Available online at www.sciencedirect.com

ScienceDirect

Mendeleev Commun., 2024, **34**, 788–791

Mendeleev Communications

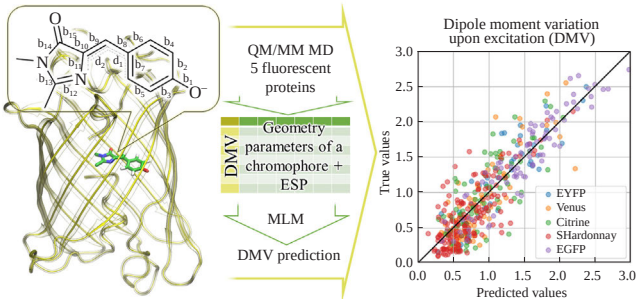
Towards machine learning prediction of the fluorescent protein absorption spectra

Roman A. Stepanyuk,[a,b] Igor V. Polyakov,[a,c] Anna M. Kulakova,[a] Ekaterina I. Marchenko[d] and Maria G. Khrenova*[a,b]

[a] *Department of Chemistry, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation. E-mail: khrenovamg@my.msu.ru*
[b] *A. N. Bach Institute of Biochemistry, Federal Research Centre 'Fundamentals of Biotechnology' of the Russian Academy of Sciences, 119071 Moscow, Russian Federation*
[c] *N. M. Emanuel Institute of Biochemical Physics, Russian Academy of Sciences, 119334 Moscow, Russian Federation*
[d] *Department of Materials Science, M. V. Lomonosov Moscow State University, 119991 Moscow, Russian Federation*

DOI: 10.1016/j.mencom.2024.10.007

We demonstrate that machine learning models trained on a set of features obtained from QM/MM molecular dynamic trajectories of fluorescent proteins can be used to predict the chromophore dipole moment variation upon excitation, the quantity related to the electronic excitation energy. Linear regression, gradient boosting, and artificial neural network-based models were considered using cross-validation on the training dataset. Gradient boosting approach proved to be the most accurate for both internal (R^2 = 0.77) and external (R^2 = 0.7) test sets.



Keywords: machine learning, fluorescent proteins, QM/MM molecular dynamics, dipole moment variation upon excitation.

Fluorescent proteins are popular tools for visualization of processes in living cells and tissues,[1,2] including super-resolution microscopy.[3,4] A chromophore group is formed autocatalytically in the protein β-barrel from three amino acid residues of the protein (Figure 1). The first discovered fluorescent protein GFP has green fluorescence, but now, a fluorescent protein palette covers whole visible regions.[5] One of the practical tasks in this field is to modify the protein to get advanced photochemical properties in the required spectral range. There is already a number of works that utilize artificial intelligence (AI) for fluorescent protein development.[6–11] Alternatively, predictive models based on physical background can be suggested. Fluorescent proteins exhibit a second-order Stark effect, which manifests itself in a quadratic dependence of the electronic excitation energy on the dipole moment variation (DMV) upon excitation.[12–14] Such dependences were also demonstrated for green and red fluorescent protein families in QM/MM based simulations: the energy corresponding to the maximum of the experimental absorption band demonstrates a quadratic dependence on the calculated DMV.[15,16] By now, such calculations were performed exclusively for minima on the potential energy surface. Calculation of the DMV along molecular dynamic (MD) trajectories can derive to the DMV distribution that can, principally, be converted to the absorption spectrum band. To get a representative distribution, one needs to calculate the DMV for a set of thousands points. The DMV calculation even at the TDDFT level of theory is a time-consuming task. In this regard, the attractive alternative is to obtain correlations between the DMV and ground electronic state geometry parameters that can be extracted on the fly from the MD trajectory.

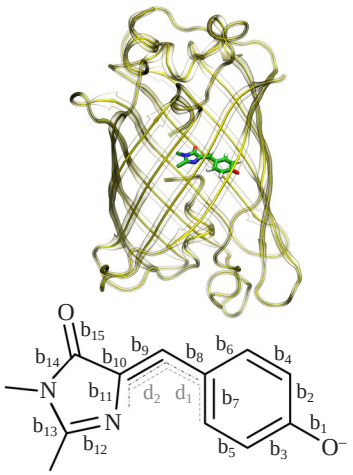


Figure 1 β-Barrel structure of the fluorescent protein: chromophore is shown by sticks (upper panel). Chemical structure of the anionic GFP chromophore; b and d notations correspond to bonds and dihedral features utilized in machine learning models (lower panel).

Herein, we aim to solve the regression problem of predicting DMV values using machine learning methods (MLMs). The dataset is composed of descriptors derived from MD simulations with the combined quantum mechanics/molecular mechanics

(QM/MM) potentials of model systems representing five fluorescent proteins: EGFP, EYFP, Venus, Citrine, and SHardonnay with the same 4-(*p*-hydroxybenzylidene)-5-imidazolinone chromophore. We compare different MLMs to discriminate the most suitable one and to reveal descriptors mostly contributing to the DMV values.

We performed QM/MM MD simulations† for all considered systems and obtained sets of states represented by MD frames for each protein. As an initial set of descriptors, we identified bond lengths in the fluorophore fragment and dihedral angles between the planes of fluorophore rings (see Figure 1). All geometry parameters demonstrate normal distributions (Figure S1 and S2, see Online Supplementary Materials). The effect of protein was taken into account by electrostatic potential (ESP) calculated on the chromophore atoms (Figure S3). ESP values include contributions from all protein atoms and solvents that could be the measure of the discrimination of a particular protein. ESP values calculated at different atoms highly correlate with each other (Figure S4). Therefore, we utilize only the sum of ESP values calculated on all chromophore atoms at each MD frame.

† *Computational protocol*. Full-atom model systems of 5 fluorescent proteins were constructed using the available X-ray structures of proteins: EGFP (PDB ID: 4EUL),[18] SHardonnay (PDB ID: 3V3D)[19] and its variant EYFP (Shardonnay with the F203Y amino acid substitution), Venus (PDB ID: 1MYW),[20] and Citrine (PDB ID: 1HUY).[21] Hydrogen atoms were added using the Reduce program[22] and manually checked. The model system was solvated in a rectangular cell and neutralized. The CHARMM36 force field[23,24] was used to describe the chromophore and protein macromolecule, and TIP3P[25] was used for water molecules. Classical molecular dynamic simulations were performed for the initial system equilibration in the NAMD program for 5 ns.[26] All MD calculations with classical and combined potentials were performed in the canonical NPT ensemble at $p = 1$ atm and $T = 300$ K with a 1 fs integration step. A representative frame from the classical trajectory was selected for each model system and used as a starting structure for molecular dynamic simulations with QM/MM potentials. The length of each QM/MM MD trajectory was 11 ps and the first 1 ps was excluded from the following analysis. The QM subsystem consisted of a chromophore, neighboring amino acid residues, and water molecules of the chromophore-containing pocket. The Kohn–Sham DFT approach was applied for the QM part with the hybrid functional PBE0[27] with empirical corrections for dispersion interactions D3[28] and a cc-pvdz basis set. Calculations of energies and forces in the QM subsystem were carried out in the TeraChem program;[29] the MM part and MD step calculations were performed in the NAMD program using a special interface.[30]

TDDFT with a hybrid functional recommended for calculations of electronic transitions ωB97X-D3[31] and cc-pvdz basis set was utilized to calculate the dipole moment variation upon excitation (DMV). Calculations of excited states were carried out using the ORCA program.[32] The DMV was calculated for the transition from the ground singlet state to the lowest excited state with a large oscillator strength. To calculate the dipole moment variation upon excitation, 400 frames of QM/MM MD trajectories were selected from the last 10 ps (every 25 fs). QM/MM calculations in the ground state and vertical electron transitions were carried out in the electron embedding variant, *i.e.*, the charges of the MM environment contributed to the one-electron part of the QM Hamiltonian.

Models were trained with three different dataset splits: on individual proteins with a test set (20% points of the protein data); a training set of 5 proteins with a pre-selected internal test set (20% of each protein); a training set of 4 proteins with an external test set sampling (EGFP protein). For linear models the data was scaled through StandardScaler. The 4-fold cross validation for all models is used. The hyperparameter optimization was performed through GridSearchCV. Objects necessary for training linear models were taken from the scikit-learn library.[33] LGBMRegressor from LightGBM[34] and TabNetRegressor from pytorch-tabnet[35] were used. Machine learning protocol code is deposited on Zenodo, https://doi.org/10.5281/zenodo.11393469.

The dataset for each protein contains 400 points with calculated descriptors and DMV values. This number of data points is not enough to utilize sophisticated methods for each individual protein. Thus, we utilize scaling and compare the linear regression (LR) model. We perform these calculations to demonstrate that selected features contribute to the DMV prediction. The ESP feature proved to be irrelevant for the linear models of the individual proteins (Figure S5). The basic model, such as linear regression, can predict the DMV with a selected feature set with great success ($R^2 > 0.9$) for EGFP and limited ($R^2 < 0.5$) for Venus. Application of more complex models requires a larger dataset. Also, the unified model for prediction of the DMV for different proteins with the same chromophore is required. Therefore, we combined datasets for all considered systems in a single dataset and performed the following MLM analysis.

For a unified model based on 5 proteins, the 4-fold cross-validation MAE for linear models is about 0.288. Approaches with regularization do not provide any improvement on this dataset. Two bridging bonds (bond 8, bond 9) and neighboring bonds including two C–C bonds from the 6-membered ring (bond 6, bond 7), and one C–C bond from the 5-membered ring (bond 10) are the most important features of the linear model (Figures 1, S6). This is in line with the previous observations that for stationary points the excitation energy is mainly determined by bridging bond lengths.[13,16,17] The ESP sum is not among important features similarly to individual models. Gradient boosting with the LightGBM (LGBM) yields cross-validation MAE values of 0.242 without ESP, and a slightly smaller MAE of 0.235 after adding the ESP feature. Further testing of the LGBM model on the internal test set yields even better metric values: MAE = 0.23 and $R^2 = 0.77$. The internal test results for each protein are shown in Figure 2. It demonstrates that quality of the LGBM model prediction is not the same for all the proteins which is probably related to the DMV distribution of individual proteins.

The SHAP analysis of feature importance for LGBM (Figure 3) demonstrates agreement with the linear model coefficients, but the ESP sum feature is on the 4th place in importance, and improves the quality of the unified 5-protein LGBM model.

To test the ability of DMV prediction for new protein variants with the same chromophore and amino acid substitutions in the protein β-barrel, an external test was conducted with 4 proteins, EYFP, Venus, Citrine, and SHardonnay, as training and validation dataset and the EGFP data as a test. This division is of practical interest, as the model is trained on yellow fluorescent proteins with additional stacking interactions between the chromophore and a side chain of an aromatic amino acid residue and the EGFP lacks these interactions. Figure 4 depicts distributions of the

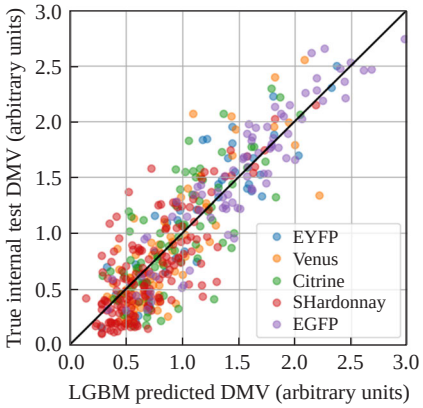


Figure 2 LGBM model prediction for internal test sets for 5-protein LGBM model.

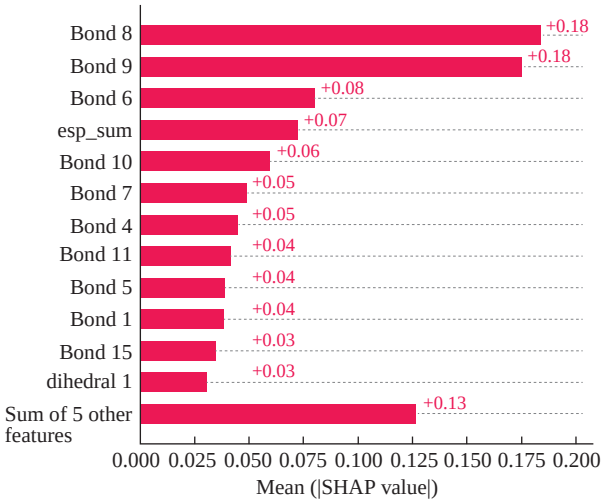


Figure 3 SHAP analysis for the unified 5-protein LGBM model.

DMV in all datasets. The DMV distribution of the EGFP is shifted relative to other distributions. From the machine learning side, this indicates that the DMV prediction for the EGFP from the model trained on yellow proteins might be not accurate as the predicting value distribution for the test protein does slightly overlap with the distributions for the training proteins. Still, the model was able to predict the shift of the EGFP distribution to the larger values that correspond to the larger excitation energies according to the quadratic Stark effect. This is in line with the experimental observation that the absorption band of the EGFP is shifted to the shorter wavelength region compared with yellow proteins.

In cross-validation, LBGM has a lower MAE value of 0.25 compared to LR (MAE = 0.3). On the validation set, LGBM with the additional ESP feature did not provide any improvement precisely because its distribution for the test protein (external test) practically does not overlap with the ESP distribution area of the training set (Figure S3). Nevertheless, even the LGBM model based only on geometry features gives quite satisfactory results on the external test set (MAE = 0.29 and R^2 = 0.70). These predicting quantities are slightly worse compared with the prediction for the internal dataset performed on 5-protein model (R^2 = 0.77). A comparison of predictions for the external and internal test sets is illustrated in Figure 5.

We do understand that for such a small feature and dataset artificial neural networks (ANNs) are not the best modeling tools, yet we try an ANN model for external set value prediction. We employ the TabNet neural network architecture designed for tabular learning. The cross-validation results were close to the LGBM model (MAE = 0.25), but the test MAE = 0.429 and

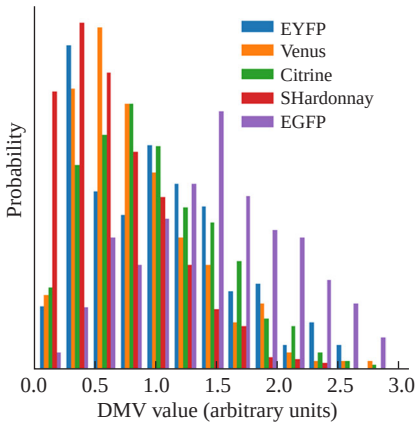


Figure 4 DMV distribution in training datasets.

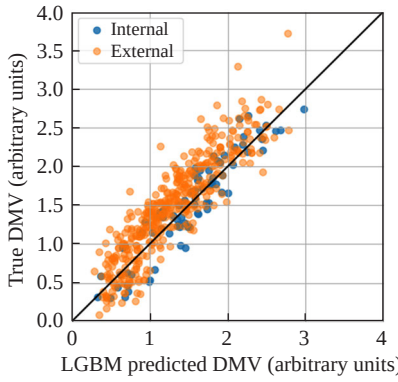


Figure 5 Results of LGBM test prediction. The internal test set refers to a 5-protein model. The external test set (dataset for EGFP protein) predictions were performed with the 4-protein model trained on proteins except EGFP.

R^2 = 0.327 proved that such model is not applicable. Nevertheless, we suppose that in the future, given larger datasets of protein variants and more features attributed to local chromophore environment, the ANN models would be more useful to predict DMV distributions and excitation spectra for new protein variants. Local environment features can include stacking interactions with the chromophore (distance between centers of rings, dihedral angle between ring plains), positions of the most important charged residues relative to the chromophore, number and average distance of hydrogen bonds to oxygen and nitrogen atoms of the chromophore, *etc*.

To conclude, we demonstrate that machine learning approaches based on geometry and ESP descriptors can predict the DMV values of fluorescent proteins, which is applicable in the field of rational design of new fluorescent proteins. We obtained high accuracy gradient boosting model. Based on the metrics obtained on the internal test set, such a model can be utilized to predict additional DMV values from the available geometry parameters and the ESP on the chromophore atoms from molecular dynamic trajectories. This reduces the computational cost of simulations as it requires only the ground electronic state QM/MM MD trajectory and a relatively small training set with explicitly calculated DMV values. Prediction of DMV values for new proteins is also possible, still it has potentially more challenges in both obtaining stable models and determining their applicability domain. Now, we see promise in introducing additional descriptors of the local environment, which would allow us to take into account the influence of a specific protein environment on a chromophore fragment more accurately. Additionally, an increase in datasets will broaden the variety of applicable machine learning methods.

This work was supported by Interdisciplinary Scientific and Educational School of Moscow State University ‘Brain, cognitive systems, artificial intelligence’ (#23-Sh03-04). The research was carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2024.10.007.

References

1 H. Shinoda, M. Shannon and T. Nagai, *Int. J. Mol. Sci.*, 2018, **19**, 1548.
2 R. N. Day and M. W. Davidson, *Chem. Soc. Rev.*, 2009, **38**, 2887.
3 K. I. Willig, W. Wegner, A. Müller, V. Clavet-Fournier and H. Steffens, *Cell Rep.*, 2021, **35**, 109192.
4 J. Lippincott-Schwartz and G. H. Patterson, *Trends Cell Biol.*, 2009, **19**, 555.

5 A. Acharya, A. M. Bogdanov, B. L. Grigorenko, K. B. Bravaya, A. V. Nemukhin, K. A. Lukyanov and A. I. Krylov, *Chem. Rev.*, 2017, **117**, 758.
6 Y. Saito, M. Oikawa, H. Nakazawa, T. Niide, T. Kameda, K. Tsuda and M. Umetsu, *ACS Synth. Biol.*, 2018, **7**, 2014.
7 B. J. Wittmann, K. E. Johnston, Z. Wu and F. H. Arnold, *Curr. Opin. Struct. Biol.*, 2021, **69**, 11.
8 K. K. Yang, Z. Wu and F. H. Arnold, *Nat. Methods*, 2019, **16**, 687.
9 C. Tam and K. Y. J. Zhang, *Proteins: Struct., Funct., Bioinf.*, 2022, **90**, 732.
10 L. Gonzalez Somermeyer, A. Fleiss, A. S. Mishin, N. G. Bozhanova, A. A. Igolkina, J. Meiler, M.-E. Alaball Pujol, E. V. Putintseva, K. S. Sarkisyan and F. A. Kondrashov, *eLife*, 2022, **11**, e75842.
11 M. Li, L. Kang, Y. Xiong, Y. G. Wang, G. Fan, P. Tan and L. Hong, *J. Cheminf.*, 2023, **15**, 12.
12 M. Drobizhev, S. Tillo, N. S. Makarov, T. E. Hughes and A. Rebane, *J. Phys. Chem. B*, 2009, **113**, 12860.
13 M. Drobizhev, P. R. Callis, R. Nifosì, G. Wicks, C. Stoltzfus, L. Barnett, T. E. Hughes, P. Sullivan and A. Rebane, *Sci. Rep.*, 2015, **5**, 13223.
14 C.-Y. Lin, M. G. Romei, L. M. Oltrogge, I. I. Mathews and S. G. Boxer, *J. Am. Chem. Soc.*, 2019, **141**, 15250.
15 R. Nifosì, B. Mennucci and C. Filippi, *Phys. Chem. Chem. Phys.*, 2019, **21**, 18988.
16 M. G. Khrenova, F. D. Mulashkin, E. S. Bulavko, T. M. Zakharova and A. V. Nemukhin, *J. Chem. Inf. Model.*, 2020, **60**, 6288.
17 C.-Y. Lin and S. G. Boxer, *J. Am. Chem. Soc.*, 2020, **142**, 11032.
18 J. A. J. Arpino, P. J. Rizkallah and D. D. Jones, *PLoS One*, 2012, **7**, e47132.
19 E. De Meulenaere, N. Nguyen Bich, M. de Wergifosse, K. Van Hecke, L. Van Meervelt, J. Vanderleyden, B. Champagne and K. Clays, *J. Am. Chem. Soc.*, 2013, **135**, 4061.
20 A. Rekas, J.-R. Alattia, T. Nagai, A. Miyawaki and M. Ikura, *J. Biol. Chem.*, 2002, **277**, 50573.
21 O. Griesbeck, G. S. Baird, R. E. Campbell, D. A. Zacharias and R. Y. Tsien, *J. Biol. Chem.*, 2001, **276**, 29188.
22 J. M. Word, S. C. Lovell, J. S. Richardson and D. C. Richardson, *J. Mol. Biol.*, 1999, **285**, 1735.
23 R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig and A. D. MacKerell, Jr., *J. Chem. Theory Comput.*, 2012, **8**, 3257.
24 E. J. Denning, U. D. Priyakumar, L. Nilsson and A. D. Mackerell, Jr., *J. Comput. Chem.*, 2011, **32**, 1929.
25 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926.
26 J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, *J. Chem. Phys.*, 2020, **153**, 044130.
27 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158.
28 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
29 S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, C. M. Isborn, S. I. L. Kokkila-Schumacher, X. Li, F. Liu, N. Luehr, J. W. Snyder, Jr., C. Song, A. V. Titov, I. S. Ufimtsev, L.-P. Wang and T. J. Martínez, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2021, **11**, e1494.
30 M. C. R. Melo, R. C. Bernardi, T. Rudack, M. Scheurer, C. Riplinger, J. C. Phillips, J. D. C. Maia, G. B. Rocha, J. V Ribeiro, J. E. Stone, F. Neese, K. Schulten and Z. Luthey-Schulten, *Nat. Methods*, 2018, **15**, 351.
31 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615.
32 F. Neese, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 73.
33 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 É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825.
34 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, 2017, pp. 3149–3157.
35 M. Joseph, *PyTorch Tabular: A Framework for Deep Learning with Tabular Data*, 2021, https://doi.org/10.48550/arXiv.2104.13638.

Received: 31st May 2024; Com. 24/7514