Assessment of Predicting Frontier Orbital Energies for Small Organic Molecules Using Knowledge-Based and Structural Information

Zong-Rong Ye,[1] Sheng-Hsuan Hung,[1] Berlin Chen,[2,*] Ming-Kang Tsai[1,*]

[1] Department of Chemistry, National Taiwan Normal University, Taipei, 11677, Taiwan

[2] Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, 11677, Taiwan

Email: berlin@csie.ntnu.edu.tw and mktsai@ntnu.edu.tw

Abstract

A systematic comparison is demonstrated for the predictions of frontier orbital energies – HOMO (E_H), LUMO (E_L), and energy gap (ΔE_{HL}) of the molecules in QM9 dataset, where it contains 120k-plus three-dimensional organic molecule structures determined by first-principle simulations. The target molecular properties (E_H, E_L, and ΔE_{HL}) are predicted using the linear regression (LR), machine learning (random forest, RF), and continuous-filter convolutional neural network (SchNET) approaches. LR and RF models built upon various knowledge-based descriptors, being derived from SMILES of the molecules, can provide predictivity of the target properties with the mean-absolute-errors (MAEs) at 4-6 times of chemical accuracy (0.043 eV). The best approach – SchNET, using the graph representation derived from molecular Cartesian coordinates, is confirmed to provide MAEs of E_H, E_L, and ΔE_{HL} at 0.051, 0.041, and 0.076 eV, respectively. With the introduction of bond-step matrix representation with SchNET model, the computational cost of dataset preparation can be substantially reduced, and the corresponding MAEs increases moderately to 2-3 times of chemical accuracy. The chemical interpretation of the important descriptors identified in the LR and RF models appear to align with the chemical knowledge of describing these molecular electronic properties, however, being accompanied with tolerable prediction errors. The combination of bond-step representation and SchNET model can provide an assessable-and-balanced option for the high-throughput screening of organic molecules and the preparation of data science approach.

Introduction

The energy difference between the highest molecular orbital (HOMO annotated as E_H) and the lowest unoccupied molecular orbital (LUMO annotated as E_L), being abbreviated as ΔE_{HL}, is commonly used to characterize the fundamental electronic properties of molecules. For instance, the electrical resistivity of molecules is generally considered to be directly proportional to ΔE_{HL}. The size of ΔE_{HL} could be quantitatively determined by the various experimental or theoretical approaches – the electrochemical oxidation/reduction potential measurements or the optimized wavefunctions governed by *ab initio* electronic structure calculations. The absolute values of E_H and E_L are derived by the explicit *inter-particle* interactions between electrons and atomic nuclei. Synthetic chemistry has a long history in developing the molecular structure diversity, leading to the ideal molecular electronic properties for the specific chemical applications. Understand the interplay between molecular structures and electronic properties plays a critical role to the pace of these scientific developments. Despite the quantum chemical calculation based approaches have been popularly adopted to investigate the insights of the molecular structure diversities, a computationally efficient-and-interpretable approach is the interest of chemistry community for addressing the unlimited possibility of molecular architectures.

Machine learning based approaches for describing the electronic properties of molecules and materials have been recently approached by several pioneering reports in the literature.[1-36] Pereira et al. used various nonlinear regression models including neural network (NN) method to predict the orbital energies of 111k molecules consisted of several main group elements.[8] Ramakrishnan and Lilienfeld introduced a property-invariant kernel for the machine learning models in predicting the various electronic and thermodynamic properties, including E_H, E_L, and ΔE_{HL}, out of 110k organic molecules.[3] Both Coulomb matrix[1] and bag-of-bonds[2] descriptors, being in conjunction with supplying three dimensional molecular structures, were adopted to represent the chemical space, and the results were close to the level of chemical accuracy at 1 kcal/mol (~0.043 eV). Huan et al. developed a class of hierarchal motif-based fingerprints to represent the molecular structures, and the fingerprints were classified as the zeroth- to third-order expressions, as being described in the format of multi-dimensional vectors.[4] The average predicted error of ΔE_{HL} was reported to be about 0.2 eV.[4] Browning et al. introduced the generic algorithm optimization of training set approach, in which the training set was categorized as 10 classes subject to the targeted property, and the average predicted error was improved as 0.173, 0.243, 0.317 eV for the energies of E_H, E_L, and ΔE_{HL}, respectively.[6] Faber et al. reported a comprehensive comparison using the various combination of regressor/representation/property, and the predicted results were shown to outperform the hybrid functional of Density Functional Theory (DFT) for describing the electronic ground-state properties of organic molecules in QM9 dataset.[7] Among their predicted 13 electronic properties, the mean absolute error (MAE) of E_H, E_L, and ΔE_{HL} were up to 0.228, 0.373, 0.441 eV, respectively, using linear model with elastic net regularization (0.221/0.367/0.430 eV with linear Bayesian ridge regression model).[7] It should be noted that MAE of E_H, E_L, and ΔE_{HL} can be generally reduced to half in respect to the linear

results if the non-linear models (Kernel ridge regression or random forest) were adopted while the use of NN approach was shown to go well beyond the accuracy of hybrid DFT.[7]

Gilmer et al. reported the message passing neural network (MPNN) models for the predictions of quantum chemistry properties.[37] Schütt et al. introduced a deep learning model (SchNET) and provided the predictions of E_H, E_L, and ΔE_{HL} to satisfy the level of chemical accuracy.[38] Such an apparent improvement required the use of three dimensional Cartesian coordinates of molecules for generating atom embedding, and these atomistic coordinates were calculated at Density Functional Theory – a computational demanding theory level for constructing a dataset of thousands molecules.

Ye et al. introduced a combinatorial quantitative structure-activity relationship (QSAR) and machine learning approach to predict the emission wavelength, the experimental measurable qualitatively equivalent to E_{HL}, of 11,460 organic fluorescent molecules against the corresponding experimental measurements.[15] The authors reported the training results of R^2 = 0.663 (MAE = 0.2449 eV) and R^2 = 0.923 (MAE = 0.1253 eV) using the linear and nonlinear (random forest) models, respectively. The emission wavelength prediction was further refined by the inclusion of solvent effect with using 3000 distinct experimentally-recorded compounds.[23] Based upon these aforementioned studies, one can see that the nonlinear models always provide more accurate electronic property predictions than the linear models. However, the corresponding interpretation generated by these property predictions still cannot be straightforwardly presented in terms of the intuitive chemical terminology, and that is due to the complex formulation in these models. Having the pros and cons addressed in the literature, we aim to demonstrate a progressive comparison in terms of model complexity, accuracy of the predictions, and the results interpretability. Such a comparison could provide an insightful perspective to benefit the field of virtual molecular design.

Methods

Sample generation

The present study totally used 132,180 molecules out of QM9 dataset,[39, 40] and these molecules were randomly partitioned as 88560 and 43620 for the training and testing sets (about 2:1 ratio), respectively. The details of molecule selection are noted in the electronic supplementary information (ESI). All of the molecular structures were previously optimized at B3LYP/6-31G(2df,p) level of theory in the vacuum. These molecules contain up to nine heavy atoms, i.e. carbon, oxygen, nitrogen, and fluorine. The molecular electronic structure properties – E_H, E_L, and ΔE_{HL}, being predicted at DFT level are adopted as the target properties. The SMILES files and the DFT optimized Cartesian coordinates were used as the input information for the subsequent machine learning and deep learning NN approaches. The energetics distributions of E_H, E_L, and ΔE_{HL} are schematically shown in Figures 1a-1c. E_H appears to be a symmetric

distribution among these three properties while E_L and ΔE_{HL} generally contains three maximum peaks due to the linear relationship of $\Delta E_{HL} = E_L - E_H$. In Figure 1d, the sorted E_H is plotted with the corresponding E_L, and that distribution appears to suggest the independence between E_H and E_L.

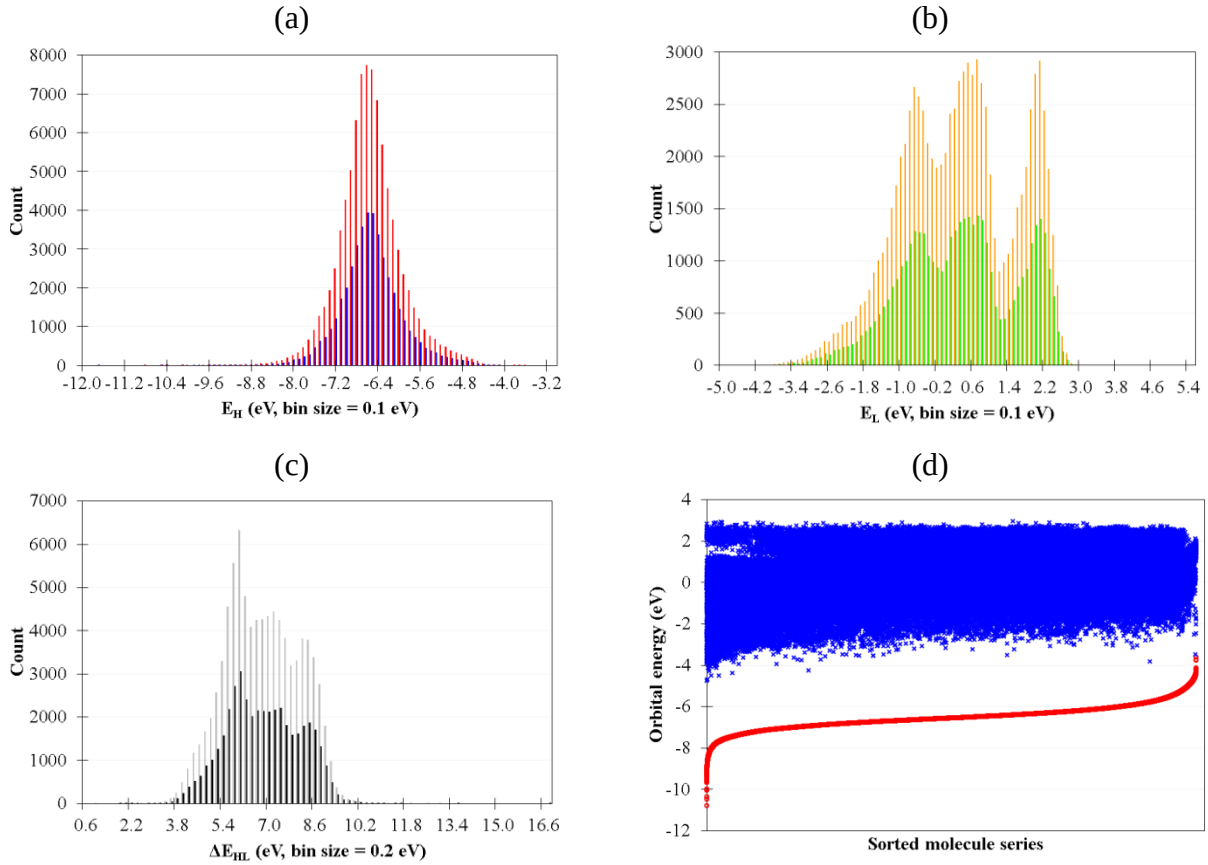


Figure 1. (a-c) E_H, E_L, and ΔE_{HL} distribution of the training and testing sets. (d) The energetic distribution of E_L in respect to the sorted E_H.

Descriptor Generation

The SMILE file of each molecule is used as the input for the descriptor generation using PaDEL,[41] and each molecule is described by 17,957 descriptors including 1444 1D & 2D descriptors, 431 3D descriptors, and 16092 fingerprints. All of these descriptors are, however, considered as the discrete chemical knowledge elements that have been introduced in the numerous early literatures. This descriptor generation scheme requires the interatomic valence-bonding connectivity being represented by SMILES format, not requiring the 3D molecular structure information (3D molecular structures are typically prepared in advance by the empirical

potentials or assessable quantum chemistry calculations). The details for the categories of the descriptors are summarized in Table S1 of ESI.

Mathematical Model Specification

Each linear regression model is solved by ordinary least squares method of sklearn with the default convergence criteria. Each random forest model is consisted of 100 decision tress where each tree could grow up to 15 layers. The minimum number of samples in a leaf is set to 1, and the minimum number of samples in a branch is set to 2. The minimum of impurity is set to 0 for the stop of the branch growth. For the SchNET model, exactly same parameter setup was used as the original report,[38] except the interatomic distance cutoff is set to 10Å .

Results and Discussion

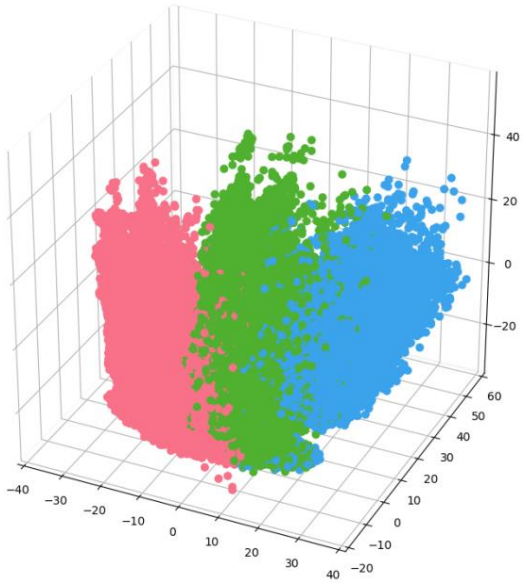
Descriptor dimension reduction

The variance threshold selection (VTS) approach is applied to remove the descriptors containing negligible variance $\sigma^2 < 0.01$, and the descriptor dimension is reduced from 17957 down to 4533. The 4533 descriptors, being denoted as VTS ensemble, include 938 topological descriptors and 3609 fingerprints. In order to reduce the dimension of VTS ensemble, the molecules of training sets were categorized into three distinct sample-subsets using mean shifted clustering method as shown by the color dots in the inset of Table 1. Each sample-subset was treated with the least absolute shrinkage and selection operator (Lasso) regression, being subject to the selection of target properties and the pre-determined penalty parameter, for the extraction of the target-property-dependent representative descriptors from VTS ensemble. For a particular target property, these extracted descriptors from three sample-subsets were emerged together (without double counting the duplicated ones) and formed the final X_LasY descriptor ensembles as summarized in Table 1, where X = H, L, or G denoting the target property – E_H, E_L, or ΔE_{HL} and Y denotes the value of penalty parameter (0.1 or 0.5). The penalty parameter of 0.1 resulted in the dimension of descriptor space with more than 1200 descriptors for all target properties, and the other case – 0.5 penalty could reduce VTS ensemble down to about 500 descriptors.

Table 1. The number of extracted descriptors by Lasso regression from each sample-subset (red, green, and blue) classified by mean-shift clustering.

Subgroups	H_Las01	L_Las01	G_Las01
Red	316	388	447
Green	426	577	593
Blue	741	794	894
[2]Resultant descriptors	1250	1431	1559
Subgroups	H_Las05	L_Las05	G_Las05
Red	119	150	184
Green	139	211	223
Blue	292	261	313
[2]Resultant descriptors	477	531	625

Colored subsets
by mean shifted clustering



[1]H, L, and G labels denote the target properties – E_H, E_L, and ΔE_{HL}, respectively. The Las01 and Las05 labels denote the penalty parameter at 0.1 or 0.5 of Lasso regression. The larger penalty results in a smaller descriptor space.
[2]Duplicated descriptors found in the three sample-subsets were merged.

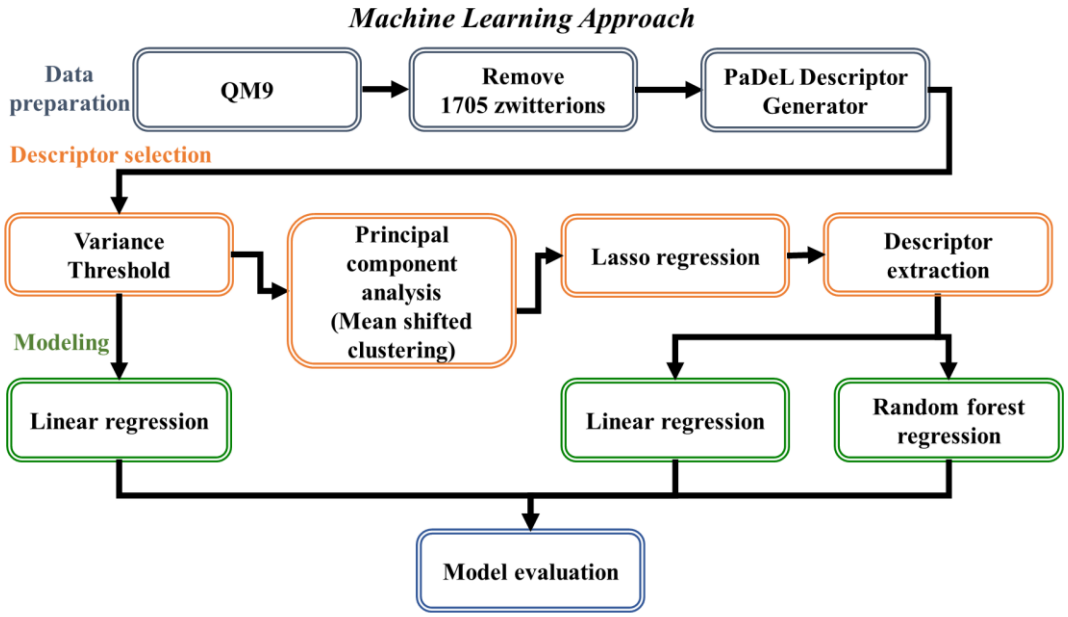
Linear regression (LR) and random forest (RF) methods were applied, being with VTS and X_LasY descriptor ensembles, to predict E_H, E_L, and ΔE_{HL}. The mean absolute errors (MAE) of the testing sets are summarized in Tables 2 and S2, and the corresponding training set results are provided in Table S3. The nonlinear RF models using X_Las05 descriptor ensembles, denoted as X_Las05_RF, appear to provide better MAEs in all three target properties than the corresponding linear models (X_Las05_LR). Current X_Las05_RF models also provide comparable predictivity in respect to the early RF results using extended connectivity fingerprints (ECP4) representation, however, demonstrated higher learning efficiency with using 88k training molecules (vs. 118k molecules for the ECP4_RF model).[7] In Table 2, current linear models labeling as X_Las05_LR appear to be noticeably more accurate than the previous linear models using Elastic Net and Baysian Ridge methods coupled with ECP4 representations.).[7]

Table 2. The mean absolute errors of testing set (in eV) for E_H, E_L, and ΔE_{HL} predictions using the linear regression and random forests models

Models	This study (X_Las05)[1]		Previous study[2]		
	LR	RF	EN	BR	RF
E_H	0.161 (0.88)	0.141 (0.90)	0.224	0.224	0.143
E_L	0.198 (0.96)	0.151 (0.97)	0.344	0.344	0.145
ΔE_{HL}	0.246 (0.93)	0.177 (0.96)	0.383	0.383	0.166

[1]R2 values are reported in parentheses. The present study contains 88560 and 43610 compounds for the training and testing sets, respectively.
[2]The training set used ~118k compounds in reference 7. EN and BR denotes Elastic Net linear model and Baysian Ridge regression model, respectively.

With lowering the penalty parameter to 0.1 during the process of descriptor extraction, one can generate substantially larger descriptor ensembles, being labeled as X_Las01, over X_Las05 as shown in Table 1. Consequently, all X_Las01_LR and X_Las01_RF models provide enhanced predictivity than the corresponding models using X_Las05 descriptor ensembles due to larger degrees of freedom in the descriptor space (see Table S2). By examining the difference between R^2 and Q^2 of the training results in Table S2, all LR and RF models using VTS, X_Las01, and X_Las05 ensembles can be considered statistically meaningful due to the absence of overfitting ($|R^2 - Q^2| < 0.1$).[42]



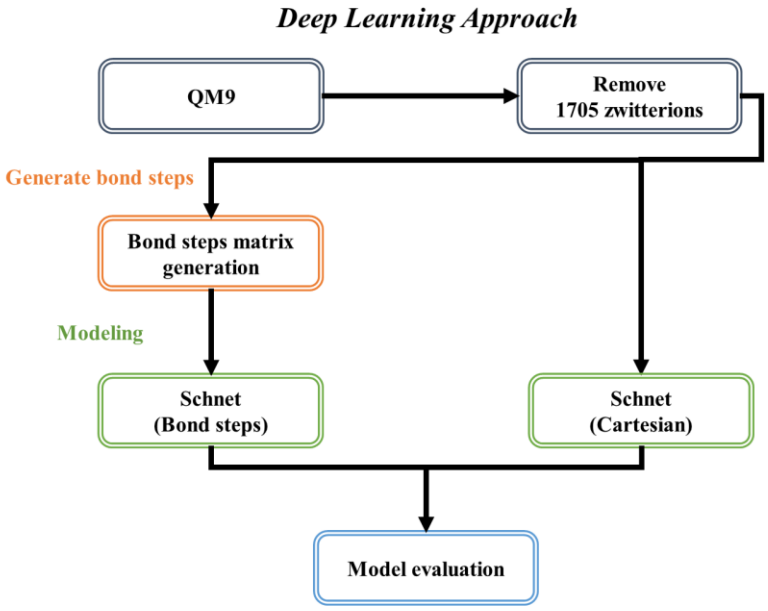


Figure 2. The flow charts of machine learning and deep learning approaches using QM9 dataset

Despite all LR and RF models provide reasonable predictions for E_H, E_L, and ΔE_{HL} of the organic molecules in QM9 dataset, the conventional chemical accuracy was still not reached using these linear and RF approaches. The current LR and RF models can achieve 3-6 times of chemical accuracy in terms of MAEs of E_H, E_L, and ΔE_{HL} predictions. Faber et al. demonstrated that combining NN models with the graph representations, being derived from the three dimensional (3D) Cartesian coordinates, could significantly reduce MAEs of E_H, E_L, and ΔE_{HL} predictions down to less than 0.1 eV.[7] In Table 3, the results reported by Schütt et al. using a deep learning NN model (SchNET) provided the most accurate results for the E_H, E_L, and ΔE_{HL} predictions where MAEs of E_H and E_L were less than 0.043 eV, except that of ΔE_{HL} was at 0.063 eV. In this study, we reproduced the SchNET approach using fewer molecules for the training set (88k vs. the original size of 110k), and the corresponding models (annotated as Schnet-3d) can provide MAEs of E_H, E_L, and ΔE_{HL} at 0.051, 0.041, 0.075 eV, respectively. Only minor deterioration was observed for the performance of these models in comparison with the original case.

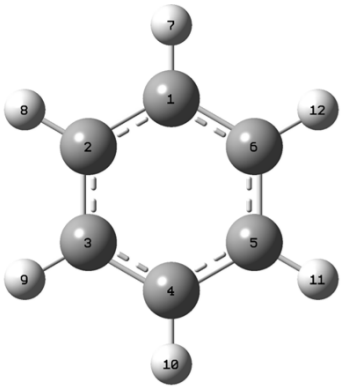
Table 3. The mean absolute errors of testing set (in eV) for E_H, E_L, and ΔE_{HL} predictions using SchNET deep learning model with graph and bond-steps representations

Models	This study[1]		Early studies			
	Schnet		GGNN[2]	GCNN[2]	MPNN[3a]	Original SchNET[3b]
	3D	Bond-step	3D	3D	3D	3D

E_H	0.051 (0.99)	0.090 (0.96)	0.057	0.055	0.043	0.041
E_L	0.041 (1.00)	0.088 (0.99)	0.063	0.062	0.037	0.034
ΔE_{HL}	0.076 (0.99)	0.125 (0.98)	0.088	0.087	0.069	0.063

[1] R^2 values are reported in parentheses. The present study contains 88560 and 43610 compounds for the training and testing sets, respectively.
[2] The training set used ~118k compounds of QM9 dataset in reference 7. GG and GC denote gated graph neural network and graph convolutional neural network models, respectively.
[3] Both models used training set of ~110k molecules of QM9 dataset in reference 41 (MPNN) and reference 38 (SchNET).

In order to reduce the computational expense in dataset preparation, SchNET model combining a bond-step representation is introduced in this study (annotated as Schnet-bs models), where the interatomic distances of all molecules in QM9 dataset are replaced by the bond counting rule. The interatomic bond-step matrixes can be directly generated from SMILES files without pre-determined three-dimensional molecular coordinates by classical or quantum mechanics simulations. A schematic presentation of bond-step representation is demonstrated by the example of C_6H_6 in Figure 3. The Schnet-bs models appear to provide better predictivity than all X_LasY_RF models, reaching the accuracy close to 0.1 eV levels, however, generating about two-times larger MAEs in respect to the original SchNET model. The schematic representations for systematically comparing X_Las05_LR, X_Las05_RF, Schent-3d, and Schnet-bs models are summarized in Figure 4.



C6H6	X	Y	Z
C1	0.000	1.395	0.000
C2	-1.208	0.697	0.000
C3	-1.208	-0.697	0.000
C4	0.000	-1.395	0.000
C5	1.208	-0.697	0.000
C6	1.208	0.697	0.000
H7	0.000	2.465	0.000
H8	-2.135	1.232	0.000
H9	-2.135	-1.232	0.000
H10	0.000	-2.465	0.000
H11	2.135	-1.232	0.000
H12	2.135	1.232	0.000

	C1	C2	C3	C4	C5	C6	H7	H8	H9	H10	H11	H12
C1	0.000											
C2	1.395	0.000										
C3	2.416	1.395	0.000									
C4	2.790	2.416	1.395	0.000								
C5	2.416	2.790	2.416	1.395	0.000							
C6	1.395	2.416	2.790	2.416	1.395	0.000						
H7	1.070	2.141	3.385	3.860	3.385	2.141	0.000					
H8	2.141	1.070	2.141	3.385	3.860	3.385	2.465	0.000				
H9	3.385	2.141	1.070	2.141	3.385	3.860	4.269	2.465	0.000			
H10	3.860	3.385	2.141	1.070	2.141	3.385	4.930	4.269	2.465	0.000		
H11	3.385	3.860	3.385	2.141	1.070	2.141	4.269	4.930	4.269	2.465	0.000	
H12	2.141	3.385	3.860	3.385	2.141	1.070	2.465	4.269	4.930	4.269	2.465	0.000

	C1	C2	C3	C4	C5	C6	H7	H8	H9	H10	H11	H12
C1	0											
C2	1	0										
C3	2	1	0									
C4	3	2	1	0								
C5	2	3	2	1	0							
C6	1	2	3	2	1	0						
H7	1	2	3	4	3	2	0					
H8	2	1	2	3	4	3	3	0				
H9	3	2	1	2	3	4	4	3	0			
H10	4	3	2	1	2	3	5	4	3	0		
H11	3	4	3	2	1	2	4	5	4	3	0	
H12	2	3	4	3	2	1	3	4	5	4	3	0

Figure 3. Schematic comparison of interatomic distance (black) and bond-step (red) matrixes. The empty matrix elements are internally omitted due to the symmetric nature of these matrixes. The original Cartesian coordinates of C_6H_6 are in Å .

Interplay of ΔE_{HL} and real-world emission wavelength

Chemical-intuitively, ΔE_{HL} can qualitatively align with the photon emission energies of fluorescent molecules. Ye et al. reported a RF model combining the knowledge-based molecular representations from PaDEL and predicted the experimental emission wavelengths of 11k organic fluorescent molecules where MAE(testing) and R^2(testing) were reported at 0.222 eV and 0.70, respectively (Table S4) without taking into account any solvent description. Without the necessity of describing solvation effect for QM9 dataset, current G_Las05_RF model improves the predictivity in ΔE_{HL} of QM9 molecules with MAE(testing) and R^2(testing) at 0.177 eV and 0.96, respectively. This enhancement suggests that RF models can still provide reasonable predictivity and learning efficiency if the necessary information had been included in the descriptor space. However, the boundary of descriptor space needs to be pre-determined in according to mankind's domain knowledge. The importance of solvent descriptor for successfully predicting the experimental optical properties has been demonstrated by the recent machine learning and deep learning approaches.[23, 43]

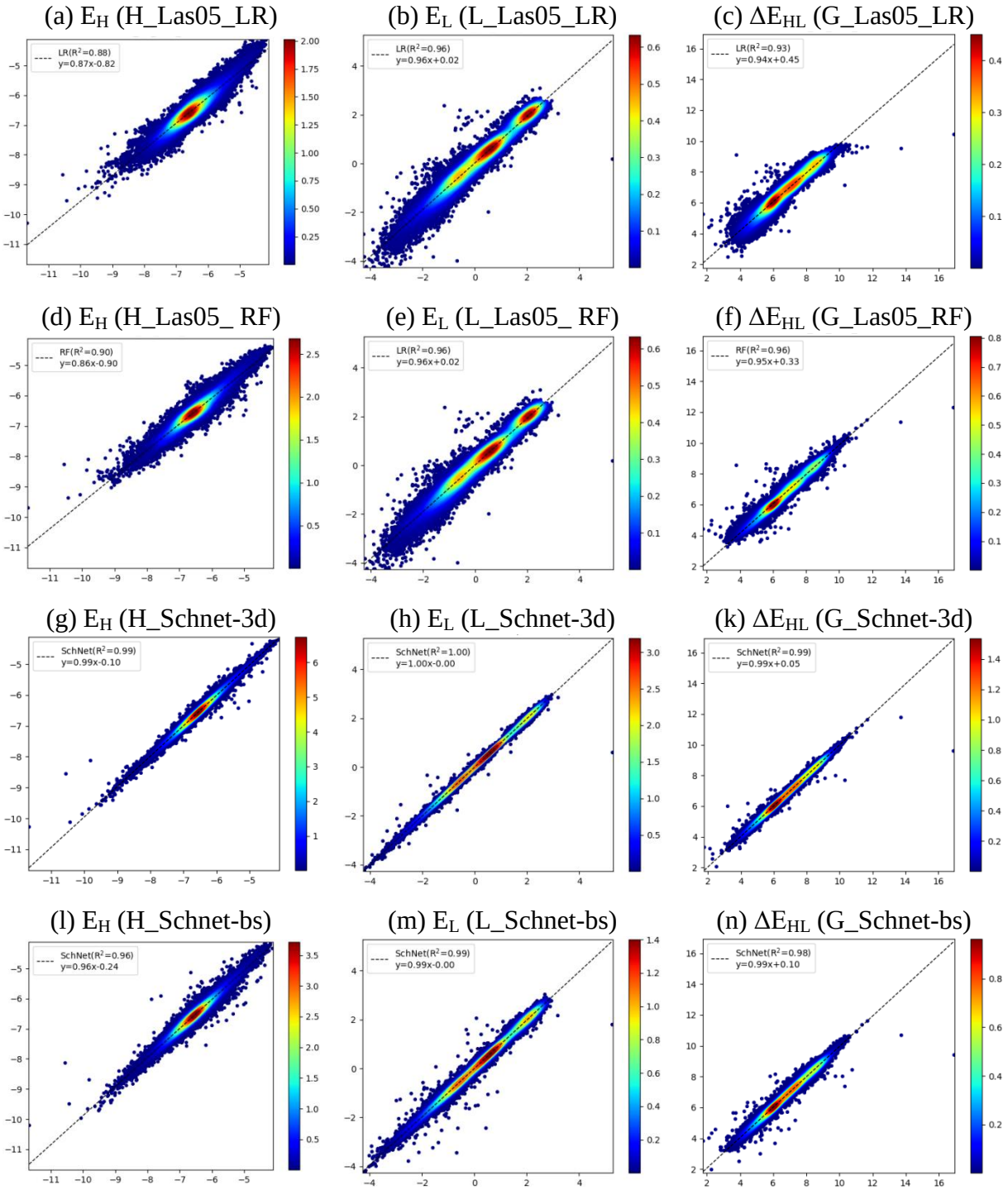


Figure 4. The prediction vs. reference comparison using the linear (a-c), and random forest X_Las05 models (d-f), Schnet-3d(g-k), and Schnet-bs(l-n) for predicting testing sets of E_H, E_L, and ΔE_{HL}, respectively. The corresponding linear-fitting equations of predicted-vs-reference are noted in the insets.

Chemical Interpretability

Despite the apparent success of deep learning models in predicting the molecular electronic properties as shown above, the chemical essence embedded in the complex neural networks is still generally not easy to interpret. Comparatively, the important descriptors adopted in RF models can be described by Gini importance – the number of a particular descriptor used to split a node, being weighted by the number of the corresponding samples. Higher Gini scores imply more presence of the descriptor determining the classification of samples. Pubchem FP416 denotes the presence of C=C fingerprint in the sample molecules, and that straightforwardly represents the existence of π-orbitals, resulting in increasing HOMO energy, between similar saturated and unsaturated molecular frameworks. SubFPC300 represents the count of 1,3-tautomerizable conjugation, adding more subtle criteria for describing conjugated molecular structures. GATS1c denotes Geary autocorrelation weighted by atomic charges at topological distance of 1.

For the linear dependent E_L and ΔE_{HL}, both L_Las05_RF and G_Las05_RF models independently identified nAtomP and R_TpiPCTPC as the important descriptors, the former one denoting the numbers of atoms in the largest π system and the latter denoting the ratio of total conventional bond order with total path count. These are straightforwardly consistent with the conventional chemical knowledge that a larger π system can result in lower E_L energy. ETA_Beta_ns denotes the measure of electron-richness of the molecule. SubFPC287, as a similar descriptor like SubFPC300 identified for E_H prediction, denotes the count of conjugated double bonds and poses a contribution of E_H to the ΔE_{HL} predictions.

The chemical interpretability may be discretely represented by the use the interpretable descriptors as listed in Table 4. However, the interplay between these descriptors, being represented in the numerous tree structures, is too complicated to formulate analytically. Figures S1-S3 summarizes the top 50 descriptors of X_Las05_LR models where substantial cancellation effects are observed in these readable quantitative formulations. For the most accurate deep learning models, the underlying networks are even more challenging to connect with the conventional chemical knowledge.

Table 4. The feature importance of X_Las05_RF models for E_H, E_L, and ΔE_{HL} predictions[1]

Property	E_H	E_L	ΔE_{HL}
Feature importance	PubchemFP416 (0.191)	nAtomP (0.199)	nAtomP (0.310)
	SubFPC300 (0.083)	ETA_Beta_ns (0.197)	R_TpiPCTPC (0.239)

GATS1c (0.079)	R_TpiPCTPC (0.149)	SubFPC287 (0.083)

[1]Feature importance is estimated by Gini importance.

Conclusion

In this study, we demonstrated a fine balance between chemical property predictivity and chemical knowledge interpretability using linear, random forest, deep learning models. We applied a systematic approach to extract a few hundred critical knowledge-based molecular descriptors, from 17k-plus descriptors generated by PaDEL, using the linear and random forest models for predicting E_H, E_L, and ΔE_{HL} of the organic molecules in QM9 dataset. Such a simple approach only requires the valence bond connectivity of the molecules, being encrypted in SMILES files, and does not require molecular structure optimizations, commonly being achieved by empirical potentials or quantum chemistry calculations. The predictivity provided by the current linear and random forest models can achieve MAEs at 4-6 times of chemical accuracy while the qualitative chemical interpretation, being represented by the identification of the leading descriptors, can be extracted from these numerical formulations.

The MAEs of E_H, E_L, and ΔE_{HL} predictions can be reduced to generally match the level of chemical accuracy if the complex deep learning model – SchNET is adopted. Nonetheless, the success of SchNET model builds upon the collection of physical-meaningful molecular structures (Cartesian coordinates), being typically optimized by quantum chemistry calculations. With replacing the molecular Cartesian coordinates by bond-step matrixes, MAEs of SchNET(bs) models are about 2-times of chemical accuracy in comparison with those of SchNET(3D) cases. Such a moderate demotion of model predictivity may be worthwhile compensated by the reduced workload of dataset preparation. This SchNET(bs) approach is finally recommended to couple with chemical graph generators for the virtual screening of new organic molecules containing novel properties.

Conflicts of interest

There are no conflicts to declare.

Electronic Supplementary Information

Electronic Supplementary Information (ESI) available: All of the descriptor categories from PaDEL, details of other linear and random forest models, and summary of leading descriptor information. See DOI: 10.1039/x0xx00000x.

Acknowledgements

This study is supported by the Ministry of Science and Technology of Taiwan (110-2113-M-003 -015 and 110-2124-M-003 -001) and the Innovation-Oriented Trilateral Research Fund for Young Investigators of NTU system. The authors are grateful for the computational resources provided by the National Center for High-Performance Computing of Taiwan and the Center for Cloud Computing in National Taiwan Normal University.

Reference

1. O. A. von Lilienfeld, *Int. J. Quan. Chem.*, 2013, 113, 1676-1689.
2. K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, 6, 2326-2331.
3. R. Ramakrishnan and O. A. von Lilienfeld, *CHIMIA*, 2015, 69, 182-186.
4. T. D. Huan, A. Mannodi-Kanakkithodi and R. Ramprasad, *Phys. Rev. B*, 2015, 92.
5. E. O. Pyzer-Knapp, K. Li and A. Aspuru-Guzik, *Adv. Func. Mater.*, 2015, 25, 6495-6502.
6. N. J. Browning, R. Ramakrishnan, O. A. Von Lilienfeld and U. Roethlisberger, *J. Phys. Chem. Lett.*, 2017, 8, 1351-1359.
7. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. Von Lilienfeld, *J. Chem. Theory Comput.*, 2017, 13, 5255-5264.
8. F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang and J. Aires-de-Sousa, *J. Chem. Inf. Model.*, 2017, 57, 11-21.
9. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, 4, 268-276.
10. J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik and Y. Jung, *Matter*, 2019, 1, 1370-1384.
11. R. S. da Silva, L. F. Marins, D. V. Almeida, K. dos Santos Machado and A. V. Werhli, *Comput. Bio. Chem.*, 2019, 83, 107089.
12. K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari and P. Rinke, *Adv. Sci.*, 2019, 6, 1970053.
13. L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, 247, 106949.
14. K. Rossi and J. Cumby, *Int. J. Quan. Chem.*, 2020, 120, e26151.
15. Z.-R. Ye, I. S. Huang, Y.-T. Chan, Z.-J. Li, C.-C. Liao, H.-R. Tsai, M.-C. Hsieh, C.-C. Chang and M.-K. Tsai, *RSC. Adv.*, 2020, 10, 23834-23841.
16. G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva and M. G. Quiles, *J. Phys. Chem. A*, 2020, 124, 9854-9866.
17. W. A. Saidi, W. Shadid and I. E. Castelli, *Npj Comput. Mater.*, 2020, 6, 36.
18. R. Nagai, R. Akashi and O. Sugino, *Npj Comput. Mater.*, 2020, 6, 43.
19. S. Kiyohara, M. Tsubaki and T. Mizoguchi, *Npj Comput. Mater.*, 2020, 6, 68.
20. A. Dunn, Q. Wang, A. Ganose, D. Dopp and A. Jain, *Npj Comput. Mater.*, 2020, 6, 138.
21. N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler and S. P. Russo, *Npj Comput. Mater.*, 2020, 6, 166.
22. C.-I. Wang, I. Joanito, C.-F. Lan and C.-P. Hsu, *J. Chem. Phys.*, 2020, 153, 214113.
23. C.-W. Ju, H. Bai, B. Li and R. Liu, *J. Chem. Inf. Model.*, 2021, 61, 1053-1065.
24. K. L. Woon, Z. X. Chong, A. Ariffin and C. S. Chan, *J. Mol. Graph Model.*, 2021, 105, 107891.
25. W. F. Reinhart, *Comput. Mater. Sci.*, 2021, 196, 110511.
26. E. T. Chenebuah, M. Nganbe and A. B. Tchagang, *Mater. Today Commun.*, 2021, 27, 102462.

27. P. Omprakash, B. Manikandan, A. Sandeep, R. Shrivastava, V. P and D. B. Panemangalore, *Comput. Mater. Sci.*, 2021, 196, 110530.
28. J. T. Margraf and K. Reuter, *Nat. Commun.*, 2021, 12, 344.
29. Y. Wan, F. Ramirez, X. Zhang, T.-Q. Nguyen, G. C. Bazan and G. Lu, *Npj Comput. Mater.*, 2021, 7, 69.
30. V. Fung, G. Hu, P. Ganesh and B. G. Sumpter, *Nat. Commun.*, 2021, 12, 88.
31. A. Ihalage and Y. Hao, *Npj Comput. Mater.*, 2021, 7, 75.
32. J. Lan, V. Kapil, P. Gasparotto, M. Ceriotti, M. Iannuzzi and V. V. Rybkin, *Nat. Commun.*, 2021, 12, 766.
33. C. E. Belle, V. Aksakalli and S. P. Russo, *J. Cheminformatics*, 2021, 13, 42.
34. Z. Wang, S. Ye, H. Wang, J. He, Q. Huang and S. Chang, *Npj Comput. Mater.*, 2021, 7, 11.
35. T. W. Ko, J. A. Finkler, S. Goedecker and J. Behler, *Nat. Commun.*, 2021, 12, 398.
36. M. Arrigoni and G. K. H. Madsen, *Npj Comput. Mater.*, 2021, 7, 71.
37. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, presented in part at the Proceedings of the 34th International Conference on Machine Learning - Volume 70, Sydney, NSW, Australia, 2017.
38. K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, 148, 241722.
39. R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Scientific Data*, 2014, 1, 140022.
40. L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, 52, 2864-2875.
41. C. W. Yap, *J. Comput. Chem.*, 2011, 32, 1466-1474.
42. N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2012, 52, 2044-2058.
43. J. F. Joung, M. Han, J. Hwang, M. Jeong, D. H. Choi and S. Park, *JACS Au*, 2021, 1, 427-438.