Interpreting machine learning models to investigate circadian regulation and facilitate exploration of clock function

Laura-Jayne Gardiner[a,1], Rachel Rusholme-Pilcher[b], Josh Colmer[b], Hannah Rees[b], Juan Manuel Crescente[a,c], Anna Paola Carrieri[a], Susan Duncan[b], Edward O. Pyzer-Knapp[a], Ritesh Krishna[a], and Anthony Hall[b,d]

[a]IBM Research Europe, The Hartree Centre, Warrington WA4 4AD, United Kingdom; [b]Earlham Institute, Norwich NR4 7UZ, United Kingdom; [c]Consejo Nacional de Investigaciones Científicas y Técnicas, C1425FQB Buenos Aires, Argentina; and [d]School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, United Kingdom

Edited by Gloria M. Coruzzi, New York University, New York, NY, and approved June 29, 2021 (received for review February 15, 2021)

The circadian clock is an important adaptation to life on Earth. Here, we use machine learning to predict complex, temporal, and circadian gene expression patterns in *Arabidopsis*. Most significantly, we classify circadian genes using DNA sequence features generated de novo from public, genomic resources, facilitating downstream application of our methods with no experimental work or prior knowledge needed. We use local model explanation that is transcript specific to rank DNA sequence features, providing a detailed profile of the potential circadian regulatory mechanisms for each transcript. Furthermore, we can discriminate the temporal phase of transcript expression using the local, explanation-derived, and ranked DNA sequence features, revealing hidden subclasses within the circadian class. Model interpretation/explanation provides the backbone of our methodological advances, giving insight into biological processes and experimental design. Next, we use model interpretation to optimize sampling strategies when we predict circadian transcripts using reduced numbers of transcriptomic timepoints. Finally, we predict the circadian time from a single, transcriptomic timepoint, deriving marker transcripts that are most impactful for accurate prediction; this could facilitate the identification of altered clock function from existing datasets.

explainable AI | circadian | transcriptome | regulation | function

The circadian clock is an internal molecular 24-h timer that is a critical adaptation to life on Earth. It temporally orchestrates physiology, biochemistry, and metabolism across the day/night cycle. As a result, it regulates many traits associated with fitness and survival (1, 2). The clock is a well-characterized transcriptional regulatory network, which drives complex, widespread, and robust patterns of temporal gene expression (3, 4). However, our understanding of such complex transcriptional regulatory systems is limited by our ability to assay them, requiring the generation of long, high-resolution time series datasets.

In plants, much of our understanding of circadian regulation comes from our study of the model plant *Arabidopsis thaliana.* This has yielded a plethora of public, multiomic resources (5–7) that can be reanalyzed to give additional insights into the roles and functions of complex regulatory networks. In this study, we use newly generated datasets, published temporal datasets (8–10) (*SI Appendix*, Table S1), and *Arabidopsis* genomes, in combination with machine learning (ML) approaches (see *SI Appendix*, *Glossary* for definitions of terms), to make predictions about circadian gene regulation and expression patterns. ML models are frequently described as "black boxes," meaning that because of their complexity their inner logic is not easily understood by a human. Critically, we advance existing approaches using explainable AI algorithms and interpretation of our models to illuminate what is inside the black box (*SI Appendix*, *Glossary*), such methods help us to understand the predictions made by ML models. There are many model interpretation strategies in which methods can identify important patterns and/or features that underly an ML model (11). For example, model interpretation has been successfully implemented in drug discovery to enable the mechanistic interpretation of drug action and drug response (12–15). We use such approaches to give insight into circadian biology and experimental design, alongside our predictions. Clarity with respect to how a model makes its predictions, we propose, will also generate confidence and trust in the model, promoting its usage. We use the *Arabidopsis* circadian clock as an example of a complex transcriptional regulatory network since some of its key regulatory elements are already known, allowing the validation of our findings with experimental evidence.

Circadian gene expression rhythms reflect a variety of waveform shapes with a characteristic periodicity of ∼24 h (16). Recent computational methods for identifying these rhythms from transcriptomic time course datasets have achieved circadian gene classification with as few as 3 to 6 timepoints (saving time for sampling and money for sequencing) (17). However, some of the most popular approaches describe optimal sampling strategies for the identification of rhythms running with >3 d of data and 2-h sampling resolution (18, 19). This is partly due to concern for the loss of information and accuracy, as a result of

Significance

The circadian clock is an internal molecular 24-h timer that is critical to life on Earth. We describe a series of artificial intelligence (AI)– and machine learning (ML)–based approaches that enable more cost-effective analysis and insight into circadian regulation and function. Throughout the manuscript, we illuminate what is inside the ML "black box" via explanation or interpretation of predictive ML models. Using this interpretation of our models, we derive biological insights into why a prediction was made, alongside accurate predictions. Most innovatively, we use only DNA sequence features for accurate circadian gene expression prediction. Using explainable AI, we define possible, responsible regulatory elements as we make these predictions; this critically requires no prior knowledge of regulatory elements.

Author contributions: L.-J.G., H.R., E.O.P.-K., R.K., and A.H. designed research; L.-J.G., R.R.-P., J.C., J.M.C., and S.D. performed research; L.-J.G., R.R.-P., J.C., H.R., J.M.C., and A.P.C. analyzed data; and L.-J.G., R.R.-P., J.C., H.R., J.M.C., A.P.C., R.K., and A.H. wrote the paper.

Competing interest statement: L.-J.G., R.K., A.P.C., and E.O.P.-K. are listed as coinventors on a patent application that has been filed.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: laura-jayne.gardiner@ibm.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2103070118/-/DCSupplemental.

Published August 5, 2021.

downsampling. Since the cost implications of this are high, our focus is on designing trusted downsampling strategies for capturing circadian oscillations using a nonoptimal number of timepoints and on improving accuracy compared to existing methods to minimize the impact of information loss. Firstly, we develop ML models that classify circadian expression patterns using an iteratively lower numbers of transcriptomic timepoints, improving accuracy compared to the state of the art. Moreover, we use model interpretation to quantify the best, transcriptomic timepoints for sampling. We believe that predictive insight on when to sample will be a valuable reference for experimental biologists when planning experiments.

Next, we redefine the field, developing ML models that distinguish circadian transcripts using no transcriptomic timepoint information and instead using DNA sequence features (*SI Appendix*, *Glossary*). The theory supporting this is that a major mechanism of (circadian or otherwise) gene expression control is through transcription factor (TF) binding to the regulatory DNA sequence. Considering previous work in *Arabidopsis*, it is likely that the promoter, 5′ untranslated region (UTR), and the first part of the coding region are the most useful locations for TF binding site (TFBS) detection (20). Genes expressed with similar patterns are more likely to be controlled by similar sets of TFBSs. In addition, small RNAs (sRNAs), comprising microRNAs (miRNAs) and small interfering RNAs, are thought to affect transcript abundance via posttranscriptional regulation of messenger RNA (mRNA) (21). Plant miRNAs predominately bind to the coding regions of mRNA and, to a lesser extent, 5′ UTR and 3′UTR regions (22, 23). As such, we consider both coding and noncoding regions to classify circadian genes using DNA sequence. Our DNA sequence features are profiles of *k*-mer–based motif representations that are identified de novo and embody a comprehensive picture of TFBS, sRNA/RNA binding sites, and other sequence-based regulatory elements, since we incorporate the promoter 5′UTR, 3′UTR, and coding regions.

A key strength of our DNA sequence–based approach is that we classify circadian transcripts using *k*-mer–based motif representations generated from preexisting, public, and genomic resources, facilitating downstream application of our methods with no experimental work or prior knowledge of regulatory elements needed. Computational regulatory motif discovery methods typically search for overrepresented words across DNA sequences using methods such as expectation maximization and Gibbs sampling (24–27). Approaches are typically limited by a requirement for input information [e.g., coexpressed genes, site abundance, or a fixed motif length (28–30)]. Artificial intelligence (AI) has been used to take DNA sequence information as an input to predict outputs that likely impact DNA function. Examples include predicting variant effects on chromatin features, such as TF binding, and histone profiles [e.g., DeepSEA (31) and DanQ (32)]. Furthermore, AI has been used to predict transcriptomic profiles directly using features such as DNA sequence or epigenetic marks. These features typically include representations of TFBS [e.g., Xpresso (33, 34)], enhancers [e.g., McEnhancer (35)], histone modifications [e.g., DeepChrome (36)], open chromatin regions (37), or promoters (38). However, these approaches typically require experimental data beyond training, prior knowledge of regulatory elements that our approach does not need, or they focus on single-gene expression states and do not consider complex patterns, as our methods do.

Traditionally, AI work to predict expression has lacked comprehensive model explanation (39). Increasingly, efforts focus on developing interpretive methods for expression prediction models (34–36, 38, 40). For example, for DNA sequence–based models, the studies (33, 38, 40) evaluate feature relevance or importance and derive predictive DNA regions by aligning differential nucleotide importance with differentially expressed genes; these regions can then be bioinformatically analyzed to identify regulatory motifs. Here, alongside our DNA sequence–based predictive model, we use explainable AI to derive regulatory motifs directly from the ML model and explore their functional consequences. We exploit model explanation to identify, on a transcript-by-transcript basis, the ranked regulatory sequences that guide the classification of its expression pattern as circadian. We identify both small and larger combinations of regulatory elements that, in combination, give a larger, overall impact on gene classification. These regulatory sequences are candidate genetic features that could control gene expression and allow us to understand the regulatory mechanisms governing circadian expression patterns and even to manipulate regulation. Ultimately, we use model explanation to generate and validate hypotheses in silico, facilitating both gene expression prediction and derivative regulatory element discovery.

Finally, assaying circadian clock function, as opposed to identifying transcript rhythmicity, has been a challenge for the study of the circadian regulation in organisms ranging from mammals to plants. Recent work applied ML to circadian time course transcriptomic datasets from human blood, to predict the phase of the endogenous circadian clock (circadian time, CT), using a single timepoint from a set of marker genes (41, 42). This allows the use of one timepoint to identify altered clock function (e.g., due to disease or environmental conditions). An equivalent major challenge exists in plant sciences. As such, we use ML to predict the circadian time in *Arabidopsis* from a single, transcriptomic timepoint using marker genes. To advance previous offerings, we identify marker genes as part of our interpretable approach, ensuring that they represent a diverse range of temporal patterns with consistent amplitudes across datasets to facilitate accurate and robust phase prediction, irrespective of sample phase. Counter intuitively, our marker genes do not include the core clock genes used in previous studies for time prediction (43). Taken together, these tools constitute a suite of informative resources for both experimental biologists and the interpretation of further circadian datasets.

Results and Discussion

ML Interpretation Optimizes Timepoint Downsampling to Define Circadian Transcripts. We used MetaCycle for the detection of circadian signals in dense time series transcriptomic data (18). MetaCycle is one of the most well-maintained and accessible tools within the community, incorporating a variety of the most widely used methods, ARSER (44), JTK_CYCLE (45), and Lomb–Scargle (46), and integrating their results so that rhythmic prediction is a cumulation of different statistical approaches. We ran MetaCycle (see *Materials and Methods*) on the *Arabidopsis* time series transcriptomic dataset generated by ref. 8, which was sampled every 4 h for 48 h (*SI Appendix*, Table S1). The data were processed to produce normalized counts per transcript (see *Materials and Methods*). MetaCycle classified 9,394 out of 44,963 transcripts as circadian ($q < 0.05$), with 7,734 denoted as high confidence ($q < 0.02$) (*SI Appendix*, Note S1). We trained a series of ML classifiers to predict if a transcript was circadian or noncircadian in a binary classification using 7,734 of the least likely candidates to be circadian ($q > 0.99$), labeled by MetaCycle alongside the 7,734 highly circadian transcripts ($q < 0.02$) (see *Materials and Methods* and *SI Appendix*, *Glossary* and Note S2). For ML models, we report the F1 scores that measure the accuracy of the model on a scale of 0 to 1, with 1 being most accurate (*SI Appendix*, *Glossary*). Considering all 12 transcriptomic timepoints, the best model was generated with Light Gradient-Boosting Machine (LightGBM) after optimization (*Materials and Methods* and *SI Appendix*, Fig. S1*A* and Table S2), with an F1 score of 0.999 on the training data, 0.955 on the (held out) test data, and a mean F1 cross validation score of 0.939 (*SI Appendix*, *Glossary*). Our confusion matrix (*SI Appendix*, Fig. S1*B* and *Glossary*) highlights consistently high-model accuracy, irrespective of the class that is being predicted (circadian/noncircadian).

Our best ML model (LightGBM) assigned a matching, circadian/noncircadian label to the majority of the transcripts that MetaCycle labeled. However, the overlap was not 100%, so we examined the small proportion of transcripts that were "inaccurately" classified. We found that the inaccurately classified cases by our ML model were often intermediate or borderline cases for MetaCycle (Fig. 1) or edge cases (e.g., with slightly longer period lengths [*SI Appendix*, Fig. S1]). We deduced this because cases rejected by MetaCycle as circadian but accepted by the ML (false positives [FPs]) had significantly lower (MetaCycle derived) P values than the cases that were rejected by both MetaCycle and ML (true negatives [TNs]) ($P < 0.0001$, t = 6.8795, df = 7753). Conversely, cases accepted by MetaCycle as rhythmic but rejected by ML (false negatives [FNs]) had higher (MetaCycle derived) P values than cases categorized as rhythmic by both MetaCycle and ML (true positives [TPs]) ($P < 0.0001$, t = 5.7744, df = 7711) (Fig. 1*A*). Additionally, cases rejected by MetaCycle as circadian but accepted by the ML (FP) have significantly lower relative amplitudes (rAMPs) compared to the TP calls in which both methods agree ($P < 0.0001$, t = 8.3845, df = 7732). Conversely, cases accepted by Metacycle as rhythmic but rejected by ML (FN) had a significantly higher rAMP than the TN calls ($P = 0.036$, t = 2.0924, df = 7732) (Fig. 1*B*). Therefore, the ML model is not simply using high- and low-expression levels to discriminate the circadian and noncircadian status of transcripts, and interestingly, the distribution of rAMPs for the FNs reflects that of the TPs far more closely than that of the TN calls.

We assessed the effect of reducing the number of transcriptomic timepoints on the accuracy of our classification of circadian/noncircadian transcripts. For our best ML model (using 12 timepoints), we reduced the number of timepoints (or features) sequentially from 12 down to 3. To obtain each interim-reduced set of timepoints from 12 to 3, we used well-known feature selection tools χ^2 and eli5 (*SI Appendix*, *Glossary*) and compared these against testing every possible feature combination for the timepoint number (see *Materials and Methods*). The method of trialing every possible feature combination for each reduced timepoint number enabled us to most accurately classify transcripts as circadian/noncircadian (Fig. 2*A*). Using this approach with six timepoints, we achieved a mean classification F1 score of 0.886 on cross-validation and of 0.792 using only three timepoints (*SI Appendix*, Table S3). The mean F1 scores on cross-validation varied by 0.09 between our best and least predictive six timepoints, likewise they varied by 0.06 between our best and least predictive three timepoints, highlighting the impact of timepoint selection. *SI Appendix*, Table S3 highlights that we have consistent accuracy, irrespective of the class that is being predicted (circadian/noncircadian). Using model interpretation (i.e., identifying the combinations of features that gave the highest F1 scores), we defined the most optimal sampling strategies for the different number of timepoints. Selecting six or more timepoints, the best combinations tended to be consecutive timepoints extending across the intersect of day 1 and day 2. In contrast, when selecting low numbers of timepoints, more accurate classifications were made when timepoints were spaced across a single day (Fig. 2*B*), and there was a distinct bias for selecting certain timepoints with others appearing to be much less informative (e.g., ZT28, ZT40, ZT52, and ZT64 that were never selected). Fig. 2*C* shows the best combination of reduced timepoints in each category 12 to 3 for the example transcript phytochrome A (PHYA). When we followed the same strategy, creating transcriptomic ML circadian classification models for wheat, a divergent plant species from *Arabidopsis* (*SI Appendix*, Table S1), we saw similar trends for the best combinations of reduced timepoints (*SI Appendix*, Note S3 and Fig. S2).

To test how generalizable our model is on unseen data (*SI Appendix*, *Glossary*), we used the most accurate three-timepoint model (timepoints 36, 48, and 60) for the binary classification of, firstly, a second *Arabidopsis* transcriptomic time series dataset developed by ref. 9 and secondly, a wheat transcriptomic dataset representing a divergent plant species from *Arabidopsis* (*SI Appendix*, Table S1). These test datasets represent different sampling strategies and experimental setups (see *Materials and Methods*). Both test datasets were processed bioinformatically as per our original (8) dataset, in which we generated ground truth circadian/noncircadian labels for transcripts using MetaCycle

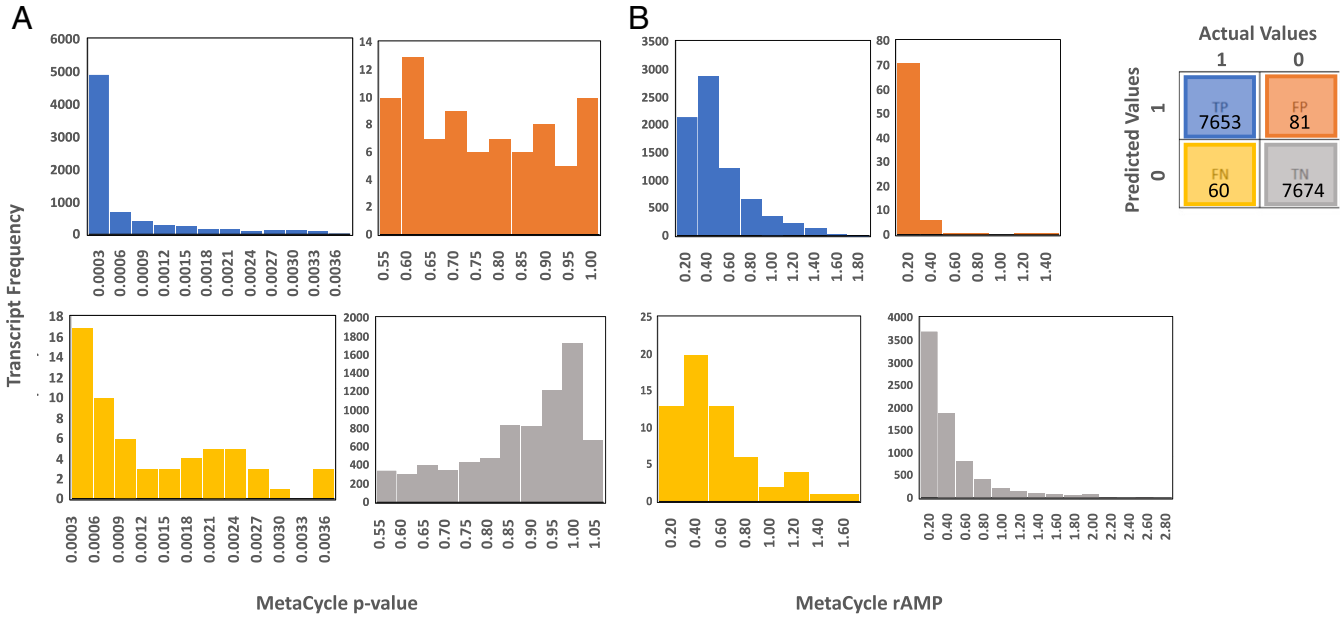
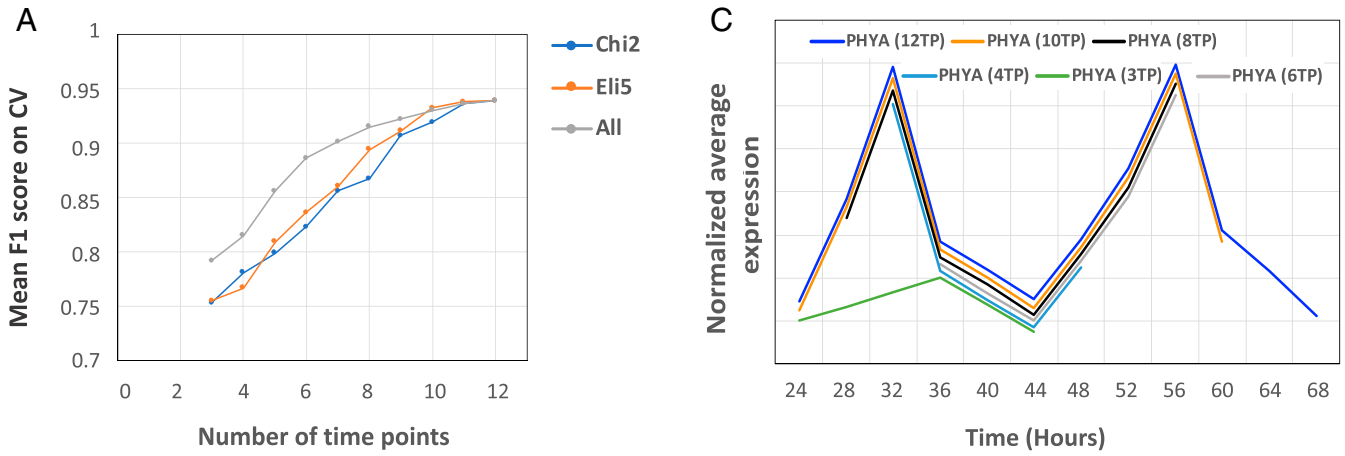


Fig. 1. *Arabidopsis* circadian/noncircadian comparative ML binary classification with 12 transcriptomic timepoints. Class 0 = Noncircadian and Class 1 = Circadian. Histograms in *A* and *B* relate to the best model from *SI Appendix*, Fig. S1*A* generated using LightGBM; the histograms are color coded as per the confusion matrix shown to the right (i.e., showing where our model assigned TP labels, FP labels, FN labels, and TN labels). The histograms show bins representing the frequency of transcripts that had various P values (*A*) or rAMPs (*B*) assigned to them by Metacycle.



B

Time points	avexp_24	avexp_28	avexp_32	avexp_36	avexp_40	avexp_44	avexp_48	avexp_52	avexp_56	avexp_60	avexp_64	avexp_68
N3	4	0	3	5	0	5	3	0	3	4	0	3
N4	3	1	7	7	2	8	5	0	3	3	0	1
N5	3	2	5	7	6	9	6	3	5	3	0	1
N6	3	3	5	9	8	9	7	6	6	2	1	1
N7	2	4	6	9	9	10	9	6	7	5	1	2
N8	4	5	8	10	9	10	9	7	9	5	1	3
N9	5	6	9	10	10	10	10	8	9	7	2	4
N10	7	7	10	10	10	10	10	9	9	8	5	5
N11	9	9	9	9	9	10	10	9	9	9	9	9

Fig. 2. *Arabidopsis* circadian/noncircadian ML binary classification to reduce the number of transcriptomic timepoints. For our best ML model, we reduced the number of timepoints sequentially from 12 to 3. (*A*) To obtain each reduced timepoint set, we compare using χ^2 (Chi2) and eli5 (Eli5) feature selection, with the best set comparing every possible random feature combination (All). We show the best F1 score after fivefold cross validation (CV) for each reduced timepoint set. (*B*) Detailing the 10 best combinations of features that gave the highest F1 score for each reduced set of timepoints. Labels N3 to N11 show the number of reduced timepoints. Labels avexp_24 to avexp_68 show sampling times. Counts 0 to 10 represent the number of times each timepoint appeared in the 10 best feature combinations. (*C*) For the example gene *PHYA*, a line plot of the gene's expression across the best combination of reduced timepoints in each set 12 to 3. Expression values are reduced by ~5% for each timepoint combination to allow the separation of lines for visualization.

with all available timepoints (see *Materials and Methods*). For the *Arabidopsis* (9) dataset, the timepoints did not match those used to train our model; sampling started 2 h after exposure to constant light (rather than 24 h after), and samples were taken every 3 h instead of every 4 h. As such, we selected the closest times to those that were used to train our model according to time of day relative to dawn (timepoints 11, 23, and 35). Unmatched timepoints are likely to have a negative effect on performance, which we observed with an F1 score for the classification of this gene set of 0.714, amounting to a decrease in accuracy of 0.08, compared to the dataset that the model was trained on. For the wheat dataset, sampling started 24 h after exposure to constant light, and measurements were taken every 2 h instead of every 4 h. Therefore, here, matching the time of day relative to dawn, we could select equivalent timepoints (12, 24, and 36 h), and the F1 score was slightly higher at 0.769, amounting to a decrease of only 0.02 on a highly divergent species. The model therefore generalizes well, irrespective of the sample's species, and we observe a much similar performance by matching the timepoints that are used (relative to dawn), as we show for wheat.

We compared our timepoint reduction analysis using ML to a range of analyses representing the state of the art across the different timepoint numbers. MetaCycle requires a minimum of six timepoints for circadian analysis and benefits from these timepoints being evenly sampled across the chosen time period (18). As such, we reduced timepoints from 12 to 6 to enable comparison, including evenly spaced sampling patterns: 4 h/1 d and 8 h/2 d versus the best suggested sampling times from our ML analysis (4 h/1 d from 36 to 56 h from Fig. 2 *B* and *C*). The reduction to six timepoints significantly decreased the number of positive circadian gene calls by MetaCycle that were conserved with the 12-timepoint analysis, independently of the sampling technique used. The highest proportion of the 9,394 circadian genes, identified with 12 timepoints by MetaCycle that were also identified with six timepoints ($P < 0.05$), was 63.7% (*SI Appendix*, Table S4). This accuracy is ~25% lower than the F1 score we achieved with six timepoints and our best ML model (*SI Appendix*, Table S3). Furthermore, when comparing the F1 score of our three-timepoint ML model, it was more appropriate to use a three-timepoint state-of-the-art analysis performed by Spörl et al. (17). *SI Appendix*, Table S4 highlights that we achieve a

12% higher accuracy with only three timepoints in a like-for-like comparison with Spörl et al. (17). This improvement is in addition to the experimental design insight that we provide.

Circadian Genes Can Be Classified from DNA Sequence. We next eliminated transcriptomic timepoints and used DNA sequence features alone to classify transcripts as circadian/noncircadian. To achieve this, we generated *k*-mer profiles (counts) de novo for the mRNA and promoter sequences associated with each transcript, comparing a range of *k*-mer lengths (see *Materials and Methods* and *SI Appendix*, *Glossary*). We trained a series of ML classifiers to predict if a transcript was circadian or noncircadian in a binary classification using the derived *k*-mer profiles for the same set of transcripts and MetaCycle-derived labels used previously (for the transcriptomic ML model). Across the range of *k*-mers, the best models were consistently generated with the classifier LightGBM, and the most accurate model used a *k*-mer length of six, with separate feature sets for promoter and mRNA regions (8,192 features of *k*-mer counts per transcript) that were both inputted into the model (see *Materials and Methods*). This best optimized model showed the following (Fig. 3*A* and *SI Appendix*, Table S2): a mean F1 score of 0.766 on cross-validation (SD 0.006) and a test F1 score of 0.751 on class 0 (noncircadian) and 0.804 on class 1 (circadian). Our accuracy was largely balanced between the classes. An optimal *k*-mer length of 6 bp for this analysis could reflect this being the smallest length that we would not expect to simply occur by chance, therefore, giving ideal resolution. Because of the large number of features created using feature selection, we tested the accuracy of our rhythmic classification when subsets of the feature set were used (see Fig. 3*B* and *Materials and Methods* and *SI Appendix*, *Glossary*). We can reduce the feature number to ∼200 and maintain an F1 score above 0.7, but the highest accuracy was achieved with all 8,192 features, and as such, for downstream investigations, we used the full feature set.

Our de novo *k*-mer generation approach allows the downstream identification and investigation of both known and previously unknown sites, with only the annotation of the TSS/TTS of a transcript required. Our short *k*-mers (6 bp) from promoter/UTR regions should mainly represent regulatory elements such as TFBSs. However, our inclusion of coding regions could encompass additional regulators (e.g., miRNA binding sites). Although miRNAs tend to be 20 to 24 bp in length, our *k*-mers may represent miRNA seed regions that are typically ∼6 bp in length and perfectly/near perfectly match targets (22).

Explanation of DNA Sequence–Based ML Model Links to Circadian Regulation. We next wanted to explain our model, to identify which *k*-mer's were most influential in guiding it to predict transcripts as circadian, since these *k*-mer's could represent the most critical regulatory elements for circadian regulation. If we observe known circadian regulatory elements in this process, this is also a means of validation of the model. We used SHAP (Shapley Additive exPlanations) to explain our best DNA sequence–based model's predictions by computing the contribution of each feature or *k*-mer to that prediction (i.e., ranked feature impact on classification) (*SI Appendix*, *Glossary*) (47). We did this firstly at a global level, looking at the top 30 most impactful features across all of the transcripts for distinguishing class 1 (circadian) from class 0 (noncircadian) (*SI Appendix*, *Glossary* and Fig. 3*C*). Approximately half of the most impactful *k*-mers in Fig. 3*C* show a positive correlation between *k*-mer frequency and the SHAP value or feature impact on the model. Higher frequencies of these *k*-mers for a transcript indicate a higher impact on it being classified circadian. Correlations between *k*-mer count and SHAP impact value for the top four most impactful *k*-mers from Fig. 3*C* are all highly significant ($r > 0.7$, $P < 0.001$; *SI Appendix*, Fig. S3). Of the positively correlated top 30 *k*-mers, 55% of those that contributed to the circadian classification of a transcript were predominantly in the promoter/UTR. We hypothesized that these *k*-mers represent TFBSs for TFs linked to circadian regulation.

To investigate if our most impactful promoter/UTR *k*-mers for prediction were TFBSs, we aligned known *Arabidopsis* TFBSs to each *k*-mer and filtered the most significant matches (*SI Appendix*, Table S5 and *Materials and Methods*). We then validated the *k*-mers that match/likely represent TFBSs using experimental evidence or insight from the literature, many closely associated with circadian regulation or circadian related processes. *k*-mers of interest included (*k*-mer 1; *SI Appendix*, Table S5) matches to TFBS for two photo-responsive TFs (AT3G58630 and AT5G05550) (*P* value 0.0002, e-value 0.18), which form interactions with a number of circadian-related proteins [e.g., LIGHT INSENSITIVE PERIOD1 (LIP1), CONSTANS-Like (Col) 11 (48) and REVEILLE 2 (49)]. Another *k*-mer (*k*-mer 7; *SI Appendix*, Table S5) matched a motif bound by several ethylene-responsive binding proteins ($P = 0.00003$, $e = 0.02$); ethylene synthesis is known to be both circadian controlled and a moderator of the circadian clock (50, 51). We also found matches for binding sites of known circadian TFs, including LUX ARRYTHMO (LUX) (52), CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) (53), and LATE ELONGATED HYPOCOTYL (LHY) (54), alongside motifs associated with light-induced or -repressed sequences (*SI Appendix*, Table S5).

Four of the positively correlated top 30 most impactful *k*-mers defined by SHAP were observed primarily in coding regions rather than promoter/UTRs across the circadian, predicted transcripts. Since miRNAs are thought to influence circadian controlled processes (55, 56) and are common in coding regions, we tested the possibility that these *k*-mers could represent miRNAs by aligning them (plus the surrounding sequence) to mature ath-miRNA sequences to identify matches (see *Materials and Methods*). Two of the four *k*-mers matched miRNA sequences that were associated with developmental timing (57) and chloroplast biogenesis (58). Therefore, for a subset of transcripts, the *k*-mers could represent putative miRNA binding sites that have been experimentally linked to circadian, regulated processes, although this only accounts for a small proportion of the transcripts (*SI Appendix*, Table S5). We next investigated if these *k*-mers could represent RNA-binding motifs (see *Materials and Methods*). RNA-binding proteins are key regulators of gene expression in eukaryotes, and because of strong sequence conservation, their recognition preferences can be inferred from RNA-binding motifs. We validated two of the four coding sequence *k*-mers, linking them to RNA-binding motifs associated with circadian, related processes (*SI Appendix*, Table S5; $P < 0.05$). The first motif is targeted by the RNA-binding protein serine and arginine-rich splicing factor 7 (SRSF7). This links to circadian processes since circadian temperature cycles are known to drive rhythmic SR protein phosphorylation to control alternative splicing (59). The *Arabidopsis* protein RSZ22 is a known ortholog of the SRSF7 SR factor that this alignment could represent (60). The second *k*-mer–matched motif is targeted by the RNA-binding protein LIN28A (*Homo sapiens*). The *Arabidopsis* protein cold shock protein 1 (CSP1) is a known homolog of LIN28A, with a similar, functional role in reprogramming that this alignment could represent (61). CSP1 has been implicated in seed germination timing that is clock related (62).

For the remaining *k*-mers in the top 30 most impactful (Fig. 3*C*), those not associated with promoters/UTRs/miRNAs/RNA binding sites, we investigated their spatial distribution across the transcripts (*SI Appendix*, Fig. S4). Strikingly, there was a clear tendency for them to appear near to the start or else in the first half of the transcript that includes the first exon. By comparison, when we look at the spatial distribution of the promoter-derived *k*-mers from the top 30 most impactful features (*SI Appendix*, Fig. S5), they were distributed more uniformly across the promoters which they occurred in. We investigated if there were any changes in nucleotide composition between our most predictive *k*-mers

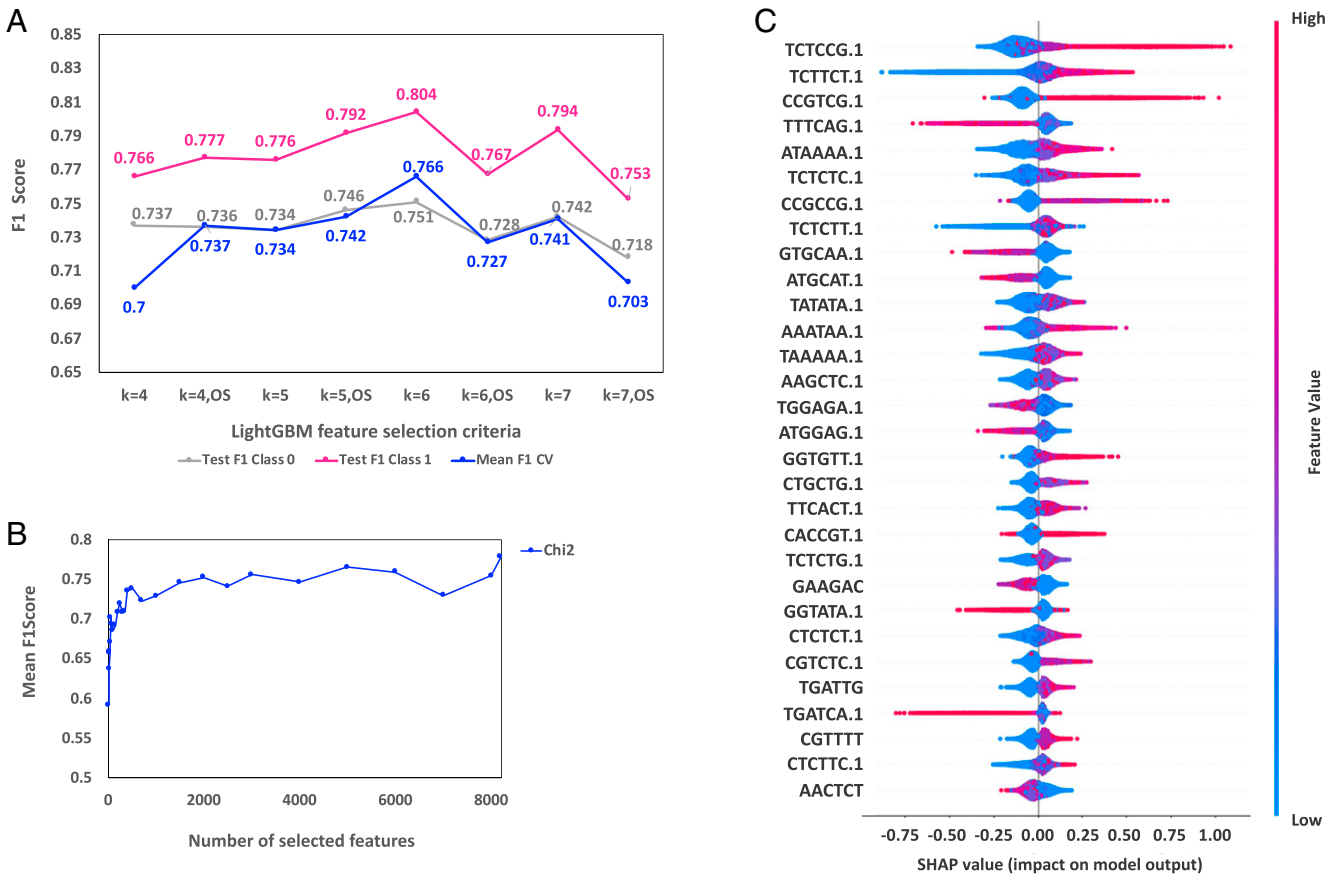


Fig. 3. *Arabidopsis* circadian/noncircadian ML binary classification using *k*-mer profiles. (*A*) For our best performing classifier LightGBM, we compare F1 scores for the test data and after cross validation (CV). These f1 scores were generated using different *k*-mer lengths (4 to 7 bp), with or without the use of oversampling (OS) since our classes are not perfectly balanced (*SI Appendix, Glossary*). (*B*) To obtain each reduced set of *k*-mers, we use χ^2 (Chi2) feature selection. We show the best F1 score after fivefold cross-validation for each set of reduced features. (*C*) The top 30 most impactful features for predicting class 1 (circadian), considering all samples in training and test as calculated using SHAP (*SI Appendix, Glossary*). Feature value denotes the frequency of a *k*-mer per transcript. When the frequency of a *k*-mer per transcript is high (red) and it has a positive SHAP value, this high frequency is driving the prediction of a circadian transcript. This is often coupled to the lower frequency of the same *k*-mer per transcript (blue) having a negative SHAP value, so the absence of the *k*-mer is driving the prediction of a noncircadian transcript. On the contrary, when the frequency of a *k*-mer per transcript is high (red) and has a negative SHAP value, the high frequency is driving the prediction of a noncircadian transcript. This is often coupled to the lower frequency of the *k*-mer per transcript (blue) that has a positive SHAP value, so the absence of the *k*-mer is driving the prediction of a circadian transcript. Features (e.g., the *k*-mer TATTGC) are labeled as "TATTGC" for counts from the promoter and "TATTGC.1" for counts from the mRNA. The corresponding plot for the class 0 (noncircadian) transcripts contains the same list of *k*-mers, but the SHAP value plots will be the exact inverse of this figure.

compared to our nonpredictive *k*-mers (*SI Appendix*, Fig. S4*N*). As a baseline for comparison, we compared these *k*-mer groups to the mRNA and promoter sequences that were used to generate *k*-mers to train the *Arabidopsis* DNA sequence–based model. *SI Appendix*, Fig. S4*N* highlights that promoter sequences show a slightly higher percentage of GC content than mRNA sequence (34% versus 39%, respectively). Our most predictive *k*-mers have a percentage of GC content of 38% that falls in between the baseline for mRNA and promoter, while our nonpredictive *k*-mers significantly deviate from this profile showing a percentage of GC content of 70%. It appears that a high percentage of GC content is more likely to result in a nonpredictive *k*-mer.

Transcript-Specific Explanations Reveal Subclasses within the Binary Circadian Class. Our DNA sequence–based model discriminated transcripts under circadian regulation from those that are not, which is useful to identify circadian regulatory elements from model explanations. However, circadian rhythms reflect a variety of waveform shapes. As such, we bioinformatically identified coexpression modules (*SI Appendix*, *Glossary*) from the transcriptomic profiles of the circadian transcripts that were used to train our ML models using weighted gene coexpression network analysis (WGCNA) (63). This resulted in eight modules with distinct circadian expression profiles. These modules embody groups of transcripts that can be differentiated by their phase of expression with the following phases/groups observed (*SI Appendix*, Fig. S6): morning phases of 0 h (cluster 7) and 4 h (cluster 5/6), day phase of 8 h (cluster 3), day/evening phase of 12 h (cluster 2), evening phase of 16 h (cluster 1), and night phase of 20 h (cluster 4/8).

We next sought to group our circadian transcripts into subgroups representative of different phases of expression, but rather than using transcriptomic information, we wanted to use the SHAP impact values of their *k*-mers. This effectively divides our DNA sequence–based model's binary class circadian into multiple subclasses, providing further insight into transcript rhythmicity. To enable this, we used model explanation of our best DNA sequence–based predictive model, but rather than identifying the most impactful *k*-mers in general (global explanation) for predicting class 1 (circadian), as previously, we identified the most impactful *k*-mers for the classification of each circadian transcript individually (local explanation) (*SI Appendix*, *Glossary*). For this, we focus on the TP circadian transcripts in which MetaCycle and

our ML model both predict the transcripts as circadian. Local explanations are transcript specific and could highlight *k*-mers that are regulating each transcript's expression. Each transcript has a calculated SHAP impact value per feature (8,192 *k*-mers), and with this set of values, we denote the SHAP value profile for a transcript. In an SHAP value profile, the *k*-mer with the highest SHAP value is the most influential on the transcript's classification as circadian. Comparison of these profiles allows us to compare and subdivide the transcripts using DNA sequence composition related to gene regulation, rather than transcriptomic profile.

After deriving local explanations, we filtered the most circadian transcripts according to their SHAP explanation ("most positive cumulative SHAP value"; see Fig. 4*A* and *Materials and Methods* and *SI Appendix*, *Glossary*). Then, we focused on known circadian genes that were within this set (i.e., experimentally validated and known TP genes from previous studies). We clustered the derivative transcripts of these genes based on the similarity of their SHAP value profiles, which represent the relative impact of the *k*-mers on their classification as circadian (Fig. 4*B*). In groups to the right of the dendrogram (purple), 85% of transcripts peak in their expression in the morning/day, whereas in groups to the left 77% of transcripts peak in the evening/night (phases determined by MetaCycle). Therefore, circadian transcripts with more similar *k*-mer SHAP value profiles also had similar expression phases, thus dividing our circadian class into subclasses representing phases of rhythmicity using *k*-mer information. Since we selected the top five most impactful *k*-mers per transcript for clustering, clustered transcripts represent those with similar combinations of *k*-mers that we hypothesize to be guiding expression phase. For example, *PRR3* and *LUX* had similar SHAP value profiles, and we validated this by observing their similar transcriptomic expression profiles, with the evening phases of expression of ZT15 and ZT13, respectively. Exceptions included the two LNK genes, which have morning phase transcript expression profiles but have SHAP profiles similar to evening- and night-expressed genes. This suggests that *LNK1*/*LNK2* may be regulated by a separate mechanism to that regulating other dawn-expressed genes. We also observed *TIC*, which peaks at dusk in the transcriptomic data, in the morning/day cluster; previously, rhythmicity of *TIC* was not detected in whole seedlings (64), whereas here we confidently classify this transcript as circadian from aerial tissue (MetaCycle q = 0.004). Previous work concluded that *TIC* functions in the late evening (65) but plays a role regulating *LHY* that is in the same morning/day cluster as *TIC*; this may explain its appearance here (64). Finally, we also see the night gene *PHYB* in the morning/day cluster, perhaps because of the presence of the similarly regulated *PHYA* in this cluster (66).

From our transcript SHAP value profile clustering (Fig. 4*B*), for subclasses of transcripts with similar expression phases, the most impactful *k*-mers per subclass could represent sequences that are regulating time-of-day–specific expression. Identifying these using model explanation could facilitate the estimation of circadian expression phase without the need for a transcriptomic time course. To test this, we split the transcripts into morning/day/evening/night and investigated which *k*-mers differentiated the groups. We identified the top 30 most variable *k*-mers between the four groups' consensus SHAP explanations; these *k*-mers vary most in their impact between the groups (see *Materials and Methods*) (*SI Appendix*, Table S6). Since we are comparing *k*-mers that differentiate groups of transcripts that are separated by their phase of expression, we validated our hypothesis by matching the *k*-mers to binding sites that have been experimentally associated with specific times of day. For example, the late night–specific Telo-box (67), a G-box–related sequence thought to associate with late night and dawn genes (68), and the evening element (EE) (69) that appeared twice in the top 30 with two *k*-mers matching it. When we compared the importance of these *k*-mers between the morning, day, evening, and night groups, the EE had a higher impact on model prediction in the evening group than in the other three groups, and this difference was statistically significant compared to both morning and night (Fig. 4*C* and *SI Appendix*, Table S7). Additionally, the Telo-box had a higher impact on model prediction when observed in the night group compared to all other groups, and this difference was statistically significant compared to day and evening, fitting with its late night specificity (Fig. 4*D* and *SI Appendix*, Table S7).

Case Study: Explanation for PHYA to PHYE Guides Reclassification of PHYC. The PHYTOCHROME (*PHY*) genes encode red and far-red photoreceptors directly involved in setting the clock. Previous studies have identified circadian regulation of PHY transcripts A to E as rhythmic with patterns ranging from strong to weak (70, 71). Here, from the (8) transcriptomic data PHYC/PHYD/PHYE were all called noncircadian by MetaCycle with q-values of 0.99, 0.60, and 0.13, respectively, while, from the same dataset, the software BooteJTK (72) (that had few differentially annotated transcripts compared to MetaCycle; *SI Appendix*, Note S1) classed PHYD/E as circadian and only PHYC as noncircadian (q-value = 0.2), with some evidence of a weak, cyclic pattern. From this and previous work, these genes could be rhythmic, but this may not be clear in the transcriptomic data, likely because of reported low-amplitude expression patterns and dependent on testing conditions, particularly for PHYC/E (*SI Appendix*, Fig. S7*A*). These genes were missing from our ML analysis and can be used as unseen test datapoints (*SI Appendix*, *Glossary*) for the ML models. The mixture of strong (PHYA/B) and weak cyclical patterns (PHYD/E and potentially PHYC) are ideal for a test of the limits of the ML models. Working with the assumption that all of the PHY primary transcripts A to E are circadian as a baseline for comparison, for the PHY primary transcripts A to E, *SI Appendix*, Table S8 highlights MetaCycle's 40% accuracy, only classifying PHYA/B as circadian, compared to our ML (12 timepoint) model's 80% accuracy, since we additionally classify PHYD/E as circadian. This is supported by the BooteJTK analysis and visually evident rhythmic expression in the transcriptomic data for PHYE and to a lesser extent for PHYD (*SI Appendix*, Fig. S7*A*). We maintain our 80% accuracy when we generate *k*-mer profiles for the PHY transcripts A to E and use our DNA sequence– or *k*-mer–based ML model to predict circadian/noncircadian. Both of our ML models (transcriptomic and DNA sequence–based) classify PHYC as noncircadian, with the other primary PHY transcripts predicted circadian. Even the DNA sequence–based ML model discriminated PHYC from the other PHY transcripts to align with the transcriptomic information, despite sequence similarity between them. Moreover, the transcriptomic expression profile for PHYC provides an unconvincing, circadian rhythm, with an amplitude tending toward zero (0.02), compared to the other transcripts (*SI Appendix*, Fig. S7*A*). This may reflect previous work that concluded a weak or noncycling steady state of PHYC mRNA potentially due to posttranscriptional, circadian regulation (70, 71).

We used the SHAP explanations for the PHY transcripts A to E to identify the regulatory elements that were most impactful in guiding their classifications using the DNA sequence–based model. We compared the SHAP impact values between each of the PHY transcripts A/B/D/E (circadian) and PHYC (noncircadian) to identify those *k*-mers or regulatory elements that are most impactful in predicting PHYA/B/D/E to be circadian but also in predicting PHYC to be noncircadian (six identified in *SI Appendix*, Table S9). The change in frequency of these *k*-mers (within the transcript) is most likely to be responsible for the circadian/noncircadian, predictive differences between the transcripts according to our model (*SI Appendix*, Note S4, Figs. S7 *B* and *C*, and S8). To investigate if altering any of the six identified *k*-mers (*SI Appendix*, Table S9) had more or less potential to induce rhythmicity in PHYC, we sequentially evolved the

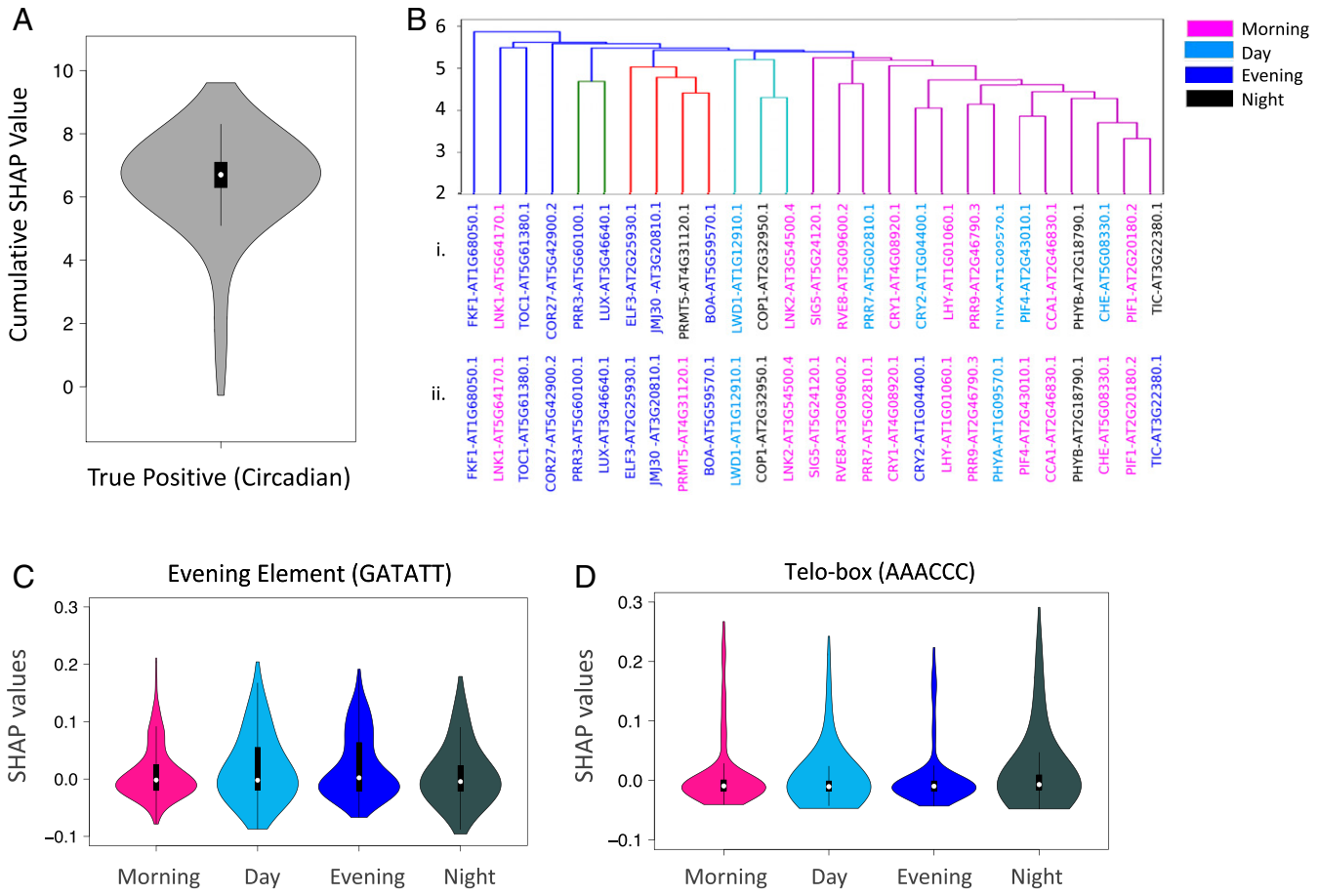


Fig. 4. Investigating *Arabidopsis* circadian TP transcripts after ML DNA sequence–based classification. This figure relates to the LightGBM ML model that we selected as our best classifier. (*A*) Violin plot showing the range of SHAP values across TP transcripts (correctly predicted circadian). A positive SHAP value for a *k*-mer, for a transcript, indicates the *k*-mer is driving a circadian prediction, while a negative SHAP value indicates the *k*-mer is driving the prediction of noncircadian for that transcript. SHAP values are summed for each transcript to produce a cumulative SHAP value. (*B*) Dendrogram clustering known core circadian transcripts according to their profiles of SHAP values if the transcripts were also present in Q1 to Q3 of *A*. We clustered transcripts using hierarchical clustering with average linkage and Euclidean distance (see *Materials and Methods*). Dendrogram branches are colored using a color threshold to color all descendant links below a cluster node k the same color if k is the first node below the cut threshold t (~5). Dendrogram labels colored according to peak phases of expression: morning (0 to 5.99999 h), day (6 to 11.99999 h), evening (12 to 17.99999 h), and night (18 to 23.99999 h), as determined by 1) MetaCycle or 2) the module of origin of the transcript from our eight WGCNA-generated modules. (*C* and *D*) Violin plots show the range of SHAP values across all TP transcripts in groups morning/day/evening/night for the *k*-mers GATATT (*C*) (EE) and AAACCC (*D*) (Telo-box).

spectrum of PHYC, one *k*-mer at a time, to mimic the robustly rhythmic PHYA/B transcripts more and more with each iteration. We used our DNA sequence–based ML model to classify the evolved transcripts. Firstly, removing *k*-mers GGTAGA then TTTCTG sites resulted in predictive probabilities for the circadian class of 0.42 and 0.48, respectively (increasing from 0.38). Secondly, adding AAATAA increased the predictive probability of circadian class membership to 0.58. Finally, adding TCTCCG resulted in a circadian class predictive probability of 0.75, placing this transcript's classification confidently as circadian. Some potential regulatory elements were more important than others, having a larger effect on the classification of the transcript; for example, *k*-mers in the 5′UTR had a larger effect. Additionally, we show that multiple elements combine to have a greater impact on transcript classification and potentially regulation.

We aligned known *Arabidopsis* TFBSs to the UTR-based *k*-mers from PHYA/B that most positively impacted PHYCs circadian reclassification during our evolution to suggest biological reasons why these sites may be having such a large effect. AAATAA aligned to the TFBS of MYB56 that is involved in the regulation of anthocyanin levels in response to circadian rhythms (73) (*SI Appendix*, Table S5). While TCTCCG-matched TFBS of AT3G58630 that has a protein–protein interaction with LIP1, a gene known to function in the clock regulating light input downstream of photoreceptors such as PHYB (74).

We collated a further 41 known key circadian genes, with published evidence of rhythmic expression from the literature and compared the classification accuracy of their associated primary transcripts between MetaCycle, our transcriptomic ML model, and our DNA sequence ML model (*SI Appendix*, Table S10). MetaCycle shows an overall accuracy of 80.49%, classifying the 41 transcripts as circadian, compared to 95.12% with the ML transcriptomic model (*SI Appendix*, Table S11). Approximately 10 of the 41 genes (*SI Appendix*, Table S10) were not used to train either of our ML models and were unseen datapoints, mainly from MetaCycle not assigning a highly confident classification to their transcripts ($q < 0.01$) because of low-amplitude expression profiles. These are problematic transcripts for classification and measure the worst-case scenario for predictions. Using 12 transcriptomic timepoints, our ML model was more accurate at correctly classifying these transcripts as circadian, despite their problematic, low-amplitude rhythms (80% accuracy versus 20%

for MetaCycle). This suggests that our model can generalize well to unseen transcripts. Interestingly, our model that used DNA sequence achieved a higher accuracy of 90% on the unseen datapoints, which was close to its recorded accuracy on all 41 genes (92.68%), sidestepping the problems associated with low amplitudes using genetic sequence features.

Predictions Using DNA Sequence Generalize to Other *Arabidopsis* Ecotypes. Our ML model (using DNA sequence) can accurately make predictions on unseen datapoints. We assessed this in both our initial testing (with held out test data; *SI Appendix*, *Glossary*) and in our case study analysis of known circadian genes. We next assess how well our model performs on a different source to that used for model training (Col-0). We selected the *Arabidopsis* ecotype Ws-2 primarily for this test but also to highlight the suitability of our approach for a divergent species like wheat (details in *SI Appendix*, Note S5 and Fig. S9). For Ws-2, we firstly generated *k*-mer spectra for related transcripts and used the two-timepoint, transcriptomic dataset generated by ref. 10 to label Ws-2 transcripts circadian/noncircadian to gauge accuracy (*SI Appendix*, Table S1, Note S6, and Fig. S10). From this analysis, 71.4% of Ws-2 DNA sequence–based classifications matched their labels derived from (10) transcriptomic data. This is ∼5% lower than the accuracy given by the DNA sequence–based model using Col-0 (mean F1 score of 0.766 on cross-validation) (*SI Appendix*, Note S7). The 5% decrease may be impacted by imperfect Ws-2 labeling; our Ws-2 labeling strategy using two transcriptomic timepoints was >90% accurate for Col-0 (*SI Appendix*, Note S6 and Fig. S10), and from previous work, we expect an additional dropout in performance between species (0.02 for matched timepoints).

We used our DNA sequence–based model to identify transcripts that differentiated in rhythmicity between *Arabidopsis* ecotypes. Then, we used model explanation to explain which regulatory elements influenced this to validate findings. Such functionality gives tremendous power for downstream gene expression manipulation, even in the absence of transcriptomic information. As an initial proof of concept for this, we ranked the transcripts according to the predictive probability of them being circadian for Col-0 and the corresponding predictive probability of them being noncircadian for Ws-2. We identified 12 transcripts that were classified as circadian for Col-0 but noncircadian for Ws-2 by the DNA sequence–based model (predictive probability >0.8) (*SI Appendix*, Table S12). Our most confident or top ranked transcript was AT1G58602.1-RECOGNITION OF PERONOSPORA PARASITICA 7 (*RPP7*) (i.e., most probable circadian transcript in Col-0 [probability 0.999] and most probable noncircadian in Ws-2 [probability 0.991]). RPP genes confer resistance to races of *Peronospora parasitica* in an ecotype-specific manner. Functional RPP7 is thought to mediate resistance to infection by the *Peronospora* isolate Hiks1. Work by ref. 75 found that while Col-0 has a functional RPP7 and is resistant to Hiks1, Ws-2 is susceptible to attack by this pathogen. This coincides with our DNA sequence predictions suggesting that the circadian regulation of RPP7 is important for defense functionality. This conclusion is supported in the experimental, transcriptomic data, in which *RPP7* in Ws-2 shows consistent low expression but in Col-0 is expressed rhythmically at higher levels (*SI Appendix*, Fig. S11*A*) (75). RPP7 has been linked to circadian regulation: firstly because resistance (R)-genes in the RPP family were reported to be under CCA1 control (76) and secondly via RPP7's required interactor EDM2 that is involved in the promotion of floral transition by regulating the floral repressor *FLC* (77).

Previous evidence supports our observed differentiation in the rhythmicity of RPP7 between Col-0 and Ws-2. We next use model explanation to understand which elements differ between Col-0 and Ws-2; in this example, in Ws-2, this could represent the elements to change to render it resistant to Hiks2. As such, for each *k*-mer, we compared the SHAP impact values from the DNA sequence–based model between the Col-0 and Ws-2 homologs of AT1G58602.1 (*RPP7*). We ranked the *k*-mers in ascending order, as the difference in SHAP impact values between the homologs increased, to highlight the regulatory elements that were most impactful in guiding the differential, circadian/noncircadian predictions (*SI Appendix*, Fig. S11). The top five ranked *k*-mers closely linked either to the circadian clock or to disease resistance mechanisms or both (*SI Appendix*, Note S8). We then sequentially evolved the *k*-mer spectrum for AT1G58602.1 in Ws-2, a *k*-mer, at a time to match Col-0 more and more with each iteration. Each iterative, evolved transcript was classified using the DNA sequence–based model, in which we observed that the predictive probability of the circadian class for each evolved transcript quickly increased (*SI Appendix*, Fig. S11*B*). The adaptation of 26 Ws-2 *k*-mers to match Col-0 changed the prediction for Ws-2 from noncircadian to circadian, and the adaptation of 124 Ws-2 *k*-mers was needed to reach the maximum predictive probability of 0.999. The predictive probability of the circadian class for Ws-2 was highly positively correlated (0.676), with the difference in SHAP values between the Col-0 and Ws-2 *k*-mers (*SI Appendix*, Fig. S11*C*). Our analysis shows that some regulatory elements have a larger effect on transcript classification than others and that this effect is quantifiable using model explanation. We show the potential for large combinations of regulatory elements to work together, potentially each contributing a small amount, to result in a large overall impact on gene classification and potentially regulation (e.g., the 26 *k*-mers that we changed here to convert Ws-2 to be classified as circadian).

AT1G58602.1 showed no expression in Ws-2 across the two transcriptomic timepoints (*SI Appendix*, Fig. S11*A*). To highlight that our DNA sequence–based model was classifying circadian Col-0/noncircadian Ws-2 transcripts, not simply expressed/non-expressed, we investigated other transcripts from *SI Appendix*, Table S12. We identified four additional transcripts in our ranked top 10, in which we observe expression of both Ws-2 and Col-0 consistently across the two transcriptomic timepoints (*SI Appendix*, Fig. S12). Here, the expression profile of Ws-2 is still seen to be largely flat versus potentially cyclical profiles of Col-0.

Identifying Transcriptional Biomarkers that Predict Internal Circadian Time. Here, we use ML to determine the circadian time of sampling (i.e., predicting the phase of the endogenous circadian clock) using a set of transcriptional biomarkers from any single, transcriptomic timepoint. Previous studies have developed such models for human and mammalian transcriptome datasets (41–43, 78, 79). We developed a method that we applied to plant data that innovatively uses model interpretation to identify *Arabidopsis* biomarker transcripts to guide predictions. This incorporates biomarker selection from across circadian phases to increase accuracy and robustness.

To train our model, we used the transcripts per million–normalized, circadian dataset described earlier (8) and the two further transcriptomic datasets (9, 10) for validation and testing (see *Materials and Methods* and *SI Appendix*, *Glossary*). Firstly, we aggregated a selection of metrics to rank and select transcript subsets from ref. 8 according to their confidence of rhythmicity for model training (see *Materials and Methods*). *SI Appendix*, Table S13 highlights the mean absolute errors (MAEs) of the predictions of circadian time without hyperparameter optimization (*SI Appendix*, *Glossary*) on the three temporal, transcriptomic datasets using different sized subsets of the highest-ranked, rhythmic genes. The lowest MAE, based on the (10) test dataset, was 104 min and was observed with a selected subset of 50 transcripts. Using the confidence of rhythmicity for transcript prioritization, we noted that the representation of our subsets of transcripts across the eight coexpression modules, generated by the WGCNA gene coexpression network analysis, was not uniform (*SI Appendix*, Fig. S13*A* and *Glossary*). This reflects an uneven representation across the phases of rhythmic expression.

Therefore, we prioritized the selection of transcripts using model interpretation in the form of feature selection to make the frequency distribution across the modules more uniform (see *Materials and Methods* and *SI Appendix*, *Glossary*). Optimizing performance based on the validation dataset, our best performing model overall used a final subset of 15 transcripts (*SI Appendix*, Table S14) and had an MAE of 21 min on the training data, 56 min on the (9) validation data, and 46 min on the test data from ref. 10. *SI Appendix*, Fig. S13 *B* and *C* highlight that after feature selection there was a decrease in the generalization error on average across the (10) test dataset, with the improvements in MAE decreasing as the number of genes increased. This supports the theory that features containing different, temporal patterns of varying strengths outperform features containing strong but highly correlated patterns.

The performance of our best model (15 transcripts with an MAE of 46 min on the test data) is in line with the ∼1-h test error reported by ref. 78 using their state-of-the-art method ZeitZeiger. As such, we applied ZeitZeiger to our datasets (8–10) to compare directly with our model. To reflect our previous approach, the dataset (8) was used to fit ZeitZeiger, with predictions then being generated on the validation (9) and testing (10) datasets to compare with the predictions generated by our method. Our approach significantly outperformed ZeitZeiger on the test dataset (MAE of 46 compared to 143 min; *SI Appendix*, Fig. S14), demonstrating our efficacy at generating highly accurate predictions for circadian time. We also noted a large disparity in training, validation, and test errors by ZeitZeiger (MAE of 6 min on training, 119 on validation, and 143 on test) that suggests overfitting (*SI Appendix*, *Glossary*). We hypothesized that our selection of biomarker transcripts, to ensure even representation across the phases of rhythmic expression, would yield a more robust or generalizable mapping from expression data to internal, circadian time (i.e., less overfitting); this analysis supports this hypothesis.

The 15 transcripts in our final subset act as a small subgroup of biomarker transcripts that are sufficient to predict the circadian time (*SI Appendix*, Table S14). Interestingly, the 15 transcripts did not include any core clock genes. This analysis was conducted using the ecotype Col-0. However, using the Ws-2 data (10), an MAE on this ecotype of only 53 min was observed (5 min lower than for Col-0 on which the model was trained). Generally, we observed no relationship between circadian time and prediction error, except for in the training dataset, in which errors at the 20-h timepoint were significantly larger than the other times (*SI Appendix*, Fig. S13*D*). However, variation in error across the timepoints typically stayed under 90 min, allowing the sufficient resolution of circadian time, given that typical sampling strategies are between 2 to 4 h.

Conclusions

We describe a series of ML-based approaches that enable cost-effective analysis and insight into circadian regulation in *Arabidopsis*. One of the drawbacks of ML is a lack of clarity as to why it makes specific predictions. We illuminate what is inside the black box via an explanation or interpretation of ML models. Although we demonstrate this for circadian rhythms, this approach has widespread implications for other complex gene expression patterns.

When we predict circadian transcripts using low numbers of mRNA sequencing (mRNA-seq) timepoints. Although there is an information loss and resultant drop in F1 score when selecting as few as three timepoints, not only do we improve accuracy compared to existing methods but we also use model interpretation to optimize sampling strategies. Some of the most accurate, reduced sampling strategies that we identify align with existing approaches (e.g., timepoints spaced evenly across a day to most effectively capture the sine wave [up-down-up/down-up-down] profile). But our observed bias for certain timepoints, with others appearing to be much less informative, provides insight into how to most effectively downsample. Our other identified, reduced sampling strategies were unexpected (e.g., consecutive timepoints or those across the intersect of day 1 and 2).

Most significantly, we use only DNA sequence features for accurate, circadian classification, requiring no prior knowledge of regulatory elements or transcriptomic data. This offers advantages over existing methods to not only predict expression but to decipher regulation at the same time since, using an explainable AI algorithm, we define regulatory elements on the fly as we make predictions. Automated definition and prioritization of these feature profiles for transcripts, de novo using AI, has the potential to support functional annotation of genomes and precision agriculture. This application could redefine how we generate testable hypotheses to understand gene expression control. Our predictive accuracy is possibly higher than our current estimates, as our DNA-based approach scores the potential of a gene to be circadian regulated. However, it is possible that this regulation may be restricted to specific tissue types or developmental stages. Therefore, our experimental, generated labels may be underestimating the number of rhythmic genes.

Finally, we predict circadian time, while using model interpretation to derive *Arabidopsis* marker transcripts. These selected transcripts could be used to test single datapoints in existing and emerging *Arabidopsis* datasets to investigate how genotypes, treatments, and environmental conditions affect circadian clock function. Additionally, since transcriptomic datasets are typically expensive in terms of both time and money, the reduction of profiling to marker genes within a single timepoint could yield a huge saving in resources.

Materials and Methods

More detailed information on the materials and methods used in this study are provided in *SI Appendix*, *Materials and Methods*.

Data Generation and Processing. The datasets used in this analysis are detailed in *SI Appendix*, Table S1. Previously published *Arabidopsis* datasets have details for data generation in the relevant associated publication. For the wheat time course, detailed data generation methods are available in *SI Appendix*. Detailed methods for the bioinformatic processing, MetaCycle analysis, and clustering of mRNA-seq datasets are in *SI Appendix*.

Binary Classification: ML Model Training, Tuning, and Validating. We used Scikit Learn (version 3.7) for the ML binary classification analysis to predict if a gene was circadian or not with either transcriptomic or DNA sequence–based feature sets (80). The following classifiers were tested: Logistic Regression, Gaussian process, Random Forest, XGBoost, LightGBM, Support Vector Machine (linear kernel), Decision Tree, and K nearest neighbors. Detailed methods for the feature generation, normalization, feature selection, model training (*SI Appendix*, Table S15), model testing, model explanation [using SHAP (47, 81)], validations of explanations, and analyses associating subclasses with phase of expression are in *SI Appendix*.

Identifying Marker Genes to Tell the Circadian Time Using a Single Transcriptomic Timepoint. We developed an ML-based pipeline to predict the circadian time (phase) at any single, transcriptomic sampling timepoint using gene expression data from a set of marker genes using an artificial neural network in TensorFlow (version 2.0.0) (82). We provide the code for this in a Jupyter Notebook and instructions to run this code at https://github.com/AHallLab/PredictingCircadianTime. The three transcriptomic datasets used previously (*SI Appendix*, Table S1) from refs. 8 to 10 were used for training, validation, and testing, respectively. Detailed methods of feature generation, normalization, feature selection, model training, testing, and validation are in *SI Appendix*.

Data Availability. All of the datasets used in this analysis are detailed in *SI Appendix*, Table S1. All previously published datasets have details for data generation in the relevant associated publication. For the wheat time course, reads are available from the European Nucleotide Archive (ENA) under project name PRJEB40948 at https://www.ebi.ac.uk/ena/browser/view/PRJEB40948 (83). The algorithms and hyperparameters used for the detailed ML models are statedin *SI Appendix*, Table S2. The data used as input into the ML models is available as supplementary data files to this submission as follows: File S1 is a csv file that includes transcript names, the 12 transcriptomic timepoints (8)

used to train our *Arabidopsis* transcriptomic ML models, and the last column denotes the circadian/noncircadian labels for each transcript (File_S1-Arabidopsis_thaliana.TAIR10_12TP.csv). Similarly, File S2 is a csv file that includes transcript names, the 24 transcriptomic timepoints used to train our wheat transcriptomic ML models, and the last column denotes the circadian/noncircadian labels for each transcript [File_S2-Wheat_24TP.csv]. For our circadian time prediction, custom code is required, and as such, we provide the code in a Jupyter Notebook and instructions to run this code at https://github.com/AHallLab/PredictingCircadianTime. All other study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank Ben White (Earlham institute), Rob Tracey (IBM Research), and Vadim Elisseev (IBM Research) for their help transferring sequencing reads to the European Nucleotide Archive. This work was supported by the Science and Technology Facilities Council Hartree Centre's Innovation Return on Research program, funded by the Department for Business, Energy, and Industrial Strategy.

1. A. N. Dodd *et al*., Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* **309**, 630–633 (2005).
2. T. P. Michael *et al*., Enhanced fitness conferred by naturally occurring variation in the circadian clock. *Science* **302**, 1049–1053 (2003).
3. M. F. Covington, J. N. Maloof, M. Straume, S. A. Kay, S. L. Harmer, Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol.* **9**, R130 (2008).
4. S. L. Harmer *et al*., Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* **290**, 2110–2113 (2000).
5. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
6. 1001 Genomes Consortium, 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
7. A. Athar *et al*., ArrayExpress update - From bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).
8. A. Romanowski, R. G. Schlaen, S. Perez-Santangelo, E. Mancini, M. J. Yanovsky, Global transcriptome analysis reveals circadian control of splicing events in *Arabidopsis thaliana*. *Plant J.* **103**, 889–902 (2020).
9. Y. Yang, Y. Li, A. Sancar, O. Oztas, The circadian clock shapes the *Arabidopsis* transcriptome by regulating alternative splicing and alternative polyadenylation. *J. Biol. Chem.* **295**, 7608–7619 (2020).
10. A. Graf *et al*., Parallel analysis of *Arabidopsis* circadian clock mutants reveals different scales of transcriptome and proteome regulation. *Open Biol.* **7**, 160333 (2017).
11. C. B. Azodi, J. Tang, S.-H. Shiu, Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.* **36**, 442–455 (2020).
12. N. L. Patel-Murray *et al*., A multi-omics interpretable machine learning model reveals modes of action of small molecules. *Sci. Rep.* **10**, 954 (2020).
13. J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
14. Z. Wang, N. R. Clark, A. Ma'ayan, Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* **32**, 2338–2345 (2016).
15. L.-J. Gardiner *et al*., Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. *Sci. Rep.* **10**, 9522 (2020).
16. T. C. Mockler *et al*., The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harb. Symp. Quant. Biol.* **72**, 353–363 (2007).
17. F. Spörl *et al*., Krüppel-like factor 9 is a circadian transcription factor in human epidermis that controls proliferation of keratinocytes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10903–10908 (2012).
18. G. Wu, R. C. Anafi, M. E. Hughes, K. Kornacker, J. B. Hogenesch, MetaCycle: An integrated R package to evaluate periodicity in large scale data. *Bioinformatics* **32**, 3351–3353 (2016).
19. T. Zielinski, A. M. Moore, E. Troup, K. J. Halliday, A. J. Millar, Strengths and limitations of period estimation methods for circadian data. *PLoS One* **9**, e96462 (2014).
20. C.-P. Yu, J.-J. Lin, W.-H. Li, Positional distribution of transcription factor binding sites in *Arabidopsis thaliana*. *Sci. Rep.* **6**, 25164 (2016).
21. F. Borges, R. A. Martienssen, The expanding world of small RNAs in plants. *Nat. Rev. Mol. Cell Biol.* **16**, 727–741 (2015).
22. J. Ding, S. Zhou, J. Guan, Finding microRNA targets in plants: Current status and perspectives. *Genomics Proteomics Bioinformatics* **10**, 264–275 (2012).
23. J. Hausser, A. P. Syed, B. Bilen, M. Zavolan, Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* **23**, 604–615 (2013).
24. S. Dvir *et al*., Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E2792–E2801 (2013). Correction in: *Proc. Natl. Acad. Sci. U.S.A.* **116**, 13701 (2019).
25. L. Narlikar, I. Ovcharenko, Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomics Proteomics* **8**, 215–230 (2009).
26. F. P. Roth, J. D. Hughes, P. W. Estep, G. M. Church, Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).
27. G. Thijs *et al*., A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**, 1113–1122 (2001).
28. W. Thompson, M. J. Palumbo, W. W. Wasserman, J. S. Liu, C. E. Lawrence, Decoding human regulatory circuits. *Genome Res.* **14**, 1967–1974 (2004).
29. S. T. Jensen, J. S. Liu, BioOptimizer: A Bayesian scoring function approach to motif discovery. *Bioinformatics* **20**, 1557–1564 (2004).
30. X. Liu, D. L. Brutlag, J. S. Liu, BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138 (2001).
31. J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
32. D. Quang, X. Xie, DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).
33. V. Agarwal, J. Shendure, Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
34. M. A. Beer, S. Tavazoie, Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
35. D. Hafez *et al*., McEnhancer: Predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome Biol.* **18**, 199 (2017).
36. R. Singh, J. Lanchantin, G. Robins, Y. Qi, DeepChrome: Deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32**, i639–i648 (2016).
37. A. Natarajan, G. G. Yardimci, N. C. Sheffield, G. E. Crawford, U. Ohler, Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* **22**, 1711–1722 (2012).
38. J. D. Washburn *et al*., Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5542–5549 (2019).
39. A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile. *AAAI* **33**, 3681–3688 (2017).
40. J. Zrimec *et al*., Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141 (2020).
41. J. J. Hughey, Machine learning identifies a compact gene set for monitoring the circadian clock in human blood. *Genome Med.* **9**, 19 (2017).
42. R. Braun *et al*., Universal method for robust detection of circadian state from gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9247–E9256 (2018).
43. F. Agostinelli, N. Ceglia, B. Shahbaba, P. Sassone-Corsi, P. Baldi, What time is it? Deep learning approaches for circadian rhythms. *Bioinformatics* **32**, 3051 (2016).Correction in: *Bioinformatics* **32**, i8–i17 (2016).
44. R. Yang, Z. Su, Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioinformatics* **26**, i168–i174 (2010).
45. M. E. Hughes, J. B. Hogenesch, K. Kornacker, JTK_CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J. Biol. Rhythms* **25**, 372–380 (2010).
46. E. F. Glynn, J. Chen, A. R. Mushegian, Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics* **22**, 310–316 (2006).
47. S. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions" in *Advances in Neural Information Processing Systems*, I. Guyon *et al*., Eds. (Curran Associates, Inc., New York, 2017), vol. 30, pp. 4768–4777.
48. A. Samach, A. Gover, Photoperiodism: The consistent use of CONSTANS. *Curr. Biol.* **11**, R651–R654 (2001).
49. N. Nakamichi, Molecular mechanisms underlying the *Arabidopsis* circadian clock. *Plant Cell Physiol.* **52**, 1709–1718 (2011).
50. J. Grundy, C. Stoker, I. A. Carré, Circadian regulation of abiotic stress tolerance in plants. *Front Plant Sci* **6**, 648 (2015).
51. C. S. Moffat *et al*., ERF5 and ERF6 play redundant roles as positive regulators of JA/Et-mediated defense against *Botrytis cinerea* in *Arabidopsis*. *PLoS One* **7**, e35995 (2012).
52. C. S. Silva *et al*., Molecular mechanisms of evening complex activity in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6901–6909 (2020).
53. R. M. Green, E. M. Tobin, The role of CCA1 and LHY in the plant circadian clock. *Dev. Cell* **2**, 516–518 (2002).
54. R. Rawat *et al*., REVEILLE8 and PSEUDO-REPONSE REGULATOR5 form a negative feedback loop within the *Arabidopsis* circadian clock. *PLoS Genet.* **7**, e1001350 (2011).
55. G. R. Prabu, A. K. Mandal, Computational identification of miRNAs and their target genes from expressed sequence tags of tea (*Camellia sinensis*). *Genomics Proteomics Bioinformatics* **8**, 113–121 (2010).
56. F. Borges, P. A. Pereira, R. K. Slotkin, R. A. Martienssen, J. D. Becker, MicroRNA activity in the *Arabidopsis* male germline. *J. Exp. Bot.* **62**, 1611–1620 (2011).
57. G. Wu *et al*., The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. *Cell* **138**, 750–759 (2009).
58. B.-B. Anna, B. Grzegorz, K. Marek, G. Piotr, F. Marcin, Exposure to high-intensity light systemically induces micro-transcriptomic changes in *Arabidopsis thaliana* roots. *Int. J. Mol. Sci.* **20**, 5131 (2019).
59. M. Preußner *et al*., Body temperature cycles control rhythmic alternative splicing in mammals. *Mol. Cell* **67**, 433–446.e4 (2017).
60. A. S. N. Reddy, G. Shad Ali, Plant serine/arginine-rich proteins: Roles in precursor messenger RNA splicing, plant development, and stress responses. *Wiley Interdiscip. Rev. RNA* **2**, 875–889 (2011).
61. C. Li *et al*., A Lin28 homologue reprograms differentiated cells to stem cells in the moss *Physcomitrella patens*. *Nat. Commun.* **8**, 14242 (2017).
62. K. Sasaki, M.-H. Kim, R. Imai, *Arabidopsis* COLD SHOCK DOMAIN PROTEIN 2 is a negative regulator of cold acclimation. *New Phytol.* **198**, 95–102 (2013).
63. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

64. Z. Ding, A. J. Millar, A. M. Davis, S. J. Davis, TIME FOR COFFEE encodes a nuclear regulator in the *Arabidopsis thaliana* circadian clock. *Plant Cell* **19**, 1522–1536 (2007).
65. A. Hall *et al*., The TIME FOR COFFEE gene maintains the amplitude and timing of *Arabidopsis* circadian clocks. *Plant Cell* **15**, 2719–2729 (2003).
66. J. Lim *et al*., Antagonistic roles of PhyA and PhyB in far-red light-dependent leaf senescence in *Arabidopsis thaliana*. *Plant Cell Physiol.* **59**, 1753–1764 (2018).
67. M. Spensley *et al*., Evolutionarily conserved regulatory motifs in the promoter of the *Arabidopsis* clock gene LATE ELONGATED HYPOCOTYL. *Plant Cell* **21**, 2606–2623 (2009).
68. D. Ezer *et al*., The G-Box transcriptional regulatory code in *Arabidopsis*. *Plant Physiol.* **175**, 628–640 (2017).
69. T. P. Michael *et al*., Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet.* **4**, e14 (2008).
70. R. Tóth *et al*., Circadian clock-regulated expression of phytochrome and cryptochrome genes in *Arabidopsis*. *Plant Physiol.* **127**, 1607–1616 (2001).
71. T. Clack, S. Mathews, R. A. Sharrock, The phytochrome apoprotein family in *Arabidopsis* is encoded by five genes: The sequences and expression of PHYD and PHYE. *Plant Mol. Biol.* **25**, 413–427 (1994).
72. A. L. Hutchison, R. Allada, A. R. Dinner, Bootstrapping and empirical Bayes methods improve rhythm detection in sparsely sampled data. *J. Biol. Rhythms* **33**, 339–349 (2018).
73. C. Y. Jeong *et al*., AtMyb56 regulates anthocyanin levels via the modulation of *AtGPT2* expression in response to sucrose in *Arabidopsis*. *Mol. Cells* **41**, 351–361 (2018).
74. K. Terecskei *et al*., The circadian clock-associated small GTPase LIGHT INSENSITIVE PERIOD1 suppresses light-controlled endoreplication and affects tolerance to salt stress in *Arabidopsis*. *Plant Physiol.* **161**, 278–290 (2013).
75. A. Nemri *et al*., Genome-wide survey of *Arabidopsis* natural variation in downy mildew resistance using combined association and linkage mapping. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10302–10307 (2010).
76. W. Wang *et al*., Timing of plant immune responses by a central circadian regulator. *Nature* **470**, 110–114 (2011).
77. T. Tsuchiya, T. Eulgem, The *Arabidopsis* defense component EDM2 affects the floral transition in an FLC-dependent manner. *Plant J.* **62**, 518–528 (2010).
78. J. J. Hughey, T. Hastie, A. J. Butte, ZeitZeiger: Supervised learning for high-dimensional data from an oscillatory system. *Nucleic Acids Res.* **44**, e80 (2016).
79. N. Wittenbrink *et al*., High-accuracy determination of internal circadian time from a single blood sample. *J. Clin. Invest.* **128**, 3826–3839 (2018).
80. F. Pedregosa *et al*., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
81. A. P. Carrieri *et al*., Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. *Sci. Rep.* **11**, 4565 (2021).
82. M. Abadi *et al*., TensorFlow: Large-scale machine learning on heterogeneous distributed systems, v2.0.0. https://www.tensorflow.org/. Accessed 19 July 2021.
83. H. Rees, S. Duncan, R. Rusholme-Pilcher, A. Hall, Circadian RNA-seq time series data for wheat. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB40948. Deposited 23 October 2020.