

RamanPFN: learning from Raman spectral structure with a tabular foundation model

Xingyu Pan^{1,2}, Huan Wang^{2,*}, Jinjia Guo²,
Zhenlin Zhao², Siming Dong², Jixi Lu¹

¹Beihang University ²Cleer Science

*Correspondence: wh.2021@tsinghua.org.cn

panxingyu2003@buaa.edu.cn, wh.2021@tsinghua.org.cn

guojj.jackey2024@gdhfi.com, zhenlinzhao@cleerlab.com

william.d@cleerlab.com, lujixi@buaa.edu.cn



北京航空航天大学
BEIHANG UNIVERSITY



清析智研
Cleer Science

Abstract

Raman spectroscopy enables non-destructive, label-free molecular characterization across materials science, biomedicine and process monitoring. Predictive Raman datasets often contain few labelled spectra and thousands of ordered wavenumbers, with informative variation within bands and across distant spectral regions. Latent-variable chemometrics accommodates collinear small-sample data but can obscure fine peak morphology, whereas deep spectral networks resolve this structure only after task-specific training. TabPFN avoids task-specific parameter fitting through pretrained in-context inference, but processes very wide inputs as feature-subsampled views that do not preserve joint visibility of related bands. We present RamanPFN, a spectral representation framework that encodes these dependencies before TabPFN inference. Global Compositional Unmixing constructs non-negative coordinates over the complete spectrum so that distant bands with shared latent variation occupy a common predictive axis. Local Vibrational Subspace Encoding represents contiguous wavenumber regions with multiple orthogonal modes that retain independent changes in peak shape, intensity and position. The representations are evaluated separately and combined at the prediction level. Evaluation covered 150 tasks from 74 public Raman datasets. RamanPFN reduced root-mean-square error by 19.6% on average across 129 regression targets relative to direct TabPFN inference and further reduced the remaining classification error by 9.0% across 21 classification tasks. These results establish explicit spectral representation as an effective interface between high-dimensional Raman measurements and reusable tabular inference.

Introduction

Raman spectroscopy has become a general analytical readout of chemical composition across materials, living systems and industrial processes. Applications include nanoscale materials characterization, high-speed biochemical imaging, disease diagnosis and label-free mapping of cells and tissues [1–5]. Advances in acquisition speed and instrument miniaturization continue to widen its experimental reach [2, 6]. The molecular specificity that gives Raman spectroscopy this reach also

makes its data structurally demanding: peak positions report vibrational modes, while peak shapes and relative intensities reflect chemical state, concentration and mixture composition [7, 8]. The relevant information is distributed over two physical scales. Neighbouring wavenumbers define a band’s envelope, shoulders and small shifts, while one chemical component can generate coordinated bands far apart on the spectral axis. Predictive Raman analysis must therefore read local vibrational morphology and full-spectrum composition together. This requirement is most difficult in the common high-dimensional, small-sample setting, where thousands of channels are measured for tens to hundreds of labelled spectra. The 74 public Raman datasets assembled in RamanBench show that this high-dimensional, small-sample regime is characteristic of modern Raman prediction [9].

Raman modelling has evolved by changing how this spectral structure is represented. Partial least-squares regression and discrimination reduce correlated channels to supervised latent directions and remain competitive when data are scarce [10, 11]. Spectral unmixing makes mixture structure more explicit by separating full-spectrum variation into components and abundance-like coordinates, often under physically motivated constraints [12–14]. These compact global representations are efficient in small datasets, although they may under-resolve nonlinear peak morphology. Deep spectral networks expanded the representation itself, learning nonlinear and local patterns for molecular identification, imaging and diagnosis [4, 7, 15–18]. Self-supervised learning has further reduced dependence on manual preprocessing and annotations [19]. These richer representations are usually learned again for each dataset, where limited labels make it harder to separate chemistry from instrument and sample variation.

Tabular foundation models offer a different route by acquiring the learning algorithm during pretraining. TabPFN performs in-context prediction on a new dataset, and large-scale comparisons place it among the strongest small-data learners [20–22]. TabPFN- and TabICL-family models are already among the strongest methods for Raman prediction, even though most regression datasets and all classification datasets in the collection exceed TabPFN’s recommended feature range [9, 23]. The same interface that enables transfer across arbitrary tables also creates a mismatch with Raman spectra. Generic tabular models treat columns as exchangeable variables; when feature counts exceed the model’s input capacity, spectra must be divided among feature subsets, compressed or addressed through further pretraining [24–29]. Across several forward passes, every wavenumber may reach the model while only fragments of a peak or a distributed molecular signature are seen together. Near-infrared calibration studies identify the same need for spectroscopy-specific structure when tabular foundation models operate on ordered measurements [30]. For high-dimensional Raman spectra, coverage is not context.

The physical organization of Raman spectra determines which channels should be interpreted together. A compositional representation must span the complete axis because one latent source can coordinate distant bands. A vibrational representation must remain local because peak envelopes, neighbouring-band ratios and small shifts derive their meaning from spectral adjacency. These scales are related but not interchangeable. Foundation models in other scientific domains make such organizing variables explicit. Materials models encode atomic geometry and interactions [31–33], forecasting models retain temporal order and scale [34, 35], and Earth-observation models combine spectral and spatial structure [36–38]. Related work has adapted tabular foundation models through domain-aware front ends, compact descriptors and ordered tokenization [29, 30, 39, 40]. For Raman prediction, a full-spectrum decomposition can recover composition while smoothing local peak morphology, whereas a local encoding can preserve vibrational detail while leaving distant bands disconnected. Their complementarity motivates two representations that meet only at prediction.

Here we introduce RamanPFN, a dual-scale spectral framework that implements this representation before tabular foundation-model inference. The complete workflow in Figure 1 first forms two coordinate systems from each spectrum. Global Compositional Unmixing follows intermediate

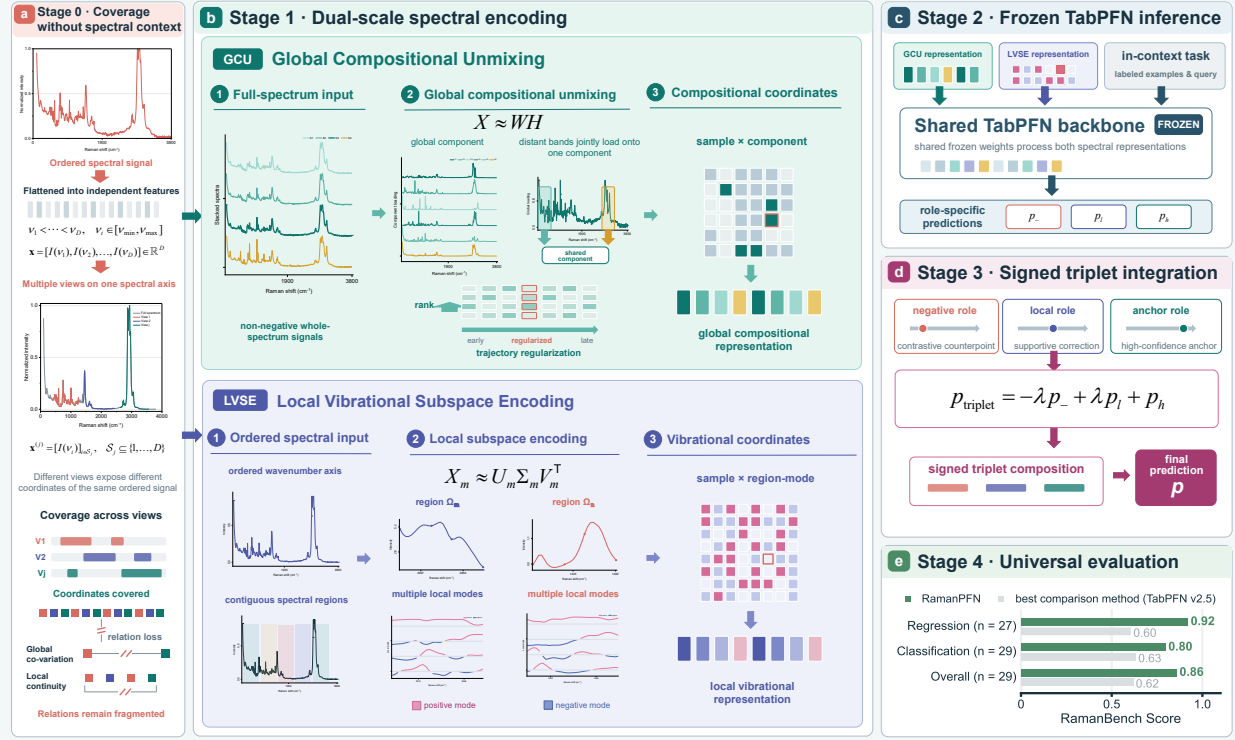


Figure 1: **RamanPFN constructs a dual-scale spectral interface between Raman spectra and a tabular foundation model.** **a**, Generic feature coverage exposes multiple views of an ordered spectrum but fragments global co-variation and local continuity. **b**, Global Compositional Unmixing (GCU) encodes full-spectrum compositional structure, whereas Local Vibrational Subspace Encoding (LVSE) preserves multiple modes of variation within contiguous wavenumber regions. **c**, Both representations are processed by the same TabPFN model to produce role-specific predictions. **d**, Training OOF evidence and a fixed prediction-space policy assign the negative, local and anchor roles, which signed triplet integration combines into the final prediction. **e**, The framework is evaluated across regression, classification and overall performance.

states of a non-negative full-spectrum factorization, producing compact abundance-like coordinates in which separated bands can contribute to the same axis. Local Vibrational Subspace Encoding partitions the ordered wavenumber axis and retains several singular directions within each region to preserve independent local variation. The same foundation model evaluates these coordinate systems in separate forward passes, and signed triplet integration combines the resulting predictions while keeping full-spectrum and interval-specific evidence distinct until aggregation.

Mechanistic visualization shows that GCU components span the complete spectral axis, whereas LVSE modes remain confined to contiguous intervals. Task-level ablation identifies targets improved by both branches and by each branch alone. Across the 150 tasks, regression and classification followed the same configuration ordering. This ordering persisted across continuous and categorical prediction, consistent with a representation-level effect.

Results

Dual-scale spectral representation in RamanPFN

RamanPFN changes what constitutes a feature for TabPFN. Instead of passing wavenumber intensities through partially overlapping 500-feature subsets, it recasts each spectrum as full-spectrum compositional coordinates and multi-mode encodings of contiguous spectral regions. A shared foundation-model predictor evaluates both representations, after which a signed integration rule combines their predictions into a single output. Figure 1 shows how this workflow brings global covariation and local peak morphology into the final prediction.

Global Compositional Unmixing (GCU) constructs the global representation from a non-negative matrix factorization of the training spectra. Each spectrum is projected onto a compact set of abundance-like coordinates derived from basis components that span the complete wavenumber axis. GCU uses the factorization trajectory as an implicit regularizer, controlling how far the representation progresses from dominant shared structure towards finer spectral variation. Distant bands driven by the same underlying variation can thus contribute to one coordinate, even when they would rarely occur together within a 500-feature subset.

Local Vibrational Subspace Encoding (LVSE) constructs the local representation by dividing the ordered wavenumber axis into contiguous regions and applying a separate singular value decomposition within each region. It retains several leading modes rather than reducing the region to a pooled intensity or a single dominant direction. The resulting coordinates encode changes in peak intensity, shoulder shape, neighbouring-band ratios and small spectral shifts while remaining tied to a defined spectral interval. This interval-specific encoding preserves local peak morphology without mixing information from unrelated parts of the spectrum.

After encoding, TabPFN produces role-specific predictions from the two representations. Signed triplet integration combines the anchor, local and negative roles as $p_{\text{triplet}} = -\lambda p_- + \lambda p_l + p_h$. The anchor supplies the main estimate, while $\lambda(p_l - p_-)$ adjusts it according to the contrast between the local and negative roles.

Complementary global and local spectral structure

TabPFN caps each forward pass at 500 features and, for wider inputs, increases the estimator count to a maximum of 32 so that every wavenumber is nominally covered at least once [20]. Even before this cap is reached, the number of available feature slots grows only linearly with spectrum width, whereas the numbers of possible wavenumber pairs and triplets grow quadratically and cubically. Marginal and joint coverage are both complete for spectra containing only a few hundred channels, but they progressively diverge as dimensionality increases. We quantified this divergence using TabPFN’s balanced feature-subsampling procedure across all 129 regression tasks, estimating the fraction of randomly selected pairs and triplets that appeared together in at least one estimator-specific forward pass. Figure 2a shows that joint visibility declined monotonically with spectral width despite complete marginal coverage.

The separation was most pronounced among the five widest tasks, which contained 11,567 to 11,689 channels. Only about 4.4% of random pairs and 0.19% of random triplets ever appeared together in an estimator-specific forward pass. At the median width of 1,901 channels, the corresponding fractions were 45.4% and 13.6%, while tasks with approximately 600 channels or fewer retained full co-occurrence. Complete feature coverage therefore did not preserve a common inference context for combinations of wavenumbers in the widest spectra examined.

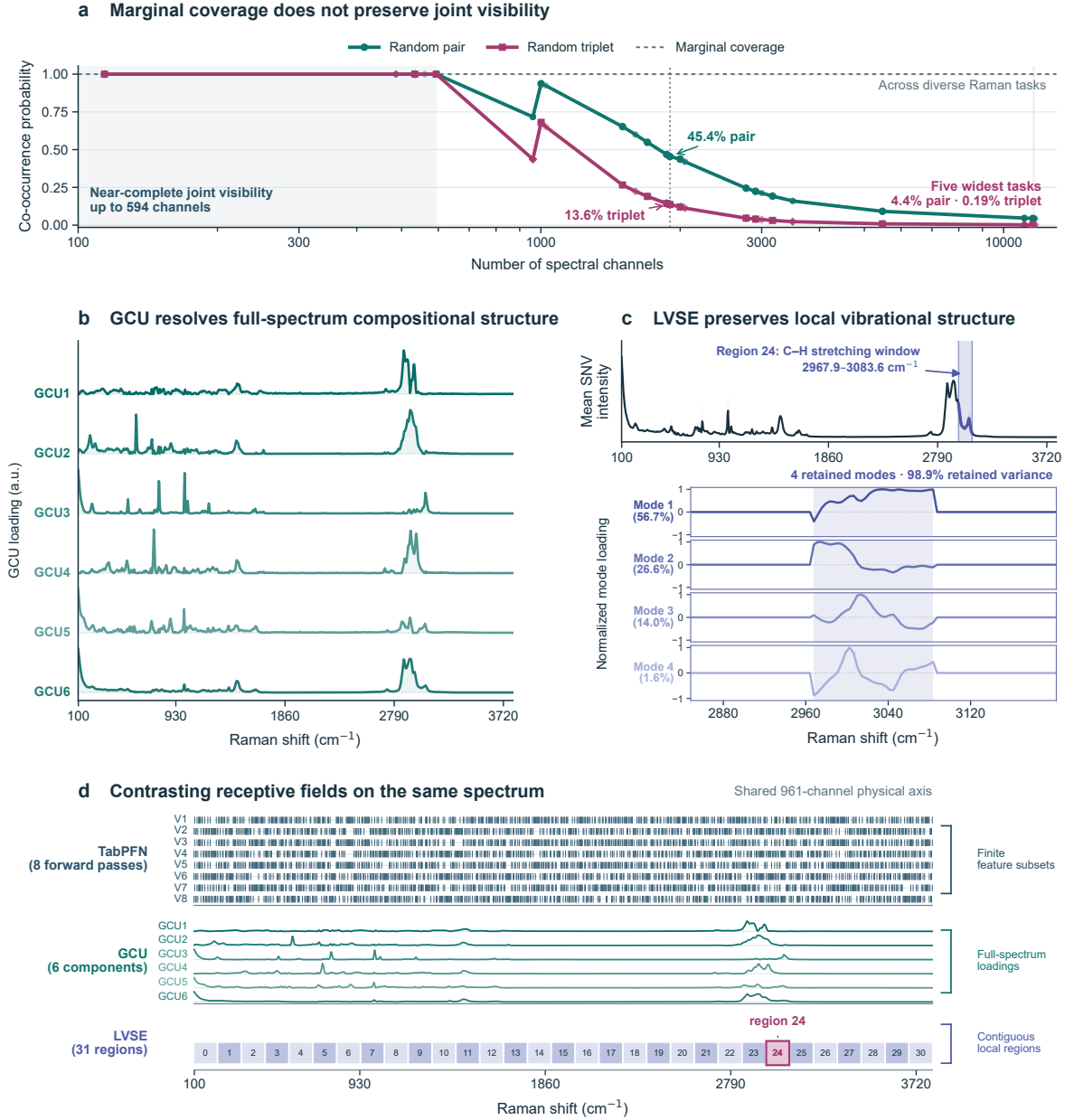


Figure 2: GCU and LVSE restore complementary spectral structure that generic feature subsampling obscures. **a**, Fractions of randomly sampled channel pairs and triplets that appeared together in at least one TabPFN forward pass across 129 regression tasks; marginal coverage remains complete as joint visibility declines with spectral width. **b**, Six high-variance GCU components for a 961-channel fuel-analysis task, each spanning the complete Raman-shift axis. **c**, Four LVSE modes within region 24, a 31-channel C–H stretching window spanning 2967.9–3083.6 cm^{-1} and retaining 98.9% of the regional variance. The configured 32-way partition realizes 31 non-empty regions. **d**, Support maps for eight TabPFN forward-pass feature subsets, six full-axis GCU components and 31 contiguous LVSE regions on the same spectrum.

To examine the global structure retained by GCU, we visualized the six components with the largest coordinate variance in the benchtop fuel task for ethyl tert-butyl ether content. The task contains 961 spectral channels, and the components placed their strongest weights in different regions while remaining defined across the complete axis. Figure 2b shows how their associated sample coordinates integrated variation from separated spectral regions without requiring those regions to occupy the same 500-feature subset. Because the decomposition was learned from the training spectra without predefined endmember annotations, the components represent data-driven axes of compositional variation rather than assigned chemical species.

To examine what LVSE preserves within a single spectral region, we visualized the four modes retained in the region with the highest retained variance among the 31 contiguous regions of this spectrum. This 31-channel region spans approximately 2,967.9 to 3,083.6 cm^{-1} within the C–H stretching envelope, and the four modes together captured 98.9% of its variance. As shown in Figure 2c, the leading mode followed a broad envelope across the interval, whereas the higher modes introduced additional sign changes and finer structure.

Figure 2d places the three support patterns on the same 961-channel axis and makes their differences explicit. The eight TabPFN forward passes contained finite, interleaved feature subsets, each GCU component remained defined across the whole spectrum, and LVSE organized the axis into 31 contiguous local regions. GCU could therefore couple distant bands through a shared coordinate, while LVSE retained several independent directions within a single vibrational window. Together, the two representations supplied full-spectrum coupling and local multi-mode resolution without requiring one 500-feature forward pass to preserve both.

Regression performance across diverse Raman datasets

The central regression test was whether one dual-scale spectral framework could remain effective across the diversity of Raman analysis. The 129 targets came from 53 public datasets [9], spanning bioprocess monitoring, fermentation, fuels, organic acids, microgels and material characterization. Training sets ranged from 6 to 6,203 spectra, while spectral width ranged from 114 to 11,689 wavenumbers. The median task contained 143 training spectra but 1,901 spectral variables. Table 1 summarizes a landscape that combines limited supervision with large changes in spectral dimensionality and scientific purpose.

Table 1: **Scientific coverage and evaluation geometry of the RamanBench dataset collection.** Panel a summarizes the four application domains using the published RamanBench metadata; evaluated targets correspond to the 129 regression targets and 21 classification tasks used in this study. Panel b reports the observed ranges and medians over the published task splits.

a Scientific coverage by application domain					
Application domain	Representative coverage	Datasets	Regression targets	Classification tasks	Spectra
Materials science	MLROD, RRUFF minerals, organic pigments and weathered microplastics	4	1	3	131,625
Biological & biotechnological	Bioprocess monitoring, fermentations, cancer cells and mutant wheat	24	60	4	67,481
Medical & clinical	Serum and saliva diagnostics, pathogenic bacteria, skin and tissue studies	13	0	13	107,367
Chemical & industrial	Fuels, organic acids, microgels, sugar mixtures and hair dyes	33	68	1	19,195
Total	74 public Raman datasets across four application domains	74	129	21	325,668

b Task and spectral geometry										
Task family	Datasets	Tasks	Training spectra		Test spectra		Wavenumbers		Output space	
			Range	Median	Range	Median	Range	Median	Range	Median
Regression	53	129	6–6,203	143	2–1,668	36	114–11,689	1,901	Continuous	
Classification	21	21	16–104,048	929	4–26,013	233	724–3,276	1,340	2–79 classes	3 classes

Across this landscape, RamanPFN lowered RMSE relative to the TabPFN reference that reads the original spectral channels directly. The mean task-level reduction was 19.6%, with a 95% task-cluster bootstrap interval of 15.8% to 23.5%. Improvements were observed for 109 of 129 targets and 49 of 53 source datasets after averaging the three repetitions. The ordered distribution in Fig. 3a shows that the gain extended across most targets rather than arising from a small group of favourable targets. The repetition-specific ranges remained predominantly below zero, and the median target-level reduction was 13.8%.

External performance was evaluated against the complete regression comparison pool of 26 independently rerun baselines. For each target, we identified the lowest three-repetition mean RMSE achieved by any baseline in this pool. Figure 3b shows that RamanPFN remained lower on 93 of 129 targets, even though the best comparison result was allowed to come from a different method for every target.

RamanPFN also achieved the best result across all six regression measures in Table 2, reaching a Score of 0.920 and an Elo of 1,808. The closest alternatives depended on the metric: TabPFN v2.5 attained the highest baseline Score at 0.605, whereas TabPFN v2 attained the highest baseline Elo at 1,488 and the lowest baseline mean RMSE at 9.574. RamanPFN’s advantage persisted after aggregation at the source-dataset level. It showed a mean symmetric RMSE advantage of 25.2% over TabPFN v2.5 and 28.2% to 57.2% over the remaining 25 baselines. Figure 3c shows that all 26 prespecified contrasts remained significant after Holm correction, with adjusted $P \leq 1.43 \times 10^{-8}$. The agreement among target-level error, best-result comparison and source-dataset-level inference establishes the breadth of the regression gain.

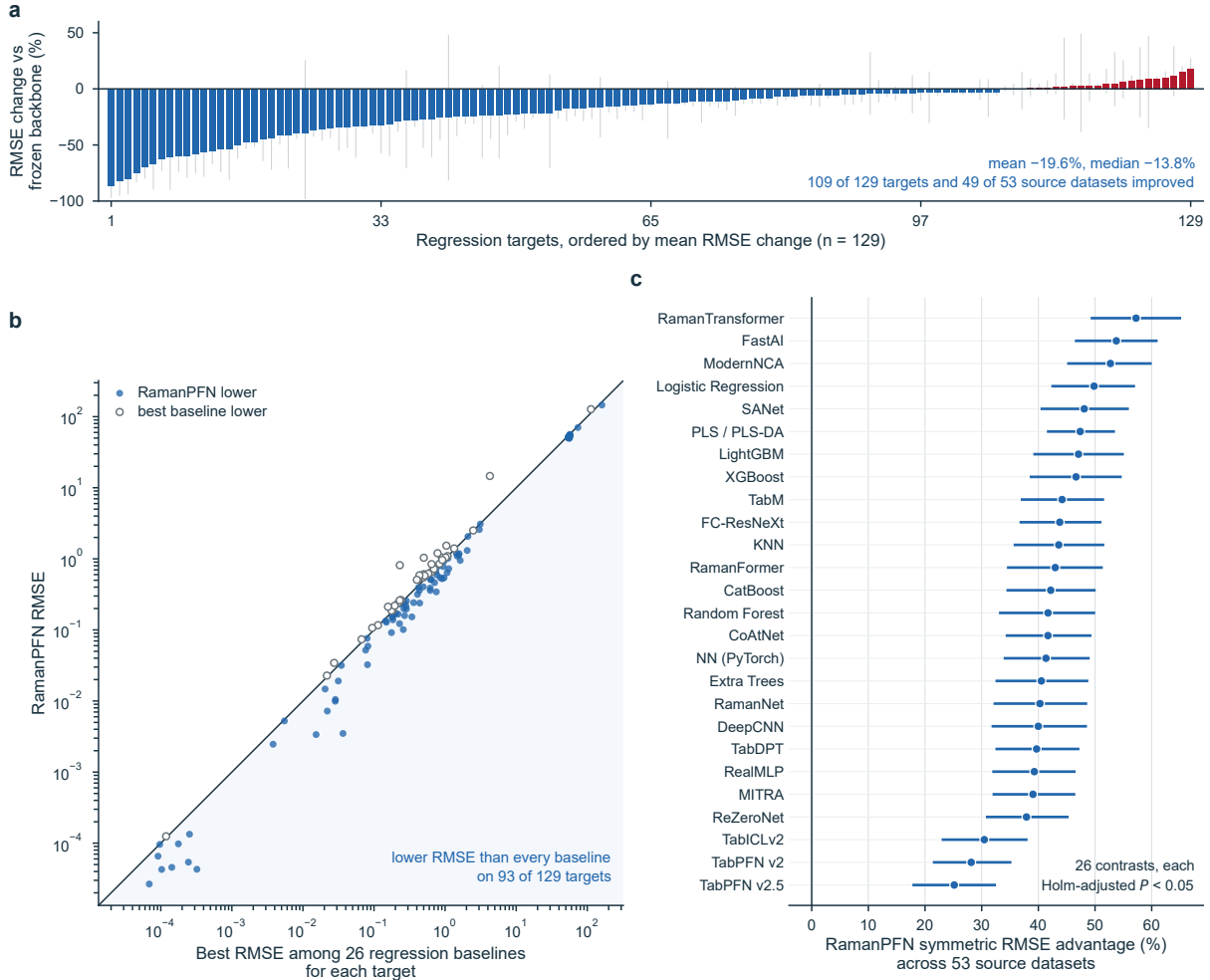


Figure 3: Breadth and statistical robustness of RamanPFN regression performance.

a, Task-level RMSE change relative to the TabPFN reference across 129 regression targets from 53 source datasets. Bars show the mean of three repetitions, and grey vertical lines span the corresponding repetition-specific values. Targets are ordered by mean change; blue indicates lower RMSE and red indicates higher RMSE. **b**, Three-repetition mean RMSE for RamanPFN against the lowest RMSE achieved by any of the 26 regression baselines on the same target. Each point denotes one target on logarithmic axes. The diagonal marks equal error; blue points indicate lower RMSE for RamanPFN and open grey points indicate a lower RMSE among the comparison methods. **c**, Source-dataset-level symmetric RMSE advantage of RamanPFN relative to each of the 26 regression baselines. Points show the mean advantage after target-level effects were averaged within each of the 53 source datasets; horizontal lines show percentile 95% bootstrap intervals for that mean. Two-sided Wilcoxon signed-rank tests use the Pratt treatment of zeros, with Holm correction across the 26 prespecified RamanPFN–baseline contrasts.

Table 2: **RamanPFN performance across diverse prediction paradigms.** RamanPFN is compared with 26 methods evaluated under the same evaluation protocol across 129 regression tasks. RMSE, MAE, gap to best and average rank are lower-is-better; gap to best is the target-wise mean of $100(1 - \text{RMSE}_{\text{best}}/\text{RMSE}_m)$. Score and Elo are higher-is-better and follow the RamanBench scoring protocol over the full comparison set. Values aggregate the three repetitions. Bold type marks the best value in each column; bold underlining marks the second best.

Method	Predictive error			Aggregate comparison		
	RMSE↓	MAE↓	Gap to best↓	Average rank↓	Score↑	Elo↑
<i>Dual-scale spectral framework</i>						
RamanPFN	8.869	6.673	4.4	2.03	0.9202	1808
<i>Tabular foundation models</i>						
TabPFN v2	<u>9.574</u>	7.268	33.5	<u>5.36</u>	0.5924	<u>1488</u>
TabPFN v2.5	9.621	<u>7.227</u>	<u>30.8</u>	5.40	<u>0.6046</u>	1,482
TabICL v2	10.141	7.642	36.9	7.19	0.4926	1,386
MITRA	11.049	8.628	45.4	11.29	0.3263	1,185
TabDPT	10.997	8.388	47.5	12.55	0.2310	1,122
<i>Raman-specific neural models</i>						
ReZeroNet	11.224	8.675	44.7	11.22	0.2858	1,178
DeepCNN	11.307	8.703	46.3	12.94	0.2339	1,102
RamanNet	10.869	8.280	49.0	13.12	0.1828	1,098
RamanFormer	10.964	8.430	50.4	14.54	0.2241	1,026
SANet	12.625	10.063	57.9	18.17	0.0958	900
RamanTransformer	18.204	15.952	70.7	24.36	0.0055	563
<i>Deep tabular and neural models</i>						
RealMLP	11.198	8.465	47.8	12.95	0.2136	1,146
NN (PyTorch)	10.750	8.221	49.7	13.62	0.1580	1,103
TabM	10.812	8.270	52.1	14.25	0.1131	1,088
FC-ResNeXt	11.578	8.804	51.2	14.96	0.1668	1,049
CoAtNet	11.543	8.819	50.9	14.70	0.1509	1,045
FastAI	12.573	9.828	60.9	20.84	0.0145	786
ModernNCA	13.308	10.121	62.0	21.63	0.0207	763
<i>Tree ensembles and gradient boosting</i>						
CatBoost	10.766	8.193	52.9	14.47	0.0823	1,079
Extra Trees	10.330	7.873	52.3	14.30	0.1406	1,070
Random Forest	10.517	8.008	54.3	16.05	0.1098	1,000
LightGBM	11.385	8.781	56.8	18.12	0.0422	911
XGBoost	11.083	8.372	57.4	18.51	0.0669	899
<i>Classical machine learning and chemometrics</i>						
Linear regression	18.047	13.536	54.0	14.51	0.2034	1,084
PLS	13.242	10.139	52.8	15.11	0.1750	1,062
KNN	10.951	8.032	53.2	15.80	0.1394	1,022

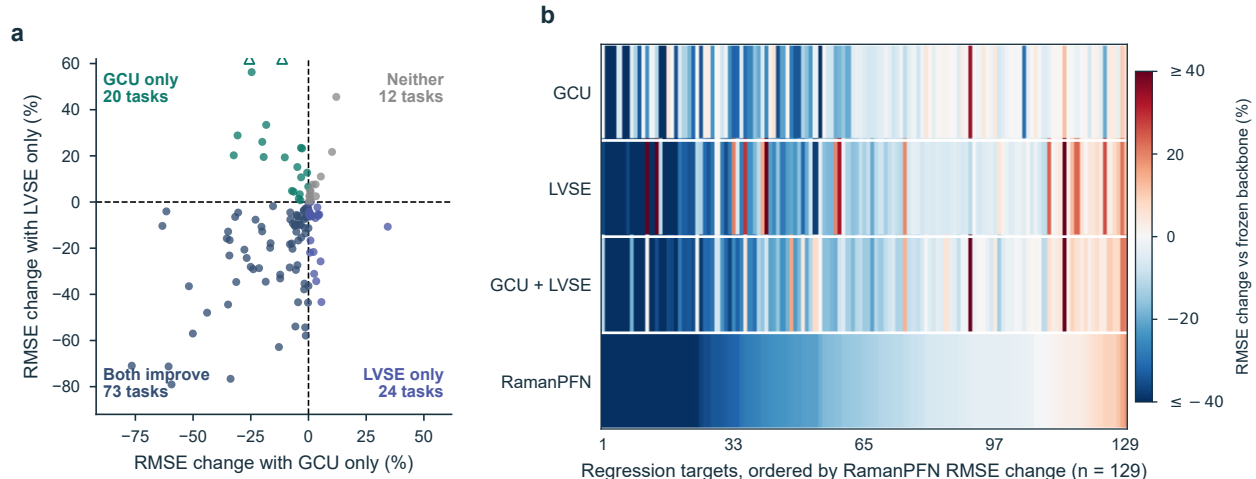


Figure 4: **Complementary task-level contributions of GCU and LVSE.** **a**, Mean RMSE change produced by GCU alone and LVSE alone for each of the 129 regression targets across three repetitions. Negative values indicate lower prediction error relative to the TabPFN reference. Point colours identify targets improved by both representations, by GCU alone, by LVSE alone or by neither representation; dashed lines mark zero change. Open triangles denote two targets outside the displayed plotting range. **b**, Task-level RMSE changes for GCU, LVSE, their direct combination and complete RamanPFN relative to the same reference. Columns preserve the target order defined by complete RamanPFN. Blue indicates lower RMSE and red indicates higher RMSE. The colour scale is saturated at $\pm 40\%$; 63 of 516 cells extend beyond these limits.

Independent and joint contributions of GCU and LVSE

The method comparison established that RamanPFN performs consistently across diverse Raman datasets, but did not reveal whether its advantage arose from one dominant representation or from the coexistence of two physical views. We therefore evaluated five configurations comprising the TabPFN reference, GCU alone, LVSE alone, both representations together and the complete RamanPFN framework. Every configuration was assessed on the same 129 regression targets under three repetitions. Figure 4 and Table 3 use this performance sequence to test the internal logic of the framework directly.

GCU and LVSE each produced a substantial gain in isolation. GCU reduced mean RMSE by 10.2%, while LVSE reduced it by 10.7%, and both bootstrap intervals excluded zero. Figure 4a resolves the different task profiles beneath these similar averages: 73 targets improved with both representations, 20 improved only with GCU and 24 improved only with LVSE, leaving just 12 targets without a gain from either branch. The global and local representations therefore recovered different task-relevant information from the same spectra.

The task-level progression in Fig. 4b follows these complementary responses through the full framework. Directly combining GCU and LVSE increased the mean RMSE reduction to 13.3%, exceeding either representation alone. Complete RamanPFN further extended the reduction to 19.6%. The increasingly coherent blue pattern across the final two rows shows that the complete framework broadens the gain across the regression targets.

Table 3: **Regression mechanism ablation of RamanPFN.** Filled symbols denote active components. The task-level RMSE ratio averages each task-repetition RMSE divided by its baseline counterpart and is the primary scale-normalized comparison. Confidence intervals are 95% task-cluster bootstrap intervals for RMSE reduction; all values are aggregated over three repetitions.

Configuration	Active mechanism			Regression performance			
	GCU	LVSE	Triplet	Mean RMSE↓	Mean task-level RMSE ratio↓	Task-level RMSE reduction (%)↑	95% CI of reduction
TabPFN baseline	○	○	○	9.259	1.000	0.0	0.0–0.0
Global representation	✓	○	○	9.233	0.898	10.2	7.4–13.2
Local representation	○	✓	○	8.885	0.893	10.7	6.2–15.2
Dual representation	✓	✓	○	9.087	0.867	13.3	9.6–17.0
RamanPFN	✓	✓	✓	8.869	0.804	19.6	15.8–23.5

The ablation resolves why the dual-scale interface works. GCU and LVSE serve distinct roles: each recovers targets missed by the other, and retaining their separate predictions allows the integration step to extract more value than the direct dual-representation configuration. RamanPFN’s predictive gain follows from this preserved complementarity between full-spectrum composition and local vibrational structure.

Extension to Raman classification

Regression and classification ask different questions of a spectrum. Regression estimates continuous quantities, whereas classification separates discrete identities or states. We used the 21 classification tasks in the collection to test whether the same dual-scale representation could support this second setting. GCU, LVSE and signed integration were evaluated under three repetitions. Figure 5 shows the task-level changes, and Table 4 reports the mechanism gradient.

Figure 5a shows that the complete framework increased mean weighted F1 from 0.833 to 0.848 and improved 16 of the 21 tasks. Relative to the remaining error, defined as $1 - F1_w$, this corresponds to a reduction of 9.0%. The gain extended across bacterial, microplastic, mineral and biomedical identification tasks. Among the five tasks without a gain, two began at a reference weighted F1 of 0.997 or above and therefore had little available headroom, two contained only four test spectra, and the remaining task declined by 0.35 percentage points. The largest and most reproducible improvement occurred on bacterial identification, where 15,700 test spectra yielded 12.6 to 12.9 percentage points in each repetition.

Figure 5b shows that mean weighted F1 increased in every repetition, by 0.76, 3.00 and 0.77 percentage points, respectively. The spread across these repetitions originated in four skin-measurement tasks that contain four test spectra each, for which a single reclassified spectrum shifts weighted F1 by more than ten percentage points; these four tasks account for all five cells beyond the colour limits. Restricting the comparison to the 17 tasks with more than four test spectra gave repetition-specific increases of 1.72, 1.55 and 2.13 percentage points and raised mean weighted F1 from 0.944 to 0.962, improving 14 of those 17 tasks.

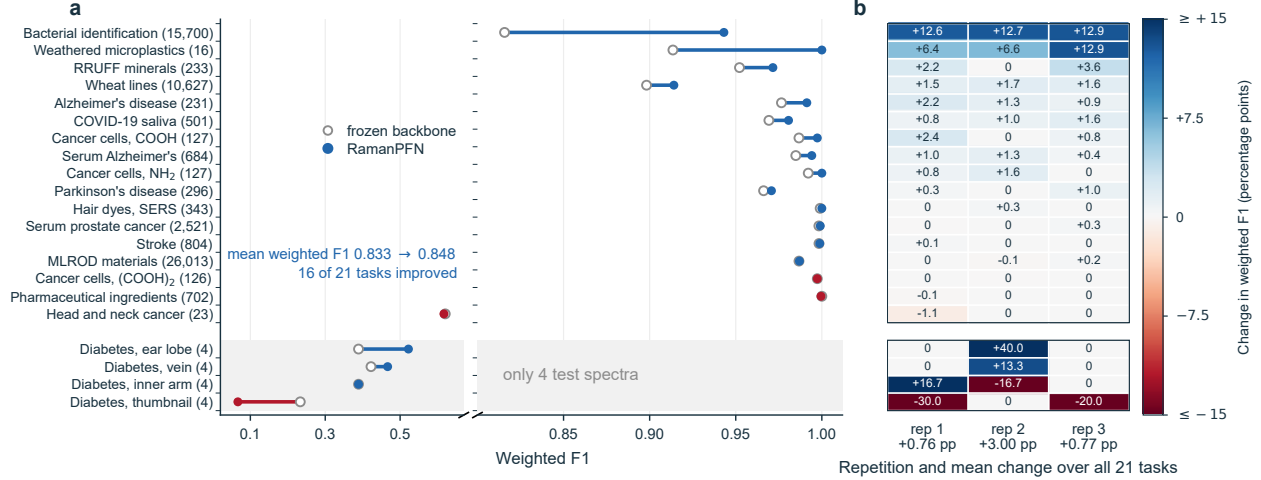


Figure 5: **Transfer of the dual-scale representation to Raman classification.** **a**, Mean weighted F1 for the TabPFN reference and RamanPFN across 21 classification tasks, averaged over three repetitions. Open grey circles denote the reference and filled circles denote RamanPFN. Blue connecting segments indicate higher weighted F1 with RamanPFN, whereas red segments indicate lower weighted F1. Tasks are ordered by mean change within each block, and the broken axis separates lower-scoring tasks from the rest. Parentheses after each task name give the number of test spectra. The shaded block holds the four diabetic-skin tasks that contain four test spectra each, for which a single reclassified spectrum shifts weighted F1 by more than ten percentage points. **b**, Repetition-specific changes in weighted F1 for the same tasks. Rows align with panel a and columns represent the three repetitions; printed values are percentage-point changes, and column labels report the mean change across all 21 tasks. Blue indicates an increase and red indicates a decrease. The colour scale is saturated at ± 15 percentage points; 5 of 63 cells extend beyond these limits, and all five lie in the four-spectrum block. Across the remaining 17 tasks the three repetitions gave increases of 1.72, 1.55 and 2.13 percentage points.

Table 4: **Classification mechanism ablation of RamanPFN.** The five configurations are evaluated on 21 classification tasks. Classification error reduction is measured on $1 - F1_w$. All values are means over three repetitions.

Configuration	Active mechanism			Classification performance		
	GCU	LVSE	Triplet	Weighted F1 $\uparrow$	$\Delta F1$ pp $\uparrow$	Error reduction % $\uparrow$
TabPFN baseline	○	○	○	0.8328	0.00	0.00
Global representation	✓	○	○	0.8440	1.12	6.72
Local representation	○	✓	○	0.8447	1.19	7.11
Dual representation	✓	✓	○	0.8454	1.26	7.51
RamanPFN	✓	✓	✓	0.8479	+1.51	9.03

Table 4 shows that the five configurations followed the same ordered pattern observed in regression. GCU and LVSE each improved the TabPFN reference independently, their joint use increased weighted F1 further, and complete RamanPFN gave the highest result. The agreement between regression and classification places the gain at the level of spectral representation rather than a single prediction objective.

RamanPFN also led all 28 external comparison methods in weighted F1, balanced accuracy, accuracy and Matthews correlation coefficient in Table 5. It obtained a classification Score of 0.797 and an Elo of 1,582. TabPFN v2.5, the strongest alternative, reached 0.631 and 1,472. When regression and classification were considered together, RamanPFN achieved an overall Score of 0.859 and an Elo of 1,742, exceeding all comparison methods on both measures in Table 6.

The classification results extend the central finding beyond continuous prediction. The same global and local coordinates support continuous quantity estimation and discrete class identification, establishing one dual-scale foundation-model interface across the principal predictive settings in Raman analysis.

Table 5: **Classification performance across diverse prediction paradigms.** RamanPFN is compared with 28 methods evaluated under the same evaluation protocol across 21 classification tasks. Weighted F1, balanced accuracy, accuracy, Matthews correlation coefficient, Score and Elo are higher-is-better. Score and Elo follow the RamanBench scoring protocol over the full comparison set, and all values aggregate the three repetitions. Bold type marks the best value in each column; bold underlining marks the second best.

Method	Predictive quality				Aggregate comparison	
	Weighted F1↑	Balanced accuracy↑	Accuracy↑	MCC↑	Score↑	Elo↑
<i>Dual-scale spectral framework</i>						
RamanPFN	0.8479	0.8604	0.8656	0.7501	0.7970	1582
<i>Tabular foundation models</i>						
TabPFN v2.5	<u>0.8308</u>	<u>0.8350</u>	0.8388	0.7051	<u>0.6314</u>	<u>1472</u>
TabICL v2	0.8262	0.8323	<u>0.8393</u>	0.7002	0.5864	1,391
TabPFN v2	0.8116	0.8182	0.8280	0.6926	0.4083	1,301
TabDPT	0.8039	0.8074	0.8140	0.6599	0.2314	1,188
MITRA	0.6999	0.7128	0.7259	0.5594	0.0294	940
<i>Raman-specific neural models</i>						
ReZeroNet	0.8103	0.8170	0.8221	0.6732	0.4571	1,299
DeepCNN	0.7944	0.7991	0.8015	0.6356	0.4233	1,246
RamanNet	0.7872	0.7864	0.7976	0.6296	0.3161	1,191
SANet	0.7168	0.7530	0.7672	0.6095	0.1656	1,080
RamanFormer	0.7331	0.7383	0.7512	0.5357	0.3091	1,046
RamanTransformer	0.3242	0.3915	0.4395	0.1246	0.0254	629
<i>Deep tabular and neural models</i>						
TabM	0.7908	0.7989	0.8102	0.6571	0.3793	1,231
RealMLP	0.8023	0.8056	0.8156	0.6588	0.2872	1,228
NN (PyTorch)	0.7906	0.7956	0.8030	0.6307	0.3351	1,224
ModernNCA	0.7857	0.7888	0.7967	0.6311	0.2493	1,176
FC-ResNeXt	0.7775	0.7811	0.7880	0.6239	0.1795	1,120
CoAtNet	0.7610	0.7685	0.7717	0.6273	0.2040	1,120
FastAI	0.6000	0.6211	0.6372	0.4341	0.0514	819
<i>Tree ensembles and gradient boosting</i>						
CatBoost	0.7738	0.7726	0.7880	0.6091	0.1367	1,117
LightGBM	0.7634	0.7862	0.7979	0.6464	0.0782	1,074
XGBoost	0.7661	0.7685	0.7796	0.6070	0.0939	1,070
Random Forest	0.7477	0.7448	0.7601	0.5851	0.0636	1,000
Extra Trees	0.7359	0.7323	0.7474	0.5651	0.0458	965
<i>Classical machine learning and chemometrics</i>						
Logistic regression	0.7668	0.7788	0.7823	0.6114	0.2979	1,159
KNN	0.7368	0.7468	0.7535	0.5841	0.0920	1,023
PLS-DA	0.6606	0.6787	0.6953	0.5199	0.1507	914
<i>Time-series classification ensembles</i>						
ROCKET	0.8296	0.8287	0.8355	<u>0.7073</u>	0.4674	1,365
ARSENAL	0.8242	0.8320	0.8378	0.7056	0.3163	1,273

Table 6: **Overall performance of RamanPFN and comparison methods.** Overall Score and Elo follow the RamanBench scoring protocol for methods with complete regression and classification coverage. Values aggregate the three repetitions. Bold type marks the best value in each column; bold underlining marks the second best.

Method	Overall Score [†]	Overall Elo [†]
<i>Dual-scale spectral framework</i>		
RamanPFN	0.8586	1742
<i>Tabular foundation models</i>		
TabPFN v2.5	<u>0.6180</u>	<u>1469</u>
TabPFN v2	0.5003	1,431
TabICL v2	0.5395	1,375
MITRA	0.1779	1,152
TabDPT	0.2312	1,118
<i>Raman-specific neural models</i>		
ReZeroNet	0.3714	1,187
DeepCNN	0.3286	1,133
RamanNet	0.2495	1,113
RamanFormer	0.2666	1,054
SANet	0.1307	920
RamanTransformer	0.0155	579
<i>Deep tabular and neural models</i>		
RealMLP	0.2504	1,144
TabM	0.2462	1,110
NN (PyTorch)	0.2465	1,104
CoAtNet	0.1775	1,061
FC-ResNeXt	0.1731	1,052
ModernNCA	0.1350	848
FastAI	0.0329	806
<i>Tree ensembles and gradient boosting</i>		
CatBoost	0.1095	1,072
Extra Trees	0.0932	1,058
Random Forest	0.0867	1,000
LightGBM	0.0602	944
XGBoost	0.0804	927
<i>Classical machine learning and chemometrics</i>		
Linear / logistic regression	0.2506	1,058
PLS / PLS-DA	0.1629	1,022
KNN	0.1157	1,015
<i>Time-series classification ensembles</i>		
ROCKET [†]	–	–
ARSENAL [†]	–	–

[†]Classification-only methods cover the 21 classification tasks but none of the 129 regression tasks; Overall Score and Elo are therefore not reported.

Discussion

RamanPFN shows that a dual-scale spectral representation can extend tabular foundation-model inference to high-dimensional Raman spectra. TabPFN already ranks among the strongest methods for Raman prediction, even though nearly all datasets evaluated here exceed its recommended feature range [9, 20]. Its feature-subsampling interface gives each wavenumber access to the model, yet joint visibility falls sharply as spectral width grows. GCU and LVSE reorganize the spectrum into explicit, complementary representations of full-spectrum compositional covariation and local vibrational morphology. This reorganization reduced mean RMSE by 19.6% across 129 regression targets. The same representation logic generalized to classification, increasing weighted F1 from 0.833 to 0.848 across 21 tasks. Together with the mechanism analysis and ablation, these results support the central conclusion of this study: coverage is not context. Once a spectrum’s global composition and local vibration are made explicit at the input, a tabular foundation model can reason across the compositional and vibrational scales that organize the signal.

Raman spectra organize predictive information along two coupled scales: global compositional covariation and local vibrational structure. GCU gives each compositional coordinate a full-spectrum receptive field, allowing separated bands driven by a common latent source to enter the same representation. This follows the physical intuition of constrained Raman unmixing, in which non-negative components and abundance coordinates organize mixtures across the complete spectrum [12]. RamanPFN treats these components as predictive axes of compositional covariation. Independent endmember evidence would be required for chemical assignment. LVSE operates at the complementary scale. Previous locality-preserving encoders have grouped adjacent channels or retained one leading direction from each ordered segment [29, 38]. LVSE retains several modes within every region and preserves independent changes in local band morphology. The task-level ablation supports this division of labour because each branch improved prediction independently and their relative advantage changed across targets. The signed triplet integration rule then uses the contrast between role-specific predictions to express both scales in the final estimate. The resulting pattern identifies global compositional structure and local vibrational structure as distinct, jointly useful sources of predictive information.

RamanPFN brings together two lines of research that have so far advanced largely in parallel. Spectroscopic learning has used physical constraints, chemometric preprocessing and decomposition to expose chemically meaningful variation, with Raman unmixing and recent near-infrared calibration studies demonstrating the value of structure-aware spectral representations [12, 30]. The latter study also identified spectroscopy-specific priors as an important next step for tabular foundation models. A second line of work has adapted TabPFN to scientific settings by changing the information presented to it. MFTabPFN introduces an adapter for multi-fidelity engineering variables, while ICL-FM uses compact composition descriptors and learned structure embeddings for materials prediction [39, 40]. RamanPFN connects these directions at the level of a single high-dimensional measurement. Its representation is constructed directly from the physically ordered wavenumber axis, and its context problem occurs within each sample as thousands of channels compete for a bounded feature budget. GCU and LVSE turn spectroscopy-specific structure into the input coordinates used for in-context learning. This connection positions RamanPFN as a foundation-model framework in which the scientific representation carries the compositional and vibrational organization of an ordered spectrum.

The broader implication of these results is that the representation interface forms a scientific modelling layer in its own right. General-purpose tabular foundation models obtain transferability by operating across arbitrary feature schemas within bounded input contexts [20, 23]. Scientific foundation models increasingly encode the organizing structure of their domain, including atomic

geometry in materials models [41–43] and temporal order in time-series models [44–46]. RamanPFN applies this principle at the level of a measured spectrum by concentrating spectral physics in the coordinates presented to the predictor. The scientific representation defines the relevant receptive fields, while the tabular prior supplies reusable small-data inference. Remote-sensing foundation models likewise encode spectral, spatial and multimodal organization across sensors and tasks [47–51]. The same design principle may prove useful for near-infrared and infrared spectra, hyperspectral measurements, chromatography, diffraction and other ordered signals in which local structure coexists with long-range covariation [30, 38]. The present evidence establishes this principle across Raman regression and classification. Whether the same physical abstractions transfer across measurement modalities remains an open question.

Several questions now become experimentally accessible. GCU coordinates could be coupled to reference spectra or molecular constraints to test when predictive factors acquire chemical identity. Peak-aware or multiresolution partitions could refine LVSE where peak density and instrumental resolution vary across the wavenumber axis. Cross-instrument transfer, batch drift, calibration extrapolation and prospective deployment provide stronger tests of whether the constructed context remains stable beyond the present data distributions. RamanPFN was evaluated across all 150 tasks without dataset-specific hyperparameter search. This shared dual-scale design remained effective across Raman measurements that vary widely in sample type, dimensionality and acquisition setting. These boundaries define a concrete research agenda. The central finding remains clear: a dual-scale spectral representation substantially expands what a tabular foundation model can extract from an ordered spectrum. RamanPFN places global composition and local vibration at the centre of foundation-model learning from Raman spectra.

Methods

Study design and evaluation protocol

Experiments used the public RamanBench dataset collection and its standardized task definitions [9]. The regression study comprised 129 prediction targets from 53 public Raman datasets, with each target treated as a separate supervised task. The classification study comprised 21 tasks covering discrete sample identities or states. We retained the complete set of tasks meeting the published inclusion criteria and applied no additional task filtering. The resulting collection spans materials, biological, clinical, chemical and industrial applications, with substantial variation in sample count and spectral dimensionality. Its quantitative composition is summarized in Table 1.

All data partitions followed the deterministic 80/20 protocol of the published RamanBench release. Each task was evaluated using random seeds 0, 1 and 2, which formed the three experimental repetitions. Classification splits were stratified by class. Datasets containing experimental groups were divided with group-aware sampling, while the remaining tasks used seeded random sampling. Representation fitting and candidate evaluation used the training partition of each repetition, with OOF predictions preserving experimental groups where available and class proportions for classification.

We quantified joint feature visibility using the balanced feature-subsampling procedure implemented by TabPFN [20]. The analysis used the observed spectral width and the ensemble configuration applied in the regression experiments. Each estimator-specific forward pass contained at most 500 features, while automatic estimator scaling increased the effective estimator count when required, subject to a maximum of 32. For each of the 129 regression tasks, we ran 3,000 Monte Carlo trials. Each trial sampled 128 random feature pairs and 128 random feature triples and recorded whether all members appeared together in at least one estimator-specific subset. Task-level

co-occurrence probabilities were obtained by averaging these indicators across trials. This analysis depended only on feature indices and the TabPFN sampling procedure and was independent of target values and RamanPFN representations.

RamanPFN formulation

Let $X_{\text{tr}} \in \mathbb{R}^{n \times d}$ denote a training set of n Raman spectra measured over d ordered spectral channels, with corresponding targets y_{tr} . The held-out spectra are denoted by $X_{\text{te}} \in \mathbb{R}^{m \times d}$. The channel positions are represented by $\nu = (\nu_1, \dots, \nu_d)$ and correspond to physical Raman shifts when these are available, or to the original channel order otherwise. Regression targets take continuous values, while classification targets belong to a finite class set.

RamanPFN constructs two representations of each spectrum. GCU produces compositional coordinates whose basis components extend across the complete spectral axis. LVSE produces coordinates associated with multiple modes of variation within contiguous spectral regions. For representation mechanism $r \in \{g, l\}$ and a structurally defined setting q , the representation map is fitted from the training spectra and then applied to both partitions:

$$\hat{\Phi}_{r,q} = \text{Fit}_{r,q}(X_{\text{tr}}, \nu), \quad Z_{r,q}^{\text{tr}} = \hat{\Phi}_{r,q}(X_{\text{tr}}), \quad Z_{r,q}^{\text{te}} = \hat{\Phi}_{r,q}(X_{\text{te}}).$$

For each candidate representation, decomposition bases and standardization statistics were estimated from the training partition. Sample-wise transformations required no target information. Both mechanisms draw from shared multiresolution portfolios. Within each task and repetition, protocol-matched OOF predictions rank the GCU and LVSE candidates for the current training distribution; candidate representations used at inference are then refitted on the complete training partition.

The same TabPFN model evaluated every representation:

$$p_{r,q} = f_{\theta}(Z_{r,q}^{\text{tr}}, y_{\text{tr}}; Z_{r,q}^{\text{te}}),$$

where θ denotes the fixed model parameters [20]. Regression used a TabPFN regressor and classification used the corresponding classifier. Both used eight initial estimators with automatic estimator scaling, and the model random state matched the repetition seed. RamanPFN therefore changes the spectral coordinates presented to TabPFN while preserving its pretrained architecture and parameters. The resulting scalar predictions or class-probability estimates provide the inputs to the dual-scale prediction integration described below.

Global Compositional Unmixing

Global Compositional Unmixing represents coordinated variation across the complete Raman spectrum. Raman intensities were first mapped to a non-negative domain using deterministic shift and scale operations that required no target information. This construction permits an additive decomposition in which latent spectral components and their sample-specific coordinates remain non-negative, following the established interpretation of spectral unmixing [12, 13, 52].

For a chosen rank ρ , GCU factorizes the transformed training spectra X_{tr}^+ by solving

$$\min_{W_{\text{tr}} \geq 0, H \geq 0} \|X_{\text{tr}}^+ - W_{\text{tr}} H\|_F^2,$$

where $W_{\text{tr}} \in \mathbb{R}_+^{n \times \rho}$ contains the sample coordinates and $H \in \mathbb{R}_+^{\rho \times d}$ contains the spectral components. Each row of H is defined over all d spectral channels. A coordinate in W_{tr} can therefore respond jointly to bands separated along the wavenumber axis.

The factorization was initialized from a non-negative singular-value decomposition [53] and optimized with hierarchical alternating least squares [54]. Intermediate factorization states yield a multiresolution family of full-spectrum compositional coordinates. At iteration checkpoint t , the pair $(W_{\text{tr}}^{(t)}, H^{(t)})$ defines one such representation.

Test spectra were projected onto the training-derived components while keeping $H^{(t)}$ fixed:

$$W_{\text{te}}^{(t)} = \arg \min_{W \geq 0} \|X_{\text{te}}^+ - WH^{(t)}\|_F^2.$$

The GCU token matrices supplied to TabPFN were therefore $Z_{g,q}^{\text{tr}} = W_{\text{tr}}^{(t)}$ and $Z_{g,q}^{\text{te}} = W_{\text{te}}^{(t)}$. Here, q indexes this multiresolution family. The components were treated as latent axes of compositional covariation. No chemical endmember identities were assigned without independent reference evidence.

Local Vibrational Subspace Encoding

Local Vibrational Subspace Encoding preserves variation within contiguous regions of the ordered spectral axis. LVSE applies a deterministic spectral transformation, denoted by T_a . Feature-wise means μ_a and standard deviations σ_a were estimated from the transformed training spectra and applied to both partitions [55]:

$$\tilde{X}_{\text{tr}} = \frac{T_a(X_{\text{tr}}) - \mu_a}{\sigma_a}, \quad \tilde{X}_{\text{te}} = \frac{T_a(X_{\text{te}}) - \mu_a}{\sigma_a}.$$

Where used, standard normal variate normalization was applied independently to each spectrum and required no target information.

The transformed spectral axis was divided into contiguous intervals $\mathcal{I}_1, \dots, \mathcal{I}_S$.

A separate centered singular value decomposition was fitted within each interval [56]. For interval s , the training and test matrices were centered using the corresponding training-region mean:

$$A_s^{\text{tr}} = \tilde{X}_{\text{tr}}[:, \mathcal{I}_s] - \bar{x}_s, \quad A_s^{\text{te}} = \tilde{X}_{\text{te}}[:, \mathcal{I}_s] - \bar{x}_s.$$

The training matrix was then decomposed as

$$A_s^{\text{tr}} = U_s \Sigma_s V_s^{\text{T}}.$$

Retaining k_s right singular vectors produced the local coordinates

$$Z_s^{\text{tr}} = A_s^{\text{tr}} V_{s,1:k_s}, \quad Z_s^{\text{te}} = A_s^{\text{te}} V_{s,1:k_s},$$

where k_s did not exceed the available rank of the interval.

The complete LVSE representation concatenated the coordinates from all intervals:

$$Z_{l,q} = [Z_1 \mid Z_2 \mid \dots \mid Z_S].$$

The concatenated coordinates define a multiresolution family indexed by q . Every LVSE coordinate remains associated with a defined spectral interval, while retaining multiple components exposes independent directions of variation within that interval. This construction builds on contiguous spectral or channel grouping in high-dimensional scientific signals [38] and extends single-direction segment compression by retaining multiple local directions within each region [29].

Dual-scale prediction integration

GCU and LVSE provide a dual-scale set of candidate predictions from the same spectrum. Training OOF evidence constructs a reference prediction p_{ref} , while a fixed prediction-space policy assigns candidates from either representation to negative (p_-), local (p_l) and anchor (p_h) roles. For retained triplet j , the signed estimate is

$$p_{\text{triplet}}^{(j)} = -\lambda_j p_-^{(j)} + \lambda_j p_l^{(j)} + p_h^{(j)}, \quad \lambda_j \in \{1, 0.75, 0.5\}.$$

Eight triplets are retained and combined relative to the OOF reference:

$$\hat{p} = p_{\text{ref}} + \alpha \left[\mathcal{A} \left(\left\{ p_{\text{triplet}}^{(j)} \right\}_{j=1}^8 \right) - p_{\text{ref}} \right].$$

The policy prescribes $\mathcal{A}$ as a mean, median or symmetrically trimmed mean. The correction magnitude α is fixed or determined analytically from the training-target scale, the radius of the selected correction and agreement among the retained triplets. For classification, the same operation is applied coordinate-wise to class-aligned probability vectors, followed by projection onto the probability simplex.

Ablation and comparative evaluation

The TabPFN model applied directly to the original spectral channels served as the reference configuration. The ablation followed the logical sequence of RamanPFN. We evaluated five controlled configurations: the TabPFN reference, TabPFN with GCU, TabPFN with LVSE, TabPFN with both spectral representations, and complete RamanPFN with signed triplet integration. The single-representation configurations measured the independent contributions of global compositional unmixing and local vibrational subspace encoding. The dual-representation configuration tested whether the two spectral scales remained complementary before signed integration, and the complete framework measured the additional contribution of output-level integration. All five configurations were evaluated on the 129 regression targets and 21 classification tasks under the same three repetitions.

The external comparison used the complete 29-method comparison pool: 26 baselines for regression and 28 baselines for classification, with ARSENAL and ROCKET available only for classification. The pool spans tabular foundation models, Raman-specific neural models, general deep tabular learners, chemometric references and tree-based ensembles [57–60]; the complete method roster and task-level results are provided in the Source Data. Every method was executed on identical task definitions, train-test partitions and repetition seeds. Each retained its native preprocessing and fitting procedure, and no external comparator received GCU or LVSE coordinates.

Score and Elo were calculated with the RamanBench scoring implementation. Regression and classification were scored independently, with RamanPFN and all reproduced methods available for the corresponding prediction type evaluated together in one pool. The overall Score assigned equal weight to the two prediction types, whereas overall Elo followed the target-wise pairing defined by the scoring protocol. Aggregate results were calculated after complete task and repetition coverage had been verified.

Evaluation metrics and statistical analysis

Regression performance was assessed primarily by root-mean-square error (RMSE), with mean absolute error (MAE) providing a complementary measure of absolute deviation. Both metrics were

calculated separately for every target and repetition. Because the regression targets span different physical quantities and numerical scales, mechanism comparisons used the RMSE change relative to the TabPFN reference:

$$\Delta_{i,s}^{(m)} = 100 \left(\frac{\text{RMSE}_{i,s}^{(m)}}{\text{RMSE}_{i,s}^{(0)}} - 1 \right),$$

where i denotes the target, s the experimental repetition, m the evaluated configuration and 0 the TabPFN reference. Negative values indicate lower prediction error. Framework-level changes were averaged over the complete set of 387 target–repetition outcomes, whereas the number of improved targets was calculated after averaging the three repetitions within each of the 129 targets.

Classification performance was summarized primarily by support-weighted F1 score. Balanced accuracy, accuracy and Matthews correlation coefficient provided complementary measures of class-balanced recall, overall correctness and agreement across the complete confusion matrix. Each metric was computed independently for every task and repetition before aggregation. Classification ablations report both the absolute change in weighted F1 and the reduction in the remaining classification error:

$$R_{\text{err}}^{(m)} = 100 \frac{(1 - \text{F1}_w^{(0)}) - (1 - \text{F1}_w^{(m)})}{1 - \text{F1}_w^{(0)}}.$$

Four classification tasks measured diabetic skin at different anatomical sites and contained 16 training and four test spectra each. They were retained as defined by RamanBench, so their task-level changes are quantized in steps of several percentage points, and classification results are reported both with and without them.

Uncertainty in the regression ablation was quantified at the target level. The three repetition-specific RMSE changes were first averaged within each target. We then generated 20,000 bootstrap samples by resampling the 129 target means with replacement. The 2.5th and 97.5th percentiles of the resulting distribution define the reported 95% task-cluster confidence intervals. Two-sided Wilcoxon signed-rank tests were additionally applied to the target-level mean log RMSE ratios relative to the TabPFN reference. For omnibus comparison across RamanPFN and the 26 external regression baselines, repetitions were averaged within each target before applying the Friedman test. A source-dataset-level analysis, obtained by averaging the ranks of targets originating from the same dataset, examined the sensitivity of the comparison to datasets containing multiple regression targets. Pairwise method comparisons used the symmetric RMSE advantage $100(\text{RMSE}_b - \text{RMSE}_a) / \max(|\text{RMSE}_a|, |\text{RMSE}_b|)$ after averaging repetitions within each target. Target-level advantages were then averaged within each of the 53 source datasets, followed by two-sided Wilcoxon signed-rank tests with the Pratt treatment of zeros and Holm correction across the 26 prespecified RamanPFN–baseline contrasts.

Score and Elo followed the RamanBench scoring protocol [9]. For each target, the three repetitions were first averaged. Score maps the median method performance to zero and the best performance to one, clips values below the median at zero and then averages the normalized values across targets. Regression uses negative RMSE as the ranking evidence, whereas classification uses weighted F1. Elo is derived from target-wise pairwise wins over 200 randomized target orderings, using a K -factor of 32 and a logistic scale of 400, and is calibrated against the fixed RamanBench reference point. Overall Score is the equal-weight mean of the regression and classification Scores, while overall Elo is calculated from the pooled regression and classification targets.

Implementation and reproducibility

RamanPFN was implemented in Python using the official TabPFN models. The accompanying implementation reproduces the representation portfolios, grouped OOF adaptation and prediction-space integration used in all reported experiments. The exact numerical configuration is retained in machine-readable form with the implementation.

References

- [1] Guang Yang, Xin Li, Yongqiang Cheng, Mingchao Wang, Dong Ma, Alexei P. Sokolov, Sergei V. Kalinin, Gabriel M. Veith, and Jagjit Nanda. Distilling nanoscale heterogeneity of amorphous silicon using tip-enhanced Raman spectroscopy (TERS) via multiresolution manifold learning. *Nature Communications*, 12(1):578, 2021. doi: 10.1038/s41467-020-20691-2.
- [2] Sicheng Li, Haozheng Li, Yiran Li, Qi Zhang, Shuai Wang, Xin Lv, Shuai Yan, Zhiliang Huang, Xingbo Liu, Qipei Zhou, Bi Zhang, Long Xiao, Yage Chen, Zhe Wang, Wanjuan Lu, Aiguo Shen, Jianfeng Liu, and Ping Wang. Photon-counting Raman spectroscopy at a MHz spectral rate for biochemical imaging of an entire organism. *Nature Communications*, 16(1):3808, 2025. doi: 10.1038/s41467-025-59030-8.
- [3] Marlous Kamp, Jakub Surmacki, Marc Segarra Mondejar, Tim Young, Karolina Chrabaszcz, Fadwa Joud, Vincent Zecchini, Alyson Speed, Christian Frezza, and Sarah E. Bohndiek. Raman micro-spectroscopy reveals the spatial distribution of fumarate in cells and tissues. *Nature Communications*, 15(1):5386, 2024. doi: 10.1038/s41467-024-49403-w.
- [4] Liping Huang, Hongwei Sun, Liangbin Sun, Keqing Shi, Yuzhe Chen, Xueqian Ren, Yuancai Ge, Danfeng Jiang, Xiaohu Liu, Wolfgang Knoll, Qingwen Zhang, and Yi Wang. Rapid, label-free histopathological diagnosis of liver cancer based on Raman spectroscopy and deep learning. *Nature Communications*, 14(1):48, 2023. doi: 10.1038/s41467-022-35696-2.
- [5] Yirui Zhang, Kai Chang, Babatunde Ogunlade, Liam Herndon, Loza F. Tadesse, Amanda R. Kirane, and Jennifer A. Dionne. From genotype to phenotype: Raman spectroscopy and machine learning for label-free single-cell analysis. *ACS Nano*, 18(28):18101–18117, 2024. doi: 10.1021/acsnano.4c04282.
- [6] Oleksii Ilchenko, Yurii Pilhun, Andrii Kutsyk, Denys Slobodianiuk, Yaman Goksel, Elodie Dumont, Lukas Vaut, Chiara Mazzoni, Lidia Morelli, Sofus Boisen, Konstantinos Stergiou, Yaroslav Aulin, Tomas Rindzevicius, Thomas Emil Andersen, Mikael Lassen, Hemanshu Mundhada, Christian Bille Jendresen, Peter Alshede Philipsen, Merete Hædersdal, and Anja Boisen. Optics miniaturization strategy for demanding Raman spectroscopy applications. *Nature Communications*, 15(1):3049, 2024. doi: 10.1038/s41467-024-47044-7.
- [7] Xiaqiong Fan, Yue Wang, Chuanxiu Yu, Yuanxia Lv, Hailiang Zhang, Qiong Yang, Ming Wen, Hongmei Lu, and Zhimin Zhang. A universal and accurate method for easily identifying components in Raman spectroscopy based on deep learning. *Analytical Chemistry*, 95(11):4863–4870, 2023. doi: 10.1021/acs.analchem.2c03853.
- [8] Manuel Sigle, Anne-Katrin Rohlfing, Martin Kenny, Sophia Scheuermann, Na Sun, Ulla Graefner, Verena Haug, Jessica Sudmann, Christian M. Seitz, David Heinzmann, Katja Schenke-Layland, Patricia B. Maguire, Axel Walch, Julia Marzi, and Meinrad Paul Gawaz.

- Translating genomic tools to Raman spectroscopy analysis enables high-dimensional tissue characterization on molecular resolution. *Nature Communications*, 14(1):5799, 2023. doi: 10.1038/s41467-023-41417-0.
- [9] Mario Koddenbrock, Christoph Lange, Robin Legner, Martin Jäger, Martin Kögler, Mariano N. Cruz Bournazou, Peter Neubauer, Felix Bießmann, and Erik Rodner. RamanBench: a large-scale benchmark for machine learning on Raman spectroscopy. *arXiv preprint arXiv:2605.02003*, 2026.
 - [10] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. doi: 10.1016/S0169-7439(01)00155-1.
 - [11] Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003. doi: 10.1002/cem.785.
 - [12] Dimitar Georgiev, Álvaro Fernández-Galiana, Simon Vilms Pedersen, Georgios Papadopoulos, Ruoxiao Xie, Molly M. Stevens, and Mauricio Barahona. Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders. *Proceedings of the National Academy of Sciences*, 121(45):e2407439121, 2024. doi: 10.1073/pnas.2407439121.
 - [13] Behnood Rasti, Alexandre Zouaoui, Julien Mairal, and Jocelyn Chanussot. Image processing and machine learning for hyperspectral unmixing: an overview and the HySUPP python package. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–31, 2024. doi: 10.1109/TGRS.2024.3393570.
 - [14] Pengrui Wang, Rong Liu, and Liangpei Zhang. MAT-Net: multiscale aggregation transformer network for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2024.3494795.
 - [15] Conor C. Horgan, Magnus Jensen, Anika Nagelkerke, Jean-Philippe St-Pierre, Tom Vercauteren, Molly M. Stevens, and Mads S. Bergholt. High-throughput molecular imaging via deep-learning-enabled Raman spectroscopy. *Analytical Chemistry*, 93(48):15850–15860, 2021. doi: 10.1021/acs.analchem.1c02178.
 - [16] Haonan Lin, Hyeon Jeong Lee, Nathan Tague, Jean-Baptiste Lugagne, Cheng Zong, Fengyuan Deng, Jonghyeon Shin, Lei Tian, Wilson Wong, Mary J. Dunlop, and Ji-Xin Cheng. Microsecond fingerprint stimulated Raman spectroscopic imaging by ultrafast tuning and spatial-spectral learning. *Nature Communications*, 12(1):3052, 2021. doi: 10.1038/s41467-021-23202-z.
 - [17] Jianhui Wan, Yuheng Wang, Weile Zhu, Weina Zhang, and Liyun Zhong. Deep learning-decoded Raman spectroscopy for hour-scale iPSC pluripotency assessment via lipid–protein biomarkers. *Analytical Chemistry*, 97(46):25488–25499, 2025. doi: 10.1021/acs.analchem.5c03560.
 - [18] Longji Zhu, Yunan Yang, Fei Xu, Xinyu Lu, Mingrui Shuai, Zhulin An, Xiaomeng Chen, Hu Li, Francis L. Martin, Peter J. Vikesland, Bin Ren, Zhong-Qun Tian, Yong-Guan Zhu, and Li Cui. Open-set deep learning-enabled single-cell Raman spectroscopy for rapid identification of airborne pathogens in real-world environments. *Science Advances*, 11(2):eadp7991, 2025. doi: 10.1126/sciadv.adp7991.
 - [19] Jiaqi Hu, Gina Jinna Chen, Chenlong Xue, Pei Liang, Yanqun Xiang, Chuanlun Zhang, Xiaokeng Chi, Guoying Liu, Yanfang Ye, Dongyu Cui, De Zhang, Xiaojun Yu, Hong Dang,

- Wen Zhang, Junfan Chen, Quan Tang, Penglai Guo, Ho-Pui Ho, Yuchao Li, Longqing Cong, and Perry Ping Shum. RSPSSL: a novel high-fidelity Raman spectral preprocessing scheme to enhance biomedical applications and chemical resolution visualization. *Light: Science & Applications*, 13:52, 2024. doi: 10.1038/s41377-024-01394-5.
- [20] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.
- [21] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, volume 36, pages 76336–76369. Curran Associates, Inc., 2023. doi: 10.52202/075280-3337.
- [22] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: a living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.
- [23] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: a tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50817–50847. PMLR, 2025.
- [24] Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maks Volkovs. TabDPT: scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, volume 38, pages 172692–172722. Curran Associates, Inc., 2025.
- [25] Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at TabPFN v2: understanding its strengths and extending its capabilities. *arXiv preprint arXiv:2502.17361*, 2025.
- [26] Si-Yang Liu and Han-Jia Ye. TabPFN unleashed: a scalable and effective solution to tabular classification problems. *arXiv preprint arXiv:2502.02527*, 2025.
- [27] Christopher Kolberg, Jules Kreuer, Jonas Huurdeman, Sofiane Ouaari, Katharina Eggersperger, and Nico Pfeifer. TabPFN-Wide: continued pre-training for extreme feature counts. *arXiv preprint arXiv:2510.06162*, 2025.
- [28] Benjamin Feuer, Robin Tibor Schirrmeister, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. TuneTables: context optimization for scalable prior-data fitted networks. In *Advances in Neural Information Processing Systems*, volume 37, pages 83430–83464, 2024. doi: 10.52202/079017-2654.
- [29] Al Zadid Sultan Bin Habib, Md Younus Ahamed, Prashnna Kumar Gyawali, Gianfranco Doretto, and Donald A. Adjeroh. GOTabPFN: from feature ordering to compact tokenization for tabular foundation models on high-dimensional data. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. To appear.
- [30] Robin Reiter, Denis Cornet, Fabien Michel, Lauriane Rouan, and Gregory Beurier. Tabular foundation models for robust calibration of near-infrared chemical sensing data. *arXiv preprint arXiv:2605.21544*, 2026.

- [31] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.
- [32] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023. doi: 10.1038/s41586-023-06735-9.
- [33] Claudio Zeni et al. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5.
- [34] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167. PMLR, 2024.
- [35] Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL <https://openreview.net/forum?id=gerNCVqqtR>.
- [36] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.
- [37] Bingsen Xue, Xinyuan Bi, Zheyi Dong, Yunzhe Xu, Minghui Liang, Xin Fang, Yizhe Yuan, Ruoxi Wang, Shuyu Liu, Rushi Jiao, Yuze Chen, Weitao Zu, Chengxiang Wang, Jianhao Zhang, Jiang Liu, Qin Zhang, Ye Yuan, Midie Xu, Ya Zhang, Yanfeng Wang, Jian Ye, and Cheng Jin. Deep spectral component filtering as a foundation model for spectral analysis demonstrated in metabolic profiling. *Nature Machine Intelligence*, 7(5):743–757, 2025. doi: 10.1038/s42256-025-01027-5.
- [38] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. HyperSIGMA: hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025. doi: 10.1109/TPAMI.2025.3557581.
- [39] Yan Shi, Cheng Liu, Aodi Yu, Zhenzhou Lu, Said Elias, Kai Cheng, Jiaqing Kou, Xin Chen, Yu Liu, Hong-Zhong Huang, and Michael Beer. A multi-fidelity tabular prior-data fitted network model for accurate prediction and uncertainty quantification. *Nature Communications*, 2026. doi: 10.1038/s41467-026-75163-w.
- [40] Qinyang Li, Rongzhi Dong, Nicholas Miklaucic, Jeffrey Hu, Sadman Sadeed Omee, Lai Wei, Sourin Dey, Ming Hu, and Jianjun Hu. In context learning foundation models for materials

- property prediction with small datasets. *npj Computational Materials*, 12(1):222, 2026. doi: 10.1038/s41524-026-02089-8.
- [41] Ilyes Batatia et al. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163:184110, 2025. doi: 10.1063/5.0297006.
 - [42] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. doi: 10.1038/s43588-022-00349-3.
 - [43] Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11(1):178, 2025. doi: 10.1038/s41524-025-01650-1.
 - [44] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: a family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16115–16152. PMLR, 2024.
 - [45] Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32369–32399. PMLR, 2024.
 - [46] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR, 2024.
 - [47] Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, Qi Zhu, Jian Wang, Ming Yang, Jingdong Chen, Yongjun Zhang, and Yansheng Li. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, 7(8):1235–1249, 2025. doi: 10.1038/s42256-025-01078-8.
 - [48] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: a vision-language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3390838.
 - [49] Weili Kong, Baisen Liu, Xiaojun Bi, Changdong Yu, Xinyao Li, and Yushi Chen. HyperSL: a spectral foundation model for hyperspectral image interpretation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025. doi: 10.1109/TGRS.2025.3566205.
 - [50] Xin He, Yushi Chen, Lingbo Huang, Danfeng Hong, and Qian Du. Foundation model-based multimodal remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. doi: 10.1109/TGRS.2023.3344698.
 - [51] Lingbo Huang, Yushi Chen, and Xin He. Foundation model-based spectral-spatial transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–25, 2024. doi: 10.1109/TGRS.2024.3456129.

- [52] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. doi: 10.1038/44565.
- [53] Christos Boutsidis and Efstratios Gallopoulos. SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008. doi: 10.1016/j.patcog.2007.09.010.
- [54] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *Independent Component Analysis and Signal Separation*, volume 4666 of *Lecture Notes in Computer Science*, pages 169–176. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-74494-8_22.
- [55] R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5):772–777, 1989. doi: 10.1366/0003702894202201.
- [56] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970. doi: 10.1007/BF02163027.
- [57] Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: fast convergence at large depth. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1352–1361. PMLR, 2021. URL <https://proceedings.mlr.press/v161/bachlechner21a.html>.
- [58] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.
- [59] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025.
- [60] Han-Jia Ye, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Revisiting nearest neighbor for tabular data: a deep tabular baseline two decades later. In *International Conference on Learning Representations*, 2025.