ATTENTION-GUIDED TRAINING

Combining domain priors and explainability methods for improved trustworthiness and performance

Jesco Talies, David Melching, Eric Breitbarth

DLR Institute for Frontier Materials on Earth and in Space

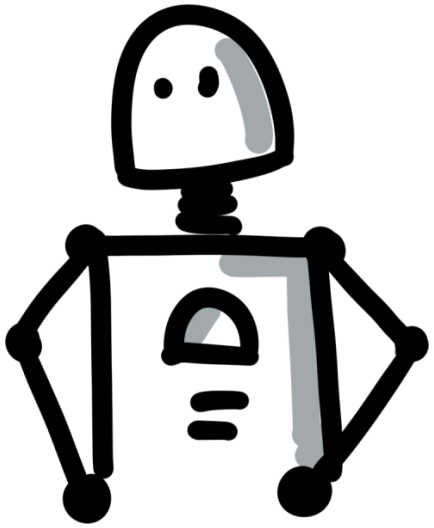
Illustration by Martina Floßdorf

A brief motivation…

Thought experiment

Possible Inputs

?



Possible Outputs:

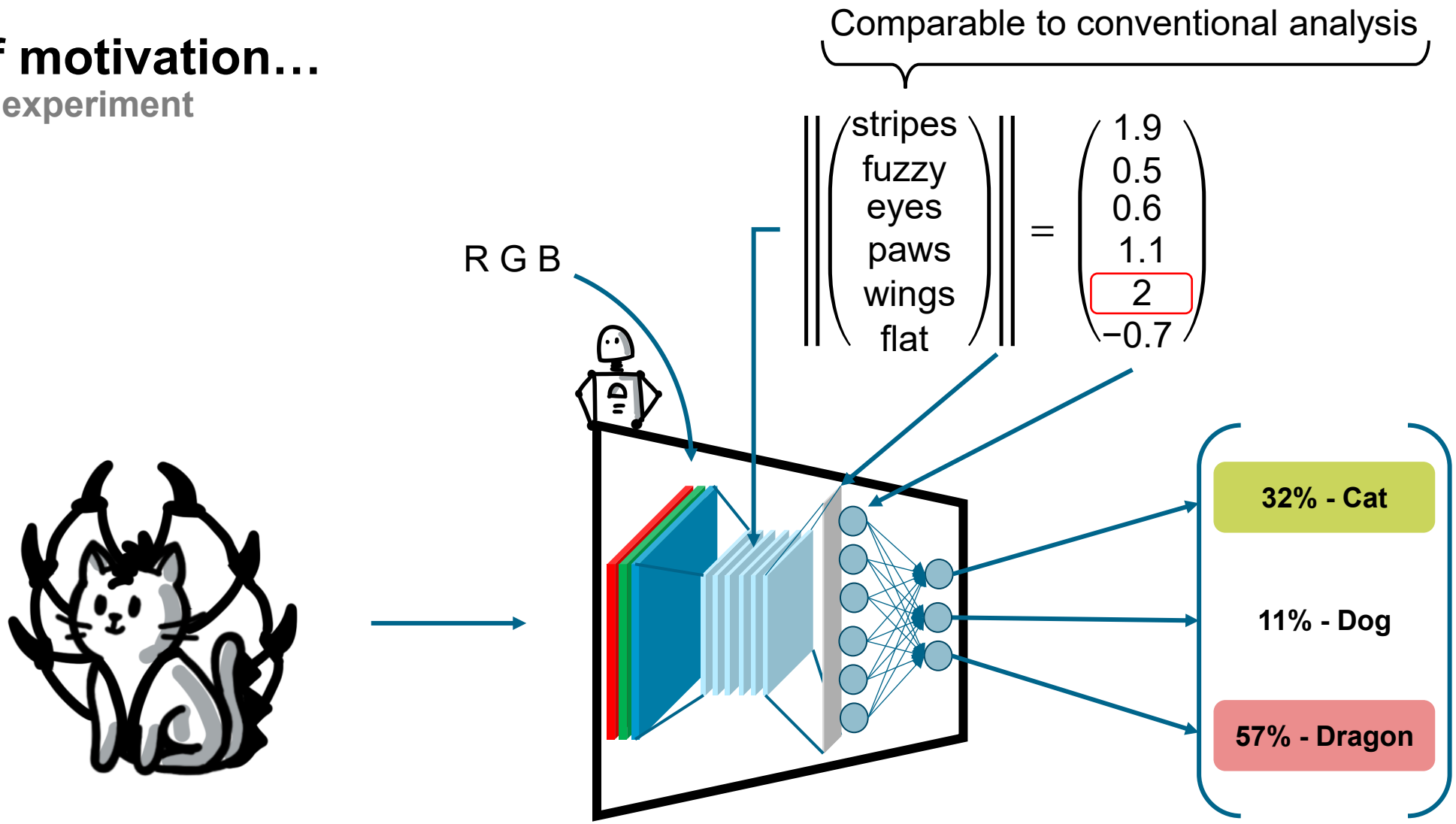
32% - Cat

11% - Dog

57% - Dragon

A brief motivation…

Thought experiment

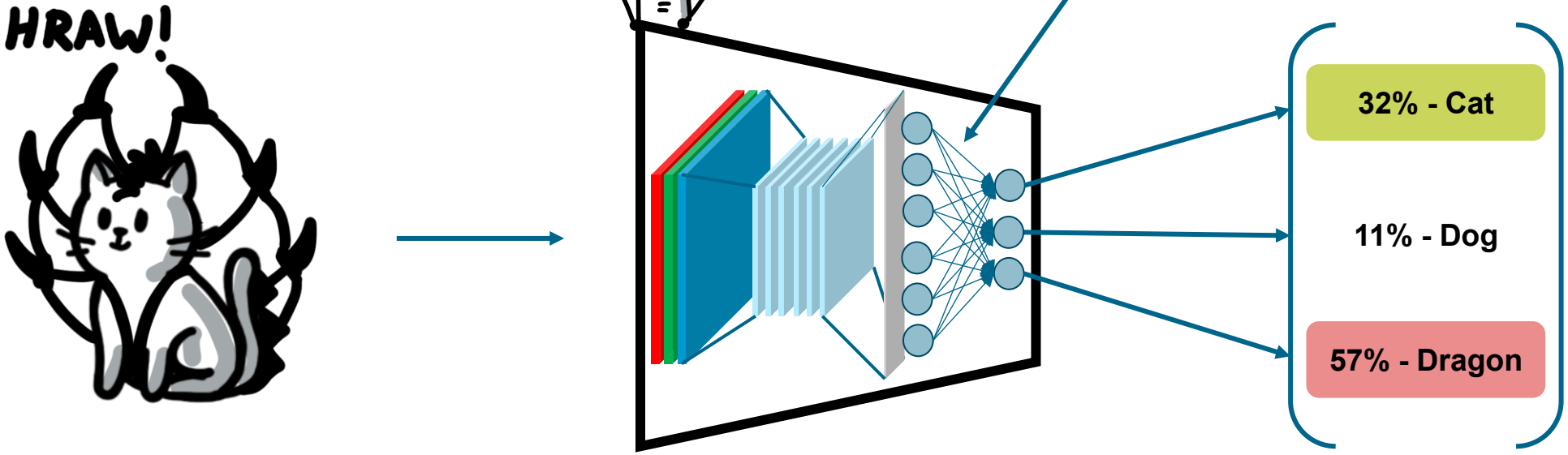


A brief motivation…

Thought experiment

	cat	dog	dragon
stripes	0.7	−1.1	−0.3
fuzzy	0.7	0.9	−0.4
eyes	0.3	1.2	0.7
paws	1.8	1.1	−0.2
wings	−0.2	−0.6	2
flat	0.1	0.1	−0.4

Weight Matrix

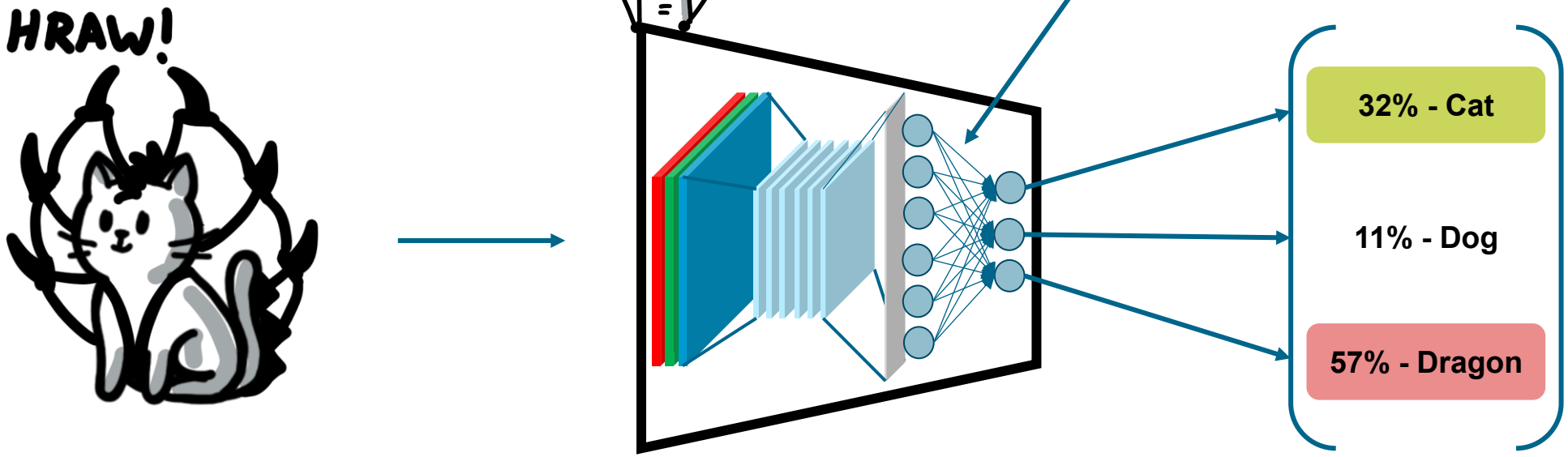


A brief motivation…

Thought experiment

	cat	dog	dragon
stripes	0.7	−1.1	−0.3
fuzzy	0.7	0.9	−0.4
eyes	0.3	1.2	0.7
paws	1.8	1.1	−0.2
wings	−0.2	−0.6	2
flat	0.1	0.1	−0.4

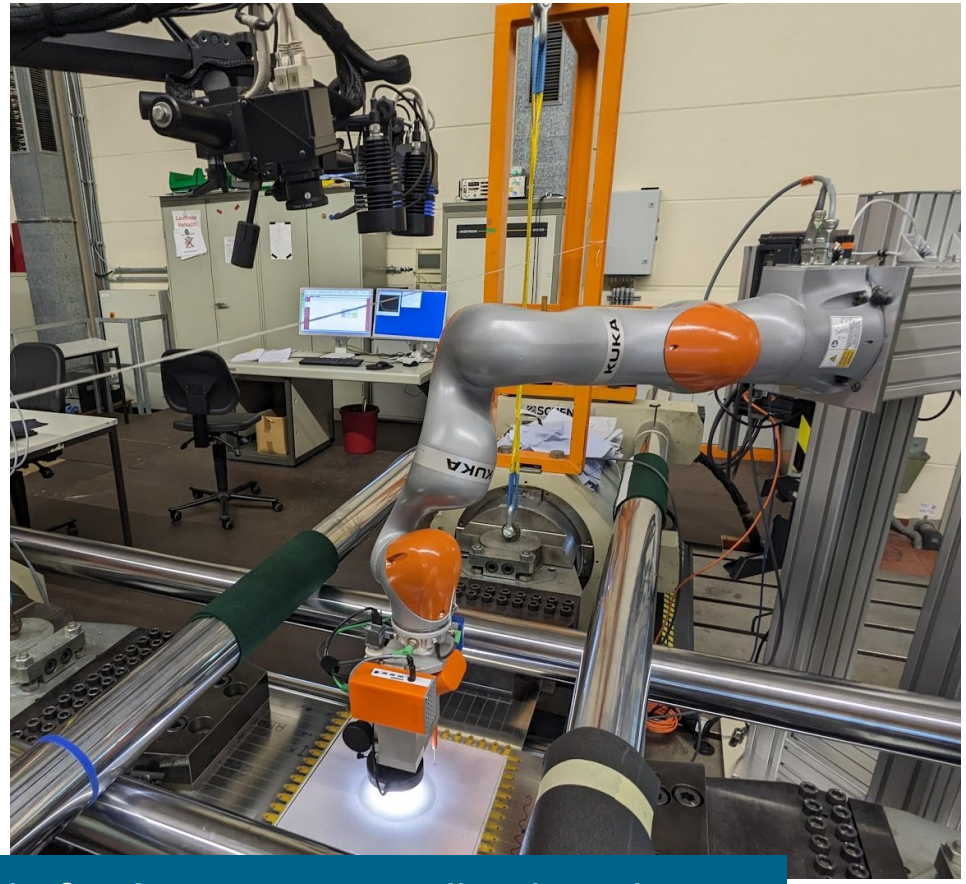
Weight Matrix



X-AI techniques can be used to gain insights into modes of failure or success in black-box models!

A brief motivation…

Connection to the real world

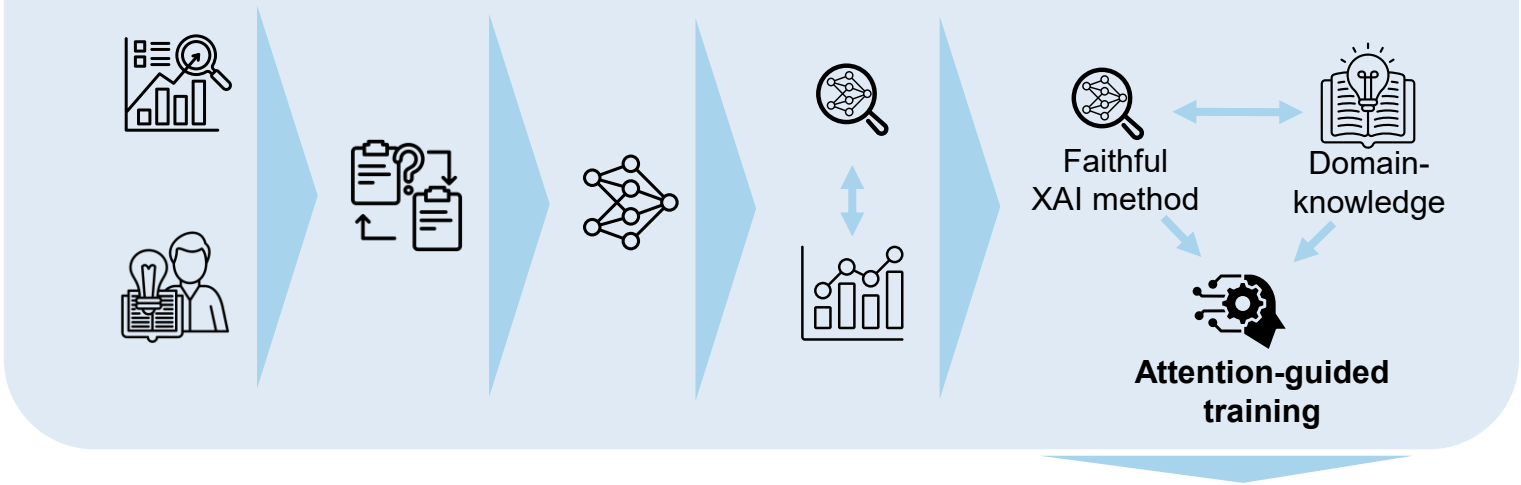


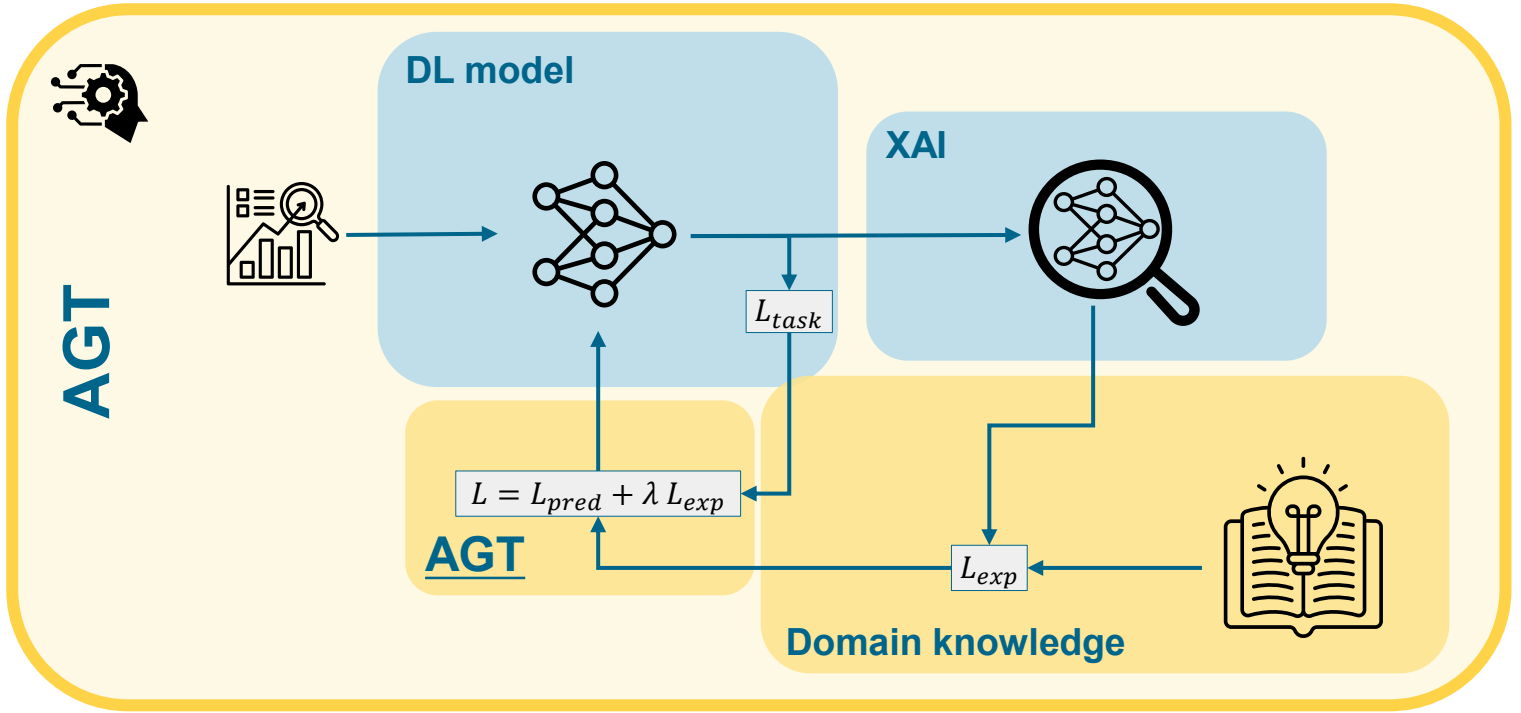
Developing and certifying novel materials for Aerospace applications is a safety critical field.

Understanding and improving our tools is a crucial part of the process!

Paysan, F., Dietrich, E., & Breitbarth, E. (2023) *Experimental Mechanics*

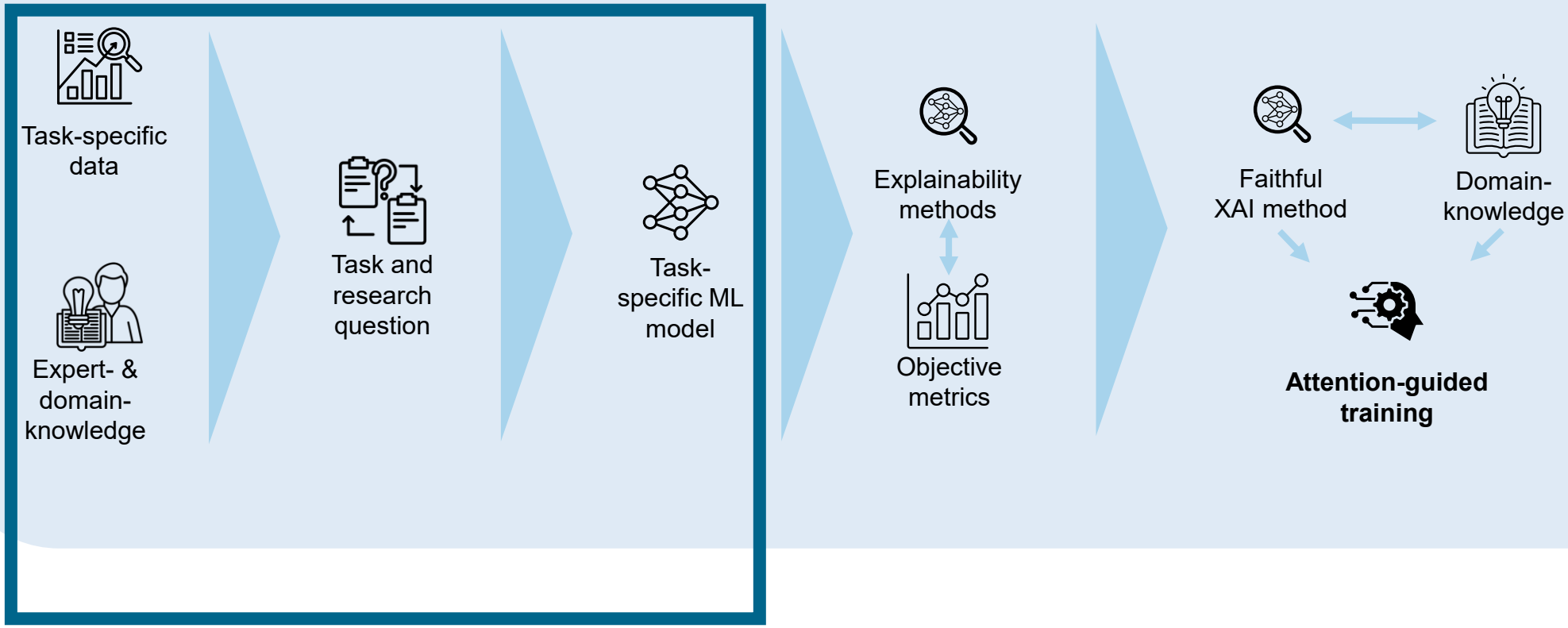
Attention-guided training framework





Towards Trustworthy AI in Materials Mechanics through Domain-Guided Attention
Talies, J., Breitbarth, E., & Melching, D. (2025) - arXiv preprint

Attention-guided training



Part 1:

Task and data and model choice

Domain knowledge

Part 2:

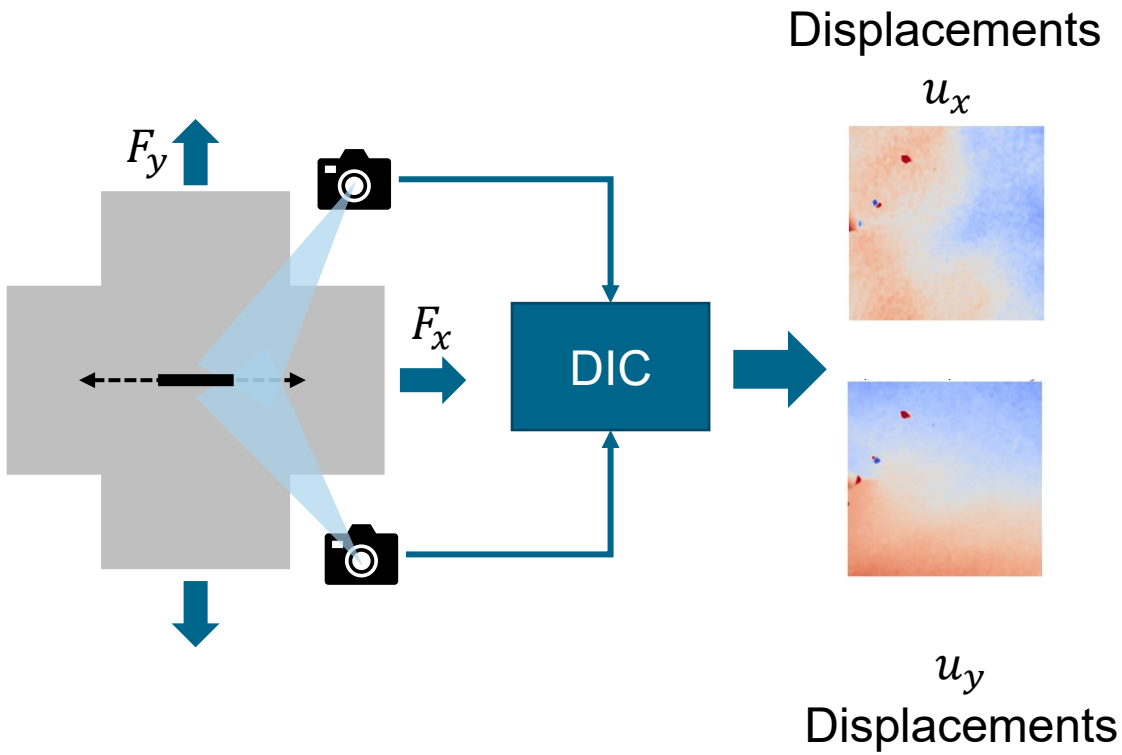
XAI Zoo and objective metrics

Part 3:

Revised Attention-guided training loop

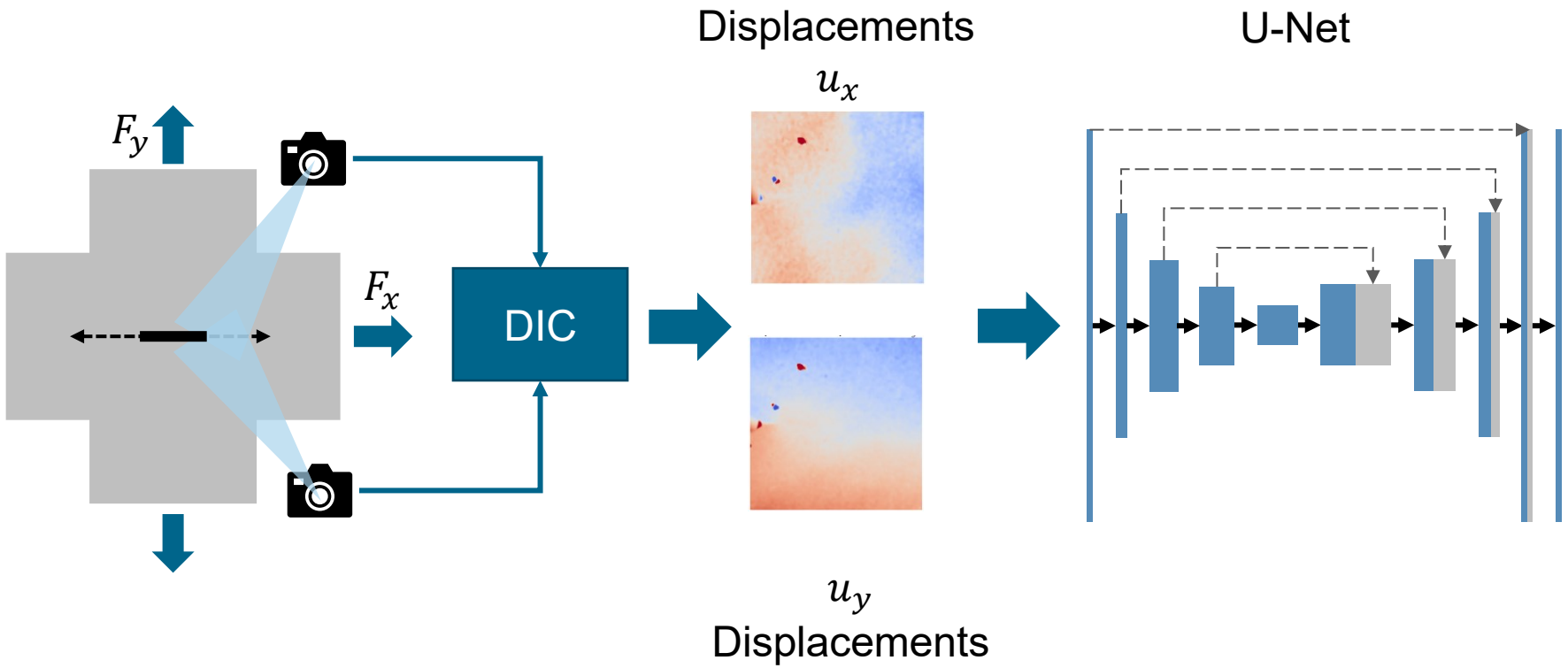
Towards Trustworthy AI in Materials Mechanics through Domain-Guided Attention
Talies, J., Breitbarth, E., & Melching, D. (2025) - arXiv preprint

Part 1 – Research task, data, and model



Fatigue Crack Growth Experiments → Digital Image Correlation

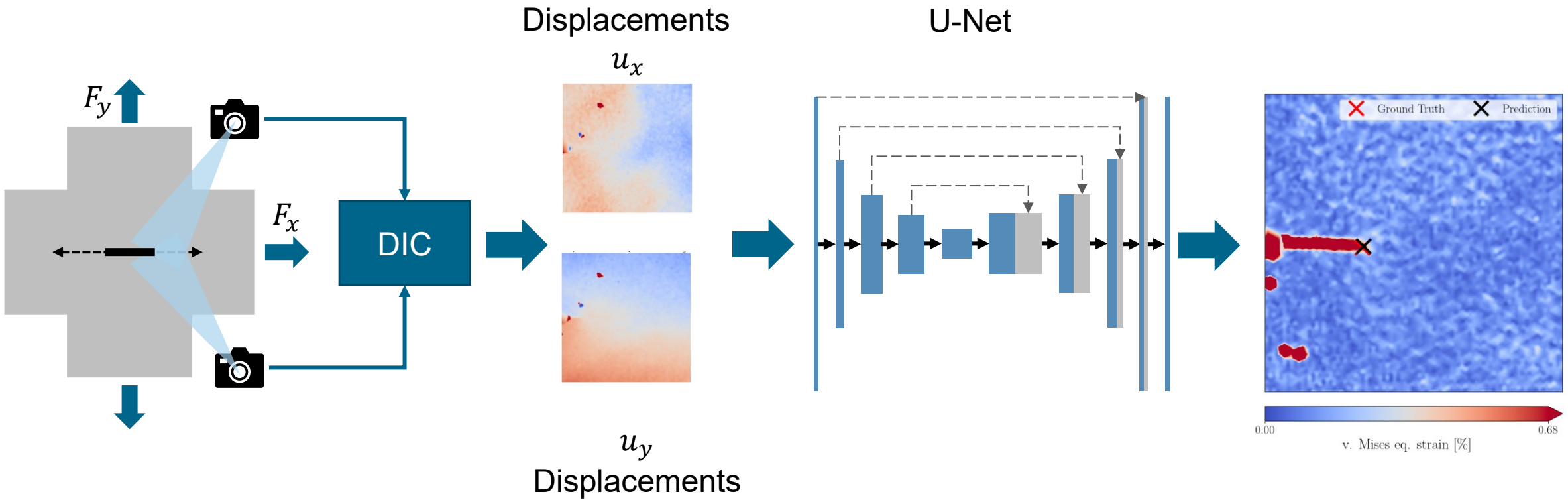
Part 1 – Research task, data, and model



Fatigue Crack Growth Experiments → Digital Image Correlation → Crack Tip Segmentation

O. Ronneberger et al. (2015) *MICCAI*

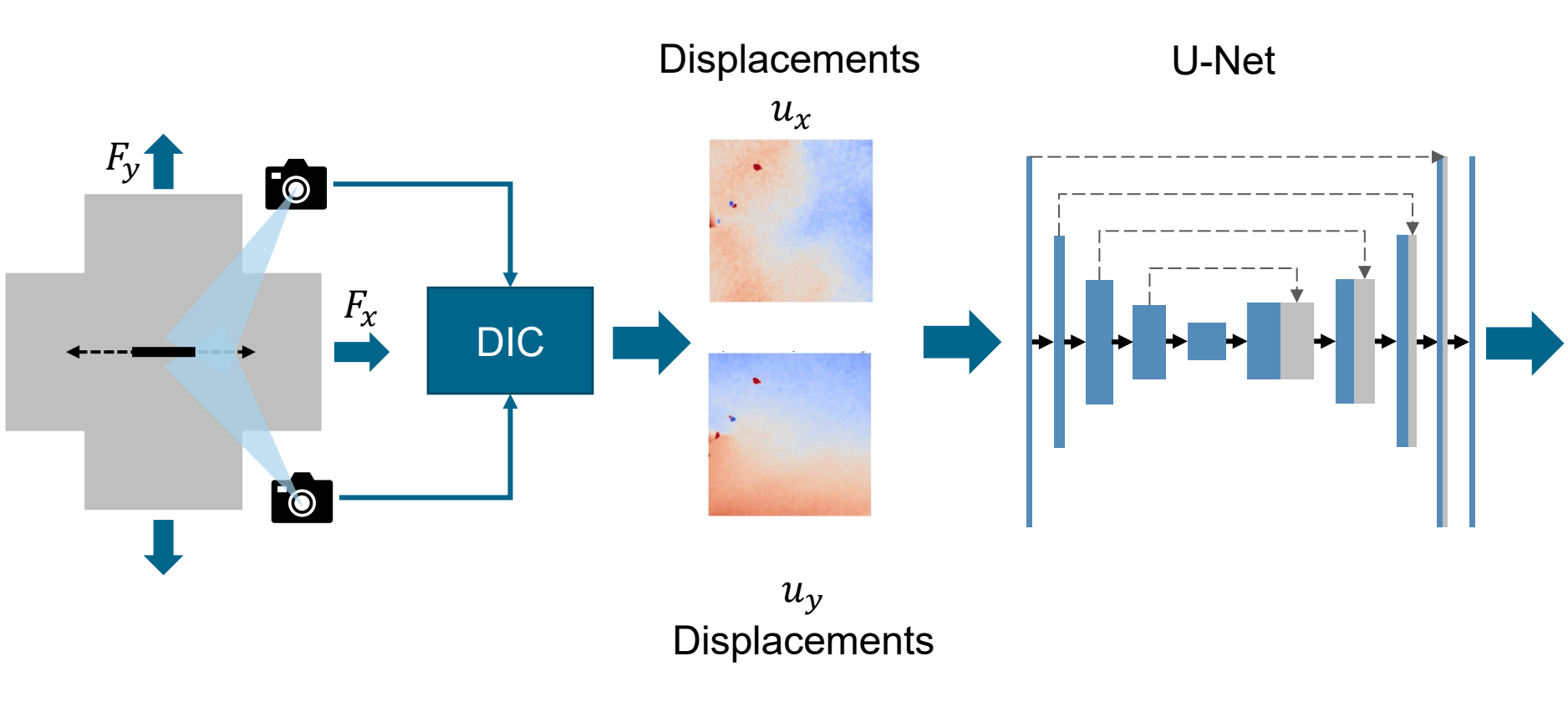
Part 1 – Research task, data, and model



Fatigue Crack Growth Experiments → Digital Image Correlation → Crack Tip Segmentation → Analysis, Feedback, etc.

O. Ronneberger et al. (2015) *MICCAI*

Part 1 – Research task, data, and model



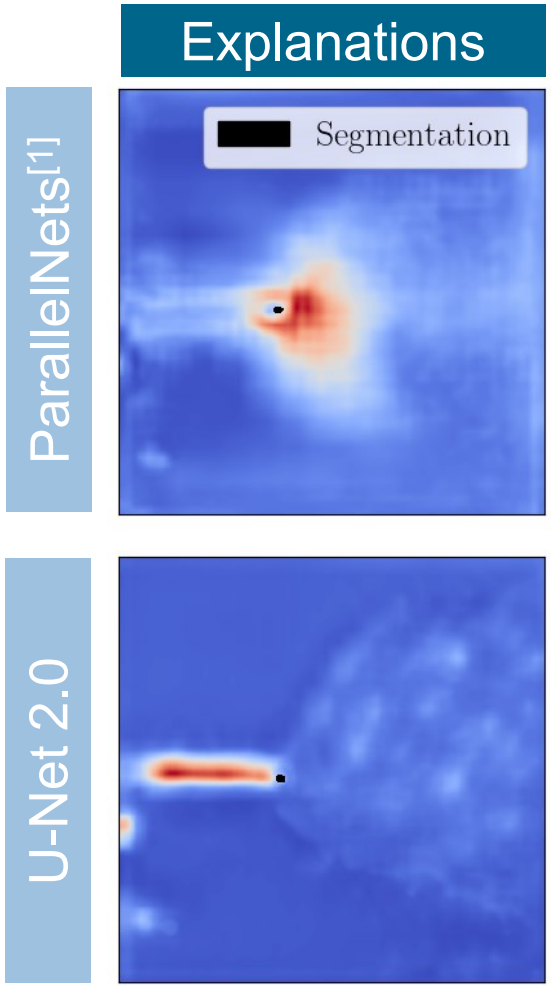
Fatigue Crack Growth Experiments → Digital Image Correlation → Crack Tip Segmentation → Analysis, Feedback, etc.

Improve

Supplement

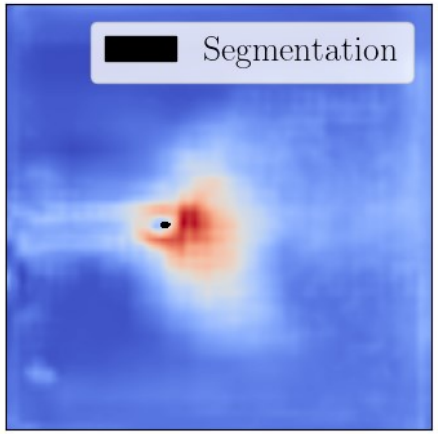
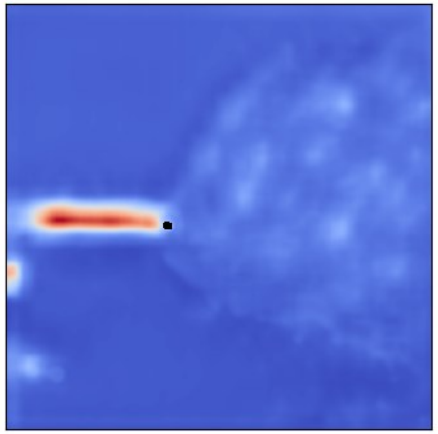
O. Ronneberger et al. (2015) *MICCAI*

Part 1 – Domain knowledge



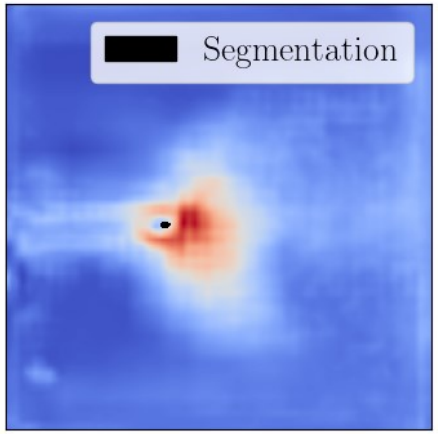
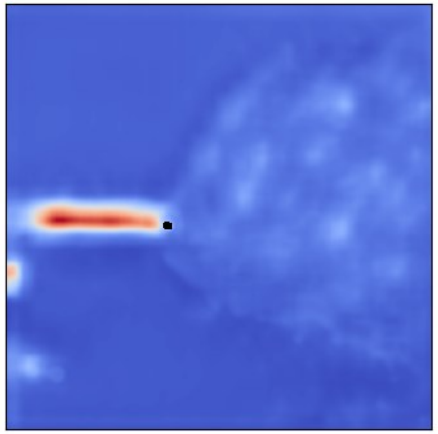
[1] D. Melching et al. (2022) *Scientific Reports*
[2] Strohmann et al. (2022) – CrackPy
[3] T. L. Anderson, Fracture mechanics

Part 1 – Domain knowledge

	Explanations	Model Benchmarks[1]
ParalleINets[1]		Validation Dice 0.33 Fraction of valid predictions 0.91
U-Net 2.0		Validation Dice 0.31 Fraction of valid predictions 0.90

[1] D. Melching et al. (2022) *Scientific Reports*
[2] Strohmann et al. (2022) – CrackPy
[3] T. L. Anderson, Fracture mechanics

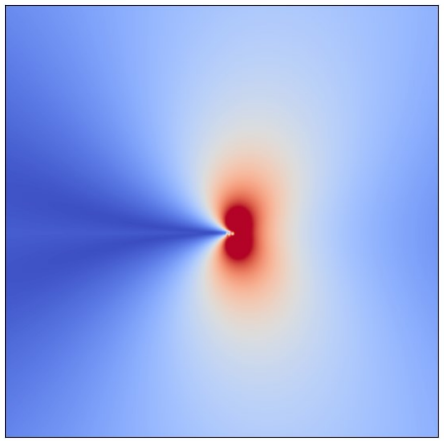
Part 1 – Domain knowledge

	Explanations	Model Benchmarks[1]
ParalleINets[1]		Validation Dice 0.33 Fraction of valid predictions 0.91
U-Net 2.0		Validation Dice 0.31 Fraction of valid predictions 0.90

Theoretical Knowledge[2][3]

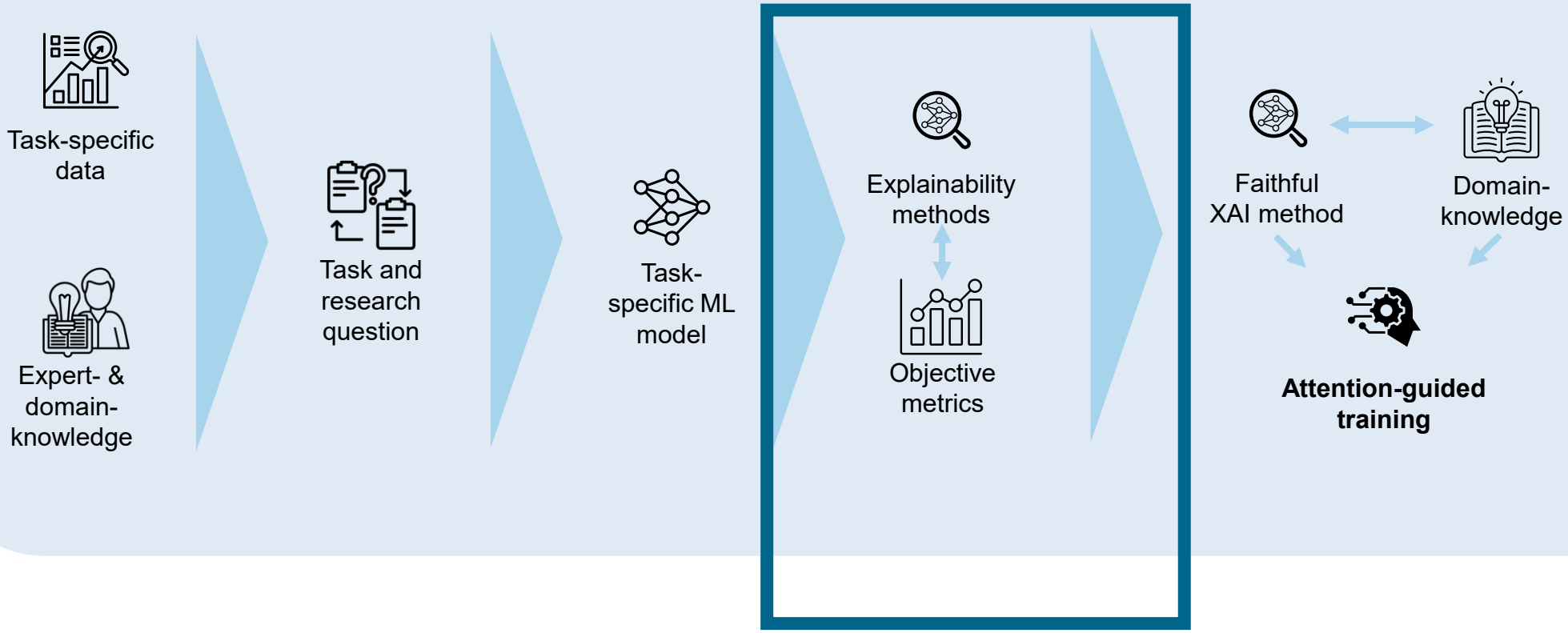
Williams Series

$$\sigma_{ij} = \frac{k}{\sqrt{r}} f_{ij}(\theta) + \sum_{m=0}^{\infty} A_m r^{\frac{m}{2}} g_{ij}^{(m)}(\theta)$$



[1] D. Melching et al. (2022) *Scientific Reports*
[2] Strohmann et al. (2022) – CrackPy
[3] T. L. Anderson, Fracture mechanics

Attention-guided training framework



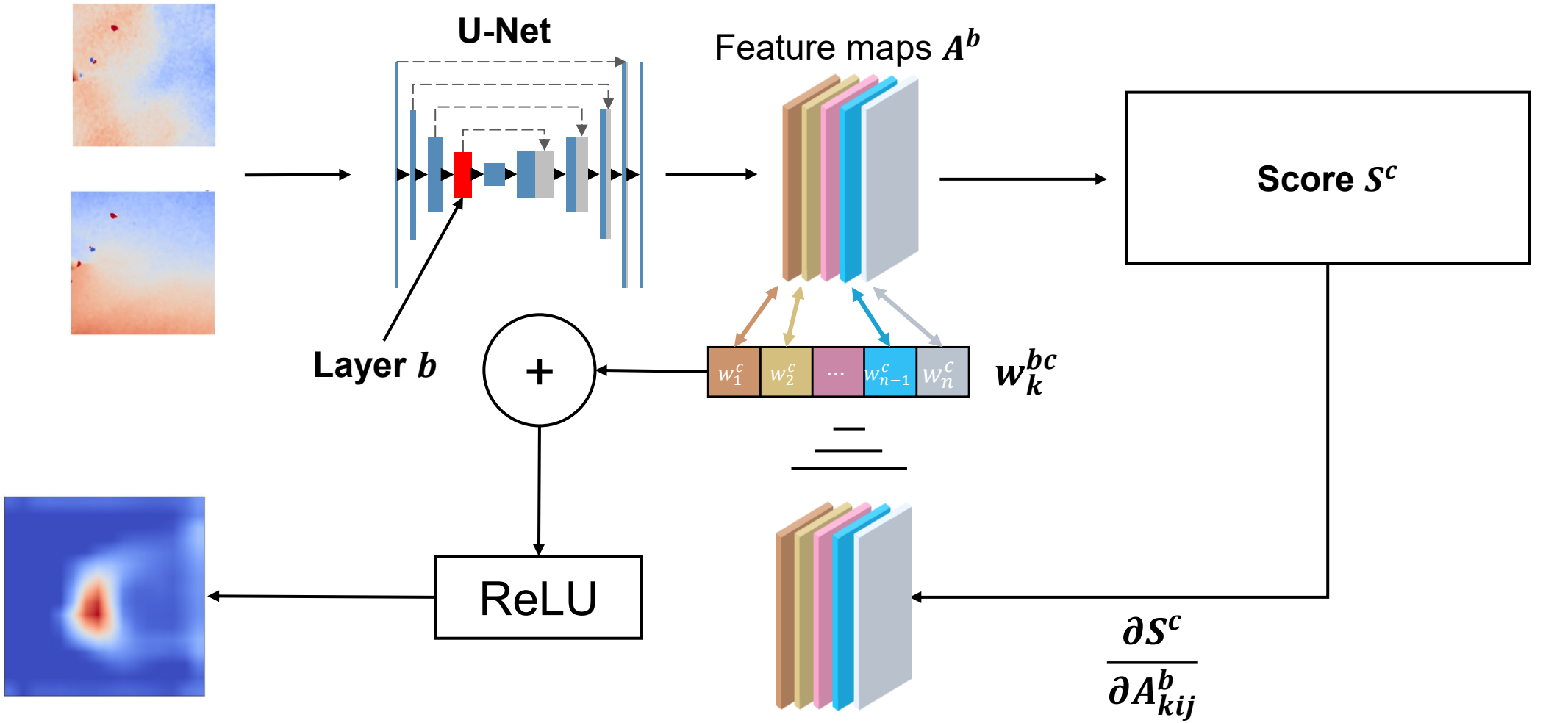
Part 1:
Task and data and model choice

Domain knowledge

Part 2:
XAI Zoo and objective metrics

Part 3:
Revised Attention-guided training loop

Part 2 – Objective choice of XAI approach
Grad-CAM

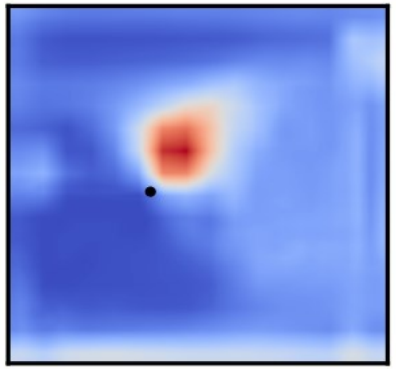


modified from R. R. Selvaraju et al. (2016) *international Journal of Computer Vision*

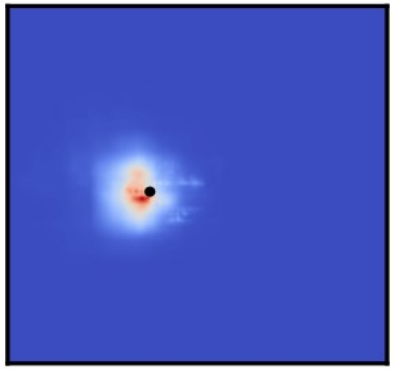
Part 2 – Objective choice of X-AI approach

A subset of the class activation mapping zoo

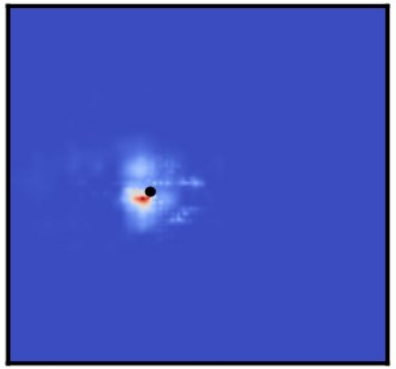
Grad-CAM



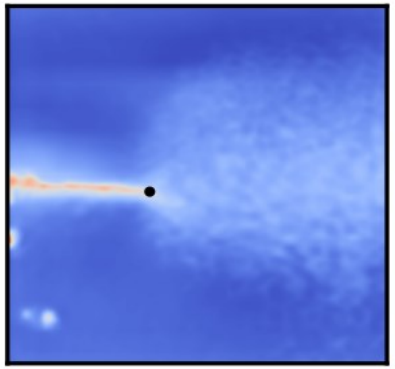
Layer-CAM



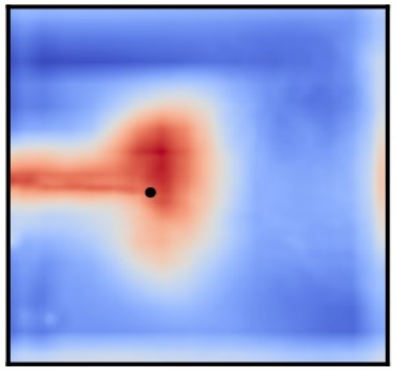
Grad-CAM-EW



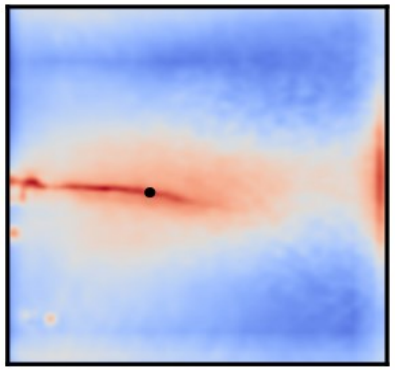
Grad-CAM++



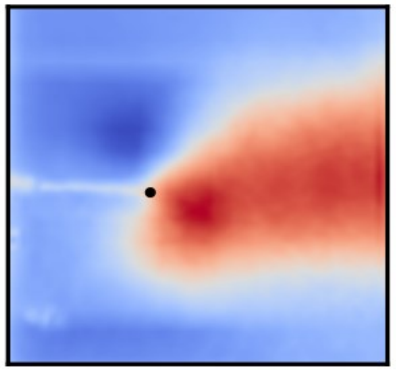
Grad-CAM^{+}



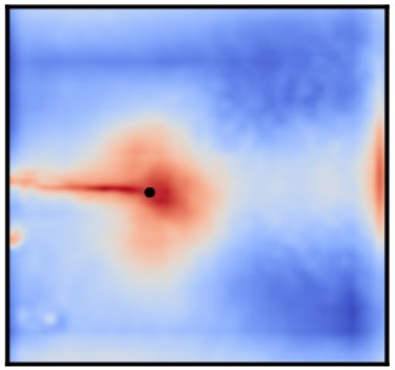
Score-CAM



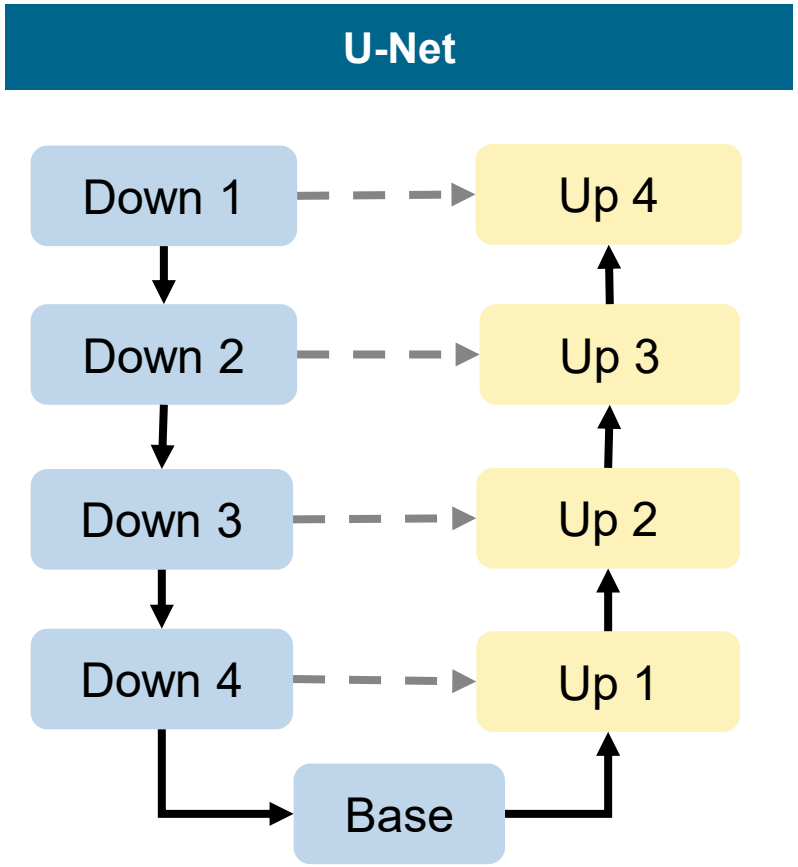
Eigen-CAM

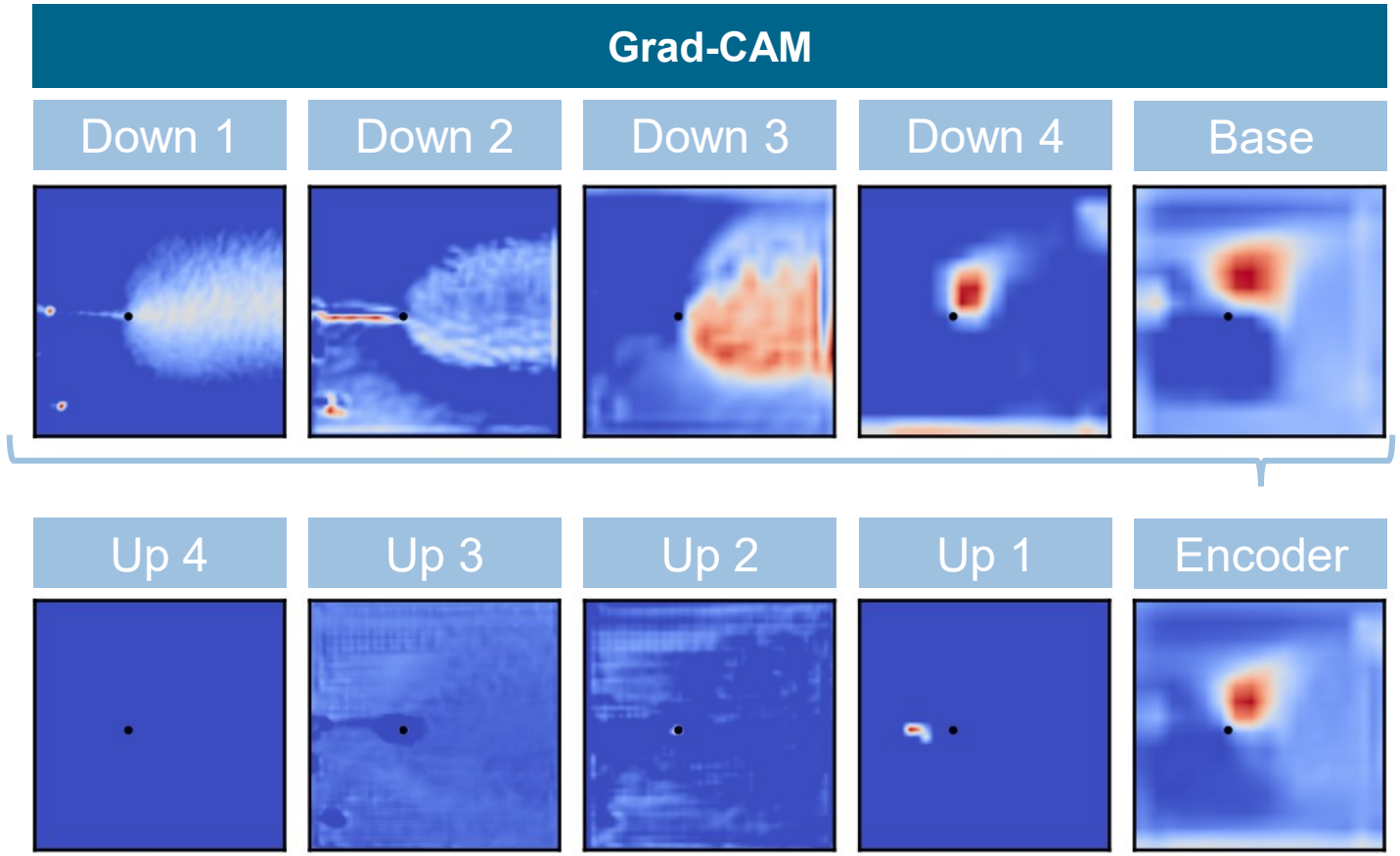


Ablation-CAM



Part 2 – Objective choice of XAI approach





Part 2 – Objective choice of XAI approach

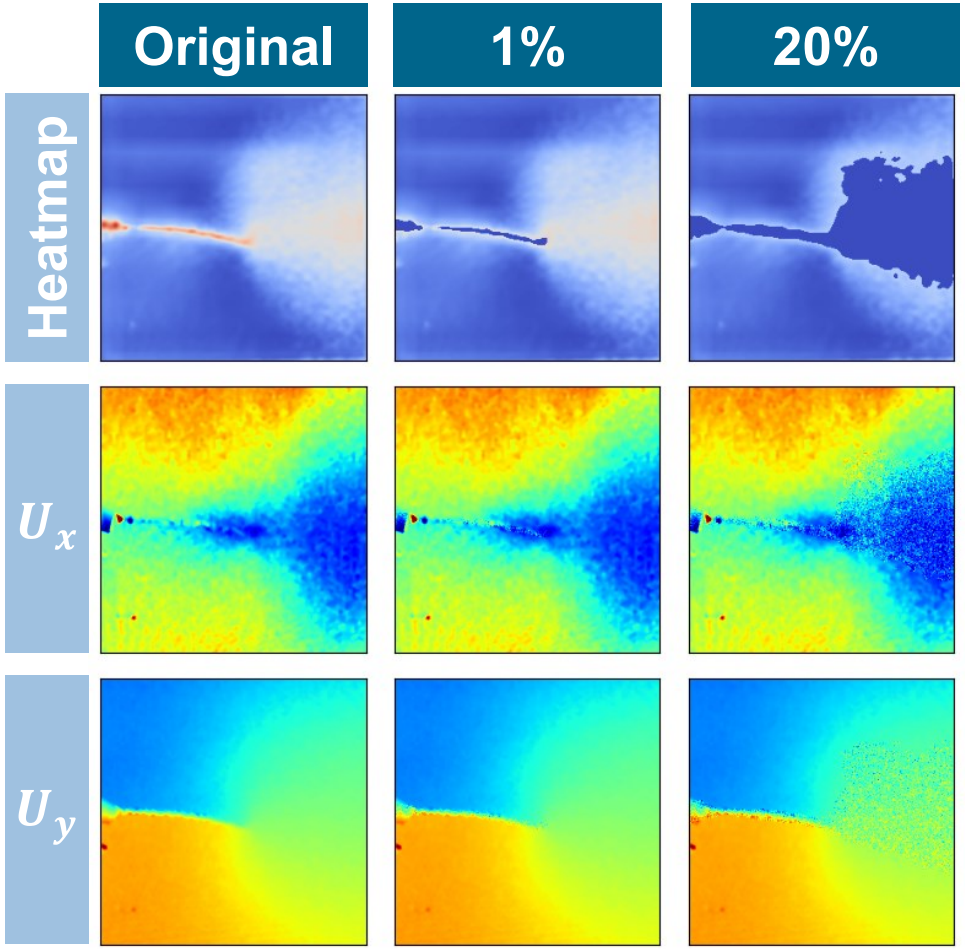
Co12 criteria

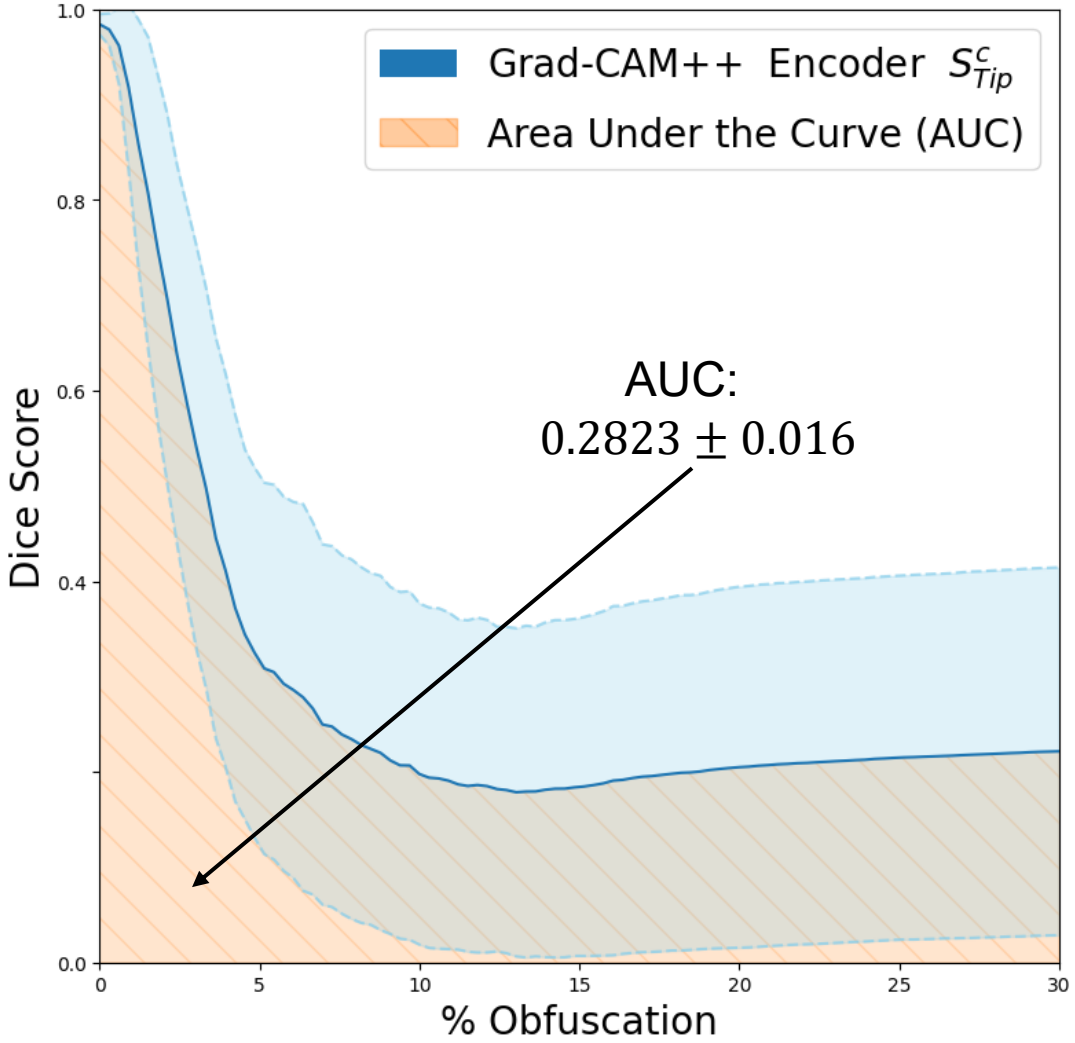
Correctness How faithful is my explanation?	1	**Compactness** Is the explanation concise?	7
Completeness Does it cover all relevant parts?	2	**Composition** How well is the result presented?	8
Consistency Do identical predictions have identical explanations?	3	**Confidence** How big is my uncertainty?	9
Continuity Is the explanation stable towards small changes?	4	**Context** How well does it suit the task and topic?	10
Contrastivity Does it reflect the contrast between classes?	5	**Coherence** Does it align with domain knowledge?	11
Covariate Complexity How difficult is the result to understand?	6	**Controllability** How well the explanation be adjusted on demand?	12

M. Nauta et al. (2023) *ACM Comput. Surv.*

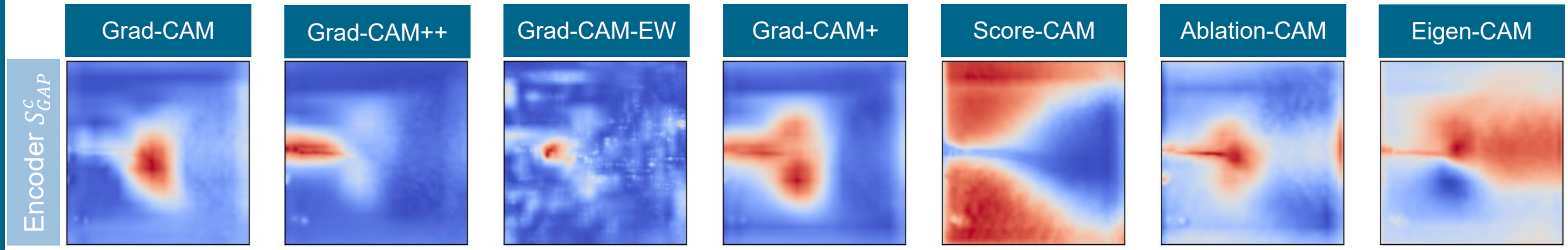
Part 2 – Objective choice of XAI approach

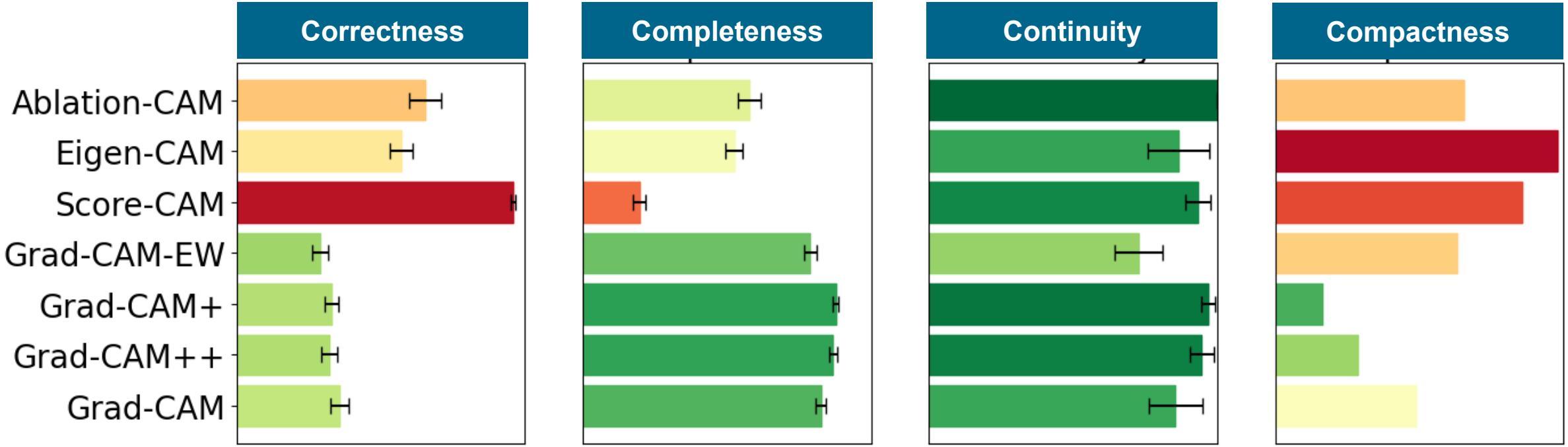
Co1 - correctness





Part 2 – Objective choice of XAI approach





Attention-guided training framework

Part 1:
Task and data and model choice

Domain knowledge

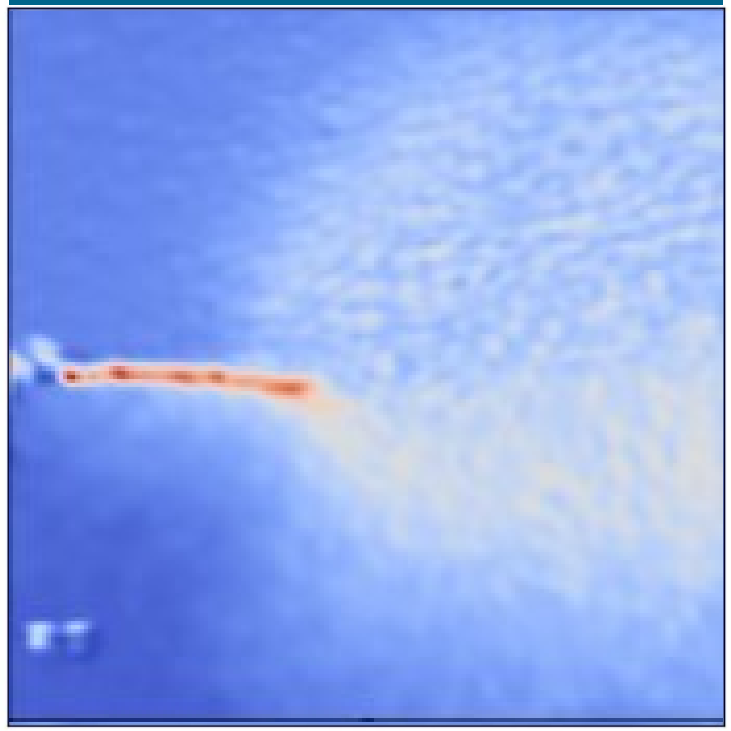
Part 2:
XAI Zoo and objective metrics

Part 3:
Revised Attention-guided training loop

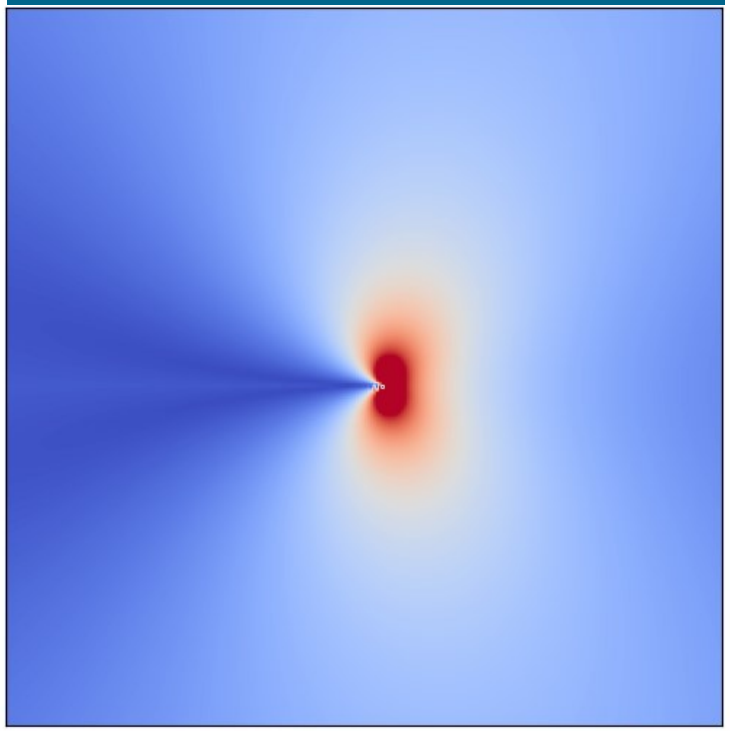
Part 3 – Attention-Guided Training

Co11 - Coherence

Grad-CAM++



Williams field[1]

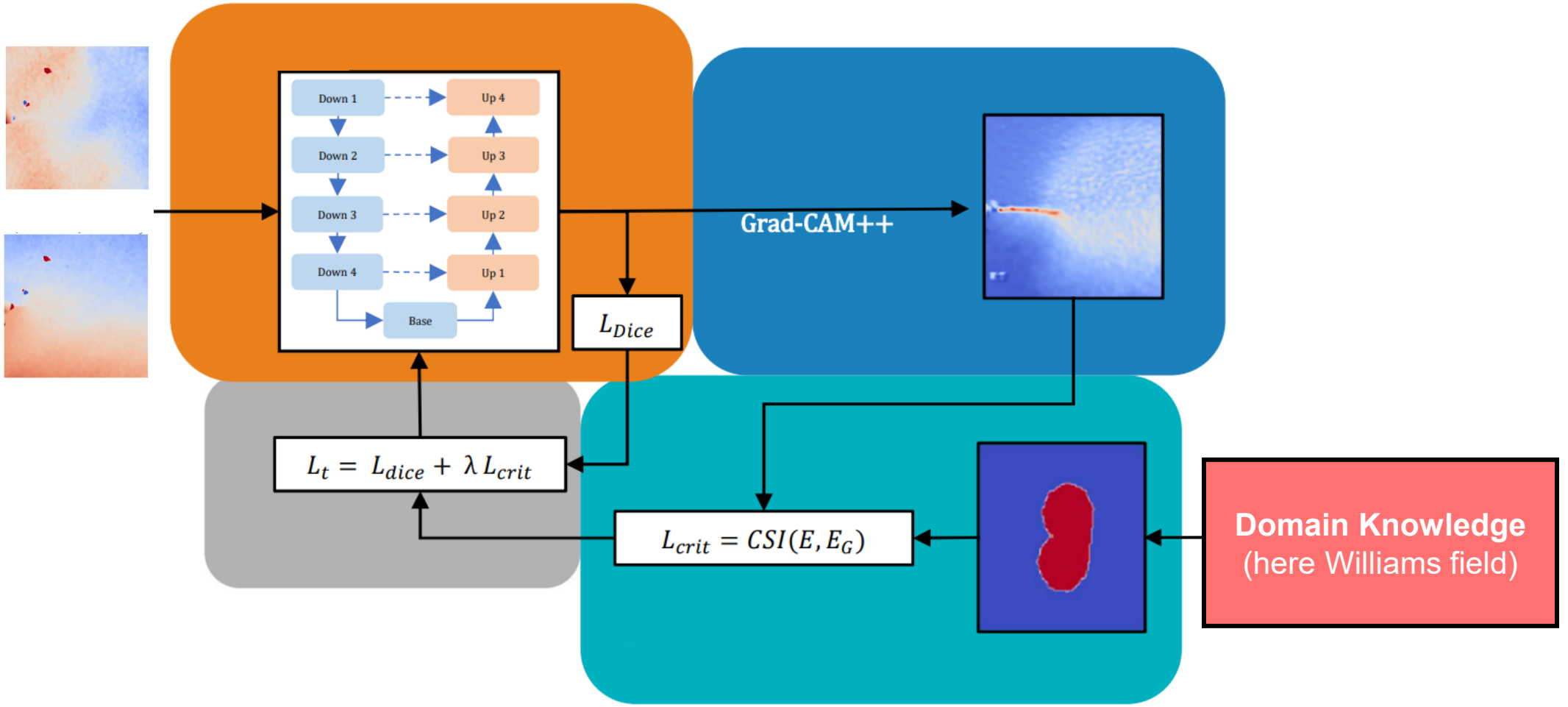


Similarity

- Structural similarity[2]
- Cosine Similarity[3]
- Mean squared error
- …

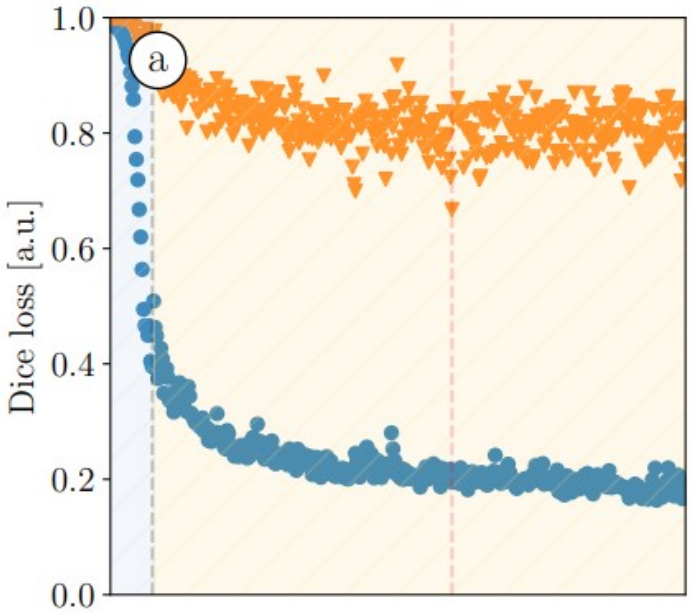
[1] Strohmann et al. (2022) – CrackPy
[2] Z. Wang et al. (2004) *IEEE transactions on image processing*
[3] G. Salton et al. (1983) "Introduction to modern information retrieval"

Part 3 – Attention-Guided Training

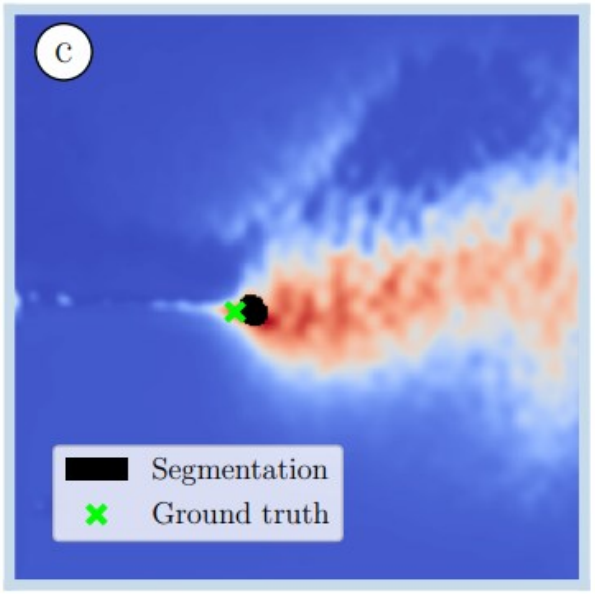


$$L_{total} = \mathrm{Dice}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + \lambda\, \mathrm{CSI}(\boldsymbol{\Phi}(\boldsymbol{y}), \hat{\boldsymbol{\Phi}}(\hat{y}))$$

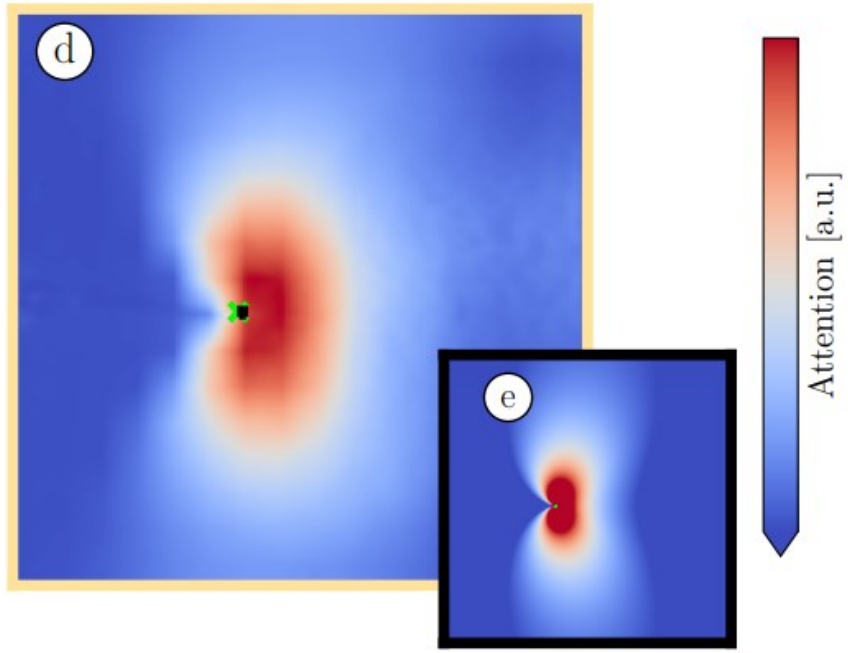
Part 3 – Attention-Guided Training



Conventional training



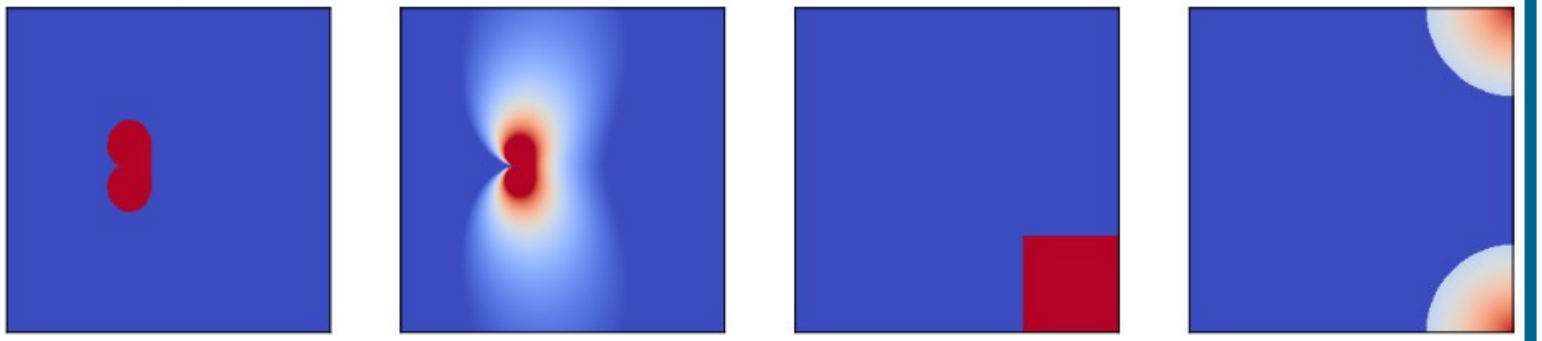
After AGT



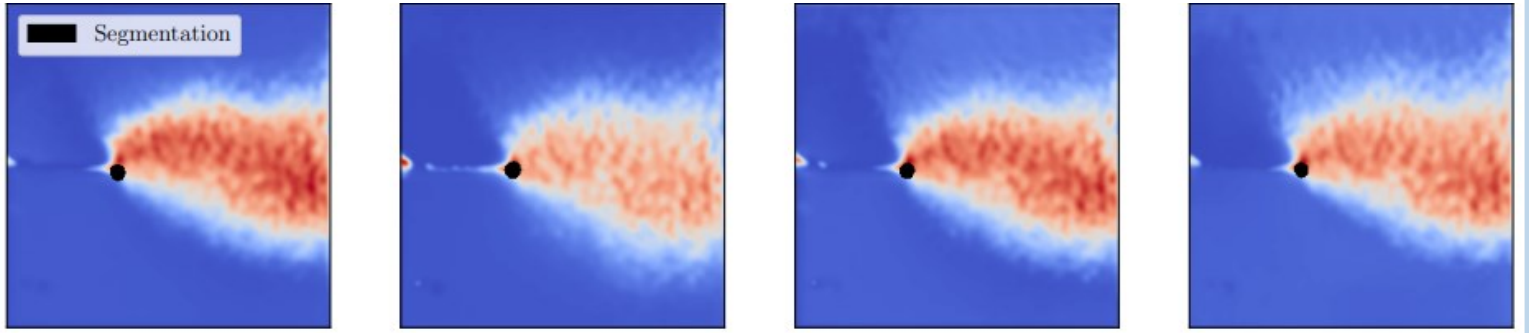
Towards Trustworthy AI in Materials Mechanics through Domain-Guided Attention
Talies, J., Breitbarth, E., & Melching, D. (2025) - arXiv preprint

Part 3 – Attention-Guided Training

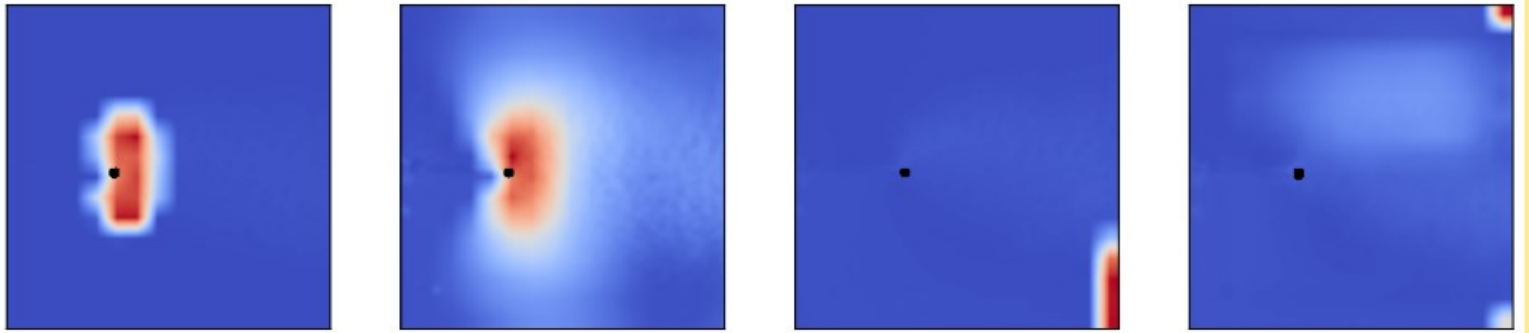
Varying physical and non-physical domain priors



Consistently non-physical explanations during regular training



Almost perfect alignment for physical priors and strong conceptual alignment for non-physical priors

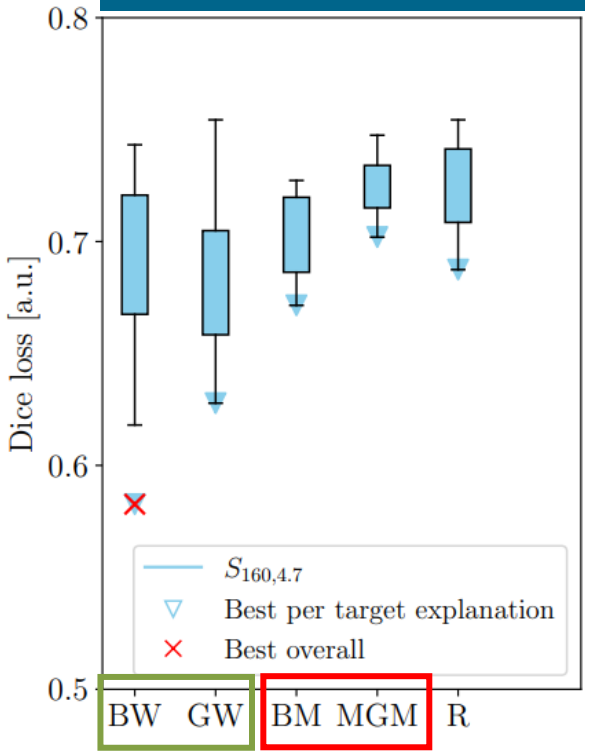


Part 3 – Attention-Guided Training

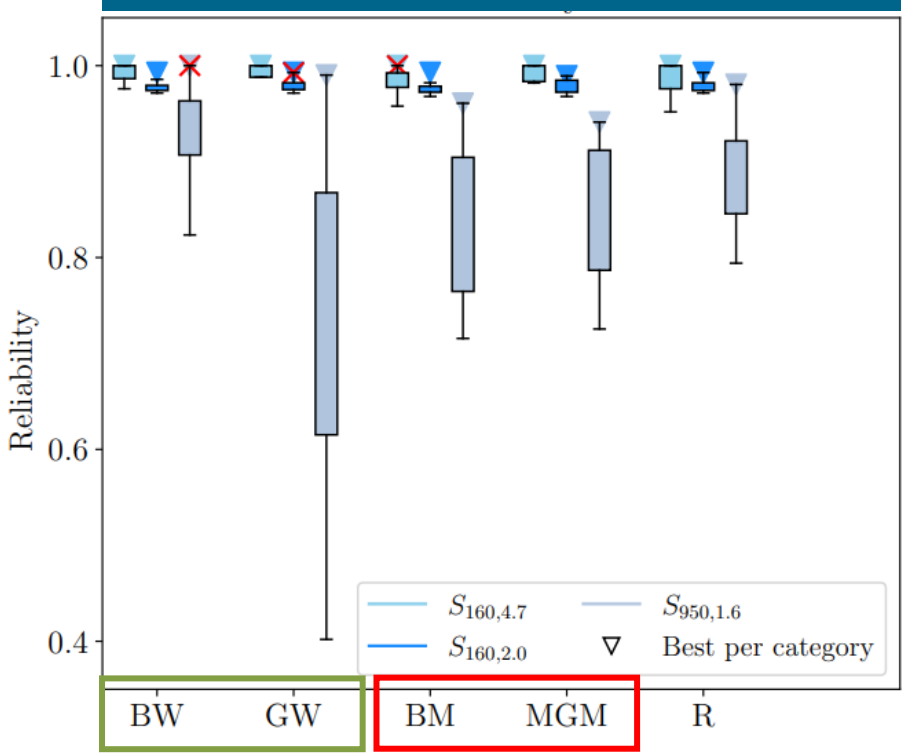
Physical explanation targets

Misleading explanation targets

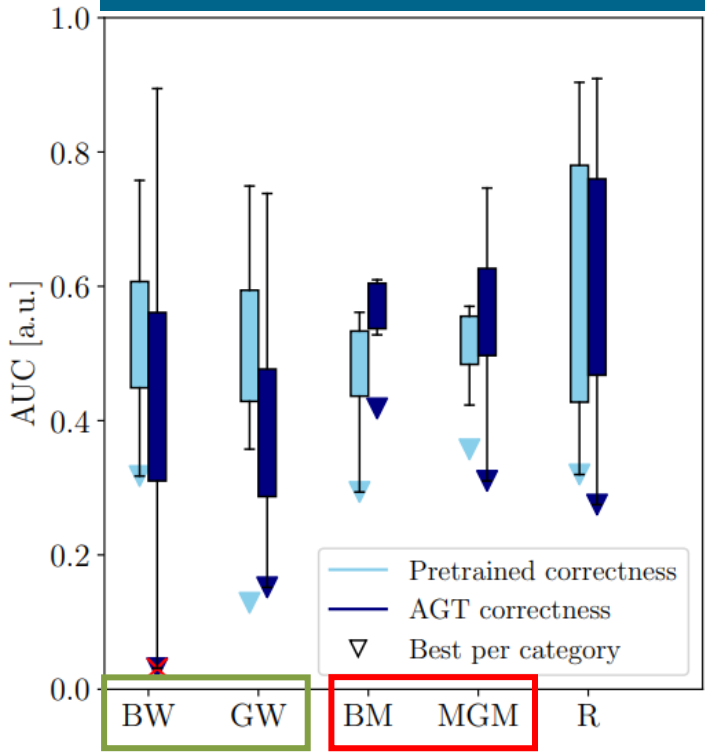
Validation Loss



Fraction of valid predictions



Explanation correctness



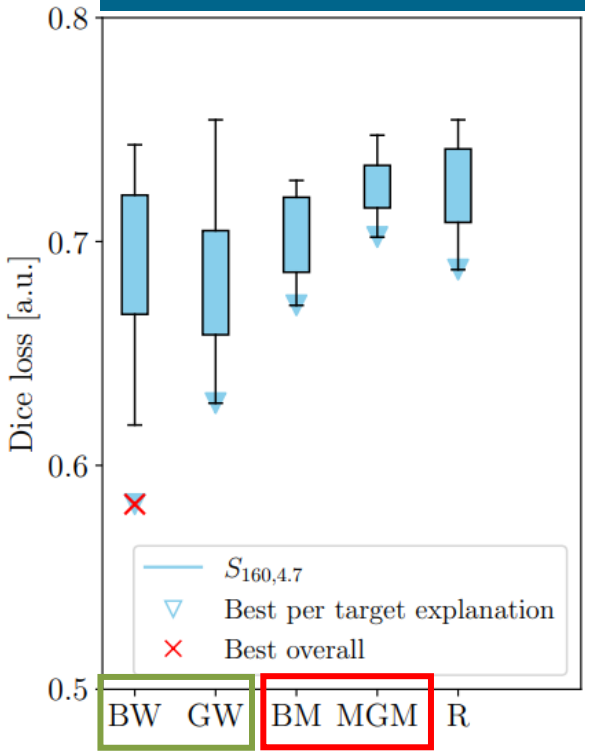
Towards Trustworthy AI in Materials Mechanics through Domain-Guided Attention
Talies, J., Breitbarth, E., & Melching, D. (2025) - arXiv preprint

Part 3 – Attention-Guided Training

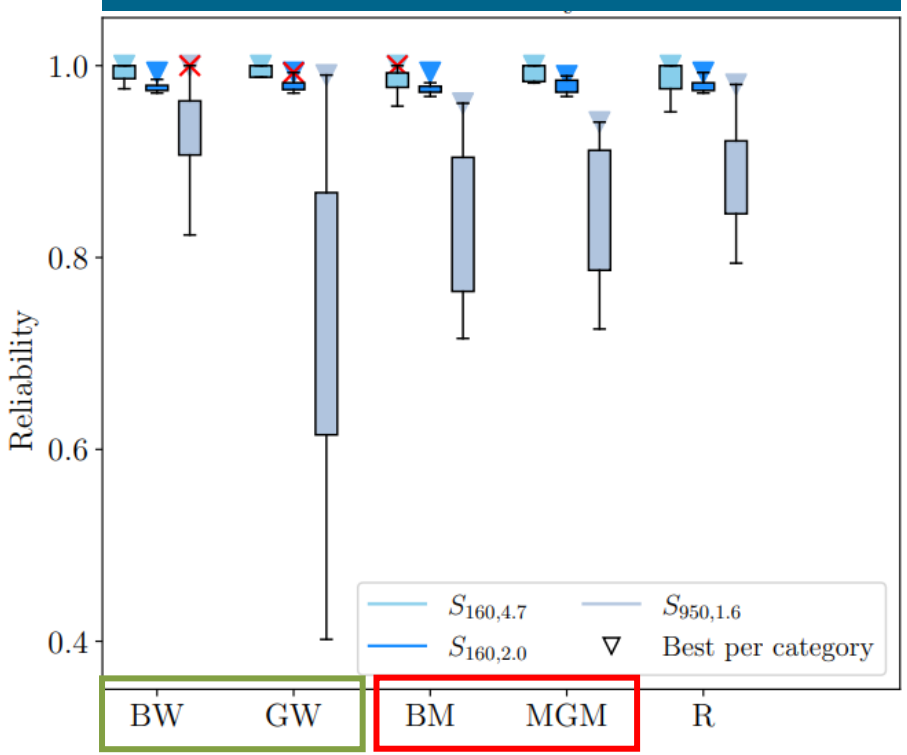
Physical explanation targets

Misleading explanation targets

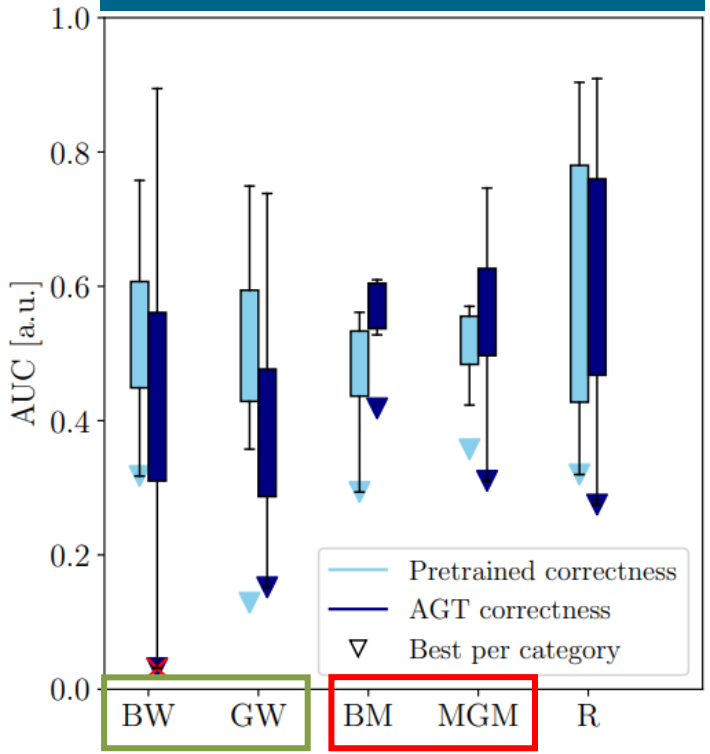
Validation Loss



Fraction of valid predictions



Explanation correctness

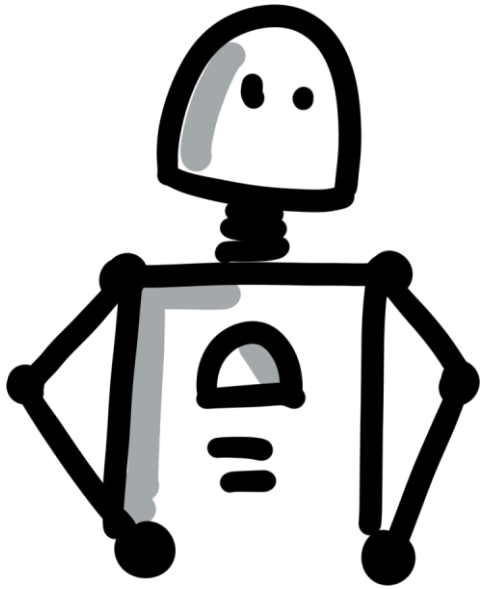


Clear correlation between choice of domain prior and model performance.

Simultaneously no evidence of the model "gaming" the additional metric.

Towards Trustworthy AI in Materials Mechanics through Domain-Guided Attention
Talies, J., Breitbarth, E., & Melching, D. (2025) - arXiv preprint

THANK YOU VERY MUCH FOR YOUR ATTENTION!



GitHub