

RESEARCH ARTICLE



An AI-Driven Multi-Objective Framework for Interpretable Near-Infrared Spectral Preprocessing

Ali Khumaidi^{1,*} and Ridwan Raafi'udin²

¹ Doctoral Program in Informatics, Universitas Nusa Mandiri, Indonesia

² Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia

Abstract: Preprocessing has a large, dataset-dependent influence on near-infrared (NIR) calibration, yet analysts often choose transformation sequences by trial and error. This makes the choice difficult to reproduce and may add steps that yield little practical benefit. We therefore cast preprocessing selection as a constrained bi-objective search in which calibration error and the number of active operations are minimized together. Candidate pipelines combined standard normal variate correction, multiplicative scatter correction, detrending, and Savitzky–Golay smoothing or derivatives, subject to chemometric ordering constraints. Random Search provided a non-guided control; Non-dominated Sorting Genetic Algorithm II (NSGA-II) and Multi-objective Tree-structured Parzen Estimator (MOTPE) were evaluated with an identical budget of 200 trials on six public NIR and visible-NIR datasets. All search decisions used calibration data, after which one Pareto-efficient pipeline was refitted and tested once on the held-out subset. NSGA-II gave the better mean validation-root mean square error of prediction (RMSEP) rank (1.33), whereas MOTPE gave the better mean normalized-hypervolume rank (1.25). Compared with the strongest manual reference for each dataset, the optimized pipelines lowered validation RMSEP in five cases by between 3.76% and 16.88%. Wheat was the exception: manual Standard normal variate (SNV) generalized better. Across 10 seeds, held-out RMSEP varied little for most datasets even when selected pipelines differed. Nonlinear calibration was chiefly useful for Mango. Adding wavelength selection reduced Wheat RMSEP to 0.5829 and removed 28.7% of its spectral variables. These findings position multi-objective search as an auditable aid for choosing concise preprocessing pipelines, not as a substitute for independent validation or chemometric judgment.

Keywords: automated preprocessing, multi-objective optimization, near-infrared spectroscopy, Pareto optimization, interpretable machine learning

1. Introduction

Near-infrared (NIR) spectroscopy enables rapid, nondestructive, and comparatively low-cost assessment of agricultural, food, pharmaceutical, and environmental materials [1–3]. Spectra can be collected with little sample preparation, which supports laboratory analysis, routine quality control, and field sensing [4]. That convenience does not make calibration straightforward. Chemical absorption is recorded together with scatter, baseline displacement, overlapping bands, instrument effects, and physical differences among samples [5–7]. Variation from these sources may conceal the spectral response of the property of interest and, if left untreated, weaken prediction reliability.

Partial least squares regression (PLSR) is a common calibration choice for NIR data because it can model many collinear wavelengths through a relatively interpretable latent-variable representation [8]. Its performance, however, is closely tied to the

spectra supplied to the model. SNV and multiplicative scatter correction (MSC) are used to correct scatter-related offsets and scaling, whereas detrending addresses baseline structure. Savitzky–Golay smoothing attenuates high-frequency noise, and its derivatives can sharpen local spectral features [5, 7, 9, 10]. None of these operations is uniformly advantageous. Their effects vary with the sample matrix, wavelength range, instrument, reference property, and validation design; a useful transformation for one dataset can be detrimental to another [5, 11, 12].

In applied NIR work, preprocessing is still often selected by trying a familiar set of transformations and retaining the sequence with the most favorable calibration or cross-validation statistic. This hands-on procedure is easy to understand, but it is difficult to reproduce and becomes cumbersome as the number, order, and parameter settings of candidate operations increase. It can also favor a long pipeline whose apparent calibration gain is too small to justify its operational burden. The issue is particularly relevant to portable, embedded, and real-time instruments, for which an acceptable pipeline must remain auditable and computationally manageable [4, 13]. In that setting, a shorter sequence with

*Corresponding author: Ali Khumaidi, Doctoral Program in Informatics, Universitas Nusa Mandiri, Indonesia. Email: alikhumaidi.kmd@nusamandiri.ac.id

comparable held-out performance may be preferable to a more elaborate one.

Several approaches have been proposed to make preprocessing choice less ad hoc. Response-oriented ensembles retain information from more than one preprocessed representation, and recent ensemble work has combined preprocessing with covariate selection [14, 15]. Piecewise preprocessing takes another route by assigning transformations to selected wavelength intervals instead of imposing one operation on the full spectrum [11]. Such methods can exploit complementary spectral information, but their output is not necessarily a short, explicit sequence that can be transferred directly to a constrained device.

Search-based schemes are closer to the selection problem considered here. Published implementations include automated open-source workflows, genetic and combinatorial searches, Bayesian optimization, and order-aware frameworks [16–21]. They reduce the amount of manual screening and make the search more repeatable. Even so, prediction error generally remains the primary objective; simplicity is either enforced through fixed constraints or absorbed into a penalty. Consequently, the analyst receives limited information about the actual trade-off between predictive gain and the structure of the resulting pipeline.

Preprocessing selection therefore involves two legitimate questions: how well does a candidate predict, and how complicated is it to implement and inspect? A single weighted score hides part of that choice. Pareto optimization instead retains candidates for which improvement in one objective would require a loss in the other. Evolutionary search through NSGA-II and Bayesian search through MOTPE offer distinct ways of exploring such non-dominated solutions [22–24]. Although these algorithms are established in multi-objective learning and engineering design, their explicit use for interpretable, deployment-aware NIR preprocessing has received much less attention than error-only search.

To address this gap, we developed a constrained multi-objective framework for selecting interpretable NIR preprocessing pipelines. It minimizes root mean square error of cross-validation (RMSECV) together with the number of active operations and permits only chemically defensible orders of SNV, MSC, detrending, and Savitzky–Golay transformations. Random Search provides a non-guided reference, while NSGA-II and MOTPE are compared under the same evaluation budget. PLSR was chosen as the primary calibration model so that gains from preprocessing could be examined without introducing an opaque predictor. The evaluation uses six public NIR and visible-NIR datasets, fixes the validation subsets before search, and compares the selected pipelines with manual references. Additional experiments examine optimizer stability over 10 seeds, nonlinear calibration, multi-target prediction, and joint wavelength selection. This design allows the framework to be assessed both when automated search helps and when a simple manual treatment remains the stronger choice.

2. Materials and Methods

2.1. Proposed framework and methodological contributions

Given calibration spectra X_{cal} and reference responses y_{cal} , the proposed framework generates ordered and parameterized preprocessing sequences, evaluates every feasible candidate through leakage-aware cross-validation, and returns a Pareto set that balances predictive error and structural pipeline complexity.

The framework output is therefore not a single opaque optimum but an auditable set of non-dominated alternatives. A representative solution is selected using calibration-derived objectives only, refitted on the complete calibration subset, and evaluated once on the held-out data; validation responses are not used to construct the Pareto front or choose the final pipeline.

The proposed methodology contributes four linked elements. First, an order-sensitive chemometric representation excludes repeated operations, mutually incompatible scatter corrections, and redundant Savitzky–Golay transformations. Second, simplicity is modeled explicitly as a competing objective rather than considered only after error minimization. Third, Random Search is retained only as a non-guided control, whereas NSGA-II and MOTPE provide matched evolutionary and Bayesian multi-objective comparisons. Fourth, the experimental protocol integrates Pareto quality, independent root mean square error of prediction (RMSEP), structural complexity, residual behavior, predefined manual references, a matched advanced-optimizer comparison, and methodological positioning against recent automated preprocessing approaches within one validation-aware assessment. Figure 1 summarizes the workflow and the separation between calibration-only optimization and final held-out testing.

2.2. Experimental protocol

The experiment comprised three phases (Figure 1). Phase I established six public spectral benchmarks and separated calibration from validation samples before any search was performed. Phase II generated feasible pipelines containing up to four ordered active operations and evaluated them using RMSECV and structural complexity. Random Search, NSGA-II, and MOTPE each received the same budget of 200 candidate evaluations per dataset. Phase III constructed the calibration-derived Pareto set, selected the non-dominated pipeline with the lowest RMSECV (using lower complexity as the tiebreaker), refitted the complete preprocessing-PLSR pipeline on all calibration samples, and assessed RMSEP and R^2 once on the held-out subset. The reported representative pipeline therefore follows an error-prioritized decision rule, while the complete Pareto set preserves alternative accuracy–complexity choices.

The same data partition, preprocessing library, PLSR configuration, and evaluation budget were used for all search methods. Corn was partitioned with group awareness so spectra from the same physical sample could not occur in both subsets, whereas the original CGL_NIR calibration-test division was retained. For the other datasets, calibration and validation sets were fixed before optimization. Data-dependent quantities, including the MSC reference spectrum, were estimated from training data only. These controls separate optimizer performance from final generalization and reduce information leakage [25].

2.3. Benchmark datasets

We selected six public NIR or visible-NIR datasets that differ in matrix, wavelength coverage, instrumentation, and response variable (Table 1). The collection spans corn grain, a controlled synthetic mixture, wheat, mango fruit, soil, and raw milk. Using heterogeneous benchmarks was important: a search procedure that works for only one matrix would provide little evidence that its accuracy–complexity choices are useful under other spectral conditions.

Corn contains measurements from several NIR instruments and therefore provides a test of instrument-related variation.

Figure 1
Experimental workflow of the proposed AI-driven multi-objective framework for interpretable NIR spectral preprocessing

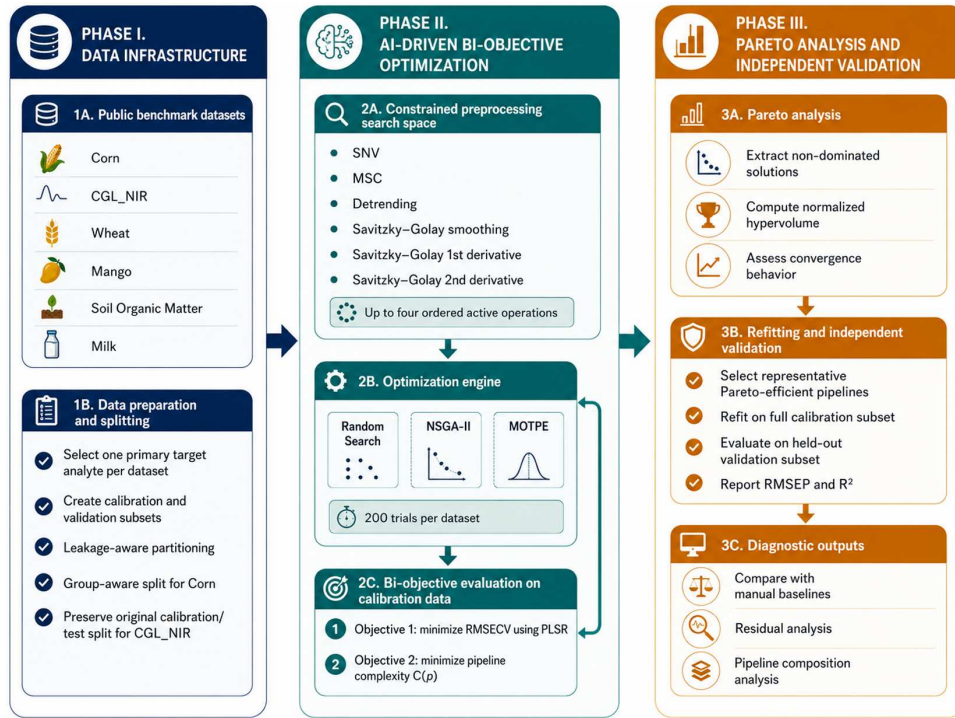


Table 1
Characteristics of the public benchmark datasets used for evaluating the proposed preprocessing framework

Dataset	Matrix	Spectra	Spectral range (nm)	Variables	Primary target	Instrument information
Corn	Grain	240	1100–2498	700	Moisture	m5, mp5, mp6
CGL_NIR	Synthetic mixture	231	1104–2495	117	Casein	NIRSystems 6500
Wheat	Cereal grain	149	1100–2500	700	Moisture	Foss NIRSYSTEM 6500/5000
Mango	Fruit	1000	309–1149	306	Dry matter	Portable fruit spectrometer
Soil Organic Matter	Soil	190	901–1700	230	Organic matter	Homemade NIR spectrometer
Milk	Dairy	1224	960–1690	256	Fat	Zeiss NIR 1.7-256

CGL_NIR is a laboratory mixture of casein, glucose, lactate, and moisture. The Wheat and Mango datasets use moisture and dry matter, respectively, as the primary responses. Soil Organic Matter was measured with a homemade NIR spectrometer, whereas the Milk spectra were acquired in transmittance mode and modeled for fat content [26]. Where a dataset supplied several response variables, one was designated in advance for the main experiment so that the optimization protocol remained comparable. Dataset locations and access details are reported in the Data Availability Statement.

2.4. Pipeline representation and constrained search space

The search library comprised six operations used routinely in NIR chemometrics: SNV, MSC, detrending, Savitzky–Golay smoothing, and first- or second-derivative Savitzky–Golay filtering. Together they address the principal disturbances considered

in this work, namely, scatter, baseline variation, high-frequency noise, and poorly resolved local features [5, 7, 9, 10]. Table 2 lists their roles, parameter ranges, and search constraints.

Each candidate was encoded as an ordered list of no more than four operation-parameter pairs. An operation could appear only once. Because SNV and MSC address related scatter effects, the two were not allowed in the same candidate. Likewise, a pipeline could contain only one Savitzky–Golay transformation. These restrictions removed redundant or difficult-to-defend sequences while leaving the order of compatible operations available for search [17, 19].

For a Savitzky–Golay operation, the search selected an odd window length from 5 to 25 and a polynomial order from 2 to 4. A proposed window was rejected when it exceeded the number of wavelengths available in the dataset. Candidates that produced non-finite values were not allowed to compete normally; they received a large error value. This handling kept numerical failures inside the recorded search while preventing them from being selected.

Table 2
Candidate preprocessing operations and constraints used in the constrained search space

Operation	Main purpose	Optimized parameter	Constraint
None	Skip a pipeline position	Not applicable	Allowed in any position
Standard normal variate	Reduce multiplicative scatter and scaling differences	Not applicable	Maximum once; not combined with MSC
Multiplicative scatter correction	Correct additive and multiplicative scatter effects	Reference spectrum estimated from training data	Maximum once; not combined with SNV
Detrending	Remove low-order baseline curvature	Polynomial order fixed at 2	Maximum once
Savitzky–Golay smoothing	Reduce high-frequency spectral noise	Window length; polynomial order	Maximum one SG-based operation per pipeline
Savitzky–Golay first derivative	Reduce baseline variation and enhance local spectral features	Window length; polynomial order	Not combined with other SG transformations
Savitzky–Golay second derivative	Enhance overlapping bands and curvature-related features	Window length; polynomial order	Not combined with other SG transformations

2.5. Bi-objective optimization and Pareto-solution selection

We treated preprocessing selection as the bi-objective minimization problem in Equation (1) [27]. One objective is the cross-validated prediction error, and the other is pipeline length. The second objective matters because an extra transformation is useful only when its predictive benefit warrants the added implementation and interpretation burden. Thus, a simpler candidate remains competitive whenever its validation-oriented calibration performance is close to that of a longer sequence.

Let p be a feasible preprocessing pipeline from the constrained set P . Equation (1) gives the joint minimization. Its error term, $f_1(p)$, is RMSECV as defined in Equation (2); its structural term, $f_2(p)$, is the complexity measure in Equation (3).

$$\min_{p \in \mathcal{P}} f(p) = [f_1(p), f_2(p)] \quad (1)$$

$$f_1(p) = \text{RMSECV}(p) \quad (2)$$

$$f_2(p) = C(p) \quad (3)$$

Here $C(p)$ is simply the number of active preprocessing operations: raw spectra have $C(p) = 0$ and a two-operation pipeline has $C(p) = 2$. It is a structural measure, not a timing benchmark. We used it because sequence length can be read directly, audited, and translated to a deployment workflow. Optimization returns a set of non-dominated candidates rather than one weighted optimum. For reporting the held-out result, we chose the Pareto candidate with the smallest calibration RMSECV and used $C(p)$ only to break a tie. The validation responses played no part in that choice.

All three searches received 200 evaluations for each dataset. Random Search sampled feasible candidates without guidance and served as the control. NSGA-II ranked a population by non-dominated sorting and maintained spread through crowding distance [22]. MOTPE instead estimated promising and less promising regions of the search space and used those distributions to propose subsequent trials [23]. Equal budgets were essential because they made differences in the discovered fronts attributable to search behavior rather than to unequal numbers of candidate evaluations.

2.6. Leakage-aware calibration and independent validation

PLSR served as the base predictor. It remains widely used for high-dimensional, collinear NIR spectra and retains a latent-variable structure that can be inspected [8, 28]. We capped the model at 10 latent variables and reduced that limit automatically when the calibration sample count or wavelength dimension required it. This gave the model enough flexibility for the benchmark tasks without permitting an unrestricted latent space.

Every preprocessing parameter was estimated inside the corresponding cross-validation fold. SNV, detrending, and Savitzky–Golay filters act on individual spectra and were applied within the fold workflow. MSC required additional care: its reference spectrum was computed from the training fold and then used unchanged for that fold’s validation spectra. At no point was a validation-fold spectrum included in an estimated preprocessing reference [25, 29].

Once the calibration Pareto front had been formed, the selected transformation and PLSR model were fit again using the entire calibration subset. That fitted pipeline was then evaluated once on the untouched held-out subset. Manual reference pipelines followed exactly the same split and fitting procedure. The reported RMSECV therefore describes selection within calibration data, whereas RMSEP measures subsequent generalization. Keeping those stages separate prevents the final validation result from steering pipeline choice.

2.7. Evaluation metrics and comparative protocol

Prediction performance during optimization was measured using RMSECV, while final generalization performance was evaluated using the RMSEP on the held-out validation subset. RMSEP was used as the main metric for comparing optimized pipelines with manual preprocessing baselines because it reflects prediction performance on samples that were not used during optimization. The coefficient of determination (R^2) was also reported to describe the proportion of validation-set variance explained by the model.

Pareto-front quality was evaluated using normalized hypervolume and Pareto-set size [24, 30]. Within each dataset, the same objective normalization and reference definition were applied

to Random Search, NSGA-II, and MOTPE. A larger normalized hypervolume indicates that the discovered front is closer to the desirable objective region and/or provides broader useful coverage, but it does not replace independent validation.

Equation (4) expresses the RMSEP change relative to the strongest manual reference. For the controlled algorithm comparison, NSGA-II and MOTPE used the same datasets, partitions, PLSR settings, search rules, and 200-trial budget; Random Search remained a non-guided control. We also compared the framework's design with recent automated preprocessing studies [16–19, 21]. Their published errors were not combined because the underlying datasets, target variables, operation libraries, calibration models, and validation protocols differ. Pooling those values would imply a numerical comparison that the study designs do not support.

$$\text{Improvement (\%)} = \frac{RMSEP_{\text{manual}} - RMSEP_{\text{optimized}}}{RMSEP_{\text{manual}}} \times 100 \quad (4)$$

2.8. Implementation and reproducibility

The experiments were run in Python. NumPy and pandas handled arrays and data organization, SciPy and project-specific functions implemented the spectral operations, and scikit-learn supplied PLSR, cross-validation, and evaluation metrics. Optuna was used to organize the Random Search, NSGA-II, and MOTPE studies [31]. The same fixed partitions, preprocessing rules, PLSR limits, and trial budgets were passed to each search. The extension analyses used scikit-learn's radial-basis-function support vector regression (SVR) and XGBoost regression. Processed results, optimization histories, and source code will accompany the public repository so that candidate generation and selection can be reproduced.

2.9. Robustness and extension experiments

We ran four extensions while leaving every held-out partition unchanged. The first examined stochastic stability. NSGA-II

and MOTPE were rerun with seeds 11, 22, 33, 44, 55, 66, 77, 88, 99, and 111; each run again contained 200 evaluations. For every optimizer-dataset pair, we recorded held-out RMSEP, R^2 , normalized hypervolume, pipeline complexity, Pareto-set size, and the frequency of the selected pipeline. For this analysis only, hypervolume was normalized jointly across the 10 seeds and both optimizers within each dataset. It therefore measures variation between runs and should not be compared numerically with the single-run hypervolumes in Table 3.

The second extension held preprocessing fixed and changed the calibration model. We took the most frequent pipeline from the 10-seed experiment and paired it with PLSR, radial-basis-function SVR ($C = 10$, $\epsilon = 0.1$, $\gamma = \text{scale}$), or XGBoost (250 trees, maximum depth 4, learning rate 0.05, subsample 0.8, and column subsample 0.8). Because no new preprocessing search was run for these comparisons, any change in held-out error can be attributed to the calibration model rather than to a different transformation sequence.

Third, datasets containing multiple response variables were re-optimized in a multi-target setting using PLS2. Because response variables have different units, the optimization objective was the macro-average normalized root mean square error (RMSE), obtained by dividing each target RMSE by its calibration-set standard deviation before averaging. Corn used moisture, oil, protein, and starch; CGL_NIR used casein, glucose, lactate, and moisture; Wheat used moisture and protein; Soil Organic Matter used organic matter and total carbon; and Milk used fat, protein, and lactose. Mango was not included because only dry matter was available in the analyzed dataset. A fixed budget of 120 evaluations per optimizer was used for this extension.

Fourth, wavelength selection was integrated as an exploratory third objective together with cross-validated error and preprocessing complexity. The selector searched an interpretable contiguous wavelength interval with a width between 25% and 90% of the available variables, while also allowing the full spectrum. The selected wavelength fraction was minimized as the third objective, and each optimizer received 120 evaluations per

Table 3
Pareto-front quality under constrained preprocessing optimization

Dataset	Algorithm	Normalized hypervolume	Pareto size
Corn	Random Search	0.7434	2
	NSGA-II	0.7500	1
	MOTPE	0.7444	4
CGL_NIR	Random Search	0.7410	2
	NSGA-II	0.7491	1
	MOTPE	0.7491	2
Wheat	Random Search	0.7228	3
	NSGA-II	0.7175	2
	MOTPE	0.7291	4
Mango	Random Search	0.7495	2
	NSGA-II	0.7497	3
	MOTPE	0.7498	4
Soil Organic Matter	Random Search	0.7500	1
	NSGA-II	0.7469	2
	MOTPE	0.7481	3
Milk	Random Search	0.7446	2
	NSGA-II	0.7421	2
	MOTPE	0.7443	4

dataset. Final assessment again used the unchanged held-out subset. To investigate the Wheat exception, sample-wise multiplicative slope and additive offset dispersion were estimated relative to the mean calibration spectrum before and after SNV, and Raw, manual SNV, and the most frequently selected optimized Wheat pipeline were compared under PLSR, SVR, and XGBoost.

3. Results and Discussion

3.1. Pareto-front quality

A non-dominated preprocessing set was obtained for every benchmark (Figure 2). Each point is one pipeline, located by its RMSECV and number of active operations; connecting lines identify points returned by the same search. The fronts are generally concentrated at low complexity. In practical terms, one or two operations were enough to reach most of the useful error-complexity region. Adding further transformations rarely delivered a commensurate calibration gain. This pattern is chemically plausible because preprocessing is intended to remove a dominant nuisance effect, not to accumulate transformations without a clear spectral purpose [5, 7].

Table 3 reports normalized hypervolume and the number of Pareto solutions. The three algorithms often produced similar hypervolumes, although the ordering changed by dataset. NSGA-II was highest for Corn and shared the highest value with MOTPE for CGL_NIR. MOTPE led for Wheat and Mango, whereas Random Search was highest for Soil Organic Matter and Milk. No method therefore controlled the full set of benchmarks. The small gaps also suggest that the chemical constraints removed much of the unproductive search space, leaving several methods able to reach useful low-complexity candidates.

The absence of a universal winner is consistent with how the two guided optimizers search. NSGA-II keeps a population and protects diversity through sorting and crowding distance [22]. MOTPE learns probability models from completed trials and samples regions judged more promising [23]. Which mechanism is favored depends on the interaction among the spectral matrix, allowed preprocessing orders, and PLSR response surface; that interaction was different for each dataset.

Hypervolume alone did not identify the pipeline that generalized best. It summarizes coverage and dominance in the calibration objective space, while validation RMSEP answers a different question: how the selected pipeline predicts samples excluded from search. In several datasets, the optimizer with the larger front indicator did not produce the lowest held-out error. We therefore interpret Pareto quality alongside, rather than in place of, independent validation [24, 25].

3.2. Independent validation performance

Table 4 compares the two guided multi-objective optimizers under the controlled held-out protocol. NSGA-II gave the lower RMSEP for Corn, Mango, and Soil Organic Matter; MOTPE gave the lower value for Wheat; and the two methods tied for CGL_NIR and Milk. When ranks were averaged across these six datasets, NSGA-II had the better validation-RMSEP rank (1.33 versus 1.67), while MOTPE had the better normalized-hypervolume rank (1.25 versus 1.75). NSGA-II was best or tied for RMSEP in five datasets and MOTPE in three. The difference between these rankings indicates that Pareto-front coverage and the generalization of the selected representative pipeline capture different aspects of optimizer performance.

The validation results varied substantially across datasets. CGL_NIR produced an almost perfect validation (R^2), which

Figure 2
Pareto fronts obtained from constrained multi-objective preprocessing optimization across six benchmark datasets: (a) Corn, (b) CGL_NIR, (c) Wheat, (d) Mango, (e) Soil Organic Matter, and (f) Milk

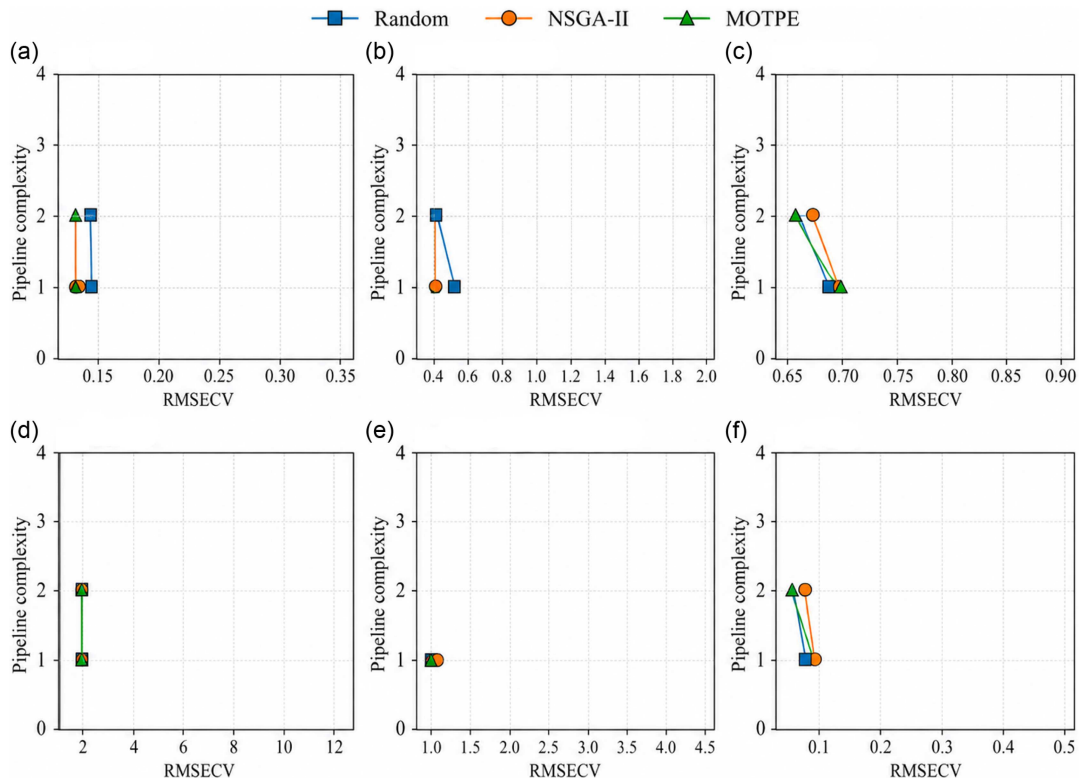


Table 4
Independent validation performance of optimized preprocessing pipelines

Dataset	Algorithm	RMSECV	RMSEP	(R^2)	Complexity	Selected pipeline
Corn	NSGA-II	0.1362	0.1333	0.8626	1	SG first derivative
	MOTPE	0.1374	0.1366	0.8557	2	SG first derivative → detrend
CGL_NIR	NSGA-II	0.5035	0.5202	0.9993	1	SG second derivative
	MOTPE	0.5035	0.5202	0.9993	1	SG second derivative
Wheat	NSGA-II	0.6729	0.7099	0.7114	2	SG first derivative → detrend
	MOTPE	0.6627	0.7062	0.7144	2	Detrend → SG first derivative
Mango	NSGA-II	1.9502	2.1355	0.3170	2	SG smoothing → detrend
	MOTPE	1.9492	2.1448	0.3111	2	SG smoothing → detrend
Soil Organic Matter	NSGA-II	0.8870	0.8490	0.7236	1	SG smoothing
	MOTPE	0.8805	0.8588	0.7172	1	SG smoothing
Milk	NSGA-II	0.0797	0.1169	0.9796	2	MSC → SG smoothing
	MOTPE	0.0797	0.1169	0.9796	2	SG smoothing → MSC

is plausible because it represents a controlled mixture system with well-defined constituent variation. Milk also showed strong predictive performance, with validation (R^2) close to 0.98. The selected Milk pipelines combined smoothing and scatter correction, which is consistent with the need to stabilize transmittance-mode milk spectra affected by scattering and physical variation in dairy constituents [16, 26].

Corn and Soil Organic Matter yielded moderate-to-strong held-out performance, but for different apparent reasons. On Corn, the NSGA-II solution was a single Savitzky–Golay first derivative and gave the smallest RMSEP. Removing baseline structure without a longer sequence was sufficient even though the data include multiple instruments. Soil Organic Matter instead favored Savitzky–Golay smoothing, which points to noise attenuation rather than derivative enhancement as the useful correction for that dataset. The contrast agrees with soil spectroscopy evidence that preprocessing effects depend on spectral coverage, matrix, and target property [32, 33].

Wheat and Mango remained more difficult. The lower validation performance for Mango suggests that constrained linear preprocessing combined with PLSR did not fully capture the variability of fruit spectra. Fruit spectroscopy is often affected by biological heterogeneity, tissue structure, maturity stage, moisture distribution, and scattering effects; in some cases, preprocessing may even remove variation that is physically related to the target quality attribute [12, 34, 35]. This result supports the view that preprocessing optimization should not be evaluated

only as a calibration-improvement procedure, but as part of a broader modeling workflow that may also require nonlinear models, wavelength selection, or external covariates. The additional model and wavelength-selection experiments in Section 3.7 test these possibilities directly. The optimized pipelines were generally short. Corn, CGL_NIR, and Soil Organic Matter required only one preprocessing operation, while Wheat, Mango, and Milk required two operations. The result is notable because the search did not gain accuracy simply by accumulating transformations; it repeatedly selected compact pipelines that balance predictive performance and interpretability. Such behavior is desirable for portable and field-oriented NIR systems, where computational simplicity, auditability, and ease of deployment are practical requirements [4, 13].

3.3. Comparison with manual baselines

Table 5 compares the best optimized pipeline with the best manual preprocessing baseline for each dataset. Positive values indicate that the optimized pipeline reduced validation RMSEP relative to the best manual baseline, while negative values indicate that the manual baseline generalized better. The optimized pipelines reduced validation error in five of the six datasets. The largest relative improvement was obtained for Milk, where the optimized pipeline reduced RMSEP by 16.88% compared with the best manual baseline. CGL_NIR and Corn also showed clear improvements, with RMSEP reductions of 12.30% and 10.32%,

Table 5
Comparison between the best manual baseline and the best optimized preprocessing pipeline

Dataset	Best manual baseline	Best manual RMSEP	Best optimized algorithm	Best optimized RMSEP	RMSEP change (%)	Best optimized pipeline
Corn	Raw	0.1486	NSGA-II	0.1333	10.32 %	SG first derivative
CGL_NIR	SNV + SG first derivative	0.5932	NSGA-II	0.5202	12.30 %	SG second derivative
Wheat	SNV	0.6000	MOTPE	0.7062	−17.70 %	Detrend → SG first derivative
Mango	Raw	2.2191	NSGA-II	2.1355	3.76 %	SG smoothing → detrend
Soil Organic	Raw	0.8861	NSGA-II	0.8490	4.20 %	SG smoothing
Milk	SNV	0.1407	MOTPE	0.1169	16.88 %	SG smoothing → MSC

respectively. Mango and Soil Organic Matter showed more modest improvements, indicating that the benefit of preprocessing optimization depends on how well the available transformations address the dominant source of irrelevant spectral variation.

A strong manual comparator is necessary for interpreting these gains. Some automation studies report improvement over only a narrow set of conventional treatments. Open-source screening, genetic search, Bayesian optimization, Combinatorial Order Pre-processing Search (COPS), and other automated frameworks have already reduced manual effort and improved repeatability [16–19]. The present experiment adds a different comparison: candidates compete on both error and pipeline length, and the selected result is still tested against the best predefined manual baseline.

Wheat is the informative exception. Manual SNV produced a lower validation RMSEP than any selected optimized pipeline. Rather than treating this outcome as a failure to be hidden, we use it to define the framework’s role: it supports a preprocessing decision but does not guarantee that automation will beat a simple treatment. For these spectra, SNV appears to remove the principal scatter component without the additional alteration introduced by a derivative or detrending step. Similar reversals have been reported when preprocessing interacts unfavorably with a dataset or its validation split [5, 7, 12].

The diagnostic calculations support that interpretation. Across the 10-seed runs, the most frequent Wheat solution was Savitzky–Golay smoothing with window 25 and polynomial order 3; with PLSR, it gave RMSEP = 0.6167 and $R^2 = 0.7822$. Manual SNV gave RMSEP = 0.6000 and $R^2 = 0.7939$. After SNV, the between-sample standard deviation of multiplicative slope fell from 0.05973 to 0.000852, and that of the additive offset fell from 0.006246 to approximately 2.3×10^{-16} . Scatter was therefore a major nuisance component. Because moisture has a broad NIR response, adding smoothing, differentiation, or detrending after scatter correction may redistribute useful low-frequency structure. Changing the model did not overturn the result: SNV+SVR gave RMSEP = 0.6322, and SNV+XGBoost gave 0.8093, both above SNV+PLSR [7, 10].

This case also shows why cross-validation and a Pareto front are not sufficient on their own. Had the manual baseline or held-out subset been omitted, the optimized Wheat pipeline

would have looked favorable despite its weaker generalization. The independent comparison exposes that difference and guards against the optimistic interpretation that can follow leakage or repeated use of validation data [25].

3.4. Matched optimizer comparison and methodological positioning

The matched experiment directly compares the two advanced multi-objective optimizers instantiated within the proposed framework. As summarized in Table 6, NSGA-II favored the held-out performance of the selected representative solution, whereas MOTPE provided stronger average Pareto-front coverage. Random Search is reported in Table 3 solely as a non-guided control and is not treated as an external automated preprocessing comparator.

Table 6 places the framework beside recent Bayesian, combinatorial, parsimonious, and multi-objective preprocessing searches [16, 17, 19, 21]. This is a design-level comparison, not a pooled performance test. The published studies use different spectra, response variables, operation libraries, regression models, and validation protocols, so their RMSEP values do not share a common denominator. The table accordingly separates evidence generated under the matched experiment from methodological similarities reported across studies.

3.5. Convergence and multi-seed stability

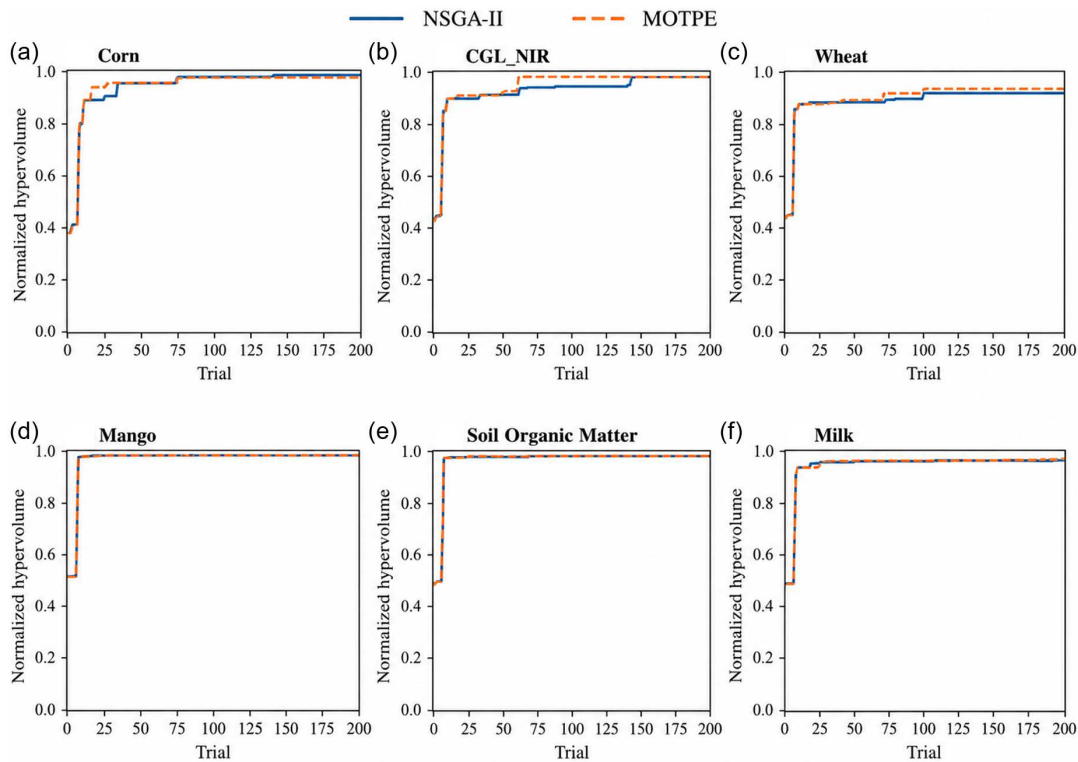
The selected and Pareto-efficient pipelines remained dataset-specific. Savitzky–Golay first derivatives were prominent for Corn and Wheat, the second derivative characterized CGL_NIR, smoothing appeared for Mango, Soil Organic Matter, and Milk, MSC was mainly associated with Milk, and detrending appeared for Wheat and Mango. This composition confirms that the search did not converge to a universal preprocessing recipe. Figure 3 presents cumulative normalized hypervolume over 200 evaluations. Most curves improved rapidly and then approached a plateau, indicating that the constrained space was tractable under the present configuration.

Convergence remained dataset-dependent, reflecting differences in sample heterogeneity, target structure, and preprocessing-model interactions. To quantify stochastic repeatability, both

Table 6
Matched optimizer comparison and methodological positioning against recent automated preprocessing approaches

Approach	Search formulation	Simplicity treatment	Comparison evidence	Validation scope
Framework with NSGA-II	Evolutionary Pareto search	Explicit structural $C(p)$	RMSEP rank 1.33; HV rank 1.75	Six datasets; matched budget; held-out validation
Framework with MOTPE	Bayesian Pareto search	Explicit structural $C(p)$	RMSEP rank 1.67; HV rank 1.25	Six datasets; matched budget; held-out validation
Bayesian optimization [16]	Bayesian single-objective search	Prediction-oriented; no explicit Pareto complexity	Cross-study positioning	Milk spectroscopy study
COPS [17]	Order-aware combinatorial search	Chemical and sequence constraints	Cross-study positioning	Process analytical technology data
Automated framework [19]	Automated parsimonious search	Complexity handled through search design/penalty	Cross-study positioning	Reported NIR benchmark studies
MOBO [21]	Multi-objective Bayesian search	Multiple preprocessing objectives	Cross-study positioning	Application-specific NIR data

Figure 3
Hypervolume convergence during constrained multi-objective preprocessing optimization: (a) Corn, (b) CGL_NIR, (c) Wheat, (d) Mango, (e) Soil Organic Matter, and (f) Milk



guided optimizers were rerun with 10 fixed seeds using the same 200-evaluation budget. Held-out RMSEP mean $\pm$ SD values for NSGA-II/MOTPE were 0.1341 ± 0.0018 / 0.1350 ± 0.0023 for Corn, 0.5201 ± 0.0157 / 0.4995 ± 0.0390 for CGL_NIR, 0.6151 ± 0.0447 / 0.6137 ± 0.0447 for Wheat, 2.1416 ± 0.0395 / 2.1596 ± 0.0372 for Mango, 0.8419 ± 0.0091 / 0.8379 ± 0.0166 for Soil Organic Matter, and 0.1182 ± 0.0016 / 0.1173 ± 0.0010 for Milk. Thus, predictive performance was highly stable for Corn and Milk, remained reasonably stable for Soil Organic Matter and Mango, and showed greater seed sensitivity for CGL_NIR and Wheat.

Pareto-front quality also varied across seeds, with normalized-hypervolume standard deviations ranging from 0.015 to 0.222 depending on dataset and optimizer. These stability hypervolumes used a common within-dataset normalization across all seeds and are therefore not numerically comparable with Table 3. Exact representative-pipeline repeatability ranged from 20% to 80%: the most frequent Corn NSGA-II pipeline was recovered in 70% of runs and the most frequent Milk MOTPE pipeline in 80%, whereas Wheat and Mango showed 20% repeatability. The stronger stability of RMSEP than of exact pipeline identity indicates that multiple preprocessing sequences can occupy similarly useful regions of the Pareto front, which is consistent with the decision-support interpretation of the framework.

3.6. Residual analysis

Figure 4 plots held-out residuals against reference values for the NSGA-II and MOTPE pipelines. The dashed horizontal line marks zero error. CGL_NIR and Milk form tight clouds around that line, consistent with their high validation R^2 values. Corn and Soil Organic Matter show a broader, but still moderate,

spread. Wheat and Mango differ: their residuals are wider and display visible structure rather than random scatter. The graphical evidence therefore agrees with the RMSEP comparison in Table 4 and also shows where similar summary errors conceal different residual patterns.

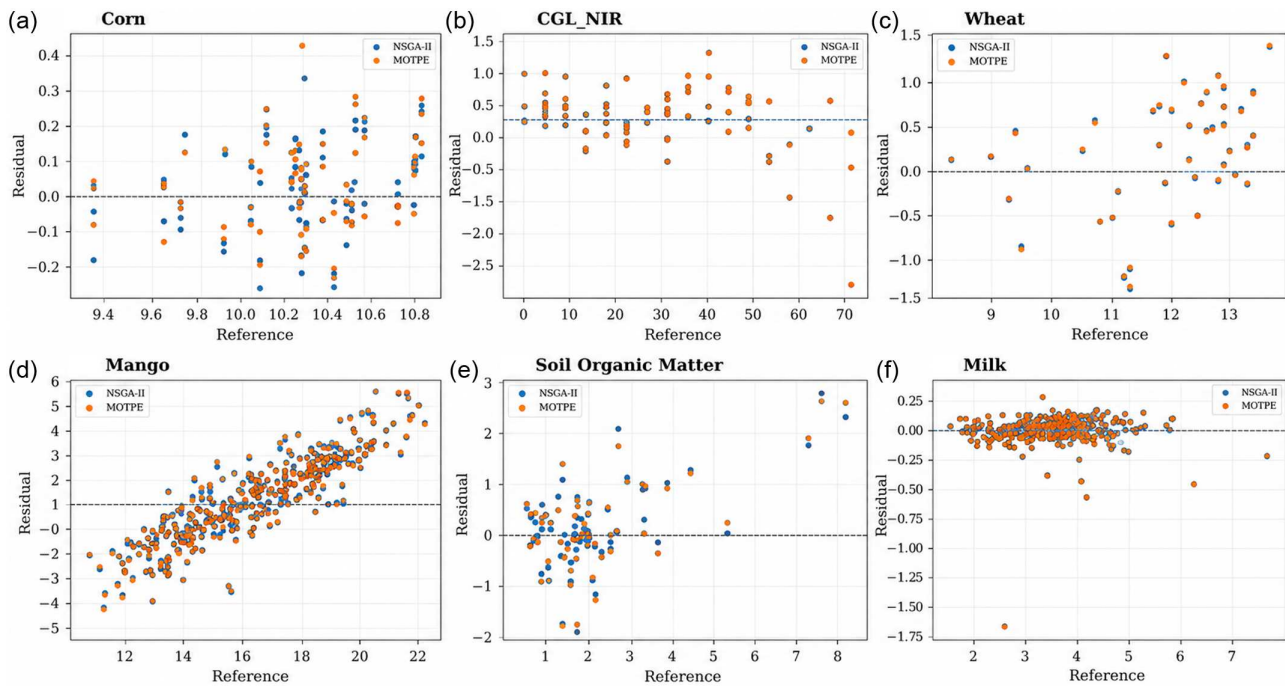
The residual plots also clarify why preprocessing cannot be prescribed uniformly. For the controlled CGL_NIR mixture and for Milk, the selected operations appear to suppress nuisance variation while retaining the response signal. Mango retains a trend across the reference range, suggesting a modeling limitation beyond the order of preprocessing alone. This is credible for intact fruit, where path length, skin and surface properties, internal tissue, and maturity vary together [12, 34, 35]. Wheat shows a different problem. Its residual behavior and the lower error from manual SNV indicate that direct scatter correction was more robust on the chosen split than the derivative-based solutions returned by search.

Taken together, the held-out results favor the optimized pipeline for five datasets, most clearly for Milk, CGL_NIR, and Corn. Gains for Mango and Soil Organic Matter are smaller, and Wheat favors the manual reference. The framework should therefore be read as a way to expose and test concise accuracy-complexity choices. It does not remove the need for chemometric judgment, nor does it justify selecting a pipeline from calibration indicators without an independent check.

3.7. Nonlinear, multi-target, and wavelength-selection extensions

Using the most frequent preprocessing pipeline from the stability experiment, PLSR remained the best held-out calibration model for Corn (RMSEP = 0.1332), CGL_NIR (0.4809),

Figure 4
Residual analysis of optimized preprocessing pipelines on independent validation subsets: (a) Corn, (b) CGL_NIR, (c) Wheat, (d) Mango, (e) Soil Organic Matter, and (f) Milk



Wheat (0.6167), Soil Organic Matter (0.8383), and Milk (0.1169). Mango was the clear exception: with the MOTPE-derived SG second-derivative $\rightarrow$ detrend pipeline, SVR and XGBoost reduced RMSEP from 2.1896 with PLSR to 1.4956 and 1.4946, respectively, with XGBoost reaching $R^2 = 0.6504$. This result confirms a meaningful preprocessing-model interaction for the biologically heterogeneous fruit dataset, while also showing that nonlinear models are not universally advantageous.

The multi-target extension also preserved compact preprocessing solutions. Validation macro-averaged normalized root mean square error (macro-NRMSE) for NSGA-II/MOTPE was 0.3997/0.3989 for Corn, 0.0258/0.0282 for CGL_NIR, 0.4124/0.4124 for Wheat, 0.5124/0.5308 for Soil Organic Matter, and 0.5810/0.5730 for Milk, with selected pipelines containing only one or two operations. These results demonstrate that the optimization can be applied to multiple chemical responses simultaneously rather than only to a single target, although target-specific difficulty remained evident, particularly for the Milk protein and lactose responses.

The tri-objective wavelength-selection extension produced strongly dataset-dependent dimensionality reductions. The best held-out configurations included CGL_NIR with 92/117 wavelengths (21.4% reduction, NSGA-II RMSEP = 0.3520), Wheat with 499/700 wavelengths (28.7% reduction, MOTPE RMSEP = 0.5829, $R^2 = 0.8055$), Mango with 82/306 wavelengths (73.2% reduction, MOTPE RMSEP = 1.0307, $R^2 = 0.8337$), and Milk with 92/256 wavelengths (64.1% reduction, MOTPE RMSEP = 0.1083, $R^2 = 0.9825$). Corn under NSGA-II and Soil Organic Matter under both optimizers retained the full spectral range, showing that aggressive wavelength reduction is not universally beneficial. Notably, the Wheat wavelength-selection result surpassed the manual SNV reference (RMSEP = 0.6000), suggesting that its original preprocessing-only limitation arose partly from spectral redundancy rather than from a failure of multi-objective optimization.

The extensions change the scope of the conclusion without making it universal. The same decision framework can accommodate another regression family, several responses, or a wavelength-range objective, but the best combination remained specific to the dataset. Added flexibility brought substantial gains in some cases and none in others. This uneven response is precisely why the held-out partition was kept fixed for every extension.

4. Conclusion

We formulated NIR preprocessing as a constrained choice between calibration error and pipeline length. Across six public NIR and visible-NIR datasets, the resulting search usually selected one or two operations and reduced held-out RMSEP relative to the strongest manual reference in five cases. The benefit was most pronounced for Milk, CGL_NIR, and Corn and more modest for Mango and Soil Organic Matter. Wheat did not follow that pattern: direct SNV generalized better than the preprocessing-only search results. This counterexample is central to the interpretation. Automated search is useful when it exposes defensible alternatives, but its calibration optimum cannot be assumed to outperform a simple baseline on new samples.

NSGA-II and MOTPE also emphasized different aspects of performance. NSGA-II had the better mean rank for validation RMSEP, whereas MOTPE had the better mean hypervolume rank. Coverage of the calibration Pareto front and generalization of one selected pipeline are therefore complementary rather than interchangeable criteria. The practical contribution of the framework is the combination of order-aware chemometric constraints, an explicit error-complexity front, leakage-controlled fitting, and a final comparison with held-out data and manual treatments.

The robustness experiments add two qualifications. First, held-out error was often stable even when the exact selected sequence changed across seeds, indicating that several short

pipelines can perform similarly. Second, extra modeling flexibility was useful only where the data supported it: nonlinear calibration chiefly improved Mango, while joint wavelength selection reduced the Wheat RMSEP to 0.5829, below the manual SNV value of 0.6000. Multi-target searches likewise retained pipelines containing only one or two operations.

Further work should evaluate the framework under external instrument and field shifts rather than only fixed benchmark partitions. Joint tuning of preprocessing and nonlinear-model parameters, repeated-seed experiments with larger budgets, and chemically informed non-contiguous wavelength selection are also needed. Such tests would show whether the present stability and dimensionality reductions persist under conditions closer to deployment.

Acknowledgment

The authors are grateful to the researchers who made the six benchmark datasets openly accessible. Those resources allowed the same constrained search and held-out protocol to be examined across instruments and sample matrices. We also acknowledge the maintainers of the open-source numerical libraries used to run and document the experiments.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the Eigenvector Research Inc. Standardization Benchmark Archives (Corn dataset) at <http://www.eigenvector.com/data/Corn>, in the Eigenvector Research Data Repository (CGL_NIR dataset) at <https://eigenvector.com/resources/data-sets/>, in Zenodo (Wheat dataset) at <https://doi.org/10.5281/zenodo.16759587>, in Mendeley Data (Mango dataset) at <https://doi.org/10.17632/46htwnp833.3>, in Mendeley Data (Soil Organic Matter dataset) at <https://doi.org/10.17632/yt78nwnhbd.1>, and in Data in Brief (Milk dataset) at <https://doi.org/10.1016/j.dib.2023.109767>.

Author Contribution Statement

Ali Khumaidi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Supervision, Project administration. **Ridwan Raafi'udin:** Software, Validation, Formal analysis, Data curation, Writing – review & editing.

References

- [1] Gao, Y., & Chu, X. L. (2026). Advances in band assignment in near-infrared spectroscopy: Principles, Methods, and Applications. *Talanta*, 308, 129851. <https://doi.org/10.1016/j.talanta.2026.129851>
- [2] da Costa Barbon, A. P. A., Nolasco-Perez, I. M., Barbon Jr, S., Máquina, A. D. V., Su, W. H., & Barbin, D. F. (2026). Process analytical technologies applied to quality control of emerging alternative protein food products: Challenges and future trends. *Journal of Pharmaceutical and Biomedical Analysis Open*, 7, 100105. <https://doi.org/10.1016/j.jpba.2026.100105>
- [3] Wang, D., Zhang, T., Ding, K., Tang, H., & Li, H. (2025). Near-infrared spectroscopy for energetic materials: Instruments, chemometrics, and process analytics (2001–2025). *TrAC Trends in Analytical Chemistry*, 196, 118631. <https://doi.org/10.1016/j.trac.2025.118631>
- [4] Yan, H., Neves, M. D., Wise, B. M., Moraes, I. A., Barbin, D. F., & Siesler, H. W. (2023). The application of hand-held near-infrared spectroscopy and Raman spectroscopic imaging for the identification and quality control of food products. *Molecules*, 28(23), 7891. <https://doi.org/10.3390/molecules28237891>
- [5] Yan, C. (2025). A review on spectral data preprocessing techniques for machine learning and quantitative analysis. *IScience*, 28(7), 1–42. <https://doi.org/10.1016/j.isci.2025.112759>
- [6] Beć, K. B., Grabska, J., & Huck, C. W. (2025). Interpretability in near-infrared (NIR) spectroscopy: Current pathways to the long-standing challenge. *TrAC Trends in Analytical Chemistry*, 189, 118254. <https://doi.org/10.1016/j.trac.2025.118254>
- [7] Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10), 1201–1222. <https://doi.org/10.1016/j.trac.2009.07.007>
- [8] Wahab, M., Zia, A., Ahmad, W., Zakarya, M., & Haleem, M. (2026). Optimizing NIR spectral preprocessing for PLS predictive accuracy: A systematic review (2019–2024). *Discover Applied Sciences*, 8(5), 560. <https://doi.org/10.1007/s42452-026-08608-y>
- [9] Esquivel, F. J., Romero-Béjar, J. L., & Esquivel, J. A. (2024). Preprocessing of spectroscopic data to highlight spectral features of materials. *Analytical Science Advances*, 5(9–10), e2400018. <https://doi.org/10.1002/ansa.202400018>
- [10] Roger, J. M., Mallet, A., & Marini, F. (2022). Preprocessing NIR spectra for aquaphotomics. *Molecules*, 27(20), 6795. <https://doi.org/10.3390/molecules27206795>
- [11] Yang, W., Xiong, Y., Xu, Z., Li, L., & Du, Y. (2022). Piecewise preprocessing of near-infrared spectra for improving prediction ability of a PLS model. *Infrared Physics & Technology*, 126, 104359. <https://doi.org/10.1016/j.infrared.2022.104359>
- [12] Mishra, P., Rutledge, D. N., Roger, J. M., Wali, K., & Khan, H. A. (2021). Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. *Talanta*, 229, 122303. <https://doi.org/10.1016/j.talanta.2021.122303>
- [13] Khumaidi, A., Raafi'udin, R., Saludin, S., Susanti, D., & Budiarto, R. (2025). Embedded system design and control of a portable Vis/NIR spectroscopy with hybrid XGBoost-ANFIS for soybean seed quality prediction. *International Journal of Robotics and Control Systems*, 6(1), 143–172. <https://doi.org/10.31763/ijrcs.v6i1.2392>
- [14] Mishra, P., Roger, J. M., Marini, F., Biancolillo, A., & Rutledge, D. N. (2022). Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): A step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling. *Chemometrics and Intelligent Laboratory Systems*, 222, 104497. <https://doi.org/10.1016/j.chemolab.2022.104497>

- [15] Wu, Y., Zhou, Y., Chen, X., Xie, Z., Ali, S., Huang, G., . . . , & Zhang, L. (2025). Pre-processing ensemble modeling based on faster covariate selection calibration for near-infrared spectroscopy. *Applied Sciences*, 15(21), 11325. <https://doi.org/10.3390/app152111325>
- [16] Babatunde, H. A., McDougal, O. M., & Andersen, T. (2025). Automated spectral preprocessing via Bayesian optimization for chemometric analysis of milk constituents. *Foods*, 14(17), 2996. <https://doi.org/10.3390/foods14172996>
- [17] Cardoso, W., Roque, J. V., Jansen, J. J., Teng, S. Y., & Teófilo, R. F. (2025). Combinatorial Order Pre-processing Search (COPS): A new pre-processing strategy for large-scale interpretable data analysis in process analytical technologies. *Computers & Chemical Engineering*, 192, 108892. <https://doi.org/10.1016/j.compchemeng.2024.108892>
- [18] Sitorus, A., & Lapcharoensuk, R. (2024). Development of automatic tuning for combined preprocessing and hyperparameters of machine learning and its application to NIR spectral data of coconut milk adulteration. *Food Chemistry*, 457, 140108. <https://doi.org/10.1016/j.foodchem.2024.140108>
- [19] Chen, X., Xie, Z., Tauler, R., He, Y., Nie, P., Peng, Y., . . . , & Yuan, L. (2025). An automated preprocessing framework for near infrared spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 267, 105542. <https://doi.org/10.1016/j.chemolab.2025.105542>
- [20] Du, Y., Wang, H., Hou, B., Zan, J., Wu, J., Sun, T., & Song, L. (2026). Non-destructive detection of internal mold in Torreyia grandis seeds using NIR spectroscopy and deep learning with integrated preprocessing and variable selection. *Journal of Food Composition and Analysis*, 150, 108899. <https://doi.org/10.1016/j.jfca.2026.108899>
- [21] Ziyang, P., Fu, H., Zhao, J., & Wang, N. (2026). A global optimization strategy for the preprocessing combination of near-infrared spectroscopy data based on Multi-Objective Bayesian Optimization (MOBO). *Analytical Letters*, 59(10), 1791–1806. <https://doi.org/10.1080/00032719.2025.2568906>
- [22] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197. <https://doi.org/10.1109/4235.996017>
- [23] Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M., & Onishi, M. (2022). Multiobjective tree-structured parzen estimator. *Journal of Artificial Intelligence Research*, 73, 1209–1250. <https://doi.org/10.1613/jair.1.13188>
- [24] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., & Da Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on evolutionary computation*, 7(2), 117–132. <https://doi.org/10.1109/TEVC.2003.810758>
- [25] Lopez, E., Gorla, G., Etxebarria-Elezgarai, J., Amigo, J. M., & Seifert, A. (2025). The importance of choosing a proper validation strategy in predictive models. Part 2: Recipes for (avoiding) overfitting. *Analytica Chimica Acta*, 1384, 344838. <https://doi.org/10.1016/j.aca.2025.344838>
- [26] Diaz-Olivares, J. A., van Nuenen, A., Gote, M. J., Díaz, V. F., Saeys, W., Adriaens, I., & Aernouts, B. (2023). Near-infrared spectra dataset of milk composition in transmittance mode. *Data in Brief*, 51, 109767. <https://doi.org/10.1016/j.dib.2023.109767>
- [27] Deeb, A., Khokhlovskiy, V., & Shkodyrev, V. (2026). Hierarchical multi-objective control of nonlinear systems with dynamical input constraints. *Artificial Intelligence and Applications*, 4(1), 57–67. <https://doi.org/10.47852/bonviewAIA52024314>
- [28] Deryck, A., Pierna, J. A. F., Baeten, V., & Lesnoff, M. (2026). A step-by-step tutorial to local partial least squares analyses for near-infrared spectroscopic data analysis. *Analytica Chimica Acta*, 1401, 345242. <https://doi.org/10.1016/j.aca.2026.345242>
- [29] Irshad, A., Muncan, J., & Omar, A. F. (2026). Principal component analysis (PCA)-integrated machine learning (ML) pipelines for spectral modelling of Harum Manis mangoes. *European Food Research and Technology*, 252(6), 224. <https://doi.org/10.1007/s00217-026-05148-y>
- [30] White, S. E., Witing, F., Wittekind, C. I., Volk, M., & Strauch, M. (2025). Distilling the Pareto optimal front into actionable insights. *Environmental Modelling & Software*, 191, 106508. <https://doi.org/10.1016/j.envsoft.2025.106508>
- [31] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631. <https://doi.org/10.1145/3292500.3330701>
- [32] Liu, S., Chen, J., Guo, L., Wang, J., Zhou, Z., Luo, J., & Yang, R. (2023). Prediction of soil organic carbon in soil profiles based on visible–near-infrared hyperspectral imaging spectroscopy. *Soil and Tillage Research*, 232, 105736. <https://doi.org/10.1016/j.still.2023.105736>
- [33] Naibo, G., José, De São., B. J. F., Zanotelli, C. C., Pesini, G., Lisboa, B., Kayser, L., Tiecher, T., . . . Tiecher, T. (2025). Near-infrared spectroscopy and machine learning to estimate the physical and chemical properties of soils cultivated with *Ilex paraguariensis*. *Environmental Technology & Innovation*, 40, 104409. <https://doi.org/10.1016/j.eti.2025.104409>
- [34] Khumaidi, A., Purwanto, Y. A., Sukoco, H., & Wijaya, S. H. (2022). Using fuzzy logic to increase accuracy in mango maturity index classification: Approach for developing a portable near-infrared spectroscopy device. *Sensors*, 22(24), 9704. <https://doi.org/10.3390/s22249704>
- [35] Munawar, A. A., Erika, C., & Pawelzik, E. (2024). Fast and simultaneous prediction of inner quality parameters on intact mangos by near infrared spectroscopy: Impact of spectra preprocessing on prediction accuracy. *Future Foods*, 10, 100463. <https://doi.org/10.1016/j.fufo.2024.100463>

How to Cite: Khumaidi, A., & Raafi'udin, R. (2026). An AI-Driven Multi-Objective Framework for Interpretable Near-Infrared Spectral Preprocessing. *Artificial Intelligence and Applications*. <https://doi.org/10.47852/bonviewAIA620210969>