AN ATTEMPT TO BOOST MOLECULAR DESCRIPTORS WITH QUANTUM-DERIVED FEATURES IN PREDICTION OF MAXIMUM EMISSION WAVELENGTHS OF CHROMOPHORES.

A PREPRINT

Bartłomiej Fliszkiewicz
Faculty of New Technologies and Chemistry
Military University of Technology
Warsaw, Poland
bartlomiej.fliszkiewicz@wat.edu.pl

May 4, 2021

ABSTRACT

The following research assesses the capability of machine learning in predicting maximum emission wavelength of organic compounds. The predictions are based on structure descriptors and fingerprints widely applied in cheminformatics. In an attempt to further improve accuracy, developed machine learning models were enriched with quantum mechanics derived features. Multi linear, gradient boosting and random forest regressions were applied. Computers were trained and tested with database of experimental data of optical properties.

***K*eywords** Machine learning · Molecular descriptors · Quantum chemistry descriptors · QSPR

1 Introduction

Machine learning gains a lot of focus these days. The wide spectrum of tools[1][2] and rushing growth of database volumes cause machine learning to influence every aspect of life[3]. Algorithms are provided with databases containing so called features that describe observations as well as properties that will further be associated with the features. As a consequence the machine is capable of predicting properties based on the features it was trained on.

Quantitative structure - activity/property relation (QSAR/QSPR) methods and algorithms are founded on an assumption that molecular structure is correlated with molecule's properties. In order to find similar compounds from vast databases a wide range of molecular descriptors were developed[4] alongside with molecular fingerprints like Morgan fingerprint[5] and MACCS keys. Cherkasov et al [6], Varnek and Baskin [7] give a wide overview of the core of QSAR/QSPR and cheminformatics, their history, advances and perspectives.

With the growing number of freely accessible databases and open source tools(eg. Python[8], RDKit[9], Scikit Learn[10], Matplotlib[11] and Seaborn[12]) it is easy to learn and apply machine learning or at least conduct data driven research. The cheminformatics gain also from openness of researchers that publish their code and data pipeline[13], simplifying the knowledge acquiring process and making QSAR/QSPR and machine learning adaptable to other problems.

2 Material and Methods

2.1 Workflow

Workflow diagram is presented in figure 1. After preprocessing and generating features the data was split into training and test subsets. Machine learning algorithms were introduced with training sets and their performance was checked

by making predictions of maximum emissions of compounds from test sets. The predictions were checked with real values from test sets and the errors were studied. Best performing machine learning models were chosen based on mean absolute error, mean squared error, maximum error and R^2 parameter. Chosen models are to be validated with compounds from laboratory.

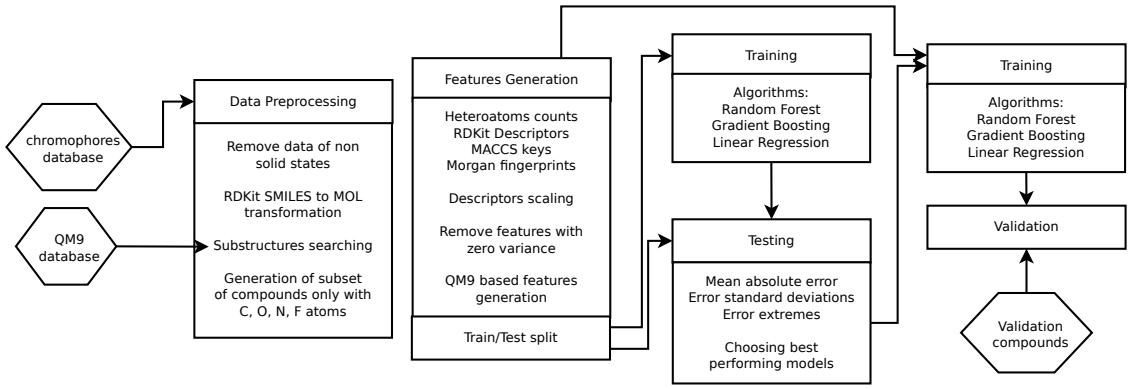

Figure 1: Workflow diagram.

2.2 Datasets

The organic compounds optical properties database[14] contains over 20,000 rows which are combination of 7,016 chromophores in 365 solvents and 17 solid matrices or in solid states. Chromophores included in the database consist of maximum 150 atoms (except H) of C, N, O, S, F, Cl, Br, I, Se, Te, Si, P, B, Sn, Ge. Out of these chromophores there are 956 that have reported properties in solid states (column Solvent and Chromophore are equal). 897 of solid state chromophores have non null value of maximum emission wavelength (nm)(dataset 1). Only solid state compounds were taken into account to avoid bias caused by solvent effects on maximum emission. A subset of compounds containing only C, O, N, F, H atoms was also examined in the study (dataset 2).

The QM9 database[15][16] contains 133,885 small organic compounds of up to 9 atoms (except H) of C, O, N, F. These compounds are a subset of GDB-17 chemical universe database[17] containing 166 billion of organic compounds. The subset includes various density functional theory(DFT) calculated quantum quantities (eg. HOMO and LUMO eigenvalues). The QM9 database was downloaded from MoleculeNet[18] since it is packed into .csv format.

The extent of maximum emission of datasets is shown on figure 2.

2.3 Machine learning models

Random Forest Regression(RFR), Multi Linear Regression(MLR) and Gradient Boosted Regression(GBR) models from Scikit-Learn[10] Python module were taken into the studies.

2.4 Feature engineering

All of available RDKit molecular descriptors(208), MACCS keys(167) and Morgan fingerprints(1024) were calculated for every chromophore. Numbers of heteroatoms were also calculated(14 for dataset 1 and 3 for dataset 2). Values of molecular descriptors were further scaled. Features that did not change across datasets were deleted.

Molecular descriptors, MACCS keys, Morgan fingerprints and numbers of heteroatoms were applied to all of tested models and they will be further referenced as universal features. After all basic data processing procedures applied the dataset 1 contained 896 chromophores and 1312 features and dataset 2 contained 523 chromophores and 1127 features.

Compounds from chromophores database were examined if they contain substructures from QM9 database using RDKit built-in function of substructure recognition. In order to provide machine learning models with more features, in an attempt to improve their prediction capabilities various additional quantities were calculated from substructures quantum properties. Karelson et al [19] covered usage of quantum modelling calculations as descriptors in QSAR/QSPR research, although the calculations were prosecuted with whole molecules not their fragments.

Finally 14 different models were trained and tested with following features.

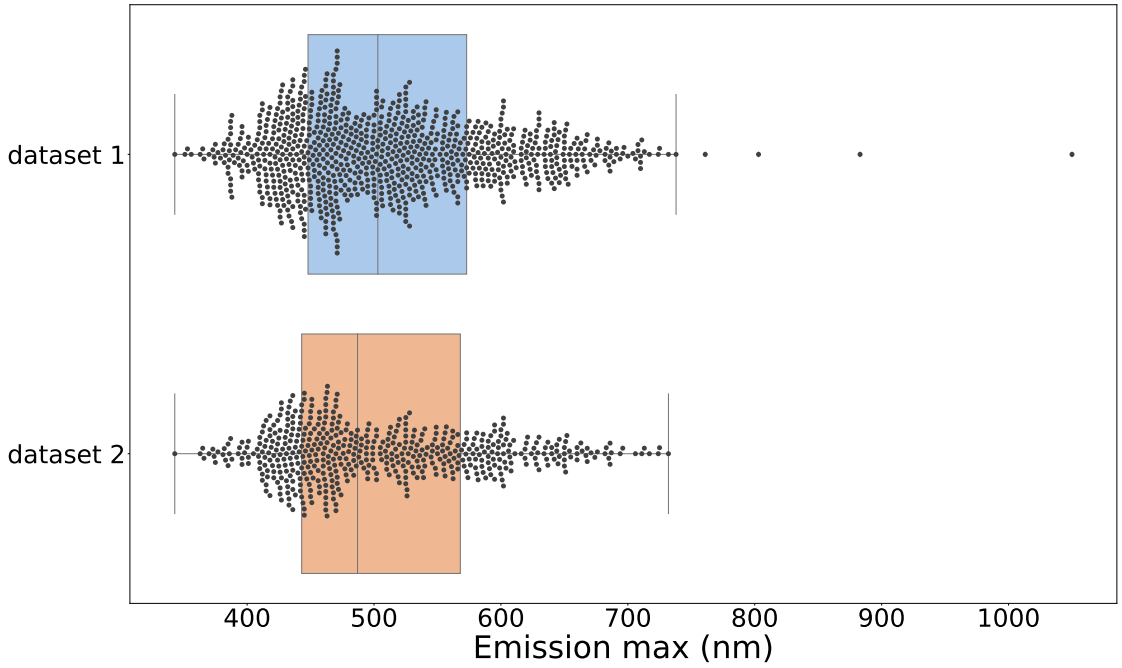

Figure 2: Emission distribution.

2.4.1 Model 1

No QM9 based features were calculated. Only universal features were applied to ML models. This approach demands the least computational time of all models covered in this paper as it does not need searching of QM9 database substructures.

2.4.2 Model 2

The sum of all quantum features from QM9 database was calculated multiplied by number of pattern(substructure) occurrences.

$$\sum_{i}^{N} n_i \epsilon_{HOMO_i}, \sum n_i \epsilon_{LUMO_i}, ... \tag{1}$$

where i - index of recognised pattern, n - number of pattern occurences.

2.4.3 Model 3

Only QM9 based features from model 2 were input into ML algorithms. With this approach it is possible to assess if non standard features are competitive to traditional ones.

2.4.4 Models 4 - 14

Features generated in these models are result of various mathematical operations of mostly eigenvalues ϵ_{HOMO}, ϵ_{LUMO}, polarizability, α, dipole moment, μ, zero point vibrational energy, $zpve$ and electronic spatial extent, $\langle R^2 \rangle$. They were developed in the beginning of the research, before applying molecular descriptors and fingerprints.

It is worth noting that the RDKit built-in method to detect substructures sometimes makes mistakes. In figure 3 is a chromophore from the database and substructures from QM9 database that were detected in the molecule. The last of detected substructures (circled) is not present in the molecule from database.

In the study the faultily detected substructures were not removed from collections of substructures and were taken into account when features were generated.

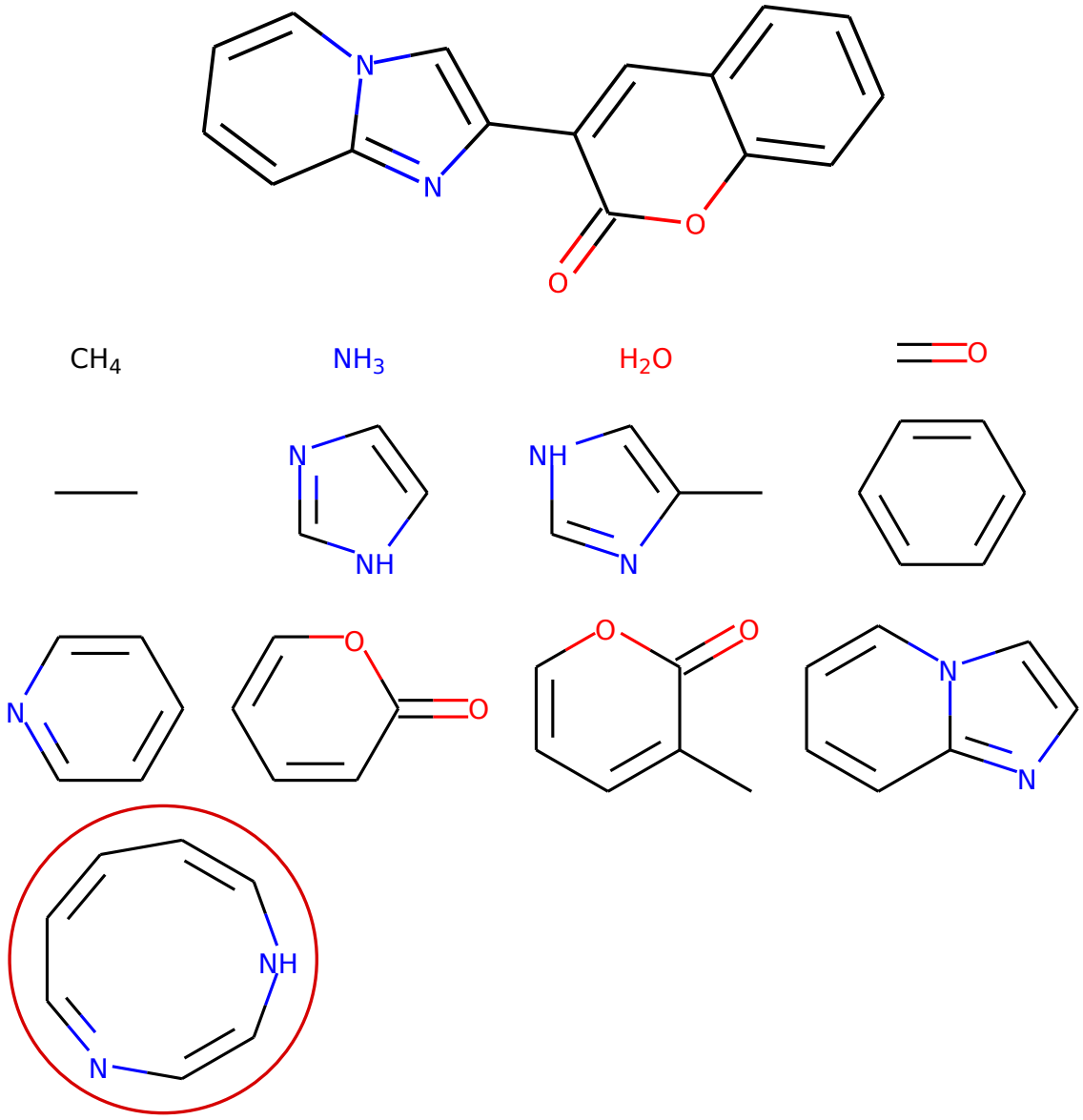

Figure 3: Molecule with its detected substructures. The circled one is a mismatch.

3 Results and Discussion

All developed machine learning models were scored using mean absolute error(MAE) and mean squared error(MSE). To further assess models' performance, R^2 and maximum errors were calculated. Scoring values are presented in tables 1 and 2 regarding dataset 1 and dataset 2 accordingly. Since ensemble algorithms (random forest and gradient boosting) outperformed linear regression they will be covered separately.

3.0.1 Gradient Boosting and Random Forest algorithms

When quantum derived features are applied there are minimum changes in prediction accuracy. The decreased performance of model 3 is a result of exclusion of molecular descriptors in training and predicting process.

Mean absolute error indicates that RFR perform better than GBR (fig. 4), particularly when trained on dataset 1. There is also improvement in performance when models are trained on dataset 2.

Table 1: Scoring values of dataset 1. ME - maximum error, MAE - mean absolute error, MSE - mean squared error.

algorithm	Gradient Boosting				Random Forest			
scoring type	ME	MAE	MSE	R^2	ME	MAE	MSE	R^2
model								
1	156.5	31.0	1734	0.759	165.0	30.1	1815	0.749
2	153.3	31.1	1749	0.756	163.5	29.9	1790	0.753
3	251.8	58.2	5584	0.223	241.1	53.2	5047	0.296
4	153.4	31.0	1747	0.757	163.3	30.0	1805	0.751
5	156.0	31.1	1762	0.754	164.3	30.0	1796	0.752
6	153.0	31.6	1790	0.751	164.0	30.1	1805	0.750
7	152.9	31.3	1774	0.752	163.4	30.2	1815	0.749
8	154.7	30.9	1745	0.757	165.2	30.3	1831	0.747
9	153.2	30.9	1740	0.757	166.0	30.3	1833	0.747
10	153.1	31.2	1739	0.757	165.5	30.3	1817	0.749
11	158.6	31.3	1781	0.751	165.8	30.3	1829	0.747
12	153.1	31.1	1765	0.754	167.0	30.0	1813	0.750
13	155.1	31.1	1759	0.754	163.4	30.1	1811	0.750
14	159.5	31.1	1761	0.754	164.1	30.2	1821	0.748
algorithm	LM1				LM2			
model								
1	>500	>100	>10000	<-1	-	-	-	-
2	>500	47.2	>10000	<-1	275.0	59.0	6070	0.133
3	275.0	59.0	6070	0.133	275.0	59.0	6070	0.133
4	>500	48.7	>10000	<-1	250.0	67.2	6837	0.058
5	>500	48.6	>10000	<-1	363.5	69.2	9055	-0.307
6	>500	48.7	>10000	<-1	256.9	68.0	7042	0.027
7	>500	49.1	>10000	<-1	253.5	65.9	6724	0.071
8	>500	47.9	>10000	<-1	278.1	65.8	6775	0.062
9	>500	49.1	>10000	<-1	255.9	62.0	6211	0.133
10	>500	50.2	>10000	<-1	254.2	66.7	6835	0.052
11	>500	50.5	>10000	<-1	255.3	67.2	6910	0.042
12	>500	50.1	>10000	<-1	259.0	69.0	7196	0.007
13	408.8	49.3	7207	-0.080	451.5	69.6	>10000	-0.872
14	>500	49.8	>10000	<-1	238.4	62.3	6061	0.155

Models trained and tested on dataset 2 perform about 3nm better on average (MAE). Most probably this phenomenon is caused by better homogenity in compounds classes in dataset 2. The dataset of compounds composed only of C, O, N, F atoms also lacks maximum emission outliers which could affect the performance of prediction.

In the opposition to MAE, the values of mean squared error (fig. 5) imply that gradient boosting performs better than random forest algorithm.

The further evidence of GBR's more accurate predictions fall to maximum error (fig. 6). The trend is that GBR perform better than RFR and the first's worst predictions are about 9nm more accurate then the second's. There is also about 35nm difference in maximum error between predictions with models trained on different datasets.

Figure 7 shows values of R^2 scoring indicator. The difference between models is very slight but the advantage of models trained on dataset 2 is further acknowledged.

3.1 Multi Linear Models

Since scoring values of multi linear regression in most cases were inapplicable when fed with the datasets with all features, 2 alternative approaches were employed. Linear regression algorithm was provided with both features from molecular descriptors and generated from pattern recognition from QM9 database (further referenced as LM1) or only features generated from QM9 database(LM2). In this new approach LM1 model 3 is the same as LM2 models 2 and 3. Except for model 3, LM1 scoring results disqualified this prediction method.

Most scoring indicators calculated in this research imply that in case of linear regression the best predicion accuracy is achieved for model 3. It is also worth noting that to obtain somehow applicable results linear regression models should be provided with features excluding molecular descriptors.

Table 2: Scoring values of dataset 2.

algorithm			Gradient Boosting				Random Forest	
scoring type	ME	MAE	MSE	R^2	ME	MAE	MSE	R^2
model								
1	118.2	27.4	1373	0.782	127.7	27.3	1425	0.773
2	117.9	27.4	1370	0.783	129.2	27.6	1461	0.768
3	187.1	54.5	4741	0.255	179.0	52.3	4517	0.291
4	118.8	27.9	1402	0.778	127.0	27.7	1452	0.770
5	118.8	27.3	1383	0.780	127.2	27.3	1425	0.773
6	117.2	27.7	1391	0.779	126.5	27.6	1436	0.772
7	120.3	27.6	1397	0.778	126.9	27.6	1458	0.768
8	116.5	27.5	1358	0.784	127.6	27.9	1476	0.766
9	117.1	27.8	1389	0.779	126.4	27.5	1438	0.771
10	126.0	27.7	1437	0.772	127.3	27.2	1433	0.771
11	123.5	27.6	1409	0.776	128.2	27.3	1437	0.771
12	120.3	27.5	1392	0.779	126.3	27.3	1433	0.772
13	120.9	27.3	1376	0.782	128.7	27.5	1447	0.770
14	123.1	27.7	1400	0.777	127.1	27.5	1448	0.769
			LM1				LM2	
model								
1	>500	>100	>10000	<-1	-	-	-	-
2	>500	51.5	>10000	-0.937	187.8	57.6	5042	0.215
3	187.8	57.6	5042	0.215	187.8	57.6	5042	0.215
4	469.9	48.8	>10000	-0.681	223.5	64.1	6419	0.011
5	494.8	48.9	>10000	-0.908	298.4	67.8	9253	-0.373
6	460.7	48.9	9551	-0.536	197.9	63.5	5980	0.070
7	346.7	47.1	6727	-0.070	198.5	61.7	5748	0.107
8	396.6	47.5	7994	-0.293	191.2	61.0	5554	0.135
9	311.5	46.1	5893	0.075	206.0	57.0	5138	0.204
10	318.1	46.5	6151	0.027	201.3	63.2	5920	0.085
11	325.3	46.6	6292	0.002	201.1	63.8	6003	0.071
12	416.7	47.5	8574	-0.393	198.2	65.8	6232	0.036
13	>500	49.2	>10000	<-1	386.4	68.2	>10000	<-1
14	421.1	47.1	8979	-0.478	209.2	59.7	5426	0.159

4 Conclusions

Presented method of predicting maximum emissions of organic compounds has limited functionality and gives loose insight into the property. There is possibility to polish the method to give better predictions.

The database of chromophores is too small or contains compounds that represent too wide number of organic compounds classes. Since its vulnerability to database correspondence to compound being assessed, the method should be provided with proper database. Alternatively machine could be trained on the go with a subset of bigger database chosen on compound's similarity. With highly probable emergence of new datasets, machine learning based approaches to QSPR will undoubtedly improve their performance.

Induction of quantum properties of compounds' substructures did not improve the accuracy of prediction of emission with RFR and GBR. Although outperformed, MLR was able to give sensible results when fed only with quantum-derived descriptors. With development of new features or with alternative fragments based approach these quantum-chemistry descriptors may play some role in prediction capabilities.

There is also possibility to apply other machine learning algorithms or change algorithms' parameters used in the research to better fulfil the aim of the studies. Only 2D molecular descriptors were utilized to train machine learning models. There are fields that 3D molecular descriptors perform better then 2D ones[20]. Some other applications of machine learning in predictions of organic compounds emission wavelengths were published[21][22].

The biggest advantage of proposed method is its speed. When introduced into web based service with locally created database in research laboratory or for specific team, it offers rapid assessment of emission property of projected compounds.

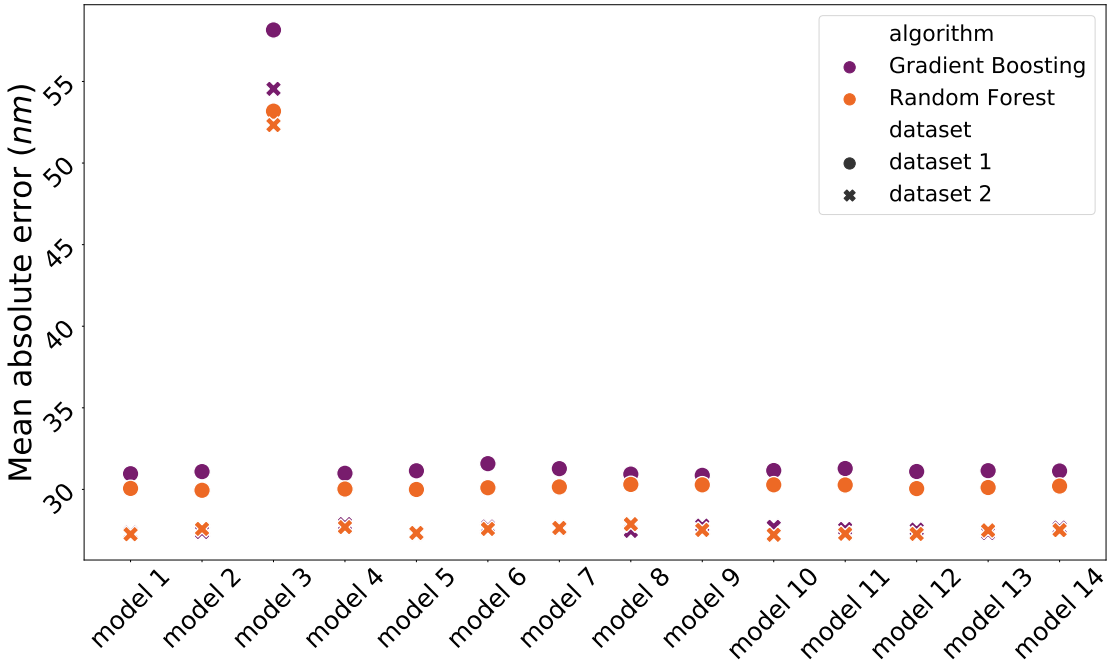

Figure 4: Mean absolute error across models and datasets.

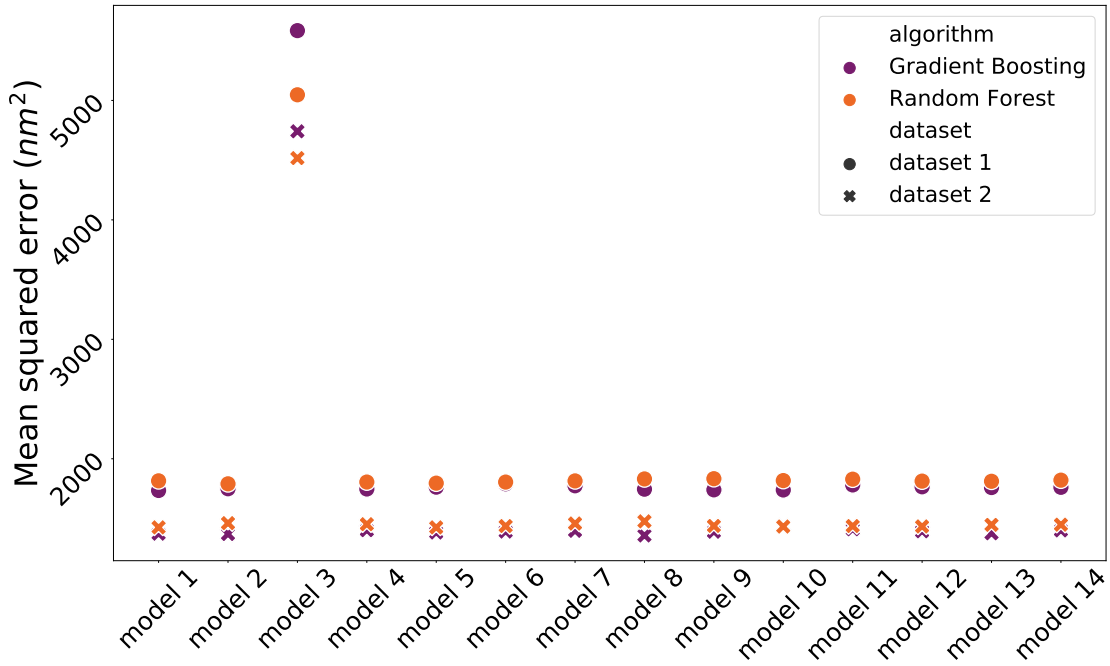

Figure 5: Mean squared error across models and datasets.

5 Data avability statements

Datasets are available at:

- QM9 dataset - `moleculenet.ai/datasets-1`
- database of chromophores - `figshare.com/articles/dataset/DB_for_chromophore/12045567`

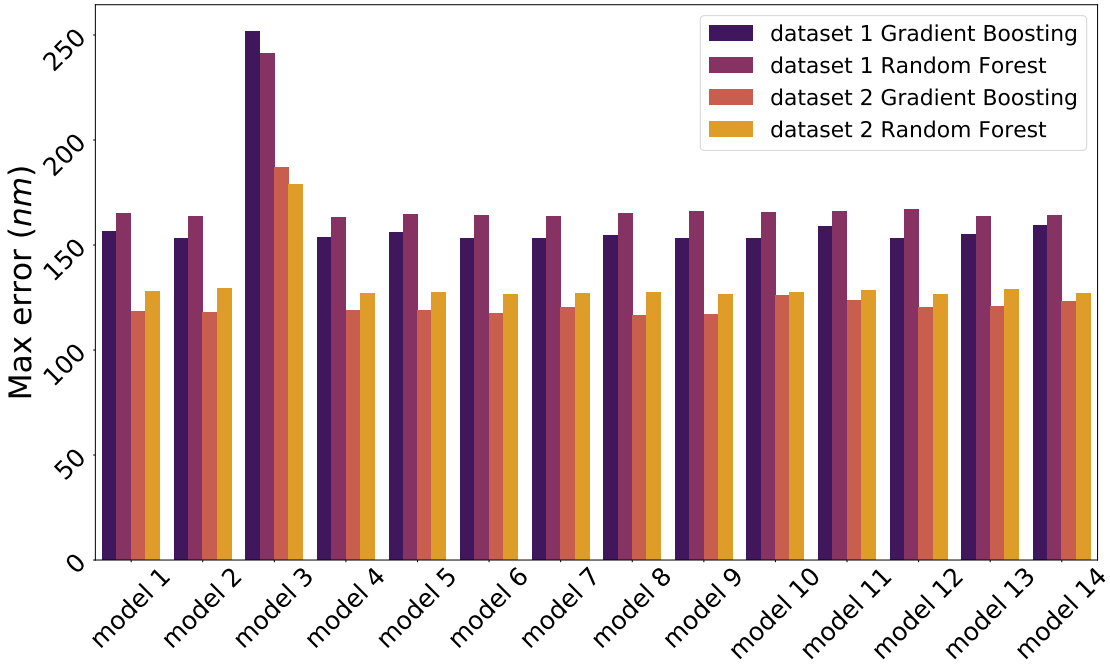

Figure 6: Maximum error across models and datasets.

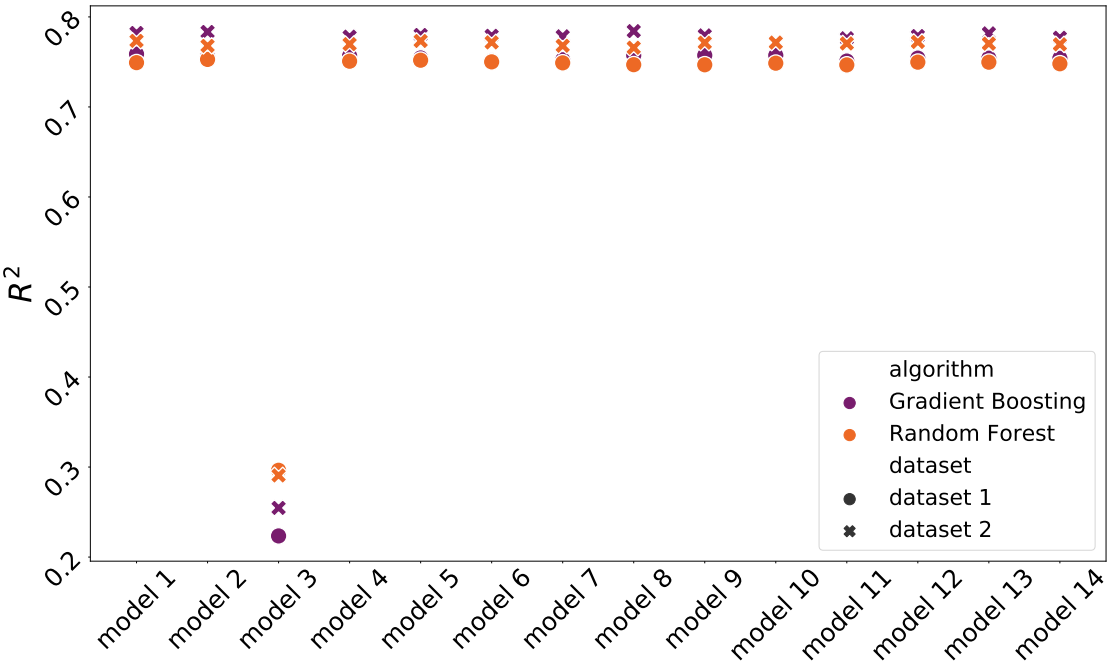

Figure 7: R^2 across models and datasets.

The code of data processing, features generation, models development and scoring and scoring visualizations is available at both and `https://doi.org/10.6084/m9.figshare.14533929.v1` and `https://github.com/BartlomiejF/articles-molecular-quantum-descriptors`.

Please be aware that different versions of RDKit may output results that vary from those presented in this paper.

Eventually at the aforementioned code resources there will also be available Raspberry Pi deployable web application utilizing models developed in the research.

References

[1] Himanshi Bansal and K. Sharma. A review study on various algorithms of machine learning. *Journal of emerging technologies and innovative research*, 2020.

[2] S. Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39, 2019.

[3] Annina Simon, Mahima Deo, Venkatesan Selvam, and Ramesh Babu. An overview of machine learning and its applications. *International Journal of Electrical Sciences & Engineering*, 1:22–24, 01 2016.

[4] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*, volume 1, page 1252. 07 2009.

[5] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965.

[6] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. Qsar modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, Jun 2014.

[7] Alexandre Varnek and Igor Baskin. Machine learning methods for property prediction in chemoinformatics: Quo vadis? *Journal of Chemical Information and Modeling*, 52(6):1413–1437, Jun 2012.

[8] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[9] RDKit: Open-source cheminformatics. `http://www.rdkit.org`. [Online; accessed 11-February-2021].

[10] 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 E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[12] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[13] Gabriel Idakwo, Sundar Thangapandian, Joseph Luttrell, Yan Li, Nan Wang, Zhaoxian Zhou, Huixiao Hong, Bei Yang, Chaoyang Zhang, and Ping Gong. Structure–activity relationship-based chemical classification of highly imbalanced tox21 datasets. *Journal of Cheminformatics*, 12(1):66, Oct 2020.

[14] J F Joung, M Han, M Jeong, and S Park. Experimental database of optical properties of organic compounds. *Scientific Data*, 7:295, 09 2020. https://doi.org/10.1038/s41597-020-00634-8.

[15] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, Aug 2014.

[16] Quantum machine. `http://quantum-machine.org/datasets/`. [Online; accessed 20-January-2021].

[17] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, Nov 2012.

[18] Moleculenet. `http://moleculenet.ai/datasets-1/`. [Online; accessed 20-January-2021].

[19] Mati Karelson, Victor S. Lobanov, and Alan R. Katritzky. Quantum-chemical descriptors in qsar/qspr studies. *Chemical Reviews*, 96(3):1027–1044, Jan 1996.

[20] Kaifu Gao, Duc Duy Nguyen, Vishnu Sresht, Alan M. Mathiowetz, Meihua Tu, and Guo-Wei Wei. Are 2d fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.*, 22:8373–8390, 2020.

[21] Zong-Rong Ye, I.-Shou Huang, Yu-Te Chan, Zhong-Ji Li, Chen-Cheng Liao, Hao-Rong Tsai, Meng-Chi Hsieh, Chun-Chih Chang, and Ming-Kang Tsai. Predicting the emission wavelength of organic molecules using a combinatorial qsar and machine learning approach. *RSC Adv.*, 10:23834–23841, 2020.

[22] Cheng-Wei Ju, Hanzhi Bai, Bo Li, and Rizhang Liu. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission wavelengths and quantum yields. *Journal of Chemical Information and Modeling*, 61(3):1053–1065, Mar 2021.