SPECTRAL-TCAV FOR PAIRWISE CONCEPT-BASED INTERPRETABILITY IN RAMAN CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Testing with Concept Activation Vectors (TCAV) interprets neural network predictions in terms of human-defined concepts, but has not been applied to Raman spectroscopy, where interpretability is usually limited to individual spectral bands or perturbation responses. We introduce Spectral-TCAV, adapting TCAV to Raman spectra for bacterial species classification, and show that it recovers spectrally meaningful concepts underlying model predictions. We further extend the framework from single spectral regions to pairs of regions, decomposing their interaction into three contrasts: Sum (joint magnitude), Diff (relative dominance), and Tradeoff (antagonistic behavior, isolated by masking for opposite-sign changes). Each contrast is binarized and evaluated with independent linear CAVs and TCAV directional-derivative scores tested against null distributions from randomly sampled regions. Across classes, the same region pair can diverge in significance and direction across contrasts, showing that Sum, Diff, and Tradeoff capture complementary information beyond a single concordant-versus-discordant label.

1 INTRODUCTION

Deep learning has become the dominant approach for analyzing vibrational spectroscopy data, such as Raman spectra, supporting tasks from bacterial identification to disease diagnosis and material characterization Cai et al. (2025); Chen et al. (2024); Westermayr & Marquetand (2025). Compared with classical chemometric methods such as partial least squares or support vector machines, deep networks extract discriminative features directly from raw spectra, removing the need for manual feature engineering and typically improving performance. This gain, however, comes at the cost of transparency: a deep network's internal decision process is not directly accessible to domain experts, which is problematic in biomedical and chemical applications, where predictions must be validated against established chemical knowledge before informing clinical or analytical decisions

Explainable artificial intelligence (XAI) aims to make model predictions interpretable. While XAI has been studied extensively for images and tabular data, its adaptation to spectroscopy has received comparatively little attention, as discussed in Section 2, existing spectral XAI methods do not adequately ground their explanations in the biochemical vocabulary domain experts use to reason about spectra Contreras & Bocklitz (2025a). To address this, we introduce Spectral-TCAV, a concept-based framework that explains predictions in terms of biochemically meaningful spectral bands (Figure 1b and Table 1a), extended to jointly defined concepts over pairs of regions to capture relationships that single-region analyses overlook.

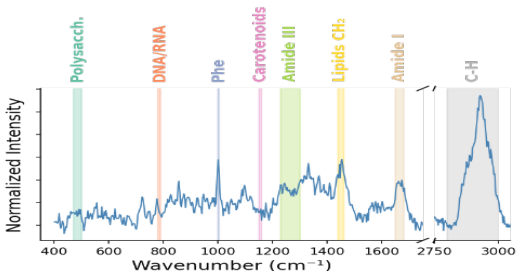
In summary, our contributions are:

- We introduce Spectral-TCAV, an adaptation of concept activation vectors to Raman spectroscopy that defines concepts as biochemically meaningful spectral bands rather than individual intensity values.
- We extend concept-based interpretability from single spectral regions to pairwise, jointly defined concepts, decomposed into three complementary contrasts—Sum, Diff, and Tradeoff—that disentangle distinct modes of interaction between spectral bands.
- We validate both single- and pairwise-concept variants on a bacterial species classification task, showing that they recover spectrally and biochemically meaningful patterns, including relationships between regions that single-band analyses overlook.

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.

Spectral concept	Wavenumbers (cm^{-1})
Polysaccharides	470–500
DNA/RNA	780–790
Phenylalanine	1000–1005
Carotenoids	1150–1160
Amide III	1230–1300
Lipids (CH_2)	1440–1460
Amide I	1650–1680
C–H stretching	2800–3000

(a)



(b)

Figure 1: (a) (b) Spectral concepts defined as user-specified spectral regions.

2 Related work

Existing approaches largely adapt methods originally developed for these domains, with SHAP Lundberg & Lee (2017), LIME Dieber & Kirrane (2020), and class activation mapping Selvaraju et al. (2017) among the most widely used, due to their model-agnostic nature and their ability to explain predictions without modifying the trained network Contreras & Bocklitz (2025a). In vibrational spectroscopy, such methods are used almost exclusively to attribute importance to individual intensity values Johnson & Tipirneni-Sajja (2024); Tokgoz et al. (2024); Ahmed et al. (2025), limiting interpretation since isolated data points often lack direct biochemical significance and produce explanations that are hard for domain experts to trust. Zone-SHAP Contreras et al. (2024b) addresses this by grouping intensities into contiguous zones that are perturbed jointly, yielding importance values tied to interpretable regions, such as characteristic peaks, rather than isolated data points.

Spectral XAI methods also rarely capture the strong correlations between neighboring wavenumbers arising from biochemical or physical properties, risking atomistic explanations that overlook collective patterns. XAI-2DCOS Contreras & Bocklitz (2025b) addresses this by repurposing two-dimensional correlation spectroscopy, organizing spectra by class probability and correlating spectral changes with model outputs to reveal which wavenumbers drive predictions and how their covariation shapes behavior. However, it relies on correlation maps over the full spectral range, computationally demanding and requiring enough data to cover the full probability range; a conditional generative adversarial network Mirza & Osindero (2014) can mitigate this by generating additional spectra, at the cost of expert configuration and unreliable results if poorly trained.

Another approach comes from attribution methods in image classification: there, low-level features, such as pixels, rarely correspond to the high-level concepts humans use to reason about predictions. Testing with Concept Activation Vectors Kim et al. (2018) (TCAV) addresses this by defining a concept through examples and training a linear classifier to separate their activations from random counterexamples at an intermediate layer; the vector orthogonal to the resulting boundary, the concept activation vector (CAV), captures the concept's direction in activation space, and a statistical test against randomly sampled concepts retains only consistently predictive ones. Other frameworks take a different approach: Concept Bottleneck Models Koh et al. (2020) and their post-hoc variant Yuksekgonul et al. (2022) constrain predictions to pass through an explicit concept layer, while Automated Concept-based Explanations Ghorbani et al. (2019) discovers concepts from data.

We adapt TCAV to Raman models by defining each concept as a contiguous spectral band tied to a user-defined biochemical vibration. Spectral CAVs thereby link model behavior to domain knowledge without modifying or retraining the network. Single-concept analyses treat each band in isolation revealing reliance on band A, but not on interactions between bands A and B, even though diagnostic information in spectroscopy is often encoded in ratios and differences between pair of peaks rather than in individual peaks. We therefore extend Spectral-TCAV to concepts defined jointly over pairs of regions, using three complementary contrasts rather than a broad concordant versus discordant split: Sum captures joint magnitude when both regions co-vary; Diff captures how consistently one band dominates the other; and Tradeoff captures cases where one band's increasing influence is offset by the other's decrease. Therefore, these contrasts give a more mechanistically informative view of how models use relationships between spectral bands.

3 METHOD

3.1 THE TCAV FRAMEWORK

We adopt the notation of Kim et al. (2018). Let l denote a fixed intermediate layer of a trained classifier and let $f_l : \mathbb{R}^d \to \mathbb{R}^m$ map an input to its activation vector at that layer. In this work, l is the global-average-pooling layer immediately preceding the classification head, denoted `gap features` in the implementation, so that d is the number of spectral wavenumbers and m is the number of feature maps at that layer. Let $h_{l,y} : \mathbb{R}^m \to \mathbb{R}$ be the function mapping this activation vector to the logit of class y, so that the full model computes $(h_{l,y} \circ f_l)(x)$ for an input spectrum x.

3.2 SPECTRAL CONCEPT DEFINITION

In the original implementation, concepts were defined as sets of examples. Here, we extend this approach by defining concepts as user-specified spectral regions. The concepts considered in this study include spectral regions associated with polysaccharides, DNA/RNA, phenylalanine, carotenoids, amide bands, lipids, and C–H stretching. Formally, a spectral concept is defined as a contiguous wavenumber band $[w_{\min}, w_{\max}]$ associated with a known biochemical vibration or user-defined band (Figure 1b and Table 1a). The per-sample concept intensity is the mean spectral intensity within that band,

$$s(x) = \frac{1}{|I|} \sum_{i \in I} x_i, \qquad I = \{\, i : w_{\min} \le w_i \le w_{\max} \,\}. \tag{1}$$

Because the input spectra are already baseline-corrected and calibrated prior to this step (Appendix A.1), this simple mean is computed on baseline-free intensities rather than confounded by broad physical artifacts such as fluorescence background; residual bleed from the tails of closely spaced adjacent bands is still possible and is not separately modeled.

3.3 CONCEPT ACTIVATION VECTOR (CAV) TRAINING

For a concept C defined as a spectral region, we derive a binary labeling function $g_C(x)$ from the spectral intensity within that region. Specifically, we first compute the median intensity of C across the dataset. We then assign an input spectrum x the value $g_C(x) = 1$ if its mean intensity within the concept region is above the dataset median and $g_C(x) = 0$ otherwise,

$$g_C(x) = 1 \quad [\, s(x) \ge \mathrm{median}(s) \,], \tag{2}$$

$$g_C(x) = 0 \quad [\, s(x) < \mathrm{median}(s) \,]. \tag{3}$$

These binary labels and the activation vectors $f_l(x)$ are then used to train a binary linear classifier that separates concept-positive and concept-negative examples in the activation space of layer l. The normal vector to the resulting separating hyperplane is defined as the concept activation vector (CAV), denoted by $v_C^l \in \mathbb{R}^m$. Thus, the CAV provides a direction in the layer-l activation space associated with the user-defined spectral concept.

Specifically, we used logistic regression as the binary linear classifier, with the LBFGS solver, a maximum of 1,000 iterations, and ℓ_2 regularization. Rather than fitting a single classifier, we repeated the procedure across $n = 5$ independent, stratified 70/30 train/validation splits. Each split yields its own decision-boundary weight vector. The final CAV is the mean of these five vectors, which are then renormalized to unit length,

$$v_C^{(l)} = \frac{\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} w_i^{(l)}}{\left\| \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} w_i^{(l)} \right\|_2}. \tag{4}$$

Here $w_i^{(l)}$ is the coefficient vector from split i; averaging across splits reduces the CAV's sensitivity to any single random partition. The reported CAV accuracy is the mean classification accuracy

held out across the five splits, indicating how locally linearly separable the concept is within the l-th layer representation. Low accuracy indicates that the concept is not well represented at that layer and that the corresponding CAV should be interpreted with caution or discarded. For the significance-testing procedure described in Sections 3.4 and 3.8, each of the N randomly sampled control regions is processed the same way, but with three rather than five splits. This reduced count keeps the cost of generating $N = 20$ null CAVs per concept comparable to that of the single true-concept CAV; since it is applied identically to every control region, it does not systematically bias the null distribution relative to the true score, though a matched split count would be preferable for maximal comparability.

3.4 TCAV Score Computation and Significance Testing

Methods of interpreting models, such as saliency maps, typically measure the sensitivity of a model's output to individual input features through gradients. In contrast, TCAV measures how sensitive a model prediction is to changes in the activation representation in the direction of a concept. Given a unit CAV v_C^l and the activation vector $f_l(x)$ at layer l, the conceptual sensitivity of the class-y logit to concept C is defined by the directional derivative

$$S_{C,y,l}(x) = \nabla h_{l,y}\big(f_l(x)\big) \cdot v_C^l. \tag{5}$$

This quantity measures whether moving the layer-l representation in the direction associated with concept C locally increases or decreases the class-y logit. Unlike feature-level attribution methods, it provides a scalar measure of sensitivity to an entire concept for a given input.

TCAV aggregates this conceptual sensitivity over a set of samples X_y belonging to class y. The TCAV score is defined as

$$\mathrm{TCAV}_{Q_{C,y,l}} = \frac{\big|\{\, x \in X_y : S_{C,y,l}(x) > 0 \,\}\big|}{|X_y|}, \tag{6}$$

where $Q_{C,y,l}$ follows the TCAV_Q notation of Kim et al. (2018). The TCAV score represents the fraction of class-y samples for which the concept has a positive directional influence on the target logit at layer l, with values ranging from 0 to 1, computed once over the full set X_y rather than per individual sample, and depends only on the sign of the directional derivative. This provides a global measure of the model's sensitivity to the concept across the selected class.

For statistical assessment, each concept's TCAV score is compared against a null distribution obtained from $N = 20$ randomly located control regions without assigned biochemical meaning. Each control region is generated with random widths (Δ) between 5 and 29 wavenumber indices and random starting positions. The resulting null scores form the null distribution. Let N' denote the number of valid null scores and let μ_{null} and σ_{null} denote their sample mean and standard deviation, respectively. The observed TCAV score is compared against this distribution using a one-sample t-test,

$$t = \big(\mathrm{TCAV}_{Q_{C,y,l}} - \mu_{\mathrm{null}}\big) \,/\, \left(\frac{\sigma_{\mathrm{null}}}{\sqrt{N'}}\right). \tag{7}$$

The resulting p-value is compared against a significance level of $\alpha = 0.05$; a concept is considered statistically significant when $p < \alpha$.

3.5 From Single Concepts to Concept Pairs

Single-concept TCAV evaluates the model's sensitivity to individual spectral regions. To assess whether the model also relies on relationships between spectral regions, we extend the analysis to concept pairs, which are defined by two user-specified wavenumber ranges, A and B. Instead of evaluating each region independently, we represent the pair by its per-sample intensities and subsequently combine them to characterize the relationship between the two regions.

Since spectral regions can differ substantially in their typical intensity scales, we standardize their intensities independently before combining them. For each sample x, let $s_A(x)$ and $s_B(x)$ denote the mean spectral intensities within regions A and B, respectively, as defined in Eq. equation 1. Each region is standardized to zero mean and unit variance across the analyzed sample set,

$$z_A(x) = \big(s_A(x) - \mu_{s_A}\big)/\sigma_{s_A}, \tag{8}$$

with an analogous definition for $z_B(x)$. This standardization places both regions on a common scale, preventing differences in their absolute intensity ranges from dominating their combination and permitting the examination of the relationship between two spectral regions from different perspectives. We note that dividing by σ_{s_A} can amplify measurement noise for concepts with intrinsically low intensity variance; the eight concepts studied here (Table 1a) all showed sufficient variance in practice, but a variance or signal-to-noise threshold prior to standardization would be advisable for concepts with near-constant intensity.

3.6 The Three Pairwise Contrasts: Sum, Diff, and Tradeoff

Instead of reducing a concept pair to a single binary label, we define three complementary contrasts: Sum, Diff, and Tradeoff. Each contrast reflects a different hypothesis about how the model combines information from regions A and B. We then treat each contrast as an individual concept and analyze it using the CAV and TCAV procedures described above.

Sum. The sum contrast measures the joint magnitude of the two regions. High values indicate samples in which both regions tend to have relatively high intensity, whereas low values indicate the opposite. It is symmetric with respect to A and B and is computed over the full sample set,

$$\mathrm{Sum}(x) = z_A(x) + z_B(x). \tag{9}$$

The resulting values are converted into binary labels using the median-split rule defined in Eqs. equation 2–equation 3.

Diff. The difference contrast quantifies the relative dominance of one region over the other,

$$\mathrm{Diff}(x) = z_A(x) - z_B(x). \tag{10}$$

Positive values indicate that region A is relatively more intense than region B, whereas negative values indicate the opposite. The contrast is computed over the full sample set and binarized using the median-split rule defined in Eqs. equation 2–equation 3.

Sum and Diff form an orthogonal reparameterization of the standardized pair,

$$z_A = \frac{\mathrm{Sum} + \mathrm{Diff}}{2}, \tag{11}$$

$$z_B = \frac{\mathrm{Sum} - \mathrm{Diff}}{2}. \tag{12}$$

However, Diff alone cannot distinguish between a simultaneous increase in both regions, with region A remaining stronger, and a genuine Tradeoff, in which region A increases while region B decreases.

Tradeoff. The Tradeoff contrast specifically tests whether the model responds to opposing changes between the two regions. Unlike Sum and Diff, it is evaluated only on samples in which the standardized intensities have opposite signs. We therefore define the Tradeoff mask as

$$M = \{\, x : \mathrm{sign}(z_A(x)) \neq \mathrm{sign}(z_B(x)) \,\}. \tag{13}$$

Within this subset, each sample is assigned to one of two complementary directions according to which region is above its mean. Thus, the two classes represent "A up, B down" and "A down, B up," with the former assigned the positive label used in CAV training. Since the Tradeoff contrast is defined by the mask itself, no additional median split is necessary. A minimum sample count of $m_{\min} = 10$ is imposed on M; pairs with fewer than $m_{\min}$ opposite-sign samples are excluded from the Tradeoff analysis to ensure reliable CAV training. We note that $m_{\min} = 10$ is a lenient floor relative to the 32-dimensional GAP activation space (Appendix A.2): with a 70/30 split this can leave as few as 7 training samples, an underdetermined regime in which perfect linear separability can arise by chance. The reported CAV accuracies (Table 2) should therefore be read as a lower bound on how well-represented a pairwise concept is, with pairs near the $m_{\min}$ threshold interpreted with additional caution.

3.7 CAV Training and TCAV Scoring for Pairwise Contrasts

CAV training and TCAV scoring follow the procedure described in Sections 3.3–3.4. For Sum and Diff, a single CAV is trained once on the full dataset X (all classes together); the resulting CAV is then used to compute a separate TCAV score for each target class y over X_y, as in Eq. equation 6. For the Tradeoff contrast, the samples are first restricted to those that satisfy the mask M (Eq. equation 13): a single CAV is trained on $X \cap M$, and the TCAV score for each class is computed over $X_y \cap M$; the corresponding precomputed layer-l activations are restricted in the same way. This procedure is applied independently to each contrast and concept pair, with the resulting CAV shared across all target classes within that contrast and pair. Thus, each combination of pair, contrast, and class produces a separate TCAV score and significance test.

3.8 Significance Testing Under Masking

For Sum and Diff, statistical significance is assessed using the null procedure described in Section 3.4, based on randomly located control regions evaluated over the full sample set. Tradeoff, however, requires a different null construction because it is evaluated only on samples exhibiting opposite-sign changes between the two regions; its null distribution applies the same masking procedure to each randomly generated control pair. For each of $N = 20$ null draws, two random control regions are sampled and processed independently using the same standardization and contrast definitions as the real concept pair. For the Tradeoff contrast, the resulting samples are filtered using the corresponding opposite-sign mask. Only draws retaining at least $m_{\min}$ samples are considered valid. A CAV is then trained on each valid masked subset using $K = 3$ splits, and the corresponding TCAV score is obtained. The observed TCAV score is compared with the resulting null distribution using the one-sample t-test defined in Eq. equation 7, with $\alpha = 0.05$. Significance testing is performed independently for each pair and contrast.

4 Results

4.1 Single Concept Attribution by Class

We first computed single-concept Spectral-TCAV scores (Section 3.4) for all eight spectral concepts (Table 1a) across the six bacterial classes, using $N = 20$ random control regions per concept for the significance test ($\alpha = 0.05$). Figure 2 shows the resulting concept-per-class TCAV score matrix. Scores range from 0 to 1, regardless of statistical significance, and take on directional meaning only when the effect passes the significance test against the null distribution from random regions. Cells corresponding to non-significant effects ($p \geq 0.05$) are left blank. For significant effects, a low TCAV score indicates that lower concept intensity is associated with an increase in the class logit, whereas a high TCAV score indicates that higher concept intensity is associated with an increase in the class logit. Table 1 summarizes, for each concept, the number of classes for which it reached significance and the corresponding CAV accuracy, which indicates how effectively each concept's high/low intensity split is separated in the GAP feature space. Forty-one of the 48 concept–class combinations (85.4%) were statistically significant.

4.2 Single Concept Class Comparisons

Lipids CH_2, Phenylalanine, Polysaccharides, and C–H stretching were significant for all six classes, whereas DNA/RNA was the least informative concept, reaching significance for only three classes (*P. stutzeri*, *R. terrigena*, and *S. cohnii*). CAV accuracies ranged from 0.792 (Lipids CH_2) to 0.887 (C–H stretching), with a mean of 0.852 across all eight concepts, above the random level (0.5) for every concept, indicating that each spectral region corresponds to a well-separated direction in the model's intermediate representation.

C–H stretching (2800–3000 cm^{-1}) was significant for every class and produced a near-binary TCAV pattern, with three classes driven by high C–H intensity (*L. innocua* $= 0.999$, *S. cohnii* $= 0.996$, *S. warneri* $= 0.999$) and the remaining three driven by low C–H intensity (*E. coli* $= 0.004$, *P. stutzeri* $= 0.000$, *R. terrigena* $= 0.000$). Sensitivity to single concepts was also uneven across classes: *R. terrigena* was significantly sensitive to all eight concepts, while *S. warneri* reached significance for

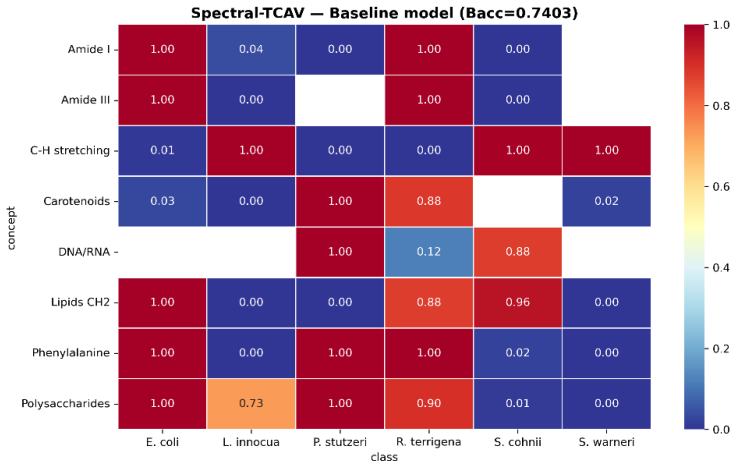


Figure 2: Single-concept Spectral-TCAV scores, for all eight spectral concepts (rows) and six bacterial classes (columns). Cells are masked (left blank) where the concept was not statistically significant for that class ($p \geq 0.05$); colored values close to 1 indicate that high concept intensity increases the class logit, and values close to 0 indicate that low concept intensity increases it.

only five of the eight, suggesting the model relies on a narrower, more specific subset of bands for this class.

Four concepts, Amide I, Phenylalanine, Polysaccharides, and (with one gap) Amide III—show high TCAV scores for *E. coli*, *P. stutzeri*, and *R. terrigena*, contrasting with the pattern observed for C–H stretching, which shows low scores for this trio and high scores for *L. innocua*, *S. cohnii*, and *S. warneri*. In contrast, DNA/RNA for *P. stutzeri* and *S. cohnii* are both significantly high (0.999 and 0.878), which may indicate they capture variation that is more species-specific than group-level. *E. coli* and *R. terrigena* differ in two places: Carotenoids, where they diverge in direction (*E. coli* $= 0.028$ low vs. *R. terrigena* $= 0.885$ high), and DNA/RNA, where *R. terrigena* shows a significantly lower TCAV score (0.121), while *E. coli* shows no significant effect (TCAV $= 0.606$), which should therefore not be interpreted as an intermediate effect.

S. cohnii and *S. warneri* show a strong dependence on high C–H stretching (0.996 and 0.999, respectively), but they differ most clearly in lipids CH_2 and DNA/RNA. The lipid CH_2 value is significantly higher for *S. cohnii* (0.961) and significantly lower for *S. warneri* (0.000). The DNA/RNA value is significantly higher for *S. cohnii* (0.878) and not significant for *S. warneri*.

We present these observations for discussion, rather than as evidence of a confirmed mechanism. As emphasized above, these analyses should be interpreted as characterizing the decision function learned by a specific model and training run rather than as unique architecture-level attributions or intrinsic data characteristics. The within-class differences illustrate the complementary information individual concepts can reveal. The pairwise Sum, Diff, and Tradeoff decomposition (Section 4.3) provides a more direct characterization of their joint contributions.

4.3 Pairwise Concept Contributions

In this section, we evaluate concepts in pairs rather than individually. For each pair of the eight spectral concepts, we determine whether the two regions reinforce each other, one dominates the other, or they exhibit a Tradeoff. The complete pairwise analysis considers all 28 concept pairs under the Sum and Diff contrasts across the full sample set and under the Tradeoff contrast within each pair's opposite-sign subset (Section 3.6). Across the six classes, this results in 504 pair $\times$ contrast $\times$ class combinations. Table 2 summarizes the statistical significance and CAV classification accuracy for each contrast.

Of the 504 combinations, 433 (85.9%) were statistically significant, which is slightly higher than the rate observed for the single-concept panel (85.4%). These rates are not corrected for multiple comparisons; with 504 (and 48, respectively) tests at $\alpha = 0.05$, some fraction of the significant

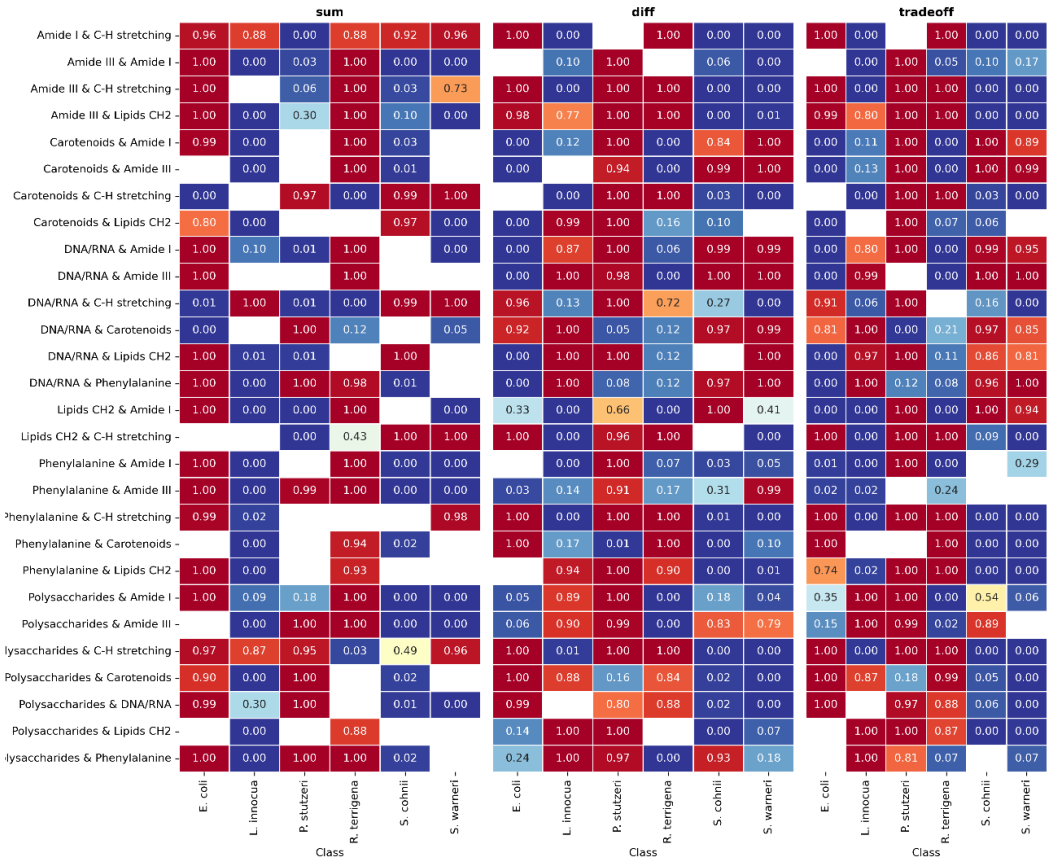


Figure 3: Pair-concept Spectral-TCAV scores, for the baseline model, for all spectral pair concepts (rows) and six bacterial classes (columns). Each block corresponds to the Sum, Diff, and Tradeoff contrasts. Cells are masked (left blank) when the concept was not statistically significant for that class ($p \geq 0.05$); values near 1 indicate that high concept intensity increases the class logit, and values near 0 indicate that low concept intensity increases it.

results is expected by chance alone, so the reported percentages should be read as descriptive of overall sensitivity rather than as family-wise error-controlled confirmatory findings. This result suggests that joint concept representations are as informative as individual spectral regions. Of the three contrasts, Diff has the highest proportion of significant results (92.9%). This suggests that for most pair–class combinations, the model was sensitive to the relative dominance of the two regions. This was true even when the model was less sensitive to their combined magnitude (Sum: 75.6%) or opposing behavior (Tradeoff: 89.3%). CAV accuracy was consistently higher for the Tradeoff contrast (mean $=$ 0.919) than for the Sum contrast (mean $=$ 0.837) or the Diff contrast (mean $=$ 0.831). This reflects the greater linear separability of the masked, opposite-sign subpopulation used to train the Tradeoff CAVs. Across the 28 concept pairs, the Tradeoff mask retained between 896 and 3,245 of the 4,324 training spectra (20.7–75.0%).

Three pairs—Amide III & Lipids CH_2, Polysaccharides & Amide I, and Polysaccharides & C–H stretching—reached significance in all 18 (contrast $\times$ class) tests, the most consistently informative pairs. As at the single-concept level (Section 4.2), *R. terrigena* showed the richest pairwise profile, reaching significance for most concept pairs across all three contrasts. Table 6 and Figure 3 show that no single contrast captures the others. Coordinated contribution (Sum), relative dominance (Diff), and opposing contributions (Tradeoff) capture distinct, only partially overlapping forms of model sensitivity to a concept pair. This motivates a closer examination of specific pairs for which the three contrasts diverge (Section A.7), summarized in Section 5: CAV-based Tradeoff comparisons chain into a transitive ordering for *S. cohnii*, and an independent perturbation-based validation (Appendix A.5) confirms the predicted directions for every case examined.

5 CONCLUSIONS

The spectral representations learned by a deep network are not necessarily unique across random initializations or training configurations, especially given the limited number of spectra available. Therefore, the TCAV scores in Figures 2–3 should be interpreted as characterizing this specific model and training run rather than as unique architecture-level attributions. Differences across similarly performing models do not necessarily indicate an error and may instead reflect genuine differences in what was learned. This suggests a potential use in model selection. Among models with comparable accuracy, concept attributions could serve as an additional criterion for whether the learned decision process aligns with domain knowledge. We note this as hypothesized rather than demonstrated since our experiments evaluate a single trained model.

We introduced Spectral-TCAV, which explains deep spectral classifiers using biochemically meaningful bands rather than individual intensities. Applied to a CNN classifying six bacterial species from Raman spectra, it recovered significant, directionally interpretable associations for 85.4% of concept–class combinations, showing that the model's decision function is organized around recognizable biochemical bands rather than arbitrary input points.

Extending the framework to concept pairs revealed that the relationship between two regions cannot be reduced to a single concordant-versus-discordant label. The Sum, Diff, and Tradeoff contrasts captured distinct and only partially overlapping aspects of pairwise sensitivity. This allowed TCAV to distinguish how a model uses a region pair rather than treating it as undifferentiated. The Phenylalanine/Amide III pair and the *S. cohnii* case study (Appendix A.7) illustrate this: regions that appear jointly important under Sum can diverge under Diff and Tradeoff. Tradeoff comparisons reveal a transitive ordering of relative importance (C–H stretching $>$ DNA/RNA $>$ lipid CH_2) that neither single-concept analysis nor Sum alone could reveal. An independent perturbation-based validation (see Appendix A.5) confirmed these contrasts are not artifacts of the linear CAV procedure: directly perturbing the input spectrum reproduced the predicted Tradeoff direction for every case examined and the Sum contrast's joint increase/decrease pattern across all pairs tested.

Two scope limitations apply. First, we did not benchmark Spectral-TCAV against other spectral or concept-based XAI methods (e.g., Zone-SHAP, XAI-2DCOS, concept bottleneck models); this is an important direction for future work. Second, although we implemented the formulation for three-band combinations, it does not scale directly to three or more bands: the number of possible combinations grows combinatorially, and the Tradeoff mask retains a shrinking sample fraction as more regions are added. For three or more bands, the practical approach is therefore to target specific trios or combinations of interest rather than exhaustively evaluating all of them.

AI USE STATEMENT

In this work, we used generative AI tools for grammar and spell checking. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The dataset and its preprocessing are described in Appendix A.1; the CNN architecture and training configuration are described in Appendix A.2. The CAV training procedure, TCAV score computation, and significance testing are fully specified in Section 3 (Eqs. equation 1–equation 7), and their extension to concept pairs in Sections 3.5–3.8. The perturbation-based behavioral validation is described in Appendix A.5, including the exact perturbation magnitudes used (Appendix A.6).

AUTHOR CONTRIBUTIONS

The author conducted the discussion and interpretation of the results, manuscript draft preparation, writing, review and editing, and preparation of tables and figures. The author has read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

[Omitted for double-blind review.]

REFERENCES

Fahad Ahmed, Naila Sammar Naz, Sunawar Khan, Ateeq Ur Rehman, Waleed M Ismael, and Muhammad Adnan Khan. Explainable artificial intelligence (xai) in medical imaging: a systematic review of techniques, applications, and challenges. *BMC medical imaging*, 26(1):37, 2026.

Md Toukir Ahmed, Md Wadud Ahmed, and Mohammed Kamruzzaman. A systematic review of explainable artificial intelligence for spectroscopic agricultural quality assessment. *Computers and Electronics in Agriculture*, 235:110354, 2025.

Nairveen Ali, Sophie Girnus, Petra Rösch, Jürgen Popp, and Thomas Bocklitz. Sample-size planning for multivariate data: a raman-spectroscopy-based example. *Analytical chemistry*, 90(21):12485–12492, 2018.

Yalu Cai, Yuduan Lin, Honghao Cai, and Hui Ni. Deep learning in vibrational spectroscopy: Benefits, limitations, and recent progress. *Journal of the Chinese Chemical Society*, 72(6):611–626, 2025.

Yuzhuo Chen, Sebastian V Pios, Maxim F Gelin, and Lipeng Chen. Accelerating molecular vibrational spectra simulations with a physically informed deep learning model. *Journal of Chemical Theory and Computation*, 20(11):4703–4710, 2024.

Jhonatan Contreras and Thomas Bocklitz. Explainable artificial intelligence for spectroscopy data: a review. *Pflügers Archiv-European Journal of Physiology*, 477(4):603–615, 2025a.

Jhonatan Contreras and Thomas Bocklitz. Xai-2dcos: Enhancing interpretability in spectral deep learning models through 2d correlation spectroscopy. *Journal of Chemometrics*, 39(7):e70045, 2025b.

Jhonatan Contreras, Sara Mostafapour, Jürgen Popp, and Thomas Bocklitz. Siamese networks for clinically relevant bacteria classification based on raman spectroscopy. *Molecules*, 29(5):1061, 2024a.

Jhonatan Contreras, Andreea Winterfeld, Jürgen Popp, and Thomas Bocklitz. Spectral zones-based shap/lime: enhancing interpretability in spectral deep learning models through grouped feature analysis. *Analytical Chemistry*, 96(39):15588–15597, 2024b.

Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Hayden Johnson and Aaryani Tipirneni-Sajja. Explainable ai to facilitate understanding of neural network-based metabolite profiling using nmr spectroscopy. *Metabolites*, 14(6):332, 2024.

Khushi Kalasampath, KN Spoorthi, Sreeparvathy Sajeev, Sahil Sarma Kuppa, Kavya Ajay, and Angulakshmi Maruthamuthu. A literature review on applications of explainable artificial intelligence (xai). *IEEE access*, 13:41111–41140, 2025.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ruihao Luo, Juergen Popp, and Thomas Bocklitz. Deep learning for raman spectroscopy: A review. *Analytica*, 3(3):287–301, 2022.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Gorkem Tokgoz, K Kubra Kirboga, Faik Ozel, Serkan Yucepur, Isa Ardahanli, and Rafig Gurbanov. Spectrochemical and explainable artificial intelligence approaches for molecular level identification of the status of critically ill patients with covid-19. *Talanta*, 279:126652, 2024.

Julia Westermayr and P Marquetand. Machine learning spectroscopy to advance computation and analysis. *Chemical Science*, 16(46):21660–21676, 2025.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Xuguang Zhou, Xiangnan Chen, Haitao Song, Xiaoyi Lv, Jin Gu, Chen Chen, and Cheng Chen. Deep learning multimodal bilinear fusion network based on vibrational spectroscopy for diagnosis of benign and malignant thyroid tumors. *Microchemical Journal*, 209:112870, 2025.

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science*, 3(11):957–964, 2023.

A Appendix

A.1 Dataset and Sample Preparation

We use the bacterial Raman dataset originally developed by Ali et al. (2018) for sample-size planning in multivariate Raman spectroscopy. The dataset comprises Raman spectra from six bacterial species: *E. coli* (DSM 423), *L. innocua* (DSM 20649), *P. stutzeri* (DSM 5190), *R. terrigena* (DSM 2687), *S. cohnii* (DSM 20261), and *S. warneri* (DSM 20316). These were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) and cultivated in nine independent biological replicates. The spectra were acquired using a Raman microscope (BioParticleExplorer, rap.ID Particle Systems GmbH) over a wavenumber range of 240–3190 cm^{-1}. After removing cosmic rays and outliers, calibrating the wavenumbers, correcting the baseline, and normalizing the spectra, each spectrum is represented by 584 wavenumber points. The dataset contains 5,420 spectra, approximately balanced across the six classes (900 spectra per class). For the reported analysis, the spectra were partitioned into training, validation, and a held-out test set, with the latter comprising one complete batch. The resulting partitions contained 4,324 training spectra, 481 validation spectra, and 607 test spectra.

A.2 CNN Architecture and Training

Since the goal of this study is to interpret a trained classifier rather than to optimize classification performance, we use a standard one-dimensional (1D) convolutional neural network (CNN) as the model under analysis. The network takes the full Raman spectrum as input and comprises four 1D convolutional blocks with 32, 64, 128, and 32 filters, respectively. Each block uses a kernel size of 7, ReLU activation, and batch normalization. This is followed by a global average pooling (GAP) layer and a dense softmax classification head for the six bacterial classes. The GAP layer serves as the intermediate representation l defined in Section 3.1, and all CAVs are trained using this representation.

The model was trained end-to-end using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy loss, and early stopping. On the held-out test set, the resulting classifier achieved an overall accuracy of 74.3%.

The CNN architecture and its hyperparameter configuration are not the focus of this study. Rather, the trained network serves as a fixed subject for interpretability analysis. In this setting, Spectral-TCAV and its pairwise extension can be applied to any differentiable classifier for which an intermediate activation representation is available. Thus, subsequent analyses focus on characterizing the individual and pairwise spectral concepts associated with the model's predictions rather than further optimizing its classification performance.

A.3 Additional results

A.4 Baseline Classification Performance

The trained CNN (Section A.2) achieved an overall balanced accuracy of 74.3% on the held-out test partition. This performance is clearly above randomness, though it may not represent state-of-the-art classification accuracy; it is sufficient for the objectives of the present study, which are not to maximize classification accuracy but to characterize the spectral concepts the model relies on, both individually and in combination, when making predictions.

A.5 Perturbation-Based Behavioral Validation

TCAV scores quantify concept sensitivity through the sign of a local directional derivative in activation space, making them local and linear in the CAV. To validate the three pairwise contrasts introduced in Section 3.6, we directly perturbed the input spectrum by controlling the intensities within a selected concept pair (A, B) and evaluated the resulting classifier outputs without using the CAV/TCAV procedure.

This analysis tests whether the behavioral effects predicted by the three pairwise contrasts are reflected in the classifier's predictions. Starting from a baseline spectrum (e.g., a class-mean spec-

Table 1: Number of classes (out of six) for which each single concept was statistically significant, and the CAV accuracy for that concept.

Spectral concept	Classes significant (of 6)	CAV accuracy
Amide I	5	0.883
Amide III	4	0.854
C–H stretching	6	0.887
Lipids CH_2	6	0.792
Phenylalanine	6	0.852
Carotenoids	5	0.813
Polysaccharides	6	0.857
DNA/RNA	3	0.870

Table 2: Significance and CAV accuracy summary across the full pairwise sweep (28 pairs $\times$ 6 classes $=$ 168 tests per contrast).

Contrast	Significant (of 168)	%	Mean CAV accuracy
Sum	127	75.6%	0.837
Diff	156	92.9%	0.831
Tradeoff	150	89.3%	0.919

trum), we generated synthetic spectra by independently increasing or decreasing the intensities within the two spectral regions that define the concept pair (A, B). Five perturbation conditions were considered (Table 3).

Table 3: Perturbation conditions used for behavioral validation.

Condition	(d_A, d_B)
Baseline	$(0, 0)$
Both up	$(+1, +1)$
Both down	$(-1, -1)$
A up / B down	$(+1, -1)$
A down / B up	$(-1, +1)$

For a replicate with sign pattern (d_A, d_B), the perturbed spectrum is

$$x'(i) = x_{\text{baseline}}(i) + d_A\,\Delta_A \cdot \mathbf{1}[i \in I_A] + d_B\,\Delta_B \cdot \mathbf{1}[i \in I_B], \tag{14}$$

where I_A, I_B are the wavenumber index sets of regions A and B, and the perturbation magnitude $\Delta_A = k \cdot \sigma_{s_A}$ is set to k standard deviations of region A across the reference dataset (analogously for Δ_B). Only the two regions defining the pair under study are modified; all other values retain baseline values, and resulting intensities are clipped to non-negative values. For each condition, n replicates are generated by adding independent Gaussian noise to the baseline spectrum before applying the perturbation, which is the source of the across-replicate variance reported in Tables 4 and 5. We note that this flat, band-limited shift introduces a sharp discontinuity at the edges of I_A and I_B rather than a smooth physical transition; while sufficient for the directional behavioral check performed here, a smoothed (e.g., Gaussian-tapered) perturbation would better preserve spectral continuity and is a natural refinement for future work.

The resulting five-condition batch was passed through the trained CNN, and the mean and standard deviation of the predicted class probability were computed per condition and per class. This provides a direct behavioral read-out of how the model's output changes as the two regions are pushed jointly up, jointly down, or in opposite directions, independent of the CAV/TCAV procedure, which we compare against the contrasts predicted by TCAV scores as a validation step.

A.6 Perturbation-Based Validation: Case Study Results

As an additional behavioral validation of the CAV-derived contrasts, we applied the perturbation protocol described in Section A.5 directly to the input spectrum for two cases that had already been analyzed in Section A.7 via CAVs: the three pairs underlying the *S. cohnii* transitive-ordering case study and the Phenylalanine and Amide III class-divergent pair. We validated each pair using a perturbation magnitude of $\Delta = 2.2$ standard deviations per region.

***S. cohnii* transitive-ordering case study.** Table 4 reports the perturbation-based validation for *S. cohnii* for the three pairs described in Section A.7 (C–H stretching > DNA/RNA > lipid CH_2).

Table 4: Perturbation-based validation for the *S. cohnii* case study pairs ($\Delta = 2.2$ SD).

Pair	Condition	Mean prob.	Std.
DNA/RNA & C–H stretching	baseline	0.964	0.029
	both up	0.960	0.027
	both down	0.797	0.127
	DNA/RNA ↑ / C–H ↓	0.904	0.070
	DNA/RNA ↓ / C–H ↑	0.945	0.036
DNA/RNA & Lipids CH_2	baseline	0.964	0.029
	both up	0.983	0.013
	both down	0.940	0.051
	DNA/RNA ↑ / Lipids CH_2 ↓	0.971	0.022
	DNA/RNA ↓ / Lipids CH_2 ↑	0.960	0.031
Lipids CH_2 & C–H stretching	baseline	0.964	0.029
	both up	0.975	0.017
	both down	0.723	0.189
	Lipids CH_2 ↑ / C–H ↓	0.913	0.054
	Lipids CH_2 ↓ / C–H ↑	0.955	0.038

All three pairs reproduce the direction predicted by the CAV-based Tradeoff contrast. DNA/RNA decrease / C–H stretching increase (0.945) exceeds DNA/RNA increase / C–H stretching decrease (0.904), matching the predicted preference for C–H stretching over DNA/RNA. DNA/RNA upward / lipids CH_2 downward (0.971) exceeds DNA/RNA downward / lipids CH_2 upward (0.960), matching the predicted preference for DNA/RNA over lipids. Lipids CH_2 down / C–H stretching up (0.955) exceeds lipids CH_2 up / C–H stretching down (0.913), matching the predicted preference for C–H stretching over lipids CH_2. The sum contrast is corroborated again for all three pairs: increasing both regions jointly raises the probability above the baseline with reduced variance. Conversely, decreasing both regions substantially decreases the probability with increased variance, consistent with Section A.7.

Phenylalanine and Amide III class-divergent pair. Table 5 reports the five perturbation conditions for Phenylalanine & Amide III.

For *R. terrigena*, the perturbation confirms the CAV-predicted Tradeoff direction (Phenylalanine decreases and Amide III increases favor the class; see Table 6): 0.869 versus 0.734. This direction remains consistent across magnitude. For *E. coli*, all five conditions fall around 0.99. The consistency across the perturbation magnitude tested suggests a ceiling effect: for this class, changes in this pair do not produce detectable changes in class probability. For *P. stutzeri*, the CAV-based Tradeoff was not significant. The largest increases are observed for both-down (0.9995) and Phenylalanine ↑ / Amide III ↓ (0.9997), with nearly identical effects. In contrast, the same Phenylalanine/Amide III pair produces the clearest decrease (0.9922) under the opposite perturbation (Phenylalanine ↓ / Amide III ↑). This pattern suggests sensitivity primarily to the absolute level of Amide III rather than a genuine Phenylalanine-versus-Amide III opposition, consistent with the ambiguity reported for this class in Section 4.3.

Table 5: Perturbation-based validation for Phenylalanine & Amide III ($\Delta = 2.2$ SD).

Class	Condition	Mean prob.	Std.
E. coli	baseline	0.99983	0.0001
	both up	0.9995	0.00027
	both down	0.9997	0.00017
	Phenylalanine ↑ / Amide III ↓	0.99983	0.00008
	Phenylalanine ↓ / Amide III ↑	0.9992	0.00045
P. stutzeri	baseline	0.997	0.0017
	both up	0.9963	0.0022
	both down	0.9995	0.0003
	Phenylalanine ↑ / Amide III ↓	0.9997	0.0002
	Phenylalanine ↓ / Amide III ↑	0.9922	0.0045
R. terrigena	baseline	0.767	0.117
	both up	0.907	0.061
	both down	0.695	0.112
	Phenylalanine ↑ / Amide III ↓	0.734	0.121
	Phenylalanine ↓ / Amide III ↑	0.869	0.066

A.7 Case Study: Trade-off Analysis

***S. cohnii* transitive-ordering case study.** We examined the three informative single-concept combinations identified in Section 4.2 (DNA/RNA, C–H stretching, and lipid CH_2), using the Sum, Diff, and Tradeoff contrasts introduced in Section 3.6.

The Sum contrast was highly significant for all three pairs (Sum score ≥ 0.99, $p < 1 \times 10^{-5}$, $n = 4324$), indicating that the two regions within each pair tend to vary in the same direction. This is consistent with the single-concept results, in which increased intensity in each of the three regions was associated with higher class logits. However, the sum contrast alone cannot determine whether this joint variation reflects redundant or co-varying information, nor can it determine whether one region contributes more strongly than the other.

The Tradeoff contrast, evaluated on samples in which the two regions varied in opposite directions, provided additional insight into their relative contributions. For DNA/RNA and C–H stretching, the dominance contrast revealed a significant preference for C–H stretching over DNA/RNA (Diff score $= 0.27$, $p = 2.1 \times 10^{-2}$). Similarly, the Tradeoff contrast revealed that among samples with opposite-sign changes ($n = 2{,}590$), a decrease in DNA/RNA accompanied by an increase in C–H stretching favored the class (score $= 0.16$, $p = 7.2 \times 10^{-3}$). For DNA/RNA and lipid CH_2, the dominance contrast was not significant. Nevertheless, the Tradeoff contrast revealed a strong, highly significant directional effect (score $= 0.86$, $p = 5.6 \times 10^{-7}$, $n = 2{,}570$), favoring the class with increasing DNA/RNA and decreasing lipid CH_2. Similarly, no significant dominance effect was observed for lipid CH_2 versus C–H stretching. However, the Tradeoff contrast revealed that decreasing lipid CH_2 while increasing C–H stretching favored the class within the opposite-sign subset ($n = 2{,}488$) (score $= 0.09$, $p = 2.0 \times 10^{-2}$).

The three Tradeoff directions are mutually consistent and establish a transitive ordering of relative contributions: C–H stretching > DNA/RNA > lipid CH_2. Thus, when two regions vary in opposite directions, the model consistently favors C–H stretching over both DNA/RNA and lipid CH_2, as well as DNA/RNA over lipid CH_2. These results suggest that, beyond the shared global co-variation captured by the sum contrast, the model is sensitive to a compositional axis along which the three regions compete in their relative contribution to class prediction. Therefore, C–H stretching emerges as the strongest contributor among the three concepts, while lipid CH_2 contributes the least.

Phenylalanine and Amide III class-divergent pair. Section 4.2 noted that Phenylalanine and Amide III take significantly high values for *E. coli*, *P. stutzeri*, and *R. terrigena*, alongside Amide I and Polysaccharides. The pairwise decomposition of Phenylalanine & Amide III shows that this

apparent three-way similarity conceals two distinct relational patterns (Table 6). Single-concept Spectral-TCAV, and even the sum contrast on its own, would have missed this distinction entirely.

Table 6: Sum, diff, and Tradeoff TCAV scores for Phenylalanine & Amide III, for the three classes with significantly high single-concept scores for both bands (Section 4.2).

Class	**Contrast**	**TCAV score**	p**-value**	n	**Reading**
E. coli	Sum	1.000	0.0150	4324	both rise together
	Diff	0.032	0.0002	4324	Amide III dominates
	Tradeoff	0.021	0.0001	1183	confirmed: Phenylalanine ↓ / Amide III ↑
P. stutzeri	Sum	0.990	0.0095	4324	both rise together
	Diff	0.914	0.0018	4324	Phenylalanine dominates
	Tradeoff	0.669	0.2103	1183	not significant
R. terrigena	Sum	1.000	0.0029	4324	both rise together
	Diff	0.167	0.0087	4324	Amide III dominates
	Tradeoff	0.241	0.0001	1183	confirmed: Phenylalanine ↓ / Amide III ↑

For all three classes, the sum is significant and close to 1, replicating the joint high-intensity pattern already visible for each concept at the pair level. However, Diff and Tradeoff split the trio into two groups. For *E. coli* and *R. terrigena*, Diff shows that Amide III dominates Phenylalanine (0.032 and 0.167, respectively). Tradeoff confirms an antagonistic relationship in the 1183 opposite-sign samples: Phenylalanine decreases while Amide III increases (*E. coli*: TCAV $= 0.021$, $p = 0.0001$; *R. terrigena*: TCAV $= 0.241$, $p = 0.0001$). For *P. stutzeri*, Diff points in the opposite direction, indicating that Phenylalanine dominates over Amide III (TCAV $= 0.914$). Critically, Tradeoff is not significant on the same subset of opposite-sign samples (TCAV $= 0.669$, $p = 0.210$).

As anticipated in Section 3.6, Diff alone cannot distinguish between a simultaneous rise in both regions, with one region leading, and a genuine Tradeoff. *P. stutzeri* is this ambiguous case precisely: the two regions rise together, and Phenylalanine tends to be relatively more intense. However, when samples are restricted to those in which the two regions actually move in opposite directions, no reliable directional effect remains. For this class, the model's sensitivity to Phenylalanine dominance does not descend from a confirmed opposing relationship between the two peaks, unlike with *E. coli* and *R. terrigena*, where the Tradeoff is verified independently.