Supporting Information

Compact CLIP Model: Predicting Spectral Properties of AgNCs Using DNA Template

Xun Zhang‡[a], Huiting Wang‡[b], Xin Liu[a], Xiaokang Zhang[a], Shuang Cui[a], Yao Yao[a], Bin Wang[c] and Qiang Zhang[a]*

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

[b] Allied Health Department, Osaka University, Osaka, 565-0871, Japan.

[c] Key Laboratory of Advanced Design and Intelligent Computing, Dalian University, Dalian, 116622, China.

‡ These authors contribute equally to this work

Correspondence: zhangq@dlut.edu.cn (Q.Z.); Tel: +86 0411 87402106;

Contents:

Table S1. DNA Sequences Design and Modifications

Table S2. DNA Sequences information for the AgNCs database

Table S3. The prediction accuracy of the compact CLIP model under different batch sizes

Table S4. The prediction accuracy of the compact CLIP model under different test set split ratios

Table S5. Comparison of the prediction accuracy of the spectral properties of AgNCs

Figure S1. Eliminate the background noise fluorescence caused by the sample plate

Figure S2. The relationship between the proportion of bases in the DNA template and the position of fluorescence peaks

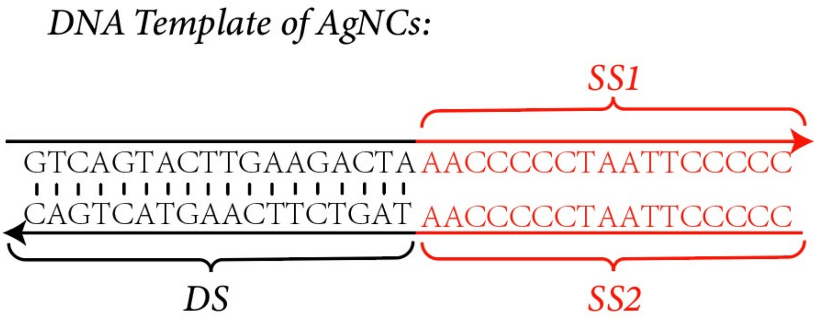
Figure S3. Real-time fluorescence data of single-layer molecular perceptron with different weight parameters

Table S1. DNA Sequences information for dsDNA and ssDNA research

The DNA template sequence information required for the study of the impact of spectral signals on ssDNA and dsDNA sections is as follows. The black part represents the double-stranded sequence, and the red part represents the overhang region sequence.

Oligonucleotides	Sequences (5 '-3 ')
Ds-Base:	GTCAGTACTTGAAGACTAAACCCCCTAATTCCCCC CCCCCTTAATCCCCCAATAGTCTTCAAGTACTGAC
DS-Base-23nt	GTCAGTACTTGAAGATCAGTCTAAACCCCCTAATTCCCCC CCCCCTTAATCCCCCAATAGACTGATCTTCAAGTACTGAC
DS-Base-28nt	GTCAGTACTTGAAGATCAGTCTAAACCCCCTAATTCCCCC CCCCCTTAATCCCCCAATAGACTGATCTTCAAGTACTGAC
DS-Base-33nt	GTCAGAGTTCTACTTACGTAGAAGATCAGTCTAAACCCCCTAATTCCCCC CCCCCTTAATCCCCCAATAGACTGATCTTCTACGTAAGTAGAACTCTGAC
Ss-5’-0nt	GTCAGTACTTGAAGACTAAACCCCCTAATTCCCCC TAGTCTTCAAGTACTGAC
Ss-5’-2nt	GTCAGTACTTGAAGACTAAACCCCCTAATTCCCCC AATAGTCTTCAAGTACTGAC
Ss-5’-7nt	GTCAGTACTTGAAGACTAAACCCCCTAATTCCCCC CCCCCAATAGTCTTCAAGTACTGAC
Ss-5’-12nt	GTCAGTACTTGAAGACTAAACCCCCTAATTCCCCC TTAATCCCCCAATAGTCTTCAAGTACTGAC
Ss-3’-0nt	GTCAGTACTTGAAGACTA CCCCCTTAATCCCCCAATAGTCTTCAAGTACTGAC
Ss-3’-2nt	GTCAGTACTTGAAGACTAAA CCCCCTTAATCCCCCAATAGTCTTCAAGTACTGAC
Ss-3’-7nt	GTCAGTACTTGAAGACTAAACCCCC CCCCCTTAATCCCCCAATAGTCTTCAAGTACTGAC
Ss-3’-12nt	GTCAGTACTTGAAGACTAAACCCCCTAATT CCCCCTTAATCCCCCAATAGTCTTCAAGTACTGAC

Table S2. DNA Sequences information for the AgNCs database


DNA Template of AgNCs:
SS1
GTCAGTACTTGAAGACTA AACCCCCTAATTCCCCC
CAGTCATGAACTTCTGAT AACCCCCTAATTCCCCC
DS
SS2

Maintain the DNA double-stranded sequence as “GTCAGTACTTGAAGACTA-TAGTCTTCAAGTACTGAC” unchanged, and alter the sequence information of SS1 and SS2 in the FW1-62 group sequences to conduct orthogonal experiments. The fluorescence heatmap and data between the DNA text sequences can be referred to in the database on GitHub.

Oligonucleotides	Sequences (5 '-3 ')
FW1	CCCCCCCCCCCCCCCCCCCC
FW2	CACCCCCCCCCCCCCCCGTC
FW3	GCATTATCCCCACCCCTCCC
FW4	TTATCAGCGCCTCGACCTTA
FW5	CGCTTCACTATGCGCTTATA
FW6	TCTATACTCAGGCGTCGAGT
FW7	AGCTAATCGCGAGGGGACCG
FW8	ATTGAAGTTCGGATGCACCG
FW9	GCTCACTTTTTTGTATTGTA
FW10	TATATGGGTTAATTTTTGGA
FW11	CCCCCCCCCCCCCCCCC
FW12	ATCCCCCCCCCCCCCCC
FW13	CCTCCCCCCAGGACCCC
FW14	AACCCCCTAATTCCCCC
FW15	ATCGCCAGGCTACCCTA
FW16	CCTCAAAATGCTCCTGG
FW17	TGCCCGAAAGTTAGACC
FW18	GAGTACCAATTGCTCAT
FW19	AGATTTGTAAATTAGCG

FW20	AGATGTATTGATTATTA
FW21	CCCCCCCCCCCCCCCC
FW22	CTACCCCCCCCCCCCC
FW23	ATGCCCTCCCTTGCCC
FW24	CTCTACGCCGCAGCGA
FW25	CAGCGGAGAGCGACCC
FW26	TCTCTTCCGCAAATTT
FW27	AGAGACCACAATGTGC
FW28	TATCGCGAGGGCAAGT
FW29	CCCAGGATGTTGATAG
FW30	ATGAGAGATATGAGAT
FW31	CCCCCCCCCCCCCCC
FW32	GGGCCCCTACCCCCC
FW33	CCTTATCCCTACCTG
FW34	TCCAGAATTCCCAGT
FW35	TCATTAACCGGCATC
FW36	GGTAGTTCCCACATA
FW37	GCGGTTCCGTTTGCT
FW38	AGTGTTTCCTTAGAC
FW39	GGTGGTGCGGTTCTG
FW40	AAAGTTGGGATTGTA
FW41	CCCCCCCCCCCC
FW42	AACCCGCCCCCC
FW43	TACCCTTGTCCT
FW44	CAGGTAGCACAA
FW45	AATTTCAGCTCT
FW46	TGGTACATTTGT
FW47	TTAAGCCGTGAG
FW48	GCCTCAGGAAAG
FW49	GGGTTGTTCATC
FW50	ATAGTTTGGTAA

FW51	GTACGTTAGCC
FW52	GGGCACCTAG
FW53	CAAAGATGC
FW54	GTCGGCGC
FW55	TGACTAG
FW56	TCTGTT
FW57	TAGGT
FW58	TGGC
FW59	CCG
FW60	TG
FW61	T
FW62	Null

Table S3. The prediction accuracy of the compact CLIP model under different batch sizes

Batch size	Training loss	Training accuracy	Test loss	Test accuracy	Random	Multiplication factor
8	0.0631	0.9740	0.1634	0.7298	0.0470	15.53
16	0.0711	0.9671	0.1450	0.5766	0.0330	17.47
32	0.0887	0.9586	0.1128	0.4485	0.0282	15.90
64	0.1347	0.9394	0.0828	0.3370	0.0292	11.54

For the method of using 32 samples as a batch, we have supplemented gradient tests for batch size, testing batch sizes of 8, 16, 32, and 64. The results obtained have been filled into Supplementary Material Table S3, as shown in the table below. As long as the batch size is controlled within the range of 8-32, the accuracy has increased by more than 15 times compared to random prediction (Multiplication factor).

Table S4. The prediction accuracy of the compact CLIP model under different test set split ratios

The size ratio of test and train set	Training loss	Training accuracy	Test loss	Test accuracy	Peak position accuracy	Peak intensity accuracy
1:9	0.0887	0.9586	0.1128	0.4485	0.80	0.68
2:8	0.0802	0.9490	0.1306	0.4209	0.80	0.67
3:7	0.0779	0.9450	0.1320	0.4197	0.73	0.67

Regarding the proportion of the test set, we have also supplemented gradient tests. We tested the compact CLIP model with test-to-training set ratios of 1:9, 2:8, and 3:7. The accuracy results obtained have been filled into Supplementary Material Table S4, as shown in the table below. It can be observed that by changing the proportions of the test and training sets, the precision of our spectral peak and fluorescence intensity predictions did not change significantly. To further improve precision, we may need to establish data standardization methods between different fluorescence databases in the future to expand the scale of the database and address this issue.

Table S5. Comparison of the prediction accuracy of the spectral properties of AgNCs

Model	Prediction Results	
	Peak position accuracy (color)	Peak intensity accuracy
Compact CLIP	0.68	0.80
RNN	0.56	/
SVM	/	0.80
MERCI	0.90	/

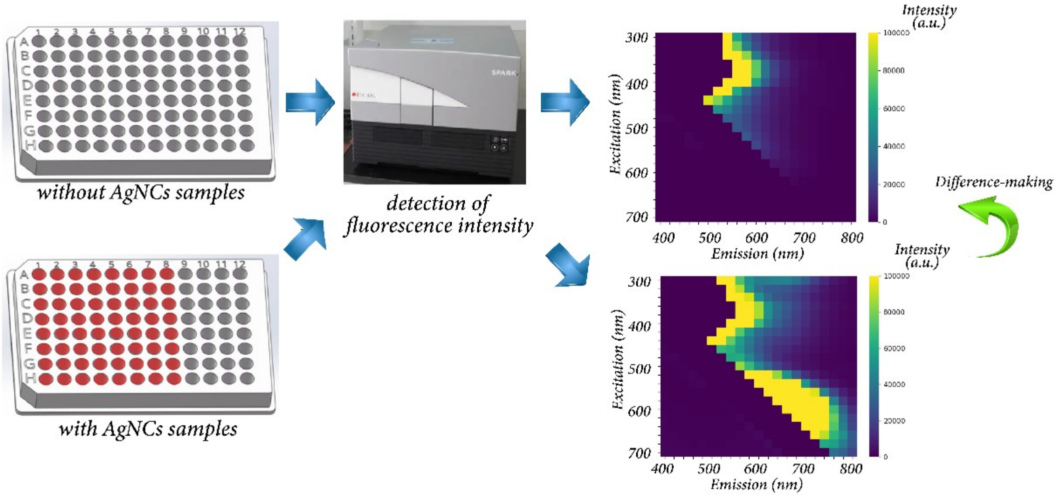
We reviewed relevant papers on the prediction of AgNCs spectra and compared our compact CLIP model with RNN[1], SVM[2], and MERCI[3]. Unlike traditional classification of AgNCs spectra based on color and fluorescence intensity, the advantage of compact CLIP lies in its ability to simultaneously predict the numerical values of fluorescence peak positions and intensities, achieving relatively good accuracy.

[1] Zhai, F., Guan, Y., Li, Y., Chen, S., & He, R. Predicting the fluorescence properties of hairpin-DNA-templated silver nanoclusters via deep learning. ACS Applied Nano Materials.2022, 5(7), 9615-9624.

[2] Copp, S. M., Bogdanov, P., Debord, M., Singh, A., & Gwinn, E. Base motif recognition and design of DNA templates for fluorescent silver clusters by machine learning. Advanced Materials. 2014, (33), 5839-5845.

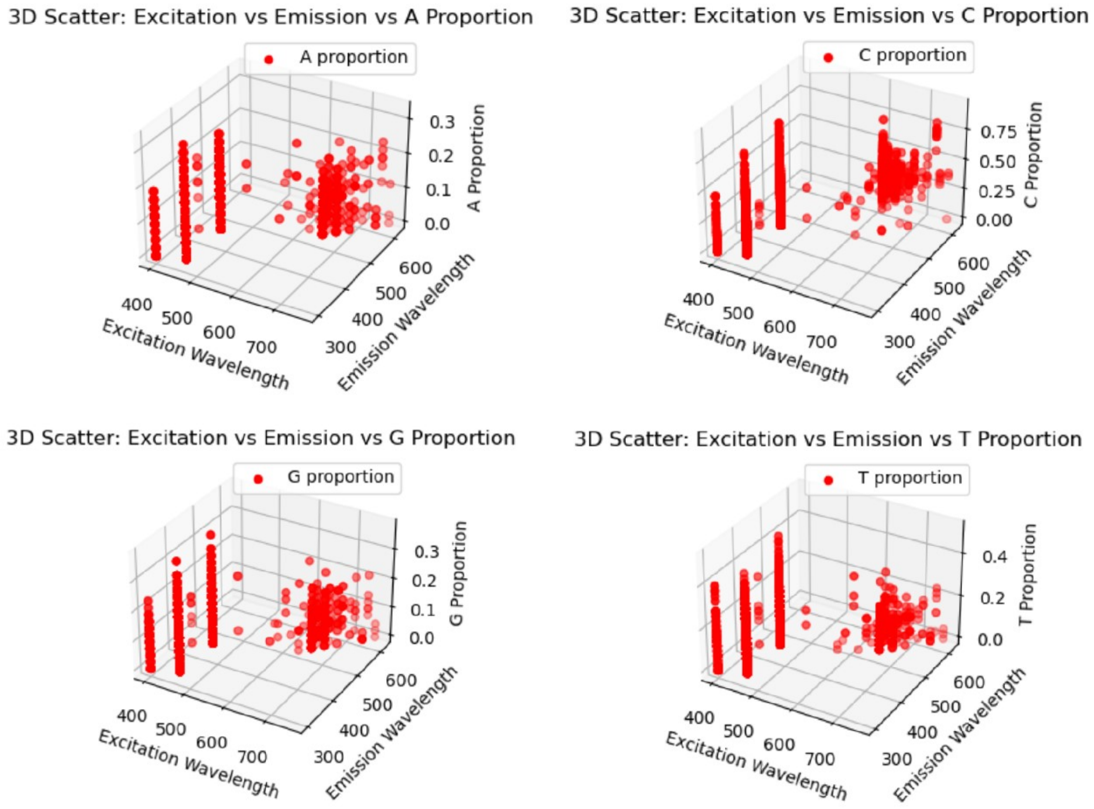
[3] Copp, S. M., Gorovits, A., Swasey, S. M., Gudibandi, S., Bogdanov, P., & Gwinn, E. G. Fluorescence color by data-driven design of genomic silver clusters. ACS nano. 2018, 12(8), 8240-8247.

Figure S1. Eliminate the background noise fluorescence caused by the sample plate



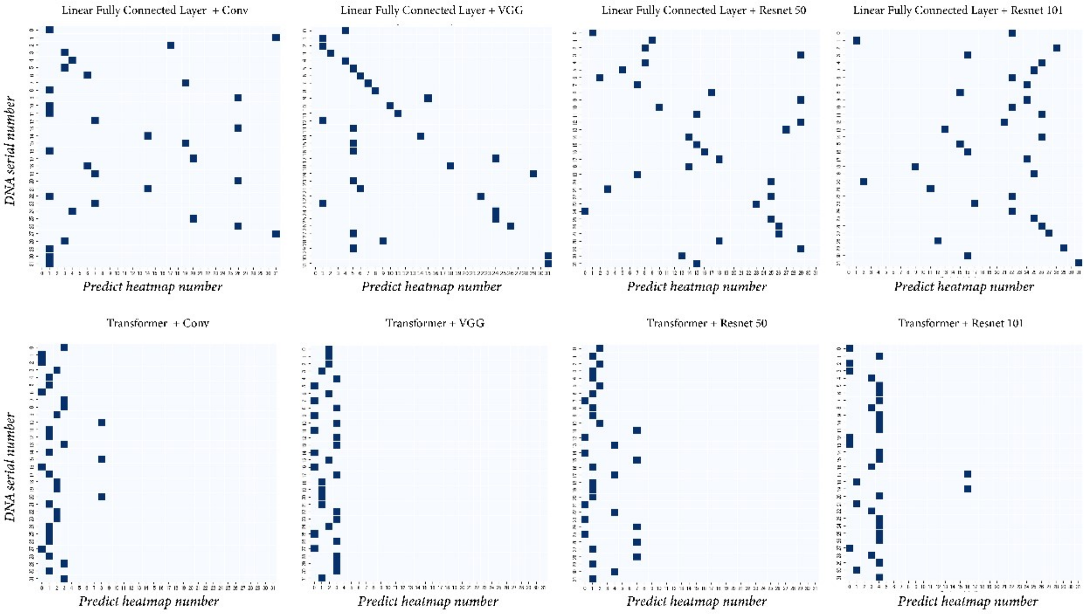
We use a 96-well sample plate as a carrier and often encounter background noise when using a Tecan microplate reader. To obtain spectral signals that are solely produced by the samples themselves, we employ a differential method to process the data. Specifically, we subtract the fluorescence of the sample plate containing buffer solution from the fluorescence of the sample plate containing AgNCs. The final difference is used as the spectral data collected from our samples.

Figure S2. The relationship between the proportion of bases in the DNA template and the position of fluorescence peaks



Plotting the EM-EX positions of the fluorescence peaks and the proportions of the four bases from 3844 sets of data in a 3D scatter plot reveals no apparent relationship between the base proportions and the peak positions.

Figure S3. Confusion matrix of the compact CLIP model



In the compact CLIP model, the use of a linear fully connected layer for the text encoder outperforms the transformer in prediction accuracy, showing a significant improvement in accuracy on the test set, with most predicted points distributed along the main diagonal of the confusion matrix. The primary reason might be that the DNA sequence texts and fluorescence heatmaps in the AgNCs database we used are not complex, so the attention mechanism carried by the transformer does not bring significant performance improvement. On the contrary, its large hyperparameter space may reduce the model's learning ability.