Supplementary Materials: Comparative Analysis of Image-Based Deep Learning and Genomic Models for Yield and Protein Content Prediction in Winter Wheat

S1. Year-Specific Analysis

We extended year-specific analysis to include evaluation on a single-year using baseline methods trained with image time-series data of all available years. We summarize the results are summarized in Table ST1 and ST2.

On the task of grain yield prediction, we observe slightly worse performance on the years 2017, 2021, and 2022, and modest improvements on 2016 and 2018. These changes in performance are relatively small, suggesting that aggregating data across years does not drastically affect within-year generalization. However, the most notable observation is that models trained on all years significantly outperform those trained on a single year in the Test E and Test G+E splits. This indicates that including a broader range of environmental variation in the training data is crucial for improving generalization to unseen environments.

Oh the other hand, for the task of grain protein content prediction, incorporating greater variation in both environments and genotypes leads to improved performance in terms of both correlation and MAPE. Specifically, we observe significant improvements in the years 2016, 2021, and 2022 on the Test G and Test G+E sets, and moderate improvement on the Test E set.

Years	Test G		Test E		Test G+E	
	Corr↑	MAPE↓	Corr↑	MAPE↓	Corr↑	MAPE↓
2016	0.492	10.413	/	/	/	/
2017	0.100	7.092	/	/	/	/
2018	**0.588**	**5.237**	/	/	/	/
2019*	/	/	0.336	7.658	0.419	9.990
2021	0.350	8.276	/	/	/	/
2022	0.432	8.952	/	/	/	/

Table ST1. **Yield prediction with the best-performing baseline model (ConvNext w/o height).** Models are obtained by training with image time series of all the years of 2016, 2017, 2018, 2021 and 2022 and tested on one of the years, including year 2019 as a test set, as shown in Test E and Test G+E. * The year of 2019 includes unseen environment and unseen genotype and is not included in the training set or used for model training.

S2. Benchmarking Results with Test P

We further evaluate the image-based methods on Test P. The training procedure is described in Sec. 3.6. Since Test P contains seen genotypes and environments, we assume its image time series are drawn from the same distribution as the training and validation splits. Therefore, it is used to assess the models with parameters tuned on the validation set. On Test P, ConvNeXt remains the best-performing model. Overall, the performance of the baseline methods on Test P is not substantially higher than on the other test splits, indicating that the models can generalize to unseen genotypes and environments, although the tasks remain challenging.

Years	Test G		Test E		Test G+E	
	Corr↑	MAPE↓	Corr↑	MAPE↓	Corr↑	MAPE↓
2016	0.465	4.779	/	/	/	/
2018	0.718	**3.610**	/	/	/	/
2019*	/	/	0.200	9.688	0.407	6.458
2021	0.469	4.285	/	/	/	/
2022	**0.782**	4.329	/	/	/	/

Table ST2. **Protein content prediction with the best-performing baseline model (ConvNext w/o height).** Models are obtained by training with image time series of all the years of 2016, 2018, 2021 and 2022 and tested on one of the years, including year 2019 as a test set, as shown in Test E and Test G+E. * The year of 2019 includes unseen environment and unseen genotype and is not included in the training set or used for model training.

S3. interpretability Analysis

S3.2 Temporal dropout

We presented the full visualization of each year for grain yield prediction and protein content prediction as described in Sec. 4.4 in Figure SF3 and SF4.

S3.1 Visualization of Attention Weights

As attention maps learned by transformers are often used to interpret interactions between input entries, we employ them as an additional approach to analyze the temporal information captured by the image-based models. Specifically, we visualize the attention weights from the aggregation model, using ConvNeXt as the image embedding model. We extract the attention weights from all transformer layers, resulting in tensors of shape (B, N, H, T, T), where B is the batch size, N is the number of transformer encoding layers, H is the number of attention heads, and T is the number of available time points. We then average the weights across the B, N, and H dimensions to obtain a mean attention map of shape (T, T). The attention maps for each year, for both yield and protein content prediction, are shown in Figure SF1 and SF2.

Modality	Models	Test P		Test G		Test E		Test G+E	
		Corr↑	MAPE↓	Corr↑	MAPE↓	Corr↑	MAPE↓	Corr↑	MAPE↓
	ResNet-50 w/o height	-0.028	13.410	-0.173	13.124	0.188	17.701	0.194	19.263
Image time series	ConvNeXt w/o height	**0.438**	**7.593**	**0.392**	7.994	**0.336**	**7.658**	**0.419**	9.990
	DINOv2 w/o height	0.298	8.350	**0.392**	**7.794**	0.245	13.851	0.237	13.855
Image time series	ResNet-50 w/ height	0.090	13.226	-0.172	14.003	0.025	13.627	-0.008	13.904
+	ConvNeXt w/ height	0.273	9.401	0.375	10.399	0.323	7.697	0.296	**9.985**
plant heights	DINOv2 w/ height	0.219	9.584	0.288	9.201	0.125	23.145	0.013	21.436

Table ST3a: **Yield prediction with baseline methods.** Test P: seen genotypes and seen environment during training, Test G: unseen genotypes during training, (b) Test E: unseen environments during training, and (c) Test G+E: unseen genotypes in unseen environments during training.

Modality	Models	Test P		Test G		Test E		Test G+E	
		Corr↑	MAPE↓	Corr↑	MAPE↓	Corr↑	MAPE↓	Corr↑	MAPE↓
	ResNet-50 w/o height	-0.001	8.727	0.133	9.534	0.107	6.543	-0.191	5.865
Image time-series	ConvNeXt w/o height	**0.455**	**4.308**	**0.608**	**4.251**	0.200	9.688	0.407	6.458
	DINOv2 w/o height	0.291	4.840	0.537	4.784	0.258	6.471	0.115	5.066
Image times-series	ResNet-50 w/ height	0.194	10.465	-0.079	11.347	-0.131	16.939	-0.175	19.836
+	ConvNeXt w/ height	0.398	4.957	**0.608**	4.806	**0.360**	7.239	**0.440**	**4.016**
plant heights	DINOv2 w/ height	0.292	5.745	0.461	5.812	0.206	**6.250**	0.013	5.395

Table ST3b: **Protein content prediction with baseline methods.** Test P: seen genotypes and seen environment during training, Test G: unseen genotypes during training, Test E: unseen environments during training, and Test G+E: unseen genotypes in unseen environments during training.

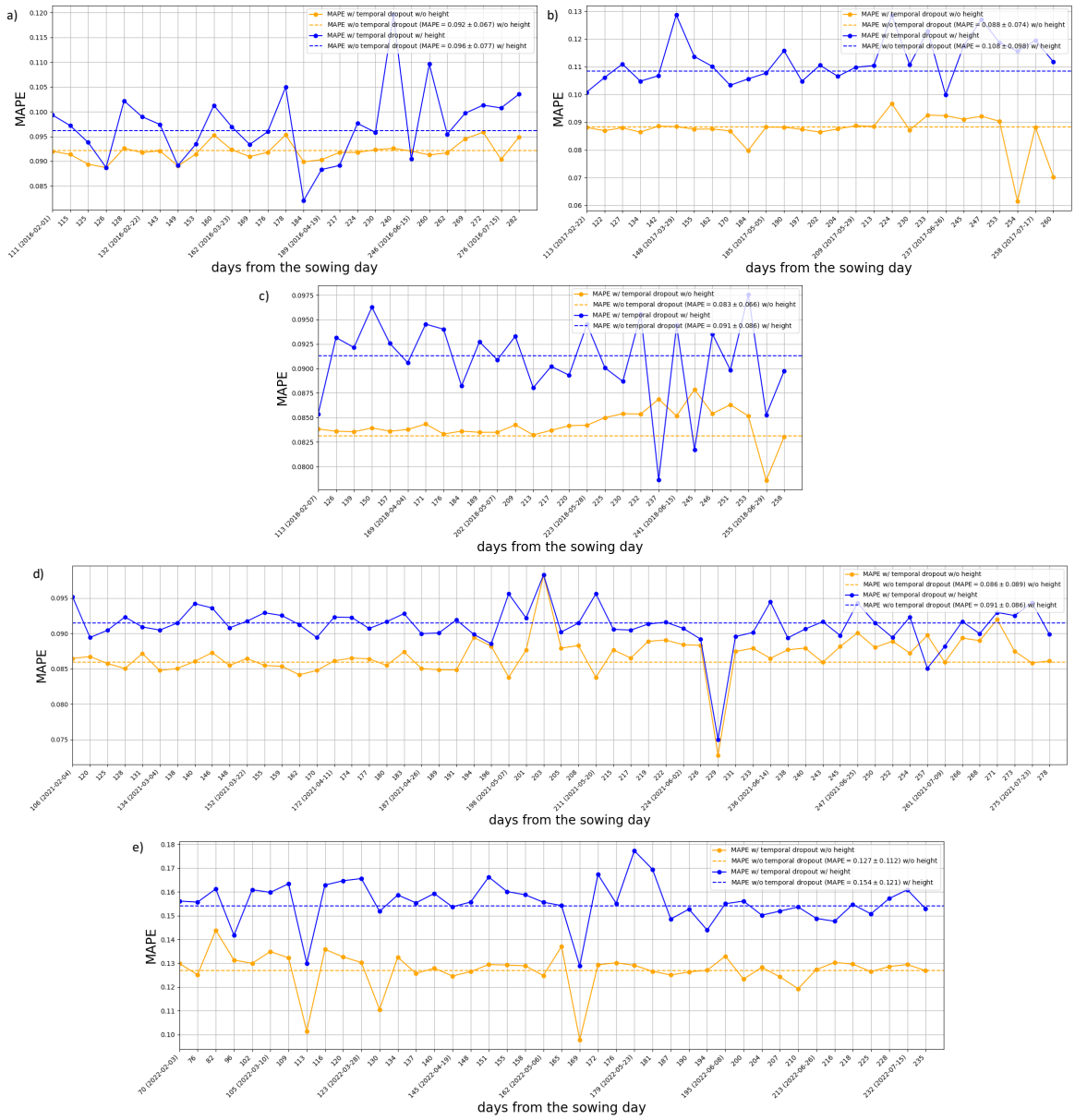

Figure SF3. **Visualizing importance of time points for grain yield prediction by temporal dropout.** The temporal importance is evaluated with the best-performing image time-series model of ConvNext for the years of 2016 (a), 2017 (b), 2018 (c), 2021(d) and 2022(e). The x-axis represents the number of days since sowing, with specific calendar dates displayed at intervals to indicate the corresponding time of year. The y-axis shows MAPE values for different configurations: ConvNeXt without plant heights (yellow) and ConvNeXt with plant heights (blue). Horizontal lines mark the MAPE values for the two models without temporal dropout.

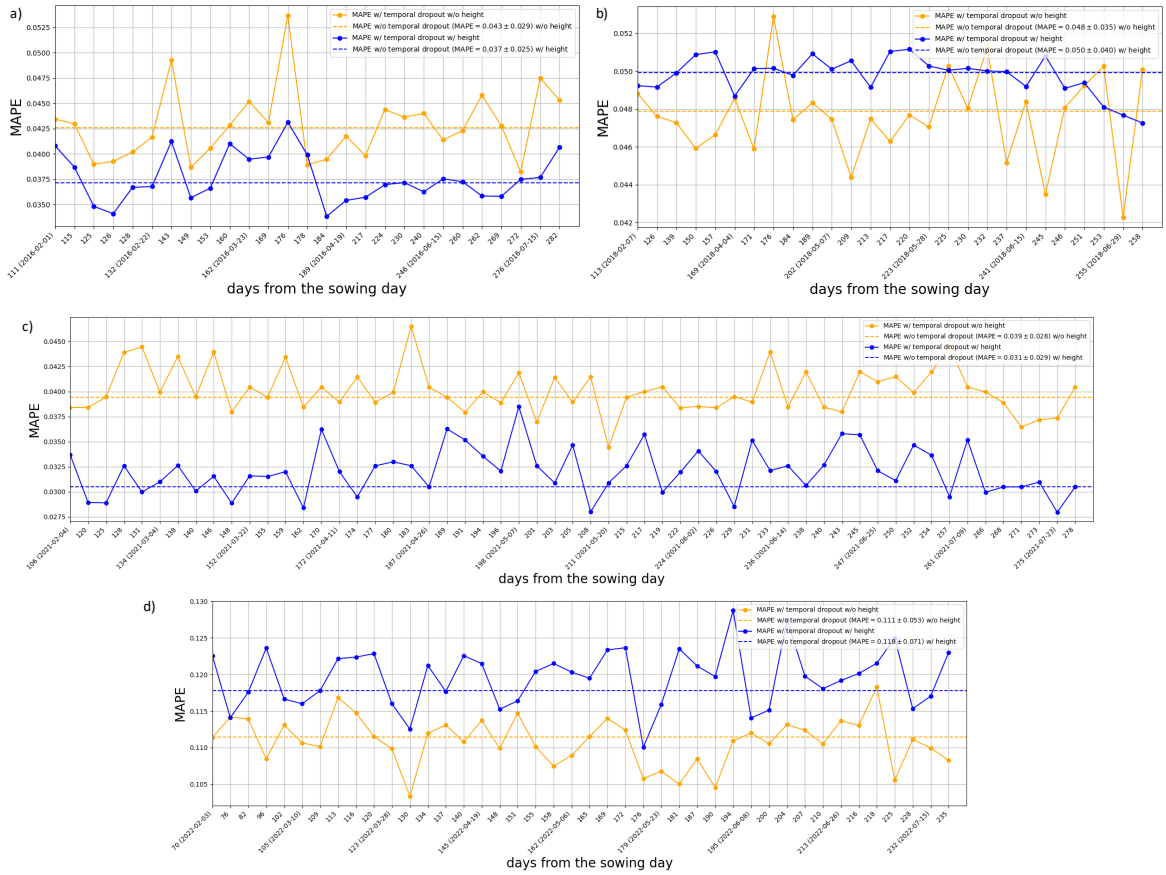

Figure SF4. **Visualizing importance of time points for grain protein content prediction by temporal dropout.** The temporal importance is evaluated with the best-performing image time-series model ConvNext for the years of 2016 (a), 2018 (b), 2021(c) and 2022(d). The x-axis represents the number of days since sowing, with specific calendar dates displayed at intervals to indicate the corresponding time of year. The y-axis shows MAPE values for different configurations: ConvNeXt without plant heights (yellow) and ConvNeXt with plant heights (blue). Horizontal lines mark the MAPE values for the two models without temporal dropout.

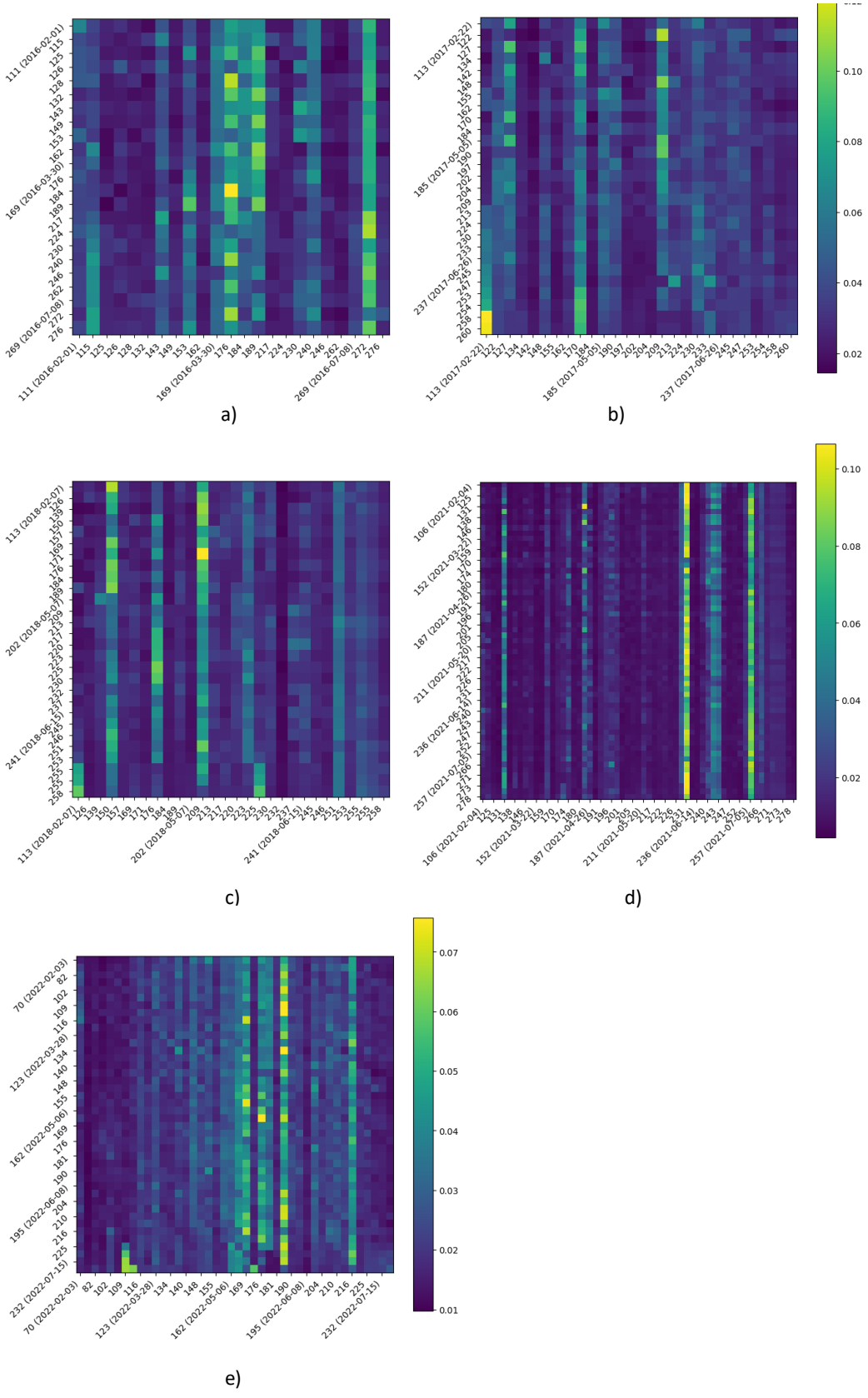

Figure SF1. **Visualizing importance of time points for grain yield prediction by learned attention weights.** The temporal importance is evaluated with the aggregation model from onvNext without plant heights for the years of 2016 (a), 2017 (b), 2018 (c), 2021 (d) and 2022 (e).

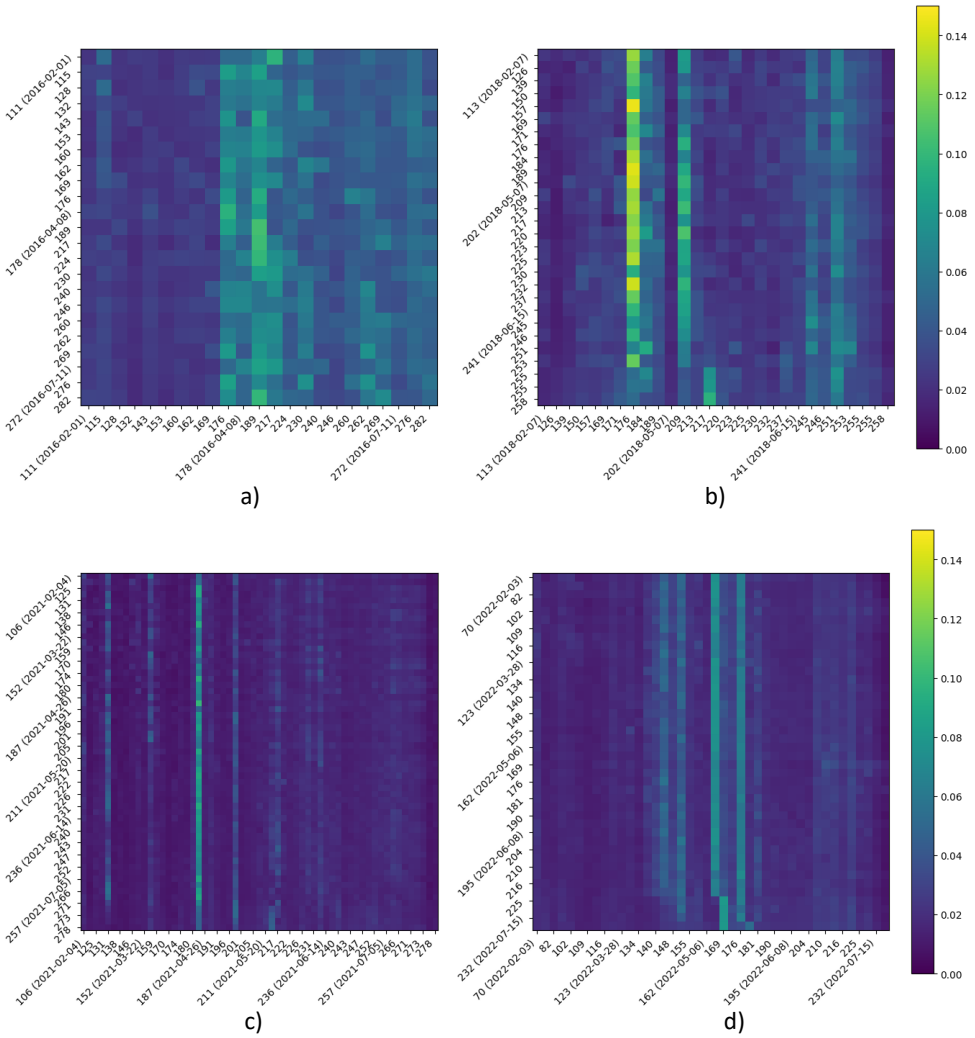

Figure SF2. **Visualizing importance of time points for grain protein content prediction by learned attention weights.** The temporal importance is evaluated with the aggregation model from ConvNext without plant heights for the years of 2016 (a), 2018 (b), 2021(c) and 2022(d).