PNAS

Supporting Information for
Machine learning estimation of human body time using metabolomic profiling

Tom Woelders[1,5], Victoria L. Revell[2,6], Benita Middleton[2], Katrin Ackermann[3,7], Manfred Kayser[3], Florence I. Raynaud[4], Debra J. Skene[2], Roelof A. Hut[1]

Affiliations:

1) Chronobiology unit, Groningen Institute of Evolutionary Life Sciences, University of Groningen, 9700CC Groningen, the Netherlands
2) Chronobiology, Faculty of Health and Medical Sciences, University of Surrey, Guildford GU2 7XH, United Kingdom
3) Department of Genetic Identification, Erasmus University Medical Center Rotterdam, 3000 CA Rotterdam, the Netherlands
4) Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London SM2 5NG, United Kingdom
5) Current address: Division of Neuroscience and Experimental Psychology, School of Biology, Faculty of Biology Medicine and Health, University of Manchester, Manchester, M13 9PT, United Kingdom
6) Current address: Surrey Sleep Research Centre, Faculty of Health and Medical Sciences, University of Surrey, Guildford GU2 7XP, United Kingdom
7) Current address: Biomedical Sciences Research Complex & Centre of Magnetic Resonance, University of St Andrews, St Andrews, KY16 9ST, United Kingdom

Corresponding author: Roelof A. Hut
Email: r.a.hut@rug.nl

This PDF file includes:

Supporting text
Figures S1 to S4
Tables S1 to S2

Other supporting materials for this manuscript include the following:

Woelders etal 2023 SuppInfo Data & Rscripts.zip
https://doi.org/10.6084/m9.figshare.22567783.v1

Supporting Information Text

Eligibility and screening

Participants were 18 - 35 years old and were either males (protocol 1) or females (protocol 2) taking combined oral contraceptive pills and being on the active phase during the laboratory sessions. Participants had to meet a defined set of inclusion/exclusion criteria including scoring ≤ 5 on the Pittsburgh Sleep Quality Index, <11 on the Epworth Sleepiness Scale, and <10 on the Beck Depression Inventory. They also could not be extreme morning or evening types on the Horne-Ostberg questionnaire. Participants had to pass a medical examination including having biochemistry and haematology blood sample assessments. Participants could not be smokers and had to pass alcohol breath, drugs of abuse, and cotinine (protocol 2 only) screening at each study visit. They had to have a habitual, regular sleep-wake cycle that involved going to bed between 22:00 and 24:00 h, and getting up between 06:00 and 08:00 h with 6 – 8 h in bed, and had to agree to keeping a regular sleep/wake schedule , wearing actiwatches and keeping sleep diaries for the duration of the study. Participants had to agree to refrain from alcohol, caffeine, heavy exercise and bright light 72 hours before and during the in-laboratory session. In addition they were asked to avoid non-steroidal anti-inflammatory drugs for 72-hours (protocol 1) or 7 days (protocol 2) before the laboratory session.

Participants were excluded if it was not safe for them to participate. They could not have a history of any systemic, psychiatric or neurological disease or drug and alcohol abuse. In addition, they could not be taking regular medication that affects melatonin synthesis or circadian rhythms (e.g., antihypertensive drugs, non-steroidal anti-inflammatory drugs, hypnotic drugs, benzodiazepines, antidepressants, antipsychotic drugs, barbiturates, antiepileptic drugs). They could not have donated > 400 ml blood in the preceding 3 months, and could not have a body mass index (BMI) < 19 or > 33 kg/m^2 or a total body weight < 50 kg. Participants could not have a clinically significant history of sleep disorders (protocol 2), and could not drink >21 units (protocol 1) or >14 units (protocol 2) of alcohol per week. Participants were not enrolled if they were regularly working evening, early morning or night shifts or had travelled across more than two time zones, within one month of and throughout the study. They could not have any history of severe allergies or any hypersensitivity to heparin which was used during the laboratory session during blood sample collection while participants were sleeping.

Baseline-at-home

For 7 days prior to the laboratory session, participants were required to maintain a regular sleep/wake schedule going to bed at 23:00 h and getting up at 07:00 h each day. They could not deviate from these times by more than 15 min. During this time they were required to fill out a daily sleep and nap diary, call a time-stamped voicemail upon awakening and before they go to bed, and wear an Actiwatch (AWL) (Cambridge Neurotechnology Ltd., UK). Each morning they were required to go outside for at least 15 min between 07:00 and 08:30 h.

Laboratory study

Participants attended the laboratory for a residential session. Day 1/Night 1 was an adaptation night followed by blood sampling during a day/night cycle (Day 2/Night

2/Day 3). Posture, meals and environmental lighting were strictly controlled throughout the laboratory session. On day 2 and 3, participants remained semi-recumbent in < 5 lux between 18:00 and 23:00 h, and between 07:00 and 09:00 h. On Night 1 participants were semi-recumbent in < 5 lux as soon as possible after admission. On Night 1 and Night 2, participants slept in a supine position in 0 lux between 23:00 and 07:00 h. Between 09:00 and 18:00 h each day participants were free to move about in normal room lighting (~100 lux). Participants were provided with identical meals on all days of the study; breakfast was 10 min after waking, lunch 6 hours later and dinner 5 hours after that. Hourly blood samples were collected from 12:00 h on Day 2 until 23:00 h on Day 3; on Night 2 a specialised blood collection system was used that allowed samples to be taken without entering the room.

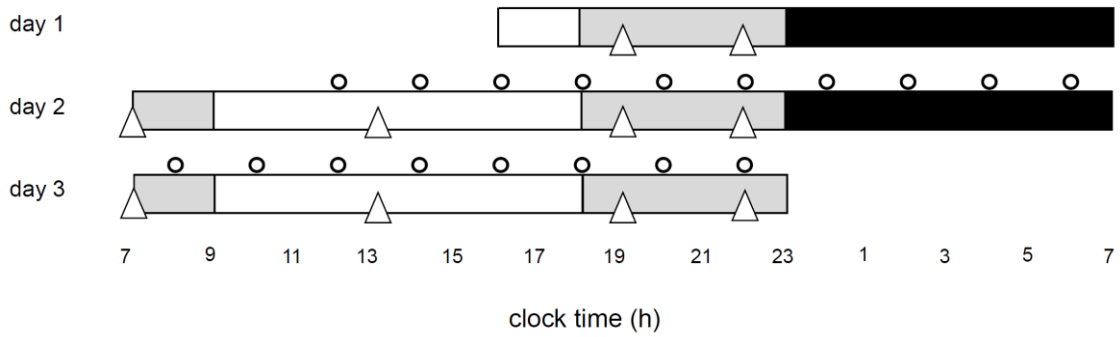

Figure S1: Sampling protocol. White bars indicate wake periods with free movement and ~100 lux white light intensity. Grey bars indicate wake periods with semi-recumbent position and <5 lux of white light. Black bars indicate sleep periods, supine at 0 lux. Triangles indicate standardized meals; circles indicate blood sampling for metabolomics. Study protocol was identical for females and males (see Supplementary text for full details).

Table S1: Model prediction errors (MdAE in h) performed for different permutations of the model input: Cortisol only (Cort), Melatonin only (Mel), Metabolites only (Met), and all combinations of these three input data. Optimal timing of the samples outperforms random timing by ~65%. Interestingly, the metabolites only based analysis seems on average to outperform the Cort, Mel, and Mel & Cort for the random samples, but not for the timed samples. The average improvement of the Metabolites based model is about 12% when Cort and Mel are added (6% for the optimally timed samples). The metabolites only approach works best in most cases, but adding Melatonin and Cortisol to the Metabolites will decrease random variation and thus likely increase the robustness of the model. Model performance using Cort, Mel and Met together increases in most scenarios, especially when three samples can be taken. Clearly, optimal timing of the samples outperforms random timing by ~65%. Interestingly, the metabolites only based analysis seems on average to outperform the Cort, Mel, and Mel & Cort for the random samples, but not for the timed samples. Nonetheless, the average improvement of the Metabolites based model is about 12% when Cort and Mel are added (6% for the optimally timed samples). We think that the metabolites only approach works best in most cases, but adding Melatonin and Cortisol to the Metabolites will improve the model in most scenarios, especially when three samples can be taken. In addition, more input variables will increase the robustness of the model against outliers.

Cort = cortisol in analysis Mel = melatonin in analysis Met = metabolites in analysis																Mean
MdAE(h)		Cort only		Mel only		Met only		Mel & Cort		Met & Cort		Met & Mel		Met & Mel & Cort		
# samples		F	M	F	M	F	M	F	M	F	M	F	M	F	M	
random	1	3.07	3.33	3.33	3.67	1.88	1.52	2.07	2.37	1.68	1.43	1.78	1.59	1.54	1.48	1.79
	2	2.56	2.78	2.64	2.73	1.47	1.23	1.43	1.84	1.28	1.17	1.37	1.19	1.16	1.15	
	3	2.21	2.32	2.18	2.32	1.22	1.09	1.19	1.47	1.08	1.01	1.21	1.04	1.02	1.01	
timed	1	1.11	0.78	0.73	0.80	0.90	0.72	0.75	0.81	0.92	0.95	0.92	0.97	1.02	0.96	0.62
	2	0.77	0.64	0.72	0.56	0.72	0.48	0.49	0.45	0.70	0.44	0.67	0.49	0.45	0.60	
	3	0.54	0.56	0.43	0.49	0.50	0.41	0.33	0.51	0.40	0.28	0.44	0.26	0.30	0.26	
mean all		1.72		1.72		1.01		1.14		0.95		0.99		0.91		
mean timed		0.73		0.62		0.62		0.56		0.62		0.62		0.60		

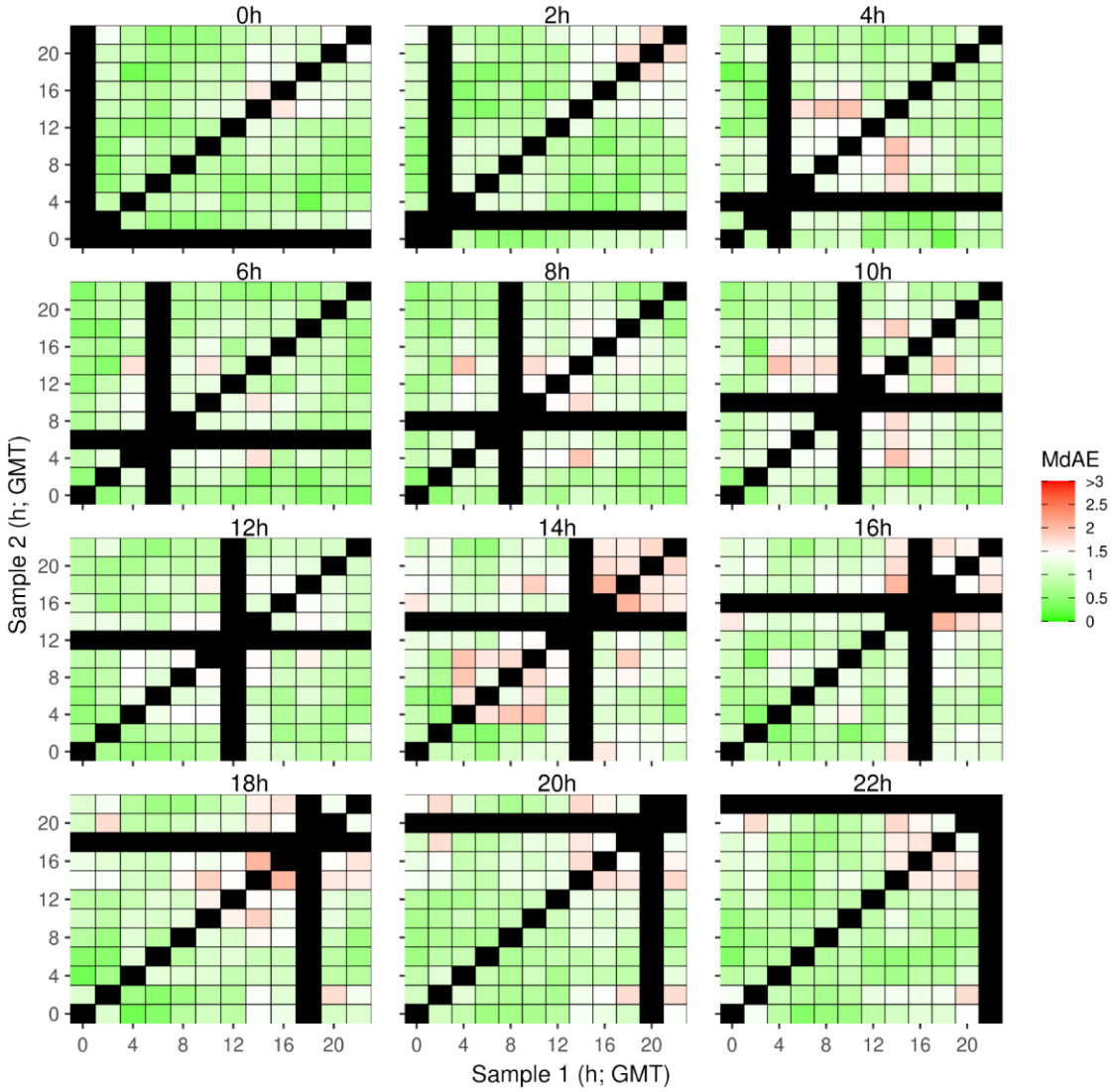

Figure S2: Median absolute errors landscape for DLMO estimation based on three blood samples in the female data set.

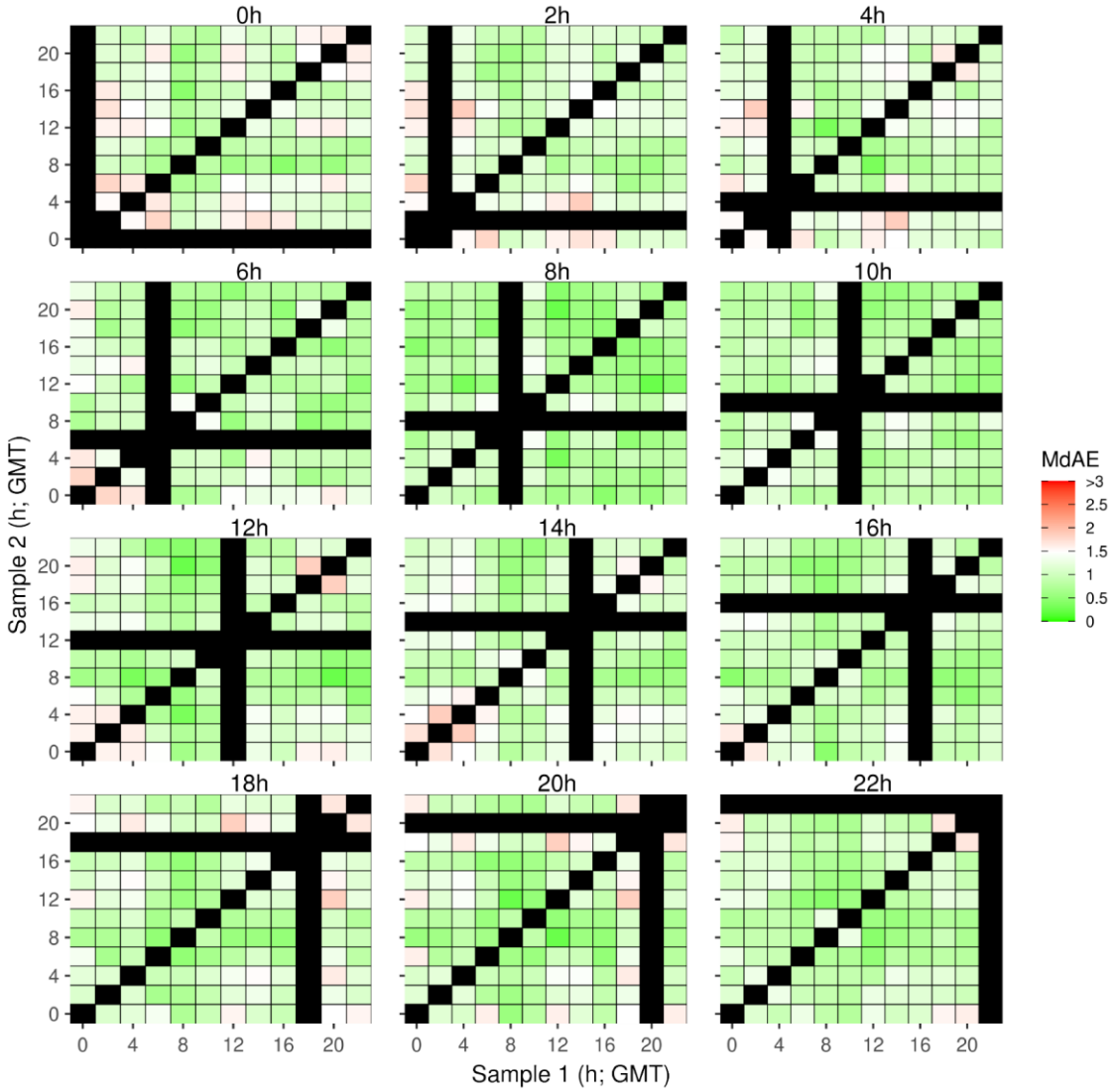

Figure S3: Median absolute errors landscape for DLMO estimation based on three blood samples in the male data set.

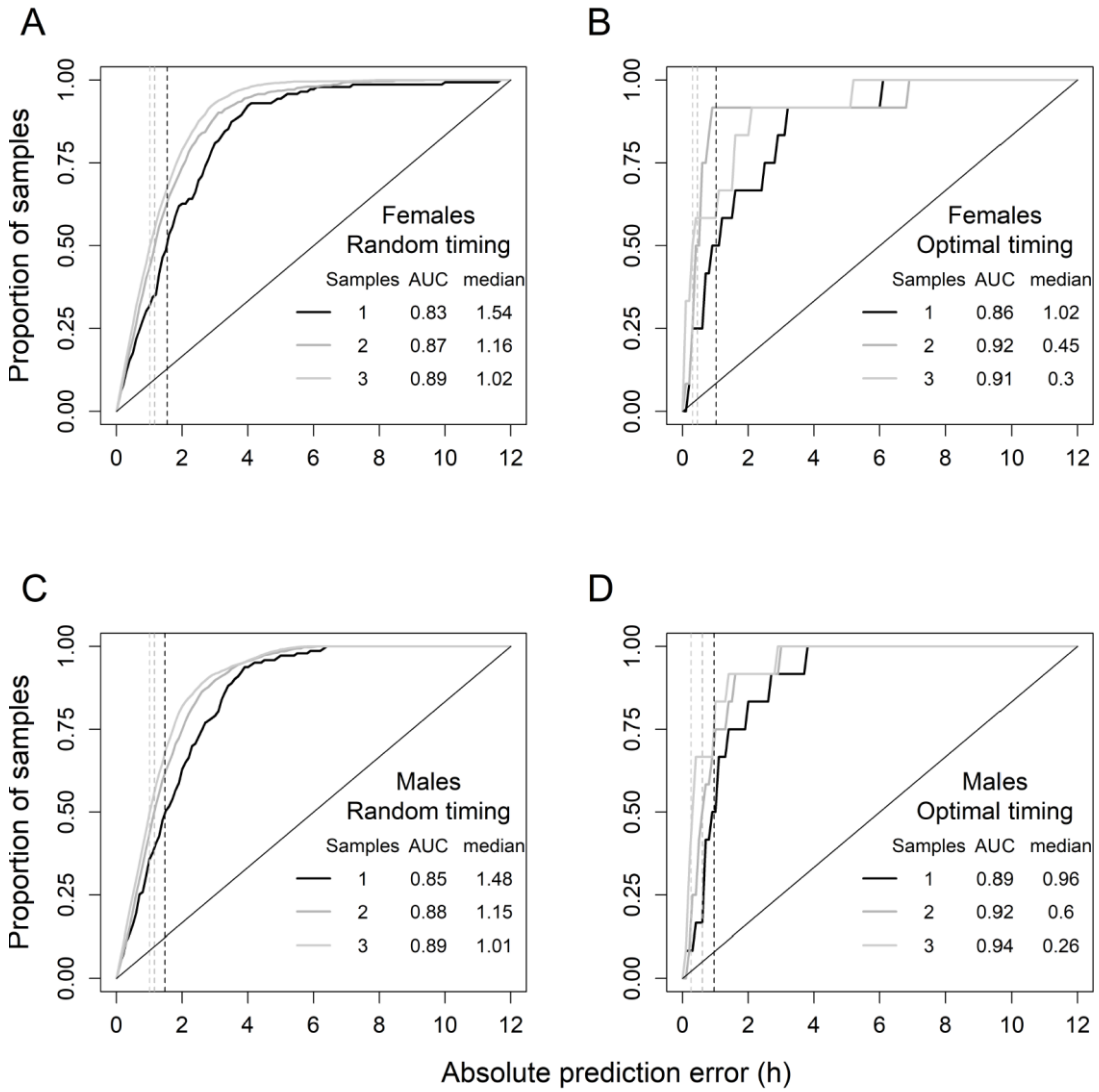

Figure S4: Receiver Operating Characteristic curves for female (A, B) and male (C, D) data sets, for randomly timed 1, 2, and 3 sample methods or optimally timed 1, 2, and 3 sample methods.

Table S2: Pathway analysis results. Metabolites that contributed to the S or C component for the female or male models (Model) were used as inputs for the KEGG pathway analysis. The pathways that were identified as significant are listed (Pathway) with the respective analysis parameters: total number of compounds in the pathway (Total) and the expected contribution (Exp); the actual matched number from the uploaded data (Hits); the original p value calculated from the enrichment analysis (Raw p); p value adjusted for False Discovery Rate for multiple testing (FDR p); the pathway impact value calculated from pathway topology analysis (Impact).

Model	Pathway	Total	Exp	Hits	Raw p	FDR p	Impact
S_females	Aminoacyl-tRNA biosynthesis	48	0.53	11	$8 \cdot 10^{-14}$	$7 \cdot 10^{-12}$	0.000
	Arginine biosynthesis	14	0.15	3	$4 \cdot 10^{-4}$	0.016	0.178
	Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.04	2	0.0007	0.019	1.000
	Nitrogen metabolism	6	0.07	2	0.0017	0.028	0.000
	D-Glutamine and D-glutamate metabolism	6	0.07	2	0.0017	0.028	0.500
	Glutathione metabolism	28	0.31	3	0.0030	0.037	0.108
	Valine, leucine and isoleucine biosynthesis	8	0.09	2	0.0031	0.037	0.000
	Glyoxylate and dicarboxylate metabolism	32	0.35	3	0.0045	0.045	0.106
	Phenylalanine metabolism	10	0.11	2	0.0048	0.045	0.357
C_females	Aminoacyl-tRNA biosynthesis	48	0.59	10	$2 \cdot 10^{-11}$	$2 \cdot 10^{-9}$	0.000
	Arginine biosynthesis	14	0.17	3	0.0005	0.022	0.178
	Nitrogen metabolism	6	0.07	2	0.0021	0.044	0.000
	D-Glutamine and D-glutamate metabolism	6	0.07	2	0.0021	0.044	0.500
S_males	Aminoacyl-tRNA biosynthesis	48	0.46	6	$3 \cdot 10^{-6}$	$2 \cdot 10^{-4}$	0.000
	Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.04	2	0.0005	0.022	1.000
C_males	Aminoacyl-tRNA biosynthesis	48	0.62	12	$2 \cdot 10^{-14}$	$2 \cdot 10^{-12}$	0.000
	Arginine biosynthesis	14	0.18	3	0.0006	0.026	0.193
	Phenylalanine, tyrosine and tryptophan biosynthesis	4	0.05	2	0.0009	0.026	1.000
	Nitrogen metabolism	6	0.08	2	0.0023	0.039	0.000
	D-Glutamine and D-glutamate metabolism	6	0.08	2	0.0023	0.039	0.500

Woelders etal 2023 SuppInfo Data & Rscripts.zip
https://doi.org/10.6084/m9.figshare.22567783.v1

Supplementary Datasets and Rscripts – Meta Information:

Instructions for the R scripts to analyse the data presented in the Woelders et al. (2023) paper.

Input data:
Mel_Cort_Entrained_F.csv, Mel_Cort_Entrained_M.csv - *Melatonin and cortisol concentrations per decimal sampling time for females and males*

Targeted_Entrained_F.csv, Targeted_Entrained_M.csv - *Targeted metabolite abundances (unnormalized) per decimal sampling time for females and males*

Y_Entrained_F.csv, Y_Entrained_M.csv - *Circadian time and cartesian coordinates per decimal sampling time for females and males (ie Y-matrix)*

DLMO_Entrained_F.csv, DLMO_Entrained_M.csv - *Estimated DLMO for females and males separately (for the two days in the lab).*

*R scripts *:*
1. Figure 1.R
Input: Data/Y_Entrained_F.csv, Y_Entrained_M.csv, Mel_Cort_Entrained_F.csv, Mel_Cort_Entrained_M.csv, Targeted_Entrained_F.csv, Targeted_Entrained_M.csv
Output: Figures/Figure 1(ABCD).eps

2. PLSR_EntrainedFemales.R
Input: Data/Y_Entrained_F.csv, Mel_Cort_Entrained_F.csv, Targeted_Entrained_F.csv
Output: Model output/dataSubset**/Data_usedByPLSR_Females.csv *(the data as was used by the PLSR algorithm),*
Model output/dataSubset**/coefficients_S_Females.csv *(the coefficients for the S model; mean, sd and significance),*
Model output/dataSubset**/coefficients_C_Females.csv *(the coefficients for the C model; mean, sd and significance),*
Model output/dataSubset**/PLSR_LOOData_Females.csv *(Y matrix and predictions per left out participant; s, c, angle and CT)*

3. PLSR_EntrainedMales.R
Input: Data/Y_Entrained_M.csv, Mel_Cort_Entrained_M.csv, Targeted_Entrained_M.csv
Output: Model output/dataSubset**/Data_usedByPLSR_Males.csv,
Model output/dataSubset**/coefficients_S_Males.csv,
Model output/dataSubset**/coefficients_C_Males.csv,
Model output/dataSubset**/PLSR_LOOData_Males.csv

4. Figure 2(ABC).R
Input: Model output/dataSubset**/PLSR_LOOData_Females.csv
Output: Figures/dataSubset**/Figure 2(ABC).eps

5. Figure 2(DEF).R
Input: Model output/dataSubset**/PLSR_LOOData_Males.csv
Output: Figures/dataSubset**/Figure 2(DEF).eps

6. Figure 3(ABC)_Figure S2_S4_Stats_Females.R
Input: Model output/dataSubset**/PLSR_LOOData_Females.csv

Output: Figures/dataSubset**/Figure 3(ABC).eps, Figure S2.eps, Figure_S4_ROC_Females.png, Statistics/dataSubset**/stats_Females.xlsx

7. Figure 3(DEF)_Figure S3_Stats_Males.R
Input: Model output//dataSubset**/PLSR_LOOData_Males.csv
Output: Figures/dataSubset**/Figure 3(DEF).eps, Figure S3.eps, Figure_S4_ROC_Males.png, Statistics/dataSubset**/stats_Males.xlsx

8. DLMOStats.R
Input: DLMO_Entrained_M.csv and DLMO_Entrained_F.csv
Output: Statistics/DLMO_stats.xlsx

Model output folder contents (One folder for each subset of the data)
For all models, the beta coefficients; mean, sd and significance are produced. For convenience (not used by any script), the full dataset as used by the PLSR procedure (i.e. X and Y matrices). Note that the folders for the one-only subsets (Only Mel or Cort), the single feature is repeated once. This is because the pls regression does not run with an X matrix of only one column. PLSR can handle this artificial colinearity (partly the reason why we are using the method).

dataSubset**/Data_usedByPLSR_Females.csv, Data_usedByPLSR_Males.csv
dataSubset**/coefficients_C_Females.csv, coefficients_C_Males.csv, coefficients_S_Females.csv, coefficients_S_Males.csv
dataSubset**/PLSR_LOOData_Females.csv, PLSR_LOOData_Males.csv
Y matrix and predictions per left out participant; s, c, angle and CT.

Statistics folder contents (One folder for each subset of the data):
These excel documents contain the MdAE values in the first sheet and the F-table of the statistical test used in the second sheet.

dataSubset**/stats_Females.xlsx and stats_Males.xsls.

Notes:
* *Scripts 4-7 use the outputs from scripts 2 and 3 so it is wise to just run the scripts in order. To run them, simply change the working directory on top of each script accordingly and make sure the required libraries are installed*

** *Scripts 2-7 are run for all combinations of melatonin, cortisol and metabolites included (7). Therefore, the figures and stats that are presented in the manuscript are also created for different subsets of the data to get a feeling of how melatonin, cortisol and metabolites (and any combination of these 3) perform*