Appendix

A Additional training details

Code was written with PyTorch Lightning. All model training was executed on NVIDIA Tesla V100S GPUs with 32 GB of memory. In the multimodal model, we used 8 GPUs for pretraining, 8 GPUs for full finetuning, and 4 GPUs for linear finetuning. For additional efficiency, we use Lightning's automatic mixed-precision training. Pretraining to 2000 epochs took around 7-10 days (depending on the exact pretraining strategy). To run 200 epochs of full finetuning and linear finetuning on 500 patients in the training set, it took around 6 hours and 4 hours, respectively.

B Comparison to prior baselines in the PhysioNet18 dataset

How does our model fare compare to other works that have used this dataset? An exact comparison is difficult since prior works with the PhysioNet18 dataset use different splits and different combinations of modalities or channels. However, we discuss a few examples here, all of which are concerned with the sleep staging task.

Banville et al. [2021] use the F3-M2 and F4-M1 channels of EEG in a model pretrained with contrastive learning. The authors were also interested in limited training data. They found that, with 595 patients in the training data, the balanced accuracy achieved by their model on their test set was 72.3%. Phan et al. [2021] use one channel each of EEG, EOG, and EMG. They train bidirectional RNNs on 944 patients and report the 5-fold cross validated score as a Cohen's Kappa score of 0.847. Perslev et al. [2019] also use 944 patients and report the 5-fold cross validated score as a F1 score of 0.77. We note that, in our hands, MultiMAE + input modality drop trained on 500 patients achieve a F1 score of 0.72 on our test set. Further examples from other works can be found in [Phan and Mikkelsen, 2022].

C Selecting pretraining hyperparameters

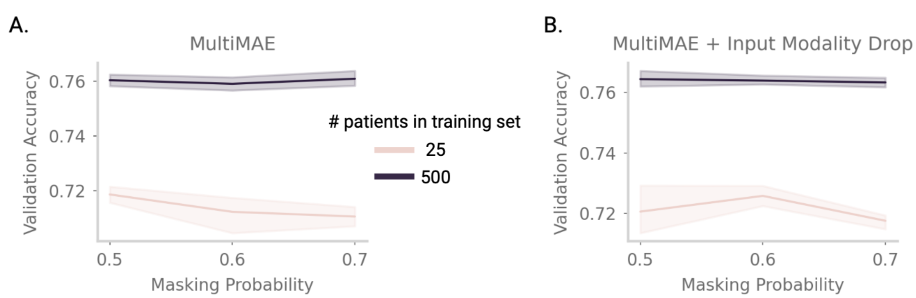

Figure 3: Hyperparameter selection in MAE models. **A.** Validation set accuracy score in the sleep staging task, with full-finetuning. Here, we show models pretrained with only MultiMAE-only. The x-axis shows the masking probability. **B.** As in (A), but for the model pretrained with MultiMAE and input modality drop.

We show validation scores of the MultiMAE model (Figure 3A) and the MultiMAE + input modality drop model (Figure 3B).

We begin with selecting the masking ratio for the model with input modality drop. If the masking ratio is p the overall masking ratio is $\frac{1}{N} + \frac{N-1}{N}p$, where N is the number of modalities. This expression arises from the modality dropping that occurs at each batch. By examining Figure 3B, it appears that using a masking probability of 0.6 is best, although this difference in performance across masking ratios is only visible in the low data regime (i.e., 25 patients in the training set). Thus, the overall masking ratio is 0.7.

Thus we select a masking ratio of 70% for the MultiMAE model (Figure 3A) as it allows for clear comparison with MultiMAE + input modality drop. We also note that, with 500 patients in the training set, the choice of masking probability does not clearly affect the downstream task performance of the MultiMAE model. Thus, both models have the same amount of tokens masked during pretraining, with the only difference due to the distribution of masking across tokens.

For visualization of the pretraining performance, we show example reconstructions made by the MultiMAE model on two samples from the training set (Fig 4). The model clearly struggles the most with reconstructing the EMG signal (this is also reflected in the mean squared error values, although those are not shown here).

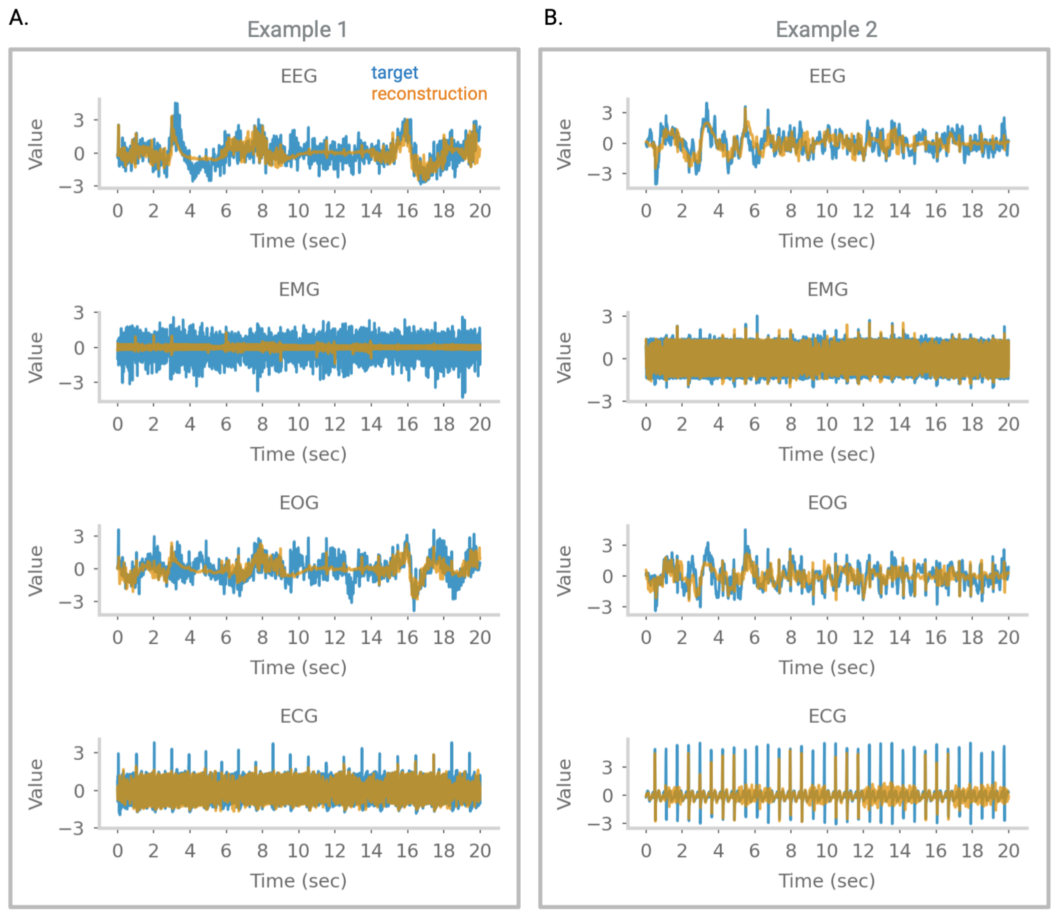

Figure 4: Reconstruction performance of MultiMAE model with 70% masking. **A.** A random sample from the training data, with target signals in blue and reconstructed signals in orange. Plot is truncated at 20 seconds for visualization purposes. **B.** As in (A), but for another random sample

D Additional results: limited training data during finetuning

Table 3: *Unimodal vs multimodal performance, with limited training data.* As in Table 1, but with 25 patients in the training data set for each of the three classification tasks.

		Sleep	Age	Arousal	Aggregate
Random	–	0.2	0.5	0.5	0.0
Pretrained	EEG	0.637 ± 0.039	**0.616 ± 0.007**	0.52 ± 0.077	0.819 ± 0.096
	EMG	0.332 ± 0.012	0.524 ± 0.026	0.512 ± 0.017	0.244 ± 0.014
	EOG	0.635 ± 0.004	0.605 ± 0.005	0.595 ± 0.05	0.859 ± 0.049
	ECG	0.255 ± 0.006	0.545 ± 0.021	0.511 ± 0.022	0.129 ± 0.008
	All	**0.688 ± 0.002**	0.571 ± 0.004	**0.597 ± 0.062**	**0.925 ± 0.067**

E Additional results: full finetuning

Table 4: *Unimodal vs multimodal performance, with full-finetuning.* As in Table 1, but parameters of the encoder are also finetuned along with training of the classification head.

		Sleep	Age	Arousal	Aggregate
Random	–	0.2	0.5	0.5	0.0
Pretrained	EEG	**0.747 ± 0.003**	0.656 ± 0.01	0.58 ± 0.02	1.069 ± 0.013
	EMG	0.457 ± 0.009	0.618 ± 0.006	0.562 ± 0.006	0.549 ± 0.018
	EOG	0.733 ± 0.001	0.637 ± 0.005	0.581 ± 0.011	1.033 ± 0.009
	ECG	0.341 ± 0.01	0.67 ± 0.018	0.55 ± 0.014	0.382 ± 0.01
	All	0.746 ± 0.005	**0.694 ± 0.009**	**0.588 ± 0.023**	**1.098 ± 0.015**

F Contrastive Pretraining

Here, we give more details of the contrastive pretraining strategy.

A.

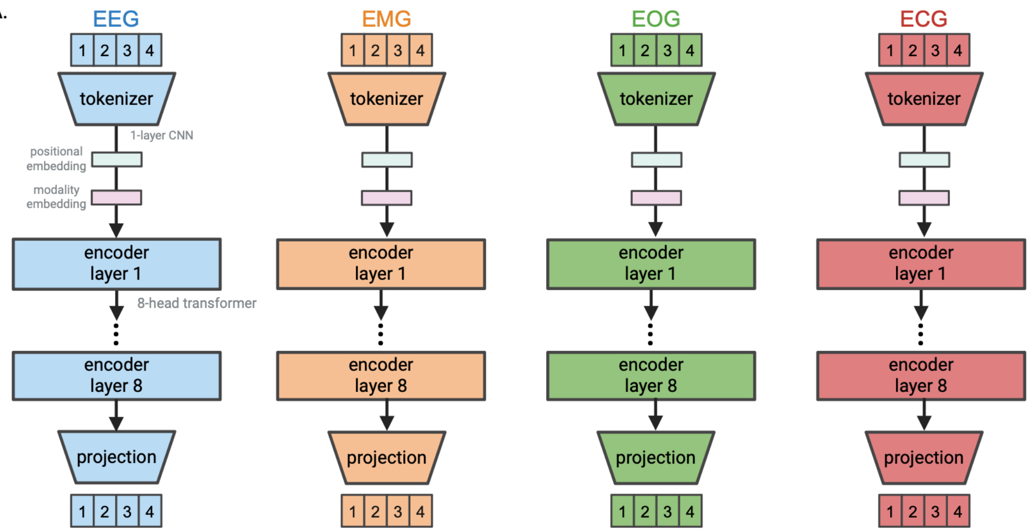

Figure 5: Contrastive learning architecture.

F.1 SimCLR-style

We use the loss functions introduced in Raghu et al. [2022]. We note a few differences from our implementations and that of Raghu et al. [2022]. One is that we use a transformer architecture to allow for comparisons with the MultiMAE models we tested. We also do not have structured data or static features. We use the same fixed temperature as in Raghu et al. [2022]. Furthermore, we also add a MLP projection head before the embeddings are passed to the contrastive loss. As before, the encoder outputs are 512-dimensional. The projection head consists of a hidden layer of dimension 256 before projection into 128 dimensions. For a given data sample, the representation we use for contrastive learning is the concatenated representations across the four modalities. Specifically, the representation is the concatenation of the output of the four projection layers. These are the representations used in the similarity calculations. Finally, As in Raghu et al. [2022], the projection head is discarded after pretraining.

F.2 CLIP-style

We use the loss functions introduced in the SleepFM paper of Thapa et al. [2024]. The architecture we use is the same as in the SimCLR-style models. Besides the difference of our transformer architecture, another difference between our implementation and that of Thapa et al. [2024] is that we use (as in the SimCLR-like model) a fixed temperature parameter and MLP projection head. We found the use of a fixed temperature and projection head led to better validation set performance in the downstream tasks, which is why we introduce these extra details.

G Raw attention matrices

Raw attention matrices of the models shown in Figure 2A-C. These matrices correspond to W_l in the expression for attention rollut given in §4.4 The matrix for each layer is averaged over the 8 heads.

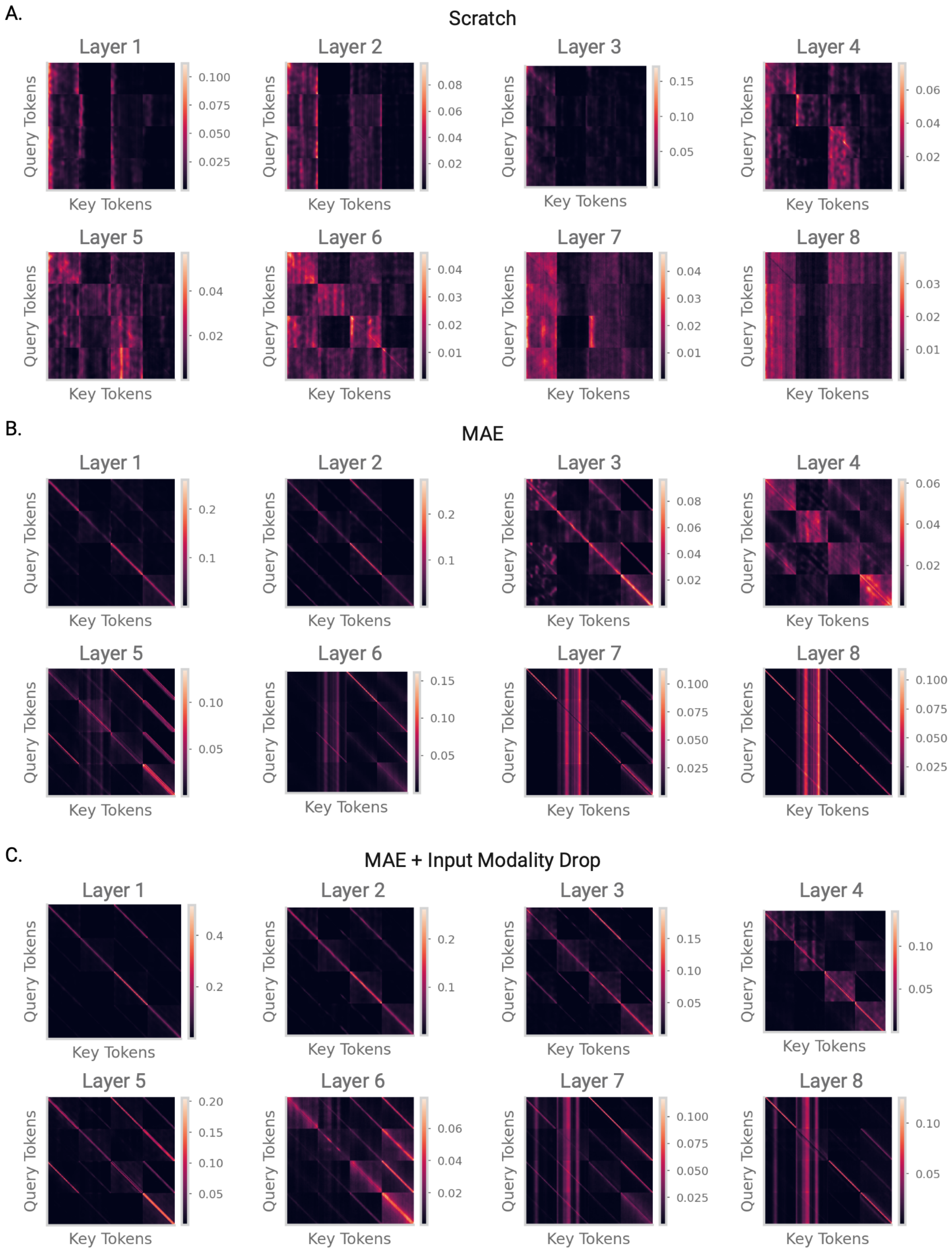

Figure 6: Raw attention matrices.

H Additional Evaluation Scores

Same as in Table 2, but with additional metrics.

Table 5: Sleep

Pretraining Strategy	Balanced Acc.	Cohen Kappa	F1
Contrastive CLIP-style (LOO)	0.708 ± 0.0	0.572 ± 0.001	0.67 ± 0.001
Contrastive CLIP-style (Pairwise)	0.703 ± 0.001	0.559 ± 0.001	0.658 ± 0.001
Contrastive SimCLR-style	0.656 ± 0.001	0.52 ± 0.001	0.632 ± 0.001
MultiMAE + Modality Drop	**0.744 ± 0.001**	**0.63 ± 0.002**	**0.718 ± 0.002**

Table 6: Age

Pretraining Strategy	Balanced Acc.	AUROC	F1
Contrastive CLIP-style (LOO)	0.643 ± 0.004	0.705 ± 0.006	0.655 ± 0.004
Contrastive CLIP-style (Pairwise)	0.646 ± 0.0	0.698 ± 0.0	0.655 ± 0.0
Contrastive SimCLR-style	0.624 ± 0.009	0.673 ± 0.015	0.635 ± 0.009
MultiMAE	0.684 ± 0.001	0.758 ± 0.001	0.694 ± 0.001
MultiMAE + Modality Drop	**0.719 ± 0.002**	**0.785 ± 0.002**	**0.728 ± 0.001**

Table 7: Arousal

Pretraining Strategy	Balanced Acc.	AUROC	F1
Contrastive CLIP-style (LOO)	**0.71 ± 0.002**	**0.776 ± 0.001**	0.638 ± 0.002
Contrastive CLIP-style (Pairwise)	0.708 ± 0.002	0.772 ± 0.001	0.627 ± 0.002
Contrastive SimCLR-style	0.585 ± 0.048	0.616 ± 0.07	0.524 ± 0.027
MultiMAE	0.604 ± 0.089	0.638 ± 0.136	0.613 ± 0.172
MultiMAE + Modality Drop	0.637 ± 0.081	0.677 ± 0.128	**0.641 ± 0.139**