Supplementary Information

1 Model and Data Differences between Finetuning and Reprogramming

Figure S1 illustrates the differences between the data domains that conventional finetuning and R2DL rely on for training. In-domain training schemes, in the context of protein knowledge tasks, rely on massive databases of biological sequences such as UniProt[40] and Uniref-50[44]. In contrast, out-of-domain training schemes (R2DL) rely on widely available language data, such as[45], and only a limited number of in-domain data samples (in the context of protein knowledge tasks, biological sequences). This makes R2DL an efficient cross-domain approach to train models in low-resource data settings.

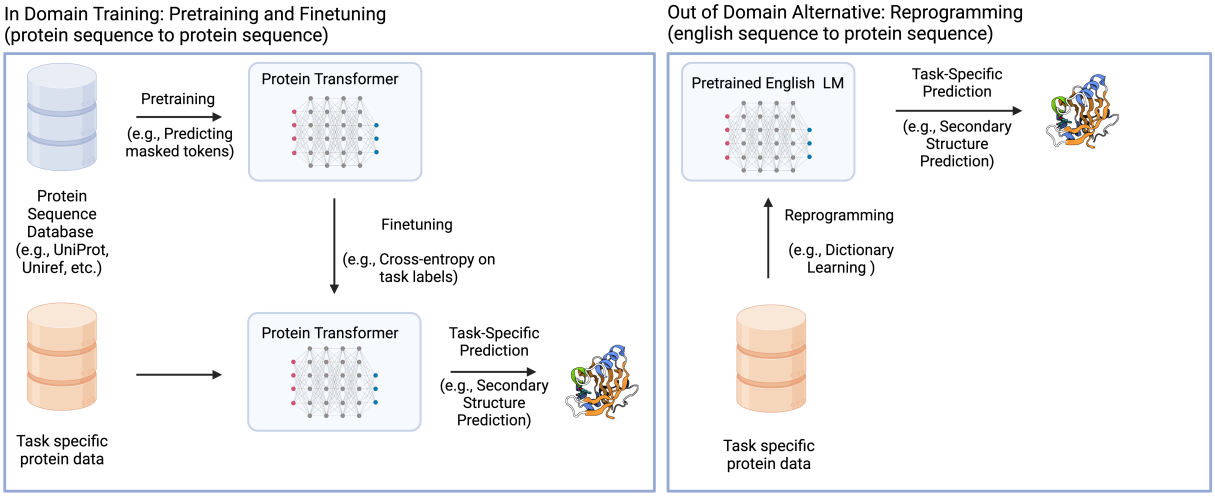


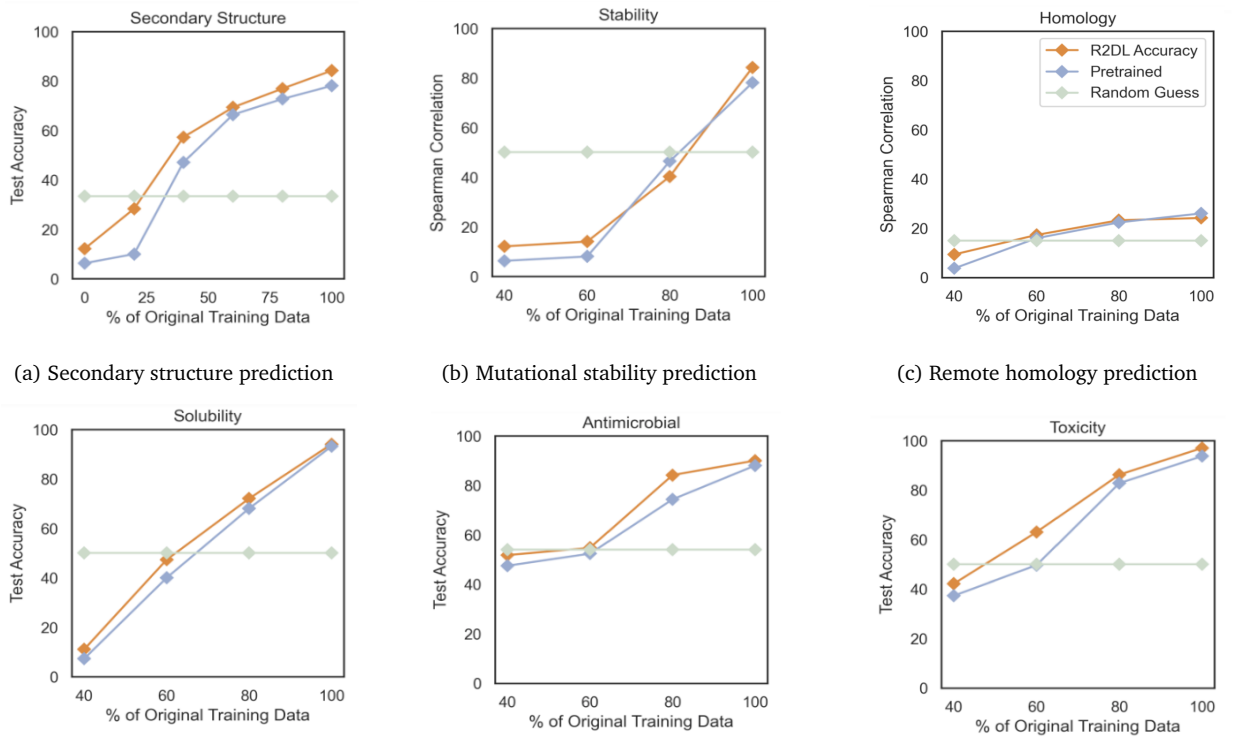
Fig. S1: In-domain pretraining vs R2DL. A typical pretrained model follows an in-domain training scheme: first the model undergoes pretraining with an objective like masked language modeling[18], and is finetuned with task-specific data. R2DL, unlike traditional pretraining schemes, operates on a small set of in-domain data, having been pretrained on out-of-domain data. In the process of R2DL, we still map between the same modalities, of sequence to sequence, as in other pretraining methods[18,20,31].

2 R2DL Performance vs. In-domain Pretraining Performance in Low Data Settings

Motivated by the data efficiency of R2DL as a framework, we tested the task-specific predictive performance of R2DL in reduced-data training settings using the pretrained English BERT model. We compared these results to the performance of task-specific baseline methods, when trained and tested in the same restricted data setting. In Figure 4, we show the performance of the R2DL model and the baselines when trained on 100%, 80%, 60%, and 40% of a specific task dataset. We show results for the Antimicrobial, Toxicity, Secondary Structure, Stability, Homology, and Solubility prediction tasks in Figure S2 and compare the performance of R2DL and in-domain pretrained models against the performance of a random guess. We observe, that for downstream tasks of Toxicity, Secondary Structure, Homology, and Solubility, R2DL always performs better than a pretrained protein language model across the size range of the limited datasets. Furthermore, we observe that, except in the stability task, the rate of failure to perform better than a random guess is higher for the in-domain pretrained models than for R2DL. In both cases, R2DL outperforms in-domain pretraining until the cutoff point which is the intersection of the random guess curve with the accuracy curves (the point at which the model is not learning any meaningful representation).

3 Baseline Model Reported Performance

For each downstream task that we train R2DL, we follow the train/test splits established by the task-specific benchmark methods. The R2DL perfomance and the task-specific benchmark model performance are reported in this section (Table S1 to Table S6).



(a) Secondary structure prediction

(b) Mutational stability prediction

(c) Remote homology prediction

(d) Membrane solubility prediction

(e) Antimicrobial-nature prediction

(f) Toxicity prediction

Fig. S2: Results of the R2DL (based on the pretrained English BERT model) and baseline methods for each downstream task in reduced training data settings. R2DL offers an extremely competitive alternative to finetuning existing pretrained approaches on each downstream protein task, achieving comparable or superior performance despite the absence of domain-specific pretraining. These results highlight R2DL's potential as a practical and resource-efficient strategy in low-data regimes where finetuning large, protein-specific models may be infeasible.

Protein Task	Source Model	Baseline Method	R2DL (out-of domain pretraining)			Pretraining (in-domain pretraining)			Train from Scratch (in-domain supervised training)		
			In-Domain Training Samples	Accuracy	Data Efficiency	In-Domain Training Samples	Accuracy	Data Efficiency	In-Domain Training Samples	Accuracy	Data Efficiency
Secondary Structure	BERT	ESM-1b	8678	0.841	9.69E-05	3.10E+07	0.800	2.58E-08	8678	62.34	7.18E-05
Stability	BERT	TAPE*	21146	0.849	4.01E-05	3.10E+07	0.738	2.38E-08	21146	65.98	3.08E-05
Homology	BERT	TAPE*	12312	0.241	1.96E-05	3.10E+07	0.265	8.55E-09	12312	24.50	1.99E-05
Solubility	BERT	TAPE*	16253	0.943	5.80E-05	1.70E+06	0.872	5.13E-07	16253	85.64	5.27E-05
Antimicrobial	BERT	PepWAE	6489	0.900	1.39E-04	1.70E+06	0.883	5.19E-07	6489	87.40	1.35E-04
Toxicity	BERT	PepWAE	8153	0.961	1.18E-04	1.70E+06	0.937	5.51E-07	8153	68.90	8.45E-05
Antibody Affinity	T5	EmiPareto	4000	0.9456	2.36E-04	1.70E+06	0.958	5.64E-07	4000	0.928	2.32E-04
Protein-Protein Interaction	T5	EDLMPPI	1368	0.852	6.23E-04	1.70E+06	0.852	5.01E-07	1368	0.433	3.2E-04

Table S1: We report the choice of source model that gives the best R2DL performance. We also report the baseline method which is the state-of-the-art performance for the given downstream protein task. *The TAPE benchmark refers to the pretrained transformer model by downstream task as reported in[11].

Attribute	Data-Split			Accuracy	
	Train	Test	Valid	Majority Class	Test
Toxicity	8153	1019	1020	0.82	0.93
Antimicrobial	6489	811	812	0.82	0.88

Table S2: Toxicity and Antimicrobial-nature baselines as reported in[9].

Task	Model	Accuracy Metric	Test Accuracy
Secondary Structure Prediction	One Hot + Alignment	Accuracy (3-class)	0.80
Remote Homology Detection	LSTM	Top 1 Accuracy	0.26
Stability	Transformer	Spearman's Rho	0.73

Table S3: Structure prediction, Remote Homology, Stability baselines as reported in [11].

Task	Model	Test Accuracy
Solubility	ProtT5-XL-UniRef50	0.91

Table S4: Solubility baselines as reported in [42]

Task	Model	Test Accuracy
Antibody Affinity	Linear Discriminant Analysis	0.92

Table S5: Antibody Affinity Binding as reported in [12].

Task	Model	Test Accuracy
Protein-Protein Interaction	Ensemble Deep Learning Model	0.858

Table S6: Protein-Protein Interaction binding site identification as reported in [13].

4 R2DL Performance with k-SVD

We study the effect of k-SVD iterations in the following tables (Table S7 to Table S12).

Source Model	Antimicrobial Sequence Samples	k-SVD Iterations	Training Accuracy	Test Accuracy
BERT	6489	100	87.12	85.64
BERT	6489	250	85.67	82.33

Table S7: R2DL: Antimicrobial Classification

Source Model	Antimicrobial Sequence Samples	k-SVD Iterations	Test Accuracy
BERT	8153	100	87.23
BERT	8153	250	86.93

Table S8: R2DL: Antimicrobial Prediction

Source Model	Training Samples	k-SVD Iterations	Training Accuracy	Test Accuracy
BERT	8,678	10000	71.47	63.65
BERT	8,678	15000	74.34	69.91
BERT	8,678	20000	76.32	74.92

Table S9: R2DL: Secondary Structure Prediction

5 R2DL Performance in Low Resource Settings

To further investigate the efficacy of the R2DL cross-domain learning approach, we compare the performance of R2DL versus models trained from scratch on task-specific protein sequences, with a restricted training data set. The test accuracy across tasks indicates that R2DL performs better when fewer labeled training data samples are available. Below 25%

Source Model	Training Samples	k-SVD Iterations	Training Accuracy	Test Accuracy
BERT	12,312	10000	11.34	10.76
BERT	12,312	15000	16.45	15.67
BERT	12,312	20000	26.23	24.50

Table S10: R2DL: Remote Homolgy Detection (Top-1 Accuracy)

Source Model	Training Samples	k-SVD Iterations	Training Accuracy	Test Accuracy
BERT	53,679	10000	60.23	61.89
BERT	53,679	15000	68.62	67.20
BERT	53,679	20000	70.78	69.73

Table S11: R2DL: Stability (Spearman's Rho)

Source Model	Training Samples	k-SVD Iterations	Training Accuracy	Test Accuracy
TinyBERT	6623	10000	68.93	69.82
TinyBERT	6623	15000	87.22	89.3
TinyBERT	6623	20000	92.85	93.21

Table S12: R2DL: Solubility

of training data samples, both methods approximately do worse than the random prediction, so we do not reduce the training data to evaluate performance after this threshold.

Task	Training Samples	R2DL Test Accuracy	Bi-LSTM Test Accuracy
Toxicity Prediction	5000	42.12	37.34
Toxicity Prediction	6000	62.98	49.62
Toxicity Prediction	7000	86.23	82.78
Toxicity Prediction	8153	89.34	93.7

Table S13: Restricted Data Setting: Toxicity Prediction

Task	Training Samples	R2DL Test Accuracy	Bi-LSTM Test Accuracy
Antimicrobial	3500	59.82	64.52
Antimicrobial	4500	72.76	68.41
Antimicrobial	5500	84.17	74.34
Antimicrobial	6489	90.01	88.0

Table S14: Restricted Data Setting: Antimicrobial

Task	Training Samples	R2DL Test Accuracy	Bi-LSTM Test Accuracy
Structure Prediction	3378	12.09	06.23
Structure Prediction	4478	34.26	37.93
Structure Prediction	6678	69.28	66.34
Structure Prediction	8678	84.14	78.0

Table S15: Restricted Data Setting: Secondary Structure Prediction (SSP)

For every downstream protein task, we perform additional analysis on the robustness of R2DL in a low-resource setting. To do this, we iteratively reduce the number of in-domain labeled samples used in training R2DL for each protein

Task	Training Samples	R2DL Test Accuracy	Bi-LSTM Test Accuracy
Homology	4312	09.35	03.69
Homology	8312	17.26	15.93
Homology	10312	23.23	22.34
Homology	12312	24.14	26.0

Table S16: Restricted Data Setting: Remote Homology Detection

Task	Training Samples	R2DL Test Accuracy	Bi-LSTM Test Accuracy
Fluorescence	10769	12.09	06.23
Fluorescence	25769	34.26	37.93
Fluorescence	45769	69.28	66.34
Fluorescence	53769	66.34	68.0

Table S17: Restricted Data Setting: Fluorescence

Task	Training Samples	R2DL Test Accuracy	Bi-LSTM Test Accuracy
Solubility	2500	011.0	07.23
Solubility	4000	47.26	39.93
Solubility	5200	85.23	87.34
Solubility	6623	94.0	93.1

Table S18: Restricted Data Setting: Solubility Prediction

prediction task. We find that in comparison to models trained from scratch on in-domain data, R2DL maintains a higher prediction accuracy for each protein task at each low resource setting. For each prediction task, we report the number of training samples in each low resource setting, the test accuracy of R2DL, and a trained-from-scratch model.

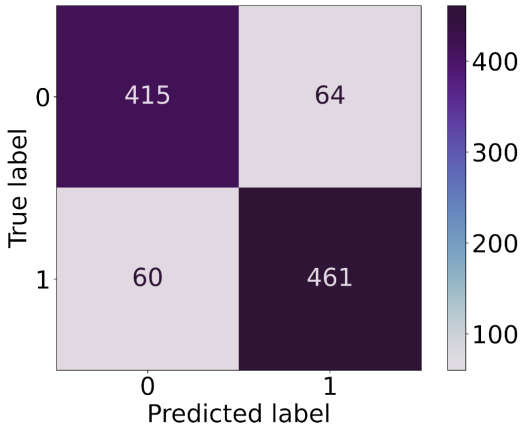
6 Classification Analysis

In Figure S3, we show a comparison between the performance of a linear discriminant analysis (LDA) model in [12] and R2DL on the antigen affinity prediction task for antibody variant sequences. The LDA model is a binary classifier that finds the optimal classification boundary by projecting the data onto a one-dimensional feature space and finding a threshold. The antibody affinity dataset consists of 4,000 labeled protein sequences, with labels {1 (on-target binding), 0 (off-target binding)}. R2DL achieves a predictive accuracy of 95.5% compared to the LDA model performance of 92.8%. R2DL achieves a higher predictive accuracy than the baseline LDA model by 3% and with a higher classification accuracy with imbalanced datasets. The antibody affinity task dataset has the following distribution on target: 1516, off-target: 2484. For 37% to 62% class-imbalance ratio of labels, we show that the R2DL model has a better classification accuracy than the LDA model. The learned representations can therefore be inferred to be more accurate in our model than in the baseline model. This is important, as in many real-world prediction tasks, the dataset is found to be class-imbalanced.

7 Additional Results of R2DL

Table S19 shows the number of training samples, the accuracy metric (mean and standard deviation with 5 independent runs), and the data efficiency of R2DL, pretrained, and supervised models. It shows the sensitivity analysis of R2DL in terms of the error bars on all 8 downstream tasks. We find that tasks with fewer in-domain training samples are more sensitive as compared to protein tasks with an order of magnitude more in-domain training samples (see Table S19).

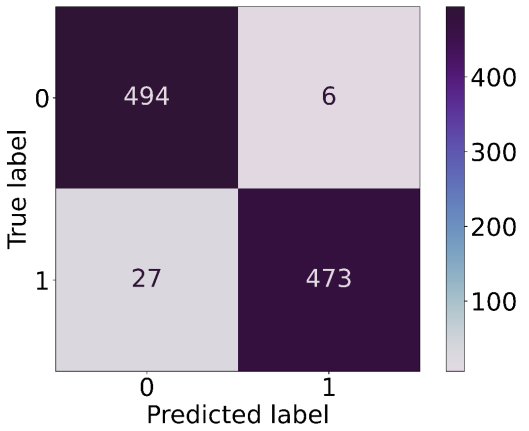
Table S20 compares the performance of R2DL with different pretrained English models, including BERT [18], roBERTa [31], T5 [20], and PubMedBERT [46]. In general, we find that using a larger-sized general-purpose large language model (e.g. T5 vs BERT) trained on web-scale text data can further improve three out of the eight considered protein sequence learning tasks, while a language model trained on domain-relevant corpora (e.g. PubMedBERT) has less benefit. We hypothesize that better source English language models enable the learned dictionary to capture more fine-grained representations of amino acid sequence distributions in the downstream protein task dataset. This hypothesis is also consistent with the theoretical justification in [28] that a more advanced source model can lead to a smaller error upper bound in the considered



$$\text{False Positive Rate}_{LDA} = \frac{\text{False Negative}}{\text{False Negatives + True Positives}} = \frac{60}{60+461} = 1$$

$$\text{True Positive Rate}_{LDA} = \frac{\text{True Positives}}{\text{False Negatives + True Positives}} = \frac{461}{60+461} = 8$$

(a) Confusion matrix of the baseline model trained in[12] for the antibody affinity prediction task.



$$\text{False Positive Rate}_{R2DL} = \frac{\text{False Negative}}{\text{False Negatives + True Positives}} = \frac{27}{27+473} = 5.4\,\%$$

$$\text{True Positive Rate}_{R2DL} = \frac{\text{True Positives}}{\text{False Negatives + True Positives}} = \frac{473}{27+473} = 94.6\,\%$$

(b) Confusion matrix of the R2DL model for the antibody affinity prediction task.

Fig. S3: R2DL performance on the antigen affinity prediction task for antibody variant sequences and its comparison with the baseline Linear Discriminant Analysis model reported in[12].

model reprogramming loss.

Table S21 summarizes the performance of the considered protein sequence learning tasks reported in the literature. Table S22 compares R2DL to conventional finetuning methods on the same pretrained English language model. The significant and consistent performance improvement in all tasks observed in R2DL over these methods demonstrates the effectiveness of R2DL as an efficient cross-domain finetuning method.

Protein Task	R2DL (out-of-domain pretraining)			Pretraining (in-domain pretraining)			Train from Scratch (in-domain supervised training)		
	In-Domain Training Samples	Accuracy	Data Efficiency	In-Domain Training Samples	Accuracy	Data Efficiency	In-Domain Training Samples	Accuracy	Data Efficiency
Secondary Structure	8678	0.841 ± 0.218	9.69E-05	3.10E+07	0.801 ± 0.035	2.58E-08	8678	0.623 ± 0.139	7.18E-05
Stability	21146	0.849 ± 0.141	4.01E-05	3.10E+07	0.738 ± 0.028	2.38E-08	21146	0.659 ± 0.0842	3.08E-05
Homology	12312	0.241 ± 0.129	1.96E-05	3.10E+07	0.265 ± 0.019	8.55E-09	12312	0.245 ± 0.285	1.99E-05
Solubility	16253	0.943 ± 0.087	5.80E-05	1.70E+06	0.872 ± 0.046	5.13E-07	16253	0.856 ± 0.303	5.27E-05
Antimicrobial	6489	0.902 ± 0.042	1.39E-04	1.70E+06	0.883 ± 0.112	5.19E-07	6489	0.874 ± 0.097	1.35E-04
Toxicity	8153	0.961 ± 0.018	1.18E-04	1.70E+06	0.937 ± 0.175	5.51E-07	8153	0.689 ± 0.273	8.45E-05
Antibody Affinity	4000	0.9456 ± 0.134	2.36E-04	1.70E+06	0.958 ± 0.088	5.64E-07	4000	0.928 ± 0.171	2.32E-04
Protein-Protein Interaction	1368	0.852 ± 0.025	6.23E-04	1.70E+06	0.852 ± 0.113	5.01E-07	1368	0.433 ± 0.261	3.2E-04

Table S19: R2DL (out-of-domain pretraining) versus Pretraining (in-domain pretraining) performance. In-domain pretraining leverages learned specific features from the protein sequences. Out-of-domain pretraining (R2DL) leverages biologically relevant grammar, which boosts performance when applied on a downstream task. R2DL results here are reported with the source model that resulted in the highest downstream task accuracy. Pretraining performance is reported with the highest accuracy pretrained model, finetuned on the in-domain training samples available for each protein task. The details on the choice of source model and protein language model are described in Appendix 7. In-domain samples are the total number of amino acid sequences used in training (including pretraining, supervised training, or finetuning processes). Data efficiency is defined as the ratio of the downstream protein task accuracy to the number of in-domain training samples.

Pretrained English Model (Source Model)	Model Parameters		R2DL Accuracy for Downstream Protein Task							
	Source Model	R2DL	Secondary Structure	Stability	Homology	Solubility	Antimicrobial	Toxicity	Antibody Affinity	Protein-Protein Interaction
BERT	110M	96M	0.841	0.849	0.241	0.943	0.900	0.961	0.9456	0.852
roBERTa	123M	96M	0.899	0.826	0.210	0.899	0.879	0.978	0.893	0.793
T5	220M	96M	0.879	0.724	0.315	0.9467	0.832	0.941	0.934	0.818
PubMedBERT	110M	96M	0.821	0.651	0.218	0.823	0.841	0.892	0.794	0.724

Table S20: R2DL prediction accuracy by protein downstream task. We report the performance of each instance of R2DL, when reprogramming a different source models.

Protein Language Model	Pretraining Corpus Size	Model Parameters	Downstream Protein Task Prediction Accuracy							
			Secondary Structure	Stability	Homology	Solubility	Antimicrobial	Toxicity	Antibody Affinity	Protein-Protein Interaction
LSTM (TAPE)	31M	38M	84.1	84.9	24.1	94.3	—	—	—	—
ESM1-b	250M	650M	71.6	—	—	—	—	—	—	—
ProtBERT	45M	420M	—	—	—	—	—	—	—	—
PepWAE	—	1.7M	—	—	—	—	88.0	93.7	—	—
EDLMPPI	—	—	—	—	—	—	—	—	—	0.858
ProtT5	45M	3B	—	0.81	—	—	0.91	—	—	0.852
EmiPareto	—	—	—	—	—	—	—	—	0.93	—

Table S21: Protein Downstream Task Baselines, as reported in[47]. For the protein language models we consider as baselines, where publicly reported, we report the size of the pretraining corpus (number of amino acids), the number of model parameters, and the prediction accuracy for each downstream task as reported by the individual models. Where some protein language models are not benchmarked (reported) for certain protein tasks, we leave out the accuracy and mark the result as "–" (which means not available).

Method	Downstream Protein Task Prediction Accuracy							
	Secondary Structure	Stability	Homology	Solubility	Antimicrobial	Toxicity	Antibody Affinity	PPI
R2DL	0.841	0.849	0.241	0.943	0.900	0.961	0.9456	0.852
Partial Finetuning	0.825	0.782	0.263	0.768	0.872	0.765	0.826	0.791
Linear Head	0.673	0.673	0.192	0.521	0.851	0.723	0.457	0.823

Table S22: Comparison to conventional finetuning alternatives. We show the results of a standardized instance of R2DL, where BERT is the source model that is reprogrammed for the downstream protein tasks. We report the performance for partial finetuning on BERT, where the last layer 4 layers are finetuned on the in-domain data set for the downstream protein task. We also report the prediction accuracy for a linear head trained on BERT embeddings. We demonstrate that R2DL outperforms both alternatives to conventional finetuning across all downstream protein prediction tasks.