

Deep neural network model for highly accurate prediction of BODIPYs absorption

Alexander A. Ksenofontov¹, Michail M. Lukanov^{1,2}, Pavel S. Bocharov^{1,2}, Michail B. Berezin¹, Igor V. Tetko^{1,3,4}

¹ *G.A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences, Akademicheskaya Street, 153045, Ivanovo, Russia*

² *Ivanovo State University of Chemistry and Technology, 7, Sheremetevskiy Avenue, Ivanovo, 153000, Russia*

³ *Helmholtz Zentrum München-German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany*

⁴ *BIGCHEM GmbH, Valerystr. 49, 85716 Unterschleißheim, Germany*

Abstract

A possibility to accurately predict the absorption maximum wavelength of BODIPYs was investigated. We found that previously reported models had a low accuracy (40-57 nm) to predict BODIPYs due to the limited dataset sizes and/or number of BODIPYs (few hundreds). New models developed in this study were based on data of 6000-plus fluorescent dyes (including 4000-plus BODIPYs) and the deep neural network architecture. The high prediction accuracy (five-fold cross-validation root mean squared error (RMSE) of 18.4 nm) was obtained using a consensus model, which was more accurate than individual models. This model provided the excellent accuracy (RMSE of 8 nm) for molecules previously synthesized in our laboratory as well as for prospective validation of three new BODIPYs. We found that solvent properties did not significantly influence the model accuracy since only few BODIPYs exhibited solvatochromism. The analysis of large prediction errors suggested that compounds able to have intermolecular interactions with solvent or salts were likely to be incorrectly predicted. The consensus model is freely available at <https://ochem.eu/article/134921> and can help the other researchers to accelerate design of new dyes with desired properties.

Keywords

QSPR, BODIPY, absorption maximum wavelength, deep neural networks, OCHEM

Introduction

The BODIPY (4,4-difluoro-4-bora-3a,4a-diaza-s-indacene) compounds are widely used as bio- and molecular sensors [1] due to their unique spectral properties. The BODIPY sensors are important tools in biochemistry for the precious investigation of various biosystems [2]. Therefore, it is important to synthesize BODIPYs with a desired set of physical and chemical properties to solve specific application tasks. The design of molecules with required properties is usually carried out using well-targeted synthetic strategies and based on the expertise of the respective chemists. This includes the use of fundamental knowledge about the properties of a fluorophore, the results of quantum-chemical calculations and the chemical intuition of the scientists. Such approaches are, however, time-consuming and their success critically depends on the expertise of the chemists. The rapid development of data processing and data prediction methods, such as quantitative structure-activity/property relationship (QSAR/QSPR) approaches, allows to dramatically speed up this process. These methods are based on the assumption that properties of molecules could be directly derived from molecular structures which are represented as a set of calculated molecular descriptors and properties. Different linear and non-linear machine learning methods are frequently used to identify such dependencies [3]. Amid these methods, artificial neural networks (ANNs) have received a lot of attention. A pioneering article to demonstrate the effectiveness of using this method to predict the position of the long-wavelength absorption band of symmetric cyanine dyes in an alcohol solution was published by Baskin et al. [4]. Nowadays, there is an increasing interest to predict this property for different classes of dyes by using modern machine learning methods, in particular new generation of ANNs, so-called deep learning neural networks, for various classes of dyes, such as acridines, anthraquinones, di- and triphenylmethanes, cyanines, safranins, porphyrins, phthalocyanines, etc. [5–7] as well as other spectral parameters. However, there are still only few publications with an application of QSPR methods to predict the BODIPYs main spectral characteristics (absorption and

emission band maximum positions, molar absorption coefficients, quantum yields and fluorescence lifetimes, etc.). One of the first studies in this direction was published by Schüller et al. [8] where the authors accurately predicted the absorption and emission band maximum positions of BODIPYs. The authors used stepwise multiple linear regression (MLR) and support vector regression (SVR) methods to analyze experimental data for a limited set of 288 of BODIPYs. The QSPR methods were also used to predict other important practical properties of BODIPYs. For example, Caruso et al. [9] used genetic algorithm (GA) variable subset selection (GA-VSS) and multiple linear regression (MLR) by the ordinary least square (OLS) to identify structural features correlated to phototoxicity (IC_{50}) of BODIPYs against a tumor cell line.

The descriptors used in the previous studies were calculated using traditional chemoinformatics tools. However, they can also come from the first principles. For example, using a library of PCE (power conversion efficiencies) for 58 of BODIPYs, Lu et al. [10] contributed an inverse design method for the development of BODIPY sensitizers in dye sensitized solar cell (DSSC). In a more recent study a time-dependent DFT (TDDFT) was used to predict the dominant nonradiative mechanism and fluorescence quantum yields of BODIPYs [11].

It is clearly seen that the application of QSPR methods to predict various spectral properties of BODIPYs has been gaining popularity in the past few years. Firstly, this is due to the potentially wide range of BODIPYs applications, which necessitates an increase in the screening rate for compounds with practically significant properties. However, the accuracy and reproducibility of the developed QSPR models depend on the correctly selected methods and descriptors, the number of molecules for the training set and the quality of the experimental data presented.

The previous studies were based on a limited set of compounds. Recently Ju et al. [7] used a database of about 3000 solvated fluorophore molecules of various nature to predict their maxima of absorption, emission wavelengths and fluorescence quantum yields. However, this dataset included only several hundreds BODIPYs, which could limit the accuracy of the model to this class of compounds, as it was demonstrated below.

In this article, we overcame the limitations of this model to predict the band maximum positions of BODIPYs absorption by including nearly new 9000 measurements. We also verified the prediction ability of this model on an external dataset of 26 BODIPYs synthesized and characterized by our scientific team [12–14] including three new BODIPYs.

Experimental part

Datasets

Three training datasets were used as summarized in Table 1. The TR1 (“Schüller set”) consisted of 298 BODIPYs (Fig. S1 a) from ref. [8]. The initial dataset contained 301 measurements which included one duplicated value (same solvent and same values). This value was removed. The TR2 (“Ju Set”) [7] comprised 2797 dyes (431 BODIPYs) including all measurements from TR1 (Fig. S1 b). As in the case with TR1 we identified 86 duplicates, which were removed from this set. Importantly, the authors also made publicly available their model, which was used by us to test it for prediction of new compounds.

To evaluate performance of the models we collected data of absorption measurements (8819 samples for 3935 BODIPYs) and conditions of experiments (temperature and solvents) from multiple publications and also the Reaxys® database (www.reaxys.com). The absorption maximum wavelengths for the selected BODIPYs ranged from 404 to 654 nm. After removal of data points overlapping with TR1 and TR2 we formed a test set TS1 which was used to test performance of the models. The final training dataset TR3 was obtained by merging TR1, TR2 and TS1 sets (Fig. 1).

Additional testing of the models was done with the spectral data for BODIPYs previously synthesized in our laboratory [12–14]. These data consisted of 173 absorption measurements at 19 solvents for 23 BODIPYs (TS2 set) and did not have any molecules in common with the training sets. A prospective validation of the models was done using three new compounds Br-BODIPY 1-3 (TS3 set, see experimental details in Scheme S1). The first two compounds were first synthesized and characterized in this work. While Br-BODIPY 3 was first described in ref. [15], its absorption maximum wavelength was never reported.

Once data were collected, we found that the reported temperature was 293 K for all measurements with an exception of two samples, which were measured at 298 K. Therefore, we ignored the effect of this condition and did not use it for modelling.

Table 1. Composition of the training and test sets used in the article

Dataset	Compounds (BODIPY)	Measurements (BODIPY)	Solvents
TR1 (“Schüller set”)	298 (all)	300 (all) ¹	DMSO
TR2 (“Ju set”)	2797 (431)	4166 (474) ²	51 solvents
TS1	3935 (all)	8819 (all)	130 solvents
TR3 (merge of TR2 and TS1)	6732 (4366)	13339 (9293)	130 solvents
TS2	23 (all)	173 (all)	19 solvents
TS3	3 (all)	3 (all)	chloroform

¹One duplicated sample was removed from the original set. ²86 duplicated samples were removed from the original set. TR - training set, TS - test set

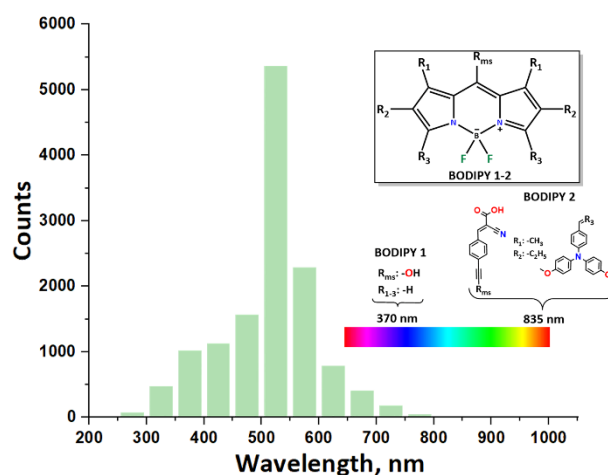


Fig. 1. Distribution of absorption maximum wavelengths of BODIPYs in TR3 (Insert: the BODIPY structures from TR3 with the smallest and largest values of the absorption maximum wavelength).

Methods and descriptors

We used methods and descriptors available and described on the online platform OCHEM (<http://ochem.eu>) [16,17]. At the first stage, five machine learning methods were initially used to build the model, including deep neural network (DNN) [18], scalable and flexible gradient boosting [19], associative neural networks [20], least squares support vector machine [21], and random forest [22]. A preliminary study using all sets indicated that DNN consistently calculated smaller Root Mean Squared Error (RMSE) for analyzed datasets. Therefore, DNN was selected for all calculations reported in this article. We used the default architecture and hyperparameters of DNN as reported in ref. [18].

We used different sets of 2D and 3D descriptors, which are described at <http://docs.ochem.eu/display/MAN/Molecular+descriptors>. Three-dimensional structures for 3D descriptors were calculated using Corina [23]. All models were validated by means of 5-fold cross-validation (5CV). A preliminary study indicated that RDKit [24], alvaDesc [25], MORDRED [26], Fragmentor [27], and PyDescriptor [28] consistently contributed models with lowest RMSE. Therefore, these descriptors were selected to develop models for all datasets. For the final models, we combined predictions of the respective models into consensus models.

In addition to the RMSE OCHEM calculated several other statistical parameters, such as squared Pearson correlation coefficient (R^2), coefficient of determination (r^2) as well as Mean Absolute Error (MAE) which were reported for individual and consensus models (Fig. S2 - S9). It should be mentioned that all reported statistical parameters were calculated between the experimental and predicted values of the absorption wavelengths using the 5CV procedure.

To account for the solvent effect we initially followed the approach of Ju et al [7] who used solvent descriptors in the form of Dimroth-Reichardt ET(30) [29] and Catalan (SP, SB, SdP, and SA) [29,30] parameters.

After development of models using descriptor-based methods, we also evaluated performance of deep-learning algorithms which are based on representation learning. These methods can be called descriptor-less since they do not use a fixed set of descriptors but discover the representations needed for the model from raw data, such as chemical graph, SMILES or chemical image.

OCHEM web site has several such algorithms, which were reviewed at [31]. The number of methods, e.g. DEEPCHEM, did not work with boron atom [B] and thus

were excluded from the consideration. Moreover, most of the analyzed methods did not allow to incorporate conditions as additional descriptors. After preliminary analysis we found that ChemProp [32] and Transformer CNN [33]. Below, we briefly describe these methods.

Graph-based ChemProp [32] is based on the directed Message Passing Neural Network (MPNN) architecture which operates on undirected chemical graphs. During the message passing stage the network transmits information across the chemical graph and builds a neural network representation of the molecule. During the readout phase the neural network uses the learned representation to make predictions of the analyzed properties.

SMILES based Transformer CNN [33] uses Transformer neural network architecture, which is one of the Natural Processing Language methods used by Google, for, e.g., translation of text between languages. The network was first learned to provide canonization of 1.7M chemical SMILES from the ChEMBL database. Once this task was completed, the internal representation of molecules (which were variable length vectors containing probabilities of characters) were used as input to Convolutional Neural Network (CNN) which is a well-known method for image processing proposed by Yann LeCun [34]. Transformer CNN uses 1D CNN neural network since the input data to this method are one-dimensional. To train and apply this algorithm we also used so-called augmented data, which were SMILES starting from a random position of an atom, which improved accuracy of this method by decreasing its variance for prediction of individual SMILES [35]. The average augmentation of $n=10$, which was found as optimal in previous studies [33,35], was used for both training and prediction of data.

Results and Discussions

Analysis of performance of existing models.

Schüller et al [8] did not make their model publicly available. That is why we redeveloped the model for this set using the approach described in the method section. Calculated 5CV RMSE was similar to the value reported in the original article (Table 2). This model calculated low accuracy for compounds from the TS2. The low accuracy obtained could be due to the small range of absorption maximum wavelength of BODIPYs in TR1 (525 - 612 nm) where 50% of all values of the absorption maxima

wavelength were in a small range (9 nm) (Fig. S10 a). Note that all BODIPYs present in the TR1 had the invariance of the BODIPY core substitution position and the substituents nature (positions in the BODIPY core: 1-4, 6, 8: -H; 5: bulky substituent; 7: -CH₃ and 1, 3: -CH₃, 2, 4, 6, 8: -H; 5: bulky substituent) [8]. The compounds from the test sets had much higher structural variability and that is why the model developed based on TR1 provided low accuracy for the unseen data.

Table 2. RMSE of models developed using different training sets

Training set	Original model	5CV	TS1 (n= 8819)	TS2 (n=173)	TS3 (n=3)
TR1	5.8 ¹	6.3±0.4	56.7±0.5	35±3	23±1
TR2 ²	21±2	-	40.5±0.5	13.5±0.9	40±3
TR2 ³		26±2	36.9±0.5	24±4	35±3
TR3		18.4±0.4	-	8.8±0.5	2.4±0.9
TR3 (solvents ignored)		18.7±0.4	-	9.7±0.6	4±1

¹Values for 10-fold CV as reported in the article. ²The performance and the results of the model published in the article was used. ³The model was redeveloped in this article using the same protocol as models for TR1 and TR3 sets.

The application of Ju et al. [7] model, which was included as a part of supplementary materials of the article, provided good results, RMSE=13.5 nm for TS2 including 23 compounds (Table 2, Fig. S10 b), which were previously synthesized in our laboratory. The model, however, failed to predict compounds from the large diverse set of compounds TS1 (RMSE=46.5 nm). This result can be attributed to the limited number of BODIPY compounds (474) in this set. To check this assumption, we developed a new model for this set using the same protocol as for the TS1 model. While this model calculated a slightly better performance for the TR1 set, its performance for prediction of this diverse set of BODIPY compounds was also low.

The model developed based on the combined set TR3 calculated very good 5CV results and provided excellent prediction of compounds from the TS2 set (Table 2, Fig. 2). Since we used the same protocol across for all three sets, such a significant

improvement in the quality of the model was achieved due to a wide variety of dye structures in TR3.

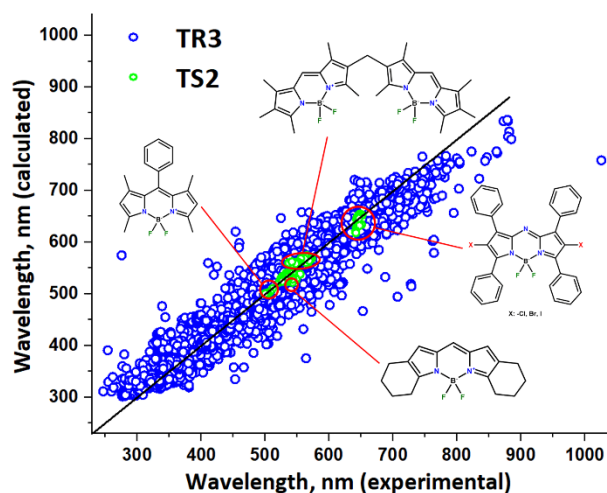


Fig. 2. Scatter plots of calculated vs. experimental absorption maximum wavelength developed using TR3 and applied to TS2. The straight line indicates a perfect fit.

We also developed a model by discarding the effect of solvent, i.e. by excluding solvent parameters from the set descriptors provided to DNN. While the performance of the model slightly decreased for both 5CV and test sets, the change in the accuracy was not significant. This result could be due to the fact that only a limited number of BODIPYs exhibit effective solvatochromism [12,36]. For example, some structures synthesized in our laboratory were measured at about 19 solvents. However, most of them had very similar absorption maximum wavelength across all solvents (Table S1). The solvent properties, nevertheless, are known to be much more important for the other spectral properties of BODIPYs (molar extinction coefficient, emission maximum wavelength, fluorescence quantum yield, and fluorescence lifetime).

We inspected samples ($n=305$, $\sim 2.3\%$ of all samples) with large prediction errors ($RMSE > 50$ nm) in the final model. The analysis showed that 60 (26%) of 231 compounds with large errors were salts while only 452 (6.9%) out of 6501 were salts in the remaining data. The salts could be sensitive to the pH of the solvent, which was not considered in this analysis, thus contributing larger errors for solvent with different acid-base properties.

In order to determine what structural features can lead to a deterioration in the quality of the compounds prediction, we analyzed the structures of the remaining compounds. Most of the samples that we have chosen as outliers are characterized by the presence of bulky substituents incl. polycyclic aromatic scaffolds, and / or partially charged functional groups capable of entering acid-base interactions, having rotary properties and/or capable of entering into intermolecular interactions with solvent molecules (Fig. 3). Also, in a number of cases, such structural features of outliers can determine the implementation of various photophysical processes (charge transfer, photoinduced electron transfer). The implementation of such processes can make a significant contribution to the change in the spectral characteristics of the luminophores. Analysis of the average values of molecular weights (MW) showed that MW of the outliers samples (excluding the salt forms) was 527 g/mol, and for all other samples, MW was 476 g/mol. Therefore, it is necessary to search for or develop new descriptors capable of taking into account the above-mentioned structural features and/or including better description of solvents (e.g., viscosity). We think that taking into account the influence of these effects will improve the quality of the prediction of the spectral characteristics of dyes and luminophores.

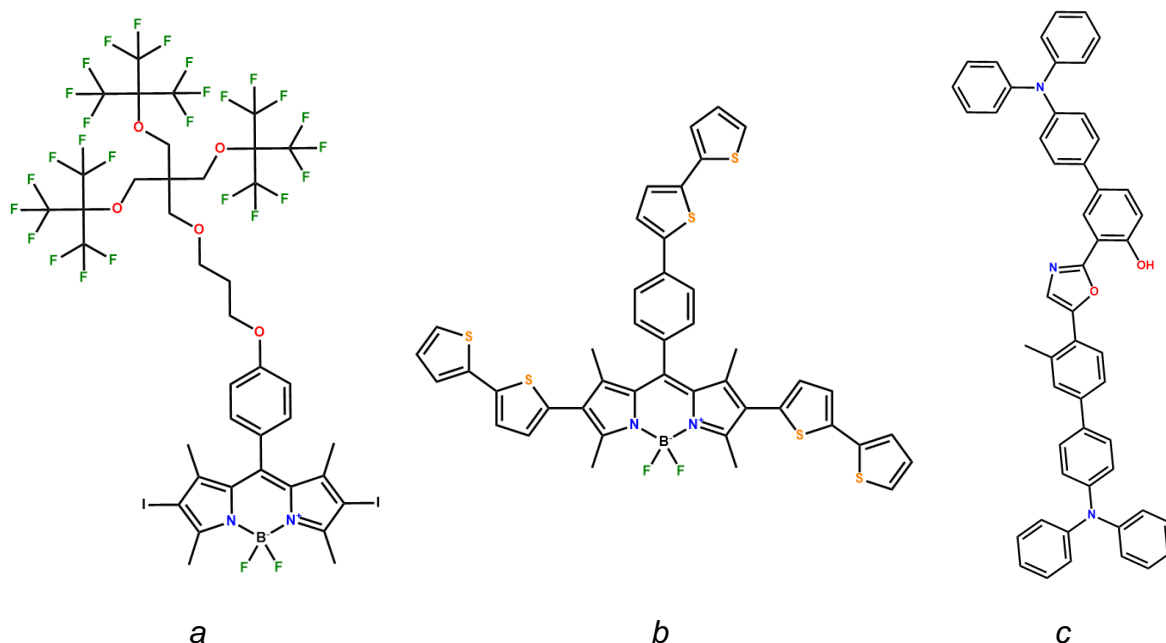
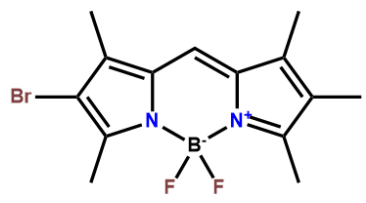
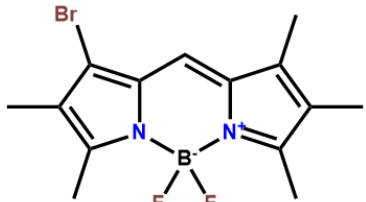


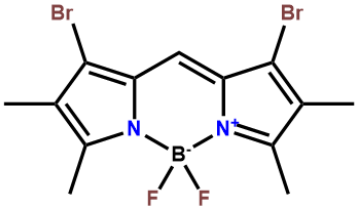
Fig. 3. Examples of outliers structures: *a* - BODIPY, characterized by the presence of multibranched fluorinated residues, *b* - D-A-D type systems based on BODIPY, *c* - 2-

methyl-4-[4-(N-phenylanilino)phenyl]-6-[5-[4-[4-(N-phenylanilino)phenyl]phenyl]-1,3-oxazol-2-yl]phenol.

We analyzed in detail the prediction accuracy of models for three newly synthesized bromo-substituted BODIPYs (Br-BODIPY 1 - 3) (Scheme S1). We applied the model provided by Ju et. Al [7] as supplementary materials (Table 2) as well as also their on-line model at <http://www.chemfluor.top>, which was re-developed by the authors using Morgan/ECFP4 fingerprints. The comparison of results (Table 3) showed that both the models calculated low prediction accuracy for Br-BODIPY compounds. Moreover, they failed to predict the tendency of the absorption maximum wavelength to shift in the red region with an increase of the core bromination degree, which was correctly captured by the new model.

Table 3. Comparative analysis of the experimental values of the Br-BODIPY 1 - 3 absorption maximum wavelength (nm) with the values predicted using QSAR models.

	Experiment ¹	new model	“Ju model”	chemfluor.top ¹
	528	530	496	517
	530	534	485	510

	534	540	493	501
-----------------------------------------------------------------------------------	-----	-----	-----	-----

¹ - values obtained in chloroform

In our further studies, new Br-BODIPY 1 - 3 will be tested for photostability and as PDT agents.

Comparison of descriptor-based and descriptor-less methods

Statistical parameters for individual and consensus models are summarized in Table 4. The individual descriptor-less methods contributed one model with a higher (ChemProp) and one (Transformer CNN) with a similar accuracy as the respective individual models based on descriptors. The consensus model developed as a simple average of these two models had a slightly higher (but not significantly different) accuracy compared to the one based on the descriptors. Interestingly, both descriptor-less methods did not use any 3D information and their results were based on SMILES with removed stereochemistry. This result indicates that the new methods based on representation learning could be very promising approaches for prediction of spectral properties of dyes.

Table 4. RMSE of models developed using different training sets.

Descriptors/ Method	Training set, 5-CV, n=13339		Retrospective validation, n=173	
	RMSE	R ²	RMSE	R ²
Descriptor-based				
RDKit [24]	21	0.93	13	0.88
alvaDesc [25]	21	0.93	13	0.89
MORDRED [26]	22	0.93	28	0.43
Fragmentor [27]	22	0.93	12	0.93
PyDescriptor [28]	22	0.93	18	0.89

Consensus	18.7 ± 0.3*	0.949 ± 0.002	9.7 ± 0.6	0.93 ± 0.1
Descriptor-less				
ChemProp [32]	19	0.94	15	0.92
Transformer CNN [33]	21	0.94	20	0.86
Consensus	18.5 ± 0.4	0.95 ± 0.002	10.2 ± 0.4	0.98 ± 0.01

*Confidence intervals were estimated based on the bootstrap procedure as described in ref. [37].

Conclusions

We contributed a QSPR model (<https://ochem.eu/article/134921>) to accurately predict the BODIPYs absorption maximum wavelength. To the best of our knowledge, this model is based on the largest dataset of BODIPY absorption maximum wavelength and will be a significant asset for scientists designing new BODIPY compounds. We also demonstrated that the accuracy of the previously reported models was limited by the diversity of used data.

The obtained results inspired us to create the improved models for predicting other spectral properties such as emission wavelength, molar absorption coefficient, fluorescence quantum yield, etc., which we plan to do in new studies. The quality, accessibility and user-friendly interface of our model will be a great asset for synthetic chemists and will speed up and facilitate the design of new BODIPYs.

In order to the convenience of the readers of this article, we have added a model manual (“Text Manual” and “Video Manual”) to the ESI.

Acknowledgments

This work received financial support of the Ministry of Science and Higher Education of the Russian Federation [No. 075-15-2021-579]. This work was carried out with the help of the center for joint use of scientific equipment “The upper Volga region center of physico-chemical research” and Ivanovo State University of Chemistry and Technology. The authors thank Alvascience Srl, ChemAxon and Molecular Networks GmbH for a possibility to use descriptors and Corina programs in their study.

References

- [1] A. Loudet, K. Burgess, BODIPY dyes and their derivatives: syntheses and spectroscopic properties, *Chem. Rev.* 107 (2007) 4891–4932. <https://doi.org/10.1021/cr078381n>.
- [2] P. Kaur, K. Singh, Recent advances in the application of BODIPY in bioimaging and chemosensing, *J. Mater. Chem. C* 7 (2019) 11361–11405. <https://doi.org/10.1039/C9TC03719E>.
- [3] Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, Alexander Tropsha, QSAR without borders, *Chem. Soc. Rev.* 49 (2020) 3525–3564. <https://doi.org/10.1039/D0CS00098A>.
- [4] I. I. Baskin, A. O. Ait, N. M. Halberstam, V. A. Palyulin, M. V. Alfimov, N. S. Zefirov, Application of Methodology of Artificial Neural Networks for Predicting the Properties of Sophisticated Molecular Systems: Prediction of the Long-Wave Absorption Band Position for Symmetric Cyanine Dyes, *Doklady Akademii Nauk* 357 (1997) 57–59.
- [5] Z.-R. Ye, I.-S. Huang, Y.-T. Chan, Z.-J. Li, C.-C. Liao, H.-R. Tsai, M.-C. Hsieh, C.-C. Chang, M.-K. Tsai, Predicting the emission wavelength of organic molecules using a combinatorial QSAR and machine learning approach, *RSC Adv.* 10 (2020) 23834–23841. <https://doi.org/10.1039/D0RA05014H>.
- [6] J.F. Joung, M. Han, J. Hwang, M. Jeong, D.H. Choi, S. Park, Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design, *JACS Au* 1 (2021) 427–438. <https://doi.org/10.1021/jacsau.1c00035>.
- [7] C.-W. Ju, H. Bai, B. Li, R. Liu, Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields, *J. Chem. Inf. Model.* 61 (2021) 1053–1065. <https://doi.org/10.1021/acs.jcim.0c01203>.
- [8] A. Schüller, G.B. Goh, H. Kim, J.-S. Lee, Y.-T. Chang, Quantitative Structure-Fluorescence Property Relationship Analysis of a Large BODIPY Library, *Mol. Inform.* 29 (2010) 717–729. <https://doi.org/10.1002/minf.201000089>.
- [9] E. Caruso, M. Gariboldi, A. Sangion, P. Gramatica, S. Banfi, Synthesis, photodynamic activity, and quantitative structure-activity relationship modelling of a series of BODIPYs, *J. Photochem. Photobiol. B* 167 (2017) 269–281. <https://doi.org/10.1016/j.jphotobiol.2017.01.012>.
- [10] T. Lu, M. Li, Z. Yao, W. Lu, Accelerated Discovery of Boron-dipyrromethene Sensitizer for Solar Cells by Integrating Data Mining and First Principle, *Journal of Materiomics* (2021). <https://doi.org/10.1016/j.jmat.2020.12.018>.

- [11] Z. Lin, A.W. Kohn, T. van Voorhis, Toward Prediction of Nonradiative Decay Pathways in Organic Compounds II: Two Internal Conversion Channels in BODIPYs, *J. Phys. Chem. C* 124 (2020) 3925–3938. <https://doi.org/10.1021/acs.jpcc.9b08292>.
- [12] L.A. Antina, A.A. Ksenofontov, A.A. Kalyagin, E.V. Antina, M.B. Berezin, I.A. Khodov, Luminescent properties of new 2,2-, 2,3- and 3,3-CH₂-bis(BODIPY)s dyes: Structural and solvation effects, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 218 (2019) 308–319. <https://doi.org/10.1016/j.saa.2019.03.117>.
- [13] L.A. Antina, A.A. Ksenofontov, A.A. Kalyagin, P.S. Bocharov, N.V. Kharitonova, A.V. Kazak, E.V. Antina, M.B. Berezin, The influence of alkylation on the photophysical properties of BODIPYs and their labeling in blood plasma proteins, *Journal of Molecular Liquids* 304 (2020) 112717. <https://doi.org/10.1016/j.molliq.2020.112717>.
- [14] G.B. Guseva, A.A. Ksenofontov, E.V. Antina, M.B. Berezin, A.I. Vyugin, Effect of solvent nature on spectral properties of blue-emitting meso-propargylamino-BODIPY, *Journal of Molecular Liquids* 285 (2019) 194–203. <https://doi.org/10.1016/j.molliq.2019.04.058>.
- [15] V. Leen, D. Miscoria, S. Yin, A. Filarowski, J. Molisho Ngongo, M. van der Auweraer, N. Boens, W. Dehaen, 1,7-Disubstituted Boron Dipyrromethene (BODIPY) Dyes: Synthesis and Spectroscopic Properties, *The Journal of Organic Chemistry* 76 (2011) 8168–8176. <https://doi.org/10.1021/jo201082z>.
- [16] Tetko IV, U. Maran, A. Tropsha, Public (Q)SAR Services, Integrated Modeling Environments, and Model Repositories on the Web: State of the Art and Perspectives for Future Development, *Mol. Inform.* 36 (2017). <https://doi.org/10.1002/minf.201600082>.
- [17] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, Baskin II, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, Tetko IV, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, *Journal of computer-aided molecular design* 25 (2011). <https://doi.org/10.1007/s10822-011-9440-2>.
- [18] S. Sosnin, D. Karlov, I.V. Tetko, M.V. Fedorov, Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space, *J. Chem. Inf. Model.* 59 (2019) 1062–1072. <https://doi.org/10.1021/acs.jcim.8b00685>.
- [19] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
- [20] I.V. Tetko, Associative Neural Network, *Neural Processing Letters* 16 (2002) 187–199. <https://doi.org/10.1023/A:1019903710291>.

- [21] J.A.K. Suykens, J. Vandewalle, Least Squares Support Vector Machine Classifiers, *Neural Processing Letters* 9 (1999) 293–300. <https://doi.org/10.1023/A:1018628609742>.
- [22] L. Breiman, Random Forests, *Machine Learning* 45 (2001) 5–32. <https://doi.org/10.1023/A:1010933404324>.
- [23] J. Sadowski, J. Gasteiger, G. Klebe, Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures, *Journal of Chemical Information and Computer Sciences* 34 (1994) 1000–1008. <https://doi.org/10.1021/ci00020a039>.
- [24] RDKit, 2019.000Z. <https://www.rdkit.org/> (accessed 11 August 2021.296Z).
- [25] A. Mauri, alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints, in: K. Roy (Ed.), *Ecotoxicological QSARs*, Humana, New York, 2020, pp. 801–820.
- [26] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, *J Cheminform* 10 (2018) 1–14. <https://doi.org/10.1186/s13321-018-0258-y>.
- [27] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov'ev, F. Hoonakker, I. Tetko, G. Marcou, ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors, *CAD* 4 (2008) 191–198. <https://doi.org/10.2174/157340908785747465>.
- [28] V.H. Masand, V. Rastija, PyDescriptor: A new PyMOL plugin for calculating thousands of easily understandable molecular descriptors, *Chemometrics and Intelligent Laboratory Systems* 169 (2017) 12–18. <https://doi.org/10.1016/j.chemolab.2017.08.003>.
- [29] C. Reichardt, T. Welton, *Solvents and solvent effects in organic chemistry*, fourthth, updated and enlarged ed., Wiley-VCH, Weinheim, 2011.
- [30] J. Catalán, Toward a Generalized Treatment of the Solvent Effect Based on Four Empirical Scales: Dipolarity (SdP, a New Scale), Polarizability (SP), Acidity (SA), and Basicity (SB) of the Medium, *The Journal of Physical Chemistry B* 113 (2009) 5951–5960. <https://doi.org/10.1021/jp8095727>.
- [31] S. Sosnin, M. Vashurina, M. Withnall, P. Karpov, M. Fedorov, I.V. Tetko, A Survey of Multi-task Learning Methods in Chemoinformatics, *Mol. Inform.* 38 (2019) 1800108. <https://doi.org/10.1002/minf.201800108>.
- [32] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction, *Journal of Chemical Information and Modeling* 59 (2019) 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>.
- [33] P. Karpov, G. Godin, I.V. Tetko, Transformer-CNN: Swiss knife for QSAR modeling and interpretation, *J Cheminform* 12 (2020) 1–12. <https://doi.org/10.1186/s13321-020-00423-w>.
- [34] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (1998) 2278–2324. <https://doi.org/10.1109/5.726791>.

- [35] I.V. Tetko, P. Karpov, E. Bruno, T.B. Kimber, G. Godin, Augmentation Is What You Need!, in: Artificial neural networks and machine learning - ICANN 2019, Cham, Springer, Cham, Switzerland, 2019, pp. 831–835.
- [36] L. Loaeza, R. Corona-Sánchez, G. Castro, M. Romero-Ávila, R. Santillan, V. Maraval, R. Chauvin, N. Farfán, Synthesis and optical properties of 1-ethyl-indol-3-yl-substituted aza-BODIPY dyes at the 1,7-positions, *Tetrahedron* 83 (2021) 131983. <https://doi.org/10.1016/j.tet.2021.131983>.
- [37] S. Vorberg, I.V. Tetko, Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM), *Mol. Inf.* 33 (2014) 73–85. <https://doi.org/10.1002/minf.201300030>.