International Journal of Advanced Research in Engineering Technology & Science ISSN: 2349-2819
Impact Factor: 7.10
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Email: editor@ijarets.org Volume-11, Issue-6 June – 2024 www.ijarets.org

PREDICTING CHEMICAL REACTIVITY USING MACHINE LEARNING-BASED MOLECULAR MODELING

Abhay Krishna1, Rashmi Dubay2, Manvendra Singh3, Tasleem Jamal4

1Dept. of Applied Sciences & Humanities, K.M.C. Language University, Lucknow (U.P) India

2Dept. of Chemistry, School of Basic Sciences, C.S.J.M. University Campus, Kanpur (U.P) India

3Dept. of Biotechnology Engineering, K.M.C. Language University, Lucknow (U.P) India

4Dept. of Computer Science Engineering, K.M.C. Language University, Lucknow (U.P) India

ABSTRACT

The prediction of chemical reactivity is one of the most significant problems facing computational chemistry. This issue enables the optimization of reaction pathways and the quick discovery of innovative compounds. Recent advancements in molecular modeling and machine learning (ML) have made it feasible to consistently predict reactivity from huge chemical datasets. This capacity was previously available only in hypothetical situations. The purpose of this study is to investigate the use of graph neural networks (GNNs), molecular descriptors, and quantum chemistry calculations in order to model reaction patterns in a variety of chemical systems. Rather of relying on costly density functional theory (DFT) computations, the technique that has been presented makes use of supervised learning algorithms in order to anticipate reaction feasibility, rate constants, and selectivity. Representations that are based on fingerprints, electronic structural properties, and topological indices provide an improvement in the interpretability of predictive models via the use of feature engineering. In light of the outstanding predictive accuracy shown in benchmarking findings in comparison to experimental datasets, it is possible that molecular modeling powered by machine learning might be beneficial to the prediction of reactions, the creation of drugs, and the design of materials. This paper highlights the revolutionary potential of artificial intelligence in the area of chemistry by highlighting the need of fostering efficient and sustainable chemical research.

Keyword: - Machine learning, molecular modeling, chemical reactivity, quantum chemistry, density functional theory.

INTRODCUTION

The identification of the primary end products of chemical reactions, taking into account the original reactants and the conditions under which the reaction took place, is an important problem in organic chemistry. Through the use of reaction prediction, which is an essential component of retro-synthetic analysis and the development of virtual libraries for drug design, we are able to get further knowledge on biochemical

catalysis and biological metabolism. When it comes to the prediction of reactions, there have been at least three primary schools of thought that have emerged: inductive machine learning, rule-based expert systems, and physical simulations of transition states that make use of various quantum mechanical and other approximations. On the other hand, none of these techniques can compete with the remarkable expertise of a human scientist.

Previous approaches and representations

As a result of the fact that it represents a simplification, on a macroscopic level, of a very intricate reality on a microscopic level, which is ultimately regulated by the principles of statistical and quantum physics, the mere concept of a "reaction" has the potential to cause misunderstanding. Even when dealing with very small systems, it is difficult to find solutions that are true to the Schrodinger equation. Therefore, in the actual world, energies are determined by using approximations that have varying degrees of precision. Some examples of these approximations are ab-initio Hartree-Fock procedures, density functional theory, semi-empirical approaches, and mechanical force fields. The fact that this is the case means that reactions may be represented by a high-dimensional potential energy surface. Stable atomic configurations can be represented as minimum energy paths, and the best path is the one that goes through the saddle point, which is the state that has the lowest energy. These approaches need precise setup and require a significant amount of computing resources; nonetheless, they are able to generalize to a wide range of chemistries and produce exceptionally accurate findings by directly simulating interactions between energies. Despite the fact that this field of computational chemistry provides invaluable resources for thorough analysis, it is not yet capable of meeting the challenge of high-throughput reactivity operations and cannot take the place of human knowledge. In sharp contrast to the representation that rule-based expert systems often utilize for high-throughput reactivity tasks, which is generic transformations over chemical graphs, this is the representation that is used the most frequently. Through the process of locating a match among a collection of acceptable graph modifications, it is possible to forecast responses. As can be seen in Figure 1, these broad transformations are just representative of the general chemical changes that occur during processes that really include a succession of transition states. There are at least four issues that arise with these rule-based techniques, which are as follows: (1) their representation is overly abstract, making it impossible to discern the underlying physical reality; (2) a substantial amount of expert knowledge needs to be manually curated; (3) at larger scales, they become unmanageable due to the need to update a large portion of existing transformations with exceptions whenever a new graph pattern is added; and (4) they are not general enough, as specific chemistries need to be encoded in order to be predicted.

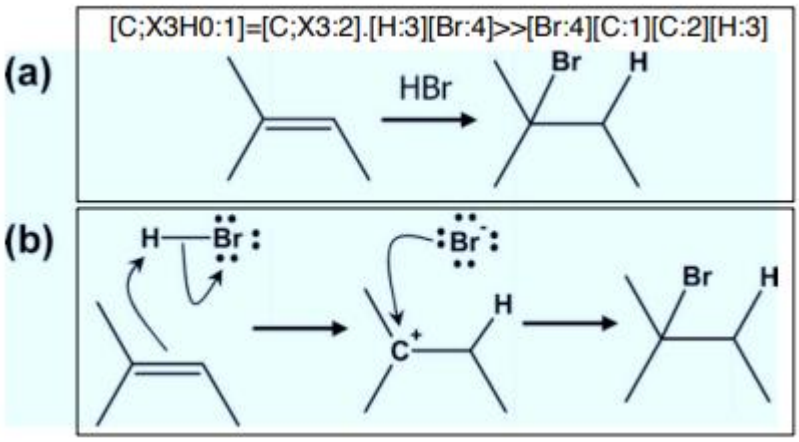

Figure 1: Alkene (hydrocarbon with double bond) to hydrobromic acid (HBr) transition and mechanistic processes. (a) illustrates the total transform as SMIRKS string pattern and graph. In a molecular network, vertices represent atoms and carbons are unlabeled. Edge count between vertices indicates bond order. +/− symbols indicate formal charge. Standard valences are filled with implicit hydrogens. (b) displays the two mechanistic reactions that make up the transformation as arrow pushing diagrams. Arrows show coordinated electron flow, whereas dots show lone pair electrons. The process begins with electrons in the carbon-carbon double bond breaking the hydrogen-bromine single bond, resulting in an anionic bromide (Br−) and a carbocation (C+). charged, electron-rich bromide attacks electron-poor carbocation to form alkyl halide.

Considerations of mechanistic reactions lie between those of low-level QM treatment and those of abstract graph-based overall transformations. The coordinated flow of electrons through a single transition state is the hallmark of a mechanistic or elementary process. You may make complete changes by combining these mechanical responses. For instance, Figure 1 depicts the two basic reactions that make up the transformation of an alkene interacting with hydrobromic acid to produce an alkyl halide, as well as the entire transformation. In a mechanistic reaction, two molecular orbitals—one serving as an electron donor and one as an electron sink—interact in an idealized form. MOs denote areas of the molecule where the electron density is high (source) or low (sink). Bonds and lone pairs of electrons make up potential electron sources, whereas vacant atomic orbitals and bonds make up potential electron sinks. In different situations, bonds may serve as sinks or sources. Due to limitations in space, we are unable to provide a comprehensive description of all the intricate molecular nuances that need to be addressed, such as the chaining involved in resonance rearrangement. See literature on mechanisms for details. It should be noted that this approach permits the complete listing of all conceivable mechanistic reactions across any set of molecules by taking into account all possible combinations of source and sink MOs. A rule-based expert system called Reaction Explorer was recently developed by Baldi and Chen. In this system, every pattern of rearrangement includes a basic reaction. In this context, the "productive" mechanical steps—the reactions that ultimately result in the main products—represent the elementary reactions. Therefore, "productive" elementary reactions are those that, while not optimal from a kinetic standpoint, lead to the final result of the thermodynamic

transformation. Though it's an improvement over earlier methods that relied on overall transformations, this rule-based system isn't perfect and has issues with generalizability, size, and curation.

Although mechanistic reaction representations are not exact solutions to the Schrödinger equation, we believe they will be more helpful than generalized transformations since they are closer to the actuality. Also, because of their simpler structure and mechanical interpretation, we anticipate that they will be simpler to forecast than total changes. Taken together, these points show that statistical machine learning methods may be easier to use and make use of when dealing with mechanical phases. So, reactions are depicted here as mechanisms, and from here on out, we'll use the word "reaction" to mean only one basic reaction. On top of that, we think the reaction prediction issue is just finding the "productive" reactions for a certain collection of reactants under a specific set of circumstances. Reaction prediction using machine learning has received surprisingly little attention. The only example is a 1990 article on a technique that was never used in a complete reaction prediction system: inductively obtaining overall transformation patterns from reaction databases. I find this circumstance rather unexpected. With the advancements in computer power and machine learning techniques over the last two decades, it is not hard to envision a machine learning system that uses reaction data to understand chemical language, for instance, using graph grammars. The dearth of relevant data may be contributing to the area's stagnation. Due to the prevalence of closed models in chemical publication, it is challenging to get access to the relevant literature. In addition, there is still no clear solution to the research challenge of how to parse scientific language and extract pertinent chemical information from text and picture data. The reactions in commercial databases like Reaxys or SPRESI are often imbalanced, not atom-mapped, and lack mechanistic clarity, albeit they do exist. Databases are either very expensive or come with a limited query interface that makes meaningful statistical data mining impossible; this is on top of the databases' extreme lack of transparency. That being said, we are still in the dark about whether or if there are any viable machine learning methods for response prediction.

OBJECTIVE

1. To create and test a machine learning-based molecular modeling framework that properly predicts chemical reactivity utilizing quantum chemistry data, molecular descriptors, and sophisticated algorithms.
2. To Evaluate suggested models against experimental and computational benchmarks for efficient reaction prediction in drug discovery, material design, and green chemistry.

METHOD

Theory of DeePHF

The energy disparity between a low-precision and a high-precision technique may be represented using the DeePHF methodology. The expression for the energy difference, Eδ, may be given in terms of the single-electron orbitals. $\{|\Psi_i\rangle\}$ from the low-precision method:

$$E_H - E_L = E_\delta \equiv E[\{|\Psi_i\rangle\}]$$

In the event where the Hartree-Fock (HF) methodology is regarded as having a low level of accuracy and the CCSD (T) method is regarded as having a high level of precision, our correlation energy is represented by the symbol Eδ. For the purpose of obtaining precise energy estimates, it is possible to create Eδ by using neural network models or other appropriate methodologies. The primary purpose of the neural network used in the DeePHF approach is to identify a correlation between the "local density matrix" and Eδ. A great number of parallels can be seen between this model and the reduced density matrix for single particles.

$$\Gamma(x, x') = \sum_i \langle x|\varphi_i\rangle\langle\varphi_i|x'\rangle = \sum_i \varphi_i^*(x')\varphi_i(x)$$

This matrix is derived by projecting the reduced density matrix of a single particle onto a set of atomic basis functions $\{\alpha^I_{nml}\}$, There are three numbers that are related with atom I: radial, azimuthal, and magnetic. These numbers are denoted by the letters n, l, and m, respectively. When it comes to deriving the single-particle reduced density matrix, the HF technique is not the only possible choice; there are other viable possibilities and alternatives as well. The expression for the "local density matrix" in its entirety is as follows:

$$(\mathcal{D}^I_{nl})_{mm'} = \sum_i \langle\alpha^I_{nlm}|\varphi_i\rangle\langle\varphi_i|\alpha^I_{nlm'}\rangle$$

The local density matrix eigenvalues serve as input descriptors for the neural network model to ensure rotational symmetry.

$$\mathbf{d}^I_{nl} = \mathrm{EigenVals}_{mm'}[(\mathrm{D}^I_{nl})_{mm'}]$$

$$E_\delta = \sum_I \mathcal{F}^{NN}(\mathbf{d}^I|\omega)$$

Through the neural network, the descriptor and energy Eδ are connected to one another. The fundamental purpose of the training technique is to bring about a reduction in the disparity that exists between the objectives and the expected levels of energy. A definition of the training loss function is as follows:

$$\mathcal{L} = \min_{\omega} \mathop{\mathbb{E}}_{\text{data}} [(E_{\text{label}} - E_{\delta}[\{\varphi_i\}|\omega])^2]$$

This loss function ensures that the model learns the energy correction Eδ with minimized prediction error.

Model Training

It is imperative that we pay attention to the training datasets before we begin training the models. Grambow and colleagues were the ones that contributed the datasets that had a high degree of accuracy. The datasets (G) that Grambow provides were divided into three sections: training, validation, and testing. This was done so that we could evaluate the effectiveness of the model. In all, there are 3,929 molecules and 1,192 reactions included in the testing set. Not only are the reactants, products, and transition structures contained in Grambow's datasets (G), but they are also included in the datasets themselves. Through the use of an additional dataset known as Transition1x (T), we are able to accentuate this. This particular dataset is made up of 10,000 reactions that were extracted from the Grambows database and then submitted to Nudged Elastic Band (NEB) simulations. The total number of chemical configurations that are included in this dataset is 9.6 million. A total of 1,000 reactions were selected for training purposes, 225 reactions were selected for validation, and 287 reactions were selected for testing from the datasets that included 8,000 NEB compounds. CCSD(T)-F12a was the method that was used in order to ascertain the energy values for each and every dataset. We train a large number of DeePHF models on a wide range of datasets by using DeePKS-kit56 and PySCF. Additionally, we determine the base energies by employing a number of baseline quantum mechanics approaches, including HF, PBE, B3LYP, M06-2X, and ὼB97M-V. Each of the three hidden layers that make up the full connect neural network model has 108 neurons. The model is fully connected. Prior to the fitting process, regularization is performed along the ridges. The batch size of a training regimen that integrates the Adam optimizer is sixteen, and the learning rate is set at three times ten to the power of four. With the passage of time, the learning rate decreased at an exponential rate, with a change occurring every 500 epochs; the decay factor was 0.96. A whole fifteen thousand epochs are designated for the training method to be carried out.

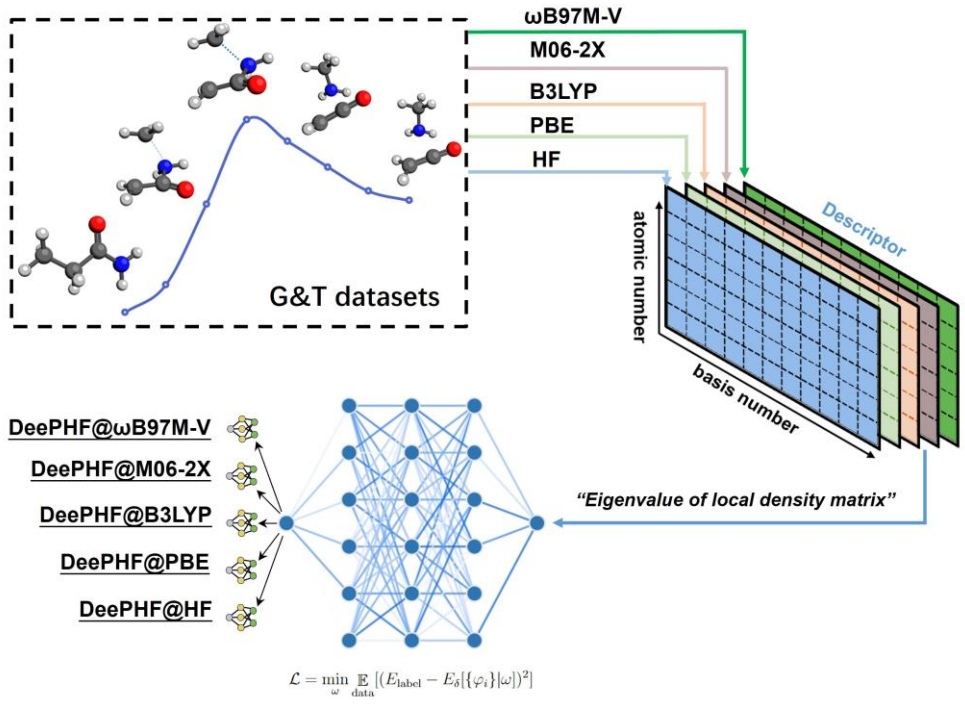

Figure 2: Procedure for DeePHF model training: Using basic QM techniques (HF, PBE, B3LYP, M06-2X, ωB97M-V), get the descriptor of Grambow's datasets (G) and Transition1x (T) in the first step. Then, train and obtain various DeePHF models.

Test Datasets Cleaning

For the purpose of ensuring the accuracy of our models, we have compiled a variety of high-precision reaction barrier datasets. These datasets include GMTKN55, BH9, Transition1x, and the recently developed comprehensive reaction and thermal property database known as Reaction Graph Depth 1 (RGD), among others. We performed a re-optimization of the RGD1 reaction structures and assessed a selection of them in order to guarantee that the model assessment was not inconsistent. This table provides a comprehensive overview of the dataset that was collected.

The GMTKN55 dataset is often used for the purpose of evaluating DFT models because to the excellent accuracy and variety it has. A broad variety of processes, ranging from hydrogen transfer to pericyclic reactions and even farther, are included in the seven categories of reaction barrier heights. By using a filter, we are able to exclude the possibility of reactions that include atoms that are not C, H, N, or O in closed-shell configurations that are neutral. Following the application of filters, the GMTKN55 dataset is comprised of 157 compounds and 84 reactions. The extensive testing of models, on the other hand, is still inadequate. The Mechanism and Catalytic Site Atlas (M-CSA) reactions are a part of the reaction benchmark dataset BH9. The major focus of the M-CSA reactions is on reactions that have biological significance. The procedure of purifying the data has resulted in the preservation of 341 structures and 97 reactions for the purpose of future testing. The growing string method (GSM) was used in the construction of Grambow's

dataset (G), which is a subset of GDB-17. This dataset covers reactions that involve up to seven heavy atoms. In this particular dataset, there were 3,929 molecules and 1,192 reactions that were used for testing purposes. The Transition1x (TEB) algorithm generates 9.6 million molecular conformations by using 10,000 processes and NEB calculations. This algorithm is built on the dataset that Grambow used. Using a random selection process, we choose the most recent NEB trajectories for 287 reactions, with eight conformers for each reaction. The RGD1 dataset was produced by Zhao et al. using data retrieved from the PubChem database. This dataset includes reactions that include a maximum of 10 heavy elements and comprises as many reactions as possible. For the purpose of constructing the test set, we choose one hundred answers at random from the datasets that have been tagged with CCSD(T)-F12a/cc-pVDZ-F12. Using the methods óB97XD3/def2-TZVP, the designs will be optimized for optimal performance. The initial geometries were used in the testing of 83 reactions, and the optimized geometries were then utilized for evaluation.

Due to the fact that our method of data cleaning and selection maintains the exceptional quality and variety of these datasets, we are in a position to carry out an evaluation of the model's performance that is both relevant and accurate. Our models will be executed on these datasets, and the results will be reported thereafter.

Table 1: Overview of test datasets in this work.

Subset	Description	Method for energy calculation	Method for geometry optimization	No.Reactions	No.Structures
BH76	Hydrogen transfer reactions	W2-F12	unknown	4	5
BHPERI	Pericyclic reactions	W1-F12 W2-F12	B3LYP/6-31G*	22	45
BHDIV10	Diverse reactions	W1-F12 W2-F12	PBEh-3c	4	10
INV24	Inversion racemisation reactions	W1-F12 W2-F12 DLPNO-CCSD(T) /CBS	B3LYP-D3(BJ) /def2-TZVPP	13	24

BHROT27	Rotation reactions	W1-F12 W2-F12	TPSS-D3(BJ) /def2-TZVPP	19	30
PX13	Proton-exchange reactions	W1-F12	B3LYP/AVTZ	8	16
WCPT18	Proton-transfer reactions	W2.2	B3LYP/AVTZ	12	19
BH9	Bio-organic organic reactions	DLPNO-CCSD(T) /CBS	(CAM-)B3LYP-D3 /6-31(G*,+G*,+G**)	97	341
Grambow(G)	Generated C,H,N,O reactions	CCSD(T)-F12a /cc-pVDZ-F12	ωB97X-D3 /def2-TZVP	1192	3929
Transition1x(T)	NEB trajectories for C,H,N,O	CCSD(T)-F12a /cc-pVDZ-F12	ωB97X /6-31G*	287	2296
RGD1	Generated C,H,N,O reactions	CCSD(T)-F12a /cc-pVDZ-F12	B3LYP-D3 /TZVP	83	249
RGD1’	Generated C,H,N,O reactions	CCSD(T)-F12a /cc-pVDZ-F12	ωB97X-D3 /def2-TZVP	83	249

*RGD1 and RGD1’ are the same reactions, but different optimization methods

RESULTS AND DISCUSSION

Performance on Train, Validation and Test sets

In order to get results that are comparable to those obtained from the CCSD(T)-F12a/cc-pVDZ-F12 reference, we will first evaluate the capability of the model to predict absolute energy by using a number of different methods. Among the three datasets, the DeePHF@M06-2X model achieves the highest mean absolute error (MAE). These datasets are Grambow(G), Transition1x(T), and Grambow and Transition1x(G&T). For the purpose of generating baseline energy and the local density matrix, this model makes use of M06-2X in conjunction with PySCF and the DeePKS-kit implementation. The chemical accuracy on the test and validation sets is achieved by all models with the exception of one, which is the DeePHF@HF model in Transition1x. The other models exhibit signs of overfitting and make a much smaller number of errors on the training set compared to the other two sets. The sample size that was utilized to train 1,000 reactions on the Transition1x dataset was rather tiny, consisting of just 8,000 molecules. This was due to the fact that the whole dataset contains 10,000 reactions and 9.6 million molecules. Consequently, overfitting is especially obvious on this dataset as a consequence of this. A consistent performance is shown by the fourth-rung functionals in Jacobs Ladder, which include the higher-accuracy models DeePHF@M06-2X, DeePHF@B3LYP, and DeePHF@ϋB97M-V. This performance is seen throughout all three sets of data, namely training, validation, and testing. The following test outcomes, on the other hand, make it clear that there are differences in their performance when subjected to more complete evaluation datasets. Despite the fact that the validation and testing MAEs of DeePHF@PBE are considerably lower than those of the aforementioned three approaches, the MAE of its training set is quite close to those of the other techniques. Training DeePHF@HF with Transition1x alone will result in validation and testing errors that are more than 1 kcal/mol both in and of itself. Nevertheless, with the use of G&T training, accuracy will be better, and performance will be improved according to the quantity of data.

Table 2: DeePHF models performance using different base methods across training, validation, and testing sets, reported as MAE (kcal/mol).

Datasets	T			G			G&T		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Number	8000	1800	2296	32691	3845	3929	40691	5645	6225
DeePHF@HF	0.12	1.21	1.01	0.08	0.41	0.40	0.10	0.52	0.49
DeePHF@PBE	0.05	0.67	0.68	0.05	0.24	0.24	0.06	0.31	0.29
DeePHF@B3LYP	0.05	0.50	0.46	0.04	0.18	0.18	0.05	0.24	0.22
DeePHF@M06-2X	0.05	0.46	0.39	0.04	0.18	0.18	0.05	0.20	0.20

DeePHF@ωB97M-V	0.05	0.48	0.39	0.04	0.19	0.19	0.05	0.21	0.21

Through a comparison of the reaction energies and barrier heights of the Grambow dataset, we were able to evaluate the performance of the model in reaction systems. In addition to our DeePHF models, we include a variety of well-known density-functional theory (DFT) methods, including πB97X-V, πB97X-D3, πB97M-V, MN15, MN15-L, and M06-2X, as well as XYG series (XYG3, XYGJ-OS) functionals, for the sake of comparison. It is important to note that all of these techniques use the def2-TZVP basis set. The DeePHF models outperform ordinary DFTs and double hybrid functionals in terms of mean absolute error (MAE) and root mean square (RMSE) for both forward and reverse barrier heights (Table 3). The exceptions to this rule are DeePHF@HF(G) and DeePHF@HF(G&T). DeePHF@HF has a root mean square error (RMSE) that is slightly larger than that of the double hybrid XYGJ-OS. This is due to the fact that the reversal barrier heights in the G dataset are 1.76 kcal/mol, whereas in the G&T dataset they are 1.80 kcal/mol. There is a significant majority of DFTs that do not work effectively when confronted with high obstacles. DeePHF models, on the other hand, provide a higher level of precision and make use of information from a number of quantum techniques about the local density matrix. In specifically, the DeePHF@M06-2X (G&T) molecule has the lowest reaction energy RMSE of 0.22 kcal/mol and the lowest barrier height RMSE of 0.84 kcal/mol for the forward and reverse reactions, respectively. Both of these values are measured in kcal/mol. With regard to its core approach M06-2X, the root-mean-squared error (RMSE) values are as follows:

Table 3: CCSD(T)-F12a/cc-pVDZ-F12 evaluation of DeePHF models and DFTs in reaction energies and barrier heights using Grambow's test datasets. Forward and reverse responses have different barrier heights. Results are MAE(kcal/mol) and RMSE. Bold indicates the best result.

	BH(Forward)		BH(Reverse)		RE	
Error(kcal/mol)	MAE	RMSE	MAE	RMSE	MAE	RMSE
ωB97X-V	4.25	6.31	3.24	5.37	2.63	3.41
MN15-L	3.94	4.77	2.96	3.86	3.26	4.24
ωB97X-D3	3.71	5.58	3.02	4.89	2.27	2.88
M06-2X	3.37	5.35	3.26	5.17	2.10	2.68
MN15	2.53	3.83	2.78	3.95	2.49	3.22
ωB97M-V	2.60	4.50	2.59	4.38	1.44	1.92

XYG3	1.52	2.00	1.43	1.91	1.39	1.77
XYGJ-OS	1.43	1.81	1.36	1.75	1.13	1.46
DeePHF@HF(G)	0.97	1.67	1.04	1.76	0.33	0.69
DeePHF@PBE(G)	0.54	0.89	0.59	0.95	0.20	0.43
DeePHF@B3LYP(G)	0.42	0.76	0.45	0.79	0.14	0.26
DeePHF@M06-2X(G)	0.43	0.83	0.47	0.87	0.14	0.27
DeePHF@ωB97M-V(G)	0.46	0.92	0.48	0.95	0.14	0.31
DeePHF@HF(G&T)	0.95	1.75	1.02	1.80	0.29	0.57
DeePHF@PBE(G&T)	0.52	0.87	0.65	1.05	0.21	0.38
DeePHF@B3LYP(G&T)	0.39	0.67	0.43	0.71	0.15	0.26
DeePHF@M06-2X(G&T)	0.41	0.84	0.43	0.86	0.12	0.22
DeePHF@ωB97M-V(G&T)	0.42	0.90	0.45	0.94	0.13	0.29

The forward and reverse barrier heights are 5.35 and 5.17 kcal/mol, respectively, while the reaction energy is 2.68 kcal/mol. While M06-2X outperforms the ϋB97M-V, the accuracy of DeePHF@ϋB97M-V (G&T) is comparable to that of DeePHF@M06-2X (G&T), indicating that DFT baselines belonging to the same functional family (or "rung" of Jacob's ladder) display comparable accuracy.

Reaction energies and barrier heights can be accurately predicted using DeePHF models, which show good transferability. Consistent with earlier results on the transferability of DeePHF, the MAE and RMSE for reaction energies are significantly low because the reactants and products are close to equilibrium structures. Extending our assessments to 287 reactions from the Transition1x dataset, we tested the models' capacity to depict combinations beyond reactants, products, and transition states. By comparing each model's relative energy profiles to CCSD(T)-F12a/cc-pVDZ-F12, we can evaluate how well each one handles the energy surface.

By subtracting the energy of the first conformation from each successive value, Figure 2(a) displays the energy profile of each reaction as the relative energy of eight molecular conformations. When we compare the relative energies, we can see whether the potential energy surface is regular. The choice response reveals that CCSD(T)-F12a coincides with DeePHF@M06-2X(G&T). Figure2 (b) shows the overall MAEs of the

287 reactions. The performance of the DeePHF models may be greatly enhanced by including Grambow's dataset. The best MAE of 0.30 kcal/mol was achieved by DeePHF@M06-2X(G&T), whereas DeePHF@B3LYP(G&T) and DeePHF@üB97M-V(G&T) gave equally good results. With the exception of DeePHF@HF(T), all MAEs of the DeeHF models are less than 1 kcal/mol. The XYGJ-OS is the only functional model with an MAE less than 1 kcal/mol, and among common DFTs, üB97M-V is the best.

Due of their intrinsic instability, structures at transition states may have relative energy errors; hence, we should be careful, even if the DeePHF@M06-2X(G&T) energy profile nearly matches the high-precision CCSD(T)-F12a/cc-pVDZ-F12 reference. We compare the relative energies of three unstable, high-energy structures at the transition state (E1 and E2 in Figure2(a)) in Figure2(c). A MAE of 0.52 kcal/mol confirms that DeePHF@M06-2X(G&T) is once again the top performer. There is a need for more refinement to improve the model's accuracy in high-energy, transition-state-adjacent areas, since all approaches demonstrate that MAE is raised in these unstable structures. The great transferability of these DeePHF models is shown even for unstable structures, as demonstrated by the comparable performance of DeePHF@M06-2X(G&T), DeePHF@B3LYP(G&T), and DeePHF@üB97M-V(G&T). With an MAE more than 1 kcal/mol, XYGJ-OS is the most functional. The DeePHF MAE at HF (T) is near to üB97X-V and M06-2X, however it is more than 2 kcal/mol. You may get the techniques' detailed value in Tables S1 and S2.

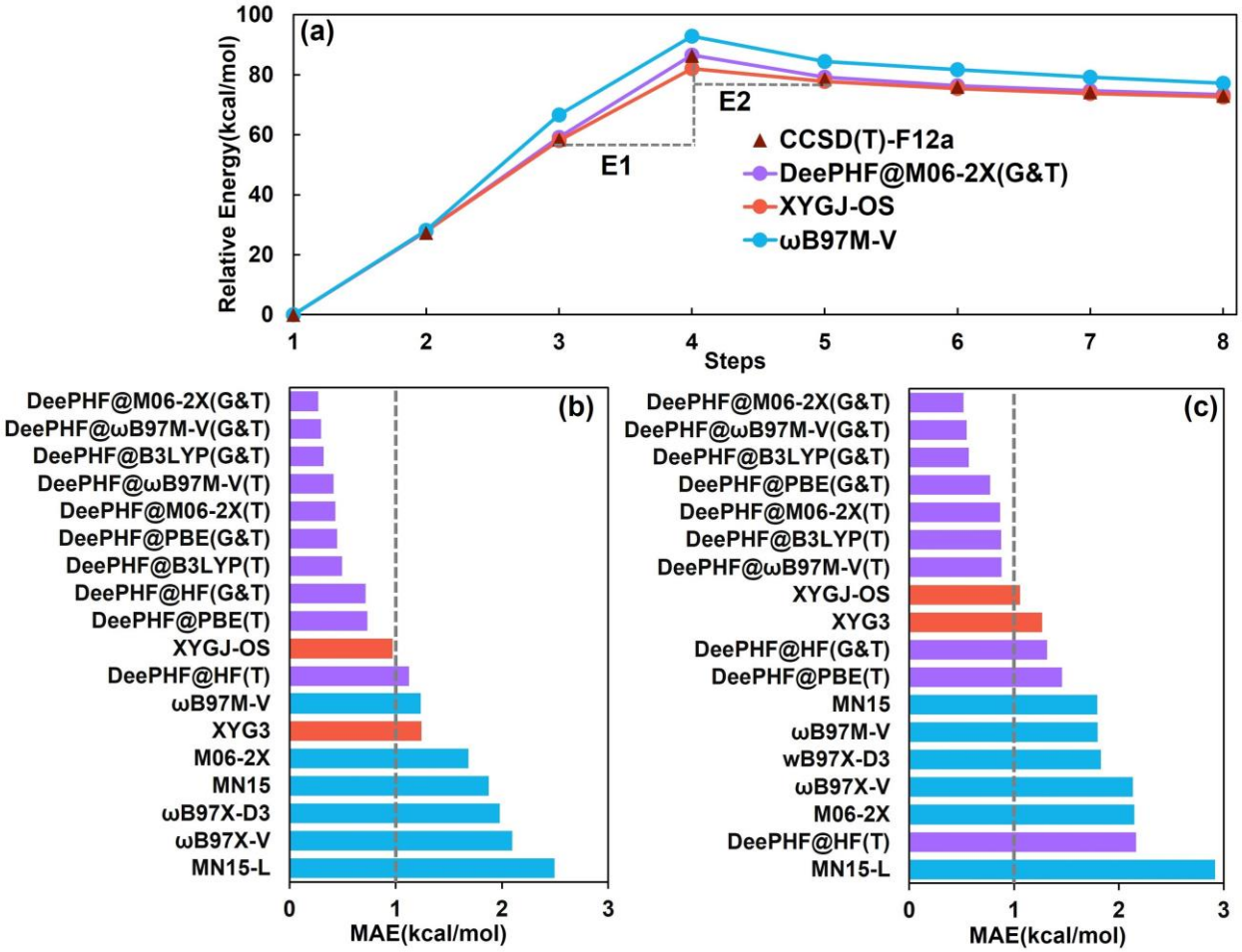

Figure 3: DeePHF and DFT performance is assessed using Transition1x datasets. (a) Reaction relative energy profile using CCSD(T)-F12a as reference. For comparison, DeePHF@M06-2X(G&T), XYGJ-OS, and ωB97M-V are used. b) MAE data for relative energy comparison across eight reaction conformations using various methodologies. Sky-blue boxes represent conventional DFTs, purple boxes DeePHF models, and red boxes double hybrid functionals. The gray line represents a 1 kcal/mol inaccuracy. (c) MAE findings for three unstable structures around the transition structure and relative energy E1/E2 comparison.

Performance on RGD1 and RGD1'

Using the YARP tool, Zhao et al. produced 176,992 reactions from the PubChem dataset in order to overcome the diversity constraints of current transition state databases. Among the 126,857 distinct reactions included in the whole dataset are 33,032 that have two or more transition structures; these reactions include molecules containing as many as 10 heavy atoms. In order to test our model apart from the training data, we randomly choose 100 responses. Nevertheless, we narrowed our focus to 83 convergent reactions in this dataset and determined their energies using CCSD(T)-F12a/cc-pVDZ-F12, taking into account both the initial and improved geometries. Figure 3 shows that out of all the geometries tested, DeePHF@B3LYP(G&T) had the

highest MAE of 1.87 kcal/mol for barrier heights in the first designs. But these mistakes are still worse than the chemical precision. DeePHF@M06-2X(G&T) demonstrated excellent transferability of DeePHF models for stable molecular geometries, as it achieved the best results for reaction energy estimates with an MAE of 0.43 kcal/mol. A combination of B3LYP-D3/TZVP optimization and GFN2-xTB intrinsic reaction coordinate (IRC) computations validated the initial structures.

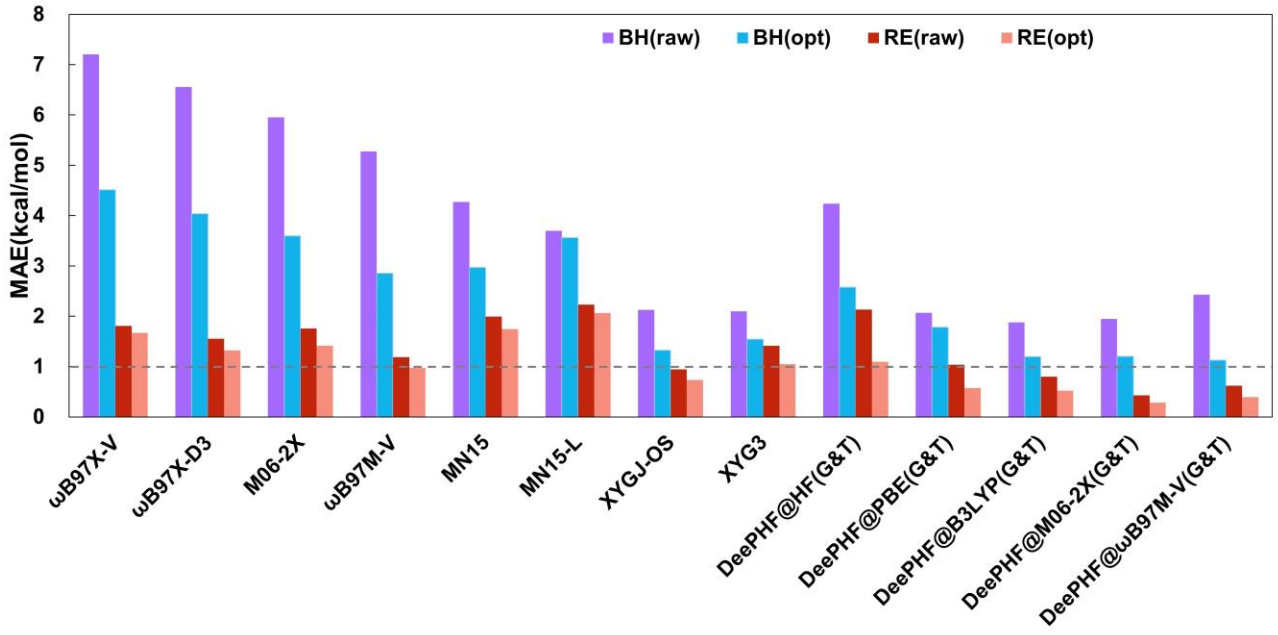

Figure 4: MAE performance of DeePHF models and DFTs in reaction energies and barrier heights for the RGD1 and RGD1' datasets. BH(raw) and RE(raw) refer to the use of original geometries, while BH(opt) and RE(opt) use optimized geometries. The gray line represents 1 kcal/mol.

We revised the reference energies CCSD (T) -F12a / cc-pVDZ-F12 and optimized these structures using ü B97X-D3 / def2-TZVP. Optimised geometry boost model performance, as shown in Figure 3. With a maximum activity excess (MAE) of 1.13 kcal/mol at barrier heights, which is close to chemical precision, DeePHF@üB97M-V(G&T) produced the best results. The performance of DeePHF@M062X(G&T)(1.20 kcal/mol) and DeePHF@B3LYP(G&T)(1.20 kcal/mol) was similar to that of DeePHF@üB97M-V(G&T). While DeePHF@HF(G&T) performed poorly in comparison to MN15 and MN15-L, DeePHF@M06-2X(G&T) and DeePHF@üB97M-V(G&T) demonstrated somewhat lower accuracy in barrier heights when using original geometries compared to DeePHF@PBE(G&T). The results show that optimized transition structures have a substantial effect on barrier height predictions, as DeePHF@M06-2X(G&T) and DeePHF@üB97M-V(G&T) fared better than DeePHF@PBE(G&T), while DeePHF@HF(G&T) performed better than MN15 and MN15-L. Despite this improvement, üB97M-V and XYGJ-OS both demonstrate less accuracy in barrier heights, with üB97M-V having an MAE of 2.86 kcal/mol and XYGJ-OS being somewhat better at 1.32 kcal/mol. Tables S3 and S4 show the result of our comparison of the front and reverse barrier heights' detailed MAE and RMSE. Most of the reactions in the RGD1 dataset are involving tiny molecules.

Further testing on larger molecules will fully evaluate the transferability of the model.

Performance on BH9

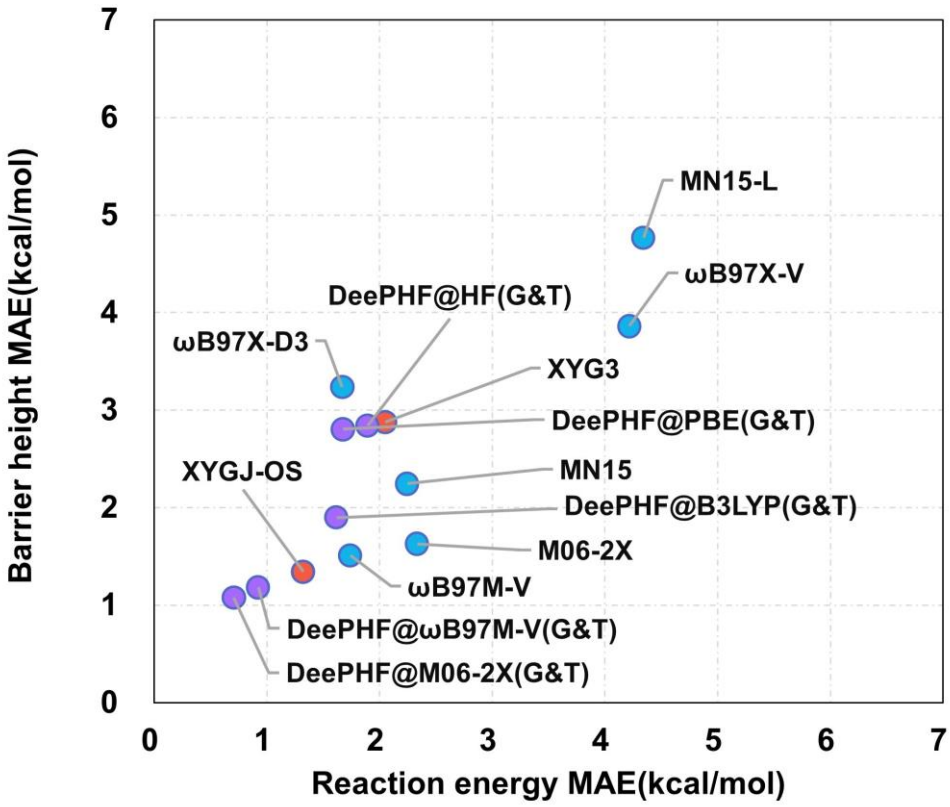

Figure 5: MAE performance of Barrier height vs Reaction energy for DFTs and models in BH9 datastes, different colors are different type methods. The purple points is DeePHF models, sky-blue is common DFTs, and red points represent double hybrid functionals

Our evaluation of the model's performance in medium-sized molecules followed our evaluation of its transferability in RGD1 datasets. We showed that the model was quite applicable to bigger molecules in our earlier study. There are several bio-organic reactions in the BH9 dataset. The filtered version has up to 31 heavy atoms. Reaction energies and barrier heights are positively correlated in the dataset, as seen in Figure 4. With an MAE of 1.08 kcal/mol for barrier heights and 0.70 kcal/mol for reaction energies, DeePHF@M062X(G&T) outperforms DeePHF@ϋB97M-V(G&T) in terms of performance. The reaction energies and barrier heights of DeePHF@B3LYP(G&T) are marginally lower, at 1.62 kcal/mol and 1.90 kcal/mol, respectively. Some more approaches that produce comparable reaction energy MAEs (1.601.75 kcal/mol) include ϋB97M-V, ϋB97X-D3, DeePHF@B3LYP(G&T), and DeePHF@PBE(G&T). However, their performance in terms of barrier height varies greatly, with ϋB97M-V producing 1.51 kcal/mol and ϋB97X-D3 producing 3.24 kcal/mol. Among all methods, the double hybrid functional XYGJ-OS comes in at number three. Its reaction energy MAE is 1.34 kcal/mol, and its barrier height is 1.32 kcal/mol. Also,

XYG3's MAEs for reaction energies are 2.05 kcal/mol and barrier heights are 2.88 kcal/mol, indicating poorer accuracy. See Table S5 for a breakdown of the 97 reactions included in the BH9 dataset into the following six types: cyclo-addition, electrocyclic, intramolecular, rearrangement, and Diels-Alder processes. Chemical correctness is typically achieved by DeePHF@M06-2X(G&T) throughout these categories, with the exception of Diels-Alder reactions, where it performs somewhat worse with a barrier height MAE of 1.40 kcal/mol. The only model that achieves a lower inaccuracy for cyclo-addition processes is DeePHF@M06-2X(G&T). Its reaction energy MAE is 0.74 kcal/mol and its barrier height MAE is 1.01 kcal/mol. Finally, we'll run tests on the seven GMTKN55 barrier height subgroups to see how well the model performs across different response types.

Performance on GMTKN55

Table 4: Performance of DeePHF models and DFTs on seven barrier height datasets from GMTKN55, with results reported as MAE and errors in kcal/mol. The best result is highlighted in bold.

Datasets	BH76	BHPERI	BHDIV10	BHROT27	WCPT18	PX13	INV24	Total1	Total2*
Number	5	22	5	19	12	8	13	84	63
MN15-L	2.35	1.91	2.79	0.84	1.78	4.66	1.05	1.86	1.50
M06-2X	1.03	1.60	1.11	0.39	2.51	6.21	1.27	1.78	1.65
ωB97X-D3	0.88	2.65	1.04	0.37	2.45	2.60	1.16	1.67	1.34
MN15	1.92	1.43	1.11	0.46	2.45	2.59	2.30	1.61	1.67
ωB97X-V	1.01	1.99	1.31	0.29	2.78	3.36	1.06	1.61	1.32
ωB97M-V	1.55	1.17	1.67	0.25	2.90	3.20	1.17	1.46	1.29
XYGJ-OS	1.39	2.17	1.54	0.32	1.79	2.65	0.72	1.43	1.35
XYG3	0.68	0.49	1.68	0.34	1.66	1.45	0.54	0.80	1.43
DeePHF@HF(G&T)	2.79	1.65	1.41	0.56	1.38	6.89	2.27	2.01	1.05
DeePHF@PBE(G&T)	1.10	1.07	0.77	0.43	2.08	6.82	2.74	1.86	0.78

DeePHF@B3LYP(G&T)	0.65	0.61	0.57	0.29	1.36	6.76	1.68	1.40	0.66
DeePHF@M06-2X(G&T)	0.69	0.60	0.81	0.20	0.98	4.83	2.86	1.33	0.58
DeePHF@ωB97M-V(G&T)	0.31	0.74	0.71	0.15	1.22	4.64	2.62	1.31	0.62

Total2∗ remove PX13 and INV24 subsets

Finally, we used the GMTKN55 filtered barrier height subsets to do one more test. From all 84 reactions in the collection, XYG3 has the best MAE of 0.80 kcal/mol, followed by DeePHF@ὼB97M-V(G&T) with 1.31 kcal/mol and DeePHF@M06-2X(G&T) with 1.33 kcal/mol, as shown in Table 4. When tested on the INV24 and PX13 subsets, the DeePHF models failed miserably. Unfortunately, our training dataset does not include PX13's proton-exchange interactions with multimer structures like trimers and hexamers. With its inversion/racemization processes mostly involving macrocyclic aromatic hydrocarbons, INV24 poses a new difficulty. We did not include either of them in our training dataset. With PX13 and INV24 removed, the DeePHF models performed much better, with the best MAE of 0.57 kcal/mol being DeePHF@M062X(G&T). The most effective common DFT, ὼB97M-V, outperforms XYGJ-OS and XYG3 with an MAE of 1.29 kcal/mol. These findings suggest that, in the presence of enough reaction data, DeePHF models may improve DFT's accuracy. On the other hand, Table 4 shows that in order to enhance the performance of the model in the future, more training data on specific response types like PX13 and INV24 is needed.

CONCLUSION

We used local density matrices produced by HF, PBE, B3LYP, M06-2X, and ὼB97M-V as descriptors to train 15 DeePHF models on three datasets (G, T, and G&T) in this work. The goal of creating these models was to accurately anticipate gas-phase reaction energy and barrier heights. To ensure the models were accurate and transferable, we put them through a battery of tests on different response datasets. The findings show that the DeePHF model's correctness is in agreement with the fundamental quantum techniques. For medium-sized molecules in the RGD1, BH9, and GMTKN55 databases, DeePHF@M06-2X(G&T) and DeePHF@ὼ97M-V(G&T), trained on ὼB97M-V and M06-2X descriptors, respectively, show impressive performance in predicting reaction energies and barrier heights.

REFERENCES

1. Guan, X.; Das, A.; Stein, C. J.; Heidar-Zadeh, F.; Bertels, L.; Liu, M.; Haghighatlari, M.; Li, J.; Zhang, O.; Hao, H.; Leven, I.; Head-Gordon, M.; Head-Gordon, T. A benchmark dataset for Hydrogen Combustion. Scientific Data 2022, 9.

2. Spiekermann, K.; Pattanaik, L.; Green, W. H. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. Scientific Data 2022, 9.

3. Grambow, C. A.; Pattanaik, L.; Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. Scientific Data 2020, 7.

4. Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. H. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. Nature Communications 2020, 11.

5. Himo, F. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. Journal of the American Chemical Society 2017, 139, 67806786.

6. Sheng, X.; Kazemi, M.; Planas, F.; Himo, F. Modeling Enzymatic Enantioselectivity using Quantum Chemical Methodology. ACS Catalysis 2020, 10, 64306449.

7. Sheng, X.; Himo, F. The Quantum Chemical Cluster Approach in Biocatalysis. Accounts of Chemical Research 2023, 56, 938947.

8. Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational prediction of chemical reactions: current status and outlook. Drug Discovery Today 2018, 23, 12031218.

9. Lam, Y.-h.; Abramov, Y.; Ananthula, R. S.; Elward, J. M.; Hilden, L. R.; Nilsson Lill, S. O.; Norrby, P.-O.; Ramirez, A.; Sherer, E. C.; Mustakis, J.; Tanoury, G. J. Applications of Quantum Chemistry in Pharmaceutical Process Development: Current State and Opportunities. Organic Process Research amp; Development 2020, 24, 14961507.

10. Bachrach, S. M. Challenges in computational organic chemistry. WIREs Computational Molecular Science 2014, 4, 482487.

11. Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. Accounts of Chemical Research 2017, 50, 539543.

12. Karton, A. A computational chemists guide to accurate thermochemistry for organic molecules. WIREs Computational Molecular Science 2016, 6, 292310.

13. Peterson, K. A.; Feller, D.; Dixon, D. A. Chemical accuracy in ab initio thermochemistry and spectroscopy: current strategies and future challenges. Theoretical Chemistry Accounts 2012, 131.

14. Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. The Journal of Chemical Physics 2013, 138.

15. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. The Journal of Chemical Physics 2013, 139.

16. Lead compound design for TPR/COX dual inhibition
A. Krishna and A. Yadav
J. Mol. Modeling, Vol.18(9), 4397-4408, 2012.
* In silico design of a dual TPR/TxS inhibitor for Venous thromboembolism and related cardiovascular diseases
A. Krishna and A.Yadav
Can. J. Chem, Vol.92(4), 284-292, 2014.