International Journal of Molecular Sciences

MDPI

Article

More Is Not Always Better: Local Models Provide Accurate Predictions of Spectral Properties of Porphyrins

Aleksey I. Rusanov [1], Olga A. Dmitrieva [1], Nugzar Zh. Mamardashvili [1] and Igor V. Tetko [1,2,3,*]

[1] G.A. Krestov Institute of Solution Chemistry of the Russian Academy of Sciences,153045 Ivanovo, Russia; rusanov.a.i@mail.ru (A.I.R.); dmitrievao.a@yandex.ru (O.A.D.); ngm@isc-ras.ru (N.Z.M.)
[2] Helmholtz Munich, Institute of Structural Biology, Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), D-85764 Neuherberg, Germany
[3] BIGCHEM GmbH, D-85716 Unterschleißheim, Germany
* Correspondence: i.tetko@helmholtz-munich.de; Tel.: +49-89-3187-3575

Abstract: The development of new functional materials based on porphyrins requires fast and accurate prediction of their spectral properties. The available models in the literature for absorption wavelength and extinction coefficient of the Soret band have low accuracy for this class of compounds. We collected spectral data for porphyrins to extend the literature set and compared the performance of global and local models for their modelling using different machine learning methods. Interestingly, extension of the public database contributed models with lower accuracies compared to the models, which we built using porphyrins only. The later model calculated acceptable RMSE = 2.26 for prediction of the absorption band of 335 porphyrins synthesized in our laboratory, but had a low accuracy (RMSE = 0.52) for extinction coefficient. A development of models using only compounds from our laboratory significantly decreased errors for these compounds (RMSE = 0.5 and 0.042 for absorption band and extinction coefficient, respectively), but limited their applicability only to these homologous series. When developing models, one should clearly keep in mind their potential use and select a strategy that could contribute the most accurate predictions for the target application. The models and data are publicly available.

Keywords: QSPR; Random Forest; local model; chromophores; porphyrins; absorbance maximum wavelength; molar extinction coefficient

Citation: Rusanov, A.I.; Dmitrieva, O.A.; Mamardashvili, N.Z.; Tetko, I.V. More Is Not Always Better: Local Models Provide Accurate Predictions of Spectral Properties of Porphyrins. *Int. J. Mol. Sci.* **2022**, *23*, 1201. https://doi.org/10.3390/ijms23031201

Academic Editor: Hanoch Senderowitz

Received: 30 December 2021
Accepted: 19 January 2022
Published: 21 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Porphyrins represent a unique class of heterocyclic tetrapyrrolic organic molecules which are classified as strong dyes (chromophore) due to their pronounced light-absorbing properties. Their unique optical properties were intensively studied in recent decades and found to have a wide range of applications in medicine [1] biological imaging [2,3], photocatalytic [4], analytical [5], industrial [6], nonlinear optics (NLO) [7], and molecular photovoltaics [8,9]. The presence of a highly conjugated system allows porphyrin to have intense absorption of light in the visible region with very unique UV-vis spectra. The main feature of porphyrin spectra is the presence of a very intense band at the 400 nm region (the so-called Soret band). It is known that the modification of the porphyrin macrocycle, namely, its meso-substitution, has a greater effect on the position and intensity of this band [10]. Consequently, the Soret band is a convenient and sensitive tool reflecting changes both in the structure of molecules and the effect of solvents on it.

The development of new chromophores frequently critically depends on the expertise of the chemist and requires a large amount of time and synthetic efforts to synthesize new compounds with the desired optical and photophysical properties. Computational methods for predicting the optical properties of new porphyrins could allow them to be estimated in advance and reduce costs of synthesis. Such methods are

actively developed in the field now, in particular based on quantum chemistry calculations, but frequently they have some significant limitations. A four-orbital model introduced by Gouterman successfully explains the presence of peaks in the absorption spectra of porphyrin and metal-free porphyrins [11]. However, this theory is unable to explain why the maximum positions in the absorption spectra remain almost unchanged when measured in different solvents for certain kinds of porphyrins. The semi-empirical quantum-chemical methods PPP-MO [12] and ZINDO/S [13] require calibration using an experimental dataset to achieve accurate wavelength predictions [14]. Time-dependent density functional theory (TD-DFT) [15–17] and ab initio calculations [18] require high level calculations to account for both dynamical and non-dynamical electron correlation, which are computational demanding and limit the practicality of such methods to single/few molecules [19]. The results of quantum-chemical calculation frequently deviate from the experimental data by 0.2–0.3 eV [20,21].

In recent years, quantitative structure–property relationship (QSPR) modeling has become a powerful tool for predicting the optical properties of chromophores [22–28]. The QSPR approach is based on the assumption that the macroscopic properties of chemical compounds depends on the calculated molecular characteristics of the compounds, which are called molecular descriptors. The advantage of this approach lies in the fact that once model is developed, it requires only the knowledge of the chemical structure and does not dependent on any experimental properties [29]. Accurate computational prediction of spectral properties of new porphyrins could allow us to design new molecules with desired properties using traditional combinatorial chemistry approaches or structures generated by deep neural networks [30]. However, since QSPRs are statistical approaches, the accuracy of developed models critically depends on the quality of data and of adequacy of the training set to the compound to be predicted. Moreover, one can use either local (by using structurally related compounds) or global models (developed with diverse sets of compounds). The advantages of each approach for the prediction of spectral properties of compounds need to be better carefully evaluated and have not been performed so far.

In this study, we tested a previous model of Joung et al. [27] as well as several new models developed with a large set of dyes and porphyrins collected from the literature to predict spectral properties of new compounds synthesized in our laboratory.

2. Material and Methods

The absorption spectra for compounds synthesised in our laboratory were obtained by spectrophotometer Cary-100 (Aglient, Santa Clara, CA, USA) in the dichloromethane (chemically pure).

2.1. Datasets

The initial analysis was performed using data from an article of Joung et al. [27] which contained optical properties of organic compounds collected from the literature which were described in [31] and were publicly available at FigShare link [32]. While Joung et al. [27] reported in their article 26,098 and 12,159 training set values for absorption band maximum position and extinction coefficient, respectively, the publicly accessible data at FigShare [32] contained only 17,294 and 8041 values for these optical properties, respectively. We excluded from these data organic compounds in the solid state, since our goal was to predict porphyrins in a liquid medium. Compounds which could not be processed by the On-line CHEmical database and Modeling environment (OCHEM) platform (very large and/or molecules with many rings for which calculation of descriptors failed) were also excluded. The remaining set (hereinafter JOUNG set) contained 6271 unique organic chromophores in 27 solvents, yielding 15,380 chromophore/solvent combinations for absorption band maximum position (Figure 1a) and 3753 unique organic chromophores in 25 solvents (7654 chromophore / solvent

combinations) for molar extinction coefficient (Figure 1b). The database included various chromophore classes, but contained only 30 porphyrins.

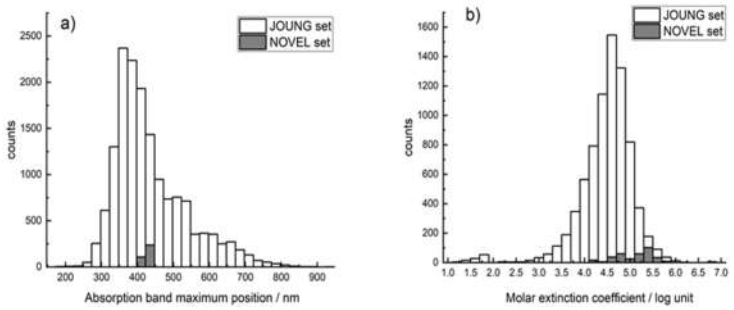

Figure 1. Histogram of the distribution of JOUNG and a NOVEL set of 335 porphyrins synthesized in our laboratory by absorption wavelengths (**a**) and the value of the extinction coefficient (**b**).

The second set (hereinafter PORPHYRINS) was collected in this work from more than 30 publications. It included data for the first absorption peak of Soret porphyrins and their analogs (2241 unique compounds in dichloromethane), as well as their values of the logarithm of the molar extinction coefficient (946 unique compounds in dichloromethane). The database included the following macroheterocycles: chlorins, protoporphyrins, porphyrins, inverted porphyrins, their metal complexes and substituted at α- and β-positions by alkyl and aryl radicals, including halogens and radicals containing heteroatoms (Supplementary Data, Table S1). The Soret absorption wavelength values were in the region of 340–500 nm, with the majority of values in the range of 410–430 nm (Figure 2a). The values of the extinction coefficient were in range from 4.15 to 5.99, with most of them being in the range from 5 to 5.8 (Figure 2b).

The third analyzed set was a combination of JOUNG and PORPHYRINS (COMBINED).

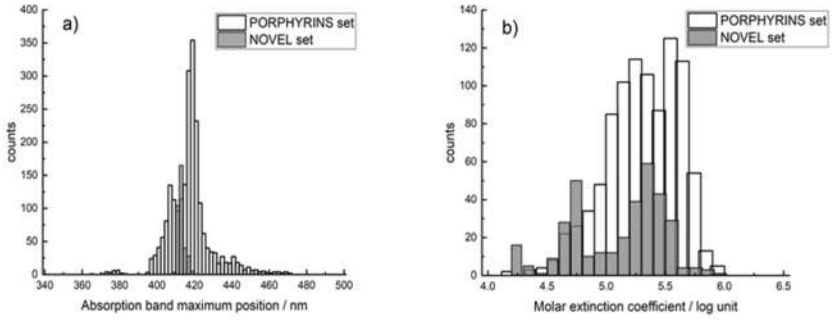

Figure 2. Histogram of the distribution of PORPHYRINS and a NOVEL set of 335 porphyrins synthesized in our laboratory by absorption wavelengths (**a**) and the value of the extinction coefficient (**b**).`

The accuracy of models was tested using cross-validation results as well as on a set of 335 newly synthesized 2,8,12,18-tetramethyl-3,7,13,17-tetraalkyl-5,15-diphenylporphyrins and 3,7,13,17-tetramethyl-2,8,12,18-tetraalkyl-5,15-diphenylporphyrins, as well as their zinc complexes, which were not present in any of the previous sets and were also not previously published by us (NOVEL set). The procedures for the synthesis of these compounds are described in the Experimental protocol section of the Supplementary Data.

2.2. Methods

Quantitative models were developed using a variety of combinations of learning methods with a different set of descriptors, which were available in On-line Chemical Database and Modeling Environment (OCHEM) [33]. The default parameters of these algorithms as specified in OCHEM were used. Amid a preliminary analysis, we found that Random Forest Regression (RFR) [34] consistently contributed better results and therefore RFR was used for all analyses reported in this study. All descriptor packages available in OCHEM were used to provide a variety of chemical structure representations for spectral properties modeling. Amid them, several packages, namely ISIDA fragmentor descriptors [35], MOLD2 descriptors [36], alvaDesc [37], and SIRMS descriptors [38] consistently contributed models with the highest performances. Most of these packages used 2D representation of chemical compounds while the Corina program [39] was used to perform 2D to 3D conversion for the alvaDesc. [37] In addition to models based on descriptors, we also used Transformer Convolutional Neural Network [40], which is a representation learning method operating directly with text representation (SMILES [41]) of chemical structures. All descriptor packages and modelling methods were used with default values of parameters as described in details on the OCHEM website [42].

Five-fold cross-validation [43] was used to develop models. Once models for individual descriptor packages were developed, we selected those with the highest performance for the training set and used them to build a consensus, which was an average of individual models following methodology developed in our earlier studies [44–46]. The statistical parameters calculated by the consensus model were used to estimate predictive performance of machine learning methods.

2.3. Statistical Parameters

The quality of models was estimated using the squared correlation coefficient (R^2) Equation (1) and root mean square error (RMSE) Equation (2):

$$R^2 = \frac{\sum_{i=1}^{n}(y_{pred,i} - \underline{y}_{pred}) \times (y_{exp,i} - \underline{y}_{exp})}{\sum_{i=1}^{n}(y_{pred,i} - \underline{y}_{pred})^2 \times \sum_{i=1}^{n}(y_{exp,i} - \underline{y}_{exp})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{pred,i} - y_{exp,i})^2} \quad (2)$$

where n is the number of data points; $y_{exp,i}$ is the experimental and $y_{pred,i}$ is the predicted value of the analyzed data point i.

3. Results and Discussion

Model Development and Testing

Our initial attempt was to predict the optical properties of the porphyrins from the NOVEL set using the model published by Joung et al. [27]. This model was accessed on the website (http://deep4chem.korea.ac.kr, accessed date is 30.12.2021). The results of predicting the positions of the Soret band maximum and the values of the extinction coefficient demonstrated low correlation between the predicted and experimental values (Figures S1 and S2) and were RMSE = 200 (R^2 = 0.01) and RMSE = 0.89 (R^2 = 0.1) for the maximum absorption and extinction coefficient of porphyrins, respectively (see Tables 1 and 2). Thus, the published model could not predict the optical properties of porphyrins.

Table 1. Statistical parameters of models developed using different training sets for prediction of absorption maximum band.

Data Set	Training Set, 5CV			Prediction of NOVEL Set, n = 335	
	n	R^2	RMSE	R^2	RMSE
Published model of Joung et al. [27]	26,098	0.926 [a]	31.6 [a]	0.01	200
JOUNG	15,380	0.904 ± 0.003	31.5 ± 0.5	0.12 ± 0.02	204 ± 2
COMBINED	17,621	0.9 ± 0.003	30.1 ± 0.5	0.03 ± 0.01	21 ± 1
COMBINED: JOUNG subset [a]	15,380	0.902 ± 0.003	31.9 ± 0.5		
COMBINED: PORPHYRINS subset [ab]	2241	0.43 ± 0.05	10.3 ± 0.7		
PORPHYRINS	2241	0.8 ± 0.01	5.4 ± 0.2	0 ± 0.005	2.26 ± 0.08
NOVEL set	335	0.93 ± 0.01	0.5 ± 0.03		

[a] The results reported by Joung et al. [27]. [b] Statistical results were calculated for a respective subset of compounds from the COMBINED set.

Table 2. Statistical parameters of models developed using different training sets for prediction of the extinction coefficient.

Data Set	Training Set, 5CV			Prediction of NOVEL Set, n = 335	
	n	R^2	RMSE	R^2	RMSE
Published model of Joung et al. [27]	12,159	0.795 [a]	0.24 [a]	0.10	0.89
JOUNG	7654	0.767 ± 0.009	0.286 ± 0.005	0.62 ± 0.02	0.84 ± 0.02
COMBINED	8600	0.806 ± 0.007	0.279 ± 0.005	0 ± 0.006	0.54 ± 0.02
COMBINED: JOUNG subset [a]	7654	0.765 ± 0.01	0.286 ± 0.005		
COMBINED: PORPHYRINS subset [ab]	946	0.49 ± 0.03	0.218 ± 0.006		
PORPHYRINS	946	0.52 ± 0.02	0.209 ± 0.006	0 ± 0.004	0.52 ± 0.02
NOVEL set	335	0.989 ± 0.002	0.042 ± 0.004		

[a] The results reported by Joung et al. [27]. [b] Statistical results were calculated for a respective subset of compounds from the COMBINED set.

As it was mentioned in the Data section, the JOUNG set contained only part of data published in Joung et al. [27]. To verify whether we can reproduce results of the original model of Joung et al. [27] with OCHEM tools, we developed QSPR models based on the JOUNG using the RFR method and different sets of descriptors. A 5-fold cross-validation was used to estimate accuracy of developed models. The initial calculations were performed with and without parameterization of the solvent using procedure described elsewhere [47]. The models with the best statistical parameters were chosen to create the consensus models as average of these individual models. We observed the same effect as in the previous study [47], namely that solvent parameterization did not provide significantly better results. For example, the mean difference between RMSE of consensus models for prediction with the parameterization of solvent and without it was 0.6 nm for the JOUNG set which was within the standard mean error of the model (Table S2). Since the difference was within the error range of the model accuracies, we decided to skip the use of solvent parameterization in the further analysis for absorption coefficient. The consensus model calculated correlation coefficient R^2 = 0.90 and RMSE = 31.5 nm, which was similar to that (R^2 = 0.926, RMSE = 31.6 nm) obtained by the authors for the test set

compounds (10% of data). It should be mentioned that results of the 5-fold cross-validation protocol used in our study (20% of data were removed from the model and predicted based on the model training with remaining 80% of compounds; procedure was repeated 5 times and results for 20% excluded compounds were averaged) were more strict than the test set protocol reported by Joung et al. (90% of compounds were used for model hyperparameter tuning, training and validation; the performance was reported for 10% of left compounds). Similar to the original model developed by the authors, the consensus model also showed a low accuracy (R^2 = 0.12 and RMSE = 204) for the NOVEL set (see also Table S2 and Figure 3a). Thus, the prediction of the absorption band based on the original model developed by Joung et al. or data from their study had a low accuracy for porphyrins.

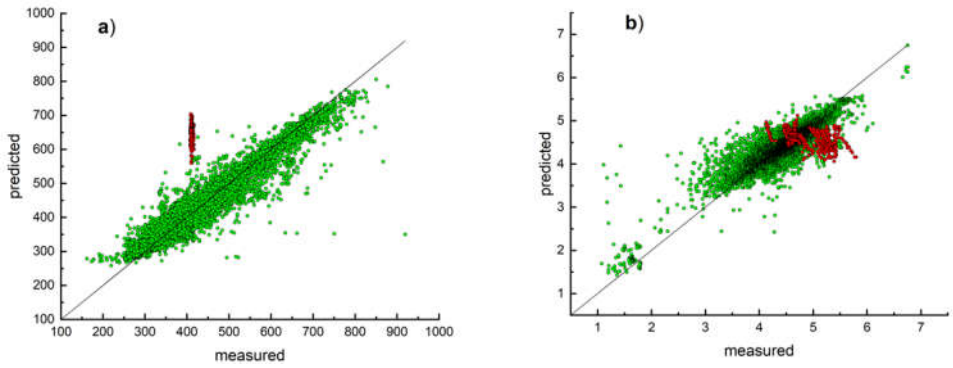

Figure 3. Distribution of the experimental and predicted values of the position of the absorption band (**a**) and values of the extinction coefficient (**b**) using models based on the JOUNG set. The green and red colors correspond to the training set data and test set data of 335 compounds, respectively.

Similar results were calculated for prediction of the extinction coefficient of chromophores and the consensus model developed using the JOUNG set provided low accuracy (R^2 = 0.62, RMSE = 0.84) for prediction of the NOVEL set compounds, which was similar to that obtained with their original model (See Table 2 and Figure 3b). Similarly for absorption coefficient, an includance of the parametrization of solvent did not improve models and was not used in further studies.

The reason for the failure of models built on the JOUNG data could be due to the low number of porphyrins in these sets (only 30 out of 15,380), which thus did not cover the chemical space of porphyrins.

To improve the prediction results, we extended the JOUNG dataset with the PORPHYRINS set to form the COMBINED set (Tables 1 and 2 and S3). Like in the study with JOUNG dataset, the models with highest accuracy for this set were used to develop the consensus models. Consensus models improved the accuracy of predicting the position of the absorption band to RMSE = 21 nm and the extinction coefficient RMSE = 0.54 for the NOVEL set. The extension of the JOUNG dataset to include porphyrins provided a global model, which was covering various classes of molecules. A combination of the JOUNG with PORPHYRINS increased the accuracy of the resulting consensus model for the JOUNG subset (we calculated statistical parameters for compounds from this subset of the COMBINED set). The accuracy of the model for the PORPHYRINS subset was higher (RMSE = 10.3 vs RMSE = 31.9) than that for the JOUNG set (Tables 1 and 2). The same tendency was observed for the extinction coefficient, but differences in statistical parameters were smaller. This result indicated that likely the quality of experimental data for the PORPHYRINS set was higher than that for the JOUNG set. By mixing low and highly accurate data, we could improve less accurate data, but at the same time, could decrease the quality of the model for more accurate ones. Therefore, we decided to develop local models using the PORPHYRINS set only.

The same methodology was used to develop models using only the PORPHYRINS. The models for both absorption band and extinction coefficients calculated higher 5CV statistical parameters than those calculated for respective subsets when they were used as part of the COMBINED set (Table S4). The developed consensus models improved the prediction of the position of the Soret band and extinction coefficients of the NOVEL set as test set compounds to RMSE = 2.26 nm (Table 1) and RMSE = 0.5 (Table 2), respectively. Thus, the development of a local model just for porphyrins as compared to the development of a global model for various dyes provided higher cross-validation accuracy for this chemical class of compounds as well as better accuracy for prediction of the NOVEL set. Although we observed an improvement of the model for prediction of extinction coefficient, the accuracy of its prediction was not satisfactory and the model with RMSE of 0.5 could hardly have any practical value.

The prediction error for the NOVEL set of 335 compounds RMSE = 2.26 nm was lower than the 5CV RMSE = 5.4 nm estimated for the PORPHYRINS set. This was a very nice result, but the experimental accuracy of the absorption band was estimated in our laboratory to be about 0.5 nm. Thus, the predicted error was about five times larger than the experimental one. For the prediction of extinction coefficient, which was typically measured with accuracy of 0.01, the discrepancy between prediction and experimental errors was about 20 times. Considering that all data for the NOVEL set were all measured in our laboratory, we were interested in determining whether we could get a better model for them.

Therefore, we used the same methodology as in the previous studies and calculated excellent consensus models for both properties for the NOVEL set (n = 335) estimated using the 5CV protocol (Tables 1 and 2 and Figures 4 and 5).

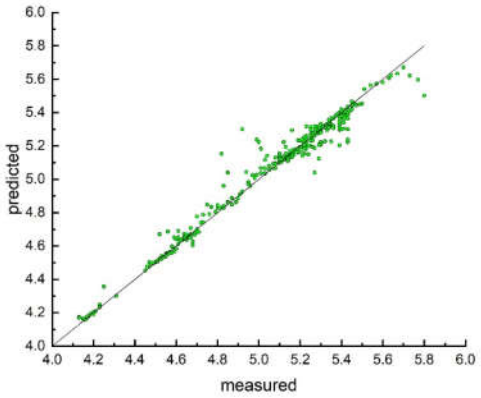

Figure 4. Distribution of the experimental and predicted values of the extinction coefficient calculated by consensus model developed with n = 335 compounds experimentally measured in this work.

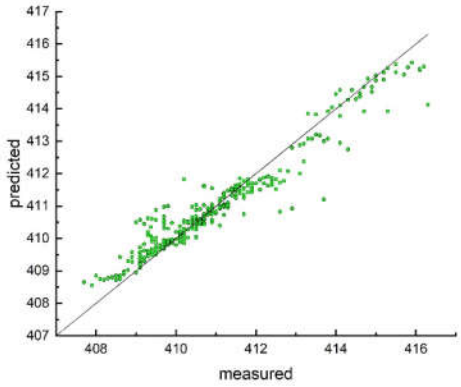

Figure 5. Distribution of the experimental and predicted values of the absorption maximum position calculated by consensus model developed with $n = 335$ compounds experimentally measured in this work.

A possible reason for such good accuracy of both these models could be the minimum noise in the data, since all measurements were performed within the same laboratory using the same equipment. On the other hand, the compounds were homologous series and just differed in functional groups in the positions of the phenyl rings, as well as in the long alkyl chains in the beta-positions.

Thus, the development of models based on the homologous series of compounds provided the best accuracy for these data. At the same time, of course, models developed with such restricted chemical series can not be used to predict compounds from other dyes, which are structurally different. The model for absorption maximum position had a range of experimental values in the 408-418 nm region (see also Figures 1 and 2) and could not extrapolate to values outside of this region. It calculated RMSE = 101.2 ± 0.8 and 14.4 ± 0.4 for prediction of dyes from JOUNG and PORPHYRINS sets, respectively. A smaller RMSE for the PORPHYRINS sets reflected a higher structural similarity of NOVEL set compounds as well as narrower range of absorption maximum position values for PORPHYRINS. Similarly, the model for extinction coefficient, which was based on the data coming from our laboratory, failed to predict these both sets too and calculated RMSE of 0.92 and 0.41 for JOUNG and PORPHYRINS, respectively. It should be mentioned that the majority of the predictions for both models were identified as out of the applicability domain [43], and thus the models correctly flagged such predictions as inconsistent with the training set data. Thus, the developed local models based on homologous series could be only applicable to these series. Contrary to that, the models developed using the PORPHYRINS sets are expected to predict a much wider class of porphyrins.

In the last study, we investigated the influence of the size of the training set for the accuracy of the model for porphyrins. Subsets of compounds were randomly sampled from respective PORPHYRIN and NOVEL sets and were used to predict the remaining compounds from the same sets that were not used for model development (see Figures 6 and 7 as well as supplementary Tables S5 and S6). With the increase of the training set size, the smaller numbers of compounds were left for testing which resulted in higher calculated errors bars. The performance of models for 100% data used as a training set was estimated using 5CV.

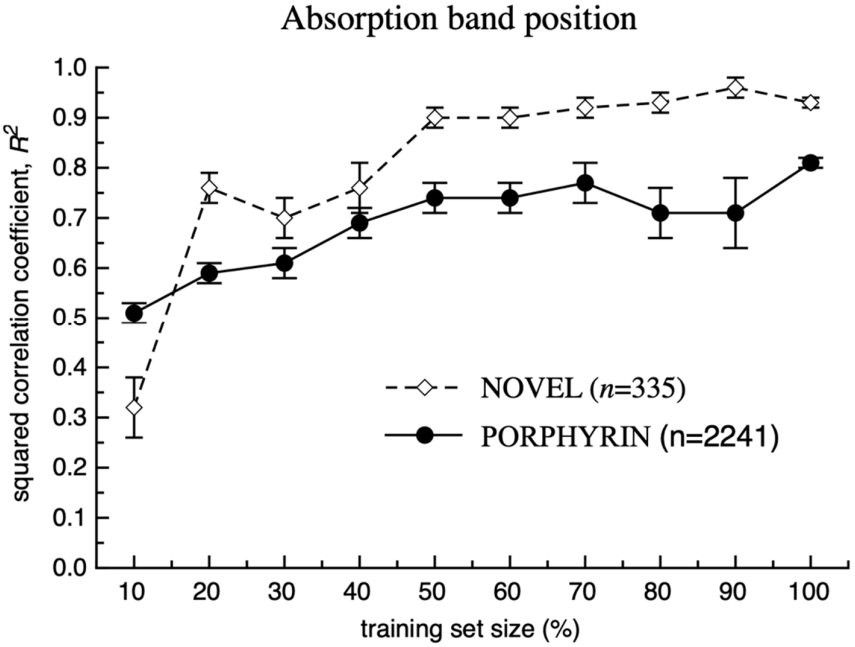

Figure 6. Statistical coefficients calculated for the prediction of the test set compounds that were not part of the respective training sets for modelling of the absorption band maximum position (see also Supplementary Data, Table S5 and S6). 5CV values were reported for 100% training set size.

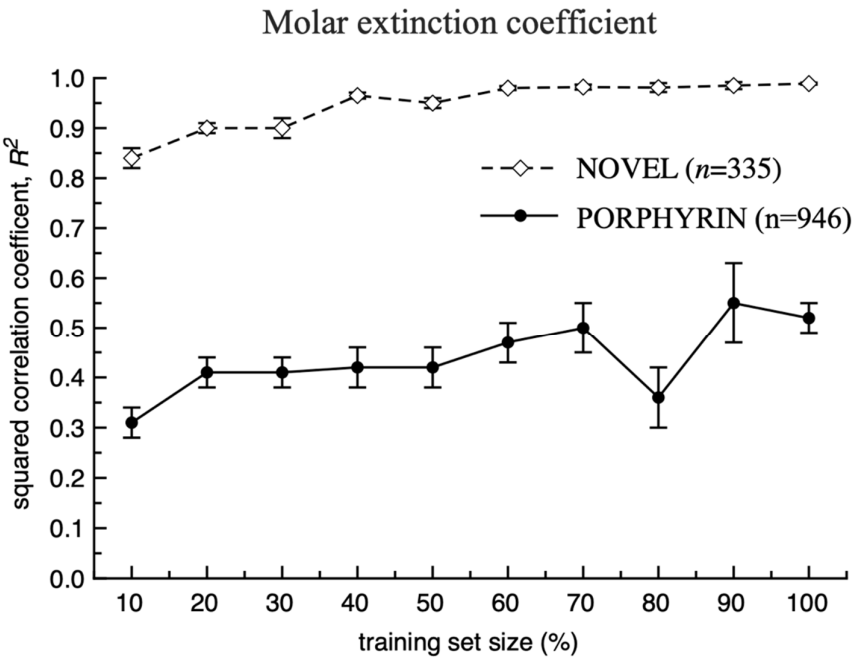

Figure 7. Statistical coefficients calculated for the prediction of the test set compounds that were not part of the respective training sets for modelling of the molar extinction coefficient (see also Supplementary Data, Table S5 and S6). 5CV values were reported for 100% training set size.

For both spectral properties, an increase of the training dataset sizes steadily increased the squared correlation coefficient, R^2 for the test sets. The accuracy of the models for the prediction of more diverse PORPHYRIN sets were lower compared to those calculated for the NOVEL set using the same percentage of the training set data. The squared

correlation coefficients for the NOVEL set using 30–40% of data were similar to those calculated using 70–100% training set data of the PORPHYRIN set. The higher values for the NOVEL set could be explained by smaller structural diversity of compounds and thus higher density of data points allowing to adequately estimate the influence of various substituents on the variation of this coefficient. Likely by further increasing the size of the PORPHYRIN set with additional data, we could reach the same values of the squared correlation coefficient obtained for the NOVEL set.

However, in the case of the extinction coefficient, there was a different behaviour and we could observe a big gap in the performances of models developed with PORPHYRIN and NOVEL sets. Thus, a further increase in the amount of literature data for this coefficient is unlikely to result in the same accuracy of the model as we calculated using the NOVEL set.

The reason for the low prediction accuracy of the molecular extinction coefficient based on the literature data could be inconsistencies and errors when collecting this parameter from various sources. These errors depend on the sensitivity of the measurement devices, e.g., type of the used spectrophotometer and the scales on which the compounds were weighed, but as well as on rounding and possibly even simple arithmetic errors when calculating the extinction coefficient from the experimental data. At the same time, if the same equipment as well as the same protocol were strictly used for its measurement within the same laboratory, one could expect much higher accuracy and consistency of data which could result in excellent models with high statistical parameters, as reported in this study.

Thus, in this work, we first analyzed the prediction accuracy of published models to predict spectral properties of porphyrins synthesized in our laboratory (NOVEL set, n = 335). We found a low performance of both published (http://deep4chem.korea.ac.kr, accessed date is 30.12.2021) as well as models re-developed by us using the publicly available data deposited by the authors (JOUNG set). The RMSE for the prediction of maximum absorption band were in range of 200 nm while for the extinction coefficient RMSE of 0.8–0.9 log units were observed. The low performance of these models was attributed to a small number of porphyrins (n = 30) in the training sets.

An extension of published sets by including porphyrins (COMBINED set) improved results for both spectral properties and RMSE = 21 and 0.54 were calculated for these properties for the NOVEL set. A development of local models using only PORPHYRINS set (n = 2241 for absorption and n = 946 for extinction coefficient) provided significant improvement of the accuracy of models (RMSE = 2.26) to predict the absorption band, but the accuracy of models for the extinction coefficient practically did not change (RMSE = 0.52).

Interestingly, a development of models using the 335 compounds from the NOVEL set contributed highly predictive models with significantly higher accuracy (RMSE = 0.5 for absorption and RMSE = 0.042 for extinction coefficient). Since models developed using NOVEL set were based on compounds with limited chemical diversity (2,8,12,18-tetramethyl-3,7,13,17-tetraalkyl-5,15-diphenylporphyrins 3,7,13,17-tetrame-thyl-2,8,12,18-tetraalkyl-5,15-diphenylporphyrins, as well as their zinc complexes), they failed to predict molecules from the PORPHYRINS and JOUNG set, since most of the predictions were outside of the applicability domain of this model.

4. Conclusions

In this study, we contributed QSPR models for predicting the optical properties of porphyrins as well as reported synthesis protocols and experimental values for n = 335 porphyrins which are publicly available at http://ochem.eu/article/140403 (accessed date is 30.12.2021). We showed that a better strategy for this chemical class was to develop local models for porphyrins rather than to extend diverse sets of dyes with additional spectral properties of these compounds. While we could successfully model the Soret band, we could not obtain models with good accuracy to predict the extinction coefficient when using literature data. The failure to model the second property could be attributed to the

experimental inconsistency of data obtained from various sources. Indeed, we obtained excellent models for both studied properties when using experimental data (NOVEL set) measured in our laboratory. Unfortunately, because of the very limited chemical diversity, models based on the NOVEL set have a limited applicability domain.

Thus, when analyzing spectral properties of chemical dyes, a possibility to develop local models to cover the studied class of molecules should not be overlooked. While such models may not cover the whole chemical space of dyes, they could be adequate to accurately predict the investigated compounds in particular for properties, such as extinction coefficient, which strongly depend on the used experimental protocol. An attempt to combine in one set inconsistent data could result in a low quality model. More is not always better!

The developed QSPR models for porphyrins can be used to predict their optical properties before they are actually synthesized. This could help to identify compounds with desired sets of properties, significantly reduce development costs, and to accelerate the development of new functional optical materials for electronic and optoelectronic applications.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/article/10.3390/ijms23031201/s1.

Author Contributions: Conceptualization, N.Z.M.; methodology, I.V.T. and N.Z.M.; validation, A.I.R.; formal analysis, O.A.D. and A.I.R.; investigation, O.A.D. and A.I.R.; resources, N.Z.M.; data curation, N.Z.M.; writing—original draft preparation, O.A.D.; writing—review and editing, I.V.T. and A.I.R.; visualization, O.A.D., A.I.R., and I.V.T.; supervision, N.Z.M.; project administration, I.V.T. and N.Z.M.; funding acquisition, I.V.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported with a grant of the Ministry of Science and Higher Education of the Russian Federation [No. 075-15-2021-579].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The supplementary materials contain synthesis protocols and experimental data for 335 novel compounds synthesized in this work as well as data tables supporting statistical analysis provided in this study.

Acknowledgments: The authors thank Alvascience Srl, ChemAxon and Molecular Networks GmbH for a possibility to use descriptors and Corina programs in their study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ptaszyńska, A.A.; Trytek, M.; Borsuk, G.; Buczek, K.; Rybicka-Jasińska, K.; Gryko, D. Porphyrins Inactivate Nosema Spp. Microsporidia. *Sci. Rep.* **2018**, *8*, 5523. https://doi.org/10.1038/s41598-018-23678-8.
1. Varchi, G.; Foglietta, F.; Canaparo, R.; Ballestri, M.; Arena, F.; Sotgiu, G.; Guerrini, A.; Nanni, C.; Cicoria, G.; Cravotto, G.; et al. Engineered Porphyrin Loaded Core-Shell Nanoparticles for Selective Sonodynamic Anticancer Treatment. *Nanomedicine* **2015**, *10*, 3483–3494. https://doi.org/10.2217/nnm.15.150.
2. Mamardashvili, G.; Mamardashvili, N.; Koifman, O. Macrocyclic Receptors for Identification and Selective Binding of Substrates of Different Nature. *Molecules* **2021**, *26*, 5292. https://doi.org/10.3390/molecules26175292.
3. Leng, F.; Liu, H.; Ding, M.; Lin, Q.-P.; Jiang, H.-L. Boosting Photocatalytic Hydrogen Production of Porphyrinic MOFs: The Metal Location in Metalloporphyrin Matters. *ACS Catal.* **2018**, *8*, 4583–4590. https://doi.org/10.1021/acscatal.8b00764.
4. Biesaga, M.; Pyrzyńska, K.; Trojanowicz, M. Porphyrins in Analytical Chemistry. A Review. *Talanta* **2000**, *51*, 209–224. https://doi.org/10.1016/S0039-9140(99)00291-X.
5. Zucca, P.; Neves, C.; Simões, M.; Neves, M.; Cocco, G.; Sanjust, E. Immobilized Lignin Peroxidase-Like Metalloporphyrins as Reusable Catalysts in Oxidative Bleaching of Industrial Dyes. *Molecules* **2016**, *21*, 964. https://doi.org/10.3390/molecules21070964.
6. Dini, D.; Calvete, M.J.F.; Hanack, M. Nonlinear Optical Materials for the Smart Filtering of Optical Radiation. *Chem. Rev.* **2016**, *116*, 13043–13233. https://doi.org/10.1021/acs.chemrev.6b00033.

7. de la Torre, G.; Bottari, G.; Sekita, M.; Hausmann, A.; Guldi, D.M.; Torres, T. A Voyage into the Synthesis and Photophysics of Homo- and Heterobinuclear Ensembles of Phthalocyanines and Porphyrins. *Chem. Soc. Rev.* **2013**, *42*, 8049. https://doi.org/10.1039/c3cs60140d.
8. Saito, S.; Osuka, A. Expanded Porphyrins: Intriguing Structures, Electronic Properties, and Reactivities. *Angew. Chem. Int. Ed.* **2011**, *50*, 4342–4373. https://doi.org/10.1002/anie.201003909.
9. Mamardashvili, N.Z.; Golubchikov, O.A. Spectral Properties of Porphyrins and Their Precursors and Derivatives. *Russ. Chem. Rev.* **2001**, *70*, 577–606. https://doi.org/10.1070/RC2001v070n07ABEH000661.
10. Nemykin, V.N.; Hadt, R.G. Interpretation of the UV−vis Spectra of the Meso(Ferrocenyl)-Containing Porphyrins Using a TDDFT Approach: Is Gouterman's Classic Four-Orbital Model Still in Play? *J. Phys. Chem. A* **2010**, *114*, 12062–12066. https://doi.org/10.1021/jp1083828.
11. Wojciechowski, K.; Szadowski, J. Effect of the Sulphonic Group Position on the Properties of Monoazo Dyes. *Dye. Pigment.* **2000**, *44*, 137–147. https://doi.org/10.1016/S0143-7208(99)00085-6.
12. Azuma, K.; Suzuki, S.; Uchiyama, S.; Kajiro, T.; Santa, T.; Imai, K. A Study of the Relationship between the Chemical Structures and the Fluorescence Quantum Yields of Coumarins, Quinoxalinones and Benzoxazinones for the Development of Sensitive Fluorescent Derivatization Reagents. *Photochem. Photobiol. Sci.* **2003**, *2*, 443. https://doi.org/10.1039/b300196b.
13. Adachi, M.; Nakamura, S. Comparison of the INDO/S and the CNDO/S Method for the Absorption Wavelength Calculation of Organic Dyes. *Dye. Pigment.* **1991**, *17*, 287–296. https://doi.org/10.1016/0143-7208(91)80021-Z.
14. Sham, L.J.; Kohn, W. One-Particle Properties of an Inhomogeneous Interacting Electron Gas. *Phys. Rev.* **1966**, *145*, 561–567. https://doi.org/10.1103/PhysRev.145.561.
15. Bauernschmitt, R.; Ahlrichs, R. Treatment of Electronic Excitations within the Adiabatic Approximation of Time Dependent Density Functional Theory. *Chem. Phys. Lett.* **1996**, *256*, 454–464. https://doi.org/10.1016/0009-2614(96)00440-X.
16. Adamo, C.; Jacquemin, D. The Calculations of Excited-State Properties with Time-Dependent Density Functional Theory. *Chem. Soc. Rev.* **2013**, *42*, 845–856. https://doi.org/10.1039/C2CS35394F.
17. Hahn, D.K.; Callis, P.R. Lowest Triplet State of Indole: An Ab Initio Study. *J. Phys. Chem. A* **1997**, *101*, 2686–2691. https://doi.org/10.1021/jp963146m.
18. Schüller, A.; Goh, G.B.; Kim, H.; Lee, J.-S.; Chang, Y.-T. Quantitative Structure-Fluorescence Property Relationship Analysis of a Large BODIPY Library. *Mol. Inf.* **2010**, *29*, 717–729. https://doi.org/10.1002/minf.201000089.
19. Grimme, S. A Simplified Tamm-Dancoff Density Functional Approach for the Electronic Excitation Spectra of Very Large Molecules. *J. Chem. Phys.* **2013**, *138*, 244104. https://doi.org/10.1063/1.4811331.
20. Heil, A. Development and Implementation of New DFT/MRCI Hamiltonians for Odd and Even Numbers of Electrons. PhD thesis, Henrich Hein University in Düsseldorf, Düsseldorf, Northrhine-Westphalia, Germany, 2019, 8 September.
21. Li, G.-Z.; Yang, J.; Song, H.-F.; Yang, S.-S.; Lu, W.-C.; Chen, N.-Y. Semiempirical Quantum Chemical Method and Artificial Neural Networks Applied for λ m ax Computation of Some Azo Dyes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2047–2050. https://doi.org/10.1021/ci049941b.
22. Li, H. Quantitative Structure—Property Relationships for Colour Reagents and Their Colour Reactions with Cerium Using Computational Neural Networks. *Talanta* **1997**, *44*, 203–211. https://doi.org/10.1016/S0039-9140(96)02034-6.
23. Shi, J.; Luan, F.; Zhang, H.; Liu, M.; Guo, Q.; Hu, Z.; Fan, B. QSPR Study of Fluorescence Wavelengths (Λex/Λem) Based on the Heuristic Method and Radial Basis Function Neural Networks. *QSAR Comb. Sci.* **2006**, *25*, 147–155. https://doi.org/10.1002/qsar.200510142.
24. Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Tansila, N.; Naenna, T.; Prachayasittikul, V. Prediction of GFP Spectral Properties Using Artificial Neural Network. *J. Comput. Chem.* **2007**, *28*, 1275–1289. https://doi.org/10.1002/jcc.20656.
25. Shedden, K.; Brumer, J.; Chang, Y.T.; Rosania, G.R. Chemoinformatic Analysis of a Supertargeted Combinatorial Library of Styryl Molecules. J. Chem. *Inf. Comput. Sci.* **2003**, *43*, 2068–2080. https://doi.org/10.1021/ci0341215.
26. Joung, J.F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D.H.; Park, S. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. *JACS Au* **2021**, *1*, 427–438. https://doi.org/10.1021/jacsau.1c00035.
27. Xu, J.; Zheng, Z.; Chen, B.; Zhang, Q. A Linear QSPR Model for Prediction of Maximum Absorption Wavelength of Second-Order NLO Chromophores. *QSAR Comb. Sci.* **2006**, *25*, 372–379. https://doi.org/10.1002/qsar.200530143.
28. Yao, X.; Wang, Y.; Zhang, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. Radial Basis Function Neural Network-Based QSPR for the Prediction of Critical Temperature. *Chemom. Intell. Lab. Syst.* **2002**, *62*, 217–225. https://doi.org/10.1016/S0169-7439(02)00017-5.
29. Xia, Z.; Karpov, P.; Popowicz, G.; Tetko, I.V. Focused Library Generator: Case of Mdmx Inhibitors. *J. Comput. Aided Mol. Des.* **2020**, *34*, 769–782. https://doi.org/10.1007/s10822-019-00242-8.
30. Joung, J.F.; Han, M.; Jeong, M.; Park, S. Experimental Database of Optical Properties of Organic Compounds. *Sci. Data* **2020**, *7*, 295. https://doi.org/10.1038/s41597-020-00634-8.
31. DB for Chromophore. Available online: https://doi.org/10.6084/m9.figshare.12045567.v2 (accessed on 29 December 2021).
32. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V.V.; Tanchuk, V.Y.; et al. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554. https://doi.org/10.1007/s10822-011-9440-2.
33. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.

34. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA—Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *CAD* **2008**, *4*, 191–198. https://doi.org/10.2174/157340908785747465.
35. Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337–1344. https://doi.org/10.1021/ci800038f.
36. Mauri, A. AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*; Roy, K., Ed.; Methods in Pharmacology and Toxicology; Springer US: New York, NY, USA, 2020; pp. 801–820. ISBN 978-1-07-160150-1.
37. Polishchuk, P.; Madzhidov, T.; Gimadiev, T.; Bodrov, A.; Nugmanov, R.; Varnek, A. Structure–Reactivity Modeling Using Mixture-Based Representation of Chemical Reactions. *J. Comput. Aided Mol. Des.* **2017**, *31*, 829–839. https://doi.org/10.1007/s10822-017-0044-3.
38. Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581. https://doi.org/10.1021/cr00023a012.
39. Karpov, P.; Godin, G.; Tetko, I.V. Transformer-CNN: Swiss Knife for QSAR Modeling and Interpretation. *J. Cheminform.* **2020**, *12*, 17. https://doi.org/10.1186/s13321-020-00423-w.
40. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. https://doi.org/10.1021/ci00057a005.
41. OCHEM Materials Home—OCHEM Materials—EADMET. Available online: http://docs.ochem.eu/ (accessed on 28 December 2021).
42. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746. https://doi.org/10.1021/ci800151m.
43. Ghosh, D.; Koch, U.; Hadian, K.; Sattler, M.; Tetko, I.V. Highly Accurate Filters to Flag Frequent Hitters in AlphaScreen Assays by Suggesting Their Mechanism. *Mol. Inf.* **2021**, *41*, e2100151. https://doi.org/10.1002/minf.202100151.
44. Tetko, I.V.; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A.E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163,000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990–2000. https://doi.org/10.1021/ci400213d.
45. Vorberg, S.; Tetko, I.V. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Mol. Inf.* **2014**, *33*, 73–85. https://doi.org/10.1002/minf.201300030.
46. Ksenofontov, A.A.; Lukanov, M.M.; Bocharov, P.S.; Berezin, M.B.; Tetko, I.V. Deep Neural Network Model for Highly Accurate Prediction of BODIPYs Absorption. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *267*, 120577. https://doi.org/10.1016/j.saa.2021.120577.