A Journal of the Gesellschaft Deutscher Chemiker

Angewandte International Edition GDCh Chemie

www.angewandte.org

Accepted Article

Title: Prediction of major regio-, site-, and diastereoisomers¬ in Diels-Alder reactions using machine-learning: The importance of physically meaningful descriptors

Authors: Wiktor Beker, Ewa Gajewska, Tomasz Badowski, and Bartosz Grzybowski

To be cited as: *Angew. Chem. Int. Ed.* 10.1002/anie.201806920
Angew. Chem. 10.1002/ange.201806920

Link to VoR: http://dx.doi.org/10.1002/anie.201806920
http://dx.doi.org/10.1002/ange.201806920

WILEY-VCH

Angewandte Chemie International Edition
Angewandte Communications
10.1002/anie.201806920

Artificial Intelligence

DOI: 10.1002/anie.201((will be filled in by the editorial staff))

Prediction of major regio-, site-, and diastereoisomers in Diels-Alder reactions using machine-learning: The importance of physically meaningful descriptors **

*Wiktor Beker†, Ewa P. Gajewska†, Tomasz Badowski, Bartosz A. Grzybowski**

Abstract. *Machine learning can predict the major regio-, site-, and diastereoselective outcomes of Diels-Alder reactions better than standard quantum-mechanical methods and with accuracies exceeding 90% provided that (i) the diene/dienophile substrates are represented by "physical-organic" descriptors reflecting the electronic and steric characteristics of their substituents and (ii) the positions of such substituents relative to the reaction core are encoded ("vectorized") in an informative way.*

Recent years have brought an explosion of interest in machine learning (ML) algorithms, which are being gradually adopted by the chemical community (and chemical industry[1]) to predict biological activities,[2a] solubilities,[2b] or crystal structures[2c] of small molecules, properties of organic photovoltaics,[2d,e] NMR,[2f] mass[2g] and IR[2h] spectra, atomistic potentials,[2i] quantum-chemical parameters,[2j] or optimal reaction conditions.[2k,3g] In the context of synthetic organic chemistry, ML techniques have been explored to predict reaction outcomes or substrates[3a,b] although their applicability is naturally limited to reaction types for which the reaction statistics are abundant enough to enable meaningful learning (thousands of literature examples according to Table 4 in the SI to ref [3i]; see also [3c]). In a recent example of ML application, Segler *et al.* [3d] developed a deep neural network to distinguish literature-reported ("correct") from artificially created ("incorrect") reactions, though this was done without taking into account any reaction conditions which is a drastic oversimplification given a myriad of examples in which the same substrates can give different products depending on conditions used. In our own attempts to predict the yields and times of organic reactions,[3e] ML methods trained on ca. 0.5 million literature-reported reactions using various sets of chemical descriptors were able to correctly categorize these reactions as high-vs-low-yielding or as rapid-vs-slow only in, respectively, ~65% and ~75% of cases. As we pointed out, such rather moderate accuracies reflected the inability of commonly used structural descriptors (be it, molecular descriptors, fingerprints, or chemical-linguistic[3e,f] fragments) to capture the nuances of chemical reactivity. This point of view was substantiated in a recent study by Ahneman *et al.*,[3g] who showed that with descriptors that capture electronic effects, a random-forest classifier performed much better and was able to correctly predict the yields of 4608 C-N cross-coupling reactions to within 7.8% RMSE. Still, such success stories remain sparse and ML methods remain largely untested in the problems synthetic experts might find useful (and non-trivial to predict based on their "human knowledge"), have rarely been extended to reactions involving stereochemistry, and have not been systematically benchmarked[3h] against other existing theoretical tools. These considerations motivated our current work in which we apply ML methods to predict the major outcomes of one of the most powerful organic reactions, the Diels-Alder, DA, cycloaddition[4]. On a diverse set of examples – including transformations used in the syntheses of complex natural products – the random forest, RF, classifiers we construct achieve unprecedented accuracies: 93.6% for the prediction of regioselectivity, 91.3% for site-selectivity, and 89.2% for diastereoselectivity. Additionally, we perform a series of comparative tests that substantiate the following conclusions which we see particularly important given the current interest (and often hype) surrounding artificial intelligence: **(1)** High accuracies are achieved only if the machine is provided some chemical "insight" about the reaction (in particular, information about reaction's core and key substituents); **(2)** The key to maintaining such high accuracies beyond examples similar to those on which the system was trained is not the choice of a specific ML method (e.g., RF vs. neural networks) but rather the use of descriptors/features that capture – as in classic physical-organic chemistry – both electron donating/withdrawing propensities (here, Hammett constants[5a,b]) and steric characteristics (TSEI indices[5c]) of the substituents on the diene and the dienophile. **(3)** Remarkably, with such physically meaningful descriptors, the accuracy of ML can be significantly higher than the values obtained with standard quantum-mechanical, QM, methods (~82%). These considerations lead us to conclude that although ML models cannot possibly rival QM in terms of generality, they can provide an accurate and rapid (~0.5 s vs several hrs in QM) alternative in specific synthetic problems for which numerous literature precedents (like the DA) allow for meaningful model training. With the hope of popularizing this vision, we make our DA predictor freely available to the academic community at http://dielsalderapp.grzybowskigroup.pl/.

Diel-Alder cycloadditions are very effective in building structural complexity but in planning their synthetic use, when the substrates bear multiple substituents, one may need to consider their regio-, site- or diastereoselective outcomes (**Figure 1**). The problem of predicting regioselectivity is an old one and it is generally accepted that relative arrangement of the diene vs. dienophile is dictated by the overlap between their frontier (HOMO/LUMO) orbitals in the transition state. Accordingly, many authors have carried out QM calculations[6] with the aim of predicting regioselectivity from reactivity indices and frontier molecular orbitals; unfortunately, these studies have been limited to very small

[*] Dr. Wiktor Beker, Ms. Ewa P. Gajewska, Dr. Tomasz Badowski, Prof. B.A. Grzybowski
Institute of Organic Chemistry, Polish Academy of Sciences,
ul. Kasprzaka 44/52, 01-224, Warsaw, Poland

Prof. B.A. Grzybowski
Center for Soft and Living Matter and
Department of Chemistry, UNIST,
50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, South Korea
E-mail: nanogrzybowski@gmail.com

[**] Authors gratefully acknowledge generous support from the U.S. DARPA ("Make-It" Award 69461-CH-DRP #W911NF1610384). B.A.G. acknowledges generous personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1.

† Authors contributed equally

‡ WB thanks the Wroclaw Supercomputing Center for CPU time to perform QM calculations.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201xxxxxx.

10.1002/anie.201806920

sets (tens of examples) of structurally related substrates precluding generalization.[6b,6d] Regarding site-selectivity and diastereoselectivity, we are unaware of any theoretical models to predict such outcomes.

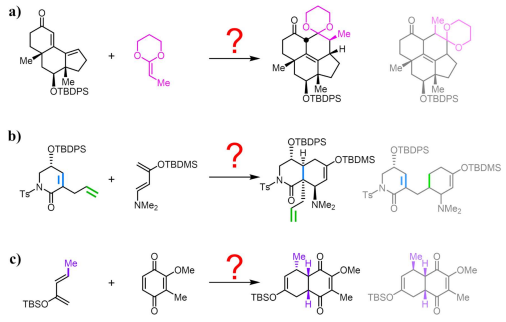


Figure 1. Examples of possible outcomes of the Diels-Alder reactions taken from some "classic" total syntheses. Products that were experimentally obtained are correctly predicted by our RF classifier. The incorrect products are shown in the rightmost column; in the first two examples, none of several possible stereoisomers was formed and so we do not specify the newly formed stereochemistry. **a)** Regioselectivity in the DA reaction used by Snyder *et. al.* in the total synthesis of Rippertenol[7a] is dictated mainly by electronic factors. **b)** Site-selectivity in the DA reaction used by Gagnon and Danishefsky en route to Xestocyclamine A.[7b] **c)** Diastereoselectivity of the DA reaction performed during the preparation of the first key intermediate in Nicolaou's total synthesis of Colombiasin A.[7c]

We studied these effects on a set of unique DA reactions available from the Reaxys repository for which different outcomes were possible (see SI, Sections S1-S3). This set comprised 6,355 reactions and divided into three partly overlapping subsets: 3,080 reactions in which two distinct regioisomers could form, 1,088 with possible site-selectivity, and 2,943 with diastereoselective outcomes. Classifiers were constructed for each set separately to reflect different expected importance of electronic vs. steric effects in each class. If more than one product was reported (in less than 400 cases in the case of the diastereoselectivity problem), only the major one was considered. Furthermore, we did not consider intramolecular DA reactions (additional 1164 examples) since their regiochemical outcomes are enforced by the geometry of the molecule, evading parametrization based on substituent effects alone.

Regioselectivity. We start our discussion with the prediction of major regioisomers by QM methods against which we will subsequently benchmark ML approaches. We followed a well-established way of understanding chemical reactivity in terms of reactivity indices.[8a] First, for all dienes and dienophiles in the reaction dataset, geometry was optimized using B3LYP/6-31G* method and basis set (all calculations were performed with Gaussian09). Next, to derive the so-called Parr functions (which are modern and improved versions of Fukui functions[8]), wave functions for neutral, cationic-radical and anionic-radical species of each diene/dienophile were calculated using B3LYP functional (open-shell for radicals) and 6-31+G** basis set (with diffusive functions on heavy atoms to better describe the third- and fourth-row elements). Parr indices (reflecting nucleophilicity or electrophilicity of each atom) were then derived from Parr functions by means of NBO population analysis. Finally, for each possible regioisomer of the DA reaction, the sum of squares of the differences between Parr indices of reacting atoms was calculated. The major isomer predicted to form was the one for which this sum was smaller. The accuracy of this prediction was 82%, which is close to 82.8% obtained in some previous works using Fukui functions (albeit for a much smaller set of 64 simple DA reactions[6b]). We note that calculations with different basis sets did not improve the accuracy perceptibly (see SI, Section S9).

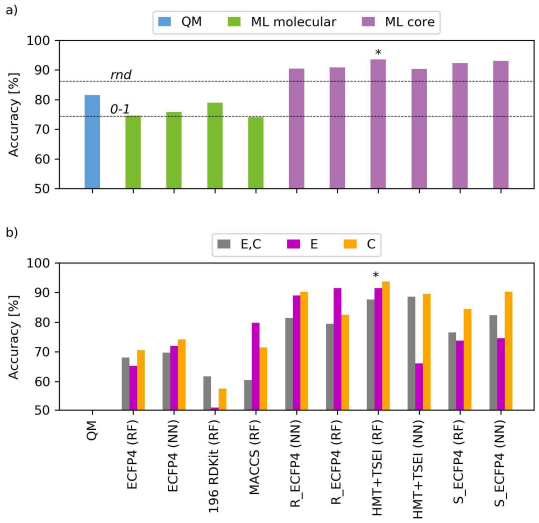


Figure 2. Performance of various ML approaches in predicting major regioisomers. Labels on the horizontal axis at the bottom are common to both graphs and specify the type of descriptors used. Prefix "R_" denotes ECFP4s of the reaction cores rather than substrate/product molecules whereas prefix "S_" means that each substituent on the diene/dienophile was described by ECFP4s. The ML method used is specified in parentheses (NN = 300x2 neural network; RF = 100-tree random forest). **a)** Results obtained for the full set of 3080 regioselective reactions. Blue bar = QM based on Parr functions, Green = NN or RF classifiers based on substrates and products but with no reaction core specified; Light purple = classifiers supplied with the information about reaction products. Dotted horizontal lines correspond to the accuracy of RF with substituents assigned 0-1 values or random numbers (see main text). **b)** Performance of NN and RF methods upon moving one or two classes of dienes with N-containing substituents (E = dienamines, C = dienamides). Legend specifies which subclasses were placed in the test set. For numerical values, see SI, Table S3.

Against these values, we compared the results of different ML approaches in which we provided the machine with various degrees of chemical "insight" by either (1) not specifying the reaction core, or (2) specifying the diene/dienophile reaction centers on input. Unless otherwise stated, all models were trained and tested with the so-called five-fold cross-validation (cf. SI, Section S7.1) and the reported accuracies are average accuracy values computed in such cross-validation (with standard errors < 1% in all cases studied).

For the first class, (1), either neural networks, NNs, or Random Forest, RF (**Figures 2** and **3**) classifiers[9a] were only "shown" the structures of substrates and possible products (i.e., the correct product observed in experiment and the incorrect one that was not observed), each represented/vectorized by various types of features available in RDKit[9b] (ECFP4,[9c] MACCS, or RDKit fingerprints, or RdKit's 196 molecular descriptors; see SI, Section 1.2 for further details). The highest prediction accuracy for a NN (with two 300-neuron layers with RELU activation function, connected to a single sigmoid output neuron, and denoted 300x2 NN; each layer with 0.3 dropout) trained on ECFP4 descriptors was 75.9% and did not improve significantly with the increase in the number of layers. For RF classifiers (100 trees), the accuracy varied between 74.1% for MACCS keys and 84.8% for the RDKit fingerprints. We also tested the performance of Jin's deep NN (DNN) based on Weisfeiler-Lehman architecture trained on 409,035 reactions (in vast majority, *not* Diels-Alder) from patents to recognize the reactivities of different types of atoms or bonds and thus predict outcomes of new reactions. [10a] Here, we wished to probe the much advertised ability of DNNs to perform the so-called transfer learning [10b]_– that is, to train on one set of problems (general patent reactions) but gain knowledge to solve another problem not much "seen" during training (regioselective outcomes of DA). As it turned out, the DNN performed very poorly (0.5% correct DA outcomes within the network's top-five predictions for

10.1002/anie.201806920

each reaction; see Figure S14) We then retrained the DNN on the DA examples – it offered 80.7% accuracy in predicting the major regioisomer being in its top five products, not significantly different from simpler NN architectures discussed above (for further discussion, see SI, Section S8). Such rather unimpressive results – summarized by green bars in **Figure 2a** – are important in so far as they suggest that learning nuances of chemical reactivity without knowing key reaction details or by training on chemically unrelated examples is problematic.

With this hindsight, we focused our attention on various representations that specify the reaction core and quantify the effects of surrounding substituents (light-violet bars in **Figure 2a**). One such representation was based on the commonly used[3a,d] subtraction of ECFP4 fingerprints of the substrates from the fingerprint of the product. When such "reaction fingerprints"[11] were used to train a 300x2 (or other) NNs, the best accuracy achieved was 90.5%. A RF classifier trained on the same representation had 90.9% accuracy. Furthermore, when separate ECFP4 fingerprints were assigned to each substituent on the diene and dienophile, the accuracies increased to 93.1% for 300x2 NN and 92.3% for RF.

A conceptually different representation (cf. SI, Section S4) was based not on fingerprints but on the combination of substituent's Hammett constants (*para* Hammett constants characterizing 306 substituents and taken from previously reported experimental studies[5a,b]) reflecting their propensities to donate/withdraw electrons, and the so-called TSEI indices[5c] capturing substituents' bulkiness. With this "stereoelectronic" representation, the 300x2 NN gave correct answer in 90% of cases whereas RF, in 93.6%. We note that inclusion of both electronic (Hammett) and steric (TSEI) information was important – with the latter omitted, the accuracy dropped to 82% for the NN and 84% for RF.

At this point it might appear that representations based on reaction (or substituent) fingerprints and on Hammett-TSEI descriptors are equally effective. However, the latter – capturing stereoelectronic effects and not only bond-connectivity and atom types – enable accurate predictions about classes of compounds not "seen" during training. To show this, we moved from the training set to the test set reactions involving dienes substituted with certain N-containing groups: dienamines (118 reactions, denoted in **Figure 2b** as E), dienamides (295 reactions, C), and both of these classes (413, reactions, E,C). When the RF or NN classifiers using reaction fingerprints or substituent fingerprints were trained on the curtailed training sets (**Figure 2b** and also SI, Section 7.2), their prediction accuracies decreased perceptibly, whereas the performance of Hammett-TSEI-based RF classifier was much less affected by such repartitioning of the dataset – we verified by the binomial test that the RF classifier using Hammett and TSEI features has higher expected accuracy than most other classifiers (with exception of set E in some representations; see SI, Section 7.5) at the confidence level of 99.9%. This result may indicate that the classifier was able to place the "newly seen" Hammett and TSEI values correctly on the electronegativity and bulkiness scales it learned during training.

To further test whether chemically-meaningful descriptors are indeed important, we considered toy representations in which (i) all non-hydrogen substituents were assigned a value of 1, while hydrogens a value of 0; or (ii) different substituents were assigned different but arbitrary (i.e., chemically meaningless) values. The RF accuracy for the former was, as could be expected, rather low, 74%. Surprisingly, however, simply assigning random numbers to substituents offered 83% accuracy (±4% depending on the set of random numbers) – that is, commensurate with the predictions of the Hammett constants alone (84%) and indicating that on this particular dataset of reactions, the classifier managed to learn some relationships between arbitrarily but still uniquely-labelled substituents. However, when the number of such "random" descriptors was increased (to substitute for both Hammett and TSEI descriptors), the accuracy was not improved (85%) and never matched that of the combined Hammett-TSEI set (93.4%); this was also the case for the predictions of diastereoselective outcomes and, especially, site-selectivity, where the combined Hammett-TSEI set was 10% better than the same number of "random" descriptors (91% vs. 81%; for all comparisons, see SI, Section S6). Overall, arbitrary representations can learn to recognize *some* substituent patterns, but not as well as those based on physically realistic features.

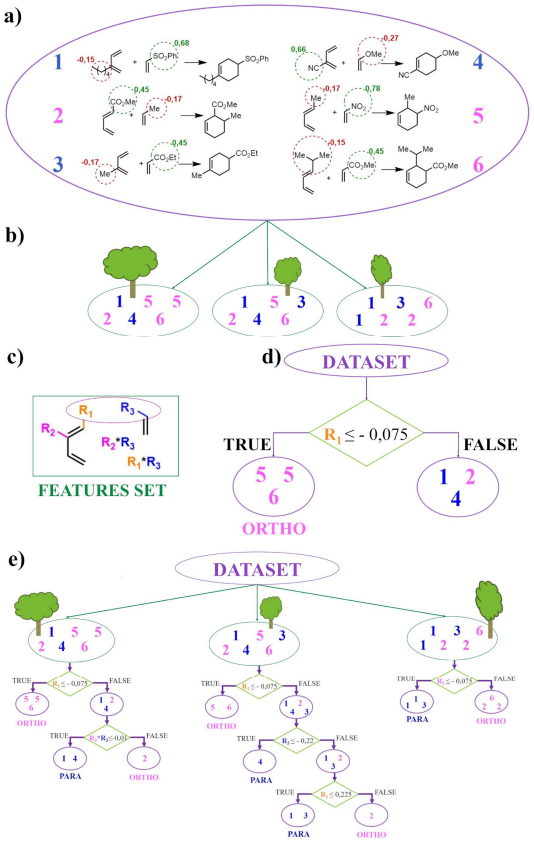


Figure 3. Training of a Random Forest, RF, classifier based on Hammett constants to predict major regioisomers. **a)** A "toy" training set of only six DA reactions in which the reactants can have only one substituent. In the product, these substituents can be oriented "ortho" (pink, reactions #2,5,6) or "para" (blue, reactions #1,3,4) to each other. For a general case of multiple substituents, see SI, Section S4. Numbers placed outside the red and green circles correspond to the Hammett constant of a given substituent; hydrogen substituents are assigned zero values (not shown). **b)** Samples (here, just three, in real analyses, 100) of the same numbers of reactions as in the training set but allowing for repetitions are drawn from the training set at random. For every sample, one decision tree is created independently. **c)** Each pair of substrates is described by a set of features (here, Hammett for substituents in positions R_1, R_2 and R_3, as well as their combinations such as R_1*R_3 and R_2*R_3). **d)** The algorithm randomly selects a subset of features and threshold values are optimized to split the dataset into subsets/"leafs" as cleanly as possible (see [9a]). Here, the $R_1 \leq -0.075$ feature/threshold are desirable because one of the subsets (5,5,6) is "pure" – that is, it consists of only reactions giving the "ortho" product. **e)** Feature selection and leaf splitting are repeated for each tree until all leafs are pure. At this point, the classifier is trained and when it encounters a new DA reaction, each tree makes its own prediction and the majority vote of all trees is ultimately taken (SI, Section S5).

Site- and diastereoselectivity. We approached these two sub-problems using Hammett/TSEI RF classifiers which, as we have seen, have proven the most accurate and robust in the comparative regioselectivity tests. Prediction of the major reaction site was a straightforward extension since the same vectors of Hammett/TSEI features as before were assigned to each diene/dienophile site. The difference was that all possible products were created (e.g., four if

10.1002/anie.201806920

two dienes and two dienophiles were present) and the RF was trained to choose from them, by a series of pairwise comparisons, the one observed experimentally. The classifier thus trained was 91.3% accurate. We note that the accuracy was not markedly decreased if only Hammett constants were used but was significantly lower with only TSEI features used for training. Interestingly, this theoretically-predicted importance of electronic effects resonates with a common experimental approach to designing a site-selective DA reactions by adjusting the electronic properties of the desired diene and-dienophile sites[4a,12].

The situation was slightly more complicated for the prediction of major diastereoisomers (i.e., the choice between two possible stereoisomeric classes, **Figure 4a**). In this case, we capitalized on the well-defined mutual orientation of the reaction substrates and encoded/vectorized it by the order of features describing the diene and the dienophile parts in the forming ring (**Figures 4b,c**). With such vectorization and the Hammett/TSEI values assigned to the substituents, the RF classifier was 89.2% correct (87.2% for Hammett features alone and 87.5% for only TSEI). We note that although the conventional ECFP4 descriptors can be also be supplemented with some degree of stereochemical information (R/S stereochemical flags, see documentation of RDKit Morgan fingerprints in ref [9b]), their performance in the current problem was visibly worse, with accuracy below 80% irrespective of whether such descriptors were used to vectorize entire products/substrates, or whether they were used as reaction fingerprints.

The last result reiterates the main message of this paper – namely, that ML can be useful in predicting outcomes of non-trivial organic reactions and can generalize to "unseen" classes of substrates when (i) descriptors carrying physically relevant information are applied, and (ii) when the machine is provided with appropriately vectorized information about the reaction core and important substituents. With these two conditions met, the methodology we described should be extendable to other reaction classes for which sufficient numbers of training examples are available.

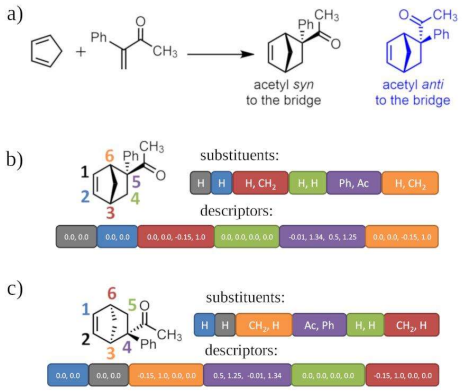


Figure 4. Prediction of major diastereoisomers and vectorization of the substituents at the reaction core. **a)** Example of a DA reaction with possible *syn* and *anti* products, of which the latter (colored blue) is obtained experimentally. **b)** Substituent-based, stereochemically-aware vectorization of product *syn*. Following the counterclockwise numbering of the atoms on the product's cyclohexene ring, substituents are sequentially added into the vector. Within each block, the substituent pointing below the plane of the ring is added first. Each substituent is then assigned specific values of the Hammett and TSEI descriptors. **c)** To avoid ambiguity of the direction in which substituents are traversed, the ring is rotated 180° around the axis bisecting the 1-2 double bond. After substituents are numbered and the descriptor values assigned, the newly created vector is concatenated with the one from (b) to give the ultimate vector representation of the reaction's diastereoselectivity. Desired classifiers are then trained on this "composite" representation (see SI, Section S4).

Received:
Published online on:

Keywords: Diels-Alder reaction, selectivity, machine learning, random forest, neural networks

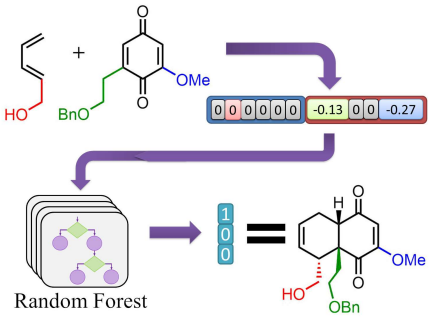
[1] R. Mullin, *Chem. Eng. News* **2017**, *95*, 26-30.
[2] a) G. S. Heck, V. O. Pintro, R. R. Pereira, M. B. de Avila, N. M .B. Levin, W. F. de Azevedo, *Curr. Med. Chem.* **2017**, *24*, 2459-2470; b) A. Lusci, G. Pollastri, P. Baldi, *J. Chem. Inf. Model.* **2013**, *53*, 1563-1575; c) A. van de Walle, *Nat. Mater.* **2005**, *4*, 362-363; d) E. O. Pyzer-Knapp, K. Li, A. Aspuru-Guzik, *Adv. Funct. Mater.* **2015**, *25*, 6495-6502; e) P. Raccuglia, K. C. Elbert, P. D. F. Alder, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature*, **2016**, *533*, 73-76; f) J. Cuny, Y. Xie, C. J. Pickard, A .A. Hassanali, *J. Chem. Theory Comput.* **2016**, *12*, 765-773; g) B. Curry, D. E. Rumelhart, *Tetrahedron Comput. Methodol.* **1990**, *3*, 213-237; h) M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci*. **2017**, *8*, 6924-6935; i) J. Behler, *Angew. Chem. Int. Ed.* **2017**, *56*, 12828-12840; j) K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, #13890; k) Z. Zhou, X. Li, R. N. Zare, *ACS Cent. Sci.* **2017**, *3*, 1337-1344.
[3] a) J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, *2*, 725-732; b) B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 1103-1113; c) S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937; d) M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604-610; e) G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, #3582; f) A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, B. A. Grzybowski, *Ang. Chem Int. Ed.* **2014**, *53,* 8108-8112; g) D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186-190; h) B. Maryasin, P. Marquetand, N. Maulide, *Angew. Chem. Int. Ed.* https://doi.org/10.1002/anie.201803562 (2018); i) Segler, M. H. S.; Waller, M. P. *Chem. Eur. J.* **2017**, *23*, 5966-5971.
[4] a) K. C. Nicolaou, S. A. Snyder, T. Montagnon, G. Vassilikogiannakis, *Angew. Chem. Int. Ed.* **2002**, *41*, 1668-1698; b) E. J. Corey, *Angew. Chem. Int. Ed.* **2002**, *41*, 1650-1667; c) R. B. Woodward, T. J. Katz, *Tetrahedron* **1959**, *5*, 70-89.
[5] a) C. Hansch, A. Leo, R. W. Taft, *Chem. Rev.* **1991**, *91*, 165–195; b) O. Exner, in *Correlation analysis in Chemistry*, Plenum Press, New York and London, 1978, p. 455-481; c) C. Chenzhong, L. Liu, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 678-687.
[6] a) S. Damoun, G. Van de Woude, F. Méndez, P. Geerlings, *J. Phys. Chem. A* **1997**, *101*, 886-893; b) G. Gayatri, G. N. Sastry, *J. Chem. Sci.* **2005**, *117*, 573-582; c) L. R. Domingo, P. Pérez, J. A. Sáez, *RSC Adv.* **2013**, *3*,1486-1494, d) H. Hirao, T. Ohwada, *J. Phys. Chem. A* **2005**, *109*, 816-824.
[7] a) S. A. Snyder, D. A. Wespe, J. M. von Hof, *J. Am. Chem. Soc.* **2011**, *133*, 8850-8853; b) A. Gagnon, S. J. Danishefsky, *Angew. Chem. Int. Ed.* **2002**, *41*, 1581-1584; c) K. C. Nicolaou, G. Vassilikogiannakis, W. Mägerlein, R. Kranich, *Angew. Chem. Int. Ed.* **2001**, *40*, 2482-2486.
[8] a) L. R. Domingo, M. Ríos-Gutiérrez, P. Pérez, *Molecules* **2016**, *21*, #748.
[9] a) T. Hastie, R. Tibshirani, J. Friedman, in *The elements of statistical learning: data mining, inference and prediction*, Springer, New York, **2009**, pp. 587-603; b) www.rdkit.org c) D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742-754.
[10] a) W. Jin, C. W. Coley, R. Barzilay, T. Jaakkola, *Predicting organic reaction outcomes with Weisfeiler-Lehman Network.* In: Advances in Neural Information Processing Systems, 2604-2613 (**2017**); b) I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, **2016**.
[11] N. Schneider, D. M. Lowe, R. A. Sayle, G. A. Landrum *J. Chem. Inf. Model.* **2015**, *55*, 39-53.
[12] R. B. Woodward, F. Sondheimer, D. Taub, K. Heusler, W. M. McLamore, *J. Am. Chem. Soc.* **1952**, *74*, 4223-4251.

10.1002/anie.201806920

Artificial Intelligence

Wiktor Beker†, Ewa. P. Gajewska†, Tomasz Badowski and Bartosz A. Grzybowski*

Prediction of major regio-, site-, and diastereoisomers in Diels-Alder reactions using machine-learning: The importance of physically meaningful descriptors.



TOC TEXT: Machine learning provides accurate predictions of the major-isomer outcomes of Dies-Alder reactions provided that the diene and dienophile are represented by vectors describing electronic and steric characteristics of their substituents.