

Supporting Information

An Agnostic Framework for the Classification/Identification of Organisms Based on RNA Post-Transcriptional Modifications

William D. McIntyre,^{1‡} Reza Nemati,^{2‡} Mehraveh Salehi,^{3‡} Colin C. Aldrich,² Molly FitzGibbon,⁴ Limin Deng,¹ Manuel A. Pazos,⁴ Rebecca E. Rose,² Botros Toro,⁴ Rachel E. Netzband,⁴ Cara T. Pager,^{4,6} Ingrid P. Robinson,⁴ Sean M. Bialosuknia,⁵ Alexander T. Ciota,⁵ and Daniele Fabris^{1,6}*

1) Department of Chemistry, University of Connecticut, Storrs, CT 06269, USA

2) Department of Chemistry, University at Albany (SUNY), Albany, NY 12222, USA

3) Department of Electrical Engineering, Yale University, New Haven, CT 06520, USA

4) Department of Biological Sciences, University at Albany, Albany, NY 12222, USA

5) School of Public Health, University at Albany, Albany, NY 12222, USA

6) RNA Institute, University at Albany, Albany, NY 12222, USA SA

[‡] These authors contributed equally to this work

* Corresponding author: email dan.fabris@uconn.edu

Table of Contents

Experimental section: Provides additional details on the materials and methods utilized in the study.

Scheme S1. Analytical workflow for rPTM analysis employed in the study.

Figure S1: Results provided by training/testing a k nearest neighbor (KNN) classifier on data from representative organisms of different kingdoms of life.

Figure S2 Results provided by training/testing a naïve Bayes (NB) classifier on data from representative organisms of different kingdoms of life.

Figure S3: Relative importance of rPTMs on gradient boosting (GB) predictions for three members of the *Enterobacteriaceae* bacilli, and correlation between the three more important rPTMs.

Figure S4: Relative importance of rPTMs on gradient boosting (GB) predictions for two different *E. coli* serotypes, and correlation between the three more important rPTMs.

Figure S5: Relative importance of rPTMs on gradient boosting (GB) predictions for wildtype *S. cerevisiae* and three different knockout strains, and correlation between the three more important rPTMs.

Figure S6: Relative importance of rPTMs on gradient boosting (GB) predictions for three types of immortalized *H. sapiens* cell lines, and correlation between the three more important rPTMs.

Figure S7: Relative importance of rPTMs on gradient boosting (GB) predictions for three types of *H. sapiens* central nervous system (CNS) cell lines, and correlation between the three more important rPTMs.

Figure S8: Representative results obtained by training and testing the *k* nearest neighbor (KNN) algorithm on the data obtained from 20 different classes.

Figure S9: Representative results obtained by training and testing the naïve Bayes (NB) algorithm on the data obtained from 20 different classes.

Table S1: rPTM profiles obtained from *E. coli* OR:H48:K, *H. salinarum*, *S. cerevisiae* BY4741, *A. thaliana*, and *H. sapiens* HeLa cells (see **Experimental**).

Table S2: Average accuracy, precision, and recall afforded by testing the *k* nearest neighbor (KNN), naïve Bayes (NB), and gradient boosting (GB) classifiers on the selected representative classes described in **Table S1**.

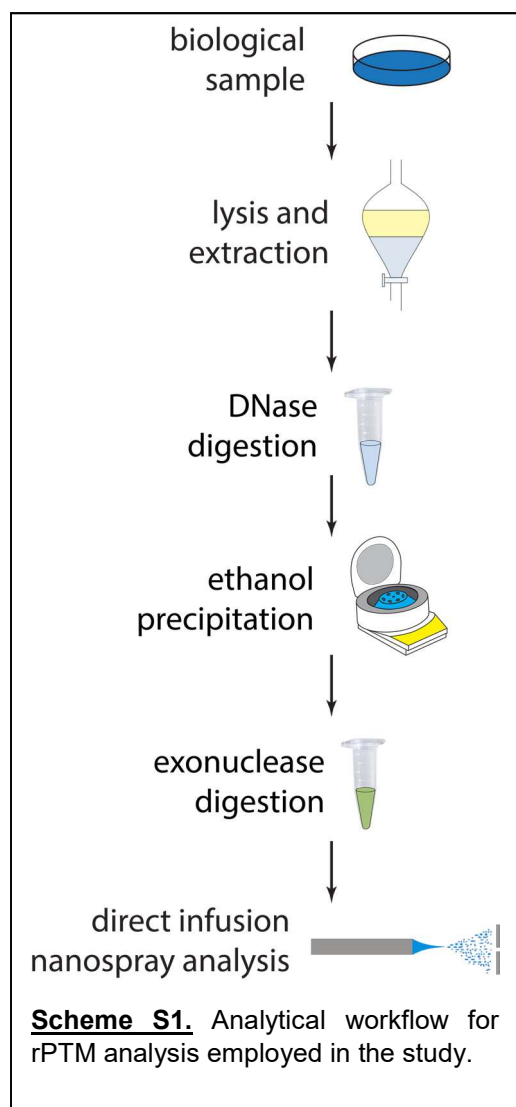
Table S3: Comprehensive summary of rPTM profiles obtained in the study.

Table S4: Average accuracy, precision, and recall afforded by testing the *k* nearest neighbor (KNN), naïve Bayes (NB), and gradient boosting (GB) classifiers on the selected representative classes described in **Table S3**.

References. Contains all citations for the entire publication, including available DOI information.

Experimental section

The study involved analyzing biological samples grown/secured under conditions that are generally considered as basal or “standard” by the respective fields. The goal of building a robust classification/identification framework called for the utilization of the most representative possible profile for each given class, which would help minimize any possible ambiguity. For this reason, we did not include samples that had been exposed to different biological/environmental conditions capable of inducing significant deviations from the basal/standard rPTM landscapes. For each class considered, all biological replicates were secured by growing the organism under exactly the same conditions and harvested at the same time/phase of growth cycle to ensure sample-to-sample reproducibility. All samples regardless of class were submitted to the same analytical workflow that was introduced and discussed at length in ref. 10-12. Briefly, the workflow consisted



of lysis/extraction to obtain a fraction containing total RNA and low levels of DNA components; DNase digestion followed by overnight ethanol precipitation to eliminate such components; exonuclease digestion to generate mono-ribonucleotide mixtures; and direct infusion nanospray analysis with no front-end chromatography (**Scheme S1**). Details and conditions for each step are provided below. As stated in the manuscript, no effort was made to separate/isolate the various types of RNAs present in each class, nor to finely tune exonuclease digestion to account for the presence of less susceptible rPTMs, or to assess the chemical stability of individual rPTMs during the entire workflow. The fact that all samples were treated in exactly the same way ensured that the analysis of total RNA produced self-consistent profiles that, regardless of bias, were considered as being properly representative of the respective class under the selected experimental conditions.

The recent acquisition of a Hamilton (Reno, NV) Microlab Nimbus liquid handler equipped with a ThermoFisher (Waltham, MA) Kingfisher Presto for

nucleic acids purification has enabled the nearly complete automation of the sample preparation workflow. The ability to work on a 96-well plate format has allowed us not only to multiplex the workup for increased throughput, but also to reduce the actual volumes handled at each step. Combined with reduced losses incurred during handling, the overall sample consumption has dropped significantly since the workflow was originally introduced. Using HeLa samples to estimate typical consumption and detection limits, we determined that approximately ~400 ng of total RNA extracted from ~10,000 cells is more than sufficient to obtain acceptable profiles containing the least abundant rPTM observed in these samples (i.e., *acp³Um/acp³Ym* with a typical relative abundance of ~0.005% AvP, *vide infra* for definition). The fact that MS analysis tends to consume only a ~4% fraction of that amount places a possible detection limit in the range of 10-20 ng of total RNA. Considering again the least abundant *acp³Um/acp³Ym*, we could estimate a detection limit of 0.5-1.0 pg. In terms of possible dynamic range, typical determinations afforded at least 3 orders of magnitude between the most and least abundant rPTMs detected in the sample (i.e., monomethyl-Gs and *acp³Um/acp³Ym* with relative abundances of ~1.078 and ~0.005% AvP, respectively). The MS section below provides additional figure of merits reported for the actual analysis itself.

Biological samples. *Escherichia coli* (*E. coli*) MG1655 (serotype OR:H48:K) and CDC EDL 933 (serotype O157:H7) were obtained from ATCC (Manassas, VA), but grown respectively in nutrient broth (NB) purchased from VWR (Radnor, PA) or tryptic soy medium (ATCC) to comply with accepted protocols. The initial cultures were streaked onto either NB or tryptic soy agar plates and incubated overnight at 37°C. Liquid cultures in the respective media were incubated at 37°C and 240 rpm of gyration, and the optical density at 600 nm (OD₆₀₀) was monitored at regular intervals on a ThermoFisher Scientific (Waltham, MA) Nanodrop 2000c spectrophotometer. The cells were harvested when an OD₆₀₀ of 0.3 units was reached. *Klebsiella aerogenes* (*K. aerogenes*) and *Salmonella typhimurium* (*S. typhimurium*) were grown in NB medium under exactly the same conditions described for *E. coli*. *Listeria monocytogenes* (*L. monocytogenes*) and *Streptococcus pneumoniae* (*S. pneumoniae*) were grown under the same conditions in brain heart infusion broth/agar (ATCC). The archaeon *Halobacterium salinarum* (*H. salinarum*) was grown in Van Neil's yeast broth (ATCC16) and incubated at 37°C on Van Neil's yeast agar plates over a three-week period. Save for the specific media requirements, all cells were cultured and harvested under the same conditions.

Saccharomyces cerevisiae (*S. cerevisiae*) strain BY4741 (defined here as wildtype, WT) and *trf4Δ::kanMX*, *rit1Δ::kanMX*, and *set1Δ::kanMX* knockouts were purchased directly from Horizon Discovery (Cambridge, UK). Each sample was grown in yeast extract, peptone, dextrose (YPD)

medium, streaked onto YPD agar, and incubated at 30°C. For all studies, an individual colony was selected from a plate and placed into an individual tube containing 20 mL of YPD. The culture was incubated at 30°C with 200 rpm gyration until an OD₆₀₀ greater than 0.3 units was reached. Liquid cultures were diluted to a final OD₆₀₀ of 0.3 units and centrifuged at 6,000×g for 5 min to obtain pellets that contained approximately the same number of cells.

Arabidopsis thaliana (*A. thaliana*) was grown at 22°C under constant light with 50-60% humidity. Plants were harvested prior to flowering. *Acineta superba* (*A. superba*) was grown in a greenhouse whose temperature was maintained at 21°C during the daytime and 12.8°C during the night. The natural light filtering through the greenhouse was regulated to reduce direct sun exposure during the summer and to maintain the desired temperature during the winter. No additional light was provided from external sources. Peter's professional fertilizer was applied every 3 weeks and tap water was supplied by an irrigation system.

Aedes aegypti (*A. aegypti*) eggs were collected in the wild in Poza Rica (Mexico) and reared to adulthood at 27°C. For these experiments, 4 to 7-day old female mosquitoes were deprived of sucrose in their diet for 18-24h. The mosquitoes were then fed with a 2.5% sucrose solution mixed into sheep's blood. This blood meal was offered through a Hemotek membrane feeding system at 37°C (Discovery Workshops, Acrinton, UK) enclosed in a porcine sausage casing membrane. After 1 hr of feeding, the mosquitoes were sedated by using CO₂, placed in separate 0.6 L cartons, and maintained at 27°C for 14 days until culling for experimental testing.

HeLa and 293T cells purchased from ATCC were cultured in Dubelcco's modified Eagle's medium (DMEM) added with 10% fetal bovine serum (FBS) and 1% glutamine. Human glioblastoma astrocytoma (U251-MG) cells from Sigma (St. Louis, MO) were cultured in Eagle's minimal essential medium (EMEM) with 10% FBS and 1% glutamine. Primary human astrocytes obtained from ThermoFischer Scientific were maintained on cell culture plates coated with Geltrex matrix (ThermoFischer Scientific), The plates were prepared by applying a 1:200 mixture of DMEM and Geltrex matrix onto the surface, incubating for 1 hr at 37°C, and then storing at 4°C for up to 1 week prior to use. Immediately before cell plating, the Geltrex matrix mixture was removed and plates were washed once with Dulbecco's phosphate buffered saline (PBS) containing calcium and magnesium chloride. Primary human neurons obtained from Neuronomics (Bethlehem, PA) were placed on plates coated with Alphabiocoat solution for 30 min at 37°C and washed twice with 5 mL of PBS immediately prior to use. These primary cells were maintained in DMEM containing 10% FBS and 1% N-2 supplement. Human primary microglia obtained from Celprogen (Torrance, CA) were maintained in complete growth medium on plates coated with

Human Microglia Cell Extracellular Matrix. All human cell lines were maintained at 37°C and 5% CO₂.

Sample preparation. Cell pellets from bacteria and archaea strains were treated with 1 mL of TRIzol reagent (Thermo Fisher Scientific) according to the vendor recommendations to achieve cell lysis and nucleic acid extraction. *S. cerevisiae* and knockout mutants were lysed by using 1 mL of TRIzol reagent in the presence of 0.5 mm glass beads (BioSpec Products, Bartlesville, OK). In the case of mammalian cells, the medium was aspirated off the plates. A 1 mL volume of TRIzol was added directly to the plates, pipetted up and down several times to detach the cells from the surfaces and achieve lysis, and finally transferred to Eppendorf tubes. In the case of plant samples, the leaves were washed thoroughly and cut to small pieces. Approximately 100 mg of freshly cut material were frozen in liquid nitrogen, homogenized by using a mortar with pestle, added with 1 mL of TRIzol, mixed vigorously, and transferred to Eppendorf tubes. Each TRIzol sample was centrifuged at 13,000×g for 15 min at 4°C to eliminate solid debris. A 200 µL volume of chloroform was then added, and the resulting mixture was shaken vigorously for 20 s to extract the nucleic acid material. After 10-15 min incubation at room temperature, samples were spun down at 12,000×g for 10-30 min at 4°C. The upper aqueous layer with a typical volume of 400-600 µL, which contains primarily RNA and low-level DNA contamination, was removed and transferred into a new tube. Each sample was added with isopropanol (Fisher Scientific, Fairlawn, NJ) in a 1:1 volume ratio, mixed, and incubated at room temperature for 10 min to induce nucleic acid precipitation. The sample was spun down at 12,000×g for 10-30 min and the supernatant discarded. The resulting pellet was washed with 70% cold ethanol twice, air dried and resuspended in 30-50 µL of chromatography-grade water (Sigma-Aldrich). The low-level DNA component was enzymatically digested with DNase I (ThermoFischer Scientific) in 1x DNase buffer (New England Bio Labs, Ipswich, MA). The solution was then submitted to overnight ethanol precipitation to separate intact RNA from the digestion products, salts, and other soluble components. The pellets were resuspended into 20 to 50 µL RNase free water (Sigma Aldrich). RNA concentrations were determined by UV reading at 260 nm and adjusted to 40 ng/µL. The sample was then treated with nuclease P1 and phosphodiesterase from snake venom (Sigma Aldrich) to obtain the desired mononucleotide (NMP) mixture for mass spectrometric analysis, as previously described.¹⁰⁻¹² It is important to note that this analytical platform is capable of correctly discriminating NMPs from their deoxy counterparts (dNMPs) based on their masses and fragmentation properties. The fact that no significant amounts of dNMPs were ever observed in these samples confirmed the ability of the ethanol precipitation step to eliminate not only the

products of DNase I digestion, but also any possible mononucleotides pre-existing in the cell extracts. This observation is consistent with the fact that ethanol precipitation tends to be significantly more efficient with polymeric than monomeric nucleic acid components, thus enabling mutual separation under proper conditions.

Mass spectrometry. Immediately before analysis, each mixture obtained by exonuclease digestion of total RNA extract was diluted to 4 ng/ μ L in 150 mM MS grade ammonium acetate (Sigma Aldrich) and 10% isopropanol. Each sample was analyzed in negative ion mode either on a Waters Synapt G2 HDMS ion mobility spectrometry mass spectrometer (IMS-MS) or a Thermo Scientific LTQ-Orbitrap Velos, as previously described.¹⁰⁻¹² Analyses were accomplished by static nanoflow electrospray (nanospray) ionization in direct infusion mode by using quartz emitters produced in house. In the absence of chromatographic or pumping system to provide a positive flow, the sample in the emitter is pulled toward the tip by capillary action driven by solution consumption at spray tip. Up to 5 μ L samples were typically loaded into each emitter by using a gel loader pipette tip and analysis consumed \sim 1-2 μ L of sample. A stainless-steel wire was inserted through the emitter's back-end to supply an ionizing voltage that ranged between 0.8 and 1.0 kV. Source temperature and desolvation conditions were adjusted by closely monitoring the incidence of ammonium adducts and water clusters.

Analyses performed on the IMS-MS instrument provided heat maps that offered comprehensive representations of the entire complement of rPTMs present in the sample.¹⁹ For these experiments, the instrument was calibrated by using a 2 mg/mL solution of cesium iodide in 50:50 water/methanol. This external calibration provided typical \sim 9 ppm accuracy.¹¹ For comprehensive mixture analysis, the Tri-WAVE region was held at a pressure of approximately 4.40 mbar (uncalibrated gauge reading) by a 90 mL/min flow of N₂ and 180 mL/min of He. It was operated with an approximately 650 m/s IMS wave velocity, a 40 V wave height, a 109 m/s transfer wave velocity, and a 2.0 V transfer wave height.

Analyses performed on the LTQ-Orbitrap provided accurate mass and fragmentation data that were used to confirm the assignments afforded by the IMS-MS experiments. For high resolution/accuracy determinations, the instrument was calibrated by using an anion mixture that contained sodium dodecyl-sulfate, sodium taurocholate, and Ultramark, or 1mg/mL of cesium iodide. We have previously demonstrated that this external calibration can afford typical \sim 200 ppb accuracy.¹⁰ For analytes the size of mono-ribonucleotides, this level of accuracy reduces considerably the number of possible elemental compositions that may fit the observed masses, thus increasing the level of confidence in the data assignments. Accurate mass determinations

were carried out at a typical ~200,000 resolution. When necessary, rPTM identity was further corroborated by performing tandem mass spectrometry (MS/MS), multistep activation experiments (MSⁿ), and consecutive reaction monitoring (CRM) in both positive and negative ion mode, as previously described.¹⁰⁻¹² In previous work, we utilized serial dilutions of a 1 mg/mL stock of commercial tRNA^{Phe} to determine a limit of detection (LOD) of $\sim 2.1 \times 10^{-11}$ M (~21 pM) with a consumption of $\sim 20.0 \times 10^{-18}$ mol (~20 attomol) and a limit of quantification (LOQ) of $\sim 2.1 \times 10^{-10}$ M (~210 pM) with a consumption of $\sim 203 \times 10^{-18}$ mol (~203 attomol), both calculated for the representative m²G modification.¹ The dynamic range's upper limit (i.e., limit of linearity, LOL) was not determined at that time, but no deviations from linearity were noted at the highest concentration examined, which consisted of the initial stock.

Data interpretation and processing. The experimental masses recorded during analysis were searched against a non-redundant database containing the masses of all known rPTMs,¹⁰ which was compiled in house from information available in the RNA Modifications (<http://mods.rna.albany.edu/>) and MODOMICS (<http://modomics.genesilico.pl/>) databases, as well as pertinent literature in the field. The relative abundance of each rPTM was expressed as Abundance versus Proxy (AvP),¹⁰ which was calculated according to the following equation:

$$\text{AvP}_x = \frac{I_x}{\sum I_c} \cdot 100\% \quad \text{Equation S1}$$

in which the signal intensity (I_x) of each rPTM was normalized against the sum of the intensities displayed in the same spectrum by the four canonical bases (I_c), pooled together to serve as a proxy internal standard. The validity of this treatment was corroborated by the negligible deviations observed between AvP values and corresponding relative abundances calculated against the sum of the intensities of all mononucleotides (i.e., canonicals and rPTMs) in the sample. In a previous study in yeast, further validation was afforded by the close match between AvP values obtained from this treatment and from absolute abundances determined according to the standard addition method.¹⁰

Replicate analyses were carried out for each type of organism/class to enable a proper assessment of the statistical significance of the results. For each class, the text explicitly indicates the total number of biological and technical replicates (N), which were respectively acquired from individual samples processed independently, or from the same sample analyzed in separate MS determinations. The replicate data were used either as individual inputs for machine learning operations, or averaged together to build summary tables in which each column constitutes the representative (average) rPTM profile of that type of organism. In such tables, a hot-cold color

gradient is used to visually represent the average AvP values.¹⁰ The left-most column contains the reference class used to perform a paired Student t-test to assess the statistical significance of the variations exhibited by the remaining classes in the table. In this way, each rPTM was assigned a different color only if the respective value differed from that of the corresponding reference with a p value of less than 0.05.

Machine Learning Models. The data afforded by rPTM analysis were used as input for selected machine learning (ML) algorithms, namely the K-nearest neighbors (KNN), naïve Bayes (NB), and gradient boosting (GB) classifier, included in the Scikit-learn v.0.23 (<https://scikit-learn.org/stable/>) package written in Python. Each input datapoint consisted of the entire collection of pairwise identity/AvP information for all the rPTMs observed in a given replicate, which was defined here as the rPTM profile obtained from the individual analysis. If a certain rPTM was not detected, it was still assigned an AvP value of 0 to express its absence from the profile. No attempt was made to weed out profiles in which low-abundance species had fallen below the detection limits of MS analysis due to subtle sample-to-sample variations. The entire available dataset was randomly split by the software in 70:30 proportions to obtain separate training and testing sets. The training process followed an iterative leave-one-out (LOO) approach in which each datapoint was in turn set aside, whereas the remaining ones were used to identify the most stringent possible rules for class differentiation. The validity of such rules was then evaluated by checking their ability to assign the set-aside datapoint to the correct class. Once the LOO cross-validation process was complete, the testing set was used to assess the performance of the classifier, which was expressed in terms of accuracy, precision, and recall. Accuracy measures the overall ability to reach a correct assignment, which is determined from the number of correct over total number of assignments. Also known as positive predictive value, precision conveys the fraction of positive assignments that were actually relevant in the context of all the instances returned as positive by the classifier, which could be either true (tp) or false (fp):

$$Precision = \frac{tp}{tp+fp} \quad \text{Equation S2}$$

Also known as sensitivity, recall conveys the fraction of relevant positives over the actual number of assignments that should have been returned as positives, but were instead split between true positives and false negatives (fn):

$$Recall = \frac{tp}{tp+f} \quad \text{Equation S3}$$

These metrics were expressed in either fractional or percentage terms to compare the overall performance of the three ML algorithms, to assess the influence of dominant features in each

class, and to investigate the effects of possible evolutionary relationships between classes. In this direction, we also tested the possibility of using the results afforded by ML algorithms to generate putative phylogenetic trees based on RNA modifications. This task was achieved, for example, by using the results filtered by the GB classifier to calculate the Euclidean distances between relevant features, which became in turn the input for an agglomerative clustering approach supported by the `sklearn.metrics.pairwise.euclidean_distances` program of the Scikit-learn v.0.23 package.

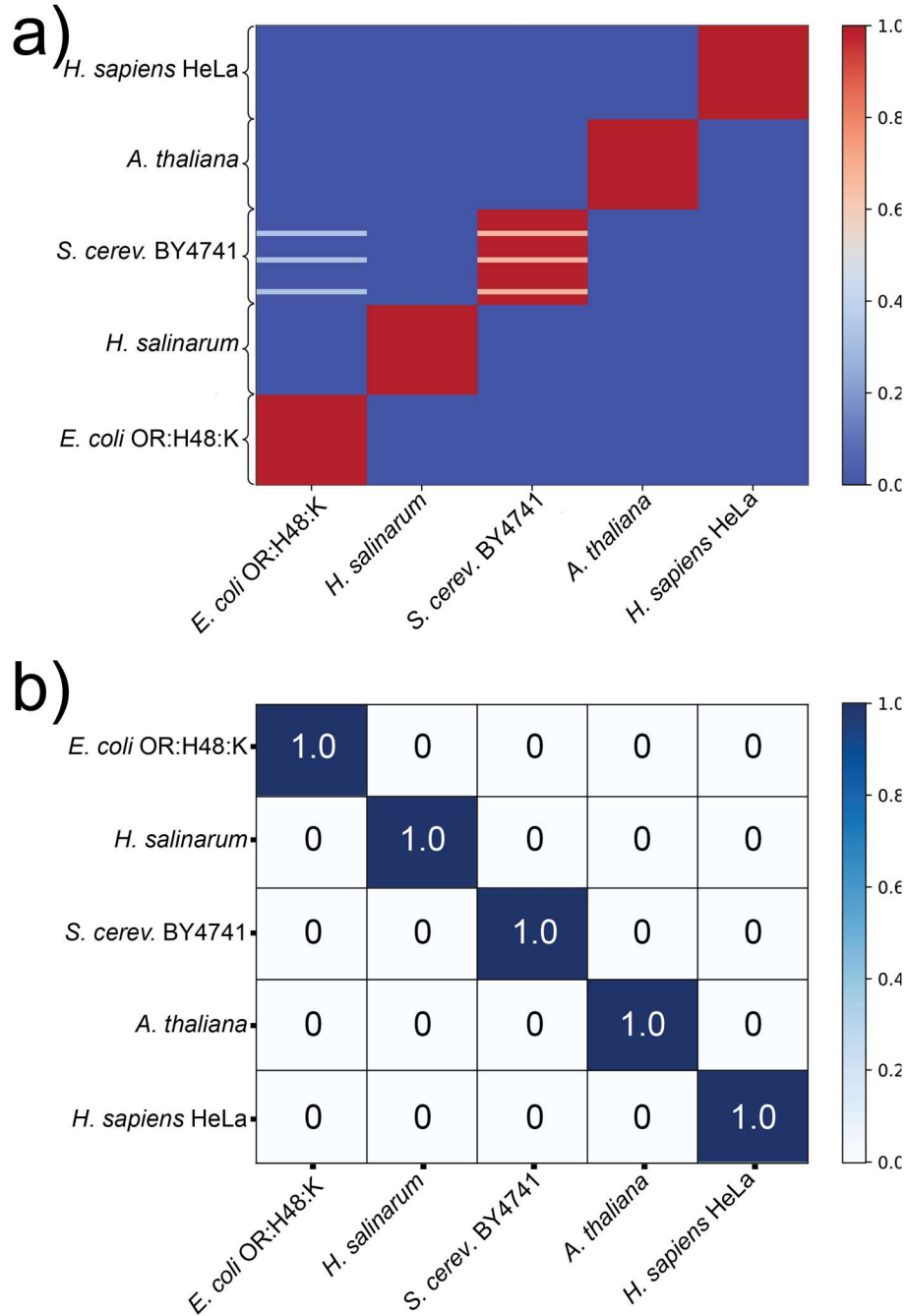


Figure S1. Results observed upon training **a)** and testing **b)** the k nearest neighbor (KNN) algorithm on the data obtained from the representative organisms summarized in **Table S1** (see **Experimental**). The initial dataset, which comprised the rPTM profiles acquired in 25 replicate analyses per organism (i.e., $N = 25$), was randomly split into training and testing sets of 70:30 proportions. **Panel a)** displays the frequency by which a certain training datapoint plotted on the y-axis was assigned to a certain class on the x-axis, during leave-one-out (LOO) iterations between rule-formulation and cross-validation steps. **Panel b)** displays the frequency by which testing datapoints of a given class plotted on the y-axis were assigned to one of the classes on the x-axis. Frequencies of assignment are conveyed by color gradients.

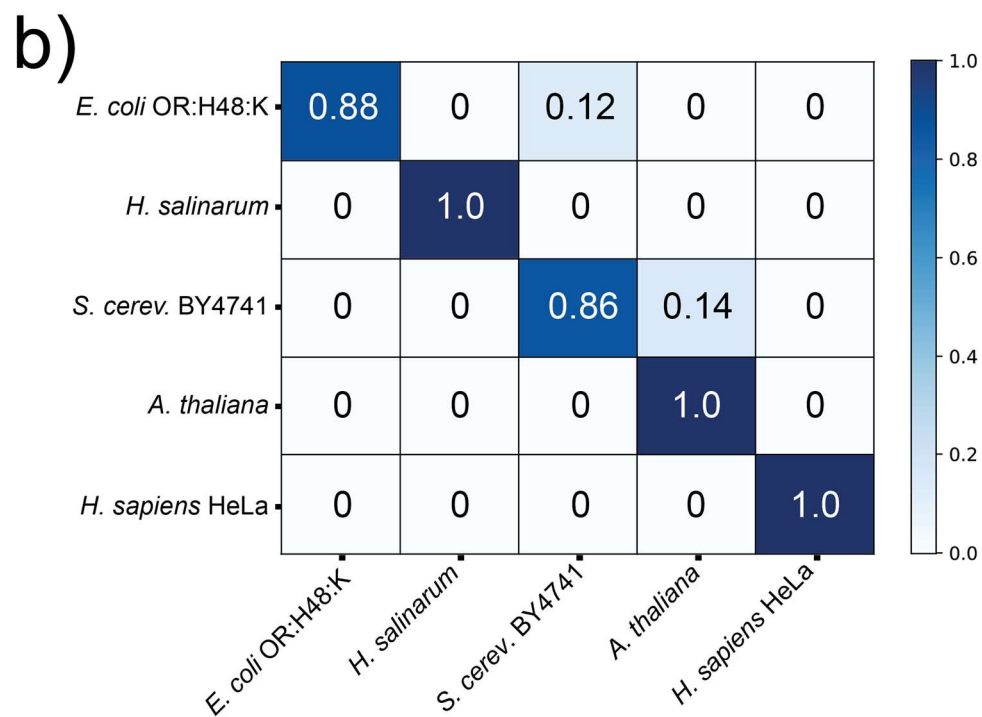
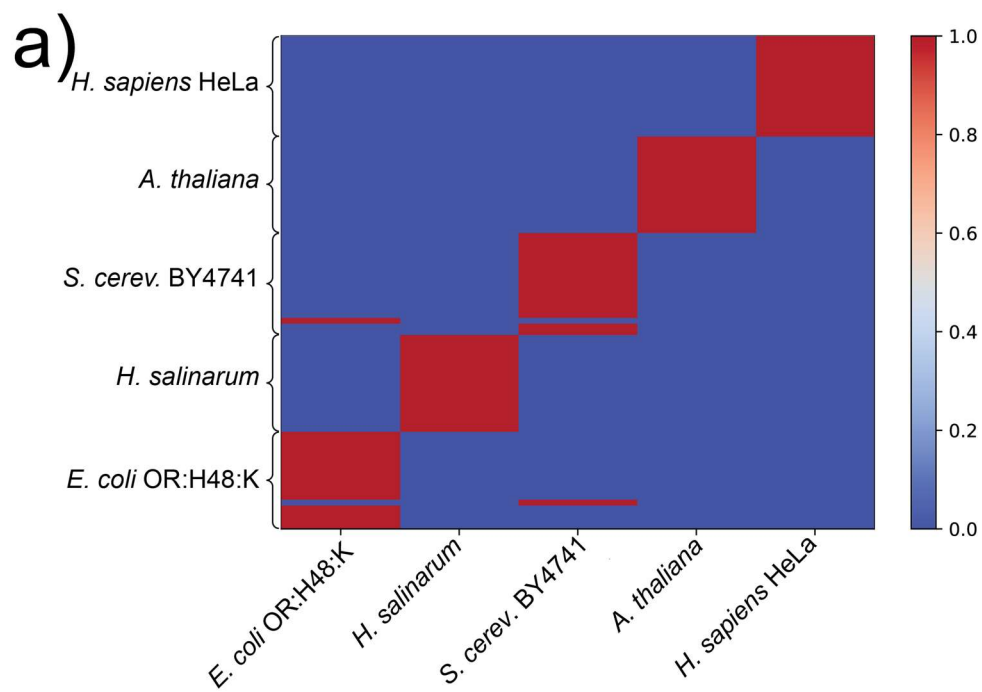


Figure S2. Results observed upon training **a)** and testing **b)** the naïve Bayes (NB) algorithm on the data obtained from the representative organisms summarized in **Table S1** (see *Experimental*). The initial dataset, which comprised the rPTM profiles acquired in 25 replicate analyses per organism (i.e., $N = 25$), was randomly split into training and testing sets of 70:30 proportions. **Panel a)** displays the frequency by which a certain training datapoint plotted on the y-axis was assigned to a certain class on the x-axis, during leave-one-out (LOO) iterations between rule-formulation and cross-validation steps. **Panel b)** displays the frequency by which testing datapoints of a given class plotted on the y-axis were assigned to one of the classes on the x-axis.

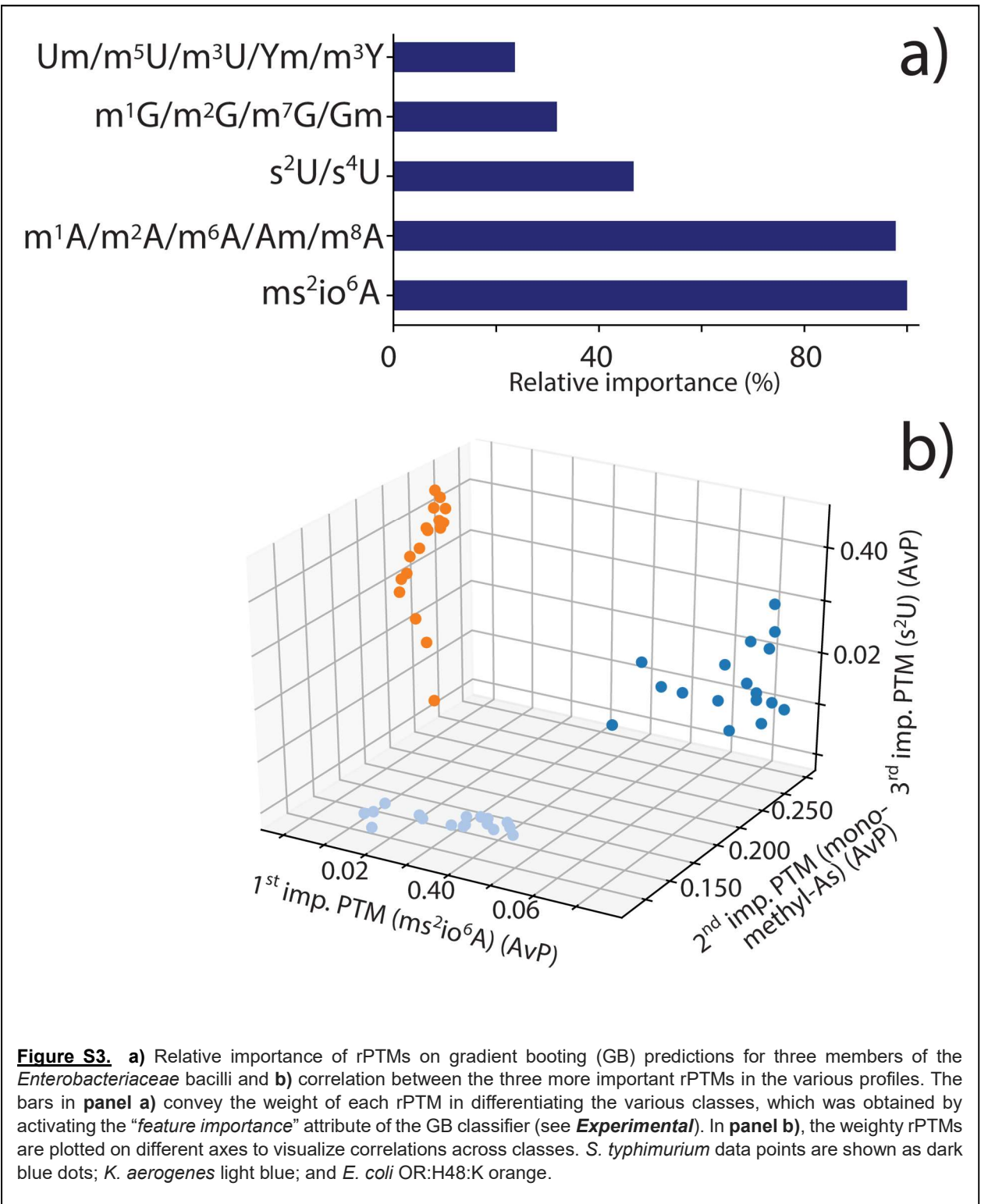


Figure S3. **a)** Relative importance of rPTMs on gradient booting (GB) predictions for three members of the *Enterobacteriaceae* bacilli and **b)** correlation between the three more important rPTMs in the various profiles. The bars in **panel a)** convey the weight of each rPTM in differentiating the various classes, which was obtained by activating the “feature importance” attribute of the GB classifier (see *Experimental*). In **panel b)**, the weighty rPTMs are plotted on different axes to visualize correlations across classes. *S. typhimurium* data points are shown as dark blue dots; *K. aerogenes* light blue; and *E. coli* OR:H48:K orange.

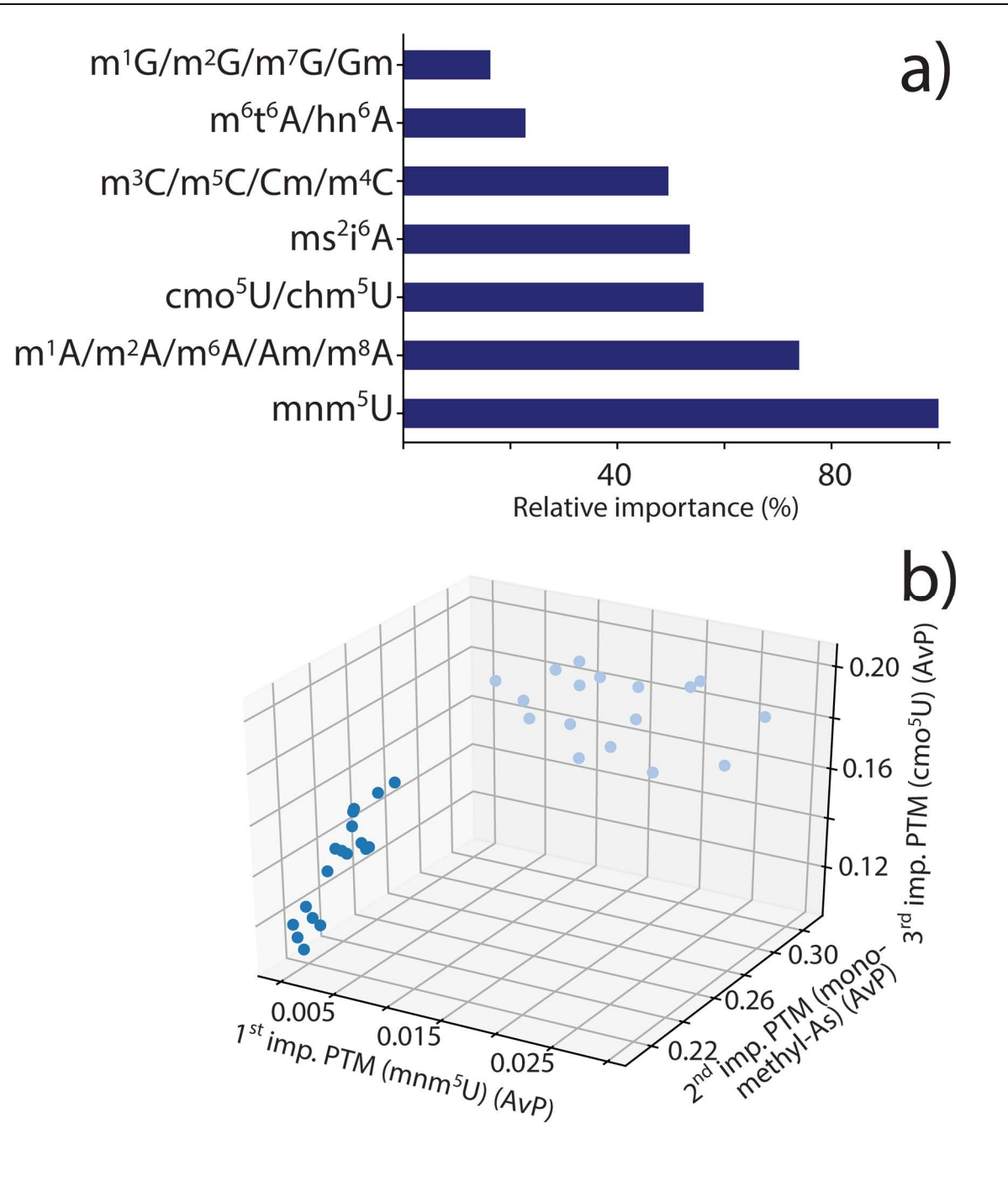
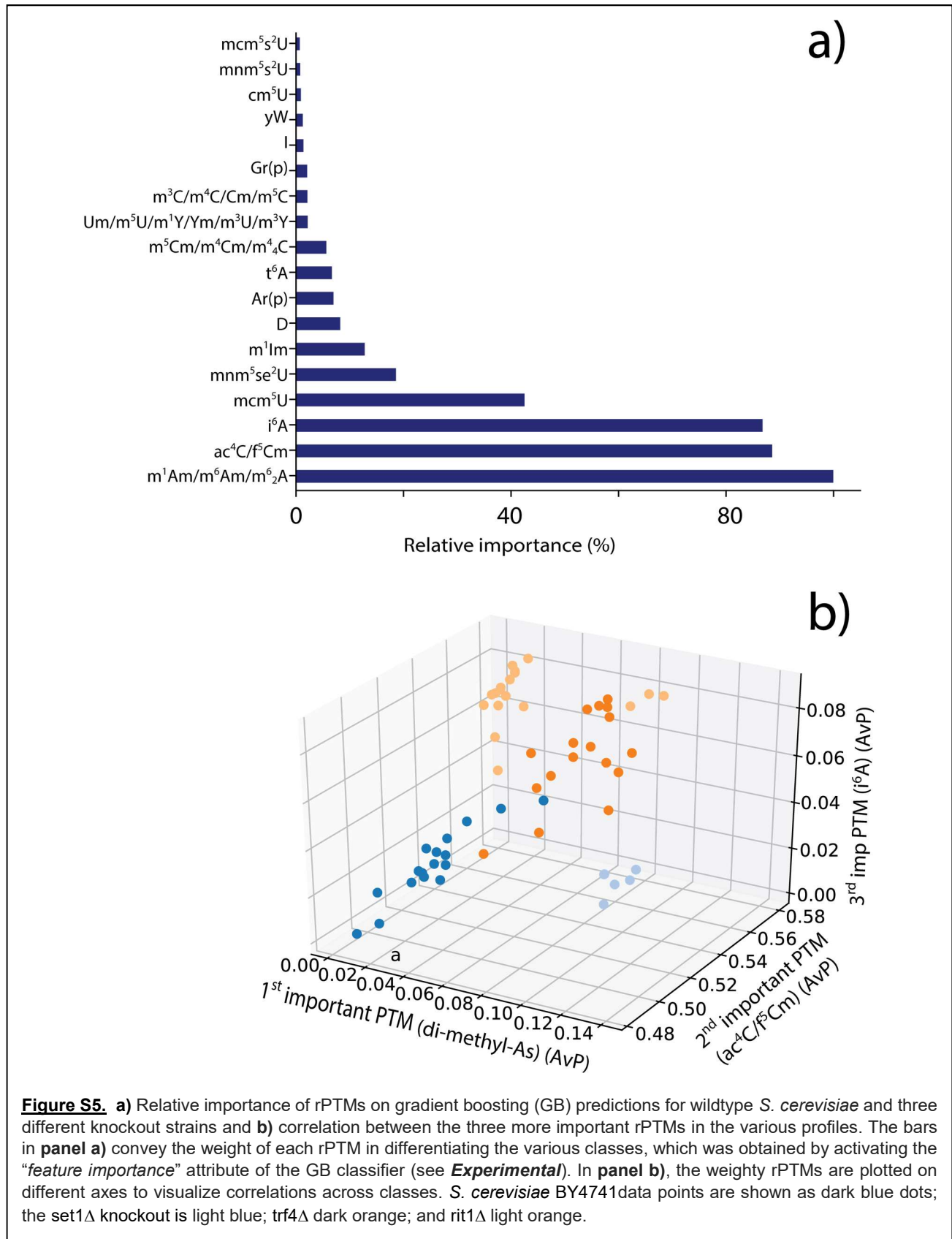


Figure S4. **a)** Relative importance of rPTMs on gradient boosting (GB) predictions for two different *E. coli* serotypes and **b)** correlation between the three more important rPTMs in the various profiles. The bars in **panel a)** convey the weight of each rPTM in differentiating the various classes, which was obtained by activating the “feature importance” attribute of the GB classifier (see **Experimental**). In **panel b)**, the weighty rPTMs are plotted on different axes to visualize correlations across classes. *E. coli* OR:H48:K and serotype O157:H7 data points are shown as dark and light blue dots, respectively.



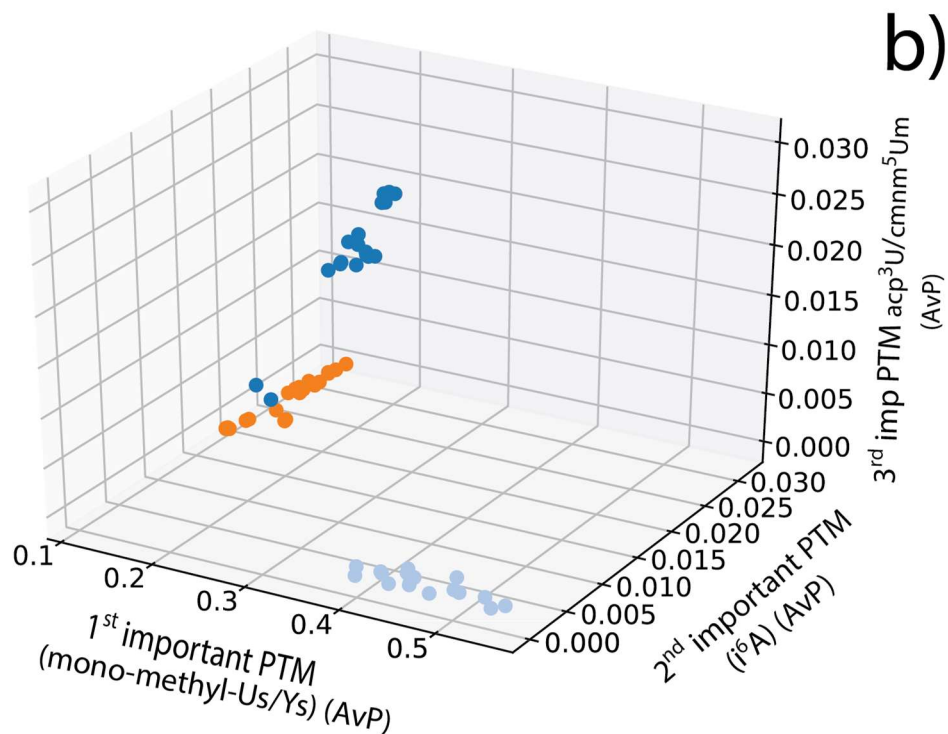
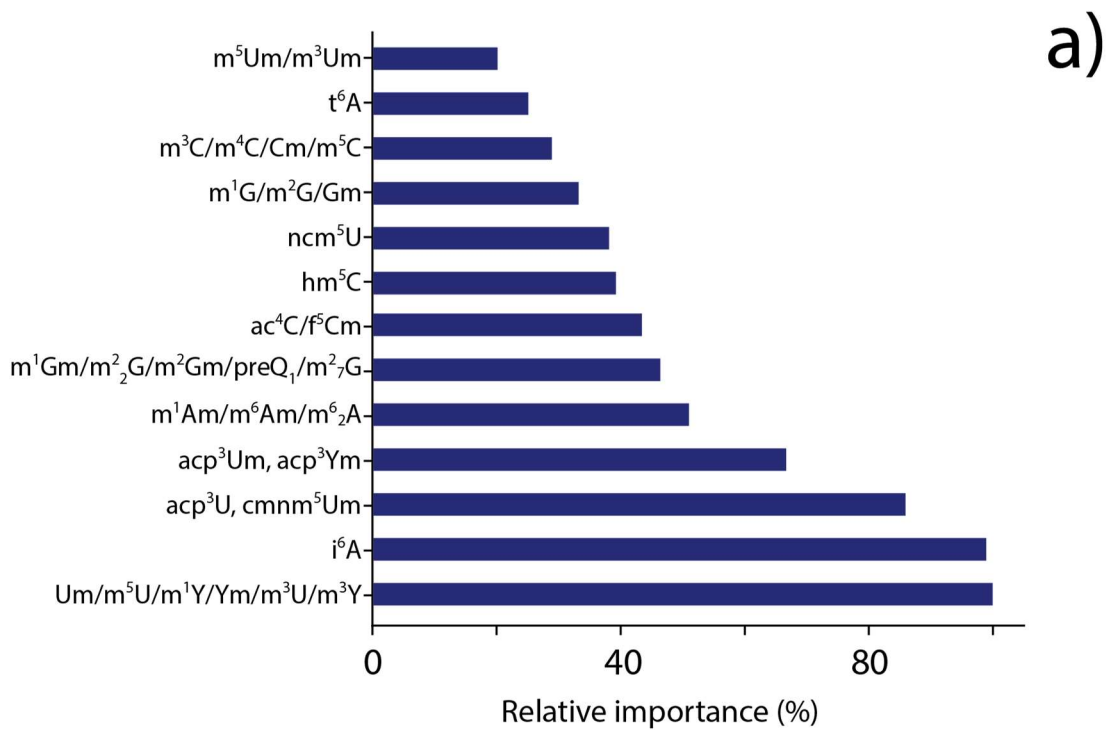
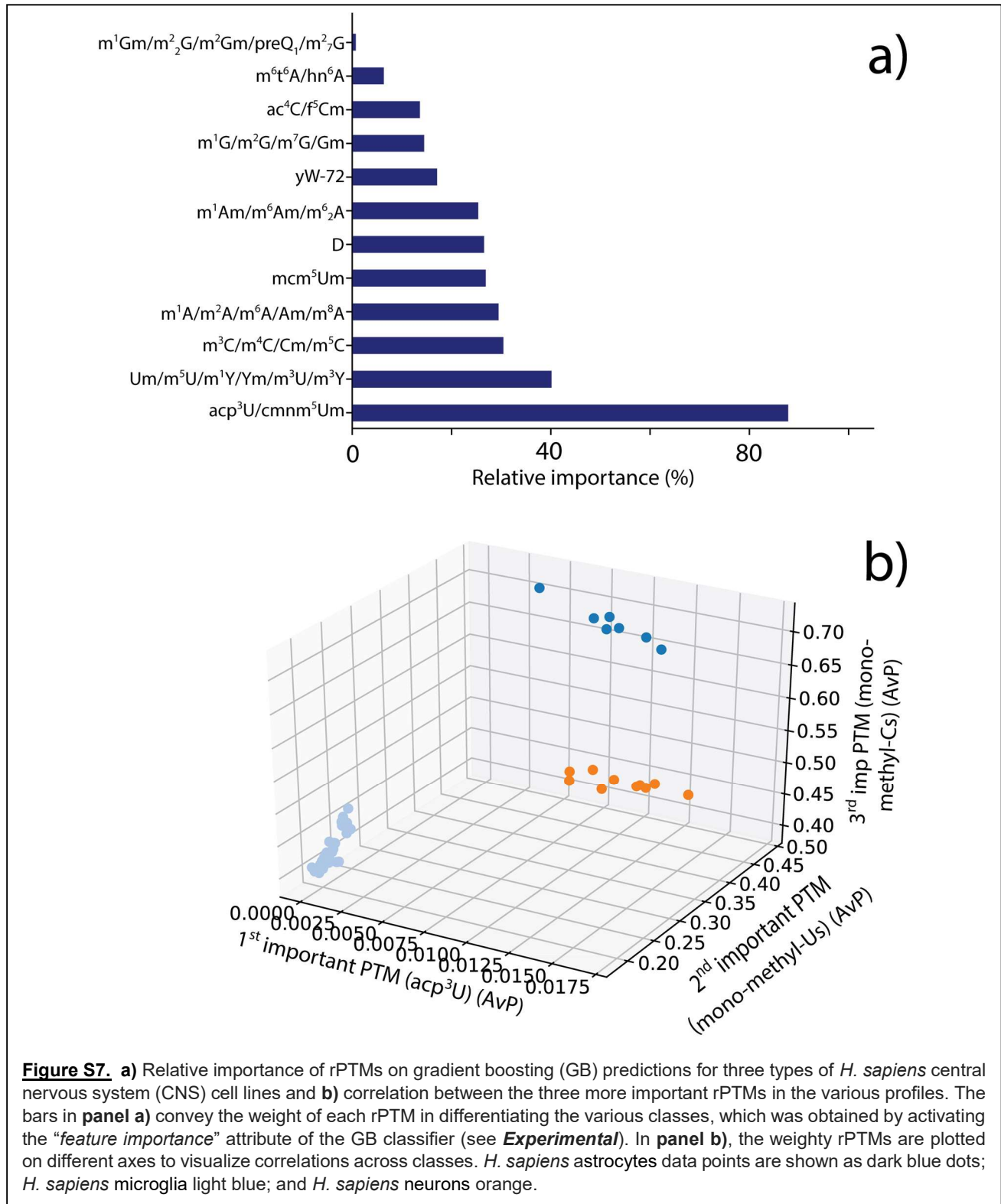


Figure S6. a) Relative importance of rPTMs on gradient boosting (GB) predictions for three types of immortalized *H. sapiens* cell lines and **b)** correlation between the three more important rPTMs in the various profiles. The bars in **panel a)** convey the weight of each rPTM in differentiating the various classes, which was obtained by activating the “feature importance” attribute of the GB classifier (see **Experimental**). In **panel b)**, the weighty rPTMs are plotted on different axes to visualize correlations across classes. *H. sapiens* HeLa data points are shown as dark blue dots; *H. sapiens* HEK 293T light blue; and *H. sapiens* U-251 MG orange.




0.001 AvP 3.6	<i>E. coli</i> OR:H48:K	<i>H. salinarum</i>	<i>S. cerevisiae</i> BY4741	<i>H. thaliana</i>	<i>H. sapiens</i> HeLa
	24	27	32	30	40
D	1.4 ± 0.1		1.5 ± 0.2	0.43 ± 0.07	0.23 ± 0.04
m ³ C, m ⁵ C, Cm, m ⁴ C		0.5 ± 0.2	0.9 ± 0.3	0.54 ± 0.04	0.51 ± 0.03
Um, m ⁵ U, m ¹ Ψ, Ψ m, m ³ U, m ³ Ψ	0.90 ± 0.07	0.24 ± 0.08	0.83 ± 0.08	0.71 ± 0.06	0.32 ± 0.01
s ² C	0.04 ± 0.01				
s ² U, s ⁴ U	0.35 ± 0.09				
I	0.023 ± 0.009		0.12 ± 0.05	0.022 ± 0.005	0.020 ± 0.003
m ¹ A, m ² A, m ⁶ A, Am, m ⁸ A	0.22 ± 0.02		0.4 ± 0.1	0.57 ± 0.02	0.65 ± 0.01
m ¹ I, Im		0.16 ± 0.06			0.04 ± 0.01
ac ⁴ C, f ⁵ Cm	0.51 ± 0.01	0.49 ± 0.02	0.52 ± 0.02	0.521 ± 0.008	0.57 ± 0.01
m ⁶ Am, m ¹ Am, m ⁶ ₂ A		0.014 ± 0.007	0.02 ± 0.01	0.034 ± 0.004	0.031 ± 0.002
m ¹ G, m ² G, Gm, m ⁷ G	0.66 ± 0.06	0.10 ± 0.03	1.0 ± 0.2		1.00 ± 0.08
ac ⁴ Cm					0.011 ± 0.002
ncm ⁵ U					0.016 ± 0.003
mnm ⁵ s ² U	0.05 ± 0.01				
m ¹ Gm, m ² ₂ G, m ² Gm, preQ1, m ² ₇ G		0.14 ± 0.05	0.2 ± 0.1	0.16 ± 0.03	0.09 ± 0.02
cmo ⁵ U, chm ⁵ U	0.13 ± 0.01				
G+		0.13 ± 0.05			
i ⁶ A			0.02 ± 0.01		
acp ³ U, cmnm ⁵ Um				0.08 ± 0.01	0.024 ± 0.004
acp ³ Um, m ¹ acp ³ Ψ					0.007 ± 0.001
t ⁶ A				0.021 ± 0.006	0.026 ± 0.007

Table S1. Representative rPTM profiles obtained from *E. coli* OR:H48:K, *H. salinarum*, *S. cerevisiae* BY4741, *A. thaliana*, and *H. sapiens* HeLa cells (see **Experimental**). Each value represents the average and standard deviation of the relative abundances (AvP) observed in five biological samples, which were each analyzed individually five times (total $N = 25$ replicates). A different color was assigned only if the rPTM's relative abundance was statistically different from that of the reference (in this case *E. coli*) with a p value not exceeding 0.05.

	KNN			NB			GB		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
<i>E. coli</i> OR:H48:K	1.00	1.00	1.00	0.95	1.00	0.88	1.00	1.00	1.00
<i>H. salinarum</i>		1.00	1.00		1.00	1.00			
<i>S. cerevisiae</i> BY4741		1.00	1.00		0.86	0.86			
<i>A. thaliana</i>		1.00	1.00		0.89	1.00			
<i>H. sapiens</i> HeLa		1.00	1.00		1.00	1.00			

Table S2. Average accuracy, precision, and recall afforded by testing the k nearest neighbor (KNN), naïve Bayes (NB), and gradient boosting (GB) classifiers on the selected representative classes described in **Table S1** (see **Experimental**). Testing was accomplished by using $N = 8$ datapoints per class, corresponding to 30% of all the rPTM profiles available for each class (see text). Each reported accuracy represents the average of the values afforded by the classifiers for the various classes under consideration.

0.001  3.6	<i>E. coli</i> OR:H48:K	<i>H. salinarum</i>	<i>E. coli</i> O157:H7	<i>K. aerogenes</i>	<i>S. typhimurium</i>	<i>S. pneumoniae</i>	<i>L. monocytogenes</i>	<i>S. cerevisiae</i> BY4741	<i>S. cerevisiae</i> Δset1	<i>S. cerevisiae</i> Δtrf4	<i>S. cerevisiae</i> Δtrf1	<i>A. thaliana</i>	<i>A. superba</i>	<i>A. aegypti</i>	<i>H. sapiens</i> HeLa	<i>H. sapiens</i> HEK 293T	<i>H. sapiens</i> U-251 MG	<i>H. sapiens</i> astrocytes	<i>H. sapiens</i> microglia	<i>H. sapiens</i> neurons
U,Y	1.6E1 ±2E-1	1.7E1 ±3E-1	1.7E1 ±2E-1	1.9E1 ±3.8E-1	1.6E1 ±3.2E-1	2.1E1 ±7.4E-1	1.2E1 ±5.4E-1	2.2E1 ±7.3E-1	2.4E1 ±1.3E0	2.1E1 ±4.1E-1	2.1E1 ±3.2E-1	2E1 ±1.9E-1	1.9E1 ±5.4E-1	1.9E1 ±6.2E-1	1.6E1 ±1.2E-1	1.5E1 ±1.9E0	1.6E1 ±5.8E-1	1.8E1 ±1.5E-1	1.8E1 ±2.4E-1	1.7E1 ±8.7E-2
D	1.4E0 ±9.7E-2		1.5E0 ±7.5E-2	1.5E0 ±2.5E-1	1.9E0 ±6.1E-2	6.9E-1 ±5.7E-2	2.1E-1 ±2.1E-2	1.5E0 ±2.2E-1	9.4E-1 ±1.8E-2	1.1E0 ±1.5E-1	1.2E0 ±1.6E-1	4.3E-1 ±7.5E-2	5.1E-1 ±1E-1	1.6E-1 ±7.6E-2	2.3E-1 ±3.7E-2	4E-1 ±1.2E-1	3.7E-1 ±1E-2	3.9E-1 ±7.9E-3	2.9E-1 ±8.6E-3	2.4E-1 ±5E-3
m ³ C, m ² C, Cm, m ² C	2.7E-2 ±1E-2	5.3E-1 ±1.9E-1	7.2E-2 ±4.9E-3	9E-2 ±8.9E-3	7.1E-2 ±1.1E-2	1.2E-1 ±6.1E-2	2E-1 ±2.5E-2	9.4E-1 ±2.9E-1	7.6E-1 ±3E-2	9.7E-1 ±5.9E-2	9.3E-1 ±1E-1	5.4E-1 ±3.6E-2	6.5E-1 ±5.7E-2	2.3E-1 ±8.4E-2	5.1E-1 ±3.3E-2	7.8E-1 ±1E-1	4.5E-1 ±4.5E-2	6.8E-1 ±1.9E-2	4.2E-1 ±1.9E-2	5.2E-1 ±7.1E-3
Um, m ² U, m ¹ Y, Ym, m ³ U, m ³ Y	9E-1 ±6.6E-2	2.4E-1 ±7.6E-2	9.9E-1 ±4.3E-2	6.5E-1 ±8.2E-2	1.1E0 ±5.3E-2	4.8E-1 ±7.6E-2	1.4E-1 ±1.8E-2	8.2E-1 ±8.3E-2	8.2E-1 ±2.3E-2	7.7E-1 ±4.7E-2	7.5E-1 ±6.1E-2	7.1E-1 ±6.3E-2	1.1E0 ±1.7E0	1.7E-1 ±6.7E-2	3.2E-1 ±1.1E-2	4.5E-1 ±4.4E-2	1.3E-1 ±1.2E-2	4.7E-1 ±1.1E-2	2.1E-1 ±1.8E-2	3.5E-1 ±4.1E-3
s ² C	3.8E-2 ±1E-2		3.9E-2 ±6.9E-3	2.9E-2 ±1.3E-2	4E-2 ±8.7E-3		1.1E-2 ±4.2E-3	4.4E-2 ±1.4E-4												
s ² U, s ⁴ U	3.6E-1 ±6.9E-2	4.6E-3 ±3.1E-3	3.3E-2 ±3.4E-2	6.4E-2 ±5.6E-2	1.5E-1 ±7.3E-2	1.8E-3 ±6.2E-4													6.1E-3 ±5.9E-3	2.2E-3 ±9.1E-4
ho ⁵ U						7.6E-2 ±1.2E-1		1.3E-2 ±1.8E-2	1.7E-2 ±7.4E-3					6.1E-2 ±9.8E-2	1E-2 ±9.1E-3		1E-2 ±1.9E-2			
m ⁵ D	1.3E-1 ±7.4E-2	1.3E-1 ±1E-1					2.7E0 ±1.2E0							6E-2 ±8.4E-2				2.3E-3 ±4.2E-9	6E-3 ±3.1E-3	
I	2.3E-2 ±9E-3		3.3E-2 ±2.1E-3	1.7E-2 ±8E-3	4.1E-2 ±8.3E-3	7E-2 ±4.2E-2		1.3E-1 ±4.5E-2	5.8E-2 ±6.6E-3	9.7E-2 ±1.3E-2	1.2E-1 ±1.2E-2	2.2E-2 ±5.4E-3	3.9E-2 ±8.6E-3	3.5E-2 ±1.5E-2	2E-2 ±3.3E-3	3.5E-2 ±1.2E-2	3.6E-2 ±1.7E-3	3.1E-2 ±5.5E-3	1.7E-2 ±2.8E-3	1.4E-2 ±2.3E-3
f ² C				3.4E-2 ±5.4E-2		4.7E-1 ±5.2E-1								1.5E-2 ±1E-2		8.9E-3 ±6.2E-3				
m ⁵ Cm, m ⁴ Cm, m ⁴ C			2.9E-3 ±1.1E-4			3.5E-3 ±1.3E-3		1.2E-2 ±3E-3				6.7E-3 ±1.6E-3		1.7E-2 ±1.5E-2	5.1E-3 ±1.7E-3			2.9E-3 ±1.4E-3		
f ² U						1.4E-2 ±1.4E-2								2.4E-2 ±3.4E-2				3E-3 ±8.9E-5		
m ⁵ Um, m ³ Um		6.3E-2 ±3.8E-2				2.2E-3 ±1.1E-3		9.7E-3 ±4.6E-3						2.4E-2 ±2.8E-2		3.2E-2 ±8.7E-3	2.2E-3 ±1.9E-3			
hm ⁵ C						9.9E-3 ±8.8E-3								3.4E-2 ±3.7E-2				2E-2 ±9.7E-3	6.2E-3 ±4.6E-3	
m ⁵ s ² U, s ² Um		5.3E-2 ±6E-2				1.3E-1 ±1.6E-1											2E-3 ±1.6E-3			
mo ⁵ U	5.8E-2 ±4.8E-2					8.2E-2 ±6.4E-2		3.8E-2 ±3.5E-4				1E-2 ±8E-3	6.3E-3 ±2.7E-3	3.5E-2 ±2.9E-2						
ho ⁵ C														9.2E-2 ±9.8E-2						
Q _{base}						2E-2 ±3.3E-2														1.1E-2 ±4.6E-3
m ¹ A, m ² A, m ⁶ A, Am, m ² A	2.3E-1 ±1.6E-2	1.2E-2 ±6.9E-3	2.8E-1 ±1.1E-2	1.2E-1 ±1.1E-2	2.6E-1 ±5.7E-3	1.6E-1 ±7.4E-2	1.4E-1 ±6.2E-3	4.4E-1 ±8.4E-2	4.5E-1 ±3.9E-2	4.6E-1 ±1.6E-2	4.8E-1 ±2.4E-2	5.7E-1 ±2E-2	5.8E-1 ±3.1E-2	3.2E-1 ±6.3E-2	6.5E-1 ±1.1E-2	3.5E-1 ±8.5E-2	5.8E-1 ±4E-2	7.2E-1 ±2.7E-2	6E-1 ±2.2E-2	5.4E-1 ±8.4E-3
m ¹ l, lm	4.2E-2 ±6.7E-3	1.8E-1 ±6.1E-2	1.2E-2 ±2.5E-4	1.9E-2 ±2.3E-3	7.8E-2 ±1.3E-2	1.4E-1 ±1.7E-1	4.7E-2 ±1.4E-2	4.7E-2 ±2.3E-2	1.3E-2 ±1.9E-6	2.6E-2 ±8.8E-3	2.6E-2 ±6.9E-3	2.3E-2 ±7.1E-3	2.1E-2 ±1.6E-2	2.8E-2 ±1.1E-2	4.4E-2 ±1.2E-2		1.9E-2 ±3.1E-3	7.6E-2 ±4.8E-2	3.8E-2 ±1.2E-2	1E-2 ±8.3E-3
cnm ⁵ U														5.4E-2 ±1.5E-2						
ac ⁴ C, f ² Cm	5.1E-1 ±1.5E-2	4.9E-1 ±2.1E-2	5.4E-1 ±9.1E-3	5.1E-1 ±2.2E-2	5.2E-1 ±2.3E-2	3.8E-1 ±6.3E-2	4.5E-1 ±6.5E-2	5.2E-1 ±1.7E-2	5E-1 ±1.2E-2	5.4E-1 ±9E-3	5.6E-1 ±7.1E-3	5.2E-1 ±7.3E-3	5.4E-1 ±3.5E-2	3.8E-1 ±4.6E-2	5.7E-1 ±1.4E-2	6.3E-1 ±3.4E-2	4.6E-1 ±2.4E-2	5.3E-1 ±6.6E-3	5.2E-1 ±1.3E-2	5.6E-1 ±7.4E-3
m ¹ 2Cm														4.4E-2 ±7.8E-2					3.2E-3 ±7.8E-4	
f ² Um		7.7E-3 ±9.7E-3						1.2E-2 ±1.2E-2	6E-4 ±2.8E-4				2.8E-2 ±1.1E-3	2.4E-2 ±2.6E-2		1.6E-3 ±1.2E-4				
mm ⁵ U			1.6E-2 ±6.8E-3	3.9E-3 ±2.2E-3	6E-3 ±3.1E-3	6.4E-2 ±1.1E-1								2.7E-2 ±2.3E-2						
f ⁵ s ² U						1E-2 ±1E-2													2.2E-3 ±3.2E-4	
nm ⁵ s ² U	4.9E-2 ±4.1E-2	6E-2 ±5.3E-2					4.8E-3 ±2.5E-3			2.2E-2 ±1.6E-2		3.3E-2 ±1.4E-2	1E-2 ±2E-3						2.2E-2 ±1E-2	8.1E-3 ±5.3E-3
m ⁶ Am, m ¹ Am, m ² A		1.4E-2 ±7E-3	5.4E-3 ±2.8E-3	2.8E-3 ±2.1E-3		3.4E-3 ±3E-3		2.3E-2 ±1.3E-2	1.3E-1 ±1.5E-2	7.8E-2 ±1.5E-2	3E-2 ±2.4E-2	3.4E-2 ±3.7E-3	2E-2 ±7E-3	4.3E-2 ±6.6E-2	3.1E-2 ±1.8E-3	2.2E-2 ±3E-3	1.6E-2 ±3E-3 N	7.5E-3 ±1.4E-3	1.7E-2 ±3.2E-3	2.7E-2 ±1.6E-3
m ¹ lm						6E-3 ±9.4E-4		2.7E-1 ±1E-1				8.7E-2 ±2.5E-1		5.3E-2 ±7.9E-2	1.8E-2 ±6.4E-4		1E-2 ±4.4E-3		5.8E-3 ±5E-3	
m ¹ G, m ² G, Gm, m ¹ G	6.6E-1 ±5.5E-2	9.9E-2 ±3E-2	9.4E-1 ±3.6E-2	6E-1 ±5.6E-2	8.9E-1 ±3.3E-2	4.6E-1 ±6.7E-2	4.1E-1 ±5.8E-2	1E0 ±1.6E-1	6.7E-1 ±1.1E-2	1.1E0 ±8.5E-2	1.2E0 ±8.7E-2	8.7E-1 ±8.9E-2	9.9E-1 ±1E-1	4.4E-1 ±1.2E-1	1E0 ±7.5E-2	6.2E-1 ±6.8E-2	6.9E-1 ±5.1E-2	9.2E-1 ±2.6E-2	6.9E-1 ±2.4E-2	7.4E-1 ±1.1E-2
ac ⁴ Cm	1.9E-1 ±3.8E-1		7.9E-3 ±1.3E-3		3.4E-3 ±1.6E-3	7E-3 ±4.2E-3	3.1E-1 ±2E-1	6.1E-3 ±2.8E-3		6.5E-3 ±2.4E-3	1.2E-2 ±3.9E-3	2.9E-3 ±2.9E-4	6.4E-3 ±2.7E-3	1.4E-2 ±1.1E-2	1.1E-2 ±1.7E-3	4.1E-3 ±1.3E-3	5.4E-3 ±1.7E-3	5.7E-3 ±2.6E-3	2E-3 ±6.9E-4	2.3E-3 ±4.7E-4
ncm ⁵ U						1.5E-2 ±2.4E-2	7.1E-1 ±3.9E-1	3.1E-2 ±1.3E-2	1.3E-2 ±2.1E-6	3.2E-2 ±5.2E-3	4.1E-2 ±8.7E-3	4.3E-3 ±5.8E-4	1.3E-2 ±5.6E-3	2E-2 ±1.9E-2	1.6E-2 ±2.8E-3	7.9E-3 ±3.1E-3	2.1E-2 ±1.1E-2		1.8E-3 ±5.4E-4	1.8E-3 ±4.3E-4
cm ⁵ U						1.7E-1 ±3E-1														
mm ⁵ s ² U	5.5E-2 ±1.2E-2		2E-2 ±1.5E-2	2.9E-2 ±1.9E-2	5.6E-2 ±9E-3	2E-2 ±3E-2		6.8E-2 ±9.9E-4		6.5E-3 ±2.4E-3										

	<i>E. coli</i> OR:H48:K	<i>H. salinarum</i>	<i>E.coli</i> O157:H7	<i>K. aerogenes</i>	<i>S. typhimurium</i>	<i>S. pneumoniae</i>	<i>L. monocytogenes</i>	<i>S. cerevisiae</i> BY4741	<i>S. cerevisiae</i> Δ set1	<i>S. cerevisiae</i> Δ trf4	<i>S. cerevisiae</i> Δ trf1	<i>A. thaliana</i>	<i>A. superba</i>	<i>A. aegypti.</i>	<i>H. sapiens</i> HeLa	<i>H. sapiens</i> HEK 293T	<i>H. sapiens</i> U- 251 MG	<i>H. sapiens</i> astrocytes	<i>H. sapiens</i> microglia	<i>H. sapiens</i> neurons
preQ0				6.9E-2 ±8.2E-2		1.6E-2 ±2E-2														
ac ⁶ A																				2.7E-3 ±1.2E-3
m ¹ Gm, m ² G, m ² Gm, preQ1, m ⁷ G		1.4E-1 ±5.3E-2	1.1E-2 ±6.8E-3			2.9E-1 ±1.1E-1		2.5E-1 ±9E-2	1.3E-1 ±8.6E-3	3.2E-1 ±3.5E-2	2.4E-1 ±8E-2	1.6E-1 ±2.9E-2	2E-1 ±3.9E-2	2.1E-2 ±2.1E-2	9.7E-2 ±1.5E-2		4.2E-2 ±5.5E-3	7.9E-2 ±4E-3	5.7E-2 ±3.8E-3	6.5E-2 ±2.7E-3
ncm ⁵ Um						3.2E-3 ±1.4E-3		9.9E-3 ±5.4E-3		5.8E-3 ±6.3E-4	2.7E-3 ±6.6E-4									
mcm ⁵ U						5.4E-3 ±3.6E-3		1.9E-2 ±8.5E-3	2.9E-2 ±3.4E-6	1.8E-2 ±6.1E-3	2.9E-2 ±7.2E-3					1.1E-2 ±9.6E-3	8.9E-4 ±3.3E-4		3.1E-3 ±1.1E-3	
cmo ⁵ U, chm ⁵ U	1.3E-1 ±1.3E-2		1.9E-1 ±1.4E-2	1.6E-1 ±2.5E-2	1.7E-1 ±1.8E-2	1.8E-2 ±3E-2		1.3E-1 ±1.1E-1												
imG-14		4.7E-3 ±3E-3				4.8E-3 ±4.2E-3	4.6E-2 ±2.9E-2													
G+		1.3E-1 ±5.2E-2				2.6E-3 ±1.5E-3									5.6E-3 ±1.9E-3	6.9E-3 ±5.7E-3		1E-2 ±3.8E-3	2.8E-3 ±1.4E-3	
m ² Gm, m ² G, m ² Gm						3.5E-3 ±2E-3		3.8E-3 ±3.5E-3							3.2E-3 ±1.1E-3		2E-3 ±7.5E-4			
mcm ⁵ Um																	1.2E-2 ±3.7E-3			5E-3 ±1.6E-3
mcm ⁵ s ² U		4.1E-3 ±2.4E-3						2.2E-2 ±2.4E-2		2.7E-2 ±1.2E-2	4.8E-2 ±1E-2				6.4E-3 ±1.5E-3					
mcmo ⁵ U, mchm ⁵ U											4.8E-3 ±2.6E-3	4.6E-3 ±8.4E-4	7.5E-3 ±3.9E-3		4.2E-3 ±1E-3	1.3E-3 ±5.9E-4	8.8E-3 ±1.5E-3			
l ⁶ A	1.3E-3 ±6.3E-4					7E-2 ±2.5E-2		2.3E-2 ±1.1E-2	4.2E-2 ±1E-2	6.3E-2 ±1.2E-2	7.6E-2 ±1.1E-2	3.7E-3 ±6.2E-4		6.1E-3 ±4.6E-3	1.2E-2 ±1.2E-3	1.6E-3 ±7.4E-4	2.3E-2 ±4.7E-3	3.7E-3 ±1.9E-3	3.3E-3 ±1.2E-3	1.9E-3 ±6.8E-4
nm ⁵ se ² U						1.6E-2 ±1.6E-2		9E-3 ±2.1E-4												
acp ³ U, cmnm ⁵ Um	1.4E-1 ±2.5E-2	3.4E-3 ±1.4E-3	1.7E-1 ±1.6E-2	1.8E-1 ±3.2E-2	1.9E-1 ±2.3E-2	2.1E-3 ±1.5E-3	2.1E-1 ±1.3E-1	1.2E-1 ±1E-1				8E-2 ±1.4E-2	7.2E-2 ±1.3E-2	1.8E-2 ±1.7E-2	2.4E-2 ±3.8E-3	5.4E-2 ±1.8E-2	8.7E-3 ±6.3E-3	9.5E-3 ±3.2E-3		1.3E-2 ±2.1E-3
cmnm ⁵ s ² U					2.9E-3 ±2.5E-3	7.6E-3 ±3.4E-3														
mn ⁵ se ² U						3.2E-3 ±1.2E-3			2.3E-2 ±1.2E-2										2E-3 ±5.4E-4	3.2E-3 ±1.9E-3
io ⁶ A						5.2E-3 ±3.4E-3						6.2E-3 ±2.1E-3								
acp ³ Um, m ¹ acp ³ Y								8E-4 ±1.4E-4					3.6E-3 ±4.4E-4		6.9E-3 ±1.4E-3				3.2E-3 ±5.2E-4	
g ⁶ A								6.3E-3 ±4.5E-3												
k ² C				9.8E-3 ±6.5E-3				2.4E-2 ±8.8E-3								9.6E-3 ±5.6E-3		4.9E-3 ±2.2E-3		
ms ² l ⁶ A	6E-2 ±1.3E-2		1.8E-1 ±1.5E-2	4E-3 ±4.6E-3	1.4E-2 ±6.4E-3	3.9E-2 ±2.1E-2		3.9E-2 ±1.2E-2												
ms ² io ⁶ A				3.1E-2 ±1.2E-2	6.6E-2 ±8.4E-3			3.9E-3 ±2.2E-3				6.6E-3 ±3.1E-3	3.3E-3 ±3.5E-4							
Q	1E-1 ±1E-2		8.5E-2 ±9E-3	9E-2 ±1.4E-2	1.3E-1 ±7.6E-3	2.8E-2 ±1.5E-2		9.3E-2 ±2.8E-4							4.2E-3 ±7.3E-4	5.9E-3 ±3E-3	9.2E-3 ±2.4E-3			
l ⁶ A	4.1E-2 ±1E-2	2E-2 ±1.4E-2	5.3E-2 ±5.9E-3	5.8E-2 ±1.3E-2	7.1E-2 ±1.5E-2	2.9E-2 ±1.7E-2		1E-1 ±2.2E-2	2.5E-2 ±3.4E-6	7.3E-2 ±1.1E-2	8.9E-2 ±1.2E-2	2.1E-2 ±5.6E-3	3.2E-2 ±9.6E-3	1.2E-2 ±8.8E-3	2.6E-2 ±6.7E-3	6.1E-2 ±1.8E-2	1.4E-2 ±3.6E-3	5E-3 ±1.5E-3	9.3E-3 ±3.3E-3	1.3E-2 ±2.3E-3
m ⁶ l ⁶ A, hn ⁶ A	3E-3 ±1.7E-3		1.3E-2 ±2.7E-3	8.5E-3 ±3.6E-3	1.4E-2 ±6.9E-3			7.4E-3 ±4.1E-3							5.7E-3 ±2E-3	6.8E-4 ±2.1E-4		1.6E-1 ±4.8E-2		
yW-72																			9.4E-2 ±2.5E-2	
ms ² l ⁶ A						2.2E-3 ±1.4E-3														
ms ² hn ⁶ A						1.5E-1 ±1.6E-1														
Ar(p)						3E-3 ±2.6E-3		8.6E-3 ±4.8E-3		6.5E-3 ±2.4E-3										
Gr(p)						1.9E-2 ±3.9E-2		2.9E-3 ±1.1E-3		3.9E-3 ±9.8E-4										
yW						1E-2 ±6.2E-3		5E-3 ±5E-3		9.1E-3 ±3.3E-3	8.3E-3 ±3.6E-3									
galQ, manQ														6.4E-3 ±8E-4	3.4E-3 ±8.1E-4	1.3E-2 ±6.2E-3	7.6E-3 ±2.2E-3			

Table S3. Comprehensive summary of rPTM profiles obtained in the study. Each value represents the average and standard deviation of the relative abundances (AvP) observed for all replicates obtained for each class, with the exception of those that afforded ambiguous assignments during training/testing of the GB classifier (see **Figure 3** of main text). A different color was assigned only if the rPTM's relative abundance was statistically different from that of the reference (*E. coli* OR:H48:K, 1st column) with a *p* value not exceeding 0.05. rPTMs with abundances greater than 3.6% we assigned a gray-scale gradient. The following classes were included: *E. coli* OR:H48:K (*N* = 23 after GB training/testing); *H. salinarum* (*N* = 25); *E. coli* O157:H7 (*N* = 25); *K. aerogenes* (*N* = 24); *S. typhimurium* (*N* = 23); *S. pneumoniae* (*N* = 85); *L. monocytogenes* (*N* = 9); *S. cerevisiae* BY4741 (*N* = 20); *S. cerevisiae* set1Δ (*N* = 10); *S. cerevisiae* trf4Δ (*N* = 21); *S. cerevisiae* rit1Δ (*N* = 22); *A. thaliana* (*N* = 23); *A. superba* (*N* = 22); *A. aegypti* (*N* = 110); *H. sapiens* HeLa (*N* = 25); *H. sapiens* HEK 293T (*N* = 23); *H. sapiens* U-251 MG (*N* = 29); *H. sapiens* astrocytes (*N* = 9); *H. sapiens* microglia (*N* = 45); and *H. sapiens* neurons (*N* = 14) (see **Experimental** for conditions).

	KNN			NB			GB		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
<i>H. salinarum</i> (N = 8)	0.95	1.00	1.00	0.91	1.00	1.00	0.97	1.00	1.00
<i>E. coli</i> OR:H48:K (N = 8)		1.00	1.00		1.00	1.00			
<i>E. coli</i> O157:H7 (N = 8)		1.00	1.00		1.00	1.00			
<i>K. aerogenes</i> (N = 8)		0.89	1.00		1.00	0.63			
<i>S. typhimurium</i> (N = 8)		1.00	1.00		1.00	0.86			
<i>S. pneumoniae</i> (N = 26)		1.00	1.00		0.93	1.00			
<i>L. monocytogenes</i> (N = 3)		1.00	1.00		1.00	1.00			
<i>S. cerevisiae</i> BY4741 (N = 8)		1.00	0.63		0.44	0.50			
<i>S. cerevisiae</i> set1Δ (N = 3)		1.00	1.00		1.00	1.00			
<i>S. cerevisiae</i> trf4Δ (N = 8)		0.71	0.71		0.70	1.00			
<i>S. cerevisiae</i> rit1Δ (N = 8)		0.70	0.88		1.00	0.88			
<i>A. thaliana</i> (N = 8)		0.89	1.00		0.78	0.88			
<i>A. superba</i> (N = 8)		1.00	0.75		1.00	0.25			
<i>A. aegypti</i> (N = 33)		0.97	1.00		0.86	0.97			
<i>H. sapiens</i> HeLa (N = 8)		1.00	0.88		1.00	1.00			
<i>H. sapiens</i> HEK 293T (N = 8)		1.00	1.00		1.00	1.00			
<i>H. sapiens</i> U-251 MG (N = 10)		1.00	1.00		1.00	1.00			
<i>H. sapiens</i> astrocytes (N = 3)		1.00	1.00		1.00	1.00			
<i>H. sapiens</i> microglia (N = 15)		1.00	1.00		1.00	1.00			
<i>H. sapiens</i> neurons (N = 5)		0.80	1.00		1.00	1.00			

Table S4. Average accuracy, precision, and recall afforded by the ML classifiers generated from samples of the following classes: *H. salinarum*, *E. coli* OR:H48:K, *E. coli* O157:H7, *K. aerogenes*, *S. typhimurium*, *S. pneumoniae*, *L. monocytogenes*, *S. cerevisiae* BY4741, *S. cerevisiae* set1Δ, *S. cerevisiae* trf4Δ, *S. cerevisiae* rit1Δ, *A. thaliana*, *A. superba*, *A. aegypti*, *H. sapiens* HeLa, *H. sapiens* HEK 293T, *H. sapiens* U251 MG, *H. sapiens* astrocytes, *H. sapiens* microglia, and *H. sapiens* neurons (see **Experimental**). The number of datapoints (N) per class refers to the rPTM profiles used during the testing phase, corresponding to 30% of the entire dataset available for each class (see text).

References

- (1) Linné, C. von; Salvius, L. Caroli Linnaei ... Species Plantarum: Exhibentibus Plantas Rite Cognitas, Ad Genera Relatas, Cum Differentiis Specificis, Nominibus Trivialibus, Impensis Laurentii Salvii, Holmiae, 1753, 1–572.
- (2) Woese, C. R. Bacterial Evolution. *Microbiol. Rev.* **1987**, 51 (2), 221–271.
- (3) D'Argenio, V.; Salvatore, F. The Role of the Gut Microbiome in the Healthy Adult Status. *Clin. Chim. Acta* **2015**, 451, 97–102.
- (4) Kitahara, K.; Miyazaki, K. Revisiting Bacterial Phylogeny: Natural and Experimental Evidence for Horizontal Gene Transfer of 16S rRNA. *Mob. Genet. Elem.* **2013**, 3 (1), e24210, 1–5.
- (5) Cantara, W. A.; Crain, P. F.; Rozenski, J.; McCloskey, J. A.; Harris, K. A.; Zhang, X.; Vendeix, F. A. P.; Fabris, D.; Agris, P. F. The RNA Modification Database: 2011 Update. *Nucleic Acids Res* **2011**, 39 (Suppl 1), D195–D201.
- (6) Boccaletto, P.; Machnicka, M. A.; Purta, E.; Piatkowski, P.; Baginski, B.; Wirecki, T. K.; de Crécy-Lagard, V.; Ross, R.; Limbach, P. A.; Kotter, A.; Helm, M.; Bujnicki, J. M. MODOMICS: A Database of RNA Modification Pathways. 2017 Update. *Nucleic Acids Res.* **2018**, 46 (D1), D303–D307.
- (7) Roundtree, I. A.; Evans, M. E.; Pan, T.; He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **2017**, 169 (7), 1187–1200.
- (8) Suzuki, T.; Nagao, A.; Suzuki, T. Human Mitochondrial tRNAs: Biogenesis, Funct., Struct. Aspects, and Diseases. *Annu. Rev. Genet.* **2011**, 45 (1), 299–329.
- (9) Brandmayr, C.; Wagner, M.; Brückl, T.; Globisch, D.; Pearson, D.; Kneuttlinger, A. C.; Reiter, V.; Hienzsch, A.; Koch, S.; Thoma, I.; Thumbs, P.; Michalakis, S.; Müller, M.; Biel, M.; Carell, T. Isotope-Based Analysis of Modified tRNA Nucleosides Correlates Modification Density with Translational Efficiency. *Angew. Chem. Int. Ed.* **2012**, 51 (44), 11162–11165.
- (10) Rose, R. E.; Quinn, R.; Sayre, J. L.; Fabris, D. Profiling Ribonucleotide Modifications at Full-Transcriptome Level by Electrospray Ionization Mass Spectrometry. *RNA* **2015**, 21, 1361–1374.
- (11) Rose, R. E.; Pazos, M. A.; Curcio, M. J.; Fabris, D. Global Profiling of RNA Post-Transcriptional Mods as an Effective Tool for Investigating the Epitranscriptomics of Stress Response. *Mol. Cell. Prot. MCP* **2016**, 15 (3), 932–944.
- (12) McIntyre, W.; Netzband, R.; Bonenfant, G.; Biegel, J. M.; Miller, C.; Fuchs, G.; Henderson, E.; Arra, M.; Canki, M.; Fabris, D.; Payer, C. T. Positive-Sense RNA Viruses Reveal the Complexity and Dynamics of the Cellular and Viral Epitranscriptomes during Infection. *Nucl Acids Res* **2018**, 46 (11), 5776–5791.
- (13) Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; Sorek, R.; Rechavi, G. Topology of the Human and Mouse M6A RNA Methylomes Revealed by M6A-Seq. *Nature* **2012**, 485 (7397), 201–206.
- (14) Motorin, Y.; Muller, S.; Behm-Ansmant, I.; Branlant, C. Identification of Modified Residues in RNAs by Reverse Transcription-Based Methods. *Methods Enzymol.* **2007**, 425, 21–53.
- (15) Trixl, L.; Rieder, D.; Amort, T.; Lusser, A. Bisulfite Sequencing of RNA for Transcriptome-Wide Detection of 5-Methylcytosine. In *Epitranscriptomics: Methods and Protocols*; Wajapeyee, N., Gupta, R., Eds.; Methods in Molecular Biology; Springer: New York, NY, 2019, 1–21.
- (16) Crain, P. F. Mass Spectrometric Techniques in Nucleic Acid Research. *Mass Spectrom Rev* **1990**, 9, 505–554.
- (17) Limbach, P. A.; Paulines, M. J. Going Global: The New Era of Mapping Modifications in RNA. *Wiley Interdiscip. Rev. RNA* **2017**, 8 (1), 1–17.
- (18) Jora, M.; Borland, K.; Abernathy, S.; Zhao, R.; Kelley, M.; Kellner, S.; Addepalli, B.; Limbach, P. A. Amination/Imination of Carbonothiolated Nucleosides During RNA Hyd.s. *Angew. Chem. Int. Ed.* **2021**, 60 (8), 3961–3966.
- (19) Bay, S. D. Combining Nearest Neighbor Classifiers Through Multiple Feature Subsets. In *ICML*; 1998, 1–8.
- (20) Islam, M. J.; Wu, Q. M. J.; Ahmadi, M.; Sid-Ahmed, M. Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers. *JCIT* **2010**, 5, 133–137.
- (21) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, 29 (5), 1189–1232.
- (22) Currie, A.; Honish, L.; Cutler, J.; Locas, A.; Lavoie, M.-C.; Gaulin, C.; Galanis, E.; Tschetter, L.; Chui, L.; Taylor, M.; Jamieson, F.; Gilmour, M.; Ng, C.; Mutti, S.; Mah, V.; Hamel, M.; Martinez, A.; Buenaventura, E.; Hoang, L.; Pacagnella, A.; Ramsay, D.; Bekal, S.; Coetzee, K.; Berry, C.; Farber, J.; Team, O. B. O. T. N. I. Outbreak of Escherichia Coli O157:H7 Infections Linked to Mechanically Tenderized Beef and the Largest Beef Recall in Canada, 2012. *J. Food Prot.* **2019**, 82 (9), 1532–1538.
- (23) Watanabe, K.; Hayashi, N.; Oyama, A.; Nishikawa, K.; Ueda, T.; Miura, K. Unusual Anticodon Loop Structure Found in E.Coli Lysine tRNA. *Nucleic Acids Res.* **1994**, 22 (1), 79–87.
- (24) Joseph, A.; Cointe, A.; Mariani Kurkdjian, P.; Rafat, C.; Hertig, A. Shiga Toxin-Ass.d Hemolytic Uremic Syndrome: A Rev. *Toxins* **2020**, 12 (2), 1–45.
- (25) Brachmann, C. B.; Davies, A.; Cost, G. J.; Caputo, E.; Li, J.; Hieter, P.; Boeke, J. D. Designer Deletion Strains Derived from Saccharomyces Cerevisiae S288C: A Useful Set of Strains and Plasmids for PCR-Mediated Gene Disruption and Other Applications. *Yeast* **1998**, 14 (2), 115–132.
- (26) Kolitz, S. E.; Lorsch, J. R. Eukaryotic Initiator tRNA: Finely Tuned and Ready for Action. *FEBS Lett.* **2010**, 584 (2), 396–404.
- (27) Boa, S.; Coert, C.; Patterson, H.-G. Saccharomyces Cerevisiae Set1p Is a Methyltransferase Specific for Lysine 4 of Histone H3 and Is Required for Efficient Gene Expression. *Yeast Chichester Engl.* **2003**, 20 (9), 827–835.
- (28) Graham, F. L.; Smiley, J.; Russell, W. C.; Nairn, R. Characteristics of a Human Cell Line Transformed by DNA from Human Adenovirus Type 5. *J. Gen. Virol.* **1977**, 36 (1), 59–72.
- (29) Landry, J. J. M.; Pyl, P. T.; Rausch, T.; Zichner, T.; Tekkedil, M. M.; Stütz, A. M.; Jauch, A.; Aiyar, R. S.; Pau, G.; Delhomme, N.; Gagneur, J.; Korbel, J. O.; Huber, W.; Steinmetz, L. M. The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3 Genes Genomes Genet.* **2013**, 3 (8), 1213–1224.
- (30) Torsvik, A.; Stieber, D.; Enger, P. Ø.; Golebiewska, A.; Molven, A.; Svendsen, A.; Westermark, B.; Niclou, S. P.; Olsen, T. K.; Enger, M. C.; Bjerkvig, R. U-251 Revisited: Genetic Drift and Phenotypic Consequences of Long-Term Cultures of Glioblastoma Cells. *Cancer Med.* **2014**, 3 (4), 812–824.
- (31) Spang, A.; Saw, J. H.; Jørgensen, S. L.; Zaremba-Niedzwiedzka, K.; Martijn, J.; Lind, A. E.; van Eijk, R.; Schleper, C.; Guy, L.; Ettema, T. J. G. Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes. *Nature* **2015**, 521 (7551), 173–179.
- (32) Eme, L.; Spang, A.; Lombard, J.; Stairs, C. W.; Ettema, T. J. G. Archaea and the Origin of Eukaryotes. *Nat. Rev. Microbiol.* **2017**, 15 (12), 711–723.
- (33) Wu, D.; Hartman, A.; Ward, N.; Eisen, J. A. An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP). *PLoS One* **2008**, 3 (7), e2566, 1–10.

