Classification of Semiconductors by Photoluminescence Spectroscopy and Machine Learning

YINCHUAN YU,[1] MATTHEW D. MCCLUSKEY[1,*]

[1]*Department of Physics and Astronomy, Washington State University, Pullman, WA 99164-2814, USA*

**mattmcc@wsu.edu*

ORCID iD (McCluskey): 0000-0002-0786-4106

Abstract

Photoluminescence (PL) spectroscopy is a non-destructive optical method that is widely used to characterize semiconductors. In the PL process, a substance absorbs photons and emits light with longer wavelengths via electronic transitions. This paper discusses a method for identifying substances from their PL spectra using machine learning, a technique that is efficient in making classifications. Neural networks were constructed by taking simulated PL spectra as the input and the identity of the substance as the output. In this paper, six different semiconductors were chosen as categories: gallium oxide (Ga_2O_3), zinc oxide (ZnO), gallium nitride (GaN), cadmium sulfide (CdS), tungsten disulfide (WS_2) and cesium lead bromide ($CsPbBr_3$). The developed algorithm has a high accuracy (>90%) for assigning a substance to one of these six categories from its PL spectrum and correctly identified a mixed Ga_2O_3/ZnO sample.

Introduction

Optical spectroscopy is a fast, nondestructive method for analyzing chemical compounds. Absorption or emission spectra provide a "fingerprint" of the material as well as its defects. Typical wavelengths range from the UV to near-IR (200 to 1100 nm). When coupled with a microscope, spectra can be obtained with high spatial resolution. Types of optical spectroscopy include transmission, reflection, fluorescence, and Raman scattering.[1]

The interpretation of spectra traditionally relies on the user's experience. A senior researcher might say, "This fluorescence spectrum has a broad green band and a sharp peak in the near-UV. That looks like zinc oxide, or maybe titania." That statement contains several concepts, including "spectrum," "green band," and "UV peak." These concepts led the researcher to assign high probabilities to zinc oxide and titania. Students or postdoctoral researchers operate in an apprentice mode and gradually become familiar with a set of spectral features relevant to their research.

In contrast to mass spectrometry and Raman spectroscopy databases,[2] there are few tools available to identify fluorescence (or photoluminescence, PL) spectra of inorganic compounds. The main problem is that room-temperature PL spectra typically consist of broad, overlapping bands. The intensity and peak positions depend on excitation intensity, wavelength, and defects in the sample. Peak fitting or principal component analysis (PCA)[3] may not be optimal for such spectra. Machine learning (ML), which can readily group images or other datasets into general categories,[4] has the potential to speed up identification of unknown substances. It can also provide a way to

organize large collections of spectra, a valuable aid for researchers in chemistry, physics, and materials science. The ML approach is rapid and can be scaled up to an arbitrary number of spectra.

In the present work, a neural network was developed to identify PL spectra. Hornik[5] showed that multilayer feed-forward architecture gives neural networks the potential of being universal approximators. Yarotsky[6] proved that deep Rectified Linear Unit (ReLU)[7] networks more efficiently approximate functions than shallow networks. We may assume there is a function for which the input is the PL spectrum and the output is the substance category. Then the task is to build a deep ReLU neural network to provide this function.

Methods

Simulated PL Spectra

The ML algorithm needs a large amount of data to get trained. To provide enough training data, simulated PL spectra were used. In general, the PL spectrum of substance has several emission bands, some of which are intrinsic features while others are caused by defects. To generate simulated PL data, the emission band wavelengths, and their relative strength and width, need to be described. In general, peaks are asymmetric, whether plotted as a function of wavelength or energy, due to factors such as phonon replicas and band-tail states. In this work, the bigaussian function is used to empirically describe each emission peak:

$$I = \begin{cases} A \exp\left(\frac{-(x-x_0)^2}{2{\sigma_1}^2}\right) & x < x_0 \\ A \exp\left(\frac{-(x-x_0)^2}{2{\sigma_2}^2}\right) & x > x_0 \end{cases} \quad (1)$$

where A is the relative peak height, x_0 is the peak center position, σ_1 and σ_2 are the left and right standard deviations that describe the asymmetric width of the peak.

An important example is gallium oxide (β-Ga_2O_3), an emerging material for power applications.[8] Because of the variability of defect populations, PL spectra from different samples show significant variations. The PL of β-Ga_2O_3 typically has broad peaks that fall into four categories. The first one is the UV band, which is caused by recombination of an electron with a self-trapped hole (STH). The peak center is located near 360 nm. A second peak near 400 nm (UV') and a third peak near 430 nm (blue) are due to donor-acceptor pair (DAP) transitions. The last peak at about 520 nm (green), also assigned to a DAP transition, is correlated with high O_2 pressure during the growth. From these peak positions, and randomized peak intensities and widths, simulated Ga_2O_3 PL spectra were generated. Four examples are shown in Fig. 1. Gaussian noise (mean 0, standard deviation 0.01) was included in the simulated spectra.

<Insert Figure 1 here>

The same idea was applied to other semiconductors with known emission properties. We restricted the data to PL spectra collected at room temperature. For ZnO, there is usually a UV peak at 378 nm due to free-exciton recombination and a broad green band near 530 nm due to defects.[9] GaN typically has a peak near 365 nm due to a band–to–band transition, and a small peak near 382 nm due to recombination of an electron in the conduction band with a hole bound to an acceptor.[10] Two broad yellow

peaks near 482 nm and 525 nm caused by carbon-related defects are also observed[11]. The main peak of CdS,[12] WS_2,[13] and $CsPbBr_3$,[14] at 508 nm, 630 nm, and 523 nm respectively, are due to exciton recombination.

Machine Learning Algorithm

Simulated and real PL spectra are represented by column vectors. We focused on the PL spectrum between 250 nm to 649 nm with 1 nm as the interval, resulting in an $n = 400$ element vector. For spectra with less range, zero padding was used, where the intensity to was set 0 where values were not reported. To normalize the spectra, the vector was divided by its magnitude.

$$x = \frac{v}{\|v\|} \tag{2}$$

where x is the unit vector and provides the input for the training networks.

The output y is a vector, the dimension of which is the number of classes of substance used. The value of the element at the position of the known substance index is 1, and the rest are all 0. When m PL spectra are stacked one by one, a $n \times m$ matrix X is formed. Since the output target is already known, this ML approach is supervised learning, and a neural network[15] is applied. There are two hidden layers. Each element in a layer is calculated by a linear combination of all elements in the previous layer plus a bias value, followed by a nonlinear activation function σ:

$$Z_l = W_l \cdot a_{l-1} + b_l \tag{3}$$

$$a_l = \sigma(Z_l) \tag{4}$$

where a_l is the vector of the l^{th} layer, W_l is a parameter matrix with rows equal to the number of elements in layer l and columns equal to the number in layer l–1, and b_l is a bias vector with same number of elements as layer l.

<Insert Figure 2 here>

The ReLU function is used as the activation function of the first two layers and a SoftMax function is used as the activation function for the final layer:

$$\text{ReLU: } \sigma(x) = \max(0, x) \tag{5}$$

$$\text{SoftMax: } \sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \quad \text{for } i = 1, 2, 3, \ldots, k \tag{6}$$

The final layer is treated as the probability of the input belonging to each class; i.e., a prediction. For a known substance, the ideal output layer should be 1 on the corresponding class element and 0 for all other elements. We denote the output layer vector $\tilde{y}$ and the vector that represents the real class y. Both $\tilde{y}$ and y are vectors, the dimension of which is equal to the total number of classes. The Loss function describes how close the prediction $\tilde{y}$ is to the target y:

$$L = -\sum_{i=1}^{k} (y_i \ln(\tilde{y}_i)) \tag{7}$$

A small L value indicates a good prediction, with a perfect prediction giving $L = 0$. Averaging over all m training data, the total cost is

$$J = \frac{1}{m} \sum_{j=1}^{m} L_j \tag{8}$$

All W and b parameters are randomly generated at the beginning, so the predictions are far from the target and the cost is high. The method of gradient decent[16] is

applied to push the Cost to 0. All parameters θ (W, b) are updated at the same time following the equation:

$$\theta := \theta - \alpha \cdot \nabla_{\theta} J \tag{9}$$

where α is the learning rate and needs to be set to an appropriately small value. Generating the Cost from the most recent parameters is a forward propagation. Updating all parameters from the most recent Cost is a backward propagation. One forward and backward propagation is one iteration of the training process. After enough iterations, the Cost converges to a constant, small value, and the parameters can be used to predict the class of a substance.

Results and Discussion

Each substance was represented by 1000 simulated spectra using bigaussian functions with the parameters shown in Table 1. The learning rate was set to be 0.01, the first hidden layer contained 50 elements, and second hidden layer contained 20 elements. After 20,000 iterations, the parameters were trained and could be used in predictions. A plot of the Cost as the function of number of iterations is shown in Fig. 3.

<Insert Figure 3 here>

Substance	Peak Position (nm)	Relative Intensity	Width (nm)
Ga_2O_3	360 400 430 520	0 – 5 0 – 5 0 – 5 0 – 1	30 – 50 40 – 60 50 – 100 25 - 100
ZnO	378 530	4 – 12 1 – 12	5 – 10 30 – 60
GaN	365 382 525	6 – 16 1 – 2 1 – 12	5 - 10 6 – 15 25 – 50
CdS	508	1	5 – 10
WS_2	630	1	8 - 9
$CsPbBr_3$	523	1	10 – 30

Table 1. Parameters used in generating simulated PL spectra.

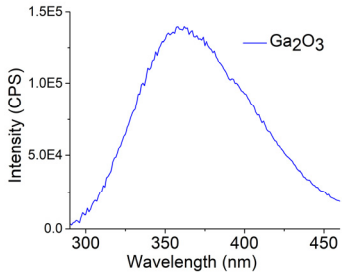
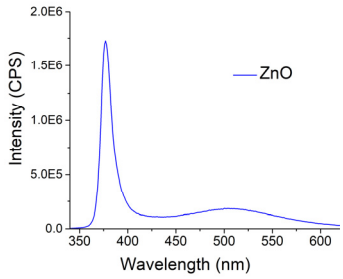
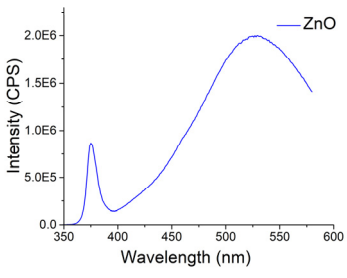
After the training process, the next task is to test how these parameters predict a new, unknown PL spectrum. The test is separated into two parts. The first part is to test with the PL spectra taken from literature, and the second part is to test with the PL collected in our lab using a Horiba Fluorolog spectrometer. By searching for papers with PL spectra, then digitalizing the PL with the GetData Graph[17] program, the PL data were collected. Applying the parameters obtained from the training process, the prediction of the PL spectra collected from literature approached 100% accuracy. Figs. 4-8 show PL spectra collected from the literature. In the Supplementary Material, Fig. S1 shows spectra of WS_2 from Gutierrez et al.,[13] Shi et al.,[18] and He et al.[19] and Fig. S2 shows spectra of $CsPbBr_3$ from Pourdavoud et al.[14] and Liu et al.[20]

<Insert Figure 4 here>

<Insert Figure 5 here>

<Insert Figure 6 here>

<Insert Figure 7 here>

Category →	Ga_2O_3	ZnO	GaN	CdS	WS_2	$CsPbBr_3$
	**98.91%**	0.09%	0.96%	0.02%	0.00%	0.02%
	0.63%	**99.29%**	0.00%	0.00%	0.00%	0.07%
	0.19%	**92.03%**	7.18%	0.00%	0.55%	0.04%

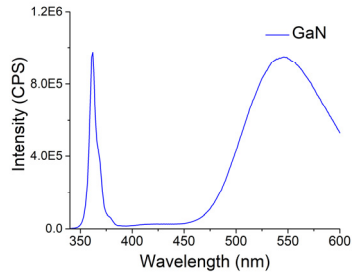
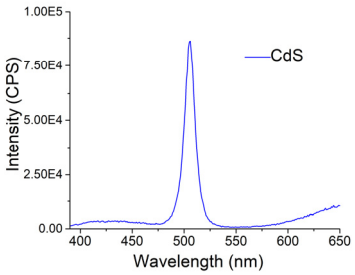
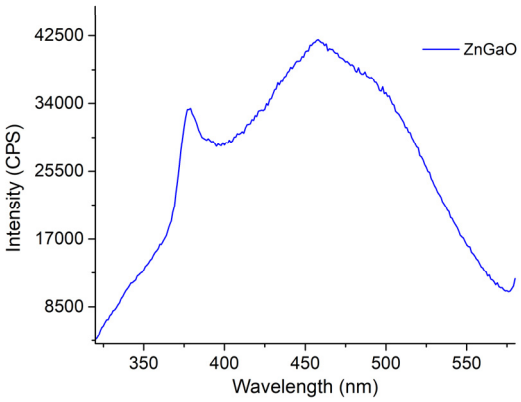
	0.02%	4.15%	**95.59%**	0.00%	0.24%	0.00%
	0.13%	0.02%	0.00%	**99.33%**	0.01%	0.51%

Table 2. PL spectra collected from Horiba Fluorolog spectrometer and predictions given by the ML algorithm.

The second part was to use machine learning to identify semiconductors measured in our lab. The following excitation wavelengths were used: 270 nm (Ga_2O_3), 320 nm (ZnO #1 and GaN), 310 nm (ZnO #2), and 370 nm (CdS). Table 2 shows the plots of the PL spectra and the suggestions (predictions) given by the ML algorithm. The predictions include the percentage for the sample belonging to each category. As shown in bold, the percentages for correct categories exceeded 90%. To further test the ML algorithm, we performed a PL measurement on a Zn-doped Ga_2O_3 sample. The PL spectrum has signatures from Ga_2O_3 and ZnO.



Ga_2O_3	ZnO	GaN	CdS	WS_2	$CsPbBr_3$
61.09%	26.09%	12.42%	0.05%	0.05%	0.30%

Table 3. PL of Zn doped Ga_2O_3 collected from Horiba Fluorolog spectrometer and the prediction given by the neural networks.

From Table 3, there are three categories that show percentages above 10%. The prediction shows the probability of Ga_2O_3 is 61.09%, which is the highest. This is expected because the main constituent of the sample is Ga_2O_3. The probability for ZnO is 26.09%, which is the second highest. This result is sensible because the Zn-doped sample also shows the PL signature of phase-separated ZnO (the 378 nm peak). Hence, ML can be used to identify mixtures of materials.

Conclusions

PL spectroscopy combined with machine learning is a potential method for identifying unknown substances or providing an organization scheme for a large number of spectra. The forward network approach enables the user to assign probabilities to a substance belonging to different categories depending on its PL spectrum. Due to the relative lack of PL spectra in the literature, simulated PL spectra must be used to train the ML algorithm. Although only six categories were used in this preliminary study, the number of categories can be readily scaled up to provide suggestions for more substances and differentiate between substances with very similar PL spectra. It is straightforward to create subcategories such as carbon-doped GaN, *n*-type GaN, etc., to provide precise identification.

Acknowledgments

The authors gratefully acknowledge Cassandra Remple and Richard Lytel for experimental assistance and helpful discussions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Research was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Materials Science and Engineering under Award No. DE-FG02-07ER46386.

Figure Captions

Fig. 1 Simulated PL spectra of Ga_2O_3 with randomized peak intensity and width.

Fig. 2. Schematic diagram of the neural network used in this project.

Fig. 3. Plot of the Cost as a function of iterations. The learning rate is 0.01 and 20,000 iterations were processed.

Fig. 4. PL spectra of Ga_2O_3 from McCluskey,[8] Liu et al.,[21] Kumar,[22] and Hou et al.[23]

Fig. 5. PL spectra of ZnO from Lyu et al.,[24] Rui Zhang et al.,[9] and X. T. Zhang et al.[25]

Fig. 6. PL spectra of GaN from Balkas et al.,[26] Chen et al.,[27] Chakrapani et al.,[28] and Freitas et al.[10]

Fig. 7. PL spectra of CdS from Hoang et al.,[12] Hu et al.,[29] and Ravindran et al.[30]

1. C. Burgess. “OPTICAL SPECTROSCOPY | Radiation Sources”. In: P. Worsfold, C. Poole, A. Townshend, M. Miró, editors. Encyclopedia of Analytical Science (Third Edition). Academic Press, Oxford, 2005. Pp. 103–109. 10.1016/B978-0-08-101983-2.00429-1.
2. P. Vandenabeele, H.G.M. Edwards, L. Moens. “A Decade of Raman Spectroscopy in Art and Archaeology”. Chem. Rev. American Chemical Society, 2007. 107(3): 675–686. 10.1021/cr068036i.
3. J.R. Beattie, F.W.L. Esmonde-White. “Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra”. Appl. Spectrosc. SAGE Publications Ltd STM, 2021. 75(4): 361–375. 10.1177/0003702820987847.
4. S.B. Kotsiantis, I.D. Zaharakis, P.E. Pintelas. “Machine learning: a review of classification and combining techniques”. Artif. Intell. Rev. 2006. 26(3): 159–190. 10.1007/s10462-007-9052-3.
5. K. Hornik. “Approximation capabilities of multilayer feedforward networks”. Neural Netw. 1991. 4(2): 251–257. 10.1016/0893-6080(91)90009-T.
6. D. Yarotsky. “Error bounds for approximations with deep ReLU networks”. Neural Netw. 2017. 94: 103–114. 10.1016/j.neunet.2017.07.002.
7. A.F. Agarap. “Deep Learning using Rectified Linear Units (ReLU)”. ArXiv180308375 Cs Stat. 2019.
8. M.D. McCluskey. “Point defects in Ga2O3”. J. Appl. Phys. 2020. 127(10): 101101. 10.1063/1.5142195.
9. R. Zhang, P.-G. Yin, N. Wang, L. Guo. “Photoluminescence and Raman scattering of ZnO nanorods”. Solid State Sci. 2009. 11(4): 865–869. 10.1016/j.solidstatesciences.2008.10.016.
10. J.A. Freitas, J.G. Tischler, N.Y. Garces, B.N. Feigelson. “Optical probing of low-pressure solution grown GaN crystal properties”. J. Cryst. Growth. 2010. 312(18): 2564–2568. 10.1016/j.jcrysgro.2010.04.013.
11. M.A. Reshchikov, J.D. McNamara, H. Helava, A. Usikov, Y. Makarov. “Two yellow luminescence bands in undoped GaN”. Sci. Rep. Nature Publishing Group, 2018. 8(1): 8091. 10.1038/s41598-018-26354-z.

12. T.B. Hoang, L.V. Titova, H.E. Jackson, L.M. Smith, J.M. Yarrison-Rice, J.L. Lensch, et al. “Temperature dependent photoluminescence of single CdS nanowires”. Appl. Phys. Lett. American Institute of Physics, 2006. 89(12): 123123. 10.1063/1.2357003.

13. H.R. Gutiérrez, N. Perea-López, A.L. Elías, A. Berkdemir, B. Wang, R. Lv, et al. “Extraordinary Room-Temperature Photoluminescence in Triangular WS2 Monolayers”. Nano Lett. American Chemical Society, 2013. 13(8): 3447–3454. 10.1021/nl3026357.

14. N. Pourdavoud, T. Haeger, A. Mayer, P.J. Cegielski, A.L. Giesecke, R. Heiderhoff, et al. “Room-Temperature Stimulated Emission and Lasing in Recrystallized Cesium Lead Bromide Perovskite Thin Films”. Adv. Mater. 2019. 31(39): 1903717. https://doi.org/10.1002/adma.201903717.

15. S.-C. Wang. “Artificial Neural Network”. In: S.-C. Wang, editor. Interdisciplinary Computing in Java Programming. Springer US, Boston, MA, 2003. Pp. 81–100. 10.1007/978-1-4615-0377-4_5.

16. S. Ruder. “An overview of gradient descent optimization algorithms”. ArXiv160904747 Cs. 2017.

17. “Digitize graphs and plots - GetData Graph Digitizer - graph digitizing software”. n.d. http://getdata-graph-digitizer.com/index.php [Mar 11 2021].

18. W. Shi, M.-L. Lin, Q.-H. Tan, X.-F. Qiao, J. Zhang, P.-H. Tan. “Raman and photoluminescence spectra of two-dimensional nanocrystallites of monolayer WS 2 and WSe 2”. 2D Mater. IOP Publishing, 2016. 3(2): 025016. 10.1088/2053-1583/3/2/025016.

19. Z. He, X. Wang, W. Xu, Y. Zhou, Y. Sheng, Y. Rong, et al. “Revealing Defect-State Photoluminescence in Monolayer WS_2 by Cryogenic Laser Processing”. ACS Nano. 2016. 10(6): 5847–5855. 10.1021/acsnano.6b00714.

20. C. Li, Z. Zang, W. Chen, Z. Hu, X. Tang, W. Hu, et al. “Highly pure green light emission of perovskite $CsPbBr_3$ quantum dots and their application for green light-emitting diodes”. Opt. Express. Optical Society of America, 2016. 24(13): 15071–15078. 10.1364/OE.24.015071.

21. H. Liu, C. Xu, X. Pan, Z. Ye. “The Photoluminescence Properties of β-Ga2O3 Thin Films”. J. Electron. Mater. 2020. 49. 10.1007/s11664-020-08134-6.

22. S. Kumar. “Nanofunctional gallium oxide (Ga2O3) nanowires/nanostructures and their applications in nanodevices”. Phys. Status Solidi RRL - Rapid Res. Lett. 2013. 10.1002/pssr.201307253.

23. Y. Hou, L. Wu, X. Wang, Z. Ding, Z. Li, X. Fu. “Photocatalytic performance of α-, β-, and γ-Ga2O3 for the destruction of volatile aromatic pollutants in air”. J. Catal. 2007. 250(1): 12–18. 10.1016/j.jcat.2007.05.012.

24. S.C. Lyu, Y. Zhang, H. Ruh, H.-J. Lee, H.-W. Shim, E.-K. Suh, et al. “Low temperature growth and photoluminescence of well-aligned zinc oxide nanowires”. Chem. Phys. Lett. 2002. 363(1): 134–138. 10.1016/S0009-2614(02)01145-4.

25. X.T. Zhang, Y.C. Liu, J.Y. Zhang, Y.M. Lu, D.Z. Shen, X.W. Fan, et al. “Structure and photoluminescence of Mn-passivated nanocrystalline ZnO thin films”. J. Cryst. Growth. 2003. 254(1): 80–85. 10.1016/S0022-0248(03)01143-6.

26. C.M. Balkaş, Z. Sitar, L. Bergman, I.K. Shmagin, J.F. Muth, R. Kolbas, et al. “Growth and characterization of GaN single crystals”. J. Cryst. Growth. 2000. 208(1): 100–106. 10.1016/S0022-0248(99)00445-5.

27. C. Chen, S. Sun, M.M.C. Chou, K. Xie. “In situ inward epitaxial growth of bulk macroporous single crystals”. Nat. Commun. Nature Publishing Group, 2017. 8(1): 2178. 10.1038/s41467-017-02197-6.

28. V. Chakrapani, C. Pendyala, K. Kash, A.B. Anderson, M.K. Sunkara, J.C. Angus. “Electrochemical Pinning of the Fermi Level: Mediation of Photoluminescence from Gallium Nitride and Zinc Oxide”. J. Am. Chem. Soc. American Chemical Society, 2008. 130(39): 12944–12952. 10.1021/ja710999r.

29. C. Hu, X. Zeng, J. Cui, H. Chen, J. Lu. “Size Effects of Raman and Photoluminescence Spectra of CdS Nanobelts”. J. Phys. Chem. C. 2013. 117(40): 20998–21005. 10.1021/jp407272u.

30. T.R. Ravindran, A.K. Arora, B. Balamurugan, B.R. Mehta. “Inhomogeneous broadening in the photoluminescence spectrum of CdS nanoparticles”. Nanostructured Mater. 1999. 11(5): 603–609. 10.1016/S0965-9773(99)00346-3.

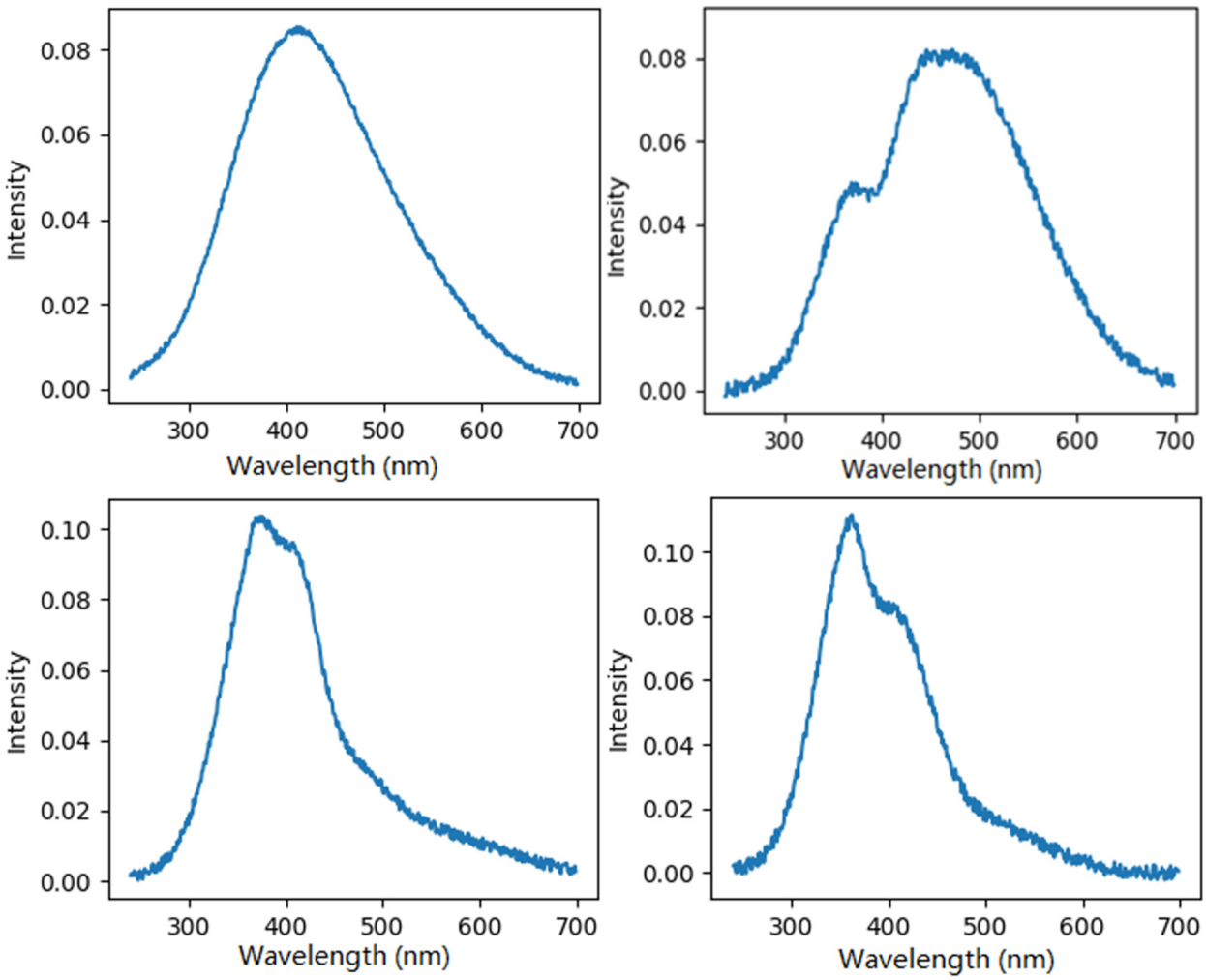

Intensity
Wavelength (nm)
0.00
0.02
0.04
0.06
0.08
0.10
300
400
500
600
700

Figure 1

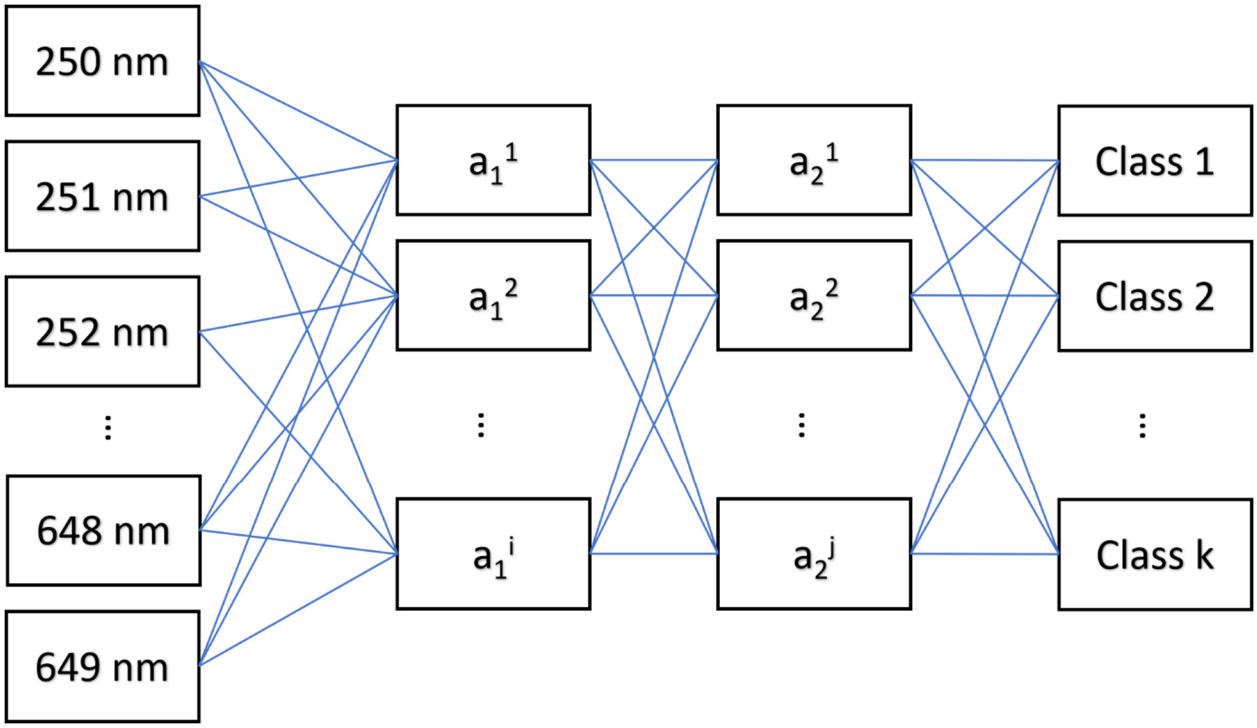

250 nm
251 nm
252 nm
⋮
648 nm
649 nm
a_1^1
a_1^2
⋮
a_1^i
a_2^1
a_2^2
⋮
a_2^j
Class 1
Class 2
⋮
Class k

Figure 2

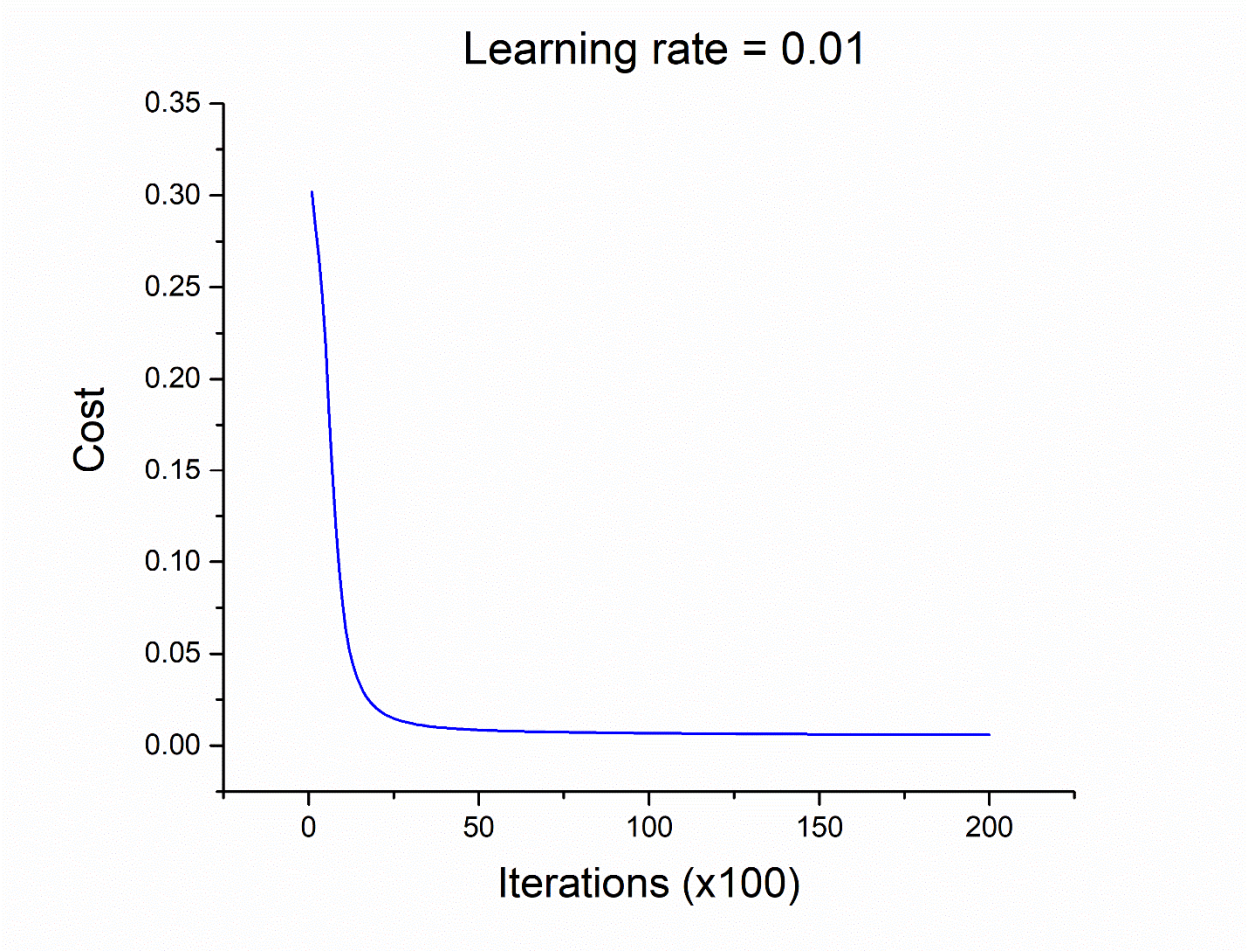

Learning rate = 0.01
Cost
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
Iterations (x100)

Figure 3

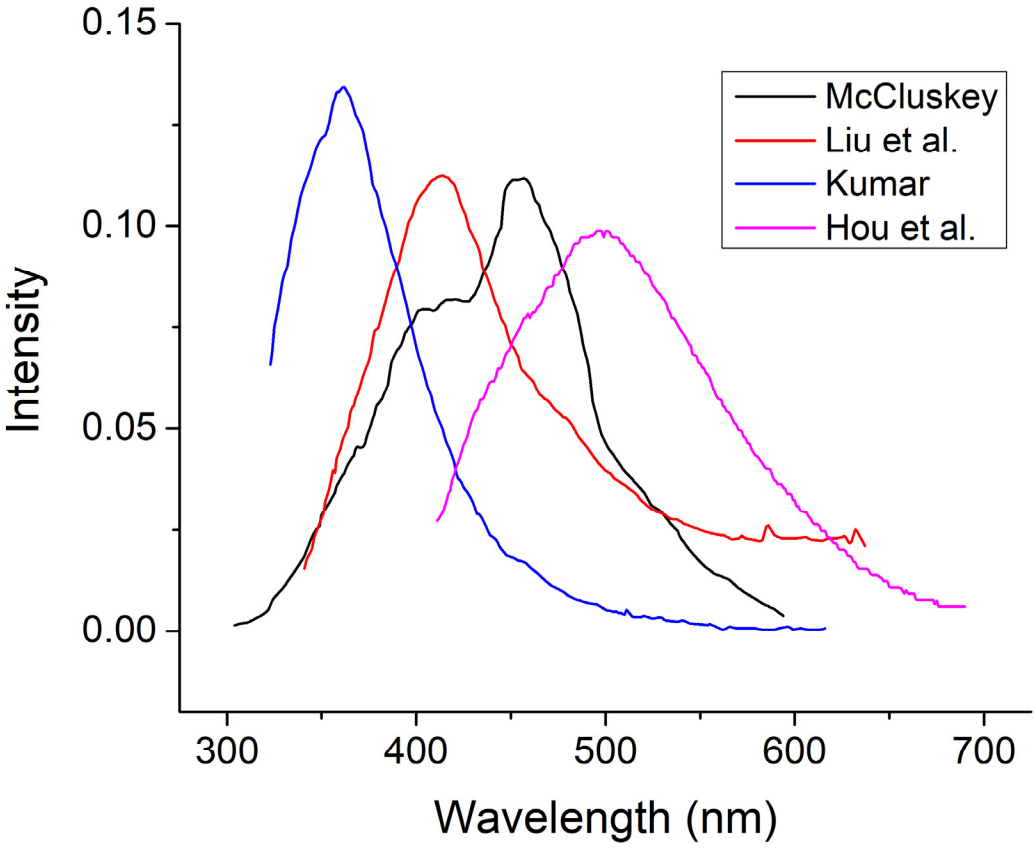

0.15
0.10
0.05
0.00
Intensity
300
400
500
600
700
Wavelength (nm)
McCluskey
Liu et al.
Kumar
Hou et al.

Figure 4

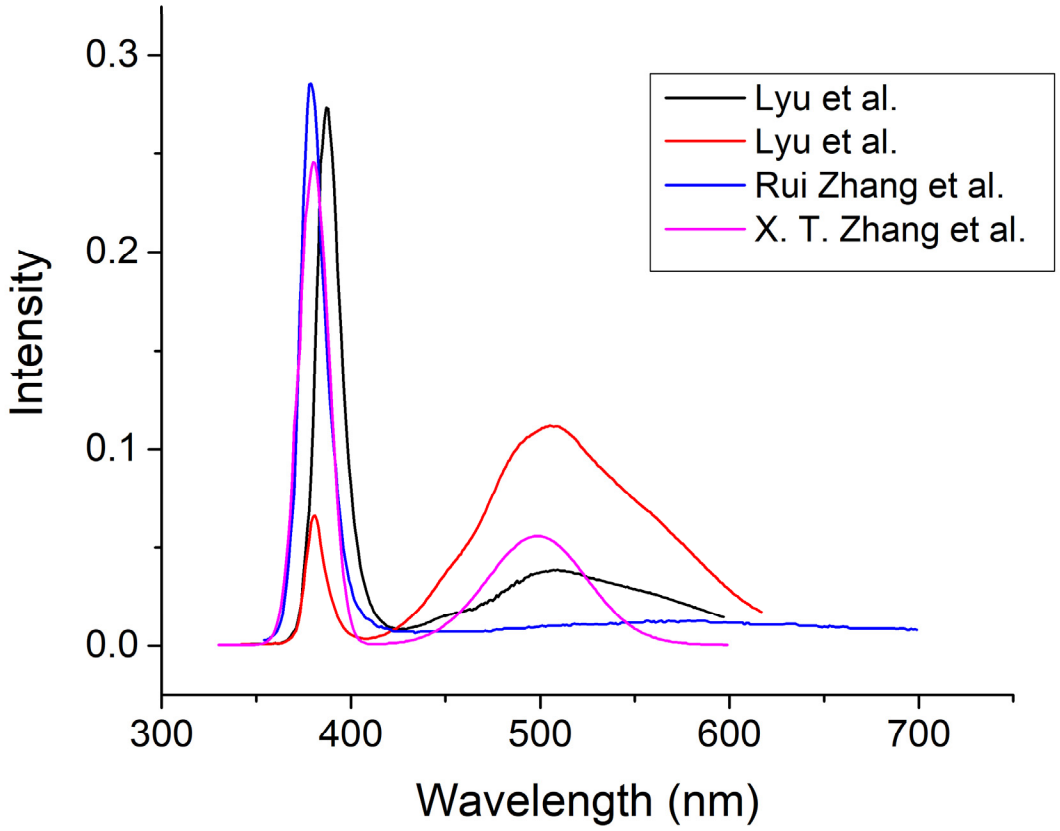

0.3
0.2
0.1
0.0
Intensity
300
400
500
600
700
Wavelength (nm)
Lyu et al.
Lyu et al.
Rui Zhang et al.
X. T. Zhang et al.

Figure 5

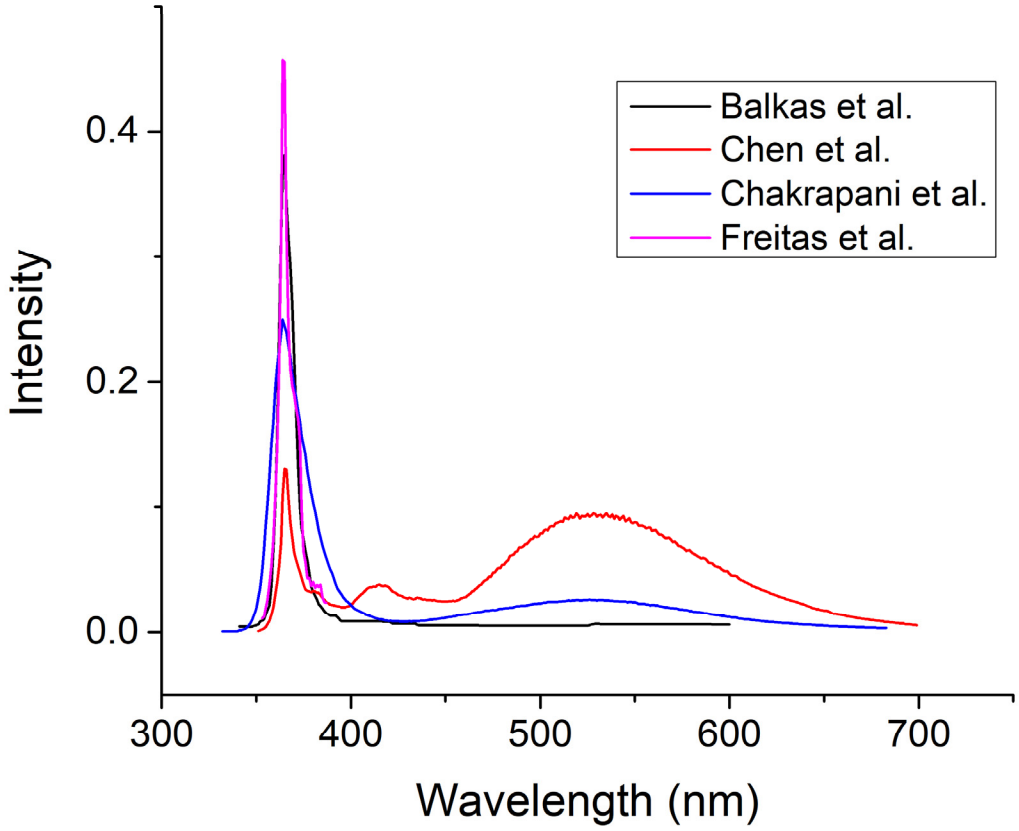

Balkas et al.
Chen et al.
Chakrapani et al.
Freitas et al.
Intensity
0.0
0.2
0.4
300
400
500
600
700
Wavelength (nm)

Figure 6

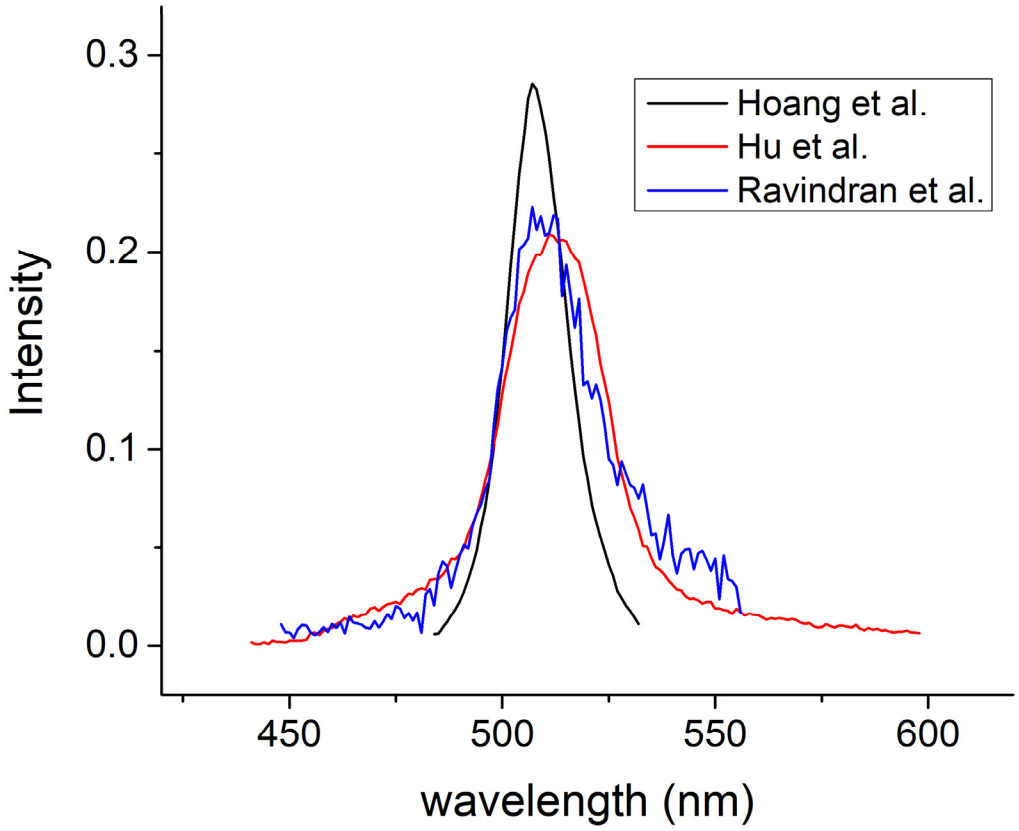

Hoang et al.
Hu et al.
Ravindran et al.
Intensity
0.0
0.1
0.2
0.3
450
500
550
600
wavelength (nm)

Figure 7