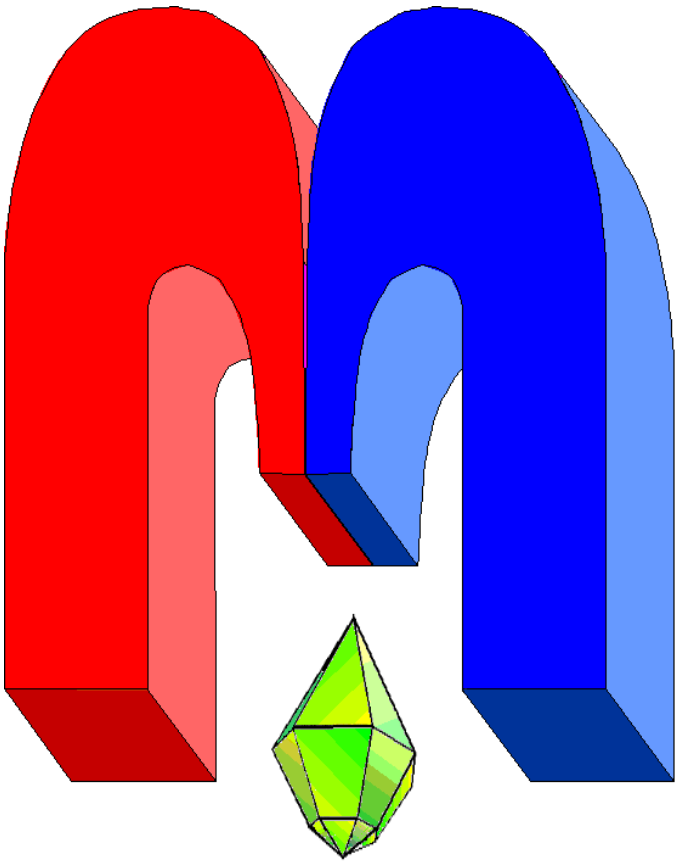
ISSN 2072-5981
doi: 10.26907/mrsej



Magnetic Resonance in Solids

Electronic Journal

Volume **22**
Issue **1**
Article No **20102**
1-11 pages
2020
doi: **10.26907/mrsej-20102**

http://mrsej.kpfu.ru
http://mrsej.ksu.ru

Established and published by Kazan University
Endorsed by International Society of Magnetic Resonance (ISMAR)
Registered by Russian Federation Committee on Press (#015140),
August 2, 1996
First Issue appeared on July 25, 1997

"***Magnetic Resonance in Solids. Electronic Journal***" (**MRS*ej***) is a peer-reviewed, all electronic journal, publishing articles which meet the highest standards of scientific quality in the field of basic research of a magnetic resonance in solids and related phenomena.

Indexed and abstracted by
Web of Science (*ESCI*, Clarivate Analytics, from 2015), *Scopus* (Elsevier, from 2012), *RusIndexSC* (*eLibrary*, from 2006), *Google Scholar*, *DOAJ*, *ROAD*, *CyberLeninka* (from 2006), *SCImago Journal & Country Rank*, etc.

Technical Editor
Maxim **Avdeev** (KFU, Kazan)

* In Kazan University the Electron Paramagnetic Resonance (EPR) was discovered by Zavoisky E.K. in 1944.

Calibrating optical spectra using machine learning algorithms

M.A. Shakirov

Kazan Federal University, Kremlevskaya 18, 420008 Kazan, Russia

E-mail: mars.shakirov@gmail.com

(Received December 30, 2019; revised January 17, 2020;
accepted January 18, 2020; published January 25, 2020)

We suggest an approach using machine learning random forest algorithms to comparing and calibrating the results of calculations of transition energies in organic molecules by ZINDO/S (Zerner's intermediate neglect of differential overlap) and TDDFT (time-dependent density-functional theory) methods. We show how our machine learning model, trained on a relatively small data set can improve the results of semi-empirical methods and obtain absorption spectra comparable to TDDFT calculations.

PACS: 76.30.-v, 68.65.-k
Keywords: machine learning, random forest, TDDFT, semi-empirical calculations, spectroscopy

1. Introduction

There are three main approaches to modeling electronic properties of molecules that, in one or another way, approximate the solution of Schrödinger equation: wavefunction based methods, DFT (density functional theory), semi-empirical methods [1, 2]. As the accuracy of calculations increases, their duration or, in other words, their computational cost increases as well [3]. Recently, a number of studies [4,5] from different fields of science, including computational chemistry [6,7], demonstrate extensive possibilities of machine learning methods to solve the problem described above. However, as a rule, they use a huge amount of data for training. The number of different molecules used varies from a few thousand [8] to hundreds of thousands [9–11]. In our work, we try to solve a less general problem of building a connection between different types of calculations using a relatively small dataset. Specifically, we model the electronic excitation spectra of molecules. For our research, we chose to calculate the first 10 excitation transition with TDDFT method by B3LYP functional [12] and the same thing with using ZINDO/S semi-empirical methods [13]. TDDFT is an extension of DFT, and can be formulated by the time-dependent Kohn-Sham equation [14, 15] for electron density $\rho(\mathbf{r}, t)$:

$$i\frac{\partial}{\partial t}\varphi_i(\mathbf{r}, t) = \left[-\frac{\nabla^2}{2} + V_{eff}[\rho](\mathbf{r}, t)\right]\varphi_i(\mathbf{r}, t). \tag{1}$$

Here $\varphi_i(\mathbf{r}, t)$ are time dependent orbitals, which can be propagated from initial condition. The effective potential $V_{eff}(\mathbf{r}, t)$ can be written as following:

$$V_{eff}(\mathbf{r}, t) = V(\mathbf{r}, t) + \int d^3r' \frac{\rho(\mathbf{r}', t)}{|\mathbf{r} - \mathbf{r}'|} + V_{xc}(\mathbf{r}, t). \tag{2}$$

First two parts are nuclear potential and electron-electron potential (Hartree) potential. Last one is the exchange correlation potential. There are number of functionals to approximate exchange-correlation potential. One of the broadly accepted examples is B3LYP (Becke-3 Parameter-Lee-Yang-Parr [12]). This functional is popular for many reasons. It was one of the first DFT methods that were a significant improvement over Hartree-Fock and fairly robust for a DFT functional. This functional is efficient for calculating UV/Vis spectra, and compared to other functionals it shows good accuracy [16]. For example, singlet and triplet transitions the accuracy

of energy is in the order of 0.1 eV [17]. The Schrödinger equation for molecules is solved by semi-empirical methods using some another level of approximation. All methods in this group can be unite by the idea that the calculation is carried out only for valence electrons; integrals of certain interactions are neglected; standard non-optimized basic functions of electronic orbits are used and some parameters obtained in the experiment are used [2, 18]. The ZINDO/S method is a version of the INDO (Intermediate Neglect of Differential Overlap) method parametrized to reproduce UV and visible optical transitions in the calculation of configuration interaction (CI) with single-part excitations [13]. It has been widely used in studies of organic molecules [19, 20]. Usually, the accuracy of ZINDO results increases with the size of the system [21]. There are many papers [22–24] in which researchers have chosen one or the other method according to their capacity and accuracy requirements. Examples of ZINDO and TDDFT calculation differences are shown in Fig. 1. We would like to show that it is possible to build a bridge between these methods through the simplest machine learning algorithms and to get corrections to ZINDO, reducing the error in comparison with TDDFT. The main goal of this work is to show that there are relatively simple ways to improve quantum-mechanical calculations results without needs of big datasets and much additional computational time.

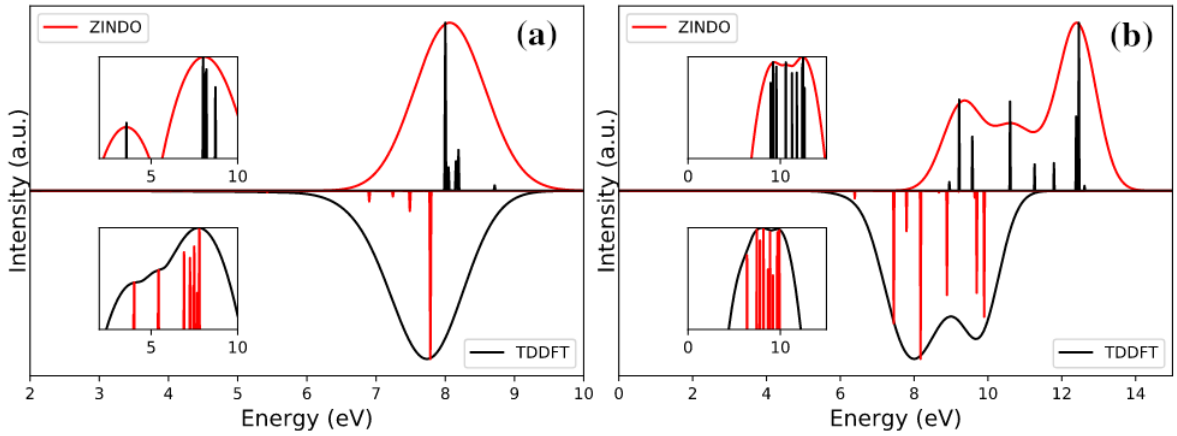


Figure 1. The spectra of two molecules (a) Glutaral and (b) Aziridine computed using TDDFT (bottom) and ZINDO (top). Insets are plotted in same energy range with intensity in logarithmic scale. The graphs show difference between two methods. The spectrum of molecule (a) is well described by ZINDO, while for other molecule (b) there is sufficient difference.

2. Data preparation

Usually working with any neural network using methods needs to compute a large number of training data. For fast fitting and thus rapid hypothesis testing we take very small number (81) different compounds. We upload all compounds using PubChem database facilities [25]. All molecules can be seen in the appendix (Fig. 13). Further preparations for the dataset are shown in the Fig. 2. Based on each optimized molecule, we generate 5 other geometries by small random displacements in the positions of the atoms. As a result of displacements, the distances between atoms are changed less than 10% relatively to initial range. We end up with $5 \times 81 = 405$ different structures and compute for each of them 10 first excitations with two methods. The Fig. 3 shows how computation time depends on the number of atoms in the structure. One can see how the molecular size growing leads to increasing the calculation time, but differently for different methods. The average value of the calculation time ratio is 72.3; the maximum ratio is about 500. The calculated 10 excitations for each of the 405 molecules form 405 pairs

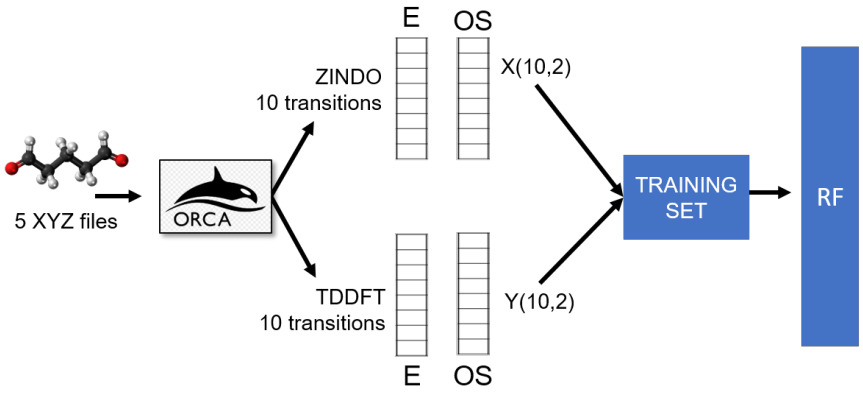


Figure 2. An illustration of how we prepare the dataset. We write the properties of each molecule into a .xyz file and remake 5 times with random deviations in the coordinates of atoms. There is 10 transitions (their energy and oscillator strength) using the TDDFT and ZINDO methods. We use the results of their calculations (matrices of size (2,10)) in the dataset as a $X-Y$ pair for further regression by the Random Forest algorithm.

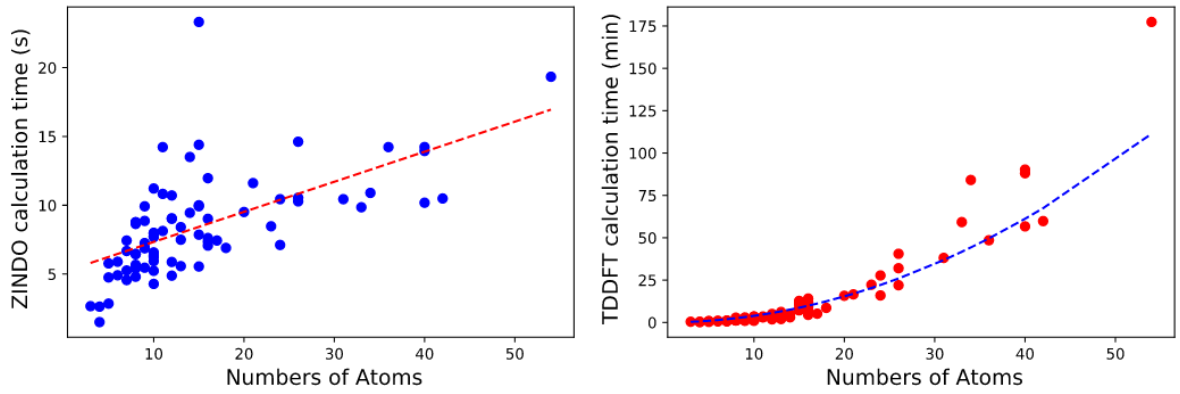


Figure 3. Evaluation of calculation time by two different methods. Depending on the size of the molecule (number of atoms) the calculation time of ZINDO method grows linearly. On the graph almost all values do not exceed 15 seconds. In case of TDDFT, the calculation time increases exponentially and is defined in minutes.

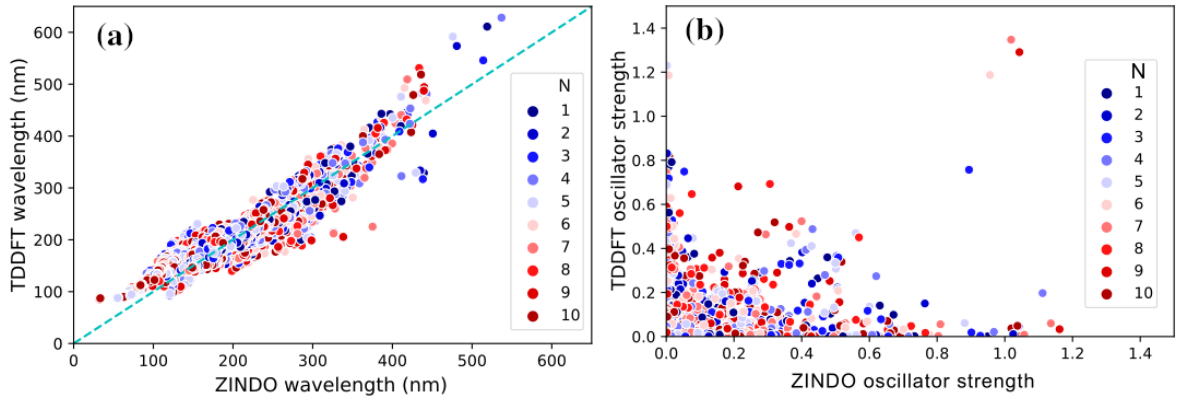


Figure 4. Graphical representation of properties of a training dataset. Each transition is marked with a point according to its oscillator force and wavelength is calculated using ZINDO and TDDFT. All points are colored according to their transition number (N). The graphs illustrate that the wavelength is predicted by ZINDO demonstrates a high correlation with the values predicted by TDDFT. In contrast, the oscillator strength looks very difficult to predict.

of input-target for our future regression algorithm. Each of the 405 input vectors contains information about the transition energy (wavelength) and intensity in terms of the oscillator strength. Fig. 4 shows the correlations between the transitions computed using ZINDO and TDDFT for each transition in the dataset. The noticeable linear dependence between them is confirmed by the large value of the correlation coefficient 0.92. The Fig 4(b) is plotted similarly, but only for the oscillator strength for each transition. There is a much lower 0.11 correlation here. This is usually interpreted as the absence of a linear relationship, but this does not exclude the possibility of finding deeper and more complex relationships between values.

3. Random Forest algorithm

For our case, we choose the simplest and most efficient machine learning algorithm at the same time. Random Forest algorithm [26] is a solid choice for nearly any prediction problem (even non-linear ones). It belongs to a larger class of machine learning algorithms called ensemble methods. Random Forest is a powerful ensemble learning method, as it relies on an ensemble of decision trees [27]. Among the advantages of this method are the following: Computationally simple and straigthforward to fit [28]; handles highly non-linear interactions and classification boundaries; the accuracy of results is competitive with the other machine learning methods. Also it is essential in natural sciences that the random forest algorithms demonstrate the stability of the output results, as they consist of a large number of random estimators [29, 30]. The algorithm was implemented by means of Python3 programming language, in particular by using Numpy [31], Scipy [32], scikit-learn libraries [33], all drawings were drawn using matplotlib library [34]. The transition properties were calculated using the ORCA program package [35].

One can check out the short Random Forest algorithm description in Fig. 5. For our task, we choose the following parameters: The number of trees in the forest is 512, and without limitations for maximal depth. The number of trees was chosen for optimal computational costs and model accuracy. For network training, we divide the original dataset by 80% to 20% for training and testing. Because the output values have a different order of magnitude, we can not apply standard score metrics. To evaluate the results, we calculate excitation in additional

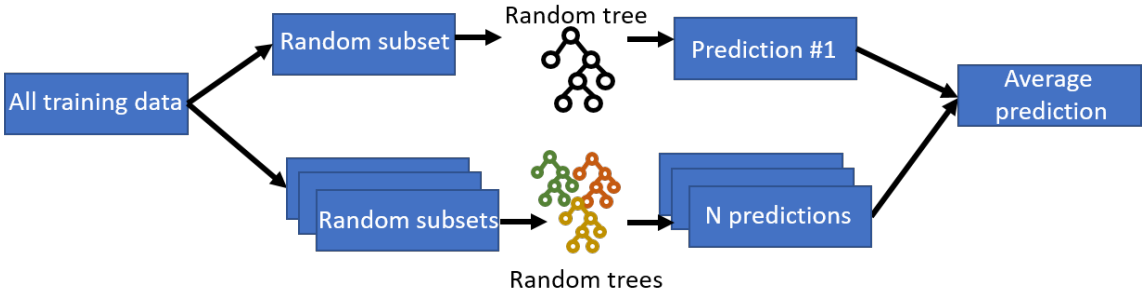


Figure 5. Random Forest Algorithm Description. From the main training set is extracted randomly subsets, on the basis of which the regression trees are built. Each of the trees makes a decision, and this decision is made with a certain weight. Then all the decisions are averaged according to their weight and a prediction is obtained.

8 molecules, which are not in the original training set. They are called from 'A' to 'H' for simplicity, they contain from 8 to 20 atoms. Full names of molecules are deciphered in Table 1.

The Fig. 6 shows how the original ZINDO error (called as initial error) is related to the error that remains after applying the random forest (RF) algorithm. For wavelengths, we consider

Table 1. Eight A-H test molecules description.

A	B	C	D
Aziridine	Arabinose	D-Glucuronic Acid	Erythrose
E	F	G	H
Aspartic acid	Dinitrotoluene	Mequinol	Muscimol

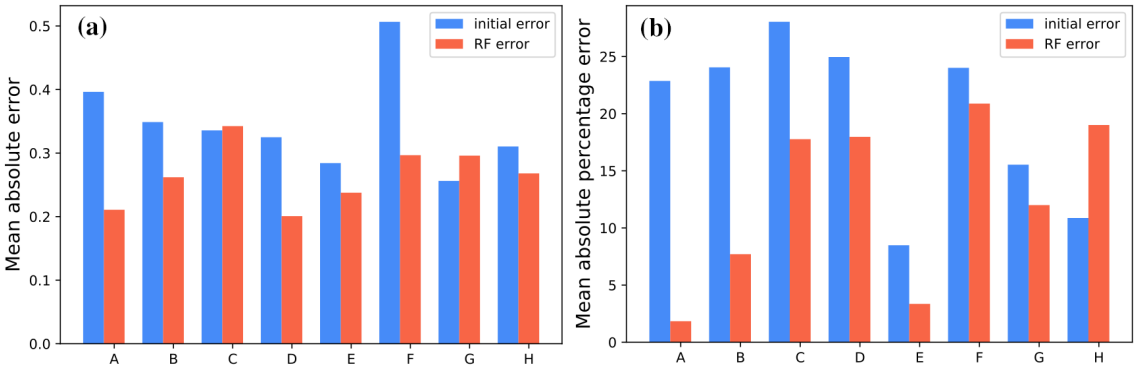


Figure 6. The bar chart represents a relative error of predictions. Due to the peculiarities of the data, the relative percentage error is selected for the wavelength (b) and the absolute error is selected for the oscillator force (a). One can see that for all connections except the latter, the RF error is less than that of ZINDO.

the mean absolute percentage error, and in Fig. 6(b) it mainly decreases as a result of the use of our approach. With the exception of the H molecule, in this case the error increases.

We apply a different metric for the oscillator strength because the percentage error can reach high values for some transitions. So, we just count the mean absolute error. With this metric, we get a decreasing of the error (see Fig 6) although we did not hope for it.

In Fig. 7 the spread shows how the ZINDO results are modified using the algorithm. Fig. 7(b) shows how the wavelength related dots is shifting towards the diagonal, which is the TDDFT value line. As for the oscillator strength, figure 7(a) does not show a better result. However, the correlation between values has increased. The initial correlation coefficient for this test set was -0.03, and then after applying the algorithm, it become 0.36.

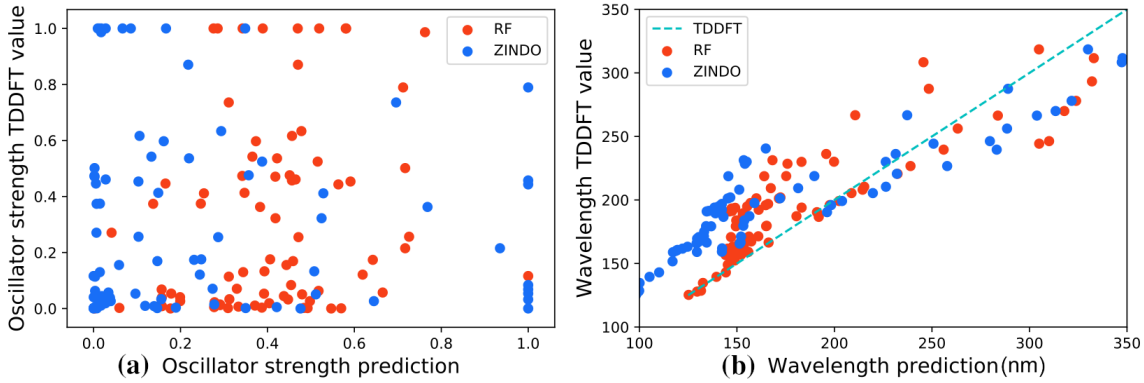
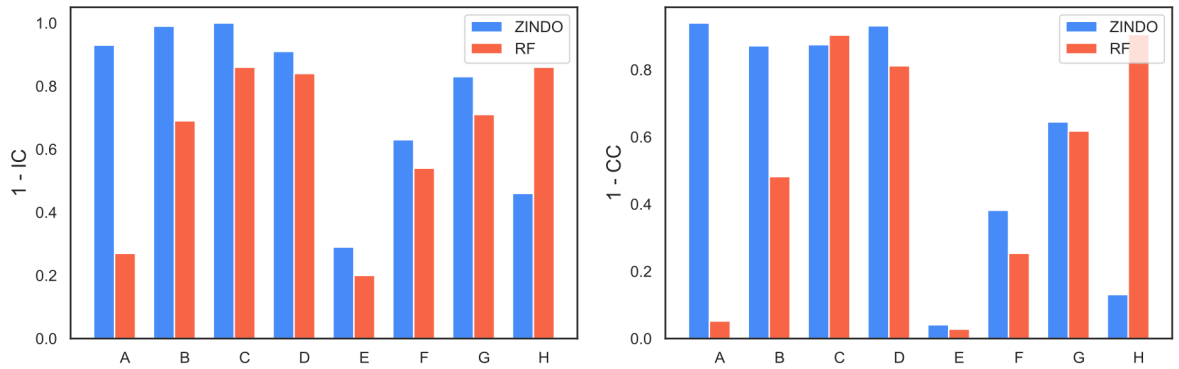


Figure 7. Comparison of wavelength and oscillator strength before and after application of the algorithm to the test molecules. The ZINDO and RF values are arranged horizontally and the corresponding TDDFT values along the vertical axes (similar to Fig. 4). The (b) shows the algorithm is close to the correct one and therefore many points are located near the diagonal.

To evaluate the correctness of the spectra profile, we introduce the IC is intersection coefficient. For this purpose, we build the spectra using TDDFT and the Gauss shape with $\sigma = 0.4$ ev. We do the same with ZINDO and RF

$$\mathrm{IC} = \frac{A_{\mathrm{TDDFT}} \cap A_{\mathrm{ZINDO}}}{A_{\mathrm{TDDFT}} \cup A_{\mathrm{ZINDO}}}, \tag{3}$$

where A_{TDDFT} is an area under TDDFT spectrum curve, and A_{ZINDO} is an area under ZINDO spectrum curve. We inverse this coefficient as $(1 - \mathrm{IC})$ to estimate error. Fig. 8(a) shows this error for different test molecules. At the same time, narrowing or expanding spectra



(a) IC error **(b)** Correlation error

Figure 8. IC (3) error indicator. Small values mean that the spectra graphs coincide. We can see that this graph correlates with Fig. 6(b).

lines always reduce the IC value. To detect such cases, we calculate the Pearson correlation coefficient between the pairs of spectra ZINDO/TDDFT and RF/TDDFT. If the correlation coefficient values are close to 1, this indicates an almost linear dependence between the point coordinates. Since the correlation coefficient (CC) meets the conditions $|\mathrm{CC}| \leqslant 1$, we draw chart $(1 - \mathrm{CC})$ on Fig. 8(b) similarly to Fig. 8(a). We select 8 molecules for the tests. The spectra of only some molecules in charts (Fig. 8) are shown in Fig. 9. The first 4 molecules have a large difference between the ZINDO and TDDFT results (Figs. 9(a),(b) and practically their spectra do not intersect. Therefore, initial IC values for these spectra are close to 0. The other 4 molecules have a non-trivial form of TDDFT (see Fig. 9(c)) or ZINDO (Fig. 9(d)) spectra. For the first two molecules (A and B) algorithm significantly improve the results of ZINDO and shift them closer to the TDDFT spectra (see Fig. 9(a),(b)). For the other two(G and H), as one can see, algorithm do not improve (9(c)) or make the result worse at all (9(d)). The obtained results allow to see the limitation of this algorithm.

4. Discrete spectra

The approach that we used in previous chapter does not separate the intensity and energy values of single transitions. As a result, the corrections to the intensities are not very precise (Fig. 6(a)), and the method relies on a high linear correlation of the energy values. In order to bypass this problem, we propose a method that allows the energy-intensity combination to be replaced by a vector that display the average intensity of the absorption spectrum at ranged energy intervals. The scheme of building a training sample is shown in Fig. 10. For all molecules in the training sample of the previous method, absorption spectra are constructed as follows: For each line in transition with corresponding intensity we construct enveloping Gaussian line with $\sigma = 0.4$, after that we sum all curves and normalize them. To create the input vector, we divide the

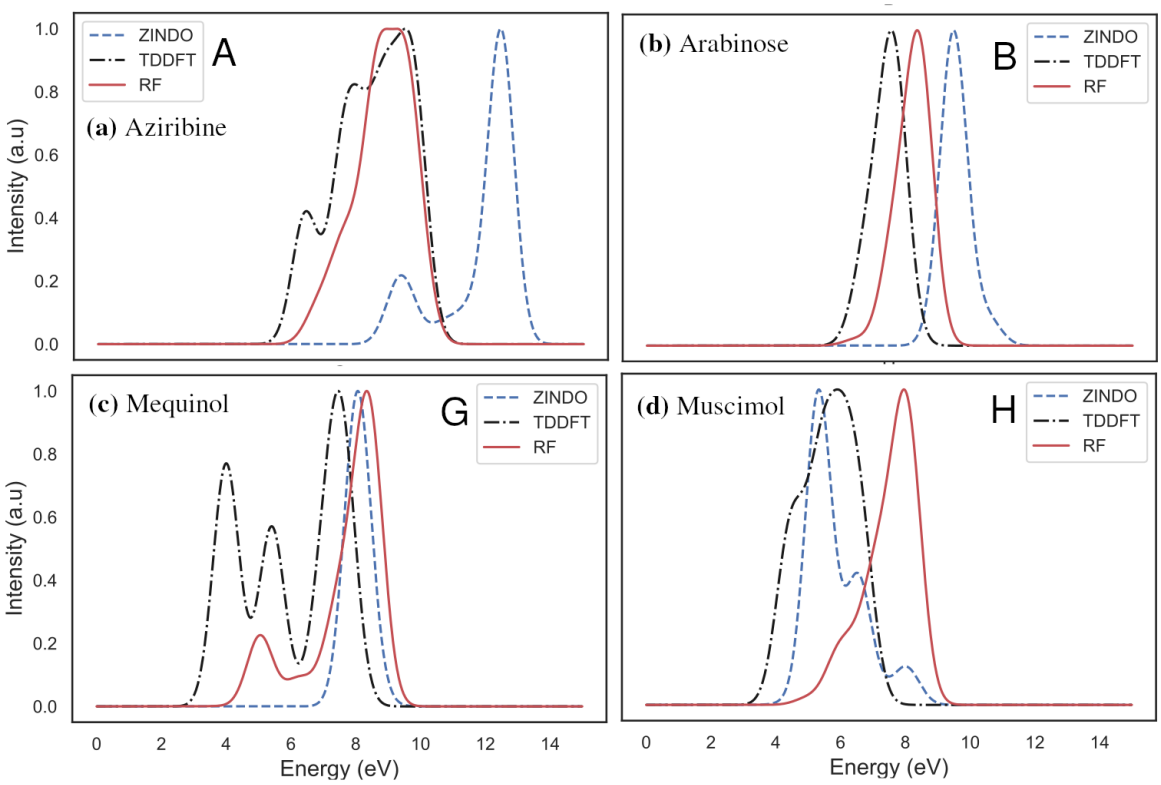


Figure 9. Comparison of absorption spectra of different molecules. There are three spectra on each panel: the calculated using TDDFT (black dash-dotted line), calculated using ZINDO (blue dashed line), calibrated using RF (solid red line). The spectra constructed with $\sigma = 0.4\,\text{eV}$. The values of intensities are normalized to 1. On each graph, there is an alphabetic code by which one can see the value of IC and CC errors in Fig. 8.

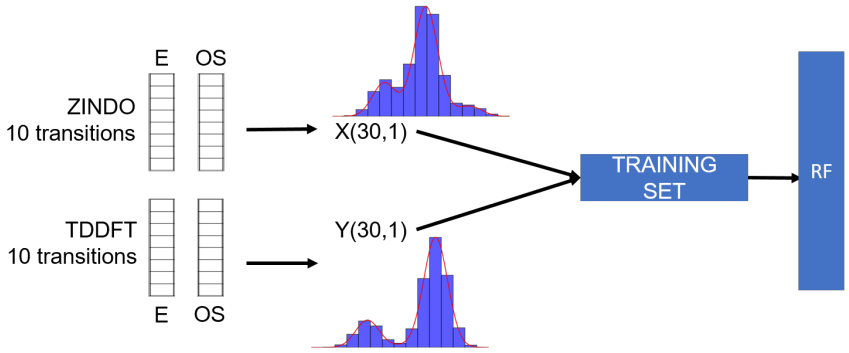
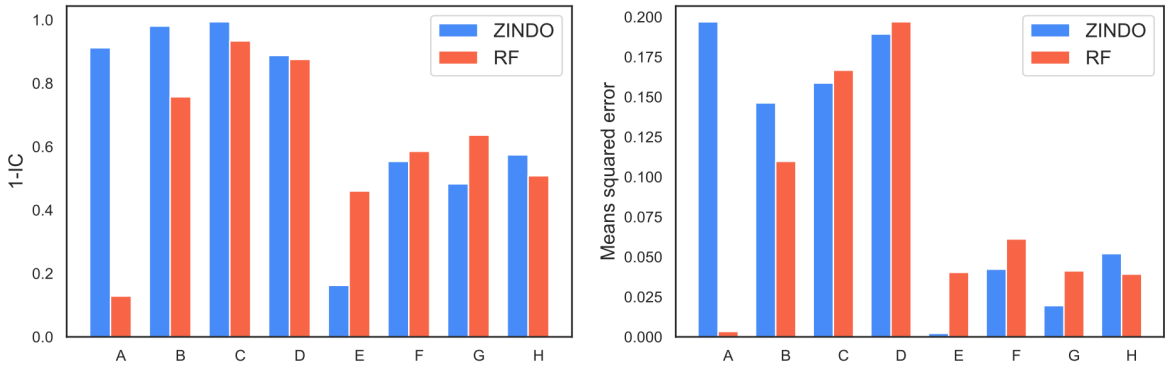


Figure 10. Illustration of the algorithm using a discrete spectra. Energy transitions and intensities are used to build up the envelope of the spectrum, after which the graph is averaged at ranged energy intervals. Each spectrum obtained from ZINDO and TDDFT values becomes the X and Y vectors for the training dataset, respectively.

obtained spectrum into energy intervals, and the calculate average intensity for each interval. We select the energy intervals from 2.8 to 16 eV by the following rule: at the range from 2.8 to 13.6 eV the width is 0.4 eV, from 13.6 to 16 eV the width is 0.8 eV. As a result of this approach, the length of the input vector becomes 30. In this case, along with the already known IC error, one can use to estimations the mean square error (MSE), which is always good for comparing vectors. Figs. 11 and 12 show, that the algorithm does not lead to significant improvements. The exception is the case (shown in Fig. 12(a)), where the large difference between the spectra of aziridine is corrected successfully.



(a) IC error **(b)** Mean squared error

Figure 11. 1 − IC error and MSE. Both assessment methods give approximately the same results. There are only small differences in the case of C and D molecules.

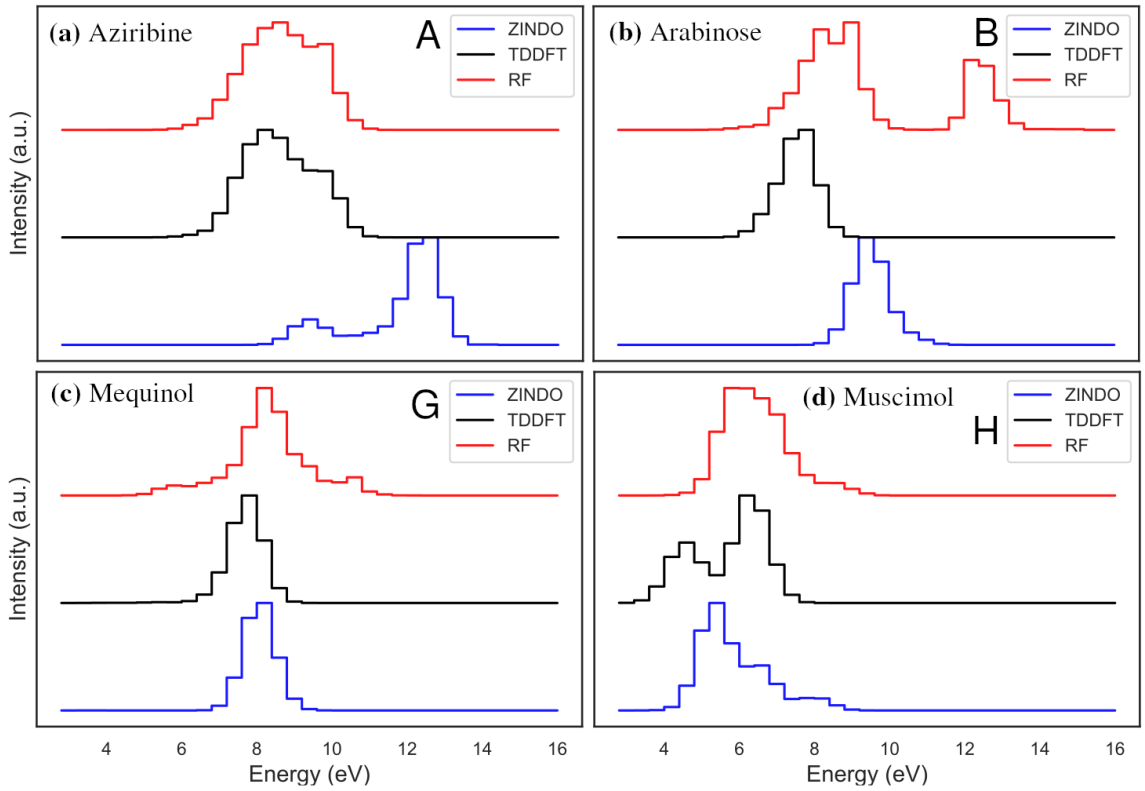


Figure 12. Illustration of the the discrete spectra algorithm results on the example of the same molecules as in the previous case (Fig. 9). Discrete spectra representations are placed one above the other for easy comparison. Colors: blue is ZINDO, black is TDDFT, red is RF.

5. Conclusions

In our paper, we managed to show that a relatively small set of training data based on a various of molecules, with a fairly general selection of parameters, can provide better results in the calculation of optical transitions. We assume that ZINDO has a systematic error when it chooses the molecular orbitals that make up the transition. It is shown that our approach is being able to adjust inaccuracies of ZINDO significantly in some cases, and at least slightly fix in other cases. We suggested two different ways to calibrate the spectra. On our dataset, the transition recalculation method showed a smaller error compared to the discrete spectra-based method.

Acknowledgments

This work was supported by the subsidy of the Ministry of Science and Higher Education of the Russian Federation (3.2166.2017) allocated to Kazan Federal University for performing the project part of the state assignment in the area of scientific activities. Author especially thankful to Dr. Semion Saikin and Prof. Dr. Yurii Proshin for valuable discussions. Author grateful to Prof. Dr. Stephan Gekle and all members of the Biofluid Simulation and Modeling group at Bayreuth University for providing an internship and sharing knowledge.

References

1. Cramer C. J. *Essentials of computational chemistry: theories and models* (John Wiley & Sons, 2013).

2. Lewars E. G. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics* (Kluwer Academic Publishers, 2003).

3. Grigorenko B. L., Mironov V. A., Polyakov I. V., Nemukhin A. V. *Supercomputing Frontiers and Innovations* **5**, 62 (2018).

4. Rodriguez-Galiano V., Sanchez-Castillo M., Chica-Olmo M., Chica-Rivas M. *Ore Geology Reviews* **71**, 804 (2015).

5. Ghasemi J. B., Tavakoli H. *Analytical Methods* **5**, 1863 (2013).

6. Montavon G., Rupp M., Gobre V., Vazquez-Mayagoitia A., Hansen K., Tkatchenko A., Müller K.-R., von Lilienfeld O. A. *New Journal of Physics* **15**, 095003 (2013).

7. Pronobis W., Schütt K. T., Tkatchenko A., Müller K.-R. *The European Physical Journal B* **91**, 178 (2018).

8. Dral P. O., von Lilienfeld O. A., Thiel W. *Journal of Chemical Theory and Computation* **11**, 2120 (2015).

9. Ramakrishnan R., Hartmann M., Tapavicza E., Von Lilienfeld O. A. *The Journal of Chemical Physics* **143**, 084111 (2015).

10. Faber F. A., Hutchison L., Huang B., Gilmer J., Schoenholz S. S., Dahl G. E., Vinyals O., Kearnes S., Riley P. F., von Lilienfeld O. A. *Journal of Chemical Theory and Computation* **13**, 5255 (2017).

11. Ghosh K., Stuke A., Todorović M., Jørgensen P. B., Schmidt M. N., Vehtari A., Rinke P. *Advanced Science* **6**, 1970053 (2019).

12. Becke A. D. *The Journal of Chemical Physics* **98**, 1372 (1993).

13. Zerner M. C. *Reviews in Computational Chemistry* **2**, 313 (1991).

14. Kohn W., Sham L. J. *Physical Review* **140**, A1133 (1965).

15. Runge E., Gross E. K. *Physical Review Letters* **52**, 997 (1984).

16. Leang S. S., Zahariev F., Gordon M. S. *The Journal of Chemical Physics* **136**, 104101 (2012).

17. Silva-Junior M. R., Schreiber M., Sauer S. P., Thiel W. *The Journal of Chemical Physics* **129**, 104103 (2008).

18. Thiel W. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4**, 145 (2014).

19. Khan M. S., Khan Z. H. *Canadian Journal of Analytical Sciences and Spectroscopy* **47**, 146 (2002).

20. Budyka M. F. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **207**, 1 (2019).

21. Silva-Junior M. R., Thiel W. *Journal of Chemical Theory and Computation* **6**, 1546 (2010).

22. Suendo V., Viridi S. *ITB Journal of Science* **44A**, 79 (2012).

23. Wegermann C. A., da Rocha J. C., Drechsel S. M., Nunes F. S. *Dyes and Pigments* **99**, 839 (2013).

24. Saraha A. R., Rakhman K. A., Sugrah N. *Asian Journal of Chemistry* **30**, 1057 (2018).

25. Kim S., Chen J., Cheng T., Gindulyte A., He J., He S., Li Q., Shoemaker B. A., Thiessen P. A., Yu B., Zaslavsky L., Zhang J., Bolton E. E. *Nucleic Acids Research* **47**, D1102 (2018).

26. Breiman L. *Machine Learning* **45**, 5 (2001).

27. Cutler A., Cutler D. R., Stevens J. R. in *Ensemble Machine Learning* (Springer, 2012) pp. 157–175.

28. Zhang W., Wu C. in *International Conference on Information Technology in Geo-Engineering* (Springer, 2019) pp. 243–255.

29. Liaw A., Wiener M. *R News* **2**, 18 (2002).

30. Svetnik V., Liaw A., Tong C., Culberson J. C., Sheridan R. P., Feuston B. P. *Journal of Chemical Information and Computer Sciences* **43**, 1947 (2003).

31. Van Der Walt S., Colbert S. C., Varoquaux G. *Computing in Science & Engineering* **13**, 22 (2011).

32. Virtanen P., Gommers R., Oliphant T. E., et al. *Nature Methods* , 1 (2020).

33. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. *Journal of Machine Learning Research* **12**, 2825 (2011).

34. Hunter J. D. *Computing in Science & Engineering* **9**, 90 (2007).

35. Neese F. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8**, e1327 (2018).

36. Willighagen E. *"SMILES Depicter,"* `https://www.simolecule.com/cdkdepict/depict.html` (2019), [Online; accessed 19-December-2019].

37. Zasso M. *"SMILES Generator,"* `http://www.cheminfo.org/flavor/malaria/Utilities/SMILES_generator___checker/index.html` (2019), [Online; accessed 19-December-2019].

Appendix

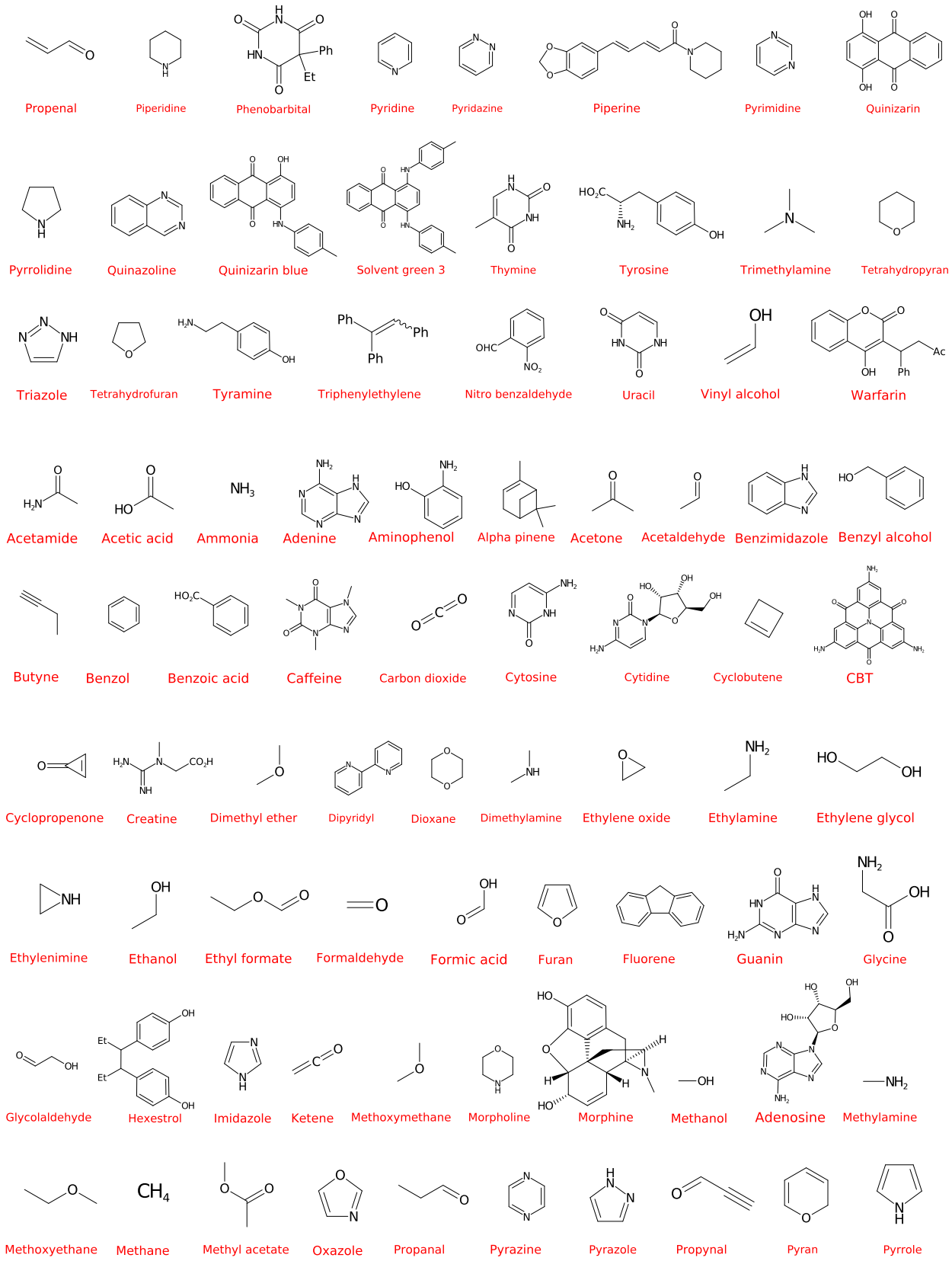


Figure 13. Two - dimensional and perspective drawing of the molecules used in the training dataset. There is a name under each compound. There are 81 compounds. This is drawn using [36,37].