A Appendix

A.1 Description of extracted pathologic data elements

Table A1: Description of the 4 extracted pathologic data elements.

Data elements	Description
Primary Gleason grade	A whole number from 1 to 5 representing the primary score given to a specimen based on the Gleason grading system to measure tumor aggressiveness.
Secondary Gleason grade	A whole number from 1 to 5 representing the secondary score given to a specimen based on the Gleason grading system to measure tumor aggressiveness.
Margin status for tumor	To evaluate surgical margins, the entire prostate surface is inked after removal. The surgical margins are designated as ”negative” if the tumor is not present at the inked margin, and ”positive” if tumor is present.
Seminal vesicle invasion	Invasion of tumor into the seminal vesicle. It is marked as ”negative” if no invasion is present in the seminal vesicle, and ”positive” if invasion is present.

A.2 Anonymized pathology report samples

- synoptic comment for prostate tumors " 1. type of tumor : adenocarcinoma small acinar type. " 2. location of tumor : both lobes. 3. estimated volume of tumor : 3. 5 ml. 4. gleason score : 4 + 3 = 7. 5. estimated volume > gleason pattern 3 : 2 ml. 6. involvement of capsule : present (e. g. slide b6). 7. extraprostatic extension : not identified. 8. status of excision margins for tumor : negative. status of excision margins for benign prostate glands : positive (e. g. slide b4). 9. involvement of seminal vesicle : not identified. 10. perineural infiltration : present (e. g. slide b11). " 11. prostatic intraepithelial neoplasia (pin) : present high - grade (e. g " slide b4). 12. ajcc / uicc stage : pt2cnxmx ; stage ii if no metastases are identified. 13. additional comments : none. final diagnosis : " a. prostate left apical margin : benign prostatic tissue. " " b. prostate and seminal vesicles resection : prostatic adenocarcinoma " gleason score 4 + 3 = 7 ; see comment.
- synoptic comment for prostate tumors - type of tumor : small acinar adenocarcinoma. - location of tumor : - right anterior midgland : slides b3 - b5. - right posterior midgland : slides b6 - b8. - left anterior midgland : slides b12 - b14. - left posterior midgland : slides b9 - b11. - left and central bladder bases : slides b16 - b17 - estimated volume of tumor : 10 cm3. " - gleason score : 7 ; primary pattern 3 secondary pattern 4. " - estimated volume > gleason pattern 3 : 40 %. " - involvement of capsule : tumor invades capsule but does not extend beyond " " capsule (slides b5 b8 b18). " - extraprostatic extension : none. - margin status for tumor : negative. - margin status for benign prostate glands : negative. - high - grade prostatic intraepithelial neoplasia (hgpin

```
) :  present ; extensive.  - tumor involvement of seminal
vesicle :  none.  - perineural infiltration :  present.  -
lymph node status :  none submitted.  - ajcc / uicc stage :
pt2cnx.  final diagnosis :  " a.  prostate left base biopsy
:  fibromuscular tissue no tumor.  " " b.  prostate radical
prostatectomy :  " " 1.  prostatic adenocarcinoma gleason
grade 3 + 4 score = 7 involving " " bilateral prostate
negative margins ; see comment.  2.  " seminal vesicles with
no significant pathologic abnormality.
```

A.3 Fine-tuning

We fine-tune the models to perform single-label classification for all tasks. We add a linear layer followed by a softmax function to the model output on the classification token. The datasets are divided into 71% training, 18% validation, and 11% test, with label distribution in each set resembling the distribution in the full datasets. Best model checkpoints are selected based on validation set performances, and are used in all experiments. For pathology reports, we evaluate the models against macro F1 as each class accounts for equal importance, while we report accuracy for MedNLI. We set the encoder sequence length to 512 tokens for pathology reports, and 256 tokens for MedNLI, which allows us to encode the full length of the majority of the datasets.

Prostate Cancer Pathology Reports We use consistent fine-tuning hyperparameters for all models and all the four tasks, as we observe the validation set performance is not very sensitive to hyperparameter selection (less than 1% F1 performance change). We use an AdamW optimizer with a 7.6e-6 learning rate, 0.01 weight decay, and a 1e-8 epsilon. We also adopt a linear learning rate schedule with a 0.2 warm-up ratio. We fine-tune for a maximum of 25 epochs with a batch size of 8 and evaluate every 50 steps on the validation set. Each model is fine-tuned on a single NVIDIA Tesla K80 GPU, and average fine-tuning time is around 3 hours.

MedNLI We use consistent fine-tuning hyperparameters for all models, as we observe the validation set performance is not very sensitive to hyperparameter selection (less than 1% accuracy change). We use an AdamW optimizer with a per-layer learning rate decay schedule (1e-4 as the starting learning rate, and 0.8 as the decay factor), 0 weight decay, 1e-6 epsilon, and a 0.1 warm-up ratio. We fine-tune for a maximum of 10 epochs with a batch size of 32 and evaluate every epoch on the validation set. Each model is fine-tuned on a single NVIDIA GeForce GTX TITAN X GPU, and the fine-tuning time on average is less than 1 hours.

A.4 Per-class accuracy on Path-PG and Path-SG

Table A2: Per-class accuracy of the five models on Path-PG and Path-SG, averaged across three runs (all stds are $< 5\%$ so we omit it to save spaces). PubMedBERT performs poorly when classifying the minority class 5 in the highly imbalanced Path-PG dataset, while it obtains descent performance across all classes in the slightly more balanced Path-SG dataset.

	Path-PG			Path-SG		
Models \Labels	3	4	5	3	4	5
BERT	0.99	0.94	1.00	0.98	0.98	0.97
TNLR	0.97	0.87	1.00	0.99	0.99	0.99
BioBERT	0.99	0.97	0.94	0.99	0.99	0.99
Clinical BioBERT	0.99	0.98	1.00	0.99	0.99	0.98
PubMedBERT	0.99	0.92	**0.67**	0.98	0.99	0.97

A.5 Fine-tuning results on MedNLI

Table A3: Per-class accuracy and overall accuracy of PubMedBERT and Clinical BioBERT on MedNLI across three runs, where three scenarios are evaluated: Balanced (‘C’:’E’:’N’=34%:33%:33%), Imbalanced (‘C’:’E’:’N’=39%:53%:8%), and Highly Imbalanced (‘C’:’E’:’N’=67%:30%:3%).

	Balanced		Imbalanced		Highly Imbalanced	
Labels \Models	PubMedBERT	Clinical BioBERT	PubMedBERT	Clinical BioBERT	PubMedBERT	Clinical BioBERT
Contradiction (’C’)	0.88 (0.03)	0.76 (0.03)	0.76 (0.03)	0.70 (0.02)	0.80 (0.01)	0.79 (0.03)
Entailment (’E’)	0.75 (0.02)	0.71 (0.02)	0.71 (0.03)	0.70 (0.02)	0.34 (0.02)	0.62 (0.05)
Neutral (’N’)	0.77 (0.05)	0.72 (0.01)	0.33 (0.16)	0.32 (0.01)	0.04 (0.03)	0.04(0.02)
Accuracy	0.83 (0.01)	0.73 (0.01)	0.70 (0.02)	0.71 (0.01)	0.71 (0.01)	0.76 (0.03)

A.6 Quantifying the closeness between pre-training data and target data

We use perplexity of pre-trained models on target tasks to define the closeness between pre-training data and target data. The lower the perplexity means the closer the two data distributions should be.

Table A4: Perplexity of the five models on pathology reports.

	BERT	TNLR	BioBERT	Clinical BioBERT	PubMedBERT
Perplexity	1.111	1.115	1.113	1.110	1.103

A.7 Auxiliary Material for Section 6.1

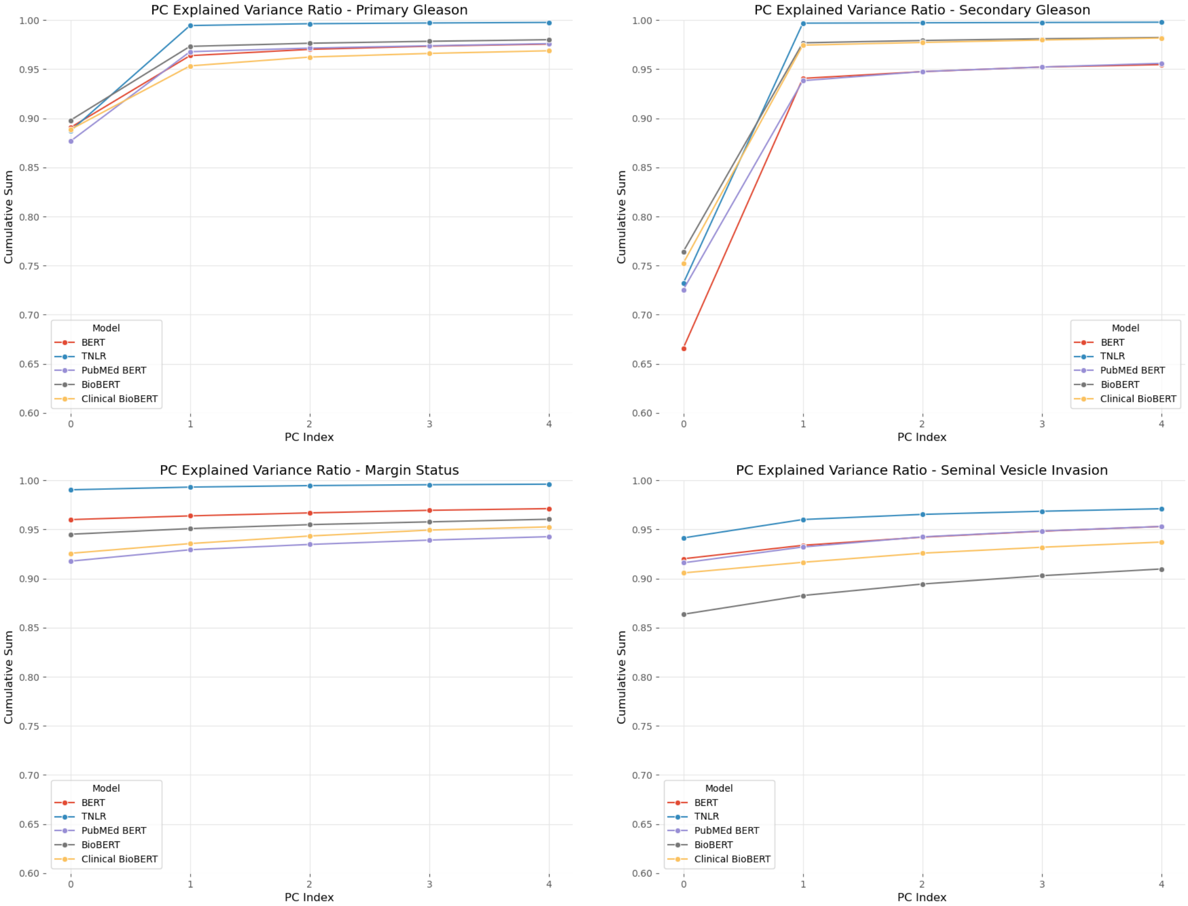

Figure A1: The first two PCs in the fine-tuned last layer classification token feature spaces of all the models explain on average 95% of the dataset variance across the 4 tasks.

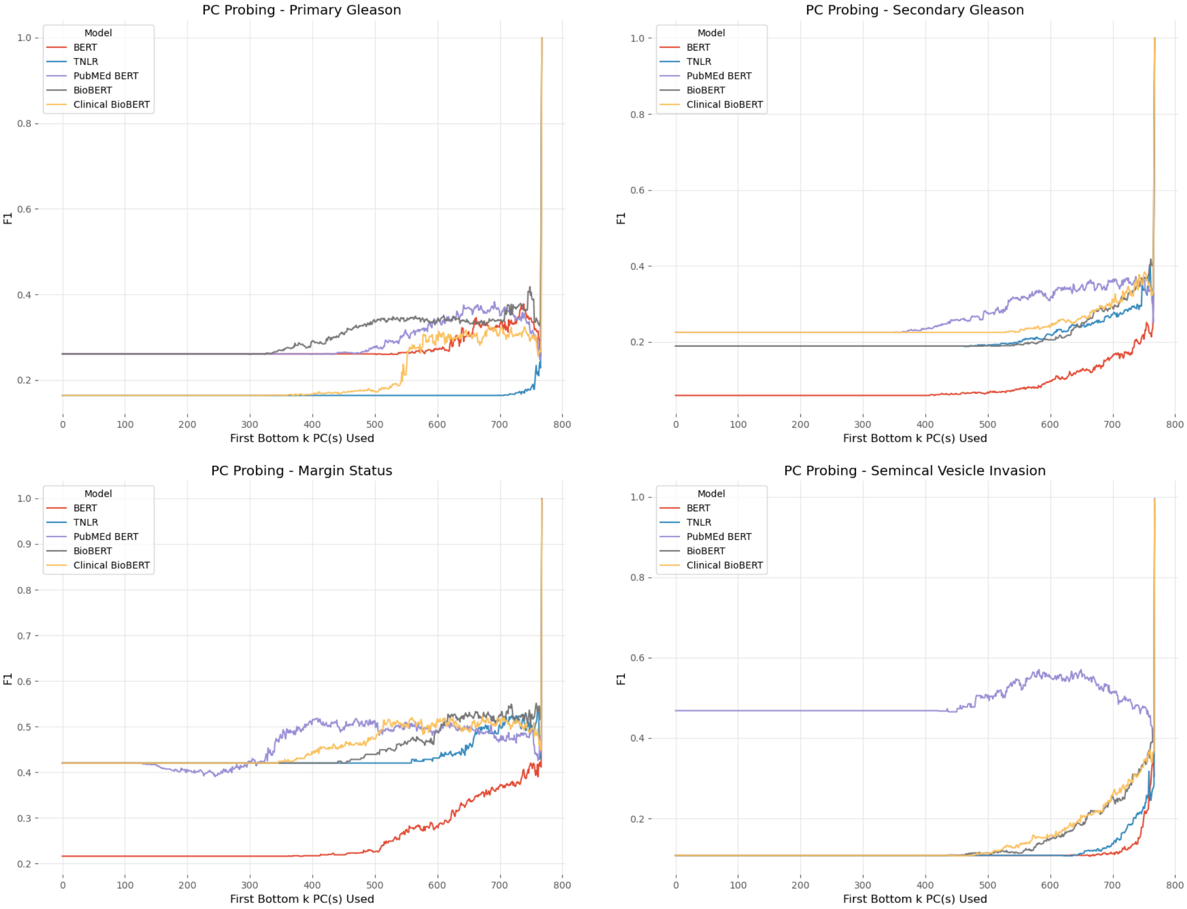

Figure A2: Filling back in the first two PCs, at the last two steps, $k = 767$ and $k = 768$, yields significant model performance gain.

A.8 Auxiliary Material for Section 6.2

The full categorization of the types of outliers obtained from our expert evaluation is provided in Table A5, while the distribution of these classes of outlier reports for each model are provided in Table A6.

Table A5: Description of categories of hard outliers

Outlier Category ID	Category Name	Category Description
1	Wrongly labeled report	These are reports for which the provided annotation is incorrect. For example, a report with `null` Gleason score corresponding to a scenario where a Gleason score cannot be assigned is wrongly included with another label.
2	Inconsistent report	For these reports, there exists inconsistent declarations of the target attribute (say, Primary Gleason score) in two different parts of the report.
3	Multiple Sources of Information	These reports contain multiple sources of information which are composed to produce one final label. One such instance of such an outlier (for the Secondary Gleason label) contained scores from five tumor nodules which were then combined to give one final composite score. A classifier must learn to distinguish the true final score from those that were used to obtain it.
4	Not reported or truncated report	These are reports for which the target attribute is either not reported or the report is truncated before entry into the database.
5	Boundary reports	These reports feature scenarios where the target attribute is hard to determine precisely or requires some interpretation of the provided information. For instance, one such report presents a Gleason score with a combined value of 7 with the other information in the report requiring the classifier to deduce that the Gleason score is $3 + 4$.

Table A6: A distribution of Hard Outliers for each model categorized according to the 5 outlier types.

Outlier Type	BERT	BioBERT	Clinical BioBERT	PubMedBERT	TNLR
1	0	0	1	1	1
2	0	1	0	1	2
3	2	0	1	1	1
4	0	1	3	5	1
5	4	0	3	3	2
Total	6	2	8	11	7

A.9 Feature dynamics

Here we present comprehensive sets of feature scatterplots along layers 1 to layer 12 (top-down) and selected epochs in the order of $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25$ (left-right) of the 5 models, as we observe the models typically show the most rapid performance gain from epoch 1 to 10, and marginal increase afterwards. We include the plots from Path-PG and Path-MS, as representatives of tasks having different number of labels to save space, but note that we observe similar trend in the results of all the 4 tasks

A.9.1 PATH-PG

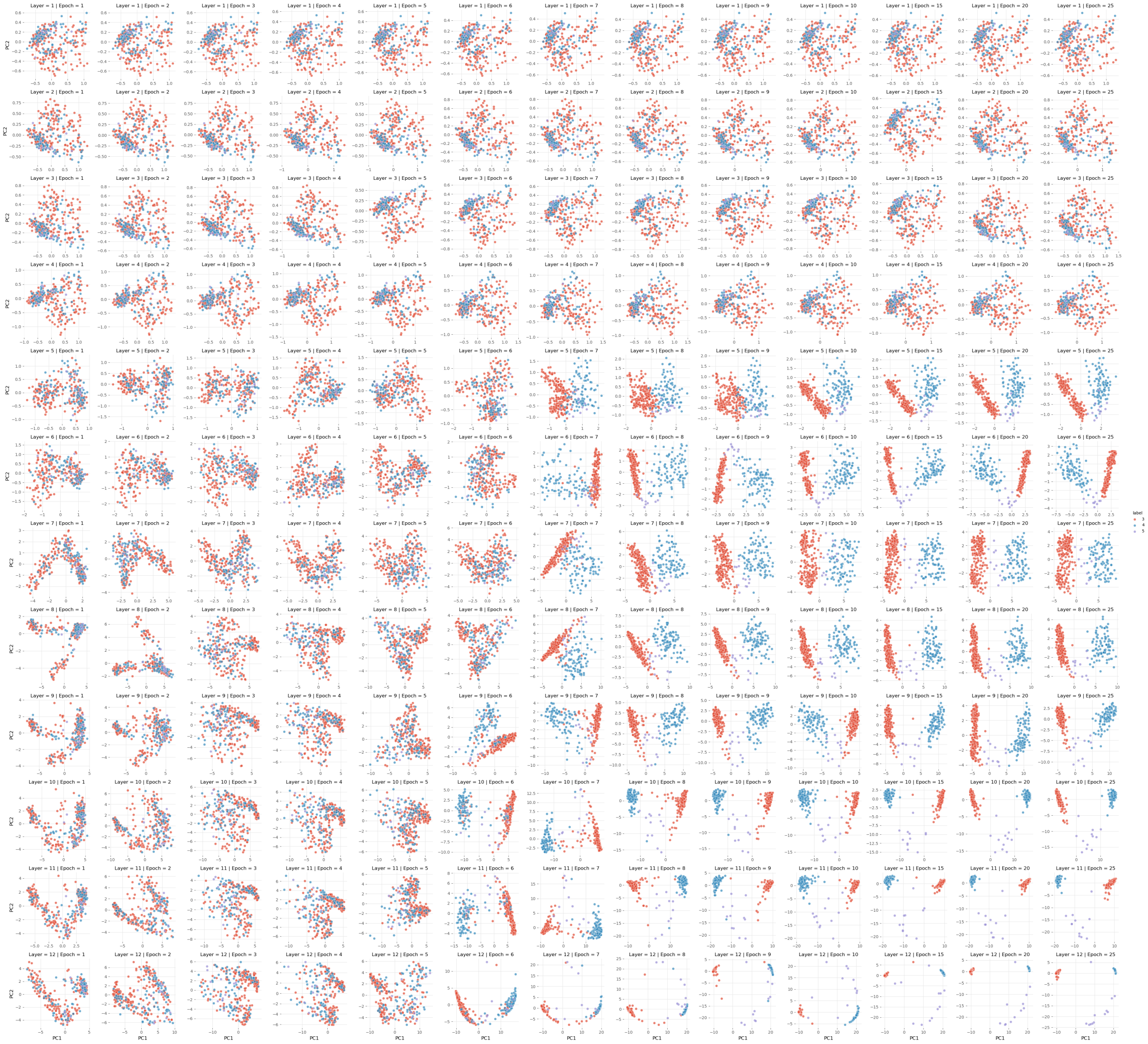

Figure A3: Path-PG: BERT

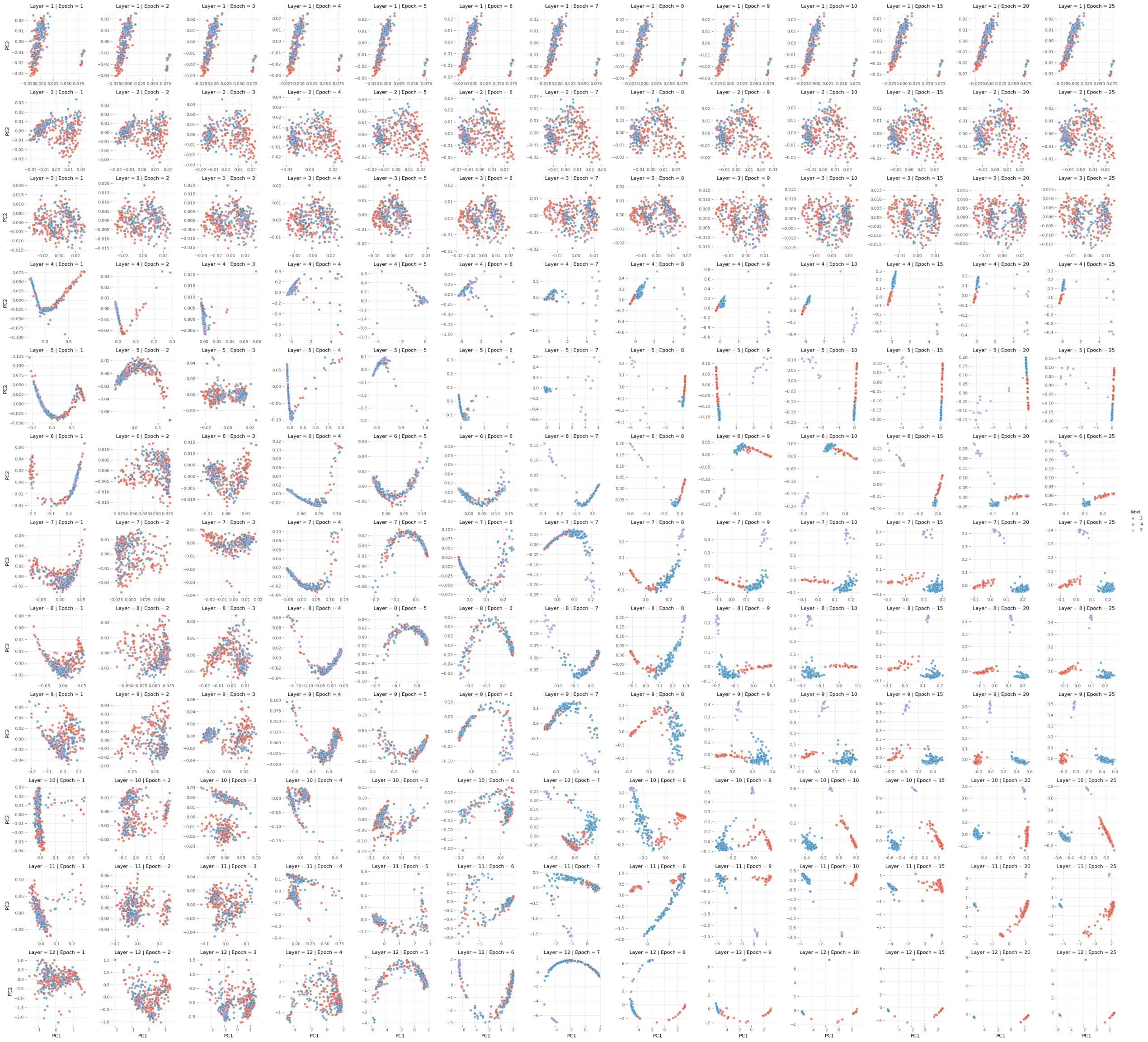

Figure A4: Path-PG: TNLR

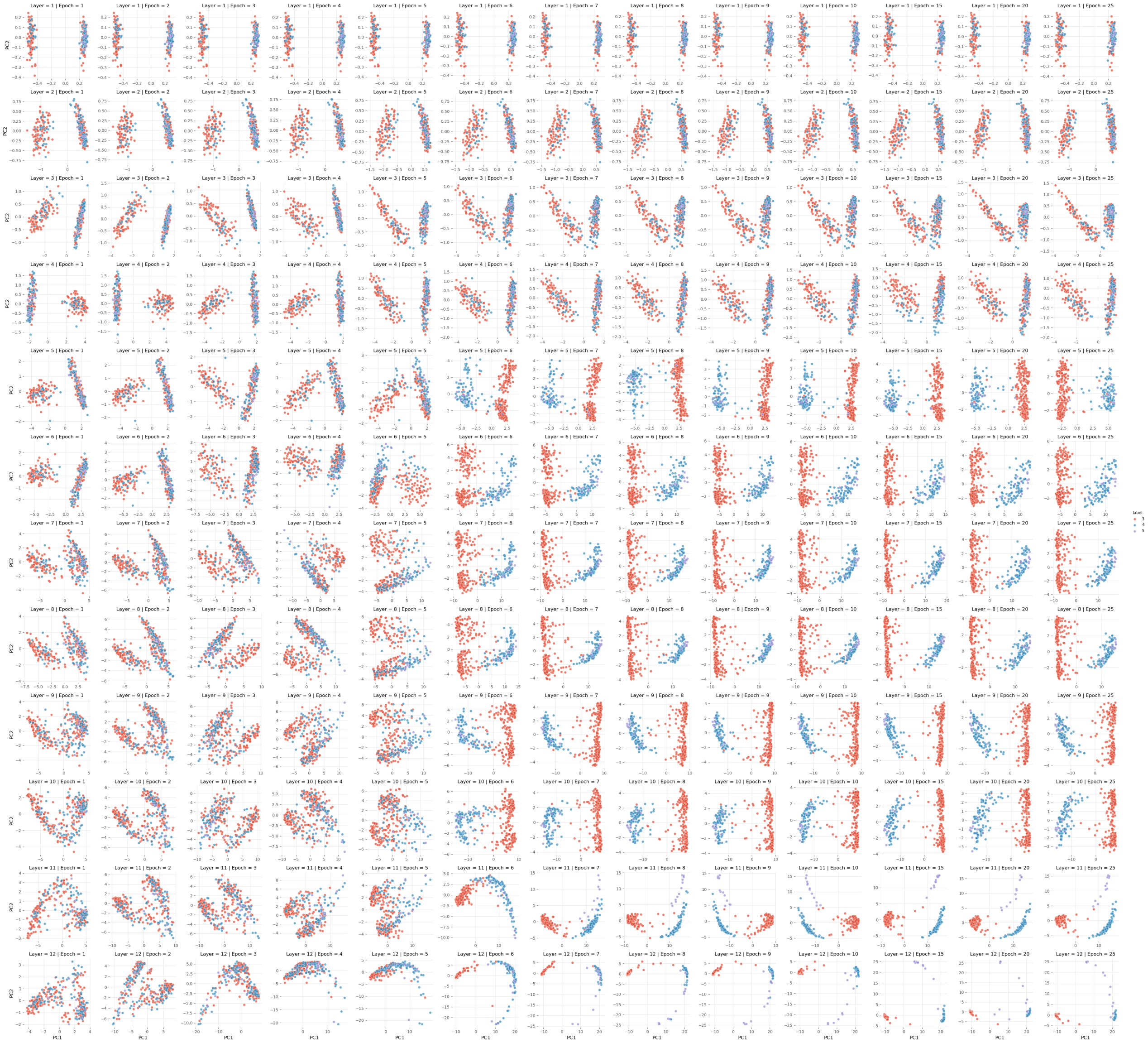

Figure A5: Path-PG: BioBERT

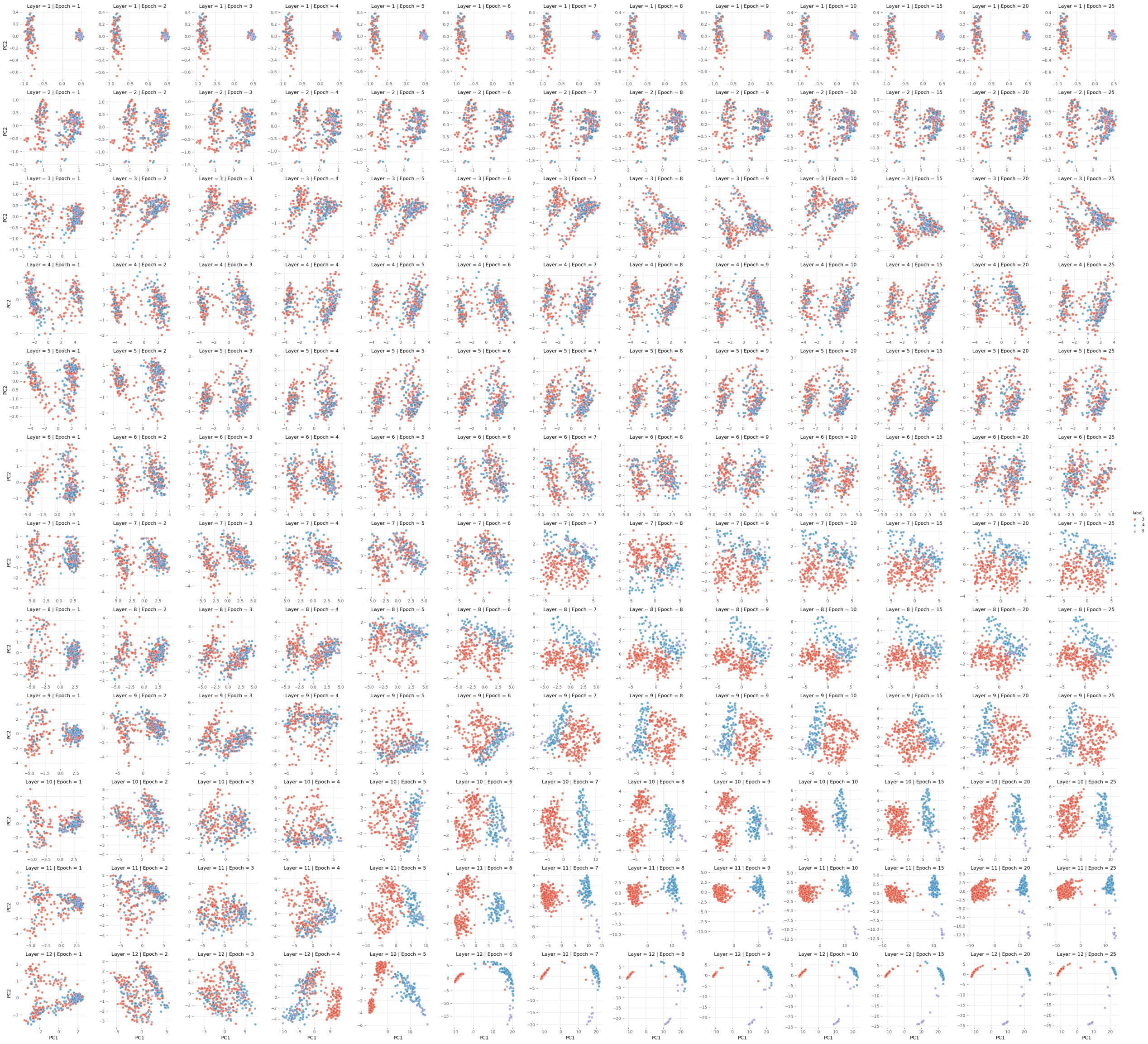

Figure A6: Path-PG: Clinical BioBERT

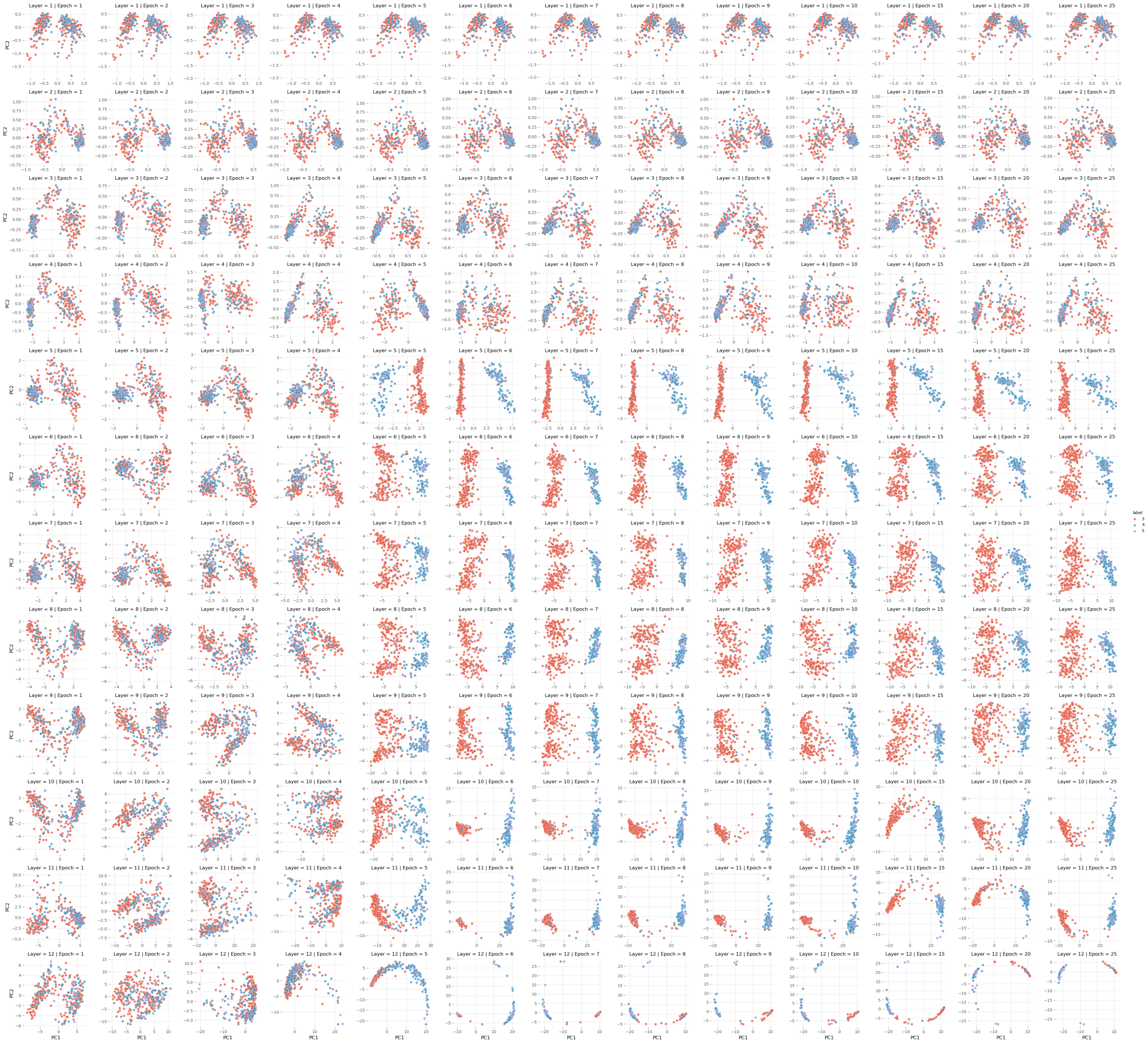

Figure A7: Path-PG: PubMedBERT

A.9.2 Path-MS

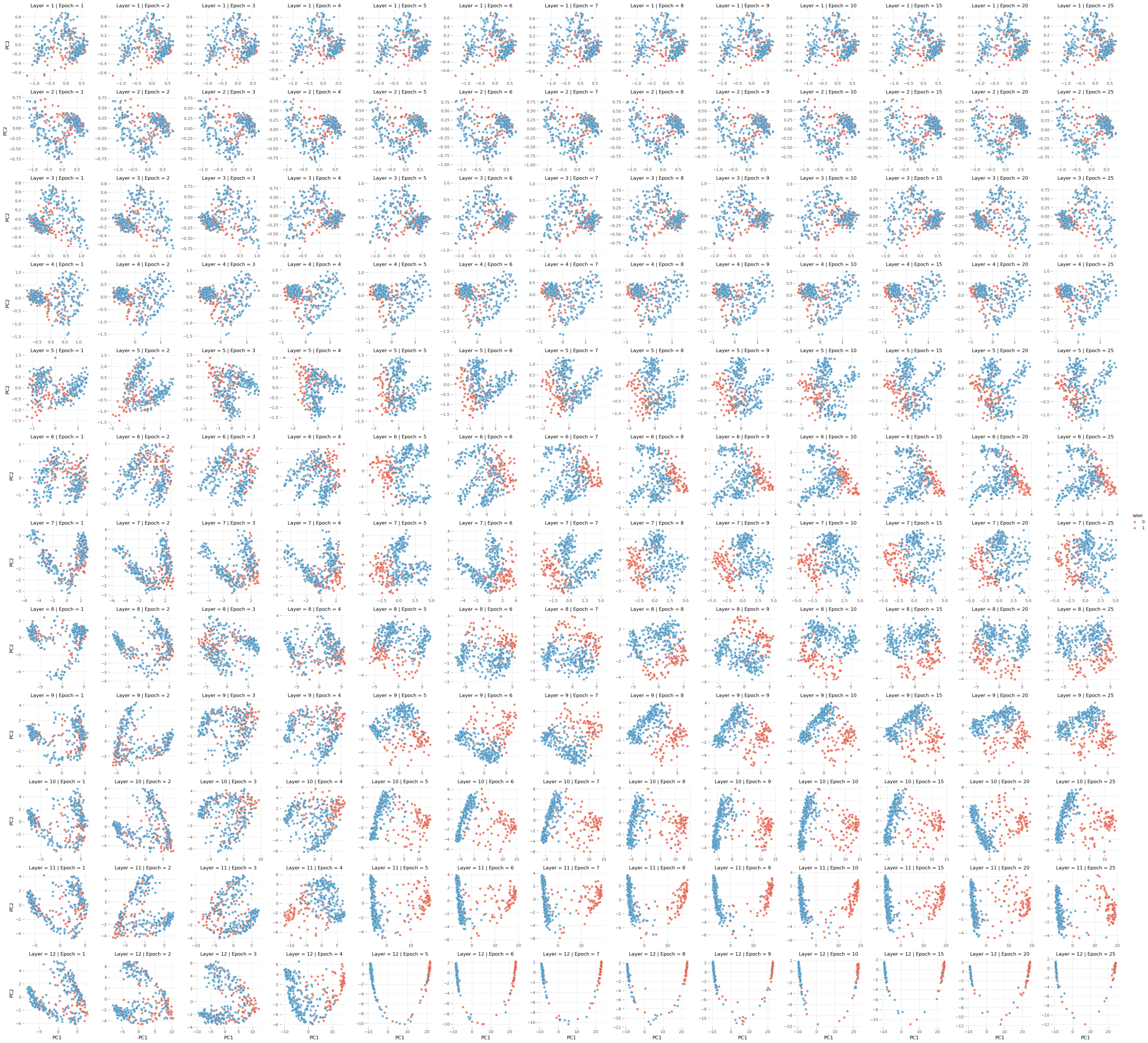
Layer = 1 | Epoch = 1, Layer = 1 | Epoch = 2, Layer = 1 | Epoch = 3, Layer = 1 | Epoch = 4, Layer = 1 | Epoch = 5, Layer = 1 | Epoch = 6, Layer = 1 | Epoch = 7, Layer = 1 | Epoch = 8, Layer = 1 | Epoch = 9, Layer = 1 | Epoch = 10, Layer = 1 | Epoch = 15, Layer = 1 | Epoch = 20, Layer = 1 | Epoch = 25
Layer = 2 | Epoch = 1, Layer = 2 | Epoch = 2, Layer = 2 | Epoch = 3, Layer = 2 | Epoch = 4, Layer = 2 | Epoch = 5, Layer = 2 | Epoch = 6, Layer = 2 | Epoch = 7, Layer = 2 | Epoch = 8, Layer = 2 | Epoch = 9, Layer = 2 | Epoch = 10, Layer = 2 | Epoch = 15, Layer = 2 | Epoch = 20, Layer = 2 | Epoch = 25
Layer = 3 | Epoch = 1, Layer = 3 | Epoch = 2, Layer = 3 | Epoch = 3, Layer = 3 | Epoch = 4, Layer = 3 | Epoch = 5, Layer = 3 | Epoch = 6, Layer = 3 | Epoch = 7, Layer = 3 | Epoch = 8, Layer = 3 | Epoch = 9, Layer = 3 | Epoch = 10, Layer = 3 | Epoch = 15, Layer = 3 | Epoch = 20, Layer = 3 | Epoch = 25
Layer = 4 | Epoch = 1, Layer = 4 | Epoch = 2, Layer = 4 | Epoch = 3, Layer = 4 | Epoch = 4, Layer = 4 | Epoch = 5, Layer = 4 | Epoch = 6, Layer = 4 | Epoch = 7, Layer = 4 | Epoch = 8, Layer = 4 | Epoch = 9, Layer = 4 | Epoch = 10, Layer = 4 | Epoch = 15, Layer = 4 | Epoch = 20, Layer = 4 | Epoch = 25
Layer = 5 | Epoch = 1, Layer = 5 | Epoch = 2, Layer = 5 | Epoch = 3, Layer = 5 | Epoch = 4, Layer = 5 | Epoch = 5, Layer = 5 | Epoch = 6, Layer = 5 | Epoch = 7, Layer = 5 | Epoch = 8, Layer = 5 | Epoch = 9, Layer = 5 | Epoch = 10, Layer = 5 | Epoch = 15, Layer = 5 | Epoch = 20, Layer = 5 | Epoch = 25
Layer = 6 | Epoch = 1, Layer = 6 | Epoch = 2, Layer = 6 | Epoch = 3, Layer = 6 | Epoch = 4, Layer = 6 | Epoch = 5, Layer = 6 | Epoch = 6, Layer = 6 | Epoch = 7, Layer = 6 | Epoch = 8, Layer = 6 | Epoch = 9, Layer = 6 | Epoch = 10, Layer = 6 | Epoch = 15, Layer = 6 | Epoch = 20, Layer = 6 | Epoch = 25
Layer = 7 | Epoch = 1, Layer = 7 | Epoch = 2, Layer = 7 | Epoch = 3, Layer = 7 | Epoch = 4, Layer = 7 | Epoch = 5, Layer = 7 | Epoch = 6, Layer = 7 | Epoch = 7, Layer = 7 | Epoch = 8, Layer = 7 | Epoch = 9, Layer = 7 | Epoch = 10, Layer = 7 | Epoch = 15, Layer = 7 | Epoch = 20, Layer = 7 | Epoch = 25
Layer = 8 | Epoch = 1, Layer = 8 | Epoch = 2, Layer = 8 | Epoch = 3, Layer = 8 | Epoch = 4, Layer = 8 | Epoch = 5, Layer = 8 | Epoch = 6, Layer = 8 | Epoch = 7, Layer = 8 | Epoch = 8, Layer = 8 | Epoch = 9, Layer = 8 | Epoch = 10, Layer = 8 | Epoch = 15, Layer = 8 | Epoch = 20, Layer = 8 | Epoch = 25
Layer = 9 | Epoch = 1, Layer = 9 | Epoch = 2, Layer = 9 | Epoch = 3, Layer = 9 | Epoch = 4, Layer = 9 | Epoch = 5, Layer = 9 | Epoch = 6, Layer = 9 | Epoch = 7, Layer = 9 | Epoch = 8, Layer = 9 | Epoch = 9, Layer = 9 | Epoch = 10, Layer = 9 | Epoch = 15, Layer = 9 | Epoch = 20, Layer = 9 | Epoch = 25
Layer = 10 | Epoch = 1, Layer = 10 | Epoch = 2, Layer = 10 | Epoch = 3, Layer = 10 | Epoch = 4, Layer = 10 | Epoch = 5, Layer = 10 | Epoch = 6, Layer = 10 | Epoch = 7, Layer = 10 | Epoch = 8, Layer = 10 | Epoch = 9, Layer = 10 | Epoch = 10, Layer = 10 | Epoch = 15, Layer = 10 | Epoch = 20, Layer = 10 | Epoch = 25
Layer = 11 | Epoch = 1, Layer = 11 | Epoch = 2, Layer = 11 | Epoch = 3, Layer = 11 | Epoch = 4, Layer = 11 | Epoch = 5, Layer = 11 | Epoch = 6, Layer = 11 | Epoch = 7, Layer = 11 | Epoch = 8, Layer = 11 | Epoch = 9, Layer = 11 | Epoch = 10, Layer = 11 | Epoch = 15, Layer = 11 | Epoch = 20, Layer = 11 | Epoch = 25
Layer = 12 | Epoch = 1, Layer = 12 | Epoch = 2, Layer = 12 | Epoch = 3, Layer = 12 | Epoch = 4, Layer = 12 | Epoch = 5, Layer = 12 | Epoch = 6, Layer = 12 | Epoch = 7, Layer = 12 | Epoch = 8, Layer = 12 | Epoch = 9, Layer = 12 | Epoch = 10, Layer = 12 | Epoch = 15, Layer = 12 | Epoch = 20, Layer = 12 | Epoch = 25

Figure A8: Path-MS: BERT

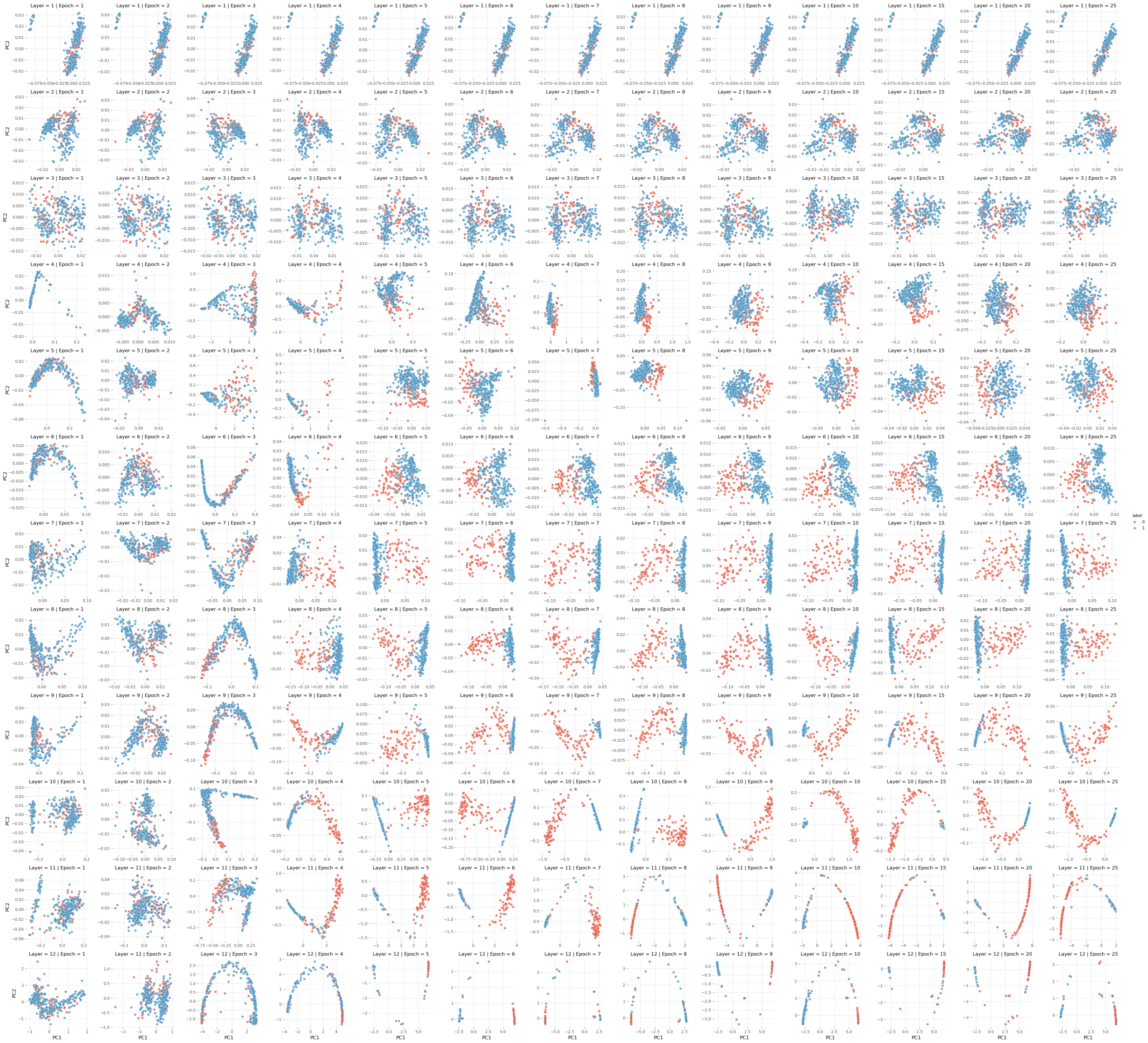

Figure A9: Path-MS: TNLR

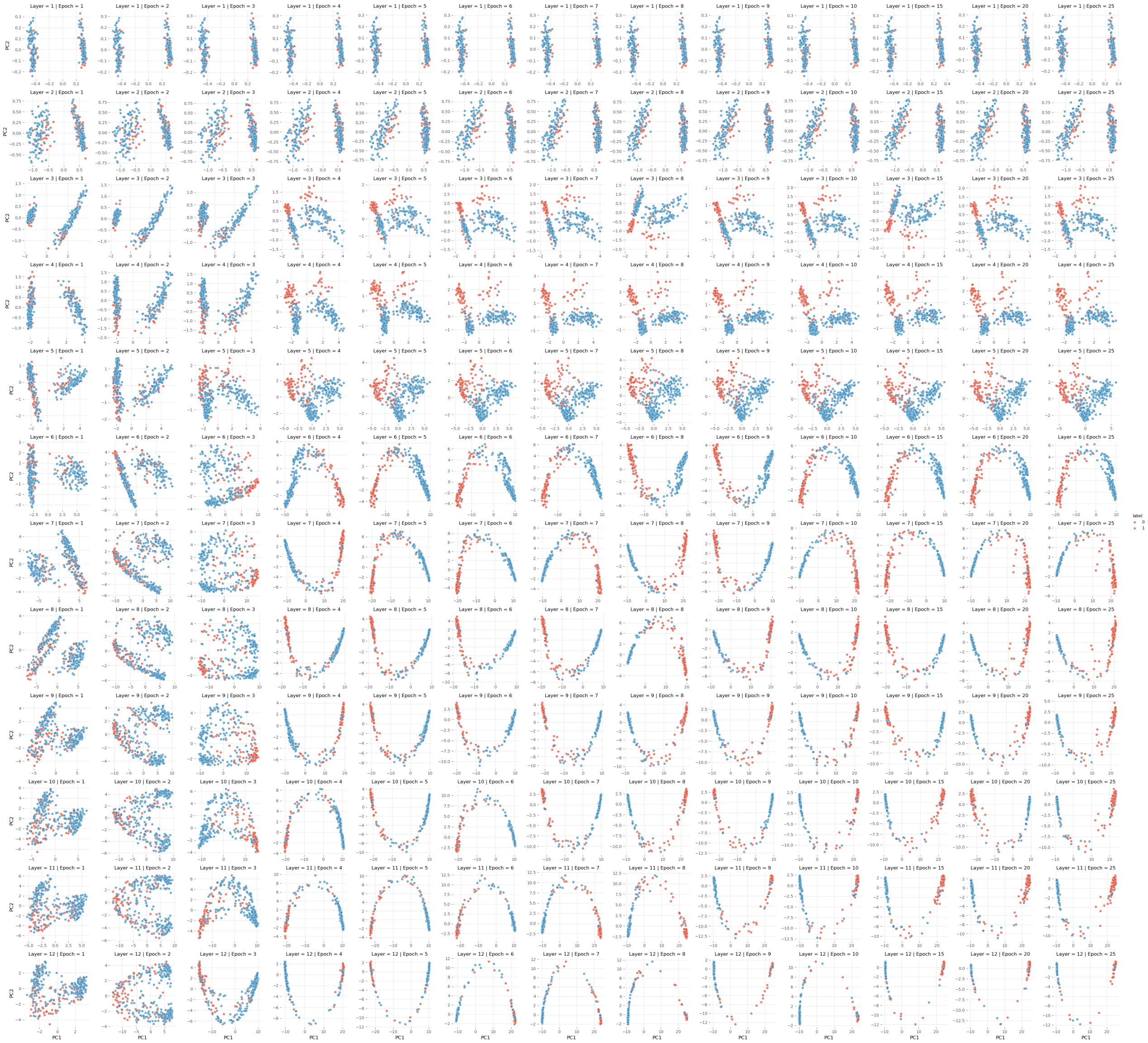

Figure A10: Path-MS: BioBERT

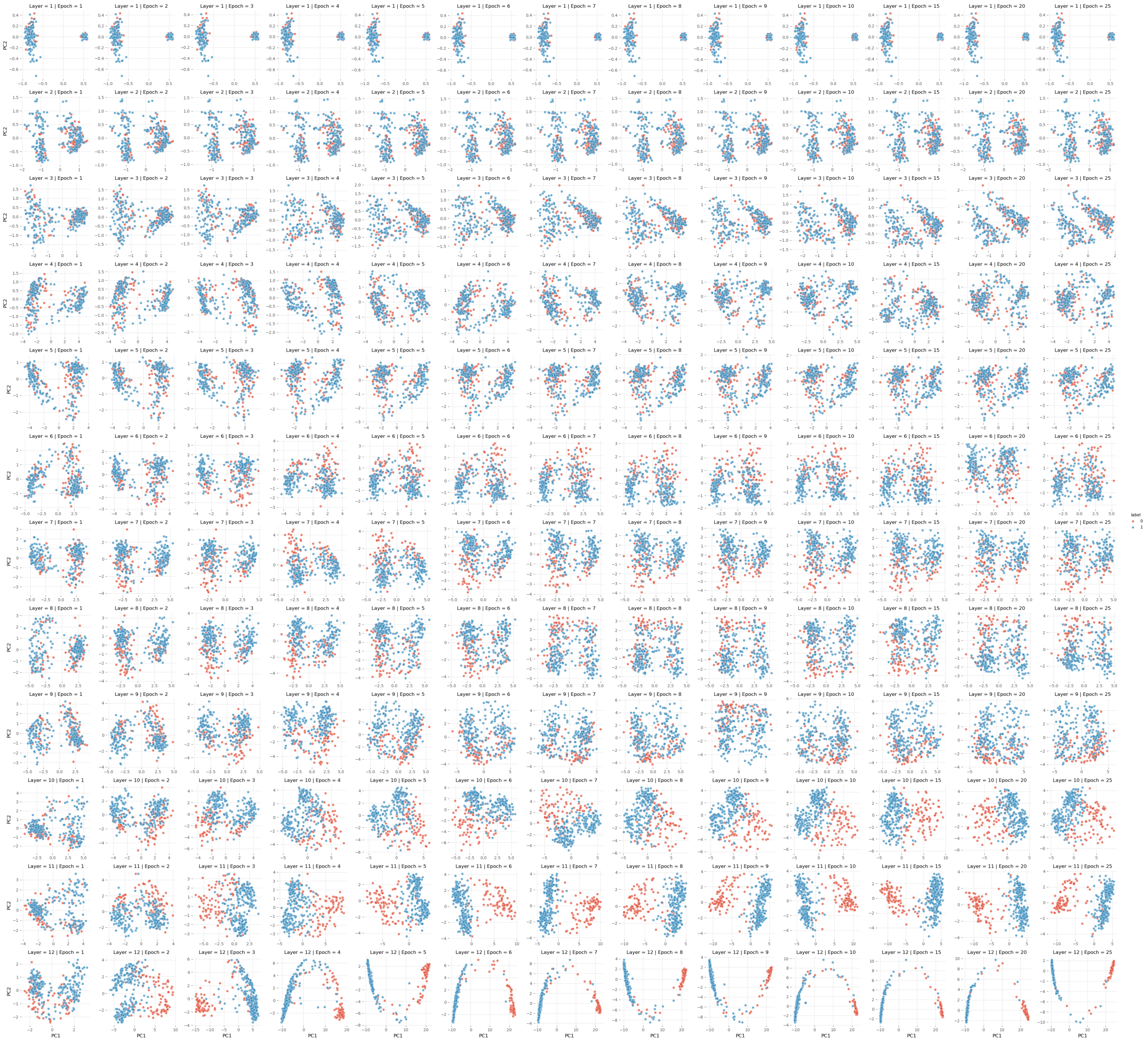

Figure A11: Path-MS: Clinical BioBERT

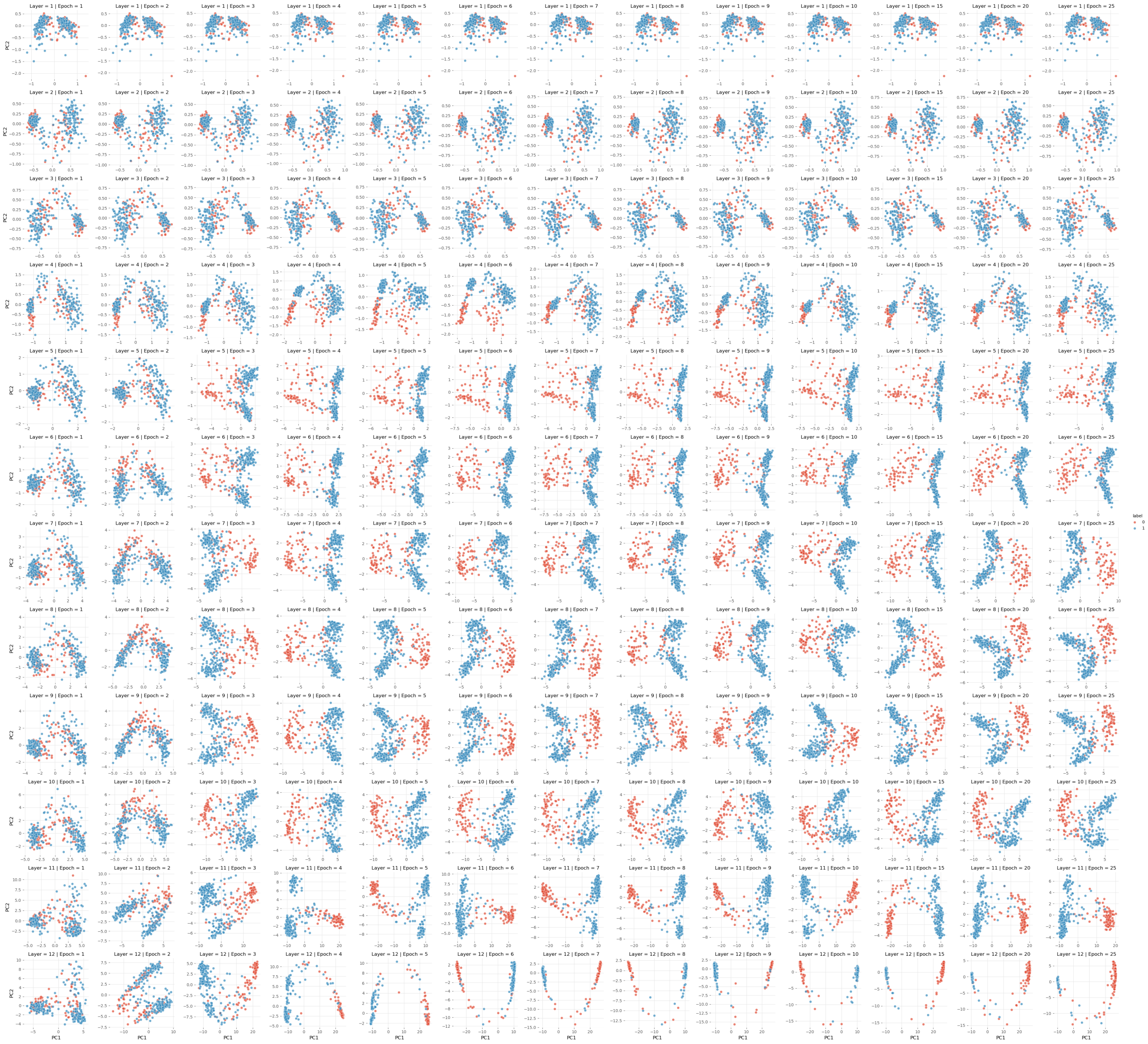

Figure A12: Path-MS: PubMedBERT