TREDL-SPEC: TEMPORAL REPLICATE-AWARE DISENTANGLING FOR LONGITUDINAL CLINICAL SPECTROSCOPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal spectroscopy can provide non-invasive monitoring of biological change, but repeated measurements introduce variability from both temporal state and measurement variation. Existing representation learning methods have separately explored static–temporal factorization, known relationships between observations, and measurement-specific variation, but these complementary sources of information have not been jointly leveraged to structure learned representations. This gap is particularly consequential for clinical spectroscopy, an increasingly explored but still comparatively underdeveloped area of biomedical AI, where substantial measurement variability can be difficult to distinguish from true biological change and can compromise the reliability and clinical trustworthiness of learned representations. We introduce TREDL-Spec (Temporal Replicate-Aware Disentangling for Longitudinal Spectroscopy), a replicate-aware framework that factorizes each spectrum into subject-stable, temporally evolving, and measurement-specific representations. Nested invariance constraints exploit repeated measurements (replicates) to explicitly associate these representations with their intended sources, while temporal dynamics are learned separately from measurement-specific nuisance variation. We evaluate the framework on controlled synthetic spectroscopy data and longitudinal human Raman spectroscopy data. TREDL-Spec achieved the highest SAP and DCI disentanglement on the synthetic dataset and the highest DCI on the human dataset, while its most predictive latent dimensions provided the strongest subject and temporal factor predictability on both datasets. TREDL-Spec also achieved the lowest future-spectrum prediction error across all evaluated metrics. These results demonstrate that explicitly structuring representations around biological and measurement-specific sources can support longitudinal spectral prediction while providing an explicit factorization of measurement variability.

1 INTRODUCTION

Spectroscopy offers a non-invasive and information-rich approach for characterizing biological systems, but interpreting changes in biological spectra is challenging because multiple sources of variation can produce similar changes in the measured signal. Raman measurements may vary because of instrumentation and sampling location, as demonstrated in skin across multiple Raman systems and measurement sites by Pence et al. (2013) and across anatomical locations in the upper gastrointestinal tract by Bergholt et al. (2011). At the same time, genuine biological heterogeneity can produce substantial spectral differences even within the same subject (Andrade et al., 2007), while physiological state can further alter spectra (Traynor et al., 2018). More broadly, vibrational spectroscopy studies have shown that sample preparation and measurement procedures can introduce additional variation (Bonnier et al., 2014; Pahlow et al., 2018). Consequently, a change in spectral measurements does not have a unique interpretation: it may reflect subject-specific characteristics, changes in biological state, or variation introduced by the measurement process. This ambiguity is particularly important in longitudinal spectroscopy, where meaningful biological changes must be distinguished from variation arising from individual measurement realizations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

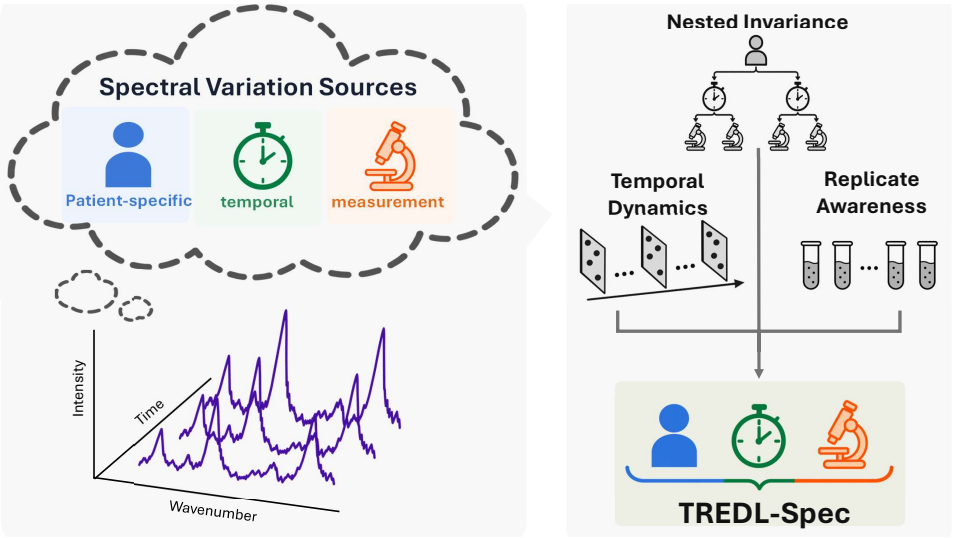


Figure 1: Longitudinal spectra mix patient-specific, temporal, and measurement-specific variation. TREDL-Spec integrates hierarchical relationships, temporal dynamics, and replicate awareness to explicitly separate these sources.

Existing representation learning methods provide mechanisms for separating multiple sources of variation, including static and temporally evolving factors, but typically do not explicitly incorporate measurement-specific variation as a distinct factor or define the biological meaning of each learned representation through known relationships between observations. For clinical spectroscopy, this distinction is important: a representation should not only separate factors of variation, but should provide an explicit correspondence between each learned factor and its intended source. This reduces the need for post-hoc inference about what a latent representation represents and provides a more direct basis for interpreting the learned features.

We therefore introduce TREDL-Spec, a replicate-aware representation-learning framework for longitudinal spectroscopy that factorizes each observation into three complementary representations: **subject-stable characteristics, temporally evolving biological state, and measurement-specific variation**. The framework exploits repeated measurements acquired from the same subject and timepoint to establish nested invariance constraints: same-time replicates share both subject-stable and temporal representations, whereas measurements from the same subject at different timepoints share only the subject-stable representation. Measurement-specific variation is represented separately and retained for reconstruction, allowing the model to distinguish changes in biological state from differences between measurement realizations. Temporal dynamics are learned exclusively in the biological latent space, separating longitudinal state prediction from measurement-specific nuisance variation.

The primary contributions of this work are:

- **Explicitly labeled factorization of longitudinal spectroscopic observations.** We define separate latent representations for subject-stable characteristics, temporally evolving biological state, and measurement-specific nuisance variation, with each factor assigned a specific role through the training objective rather than identified post hoc.
- **Nested replicate-aware invariance constraints.** We use the experimental structure of repeated measurements to impose different invariances at different levels: same-time replicates share both subject and temporal representations, while observations across time share only subject-stable information.
- **Separation of biological dynamics from measurement variation.** The temporal dynamics network uses the current temporal representation and measurement times to predict future temporal state, while measurement-specific information remains represented separately in z^N.

- **Evaluation on controlled and real longitudinal spectroscopy.** We evaluate the framework using synthetic spectroscopic data with controlled sources of variation and longitudinal human Raman spectroscopy data, assessing latent-factor disentanglement and downstream future-spectrum prediction.

2 Related Works

Disentangled Representation Learning. Disentangled representation learning seeks to represent observations using separate latent variables corresponding to distinct underlying factors of variation. Two methods are particularly relevant to our evaluation. CaRiNG (Chen et al., 2024) learns identifiable temporal causal representations under non-invertible generative processes by using temporal context to recover latent information. TopDis (Wang et al., 2024) promotes disentanglement using a differentiable topology-preserving loss over latent traversals. Neither method was originally developed for spectroscopic data, motivating their adaptation here as comparison methods for longitudinal spectral representation learning. Several recent approaches have investigated the separation of static and temporally varying factors. Berman et al. (2023) proposed a structured Koopman autoencoder that separately models static and dynamic representations, based on the assumption that the static representation can be inferred from a single sample and that static and dynamic codes can be learned separately. While this provides a mechanism for separating persistent and changing information, it does not explicitly model additional variation between replicates that should remain invariant to both subject identity and biological state. This distinction is important for spectroscopy, where replicates from the same subject and timepoint may exhibit substantial measurement-specific variation despite representing the same underlying biological state.

Berman et al. (2024) further investigated unsupervised disentanglement of static and dynamic factors, including mechanisms for limiting information leakage between the corresponding representations. Barami et al. (2025) similarly investigated multifactor sequential disentanglement and methods for associating learned latent factors with higher-level semantic properties. In these unsupervised settings, however, the learned latent factors are not necessarily assigned their semantic identities directly through known relationships between observations. Additional evaluation, classification, or semantic alignment may therefore be required to determine which learned features correspond to particular high-level factors. For clinical spectroscopy, where the distinction between biological and measurement-related variation is itself scientifically important, relying on a subsequent inference step to determine the meaning of latent factors introduces additional ambiguity.

Longitudinal Representation Learning. Longitudinal representation learning provides a complementary perspective by explicitly using repeated observations from the same subject to model persistent characteristics and changes over time. Zhao et al. (2021) introduced longitudinal self-supervised learning for neuroimaging data, using repeated measurements to identify information that remains consistent across time and information that changes longitudinally. More recent work has investigated structured latent representations of disease progression in longitudinal imaging; for example, Chen et al. (2025) modeled patient-specific disease dynamics using latent flow matching and longitudinal MRI data. These approaches demonstrate the value of repeated observations for learning representations of biological change. However, longitudinal consistency alone does not distinguish temporal biological variation from measurement-specific variation occurring between repeated measurements at the same timepoint. In spectroscopic data, where such within-timepoint variation can be substantial, these two forms of consistency provide distinct constraints.

Representation Learning for Measurement Variation and Replicate Data. A related body of work explicitly models variation associated with measurement conditions or repeated observations. Bharadwaj et al. (2022) introduced SymAE, a symmetric autoencoder that separates coherent information associated with an underlying physical state from incoherent nuisance information associated with sensing conditions across repeated measurements. Similarly, Castorena & Oyen (2023) proposed SpectroscopyNet for learning signal and noise representations from spectroscopy measurements without requiring clean reference signals. These approaches demonstrate that measurement variability can be incorporated directly into representation learning rather than treated solely as a preprocessing problem. Domain-adaptation approaches provide another related perspective: Zhang et al. (2024) proposed adversarial domain separation for near-infrared spectra, while Li et al. (2026) investigated Raman spectral calibration transfer across instruments. However, these approaches

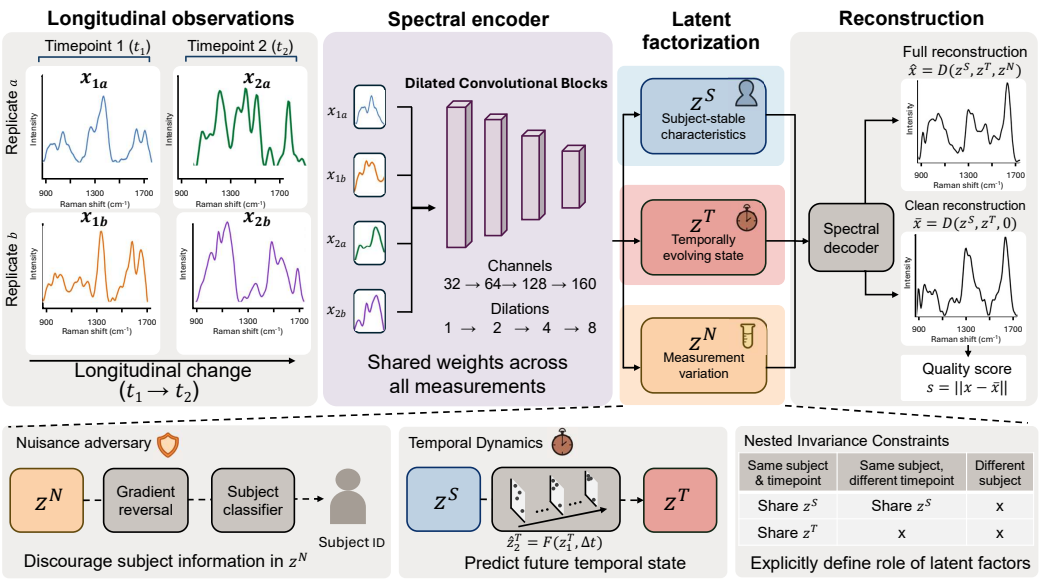


Figure 2: Overview of TREDL-Spec. Replicate longitudinal Raman spectra are encoded into three latent factors representing subject-stable characteristics (z^S), temporally evolving biological state (z^T), and measurement-specific nuisance variation (z^N). Nested invariance constraints and a nuisance adversary encourage the intended factorization, while a temporal dynamics model propagates z^T to future timepoints. The decoder reconstructs observed and nuisance-free spectra, and future spectra are predicted using the subject and predicted temporal representations with $z^N = 0$.

primarily seek invariance across sensing conditions, domains, or instruments rather than jointly separating measurement variation from subject-specific characteristics and longitudinal biological state within an individual.

3 Methods

3.1 Data

We evaluated the proposed framework on both controlled synthetic spectroscopy data and longitudinal human Raman spectroscopy data. In both settings, longitudinal examples at training consisted of two replicate spectra $\{x_{1a}, x_{1b}, x_{2a}, x_{2b}\}$ along with timepoints t_1, t_2, where the subscripts denote timepoint (1 or 2) and replicate (a or b), respectively. The model therefore receives replicate measurements of the same underlying sample at each given timepoint while explicitly distinguishing temporal and measurement-specific variation.

For the synthetic dataset, spectra were generated from experimentally acquired pure-component Raman spectra using two measurement systems. Synthetic subjects were assigned longitudinally varying component compositions, with replicate spectra generated from the same underlying composition but with system-specific variation in the component spectra. Details of the experimental spectra, preprocessing, and synthetic longitudinal data generation are provided in Appendix A.1.

The human dataset consisted of longitudinal cervical Raman spectra acquired from 59 pregnant patients. Measurements were collected at multiple gestational timepoints, with replicate measurements obtained at each visit. Details of participant recruitment, Raman instrumentation, acquisition protocol, and data preprocessing are provided in Appendix A.1.

Both datasets were processed using a standard Raman spectroscopy preprocessing pipeline (Haugen et al., 2025); detailed preprocessing procedures are provided in Appendix A.1. Data were divided into training, validation, and test sets at the subject level, preventing measurements from the same

subject from appearing in multiple partitions. Spectra were normalized using the same procedure for the proposed model and all baselines.

For future-spectrum prediction, the known clean spectrum was used as the target for the synthetic dataset. For the human dataset, where a clean ground-truth spectrum is unavailable, the prediction target was formed by averaging the two replicates at the future timepoint (Haugen et al., 2026).

3.2 Architectural Design

As illustrated in Figure 2, TREDL-Spec decomposes each observed spectrum into three complementary latent factors: subject-stable information, temporally evolving biological state, and measurement-specific (nuisance) variation. For a spectrum acquired from subject i at time t and replicate r, the model represents the observation as

$$x_{i,t,r} = D\left(z_i^S, z_{i,t}^T, z_{i,t,r}^N\right),$$

where z_i^S denotes subject-stable information, $z_{i,t}^T$ denotes the biological state, and $z_{i,t,r}^N$ denotes measurement-specific nuisance variation.

A shared dilated convolutional neural network (dCNN)-based encoder maps each spectrum to a feature representation, then three separate projection heads map the shared representation to the three latent spaces, Z^S, Z^T, and Z^N (representing the subject-stable features, temporal features, and nuisance features, respectively) using three major constraints: nested replicate relationships (Section 3.3), temporal dynamics (Section 3.4), and adversarial learning (Section 3.5).

Details of model parameters can be found in Appendix A.2.

3.3 Nested Replicate Relationships

The model exploits the nested structure of repeated longitudinal measurements. Replicate spectra acquired from the same subject at the same timepoint should share both subject-stable and temporal biological information, while measurement-specific variation is permitted to differ:

$$z_{i,t,a}^S \approx z_{i,t,b}^S, \qquad z_{i,t,a}^T \approx z_{i,t,b}^T.$$

Additionally, spectra from the same subject at different timepoints are expected to share subject-stable information while allowing their temporal representations to change:

$$z_{i,t_1}^S \approx z_{i,t_2}^S.$$

We explicitly encourage these relationships by minimizing a subject-replicate loss in the projection head to Z^S:

$$L_S = \sum_{(r,t),(r',t')} \| z_{i,t,r}^S - z_{i,r',t'}^S \|^2$$

and by minimizing a temporal-replicate loss in the projection head to Z^T

$$L_T = \| z_{1,a}^T - z_{1,b}^T \|^2 + \| z_{2,a}^T - z_{2,b}^T \|^2 .$$

Thus, the model imposes a nested hierarchy of invariances: same-subject replicates are constrained to agree in both stable and temporal representations, whereas measurements from the same subject across time are constrained only in the subject-stable representation.

3.4 Temporal Dynamics Network

To model longitudinal change, a multilayer perceptron (MLP) temporal dynamics network F predicts the future temporal representation from the temporal representation at the first timepoint and the measurement times:

$$\hat{z}_{i,t_2}^T = F(z_{i,t_1}^T, t_1, t_2, \Delta t).$$

For paired replicate inputs at t_1, the two temporal representations are averaged before being passed to the temporal dynamics network. The two subject-stable representations are averaged separately

and retained for decoding. The predicted future temporal representation is then combined with the averaged subject-stable representation to generate a future spectrum while setting the nuisance representation to zero.

For training and evaluation, the future target is the known clean spectrum for the synthetic dataset. For the human dataset, where no clean ground truth is available, the future target is the mean of the two replicate spectra at the future timepoint:

$$x_{i,t_2,\text{mean}} = \frac{1}{2}\left(x_{i,t_2,a} + x_{i,t_2,b}\right).$$

The future prediction loss therefore measures how accurately the learned longitudinal representation can predict the future spectral signal while excluding measurement-specific nuisance variation.

3.5 Adversarial Classifier

The nuisance representation $z^N_{i,t,r}$ is intended to capture variation specific to an individual spectral measurement rather than persistent subject characteristics or longitudinal biological state. To discourage subject identity from being encoded in this representation, an MLP classification model is trained to predict subject ID using standard cross-entropy loss and is applied to z^N using a gradient-reversal operation.

3.6 Spectral Autoencoder

The model has two main tasks: 1) predict a clean spectrum at a given future timepoint based on two earlier same-subject replicates, and 2) independently reconstruct each observed input spectrum. Hence, the spectral autoencoder is trained using a weighted combination of clean-prediction loss and reconstruction loss.

Clean Prediction Given two replicate spectra $x_{1,a}, x_{1,b}$ at t_1 from a single subject, the model is asked to predict a clean signal x_2 at the specified t_2 for the same subject. This is achieved by first encoding the two signals $x_{1,a}, x_{1,b}$ into the three latent spaces as described in previous sections. The subject-stable representations $z^S_{1,a}$ and $z^S_{1,b}$ are averaged to produce z^S_1, and the temporal representations $z^T_{1,a}$ and $z^T_{1,b}$ are averaged to produce z^T_1. Next, the temporal dynamics network predicts the future temporal representation,

$$\hat{z}^T_2 = F(z^T_1, t_1, t_2, \Delta t),$$

as described in Section 3.4. Finally, the decoder reconstructs the clean prediction signal

$$\bar{x}_2 = D(z^S_1, \hat{z}^T_2, 0)$$

using the subject-stable representation, predicted future temporal representation, and none of the nuisance representation. During training, the model minimizes the mean-squared error (MSE) between the predicted $\bar{x}_2$ and the target x_2. For the synthetic dataset, the ground-truth clean signal is used as the target. In the human dataset, where no ground truth is available, the average of the two future observed spectra is used as a silver standard.

Reconstruction The model is trained in parallel to reconstruct each individual measurement observation. For a single input spectrum, arbitrarily choosing $x_{1,a}$, the model reconstructs $\hat{x}_{1,a}$ directly using the measurement's own latent representations,

$$\hat{x}_{1,a} = D(z^S_{1,a}, z^T_{1,a}, z^N_{1,a}).$$

During training, this reconstruction is performed independently for each of the four input signals, and the MSE loss for the four reconstructions is averaged.

3.7 Experiments and Evaluation

We evaluated TREDL-Spec primarily on its ability to learn disentangled representations of longitudinal spectral variation. We compare TREDL-Spec against two representation-learning methods with explicit disentanglement objectives: CaRiNG (Chen et al., 2024) and TopDis (Balabin et al., 2024). CaRiNG learns temporal causal representations using temporal context and a latent transition mechanism, whereas TopDis encourages disentanglement through a topology-preserving latent

Table 1: Disentanglement results on the synthetic dataset. MIG, SAP, and DCI are computed over individual latent dimensions of the complete learned representations. Factor predictability reports the SAP predictability of the single most predictive latent dimension for each ground-truth factor. Higher values indicate better performance. Best performance is shown in bold; second-best is shown in italics. The shaded row denotes TREDL-Spec.

Model	MIG ↑	SAP ↑	DCI ↑	Subject ↑	Time ↑	Measurement ↑
CaRiNG	0.008344	*0.037558*	0.219633	0.236607	0.120890	**0.662913**
TopDis	**0.017131**	0.031113	*0.377505*	*0.352679*	*0.193225*	0.240507
TREDL-Spec	*0.015669*	**0.073325**	**0.770260**	**0.602679**	**0.411708**	*0.560563*

objective. Because neither method was originally designed for one-dimensional spectroscopic measurements, we adapted their input and output components to operate on Raman spectra while preserving their core representation-learning mechanisms. These adaptations were limited to what was required to accommodate the spectral data structure and the downstream future-spectrum prediction task. We additionally evaluated future-spectrum prediction as a downstream task to assess whether the learned representations retained information useful for longitudinal prediction.

For all experiments, models were trained using only the training subset, model selection was performed using the validation subset, and all reported metrics were calculated on held-out test data.

3.7.1 Latent-Factor Disentanglement

Latent disentanglement was quantitatively evaluated using Mutual Information Gap (MIG), Separated Attribute Predictability (SAP), and DCI disentanglement. These standard metrics were calculated at the level of individual scalar latent coordinates rather than over the complete multi-dimensional latent subspaces.

For TREDL-Spec, the complete representation was formed by concatenating the three learned subspaces,

$$z = [z^S; z^T; z^N], \tag{1}$$

giving a 96-dimensional representation. CaRiNG and TopDis produced 8- and 32-dimensional representations, respectively. Each scalar coordinate was treated independently when computing MIG, SAP, and DCI.

For the synthetic dataset, the known ground-truth factors were subject identity, temporal state, and measurement variation. Subject identity was treated as categorical, whereas time and measurement variation were treated as continuous variables. For the human dataset, only subject identity and temporal state were available as known factors; no direct ground-truth target was available for measurement-specific variability.

MIG measures whether information about a ground-truth factor is preferentially concentrated in one latent coordinate relative to the next-most informative coordinate. SAP measures the difference in held-out factor predictability between the two most predictive coordinates for each factor. DCI disentanglement evaluates whether individual coordinates selectively represent particular factors rather than mixing information from several factors. Higher values indicate stronger disentanglement.

The disentanglement metrics capture complementary properties of the learned representations (Tables 1 and 2). On the synthetic dataset, TopDis achieved the highest MIG, while TREDL-Spec achieved the highest SAP and substantially higher DCI disentanglement than either baseline. TREDL-Spec also provided the strongest SAP predictability for subject identity and temporal state, while CaRiNG achieved the strongest predictability for measurement variation. On the human dataset, TopDis achieved the highest MIG and SAP, whereas TREDL-Spec achieved the highest DCI disentanglement and the strongest subject and time predictability.

Together, these results show complementary strengths across the disentanglement metrics. TREDL-Spec achieved the strongest synthetic SAP and DCI, the strongest human DCI, and the strongest

Table 2: Disentanglement results on the human Raman dataset. MIG, SAP, and DCI are computed over individual latent dimensions of the complete learned representations. Factor predictability reports the SAP predictability of the single most predictive latent dimension for each available ground-truth factor. Higher values indicate better performance. Best performance is shown in bold; second-best is shown in italics. The shaded row denotes TREDL-Spec.

Model	MIG ↑	SAP ↑	DCI ↑	Subject ↑	Time ↑
CaRiNG	0.012663	0.007979	0.113345	0.612500	0.003985
TopDis	**0.032510**	**0.039712**	*0.354915*	*0.637500*	*0.115018*
TREDL-Spec	*0.019389*	*0.015761*	**0.623950**	**0.700000**	**0.193493**

factor-specific prediction of subject and temporal information on both datasets, while TopDis achieved the highest MIG on both datasets.

3.7.2 Qualitative Latent-Space Visualization

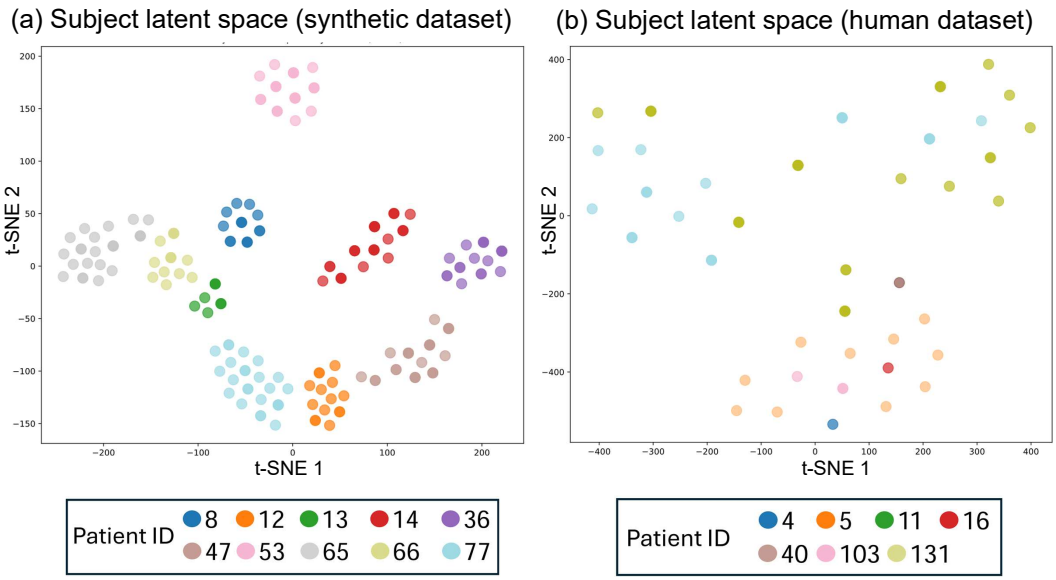


Figure 3: t-SNE visualization of the learned subject-stable representations (z^S) for the held-out (a) synthetic and (b) human Raman test subsets. Points are colored by de-identified subject ID.

To complement the quantitative disentanglement analysis, we visualized the complete subject-stable latent space z^S using t-SNE for both datasets (Figure 3). In the synthetic dataset, repeated measurements from the same subject form localized groups within the learned representation. Similar organization is observed in the human Raman dataset despite substantially greater biological and measurement variability. These visualizations provide qualitative context for the quantitative disentanglement results while illustrating structure across the complete multi-dimensional z^S representation.

3.7.3 Future-Spectrum Prediction

We additionally evaluated future-spectrum prediction as a downstream task. Given two replicate measurements at t_1 and a specified future timepoint t_2, TREDL-Spec averages the two temporal representations at t_1 and uses this representation together with t_1, t_2, and Δt to predict the future temporal representation. The two subject-stable representations are averaged separately and combined with the predicted temporal representation for decoding with $z^N = 0$. CaRiNG uses its

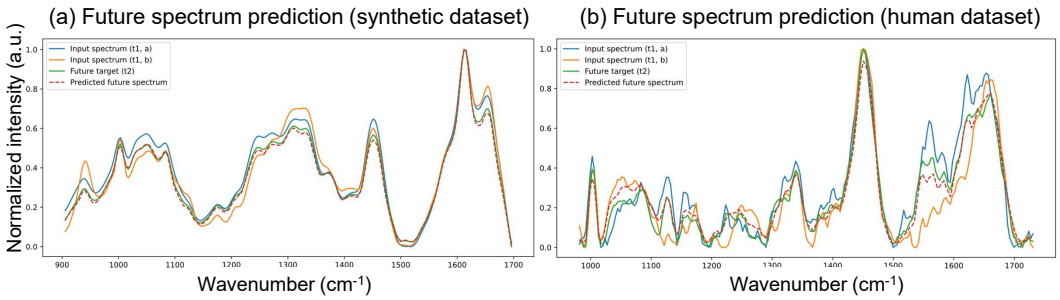


Figure 4: Future-spectrum predictions from TREDL-Spec (red dashed line) against input replicates (blue and orange) and target (green). (a) The synthetic dataset evaluation uses the known ground truth as the target, whereas (b) the human dataset evaluation uses the averaged future replicates as the silver-standard target.

learned temporal transition model, while TopDis uses a downstream prediction head trained on its frozen representation.

For the synthetic dataset, predictions were evaluated against the known clean spectrum at t_2. For the human Raman dataset, the replicate-averaged spectrum at t_2 was used as a silver-standard target. Performance was evaluated using MSE, MAE, and spectral angle mapper (SAM). TREDL-Spec achieved the lowest prediction error across all three metrics on both datasets, with MSE values of 0.000657 on the synthetic dataset and 0.013319 on the human dataset. Representative TREDL-Spec predictions are shown in Figure 4. Complete baseline comparison results are provided in Appendix A.4.

4 Conclusion

In this work, we presented TREDL-Spec, a longitudinal spectral modeling framework that explicitly disentangles three sources of variation in clinical vibrational spectra: subject-specific characteristics, temporally evolving biological state, and measurement-specific nuisance variation. Rather than relying on post-hoc interpretation of an unconstrained latent representation, TREDL-Spec assigns each latent space a defined role through replicate-aware nested constraints and the structure of the temporal dynamics.

Synthetic experiments provide a controlled setting in which the underlying biological and measurement-specific signals are known, allowing us to directly evaluate the proposed factorization. TREDL-Spec achieved the highest SAP and DCI disentanglement on the synthetic dataset and the highest DCI on the human Raman dataset. Its most predictive latent dimensions also achieved the strongest subject and temporal factor predictability on both datasets, while TopDis achieved the highest MIG on both datasets and the highest SAP on the human data. TREDL-Spec additionally achieved the lowest future-spectrum prediction error across MSE, MAE, and SAM on both datasets.

Together, these results support the use of explicitly structured latent representations as a means of making longitudinal spectral models more interpretable. By assigning distinct latent spaces to subject, temporal, and measurement-specific variation, the framework provides a direct basis for examining what information the model uses to represent an observation, how that representation changes over time, and which differences are attributable to individual measurement realizations. This explicit factorization may provide a more transparent foundation for developing predictive models from clinical spectroscopy, where distinguishing biological change from measurement variation is essential for meaningful interpretation.

AI use statement

In this work, we have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining

hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, nor interpreting results, and assisting with translation is not applicable to this work. We have used generative AI tools for improving code and paper readability, and as a supplement when searching for related literature. We have reviewed all AI-assisted work. LLM-edited code was verified and tested by the primary author and LLM-generated text was reviewed and revised as-needed by the primary author. Related literature searches were done in tandem with manual literature search, and any suggested literature was read and analyzed by the author for relevance to the proposed work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

The human dataset used in this work was obtained from a previous study which was conducted with IRB approval, and no personally-identifiable information is released.

Reproducibility statement

We have taken steps to facilitate reproducibility of our results. Detailed descriptions of synthetic data generation, data preprocessing, model architectures, training objectives, and evaluation protocols are provided in the Methods and Appendix sections. Hyperparameters and implementation details are reported in the supplementary material. Anonymized source code and synthetic data generation code is available at https://anonymous.4open.science/r/spectroscopyDisentanglement/ to facilitate independent verification and reproduction of the reported experiments.

References

P. O. Andrade, R. A. Bitar, K. Yassoyama, H. Martinho, A. M. E. Santo, P. M. Bruno, and A. A. Martin. Study of normal colorectal tissue by FT-Raman spectroscopy. *Analytical and Bioanalytical Chemistry*, 387(5):1643–1648, March 2007. ISSN 1618-2642. doi: 10.1007/s00216-006-0819-1.

Nikita Balabin, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, and Serguei Barannikov. Disentanglement Learning via Topology. 2024.

Tal Barami, Nimrod Berman, Ilan Naiman, Amos H Hason, Rotem Ezra, and Omri Azencot. Disentanglement Beyond Static vs. Dynamic: A Benchmark and Evaluation Framework for Multi-Factor Sequential Representations. In *Advances in Neural Information Processing Systems 38*, pp. 140328–140370, San Diego, California, USA and Mexico City, Mexico, 2025. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 979-8-3313-3827-5. doi: 10.52202/085713-4223. URL http://www.proceedings.com/085713-4223.html.

Mads Sylvest Bergholt, Wei Zheng, Kan Lin, Khek Yu Ho, Ming Teh, Khay Guan Yeoh, Jimmy Bok Yan So, and Zhiwei Huang. Characterizing variability in in vivo Raman spectra of different anatomical locations in the upper gastrointestinal tract toward cancer detection. *Journal of Biomedical Optics*, 16(3):037003, March 2011. ISSN 1560-2281. doi: 10.1117/1.3556723.

Nimrod Berman, Ilan Naiman, and Omri Azencot. Multifactor Sequential Disentanglement via Structured Koopman Autoencoders. March 2023. URL https://www.alphaxiv.org/abs/2303.17264.

Nimrod Berman, Ilan Naiman, Idan Arbiv, Gal Fadlon, and Omri Azencot. Sequential Disentanglement by Extracting Static Information From A Single Sequence Element. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 3539–3564. PMLR, July 2024. URL https://proceedings.mlr.press/v235/berman24a.html.

Pawan Bharadwaj, Matthew Li, and Laurent Demanet. Redatuming physical systems using symmetric autoencoders, February 2022. URL http://arxiv.org/abs/2108.02537. arXiv:2108.02537 [physics.comp-ph].

Franck Bonnier, François Petitjean, Matthew J. Baker, and Hugh J. Byrne. Improved protocols for vibrational spectroscopic analysis of body fluids. *Journal of Biophotonics*, 7(3-4):167–179, April 2014. ISSN 1864-0648. doi: 10.1002/jbio.201300130.

Juan Castorena and Diane Oyen. SpectroscopyNet: Learning to pre-process Spectroscopy Signals without clean data, January 2023. URL http://arxiv.org/abs/2110.13748. arXiv:2110.13748 [cs.LG].

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 7236–7259. PMLR, July 2024. URL https://proceedings.mlr.press/v235/chen24ai.html.

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning Patient-Specific Disease Dynamics With Latent Flow Matching For Longitudinal Imaging Generation. October 2025. URL https://openreview.net/forum?id=cuGnuOfQ4U.

Ezekiel J. Haugen, Andrea K. Locke, Lily H. Dao, Alec B. Walter, Pratheepa K. Rasiah, Justin S. Baba, Matthew A. Buendia, Hernán Correa, Girish Hiremath, and Anita Mahadevan-Jansen. Biochemical detection of pediatric eosinophilic esophagitis using high wavenumber Raman endoscopy and stimulated Raman microscopy. *Scientific Reports*, 15(1):22471, July 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-05591-z. URL https://www.nature.com/articles/s41598-025-05591-z.

Ezekiel J. Haugen, Rekha Gautam, Andrea K. Locke, and Anita Mahadevan-Jansen. In vivo Raman spectroscopy for real-time biochemical assessment of tissue pathology and physiology. *Nature Protocols*, January 2026. ISSN 1754-2189, 1750-2799. doi: 10.1038/s41596-025-01274-1. URL https://www.nature.com/articles/s41596-025-01274-1.

Qingbo Li, Yefan Zhang, Xupeng Shao, Zhiqi Zhou, Yan Zhou, Jingying Jiang, Hongbo Bao, Zeya Yan, Wenjun Tu, Ping Feng, and Lipu Zhou. Domain-adaptive Raman spectral calibration transfer for cross-instrument glioma detection. *Analytical Methods*, 18(25):5324–5335, July 2026. ISSN 1759-9660. doi: 10.1039/d6ay00651e. URL https://doi.org/10.1039/d6ay00651e.

Chad A. Lieber and Anita Mahadevan-Jansen. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. *Applied Spectroscopy*, 57(11):1363–1367, November 2003. ISSN 0003-7028, 1943-3530. doi: 10.1366/000370203322554518. URL https://journals.sagepub.com/doi/10.1366/000370203322554518.

Susanne Pahlow, Karina Weber, Jürgen Popp, Bayden R. Wood, Kamila Kochan, Anja Rüther, David Perez-Guaita, Philip Heraud, Nick Stone, Alex Dudgeon, Ben Gardner, Rohith Reddy, David Mayerich, and Rohit Bhargava. Application of Vibrational Spectroscopy and Imaging to Point-of-Care Medicine: A Review. *Applied spectroscopy*, 72(1 Suppl):52–84, September 2018. ISSN 0003-7028. doi: 10.1177/0003702818791939. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6524782/.

Isaac J. Pence, Elizabeth Vargis, and Anita Mahadevan-Jansen. Assessing Variability of in Vivo Tissue Raman Spectra. *Applied Spectroscopy*, 67(7):789–800, July 2013. ISSN 0003-7028, 1943-3530. doi: 10.1366/12-06773. URL https://journals.sagepub.com/doi/10.1366/12-06773.

Damien Traynor, Padraig Kearney, Ines Ramos, Cara M. Martin, John J. O'Leary, and Fiona M. Lyng. A study of hormonal effects in cervical smear samples using Raman spectroscopy. *Journal of Biophotonics*, 11(6):e201700240, June 2018. ISSN 1864-0648. doi: 10.1002/jbio.201700240.

Jinfeng Wang, Sifan Song, Jionglong Su, and S. Kevin Zhou. Distortion-Disentangled Contrastive Learning. pp. 75–85, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Wang_Distortion-Disentangled_Contrastive_Learning_WACV_2024_paper.html?utm_source=chatgpt.com.

Zheyu Zhang, Stavros Avramidis, Yaoxiang Li, Xiaoli Liu, Rundong Peng, Ya Chen, and Zichun Wang. A bidirectional domain separation adversarial network based transfer learning method for near-infrared spectra. *Engineering Applications of Artificial Intelligence*, 137:109140, November 2024. ISSN 0952-1976. doi: 10.1016/j.engappai.2024.109140. URL https://www.sciencedirect.com/science/article/pii/S0952197624012983.

Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M. Pohl. Longitudinal self-supervised learning. *Medical Image Analysis*, 71:102051, July 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102051. URL https://www.sciencedirect.com/science/article/pii/S1361841521000979.

A Appendix

A.1 Data Collection and Synthetic Data Generation

A.1.1 Synthetic Raman Spectroscopy Data

Synthetic longitudinal Raman spectra were generated from experimentally acquired pure-component spectra collected using two Raman spectroscopy systems. System A was a Renishaw inVia^{TM} Raman microscope equipped with a 785 nm laser and a 50×, 0.75 NA objective. The system was precalibrated using neon calibration lines during installation and regularly verified using an internal silicon reference at 520.5 cm^{-1}. System B was a fiber-optic Raman system integrated with a standard clinical endoscope, consisting of a 785 nm diode laser and a custom-designed fiber-optic probe. System B was calibrated using a neon-argon lamp, naphthalene powder, and a 4-acetaminophen tablet as wavenumber references. Pure-component spectra were collected for 11 components on System A: actin, histone, DNA, oleic acid, phosphate-buffered saline, albumin, glycogen, water, amylopectin, RNA, and collagen. Eight of these components were also available on System B: actin, histone, DNA, oleic acid, albumin, glycogen, RNA, and collagen.

All measured pure-component spectra underwent a standard Raman spectroscopy preprocessing pipeline (Haugen et al., 2025). Specifically, autofluorescence was removed using the modified polynomial fitting method (Lieber & Mahadevan-Jansen, 2003), with the polynomial degree selected according to the physical properties of the sample and measurement system. Spectra were restricted to the fingerprint region from 900 to 1700 cm^{-1} and binned to a common spacing of 3.5 cm^{-1}. Spectra from the two systems were interpolated onto a common wavenumber axis spanning the range shared by all component measurements.

Synthetic longitudinal datasets were constructed by combining the experimentally measured component spectra according to simulated subject-specific biological compositions. For each subject, a baseline composition was generated by sampling component abundances in log space and normalizing them to sum to one. The composition was then evolved over normalized longitudinal time using a controlled nonlinear trajectory with progression proportional to t^2, where $t \in [0, 1]$. The trajectory modified the relative abundance of individual components over time, including decreases in collagen, albumin, histone, glycogen, and amylopectin and increases in water, actin, DNA, RNA, and oleic acid. A subject-specific progression rate was sampled around a nominal value of 1.0 to introduce variation in the rate of compositional change between subjects.

Each subject was assigned between four and eight irregularly spaced longitudinal visits, with five replicate spectra generated at each visit. Replicates were generated from the same underlying biological composition while introducing measurement-system variation through the selection of pure-component spectra from System A, System B, or a mixture of the two systems. Replicates followed a repeating System A, System B, mixed, System A, System B pattern. For mixed replicates, each component available on both systems was independently selected from System B with probability 0.5 and from System A otherwise; components available only on System A were always drawn from System A. To further increase measurement realism, we introduced two additional replicate-specific sources of spectral variation. First, we applied Gaussian peak broadening with a randomly sampled width for each replicate to simulate variation in spectral resolution and peak width between measurements. Second, we added a smooth fluorescence baseline generated from a low-order polynomial with randomly sampled slope and curvature, with its amplitude scaled relative to the maximum

spectral intensity. Thus, replicate measurements could differ due to both system-specific spectral variation and other nuisance effects.

Each synthetic spectrum was formed as a linear combination of the selected pure-component spectra weighted by the simulated component abundances. A corresponding clean reference spectrum was generated using the same biological composition with all components drawn from System A and no added noise. The synthetic dataset contained 100 subjects with four to eight longitudinal visits and five replicates per visit.

A.1.2 Longitudinal Human Raman Spectroscopy Dataset

The longitudinal human cervix dataset consisted of Raman spectra acquired from 68 pregnant patients receiving prenatal care. The study was approved by the hospital's Institutional Review Board and conducted in accordance with the Declaration of Helsinki. Patients were eligible for enrollment if they were at least 18 years of age and able to provide legally valid prospective informed consent. Written informed consent was obtained from all participants.

Spectra were acquired using a portable Raman spectroscopy system consisting of a 785 nm diode laser (Innovative Photonics Solutions, Monmouth Junction, NJ), imaging spectrograph (Kaiser Holospec, Ann Arbor, MI), deep-depletion thermoelectrically cooled CCD camera (Princeton Instruments, Trenton, NJ), laptop computer, and fiber-optic Raman probe (EmVision LLC, Loxahatchee, FL). The probe was approximately 2.1 mm in diameter, with an external stainless-steel casing resulting in an outer diameter of 6.3 mm. It contained one optical fiber for delivery of excitation light and seven collection fibers for acquisition of backscattered Raman light. Laser power at the tissue was 80 mW.

Following informed consent, a certified women's health nurse practitioner performed a speculum examination to expose the cervix. The cervix was cleaned with a cotton swab soaked in sterile saline, after which the Raman probe was gently placed in contact with the ectocervix. Ambient room lighting was turned off during spectral acquisition. Three to five measurements were acquired from separate locations around the ectocervix, with the probe positioned on the squamous epithelium. Each measurement used a 3 second integration time, and the complete acquisition procedure required less than 10 minutes per visit. The Raman probe was disinfected between patients according to standard disinfection procedures. Measurements were acquired longitudinally at approximately monthly intervals during the first and second trimesters, and weekly during the third trimester. Postpartum measurements were also obtained but were excluded from this study.

The Raman system was wavelength calibrated daily using a neon-argon lamp. The excitation wavelength and Raman shift were determined using acetaminophen and naphthalene standards with well-characterized Raman shifts. System spectral response was initially characterized using a National Institute of Standards and Technology (NIST)-calibrated tungsten lamp and subsequently calibrated daily using a green glass standard with a similar broadband spectral response. The resulting calibration factor was applied to spectra based on the measured spectral throughput for each day. Spectra were processed using a standard Raman spectroscopy preprocessing pipeline (Haugen et al., 2025), including removal of ambient background light, Savitzky–Golay smoothing, and fluorescence correction using the modified polynomial method (Lieber & Mahadevan-Jansen, 2003). Spectra were restricted to the 930–1780 cm^{-1} spectral range.

For model development and evaluation, the human dataset was restricted to measurements from term pregnancies, and measurements with significant artifacts (such as fluorescence saturation) were excluded. After exclusions and filtering, the dataset consisted of 59 patients and 1508 total measurements. Each example contained two replicate spectra at each of two gestational timepoints. The human dataset was divided at the participant level into training, validation, and test sets using an 80/10/10 split with random seed 42, ensuring that no participant contributed spectra to more than one partition.

A.2 TREDL-Spec Model Parameters

The encoder consists of four one-dimensional convolutional blocks with channel dimensions $(32, 64, 128, 160)$ and kernel size 5. The dilation factor doubles across successive blocks, providing increasing spectral receptive fields while retaining the one-dimensional structure of the spec-

trum. Each block contains a convolution, batch normalization, ReLU activation, and dropout with probability 0.2. Global average pooling followed by a fully connected projection produces a 160-dimensional spectral representation.

For both datasets, TREDL-Spec was trained using the Adam optimizer with learning rate 10^{-4} and weight decay 10^{-5} and a batch size of 32. TREDL-Spec was trained for up to 200 epochs with an early stopping patience of 50 epochs. A learning-rate scheduler was used during training, and the model with the best validation performance was retained.

A.3 Baseline Model Parameters

We compare TREDL-Spec against CaRiNG (Chen et al., 2024) and TopDis (Balabin et al., 2024) using the same subject-level train, validation, and test partitions and the same spectral normalization used for TREDL-Spec. The original methods were not developed for spectroscopic inputs, so we adapted their input/output networks to the one-dimensional spectral setting while retaining the defining mechanisms of each method.

CaRiNG was implemented with an 8-dimensional latent representation and hidden dimension of 128. We retained its temporal-context representation and transition-prior structure while using spectral encoder and decoder components appropriate for one-dimensional Raman inputs. Representation learning jointly optimized reconstruction, temporal-transition, KL, and future-prediction objectives. The model was trained using Adam with learning rate 10^{-4} and weight decay 10^{-5}, with early stopping based on validation performance.

TopDis was implemented with a dilated convolutional encoder using channel dimensions $(32, 64, 128, 160)$, a 32-dimensional latent representation, and a decoder hidden dimension of 256. This adapts the encoder and decoder to one-dimensional Raman spectra while retaining the TopDis topology-preserving representation objective. The representation objective combined spectral reconstruction, KL regularization, and the TopDis topology-preserving loss. We used $\beta = 1.0$, TopDis loss weight 1.0, latent shift scale 0.15, and $k = 5$ neighbors for the topology loss. The representation model was trained using Adam with learning rate 10^{-4} and weight decay 10^{-5}. Because TopDis does not include a native longitudinal future-spectrum prediction mechanism, the learned representation was frozen and a separate downstream prediction head was trained using the latent representation together with t_1 and the elapsed time Δt, computed from t_1 and t_2. The future-prediction head used hidden dimension 256 and dropout 0.2 and was trained with learning rate 10^{-3}.

A.4 Future-Spectrum Prediction

Future-spectrum prediction was evaluated using MSE, MAE, and spectral angle mapper (SAM). For the synthetic dataset, the prediction target was the known clean spectrum at t_2. For the human Raman dataset, where no clean ground-truth spectrum is available, the mean of the two replicate measurements at t_2 was used as a silver-standard target.

For TREDL-Spec, the two temporal representations from the replicate spectra at t_1 were averaged and passed to the temporal dynamics network together with t_1, t_2, and Δt. The two subject-stable representations were averaged separately and combined with the predicted temporal representation for decoding with $z^N = 0$. For CaRiNG and TopDis, predictions were generated separately from each t_1 replicate using each method's prediction interface and then averaged.

Table 3: Future-spectrum prediction results on the synthetic (a) and human (b) datasets. Lower values indicate better performance. Best performance within each dataset is shown in bold; second-best is shown in italics. The shaded row denotes TREDL-Spec.

	(a) Synthetic dataset			(b) Human dataset		
Model	MSE ↓	MAE ↓	SAM ↓	MSE ↓	MAE ↓	SAM ↓
CaRiNG	0.003686	0.046882	0.104096	*0.014120*	*0.082438*	*0.279343*
TopDis	*0.001936*	*0.031312*	*0.075689*	0.016112	0.088191	0.283289
TREDL-Spec	**0.000657**	**0.018792**	**0.045590**	**0.013319**	**0.079391**	**0.265584**

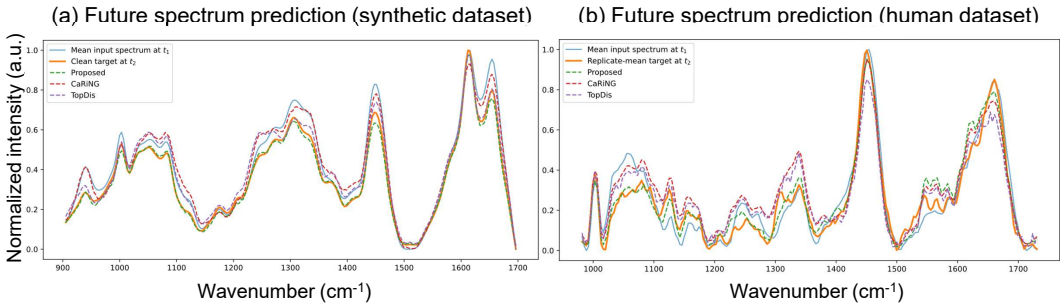


Figure 5: Future-spectrum predictions from TREDL-Spec and disentanglement baselines. (a) Representative synthetic example compared with the known clean spectrum at t_2. (b) Representative human Raman example compared with the replicate-averaged silver-standard target at t_2.

TREDL-Spec achieved the lowest future-prediction error across all three metrics on both datasets (Table 3). On the synthetic dataset, TopDis provided the second-best performance, whereas CaRiNG was the second-best model on the human data. Representative predictions from all three methods are shown in Figure 5.