A Machine Learning Classification Model for Gold-Binding Peptides

Ali Ahmadi Esfidi

May 2025

1 Gold-Binding Peptides

Gold-binding peptides are short chains of amino acids (typically 5–20 residues) that have a natural affinity for gold surfaces or nanoparticles. They are identified or designed so that specific residues (often cysteine, histidine, or aromatic amino acids) coordinate with gold atoms, allowing the peptide to stick strongly and specifically to gold.[1]

1.1 Usage

1. **Nano-templating & Nanofabrication.** Peptides guide the formation of gold nanowires, rods, or particles with controlled size and shape, serving as a "molecular mold" for electronic or optical devices.[2]

2. **Biosensing & Diagnostics.** When immobilized on electrodes or sensor surfaces, gold-binding peptides can capture target biomolecules (e.g., antibodies) in a precise orientation, improving sensitivity for medical assays.[3]

3. **Targeted Drug Delivery & Imaging.** Conjugating drugs or imaging agents to gold nanoparticles via these peptides allows for targeted delivery and enhanced imaging contrast in cancer or inflammatory disease models.[4]

4. **Surface Functionalization.** They enable simple, one-step coating of gold surfaces with proteins or other functional polymers, useful in creating antifouling coatings or bioactive interfaces.

1.2 Intensity

In spot-array binding assays, each 10-residue peptide is immobilized on a solid support and exposed to gold nanoparticles. The *intensity* is defined as the

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC10337651/
[2] https://pubs.rsc.org/en/content/articlepdf/2023/ra/d3ra04269c
[3] https://pmc.ncbi.nlm.nih.gov/articles/PMC9918321/
[4] https://www.sciencedirect.com/science/article/abs/pii/S0378517324011542

optical (colorimetric) signal measured at each peptide spot, proportional to the amount of bound nanoparticles. Peptides with stronger binding produce darker spots (higher intensity), whereas weak or non-binding sequences yield lighter spots (lower intensity).

In the dataset of Janairo *et al.*, each of the 1 720 unique 10-mer peptides was assigned an intensity value based on the median image-analysis readout from Tanaka *et al.*'s screen. To classify peptides into binders and non-binders, the median intensity of the entire set, denoted I_{med}, was used as the threshold:

$$I_{\text{med}} = 207{,}500 \quad \text{(arbitrary units)}.$$

Peptides were then dichotomized into two classes:

$$\begin{aligned} \text{Class A (strong binders)} &: I > I_{\text{med}}, \\ \text{Class B (weak/non-binders)} &: I \leq I_{\text{med}}. \end{aligned}$$

Thus, the intensity values (in arbitrary units) provide a relative measure of each peptide's adsorption of gold nanoparticles under the standardized assay conditions.

2 Formulation of Problem

- **Input:**
  1. **Peptide sequence:** A string of ten amino-acid letters.
  2. **Derived features:** Each sequence is converted into a fixed-length numeric vector, so that the model can process it.
- **Output:** A binary prediction for each input sequence:
  - "Strong binder" (class A) if the model believes the peptide's binding intensity would exceed the threshold set by the median of all measured intensities.
  - "Weak/non-binder" (class B) otherwise.

2.1 Mathematical Presentation

1. Peptide Sequences:
$$S = \{s_i\}, \quad s_i \in \mathcal{A}^{10}$$
2. Feature Mapping:
$$\forall s_i \in S, \quad \Phi(s_i) = x_i \in R^d$$
3. Binding Intensities:
$$I = \{I_i\}_{i=1}^{N}, \quad I_i \in R_{\geq 0},$$

4. Classification Labels:

$$y_i = \begin{cases} A, & I_i > T, \\ B, & I_i \leq T, \end{cases} \quad T = \text{median}\big(\{I_i\}\big)$$

5. Prediction Function:

$$\forall s_i \in S, \quad f(\Phi(s_i)) = \hat{y}_i \in \{A, B\}$$

2.2 Assumptions

1. All peptides have fixed length 10.
2. Median intensity threshold T meaningfully separates strong vs. weak binders.
3. Samples are independently and identically distributed.

3 Dataset

Dataset comprises 1,720 peptide sequences, which we classify as A or B based on their intensity values.[5]

3.1 Distribution

It's important to confirm class balance before training classifiers, as imbalanced datasets can introduce bias.

The class sizes are nearly identical; exact count: **861** for Class A and **859** for Class B.

However, as illustrated in the violin plot and histogram in Figure 1, Class A peptides are high-intensity binders with intensity values concentrated around 250,000. Class B peptides, conversely, exhibit lower intensity values, ranging widely from nearly 0 to 200,000, and display a prominent left tail.

4 Related Work

As part of his lecture series, Jose Isagani B. Janairo presents a practical workflow for classifying gold-binding peptides that elegantly demonstrates end-to-end machine-learning in R: starting with raw amino-acid sequences, deriving ten physicochemical descriptors via Kidera factors, and organizing those into a labeled dataset; next, stratified train–validation splitting and 10-fold cross-validation to tune and compare multiple classifiers (logistic regression, decision trees, k-nearest neighbors, SVMs with various kernels, and a neural network); then refining the best performer—a radial-basis SVM—by selecting only the

[5] `https://pubs.acs.org/doi/suppl/10.1021/acsomega.2c00640/suppl_file/ao2c00640_si_001.pdf`

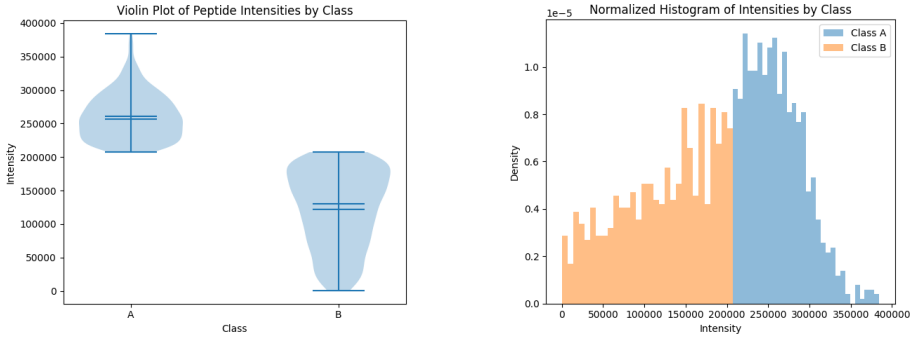

Figure 1: Overall caption for both images.

most informative descriptor subset; and finally, assessing generalization on hold-out data and employing permutation-based feature importance to reveal which peptide properties most strongly drive gold-binding predictions, all accompanied by clear visualizations of model performance and variable rankings.

On a macOS environment, the code was executed and produced results (Table 1) consistent with the related lecture (Accuracy: 0.8019).

	Reference	
	A	**B**
Prediction A	175	45
Prediction B	40	169

Table 1: Confusion Matrix

Also Table 2 reports the permutation-based importance of each Kidera factor (KF1–KF10) for the final radial-basis SVM, together with uncertainty bounds and the baseline increase in classification error when that feature is shuffled.

In practice, this tells you that the model heavily relies on KF4, KF2, and KF3—disrupting these degrades performance the most—whereas KF1 and KF8 contribute almost nothing to gold-binding predictions in the final SVM.

feature	importance	permutation.error
KF4	1.446512	0.2408985
KF2	1.395349	0.2323780
KF3	1.395349	0.2323780
KF9	1.306977	0.2176607
KF10	1.158140	0.1928737
KF5	1.148837	0.1913246
KF7	1.144186	0.1905500
KF6	1.130233	0.1882262
KF1	1.000000	0.1665376
KF8	1.000000	0.1665376

Table 2: Feature Importance