

A Appendices for “Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art”

A.1 Models and Code

Models and code will be made available at <https://github.com/facebookresearch/bio-lm>

A.2 Full Results

Table 7 shows all test results from our experiments.

Task Name	General-purpose		Specialized			Ours-base			Ours-Large		Ours-Distilled + PM + M3 + Voc.
	RoBERTa- base	RoBERTa- large	SciBERT	BioBERT	Clinical- BERT	BioMed- RoBERTa	+ PM + Voc.	+ PM + M3	+ PM + M3 + Voc.	+ PM + M3 + Voc.	
BC5CDR-C.	87.3	90.8	90.3	91.9	91.9	90.6	91.7	93.0	92.0	93.7	93.1
BC5CDR-D.	77.6	82.3	80.6	83.6	83.3	81.3	81.9	83.9	82.4	85.2	84.0
JNLPBA	79.5	80.1	80.2	80.3	80.4	79.3	80.4	80.9	80.5	81.0	80.9
NCBI-disease	84.7	87.1	86.1	86.9	87.6	86.1	88.0	88.2	87.2	89.0	88.2
BC4CHEMD	88.6	90.6	89.7	91.8	92.2	90.3	92.6	92.6	92.5	93.7	92.8
BC2GM	82.7	85.3	84.2	85.7	85.6	83.9	86.6	87.2	86.3	88.0	87.5
LINNEAEUS	79.8	87.8	84.2	84.1	86.2	84.8	85.0	85.9	82.9	88.4	85.8
Species-800	75.8	78.3	77.3	77.8	79.2	77.4	78.4	79.5	78.2	81.1	79.9
I2B2-2010	83.5	87.3	85.0	86.3	86.0	86.3	86.0	86.7	88.2	89.7	88.5
I2B2-2012	74.9	78.3	76.4	77.6	77.6	78.0	77.3	78.4	79.4	80.8	80.0
I2B2-2014	95.6	95.8	95.2	95.2	94.7	94.6	94.9	95.0	95.5	96.3	95.3
HOC	86.0	85.2	86.7	84.7	86.6	86.2	86.5	85.9	85.8	86.6	86.9
ChemProt	69.6	71.7	75.7	69.7	73.9	68.5	74.9	72.4	74.7	76.2	75.7
GAD	79.4	73.4	81.6	78.7	81.2	79.2	81.8	81.1	81.9	81.1	82.3
EU-ADR	85.0	85.0	85.0	85.5	85.0	85.1	85.0	85.0	85.0	85.0	85.0
DDI-2013	79.0	80.5	80.7	79.1	79.9	77.3	80.5	81.9	82.1	82.1	81.2
I2B2-2010-RE	72.4	75.2	75.0	69.8	74.4	74.0	74.5	73.8	76.2	78.6	75.5
MedNLI	81.4	83.3	85.1	79.7	82.5	81.8	84.7	83.2	85.7	88.5	86.8
Mean (Seq. Lab.)	82.7	85.8	84.5	85.6	85.9	84.8	85.7	86.5	85.9	87.9	86.9
Mean (Classif.)	79.0	79.2	81.4	78.2	80.5	78.9	81.1	80.4	81.6	82.6	81.9
Mean (PubMed)	81.1	82.9	83.3	83.1	84.1	82.3	84.1	84.4	84.0	85.1	84.9
Mean (Clinical)	81.6	84.0	83.3	81.7	83.0	82.9	83.5	83.4	85.0	86.8	85.2
Mean (all)	81.3	83.2	83.3	82.7	83.8	82.5	83.9	84.1	84.2	85.8	85.0

Table 7: Test set results for all our experiments. Results are medians across 5 seeds.