

Supplementary Material

Disease-Aware Multi-Relational Representation Learning and Neighborhood-Prior Decision Calibration for Chest X-ray Multi-Label Prediction

S1. Dataset Class Distributions

The positive proportion for each label was calculated as the number of samples annotated as positive divided by the total number of samples in the corresponding dataset. Because these datasets use multi-label annotations, label counts are not mutually exclusive.

Supplementary Table S1. Positive sample counts and proportions for the original CheXpert observations and the CheXlocalize held-out test set. CheXpert contains 224,316 samples, and CheXlocalize contains 668 samples.

Label	CheXpert, n (%)	CheXlocalize, n (%)
No Finding	22,528 (10.04)	109 (16.32)
Enlarged Cardiomeastinum	11,205 (5.00)	298 (44.61)
Cardiomegaly	27,243 (12.14)	175 (26.20)
Lung Opacity	106,017 (47.26)	310 (46.41)
Lung Lesion	9,201 (4.10)	14 (2.10)
Edema	52,376 (23.35)	85 (12.72)
Consolidation	14,851 (6.62)	35 (5.24)
Pneumonia	6,061 (2.70)	14 (2.10)
Atelectasis	33,634 (14.99)	178 (26.65)
Pneumothorax	19,466 (8.68)	10 (1.50)
Pleural Effusion	86,374 (38.51)	120 (17.96)
Pleural Other	3,532 (1.57)	8 (1.20)
Fracture	9,046 (4.03)	6 (0.90)
Support Devices	116,423 (51.90)	315 (47.16)

Supplementary Table S2. Positive sample counts and proportions for the NIH Chest X-rays labels among 112,120 samples.

Label	Positive samples, n (%)
Atelectasis	11,559 (10.31)
Cardiomegaly	2,776 (2.48)
Effusion	13,317 (11.88)
Infiltration	19,894 (17.74)
Mass	5,782 (5.16)
Nodule	6,331 (5.65)
Pneumonia	1,431 (1.28)

Label	Positive samples, n (%)
Pneumothorax	5,302 (4.73)
Consolidation	4,667 (4.16)
Edema	2,303 (2.05)
Emphysema	2,516 (2.24)
Fibrosis	1,686 (1.50)
Pleural Thickening	3,385 (3.02)
Hernia	227 (0.20)
No Finding	60,361 (53.84)

S2. Relation-Level Complementarity Analysis

The structured input to the auxiliary tabular encoder contains 13 binary report-derived disease fields: No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Pneumonia. The raw input dimension is 13. The encoder produces a 512-dimensional pooled embedding and 14 tokens of 512 dimensions, corresponding to a flattened token-level dimension of 7,168.

The analysis used 668 held-out samples, yielding 222,778 unordered sample pairs. Jaccard similarity between the binary disease-label vectors served as the explicit label relation. This relation was compared with cosine similarity computed from the raw disease-label vectors, pooled encoder embeddings, and flattened token-level embeddings. Pearson and Spearman correlations were computed over the upper-triangular pairwise similarities. Nonlinear explained variance was measured by fitting isotonic regression on one randomly selected half of the samples and evaluating R-squared on the remaining half. Top-10 overlap denotes the mean proportion of shared neighbors between the two relation matrices.

Supplementary Table S3. Relation-level comparison between explicit disease-label similarity and cosine similarities obtained from raw and encoded structured-label representations.

Similarity representation	Pearson	Spearman	Nonlinear explained variance	Top-10 neighbor overlap
Raw disease-label cosine similarity	0.9632	0.9976	0.9947	91.14%
Pooled structured-label embedding cosine similarity	0.4387	0.4347	0.2083	39.72%
Token-level structured-label embedding cosine similarity	0.4014	0.3606	0.1627	40.67%

Raw disease-label cosine similarity closely preserves the Jaccard relation, whereas the pooled and token-level encoder similarities show substantially lower correlation, explained variance, and neighborhood overlap. The encoded representations therefore remain related to the disease-label input while inducing a different patient-relation geometry.

S3. Controlled Relation Experiments

The controlled experiments used a fixed 60% subset of the CheXpert training data, while the validation and held-out test sets remained unchanged. All variants used the same subset, optimization settings, evaluation protocol, and random seeds (2022, 2023, and 2024). Label only retains explicit disease-label similarity, whereas Embedding only retains the learned structured-label embedding relation. Label + Embedding combines both relations. In Label + Shuffled embedding, the sample-embedding correspondence is disrupted while preserving the embedding distribution and dimensionality. In Label + Random encoder, the auxiliary encoder is randomly initialized and remains frozen throughout training.

Supplementary Table S4. Control experiments for disease-label similarity and structured-label embedding similarity using 60% of the CheXpert training data. Results are reported as mean $\pm$ standard deviation over three random seeds. The best result for each metric is highlighted in bold.

Setting	Exact Match	Macro-AUROC	Macro-AUPRC	Macro-F1	Micro-F1
Label only	0.0599 $\pm$ 0.0191	0.6970 $\pm$ 0.0488	0.3140 $\pm$ 0.0364	0.1232 $\pm$ 0.0304	0.3700 $\pm$ 0.0509
Embedding only	0.0674 $\pm$ 0.0173	0.7166 $\pm$ 0.0061	0.3255 $\pm$ 0.0015	0.1495 $\pm$ 0.0141	0.3931 $\pm$ 0.0241
Label + Embedding	0.0698 $\pm$ 0.0206	0.7234 $\pm$ 0.0265	0.3353 $\pm$ 0.0136	0.1578 $\pm$ 0.0225	0.3874 $\pm$ 0.0208
Label + Shuffled embedding	0.0569 $\pm$ 0.0078	0.6235 $\pm$ 0.0106	0.2852 $\pm$ 0.0029	0.1402 $\pm$ 0.0090	0.3797 $\pm$ 0.0266
Label + Random encoder	0.0609 $\pm$ 0.0286	0.6647 $\pm$ 0.0731	0.3023 $\pm$ 0.0560	0.1286 $\pm$ 0.0446	0.3735 $\pm$ 0.0436

The combined relation achieves the highest Exact Match, Macro-AUROC, Macro-AUPRC, and Macro-F1. Performance decreases when the embedding correspondence is shuffled or when the auxiliary encoder is randomly initialized and frozen. These controls indicate that the contribution of the encoded relation is not explained solely by representation dimensionality or an arbitrary parameterized transformation.