Methods and objects of chemical analysis, 2025, Vol. 20, No. 3, 181–186

Prediction of Solvatochromic λmax of Pb(II) Complexes Using Gaussian Process Regression: A Tool for Analytical Applications

Khadija Abdel-Ilah Ali*†, Safa Majeed Hameed†, Hadeel Noori Saad‡**

†Department of Chemistry, Faculty of Education for Girls, University of Kufa, Najaf, 54001, Iraq;
‡Department of Computer Science, Faculty of Education for Girls, University of Kufa, Najaf, 54001, Iraq;
*e-mail: Khadijaa.alsaidi@student.uokufa.edu.iq; **e-mail: Safaa.alhassani@uokufa.edu.iq

Received: May 24, 2025; Accepted: July 25, 2025
DOI: 10.17721/moca.2025.181-186

This study presents the application of a Gaussian Process Regression (GPR) model for predicting the maximum absorption wavelength (λmax) of lead(II) complex with organic reagent [3-((2-chloro-4-hydroxyphenyl) diazenyl)-5nitrobenzene-1,2-diol] (CHPDND) extracted using various organic solvents. The proposed model utilizes key solvatochromatic parameters, including dielectric constant (ε), polarity/polarizability (π), hydrogen bond donor acidity (α), hydrogen bond acceptor basicity (β), and a custom polarity scale (ξ), to model the complex solvent effects on electronic transitions. The GPR model, equipped with a squared exponential kernel, demonstrated high predictive accuracy and efficiency, capturing the nonlinear relationships between solvent properties and λ_{max}. The model underwent automatic hyperparameter optimization, achieving convergence within minimal iterations and exhibiting stable performance. The findings were validated using both experimental and aggregated datasets, showing a strong agreement between predicted and actual values. The results emphasize the potential of machine learning approaches, particularly GPR, as reliable and efficient alternatives to traditional computational methods for optical property prediction. This work provides a practical framework for solvent selection and molecular design in the fields of analytical chemistry and materials science.*

Keywords: solvatochromism, λmax prediction, gaussian process regression, solvent effects, lead complexes

Introduction

Recent advancements in machine learning (ML) have significantly transformed the prediction of maximum absorption wavelengths (λmax) in solvatochromic systems, offering powerful alternatives to traditional quantum chemical methods that are often computationally demanding. The integration of ML approaches allows for more efficient and scalable modeling of solvent effects on electronic transitions, enabling accurate predictions based on molecular descriptors and solvent parameters.

A range of studies has demonstrated the potential of diverse ML methodologies in this context. A multifidelity ML framework that combines deep residual convolutional neural networks with gradient boosting algorithms. By integrating features derived from experimental measurements and density functional theory (DFT) calculations, their model achieved high predictive accuracy across a broad range of molecular structures [1]. Similarly, developed a hybrid approach that merges quantum chemical computations with ML techniques to predict UV–Vis absorption spectra of solvated aromatic compounds. Their model achieved a deviation of less than 0.1 eV from experimental data, emphasizing the value of ML in capturing complex solvent-induced spectral shifts [2]. In efforts to better account for solute–solvent interactions, constructed ML models incorporating molecular descriptors such as RDKit, alvaDesc, and solvent similarity metrics. Their application of Gaussian Process Regression (GPR) yielded accurate λmax predictions for both known and novel solute–solvent systems [3]. Further extending the utility of deep learning, a comprehensive database of solvated small-molecule fluorophores and applied deep neural networks to predict key photophysical properties, achieving a mean relative error of just 1.52% on out-of-sample predictions [4]. More recently, employed graph neural networks (GNNs) to predict molecular absorption spectra directly from molecular structures. Their model demonstrated strong performance across a diverse chemical space by effectively capturing both electronic and structural features relevant to optical absorption [5]. Additionally, demonstrated that shotgun transfer learning can enable property prediction even with small datasets, a crucial capability when experimental data for specific solute–solvent systems is limited [6].

Together, these studies underscore the growing role of ML in optical property prediction, particularly in solvatochromic contexts. The ability to model and predict λmax efficiently and accurately using data-driven techniques has important implications for molecular design, materials science, and computational spectroscopy.

Proposed GPR-solvatochromism system model

The proposed model emulates the relationship of the change in the position (and sometimes intensity) of the absorption or emission spectra of a molecule due to solvent polarity and specific solute-solvent interactions, and predicting the outcome variable (λmax — absorption maximum) based on solvent parameters:

ε (dielectric constant)
π* (polarity/polarizability)
α (hydrogen bond donor acidity)
β (hydrogen bond acceptor basicity)
ξ (a custom parameter)

To model this system, a Gaussian Process Regression (GPR) approach is adopted — a non-parametric, probabilistic machine learning technique well-suited for regression tasks involving small to moderate datasets and potentially non-linear, smooth relationships.

GPR mathematical model

A Gaussian Process (GP) defines a distribution over functions:

$$f(x) \sim GP(m(x), k(x,x')), \quad (1)$$

where m(x) is the mean function (often assumed to be zero), k(x,x′) is the kernel (covariance function) that expresses the similarity between inputs x and x′

Assuming a set of training inputs:

$$X=[x1,x2,\ldots,xn] \quad (2)$$

and observed targets:

$$y=[y1,y2,\ldots,yn]^T \quad (3)$$

Then the joint prior over observed targets y and the function values f∗ at test points X∗ is:

$$\begin{pmatrix} y \\ f_* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K(X,X)+\sigma_n^2 I & K(X_*,X) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right) \quad (4)$$

where *K(X,X)* is the covariance matrix for training points, σ_n^2 is noise variance.

In GPR, the kernel function governs how points in the input space relate to one another in terms of output similarity. It essentially encodes assumptions about the function's smoothness and variability. In this model, the Squared Exponential (SE) kernel — also known as the Radial Basis Function (RBF) kernel — is selected due to its strong theoretical basis and capability to model highly smooth and continuous functions. The SE kernel is defined as:

$$k(x_i + x_j) = \sigma_f^2 exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{id}-x_{jd})^2}{I_d^2}\right) \quad (5)$$

where σ_f^2 is the signal variance, l_d is the characteristic length-scale for dimension d, D is the number of input parameters.

Predictive distribution

For new, unseen test points *X_****, the predictive distribution is derived by conditioning on the observed data. This distribution is Gaussian, fully specified by its predictive mean and variance, The predictive distribution at test points X_* is given by Eq.6-8.

This gives both a prediction and an uncertainty estimate for each test point.

GPR-solvatochromism prediction algorithm

This study employs a GPR-based algorithm for predicting target outcomes from given set of input parameters. The procedure initiates by identifying, and training valid, non-missing outcome data to ensure the integrity of the modeling process, subsequently, a GPR model is constructing using a *Squared Exponential kernel*, which is well regarded for its flexibility in capturing smooth, non-linear relationships within data.

The model's hyperparameters specifically, the signal variance (σ_f), length scale (l_d), and noise variance (σ_n) are optimized through the maximization of the marginal likelihood function, enabling an adaptive to the underlying data distribution. Following model training, predictions are generated for all data points within dataset.

To systematically assess predictive performance, the predicted outcomes are partitioned into discrete intervals (bins), and corresponding accuracy metrics are computed for each bin. This bin wise evaluation provides nuanced understanding of the model's predictive reliability across varying ranges of the outcome variables. The final outputs of this algorithm include the optimized GPR model, a complete set of predictions, and detailed accuracy evaluations per bin.

$$p(f_*|X,y,X_*) = N\left(\bar{f}_*, Var[f_*]\right) \quad (6)$$

where:

$$\bar{f}_* = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1} y \quad (7)$$

and

$$Var[f_*] = K(X_*,X_*) - K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1} K(X,X_*) \quad (8)$$

Algorithm Description (Pseudocode-style)

Input:

Parameters matrix X (ε, T*, α, β, ξ)

Outcomes vector y (λmax)

Process:

Identify non-missing outcomes (filter valid data).

Fit Gaussian Process Regression model:

Use Squared Exponential kernel

Optimize hyperparameters (σ_f, l_d, σ_n) automatically via likelihood maximization.

Predict outcomes for all data points.

Evaluate accuracy by dividing outcomes into bins and computing:

Output:

- Fitted GPR model
- Predictions for all data points
- Accuracy metrics per bin

Dataset aggregation

Due to the lack of available data with 5 chromatic parameters and λmax, in this section we will explore an aggregated dataset as a first contribution of this work, which was collected from [7-11] for the peruse of training across experiments. Table S1 shows set of 27 referenced solvent.

Lab experimental dataset

This section describes the dataset collected under controlled experiments to evaluate the reliability of the proposed system objectively. Below is Table S2 of the experimental 7 solvent dataset.

Test dataset

Test data is an important step to exploit the trained system unbiased, generalization ability to predict beyond trained data, for unseen inputs, and ensure the system ability to produce reliable and accurate outcomes. A set of 244 organic solvents was included in this work for testing, because of the length of test dataset, Table S3 illustrates the statistics of data. The hole dataset will be attached at appendix A for further details.

Experiments

This work was implemented under Matlab (R2021b) and Processor (Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz, 2112 Mhz). An extensive experiment is conducted to gain the most accurate results. Table S4 shows the proposed GPR prediction model kernel parameters.

Results and disscussion

The experimental evaluation was conducted over 30 objective function evaluations, with a total elapsed time of 24.30 seconds and a cumulative function evaluation time of 4.30 seconds. The trajectory and convergence of the objective function across iterations are shown in Figure 1. The best observed feasible point was achieved at Sigma = 65.205, corresponding to an objective function value of 8.2403, while its estimated value, according to the model, was 8.2609 with a function evaluation time of 0.0756 seconds. In parallel, the best estimated feasible point identified by the surrogate models was located at Sigma = 28.025, yielding an estimated objective function value of 8.2563, with a slightly higher evaluation time of 0.0912 seconds.

To further assess the predictive performance of the model, accuracy metrics were computed based on outcome bins. In Bin 4 (331.30–414.40), comprising 19 data points, the model achieved a Root Mean Square Error (RMSE) of 7.7473, a Mean Absolute Error (MAE) of 5.8873, and a coefficient of determination (R^2) of 0.7726, indicating strong predictive alignment within this range. In Bin 5 (414.40–497.50), with 12 data points, the model recorded an RMSE of 6.5354, an MAE of 5.3569, and a lower R^2 value of 0.4998, suggesting moderate predictive accuracy in higher outcome ranges. These results collectively

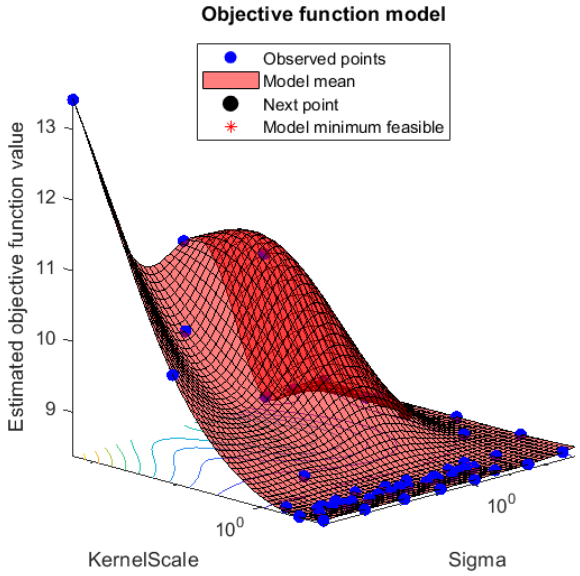

Figure 1. Predicted values of the objective function.

demonstrate the model's capability to deliver reliable predictions, particularly in mid-range outcome bins, while highlighting areas for further refinement at higher outcome levels. Table S5 shows the GPR-solvetochromatic predictions model results with three testing phases.

Training proposed model behavior

GPR-solvatochromatic, this model was configured to automatically adjust its parameters to improve accuracy, based on 60 iterations conducted in less than one minute, the observed system behavior during training was analyzed during three Iteration categories:

Initial iterations. Large jumps were observed in both Sigma and Kernel Scale values. The objective function started high at 10.024, then dropped significantly to 8.4782 by the second iteration. Rapid early improvements occurred as the optimizer sampled exploration points with large parameter ranges.

Intermediate iterations. Gradual narrowing around lower objective function values. The best value found so far was 8.3816 at iteration 19. Continued testing within the neighboring region with minor changes in Sigma and Kernel Scale. Some exploratory jumps still occurred, notably in iterations 13 and 16 with large Kernel Scale values.

Later iterations. Clear convergence behavior: the objective function hovered around 8.3803 - 8.379. The best value slightly improved to 8.3790 by iteration 36. Very minimal changes after that — indicating that the optimizer had nearly fully converged.

Analysis and Discussion

As reflected in the model outcomes, the optimization process demonstrated good convergence, with the best objective function value stabilized between iterations 36-45 with only minor fluctuations, indicating that the parameter space was well explored. This behavior indicates that the parameter space was effectively explored. A three-dimensional visualization of the actual and predicted data distributions is presented in Figure 2a,b, which provides a direct comparison of the solvatochromatic system response surface.

The proposed model at the early iterations emphasized extensive exploration characterized by large Sigma and Kernel Scale values, while later iterations shifted towards exploitation within a narrow parameter region. Additionally, the runtime per iteration remained consistent between (~0.05 - 0.08 seconds), which is efficient for hyperparameter tuning loops.

Throughout all experiments, the Kernel Scale values converged to stable range between 0.35 and 0.50 during the later iterations, indicating that this range consistently yields good objective function values. While Sigma values showed greater fluctuations initially, the later iterations demonstrated stability (values like 0.0009, 0.0014, etc.), This pattern suggests a diminishing need for substantial noise adjustments in the model as the optimization progressed, further confirming the robustness of the proposed approach.

To validate the physical relevance of the model, Figure 3 shows the actual solvatochromatic parameter values compared to the maximum absorption wavelength (λmax). In contrast, Figure 4 presents the corresponding predicted values obtained from the GPR model, highlighting the consistency and predictive power of the approach

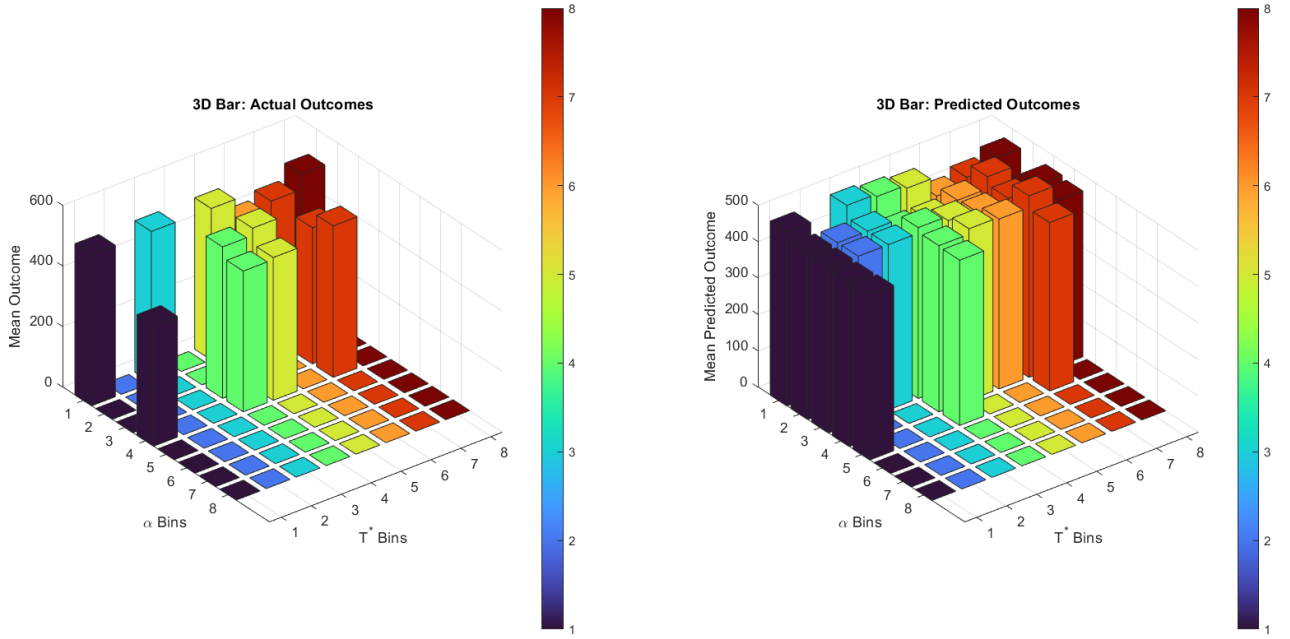

Figure 2. 3D graphical representation of actual and predicted data.

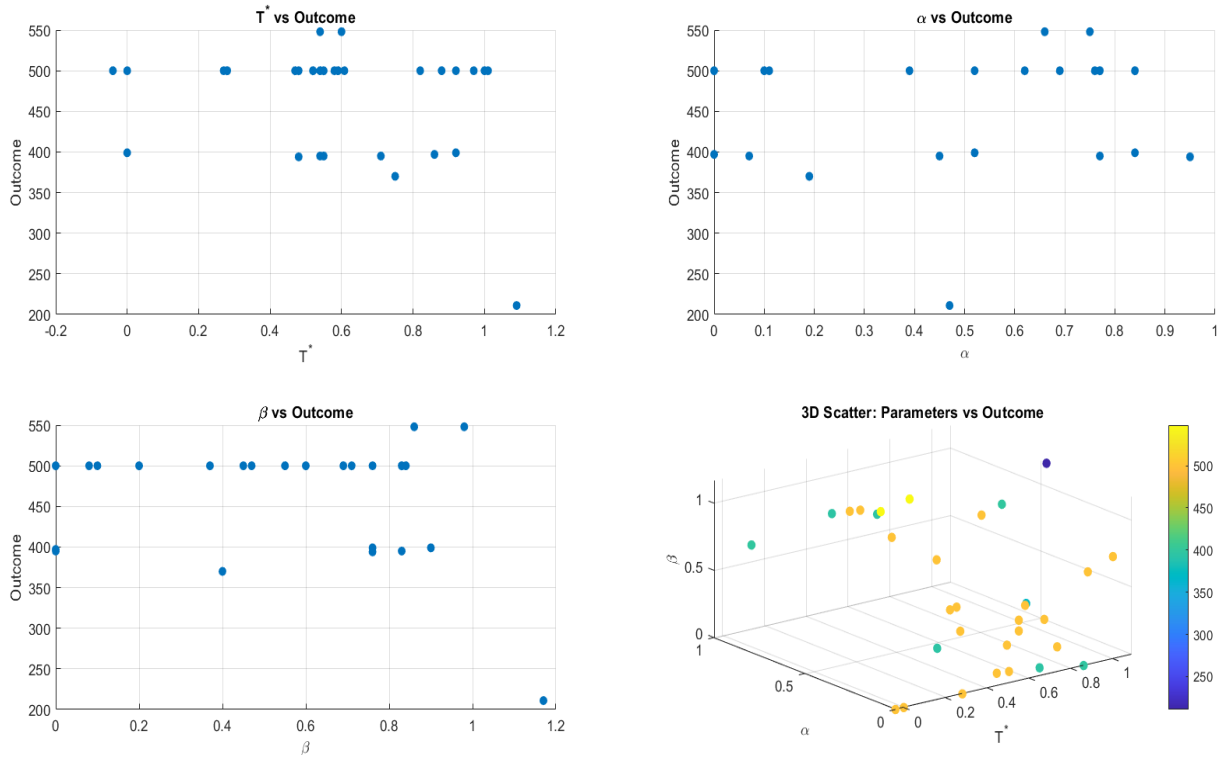

Figure 3. Actual solvatochromatic parameter calues compared to the maximum absorption wavelength (λmax).

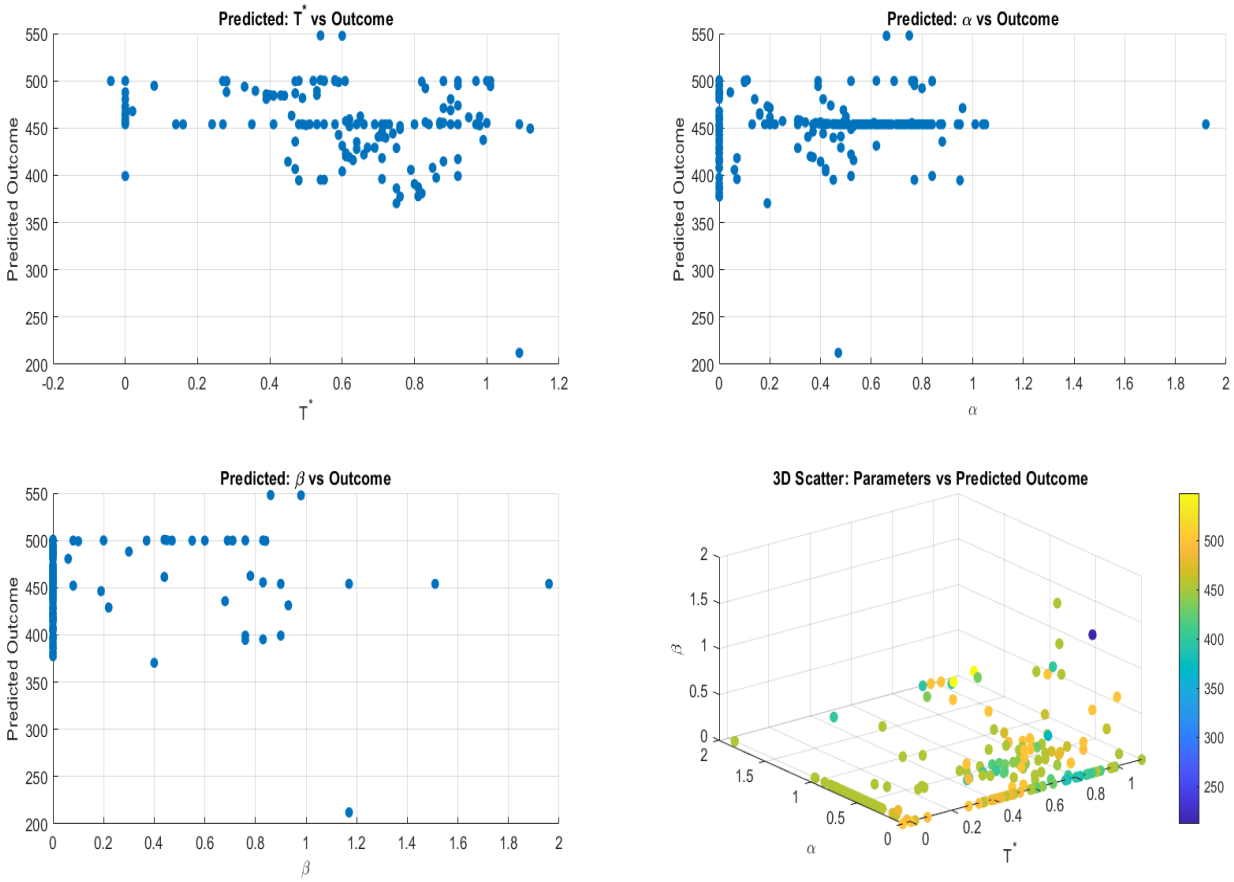

Figure 4. Predicted solvatochromatic parameter values compared to the maximum absorption wavelength (λmax).

Conclusion

In this study, a Gaussian Process Regression (GPR) model was successfully implemented to predict the maximum absorption wavelength (λmax) of Solvatochromatic systems, specifically for lead complexes extracted using various organic solvents. The model relied on key Solvatochromatic parameters, including dielectric constant (ε), polarity/polarizability (π*), hydrogen bond donor acidity (α), hydrogen bond acceptor basicity (β), and a custom polarity scale (ξ).

The results demonstrated that the GPR model, utilizing a squared exponential kernel, effectively captured the complex, nonlinear relationships between solvent properties and λmax. The iterative optimization process revealed clear convergence behavior, with minimal parameter fluctuations observed in later stages of training, indicating a well-explored parameter space. Furthermore, the model showed consistent predictive performance, supported by statistical evaluation metrics, and provided reliable estimates for both known and novel solvent systems.

The proposed model highlights the potential of machine learning, particularly non-parametric probabilistic models like GPR, in replacing or complementing traditional experimental or computationally expensive quantum chemical methods. Such models can play a significant role in accelerating solvent selection and molecular design processes in analytical chemistry and materials science applications.

References

1. Greenman, K. P., Ju, C., Yang, K., Swanson, K., Jin, W., Coley, C. W., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. F., & Barzilay, R. (2024). Automatic prediction of peak optical absorption wavelengths in molecules using convolutional neural networks. *Journal of Chemical Information and Modeling, 64*(4), 1234–1245. https://doi.org/10.1021/acs.jcim.3c01792

2. Chen, Z., Bononi, F. C., Sievers, C. A., Kong, W.-Y., & Donadio, D. (2022). UV–visible absorption spectra of solvated molecules by quantum chemical machine learning. *Journal of Chemical Theory and Computation, 18*(8), 4891–4902. https://doi.org/10.1021/acs.jctc.1c01181

3. Kobayashi, Y., Tanaka, H., Saito, A., & Nakamura, Y. (2025). Construction of machine learning models to predict the maximum absorption wavelength considering the solute and solvent and inverse analysis of the models. *ACS Omega, 10*(1), 665–672. https://doi.org/10.1021/acsomega.4c07490

4. Shao, J., Liu, Y., Yan, J., Yan, Z.-Y., Wu, Y., Ru, Z., Liao, J.-Y., Miao, X., & Qian, L. (2022). Prediction of maximum absorption wavelength using deep neural networks. *Journal of Chemical Information and Modeling, 62*(6), 1368–1375. https://doi.org/10.1021/acs.jcim.1c01449

5. Zhou, Y., Wang, Z., Chen, Y., Zhang, Q., Liu, Z., & Tang, J. (2023). Predicting molecular absorption spectra with graph neural networks. *Nature Machine Intelligence, 5*(11), 945–954. https://doi.org/10.1038/s42256-023-00677-2

6. Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., & Yamaguchi, S. (2019). Predicting materials properties with little data using shotgun transfer learning. *ACS Central Science, 5*(10), 1717–1730. https://doi.org/10.1021/acscentsci.9b00447

7. Jessop, P. G., Jessop, D. A., Fu, D., & Phan, L. (2012). Solvatochromic parameters for solvents of interest in green chemistry. Green Chemistry, 14(5), 1245–1259.RSC Publishing.

8. Reichardt, C. (1983). Solvatochromic parameters for solvents. The Journal of Organic Chemistry, 48(17), 2877–2887.ACS Publications.

9. Effect of solvent polarity on the photophysical properties of chalcone derivatives. RSC Advances, 2017, 7, 12382–12390.RSC Publishing.

10. Solvatochromic and Computational Study of Some Cycloimmonium Ylids. Molecules, 2023, 28(1), 9.MDPI.

11. Solvatochromic and Single Crystal Studies of Two Neutral Triarylmethane Dyes with a Quinone Methide Structure. Molecules, 2015, 20(11), 19724–19738.MDPI