



Improving Protein Quantification with SERS Superspectra and Machine Learning

Jiaheng Cui¹, Chenyao Feng², Xulan Chen³, Yanjun Yang⁴, Pengju Yin², Yiping Zhao⁴

1. School of Electrical and Computer Engineering, University of Georgia, Athens, GA 30602, USA
2. School of Mathematics and Physics, Hebei University of Engineering, Handan, Hebei 056038, China
3. Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA
4. Department of Physics and Astronomy, University of Georgia, Athens, GA 30602, USA



Introduction

Quantitative protein analysis by SERS remains difficult: protein adsorption on plasmonic surfaces is weak, heterogeneous, and strongly dependent on surface chemistry and molecular orientation. Individual Raman peaks therefore rarely track concentration monotonically, and conventional calibration-curve quantification becomes unreliable.

We introduce a superspectra-guided SERS strategy in which the same protein is probed under several chemically distinct surface environments. Spectra from complementary surfaces are concatenated into a single feature vector and analyzed by machine learning (ML), using bovine serum albumin (BSA) as a model protein.

Objectives

- Determine how AgNR surface chemistry modulates protein SERS fingerprints.
- Evaluate whether chemically complementary superspectra improve ML-based protein quantification relative to single substrates.

Experimental Strategy

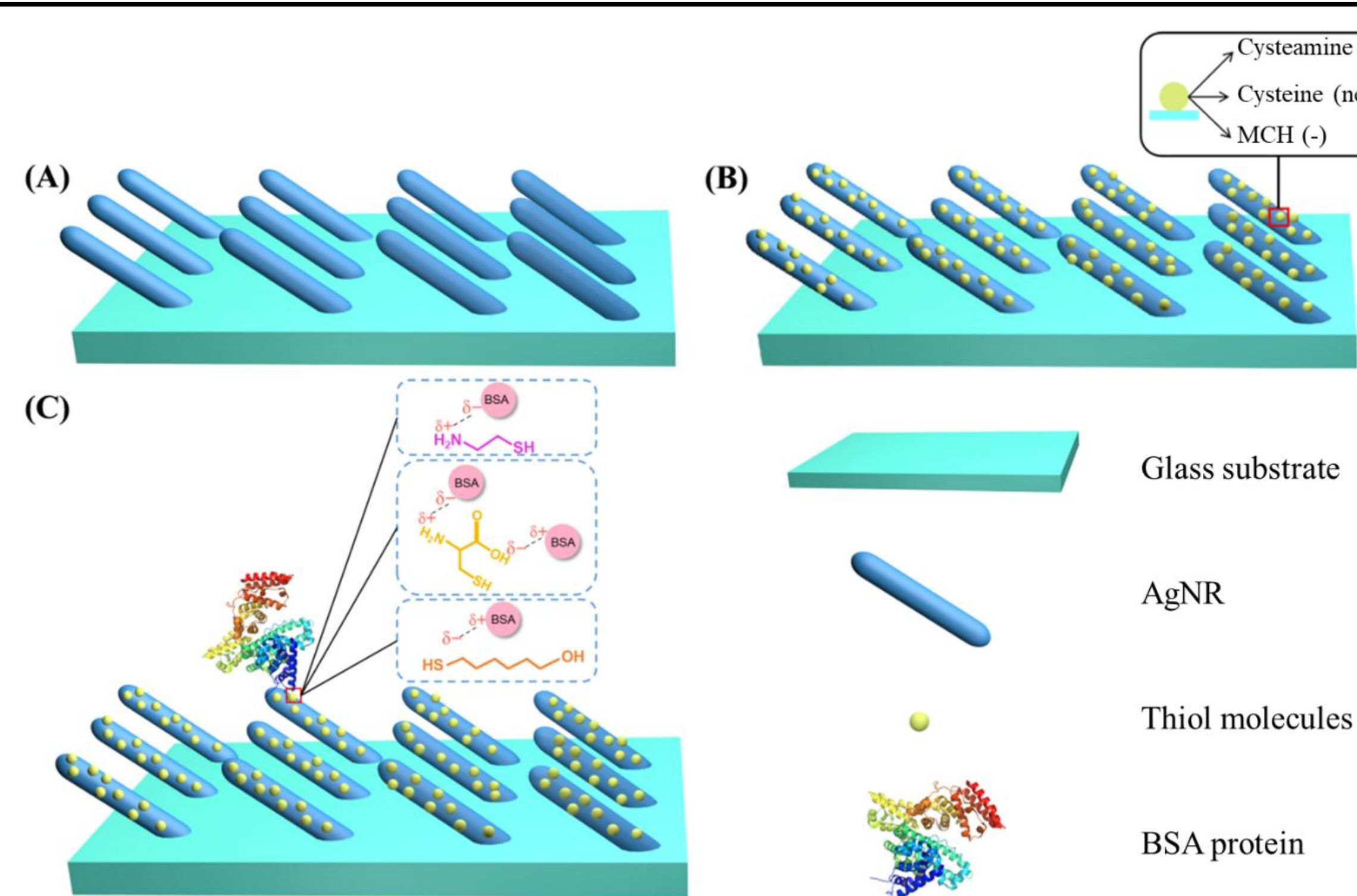


Fig. 1. Superspectra-guided SERS strategy for protein quantification. BSA was measured on bare (B), cysteamine (CM, +), cysteine (CN, zwitterionic), and MCH-functionalized AgNR substrates. Preprocessed spectra acquired at the same BSA concentration on different surfaces were concatenated into S1–S4 superspectra and analyzed by ML regression. Inset: SEM of the AgNR substrate.

- Substrates: AgNR arrays by oblique angle deposition; PDMS four-well layer; thiol SAMs of CM, CN, MCH.
- BSA: five concentrations, 8×10^{-9} – 0.05 mg/mL (five orders of magnitude); 20 μ L dried per well.
- SERS: 785 nm, 10% power, 5 $\times$ objective, 10 s; 50–60 spectra per concentration per surface.
- Preprocessing (SpectraGuru): Interpolation $\rightarrow$ Gaussian–Lorentzian baseline + airPLS $\rightarrow$ mean normalization.

Complementary SERS Fingerprints

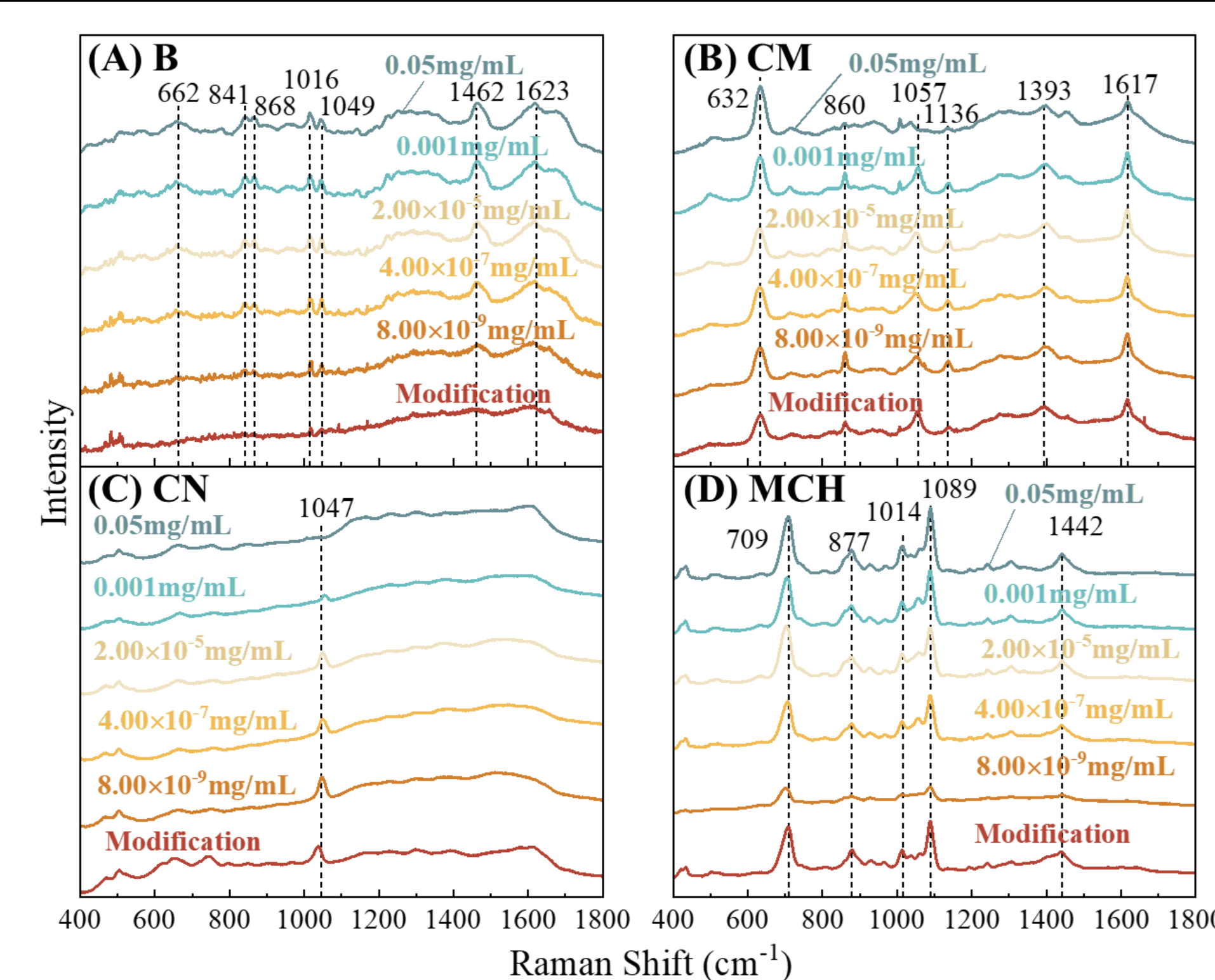


Fig. 2. Surface chemistry produces distinct BSA SERS fingerprints. Average normalized spectra on (A) bare, (B) CM-, (C) CN-, and (D) MCH-modified AgNRs at five BSA concentrations (8×10^{-9} to 0.05 mg/mL, bottom to top); the red trace is the modifier-only reference.

Bare (B): clean BSA fingerprint (662, 841, 1016, 1462, 1623 cm^{-1}).
CM (+): attracts acidic residues; amide mode at 1617 cm^{-1} enhanced.
CN ($\pm$): marker at 1047 cm^{-1} decreases.
MCH (–): weak adsorption; modifier bands dominate, protein modes suppressed.

The four surfaces do not repeat one spectrum — each encodes a different aspect of the protein–surface interaction.

Surface identification by SVM and RF: 100% test accuracy — the four surfaces carry non-redundant information.

Single-Peak Calibration Fails

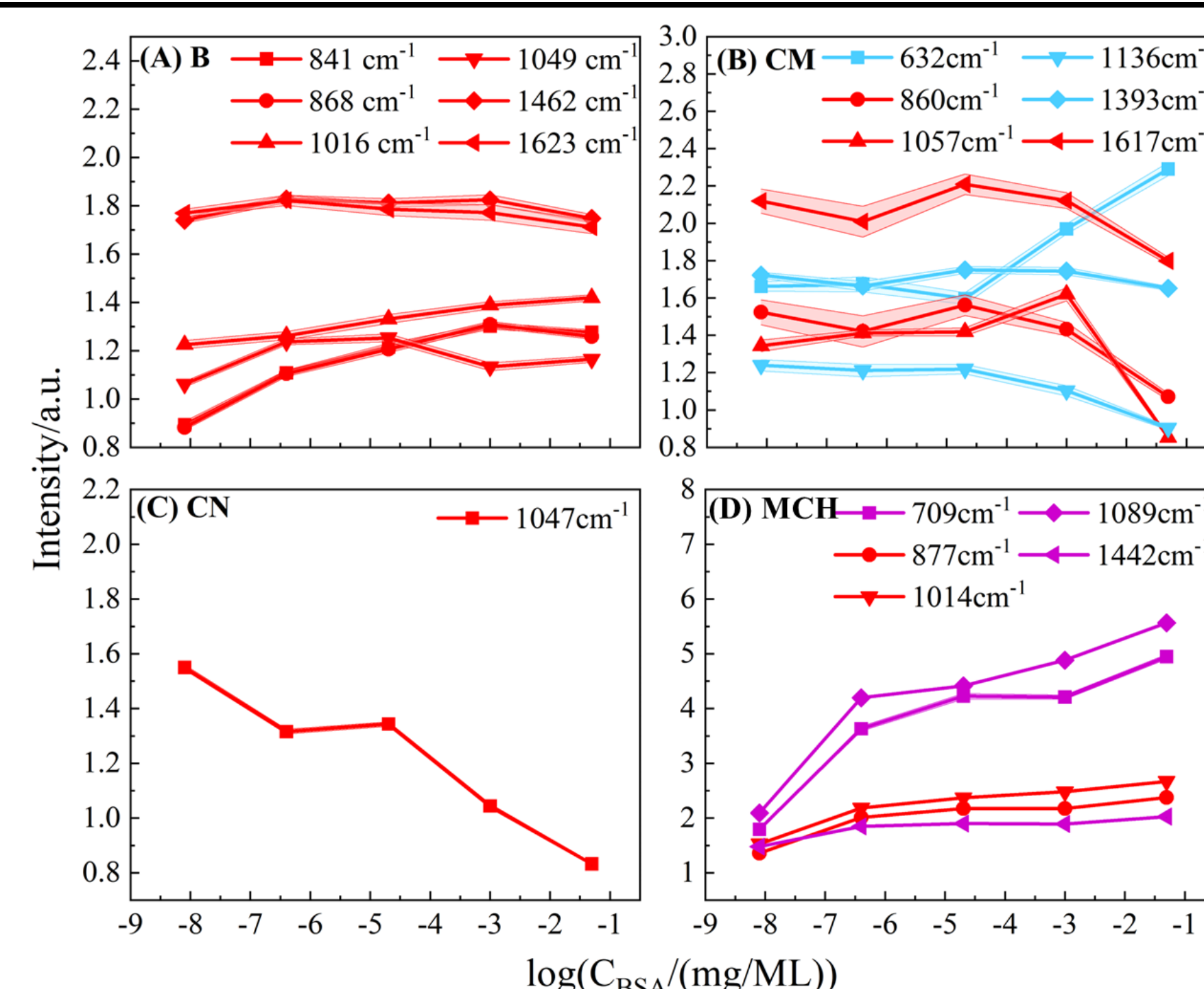


Fig. 3. Raman peak intensities show non-monotonic concentration dependence. BSA- and modifier-related peaks on (A) bare, (B) CM-, (C) CN-, and (D) MCH-modified substrates respond irregularly across five concentrations.

Competitive adsorption, surface saturation, and chemistry-dependent orientation produce nonlinear responses; quantitative information must therefore be extracted from the full spectral pattern.

Analysis Workflow

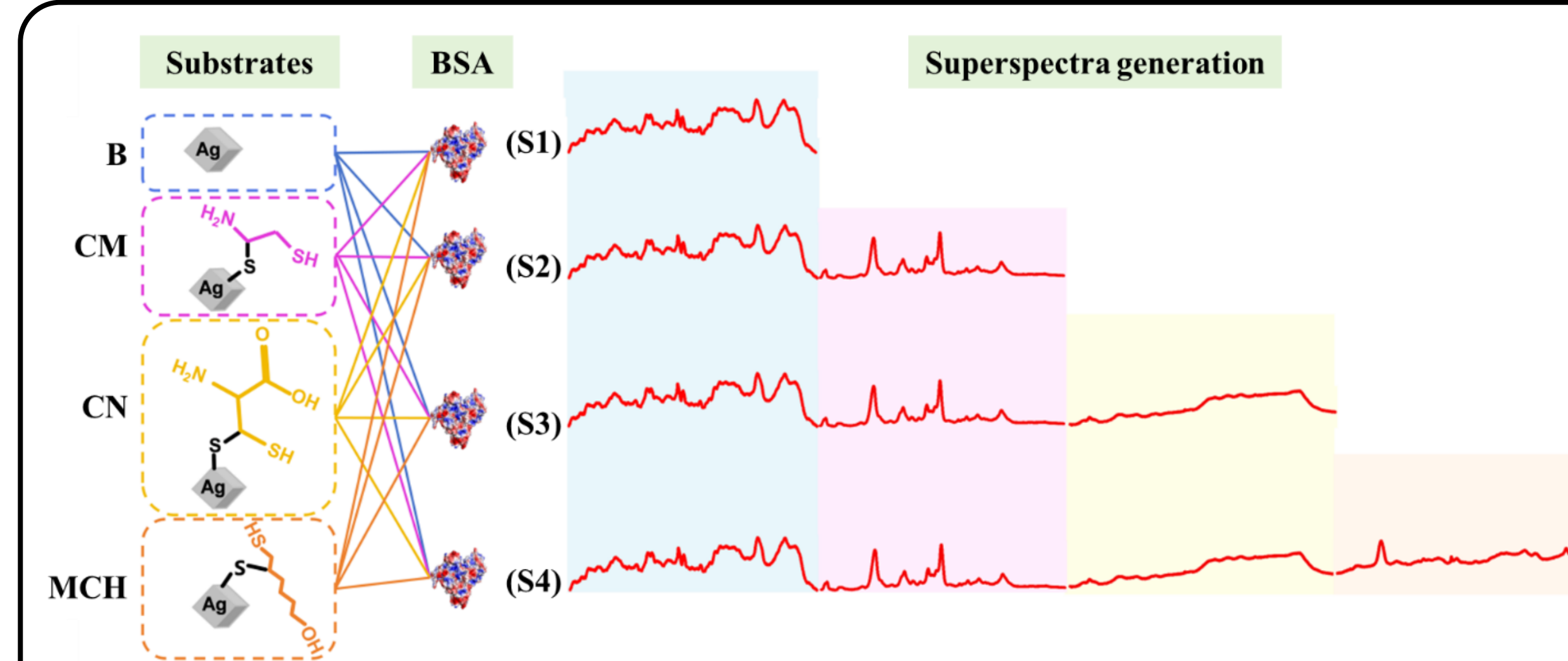


Fig. 4. Schematic illustration of superspectra generation. SERS spectra from individual surface-modified substrates—bare (B), cysteamine (CM), cysteine (CN), and MCH—are arranged in alphabetical order and concatenated to form superspectra.

- Superspectra: spectra from 1–4 substrates at the same BSA concentration concatenated in fixed order (B $\rightarrow$ CM $\rightarrow$ CN $\rightarrow$ MCH).
- Stratified 2:1 train/test split by substrate–concentration group; no data leakage.
- Regression: Support vector regression (SVR) and random forest regression (RFR), 5-fold cross-validated grid search, hyperparameters selected by lowest validation MAE.
- RFR consistently outperformed SVR (SVR weighted MAE > 3)

Superspectra Improve Quantification

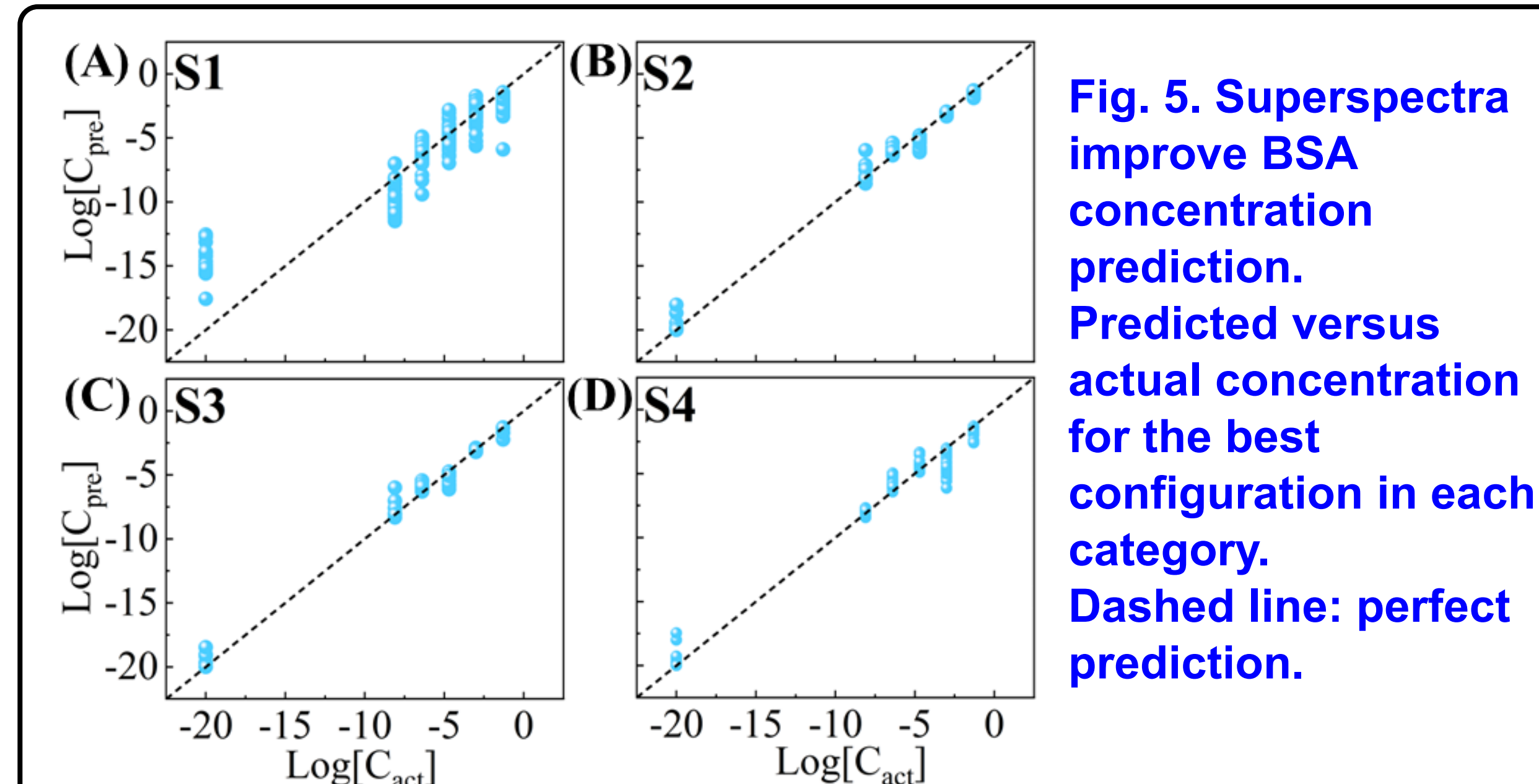


Fig. 5. Superspectra improve BSA concentration prediction. Predicted versus actual concentration for the best configuration in each category. Dashed line: perfect prediction.

Input	Best combination	MAE	R ²
S1	CN	0.477	0.966
S2	CM&CN	0.235	0.993
S3	B&CM&CN	0.254	0.993
S4	B&CM&CN&MCH	0.494	0.978

**bare substrate $\rightarrow$ best superspectrum:
MAE 1.44 $\rightarrow$ 0.24 R² 0.795 $\rightarrow$ 0.993**

- Adding chemically complementary information improves prediction; increasing the number of substrates alone does not.
- S2 (CM&CN) outperforms S4 (all four substrates) — the best model is not the largest one.

Performance Across Combinations

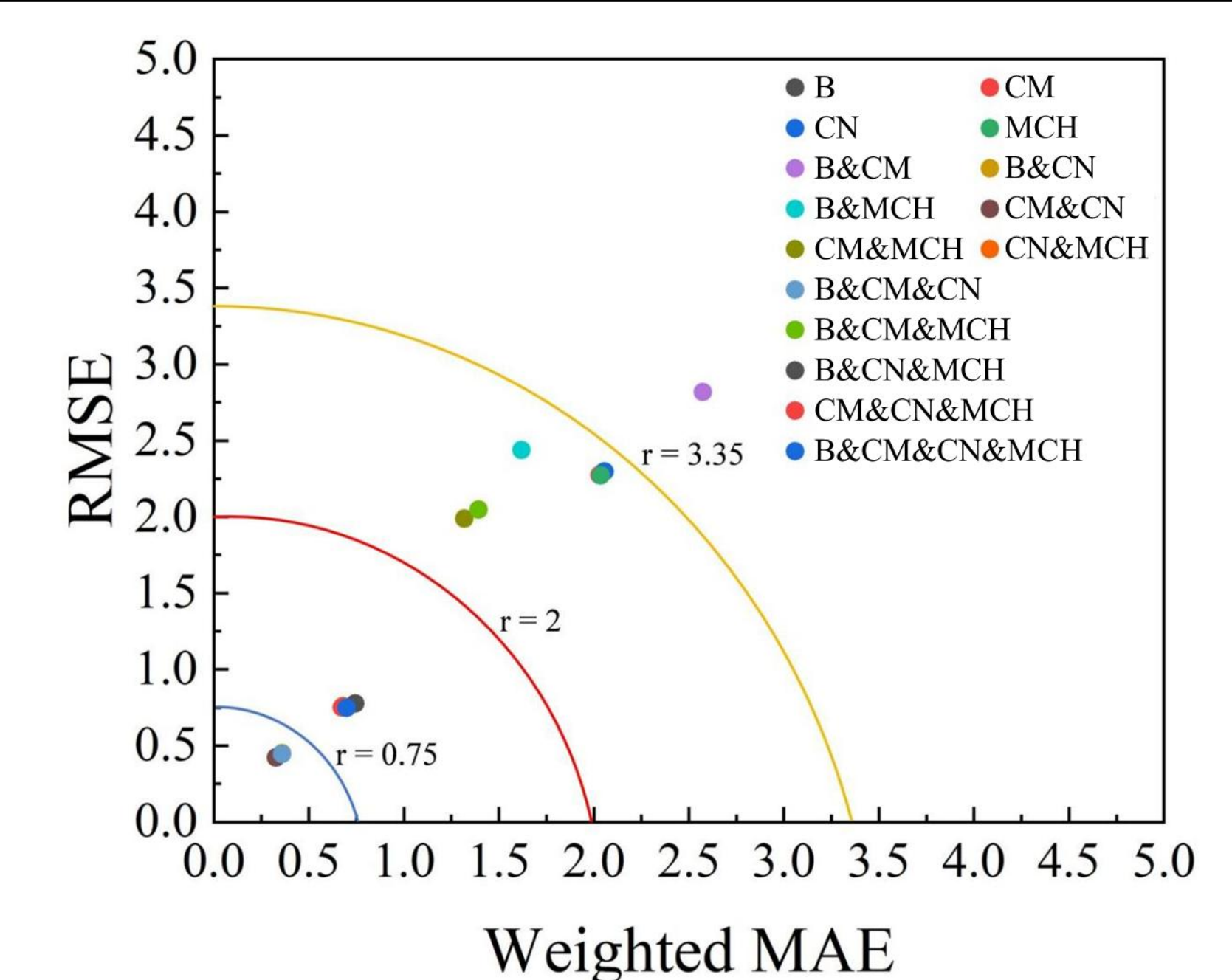


Fig. 6. Performance is governed by substrate chemistry, not substrate count. Weighted MAE versus RMSE for all RFR configurations; contours denote constant combined error. MCH-containing combinations cluster far from the origin.

- CM&CN lies closest to the origin; B&CM&CN ranks second.
- Several S2 and S3 combinations outperform the four-substrate S4 model.
- For acidic BSA (pI $\approx$ 4.7), the hydroxyl-terminated MCH surface adsorbs little protein and contributes weak or conflicting features.

Main Takeaway

- B $\rightarrow$ weak BSA information $\rightarrow$ accuracy drops $\times$
- CM (+) + CN ($\pm$) $\rightarrow$ complementary modes $\rightarrow$ high accuracy $\checkmark$
- B + CM + CN $\rightarrow$ neutral + cationic + zwitterionic $\rightarrow$ high accuracy $\checkmark$
- B + CM + CN + MCH (–) $\rightarrow$ weak/conflicting BSA information $\rightarrow$ accuracy drops $\times$ (for acidic proteins such as BSA)

Maximize chemical complementarity, not the number of spectra.

Conclusions

- Surface functionalization generates chemically distinct BSA SERS fingerprints.
- Individual Raman peaks vary non-monotonically with concentration, limiting peak-based calibration.
- CM&CN and B&CM&CN superspectra gave the highest RFR accuracy (R² = 0.993), a 4–5 $\times$ error reduction versus single substrates.
- Effective superspectra require complementary analyte–surface interactions; more substrates do not guarantee better prediction.

Acknowledgement

Funded by USDA NIFA Grant 2023-67015-39237.

Reference

Cui, Jiaheng, et al. "Improving Protein Quantification with SERS Superspectra and Machine Learning." ACS omega 11.6 (2026): 10843-10853.