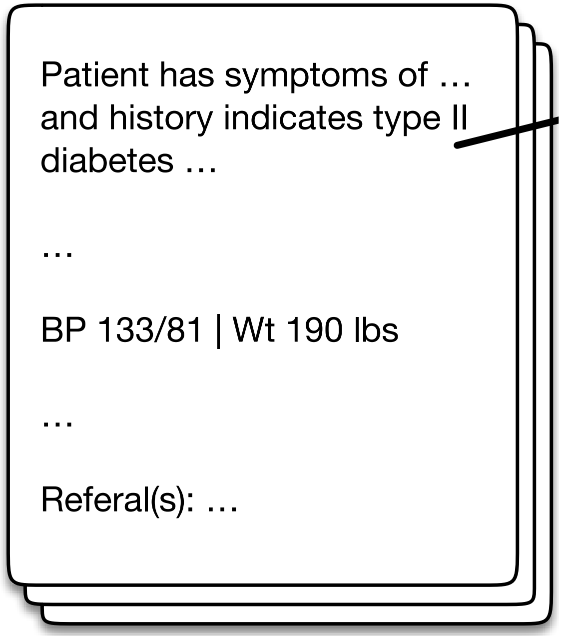
CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models

Denis Jered McInerney[1], Geoffrey Young[2], Jan-Willem van de Meent[3], Byron C. Wallace[1]
[1]Northeastern University, [2]Brigham and Women's Hospital, [3]University of Amsterdam

Presented by

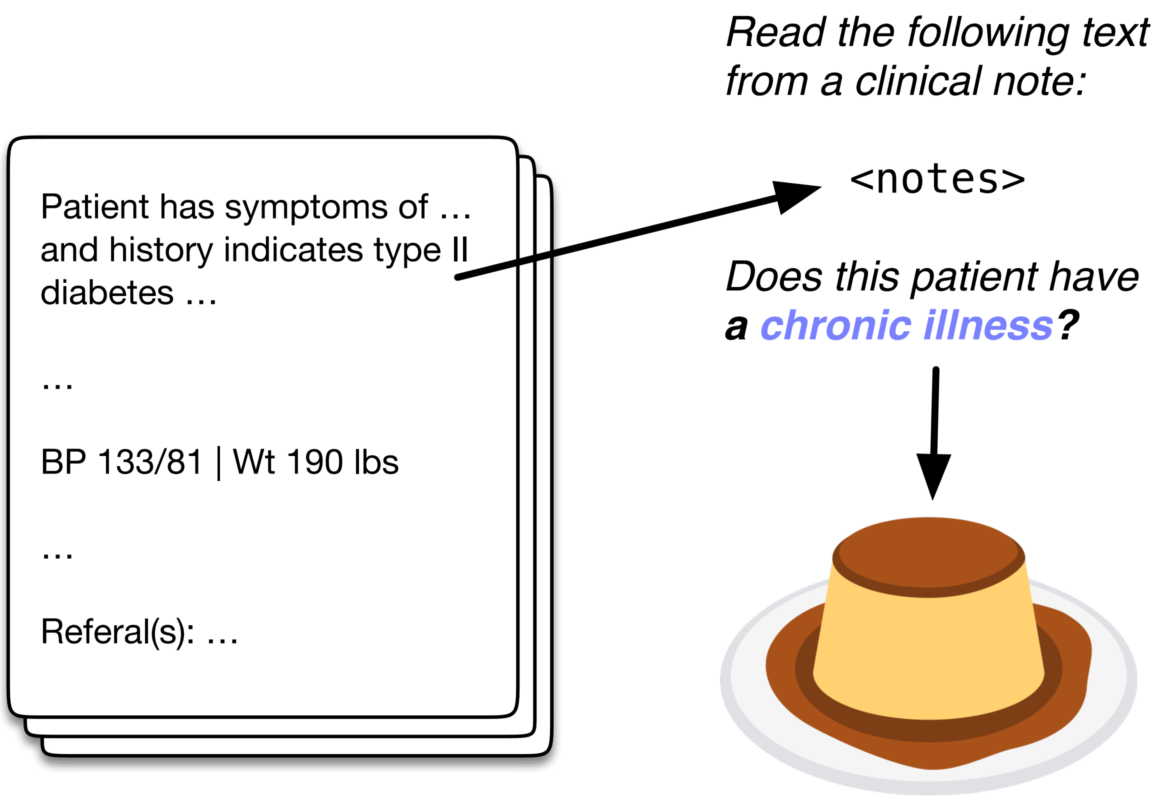
Denis Jered McInerney

Craft **Hi**gh-**L**evel **L**atents (**CHiLL**)



Raw patient notes

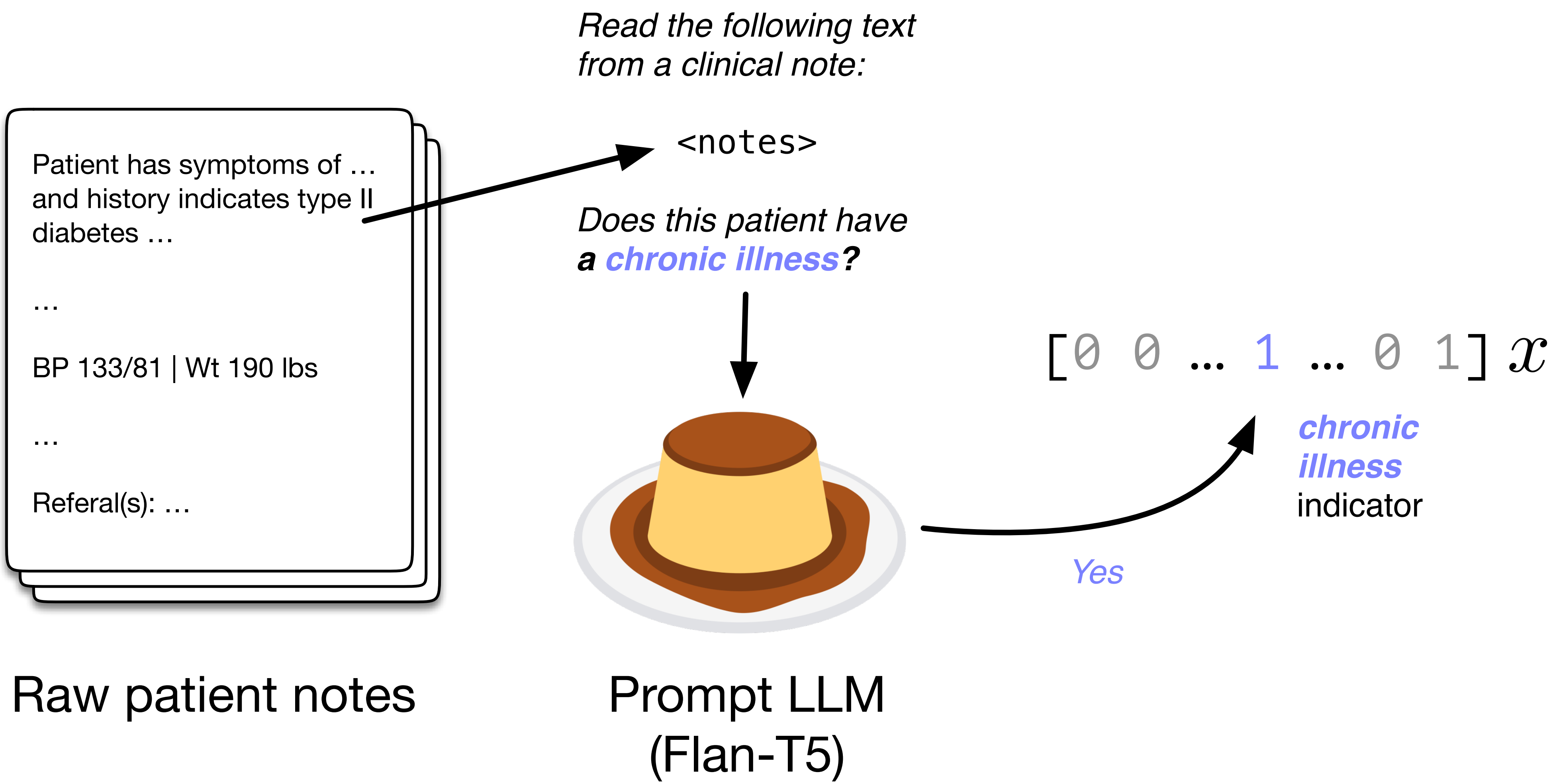
Craft **Hi**gh-**L**evel **L**atents (**CHiLL**)



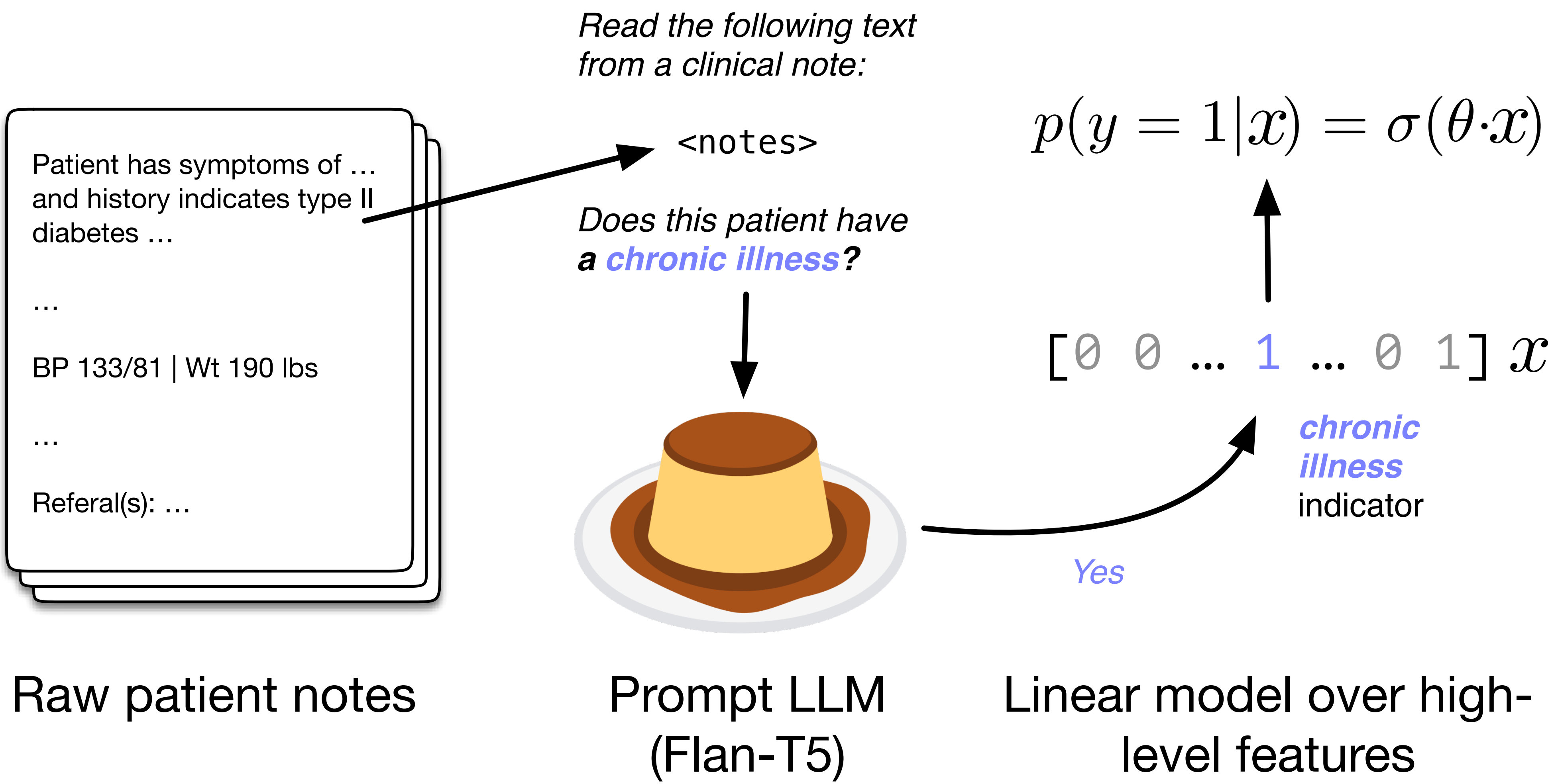
Raw patient notes

Prompt LLM
(Flan-T5)

Craft High-Level Latents (CHiLL)



Craft High-Level Latents (CHiLL)



Feature Extraction Performance

MIMIC-III Feature Extraction
(ICD Code Features)

For Example:
- Unspecified Essential Hypertension
- Congestive Heart Failure
- Atrial Fibrillation

…

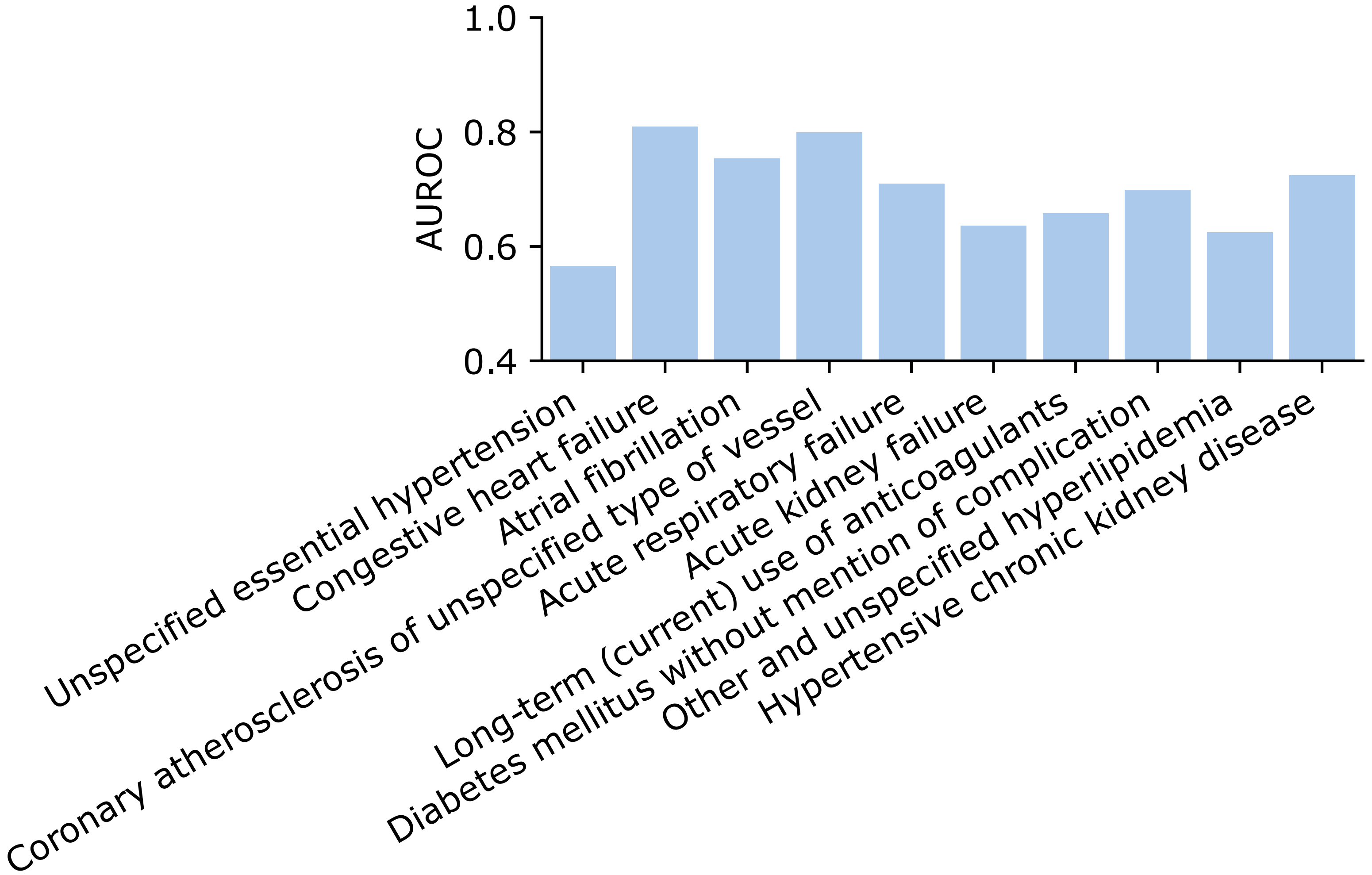
MIMIC-CXR Feature Extraction
(Expert-Crafted Features)

For Example:
- Bulging Fissures
- Decreased Lung Volumes
- Collapse of Lung

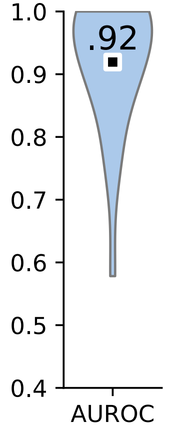
…

Feature Extraction Performance

MIMIC-III Feature Extraction
(ICD Code Features)

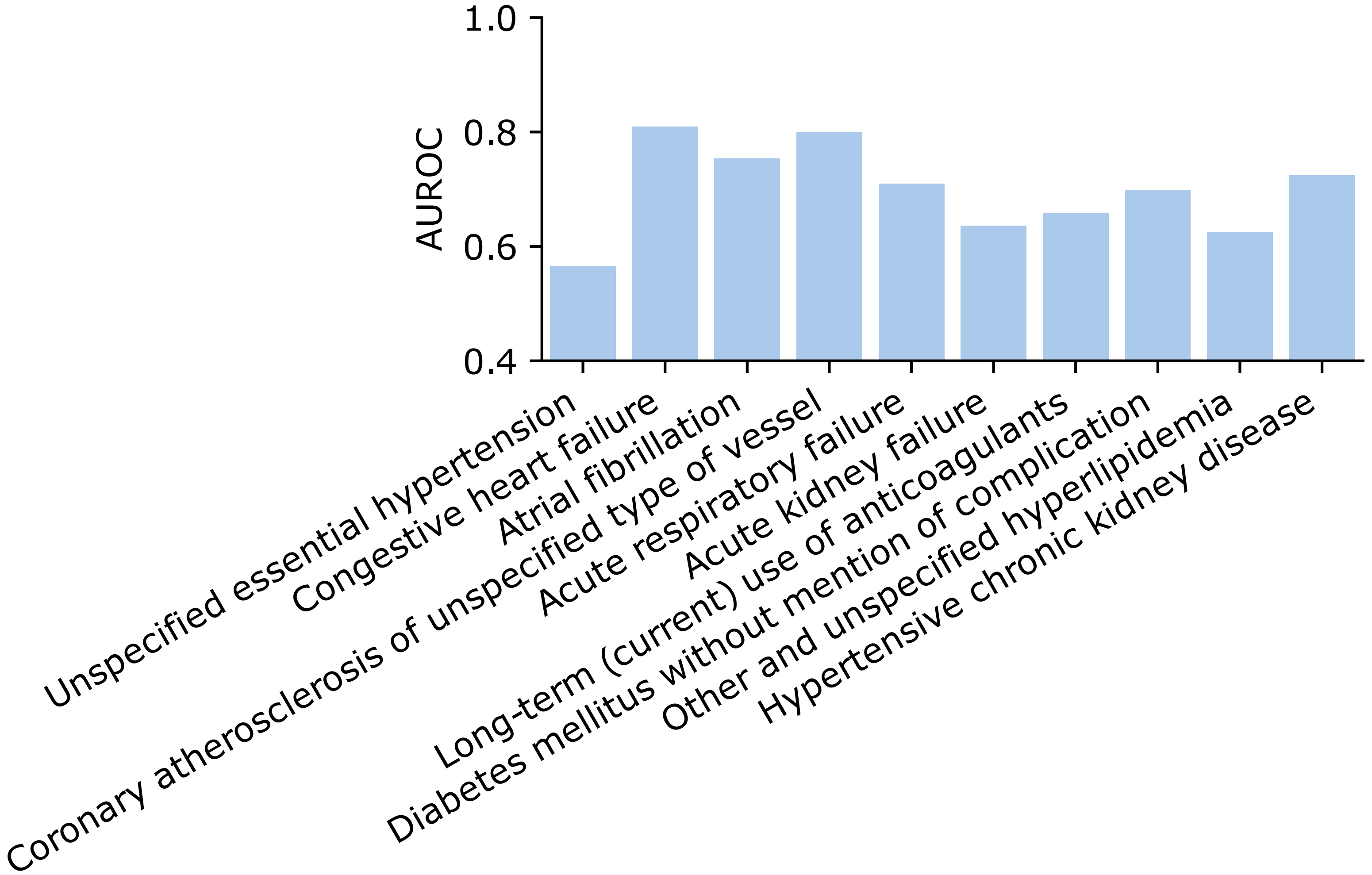


MIMIC-CXR Feature Extraction
(Expert-Crafted Features)

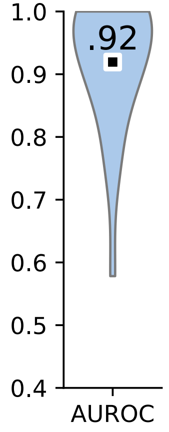


Feature Extraction Performance

MIMIC-III Feature Extraction
(ICD Code Features)

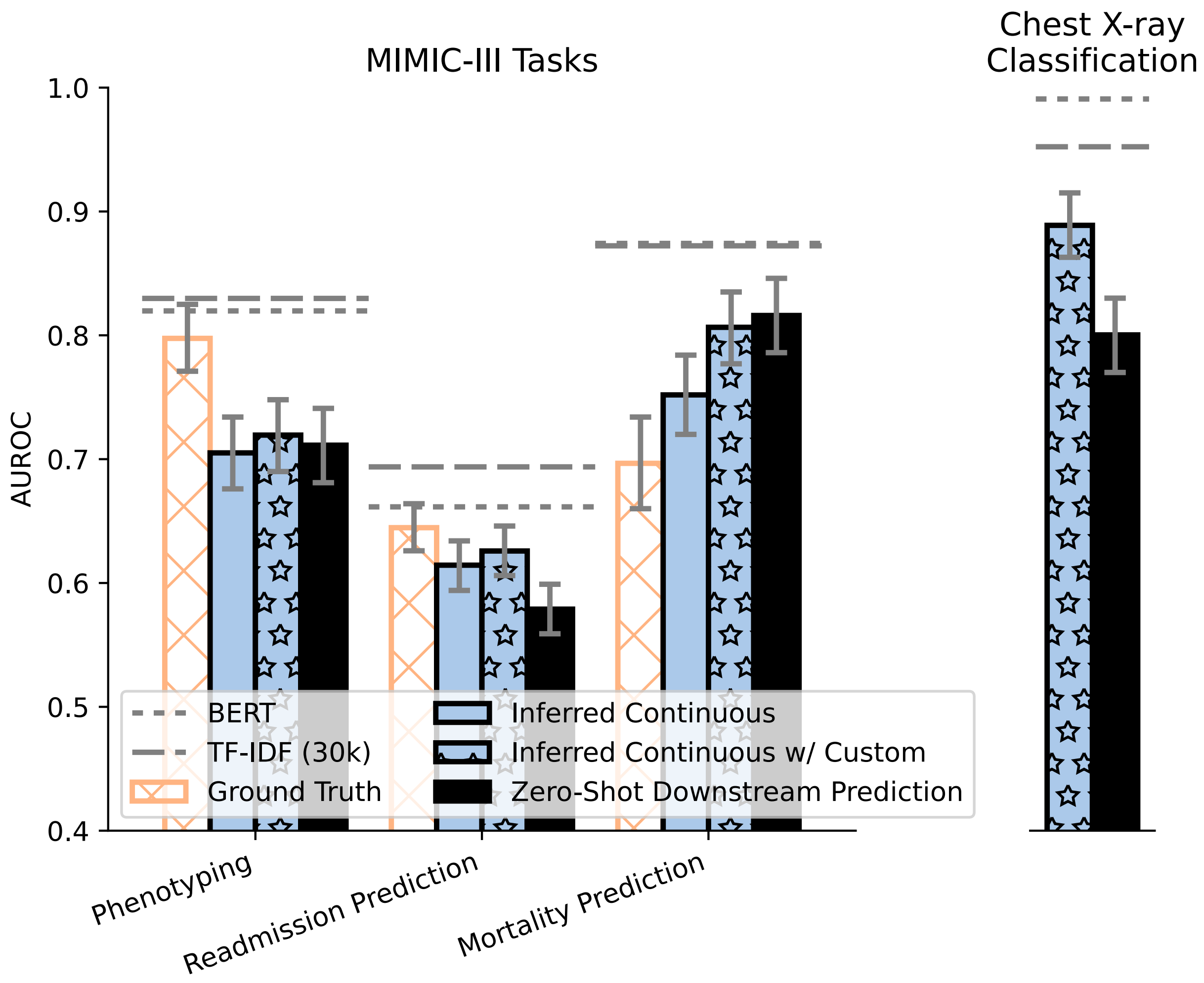


MIMIC-CXR Feature Extraction
(Expert-Crafted Features)

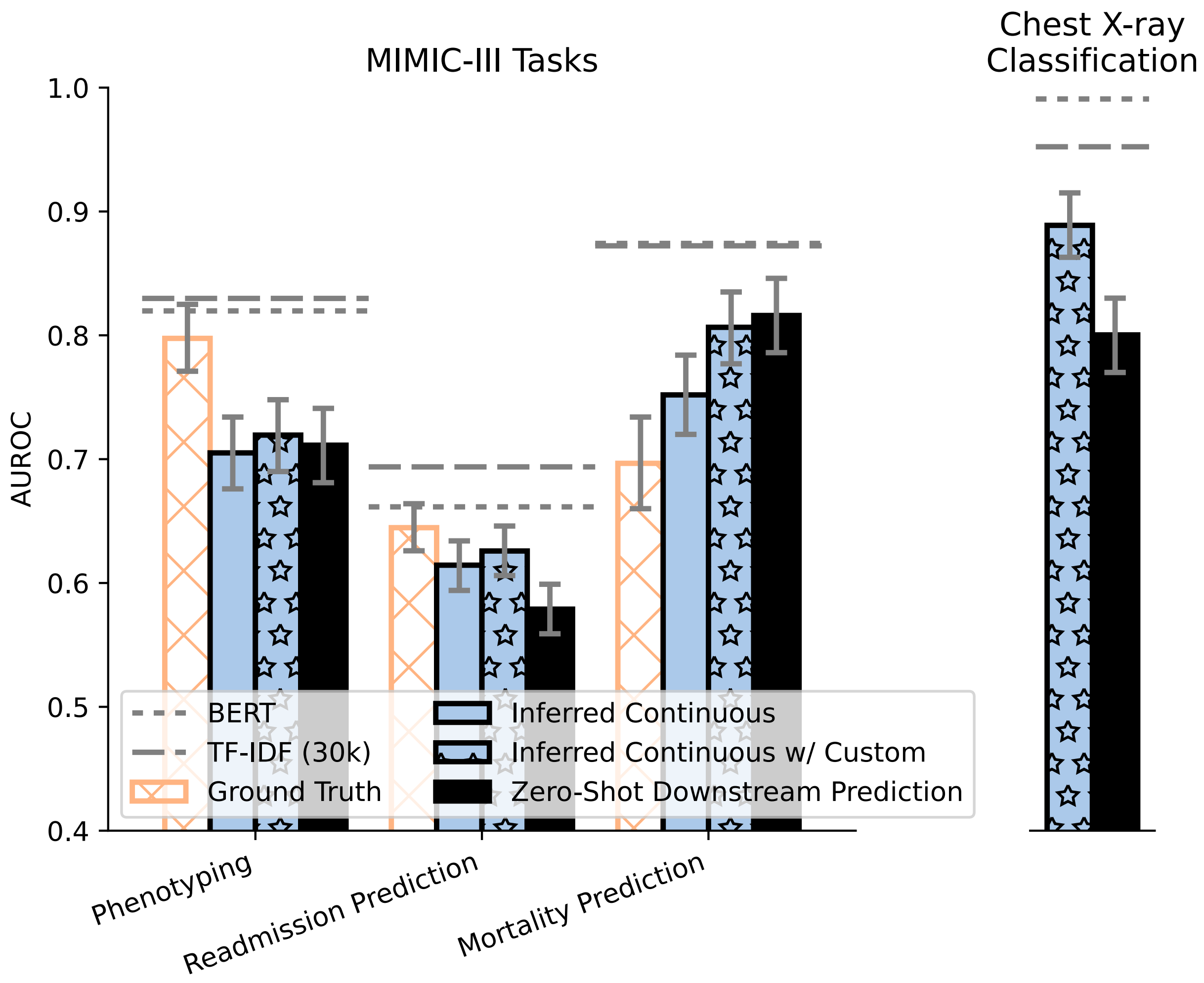


Reasonable!

Downstream Performance



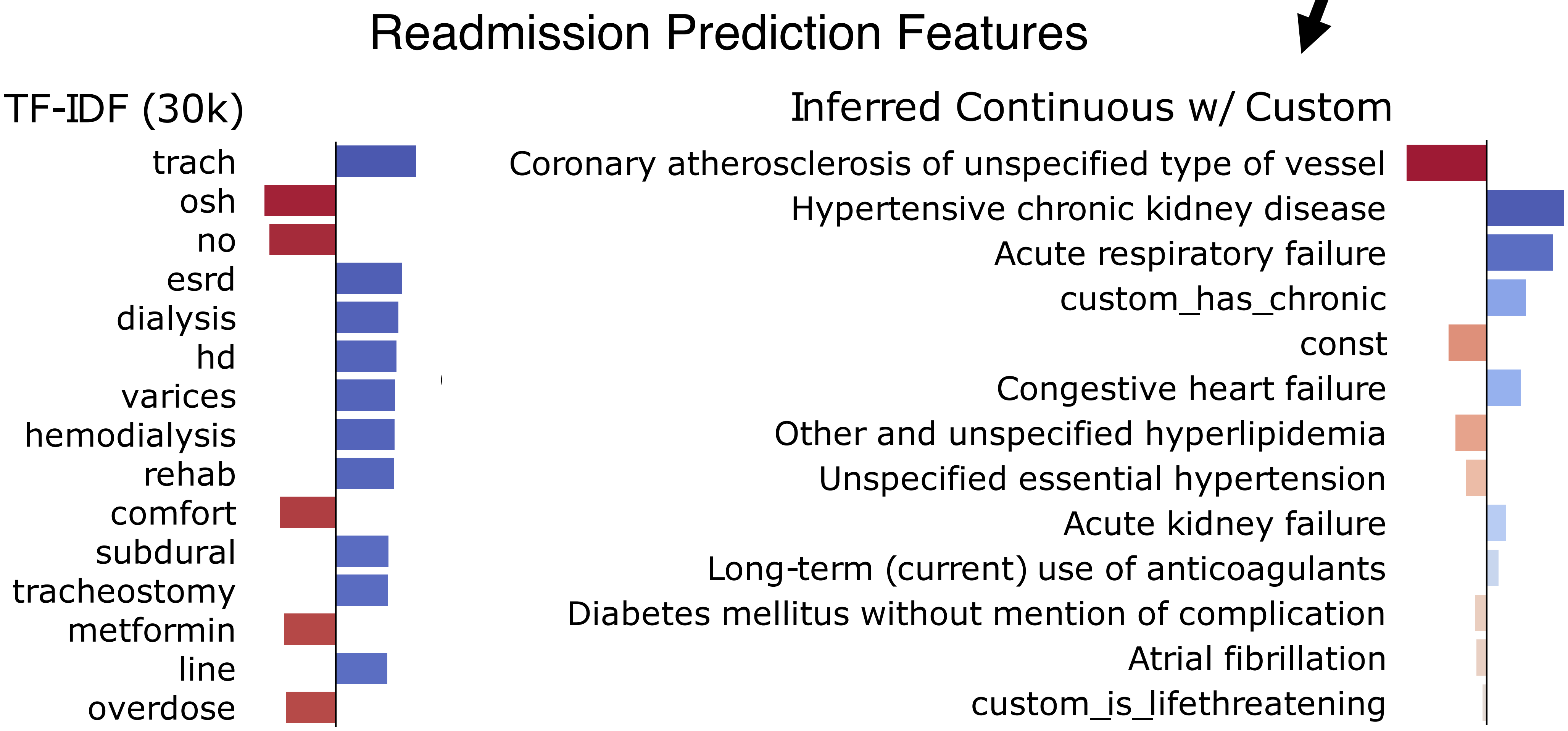
Downstream Performance



Reasonable!

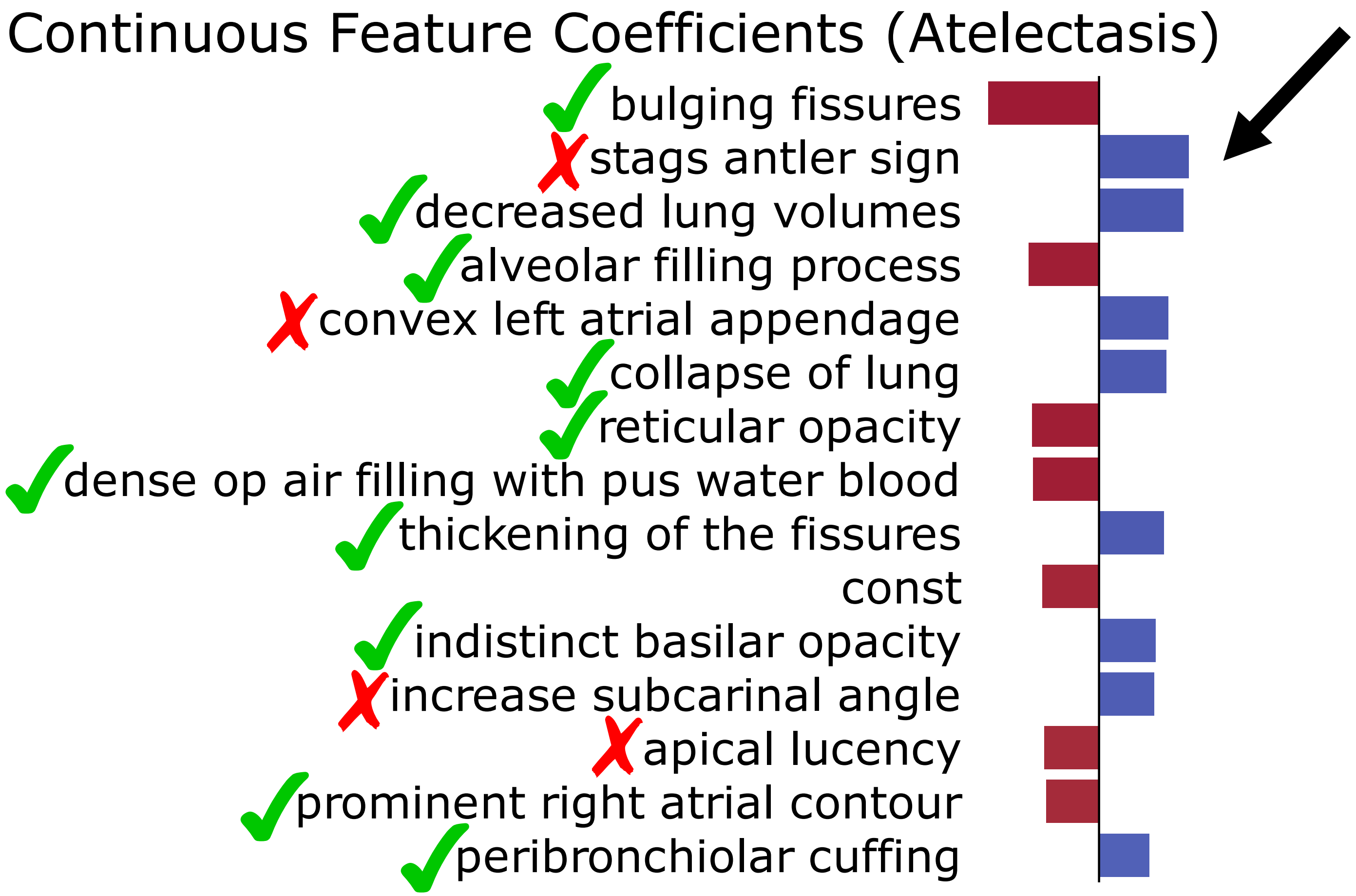
Feature Interpretability

Crafted feature names are more naturally interpretable

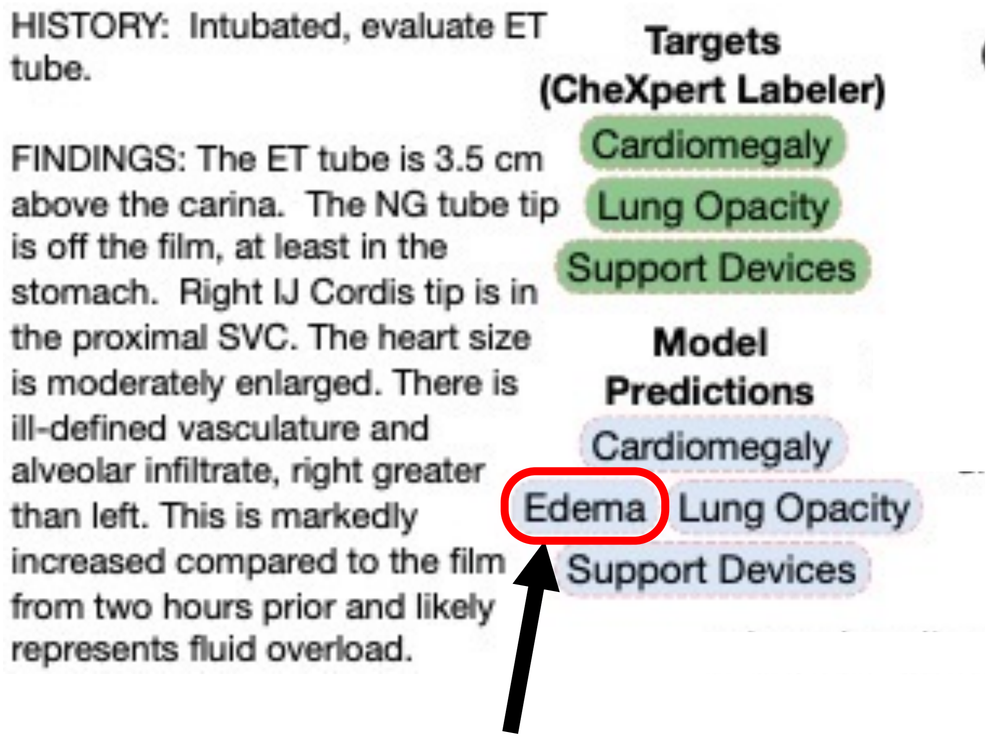


Model Interpretability

Weights align with clinical expectations

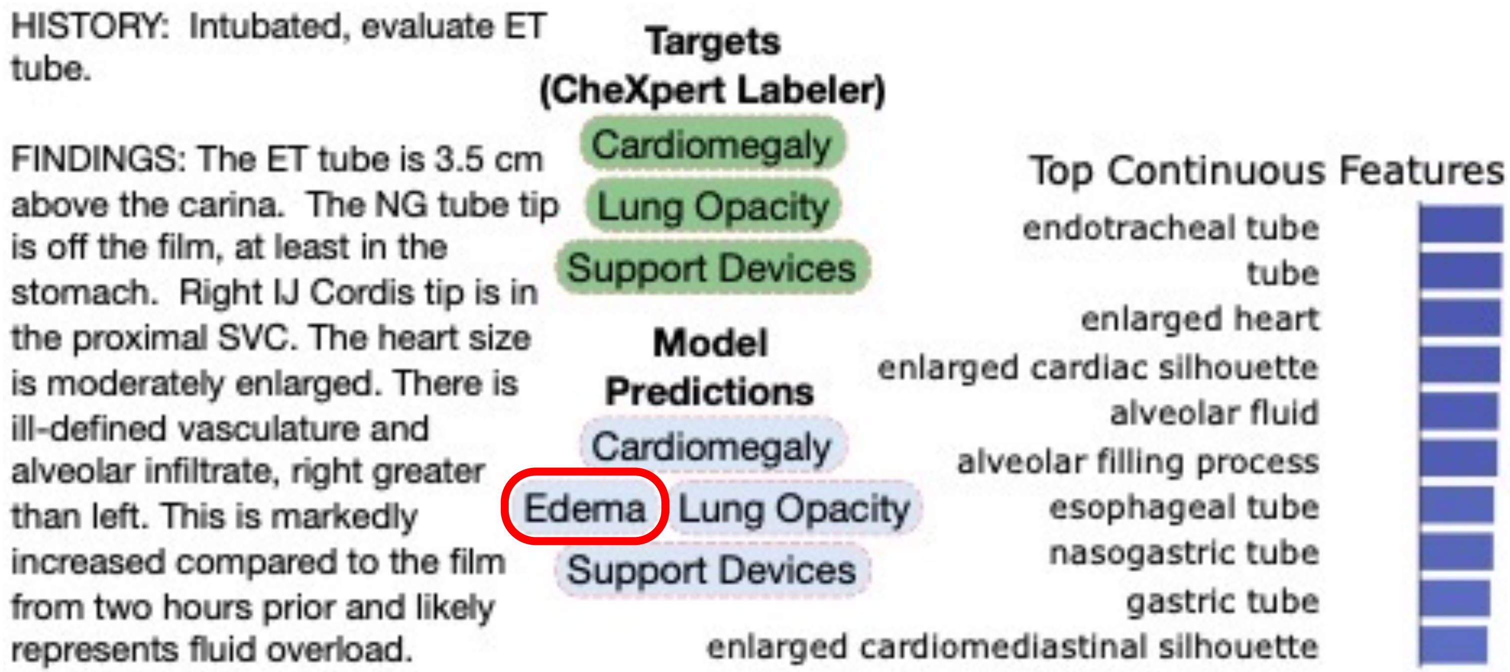


Feature Interpretability

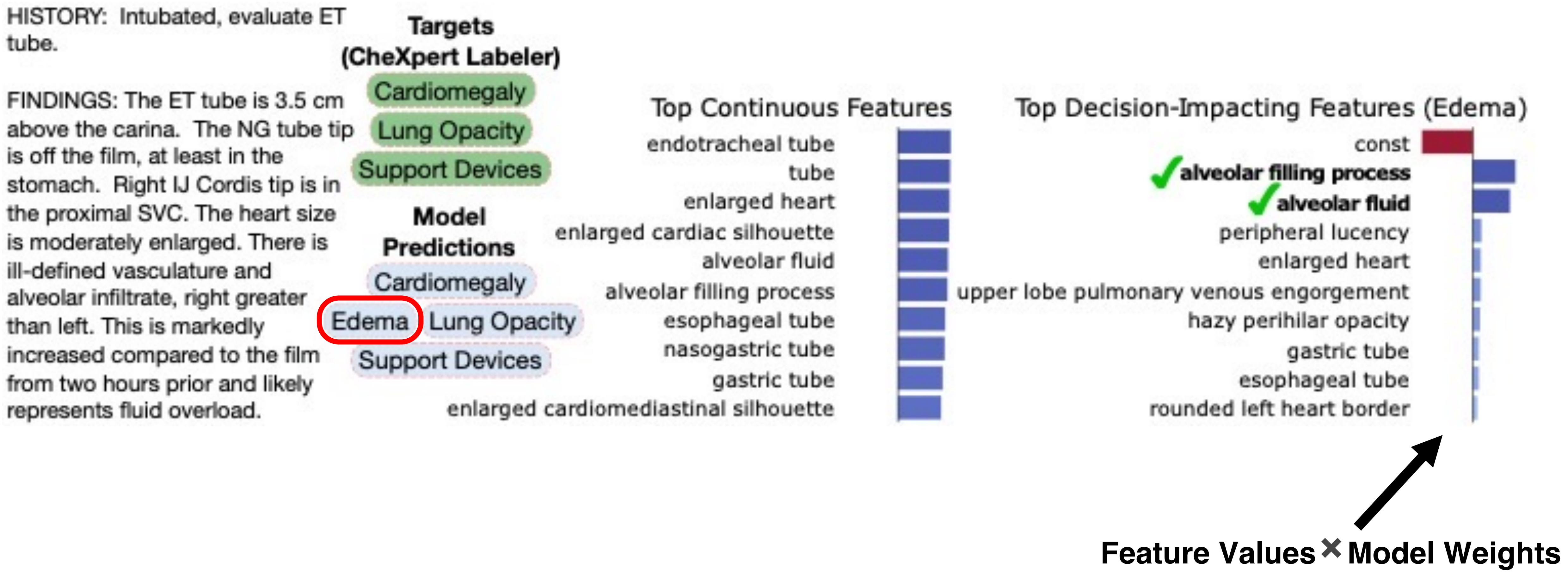


What went wrong?

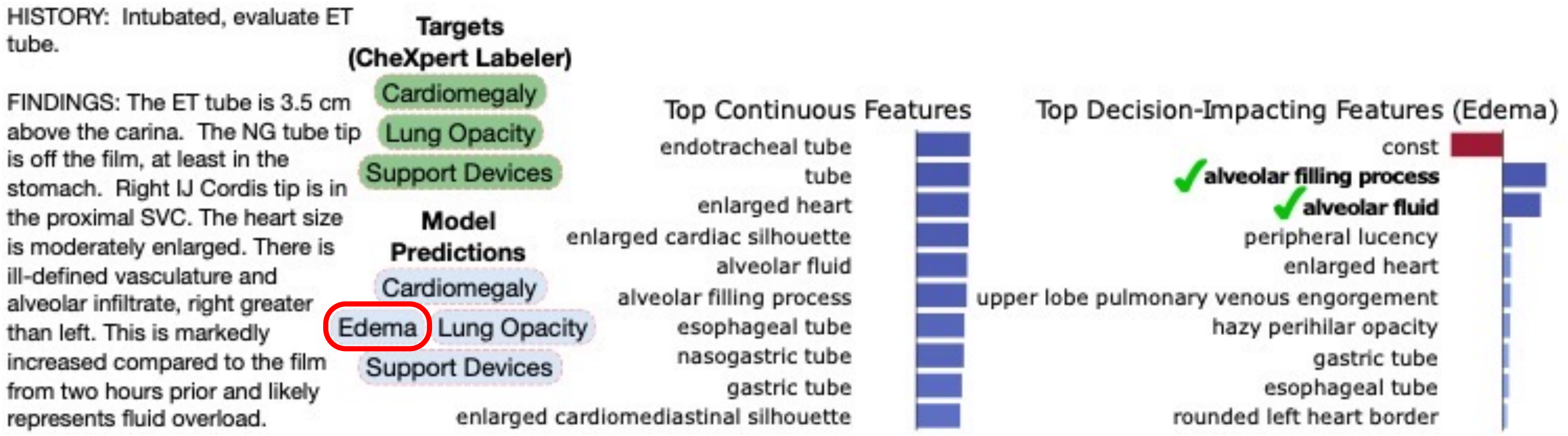
Feature Interpretability



Feature Interpretability

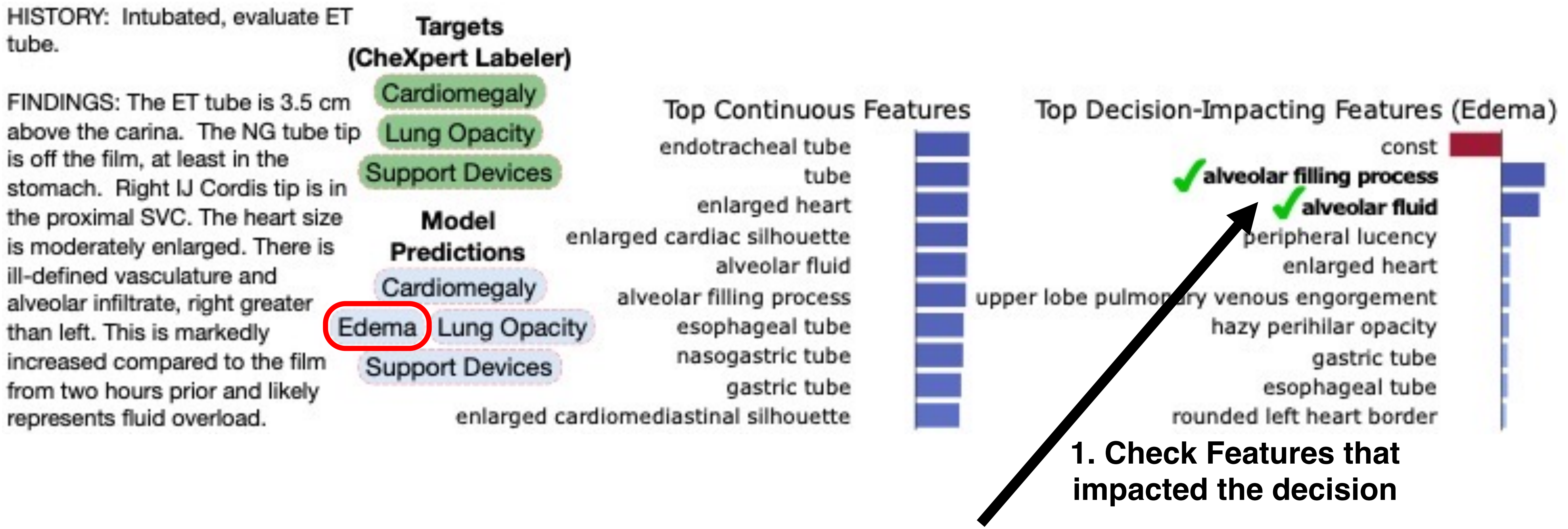


Feature Interpretability



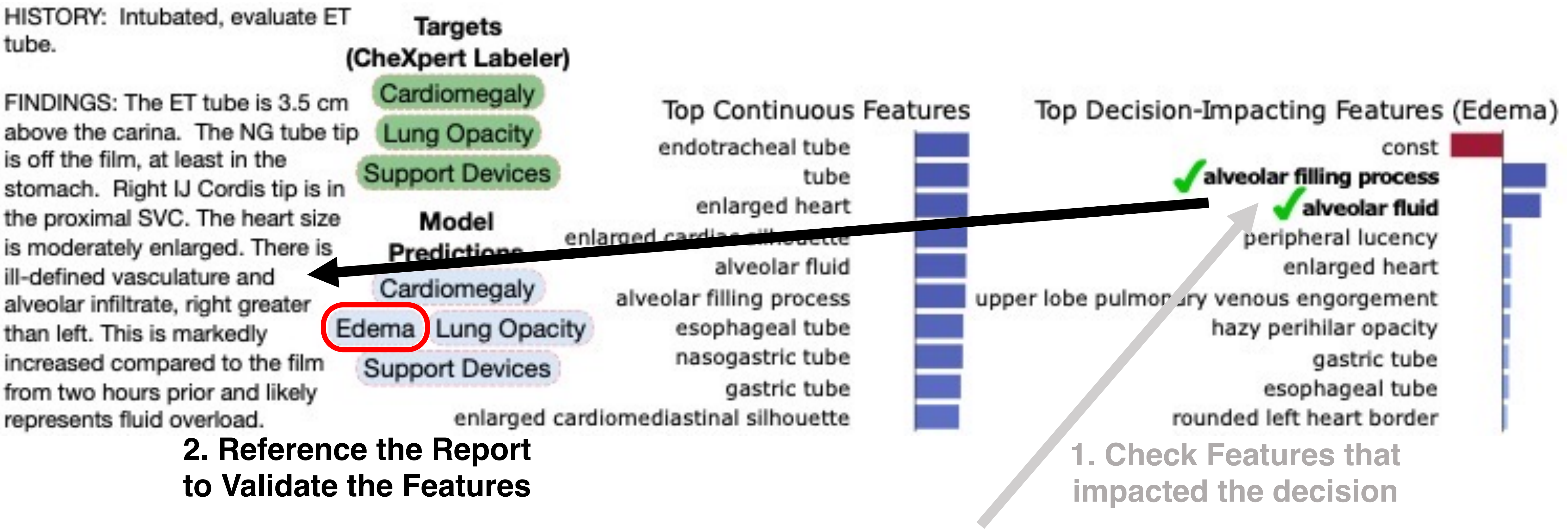
Extracted features permit easier inspection and validation of predictions on individual examples

Feature Interpretability



Extracted features permit easier inspection and validation of predictions on individual examples

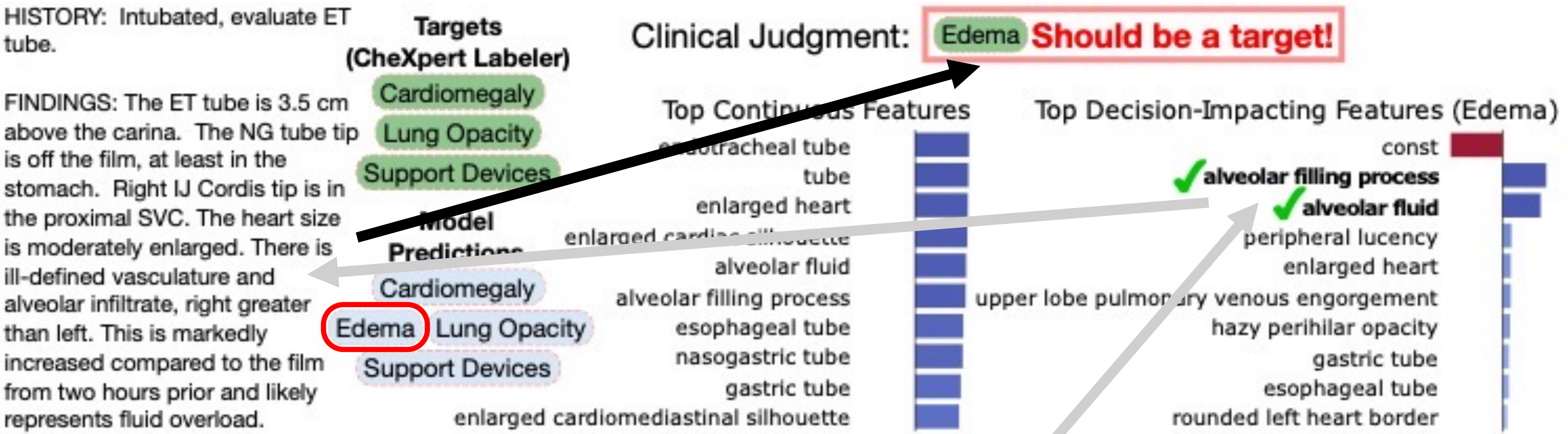
Feature Interpretability



Extracted features permit easier inspection and validation of predictions on individual examples

Feature Interpretability

3. Make a Clinical Judgment



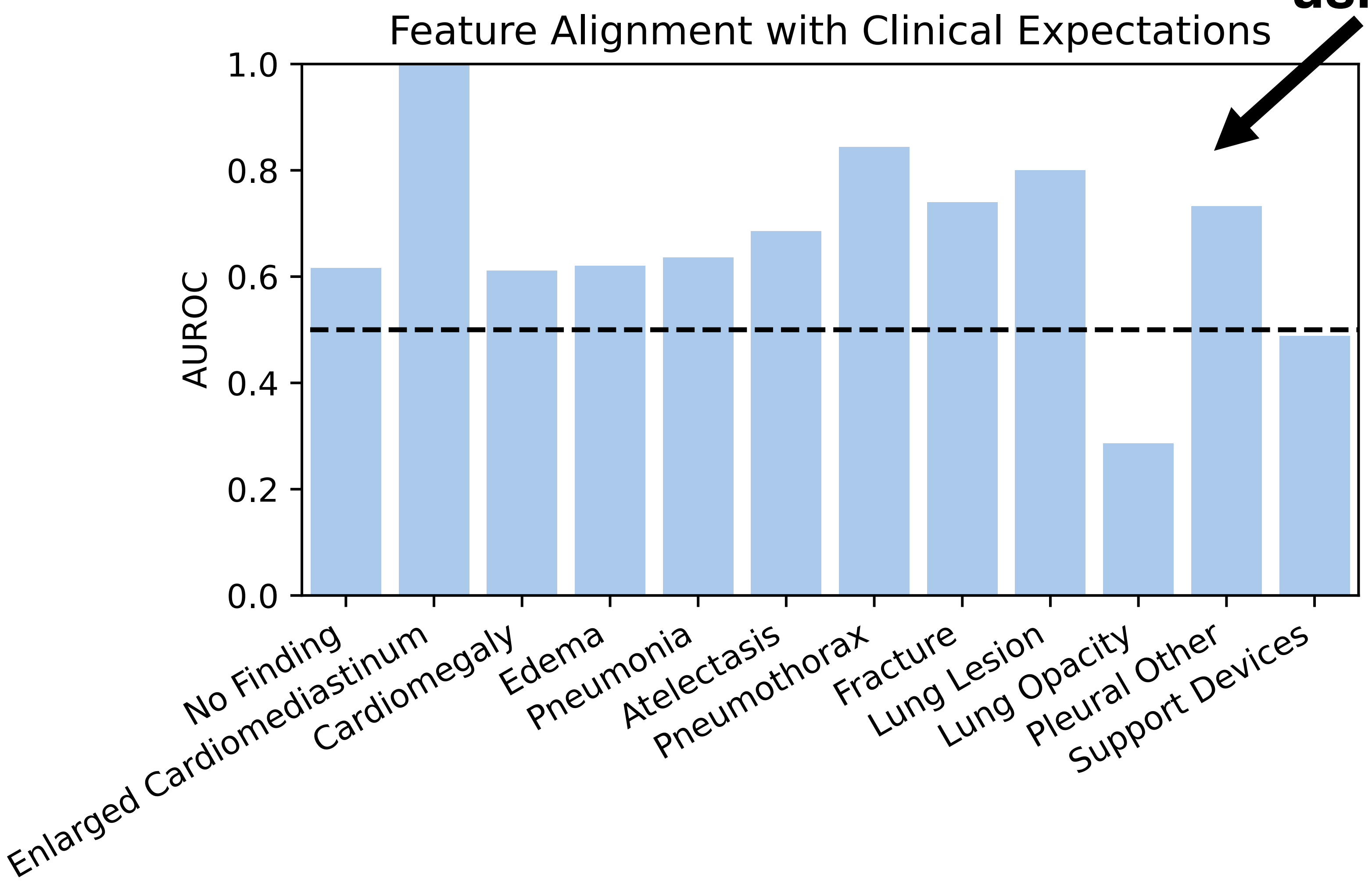
2. Reference the Report to Validate the Features

1. Check Features that impacted the decision

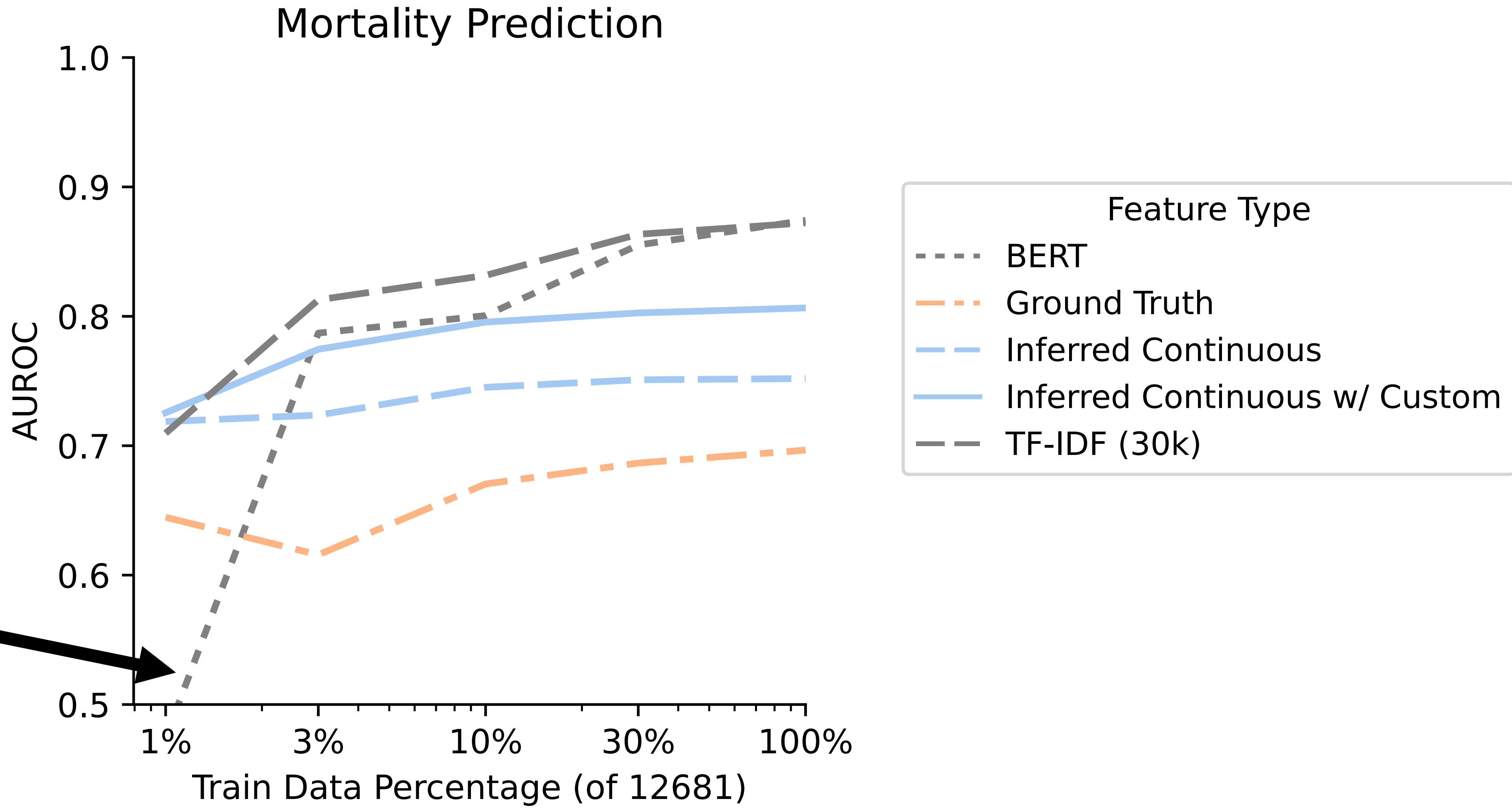
Extracted features permit easier inspection and validation of predictions on individual examples

Feature Alignment with Clinical Expectations

We can retrieve positively-contributing features using the model weights



Data Efficiency

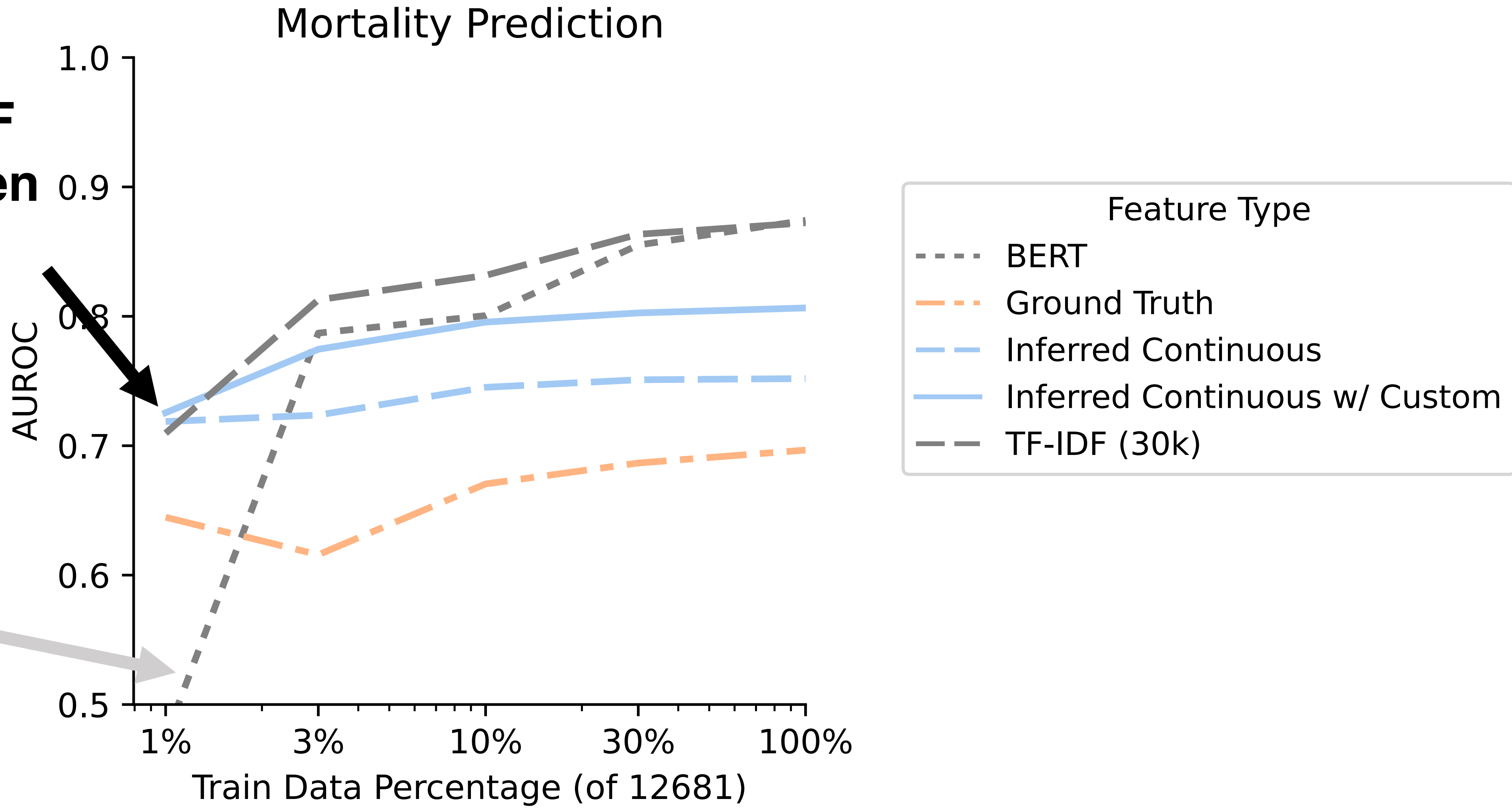


Less data-hungry than BERT

Data Efficiency

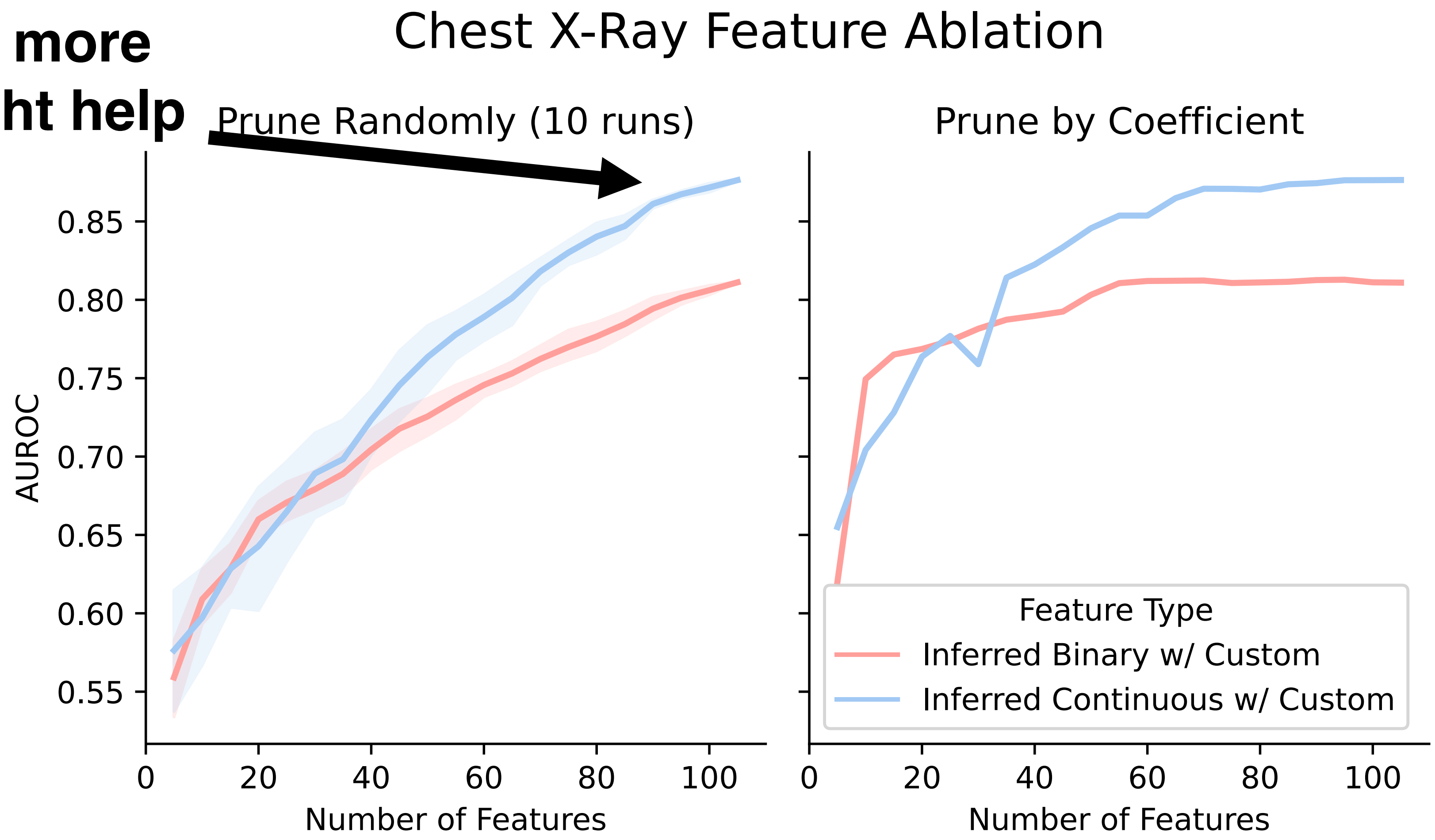
Similar to TF-IDF Performance when training data is scarce

Less data-hungry than BERT



Feature Efficiency

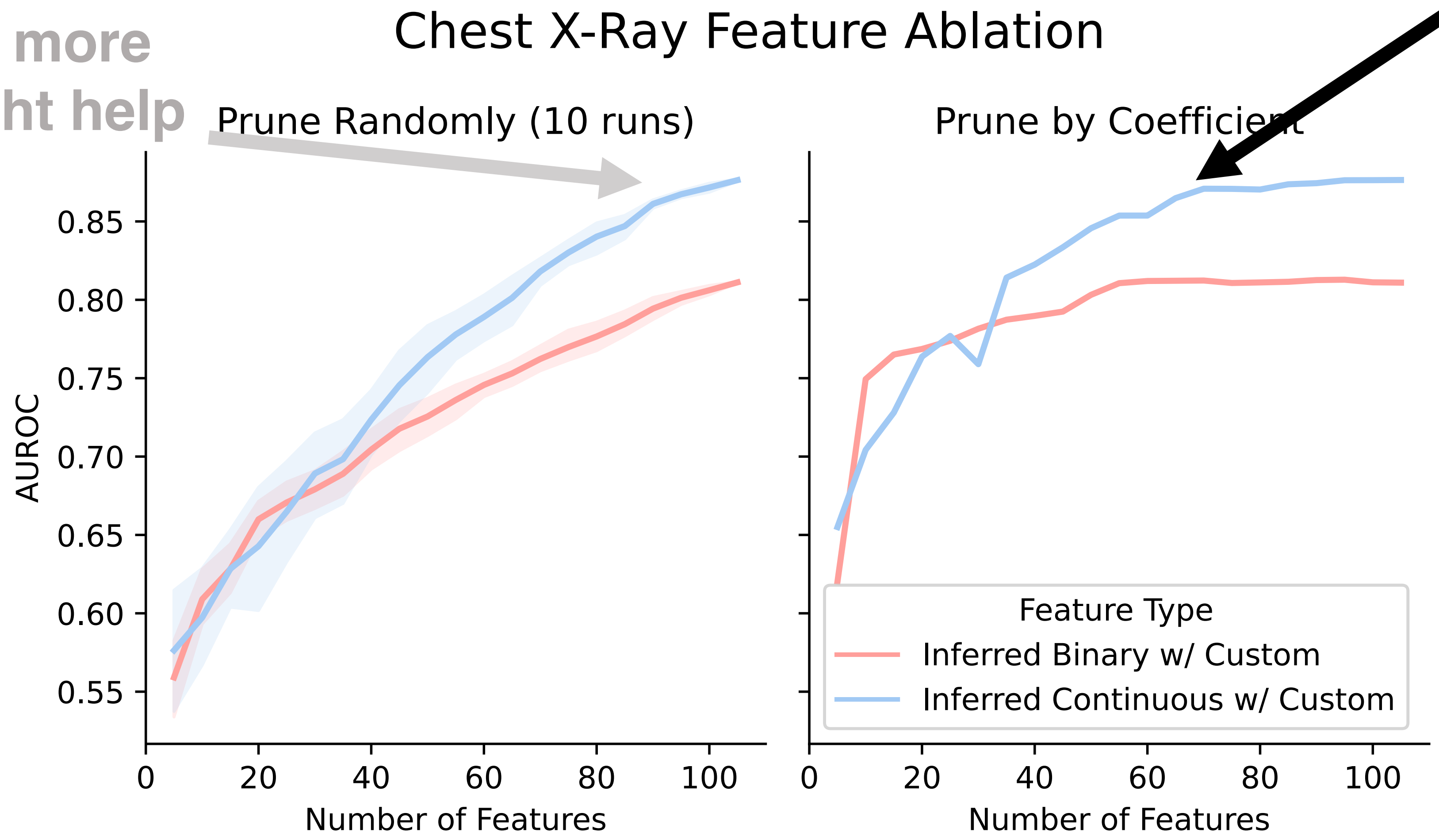
Not too much Plateau, so more features might help



Feature Efficiency

High Density of Useful Features

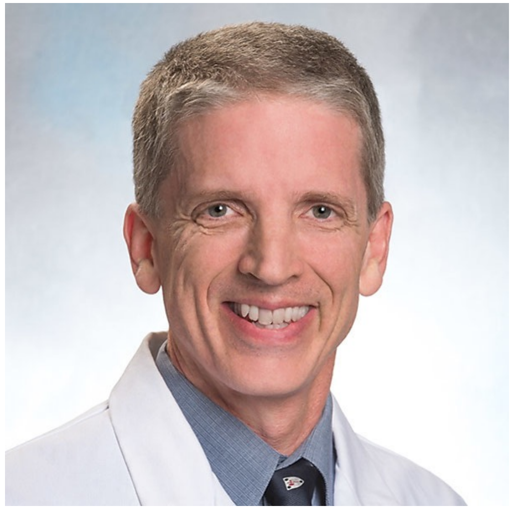
Not too much Plateau, so more features might help



Thank you

Byron C. Wallace
Northeastern University

Jan-Willem van de Meent
University of Amsterdam
Northeastern University



Geoffrey Young
Brigham and Women's Hospital

McInerney, Denis Jered, Geoffrey Young, Jan-Willem van de Meent and Byron C. Wallace. (2023). That's the Wrong Lung! Evaluating and Improving the Interpretability of Unsupervised Multimodal Encoders. Preprint.

Contact: mcinerney.de@northeastern.edu