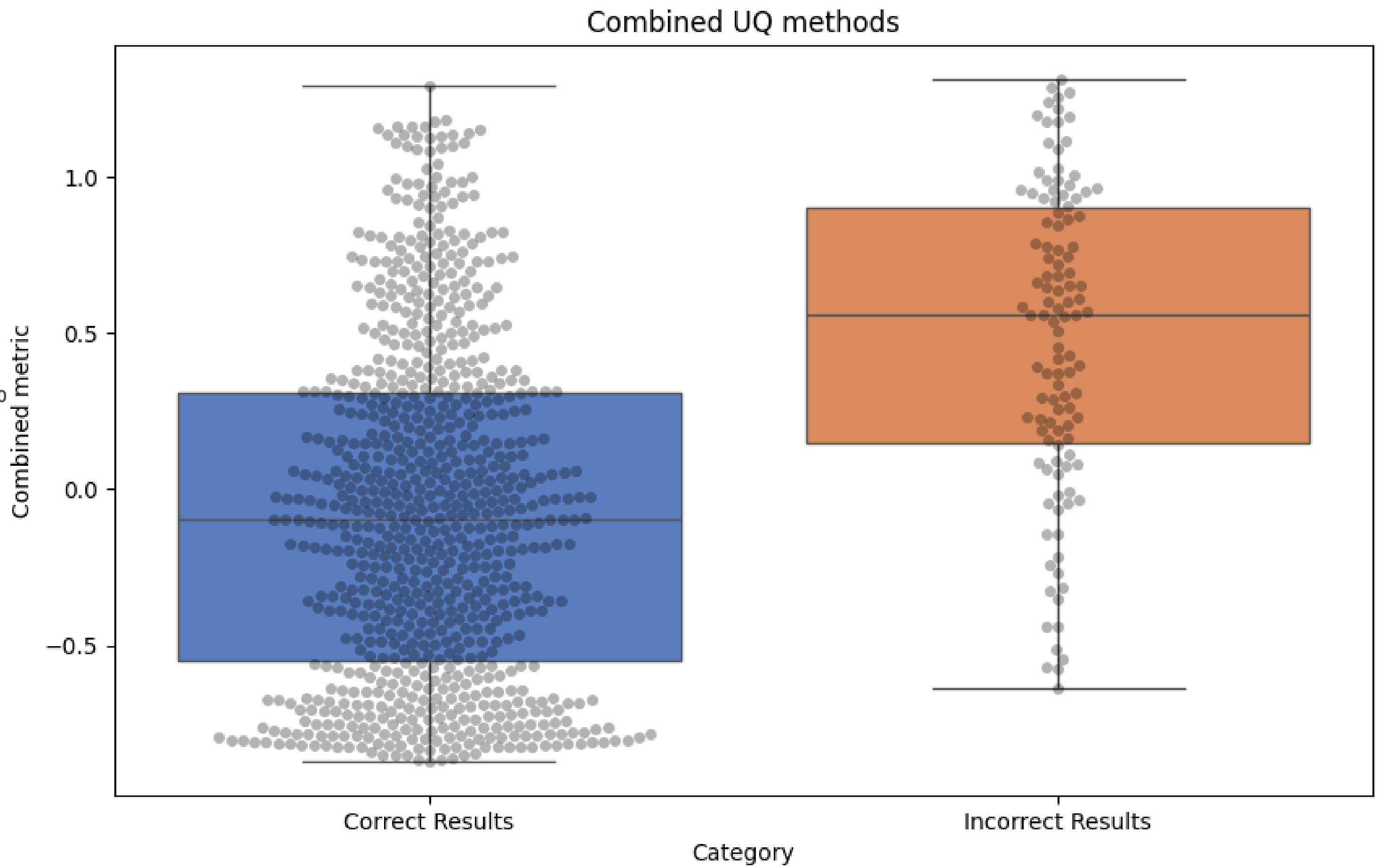
$\bar{X}$



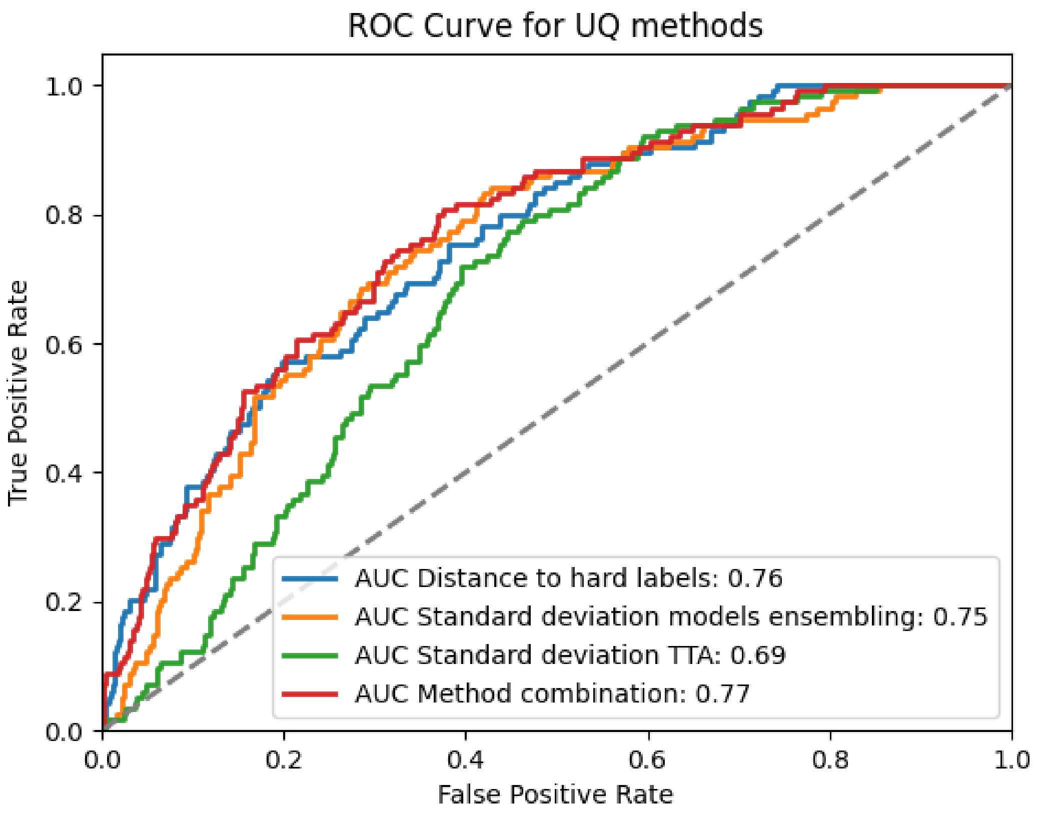
```
array([[0.04681664, 0.03284986, 0.14155591],
       [0.05550238, 0.06527425, 0.36586308],
       [0.05247971, 0.09556735, 0.25160748],
       ...,
```

```
array([[-0.4925644 , -0.62057014,  0.37572234],
       [-0.41295939, -0.32340015,  2.43150081],
       [-0.44066225, -0.04576339,  1.38434656],
       ...,
```

```
array([-2.45804069e-01,  5.65047091e-01,  2.99306973e-01,
```



# Methods evaluation



- Receiver operating characteristic curves to evaluate performance of each UQ method + Combination method
- Clear trend, but not good enough to almost systematically identify failure cases

# Perspectives

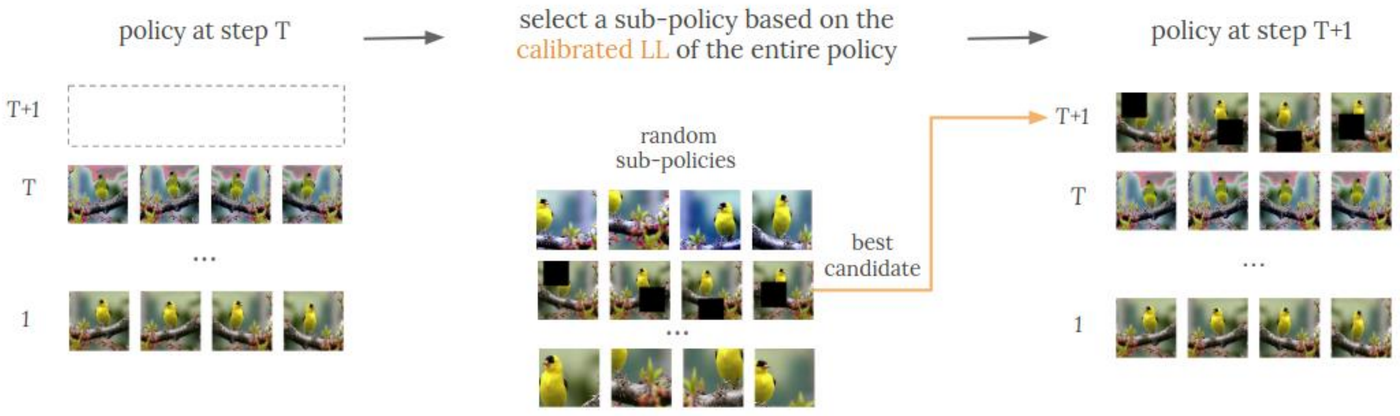
- *Uncertainty quantification:*
  - New post-hoc methods to implement and test (test-time augmentation, out of distribution detection into latent space…)
  - Evaluation of methods to combine UQ metrics
- *Use cases:*
  - Add more patients (80 labelled MRI ready) for BIRADS classification task
  - Test methods on other models available at the lab (segmentation, multi-class classification, survival prediction)

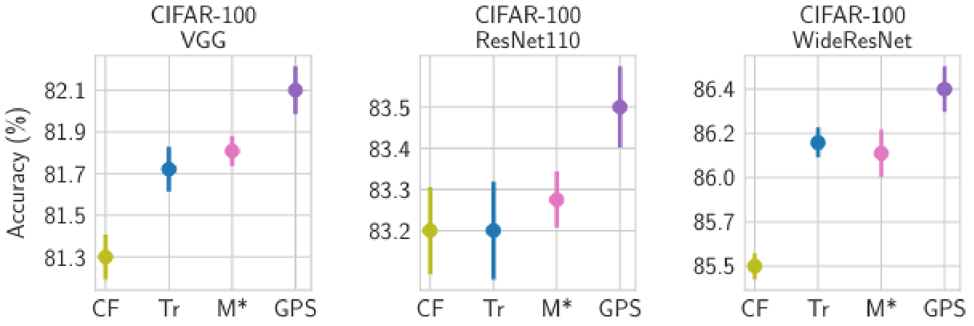
# Test-time augmentation, article review: GPS (Greedy Search Policy)

Problem: Already trained black-box model with no control on augmentation policies used → Variability involved by too aggressive augmentations can bias TTA results by modifying labels

**Needed: Black-box model and external dataset**

GPS: Learn test time augmentation policies that best improves predictive performance and UQ.





Iterative process that searches sub-policy that most improves calibrated log-likelihood at each step

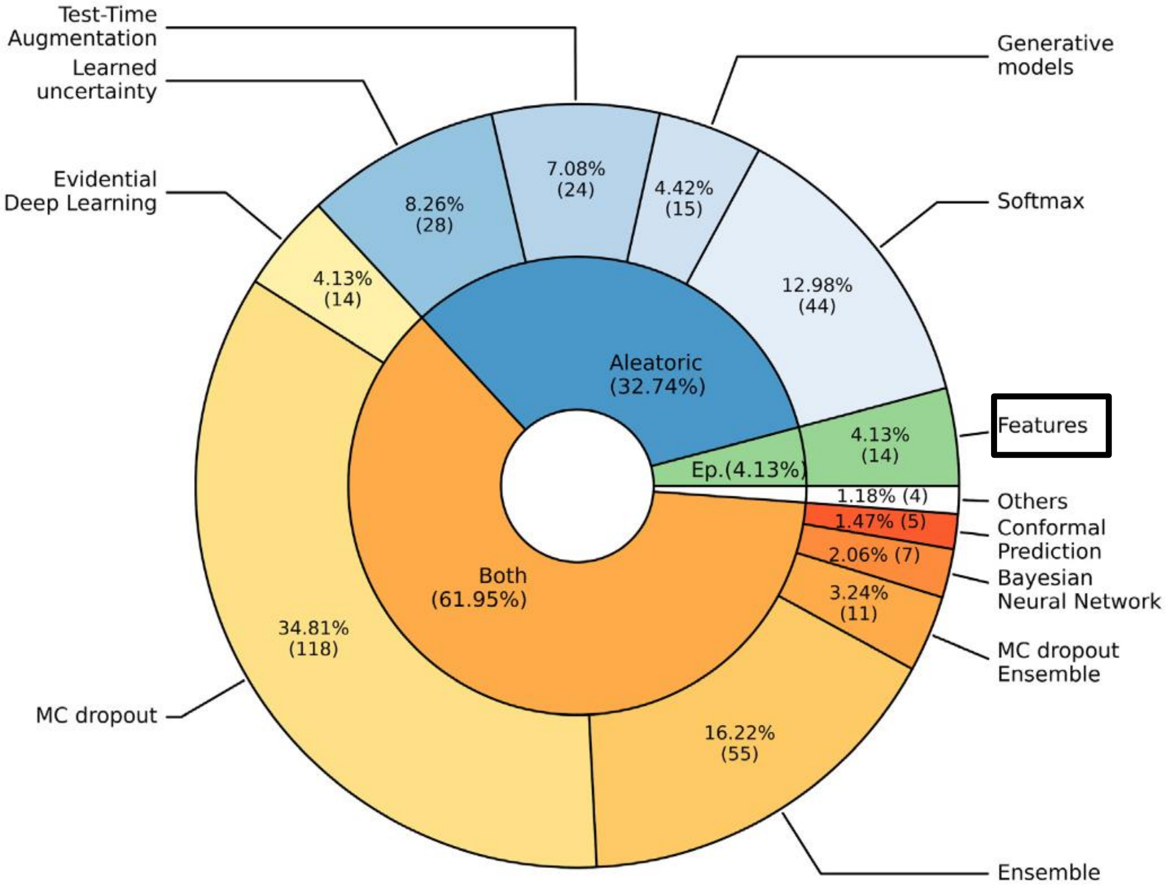
Results on varying type of tasks/models/datasets vs random crops and horizontal flips (CF), augmentation used for training (Tr) and Randaugment with optimal magnitude (M*)

https://github.com/SamsungLabs/gps-augment

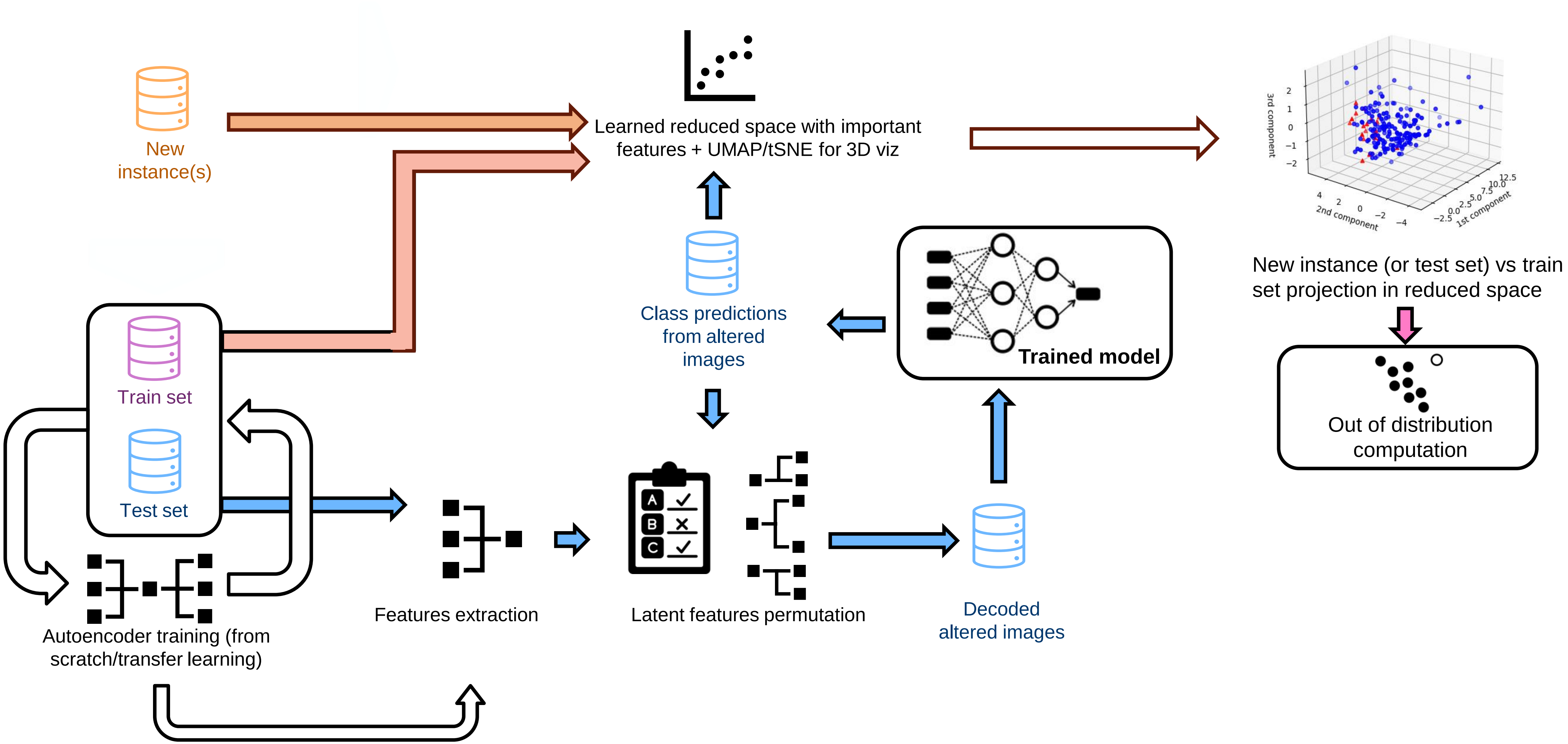
# Methods for Uncertainty Quantification

Needed:

- Trained model (w/wo data augmentation)
- Training set
- Annotated test set



# Out of distribution detection



New instance(s)

Learned reduced space with important features + UMAP/tSNE for 3D viz

New instance (or test set) vs train set projection in reduced space

Out of distribution computation

Train set

Test set

Autoencoder training (from scratch/transfer learning)

Features extraction

Latent features permutation

Decoded altered images

Trained model

Class predictions from altered images