Unraveling the Correlation Between Raman and Photoluminescence in Monolayer MoS_2 Through Machine Learning Models

機械学習モデルによる単層MoS_2のラマンとフォトルミネッセンスの相関の解明

Ang-Yu LU

Transition metal dichalcogenides (TMDCs) show strong, tunable photoluminescence (PL), advancing optoelectronic applications. Raman spectroscopy, which is crucial for analyzing 2D materials, discerns crystallinity and material variations such as doping and strain. Nonetheless, the hidden PL-Raman correlations in MoS_2 monolayers are not fully studied. This work methodically investigates PL-Raman interconnections, clarifying the underlying physical mechanisms. Employing machine learning, we differentiate strain and doping effects in Raman data. A DenseNet model predicts PL from Raman maps, while gradient-boosted trees with SHAP assess Raman features' PL influence, elucidating MoS_2's strain and doping. This research offers a machine learning-based methodology for 2D material characterization and informs the tuning of semiconductors for enhanced PL.

遷移金属ジカルコゲナイド(TMDC)は高輝度かつ変化するフォトルミネッセンス(PL)を示し，オプトエレクトロニクス応用を前進させるだろう。ラマン分光法は2次元材料の分析に不可欠であり，結晶性やドーピングやひずみなどの材料変化を解析することができる。にもかかわらず，MoS_2単分子膜における隠れたPL-ラマン相関は十分に研究されていない。本研究ではPL-ラマン相関を系統的に調べ，その奥にある物理メカニズムを明らかにする。機械学習を用いてラマンデータのひずみとドーピングを識別する。DenseNetモデルにより，ラマンマップからPLを予測し，SHAPによる勾配ブースティングツリー(GBT)によりラマンの特長，PLへの影響を評価し，MoS_2のひずみとドーピングを解明することができる。この研究は2次元材料特性評価のための機械学習ベースの方法論を提供し，PL増強のための半導体のチューニングに役立つ。

Introduction to MoS_2 Monolayers and Machine Learning Models

Two-dimensional (2D) materials, distinguished by their ultra-thin structure and high surface-to-volume ratio, exhibit unique physical and chemical characteristics. Among these, monolayer transition metal dichalcogenides (TMDCs) are 2D semiconductors known for their adjustable photoluminescence (PL). This PL can be altered through external factors like strain and doping. For instance, MoS_2, a type of TMDC, shows adjustable band structures and broad-spectrum optical absorption when subjected to strain. These properties make it highly suitable for various advanced applications, such as cutting-edge photovoltaic systems[1] and quantum information science technologies, including single-photon emission[2]. Additionally, the near-perfect PL quantum yield in MoS_2, achieved through either chemical[3] or electrostatic doping[4], paves the way for creating highly efficient light-emitting diodes[5] and lasers[6]. To analyze these external influences, Raman spectroscopy is employed as an effective and non-invasive method to measure the impact of strain and doping on the properties of MoS_2. While Raman and photoluminescence (PL) spectroscopy have been instrumental in exploring strain and doping effects in MoS_2, the majority of research has treated these effects separately. Discovering the hidden correlations between Raman and PL spectra can enable us to understand the strain and doping effects comprehensively. Recently, the rise of machine learning has revolutionized fields such as computer vision and natural language processing and has made significant inroads in diverse scientific disciplines, including biology[7], mathematics[8], and material science[9]. Although machine learning approaches have been utilized

in research on 2D materials, these efforts remain in the nascent stage[10,11] and hold potential for groundbreaking discoveries.

In this study, we leveraged an array of machine learning algorithms to uncover the hidden patterns linking Raman and PL spectra in MoS_2, providing insights into the physical mechanisms connecting PL and Raman features. Our approach started with the implementation of a DenseNet model, which demonstrated high predictive accuracy for PL features from Raman spectral maps. Subsequently, we integrated a gradient-boosted model with SHapley Additive exPlanations (SHAP) to correlate Raman and PL data, offering both global significance and local interpretability in terms of feature contributions. Lastly, we projected MoS_2 Raman features on frequency scatter plots to decompose strain and doping effects. Our findings illustrate the potent capability of machine learning tools in elucidating complex relationships across different material characterization techniques.

The conceptual illustration provided in Figure 1(a) shows the trajectory of knowledge acquisition through machine learning models (represented by the red line), commencing from the results of material characterization, progressing through established material knowledge, and culminating in the understanding of external perturbations and defect structures. This methodology enables the integration of prior investigations-those that examined changes in Raman and PL spectra due to single external effects such as strain (indicated by the green line) or doping (denoted by the blue line)-into a comprehensive understanding of MoS_2 monolayers. Additionally, the employment of statistical data analysis within our framework serves to reduce potential biases arising from sample selection and experimental setup. By integrating statistical analysis with machine learning, our model creates a robust connection between Raman and PL characteristics, the crystalline and electronic structures, as well as the effects of strain and electrostatic doping.

Synthesis and Characterization of MoS_2 Monolayers

MoS_2 monolayers are synthesized through chemical vapor deposition (CVD) on 300 nm SiO_2/Si substrates, utilizing molybdenum trioxide (MoO_3) and sulfur (S) powders, each weighing 20 mg, as source materials. Substrates are prepared with a spin-coated layer of Perylene-3,4,9,10-tetracarboxylic acid tetrapotassium (PTAS) solution, which acts as a seeding promoter. To ensure an oxygen and moisture-free environment, the CVD system is flushed with an Argon (Ar) flow of 1000 sccm for 5 minutes. The temperature of the furnace is increased to 625℃ at a rate of 30℃ per minute. Concurrently, sulfur is maintained at 180℃ in an upstream position within the system. The growth of MoS_2 monolayers occurs at 625℃ under atmospheric pressure for 3 minutes, with an Ar flow of 20 sccm and an O_2 flow ranging from 0 to 1 sccm, serving as the carrier and reactant gases, respectively. Post-growth, the furnace is allowed to cool to room temperature naturally under a continuous Ar flow of 1000 sccm to avert any additional unintended chemical reactions. The MoS_2 crystals were obtained from SPI Supplies and 2D semiconductors, and then mechanically exfoliated and deposited onto a 300 nm SiO_2/Si substrate. The SPI crystals were naturally grown, while the 2D semiconductor crystals were synthetic. The exfoliated flakes were termed natural and synthetic, accordingly.

We employed the HORIBA LabRAM HR800 spectrometer for Raman and photoluminescence (PL) characterizations, utilizing a 532 nm (2.33 eV) laser source. Due to time constraints associated with each spectral mapping,

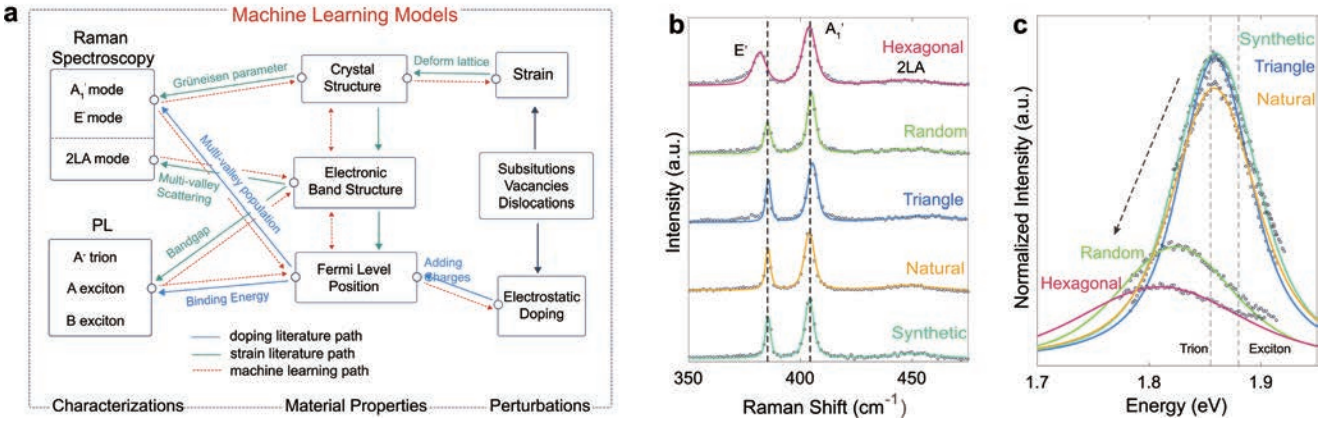

Figure 1 (a) Overview of unraveling correlation between Raman and PL with external perturbations. The green and blue lines correspond to the studies of strain and doping effects, respectively. The red dashed line indicates the discovering path by the machine learning models in this work. (b-c) Raman and PL spectra of CVD-grown MoS_2 monolayers. (b) Raman, and (c) PL spectra of CVD-grown (hexagonal, random, and triangle) and exfoliated (natural and synthetic) MoS_2. The vertical dashed lines denote the Raman E' and A'_1 frequencies for the synthetic MoS_2 in a and the MoS_2 exciton energy of 1.86 and 1.89 eV for trion and exciton, respectively.

the laser power was carefully maintained at approximately 1 mW, and the acquisition time was set to 0.5 seconds.

For our spectral maps, we collected data across 1,600 pixels, corresponding to a spatial dimension of 40 by 40. This process required approximately 30 minutes per spectral map. Consequently, the complete measurement of a single MoS_2 crystal, including the necessary system calibrations, amounted to over 1.5 hours. Our Raman and PL measurements utilized a 100X objective lens, which focuses the laser to a 1 μm diameter spot size. The normalization of Raman spectra was conducted against the intensity of the silicon (Si) peak, with calibration referring to the Si Raman frequency established at 520.6 cm^{-1}. We employed spectral gratings of 1,800 grooves per millimeter (gr/mm) for Raman and 300 gr/mm for PL measurements to ensure high-resolution spectral data. The spatial dimensions for the Raman and PL mapping were tailored to match the domain size of the MoS_2 flakes under investigation, with a spatial resolution of 1 μm achieved through the precision control of a motorized stage.

The Raman spectra of MoS_2 monolayers are characterized by three characteristic features as illustrated in Figure 1(b), which represent the in-plane E' mode at approximately ~385 cm^{-1}, the out-of-plane A'_1 mode near 405 cm^{-1}, and the second-order double resonance 2LA mode around 450 cm^{-1}. These vibrational modes are precisely defined using a Voigt profile for the extraction of key parameters: the frequency (Freq, ω), the full-width-at-half-maximum (FWHM, Γ), and the intensity (Int, I). Figure 1(b) displays the frequency distribution for these modes across the MoS_2 monolayers sampled, highlighting a trend of frequency softening for the $\omega_{A'_1}$ mode in naturally-grown MoS_2 and for the $\omega_{E'}$ mode in hexagonal-shaped MoS_2, in comparison to those exfoliated from synthetic sources. In the context of PL characterization, trions appear to dominate the PL response from MoS_2 due to the ~1 mW laser power utilized during our experiments. A closer evaluation of Figure 1(c) indicates that MoS_2 flakes with hexagonal and random shapes tend to show lower PL energy, wider FWHM, and reduced intensity in contrast to triangle-shaped and mechanically exfoliated MoS_2 crystals.

For the analysis of spectral data, a curve-fitting routine was executed within a Python environment. This routine involved the subtraction of background noise from the spectra by employing the BaselineRemoval package, specifically utilizing the ZhangFit method[12]. To accurately fit the spectral lines, we utilized the Voigt profile function, accessible from the Scipy library, which is defined by a quartet of parameters: the frequency of the Raman feature, σ (which is the standard deviation of the Gaussian component of the Voigt profile), γ (representing the half-width at half-maximum of the Lorentzian component), and the intensity of the Raman feature. For optimization purposes, the least-squares optimization function from the Scipy library was engaged. The full width at half-maximum (FWHM) is then computed by applying the following equation (1-3):

$$f_G = 2\sigma\sqrt{2ln2} \quad (1)$$

$$f_L = 2\gamma \quad (2)$$

$$\Gamma \approx 0.5346 f_L + \sqrt{0.2166 f_L^2 + f_G^2} \quad (3)$$

This equation effectively combines the contributions from both the Gaussian and Lorentzian components that make up the Voigt profile, giving a comprehensive measure of the spectral line's width at its half-maximum intensity. Upon capturing the Raman and PL spectra, we channeled the data through a curve-fitting process. Each spectral peak was modeled with a Voigt profile, which yielded three primary parameters: the peak frequencies, the FWHM, and the intensities. We then normalized all characteristic peaks using the intensity of the Raman signal at 520.6 cm^{-1} from the silicon substrates. For MoS_2 monolayers with a normalized intensity of the A'_1 mode ($I_{A'_1}$) less than 0.5, we applied a stringent threshold to demarcate regions attributed to multilayer MoS_2. Meanwhile, outliers identified in the spectral maps were subsequently eliminated using a binary opening operation.

In the development of our machine learning models - specifically XGBoost and the support vector machine (SVM)-we utilized a total of 7,023 data points. The DenseNet model's training dataset comprised all pixel data, inclusive of those with and without Raman/PL signals. Data augmentation was performed by applying a 90-degree rotation to the pixel maps, accumulating in a dataset encompassing 35,596 patched maps. Statistical analysis was conducted using Matlab and Python, incorporating libraries such as Pytorch, Scipy, and Numpy to facilitate the analysis.

Statistical Analysis for MoS_2 Monolayers

To further elucidate the characteristics of PL in MoS_2, we graphed the PL FWHM against the normalized PL intensity, as depicted in Figure 2(a). This graph demonstrates a discernible trend: higher PL intensities are associated with narrower PL FWHMs, a pattern typically observed in synthetic and triangle-shaped MoS_2 crystals, as shown in Figure 2(b). The PL FWHM (Γ_{PL}) can be represented by a reciprocal relationship with PL intensity (I_{PL}) as shown in Equation (4), which plots as the blue curve in Figure 2(a). When considering potential discrepancies

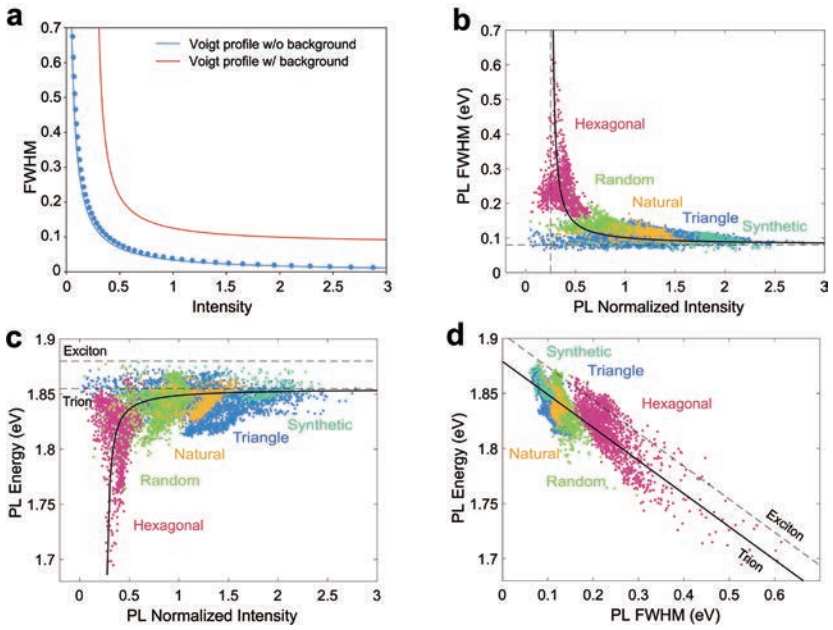

Figure 2 Scatter plots for PL features for MoS_2 monolayers. (a) The correlation between FWHM and intensity in a Voigt function for a fixed area of 0.1 without (blue, Equation (4)) and with (red, Equation (5)) the background. (b) PL FWHM as a function of the normalized intensity following the multiplicative inverse function (solid-black line) described by Equation (5). (c) PL energy as a function of the normalized intensity following the reverse multiplicative function (solid-black line) described by Equation (6). (d) PL energy as a function of the PL FWHM following a linear function (solid black) described by Equation (7).

due to imperfect background subtractions in the spectral data, we adjusted the equation to include an offset for both intensity and FWHM, resulting in Equation (5). This adjusted model corresponds to the red curve seen in Figure 2(a) and aligns well with the distribution observed in Figure 2(b). In addition, we explored the relationship between PL energy (E_{PL}) and intensity (I_{PL}) as shown in Figure 2(c). As the PL spectrum was fitted using a Voigt profile with a fixed integrated area, a reverse reciprocal function illustrated by Equation (6) was employed, indicating that stronger PL intensities are associated with higher PL energies. Moreover, a linear relationship between PL energy and FWHM is depicted in Equation (7), and this is graphically represented in Figure 2(d).

The collective interpretation of these findings suggests that the triangular, as well as natural and synthetic MoS_2 flakes, display PL peaks that are more intense, sharper, and exhibit a blue shift, signaling a higher crystal quality when compared to the random and hexagonal MoS_2 flakes.

$$\Gamma_{PL} = 0.0157/I_{PL} \quad (4)$$

$$\Gamma_{PL} = 0.0157/(I_{PL} - 0.25) + 0.08 \quad (5)$$

$$E_{PL} = 1.855 - 0.0047/(I_{PL} - 0.25) \quad (6)$$

$$E_{PL} = 1.879 - 0.3 * \Gamma_{PL} \quad (7)$$

High Performance of DenseNet

To unravel the hidden relationships within our Raman and PL data, we employed a variety of machine learning techniques, particularly focusing on revealing hidden patterns and establishing connections to underlying physical phenomena. Deep convolutional neural networks (CNNs), renowned for their proficiency in a multitude of visual recognition tasks, enable the extraction of valuable insights from diverse imaging systems, encompassing the biomedical[13] to the microscopic[14] and hyperspectral domains[15]. Viewing spectral maps as image-based datasets with multiple channels, such as the number of spectral points. Using CNNs, we correlate the Raman spectra with the corresponding PL features for the CVD-grown and exfoliated MoS_2 flakes.

We chose to deploy Dense Convolutional Networks (DenseNet)[16] for their efficiency in predicting three PL features from the Raman spectral images. DenseNet has been demonstrated to require fewer parameters and less down-sampling compared to other advanced CNN models, such as U-Net[17] and SegNet[18], while still delivering comparable accuracy. This characteristic makes DenseNet particularly advantageous for handling small datasets and small pixelated inputs, aligning perfectly with the scope of our research. The DenseNet architecture[16] implemented in our study was a tailored version of the original design, adapted to comprise two dense blocks. Preceding the entry to the first dense block, the input image undergoes a convolution with an output of 12 channels; the specifics of this step are illustrated in Figure 3(a). Each dense block is constructed with several layers: batch normalization, ReLU activation, convolutions with 1x1 and 3x3 kernel sizes, and is followed by a dropout rate set at 0.1 to prevent overfitting. The transition layer that bridges the two dense blocks includes batch

normalization, a ReLU layer, a convolutional layer with a 1x1 kernel size, and concludes with an average pooling layer. Upon the completion of the final dense block, an adaptive average pooling operation is executed, outputting three channels, which are then connected to a linear layer designed to produce three final output values. The deployment of DenseNet was facilitated through the PyTorch framework. To evaluate the performance of DenseNet, we utilized the relative absolute error, articulated by the following Equation (8):

$$RAE = \frac{[\sum_{i=1}^{n}(y_i' - y_i)^2]^{\frac{1}{2}}}{[\sum_{i=1}^{n} y_i^2]^{\frac{1}{2}}} \quad (8)$$

In this equation, y'_i represents the predicted values obtained from DenseNet, and y_i denotes the actual experimental values acquired from the PL measurements. This metric allows for the quantification of the prediction accuracy of the network relative to the true data values.

To explore the relationship between the spatial information contained within Raman patched maps and the performance of the DenseNet architecture, we experimented with various sizes of Raman patched maps. These ranged from a local spatial size of 1x1 to a more extensive spatial size of 11x11 for the intensity data of hexagonal MoS_2, an example of which is shown in Figure 3(b-d). It was observed that the smaller 1x1 patch size yielded a higher error rate of 21.48%, which can be attributed to the limited spatial information provided by the Raman maps. On the other end of the spectrum, the 11x11 patch size resulted in a marginally increased relative error of 11.86%, potentially due to zero padding implemented around the edges of the patched inputs. Out of all the patch sizes tested, the 5x5 configuration achieved the most favorable balance, exhibiting the lowest relative absolute error (RAE) of 10.31% for the PL intensity of a triangle-shaped MoS_2. This particular patch size managed to integrate adjacent Raman signals while avoiding the inclusion of extraneous spatial information.

The central columns of Figure 3(e-g) illustrate typical predictions for PL energy, FWHM, and intensity as derived from the trained DenseNet model when applied to a random-shaped MoS_2 using a 5x5 patch size. For model performance assessment, the experimentally measured PL maps were considered as the benchmark (ground truth), displayed in the left column. The relative errors computed are shown in the right column of Figure 3(e-g). The RAEs for the PL energy and FWHM were notably low, at 0.25% and 4.61% respectively. However, the RAE for PL intensity was higher, recorded at 10.93%, which may be

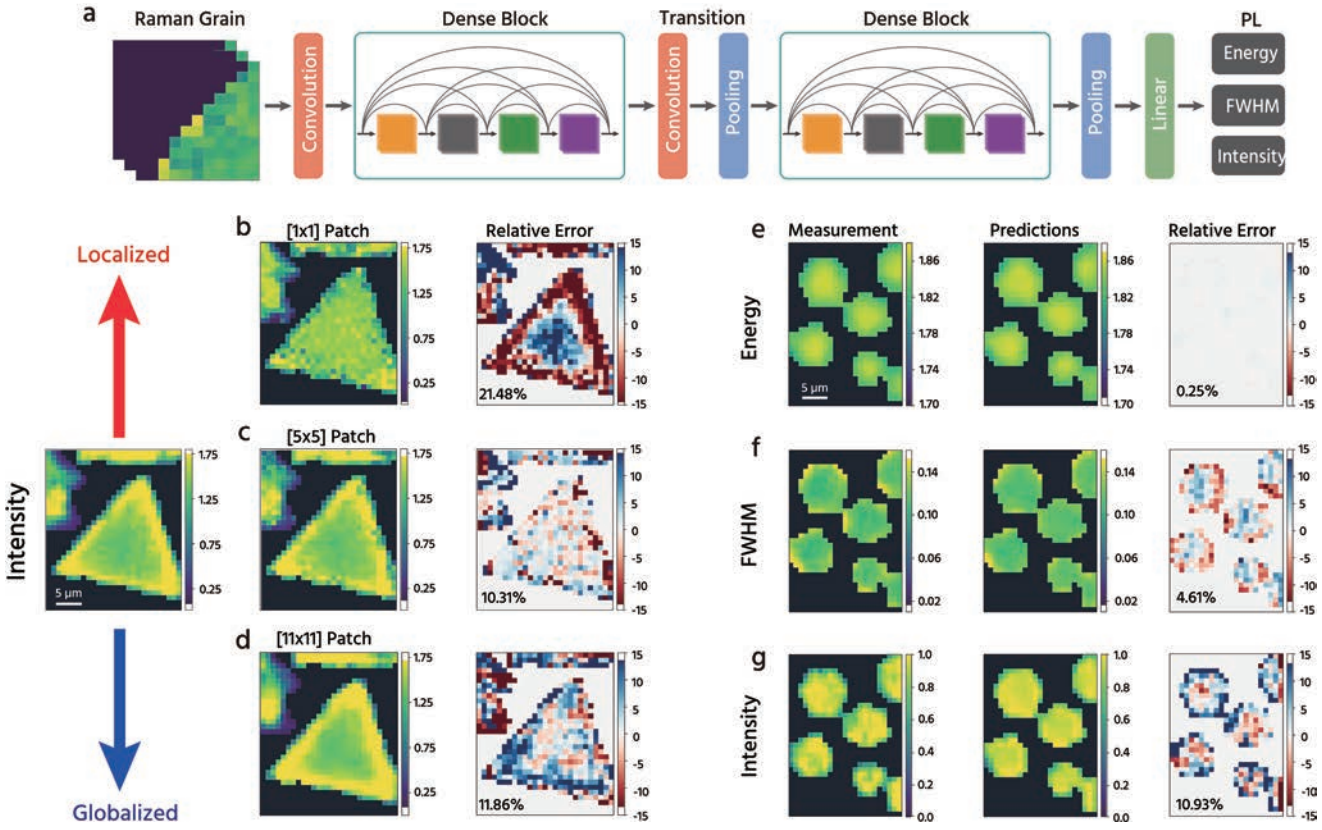

Figure 3 (a) Schematic illustration of a DenseNet model with two dense blocks. (b-d) PL predictions from the trained DenseNet. The patch size effect on the DenseNet for the predicted intensity of a triangle MoS_2. (b) 1-by-1, (c) 5-by-5, and (d) 11-by-11 with 21.48%, 10.31 and 11.86% relative absolute errors, respectively. (e-g) The PL mapping predictions of (e) energy, (f) FWHM, and (g) intensity for CVD-grown MoS_2 with random shape. Left: the measured PL maps as the ground truth for the DenseNet model. Middle: the predicted results by the trained DenseNet with 5-by-5 patch inputs. Right: the relative error between measured and predicted PL maps.

indicative of non-ideal experimental conditions or errors in the data processing stages, including spectroscopic measurements, background spectral subtraction, and the fitting procedures.

XGBoost Model with SHAP Explainer

Although CNNs represent the state-of-the-art model to make inferences on image- or spectral-based tasks, their multilayer nonlinear structures are often criticized as non-transparent and non-explainable[19]. In response to this, we transformed spectral maps into a tabular dataset comprising roughly 7000 discrete data points. An extreme gradient boosting (XGBoost) model, trained on this tabular dataset, was utilized to discern the correlations between Raman characteristics and corresponding PL features in MoS_2 monolayers.

XGBoost, an ensemble learning model constructed from decision trees[20], is widely recognized for its effectiveness in supervised learning tasks, especially when dealing with tabular datasets featuring individually significant attributes that do not incorporate temporal or spatial structures[20]. The optimization of the XGBoost regressor's hyperparameters was conducted via Bayesian Optimization[21], with the model configured to include 700 gradient-boosted trees, a learning rate of 0.05, and a maximum tree depth of 15. Root mean square log error (RMSLE) was employed as the evaluative metric to minimize the impact of outliers on error calculation.

To interpret the XGBoost model predictions and link them to their physical underpinnings, we applied Shapley Additive exPlanations (SHAP)[22], which acts as a tree

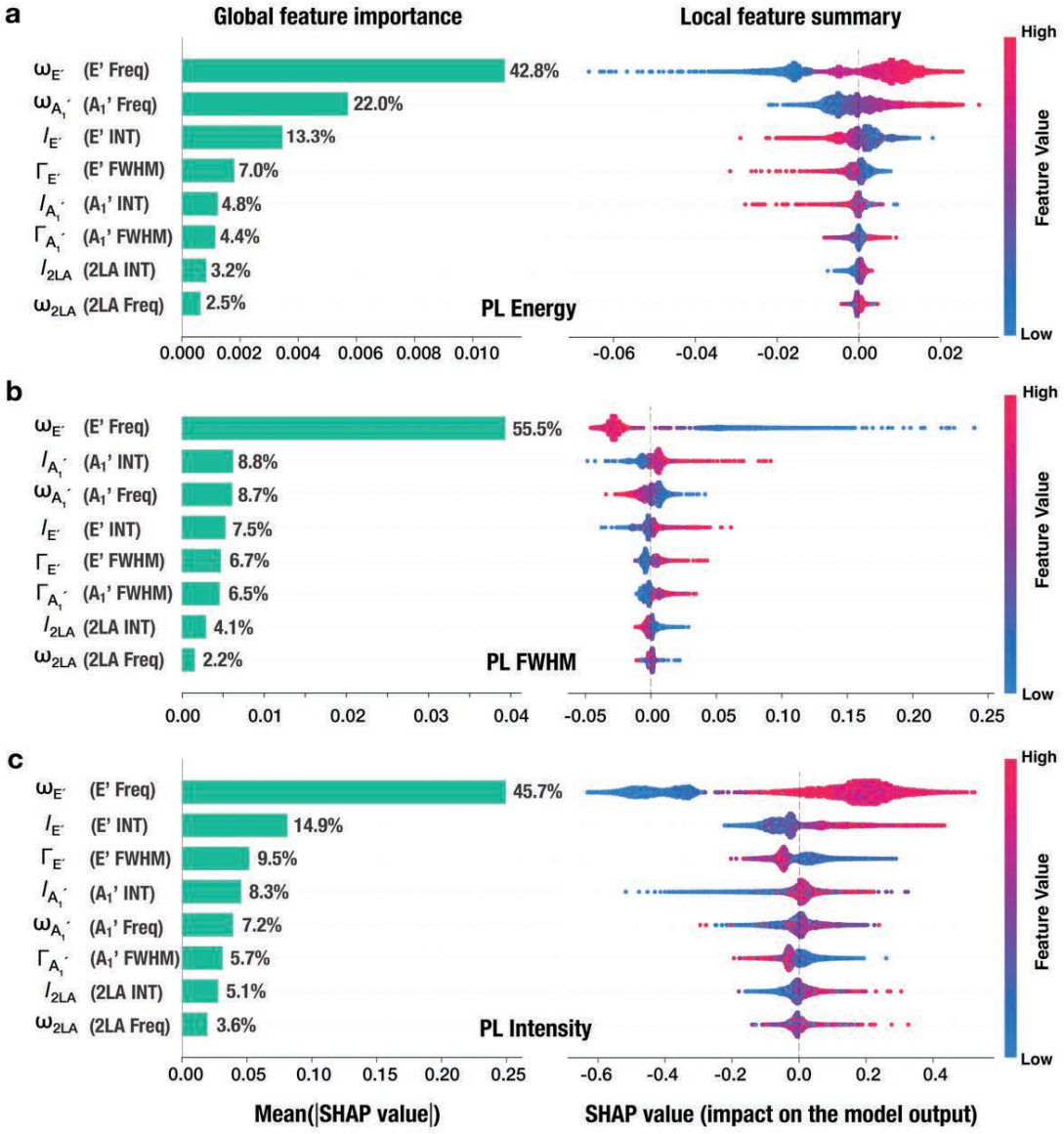

Figure 4 Correlation analysis for Raman and PL by XGBoost with SHAP values. (a) PL energy, (b) PL FWHM, and (c) PL intensity results are interpreted by the trained XGBoost model by Shapley additive explanations (SHAP). The Raman features are sorted in descending order according to global parameter importance. Left: the global importance of Raman features based on the average SHAP value magnitude for PL features. Right: a set of beeswarm plots corresponding to a single pair of Raman and PL. The vertical axis displays the sorted Raman features, while the horizontal axis shows the impact of the model output. Each data point represents a predicted output, and the color indicates the Raman features values.

explainer. SHAP provides both local and global insights based on game theory principles, elucidating the connections between Raman and PL spectra.

In Figure 4, SHAP summary plots visualize the influence of specific Raman feature values on the predicted PL features. Individual dots on these plots represent model predictions, with colors encoding the value of a particular Raman feature. For instance, a higher frequency of the E' Raman mode ($\omega_{E'}$, indicated by a red color) correlates with an increased SHAP value, suggesting a heightened PL energy. Moreover, bar charts in Figure 4 detail the SHAP importance values, offering a global perspective on the contribution of each Raman parameter to the PL features.

Analysis of the SHAP values revealed that the Raman features, $\omega_{E'}$, $\omega_{A'_1}$, and $I_{E'}$ are the most impact factors in predicting PL features. The average SHAP importance for the E', A'_1, and 2LA modes with respect to the PL features are 67.6%, 25.5%, and 6.9%, respectively. This distribution of importance is consistent with prior studies indicating that the E' Raman mode is sensitive to in-plane strain but less affected by doping[23], whereas the A'_1 mode's sensitivity is reversed, being more responsive to doping than to strain[23-25]. Given that the E' mode exhibits the most significant SHAP importance (67.6%) for PL prediction, we infer that the PL response within our dataset is predominantly influenced by strain effects rather than doping.

Scatter Plots for Decomposition of Raman Frequencies

The differentiation of strain and doping effects on the vibrational properties of graphene has been established through the shifting of G and 2D band frequencies. Extending this methodology to monolayer MoS_2[26,27], the SHAP importance results have highlighted that the frequencies of the $\omega_{E'}$ and $\omega_{A'_1}$ modes predominantly influence the PL characteristics. While similar strategies have been previously applied to MoS_2, the hidden details of these physical phenomena have not been fully discerned from the Raman frequency analyses.

In our investigation, we demonstrated the decomposition of strain and doping effects as functions of $\omega_{E'}$ and $\omega_{A'_1}$ in Figure 5(a). We begin by identifying the intrinsic point,

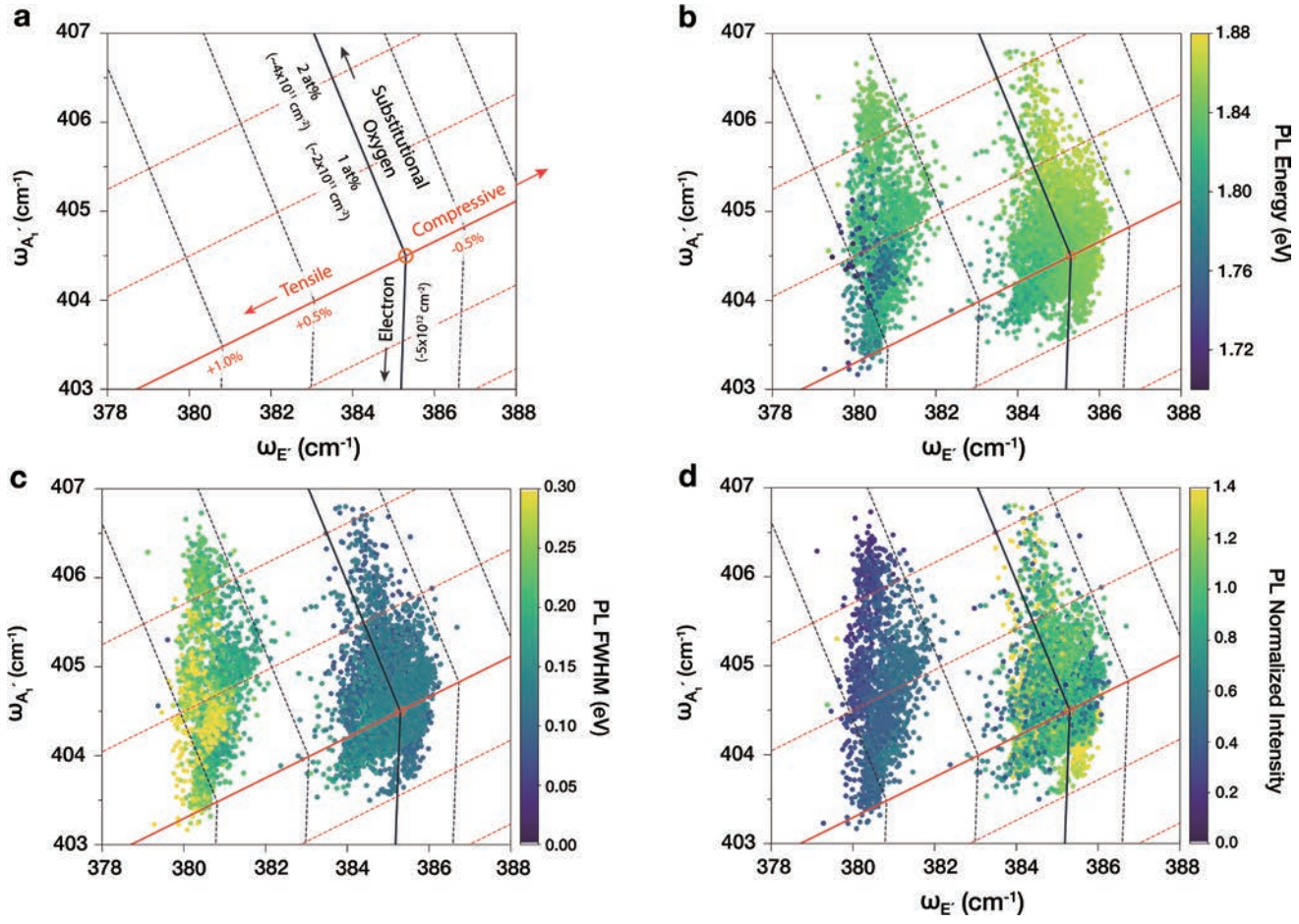

Figure 5 (a) Schematic representation of strain and doping base vectors for ($\omega_{E'}$, $\omega_{A'_1}$) coordinates. The red and black solid line corresponds to strain and doping, respectively. The orange circle is denoted as the intrinsic point, defined as the charge-neutral and unstrained state. (b-d) Scattered plots of CVD-grown and exfoliated MoS_2 monolayers on Raman features of $\omega_{A'_1}$ versus $\omega_{E'}$. The coded colors indicate (b) PL energy, (c) PL FWHM, and (d) PL intensity.

which is characterized as the undoped and unstrained state of MoS_2. The intrinsic Raman frequencies of the $\omega_{E'}$ and $\omega_{A'_1}$ modes are somewhat elusive; however, by comparing a range of E' and A'_1 data from both literature and our own studies, we find that most data align with the established frequency difference of $\omega_{A'_1} - \omega_{E'} = 19$ cm^{-1}[28], which is the recognized standard for monolayer MoS_2. We have designated the Raman frequencies from exfoliated synthetic MoS_2, located centrally within our data distribution, as the intrinsic point, marked at (385.3, 404.5) for ($\omega_{E'}$, $\omega_{A'_1}$), respectively. This point is indicated by an orange circle in our representations. Regarding strain effects, we observe that tensile strain induces shifts in the $\omega_{E'}$ and $\omega_{A'_1}$ of 4.48 and 1.02 cm^{-1}/%, respectively. This observation is aligned with the ratio of the Grüneisen parameters of the E' and A'_1 phonons[29]. The impact of compressive strain, for which Raman studies on MoS_2 are scarce, has been inferred from literature, suggesting that Raman frequency shifts due to tensile strain are 1.56 times greater than those due to compressive strain[30]. Doping effects are represented by a black line, with recent studies indicating that the $\omega_{A'_1}$ mode softens with electron accumulation but remains unaltered with hole doping[24]. Thus, the vector for electron doping in the low electron concentration region has been quantified as ($\omega_{E'}$, $\omega_{A'_1}$) is (-0.15, -1.19) cm^{-1}/1013 cm^{-2}[25] as shown in Figure 5(a). The hardening of $\omega_{A'_1}$, potentially caused by substitutional doping during the CVD growth process, is exemplified by shifts of -0.18 and 0.2 cm^{-1}/at% due to substitutional oxygen doping[31]. The scatter plot in Figure 5(b-d) of $\omega_{E'}$ and $\omega_{A'_1}$ which considers both strain and doping effects, reveals that the intrinsic point corresponds to higher PL energy, increased intensity, and reduced FWHM. This relationship is visualized with color coding that represents the PL energy, intensity, and FWHM across various data points.

Conclusion

In conclusion, we have demonstrated a framework for capturing the correlations between Raman and PL essential to tune MoS_2 optical properties by external perturbations to understand, predict, and design next-generation devices. We utilize the DenseNet model to build end-to-end connections from Raman spectral maps to photoluminescence. To gain more comprehensive insights into the physical mechanisms of strain and doping effects, we adopt the XGBoost model with the SHAP explainer and reveal that $\omega_{E'}$, $\omega_{A'_1}$, and $I_{E'}$ are the three dominant Raman characteristics for PL feature predictions, which further indicates that the strain effects govern the PL response more than the doping effects in our dataset. We further disentangle strain and doping effects and predict the location of the intrinsic point on the Raman frequency plot. The proposed methodology establishes an analysis approach to comprehensively interpret experimental observations to explore novel physics, which is suitable for Raman spectra and PL on 2D materials and for many other types of spectroscopies and condensed matter.

References

[1] J. Feng, X. Qian, C.-W. Huang, J. Li, *Nat. Photonics* **2012**, *6*, 866.

[2] A. V. Tyurnina, D. A. Bandurin, E. Khestanova, V. G. Kravets, M. Koperski, F. Guinea, A. N. Grigorenko, A. K. Geim, I. V. Grigorieva, *ACS Photonics* **2019**, *6*, 516.

[3] M. Amani, D.-H. Lien, D. Kiriya, J. Xiao, A. Azcatl, J. Noh, S. R. Madhvapathy, R. Addou, S. Kc, M. Dubey, K. Cho, R. M. Wallace, S.-C. Lee, J.-H. He, J. W. Ager, X. Zhang, E. Yablonovitch, A. Javey, *Science* **2015**, *350*, 1065.

[4] D.-H. Lien, S. Z. Uddin, M. Yeh, M. Amani, H. Kim, J. W. Ager 3rd, E. Yablonovitch, A. Javey, *Science* **2019**, *364*, 468.

[5] F. Withers, O. Del Pozo-Zamudio, A. Mishchenko, A. P. Rooney, A. Gholinia, K. Watanabe, T. Taniguchi, S. J. Haigh, A. K. Geim, A. I. Tartakovskii, K. S. Novoselov, *Nat. Mater.* **2015**, *14*, 301.

[6] O. Salehzadeh, M. Djavid, N. H. Tran, I. Shih, Z. Mi, *Nano Lett.* **2015**, *15*, 5302.

[7] J. G. Greener, S. M. Kandathil, L. Moffat, D. T. Jones, *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40.

[8] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, P. Kohli, *Nature* **2021**, *600*, 70.

[9] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.

[10] K. Tanaka, K. Hachiya, W. Zhang, K. Matsuda, Y. Miyauchi, *ACS Nano* **2019**, *13*, 12687.

[11] Y. Mao, N. Dong, L. Wang, X. Chen, H. Wang, Z. Wang, I. M. Kislyakov, J. Wang, *Nanomaterials (Basel)* **2020**, *10*, 2223.

[12] M. A. Haque, *BaselineRemoval: Python Package Code Repo for BaselineRemoval. It Has 3 Methods for Baseline Removal from Spectra for Baseline Correction, Namely ModPoly, IModPoly and Zhang Fit. The Functions Will Return Baseline-Subtracted Spectrum*, Github, **n.d.**

[13] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. Johansen, D. Johansen, M. Riegler, S. Ali, P. Halvorsen, *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2252.

[14] J. M. Ede, *Mach. Learn. Sci. Technol.* **2021**, *2*, 011004.

[15] B. Fang, Y. Li, H. Zhang, J. Chan, *Remote Sens. (Basel)* **2019**, *11*, 159.

[16] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, in 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2017**, pp. 2261-2269.

[17] O. Ronneberger, P. Fischer, T. Brox, *arXiv [cs.CV]* **2015**.

[18] V. Badrinarayanan, A. Kendall, R. Cipolla, *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481.

[19] V. Buhrmester, D. Münch, M. Arens, *arXiv [cs.AI]* **2019**.

[20] T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association For Computing Machinery, New York, NY, USA, **2016**, pp. 785-794.

[21] fernando, *BayesianOptimization: A Python Implementation of Global Optimization with Gaussian Processes*, Github, **n.d.**

[22] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, *Nat Mach Intell* **2020**, *2*, 56.

[23] C. Rice, R. J. Young, R. Zan, U. Bangert, D. Wolverson, T. Georgiou, R. Jalil, K. S. Novoselov, *Phys. Rev. B Condens. Matter Mater. Phys.* **2013**, *87*, DOI 10.1103/physrevb.87.081307.

[24] T. Sohier, E. Ponomarev, M. Gibertini, H. Berger, N. Marzari, N. Ubrig, A. F. Morpurgo, *Phys. Rev. X* **2019**, *9*, 031019.

[25] Z. Melnikova-Kominkova, K. Jurkova, V. Vales, K. Drogowska-Horná, O. Frank, M. Kalbac, *Phys. Chem. Chem. Phys.* **2019**, *21*, 25700.

[26] A. Michail, N. Delikoukos, J. Parthenios, C. Galiotis, K. Papagelis, *Appl. Phys. Lett.* **2016**, *108*, 173102.

[27] W. H. Chae, J. D. Cain, E. D. Hanson, A. A. Murthy, V. P. Dravid, *Appl. Phys. Lett.* **2017**, *111*, 143106.

[28] H. Li, Q. Zhang, C. C. R. Yap, B. K. Tay, T. H. T. Edwin, A. Olivier, D. Baillargeat, *Adv. Funct. Mater.* **2012**, *22*, 1385.

[29] H. Li, A. W. Contryman, X. Qian, S. M. Ardakani, Y. Gong, X. Wang, J. M. Weisse, C. H. Lee, J. Zhao, P. M. Ajayan, J. Li, H. C. Manoharan, X. Zheng, *Nat. Commun.* **2015**, *6*, 7381.

[30] S. Pak, J. Lee, Y.-W. Lee, A.-R. Jang, S. Ahn, K. Y. Ma, Y. Cho, J. Hong, S. Lee, H. Y. Jeong, H. Im, H. S. Shin, S. M. Morris, S. Cha, J. I. Sohn, J. M. Kim, *Nano Lett.* **2017**, *17*, 5634.

[31] J. Tang, Z. Wei, Q. Wang, Y. Wang, B. Han, X. Li, B. Huang, M. Liao, J. Liu, N. Li, Y. Zhao, C. Shen, Y. Guo, X. Bai, P. Gao, W. Yang, L. Chen, K. Wu, R. Yang, D. Shi, G. Zhang, *Small* **2020**, *16*, e2004276.

Dr. Ang-Yu LU

PhD Student,
Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology