Gebreyesus, G., Jensen, K., Andersen, L., Noel, S., Buitenhuis, A., Zaalberg, R. (2026) Phenotyping through the mess: contamination-aware NIR models for prediction of amino acid composition in live black soldier fly larvae for selective breeding applications, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 24144: https://doi.org/10.31274/wcgalp.24144
©2026 *Gebreyesus, et al.* CC BY-NC-ND

# Phenotyping through the mess: contamination-aware NIR models for prediction of amino acid composition in live black soldier fly larvae for selective breeding applications

G. Gebreyesus[1*], K. Jensen[2], L.B. Andersen[1], S.J. Noel[2], A.J. Buitenhuis[1], and R.M. Zaalberg[1]

*[1]Aarhus University, Aarhus C, Denmark; [2]Aarhus University, Tjele, Denmark; [*]Corresponding author: grum.gebreyesus@qgg.au.dk*

## Introduction

The black soldier fly (BSF; *Hermetia illucens*) has emerged as a leading species for the industrial bioconversion of organic side-streams into high-value protein and lipid biomass (Gold et al., 2018). As the industry scales, nutritional composition, specifically amino acid (AA) profiles, becomes of primary economic interest, determining feed formulation efficiency, animal growth performance, and ingredient market value (Barragán-Fonseca et al., 2017). While genetic improvement shows potential, large-scale phenotyping remains a critical bottleneck (Nawoya et al., 2024; 2025). Conventional AA determination relies on chromatographic methods that are destructive, time-consuming, and expensive, making them incompatible with breeding pipelines requiring thousands of measurements per generation. Near-infrared reflectance spectroscopy (NIRS) offers a rapid, low-cost alternative for biochemical phenotyping and is widely adopted in crop and livestock breeding. In insects, NIRS has successfully predicted protein and fat in dried meals, but live-larva phenotyping remains challenging due to the moist, heterogeneous substrates inherent to production (Ssemakula et al., 2025). Adhering feed particles introduce strong spectral contributions from lipids, carbohydrates, and fermentation products that overlap with larval tissue absorption bands. Standard NIR calibration frameworks often fail because they treat substrate signals as noise rather than chemically structured information. This limitation is recognized in dairy science, where Cattaneo and Holroyd (2013) demonstrated that milk contaminants introduce predictable spectral signatures that allow for reliable quantification when explicitly modelled. Similarly, in polymer recycling, Mhaddolkar et al. (2024) showed that food residues on plastics can be captured by multivariate models if treated as part of the spectral system. Because the BSF substrate matrix directly interacts with larval tissues, automated phenotyping systems must explicitly incorporate substrate effects. Here, we present a contamination-aware NIR framework for predicting AA composition in live BSF larvae. Our approach integrates clean larval spectra, contaminated larval spectra, and paired wet- and dry-matter diet spectra within a block-structured partial least squares (PLS) model. By training on clean larvae and evaluating on larvae with real substrate residues, we demonstrate a robust method for high-throughput, non-destructive biochemical phenotyping under production-relevant conditions.

## Materials & Methods

**Larval material and experimental design.** BSF larvae were reared on 17 experimental diets designed to maximize biochemical variation by modifying protein (2.3–89.8% dry matter) and digestible carbohydrate (8.5–96.0% dry matter) levels through casein (C7078, Sigma-Aldrich, Steinheim, Germany) or sucrose (Dansukker, Copenhagen, Denmark) additions to a base of chicken feed (Paco Start, Danish Agricultural Grocery, Fredericia, Denmark). Larvae were collected on four consecutive days (12–15 days post-hatching) across three replicates, yielding 204 pooled larval group samples (111–337 larvae per group sample) for analysis. Each larval

group sample originated from a unique cup. For each group sample, a transparent plastic cup (70:95 mm H:Ø) was prepared with 60 g mixed dry feed and 140 mL preboiled water with 1 g agar, and ~300 six-day old larvae. The cups were closed with a lid with a 65 mm Ø hole covered with a mesh for ventilation and placed in a climate chamber at 28°C and 70% RH.

**Near-infrared spectral acquisition.** Spectra (400–2,500 nm) were acquired at 0.5-nm resolution using a FOSS DS2500 spectrometer. Each sample was scanned in two states: "*CLEAN*" (rinsed and towel-dried) and "*DIRTY*" (before cleaning while contaminated with adhering substrate and moisture). Larvae were scanned twice with repacking to reduce positioning effects, then averaged into representative clean and dirty spectra. Paired wet- and dry-matter spectra were also collected for each diet to provide explicit representations of the contaminating matrix.

**Amino-acid reference phenotypes.** Following the second NIR scanning after cleaning, larval samples were immediately frozen. For chemical reference analysis, larvae were freeze-dried and homogenized by grinding, followed by acid extraction and quantification of free, underivatized amino acids using LC–MS/MS. This approach avoids hydrolysis-induced degradation and provides a physiologically meaningful measure of larval amino-acid composition. Amino acid concentrations are reported on a dry-matter basis (mg $g^{-1}$ larval dry weight). In total, 20 free amino acids, including all nutritionally essential amino acids, were quantified for each larval group and used as reference phenotypes for spectral model training and evaluation.

**Contamination-aware spectral modelling.** Predictions were performed using block-structured partial least squares (PLS) regression in Python. Spectral predictors were organised into three blocks: larval tissue spectra from uncontaminated samples (*CLEAN_spectra*), near-infrared spectra of the dry-matter diet (dry content before mixing with water), and near-infrared spectra of the wet-matter diet (dry content and water). During model training, clean larval spectra were used to define the spectral signature of larval tissue, while paired diet spectra were included to explicitly parameterise the contaminating substrate. This allowed the model to learn how larval tissue and substrate signals contribute to the observed spectrum. Model evaluation was conducted exclusively on the dirty larval spectra (*DIRTY_spectra)*, ensuring that performance was assessed under realistic rearing conditions. The number of latent components was selected by minimising the cross-validated root mean squared error of prediction (RMSEP) within the training set. Model performance on the contaminated test set was evaluated using the coefficient of determination ($R^2$) and RMSE.

## Results

Summary statistics and prediction performance for all 20 AA is summarized in Table 1. Mean concentrations ranged from 0.06 mg/g for asparagine to 6.50 mg/g for alanine, while total free amino acids (tAA) averaged 34.95 ± 17.29 mg/g. Substantial variation was observed across samples, with coefficients of variation (CV) ranging from 38.9% for threonine to over 124% for asparagine. Prediction accuracy varied markedly across the AA profile. Several traits exhibited moderate to strong predictability despite substantial substrate interference. Glutamate showed the highest performance ($R^2_{Test}$ = 0.66), followed by proline (0.48) and glycine (0.45). Aromatic amino acids also retained structured signals, specifically tryptophan (0.34) and phenylalanine (0.30). Among nutritionally essential amino acids, histidine, isoleucine, methionine, and valine achieved $R^2$ values between 0.21 and 0.26, indicating that key indicators of larval protein quality remain detectable under contaminated conditions. Conversely, several amino acids showed negligible to weak predictive power, reflecting limited signal-to-noise

ratios. Arginine demonstrated the lowest predictability ($R^2_{Test} < 0.001$), followed by glutamine (0.02), tyrosine (0.03), and lysine (0.05). Total free amino acids (tAA) were predicted with an $R^2$ of 0.17, representing the aggregate signal of both well- and poorly predicted individual components.

**Table 1.** Average concentrations (mg/g of dry sample) and predictive performance for free amino acids in black soldier fly larvae evaluated under contaminated test conditions using contamination-aware NIR models.

| Amino acid | Mean (± SD)[2] | CV (%)[3] | $R^2_{Test}$[4] | $RMSE_{Test}$[5] |
|---|---|---|---|---|
| Alanine | 6.50 (3.52) | 54.16 | 0.22 | 5.28 |
| Arginine | 1.09 (0.85) | 78.3 | 0.00071 | 2.04 |
| Asparagine | 0.06 (0.08) | 124.7 | 0.08 | 0.14 |
| Aspartate | 0.33 (0.22) | 67.21 | 0.08 | 0.48 |
| Glutamate | 2.41 (1.66) | 68.78 | 0.66 | 1.81 |
| Glutamine | 5.89 (5.46) | 92.57 | 0.02 | 20.87 |
| Glycine | 1.16 (0.74) | 63.37 | 0.45 | 0.81 |
| Histidine | 2.32 (2.46) | 105.95 | 0.26 | 3.84 |
| Isoleucine | 0.96 (0.47) | 48.82 | 0.22 | 0.64 |
| Leucine | 1.25 (0.70) | 56.08 | 0.11 | 1.55 |
| Lysine | 1.34 (0.82) | 61.52 | 0.05 | 1.35 |
| Methionine | 0.20 (0.15) | 73.18 | 0.21 | 0.22 |
| Phenylalanine | 0.64 (0.29) | 45.51 | 0.30 | 0.28 |
| Proline | 5.61 (4.45) | 79.26 | 0.48 | 5.78 |
| Serine | 0.95 (0.63) | 66.25 | 0.25 | 0.56 |
| Threonine | 0.91 (0.35) | 38.89 | 0.10 | 0.49 |
| Tryptophane | 0.63 (0.26) | 40.95 | 0.34 | 0.32 |
| Tyrosine | 1.10 (0.81) | 73.16 | 0.03 | 1.53 |
| Valine | 1.48 (0.72) | 48.55 | 0.25 | 1.40 |
| tAA[1] | 34.95 (17.29) | 49.46 | 0.17 | 28.2 |

[1]tAA = total free amino acids as the sum of all quantified amino acids (mg $g^{-1}$ sample).
[2]Mean (± SD) is calculated across all 204 larval group samples.
[3]CV = coefficient of variation (%) across samples.
[4]$R^2_{Test}$ = coefficient of determination between observed and predicted AA in the DIRTY_spectra.
[5]$RMSE_{Test}$ = root mean squared error of prediction in the DIRTY_spectra (mg/ g).

## Discussion

The results demonstrate that NIR spectra acquired from substrate-contaminated larvae contain sufficient structured biochemical information to support prediction of multiple amino acids when contamination is explicitly modelled. Rather than acting as unstructured noise, the substrate introduces reproducible spectral components that can be separated from larval tissue signals when diet spectra are incorporated as dedicated predictor blocks.

The strong performance for glutamate, proline and glycine is consistent with their high abundance and strong association with amide, N–H and C–H combination bands in the near-infrared region, which dominate protein-related absorption features. The fact that these signals remain recoverable under contamination indicates that larval biochemical variation is not lost but is masked by overlapping substrate spectra that must be modelled rather than removed. Similar principles have been demonstrated in contaminated NIR systems in food and polymer sorting, where spectral interference from adhering materials can be explicitly learned by

multivariate models rather than filtered out (Mhaddolkar et al., 2024). NIR scanning as applied here was non-destructive and of short duration; effects of repeated or prolonged NIR exposure on larval survival or adult reproductive fitness were not assessed in this study and would be a relevant question to address before deploying the method at scale.

The strict separation between clean training samples and contaminated test samples ensures that reported accuracies reflect realistic deployment conditions relevant for breeding-oriented phenotyping, where larvae are measured under practical rearing environments. By explicitly incorporating diet spectra into the model structure, the approach accounts for substrate-related spectral interference and supports prediction when larvae are measured in the presence of residual substrate. Because amino acid reference values were obtained from freeze-dried, homogenized larvae and models were trained exclusively on clean larval spectra, the observed predictive performance is consistent with recovery of larval biochemical information rather than indirect proxy prediction from substrate alone. The prediction accuracies achieved here, with $R^2$ values reaching 0.66 for the best-predicted trait and a median around 0.25 across the full amino acid profile, are below generally accepted thresholds typically for index-based selection in livestock. Improving accuracy will likely require expanding the calibration dataset, particularly with greater substrate diversity, as the spectral model is currently trained on a limited set of diet types.

## Conclusions

This study shows that block-structured PLS modelling can explicitly account for substrate interference in NIR-based phenotyping of live BSF larvae. By integrating diet spectra as structured predictors, biologically meaningful variation in amino acid composition could be recovered even when larvae were scanned in a naturally contaminated state. While the achieved prediction accuracies vary across traits and do not in themselves define breeding-ready selection criteria, they demonstrate that informative biochemical signal persists under realistic rearing conditions. The proposed contamination-aware framework therefore provides a practical foundation for non-destructive, high-throughput phenotyping of nutritional traits, supporting future efforts to integrate NIR spectroscopy into genetic evaluation pipelines as calibration datasets expand and prediction accuracy improves.

## Acknowledgements

Financial support was provided by the Aarhus University Research Foundation (AUFF–NOVA). We thank Enorm Biofactory A/S for providing the larvae and the DynaMo MS-Analytics Facility at the University of Copenhagen for performing the LC-MS/MS analysis.

## References

Gold, M., Tomberlin, J.K., Diener, S., Zurbrügg, C., & Mathys, A. (2018). Waste Management, 82, 302-318

Barragán-Fonseca K.B., Dicke M., and van Loon J.J.A. (2017). Journal of Insects as Food and Feed 3:105–120.

Nawoya, S., Ssemakula, F., Akol, R., Geissmann, Q., Karstoft, H., Bjerge, K., Mwikirize, C., Katumba, A., & Gebreyesus, G. (2024). Comput. Electron. Agric., 216, 108503.

Nawoya S., Geissmann, Q., Karstoft, H., Bjerge, K., Roseline, A., Katumba, A., Mwikirize, C., & Gebreyesus, G. (2025). Smart Agricultural Technology., 11, 100953.

Ssemakula, F., Nawoya, S., Kunyanga, C.N., Akol, R.N., Nakimbugwe, D., Bett, R., Karstoft, H., Bjerge, K., Katumba, A., Mwikirize, C., & Gebreyesus, G. (2025). Comput. Electron. Agric., 239C, 111126.

Cattaneo T.M.P., and Holroyd S.E. (2013). Food Chemistry 141:1784–1792.

Mhaddolkar S., Stoll D., Riedel M., and Kessler W. (2024). Polymers 16:2343.