Enhancing UV Spectral Prediction through Auxiliary Task, Curriculum Learning, and Curvature Limitation

Hajime Shinohara
IBM Research – Tokyo

Akihiro Kishimoto
IBM Research – Tokyo

Hiroshi Kajino
IBM Research – Tokyo

Abstract

Accurate UV spectral prediction is challenging for machine learning. UV spectra exhibit broad absorption bands characterized by the peak positions, band shapes, and curvature profiles. However, current models fail to capture these characteristics. We present Peak Position Awareness (PPA), Curriculum Learning for Interpolated Abstracted Spectra (CLIAS), and Spectrum Curvature Limitation (SCL) to handle the above characteristics, showing consistent improvements over diverse models.

1 Introduction

The successful development of accurate machine learning models for chemical or physical properties can lead to effective design of materials of interest. Despite numerous attempts to advance machine learning research in materials discovery [4, 18], predicting UV spectrum of an organic molecule remains an open problem. While training neural networks (NNs) is one way to predict UV spectra, effective training is pressed by limited data due to the difficulties and inaccuracies in both experimental measurements and theoretical calculations [1, 14] as well as limited experimental resources.

An essential challenge closely related to UV spectroscopy is also raised. A UV spectrum is characterized by (1) *peak positions* that are the wavelengths showing (local) maximum absorbance, (2) *band shapes* that are the wavelength ranges characterizing absorbance including intensities, and (3) *curvature profiles* that determine spectral curves. Although these basic characteristics are established as typical ones for qualitative and quantitative molecular analysis in UV spectroscopy [5], state-of-the-art methods fail to capture these characteristics, thus suffering from accuracy and physical realism [20, 28].

We present three methods to capture the above characteristics: **PPA** predicts peak positions as an auxiliary classification task; **CLIAS** progressively trains on interpolated data through curriculum learning [3] to effectively learn band shapes; and **SCL** introduces second-derivative limitation for realistic curvatures. With the standard benchmark [28], we show that a careful combination of our methods successfully improves the performance across different models. We also demonstrate predicted spectra that clearly illustrate the success of our methods and room for further improvement.

1.1 UV Spectral Prediction Task and Spectral Generation Model

We follow the setting of Urbina *et al.* [28] where UV range spectra are observed. For fixed range $[w, w + N - 1]$ nm of the wavelengths in the entire dataset, where w and N are integers, a training example for spectrum i consists of a pair (m_i, S_i), where m_i is a molecule in SMILES format [31] and S_i is a spectrum to learn. S_i is a sequence $y_{i,0} \to y_{i,1} \to \cdots \to y_{i,N-1}$, where $y_{i,j}$ is a real value in $[0, 1]$. Semantically, $y_{i,j}$ is an absorption rate of wavelength $(w + j)$ nm irradiated to m_i.

As is investigated in [28], a *spectral generation model* receives *fixed-sized* input on a molecular structure and yields the absorption rates for corresponding wavelengths as output. The spectral

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Design (AI4Mat).

generation model is implemented either as a multi-output regression NN model, e.g., multi-layer perceptron (MLP), or as a sequence generation model, e.g., RNN or Transformer [29]. Our methods are easily incorporated into these cases to enhance the performance of the spectral generation model.

2 Related Work

Spectroscopic knowledge is promising in other spectroscopic areas. For example, McGill *et al.* [19] predicted IR spectra directly from SMILES using message passing neural networks. Chen *et al.* [6] used absorption bands for energy applications. Applying their methods to our task remains open due to different spectral characteristics.

Machine learning has been recently applied to UV spectral prediction. Several works focused on optical peak prediction [15, 10], while Urbina *et al.* [28] developed UV-adVISor, an attention-based recurrent NN for full spectra that performed best in their performance evaluation. In principle, our methods can be combined with UV-adVISor. However, as the source code is not publicly released, an empirical comparison remains future work. McNaughton *et al.* [20] used 3D molecular geometries and TD-DFT calculations as input, but cannot be directly compared with our 2D SMILES-string-based approach.

Curriculum learning [3] progressively exposes models to increasing task complexity and has been applied to many tasks including computer vision and natural language [30] with various strategies [24]. Attempts have also been made to apply curriculum learning to scientific domains. In these attempts, physical constraints have been incorporated into various scientific domains through theory-guided data science [16], physical property regularization [13], and physics-informed NNs [7]. However, how to introduce such physical constraints to UV spectral prediction still remains unresolved.

3 Our Enhancements to UV Spectral Prediction

We give details of our methods motivated by the fundamental principles behind UV spectroscopy.

3.1 PPA

As peaks are a reliable feature in UV spectroscopy, PPA introduces a *peak classification model (PCM)* that addresses an auxiliary task of peak classification. For wavelength range $[w, w + N - 1]$ nm, the PCM receives fixed sized input on molecular structure m and predicts peak locations as a binary vector $\mathrm{pv} = (v_0, v_1, \cdots, v_{N-1})$, where $v_j \in \{0, 1\}$ indicates whether wavelength $(w + j)$ nm is a peak position. The spectral generation model is extended to receive an input concatenated m with pv, aiming at more accurate predictions with this enriched input.

PCM needs a training dataset on pv. For sufficiently large thresholds on height h_{peak}, distance d_{peak} and width w_{peak} peak positions are calculated by an algorithm to find local maxima, which is based on simple comparisons of neighboring values and implemented in the scikit-learn library [23]. Additionally, peak positions are not apriori knowledge on the target molecule. For the test dataset, the spectral generation model receives pv predicted by the PCM. In contrast, pv is available in our augmented training dataset. The spectral generation model is trained with such available pv.

3.2 CLIAS

Existing domains where curriculum learning has been applied have clear heuristic criteria to effectively order training examples, e.g., the sequence length in natural language [25, 27]. In contrast, the spectral generation model always receives a fixed-sized input, and a clear criterion of the curriculum is needed. The curriculum of CLIAS captures broad bands of UV spectra. For $[N] := \{0, 1, \cdots, N - 1\}$, CLIAS selects a subset $P \subset [N]$ to satisfy that (1) P includes all indices for peak positions and indices 0 and $N - 1$, and (2) the difference in wavelengths between two adjacent indices in P are identical among all located between closest peak positions, or index 0 or $N - 1$. After training is complete with a new dataset based on P, CLIAS trains the model with the original dataset. CLIAS can also train with k_a subsets $P_1 \subset P_2 \subset \cdots \subset P_{k_a} = [N]$, progressing from A_1 towards A_{k_a}.

While P conveys dominant, abstract band shapes, $|P|$ can be inconsistent with the input size N of the spectral generation model. To ensure the consistent input size of N, CLIAS calculates interpolated

Table 1: RMSE values. P=PPA, C=CLIAS and S=SCL. Standard deviations are shown in the parentheses. See underline and bold numbers for the best case of each architecture and of all models.

Model	Baseline	P	P + C	All	P+C→P+S
MLP	0.12130 (0.0038)	0.10870 (0.0069)	0.10570 (0.0017)	0.10490 (0.0048)	0.10100 (0.0041)
Transformer	0.14160 (0.0033)	0.11600 (0.0023)	0.10980 (0.0035)	0.11000 (0.0027)	0.10860 (0.0036)
BiLSTM	0.11750 (0.0063)	0.10810 (0.0089)	0.10380 (0.0047)	0.10610 (0.0066)	**0.10030** (0.0036)

absorption rates for the unselected wavelengths. For spectrum i and two adjacent indices $j, j+L \in P$, we define an interpolated absorption rate for unselected wavelength $(w+j+k)$ nm (i.e. $j+k \notin P$) for abstract spectrum i is defined as $((L-k)y_{i,j} + ky_{i,j+L})/L$, where $y_{i,j}$ and $y_{i,j+L}$ are the absorption rates in its corresponding original spectrum.

3.3 SCL

Peak curvature in UV spectra is physically constrained by natural line broadening mechanisms. SCL enforces realistic curvatures by penalizing excessive curvature rarely observed in practice. For point j in spectrum i in the training data, let $d^2y_{\mathrm{true},i,j}$ and $d^2y_{\mathrm{pred},i,j}$ be respectively the true and predicted curvature values calculated as the second derivative approximation: $d^2y_{i,j} \approx y_{i,j+1} - 2y_{i,j} + y_{i,j-1}$. Using standard deviation σ of $d^2y_{\mathrm{true},i,j}$ for all i and j, SCL regards predicted point j in spectrum i that satisfies $|d^2y_{\mathrm{pred},i,j}| > \sigma$ as an unrealistic one to be penalized. For unrealistic pairs $(i,j) \in V$ and the original loss function $\mathcal{L}(y_{\mathrm{pred}}, y_{\mathrm{true}})$, SCL defines the loss function $\mathcal{L}_{\mathrm{SCL}}(y_{\mathrm{pred}}, y_{\mathrm{true}})$ as:

$$\mathcal{L}_{\mathrm{SCL}}(y_{\mathrm{pred}}, y_{\mathrm{true}}) := \mathcal{L}(y_{\mathrm{pred}}, y_{\mathrm{true}}) + \lambda_{\mathrm{cur}} \sum_{(i,j)\in V} (d^2y_{\mathrm{pred},i,j} - b_{\mathrm{cur}})^2$$

where λ_{cur} is a hyperparameter that controls the strength of the constraint, and b_{cur} is a hyperparameter that controls the penalty for each violated curvature.

4 Performance Evaluation

This section performs empirical evaluations of our methods and discusses the results.

4.1 Setup

While two smaller datasets obtained by physical experimental measurements were prepared in [28], we attempted to increase the size, merging these datasets into one with common wavelength ranges. The resulting dataset comprised 3,170 UV spectra (after removing 2 invalid SMILES) with absorption rates in the 230-400 nm range at 1 nm resolution (171 data points per spectrum), relevant to pharmaceutical development, organic electronics, and photocatalytic materials. The dataset was randomly split into training:validation:test = 7:1.5:1.5.

We aim at elucidating the behavior of our enhancements with diverse models. We implemented the following architectures in Python with the PyTorch library and trained them on NVIDIA H100 80GB GPUs: (1) **MLP** that generates a multi-output, (2) **BiLSTM** [11] that generates a sequence, and (3) **Transformer** that successfully leverages attention mechanisms [29]. The peak prediction model was implemented as an MLP. See Appendix A for the other detailed configurations and implementations. As the most common evaluation metric used in machine learning, we report RMSE values averaged over three independent runs with different random seeds (42, 123 and 456). The standard deviations shown in parentheses in Table 1 are calculated across these three runs to assess the stability and reproducibility of each model configuration.

4.2 Results and Discussion

Table 1 shows the performance of each model when each method is added one by one, highlighting only important cases. When including SCL as a final enhancement, we investigated two strategies. *All* trains a model by incorporating PPA, CLIAS and SCL all at once. *P+C → P + S* further tunes the model with PPA and SCL with additional 30 epochs, after initially trained with PPA and CLIAS.

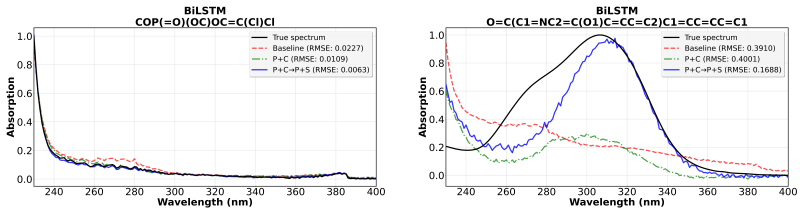

Figure 1: Spectra generated by baseline (red), P+C (green), and P+C→P+S (blue) and compared against ground truth (black).

PPA provided most fundamental benefits. This significant improvement is consistent with the importance of the peaks for deeper analysis in UV spectroscopy, where peak positions are the primary principle that directly encodes the energy gap between molecular orbitals through $E = hc/\lambda$. In contrast, baseline Transformer performed poorly to predict UV spectra. Since it receives the identical, repeated input on molecular structure, the positional encoder is the only factor that can differentiate the keys and values and queries of Transformer's self-attentions. This could be one reason Transformer failed to capture the relations between different wavelengths effectively.

PPA with CLIAS (P+C) yielded additional improvements to PPA roughly by 5.6% in the best case, which was still essential for more accurate spectral prediction. CLIAS focuses on band shapes that are a secondary feature and less crucial than peak positions. The fact that the performance improvement with P+C was modest consistently mirrors the physical reality of UV spectroscopy, such as vibrational structure and solvent effects through the Franck-Condon principle [2, 8].

P+C→P+S allowed P+C to first establish a stable solution for peak positions and an overall shape and provided P+S room with a better initialization point for curvature refinement, achieving at least a 10% improvement over each baseline model. P+C→P+S is consistent with UV spectroscopic analysis based on natural line broadening mechanisms, considered as a tertiary approach after peaks and band shapes are established. On the other hand, combining all three methods at one time tended to underperform P+C. The underperformed results of the naïve integration are attributed to the fact that the capability of SCL penalizing sharp shapes cannot effectively work when basic spectral patterns have not been learned yet, such as in an early training phase.

Figure 1 shows two representative spectra that contain single peaks. While their predictability differed drastically, the optimized BiLSTM model (i.e., P+C→P+S) generated better shapes for both cases. For the left spectrum, irrespective of with or without PPA, both models accurately predicted the peak at 230 nm. We found that the training dataset tended to contain high absorption rates for 230 nm. The models accurately predicted these high absorption rates, resulting in easily identifying the peak. Compared to the baseline model, P+C→P+S obtained the performance gain by reducing the absorption rates for the other part, clearly demonstrating the superiority of CLIAS and SCL. For the right spectrum, BiLSTM failed to correctly locate the 310 nm peak, while PPA helped accurately identify peak position. P+C→P+S further increased the peak absorption rate to be closer to the true value, demonstrating its effectiveness to address challenging cases even if molecular characteristics deviate from typical patterns. On the other hand, all models incorrectly generated high values at 230 nm caused by the characteristics of the training data, making it difficult to correct the shapes.

5 Conclusions and Future Work

As an initial step to developing an accurate spectral generation model, we introduced PPA, CLIAS and SCL, which embody fundamental physical principles. With the dataset obtained by the physical experiments [28], we showed that P+C→ P+S achieved at least a 10% improvement over the best baseline. There are numerous approaches to explore as future work. For example, more comprehensive understanding to the behavior of these methods is necessary, including a comparison and integration with the state-of-the-art UV-adVISor model.

Extending the dataset size is clearly of importance, and opens up opportunities to develop better theoretical and experimental methods in UV spectroscopy as well as new approaches that allow to train with the dataset comprising both theoretical and experimental spectra that might have non-negligible gaps. New machine learning algorithms that leverage partially observed spectra also help effectively train models with extended data.

References

[1] C. Adamo and D. Jacquemin. The calculations of excited-state properties with Time-Dependent Density Functional Theory. *Chemical Society Reviews*, 42(3):845–856, 2013.

[2] P. W. Atkins and R. S. Friedman. *Molecular Quantum Mechanics*. Oxford University Press, Oxford, UK, 5th edition, 2011.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48. ACM, 2009.

[4] J. Cai, X. Chu, K. Xu, H. Li, and J. Wei. Machine Learning-Driven New Material Discovery. *Nanoscale Advances*, 2(8):3115–3130, 2020.

[5] L. Chen, Z. Gao, Q. Li, C. Yan, H. Zhang, Y. Li, and C. Liu. A review: Comprehensive investigation on bandgap engineering under high pressure utilizing microscopic UV–Vis absorption spectroscopy. *APL Materials*, 12(3):030602, 2024.

[6] X. Chen, M. M. Singh, and P. Geyer. Utilizing Domain Knowledge: Robust Machine Learning for Building Energy Prediction with Small, Inconsistent Datasets. *Energy and Buildings*, 278:112631, 2023.

[7] Y. Chen, L. Lu, G. E. Karniadakis, and L. Dal Negro. Physics-informed neural networks for inverse problems in nano-optics and metamaterials. *Optics Express*, 28(8):11618, Apr. 2020.

[8] E. U. Condon. A Theory of Intensity Distribution in Band Systems. *Physical Review*, 28(6):1182–1201, 1926.

[9] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

[10] K. P. Greenman et al. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chem. Sci.*, 13:1152–1162, 2022.

[11] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[12] S. Honda, S. Shi, and H. R. Ueda. SMILES Transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

[13] T. Huynh, M. J. Lai, Y.-L. Liu, L. Ly, X. Gong, K. R. Rommel, and D. L. Young. Spectral Analysis Methods Based on Background Subtraction and Curvature Calculation Used in the Detection or Quantification of Hemolysis and Icterus in Blood-derived Clinical Samples. *Cureus*, 9(12):e1965, 2017.

[14] D. Jacquemin, E. A. Perpète, I. Ciofini, C. Adamo, R. Valero, Y. Zhao, and D. G. Truhlar. TD-DFT performance for the visible absorption spectra of organic dyes: Conventional versus long-range hybrids. *Journal of Chemical Theory and Computation*, 7(2):369–376, 2011.

[15] J. F. Joung et al. Deep learning optical spectroscopy based on experimental database. *JACS Au*, 1:427–438, 2021.

[16] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

[17] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[18] Y. Liu, T. Zhao, W. Ju, and S. Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

[19] C. J. McGill, M. Forsuelo, Y. Guan, and W. H. Green. Predicting Infrared Spectra with Message Passing Neural Networks. *Journal of Chemical Information and Modeling*, 61(6):2594–2609, 2021.

[20] A. D. McNaughton, R. P. Joshi, C. R. Knutson, A. Fnu, K. J. Luebke, J. P. Malerich, P. B. Madrid, and N. Kumar. Machine Learning Models for Predicting Molecular UV–vis Spectra with Quantum Mechanical Properties. *Journal of Chemical Information and Modeling*, 63(5):1462–1471, 2023.

[21] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures – a Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.*, 5(2):107–113, May 1965.

[22] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, pages 807–814, 2010.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe. Curriculum Llearning in Ddeep Nneural Nnetworks: A literature review. *Neural Networks*, 152:345–363, 2022.

[25] V. I. Spitkovsky, H. Alshawi, and D. Jurafsky. Baby Steps: How "Less is More" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 751–759, 2010.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[27] S. Subramanian, S. Rajeswar, F. Dutil, C. Pal, and A. Courville. Adversarial Generation of Natural Language. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 241–251, 2017.

[28] F. Urbina, K. Batra, K. J. Luebke, J. D. White, D. Matsiev, L. L. Olson, J. P. Malerich, M. A. Z. Hupcey, P. B. Madrid, and S. Ekins. UV-adVISor: Attention-based recurrent neural networks to predict UV–vis spectra. *Analytical Chemistry*, 93(48):16076–16085, 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In *NeurIPS*, pages 5998–6008, 2017.

[30] X. Wang, Y. Chen, and W. Zhu. A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022.

[31] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

A Detailed Setups for Performance Evaluation

The NN architectures of our baseline models are summarized as follows:

- **MLP** with 4 hidden layers sized 1024, 512, 256 and 128, activated by ReLU [22]
- **BiLSTM** [11] with 256 hidden units
- **Transformer** [29] simplified with 4 attention heads and 256-dimensional embeddings

The peak prediction model consists of a 3-layer MLP with each hidden layer size 2048, 1024, 512, and 256, followed by 171 outputs with a sigmoid operation to calculate a probability that each absorption rate is a peak.

There are several choices to encode the molecular structure, e.g., fingerprint [21] and recent pretrained NNs [9, 12]. We employ the approach of Urbina *et al.* [28], which partitions a SMILES string of a molecule to a sequence of tokens such as atoms and bonds. Paddings are added to the end of the sequence to create a fixed-sized vector passed as input to the spectral generation model.

A molecule is encoded as SMILES string at most with 150 characters, tokenized by common chemical vocabularies. The padded tokens are embedded to a vector of size 256, which is then passed to the input of each NN. When PPA is combined, another embedded vector of size 256 is created from the peak-value vector of size 171, concatenated with the embedded vector on the molecule.

We performed preliminary experiments with the validation set and grid search, and set hyperparameters $\lambda_{\mathrm{cur}} = 0.1$, $b_{\mathrm{cur}} = 0.1$, $h_{\mathrm{peak}} = 0.2$, $d_{\mathrm{peak}} = 40$ and $w_{\mathrm{peak}} = 40$.

In model training, we set early stopping with a patience of 30 epochs and the Adam optimizer [17] with optimized learning rates: 10^{-3} for MLP, 10^{-4} for Transformer, and 5.0×10^{-4} for the others. Dropout (rate=0.2) [26] was applied between fully connected layers.

For the peak prediction model, after training the model for 50 epochs, we set a threshold to 0.2, peak distance to 40 nm and peak width to 40nm to determine if a value is a peak.

For CLIAS, in our preliminary evaluation, we set k_a to at most 3 and tested three combinations of abstraction sizes: [43,171], [86,171], and [43,86,171]. The best-performing configurations were [43, 86, 171] for MLP and BiLSTM, and [43, 171] for Transformer.

Each phase except the final phase used early stopping (patience=20) and learning rate decay of 0.8 between phases, while the final phase used patience=7.

The MSE loss was used for training all the spectral generation models, with additional terms when SCL was incorporated. Using the peak one-hot vectors as binary labels, the binary cross entropy loss was used to train the peak classification model.