ChemPubSoc Europe

DOI: 10.1002/chem.201800345

CHEMISTRY
A European Journal
Full Paper

Machine Learning

Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning

Sunghwan Choi,[a, b] Yeonjoon Kim,[a] Jin Woo Kim,[a] Zeehyo Kim,[a] and Woo Youn Kim*[a, c]

Abstract: Machine learning based on big data has emerged as a powerful solution in various chemical problems. We investigated the feasibility of machine learning models for the prediction of activation energies of gas-phase reactions. Six different models with three different types, including the artificial neural network, the support vector regression, and the tree boosting methods, were tested. We used the structural and thermodynamic properties of molecules and their differences as input features without resorting to specific reaction types so as to maintain the most general input form for broad applicability. The tree boosting method showed the best performance among others in terms of the coefficient of determination, mean absolute error, and root mean square error, the values of which were 0.89, 1.95, and 4.49 kcal mol^{-1}, respectively. Computation time for the prediction of activation energies for 2541 test reactions was about one second on a single computing node without using accelerators.

Introduction

The remarkable advance of machine learning (ML) techniques provides a new fascinating strategy to deal with various chemical problems; instead of solving physical models directly, ML predicts solutions by using patterns learned from big data.[1] The main advantage of ML is its tremendous computation speed compared with solving the physical models. Therefore, they are actively being used in various chemical fields such as quantum chemistry,[2–6] virtual screening of molecular materials,[7,8] and synthetic pathways for target chemicals.[9,10]

Activation energy is a key factor in determining the reaction rate, mechanism, and products of chemical reactions.[10,11] Quantum chemical methods are able to evaluate the activation energy by identifying the corresponding transition state along a given reaction path. Despite the high reliability of such methods, heavy computational costs are inevitable to deal with complex chemical processes involving a number of reactions.[12] An acceleration technique has been devised to analyze complex chemical reaction networks, but it entails a loss of accuracy.[13,14] Group additivity and rule-based models can be used to predict kinetic parameters for specific types of reactions, but their applicability is limited.[15–20] In this context, it is valuable to know whether ML can be a solution for the accurate prediction of activation energies with broad applicability or not.

For high accuracy of ML, the following two aspects are important. First, appropriate input features that have a high correlation with the target properties should be developed. For intrinsic molecular properties such as total energies, band gaps, and so on, various structural features have been proposed.[3,5,8,21] However, activation energies are determined by reaction paths passing through transition states. Therefore, information on the reaction paths is essential for accurate prediction. Unfortunately, however, such data is hardly available. Moreover, no suitable descriptor for the reaction path has been reported yet.[22] Second, a large amount of data with high quality should be available. Fortunately, kinetic parameters for many gas-phase reactions are publicly accessible, and they are free from solvent effects, which may hinder the accurate prediction of activation energies. Moreover, the reliable prediction of activation energies of gas-phase reactions is important for industrial applications such as combustions and jet engine reactions.[11]

In this regard, we investigated the performance of various ML models for activation energy prediction. Particularly, we focused on the possibility of using only the molecular properties of reactants and products as input features without information on reaction paths. We considered the artificial neural network (ANN),[23] the support vector regression (SVR),[24] and the tree boosting (TB)[25] methods. The ANN method as a prototype of deep learning often requires a large amount of data for sufficient training while avoiding overfitting as a result of

[a] Dr. S. Choi, Dr. Y. Kim, J. W. Kim, Z. Kim, Prof. W. Y. Kim
Department of Chemistry, KAIST
291, Daehak-Ro, Yuseong-gu, Daejeon, 34141 (Republic of Korea)
E-mail: wooyoun@kaist.ac.kr

[b] Dr. S. Choi
National Institute of Supercomputing and Network
Korea Institute of Science and Technology Information
245 Daehak-Ro, Yuseong-gu, Daejeon, 34141 (Republic of Korea)

[c] Prof. W. Y. Kim
KI for Artificial Intelligence, KAIST
291 Daehak-ro, Yuseong-gu, Daejeon (Republic of Korea)

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/chem.201800345.

Part of a Special Issue to commemorate young and emerging scientists. To view the complete issue, visit Issue 47.

many parameters. The SVR and TB methods are classified as shallow learning. They have less possibility of overfitting owing to a restricted search area. We used gas-phase reactions for training and testing the ML models.

Computational Method

The input features of elementary reactions were evaluated as follows [Eq. (1)]:

$$f^i = \sum_p f^i_p - \sum_r f^i_r \quad (1)$$

in which p and r denote products and reactants, respectively, and f^i, f^i_p, and f^i_r are the *i*-th feature of a given reaction, the *i*-th molecular feature of the p-th product molecule, and the *i*-th molecular feature of the r-th reactant molecule, respectively. The difference expression explicitly indicates which molecular features are changed by the reaction. Those features consist of thermodynamic quantities, topological indices, and the Morgan molecular fingerprints[26,27] as shown in Figure 1. The thermodynamic quantities represent the change in reaction enthalpy and entropy. The topological indices are associated with global structural changes denoted by atom connectivity, whereas the Morgan fingerprints represent local substructural changes within a given radius. It should be noted that these input features are intended to maintain the most general input form without resorting to specific reaction types, so the ML models can be applicable to a wide range of chemical reactions. More details of the input features and training procedures are given in the Supporting Information.

Reaction data was obtained from the RMG-Py database.[19] It contains the kinetic information measured in both gas and solution phases, but we considered only the 12704 gas-phase reactions. We noted that the RMG-Py database contains different barrier heights from multiple sources for the same reactions. The number of unique reactions thus became 6078. As it was not clear which values were more reliable, we considered all reference values independently. We randomly chose 20% out of them as the test set (2541 reactions), whereas the rest was used for training.

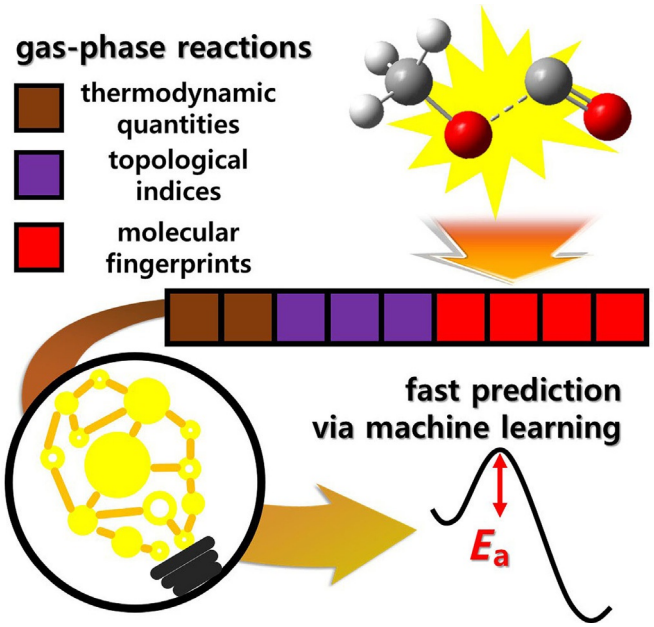


Figure 1. Schematic illustration of the prediction of activation energies by using machine learning.

Results and Discussion

Figure 2 shows the accuracy of six different ML models: SVR, TB, and ANN*n* where *n* denotes the number of hidden layers. We refer to the Supporting Information for technical details of the models. The relative performance of each model was evaluated by the following three factors: coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). Overall accuracy of the ML models in terms of MAE was comparable to the reported error range of density functional theory calculations denoted by the shaded region in Figure 2b, which is about 1.5–7.5 kcal mol^{-1} (Table 1 in Ref. [28]). This accuracy is remarkable because we did not use any information on the reaction paths. The TB model was the best for all three factors. The SVR model showed the worst performance, and the ANN models were between the TB and the SVR. The accuracy of the ANN models seemed to be saturated as the number of hidden layers increased. This result may indicate that overfitting was well prevented by the dropout and early stopping techniques. Nonetheless, the performance of the ANN models was lower than expected. We suspect that it is because the limited number of data caused insufficient training of the ANN models.

To improve the accuracy of the ML prediction, the extraction of optimal features is necessary. The dimensions of the thermodynamic quantities and the topological indices are already small enough, but that of the fingerprints seems to be large. As a result, most fingerprints of reactants and products for all reactions in the data set were identical, and hence Equation (1) gave a number of zero values. Therefore, we compressed the fingerprints to find an optimal size while minimizing information loss by a singular value decomposition method.

Woo Youn Kim completed his undergraduate degree in chemistry at POSTECH in 2004 and a Ph.D. in quantum chemistry under the guidance of Prof. Kwang S. Kim at POSTECH in 2009. He was a postdoctoral fellow in the group of Prof. E. K. U. Gross at Max Planck Institute of microstructure physics. Currently, he is an Associate Professor in the Department of Chemistry at KAIST since 2011. His research interests include the development of quantum chemical methods, machine learning techniques, and automated reaction path finders.

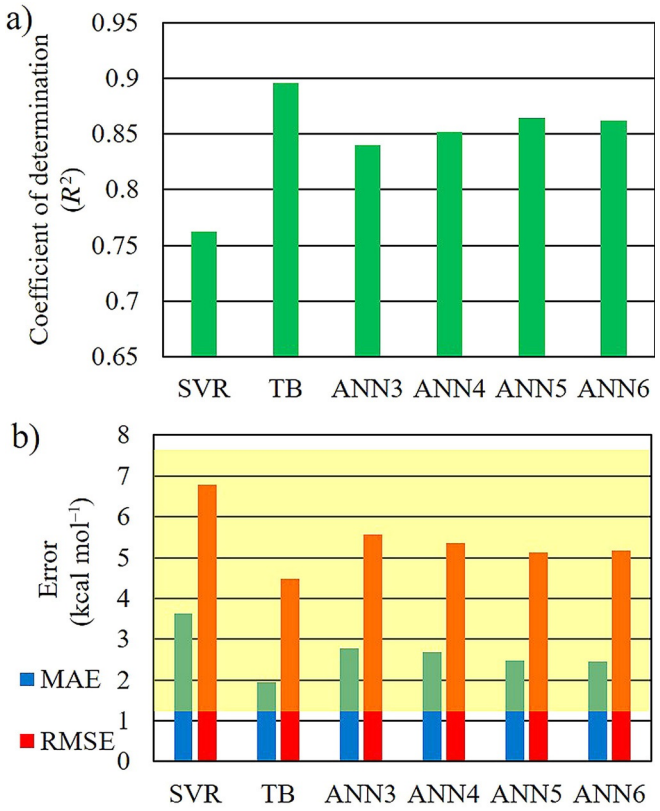


Figure 2. Performance of SVR, TB, and ANN with three different numbers of hidden layers. (a) Coefficient of determination (R^2), and (b) mean absolute error (MAE) and the root mean square error (RMSE). The shaded region in (b) denotes the MAE range of density functional theory calculations.[28]

Figure 3 shows the accuracy of the TB method in the prediction of the activation energies for the test set as a function of the feature size. Indeed, it had an optimal value showing the best performance. Specifically, R^2 and RMSE became maximum (0.89) and minimum (4.49 kcal mol^{-1}) at the dimension of 75, respectively. In this case, the corresponding MAE was 1.95 kcal mol^{-1}. Similar trends were observed in the other ML models but with different optimal sizes (Table S1 in the Supporting Information). In the TB model (and also ANN4 in Table S1 in the Supporting Information), the MAE was decreased, whereas the RMSE increased, as the feature size decreased. Statistically, inconsistent errors lead to larger increases in the RMSE than in the MAE. This means that outliers by the inconsistent errors would appear more often at smaller dimensions. In this study, we aim to examine the feasibility of the ML models rather than to improve their accuracy. For further analysis, hence, we chose 75 as the optimal size in the TB model without further improvement. The same tests were performed for the other models. Figure 2 shows their best performance with their own optimal feature sizes.

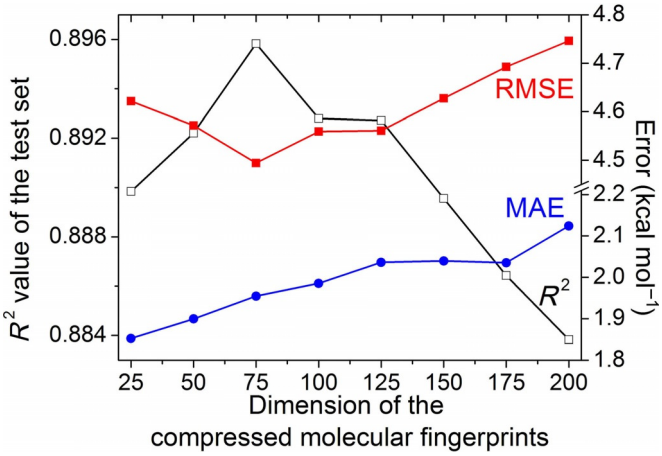


Figure 3. Performance of the tree boosting as a function of the dimension of molecular fingerprints.

Figure 4 shows the correlation between the true versus predicted activation energies for the test set. The gray shade indicates the area with errors less than 3 kcal mol^{-1}, and the blue lines indicate the boundaries of ±5 kcal mol^{-1} errors. The percent ratio of the points in the gray area and that of the points between the blue lines were 83.8% and 90.4%, respectively.

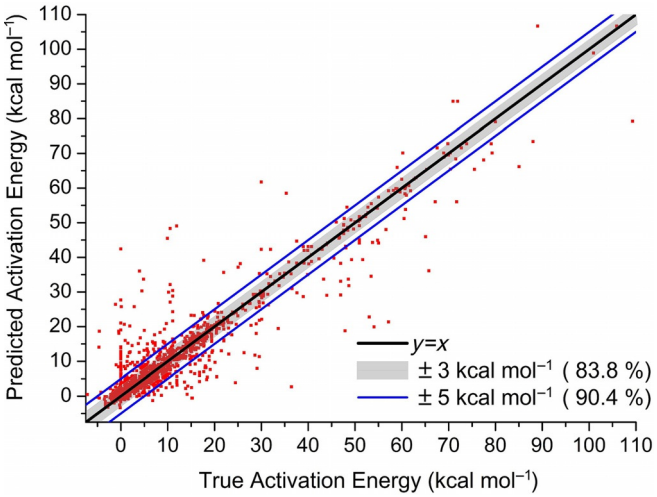


Figure 4. Correlation plot of the true versus predicted activation energies. The gray area and the area between the two blue lines indicate deviations within 3 and 5 kcal mol^{-1}, respectively. The percent ratio of the data points in each area are about 83.8% and 90.4%, respectively.

To elucidate the origin of the large errors of the remaining 9.6% molecules, we classified all reactions into two groups: the L group for molecules with errors larger than 5 kcal mol^{-1} and the S group for the rest. Table 1 shows five different factors of reactions in each of the L and S groups. Both groups showed no difference in the change of the number of radical electrons by the reactions, indicating that the electronic effect is not the main origin of the large errors in the L group. However, we found noticeable structural differences between reactions of the L and S groups. The L group has larger molecular weights and more heavy atoms (i.e., non-hydrogen atoms) than those of the S group. In particular, the L group has more product molecules than that of the S group, as indicated by the large ratio (~18%) of reactions having more than two product molecules in the L group. The other ML models showed similar trends (Table S2 in the Supporting Information),

Table 1. Average values of five different factors of reactions in the large-error (L) and the small-error (S) groups.

Parameter	L group[a]	S group[b]
Number of radical electrons changed by reaction	0.52	0.52
Sum of molecular weights of reactants or products	94.86	78.41
Number of heavy atoms participating in reactions	6.94	5.52
Percent ratio of reactions having more than two product molecules [%]	18.1	6.9
Percent ratio of reactions having more than two reactant molecules [%]	0.4	0.2

[a] Reactions with errors less than 5 kcal mol^{-1}. [b] The rest.

implying that the large errors may be caused by insufficient input features for those reactions. A molecule with more atoms is likely to have more conformers owing to the increased structural complexity. Various conformers may cause the possibility of diverse reaction paths. Also, more product molecules may provoke more complicated reaction mechanisms.

To elaborate the above analysis, we examined the trend of MAE of the predicted activation energies with respect to the number of changed bonds by reaction in the test set. The number of changed bonds was evaluated as the sum of the numbers of bond formations and bond dissociations. Multiple bond orders have been counted as one. Figure 5 shows the result. About 96% of reactions involved numbers of changed bonds less than or equal to four as indicated by the red bar. The training set has almost the same distribution (Table S4 in the Supporting Information). Therefore, the ML models may be biased by those reactions. Indeed, the MAE increases as the number of changed bonds increases as shown by the green line. Moreover, reactions having more than four bond changes likely cause complicated reaction mechanisms, which supports

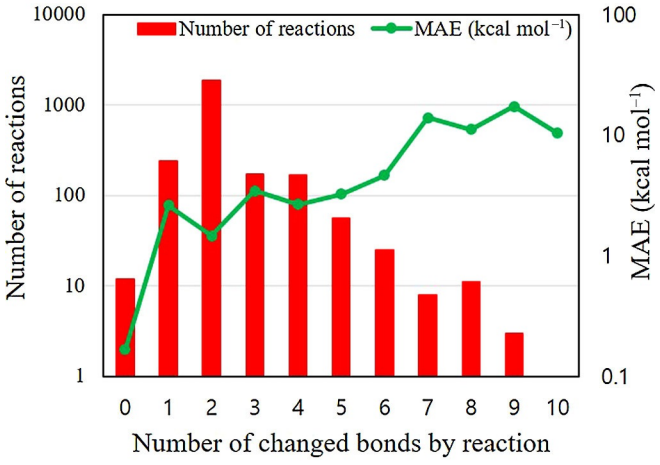
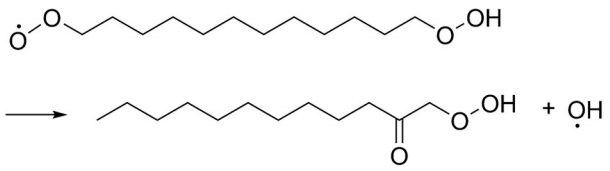


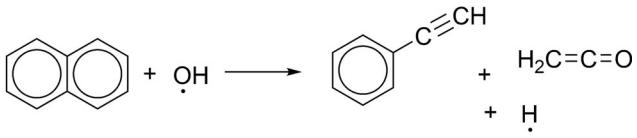
Figure 5. MAE of the predicted activation energies with respect to the change in the number of chemical bonds by reaction for the test set (see Table S3 in the Supporting Information for more details). The red bar and the green line indicate the number of reactions sharing the same number of changed bonds in the test set and the corresponding MAE values, respectively.

the argument in the previous paragraph. For instance, the three reactions with the largest errors are shown in Figure 6. Case 1 and case 2 involve seven and eight bond changes, respectively. Case 3 has two bond formations and two bond dis-

Case 1: 4 broken, 3 formed,
0.0 kcal mol^{-1} (true) , 42.4 kcal mol^{-1} (predicted)



Case 2: 5 broken, 3 formed,
10.5 kcal mol^{-1} (true) , 48.3 kcal mol^{-1} (predicted)



Case 3: 2 broken, 2 formed,
11.9 kcal mol^{-1} (true) , 49.0 kcal mol^{-1} (predicted)

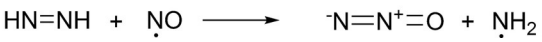


Figure 6. The three reactions with the largest errors in the prediction of activation energies for the test set.

occiations, but it must break the double bond of HN=HN and form a new double bond between N and NO at the same time. All the three cases seem to be non-elementary reactions or multiple reaction paths. In general, as the number of reaction steps increases, our core assumption, activation energy can be inferred by only difference between reactants and products, becomes uncertain. To improve the prediction accuracy, those cases may require direct information on reaction paths in the input or more reaction data with similar patterns.

We further analyzed the MAE of the test set according to reaction types as shown in Table 2. Most reactions belonged to substitution (A–B + C → A + B–C type), whereas 342 reactions in the test set were unclassified. A similar data distribution was shown in the training set (Table S5 in the Supporting Information). In particular, addition reactions (reactions with a single product from multiple reactants) showed the largest MAE. Interestingly, they also involved more than four bond changes, which is consistent with the analysis on the error source.

To complete the rate equation for kinetic study, we also tried to learn the pre-exponential factors in the RMG-Py database but failed owing to the following reason (not shown here). We note that to obtain the prefactors, some reactions used the ordinary Arrhenius equation ($Ae^{-Ea/RT}$), whereas others adopted the modified Arrhenius equation ($AT^n e^{-Ea/RT}$). Such an inconsistency probably provoked a large variation of those values in the range of [10^{-3}, 10^{20}]. Therefore, it was difficult to achieve good accuracy with our ML models. Using the logarithmic values of the prefactors did not improve the result.

Table 2. MAE of the predicted activation energies with respect to reaction types for the test set.

Reaction type	Number of reactions	Average number of changed bonds by reaction	MAE [kcal mol^{-1}]
Substitution[a]	1839	2.04	1.41
Isomerization[b]	12	0	0.17
Reductive coupling[c]	1	3	0.13
Oxidative coupling[c]	5	3	1.07
Association[d]	167	1.02	1.69
Dissociation[d]	86	1.1	5.50
Addition[e]	7	4.29	10.44
Elimination[e]	82	2.87	5.70
Others	342	4.18	3.12

[a] A–B + C↔A + B–C type. [b] Reactions where the reactants and products are identical. [c] A + B–C↔B–A–C type. [d] Reactions only with bond dissociations or bond formations. [e] Reactions with a single product from multiple reactants or multiple products from a single reactant.

In addition to the reasonable accuracy, the main advantage of the ML comes from its unrivaled computational speed. Activation energy prediction of 2541 reactions in the test set took just one second on a single computing node without using accelerators, which is a contrast to the fact that conventional methods take at least a few hours even for a single elementary reaction.

Conclusion

We examined the performance of various ML models for activation energy prediction. The TB method showed the best performance with the small MAE of 1.95 kcal mol^{-1}, followed by the ANN and the SVR methods. This result manifests the feasibility of ML models for reliable prediction. However, the relatively large RMSE (4.49 kcal mol^{-1}) and small R^2 (0.89) by significant inconsistent errors were observed. Reactions with relatively large errors involved the change of a number of chemical bonds (>4), indicating complicated reaction mechanisms. These errors may be improved with better input features implying information on the reaction paths. A deep learning approach with big reaction data may be a viable solution. Also, the feasibility of its extension to solution-phase reactions should be addressed for broad applicability.

Acknowledgments

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning (NRF-2017R1E1A1A01078109).

Conflict of interest

The authors declare no conflict of interest.

Keywords: activation energy · gas-phase reactions · machine learning · quantum chemistry

[1] R. Ramakrishnan, O. A. von Lilienfeld, in *Reviews in Computational Chemistry Vol. 30* (Eds.: A. L. Parrill, K. B. Lipkowitz), Wiley, Hoboken, **2017**, pp. 225 – 257.
[2] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 6 – 13.
[3] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. Anatole von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.
[4] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087 – 2096.
[5] T. Suzuki, R. Tamura, T. Miyazaki, *Int. J. Quantum Chem.* **2017**, *117*, 33 – 39.
[6] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **2012**, *108*, 253002.
[7] X. Li, L. Chen, F. Cheng, Z. Wu, H. Bian, C. Xu, W. Li, G. Liu, X. Shen, Y. Tang, *J. Chem. Inf. Model.* **2014**, *54*, 1061 – 1069.
[8] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*, 2810.
[9] Y. Kim, J. W. Kim, Z. Kim, W. Y. Kim, *Chem. Sci.* **2018**, *9*, 825 – 835.
[10] 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; *Angew. Chem.* **2016**, *128*, 6004 – 6040.
[11] F. Battin-Leclerc, E. Blurock, R. Bounaceur, R. Fournet, P.-A. Glaude, O. Herbinet, B. Sirjean, V. Warth, *Chem. Soc. Rev.* **2011**, *40*, 4762 – 4782.
[12] T. Lu, C. K. Law, *Prog. Energy Combust. Sci.* **2009**, *35*, 192 – 215.
[13] P. Zimmerman, *J. Chem. Theory Comput.* **2013**, *9*, 3043 – 3050.
[14] Y. V. Suleimanov, W. H. Green, *J. Chem. Theory Comput.* **2015**, *11*, 4248 – 4259.
[15] R. Sumathi, H. H. Carstensen, W. H. Green, *J. Phys. Chem. A* **2001**, *105*, 8969 – 8984.
[16] R. Sumathi, H. H. Carstensen, W. H. Green, *J. Phys. Chem. A* **2002**, *106*, 5474 – 5489.
[17] M. Saeys, M.-F. Reyniers, V. Van Speybroeck, M. Waroquier, G. B. Marin, *ChemPhysChem* **2006**, *7*, 188 – 199.
[18] R. Van de Vijver, N. M. Vandewiele, A. G. Vandeputte, K. M. Van Geem, M. F. Reyniers, W. H. Green, G. B. Marin, *Chem. Eng. J.* **2015**, *278*, 385 – 393.
[19] C. W. Gao, J. W. Allen, W. H. Green, R. H. West, *Comput. Phys. Commun.* **2016**, *203*, 212 – 225.
[20] N. M. Vandewiele, K. M. Van Geem, M.-F. Reyniers, G. B. Marin, *Chem. Eng. J.* **2012**, *207 – 208*, 526 – 538.
[21] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241 – 2251.
[22] G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuc, E. P. Gajewska, B. A. Grzybowski, A. Gambin, *Sci. Rep.* **2017**, *7*, 3582.
[23] X. Yao, *Int. J. Intell. Syst.* **1993**, *8*, 539 – 567.
[24] D. Basak, S. Pal, D. C. Patranabis, *Neuronal Inf. Process.- Lett. Rev.* **2007**, *11*, 203 – 224.
[25] T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **2016**, ACM Press, New York, USA, pp. 785 – 794.
[26] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107 – 113.
[27] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, S. Goedecker, *J. Chem. Phys.* **2013**, *139*, 184118.
[28] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289 – 320.

Manuscript received: January 23, 2018
Accepted manuscript online: February 23, 2018
Version of record online: April 24, 2018