Supplementary information

for

Single-molecule Detection of Amino Acid Phosphorylation using electron tunnelling currents: Toward Neurodegenerative Disease Diagnosis

Yuki Komoto, Wataru Takahagi, Takahito Ohshiro, Sumire Nishihata, Kosuke Fujishima, and Masateru Taniguchi*

Correspondence to: taniguti@sanken.osaka-u.ac.jp

This PDF file includes:

SI 1. Post-fabrication attachment of the PDMS pool for solution containment

Electrical measurements were performed under solution conditions. To retain the solution, a polydimethylsiloxane (PDMS) pool was attached onto a Si substrate on which an Au nanobridge had been fabricated, as described in the main text. A schematic illustration of the measurement substrate equipped with the PDMS pool is shown in Figure S1. Measurements were carried out while the PDMS pool was filled with an aqueous amino acid solution.

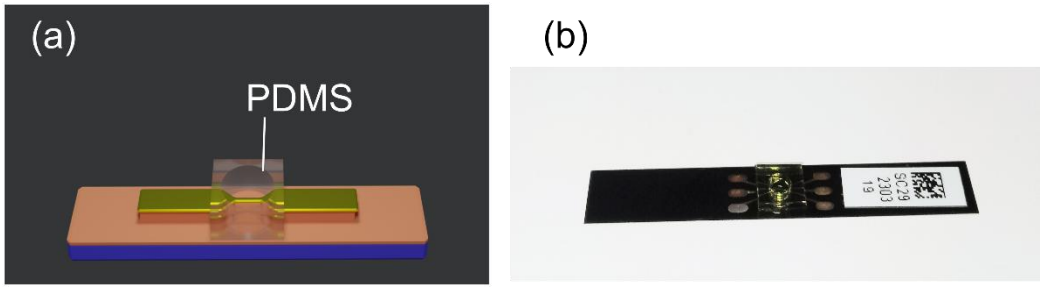

Figure S1. (a)Schematic illustration of the measurement substrate with a PDMS pool.
(b) Photograph of the MCBJ substrate with a PDMS pool.

SI 2. Estimation of gap distance

The gap distance is estimated using by following current equation of direct tunneling current

$$I = const \exp\left(-\frac{4\pi}{h}\sqrt{2mwl}\right).$$

Here, h, m,w, and l represents plank constant, electron mass, work function of gold electrode, gap distance. We used electron mass of 9.1×10^{-31} kg as m, and work function of Au(111) 5.3 eV as w. Effective mass and work function of gold nanogap not (111) surface should be used for accurate estimation. Furthermore, the inelastic gold gap broadening just after breaking atomic junction is not under consideration. Hence, the experimental gap length is larger than the target width of 0.56 and 0.58 nm .

SI 3. Electrical measurements result of blank

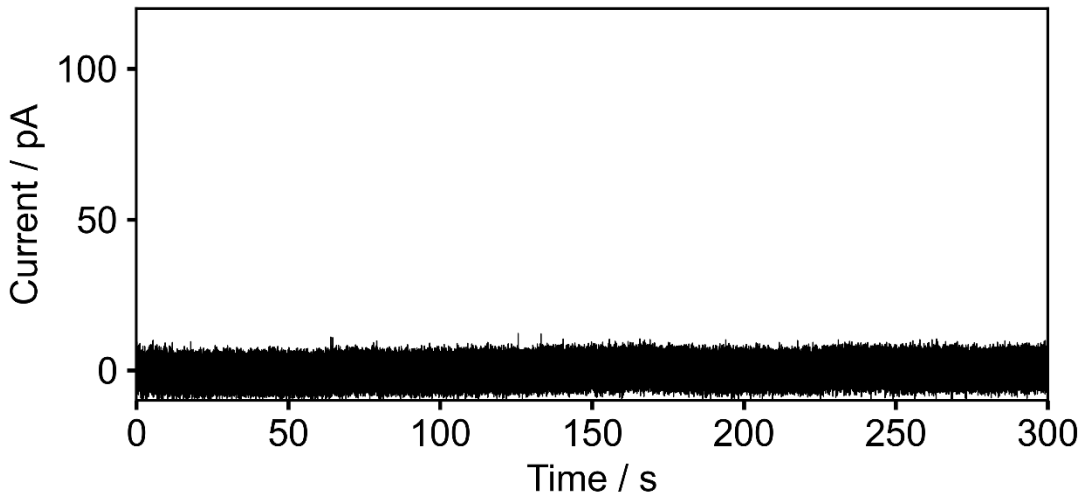

Figure S2. Current-time profiles of blank.

SI 4. Energy alignment of LUMO

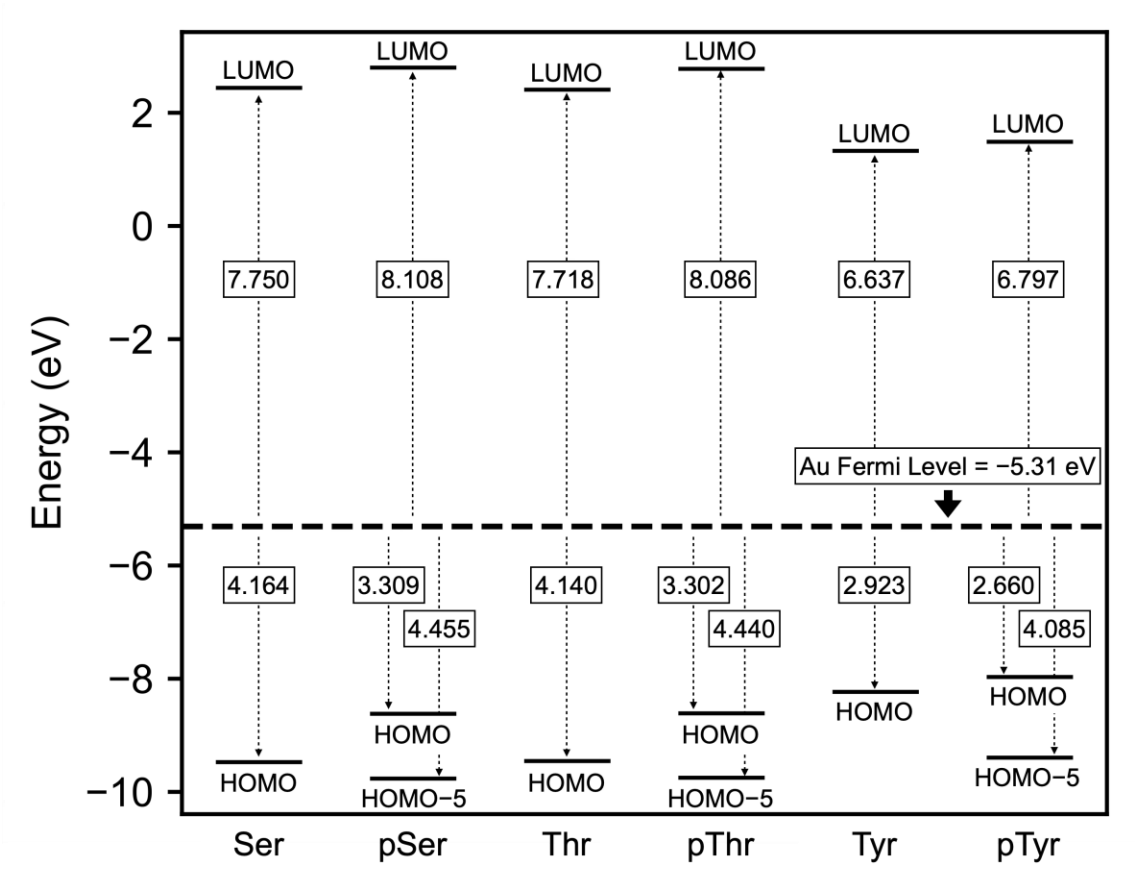

Figure S3 Energy diagram including LUMO of Ser, pSer, Thr, pThr, Tyr, and pTyr. Calculation method is described in main manuscript.

SI 5. Molecular size

Table S1 Van der Waals volume and Molecular length

Compound	van der Waals volume / Å^3	Molecular length / Å
Ser	91.90	5.10
pSer	129.76	6.72
Thr	108.47	5.41
pThr	146.44	6.83
Tyr	163.35	8.97
pTyr	203.39	11.02

The calculations were performed using DFT-optimized structures, where Ser, Thr, and Tyr were treated in their net-neutral zwitterionic forms, whereas pSer, pThr, and pTyr were treated as dianionic species (−2). Molecular length was defined as the maximum distance between any two atoms in the molecule. For the molecular volume calculation, each atom was represented as a sphere with its Bondi van der Waals radius, and the molecular van der Waals volume was defined as the union volume of all overlapping atomic spheres. The volume was evaluated numerically on a three-dimensional Cartesian grid with a spacing of 0.01 Å. Grid points located inside at least one atomic sphere were counted as occupied, and the total volume was obtained by multiplying the number of occupied voxels by the voxel volume.

SI 6. Feature importances of the classification model

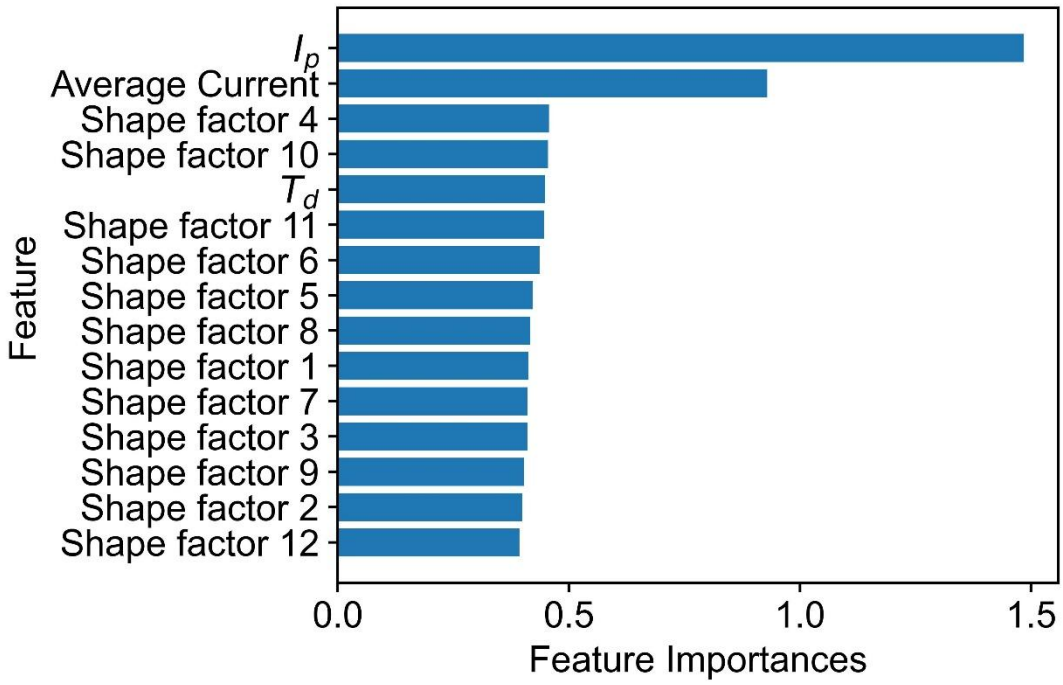

Figure S4 The importance of each feature in the 6 amino acids classification model.

Feature importances were defined as average loss reduction contributed by splits using the feature.

SI 7. Discrimination accuracy by multiple signals accumulation

The accuracies with multiple signals, 10, 20, and 30 signals were calculated as below.

For each molecule, the predicted outcomes for a given number of accumulated signals were Monte Carlo simulated based on the probabilities derived from the row corresponding to the true molecule in the confusion matrix shown in Fig. 6e (main text). For example, in the case of a simulation with 10 signals for Ser, the occurrence probabilities for Ser, pSer, Tyr, pTyr, Thr, and pThr were 59.1%, 5.0%, 10.1%, 11.3%, 4.1%, and 7.4%, respectively. Using these probabilities, 10 predicted signals were generated by random sampling. The same procedure was applied to simulations for the other molecules. For a given number of signals, the molecule with the highest frequency among the predicted results was regarded as the inferred true molecule. In the implementation, 100,000 trials were performed for each signal number and molecule, and the classification accuracy for multiple accumulated signals was calculated as the proportion of correct predictions. When performing majority voting in each trial, if the number of predictions for the true molecule was tied with that of one or more incorrect molecules, the contribution was divided by the number of tied molecular species before accumulation.

The average classification accuracy reported in the main text was calculated as the arithmetic mean of the probabilities obtained for the six molecules.