



OPEN Exploration of linear and interpretable models for quantification of cell parameters via contactless short-wave infrared hyperspectral sensing

Anjana Hevaganinge¹, Eva Lowenstein¹, Anna Filatova¹, Mihir Modak¹, Nandi Thales Mogo¹, Bryana Rowley¹, Jenny Yarmowsky¹, Joshua Ehizibolo¹, Ravidu Hevaganinge¹, Amy Musser¹, Abbey Kim¹, Anthony Neri¹, Jessica Conway¹, Yiding Yuan¹, Maurizio Cattaneo^{1,2}, Sui Seng Tee³ & Yang Tao^{1✉}

The development of optical sensors for label-free quantification of cell parameters has numerous uses in the biomedical arena. However, using current optical probes requires the laborious collection of sufficiently large datasets that can be used to calibrate optical probe signals to true metabolite concentrations. Further, most practitioners find it difficult to confidently adapt black box chemometric models that are difficult to troubleshoot in high-stakes applications such as biopharmaceutical manufacturing. Replacing optical probes with contactless short-wave infrared (SWIR) hyperspectral cameras allows efficient collection of thousands of absorption signals in a handful of images. This high repetition allows for effective denoising of each spectrum, so interpretable linear models can quantify metabolites. To illustrate, an interpretable linear model called L-SLR is trained using small datasets obtained with a SWIR HSI camera to quantify fructose, viable cell density (VCD), glucose, and lactate. The performance of this model is also compared to other existing linear models, namely Partial Least Squares (PLS) and Non-negative Matrix Factorization (NMF). Using only 50% of the dataset for training, reasonable test performance of mean absolute error (MAE) and correlations (r^2) are achieved for glucose ($r^2 = 0.88$, MAE = 37 mg/dL), lactate ($r^2 = 0.93$, MAE = 15.08 mg/dL), and VCD ($r^2 = 0.81$, MAE = 8.6×10^5 cells/mL). Further, these models are also able to handle quantification of a metabolite like fructose in the presence of high background concentration of similar metabolite with almost identical chemical interactions in water like glucose. The model achieves reasonable quantification performance for large fructose level (100–1000 mg/dL) quantification ($r^2 = 0.92$, MAE = 25.1 mg/dL) and small fructose level (<60 mg/dL) concentrations ($r^2 = 0.85$, MAE = 4.97 mg/dL) in complex media like Fetal Bovine Serum (FBS). Finally, the model provides sparse interpretable weight matrices that hint at the underlying solution changes that correlate to each cell parameter prediction.

Keywords Contactless bio-sensor, Short wave infrared (SWIR), Near infrared (NIR), Machine learning

There exist several biomedical applications for which a contactless sensor for real-time monitoring of metabolites would be invaluable^{1–8}. These applications range from cell cultivation inside bioreactors for therapeutics manufacture^{1,4,6–8} to minimally invasive disease diagnosis^{2,3}. For example, The association of fructose with Non-Alcoholic Fatty Liver Disease (NAFLD) is well established epidemiologically^{9–12}, yet technologies available to measure fructolysis remain limited. The most common methods to study cellular metabolism are mass spectrometry (MS) and nuclear magnetic resonance (NMR)^{13,14}. These methods can be destructive, bulky, and expensive and do not allow repeated measurements on the same sample. The development of inline optical sensors coupled with machine learning analysis of signals to yield metabolite concentrations within bioreactors is prevalent in the literature. Examples include NIR probes^{6,8,15–20}, Raman probes^{4,16,21–25} and even NMR

¹Fischell Department of Bioengineering, University of Maryland, College Park, USA. ²Applied Imaging Solutions, LLC, Quincy, MA, USA. ³Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, USA. ✉email: ytao@umd.edu

spectrometers^{26,27}. However, these probes are prone to noise, fouling since they must be submerged within the cell sample, and sensor drift²⁸.

As an alternative, the use of a short wave infrared hyperspectral camera (SWIR - HSI) allows for fully contactless collection of thousands of absorption spectra using a single HSI image. This wealth of signals can then be used to efficiently denoise sample spectra²⁸. It can also train interpretable deep learning models using a small dataset of ground truth metabolite samples²⁸. These properties enable instantaneous troubleshooting of model predictions or automatic filtering of drifting readouts. This HSI sensor setup can also produce sample measurements using less than 500 μL of liquid per sample measurement²⁸. A truly multiplex sensor must be capable of distinguishing and quantifying similar metabolites since several biological metabolites, such as fructose and glucose, share similar chemical structures. Most studies in the literature attempt to quantify distinct metabolite relevant to bioprocessing using probes and traditional chemometric models like partial least square (PLS)^{4,6,7}. These models tightly couple dimensionality reduction with regression analysis, so it is difficult to determine which features are both noise free and relevant to metabolite prediction²⁹. Thus, an alternative example of interpretable linear models is explored called L-SLR. The performance of this model is compared with both PLS and well known unmixing model Non-negative Matrix Factorization (NMF).

This paper addresses the following objectives (Fig. 1): (1) enable contactless, label free rapid quantification of fructose and other cell growth parameters using a short-wave infrared (SWIR) hyperspectral camera; (2) develop and demonstrate models that easily adapts to varying cell media environments, and (3) develop model that allows practitioner to troubleshoot model performance based on sparse weight matrix maps. This leads to the development of small but interpretable models that can be extended to quantify any number of metabolites important to human health and cell growth alike.

Materials and methods

Materials

Spent Cell media samples were collected and imaged as described by Hevaganing et al.²⁸. In order to collect the fructose datasets, various weights of Fructose powder (Sigma F0127) were dissolved in solutions of Glucose (Sigma 150023-021) and deionized water to produce calibration datasets for fructose. Fructose was also dissolved in Fetal Bovine Serum (FBS) (Gibco 2309286) spiked with Glucose to generate further calibration curves. Sugar samples were weighed using a precise scale (Mettler-Toledo, ME204). Calibration curves were collected for fructose at varying glucose background levels, as shown in Table 1. The ID indicates the dataset number, while G denotes background glucose concentration in mg/dL,

F_{min} , F_{max} denote fructose concentration range in mg/dL, Δ_F denotes fructose concentration increment in mg/dL, and N is the number of samples collected in each dataset. T represents test data, and V represents validation data in the Fold column of Table 1. Lines demarcate combinations of data used to evaluate each

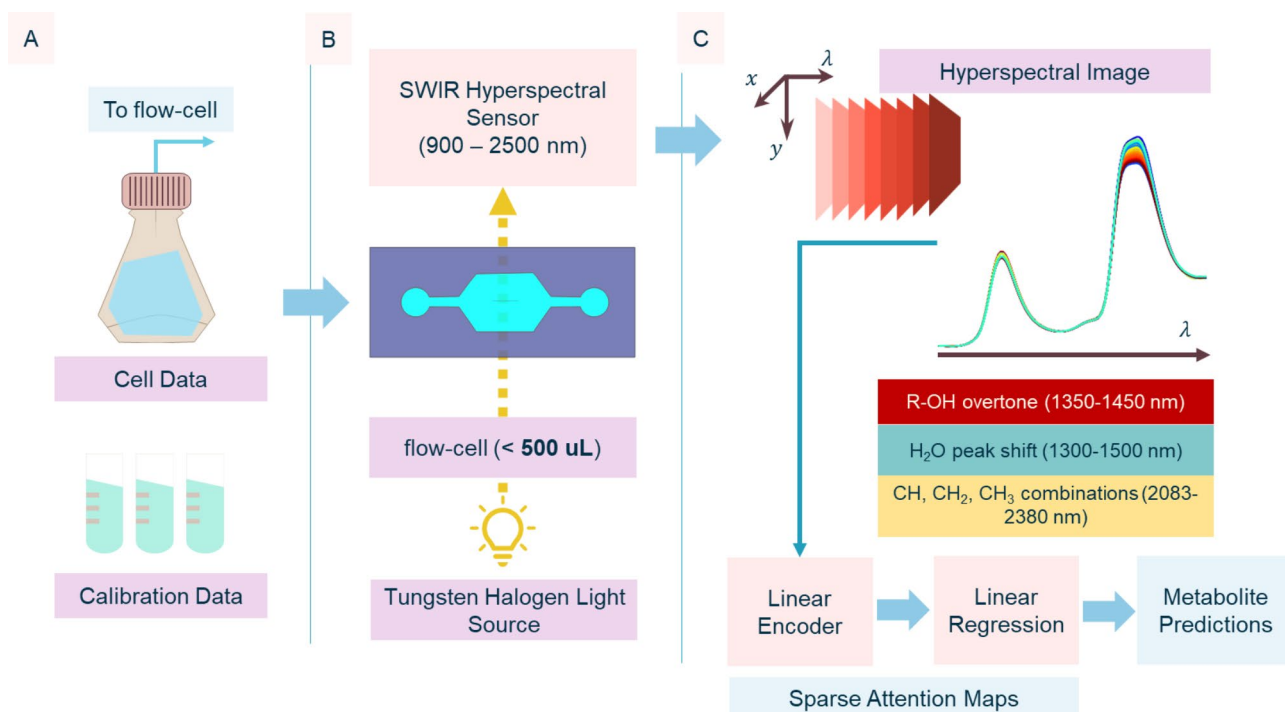


Fig. 1. An overview of the contactless, label free rapid quantification system using a short-wave infrared (SWIR) hyperspectral camera. (A) Collection of dataset of hyperspectral images and corresponding labels. (B) Collection of hyperspectral images in SWIR range for (C) training an encoder and sparse linear regression model for metabolite prediction.

ID	G	F_{min}	F_{max}	Δ_F	Bkg	N	Fold
1	0	0	0	0	Water	1	T
2	25	100	500	50	Water	16	T
3	250	100	500	50	Water	16	V
4	500	100	500	50	Water	16	T
5	78	6	60	6	Water	9	T
6	95	6	60	6	Water	9	V
7	112	6	60	6	Water	9	T
8	112	0	60	0.4	FBS	150	T
9	50	0	1000	100	FBS	10	T
10	500	0	1000	100	FBS	10	T
11	1000	0	1000	100	FBS	10	V

Table 1. Summary of experimental conditions used to collect fructose datasets.

model. Because fructose dataset was collected with varying background levels of glucose in both water and FBS (Table 1), it was possible to isolate a single glucose level from water and FBS environments respectively for k-fold cross validation and optimization of model components, while the remaining data was divided into k folds and treated as test data.

Fructose model training and analysis

Because fructose dataset was collected with varying background levels of glucose in both water and FBS (Table 1), it was possible to isolate a single glucose level from water and FBS environments respectively for k-fold cross validation and optimization of model components, while the remaining data were treated as test data. Both validation and test data from each respective glucose background were split into k-folds consisting of 2 samples to obtain final metrics. 4 separate model sets were trained by partitioning the data as shown in Table 1, with model validation datasets bolded and all remaining datasets used as test data. Datasets used as validation data are bolded.

Fructose dataset collection

To create the water solutions, glucose and fructose were weighed using a balance (Mettler-Toledo, ME204) and placed in 15mL conical tubes. Next, the appropriate volume of water was added to reach the desired concentration. The tubes were then thoroughly mixed. Serum solutions were created with heat-inactivated Fetal Bovine Serum (FBS) (Gibco, Lot 2309286) using the same procedure as for water solutions. Since glucose is present in horse serum at unknown concentrations, the resulting solutions contained concentrations of glucose augmented by both fructose and additional glucose. Sample order was randomized before image acquisition. 500 μ L of each sample was loaded into a quartz cuvette (Hellma 110-1-40) and imaged using Hyperspectral sensor (Headwall Photonics, Bolton, Massachusetts). The spatial resolution of this system was approximately 0.5 mm, which allowed collection of hundreds of transmittance readings from each cuvette photo. Each image pixel contains a transmittance spectrum which is collected in the range from 900 to 2500 nm in 8 nm increments. The quartz cuvette was washed with a single wash of water and two washes of solution of interest prior to image collection.

Cell data model training and analysis

As detailed previously²⁸, cell permeate dataset was collected and quantified from eight different cell culture flasks, such that approximately 30 samples were collected from each flask prior to data cleaning and filtering. Further, ground truth labels for cell media samples were kept unfiltered. Only data with cell viability greater than 80% and lactate levels less than 300 mg/dL were used for model training from spent cell media dataset, resulting in approximately 20 samples per flask. The first flask was chosen as the validation flask and used to determine minimum number of flasks that could be used for model training while maintaining reasonable metabolite quantification results, as well as optimal number of components for all models. Of the remaining flasks, 4 were randomly chosen as test flasks, and remaining flasks were used as training data. 20 folds of flask combinations were generated for evaluation. To minimize model extrapolation, only data with cell viability greater than 80% was used as test data for models trained using spent cell media dataset.

Data pre-processing

Hyperspectral images were processed as described by Hevaganinge et al.²⁸ for all datasets. First, min-max normalization of each transmission signal was conducted against respective dark and white background reference hyperspectral images that are taken at the start of each experiment. Next, transmittance spectra are converted to polynomial basis³⁰ based on Eq. (1), where $\phi \in \mathbb{R}^{N \times P}$, with N samples and P terms, represents set of linear and 2nd order cross terms and x_i, x_j represent transmittance values from i^{th} and j^{th} wavelength respectively. Also, B is the total number of wavelengths. All terms undergo min-max normalization based on extreme features of respective training dataset.

$$\phi = x_1 + x_2 + x_i + \dots + x_B + x_1x_2 + \dots + x_ix_j + \dots + x_{B-1}x_B \quad (1)$$

Data analysis and model training

Three models were explored for contactless quantification of various cell parameters. The respective validation dataset was used for model hyperparameter tuning, and validation data was omitted from test data evaluation as needed. Scikit-Learn implementation of Partial Least Squares with 10 components^{29,30} and polynomial features (ϕ) was used as a baseline model to compare against the performance of two source separation models. The first alternative model used a blind source separation model, specifically Scikit-Learn implementation of Non-negative Matrix Factorization (NMF)^{30,31} as the encoder which takes ϕ as input and returns compressed signal with 10 components for subsequent linear regression. Nonnegative Double Singular Value Decomposition (NDSVD) was used as the initialization method, and the solver was Multiplicative Update Solver. The architecture of the proposed L-SLR model is shown in Fig. 2. First, the transmittance signal, $T \in \mathbb{R}^{N \times B}$, with N samples and B wavelengths, was converted to a polynomial basis (2 A). Concurrently, the matrix of metabolite concentrations, $Y \in \mathbb{R}^{N \times M}$ from the training dataset with M metabolites was converted into a single integer index via Hilbert transformation³² (2B). Specifically, each metabolite label was divided by maximum concentration in training dataset, multiplied by 100 and rounded to the nearest integer. Each resulting number was treated as an index in M degree space which can be spanned using a Hilbert curve using the Hilbert curve package in python³³. These integer labels were used to pretrain an Scikit-Learn implementation of Orthogonal Matching Pursuit sparse regression model (OMP)^{30,34} (2 C) which output a sparse feature mask that was used to mask out noisy terms in original feature space that were not relevant to metabolite prediction, as shown in Eq. (2), where H_i is the sparse cross terms feature vector of sample i from ϕ after masking, and s is the sparse feature mask.

$$H_i = \phi_i[s] \quad (2)$$

Next, these integer labels and corresponding masked features were used to pre-train a Linear Discriminant Analysis (LDA) encoder³⁵ (2E). The LDA model forms a compressed latent space which helps to separate very redundant metabolite spectra from a variety of metabolite environments. Test signals were masked using the OMP method (2 C) then dimensionality reduction was done by pretrained LDA according to Eq. (3), where Z is the compressed signal, $H_i \in \mathbb{R}^{N \times P^+}$, with N samples and P^+ terms, is the masked cross terms signal, and W is the basis matrix of $n = 10$ highest eigenvectors obtained via optimization of LDA cost function (2E).

$$Z = HW \quad (3)$$

After, simple linear regression³⁰ was used to quantify metabolite levels using compressed signal from LDA model, according to Eq. (4) (2 F) where \hat{y} is the predicted metabolite reading, Θ is the sparse weights, and z is the LDA latent space signal. All ground truth labels are normalized by corresponding maximum training label prior to fitting linear regression model.

$$\hat{y} = \Theta z \quad (4)$$

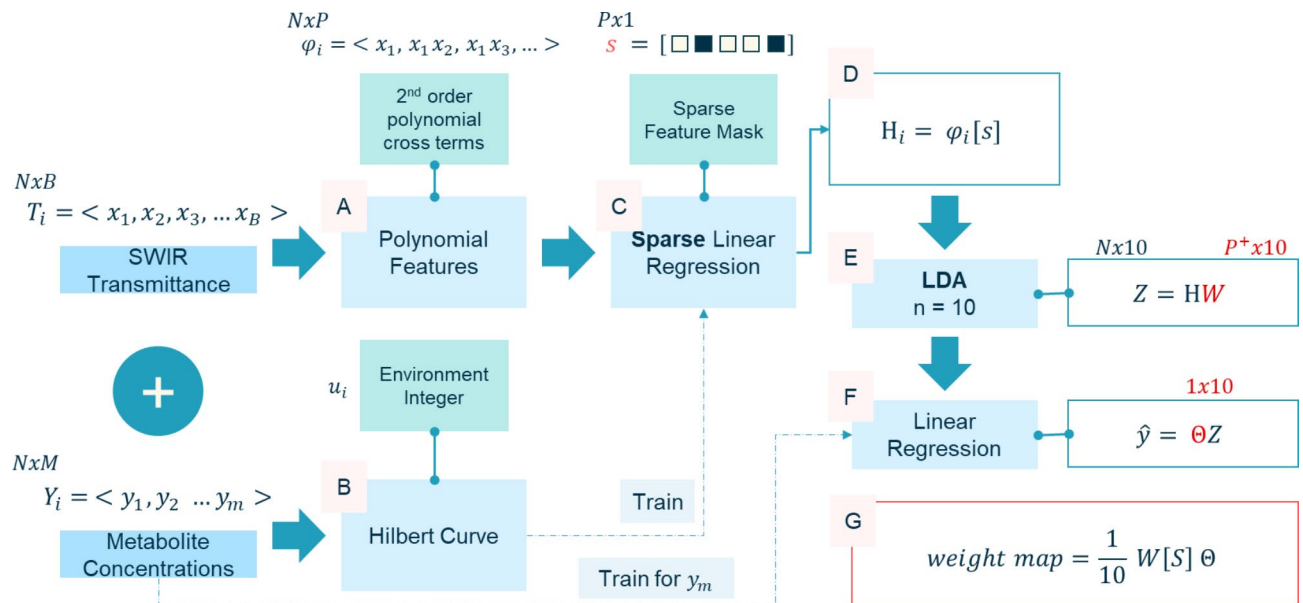


Fig. 2. Architecture of L-SLR. First, absorbance signal is converted to a polynomial basis (A), then masked based on weights from pretrained sparse linear regression model (C). Next, masked signal is passed into pretrained LDA model (E). LDA forms compressed latent space which is used to generate metabolite predictions via linear regression (F). Coefficients from LDA and SLR are then used to form weight matrix visuals (G). Models used in (C) and (E) are pretrained using a labeled dataset consisting of integer labels computed via Hilbert curve (B) that represent unique metabolite combinations.

	Validation in dataset 11				Testing in datasets 9, 10			
	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2
L-SLR (<i>n</i> = 10)	34.38	0.08	53.4	0.96	28.37	0.13	33.69	0.98
NMF Regression (<i>n</i> = 10)	94.05	0.24	116.01	0.8	76.7	0.19	109.62	0.82
PLS-R (<i>n</i> = 10)	36.31	0.11	47.03	0.97	19.43	0.11	28.94	0.99

Table 2. Quantification of large fructose concentrations in horse serum and glucose solutions.

	Validation in dataset 3				Testing in datasets 2, 4			
	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2
L-SLR (<i>n</i> = 10)	29.74	0.14	37.44	0.86	25.1	0.16	32.75	0.92
NMF Regression (<i>n</i> = 10)	47.6	0.21	53.6	0.71	78.25	0.52	108.51	0.1
PLS-R (<i>n</i> = 10)	25.62	0.12	33.1	0.89	19.7	0.12	24.81	0.95

Table 3. Quantification of large fructose concentrations in glucose solutions.

	Validation in dataset 6				Testing in dataset 5, 7				Testing in dataset 8			
	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2
L-SLR (<i>n</i> = 10)	4.78	0.21	5.85	0.82	5.84	0.34	6.79	0.81	4.97	0.38	6.88	0.85
NMF Regression (<i>n</i> = 10)	21.46	0.73	25.42	0	12.05	0.56	19.75	0	5.55	0.39	7.66	0.81
PLS-R (<i>n</i> = 10)	2.46	0.11	3.29	0.94	7.35	0.46	8.07	0.73	4.17	0.28	6.28	0.87

Table 4. Quantification of small fructose concentrations in glucose solutions and horse serum.

Finally, the sparse weight visualizations were formed by taking the weighted average of LDA weights (W) based on linear regression weights (Θ) before replacing nonzero elements of sparse feature mask s_i with the corresponding weight value. All terms used to construct weight map are highlighted in red (Fig. 2). These steps are summarized in Eq. (5), where s_p is the p^{th} term in the sparse feature mask, w_p is the p^{th} column of LDA eigenvector matrix, and Θ is the weight matrix used for metabolite quantification.

$$s_p = \frac{1}{n} w_p \Theta \tag{5}$$

L-SLR weight matrix map generation

Because L-SLR is a fully linear model, the weight matrix may be visualized and analyzed similar to an attention map. The weight matrix not only reveals which features are excluded from metabolite prediction³⁶, but also reveals which features are combined to generate final metabolite predictions. Therefore, the cumulative sum of all weight matrices generated across folds for each dataset was used as an approximation of the stable weight matrix.

Model evaluation and statistics

Predicted cell parameter concentrations were compared to ground truth readings via r^2 , root mean squared error (RMSE), mean absolute error (MAE), and relative mean absolute percentage error (MAPE) metrics. All metrics were calculated using Scikit-Learn implementations³⁰. MAE, and RMSE are reported in units of mg/dL, while the relative MAPE decimal is reported in all tables. Number of latent space components used for each model is listed as n , with 10 components used for all models.

Results and discussion

Three linear models were evaluated using a series of datasets collected in house using labeled HSI images of not only fructose solutions but also spent cell media from eight CHO cell culture flasks. Tables 2, 3 and 4 summarize validation and test performance when fructose samples in various glucose backgrounds are partitioned into k -folds each containing 2 unique fructose concentrations as test samples. In dataset number 8, 5 unique fructose concentrations are sampled for each test fold since it contains a higher number of unique fructose concentrations compared to the other fructose datasets collected. The dataset number corresponds to data which was collected under conditions listed in Table 1.

Figures 3, 4 visualize PLS-R and L-SLR correlation to fructose ground truth for test datasets in all datasets of interest for a variety of fructose solution in glucose solutions (Fig. 3) and Horse Serum solutions (Fig. 4). r^2 and MAE are listed in the top left corner of each graph. All measurements are reported in units of mg/dL.

Across all folds, PLS-R exhibits better test performance when compared to L-SLR, with better scoring performance across all metrics assessed. Nevertheless, the accuracies of both L-SLR and PLS fall within the range

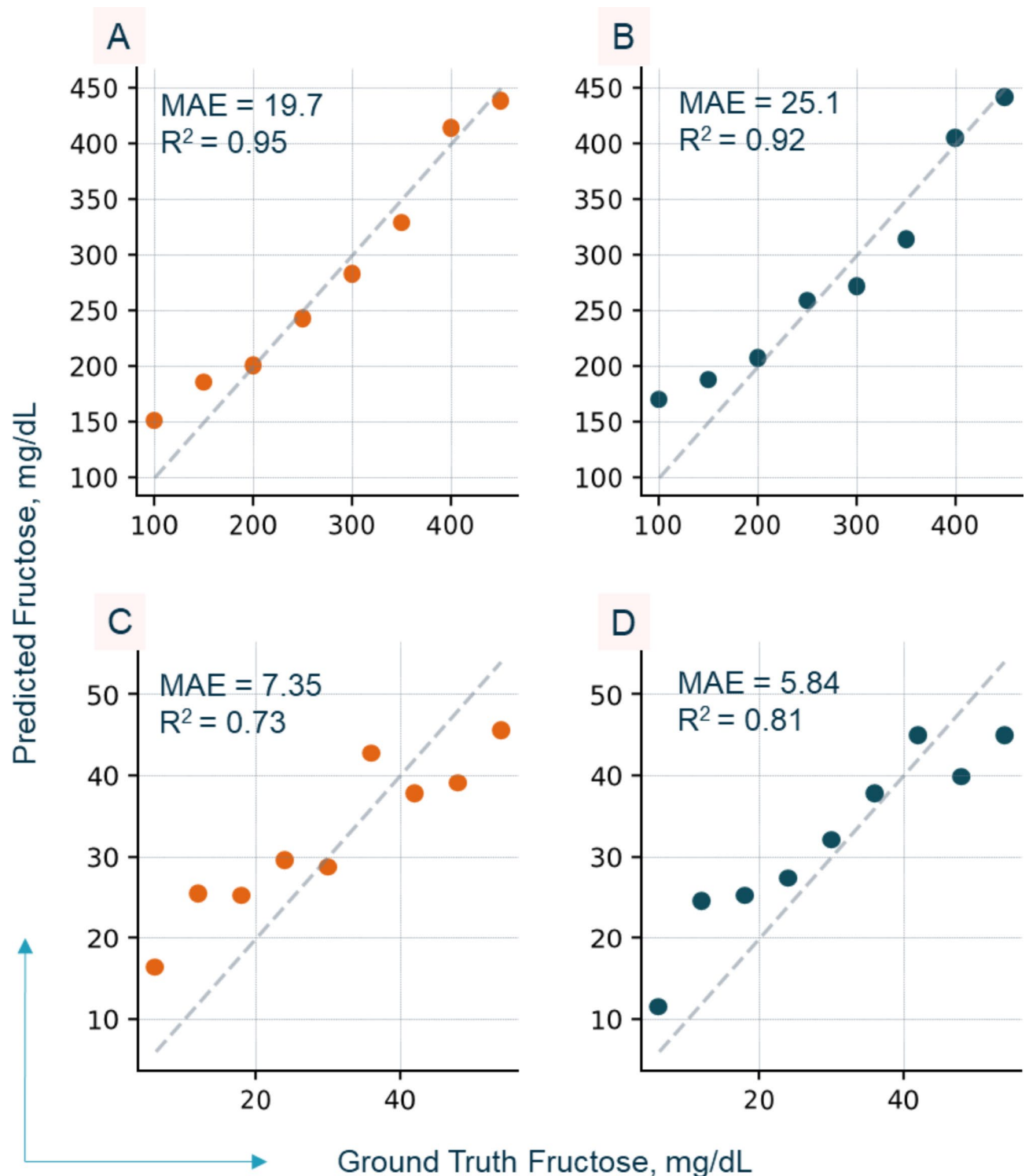


Fig. 3. Correlation Visualization between ground truth fructose and fructose predicted by PLS-R (A, C) and L-SLR (B, D) for test data collected in dataset conditions 2 and 4 (A, B) and conditions 5 and 7 (C, D). r^2 and MAE are listed in the top left corner of each graph.

of usability for biological studies of fructose^{2,12} and bioprocess monitoring^{16,23}. The performance gap between PLS-R and L-SLR widens as the fructose concentration decreases, suggesting that LDA has trouble defining discrete clusters for small fructose samples, especially in the presence of relatively high glucose backgrounds. It is also clear that folds sampled at the extreme concentrations tend to exhibit higher levels of drift compared to concentrations in the center of the fructose concentration data distribution. Perhaps more concentrations beyond the extremes of interest need to be sampled and placed as fixed training data points across all folds. In

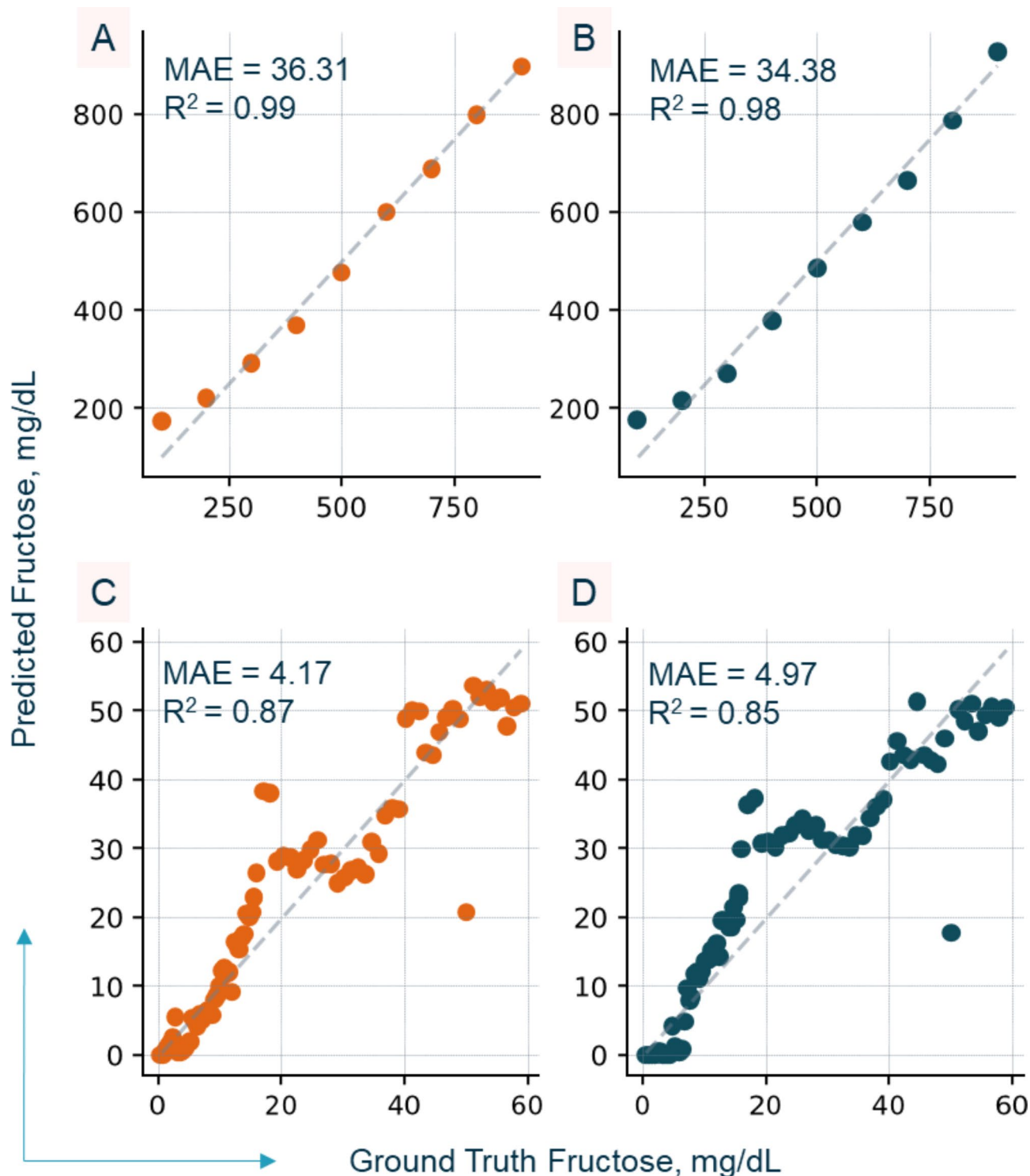


Fig. 4. Correlation Visualization between ground truth fructose and fructose predicted by PLS-R (A, C) and L-SLR (B, D) for test data collected in dataset conditions 9 and 10 (A, B) and condition 8 (C, D). r^2 and MAE are listed in the top left corner of each graph.

Fig. 4C and D, there are a few outlier points that have high test error for both PLS-R and L-SLR, while most of the other datapoints are quantified with relatively smaller error. The misclassification of these points may stem from the fact that background glucose concentration is more than twice the fructose concentrations imaged, so signal differences caused by fructose concentration changes are small, subtle, and difficult to differentiate properly. Further, the performance of unmixing regression model based on nonlinear matrix factorization (NMF) is shown in Tables 2, 3 and 4, since this is a well established model in the literature for analysis of spectroscopic

	Training				Validation in flask 1				Testing across 20 folds			
	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2
L-SLR ($n=10$)	7.06E+05	0.22	9.53E+05	0.87	1.03E+06	0.37	1.24E+06	0.84	8.60E+05	0.30	1.17E+06	0.81
NMF Regression ($n=10$)	7.72E+05	0.26	1.02E+06	0.85	1.22E+06	0.29	1.50E+06	0.76	9.88E+05	0.33	1.29E+06	0.76
PLS-R ($n=10$)	6.25E+05	0.20	8.26E+05	0.91	1.19E+06	0.32	1.42E+06	0.79	8.45E+05	0.27	1.14E+06	0.82

Table 5. Quantification of viable cell density (VCD) in spent cell media.

	Training				Validation in flask 1				Testing across 20 folds			
	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2
L-SLR ($n=10$)	28.55	0.22	37.26	0.93	41.10	0.24	58.85	0.87	36.73	0.26	49.57	0.89
NMF Regression ($n=10$)	42.98	0.40	53.76	0.86	67.34	0.37	78.96	0.76	57.17	0.43	79.48	0.70
PLS-R ($n=10$)	19.68	0.18	26.08	0.97	58.11	0.36	76.10	0.78	30.92	0.22	45.55	0.90

Table 6. Quantification of glucose in spent cell media.

	Training				Validation in flask 1				Testing across 20 folds			
	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2	MAE	MAPE	RMSE	R2
L-SLR ($n=10$)	12.83	0.09	15.89	0.95	19.69	0.18	23.00	0.94	15.08	0.11	18.89	0.93
NMF Regression ($n=10$)	16.25	0.13	20.63	0.92	20.74	0.23	24.08	0.93	20.58	0.16	26.07	0.86
PLS-R ($n=10$)	10.62	0.08	13.59	0.96	19.66	0.22	23.30	0.93	14.42	0.10	18.95	0.92

Table 7. Quantification of lactate in spent cell media.

signals^{37,38}. However, the performance of the NMF regression is worse when compared to PLS and L-SLR, most likely since NMF is an iterative model whose performance is dependent on initialization³¹. Next, to evaluate model performance on cell culture flask dataset quantification, all models were trained and evaluated on 20 folds of 4 randomly sampled flasks for training data, while using remaining cell culture flasks as test data. The validation flask was omitted from test data evaluation as needed. Tables 5, 6 and 7 summarize average viable cell density (VCD), glucose and lactate quantification performance across 20 cross validation folds when four cell culture flasks (50% of dataset) are randomly reserved for model training. This number was determined as the minimum number of flasks required to achieve comparable validation performance as using 75% (6 randomly sampled flasks) as training data and flask 1 as validation data. The use of 50% of the dataset rather than 75% of the datasets highlights the powerful generalizability and training efficiency of linear models.

Figure 5, 6, 7 visualize PLS-R and L-SLR correlation to VCD, glucose and lactate ground truth labels respectively for randomly selected test dataset folds. r^2 and MAE are listed in the top left corner of each graph. All measurements are reported in units of mg/dL.

Across all folds, L-SLR exhibits similar performance to PLS-R, although PLS-R tends to display higher performance across all metrics evaluated on the test datasets. Both models achieve best test performance with lactate, test r^2 values above 0.9, and MAE less than 20 mg/dL. This suggests that signals correlated to lactate are robust and bi-linear. Performance metrics are lowered slightly when evaluating glucose quantification, but r^2 values are still close to 0.9. After analyzing Fig. 6, it is clear that a few cell permeate images exhibit high glucose quantification error, while most of the points have relatively low error. It is important to note that SWIR spectra mainly reflect the absorbance of water, and the subtle shifts of the water peak are used as a “molecular mirror” of substrates dissolved in the hydrogen bonding matrix of water³⁹. Thus, signal shifts caused by glucose are not always linear, or even bi-linear⁴⁰. This may explain the spurious misclassification of certain glucose images by both models. Further, PLS-R exhibits significant overfitting when compared to L-SLR performance for glucose and lactate. This is highlighted by the relatively large reduction in all metrics between training and testing performance for PLS-R when compared to L-SLR. Finally, both models achieve r^2 values less than 0.9 for VCD. Since cell free permeate is imaged, perhaps hyperspectral images lack concrete signals that correlate to VCD. Most likely, the model is forced to correlate signals belonging to other cell parameters with VCD instead. It was hoped that linear unmixing methods like NMF with the help of 2nd degree cross term features could help uncover interpretable signals that directly correlate to VCD. In practice, NMF regression models display the worst performance of the models evaluated, most likely since NMF is an iterative model whose performance is dependent on initialization. Of course, L-SLR might be thought of as a unmixing model as well, since it does factor the input matrix X into eigenvectors and eigen values that best explain the variance correlated to ground truth label separation³⁵. In the future, perhaps the Fischer’s Discriminant ratio can be modified so that chosen components group similar environment samples close together.

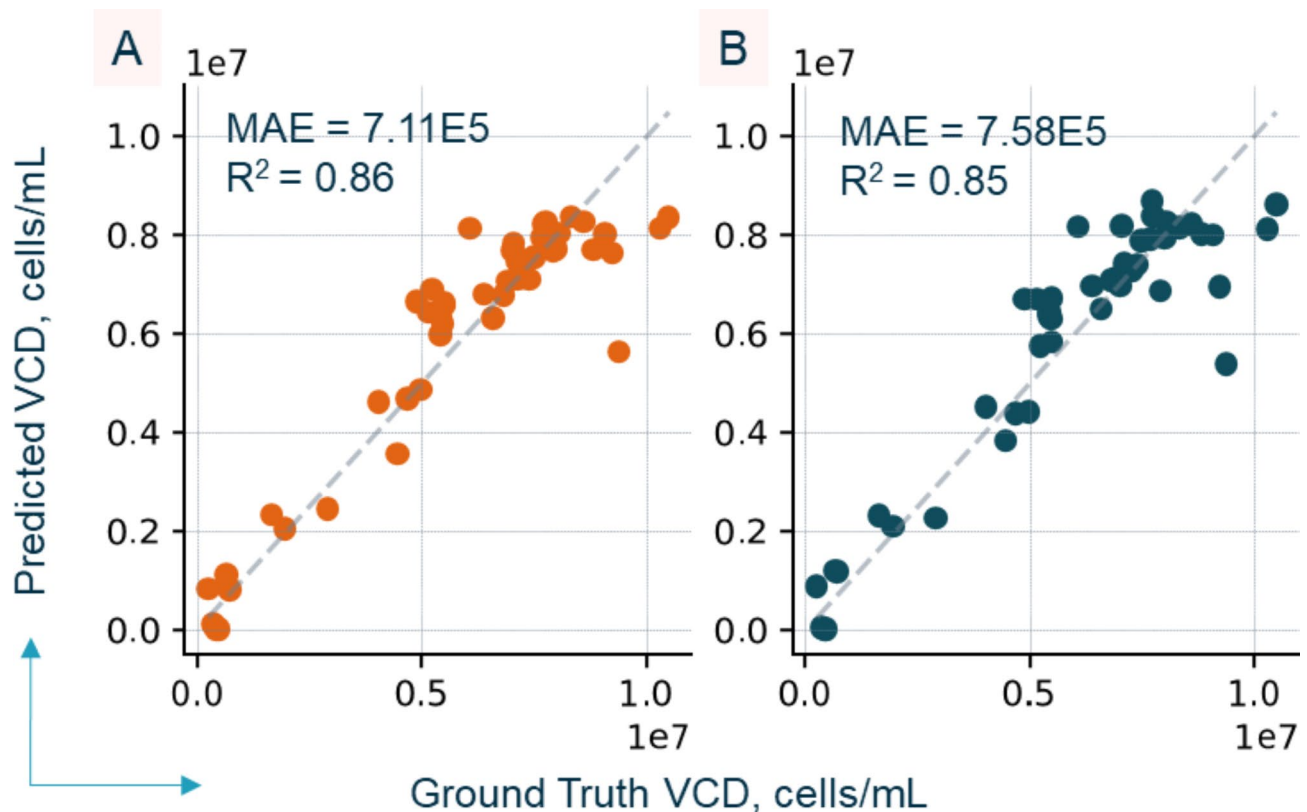


Fig. 5. Correlation visualization between ground truth VCD and VCD predicted by PLS-R (A) and L-SLR (B) respectively in spent cell media are shown for randomly selected test fold. r^2 and MAE are listed in the top left corner of each graph.

L-SLR interpretability

Unlike Mid-Infrared (MIR) absorption, SWIR absorption reflects weak signals caused by overtones and combination vibrations^{39,41–43}. These signals are difficult to simulate and interpret, but linear weight matrices help the practitioner correlate high weights with known functional group harmonics in the literature⁴¹. Linear models afford weight matrices which directly show both which features are omitted and which features are used as well as how these features contribute to the final metabolite predictions. In contrast, attention maps formed to summarize complex non linear models only afford a glimpse of which features are omitted for model prediction³⁶. Summary visualizations of all weight matrices produced across all folds evaluated are shown in Figs. 8 and 9.

It is well known that hydrogen bonding within the solution matrix impacts SWIR absorption spectra⁴¹. This phenomenon is reflected in the distinct differences between weight matrices for models trained to identify fructose within glucose solutions vs. horse serum solutions spiked with glucose (Fig. 8A–D). For example, the model trained using glucose solutions pays more “attention” to absorption signals below 1400 nm, whereas the model trained using horse serum pays more “attention” to absorption signals between 1600 and 1800 nm. The former region is dominated by overtones related to O–H stretch, whereas the latter region is dominated by C–H stretch vibrations⁴¹. This implies that the correlation between fructose concentration and SWIR absorption can be related to both changes in the hydrogen bonding matrix of water and interactions of fructose with organic compounds dissolved in the horse serum. Further only cross terms in region between bands less than 1400 and bands from 1600 to 1900 or region between 1600 and 1800 are heavily used for all models other than those used to quantify small quantities of fructose in FBS (Fig. 8B). According to Fig. 4, the corresponding model displays the largest extremes in model prediction error compared to the other models. Perhaps this distribution of weights suggests that the model was not able to clearly distinguish linear or bi-linear cross term signals that correlate with fructose concentrations, which leads to overfitting to noise in the training dataset instead. The misclassification of these points may stem from the fact that background glucose concentration is very high compared to the incremental fructose concentration changes in this dataset, so feature changes are small, subtle, and difficult to differentiate properly. A similar analysis is less clear in the PLS-R weight matrices, since non-zero weights are much less sparse compared to equivalent L-SLR model. However, points of distinction are highlighted in Fig. 8E–H.

Although three distinct cell culture parameters were predicted using the cell culture flask dataset, the weight profile of all three models is quite similar (Fig. 9A–C). In fact, the weight matrices formed by PLS-R (Fig. 9D–F) are virtually indistinguishable, so weight matrices formed by L-SLR will be analyzed for the remainder of the discussion. Notably, both the VCD model and the lactate model share similar regions of high weight, implying

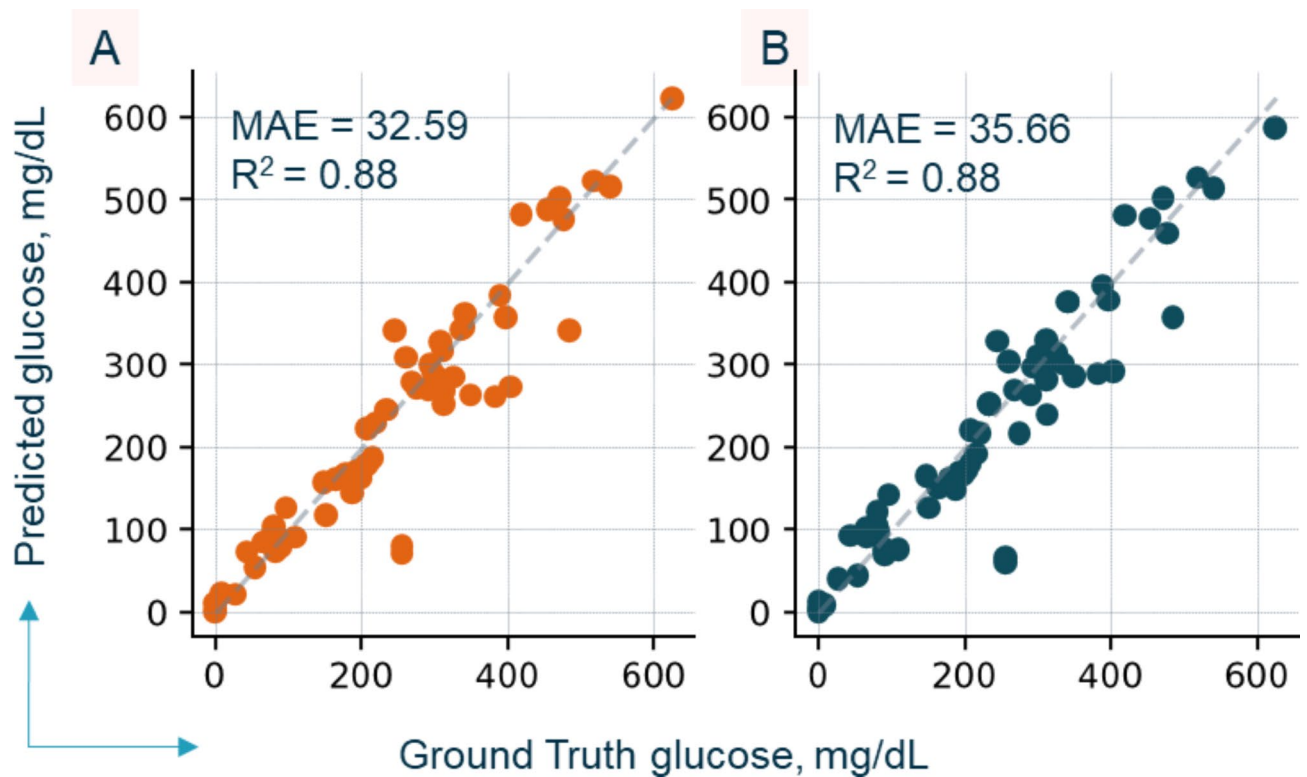


Fig. 6. Correlation visualization between ground truth glucose and glucose predicted by PLS-R (A) and L-SLR (B) respectively in spent cell media are shown for randomly selected test fold. r^2 and MAE are listed in the top left corner of each graph.

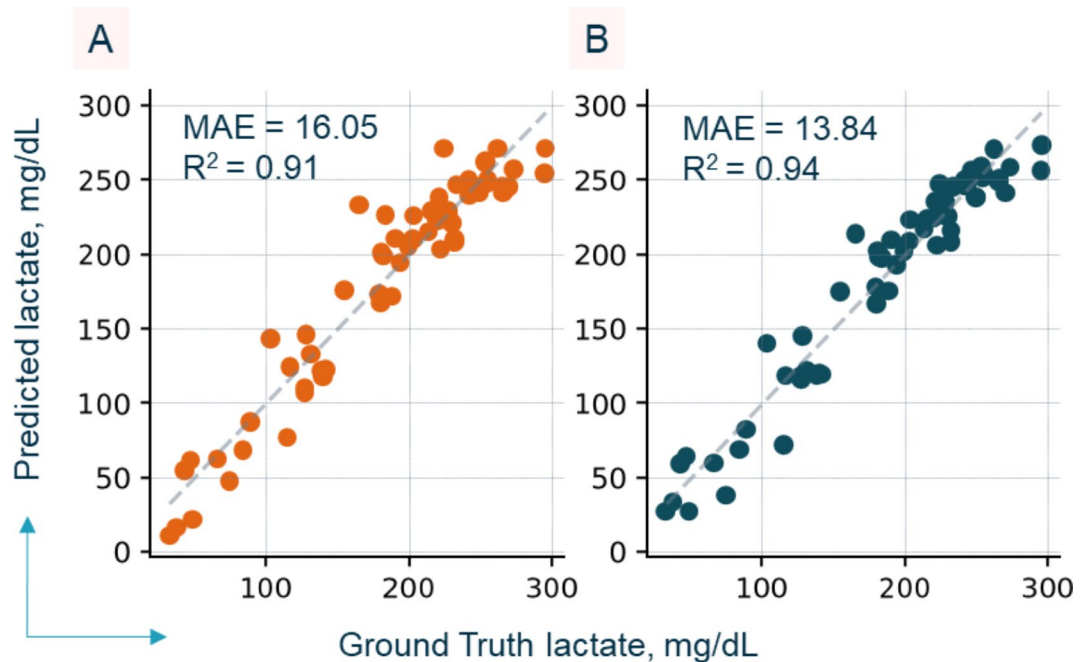


Fig. 7. Correlation visualization between ground truth lactate and lactate predicted by PLS-R (A) and L-SLR (B) respectively in spent cell media are shown for randomly selected test fold. r^2 and MAE are listed in the top left corner of each graph.

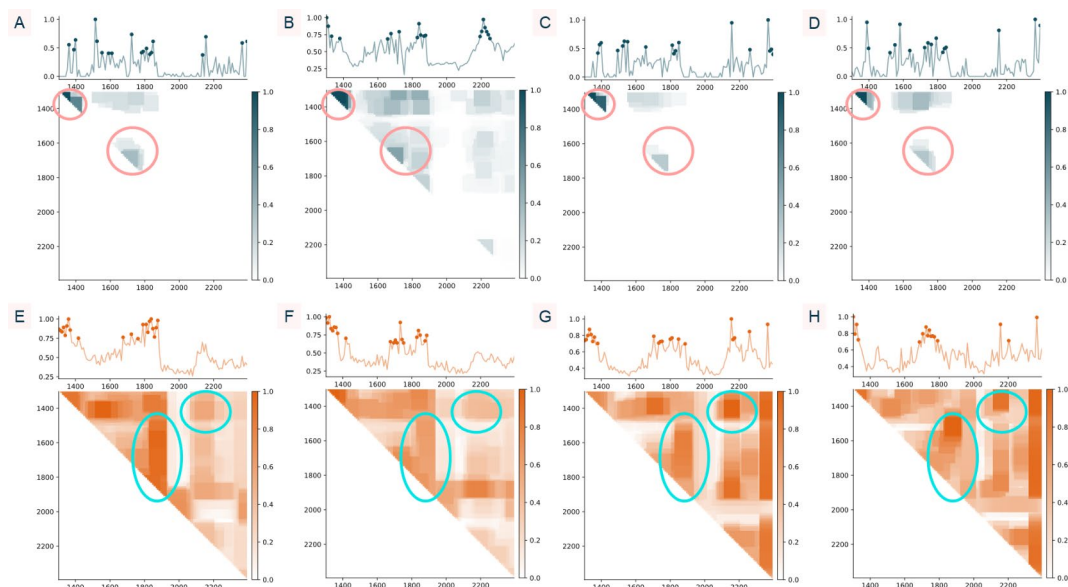


Fig. 8. Cumulative sum weight matrices produced by L-SLR (A–D) and PLS-R (E–H) for bands and cross terms across all models evaluated for fructose quantification via datasets 9–11 (A, E), dataset 8 (B, F), datasets 2–4 (C, G) and datasets 5–7 (D, H). All weights are normalized by the respective maximum of each cumulative matrix to facilitate comparison of weight matrices across models and datasets alike.

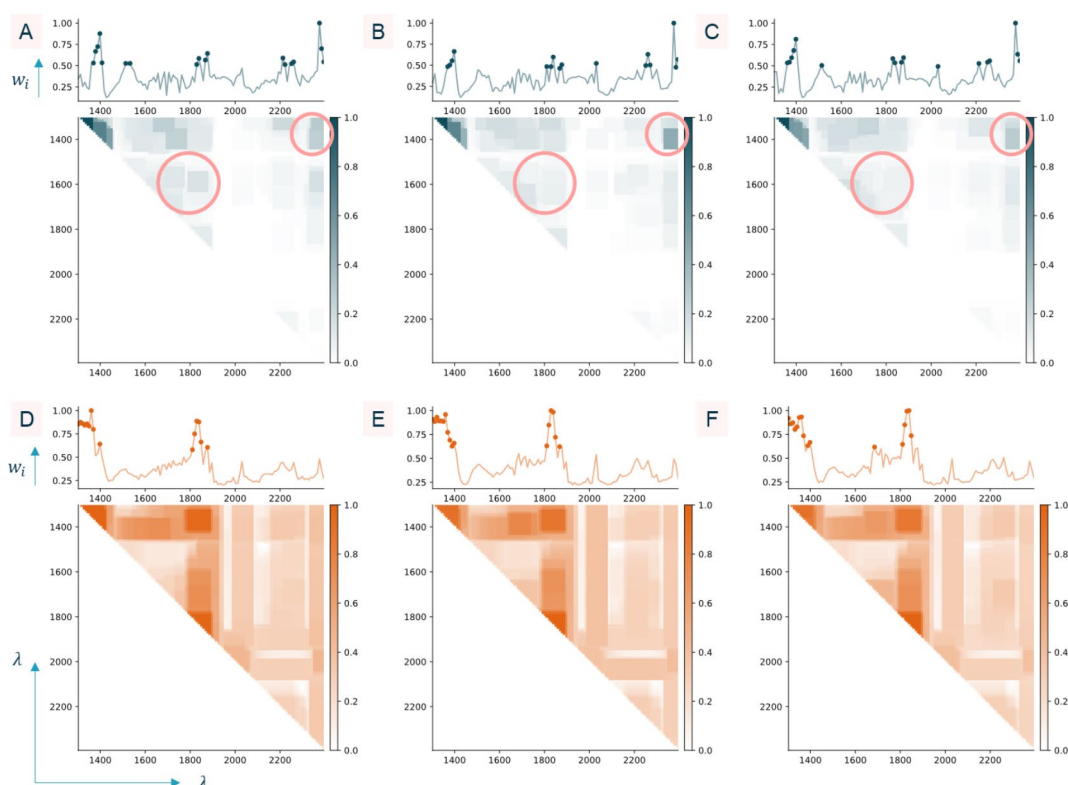


Fig. 9. Cumulative sum weight matrices produced by L-SLR (A–C) and PLS-R (D–F) for bands and cross terms across all models evaluated for quantification of VCD (A, D), glucose (B, E), and lactate (C, F) in spent cell media. All weights are normalized by the respective maximum of each cumulative matrix to facilitate comparison of weight matrices across models and datasets alike.

that changes in lactate concentration correlate with VCD changes. While both VCD and lactate models share a high weight band around 2200 nm, the glucose model does not weigh this band region quite as heavily. This region corresponds to overtone stretches caused by carboxylic acid⁴¹, which is the main functional group on lactic acid. Still, all three models pay “attention” to absorption signals below 1400, implying that changes to the hydrogen bonding matrix of water correlate with changes in all three cell parameters investigated. Still, SWIR spectra exhibit broad, overlapping peaks with non-linear relationships to dissolved molecules. Similar absorption spectra may correspond to very different solution matrix profiles, so the model confuses distinct cell parameter predictions. Because LDA is already designed to maximize separation of data points based on class³⁵, it was assumed that LDA would be able to treat different sets of metabolites as distinct classes and find an appropriate lower dimensional latent space that facilitates metabolite quantification. In the future, the cost function of LDA should be modified to distance points from different environments based on environment similarity. This smooth latent space would also allow for identification and removal of ambiguous absorption spectra which overlap multiple metabolic profiles. In contrast, it is difficult to imagine how PLS-R latent space may be modified to ensure continuous latent space, since dimensionality reduction and regression are highly coupled²⁹. Across all datasets tested, the NMF model performed the worst, most likely due to dependence on weight initialization. Unlike PLS-R and L-SLR, NMF is a blind source separation model, so it cannot take advantage of the supervised labeled dataset available. It is reasonable and arguably essential in the bioprocessing and therapeutics realm for labeled datasets to be collected prior to model training for metabolite quantification. Based on remarks in the literature, an interpretable chemometrics model in the bioprocessing and therapeutics realm should (1) produce reduced dimensionality latent space that separates spectra based on corresponding metabolite environment, (2) easily adapts to varying cell media environments, and (3) allows practitioner to troubleshoot model performance based on sparse weight matrix maps and the prediction of confidence intervals^{28,36}. So far, L-SLR has produced a reduced dimensionality latent space coupled with sparse weight matrices that highlight which bands and cross terms are used for final prediction. Both L-SLR and PLS demonstrate reasonable metabolite quantification accuracy, and the ability to generalize to test datasets using 50% of the dataset available. This implies that these models can generalize training data environments to test environments with the same cell type and media using a 1:1 ratio of train to test data. In the future, this sensor framework may be used to quantify other valuable cell parameters such as pH, amino acids, and even antibody glycosylation^{5,6,20,44}.

Conclusion

Herein, we presented the development of a contactless label free sensor coupled with an interpretable model for the quantification of fructose and other key cell parameters. This SWIR sensor system has succeeded in (1) contactless, label free rapid quantification of fructose and other cell growth parameters, and (2) easily adapts to varying cell media environments using (3) interpretable models that shed light on how the bands of the nonlinear absorption spectra contribute to metabolite quantification. Further, the accuracies of both L-SLR and PLS fall within the range of usability for biological studies of fructose^{2,12} and bioprocess monitoring^{16,22}. Because the HSI camera is capable of recording thousands of absorbance spectra within a single image, this sensor system can efficiently train and adapt to new metabolites using relatively small calibration datasets. Currently, existing invasive time consuming quantification methods like high performance liquid chromatography (HPLC) and mass spectrometry (MS) remain the gold standard measurement devices in biomedical arena. There are several biomedical applications for which real-time monitoring of cell parameter levels coupled with intermittent ground truth readings will enable both rapid diagnosis of diseases like NAFLD, and precise control of bioprocesses.

Data availability

Raw data were generated at University of Maryland, College Park. Derived data supporting the findings of this study are available from the corresponding author Dr. Yang Tao on request.

Received: 1 August 2024; Accepted: 7 January 2025

Published online: 17 January 2025

References

- Costariol, E. et al. Demonstrating the manufacture of human CAR-T cells in an automated stirred-tank bioreactor. *Biotechnol. J.* **15**, 2000177 (2020).
- Miller, M., Drucker, W. R., Owens, J. E., Craig, J. W. & Woodward, H. Metabolism of intravenous fructose and glucose in normal and diabetic subjects. *J. Clin. Invest.* **31**, 115–125 (1952).
- Kruszynska, Y. T., Meyer-Alber, A., Wollen, N. & McIntyre, N. Energy expenditure and substrate metabolism after oral fructose in patients with cirrhosis. *J. Hepatol.* **19**, 241–251 (1993).
- Schwarz, H., Mäkinen, M. E., Castan, A. & Chotteau, V. Monitoring of amino acids and antibody N-glycosylation in high cell density perfusion culture based on Raman spectroscopy. *Biochem. Eng. J.* **182**, 108426 (2022).
- Ishigaki, M. et al. Development of an amino acid sequence-dependent analytical method for peptides using near-infrared spectroscopy. *Analyst* **147**, 3634–3642 (2022).
- Hubli, G. B., Banerjee, S. & Rathore, A. S. Near-infrared spectroscopy based monitoring of all 20 amino acids in mammalian cell culture broth. *Talanta* **254**, 124187 (2023).
- Zavala-Ortiz, D. A. et al. Support vector and locally weighted regressions to monitor monoclonal antibody glycosylation during CHO cell culture processes, an enhanced alternative to partial least squares regression. *Biochem. Eng. J.* **154**, 107457 (2020).
- Thakur, G., Hebhi, V. & Rathore, A. S. An NIR-based PAT approach for real-time control of loading in Protein A chromatography in continuous manufacturing of monoclonal antibodies. *Biotechnol. Bioeng.* **117**, 673–686 (2020).
- Chiu, S. et al. Effect of fructose on markers of non-alcoholic fatty liver disease (NAFLD): A systematic review and meta-analysis of controlled feeding trials. *Eur. J. Clin. Nutr.* <https://doi.org/10.1038/ejcn.2014.8> (2014).
- Futatsugi, K. et al. Discovery of PF-06835919: A potent inhibitor of ketohexokinase (KHK) for the treatment of metabolic disorders driven by the overconsumption of fructose. *J. Med. Chem.* **63**, 13546–13560 (2020).

11. Ishimoto, T. et al. High-fat and high-sucrose (western) diet induces steatohepatitis that is dependent on fructokinase. *Hepatology* **58**, 1632–1643 (2013).
12. Tee, S. S. et al. Ketohexokinase-mediated fructose metabolism is lost in hepatocellular carcinoma and can be leveraged for metabolic imaging. *Sci. Adv.* **8**, eabm7985 (2022).
13. Moco, S. Studying metabolism by NMR-based metabolomics. *Front. Mol. Biosci.* <https://doi.org/10.3389/fmolb.2022.882487> (2022).
14. Nagana Gowda, G. A. & Djukovic, D. Overview of mass spectrometry-based metabolomics: Opportunities and challenges. *Methods Mol. Biol.* **1198**, (2014).
15. Rodriguez-Saona, L. E., Fry, F. S., McLaughlin, M. A. & Calvey, E. M. Rapid analysis of sugars in fruit juices by FT-NIR spectroscopy. *Carbohydr. Res.* **336**, 63–74 (2001).
16. Kozma, B. et al. On-line prediction of the glucose concentration of CHO cell cultivations by NIR and Raman spectroscopy: Comparative scalability test with a shake flask model system. *J. Pharm. Biomed. Anal.* **145**, 346–355 (2017).
17. Cheng, J. H. & Sun, D. W. Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food Eng. Rev.* **2016**(9), 36–49 (2016).
18. Sorvaniemi, J., Kinnunen, A., Tsados, A. & Mälkki, Y. Using partial least squares regression and multiplicative scatter correction for FT-NIR data evaluation of wheat flours. *LWT - Food Sci. Technol.* **26**, 251–258 (1993).
19. Roychoudhury, P., O’Kennedy, R., McNeil, B. & Harvey, L. M. Multiplexing fibre optic near infrared (NIR) spectroscopy as an emerging technology to monitor industrial bioprocesses. *Anal. Chim. Acta* **590**, 110–117 (2007).
20. Henriques, J. G., Buziol, S., Stocker, E., Voogd, A. & Menezes, J. C. Monitoring mammalian cell cultivations for monoclonal antibody production using near-infrared spectroscopy. *Adv. Biochem. Eng. Biotechnol.* **116**, 29–72 (2009).
21. Whelan, J., Craven, S. & Glennon, B. In situ Raman spectroscopy for simultaneous monitoring of multiple process parameters in mammalian cell culture bioreactors. *Biotechnol. Prog.* **28**, 1355–1362 (2012).
22. Rafferty, C. et al. Raman spectroscopic based chemometric models to support a dynamic capacitance based cell culture feeding strategy. *Bioprocess Biosyst. Eng.* **43**, 1415–1429 (2020).
23. Rafferty, C. et al. Analysis of chemometric models applied to Raman spectroscopy for monitoring key metabolites of cell culture. *Biotechnol. Prog.* **36**, e2977 (2020).
24. Mehdizadeh, H. et al. Generic Raman-based calibration models enabling real-time monitoring of cell culture bioreactors. *Biotechnol. Prog.* **31**, 1004–1013 (2015).
25. Hu, J., Chen, G., Yongjun, Q. I. N. & Zhou, W. A raman spectroscopy integrated perfusion cell culture system for monitoring and auto-controlling perfusion cell culture. Preprint at (2020).
26. Mehendale, N. et al. A nuclear magnetic resonance (NMR) platform for real-time metabolic monitoring of bioprocesses. *Molecules* **25**, 4675 (2020).
27. Wang, R. C. C., Campbell, D. A., Green, J. R. & Čuprovčić-Culf, M. Automatic 1D ¹H NMR metabolite quantification for bioreactor monitoring. *Metabolites* <https://doi.org/10.3390/metabo11030157> (2021).
28. Hevaganinge, A. et al. Fast-training deep learning algorithm for multiplex quantification of mammalian bioproduction metabolites via contactless short-wave infrared hyperspectral sensing. *ACS Omega* **8**, 14774–14783 (2023).
29. Wegelin, A., J. A Survey of Partial Least Squares (PLS) Methods.
30. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Preprint at <https://doi.org/10.48550/arXiv.1201.0490> (2018).
31. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
32. Anders, S. Visualization of genomic data with the Hilbert curve. *Bioinformatics* **25**, 1231–1235 (2009).
33. Altay, G. Programming the Hilbert Curve. (2024).
34. Cai, T. T. & Wang, L. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inf. Theory* **57**, 4680–4688 (2011).
35. Zhao, S., Zhang, B., Yang, J., Zhou, J. & Xu, Y. Linear discriminant analysis. *Nat. Rev. Methods Primer* **4**, 1–16 (2024).
36. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
37. Wei, J. & Wang, X. An Overview on Linear Unmixing of Hyperspectral Data. *Math. Probl. Eng.* **2020**, 3735403 (2020).
38. Palsson, B., Sveinsson, J. R. & Ulfarsson, M. O. Blind hyperspectral unmixing using autoencoders: A critical comparison. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **15**, 1340–1372 (2022).
39. Tsenkova, R., Muncán, J., Pollner, B. & Kovacs, Z. Essentials of aquaphotomics and its chemometrics approaches. *Front. Chem.* **6**, 363 (2018).
40. Czarnecki, M. A., Morisawa, Y., Futami, Y. & Ozaki, Y. Advances in molecular structure and interaction studies using near-infrared spectroscopy. *Chem. Rev.* **115**, 9707–9744 (2015).
41. Workman, J. & Weyer, L. *Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy*. (2012). <https://doi.org/10.1201/b11894>
42. Kojić, D., Tsenkova, R. & Yasui, M. Improving accuracy and reproducibility of vibrational spectra for diluted solutions. *Anal. Chim. Acta* **955**, 86–97 (2017).
43. Kojić, D., Tsenkova, R., Tomobe, K., Yasuoka, K. & Yasui, M. Water confined in the local field of ions. *ChemPhysChem* **15**, 4077–4086 (2014).
44. Fontaine, J., Hörj, J. & Schirmer, B. Near-infrared reflectance spectroscopy enables the fast and accurate prediction of the essential amino acid contents in soy, rapeseed meal, sunflower meal, peas, fishmeal, meat meal products, and poultry meal. *J. Agric. Food Chem.* **49**, 57–66 (2001).

Acknowledgements

This study was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1840340 to CMW. National Institutes of Health grant R21CA245492 (S.S.T.). We also acknowledge the support of the University of Maryland, Baltimore, Institute for Clinical & Translational Research (ICTR) and the National Center for Advancing Translational Sciences (NCATS) Clinical Translational Science Award (CTSA) grant number 1UL1TR003098. We also acknowledge the support of the National Cancer Institute-Cancer Center Support Grant (CCSG) - P30CA134274, as well as the Maryland Department of Health’s Cigarette Restitution Fund Program CH-649-CRF. The funders did not participate in research or review any details of this study.

Author contributions

A.H., E.L., A.F., M.M., N.T.M., B.R., J.Y., J.E., R.H. participated in collection of fructose data as well as experimental setup for related experiments. A.H., J.Y., J.E., R.H., A.M., A.K., J.C., Y.Y. participated in collection of data related to spent cell media (glucose, lactate, viable cell density). A.H. prepared manuscript and figures. All authors reviewed manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025