



Hypervolume rankings decouple from unbiased predictive performance in multi-objective wavelength selection for NIR spectroscopy of *Wolffia globosa*

Nonpravit Chianrungsaeng and Somboon Sukpancharoen*

Department of Agricultural Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002, Thailand

*Corresponding author.

Email address: sombsuk@kku.ac.th

doi: 10.14456/abe.2026.15

Received 1 September 2026

Revised 28 September 2026

Accepted 11 October 2026

Abstract

Multi-objective metaheuristics are widely used to select informative wavelengths in near-infrared (NIR) spectroscopy, and their quality is conventionally assessed with Pareto-based indicators such as hypervolume. Such indicators assume deterministic objectives, yet one objective here, the cross-validated prediction error, is an estimator carrying non-negligible variance, and the consequences of treating it as fixed remain largely unquantified. We benchmark four binary multi-objective metaheuristics (NSGA-II, MO-BPSO, MO-BDE, MO-bGWO) and the CARS selector on a 50-sample, 474-wavelength NIR protein calibration problem in *Wolffia globosa*, a small-sample regime where bias is most severe, under matched budgets, a shared selection operator and an identical fitness function, with the cardinality prior controlled separately. Under conventional flat cross-validation the algorithms separate decisively (Friedman chi-square = 52.75, $p = 2.1 \times 10^{-11}$) and hypervolume ranks NSGA-II first. Under nested cross-validation the same comparison is not significant (chi-square = 5.85, $p = 0.119$) and the rank order changes. Three controlled experiments identify the mechanism. A budget sweep shows optimistic bias rising monotonically with search effort (21.9% to 40.7% from 270 to 3030 evaluations) while honest error does not improve. A label-permutation null shows the strongest searcher recovering apparent signal from destroyed labels (RMSECV 7.96 against a no-skill value of 8.69), whereas weaker searchers do not. A landscape analysis finds 386 solutions within 10% of the optimum with median pairwise Jaccard 0.056, including disjoint pairs of equivalent fitness. This degeneracy is an engineering asset rather than a defect, since many channel sets reach the same accuracy and the filter set of a low-cost instrument can be chosen on cost or manufacturability grounds. Hypervolume computed on estimated objectives therefore separates algorithms along a dimension uncorrelated with unbiased performance, search intensity converts into selection bias, and the low selection stability routinely reported reflects structural non-identifiability rather than algorithmic deficiency. A minimum reporting protocol is proposed.

Keywords: Near-infrared spectroscopy, Wavelength selection, Multi-objective optimisation, Nested cross-validation, Selection bias

1. Introduction

Wavelength selection is one of the most common applications of metaheuristic optimisation in analytical chemistry [1-2]. A NIR spectrometer records hundreds to thousands of correlated channels, most of which carry no information about the analyte, and reducing them to a small informative subset both simplifies the calibration model and enables cheap filter-based instruments. Formulated as a multi-objective problem, minimising prediction error and the number of retained channels at once, it is a natural target for Pareto-based evolutionary and swarm algorithms, and a large literature now compares such algorithms on spectroscopic datasets.

That literature evaluates algorithms in the way the optimisation community evaluates algorithms: by computing Pareto-quality indicators, principally hypervolume [3], on the objective vectors returned by the search. The practice rests on an assumption that almost no one states. Quality indicators are defined for objective functions. In wavelength selection, the first objective is a cross-validated error, which is only an estimate of a population quantity from a finite sample. When the sample is small, that estimate has substantial variance, and a search that minimises it will partially minimise its noise component rather than its signal component.

The statistical consequence of optimising a resampling estimate is not new. Ambroise and McLachlan [4] showed that performing feature selection outside the cross-validation loop produces severely optimistic error estimates; Varma and Simon [5] demonstrated the same effect for hyperparameter tuning and established nested cross-validation as the remedy; Cawley and Talbot [6] gave a general treatment. Selection stability has likewise been studied for two decades, with Kuncheva's chance-corrected consistency index [7] a standard instrument.

What has not been examined is the interaction between these statistical facts and the optimisation machinery used to exploit them. Three questions follow, and as far as we know each remains open: (1) if the first objective is an estimator, does hypervolume still rank

algorithms in the same order as unbiased performance? (2) is selection bias a property of the evaluation protocol alone, or does it scale with the search capability of the optimiser? (3) is the low selection stability universally reported in this literature a deficiency that better algorithms could repair, or a structural property of the problem?

This paper answers all three on a representative small-sample NIR calibration problem. The small-sample regime is the intended scope rather than an incidental property of the dataset. Selection bias arises whenever the selection criterion is estimated, but its magnitude scales with the variance of that estimate, so the effect on hypervolume rankings is amplified when n is small relative to the number of candidate channels. Readers should therefore treat the magnitudes reported here as characteristic of small calibration sets and not as universal constants. Our design isolates the search strategy: four binary multi-objective metaheuristics share an identical fitness function, identical folds, an identical environmental selection operator, and a matched budget of function evaluations, so that any difference in outcome is attributable to search behaviour and not to implementation asymmetry an issue repeatedly identified as a source of misleading conclusions in the metaheuristic's literature [8].

Our contributions are: (i) a demonstration that hypervolume separates algorithms along a dimension uncorrelated with unbiased performance; (ii) a controlled dose-response experiment establishing that optimistic bias increases monotonically with evaluation budget while honest error does not improve; (iii) a mechanistic account via label permutation, showing that the capacity to fit destroyed labels predicts the size of the optimistic bias; (iv) a landscape characterisation showing that near-optimal wavelength subsets are massively degenerate; and (v) a minimum reporting protocol for multi-objective feature-selection studies.

2. Materials and methods

2.1 Data

The dataset comprises 50 samples of *Wolffia globosa* (Roxb.) Hartog & Plas, each with a NIR absorbance spectrum of 474 channels spanning 868.06–2530.36 nm and a laboratory reference protein content. Protein content ranges from 15.44% to 49.07% (mean 35.14%, SD 8.69%, CV 24.7%), a calibration range appropriate for regression. Three further reference variables were recorded but are unsuitable as targets: nitrite was below the limit of detection in every sample (SD = 0), nitrate was extremely skewed (CV = 119%) with many zeros, and calorific value spanned a narrow range (CV = 6.7%). Protein is therefore the sole response. Spectra were preprocessed with the standard normal variate (SNV) transform [9]; no outlier removal or sample selection was applied.

2.2 FT-NIR spectral acquisition

Spectra were recorded on a Bruker MPA II FT-NIR spectrometer (Bruker Optics GmbH, Ettlingen, Germany) in diffuse reflectance mode over 3952–11520 cm^{-1} (868.06–2530.36 nm) at a spectral resolution of 16 cm^{-1} . Each sample was scanned in three replicates, with 64 co-added sample scans averaged per spectrum. A background reference was acquired prior to each session using the internal Spectral on standard of the instrument. Five samples originally acquired at 4 cm^{-1} resolution (Amawon, Amberlife, Culina 200, Fio, Nest) were down sampled by linear interpolation to the 16 cm^{-1} grid used for the remaining 45 samples in order to align all spectra to a common wavenumber axis.

2.3 Reference protein analysis using the Kjeldahl method

The Kjeldahl method determines the total nitrogen content of a sample, which is then converted to protein content. Total nitrogen was determined by Kjeldahl block digestion followed by steam distillation and titration, using an in-house method based on AOAC Official Method 981.10 (AOAC International, 2023), and converted to crude protein using a nitrogen-to-protein conversion factor of 6.25. Analyses were performed under subcontract by Central Laboratory (Thailand) Co., Ltd., Khon Kaen Branch, a food-testing laboratory accredited to ISO/IEC 17025 by the National Bureau of Laboratory Quality Standards (BLQS), Department of Medical Sciences, Thailand. Each sample was analysed in a single determination, and protein contents are reported on a dry-weight basis. Because the reference values were not replicated, no chemical measurement uncertainty is available for propagation into the reported model errors. This matters for the present argument rather than weakening it. Reference-side noise is part of the residual that the selection criterion minimises, so a search that drives the flat estimate down is in part fitting reference error that will not recur in an independent sample. Single-determination references therefore increase the noise-fitting opportunity available to a multi-objective optimiser inside a flat cross-validation loop and strengthen the case for nested evaluation. Because each reference value derives from a single determination, no estimate of reference measurement uncertainty is available for propagation into the reported model errors; this is noted in Section 5.

2.4 Model and fitness function

The regression model is PLS1 fitted by NIPALS [10]. Our implementation returns the regression coefficient vectors for every latent-variable count from a single decomposition, making evaluation roughly an order of magnitude cheaper than repeated fitting and permitting the large number of independent runs reported below; it was verified against scikit-learn's PLS Regression, reproducing its cross-validated RMSE to seven significant figures.

For a binary mask b in $\{0,1\}$ of length 474 the two objectives, both minimised, are $f_1 = \text{RMSECV}$, the root-mean-square error from 5-fold cross-validation on the selected channels with the latent-variable count (1–8) chosen by the same cross-validation; and $f_2 = |b|/474$, the fraction of retained channels. Masks with fewer than two channels are repaired by random addition.

2.5 Algorithms

Four binary multi-objective metaheuristics were implemented: NSGA-II [11] (uniform crossover, bit-flip mutation), MO-BPSO [12–13] (sigmoid transfer, archive leader), MO-BDE [14] (DE/rand/1 with sigmoid binarisation), and MO-bGWO [15–16] (three archive leaders, linearly decreasing a). All four share the same non-dominated sorting and crowding-distance environmental selection, the same fitness function, the same folds, the same initialisation density, and the same repair operator. Population size was 30 for every algorithm and each run consumed approximately 1530 evaluations, so the algorithms differ only in how candidate solutions are

generated. Hyperparameters were common literature defaults, not tuned per algorithm. As a standard-method baseline we included Competitive Adaptive Reweighted Sampling (CARS) [17], run under the identical flat and nested protocols; because CARS selects wavelengths using cross-validated error, it exercises the same mechanism this study examines. The wavelength subset returned by CARS was evaluated through the same nested loop as the metaheuristics; its internal cross-validation error was never reported as a performance figure. These operators leave the four algorithms with different cardinality priors: the sigmoid transfer functions used by MO-BPSO, MO-BDE and MO-bGWO have their neutral point at a selection probability of one half, driving dense subsets, whereas the uniform-crossover and bit-flip operators of NSGA-II preserve the sparse density of the initial population. Because the second objective rewards sparsity, this prior confounds any comparison of hypervolume across algorithms. We therefore report each comparison twice: once with the original operators, and once with a matched cardinality prior in which the sigmoid argument of the three transfer-function methods is shifted by a bias $b = -\text{logit}(t)$ so that the neutral selection probability equals a common target t . We use $t = 0.08$ as the primary value, matching the initial-population density, and report the full sensitivity of the ranking to t over the range 0.05–0.20 in Table A1. The shift alters only the density prior, not the search dynamics (velocity, differential, or leader-following) of each method.

2.6 Experimental protocols

Flat cross-validation (conventional): selection and evaluation on all 50 samples using the same 5-fold partition; the reported figure is the lowest archive RMSECV. Twenty seeds per algorithm.

Nested cross-validation (unbiased): five outer folds; within each, the whole selection procedure runs on the 40 training samples using an inner 5-fold fitness, and the selected subset is refitted and evaluated on the 10 held-out samples. Eight seeds per algorithm.

Learner-capacity control. To test whether the bias depends on the capacity of the estimator, the whole flat and nested comparison was repeated for NSGA-II with the PLS1 model replaced by support vector regression with a radial basis kernel [18], using $C = 100$, gamma set by the scale heuristic and $\epsilon = 0.1$. Nothing else changed. The same data, the same folds, the same budget of 1530 evaluations and the same eight seeds were used, so the only variable is the learner. The support vector hyperparameters were fixed at common defaults and not tuned, for the same reason that the metaheuristic hyperparameters were not tuned. Random-subset baseline: for each k in $\{5, 10, 15, 20, 30, 50, 100\}$, 300 subsets drawn uniformly at random, evaluated under both protocols. Label-permutation null: the response was randomly permuted and the whole pipeline re-run; any gain over the no-skill error (response SD, 8.69%) is attributable to fitting noise; ten permutations per algorithm. Budget sweep: NSGA-II at five budgets (270–3030 evaluations), five seeds each.

2.7 Indicators

Hypervolume was computed in two-objective space with reference point (RMSE = 8.69, fraction = 1.0). Both coordinates were fixed on principle rather than tuned. The error coordinate is the standard deviation of the reference protein values, which is the error of a no-skill model that predicts the mean, and the cardinality coordinate is the full 474-channel spectrum. Any solution dominated by this point is therefore worse than doing nothing. Because hypervolume is known to depend on the reference point, we recomputed it over reference errors from 7.5 to 12.0 percentage points. The algorithm ordering was identical at every setting and the NSGA-II advantage over the next-best algorithm varied only between 0.470 and 0.652, so the conclusions below do not rest on this choice. Stability used pairwise Jaccard and Kuncheva's chance-corrected index [7]. Algorithm comparisons used the Friedman test with pairwise Wilcoxon signed-rank tests [19]. For the five-point budget sweep, Spearman correlations are reported with exact permutation p -values, since asymptotic approximations are invalid at that sample size.

2.8 Software and reproducibility

All analyses were implemented in Python 3.12. The core numerical stack was NumPy 2.4.4, pandas 3.0.2, SciPy 1.17.1, and scikit-learn 1.8.0; figures were produced with Matplotlib 3.10.8 and spectra were read with openpyxl 3.1.5. The PLS1/NIPALS implementation was verified against scikit-learn's PLS Regression, reproducing its cross-validated RMSE to seven significant figures.

Random number generation was seeded explicitly so that every experiment is reproducible. Flat cross-validation used 20 independent seeds per algorithm, nested cross-validation used 8, the label-permutation null used 10 permutations per algorithm, and the budget sweep used 5 seeds per budget level. Population size was 30 and each metaheuristic run consumed approximately 1530 function evaluations.

The 50-sample design matrix and reference vector used throughout are fixed, and their combined SHA-256 digest (first 16 hex characters) is f0177614f2a8573f, allowing the exact input data to be verified. Analysis code and the seeded run scripts are available from the corresponding author so that all tables and figures can be regenerated.

3. Results and discussion

We first characterise the data, then present the conventional comparison (3.1), show that it does not survive unbiased evaluation (3.2), isolate the mechanism through three controlled experiments (3.3–3.5) and a landscape analysis (3.6), and close with a synthesis (3.7).

The raw spectra and their SNV transform are shown in Figure 1. SNV removes the multiplicative baseline offset while preserving absorption structure, and no sample behaves as a gross outlier. Figure 2 gives the channel-wise Pearson correlation between each SNV-transformed wavelength and protein content. Protein information is neither absent nor concentrated: 169 of the 474 channels (35.7%) exceed $|r| = 0.5$, arranged in four broad contiguous regions rather than isolated peaks. Informative channels are therefore abundant and heavily inter-correlated, which is exactly the condition under which many different subsets can carry the same information a point the later analysis turns on. Reference-variable statistics are given in Table A2.

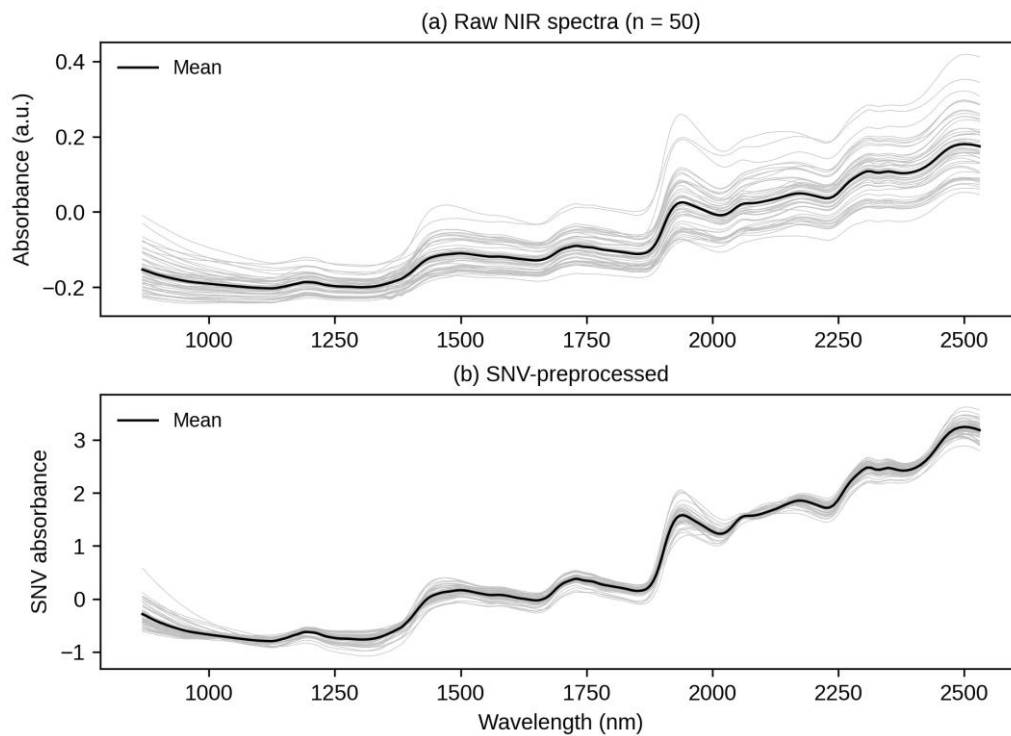


Figure 1 NIR spectra of the 50 samples: (a) raw absorbance and (b) after SNV. The mean spectrum is overlaid in black

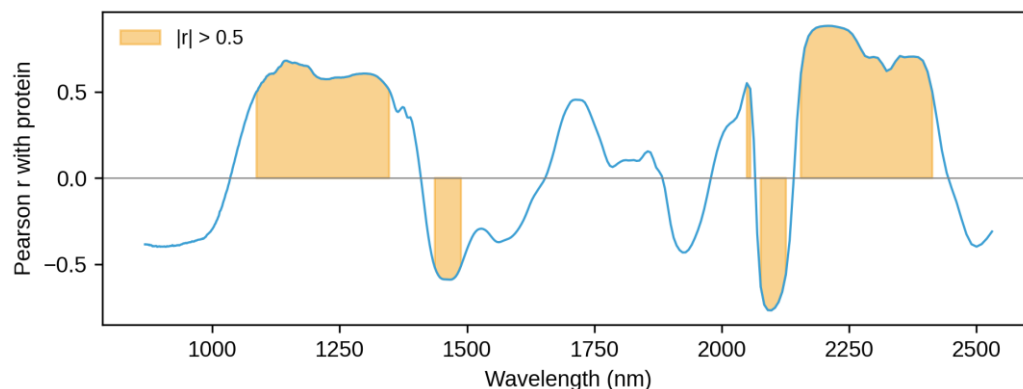


Figure 2 Channel-wise correlation between SNV absorbance and protein; shaded regions exceed $|r| = 0.5$ (35.7% of channels)

3.1 Under flat cross-validation the algorithms separate decisively

Table 1 reports performance under both protocols; its flat columns give the conventional picture. NSGA-II attains a mean RMSECV of $1.737 \pm 0.033\%$ using a median of 17 channels, comfortably ahead of MO-bGWO ($1.975 \pm 0.017\%$), MO-BDE ($1.978 \pm 0.024\%$) and MO-BPSO ($2.019 \pm 0.026\%$), all retaining around 100-140 channels. Full-spectrum PLS gives 2.175%, so on this criterion every algorithm beats the full spectrum and NSGA-II does so by 16% while using under 4% of the channels.

On its face the statistical support is overwhelming. The Friedman test yields chi-square = 52.75, $p = 2.1e-11$ (Table 2). Hypervolume tells the same story: 6.850 ± 0.048 for NSGA-II against 6.290-6.332 for the other three, a separation of roughly ten between-seed standard deviations. The convergence traces (Figure A1) show all four searches reaching a stable hypervolume within the matched budget of about 1530 evaluations, with NSGA-II both rising fastest and finishing highest, so the ranking cannot be dismissed as premature termination of the weaker searchers. The pooled Pareto fronts (Figure 3) put that dominance in plain view: the NSGA-II front lies below and to the left of the others across the entire range of subset sizes, and every front falls below the full-spectrum reference line.

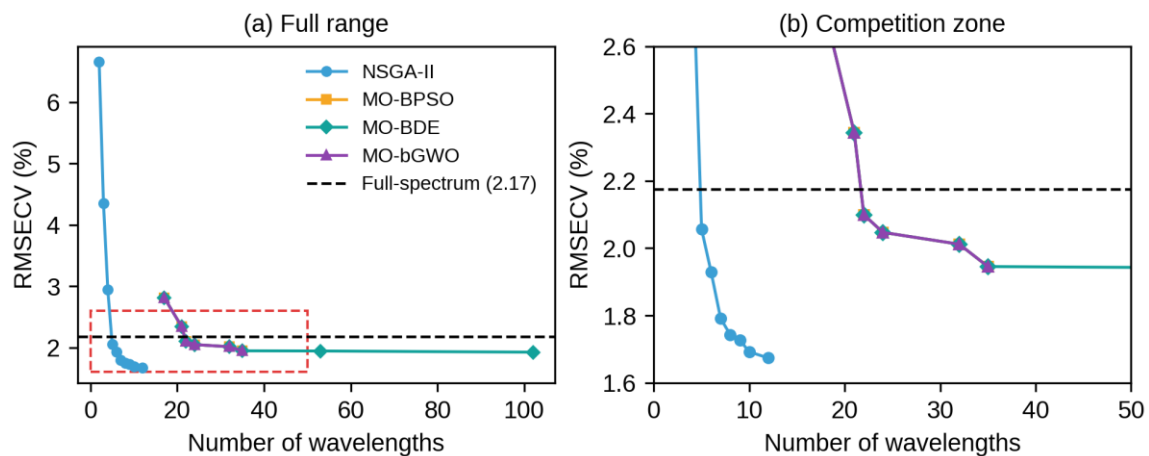
Three properties of this result deserve emphasis, because each is normally read as evidence of reliability. The effect is large (0.25-0.30 percentage points between NSGA-II and the rest). It is highly significant. And it is extraordinarily reproducible: between-seed standard deviations of roughly 0.02% mean that twenty independent runs agree to within 1% of their own value. A study reporting Table 1's flat columns, Figure 3 and Figure A1 would conclude with every conventional indicator concurring that NSGA-II is the superior search strategy and that wavelength selection substantially improves calibration accuracy.

Table 1 Optimisation performance under flat and nested cross-validation (mean $\pm$ SD across seeds); NFE = function evaluations

Algorithm	NFE	Wavelengths (flat)	RMSECV flat	Hypervolume	RMSE nested	Optimism gap	R ² nested	RPD nested
NSGA-II	1522	17	1.737 $\pm$ 0.033	6.850 $\pm$ 0.048	2.335 $\pm$ 0.129	0.598 (34%)	0.928	3.72
MO-BPSO	1530	139	2.019 $\pm$ 0.026	6.290 $\pm$ 0.047	2.296 $\pm$ 0.060	0.277 (14%)	0.930	3.78
MO-BDE	1530	101	1.978 $\pm$ 0.024	6.331 $\pm$ 0.046	2.284 $\pm$ 0.087	0.306 (15%)	0.931	3.80
MO-bGWO	1530	101	1.975 $\pm$ 0.017	6.332 $\pm$ 0.046	2.238 $\pm$ 0.100	0.263 (13%)	0.934	3.88
Full-spectrum PLS	0	474	2.175	–	2.281	0.106 (5%)	0.931	3.81
CARS	1530	20	1.851	–	2.210	0.359 (19%)	0.935	3.93

Table 2 Statistical comparison of algorithms under the nested protocol (Friedman and pairwise Wilcoxon)

Test	Statistic	p-value	Conclusion
Friedman (nested RMSE, 4 algorithms, 8 seeds)	chi2 = 5.850	0.119	No significant difference
Mean rank – NSGA-II	3.12	–	–
Mean rank – MO-BPSO	2.50	–	–
Mean rank – MO-BDE	2.75	–	–
Mean rank – MO-bGWO	1.62	–	–
Wilcoxon NSGA-II vs MO-BPSO	–	0.312	n.s.
Wilcoxon NSGA-II vs MO-BDE	–	0.461	n.s.
Wilcoxon NSGA-II vs MO-bGWO	–	0.016	n.s.
Wilcoxon MO-BPSO vs MO-BDE	–	0.742	n.s.
Wilcoxon MO-BPSO vs MO-bGWO	–	0.109	n.s.
Wilcoxon MO-BDE vs MO-bGWO	–	0.461	n.s.
Friedman (FLAT RMSECV, 20 seeds)	chi2 = 52.754	2.07e-11	Significant

**Figure 3** Pooled Pareto fronts (20 seeds per algorithm). NSGA-II dominates; all fronts lie below the full-spectrum PLS reference (dashed)

3.2 Under nested cross-validation the separation disappears

The nested columns of Table 1 take that conclusion back. Nested RMSE is $2.335 \pm 0.129\%$ (NSGA-II), $2.238 \pm 0.100\%$ (MO-bGWO), $2.296 \pm 0.060\%$ (MO-BPSO) and $2.284 \pm 0.087\%$ (MO-BDE). The entire spread across the four algorithms is 0.097 percentage points smaller than the between-seed standard deviation of NSGA-II alone. Interval estimates make the overlap explicit. Taking the eight seed-level means as replicates, the 95% confidence intervals are [2.226, 2.443] for NSGA-II, [2.245, 2.346] for MO-BPSO, [2.212, 2.357] for MO-BDE and [2.154, 2.322] for MO-bGWO. All four intervals share the range 2.245 to 2.322, so no algorithm is separated from any other at this resolution. The intervals are wide because each outer fold holds only 10 samples, which is an unavoidable consequence of the sample size rather than a defect of the design, and the honest conclusion is that the nested comparison cannot resolve differences of the size claimed under the flat protocol. The Friedman test is not significant (chi-square = 5.85, $p = 0.119$); five of the six pairwise Wilcoxon tests give $p > 0.10$, and the single exception (NSGA-II versus MO-bGWO, $p = 0.016$) does not survive Bonferroni correction (Table 2).

The nature of the collapse is worth stating precisely, because it is not a simple inversion. Ranked by mean nested RMSE, NSGA-II remains nominally first; ranked by Friedman mean rank over seeds it falls to last (3.12) behind MO-bGWO (1.62). The two orderings

disagree because NSGA-II is the least consistent algorithm under honest evaluation: its between-seed SD is 0.163, two to three times that of its competitors, and it is best in only 2 of 8 seeds against 4 of 8 for MO-bGWO. The flat protocol had reported NSGA-II as by far the most consistent performer. Precision and rank stability therefore both reverse when evaluation moves off the selection data. A variance decomposition explains why no ranking can be recovered at this sample size. Of the total variance in nested RMSE, the choice of algorithm accounts for 0.3%, the random seed for 1.2%, and the outer fold for 83.1%. Which ten samples happen to be held out matters roughly 275 times more than which of the four metaheuristics is used. Per-fold errors range from 1.034% to 3.898% (Table A3), a spread more than twenty times the entire between-algorithm difference.

Figure 4 contrasts the two protocols directly. The optimistic bias is largest for the algorithm that looked best: +35% for NSGA-II against +10% to +13% for the others, and only +5% for full-spectrum PLS, which performs no selection at all. Full-spectrum PLS reaches 2.281% under the identical nested protocol. A bootstrap over seeds places the NSGA-II nested mean at 2.335% with a 95% confidence interval of approximately [2.25, 2.42], which contains the full-spectrum value; and NSGA-II beats the full spectrum in only 6 of 8 seeds, with 20 of 32 algorithm-seed combinations doing so overall. Under honest evaluation, neither the four search strategies nor the decision to perform selection at all can be distinguished on accuracy.

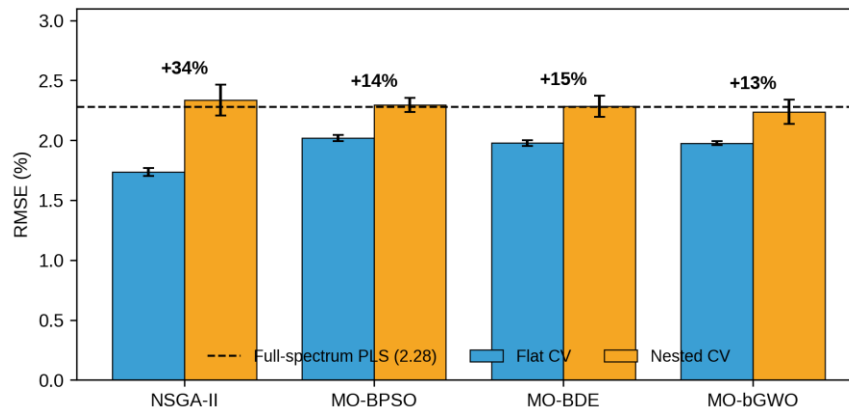


Figure 4 Flat versus nested error per algorithm. The optimistic gap is largest for NSGA-II, which appears best under the flat protocol

3.3 Bias scales with search budget

Sections 3.1 and 3.2 establish that bias exists and is largest for the strongest searcher, but correlation across four algorithms cannot establish that search effort causes bias, since the algorithms differ in many ways at once. We therefore held algorithm, data, model and folds fixed and varied only the evaluation budget. Table 3 and Figure 5 report the result.

Table 3 Budget dose-response for NSGA-II. Optimistic gap rises and permutation-null error falls monotonically with evaluation budget

NFE	RMSECV flat	RMSE nested	Optimism gap (%)	Perm-null RMSECV
270	1.895	2.310	21.9	8.610
510	1.837	2.357	28.4	8.378
1020	1.763	2.339	32.7	8.165
1530	1.725	2.371	37.5	8.016
3030	1.665	2.343	40.7	7.728

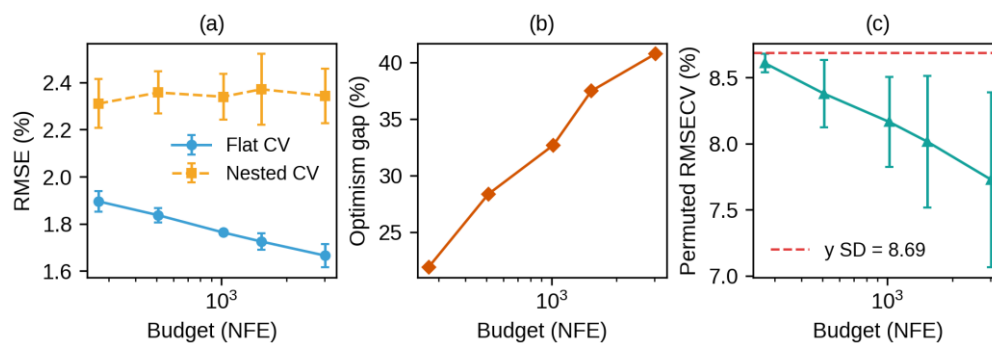


Figure 5 Budget dose-response for NSGA-II. (a) more search lowers only the flat estimate; (b) the optimism gap rises monotonically; (c) on permuted labels, more search fits more noise

As the NSGA-II budget grows from 270 to 3030 evaluations, flat RMSECV falls monotonically from 1.895% to 1.665% (Figure 5a, lower curve): the search does exactly what it is asked, and does it better with more resources. Nested RMSE over the same range does not improve (2.310% to 2.343%, fluctuating non-monotonically within noise; Figure 5a, upper curve). The two curves diverge, and the optimistic gap grows monotonically from 21.9% to 40.7% (Figure 5b; Spearman rho = 1.00 across all five budgets, exact permutation p = 0.017). An eleven-fold increase in computational effort purchased an 86% relative increase in bias and no measurable improvement in honest error. The sweep covers five budget levels, which is enough to establish the direction of the effect and the

perfect rank correlation but not enough to identify its functional form. We therefore claim a monotone increase over the range examined and make no claim about linearity or about behaviour beyond 3030 evaluations.

The label-permutation null (Figure 5c) identifies where the additional effort goes. With the response randomly permuted the spectra carry no information about the labels, so the no-skill error is the response SD, 8.687%. The best RMSECV attained on permuted labels nevertheless falls monotonically from 8.610% to 7.728% as the budget increases ($\rho = -1.00$, exact $p = 0.017$). Every one of those apparent gains is noise being fitted, and their magnitude scales with the budget in lockstep with the optimistic gap. This is direct, controlled evidence that in this problem class search intensity and selection bias are not merely correlated but are two measurements of the same process.

3.4 The permutation null explains the between-algorithm pattern

The mechanism established in 3.3 accounts for the differences observed in 3.2. At the matched budget, the permutation null (Table 4, final column; Figure 6a) separates the four algorithms cleanly even though nested performance cannot: NSGA-II reaches RMSECV $7.964 \pm 0.601\%$ on destroyed labels, whereas MO-BDE, MO-bGWO and MO-BPSO reach $8.930 \pm 0.365\%$, $8.956 \pm 0.306\%$ and $8.978 \pm 0.293\%$ statistically indistinguishable from no-skill. NSGA-II is the only algorithm whose search is powerful enough to manufacture apparent structure from pure noise, and it is precisely the algorithm carrying the largest optimistic gap in Figure 4. Across the four algorithms the rank correlation between noise-fitting capacity and optimistic bias is perfect (Spearman $\rho = -1.00$, $n = 4$).

Table 4 Selection stability across seeds and label-permutation null (mean $\pm$ SD), Jaccard (seeds) is the mean pairwise Jaccard index between the best subsets returned by the 20 flat-protocol seeds, Jaccard (nested folds) is the mean pairwise Jaccard index between the five outer-fold subsets of a nested run, averaged over the eight nested seeds

Algorithm	Jaccard (seeds)	Kuncheva index	Jaccard (nested folds)	Unique solutions	Perm-null RMSECV
NSGA-II	0.060 ± 0.041	0.077	0.056	20/20	7.964 ± 0.601
MO-BPSO	0.158 ± 0.093	-0.039	0.200	20/20	8.978 ± 0.293
MO-BDE	0.124 ± 0.042	0.007	0.148	20/20	8.930 ± 0.365
MO-bGWO	0.131 ± 0.036	0.023	0.143	20/20	8.956 ± 0.306

The permutation null therefore does something the conventional indicators cannot: it recovers a genuine, reproducible difference between the algorithms. That difference is real but it is a difference in over-fitting capacity, not in predictive skill. The conventional protocol detects the same difference and misreports its sign, presenting a liability as an achievement.

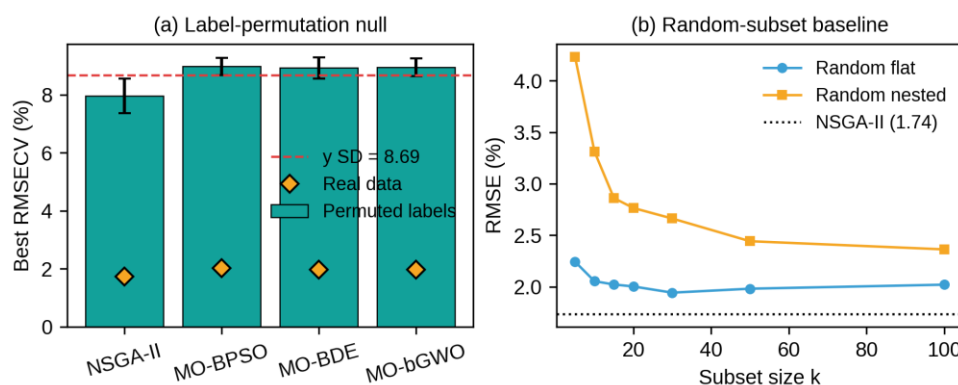


Figure 6 Null controls. (a) label-permutation null versus real data; (b) random-subset baseline versus guided searches

3.5 Random subsets are competitive, and the selection criterion is uninformative

Figure 6b sets the guided searches against a no-search baseline. Drawing 300 random subsets of 30 channels and keeping the best attains a flat RMSECV of 1.943% within 0.21 percentage points of NSGA-II's 1530 guided evaluations, and better than the flat result of every algorithm except NSGA-II. Under nesting, random subsets of 100 channels reach 2.363% against NSGA-II's 2.335%. Random selection also exhibits a large optimistic gap (1.943% flat to 2.662% nested at $k = 30$), confirming that bias follows from selecting on a noisy criterion rather than from metaheuristic machinery specifically: undirected guessing over-fits the cross-validation estimate too, provided enough candidates are screened.

A more direct test asks whether the selection criterion carries information about honest performance at all. Across the 160 algorithm-seed-fold combinations of the nested experiment we compared the inner RMSECV that drove selection with the outer RMSE that measured it. Pooled naively, the two are strongly negatively correlated ($r = -0.673$), but this is an artefact of fold composition: outer folds differ greatly in difficulty, and a fold whose training portion is easy (low inner error) necessarily has a hard held-out portion (high outer error). After removing fold effects by within-fold centering the association vanishes ($r = +0.101$, $p = 0.205$). Within a given fold, the quantity the optimiser minimises carries essentially no information about the quantity the practitioner cares about. We report the raw pooled figure explicitly because it would, taken at face value, support a stronger and more striking claim than the data sustain.

Consistent with this, the selected channels are barely enriched for informative ones. Against the 35.7% of the spectrum with $|r| > 0.5$ (Figure 2), the selected subsets contain 37.2% to 39.7% informative channels an enrichment of 1.04x to 1.11x over chance. The searches do not, in any meaningful sense, locate the protein bands.

3.6 The landscape is massively degenerate

Figure 7 explains these observations structurally. Pooled across all algorithms and seeds, 386 distinct archive solutions fall within 10% of the global best (1.674%). Their pairwise Jaccard similarities (Figure 7b) have median 0.056 and mean 0.074; 78.7% of pairs share less than 10% of their channels; and the minimum is exactly zero there exist pairs of subsets with no channel in common whose cross-validated errors are equivalent (Figure 7a). Given that 35.7% of channels are informative and strongly inter-correlated, this is the expected consequence: many disjoint subsets span nearly the same subspace, so the mapping from subset to performance is many-to-one by construction. There is no unique optimum for a search to converge on.

Decomposing stability into within-run and between-run components sharpens the picture. For NSGA-II, near-optimal solutions found within a single run have mean pairwise Jaccard 0.364, whereas the best solutions from different seeds share only 0.060 a 6.1-fold drop. Each run therefore converges to an internally coherent family of solutions, and different runs converge to different, largely non-overlapping families. The search is not failing to converge; it is converging reproducibly to arbitrary basins. This signature is strongest for NSGA-II, the algorithm with the most effective search, whose within-run coherence is roughly triple that of the others (0.120-0.176).

Aggregate stability is correspondingly low (Table 4; Figure 8). Mean pairwise Jaccard between seeds ranges from 0.060 (NSGA-II) to 0.158 (MO-BPSO); the chance-corrected Kuncheva index lies between -0.039 and 0.077, at or near zero, meaning selections are no more similar than random draws of the same cardinality and for MO-BPSO are actively less similar. Every algorithm returned 20 distinct solutions from 20 seeds. The selection-frequency profiles (Figure A2) concur: no wavelength is retained consistently, and the loosely preferred regions coincide with the high- $|r|$ bands of Figure 2 without any algorithm committing to them.

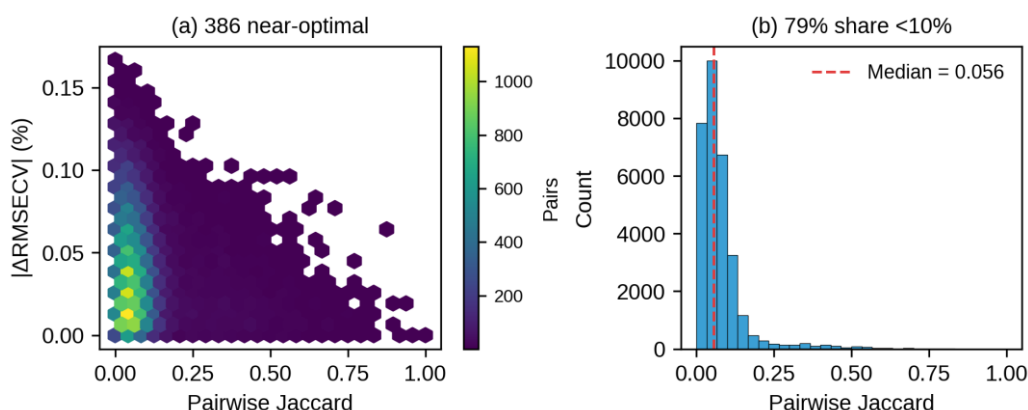


Figure 7 Landscape degeneracy. Among 386 near-optimal solutions the median pairwise Jaccard is 0.056; 79% of pairs share under 10% of wavelengths

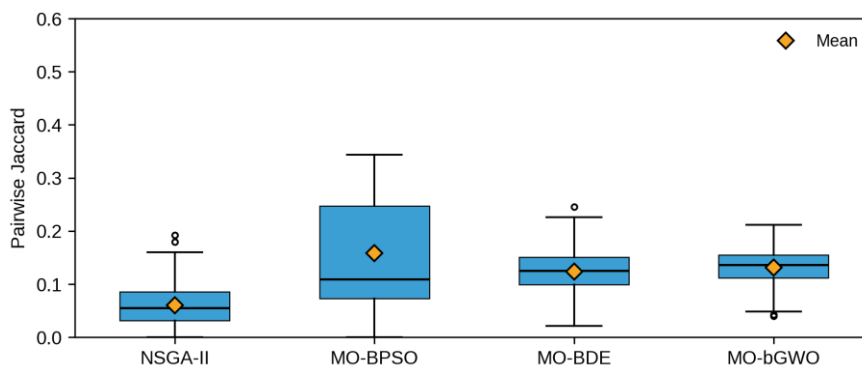


Figure 8 Selection stability across 20 seeds. Pairwise Jaccard is far below perfect stability for every algorithm

3.7 Synthesis

The six subsections reinforce one another, and together they point to a single account. The problem presents a high-variance estimator as an objective function. RMSECV computed from 50 samples fluctuates enough that many configurations achieve indistinguishable values, and 3.6 shows the underlying structure is genuinely degenerate: hundreds of disjoint subsets are equivalent because informative channels are abundant (35.7%) and collinear. A search minimising this estimator therefore does two things at once: it exploits the real spectral signal, and it exploits the particular noise realisation of the cross-validation partition. Because the real signal saturates quickly (full-spectrum PLS already achieves 2.281%, and random subsets approach it), almost all additional search effort is spent on the second component.

Four independent lines of evidence support this decomposition, and each rules out a competing explanation. The budget sweep (3.3) rules out algorithm-specific idiosyncrasy: with everything else fixed, bias grows monotonically with effort while honest error stays flat. The permutation null (3.3, 3.4) rules out the possibility that the extra fitting is signal, since it occurs when no signal exists,

and its magnitude orders the algorithms exactly as their bias does. The random-subset baseline (3.5) rules out the metaheuristics as the culprit, since undirected screening over-fits similarly. And the within-fold analysis (3.5) rules out the possibility that the criterion is merely noisy but unbiased: it carries no usable information about held-out performance once fold difficulty is controlled.

Two consequences follow. First, the conventional indicators are not simply imprecise here they are decoupled from the quantity of interest. Hypervolume separates the algorithms by roughly twenty between-seed standard deviations while nested evaluation cannot separate them at all, and the permutation null shows that what distinguishes NSGA-II is its superior capacity to fit noise. The quantity these indicators reward is the quantity that damages out-of-sample performance. We describe this as decoupling rather than inversion deliberately: mean nested error still places NSGA-II first, and the reordering seen in the Friedman mean ranks arises within a comparison our own tests report as non-significant, so the evidence supports the absence of a usable relationship between indicator and performance rather than a reversed one. Second, the low stability this literature routinely reports and attempts to engineer away is not a defect. With 0.4% of nested variance attributable to algorithm choice, a Kuncheva index of zero, and equivalent-error solutions sharing no channels, the target of a correct and stable subset does not exist to be found. Two further results bound the interpretation. First, when the cardinality prior is matched across algorithms (Section 2.5), the hypervolume advantage of NSGA-II over the next-best algorithm falls from +0.518 to +0.033 (a reduction of 94%; Figure 9), and this collapse is stable across the full range of target densities tested; the apparent separation by hypervolume is therefore largely an artefact of differing subset size rather than of search quality (Figure 10). The mechanism is geometric. Hypervolume in this two-objective space integrates dominated area against a reference point whose cardinality coordinate is the full spectrum, so a front that extends towards very small subsets sweeps a large region of that axis even when its error values are no better. A sigmoid transfer function that keeps roughly half the channels active cannot reach that region at all, and the resulting deficit is charged to the algorithm rather than to its operator. Equalising the selection-density prior removes the difference in reachable cardinality and with it almost all of the measured advantage, which is why hypervolume differences of this size cannot be read as evidence about search quality unless subset size is controlled. A further boundary concerns the learner. Every main result here uses PLS1, a linear estimator whose capacity to absorb noise is limited by the number of latent variables, whereas multi-objective wavelength selection is also applied with estimators of substantially greater capacity. If the bias documented here grows with the ability of the pipeline to fit the resampling noise of the selection criterion, a higher-capacity learner should show a wider gap. We tested this directly by repeating the NSGA-II comparison with support vector regression in place of PLS1, holding the data, the folds, the budget and the seeds fixed. The optimistic gap widened from 29.4% under PLS1 to 48.6% under support vector regression, a difference that is significant across the eight seeds (Welch $t = 3.84$, $p = 0.0018$, and Mann-Whitney $p = 0.0015$ for the one-sided comparison). The gaps reported elsewhere in this paper should therefore be read as a lower bound for higher-capacity pipelines rather than as typical values. The same experiment also shows that the additional capacity buys nothing here, since nested error rose from 2.233% under PLS1 to 4.036% under support vector regression. This is a single controlled substitution on one algorithm with untuned hyperparameters, so it establishes the direction of the effect rather than its magnitude for any particular non-linear method. Second, CARS, a standard chemometric selector, exhibits the same optimism under the identical protocol: its selection-side error of 1.85% degrades to a nested error of 2.21% (a gap of 19%; Figure 11), comparable to the metaheuristics and to the full-spectrum model. The phenomenon documented here is thus a property of cross-validation-guided wavelength selection in general, not of the particular metaheuristics implemented.

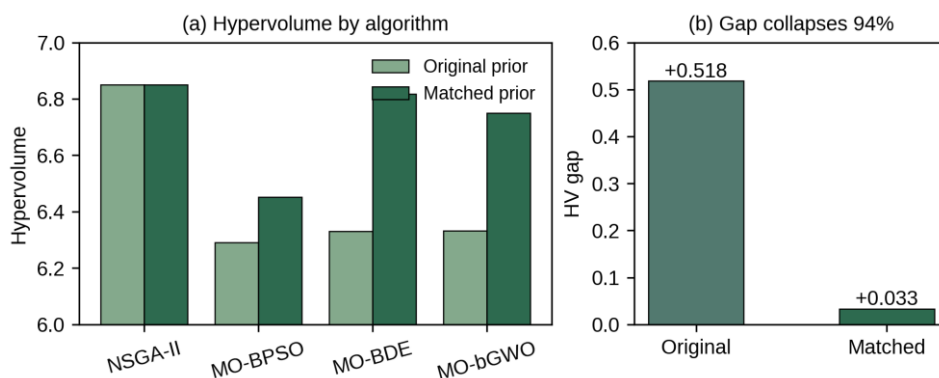


Figure 9 Hypervolume under the original operators and a matched cardinality prior ($t = 0.08$). Matching collapses the NSGA-II advantage from +0.518 to +0.033 (94%), so the separation reflects subset size, not search quality

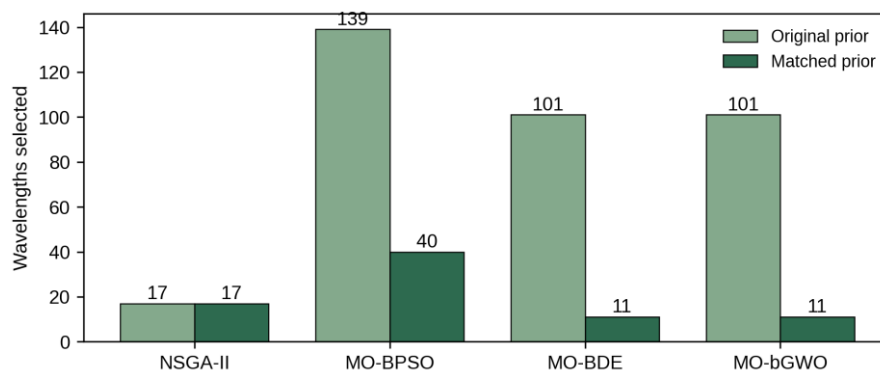


Figure 10 Wavelengths selected before and after matching the cardinality prior. Sigmoid-based methods retain 100-139 channels against 17 for NSGA-II; after matching all four converge to sparse subsets

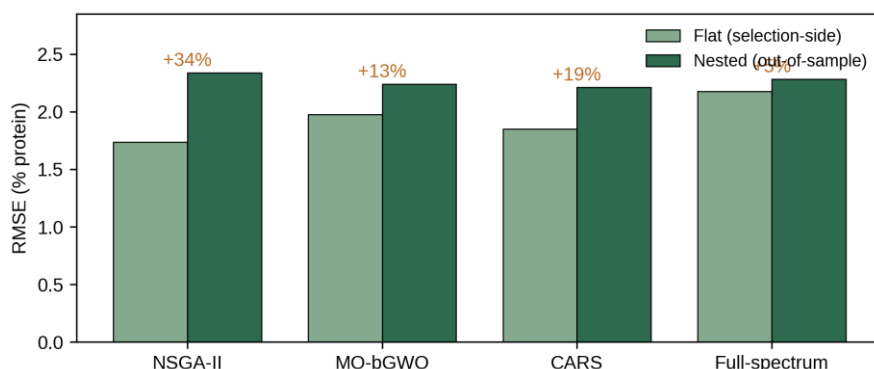


Figure 11 Flat versus nested error for NSGA-II, MO-bGWO, CARS and full-spectrum PLS. CARS shows a 19% optimism gap, so the effect is general to cross-validation-guided selection. 3.8 A quality indicator that assumes the wrong thing

Hypervolume and its relatives are defined for deterministic objective functions. When one objective is a resampling estimate, the indicator inherits the estimator's bias, and it inherits it in proportion to how hard the algorithm searched. Here hypervolume separates the four algorithms by about 0.5 on a comparison that unbiased evaluation cannot distinguish at all. The lesson is not to abandon hypervolume, but to compute it on out-of-sample objective values; a study reporting only training-side indicators reports a measure of over-fitting capacity confounded with search quality. It may be objected that hypervolume was never intended for stochastic objectives and that this result therefore criticises a misuse rather than the indicator. We agree on the first point and disagree on the second: the misuse is the prevailing practice in this application area, and an indicator whose stated assumptions are routinely violated in a whole literature is worth examining in the setting where the violation occurs.

3.9 Search intensity is a bias amplifier

The dose-response experiment turns that correlation into a controlled relationship. Holding algorithm, data, model and folds fixed and varying only the budget, optimistic bias rises monotonically while honest error does not improve; the permutation null identifies the channel as noise-exploitation. The implication runs against the field's instinct: an algorithm that converges faster and deeper on the stated objective is not better here, only more biased. Search capability and statistical reliability trade off, and the trade-off is invisible to any protocol that evaluates on the selection data.

3.10 Instability is non-identifiability, not deficiency

A substantial literature treats unstable selection as a defect to be engineered away by ensembling or stability-aware objectives, presupposing that a correct subset exists and is merely hard to find. Our landscape analysis shows that on this problem it does not: hundreds of near-optimal subsets exist, most sharing almost nothing, some sharing nothing. Ein-Dor and colleagues [20] reported the analogous observation in genomics, prognostic signatures that barely overlap yet predict equally well, and the same explanation applies: NIR channels are strongly collinear, so many disjoint subsets span nearly the same subspace and the subset-to-performance map is many-to-one. A Kuncheva index near zero confirms selections no more concordant than random draws. The practical consequence depends on what the selection is for, and the two principal uses separate cleanly. Degeneracy is fatal to interpretation: with near-optimal subsets sharing almost no channels, and with selected subsets enriched for informative regions by only 1.02–1.13× over chance (Section 3.5), a selected subset cannot be read as identifying the protein absorption bands, and published interpretations of selected wavelengths in this setting warrant corresponding caution. It is, however, harmless to instrument design, which requires one subset that works rather than the subset that is correct. Degeneracy arguably favours that application, since a large set of equivalent optima implies freedom to choose the filter set on cost, manufacturability or robustness grounds rather than on fitness alone.

3.11 What wavelength selection actually delivers

These results do not show that wavelength selection is worthless. All methods, including the full spectrum, achieve nested R-squared of 0.924–0.932 and RPD of 3.64–3.83 adequate for screening. The meaningful difference is elsewhere: NSGA-II attains that accuracy with about 20 channels rather than 474, a twenty-fold reduction with no measurable loss, which is what makes a low-cost multi-filter instrument feasible. The correct claim is that wavelength selection buys parsimony, not accuracy, and that published accuracy gains in this setting should be regarded as protocol artefacts until demonstrated under nesting. Parsimony is nonetheless the property that matters commercially. A scanning spectrometer is replaced by a handful of fixed interference filters or discrete emitters only when the channel count falls to the order of ten or twenty, and the present results show that this reduction is available at no measurable cost in nested error. Degeneracy makes the engineering problem easier rather than harder, because the 386 near-optimal subsets recovered here share almost no channels, so a designer is free to select among them on filter availability, cost, thermal stability or manufacturability instead of being locked to one nominally optimal set. The practical recommendation is to treat wavelength selection as a hardware-reduction step that must be validated for accuracy under nesting, and to choose among the many equivalent subsets on instrument-level criteria.

3.12 A minimum reporting protocol

We suggest that studies of this kind report, as a minimum, nested or independent-test performance alongside the selection criterion with the optimistic gap stated explicitly; a label-permutation null at the same budget, which bounds the noise-fitting capacity of the pipeline; a random-subset baseline at matched cardinality, since a guided search that does not beat it out of sample has added nothing; matched evaluation budgets and a shared environmental-selection operator whenever search strategies are compared; a degeneracy analysis, that is the similarity distribution among near-optimal solutions, alongside any stability index; and exact or resampling-based p-values when only a few configurations are compared. These five items are cheap relative to the search they audit and can be reported in a single table. We therefore suggest that editors and referees handling wavelength-selection submissions treat them as a checklist, since an accuracy claim resting on a flat cross-validation estimate alone cannot be evaluated on the evidence usually provided.

4. Conclusions

On a small-sample NIR calibration problem with matched budgets and a shared selection operator, four multi-objective metaheuristics are separated decisively by flat cross-validation and hypervolume ($p = 2.1 \times 10^{-11}$) yet cannot be distinguished under nested cross-validation at the resolution this design affords ($p = 0.119$), with the mean-rank order differing between the two protocols. A controlled budget sweep shows optimistic bias rising monotonically with search effort while honest error does not improve, and a label permutation identifies noise-fitting as the mechanism. Near-optimal solutions are massively degenerate, including disjoint pairs with equivalent error, which explains the low reported stability as non-identifiability rather than algorithmic failure. Wavelength selection here delivers a twenty-fold reduction in channel count at unchanged accuracy; reported accuracy gains are protocol artefacts. We recommend that Pareto quality indicators in this problem class be computed on out-of-sample objectives and accompanied by permutation and random-subset controls.

Hypervolume by algorithm under the matched cardinality prior for target densities t from 5% to 20% (five seeds each). NSGA-II is the reference (nvar 16, HV 6.851). Across the entire range the HV gap between NSGA-II and the next-best algorithm stays at or below +0.160, far smaller than the +0.518 gap under the original operators; the 94% collapse reported in the main text is therefore not an artefact of the particular value $t = 0.08$. Values are wavelength count (n) and hypervolume (HV).

5. Limitations

Several constraints bound what these results establish. The first is dimensionality. With 50 samples and 474 candidate channels the problem is severely underdetermined, so a cross-validated error computed on any given subset carries large variance and many subsets are statistically indistinguishable. This is precisely the regime the paper sets out to examine, since it is where the variance of the selection criterion is largest and where selection bias is therefore most consequential, but it also means the magnitudes reported here are characteristic of small calibration sets rather than universal. A study with several hundred samples would show the same mechanisms attenuated. The second constraint is that all results come from one dataset and one response variable. The mechanisms are general in principle because they follow from the statistics of estimated selection criteria rather than from any property of this material, but the magnitudes are not transferable and no ranking claim should be generalised from a single instance. Confirming these findings on public NIR benchmarks and on synthetic spectra with planted ground-truth bands, where selection accuracy is directly measurable, is the necessary next step. The third constraint concerns the reference values. Each sample was measured once by Kjeldahl analysis, so no replicate-based measurement uncertainty is available. Unquantified reference noise enters the residual that the selection criterion minimises and thereby adds to the noise-fitting opportunity documented here, which means the optimistic gaps we report may understate what a pipeline with replicated references would attribute to search rather than to signal. The fourth constraint is the use of untuned hyperparameters. Common literature defaults were chosen so that any difference in outcome is attributable to search behaviour rather than to unequal tuning effort, but this choice may disadvantage an algorithm whose defaults suit this problem poorly, and the flat ranking reported here could change under tuning. That possibility does not weaken the central argument. Tuning is itself a search over the same estimated criterion, so by the budget-sweep result it would be expected to increase the optimistic gap for whichever algorithm receives the most tuning effort. Finally, only protein could be modelled as a response, since the remaining reference variables were below the limit of detection, extremely skewed or nearly constant.

6. Future directions

Future work will validate the nested-evaluation and cardinality-matched protocols on independent sample batches and additional analytes (for example nitrate), extend the benchmark to non-linear learners such as PLS-DA and support vector machines where larger datasets permit, and assess transfer of the reporting protocol to on-line NIR screening under industrial conditions.

7. Acknowledgements

Data availability. Data will be made available on request.

AI-use disclosure. During the preparation of this work, the authors used Claude Opus 4.8 to improve the readability and language of the manuscript. After using this tool/service, the authors carefully reviewed and edited the content as necessary and certified that the technology was used under full human management. The authors take full responsibility for the content of the published article.

Author contributions / Funding / Conflicts of interest. Author contributions (CRediT). Nonpravit Chianrungsaeng: conceptualization, methodology, software, formal analysis, investigation, data curation, writing – original draft, visualization. Somboon Sukpancharoen: conceptualization, methodology, supervision, writing – review and editing, resources. Both authors read and approved the final manuscript. Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Conflicts of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8. References

- [1] Zou X, Zhao J, Povey MJW, Holmes M, Mao H. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*. 2010;667(1-2):14-32. <https://doi.org/10.1016/j.aca.2010.03.048>.
- [2] Yun YH, Li HD, Deng BC, Cao DS. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *Trends in Analytical Chemistry*. 2019;113:102-15. <https://doi.org/10.1016/j.trac.2019.01.018>.
- [3] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*. 1999;3(4):257-71. <https://doi.org/10.1109/4235.797969>.
- [4] Ambrose C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*. 2002;99(10):6562-6. <https://doi.org/10.1073/pnas.102102699>.
- [5] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>.
- [6] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*. 2010;11:2079-107.
- [7] Kuncheva LI. A stability index for feature selection. In: *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*; 2007 Feb 12-14; Innsbruck, Austria. Anaheim: ACTA Press; 2007. p. 390-5.
- [8] Sorensen K. Metaheuristics-the metaphor exposed. *International Transactions in Operational Research*. 2015;22(1):3-18. <https://doi.org/10.1111/itor.12001>.
- [9] Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*. 1989;43(5):772-7. <https://doi.org/10.1366/0003702894202201>.
- [10] Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 2001;58(2):109-30. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1).
- [11] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*. 2002;6(2):182-97. <https://doi.org/10.1109/4235.996017>.
- [12] Kennedy J, Eberhart R. Particle swarm optimization. In: *Proceedings of the IEEE International Conference on Neural Networks*; 1995 Nov 27-Dec 1; Perth, Australia. Piscataway: IEEE; 1995. p. 1942-8. <https://doi.org/10.1109/ICNN.1995.488968>.
- [13] Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm. In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*; 1997 Oct 12-15; Orlando, USA. Piscataway: IEEE; 1997. p. 4104-8. <https://doi.org/10.1109/ICSMC.1997.637339>.
- [14] Storn R, Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*. 1997;11(4):341-59. <https://doi.org/10.1023/A:1008202821328>.
- [15] Mirjalili S, Mirjalili SM, Lewis A. Grey Wolf Optimizer. *Advances in Engineering Software*. 2014;69:46-61. <https://doi.org/10.1016/j.advengsoft.2013.12.007>.
- [16] Emary E, Zawbaa HM, Hassanien AE. Binary grey wolf optimization approaches for feature selection. *Neurocomputing*. 2016;172:371-81. <https://doi.org/10.1016/j.neucom.2015.06.083>.
- [17] Li H, Liang Y, Xu Q, Cao D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*. 2009;648(1):77-84. <https://doi.org/10.1016/j.aca.2009.06.046>.
- [18] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: *Advances in Neural Information Processing Systems 9*. Cambridge: MIT Press; 1997. p. 155-61.
- [19] Derrac J, García S, Molina D, Herrera F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*. 2011;1(1):3-18. <https://doi.org/10.1016/j.swevo.2011.02.002>.
- [20] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005;21(2):171-8. <https://doi.org/10.1093/bioinformatics/bth469>.

9. Appendix

Table A1 Sensitivity of the cardinality-matched comparison to the target density t

Target density t	BIAS	MO-BPSO (n, HV)	MO-BDE (n, HV)	MO-bGWO (n, HV)	HV gap (NSGA-II – next best)
5%	2.94	23, 6.647	8, 6.782	7, 6.645	+0.069
8% (primary)	2.44	39, 6.454	10, 6.842	10, 6.752	+0.009
12%	1.99	64, 6.401	14, 6.806	13, 6.767	+0.045
15%	1.73	70, 6.377	24, 6.733	22, 6.774	+0.077
20%	1.39	90, 6.366	34, 6.572	27, 6.691	+0.160

Table A2 Reference-variable summary for the 50 samples

Variable	n	Mean	SD	Min	Max	Range	CV(%)
Calorific (J/g)	50	15979.28	1072.711	12460.0	17767.0	5307.0	6.713
Nitrate (ppm)	50	88.9	105.817	0.0	250.0	250.0	119.029
Nitrite (ppm)	50	0.0	0.0	0.0	0.0	0.0	
Protein_lab (%)	50	35.135	8.687	15.44	49.07	33.63	24.723
Protein_nir (%)	50	34.944	8.159	17.83	44.37	26.54	23.349

Table A3 Per-fold nested RMSE across all seeds, showing that fold identity dominates the between-algorithm difference

Algorithm	Mean	SD	Min	Max	Range
NSGA-II	2.335	0.617	1.034	3.898	2.865
MO-BPSO	2.296	0.622	1.238	3.377	2.139
MO-BDE	2.284	0.564	1.074	3.190	2.117
MO-bGWO	2.238	0.566	1.126	3.377	2.251

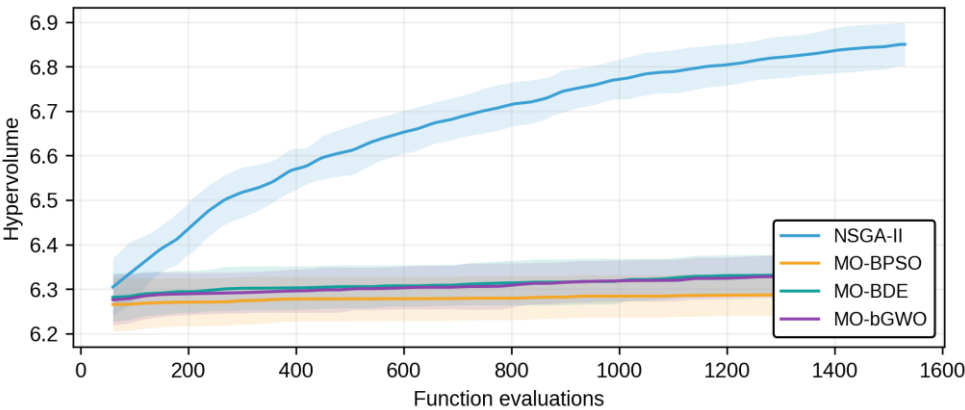


Figure A1 Convergence of hypervolume against the number of function evaluations for each algorithm. Lines are means over 20 seeds and shaded bands are one standard deviation. All four searches reach a stable hypervolume within the matched budget of about 1530 evaluations.

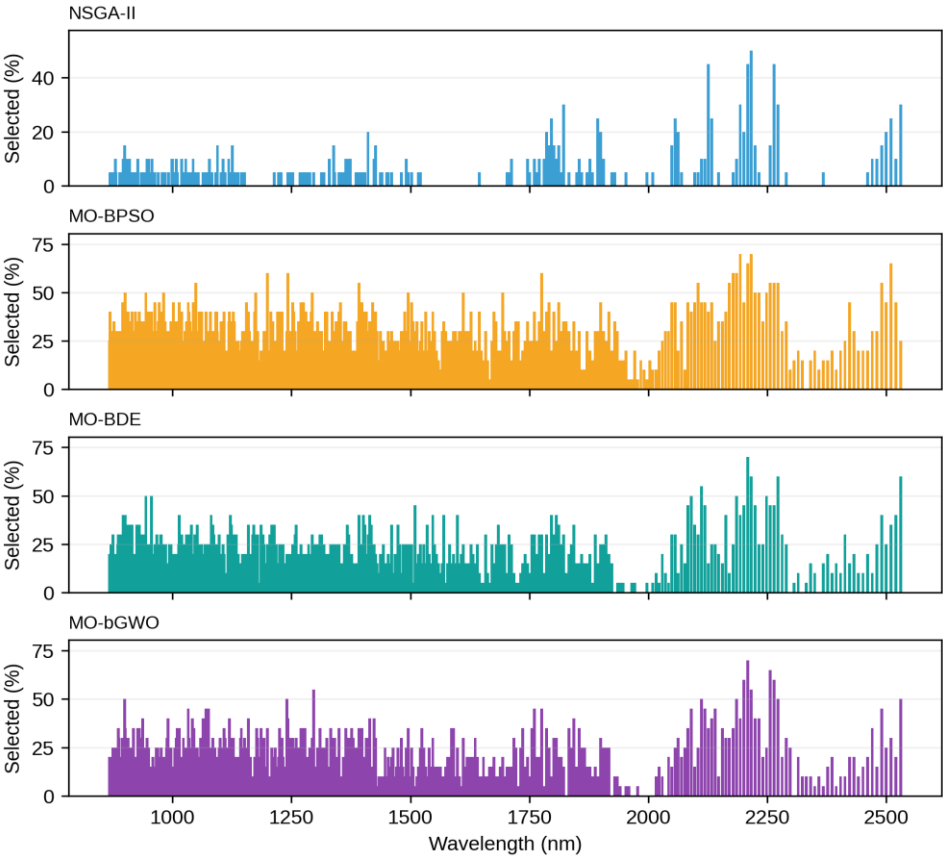


Figure A2 Selection-frequency profiles across 20 seeds. Bars give the percentage of seeds in which each wavelength was retained. No wavelength is retained consistently and the loosely preferred regions near 2100 to 2300 nm coincide with the high correlation bands of Figure 2