



When spectral libraries are too complex to search: Evolutionary subset selection for domain-adaptive calibration

Leonardo Ramirez-Lopez ^{a,b},*, Raphael Viscarra Rossel ^c, Claudio Orellano ^b, Laurens Kooijman ^b, Estefania Perez-Fernandez ^b, Alexandre M.J.-C. Wadoux ^e, Marçal Plans ^b, Timo Breure ^f, Laura Summerauer ^g, José L. Safanelli ^h, Thorsten Behrens ^d, Ricard Boqué ⁱ

^a Imperial College London, Imperial College Business School, South Kensington Campus, London, SW7 2AZ, England, United Kingdom

^b BUCHI Labortechnik AG, Department of Data Science, Meierseggsstrasse 40, Flawil, CH-9230, Switzerland

^c Curtin University, Soil and Landscape Science, School of Molecular and Life Sciences, GPO Box U1987, Perth, WA 6845, Australia

^d Swiss Competence Center for Soil, Länggasse 85, Zollikofen, CH-3052, Switzerland

^e College of Science and Engineering, James Cook University, Cairns, Queensland, Australia

^f European Commission, Joint Research Centre (JRC), Via E. Fermi, 2749, Ispra, 21027, Italy

^g Research Institute of Organic Agriculture FiBL, Department of Soil Sciences, Ackerstrasse 113, Frick, CH-5070, Switzerland

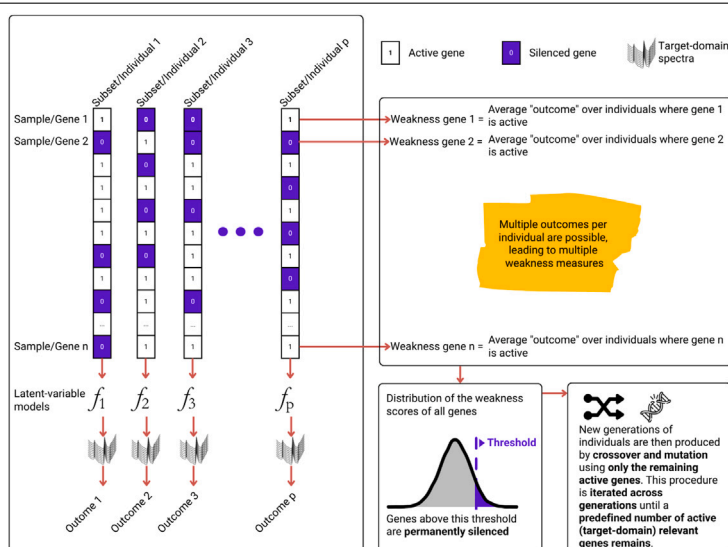
^h Woodwell Climate Research Center, 149 Woods Hole Road, Falmouth, 02540, MA, USA

ⁱ Universitat Rovira i Virgili, ChemoSens group, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, Edifici N4, C/ Marcel·lí Domingo 1, Tarragona, 43007, Catalunya, Spain

HIGHLIGHTS

- A method to exploit large spectral libraries for target-domain modelling is presented.
- This new method, called **gesearch**, is based on evolutionary sample selection.
- **gesearch** selects samples spectrally and compositionally coherent with the target domain.
- **gesearch** performs non-linear domain alignment while maximising predictive accuracy.
- **gesearch** builds interpretable target-domain models from complex spectral libraries.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Lutgarde Buydens

ABSTRACT

Background: The increasing availability of large spectral libraries offers new opportunities to reduce the costs and efforts required to develop spectroscopy-based sensing techniques for rapidly and non-destructively

* Corresponding author at: Imperial College London, Imperial College Business School, South Kensington Campus, London, SW7 2AZ, England, United Kingdom.
E-mail addresses: ramirez-lopez.l@buchi.com, leonardo.ramirez-lopez23@imperial.ac.uk (L. Ramirez-Lopez), alexandre.wadoux@jcu.edu.au (A.M.-C. Wadoux).

<https://doi.org/10.1016/j.aca.2026.345651>

Received 23 February 2026; Received in revised form 4 May 2026; Accepted 11 May 2026

Available online 14 May 2026

0003-2670/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Dataset link: <https://CRAN.R-project.org/package=resemble>, <https://github.com/l-ramirez-lopez/resemble>, <https://l-ramirez-lopez.github.io/resemble/>

Keywords:

Evolutionary search
Chemometrics
Quantitative spectroscopy
Stochastic optimisation
Cross-domain modelling
Transfer learning
Diffuse reflectance spectroscopy

estimating key properties across environmental and agricultural matrices. These libraries can provide training samples for developing models adapted to specific target domains. However, identifying which samples are most relevant for a given target domain remains challenging. This study introduces *gesearch*, a non-linear evolutionary algorithm for selecting target-domain-relevant training samples from complex spectral libraries to build accurate and interpretable quantitative models.

Results: The *gesearch* method was used to extract a subset of relevant samples from a large North American infrared soil spectral library to build predictive models of total carbon for an independent target area in the Democratic Republic of the Congo. In this challenging cross-domain test case, simple linear models built with the training samples found by *gesearch* achieved superior accuracy compared with established modelling approaches, including LOCAL, Cubist, convolutional neural networks, and global partial least squares regression.

Significance: The proposed method provides a practical framework for exploiting large heterogeneous spectral libraries when only limited target-domain reference data are available. By selecting samples that are spectrally and compositionally coherent with the target domain, *gesearch* supports accurate, compact, and interpretable calibration models. It can also operate when only unlabelled target-domain spectra are available.

1. Introduction

The sustainable management of environmental and agricultural resources increasingly depends on rapid, cost-effective methods for assessing their chemical and physical properties. Diffuse reflectance spectroscopy (DRS) has emerged as a viable method for the rapid, non-destructive, and scalable assessment of multiple properties of complex matrices such as soils [1,2]. DRS operates by analysing interactions between electromagnetic radiation and matter, with quantitative predictions typically obtained through empirical calibration models linking spectral measurements to laboratory reference data [3]. Accordingly, for quantitative analyses of complex matrices, DRS techniques such as near-infrared and infrared spectroscopy are generally regarded as secondary analytical methods because they rely on calibration against reference measurements.

A persistent challenge in DRS applications is the development of suitable training sets for specific target populations. Within such populations, the relationships between DRS features and responses are often approximately linear, although they may vary across populations [4]. When this is the case, these relationships can be captured by collecting a training set that adequately represents the target population, and building linear models that perform well in such domains is therefore relatively straightforward. In contrast, when the number of available training samples is limited, model generalisation can be severely hampered [5]. In such cases, predictive models are often constructed by leveraging large and diverse pre-existing DRS spectral libraries. However, due to their inherent heterogeneity, not all samples in a spectral library may be informative for a given target population (i.e. some may be irrelevant, noisy, or even misleading, contributing to the degradation of the predictive performance). As a result, general-purpose models trained on all available library samples do not necessarily guarantee optimal accuracy in a specific target domain [6]. In this context, Viscarra Rossel et al. [7] highlighted the critical role of transfer-learning strategies for addressing domain mismatch and data scarcity in target-domain modelling. Within this framework, library search and spectroscopic localisation methods can be viewed as instance-based, homogeneous transfer-learning approaches that mitigate domain mismatch through selective exploitation of source-domain samples.

Rather than adapting a global model to multiple target domains, an alternative is to construct domain-specific models directly from comprehensive and heterogeneous spectral libraries. Library search and spectroscopic localisation methods implement this strategy by identifying and exploiting library samples that are most relevant to a given target population, thereby enabling targeted modelling without explicit model adaptation. Most library search and spectroscopic localisation methods employed in chemometrics (e.g. [8–12]) rely primarily on spectral similarity metrics. However, spectral similarity alone does not necessarily reflect similarity in the underlying spectrum–response

relationships, particularly under domain shift, where marginal and conditional distributions may differ substantially between source and target populations.

To address some of the limitations of purely deterministic similarity-based retrieval, Lobsey et al. [4] introduced the *rs-local* algorithm (re-sampling strategy for site-specific calibrations), which formulated library search as a stochastic optimisation problem. Rather than selecting samples solely based on spectral proximity, *rs-local* iteratively identifies subsets of library samples that minimise prediction error on a small labelled target set. This data-driven strategy was shown, in some cases, to provide improved predictive performance relative to similarity-based methods when applied to large and heterogeneous spectral libraries [13]. These results indicate that, under certain conditions, predictive relevance of samples may not be fully explained by spectral similarity alone. As such, *rs-local* represents an important advance in instance-based library search for target-domain modelling under data scarcity.

Despite its demonstrated effectiveness, the optimisation strategy employed by *rs-local* is primarily driven by predictive error on a limited labelled target set. While this objective is well aligned with improving target-domain accuracy, it does not explicitly account for the representativeness of the selected samples with respect to the broader spectral variability of the target population. In particular, for variable-projection methods such as partial least squares regression, it is critical that the projected spectra preserve the structure of the spectral variance expected in the target domain. Subsets selected solely to minimise prediction error may fail to properly represent this variance, resulting in latent-space distortions which ultimately lead to model instability. In addition, *rs-local* provides no explicit mechanism to control overfitting and requires response measurements in the target set, which may be costly to obtain [4].

Taken together, these considerations suggest that instance-based transfer learning would benefit from optimisation strategies that move beyond purely error-driven selection, explicitly balancing predictive performance with preservation of target-domain spectral structure, while reducing dependence on labelled target data. In this context, this work introduces *gesearch* (genetic evolutionary search), which operationalises evolutionary optimisation within an instance-based transfer-learning framework for spectroscopic library search. Rather than relying on purely error-driven sample elimination, *gesearch* adopts a gene-centric optimisation strategy in which individual library samples are evaluated based on their contribution to reducing both marginal and conditional distribution mismatch. In doing so, *gesearch* addresses key limitations of *rs-local*, including the lack of explicit preservation of target-domain spectral variance and the reliance on response measurements in the target set. The proposed framework relaxes the requirement for target-domain reference values, while providing a systematic mechanism for exploiting large, heterogeneous spectral libraries to improve target-domain predictive performance.

The primary objectives of this work are to: (i) reformulate instance-based library search within an evolutionary-search framework through

gene-centric optimisation of sample selection; (ii) develop a sample-selection strategy that explicitly balances predictive performance with preservation of target-domain spectral structure; and (iii) assess the predictive accuracy and data efficiency of *gesearch* relative to *rs-local* and other state-of-the-art methods commonly used in DRS.

2. Theory

The method presented here aims to exploit large, diverse spectral libraries to produce accurate DRS predictive models for target domains for which limited data are available. Let

$$D_u := (\mathbf{X}_u, \mathbf{y}_u) = \{(\mathbf{x}_u^i, y_u^i)\}_{i=1}^m \quad (1)$$

denote the target set of the population of interest (where the subscript u indicates a partially “unknown” target domain), consisting of m paired observations $(\mathbf{x}_u^i, y_u^i)$. Here, $\mathbf{x}_u^i \in \mathbb{R}^d$ represents the spectrum of sample i measured at d wavelengths or wavenumbers (a d -dimensional feature vector), and $y_u^i \in \mathbb{R}$ denotes the corresponding scalar reference measurement obtained by conventional laboratory analysis. Note that y_u may be partially missing or entirely unavailable for the samples of the target domain. Similarly, let

$$D_r := (\mathbf{X}_r, \mathbf{y}_r) = \{(\mathbf{x}_r^j, y_r^j)\}_{j=1}^n \quad (2)$$

denote the spectral library (where the subscript r indicates a “reference” set), a pre-existing collection of (n) candidate samples for model training. This library is typically large and heterogeneous and does not necessarily contain observations from the target population itself, i.e. $D_u \not\subseteq D_r$. Nevertheless, it is assumed to include at least some samples that are spectrally and compositionally relevant to the target domain. Otherwise, searching the library for useful calibration samples would not be meaningful.

The available target set D_u is assumed to satisfy $n \gg m$ and to be representative of the target population, in the sense that it provides a reasonable approximation of its basic statistical properties.

Note that the assumption of having only a few reference values for the target set reflects a realistic scenario. Collecting a limited number of samples with laboratory reference measurements is typically feasible in terms of cost and time. Moreover, because *rs-local* relies entirely on samples with known reference values, a small subset of such samples was introduced to ensure a fair comparison between *rs-local* and *gesearch*.

2.1. The *gesearch* algorithm: an evolutionary search method for domain-relevant sample selection

2.1.1. Background

The *gesearch* algorithm is inspired by the *rs-local* spectral library search method of Lobsey et al. [4]. In *rs-local*, given the target set D_u and library D_r , the aim is to identify an optimal subset of D_r for calibration. The method iteratively removes irrelevant samples until a stopping condition, typically a target subset size, is met. At each iteration, random subsets of D_r are drawn. Each subset is then used to predict y_u from $\mathbf{X}_u$ and the prediction error is calculated. Each sample in D_r receives an error score, which is the average prediction error across all subsets in which it appears. Those samples with the largest error score are discarded at the end of each iteration as these are considered irrelevant, i.e. samples that consistently degrade predictive performance.

Compared to local regression methods, both *gesearch* and *rs-local* reduce reliance on neighbour selection driven by a single global similarity metric, as different similarity definitions can emphasise different aspects of the data and lead to different selections in complex and heterogeneous datasets [14].

2.1.2. *Gesearch*

The proposed *gesearch* algorithm provides an evolutionary search method to select an optimal subset of library samples that maximises predictive performance and descriptive relevance for a target domain. Unlike *rs-local*, it provides a formal optimisation basis, enables extensibility, and can operate without requiring response values from the target set.

In the *gesearch* algorithm, the spectral library D_r is conceptualised as a pool of n genes, where each gene $g_i = (\mathbf{x}_r^i, y_r^i)$ corresponds to a training sample. These genes are carried by individuals that together form a population. Here, an individual is simply a representation of a candidate subset of samples. To remain faithful to the genetic analogy, each individual is encoded as a binary vector over the full set of n genes, where genes marked with 1 are active (belong to the subset) and genes marked with 0 are silent (do not belong to the subset). Thus, all individuals share the same underlying genetic structure, but differ in which genes are expressed. Only the subset of (k) active genes contributes to model construction. Each individual defines a candidate calibration model that predicts the response variable y_u from the spectra $\mathbf{X}_u$ in the target set. Unlike standard genetic algorithms, evaluation in *gesearch* is gene-centric: the weakness of each individual (e.g. how weak a gene is in contributing to a good prediction) is first quantified, and then, for each gene g_i , an average weakness score is computed across all individuals in which g_i is active. Thus, each gene in the pool receives a weakness score quantifying its marginal effect on the performance of an individual. The score can be based on predictive accuracy, on the descriptive power of the spectra as quantified through preservation of latent spectral structure, on spectral similarity to the target domain, or on any combination of these criteria. Weak genes are identified and permanently silenced, i.e. they are irreversibly inactivated such that they remain present in the representation but are no longer expressed in subsequent generations. New generations are produced through crossover and mutation. After a stopping criterion is met, the algorithm returns a final set of genes (e.g. the samples with the highest predictive and descriptive power for the target set). Thus, *gesearch* indirectly improves individual fitness (consistent with evolutionary algorithms) by minimising gene-level weakness, i.e., silencing high-weakness genes across generations.

Fig. 1 illustrates a simplified example of *gesearch* on a fictitious library of 40 genes/samples. Each individual expresses exactly 15 genes, where blue is an active gene and black a silenced one. The first generation contains 27 individuals, with diversity arising from random silencing of genes. After evaluation, genes whose weakness exceeds the $(1 - \alpha)$ -quantile are permanently silenced for future generations (i.e. α constitutes a hyperparameter). Weakness is defined as the marginal contribution of a gene to model performance on the target set: genes whose active presence systematically degrades the predictive or descriptive accuracy of individuals receive higher weakness scores. Generation 2 offspring are formed by crossing two individuals (parents) and selecting 15 genes from the union of their active genes (crossover); if the union is short (i.e. less than 15 active genes), additional genes are sampled from the remaining active pool (mutation). Silenced genes never re-enter. The representation factor $b = 10$ targets each gene to appear in roughly 10 individuals per generation, so the row-wise counts $(h_j^{(r)})$ are close to 10. The same procedure produces Generation 3. Generation 3 already shows a pattern of diversity reduction (i.e. individuals share a larger fraction of the same active genes), and only the fittest (i.e. low-weakness) genes persist, indicating progression towards a final active set.

2.1.3. First generation of individuals and basic gene weakness principles

For *gesearch*, each element of the spectral library D_r is referred to as a gene g_j , so that

$$G = \{g_j\}_{j=1}^n = D_r. \quad (3)$$

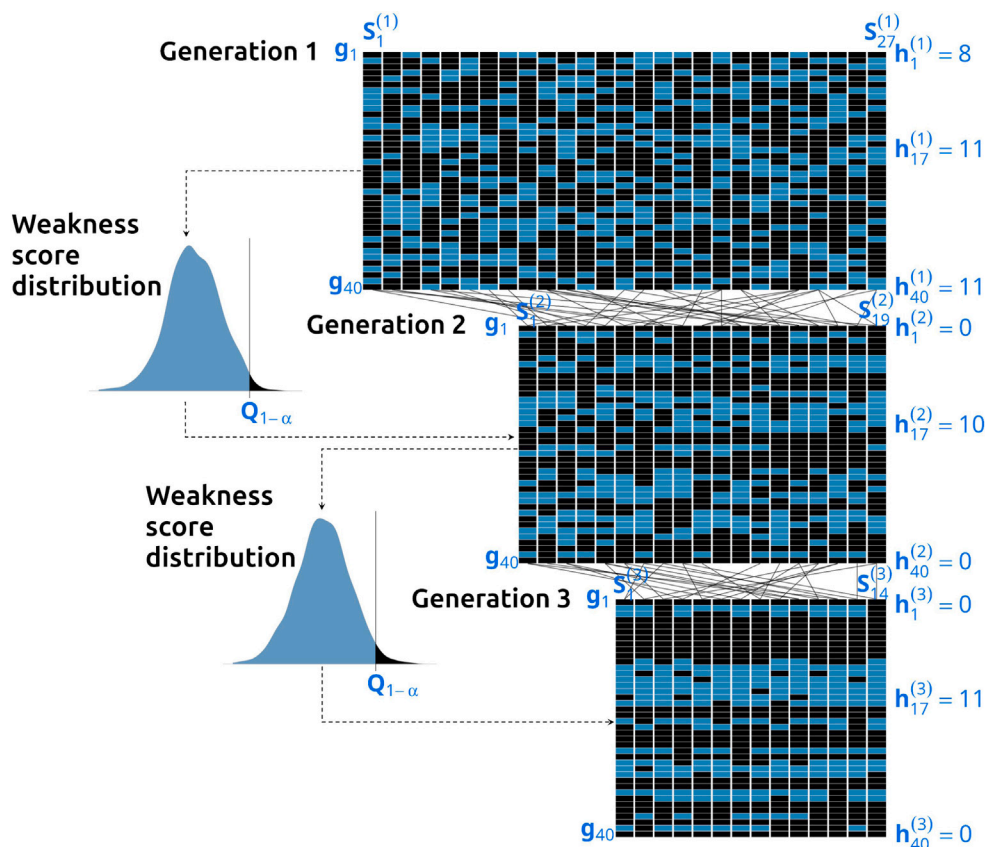


Fig. 1. Illustration of the gesearch algorithm across three generations with 40 samples, parameter values $b = 10$ and $k = 15$, and 27 individuals in the initial generation. Black cells represent silenced genes and blue cells denote active genes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

An initial population of p individuals (fixed *a priori*) is generated, where each individual corresponds to a unique subset of k active genes randomly selected from the n available ones ($k \ll n$); the remaining genes are silenced. Formally, an individual at generation γ is denoted by

$$S_i^{(\gamma)} \subset G, \quad |S_i^{(\gamma)}| = k, \quad (4)$$

where membership indicates that a gene is active in that individual. The set of individuals, denoted $S^{(1)} = \{S_i^{(1)}\}_{i=1}^p$, thus represents the entire population at the first generation. Simply, a set of individuals can be seen as different random subsets of genes (samples) drawn from the pool of non-permanently silenced genes in the spectral library. Each individual is thus characterised by its subset of active genes:

$$S_i^{(1)} = \{g_{j_1}, g_{j_2}, \dots, g_{j_k}\}, \quad S_i^{(1)} \subseteq G, \quad (5)$$

where the indices $j_1, \dots, j_k$ are sampled without replacement from $1, \dots, n$ (the spectral library). Here, $G = \{g_{j=1}^n\}$ is the full gene pool (the spectral library), and each $S_i^{(1)}$ corresponds to a unique expression pattern in which exactly k genes are active. The set $S^{(1)}$ constitutes the first generation of individuals and serves as the starting point for the evolutionary process.

The approximate number of individuals in which any given gene should be active is controlled by a fixed hyperparameter b (a gene-representation factor). For the first generation, the required population size is

$$p^{(1)} = \left\lceil \frac{b \cdot n}{k} \right\rceil \quad (6)$$

Note that the value of k should not be too large, as the contribution of each gene decreases as the number of genes within an individual

increases, making it harder to accurately assess the influence of single genes.

To identify and eliminate weak genes during the evolutionary process, gesearch considers three complementary weakness functions, each providing a different criterion for assessing the contribution of a gene. These include response weakness $w_r^{(\gamma)}(g_j)$, similarity weakness $w_s^{(\gamma)}(g_j)$, and reconstruction weakness $w_d^{(\gamma)}(g_j)$. Each of these functions is optional, but at least one must be applied.

Ensuring that each gene is expressed in about b individuals allows its average contribution to weakness to be quantified. Consider a population of 10 000 individuals (p), where each individual contains 5000 genes, of which only $k = 50$ are active. On average, each gene in the pool would therefore appear in about $b = 100$ individuals, according to Eqs. (6) and (14). The reconstruction weakness score of a gene is defined as the average reconstruction error computed across all individuals in which that gene is active. The same rationale applies to the response and similarity weakness scores.

PLS modelling basis for weakness evaluations

All weakness functions in gesearch are evaluated using partial least squares (PLS; [15]) models fitted at the individual level. For each individual $S_i^{(\gamma)}$, a PLS model $\hat{f}_i^{(\gamma)}$ is calibrated on the k active genes encoded by that individual. A fixed number of latent components κ (selected *a priori*) is used across all individuals and generations. This design choice ensures comparability of weakness scores across individuals and evolutionary iterations, and avoids distorting the weakness evaluation with changes in latent dimensionality. From the resulting PLS models, the different weakness functions defined below are derived.

Response weakness function

The response weakness quantifies the contribution of a gene to predictive accuracy in the target domain. For each gene, the weakness

score is computed by averaging a chosen performance metric over only those individuals in which the gene is active. The performance metric used here is the target-domain prediction error, estimated using a small number of target-domain genes/samples with available response values. If no reference values are available for the target domain, this weakness function cannot be evaluated. However, *gesearch* can still be applied using one or both of the other weakness functions presented below.

Let γ denote the generation index. Define the incidence count of the j th gene as:

$$h_j^{(\gamma)} = \sum_{i=1}^{p^{(\gamma)}} I_{ij}^{(\gamma)}, \quad (7)$$

$$I_{ij}^{(\gamma)} = \begin{cases} 1, & \text{if } g_j \in S_i^{(\gamma)}, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

and the response weakness of gene g_j as

$$w_r^{(\gamma)}(g_j) = \frac{1}{h_j^{(\gamma)}} \sum_{i: g_j \in S_i^{(\gamma)}} e(S_i^{(\gamma)}). \quad (9)$$

Here, $e(S_i^{(\gamma)})$ denotes the root-mean-squared prediction error obtained by applying the individual-level PLS model $\hat{f}_i^{(\gamma)}$, previously defined, to the target spectra $\mathbf{X}_u$ and their corresponding reference values $\mathbf{y}_u$. Additional implementation details are provided in [Appendix A](#).

Reconstruction weakness function

In contrast to the response weakness, which relies on target-domain reference values, the reconstruction weakness exploits the latent structure of the individual-level PLS models to evaluate gene relevance independently of response information. Because PLS is projection-based, each individual $S_i^{(\gamma)}$ defines a latent space that can be used to reconstruct the target spectra $\mathbf{X}_u$. If an individual is composed of strong, target-aligned genes, its reconstruction error is small, otherwise, it is large.

For an individual $S_i^{(\gamma)}$ using its κ latent components, let $\tilde{\mathbf{X}}_{u,\kappa}^{(i,\gamma)}$ denote the reconstruction of $\mathbf{X}_u$ obtained from its corresponding PLS model $\hat{f}_i^{(\gamma)}$. The squared Frobenius reconstruction error is defined as

$$q_i^{(\gamma)}(\kappa) = \|\mathbf{X}_u - \tilde{\mathbf{X}}_{u,\kappa}^{(i,\gamma)}\|_F^2. \quad (10)$$

The reconstruction weakness for an active gene g_j is computed as the incidence-normalised (see Eqs. (7) and (8)) average of these errors over all individuals in which the gene is expressed:

$$w_q^{(\gamma)}(g_j) = \frac{1}{h_j^{(\gamma)}} \sum_{i: g_j \in S_i^{(\gamma)}} q_i^{(\gamma)}(\kappa). \quad (11)$$

The reconstruction weakness scores can be interpreted as indicators of how well each gene contributes to reconstructing $\mathbf{X}_u$. Genes with low $w_q^{(\gamma)}(g_j)$ values are assumed to contribute positively to an accurate reconstruction of $\mathbf{X}_u$. This criterion is crucial for PLS-based modelling, as reliable reconstruction ensures that the latent projections remain representative of the target domain, thereby reducing model uncertainty. See [Appendix A](#) for additional details.

Similarity weakness function

As with the reconstruction weakness, this score can be computed without target-domain reference values, since it relies exclusively on projections of the target spectra. The similarity weakness quantifies how well the latent representation of an individual aligns with the target domain. For individual $S_i^{(\gamma)}$, let $\bar{\mathbf{T}}_{i,\kappa}^{(i,\gamma)}$ denote the centroid of its PLS score representation, and let $\bar{\mathbf{Z}}_{u,\kappa}^{(i,\gamma)}$ denote the centroid of the target spectra projected into the same latent space. The Mahalanobis distance between these centroids is defined as

$$D_i^{(\gamma)} = (\bar{\mathbf{Z}}_{u,\kappa}^{(i,\gamma)} - \bar{\mathbf{T}}_{i,\kappa}^{(i,\gamma)})^\top (\Sigma_{i,\kappa}^{(\gamma)})^{-1} (\bar{\mathbf{Z}}_{u,\kappa}^{(i,\gamma)} - \bar{\mathbf{T}}_{i,\kappa}^{(i,\gamma)}), \quad (12)$$

where $\Sigma_{i,\kappa}^{(\gamma)}$ is the covariance matrix of the individual's score representation. The similarity weakness of gene g_j is obtained by aggregating these distances over all individuals in which the gene is expressed, normalised by the incidence count $h_j^{(\gamma)}$ defined in Eqs. (7) and (8):

$$w_d^{(\gamma)}(g_j) = \frac{1}{h_j^{(\gamma)}} \sum_{i: g_j \in S_i^{(\gamma)}} D_i^{(\gamma)}. \quad (13)$$

Thus, $w_d^{(\gamma)}(g_j)$ represents the average distance between the centre of the target domain and the centres of the individuals in which gene g_j is active, measured in the latent space. Lower values indicate better alignment. See [Appendix A](#) for additional details.

2.1.4. Mechanisms of offspring generation

At generation γ , if $n^{(\gamma)} = |G^{(\gamma)}|$ denotes the number of genes allowed to be expressed, then the corresponding population size is

$$p^{(\gamma)} = \left\lceil \frac{b n^{(\gamma)}}{k} \right\rceil \quad (14)$$

Define the active genes at generation γ as $G^{(\gamma)} \subseteq G$. For the first generation, set $G^{(1)} := G$. After weakness-based gene silencing in generation γ , let $U^{(\gamma)} \subseteq G^{(\gamma)}$ be the set of silenced genes. The pool of genes available for expression in the next generation is then

$$G^{(\gamma+1)} := G^{(\gamma)} \setminus U^{(\gamma)}. \quad (15)$$

The generation- γ population is $S^{(\gamma)} = \{S_1^{(\gamma)}, \dots, S_{p^{(\gamma)}}^{(\gamma)}\}$, where each $S_i^{(\gamma)} \subseteq G^{(\gamma)}$ and $|S_i^{(\gamma)}| = k$.

To generate offspring, sample two parents $S_i^{(\gamma)}$ and $S_j^{(\gamma)}$. If $|S_i^{(\gamma)} \cup S_j^{(\gamma)}| > k$, choose k genes at random from the union (crossover); otherwise, if $< k$, add random genes from $G^{(\gamma)} \setminus U^{(\gamma)}$ to reach k (mutation). If random genes are added, these must not already be in either $S_i^{(\gamma)}$ or $S_j^{(\gamma)}$. Genes in $U^{(\gamma)}$ remain silenced. Parent pairs may produce multiple offspring.

Gene selection is performed independently for each weakness score. Any gene exceeding the $(1 - \alpha)$ quantile for any score is permanently silenced:

$$\begin{aligned} w_{r,j}^{(\gamma)} &\geq Q_{1-\alpha}(\{w_{r,\ell}^{(\gamma)}\}_{\ell=1}^{n^{(\gamma)}}) \\ \text{Silence } g_j \text{ if } &\vee w_{q,j}^{(\gamma)} \geq Q_{1-\alpha}(\{w_{q,\ell}^{(\gamma)}\}_{\ell=1}^{n^{(\gamma)}}) \\ &\vee w_{d,j}^{(\gamma)} \geq Q_{1-\alpha}(\{w_{d,\ell}^{(\gamma)}\}_{\ell=1}^{n^{(\gamma)}}). \end{aligned} \quad (16)$$

Let

$$U^{(\gamma)} = \{g_j \in G^{(\gamma)} : \text{any of the above holds}\}, \quad (17)$$

with $G^{(\gamma+1)}$ updated accordingly.

Using the quantile makes the number of silenced genes adapt to the empirical score distributions in each generation.

2.1.5. Subsequent generations

An individual at generation γ is a set with k active genes

$$S_i^{(\gamma)} = \{g_{j_1}, g_{j_2}, \dots, g_{j_k}\}, \quad S_i^{(\gamma)} \subseteq G, \quad (18)$$

At $\gamma = 1$, indices are sampled without replacement from G ; for $\gamma > 1$, $S_i^{(\gamma)}$ is formed by crossover/mutation of two parents in $S^{(\gamma-1)}$.

The procedure of irreversible gene silencing, and offspring generation via crossover and mutation are iterated to form successive generations. As weak genes are progressively silenced, the overall subset of active genes becomes smaller in each round. Since every individual must contain k active genes, this reduction in the pool of active genes naturally limits the diversity of new offspring: with fewer active genes available, individuals become increasingly similar to each other.

2.1.6. Stopping criterion

Evolution terminates when the active gene pool size falls below a specified minimum. The genes that remain active are considered relevant to the target domain. Note that this stopping criterion does not guarantee convergence of the weakness scores. This termination rule simply enforces a practical limit on the search process rather than ensuring full stability or convergence of the underlying optimisation dynamics, which can be considered a limitation, since the search may terminate before the weakness scores converge to a stable optimum.

2.1.7. Final model fitting

The active genes identified at the end of the evolutionary search are used to construct the final predictive model of y_u from X_u . This model is calibrated using PLS regression on the selected subset of samples from the spectral library. When the samples from the target domain that were used to guide the search have response values available, these are combined with the samples or genes found in D_r to produce a final enhanced model.

The final number of optimal PLS components is obtained by internal cross-validation (i.e. it is not equal to κ).

3. Methodology

To test the performance of *gesearch*, experiments were conducted on an existing large, heterogeneous infrared (IR) spectral library of soils from the United States and on an existing target set from Tshuapa Province, located in the Democratic Republic of the Congo. The objective was to obtain simple, accurate models of total carbon (TC) for the target set. The IR measurements and routine laboratory analyses of TC were performed independently for the source and target datasets (different instruments, laboratories, acquisition dates, operators, and geographic domains). For both datasets, the TC was measured via high-temperature dry combustion.

This experiment provides an opportunity to evaluate the algorithms in a scenario in which a spectral library from one continent is used to construct predictive models for an entirely independent set of samples originating from a different continent. This represents a challenging case of cross-domain model transfer, involving substantial differences in soil types, sample handling, instrumentation, and environmental conditions.

3.1. Datasets

3.1.1. Target domain: Tshuapa province, democratic Republic of the Congo

The dataset from the target domain comprises 738 soil samples from the Tshuapa Province (Democratic Republic of the Congo) spanning forest and agricultural sites. This dataset is hereafter referred to as the *Congo dataset*. These samples were collected during three independent field campaigns [16–18]. All IR spectral measurements were conducted by [19] over the range from 7500 cm^{-1} to 600 cm^{-1} , at a resolution of 2 cm^{-1} .

Outliers were screened following the procedure described by De Maesschalck et al. [20]. Pre-processed spectra were projected onto the principal component (PC) space, and components explaining less than 5% of the original spectral variance were discarded. In this way, only the first five PCs were retained, together accounting for 84% of the total variance. Samples with a Mahalanobis distance from the PC centroid greater than 3 were flagged as potential outliers and removed. This procedure led to the removal of just one sample.

The TC distribution for the target set (Table 1) is relatively narrow: mean 1.26% (median 1.16%), standard deviation 0.58%, interquartile range 0.71%, and observed range 0.37–4.74%. The skewness of 1.30 indicates the predominance of soils with low to moderate carbon contents and a limited number of higher-carbon samples. Overall, this distribution suggests relatively homogeneous soils with restricted variability in TC within the target domain.

Table 1

Distribution statistics for total carbon (%) in the Congo target set and the North American soil spectral library.

Dataset	n	mean	median	sd	IQR	min	max	skew
Congo	737	1.26	1.16	0.58	0.71	0.37	4.74	1.30
North America	60,272	5.24	1.46	10.85	2.98	0.05	59.97	3.14

3.1.2. Reference spectral library: North America

This publicly available dataset, compiled by the USDA-NRCS-NSSC Kellogg Soil Survey Laboratory (Lincoln, NE, USA) and available through the Open Soil Spectral Library (OSSL-KSSL v1; [21]), contains more than 68 000 samples predominantly from the United States. Records include IR spectra, laboratory measurements, site metadata, and quality-control metrics. To harmonise sources, entries were retained only when spectrum–response IDs matched, duplicates were removed, incomplete spectra or anomalously high absorbances were excluded, TC values reported as exactly zero were discarded (assumed below detection or unmeasured), and records outside the Americas were excluded. The resulting “North American IR spectral library” comprises 60 272 samples (all with TC), largely from North America (only 78 from Central/South America: about 60 from Costa Rica, 10 from Suriname, and 1 from Chile), with spectra over 4000–600 cm^{-1} at 2 cm^{-1} resolution.

No outlier detection was applied. Given the size of the library and its diversity, coverage of the target domain is expected to be adequate, and the methods used are robust to occasional outliers.

As expected, the North American library shows a much broader TC distribution than the Congo set, with a mean of 5.24% and a median of 1.46% (strong positive skew), spanning 0.05%–60% (Table 1). This heavy-tailed heterogeneity underscores the difficulty of directly transferring calibration models between these two domains.

3.2. Spectral data

Spectra were trimmed to 4000–605 cm^{-1} and resampled to a constant resolution of 8 cm^{-1} , returning a total of 425 spectral variables (Fig. 2). This resolution provided a balance between preserving chemically relevant information for TC prediction (see Deiss et al. [22]) and maintaining computational tractability during iterative evolutionary optimisation.

To ensure that similarity and reconstruction measures used in the evolutionary weakness functions were not dominated by baseline shifts or multiplicative artefacts, spectra were preprocessed using a second-order Savitzky–Golay derivative [23] followed by standard normal variate transformation [24]. The derivative employed a window length of 25 spectral variables (corresponding to 200 cm^{-1}) and a polynomial order of 2. This preprocessing reduced additive and multiplicative scatter effects, thereby improving the stability and interpretability of the subset optimisation process.

3.3. Optimised subset selection via evolutionary search using *gesearch*

From the Congo dataset, a small reference subset of $m = 20$ samples was randomly selected, and their TC values were retained to guide the evolutionary search. The choice of $m = 20$ reflects a realistic operational scenario in which a limited number of labelled target-domain samples can be acquired without substantial cost or time. The remaining 717 samples were reserved for independent validation with respect to the response variable. Spectra from all 738 target samples informed the reconstruction and similarity weakness scores. Accordingly, the response weakness was computed only on the $m = 20$ labelled samples, which were the sole contributors to the supervised search. The response values of the 717 held-out samples were withheld throughout the search, hyperparameter optimisation, and model-fitting procedure, and were accessed only during the final evaluation step.

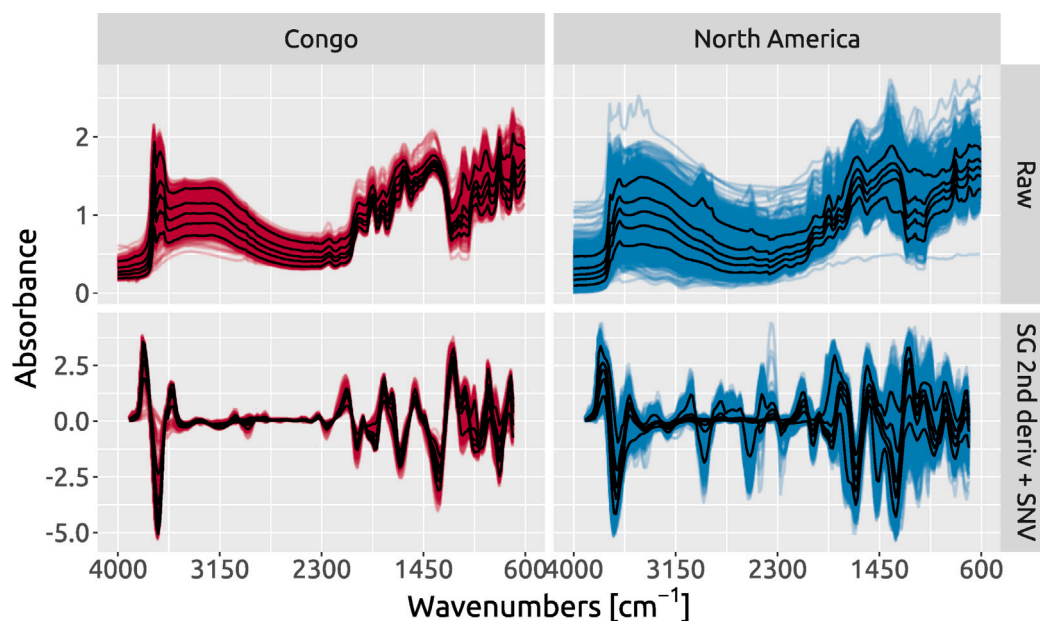


Fig. 2. Spectra of the Congo dataset and the North American soil spectral library, shown before and after pre-processing (SG 2nd deriv + SNV). Black lines represent the 5th, 25th, 50th, 75th, and 95th percentile spectra.

Consequently, the independence of the validation set was preserved with respect to the response variable. At the same time, *gesearch* is intentionally designed to exploit unlabelled target-domain spectra to characterise the target distribution, which is a practical advantage in settings where spectra are easy to acquire but reference measurements are costly. This design is consistent with transductive and semi-supervised learning settings, where the marginal distribution of unlabelled target data is routinely exploited to improve model generalisation without compromising the independence of held-out response values [25]. Furthermore, this reflects a realistic operational assumption in DRS applications, where spectra are considerably less expensive to acquire than laboratory reference measurements, making it practical and desirable to collect target-domain spectra without reference values to support model development.

Prefiltering of the reference spectral library: To reduce the computational cost of the analyses, a nearest-neighbour prefiltering step was applied to the North American library. Moving-window Pearson correlation dissimilarity ($\tau = 101$ data points; where τ is the window size) was computed between each sample in the Congo dataset and all samples in the North American dataset. The underlying assumption was that reference samples with weak spectral correlation to the target set were unlikely to be informative for modelling the target domain and could therefore be excluded as an initial screening step. Reference samples with a correlation dissimilarity above 0.5 were discarded. This reduced the North American library to 13 078 samples, which then served as the basis for the *gesearch* analyses.

Hyperparameter optimisation: *gesearch* was applied to D_t to select a final subset of θ samples most relevant to the target set. A grid search varied (k, b, α, θ) with $k \in \{50, 75, 100\}$, $b \in \{50, 75, 100\}$, $\alpha \in \{0.90, 0.95, 0.97, 0.99\}$, and $\theta \in \{120, 240\}$. During the search, the number of PLS latent factors was fixed at $\kappa = 20$. The final number of latent variables was re-optimised afterwards. For hyperparameter tuning, the $m = 20$ labelled target samples were withheld (the search used only reconstruction and similarity weakness scores). For each hyperparameter combination, a final PLS model was fitted using only the subset selected from the reference library, and its predictive performance was evaluated on the fixed set of $m = 20$ labelled target samples. The optimal combination was chosen as the one that minimised the root mean squared error (RMSE) on this small external validation set.

Validation and final evaluation: The optimal hyperparameter configuration was used to run 100 realisations of the *gesearch* algorithm. Each realisation produced a different subset of θ samples from the North American library. Model quality was assessed by analysing variability across realisations and by evaluating the stability of sample selection (i.e., which samples were chosen and how consistently across runs), as well as predictive performance, regression coefficients, and variable importance profiles (estimated using the selectivity ratio method; see [26]).

3.4. Benchmark models and comparisons

Six established modelling approaches were included to contextualise the results of evolutionary subset selection relative to commonly applied global, local, and rule-based strategies. The intent was not to argue that the proposed strategy consistently outperforms all competing methods. The benchmarked approaches represent modelling frameworks, whereas *gesearch* is a search-driven subset optimisation strategy that improves calibration through sample selection rather than through a new model architecture. The comparisons just provide a view on whether optimisation in sample space provides competitive predictive performance under cross-domain conditions.

PLS: The simplest benchmark was a global PLS regression developed exclusively with the 20 labelled samples from the Congo dataset. These were the same samples used to guide the *gesearch* procedure. This model represents the limiting case in which only a very small, domain-specific calibration set is available and no exploitation of existing libraries is attempted. Due to the very small sample size, the number of PLS components was optimised using leave-one-out cross-validation. A maximum of 15 components was set deliberately to ensure that the tuning range extended beyond the expected optimum, thereby allowing assessment of whether the cross-validation strategy detected any overfitting pattern at higher numbers of components.

PLS [Congo + North America]: This is a global PLS regression developed with the North American library after prefiltering using the moving-window correlation dissimilarity. The training data for this approach consisted of the same 13 078 prefiltered samples employed in the *gesearch* analyses. This configuration represents the absence of evolutionary subset optimisation, relying instead on direct calibration

using the full prefiltered library. To align the model as closely as possible with the target domain, the subset of 20 labelled Congo samples was withheld during fitting and used only for component selection. For each candidate number of latent variables, a PLS model was fitted on the prefiltered North American samples and then applied to the 20 Congo samples. The optimal number of components was chosen as the one that minimised the RMSE on this small labelled subset of the target domain. After selecting the optimal number of components, the final PLS model was refitted using the prefiltered North American samples together with the 20 labelled Congo samples.

rs-local [Lobsey et al. 2017]: This is the original algorithm proposed by [4], which provided the main inspiration for the *gesearch* algorithm. Because rs-local relies entirely on labelled data from the target domain, hyperparameter optimisation is more difficult. To ensure comparability, the same optimal hyperparameters identified for *gesearch* were applied here, with the sample removal rate per iteration fixed at 90%. As for *gesearch*, this method was run for 100 realisations to assess the stability of the results.

LOCAL [adjusted]: The LOCAL algorithm [9] is a memory-based learning method that has been widely and successfully applied in DRS spectroscopy [27]. LOCAL fits local PLS models for each query sample based on its nearest neighbours in the reference library. In this study, two modifications were introduced to ensure a better target-domain alignment: (i) the correlation dissimilarity used in LOCAL was replaced by the moving-window Pearson correlation dissimilarity with $\tau = 101$; (ii) the 20 labelled Congo samples were forced to be included in each of the neighbourhoods of the 717 unlabelled Congo samples, ensuring that each local model was at least partially informed by target-domain data. The number of neighbours (k) and the range of PLS components were optimised, with k varying from 50 to 500 in increments of 30 and the number of PLS components ranging from 2 to 30. A nearest-neighbour cross-validation approach [10] was used for hyperparameter optimisation. Note that in LOCAL, predictions are obtained as an ensemble across multiple PLS component counts, where higher component numbers are implicitly penalised by down-weighting models with poor reconstruction quality or unstable regression coefficients. This regularisation typically permits the use of a larger maximum number of components without degrading predictive performance.

Cubist: Cubist [28] is a rule-based regression tree algorithm that extends model trees with linear regressions at their terminal nodes. In addition to this global rule-based structure, Cubist employs a memory-based learning (MBL) component during prediction, where the model retrieves the k -nearest neighbours of a query case from the training set and uses them to adjust the rule-based regression estimate. This method has been successfully applied in various DRS applications, gaining notable popularity [29,30]. A grid search was conducted to optimise Cubist model hyperparameters, specifically the number of committees and neighbours. Committee sizes were varied from 1 to 100 in 40 evenly spaced steps, and neighbour values ranged from 0 to 9, resulting in 400 committee-neighbour combinations. For each configuration, a Cubist model was trained on the North American library after prefiltering with the moving-window correlation dissimilarity. The same 13 078 prefiltered samples used in the *gesearch* experiments were employed here. Model complexity was capped at a maximum of 1000 rules. Predictions were then generated for the 20 labelled samples from the Congo dataset. The optimal hyperparameter combination was identified as the one that minimised the RMSE of predictions on these 20 labelled samples. The final model was then retrained using the optimal hyperparameters on the entire prefiltered North American library combined with the 20 labelled Congo samples. For the final evaluation, the optimally configured Cubist model was applied to the 717 unlabelled samples from the Congo dataset.

CNN [North America, Congo adjusted]: A one-dimensional convolutional neural network (CNN) was trained on the North American library and later fine-tuned with 20 labelled Congo samples, representing transfer learning from a global to a target domain. The architecture comprised three convolutional blocks (Conv1D–BatchNorm–ReLU–

MaxPool with 32, 64, and 128 channels, kernel sizes 5, 5, and 3, respectively) followed by two fully connected layers. This configuration follows the standard 1D CNN design widely used in spectroscopy (similar to [31], which used stacked Conv1D–ReLU–pooling blocks for modelling soil spectra, and to [32], which applied comparable multi-layer 1D CNNs for fruit-quality prediction). Training employed AdamW optimisation with Smooth L1 loss, with β set to 0.5 of the interquartile range of TC. Hyperparameters were selected via a basic grid search over learning rates 10^{-3} , 3×10^{-4} , 10^{-4} , weight decays 10^{-3} , 5×10^{-4} , 10^{-4} , and batch sizes 128, 256. This restricted search was sufficient to ensure stable convergence but was intentionally limited, both due to time constraints and because the model was subsequently adapted on a very small labelled target set, which reduces the value of extensive optimisation on the source domain. During adaptation, the pre-trained convolutional layers were frozen, and only the linear layers were fine-tuned on the Congo samples, with early stopping based on validation RMSE. Final performance was assessed on the 717 held-out Congo samples.

3.5. External validation for modelling without target-domain reference values for adaptation

An additional evaluation was performed on the same external validation set under a setting in which no labelled Congo samples were available for guiding the models. This analysis was included to assess how the methods behave under a more restrictive, but practically relevant, deployment scenario in which target-domain spectra may be available whereas target-domain reference values are not. For *gesearch*, this means that the search can be regarded as unsupervised, since it relied exclusively on the reconstruction and similarity weakness scores. Although this setting is more challenging than the scenario with limited target-domain reference values for adaptation considered above, it provides a useful test of transferability in the complete absence of target-domain reference data.

All models were trained exclusively on the North American spectral library, without using target-domain reference values at any stage. Method-specific components, such as the memory-based adjustment in Cubist and the local neighbourhood construction in LOCAL, were therefore based solely on the North American training data, and hyperparameters were re-optimised under this constraint. For the PLS model and for Cubist, hyperparameters were optimised using 10-fold cross-validation. For LOCAL, hyperparameter optimisation was again performed using the nearest-neighbour cross-validation approach of Ramirez-Lopez et al. [10].

For the CNN, the convolutional backbone and regression head previously trained on the North American library were reused without further adaptation. Note that rs-local was not evaluated in this setting because it requires target-domain reference values to operate. The resulting models were then applied to the 717 Congo validation samples to assess predictive performance in the absence of labelled target-domain information.

3.6. Evaluation metrics

The predictive performance of all models was assessed using the 717 Congo validation samples. The following metrics of predictive accuracy were used: the coefficient of determination (R^2), the root-mean-squared error (RMSE), and the mean error (ME). Let $\{(y_u^v, \hat{y}_u^v)\}_{v=1}^{m_{\text{val}}}$ denote the observed and predicted response values for the m_{val} validation samples. These figures of merit were computed as

$$R^2 = 1 - \frac{\sum_{v=1}^{m_{\text{val}}} (y_u^v - \hat{y}_u^v)^2}{\sum_{v=1}^{m_{\text{val}}} (y_u^v - \bar{y}_u)^2}, \quad (19)$$

$$\text{RMSE} = \sqrt{\frac{1}{m_{\text{val}}} \sum_{v=1}^{m_{\text{val}}} (y_u^v - \hat{y}_u^v)^2}, \quad (20)$$

$$ME = \frac{1}{m_{\text{val}}} \sum_{\nu=1}^{m_{\text{val}}} (\bar{y}_u^{\nu} - y_u^{\nu}), \quad (21)$$

where $\bar{y}_u$ denotes the mean of the observed response values in the validation set. For *gesearch* and *rs-local*, these performance measures are reported as averages across the 100 realisations run for each method.

3.7. Implementation

All algorithms and experiments, except for the convolutional neural network, were implemented in R (v.4.5.1; [33]). The CNN was implemented in Python (v.3.11.9) using the PyTorch framework (v.2.5.1+cu121; [34]). To address the computational demands of the algorithms implemented in R, core routines involving intensive matrix multiplications and iterative loops were implemented in C++ using the Armadillo library through RcppArmadillo, which extends the R-C++ interface provided by Rcpp [35,36]. Spectral preprocessing was carried out with the *prospectr* package [37]. The LOCAL algorithm for benchmarking was implemented using the *resemble* package [38].

The datasets, spectral preprocessing, and benchmarking framework employed here are aligned with those used in a related study investigating a new local modelling strategy based on libraries of models (currently under review). This alignment was adopted to enable controlled comparison of distinct methodological approaches under identical experimental conditions.

4. Results and discussion

4.1. Performance with limited target-domain reference values for adaptation

4.1.1. Optimised subset selection via evolutionary search

With the optimal hyperparameters, *gesearch* ended in 16–22 generations across 100 realisations. Generation 1 evaluated 26156 candidate individuals. Thereafter, the pool shrank rapidly as weak genes/samples were silenced by the quantile threshold (α). Per realisation, about 10^5 individuals were evaluated across generations, providing extensive sampling of gene combinations and thus a broad exploration of the spectral library for identifying predictive and descriptive samples.

Hyperparameters: Grid search returned a clear optimum at $k = 50$, $b = 100$, $\alpha = 0.90$, $\theta = 120$ (RMSE = 0.080, $R^2 = 0.83$ on the 20 labelled samples). Other configurations with similar performance consistently shared $k = 50$ and $b = 100$, with the best results obtained at $\alpha = 0.90$ and with 10 PLS latent components selected.

Importantly, although the mean and the $(1-\alpha)$ -quantile of the weakness scores decreased monotonically across generations, they showed no clear evidence of stabilisation before the number of active genes reached the predefined target size (see Appendix B for an example of the evolution of the weakness-score cut-offs). This suggests that the probabilistic silencing rule may primarily drive progressive diversity reduction rather than convergence towards a stable low-weakness subset, thereby risking bias towards internally homogeneous selections that under-represent target-domain variability. The implications of this behaviour remain unclear and warrant further investigation.

Prediction performance: Table 2 summarises results on the Congo validation set, and predictive performance is discussed primarily in terms of RMSE. The best performance was delivered by *gesearch* ($R^2 = 0.843$, RMSE = 0.237%, ME = −0.015%). It also provided better results than those reported by Summerauer et al. [19] for the same region and dataset, where an RMSE of 0.319% was obtained using an African spectral library that included samples from neighbouring regions, implying less severe extrapolation than in the highly constrained scenario analysed here.

Table 2

Performance of *gesearch* and benchmark models on the 717 validation samples under two experimental settings: *with target-domain reference values for adaptation* (20 labelled Congo samples available) and *without target-domain reference values for adaptation* (no Congo reference samples available, although target-domain spectra may still be used when the method allows it). Benchmarks in the first setting include *rs-local*, LOCAL (adapted correlation dissimilarity; neighbourhoods forced to include target-domain samples when available), Cubist, PLS [Congo], PLS [Congo + North America], and a convolutional neural network (CNN; trained on North American samples and fine-tuned with 20 Congo samples). In the second setting, *rs-local* is not reported because it requires target-domain reference values.

Model	R^2	RMSE (%)	ME (%)
<i>With limited target-domain reference values for adaptation</i>			
<i>gesearch</i>	0.843	0.237	−0.015
Benchmarks:			
rs-local (Lobsey et al. 2017)	0.824	0.264	0.020
CNN [North America, Congo adjusted]	0.774	0.278	0.008
LOCAL (adapted)	0.833	0.315	0.094
Cubist	0.776	0.374	−0.001
PLS [Congo]	0.670	0.403	0.062
PLS [Congo + North America]	0.760	0.677	0.392
<i>Without target-domain reference values for adaptation</i>			
<i>gesearch</i>	0.760	0.331	0.126
Benchmarks:			
CNN	0.469	0.628	0.295
LOCAL	0.799	0.423	0.091
Cubist	0.794	0.408	0.131
PLS	0.814	0.978	0.737

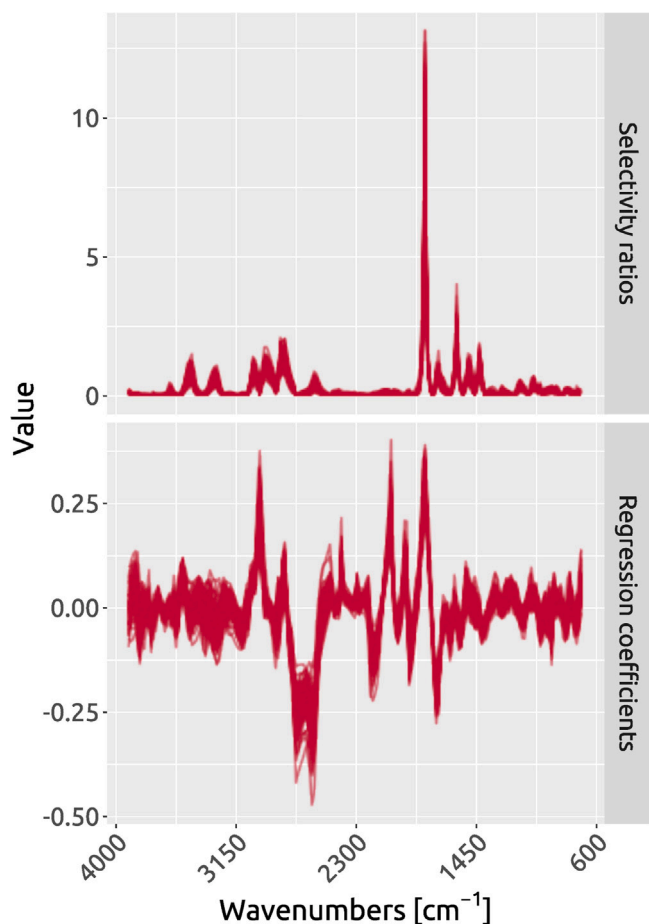
Among the benchmarks, *rs-local* produced lower performance than *gesearch*, suggesting that the evolutionary mechanisms introduced in *gesearch* produce tangible gains. The CNN (state-of-the-art deep learning approach for spectroscopy) performed reasonably and produced better results than the adapted LOCAL algorithm. In contrast, the global PLS models performed substantially worse, highlighting the limitations of relying on very small domain-specific training sets. For the global PLS model built with the 20 Congo samples, cross-validation selected 8 latent variables as optimal. For the global PLS model calibrated on the combined Congo and North American datasets, the optimal number of latent variables was 5. Interestingly, despite showing the highest prediction error (in the scenario with limited target-domain reference values for adaptation), this combined-dataset PLS model still achieved a relatively good R^2 , indicating that it preserved a substantial proportion of the variance in the target data, even though its absolute predictive accuracy was poor. See Appendix C for detailed validation plots.

For the datasets used here, the uncertainty of the laboratory reference measurements was not reported. However, Stevens et al. [39], working with mineral soils and the same reference method (high-temperature dry combustion), reported reproducibility errors between approximately 0.08% and 0.29%, increasing with carbon content. In this context, the RMSE values obtained with *gesearch* (0.237%), *rs-local* (0.264%), and CNN (0.278%) fall within the range of reproducibility errors reported for the reference method itself, which may be taken as an indication that these models achieved a reasonable degree of predictive accuracy for the conditions of this study.

Stability: A first aspect of stability is the recurrence of selected genes across realisations of the evolutionary search. Recurrently selected genes reflect stable selection behaviour, whereas infrequently selected genes are likely marginal with respect to the applied weakness criteria. Table 3 summarises recurrence across 100 independent realisations (with the optimal hyperparameters) using the filtered North American library (13 078 samples). Only 2492 samples were selected at least once, and 188 appeared at least 10 times, indicating stable algorithmic behaviour in the identification of a core subset of relevant genes.

Table 3Recurrence of the samples selected by *gesearch* across the 100 realisations.

t	Minimum frequency ($\geq t$)									
	1	2	3	4	5	6	7	8	9	10
Sample count	2492	1732	1240	897	654	476	351	269	214	188

**Fig. 3.** Variable importance (top, selectivity ratios) and regression coefficients (bottom) for 100 *gesearch* models.

A second aspect of stability is the dispersion of predictions across the 100 realisations. The median per-sample standard deviation was 0.127%, indicating minimal variability across realisations. Despite the stochastic nature of *gesearch*, the consistency of its results suggests convergence to similar solutions. Hence, uncertainty is dominated by data and inherent modelling limits ($\text{RMSE} = 0.237\%$), rather than by instability in the selection mechanism.

A third indicator of stability is the high consistency of both variable importance (selectivity ratios) and regression coefficients across the 100 realisations (Fig. 3). The most prominent peaks, corresponding to regions of high variable importance, were repeatedly identified at 1808, 1584, 1416, 2808, 2920, 1496, 3016, 3464, 1712, and 3288 cm^{-1} across iterations. Fig. 3 also shows that regression coefficients were highly consistent, especially in the region between 3150 and 1500 cm^{-1} .

Emergent geographical pattern: An interesting observation is that the samples found by *gesearch* form a cluster in the eastern United States (see Fig. 4, left), despite geography not being used in the search. A plausible (yet speculative) explanation is shared environmental characteristics with the Congo study area. This interpretation is supported by the fact that the samples retrieved by *gesearch* are also spectrally

similar to those from Congo (see Fig. 4, right), which in turn suggests an underlying compositional similarity. These findings suggest that *gesearch* can discover patterns beyond the information explicitly supplied.

4.2. Model performance without target-domain reference values for adaptation

Model performance in the absence of target-domain reference values for adaptation is summarised in Table 2. In this setting, all fitted predictive models were derived from the North American library, while methods that permit it could still exploit unlabelled target-domain spectra. Consistent with the more challenging nature of this scenario, all modelling approaches showed reduced predictive performance relative to the case in which a small number of target-domain reference values were available for adaptation.

For *gesearch*, the optimal hyperparameters in this setting were $k = 75$, $b = 75$, $\alpha = 0.95$, and $\theta = 120$, with 4 PLS latent components in the final model. Under these conditions, *gesearch* achieved the lowest RMSE in this scenario (0.331%), while maintaining a competitive R^2 (0.760). This indicates that, even without any target-domain reference values, the unsupervised search based only on reconstruction and similarity weakness scores was still able to retrieve a subset of samples that supported reasonably accurate prediction.

Among the benchmark models, LOCAL achieved the highest R^2 (0.799), followed closely by Cubist (0.794), but both showed higher RMSE values than *gesearch* (0.423% and 0.408%, respectively). The CNN showed the largest deterioration in performance, with RMSE increasing from 0.278% in the adaptation setting to 0.628% without target-domain reference values. The global PLS model produced the poorest results overall, with the highest RMSE (0.978%) and the largest mean error (0.737%), indicating that a substantial systematic bias dominated the prediction error, even though the model still captured a considerable proportion of the variance.

For the PLS model, the optimal number of latent components in this setting was 20. For the remaining algorithms, the resulting optimal hyperparameters happened to coincide with those identified in the scenario with limited target-domain reference values.

A comparison with the adaptation setting shows that even a small number of labelled target-domain samples provided a clear benefit for all methods. For example, the RMSE of *gesearch* decreased from 0.331% to 0.237%, that of CNN from 0.628% to 0.278%, that of LOCAL from 0.423% to 0.315%, and that of Cubist from 0.408% to 0.374%. These results indicate that limited target-domain reference information can substantially improve calibration transfer, particularly for methods whose adaptation mechanisms depend more strongly on labelled target-domain data.

4.3. Final considerations

The *gesearch* algorithm delivered the best predictive performance on the Congo validation set (see Table 2). This aligns with the expectation that exploiting spectral libraries through targeted selection improves transfer to new domains.

The relative contributions of soil composition and analytical procedures to systematic differences between the Congo and North American datasets cannot be disentangled. While both datasets used the same type of spectrometer and elemental TC analysis, differences in sample preparation, instrument calibration, reference materials, or analytical

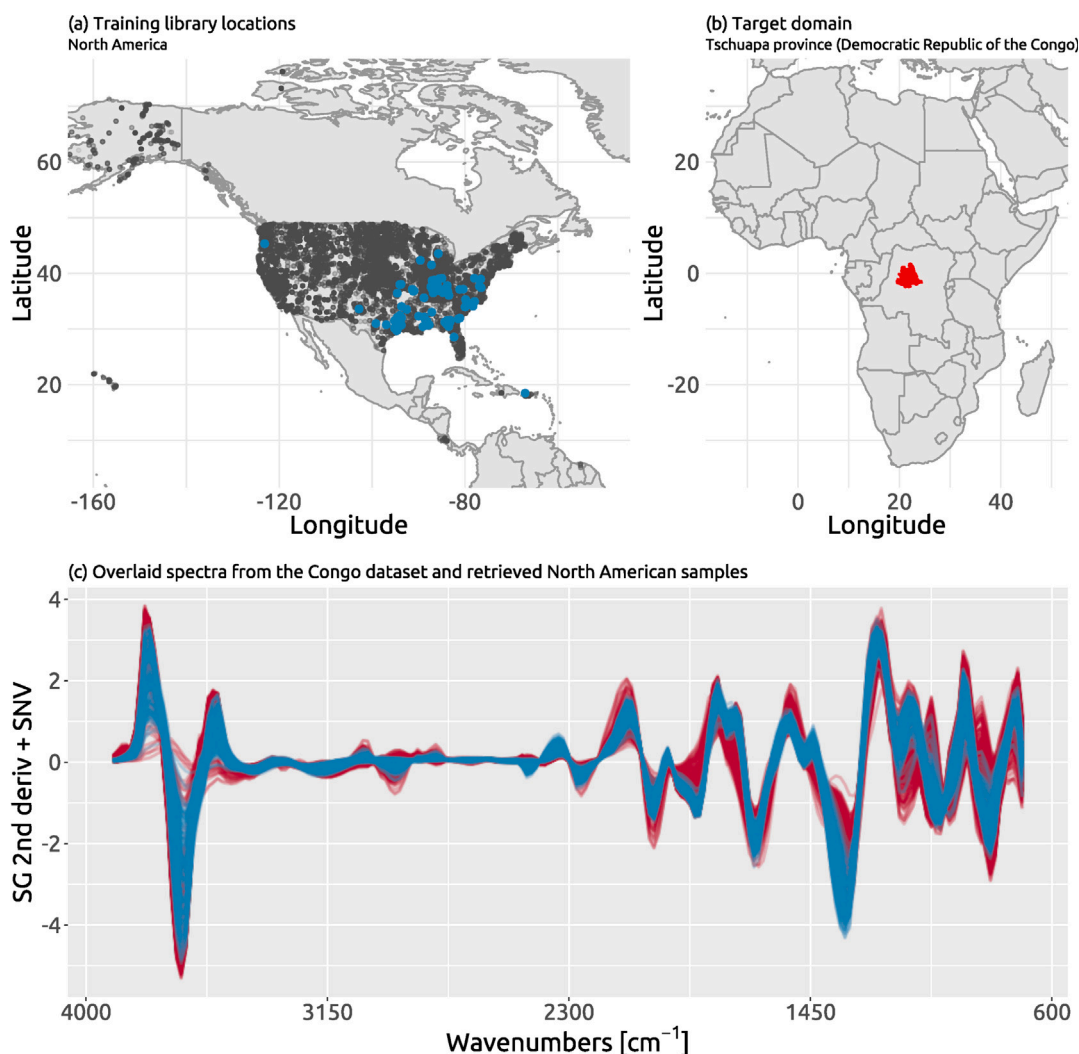


Fig. 4. Geographic distribution and spectra of samples identified by *gesearch* in the North American library across 100 realisations. (a): grey dots denote all North American library samples; blue dots mark samples repeatedly identified by *gesearch* (recurrence ≥ 10). (b): Tschuapa province (highlighted in red) in the Democratic Republic of the Congo, where the samples in the Congo dataset originate from. (c): red spectra show the Congo dataset; blue spectra show North American samples with recurrence ≥ 10 across the 100 realisations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

protocols may still introduce systematic effects, despite partial overlap in laboratory analyses. Both laboratories involved in the Congo and North American datasets participated in an interlaboratory ring test [40]. In this ring test the differences between these two laboratories were small. However, a complete dataset was not available to assess whether these effects were limited in magnitude. It could be that they are more pronounced but successfully compensated for by the *gesearch* algorithm. That is, *gesearch* assesses differences in the spectral distribution of the domains through the similarity weakness score, and partially captures aspects of the spectrum–response relationship through the supervised latent structure underlying the reconstruction and response weakness scores.

The *rs-local* method performed well and was stable across realisations, but it did not reveal clear geographical patterns like those observed with *gesearch* (see Appendix D). More importantly, *rs-local* provides no guarantee that the retrieved samples adequately reconstruct the target spectra or define a reliable latent projection for PLS, which can inflate predictive uncertainty. In contrast, *gesearch* emphasises samples that preserve both spectral structure and response behaviour in the target domain.

The CNN benchmark, despite using a standard architecture and a conservative hyperparameter search, outperformed most of the benchmarks (except for *rs-local*) when limited target-domain adaptation was possible. However, its performance deteriorated markedly when no target-domain reference values were available, indicating stronger dependence on supervised target-domain adjustment than *gesearch*. A more extensive architectural exploration could have produced a CNN variant capable of transferring more accurately to the Congo samples. However, the large number of hyperparameters makes such optimisation challenging. By contrast, *gesearch* provides straightforward ways to construct models that are simple, interpretable, and easy to update for a given target domain. From an industry perspective, this translates into lower deployment costs and reduced dependence on specialised machine-learning expertise, while still ensuring robust predictive performance.

From a deployment perspective, *gesearch* operates with compact artefacts, namely a selected calibration subset. This favours embedded, privacy-constrained, or service-oriented use cases. In practice, *gesearch* is well aligned with target domains that are spectrally and compositionally coherent, and where spectra–response relations are largely linear.

4.3.1. Assumptions and limitations of the search and predictive performance

Although *gesearch* uses a non-linear search process, it is primarily intended for domains where spectrum-response relationships are approximately linear. The predictive performance of *gesearch* is closely linked to the modelling assumptions used during the evaluation of individuals/subsets. In its present formulation, subset quality is assessed using linear regression models. Consequently, *gesearch* is expected to perform best in target domains where the relationship between spectra and response variables is approximately linear. Under such conditions, the evolutionary search can identify subsets of samples that are both spectrally coherent with the target domain and informative for estimating stable linear relationships.

A more flexible, non-linear method could in principle be used at the final model-fitting stage. Such an extension may help capture mild residual non-linearities within the subset selected by *gesearch*. However, this added flexibility does not alter the main assumption underlying the search itself, since the weakness scores that guide gene selection are still computed from linear subset evaluations.

Accordingly, when the underlying relationships between spectra and response variables are strongly non-linear and cannot be approximated locally by linear models, the performance of *gesearch* may deteriorate. In such cases, the weakness scores are computed under a misspecified model class, and samples that are informative only through non-linear effects may be incorrectly assigned high weakness and permanently silenced during the evolutionary process. A non-linear model fitted after the search may partially recover subtle departures from linearity, but it is unlikely to compensate fully for information that has already been excluded during sample selection.

These considerations suggest that the effectiveness of *gesearch* depends not only on the availability of relevant samples in the spectral library, but also on the extent to which target-domain relationships can be represented within a locally linear latent-variable framework.

Although the results obtained here without target-domain reference values were still reasonably good, this setting should be interpreted with caution. When no response data are available, the search is effectively blind to the spectrum-response relationship in the target domain, and performance may therefore depend strongly on the dataset and transfer conditions. In such cases, *gesearch* may be more appropriately viewed as a preliminary sample-filtering strategy rather than as a fully supervised sample-selection method. Further evaluation on additional datasets will be needed to determine how robust this behaviour is in broader applications.

5. Conclusions

This study introduced *gesearch*, an algorithm designed to exploit spectral libraries when target-domain data are scarce. In a demanding cross-continental setting involving a North American spectral library and Congo target samples, *gesearch* achieved the best overall predictive performance when a small number of labelled target-domain samples were available for adaptation, and it remained reasonably competitive when no target-domain reference values were available (Table 2). These findings suggest that *gesearch* can operate successfully with limited target-domain reference values for adaptation and may also provide a useful unsupervised sample-filtering strategy when only target-domain spectra are available.

Although the non-linear search process of *gesearch* is purely data-driven (i.e. it does not model the underlying spectral-chemical processes), it enhances the conditions under which mechanistic interpretation is possible. By selecting samples that are compositionally and spectrally coherent with the target domain, the final PLS model, which remains linear, can yield more physically meaningful loadings and coefficients.

This study demonstrates that *gesearch* models trained on a small subset (120 samples) retrieved from an existing publicly available

library and from a different domain can be highly accurate, and at the same time be parsimonious and interpretable. Two limitations emerged: (i) weakness scores tend to decrease monotonically and therefore without clear convergence, suggesting that quantile-based silencing may reduce diversity rather than stabilise to a minimum-weakness set; and (ii) rare or outlying samples may exert limited influence, risking under-coverage of the target domain. Two potential refinements could be: (i) adding diversity-preserving constraints (coverage penalties, redundancy discouragement) alongside convergence-aware stopping criteria; and (ii) incorporating non-linear learners (e.g. kernel-based PLS) into the weakness evaluation to reduce the linear bias of PLS-based scoring.

Overall, *gesearch* provides a reliable way to operationalise large spectral libraries for predictive modelling with minimal additional laboratory effort. By enabling the construction of parsimonious and interpretable calibration models, it reduces both the time and cost required to develop models for new target domains. Future work should focus on diversity-aware selection, calibrated uncertainty estimation, and integration with calibration workflows that support automated updating and adaptation as new data become available.

CRedit authorship contribution statement

Leonardo Ramirez-Lopez: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Raphael Viscarra Rossel:** Writing – review & editing, Software, Conceptualization. **Claudio Orellano:** Writing – review & editing, Validation, Software, Conceptualization. **Laurens Kooijman:** Conceptualization. **Estefania Perez-Fernandez:** Writing – review & editing, Validation, Conceptualization. **Alexandre M.J.-C. Wadoux:** Writing – review & editing. **Marçal Plans:** Writing – review & editing, Validation. **Timo Breure:** Writing – review & editing. **Laura Summerauer:** Writing – review & editing, Resources, Data curation. **José L. Safanelli:** Writing – review & editing. **Thorsten Behrens:** Writing – review & editing.

Declaration of competing 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.

Acknowledgements

The authors thank Dr Xiaocheng Li for supervising the first author's MSc research report at Imperial College London, Imperial College Business School, and for providing valuable feedback on early conceptual ideas that helped shape this work. The authors also thank the Soil Resources Group at ETH Zurich, Institute of Terrestrial Ecosystems (ITES), for funding the data collection in Congo as part of its own research activities and for kindly making the data available for use in this study. The authors also thank Dr Julieta Goenaga for her thoughtful and constructive criticism of the genetic analogy used in earlier versions of the algorithm; her feedback prompted a critical reassessment of this aspect of the work and contributed to improving both the conceptual coherence of the manuscript and the consistency of the algorithm with genetic theory. The USDA NRCS National Soil Survey Center and the Kellogg Soil Survey Laboratory (NSSC-KSSL) are gratefully acknowledged for curating and maintaining the North American spectral library employed in this study; the SoilSpec4GG initiative is also acknowledged for facilitating access to this resource. Finally, the first author thanks Galo and Pablo for their patience and for the joy they brought, which helped keep his weakness scores low.

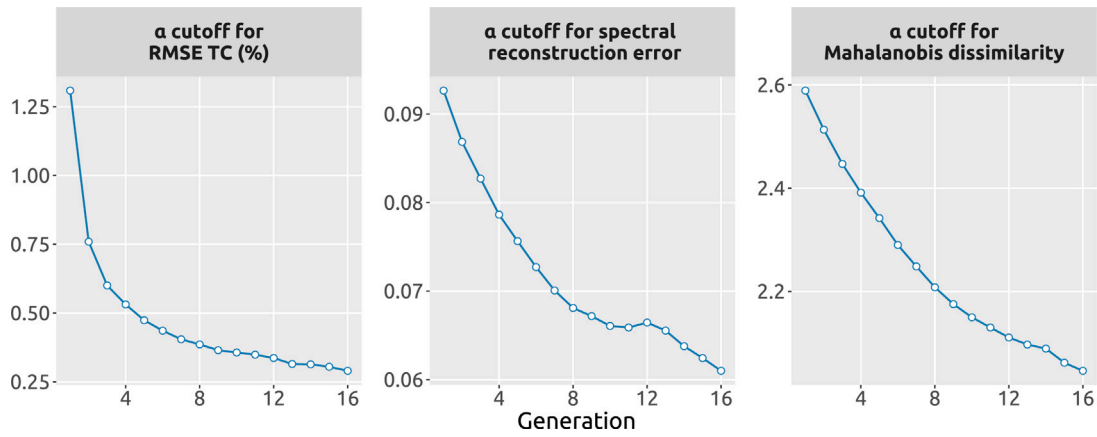


Fig. B.5. Example of the progress of the α cutoff quantiles applied to weakness scores for gene/sample silencing across generations in a single realisation of gene research.

Appendix A. Computation of the weakness scores in gene research

A.1. Response weakness score

Let γ denote the generation index. Define the response weakness of gene g_j at generation γ as $w_r^{(\gamma)}(g_j)$. It depends on the per-individual error $e(S_i^{(\gamma)})$ and on the incidence count $h_j^{(\gamma)}$ (defined in Eqs. (7) and (8)), the number of individuals in generation γ that contain g_j :

$$e(S_i^{(\gamma)}) = \sqrt{\frac{1}{m} \sum_{\ell=1}^m (y_u^\ell - \hat{f}_i^{(\gamma)}(x_u^\ell))^2}, \quad (\text{A.1})$$

Here y_u^ℓ is the reference value of the ℓ th target sample and $\hat{f}_i^{(\gamma)}(x_u^\ell)$ the prediction from the model calibrated on $S_i^{(\gamma)}$. Models $\hat{f}_i^{(\gamma)}$ are fitted by PLS [15] on the k samples expressed in $S_i^{(\gamma)}$, using a globally fixed number κ of latent components.

The response weakness is then

$$w_r^{(\gamma)}(g_j) = \frac{1}{h_j^{(\gamma)}} \sum_{i: g_j \in S_i^{(\gamma)}} e(S_i^{(\gamma)}), \quad (\text{A.2})$$

i.e. the average target-RMSE of the individuals in which g_j is active, scaled by $h_j^{(\gamma)}$ to ensure comparability across genes.

A.2. Reconstruction weakness score

For individual $S_i^{(\gamma)}$, fit PLS to $(\mathbf{X}_i^{(\gamma)}, \mathbf{y}_i^{(\gamma)})$ to obtain scores $\mathbf{T}_{i,\kappa}^{(\gamma)} \in \mathbb{R}^{k \times \kappa}$, X -loadings $\mathbf{P}_{i,\kappa}^{(\gamma)} \in \mathbb{R}^{d \times \kappa}$, weights $\mathbf{W}_{i,\kappa}^{(\gamma)} \in \mathbb{R}^{d \times \kappa}$, and residuals $\mathbf{e}_i^{(\gamma)}$, with $a \leq \min k, d$ (cf. [15]). As usual,

$$\mathbf{T}_i^{(\gamma)} = \mathbf{X}_i^{(\gamma)} \mathbf{W}_i^{(\gamma)} ((\mathbf{P}_i^{(\gamma)})^\top \mathbf{W}_i^{(\gamma)})^{-1},$$

using a pseudoinverse if necessary. Let $\kappa \in 1, \dots, a$ be the global number of retained latent factors. With $\mathbf{P}_{i,\kappa}^{(\gamma)}$ and $\mathbf{W}_{i,\kappa}^{(\gamma)}$ the first κ columns of $\mathbf{P}_i^{(\gamma)}$ and $\mathbf{W}_i^{(\gamma)}$, project and reconstruct the target spectra $\mathbf{X}_u \in \mathbb{R}^{m \times d}$:

$$\mathbf{Z}_{u,\kappa}^{(i,\gamma)} = \mathbf{X}_u \mathbf{W}_{i,\kappa}^{(\gamma)} ((\mathbf{P}_{i,\kappa}^{(\gamma)})^\top \mathbf{W}_{i,\kappa}^{(\gamma)})^{-1} \in \mathbb{R}^{m \times \kappa}, \quad \tilde{\mathbf{X}}_{u,\kappa}^{(i,\gamma)} = \mathbf{Z}_{u,\kappa}^{(i,\gamma)} (\mathbf{P}_{i,\kappa}^{(\gamma)})^\top.$$

Define the squared-Frobenius reconstruction error

$$q_i^{(\gamma)}(\kappa) = \|\mathbf{X}_u - \tilde{\mathbf{X}}_{u,\kappa}^{(i,\gamma)}\|_F^2.$$

The reconstruction weakness of active gene g_j at generation γ is

$$w_q^{(\gamma)}(g_j) = \frac{1}{h_j^{(\gamma)}} \sum_{i: g_j \in S_i^{(\gamma)}} q_i^{(\gamma)}(\kappa),$$

Normalisation by $h_j^{(\gamma)}$ makes scores comparable across genes irrespective of sampling frequency. Using a fixed κ focuses on the most

important components: individuals composed of target-aligned genes reconstruct $\mathbf{X}_u$ well (small $q_i^{(\gamma)}(\kappa)$), whereas misaligned genes yield larger errors.

A.3. Similarity weakness score

For each individual $S_i^{(\gamma)}$, fit PLS on $(\mathbf{X}_i^{(\gamma)}, \mathbf{y}_i^{(\gamma)})$ and retain a globally fixed number κ of components. Let $\mathbf{T}_{i,\kappa}^{(\gamma)} \in \mathbb{R}^{k \times \kappa}$ be its score matrix and project the target spectra $\mathbf{X}_u$ into the same score space to obtain $\mathbf{Z}_{u,\kappa}^{(i,\gamma)} \in \mathbb{R}^{m \times \kappa}$. Define the centroids

$$\bar{\mathbf{T}}_{i,\kappa}^{(\gamma)} = \frac{1}{k} \sum_{r=1}^k \mathbf{T}_{i,\kappa}^{(\gamma)}[r, :], \quad \bar{\mathbf{Z}}_{u,\kappa}^{(i,\gamma)} = \frac{1}{m} \sum_{\ell=1}^m \mathbf{Z}_{u,\kappa}^{(i,\gamma)}[\ell, :],$$

Let $\Sigma_{i,\kappa}^{(\gamma)}$ denote the empirical covariance of the rows of $\mathbf{T}_{i,\kappa}^{(\gamma)}$. For numerical stability, one may use a ridge-regularised version $(\Sigma_{i,\kappa}^{(\gamma)} + \lambda \mathbf{I})$ with $\lambda \geq 0$. The individual-level dissimilarity is the Mahalanobis distance between the two centroids in the PLS score space:

$$D_i^{(\gamma)} = (\bar{\mathbf{Z}}_{u,\kappa}^{(i,\gamma)} - \bar{\mathbf{T}}_{i,\kappa}^{(\gamma)})^\top (\Sigma_{i,\kappa}^{(\gamma)} + \lambda \mathbf{I})^{-1} (\bar{\mathbf{Z}}_{u,\kappa}^{(i,\gamma)} - \bar{\mathbf{T}}_{i,\kappa}^{(\gamma)}).$$

Aggregate to the gene level using the incidence-normalised average. Using the previously defined incidence count $h_j^{(\gamma)}$, the similarity weakness is

$$w_d^{(\gamma)}(g_j) = \frac{1}{h_j^{(\gamma)}} \sum_{i: g_j \in S_i^{(\gamma)}} D_i^{(\gamma)}.$$

Here κ is fixed globally (hence suppressed in $D_i^{(\gamma)}$ and $w_d^{(\gamma)}(g_j)$ for compactness). Lower $w_d^{(\gamma)}(g_j)$ indicates genes associated with individuals whose latent spaces are more closely aligned to the centre of the target domain.

Appendix B. Weakness cutoffs example

See Fig. B.5.

Appendix C. Validation plots

See Fig. C.6.

Appendix D. Samples selected by rs-local

See Fig. D.7.

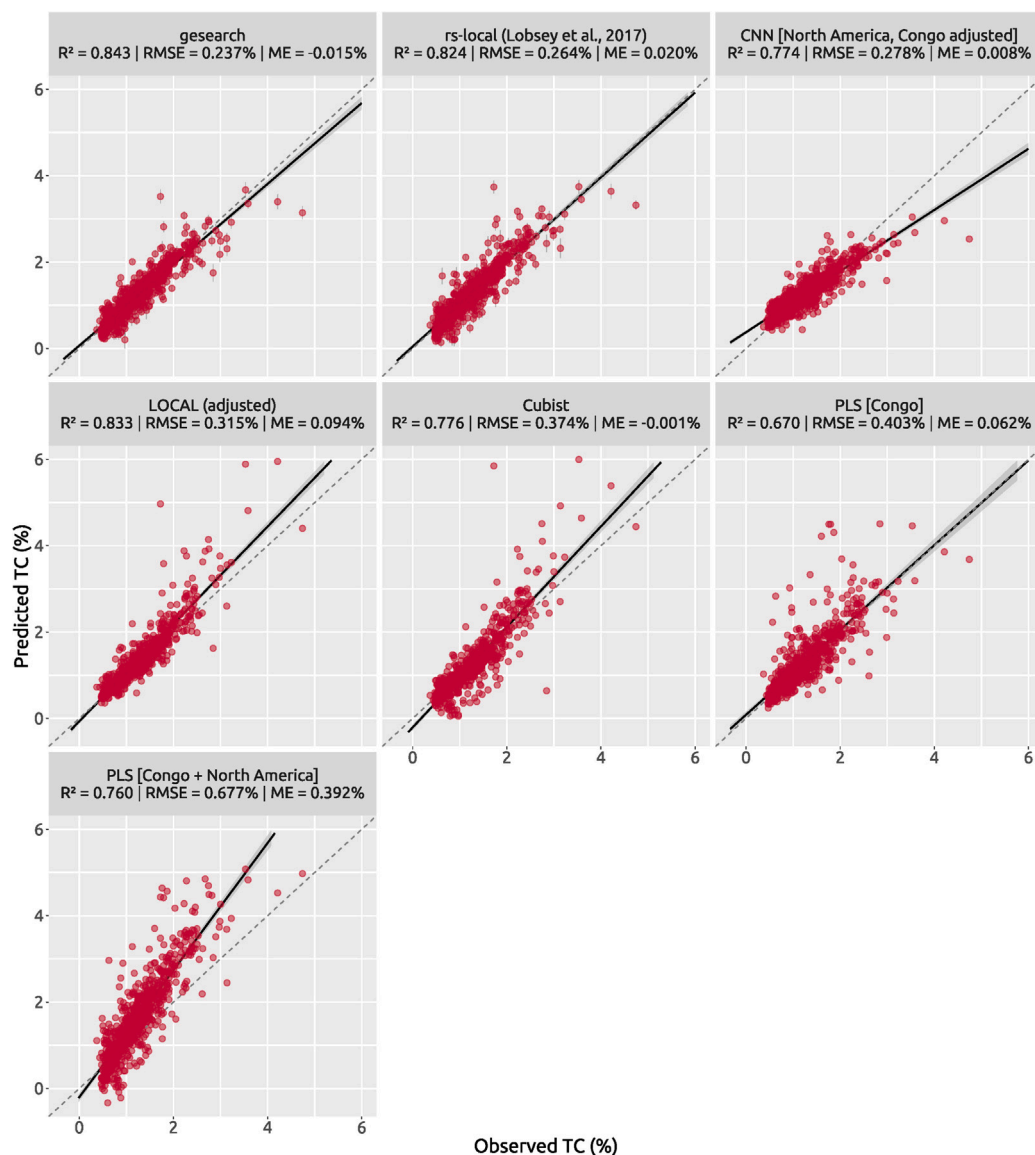


Fig. C.6. Validation plots for gesearch and benchmark models under the scenario where limited labelled samples for the target domain are available for model adaptation. For gesearch and rs-local, predictions correspond to the average across 100 realisations. Shaded bands indicate the standard deviation of the predictions across realisations; these bands are almost indistinguishable from the mean prediction curves, indicating very limited variability across realisations.

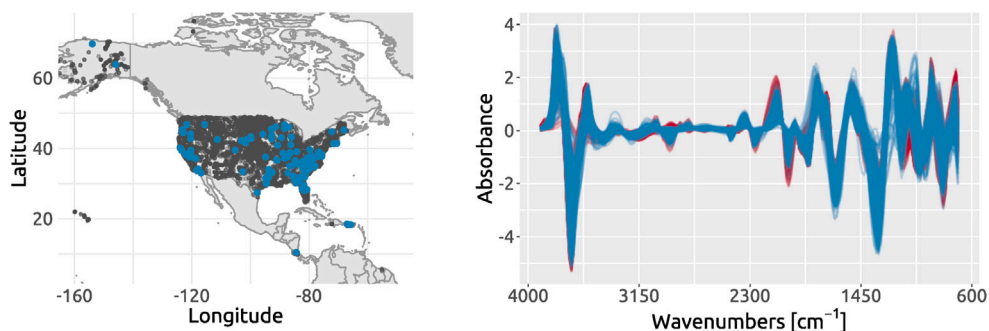


Fig. D.7. Geographic distribution and spectra of samples identified by rs-local in the North American library across 100 realisations. Left: grey dots denote all North American library samples; blue dots mark samples repeatedly identified by rs-local (recurrence ≥ 10). Right: red curves show the Congo dataset; blue curves show North American samples with recurrence ≥ 10 across the 100 realisations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Data availability

The open-source implementation of the **gesearch** algorithm is publicly available, together with detailed documentation, through the **resemble** package for the R programming environment. The stable release is available from CRAN at <https://CRAN.R-project.org/package=resemble>, and the development version is available at <https://github.com/l-ramirez-lopez/resemble>. Additional documentation and usage examples are provided at <https://l-ramirez-lopez.github.io/resemble/>.

References

- [1] C. Carletto, Better data, higher impact: improving agricultural data systems for societal change, *Eur. Rev. Agric. Econ.* 48 (2021) 719–740.
- [2] A. Fischer, G. Cameron, C. Tilus, J. Espey, S. Badiie, Revisiting the assumptions of the data revolution as an accelerator of the sustainable development goals, *Data Policy* 7 (2025) e49.
- [3] P. Williams, K. Norris (Eds.), *Near-Infrared Technology in the Agricultural and Food Industries*, American Association of Cereal Chemists, St. Paul, MN, 2001.
- [4] C. Lobsey, R.A. Viscarra Rossel, P. Roudier, C. Hedley, Rs-local data-mines information from spectral libraries to improve local calibrations, *Eur. J. Soil Sci.* 68 (2017) 840–852.
- [5] L. Ramirez-Lopez, K. Schmidt, T. Behrens, B. Van Wesemael, J.A. Demattê, T. Scholten, Sampling optimal calibration sets in soil infrared spectroscopy, *Geoderma* 226 (2014) 140–150.
- [6] P. Mishra, D. Passos, Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios, *Chemometr. Intell. Lab. Syst.* 212 (2021) 104283.
- [7] R.A. Viscarra Rossel, Z. Shen, L.R. Lopez, T. Behrens, Z. Shi, J. Wetterlind, K.A. Sudduth, B. Stenberg, C. Guerrero, A. Gholizadeh, et al., An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning, *Earth-Sci. Rev.* 254 (2024) 104797.
- [8] T. Naes, T. Isaksson, B. Kowalski, Locally weighted regression and scatter correction for near-infrared reflectance data, *Anal. Chem.* 62 (1990) 664–673.
- [9] J.S. Shenk, M.O. Westerhaus, P. Berzaghi, Investigation of a local calibration procedure for near infrared instruments, *J. Near Infrared Spectrosc.* 5 (1997) 223–232.
- [10] L. Ramirez-Lopez, T. Behrens, K. Schmidt, A. Stevens, J.A.M. Demattê, T. Scholten, The spectrum-based learner: A new local approach for modeling soil vis–nir spectra of complex datasets, *Geoderma* 195 (2013) 268–279.
- [11] M. Lesnoff, M. Metz, J.M. Roger, Comparison of locally weighted pls strategies for regression and discrimination on agronomic nir data, *J. Chemom.* 34 (2020) e3209.
- [12] D. Tanzilli, L. Strani, M. Metz, J.M. Roger, M. Lesnoff, C. Ruckebusch, M. Cocchi, R. Vitale, Locally-weighted-roboost-pls: a multivariate calibration approach to simultaneously cope with non-linearities and outliers, *Anal. Chim. Acta* (2025) 344167.
- [13] A. Helfenstein, P. Baumann, R. Viscarra Rossel, A. Gubler, S. Oechslin, J. Six, Quantifying soil carbon in temperate peatlands using a mid-ir soil spectral library, *Soil* 7 (2021) 193–215.
- [14] R. Spiers, J.H. Kalivas, Local adaptive fusion regression (lafr) for local linear multivariate calibration: Application to large datasets, *Appl. Spectrosc.* 79 (2025) 797–807.
- [15] S. Wold, M. Sjöström, L. Eriksson, Pls-regression: a basic tool of chemometrics, *Chemometr. Intell. Lab. Syst.* 58 (2001) 109–130.
- [16] M. Bateurs, O. Vercleyen, B. Vanlauwe, J. Six, B. Bonyoma, H. Badjoko, W. Hubau, A. Hoyt, M. Boudin, H. Verbeeck, et al., Long-term recovery of the functional community assembly and carbon pools in an african tropical forest succession, *Biotropica* 51 (2019) 319–329.
- [17] K. Minten, Development of a Business Plan for Production and Export of Green Coffee Beans from the Equateur Province in the Democratic Republic of the Congo (Ph.D. thesis. Master thesis), Ghent University, Ghent, Belgium, 2017, p. 68.
- [18] L. Summerauer, Sustainable Agricultural Intensification Methods of Cassava Based Systems for Improving Livelihoods and Forest Conservation in the Congo Basin (Ph.D. thesis. Master thesis), ETH Zurich, Zurich, Switzerland, 2017, p. 60.
- [19] L. Summerauer, P. Baumann, L. Ramirez-Lopez, M. Barthel, M. Bateurs, B. Bukombe, M. Reichenbach, P. Boeckx, E. Kearsley, K. Van Oost, et al., The central african soil spectral library: a new soil infrared repository and a geographical prediction analysis, *Soil* 7 (2021) 693–715.
- [20] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, *Chemometr. Intell. Lab. Syst.* 50 (2000) 1–18.
- [21] T. Hengl, J. Sanderman, L. Parente, Open soil spectral library (training data and calibration models), 2021, <http://dx.doi.org/10.5281/zenodo.5759694>.
- [22] L. Deiss, A.J. Margenot, S.W. Culman, M.S. Demyan, Optimizing acquisition parameters in diffuse reflectance infrared fourier transform spectroscopy of soils, *Soil Sci. Am. J.* 84 (2020) 930–948.
- [23] A. Savitzky, M.J. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Anal. Chem.* 36 (1964) 1627–1639.
- [24] R. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Appl. Spectrosc.* 43 (1989) 772–777.
- [25] O. Chapelle, B. Schölkopf, A. Zien (Eds.), *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.
- [26] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, *Chemometr. Intell. Lab. Syst.* 95 (2009) 35–48.
- [27] E. Pérez-Fernández, A.J. Robertson, Global and local calibrations to predict chemical and physical properties of a national spatial dataset of scottish soils from their near infrared spectra, *J. Near Infrared Spectrosc.* 24 (2016) 305–316.
- [28] J.R. Quinlan, Combining instance-based and model-based learning, in: *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 236–243.
- [29] R. Haghi, E. Pérez-Fernández, A. Robertson, Prediction of various soil properties for a national spatial dataset of scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy, *Geoderma* 396 (2021) 115071.
- [30] J.L. Safanelli, T. Hengl, L.L. Parente, R. Minarik, D.E. Bloom, K. Todd-Brown, A. Gholizadeh, W.d.S. Mendes, J. Sanderman, Open soil spectral library (oss): Building reproducible soil calibration models through open development and community engagement, *PloS One* 20 (2025) e0296545.
- [31] W. Ng, B. Minasny, M. Montazerolghaem, J. Padarian, R. Ferguson, S. Bailey, A.B. McBratney, Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra, *Geoderma* 352 (2019) 251–267.
- [32] D. Passos, P. Mishra, Deep tutti frutti: Exploring cnn architectures for dry matter prediction in fruit from multi-fruit near-infrared spectra, *Chemometr. Intell. Lab. Syst.* 243 (2023) 105023.
- [33] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2025, URL: <https://www.R-project.org/>.
- [34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, *Adv. Neural Inf. Process. Syst.* 32 (2019).
- [35] D. Eddelbuettel, J.J. Balamuta, Extending r with c++: a brief introduction to rcpp, *Amer. Statist.* 72 (2018) 28–36.
- [36] D. Eddelbuettel, C. Sanderson, Rcpparmadillo: Accelerating r with high-performance c++ linear algebra, *Comput. Statist. Data Anal.* 71 (2014) 1054–1063.
- [37] A. Stevens, L. Ramirez-Lopez, An introduction to the prospectr package, 2025.
- [38] L. Ramirez-Lopez, A. Stevens, R. Viscarra Rossel, C. Lobsey, A. Wadoux, T. Breure, Resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics, 2025, R package version 1.
- [39] A. Stevens, M. Nocita, G. Tóth, L. Montanarella, B. van Wesemael, Prediction of soil organic carbon at the european scale by visible and near infrared reflectance spectroscopy, *PloS One* 8 (2013) e66409.
- [40] J.L. Safanelli, J. Sanderman, D. Bloom, K. Todd-Brown, L.L. Parente, T. Hengl, S. Adam, F. Albinet, E. Ben-Dor, C.M. Boot, et al., An interlaboratory comparison of mid-infrared spectra acquisition: Instruments and procedures matter, *Geoderma* 440 (2023) 116724.