ACS Sustainable Chemistry & Engineering

pubs.acs.org/journal/ascecg

Letter

Accurate Thermochemistry of Complex Lignin Structures via Density Functional Theory, Group Additivity, and Machine Learning

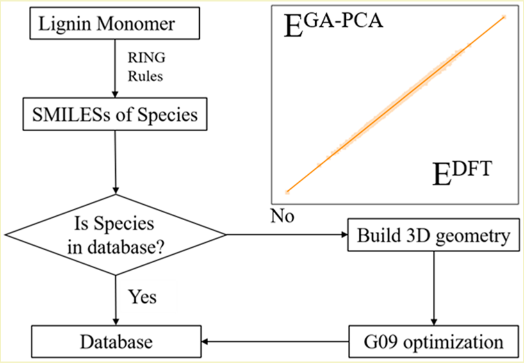
Qiang Li, Gerhard Wittreich, Yifan Wang, Himaghna Bhattacharjee, Udit Gupta, and Dionisios G. Vlachos*

Cite This: *ACS Sustainable Chem. Eng.* 2021, 9, 3043−3049

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A molecular-level understanding of lignin structures and bond dissociation energies could facilitate depolymerization technologies. Still, this information is currently limited due to the lack of databases and the simplification of surrogate models. Here, substitution effects on seven common linkages in lignin polymers are systematically investigated. An automated reaction network generator is employed to create a database of structures. A new group additivity (GA) model based on principal component analysis (PCA) descriptors is introduced and trained on gas-phase density functional theory data of 4100 species at the M06-2X/6-311++G(d,p) level. Hydrogen bonds, local steric, and nonaromatic ring contributions are also incorporated. Finally, we improve the accuracy of the group additivity model to reach the G4 theory by computing a data set of 770 species at this level and using a data fusion approach.



KEYWORDS: *Lignin, Database, Group additivity, Principal component analysis, G4 theory*

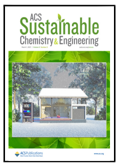
■ INTRODUCTION

Lignin accounts for 15%−30% in weight and 40% of the energy of lignocellulose[1] and is responsible for the structural rigidity of plant cell walls. Utilization of lignin is still challenging, and only less than 2% of lignin from the pulp and paper industry is used for high-value chemicals, with most lignin burned for heat generation.[2] Among several depolymerization technologies,[3] pyrolysis is widely applied at high temperatures to break the C−C and C−O bonds and make bio-oils for further upgrade.[4] Bio-oil is the main product in lignin pyrolysis and contains numerous compounds. However, the product complexity and low selectivity of pyrolysis limit its commercial potential.[5,6] A fundamental understanding of the mechanisms in lignin pyrolysis can guide the design of optimal conditions and simplify the downstream catalytic upgrade. For this purpose, density functional theory (DFT) calculations can be used. However, a challenge in modeling lignin is the combinatorics arising from its amorphous structure that makes periodic calculations inapplicable and small structures insufficient.

Lignin is a cross-linked, amorphous polymer formed by phenolic monomers. Unlike cellulose, whose structure consists of repeated C6 sugar units, lignin is polydisperse due to the various linkages, such as α−O−4, β−O−4, 4−O−5, β−1, β−5, β−β, and 5−5 (Figure 1). The three monolignols, namely, *p*-coumaryl (H), coniferyl (G), and sinapyl alcohol (S), can be linked at different positions (C$^{\alpha}$ and C$^{\beta}$ at the aliphatic tail and 4 and 5 in the aromatic unit).

The bond dissociation energy (BDE) is the crucial descriptor for understanding the bond strengths in various linkages. Previous studies using linear linkage models[7−9] suggest that the BDEs follow the order of Aryl(Ar)−Ar > Ar−C$^{\alpha}$ > α−O−4 > β−O−4. BDEs for linkages with intramolecular rings, such as β-1, β-β, and benzodioxane, have been investigated by Elder and co-workers.[10−12] Relatively small dimer structure models are mostly employed to calculate BDEs by DFT. Due to lignin structures and conformers' combinatorial complexity, the BDEs vary widely even for one specific type of linkage.[13] However, the structure models for different linkages are usually simplified by replacing the functional groups attached to either the aromatic (−OCH$_3$) or aliphatic tails (−OH), with H atoms. In Figure 1, the α−O−4, β−O−4, 4−O−5, β−1(5), β−β, and 5−5 linkage models can be represented by benzyl phenyl ether, phenethoxybenzene, diphenyl ether, 1,2-diphenylethane, 1,4-diphenylbutane, and biphenyl molecules when all the R groups are H atoms. A systematic study on the substitution effects is crucial for understanding how the bond strength is affected by

Received: December 5, 2020
Revised: January 27, 2021
Published: February 15, 2021



ACS Publications

https://dx.doi.org/10.1021/acssuschemeng.0c08856

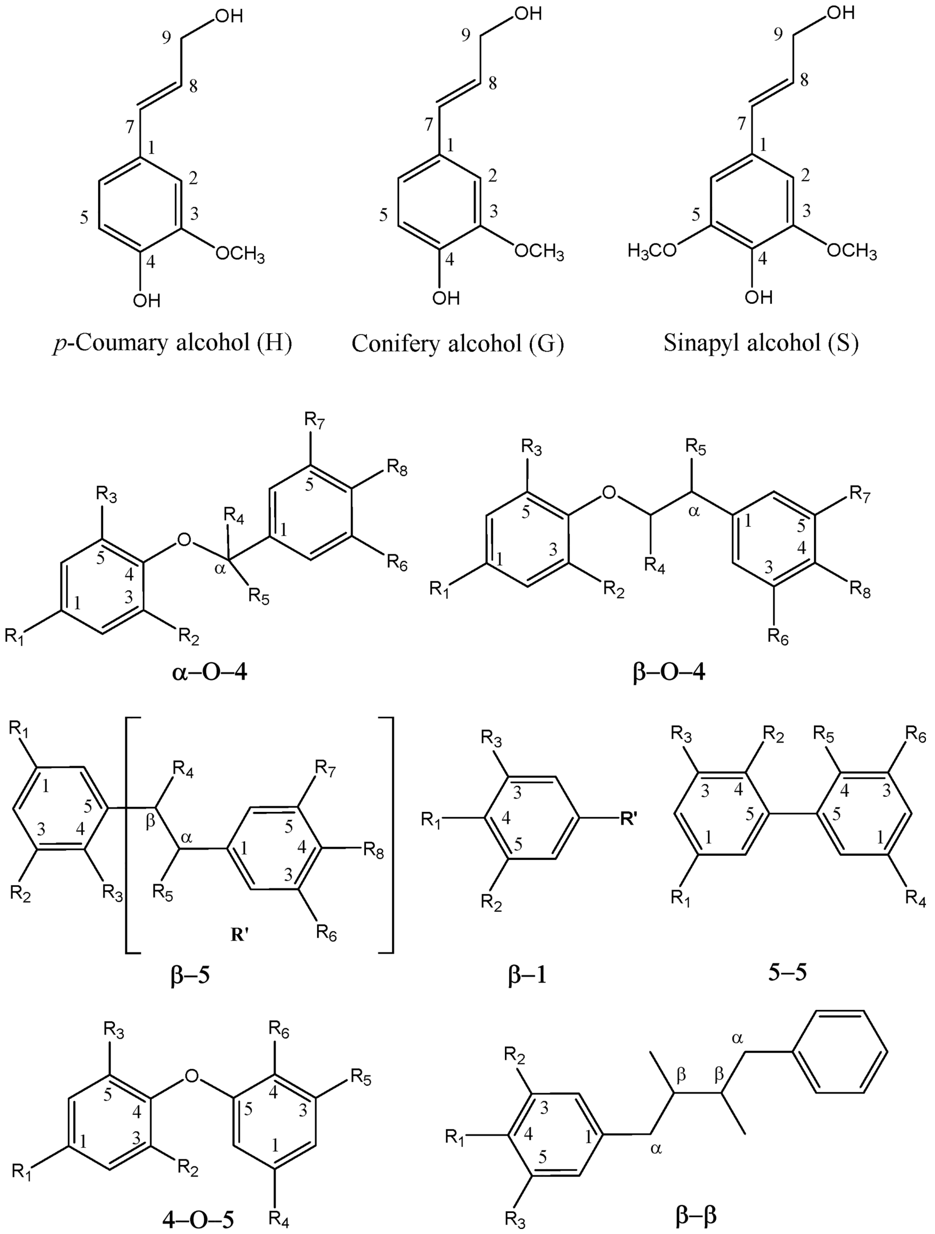


Figure 1. Structures of monolignols and typical lignin linkages.

various groups and the potential products made from different lignins.This work fills in this gap.

The computational cost increases significantly with the complexity of linkage models, and thus, a trade-off balancing the two is needed. Group additivity (GA), a method pioneered by Cohen and Benson,[14] offers the opportunity to predict the thermochemical properties of compounds based on groups whose thermochemistry values are tabulated in a database. As shown in eq 1, the property y of a target molecule is the sum of the group additivity values (GAVs) of all the groups in the molecule. The groups describing the molecular structure are identified using, for example, graph theory, the one heavy atom (C, O, N, etc.)-centered substructure from Cohen and Benson's scheme and LASSO-based subgraphs.[15] Their GAVs are estimated from a small training data set obtained from DFT calculations and/or available experimental values.

$$y = \sum_{1}^{i} n_i \mathrm{GAV}_i \tag{1}$$

Due to the large abundance of the same groups in the lignin structures, the GA method is ideal for predicting thermochemistry at a low computational cost. Extensive studies on the application of GA to gaseous hydrocarbons, oxygenates, and their radicals have been undertaken by Marin and coworkers.[16,17] GA of monocyclic aromatic species with different substituted functional groups has also been developed.[18,19] However, a comprehensive GA method for estimating lignin thermochemistry is not available due to the lack of a rich database that describes the structural complexity of lignin. Given the complexity and large size of the structures, a large abundance of species is required for constructing such a database. The effect of steric and weak interactions (e.g., hydrogen bonds) becomes progressively more important with increasing model size and cannot be ignored. Furthermore, the computational cost of DFT calculations, even for gas-phase molecules, is enormous, limiting the application of high accuracy methods.

In the present work, DFT calculations at the M06-2X/6-311++G(d,p) level have been performed to study the steric and hydrogen bond effects from the substituted groups on the BDEs and geometry. An automated reaction network generator was first employed to create a database containing 4100 lignin-related molecules and radicals. The database was then transformed using principal component analysis and was used to train a GA model. The GA model's accuracy was further improved by implementing steric and hydrogen bond effects. Finally, we compute a subset of structures at the G4 theory level[20] to enhance the entire data set accuracy using a data-fusion approach. Combined with the linear relationships, such as Brønsted−Evans−Polanyi,[21,22] thermochemical properties estimated from the GA model can be used to predict activation barriers in lignification[23,24] and pyrolysis,[25,26] and to expand our understanding of this complex biopolymer.

■ METHODS

DFT calculations were performed using Gaussian 09[27] with the M06-2X hybrid exchange-correlation functional,[28] which is suitable for calculating BDEs of lignin-related species.[7,8,29,30] Initial geometries from the literature[7,8] were optimized using the 6-31G(d,p) basis set. A larger basis set, 6-311++G(d,p), was then used for refining these structures and verifying their stability via frequency calculations with ultrafine integration grids. Zero-point energy and thermal corrections were added to compute the homolytic bond dissociation energies using eq S1. The thermodynamic properties at different temperatures were calculated using the Python Multiscale Thermochemistry Toolbox (pMuTT)[31] with a frequency scaling factor of 0.970.[32]

As shown in Figure 2, we constructed the database containing aliphatic alkanes, aromatic groups, and their radicals. The database was automatically populated using the Rule Input Network Generator (RING), discussed in Section S2.[33] In addition, $C_xH_yO_z$ species reported from the literature[7,8,29,30,34,35] were also added. The simplified molecular input line-entry system (SMILES) notation,[36] based on ASCII strings, was used to label the species. The species generated using RING and taken from the literature were converted to 1000 initial 3D coordinates using the SMILES-based methods through Open Babel[37] and Rdkit[38] and then optimized using the MMFF[39] force field. The one with the lowest energies were chosen for further optimization. H-bond corrections to the geometries, using the PM6-D3H4 method[40] in the MOPAC2016 package,[41] were employed to correct the H-bond local geometries in the initial structures. These geometries were then optimized at the M06-2X/6-311++G(d,p) level and confirmed using frequency calculations at the same theoretical level. 770 small size species were then taken for G4 calculations.

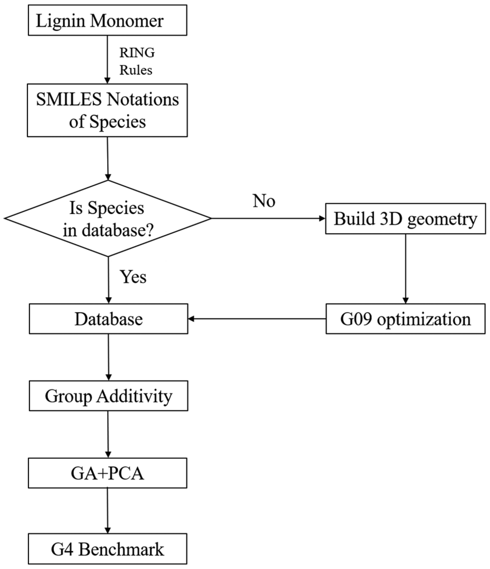


Figure 2. Computational framework for database generation and group additivity model training.

For the group additivity (GA) model, Cohen and Benson's scheme[14] was employed, and Python group additivity (pGrAdd) software[42] developed by our group was used to analyze the group information (molecular substructures and numbers of their appearances) from the species SMILES and create the configuration matrix. As shown in eq 2, the n_s^m and GAV^m stand for the number of times group *m* appears in species *s* and its GAV, respectively, and y_s is the DFT-derived thermal property (e.g., enthalpy of formation) of species *s*.

$$\begin{pmatrix} n_1^1 & \cdots & n_1^m \\ \vdots & \ddots & \vdots \\ n_s^1 & \cdots & n_s^m \end{pmatrix} \begin{pmatrix} \mathrm{GAV}^1 \\ \vdots \\ \mathrm{GAV}^m \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix} \tag{2}$$

Due to linearly dependent groups, the configuration matrix[16] is often underdetermined; i.e., the matrix rank is less than the total number of groups, leading to nonunique regressed GAVs. Collapsing groups with structural similarity and combining linearly dependent groups are standard methods to overcome this problem. However, these methods rely on analyzing the entire set and are impractical when the database size is enormous. In this work, we employ principal components analysis (PCA) to reduce the dimensionality and transform the groups into independent orthogonal PCA vectors via the Scikit-learn[43] Python package. As shown in eq 3, each PC is a vector that contains the group contributions (coefficient values), a_i^1, a_i^2, a_i^3, ..., a_i^m. For one PC in species *s* (PC_s^i), its number vector can be obtained via the linear combination of group numbers and their coefficient values as described in eq 4. The thermal property of species *s* (y_s) is the sum of the contributions from all PCs.

$$\mathrm{PC}_i = a_i^1, a_i^2, a_i^3, ..., a_i^m \tag{3}$$

$$n_{\mathrm{PC}_s^i} = a_i^1 \times n_s^1, a_i^2 \times n_s^2, a_i^3 \times n_s^3, ..., a_i^m \times n_s^m \tag{4}$$

$$y_s = \sum_1^i n_{\mathrm{PC}_s^i} \mathrm{GAV}_{\mathrm{PC}_i} \tag{5}$$

The rule of choosing number of PCs (N_{PC}) was established in four test models using the maximum positive (MPE), minimum negative (MNE), mean absolute (MAE), and root-mean-square errors (RMSE) to evaluate the prediction accuracy with different N_{PC} (Section S4.1). The hydrogen bond and steric corrections to the thermal properties were then added to the GA model (Section S4.2). To evaluate the predictive ability of the GA-PCA model, 10-fold cross-validation based on the bootstrap sampling method was performed (Figure S8). The data set was shuffled, and 90% of the data set containing the same number of groups as the whole database was used for training the GA-PCA model and the rest 10% for validation. This procedure was repeated 100 times, and the accuracy was assessed through the average RMSE (Section S4.3).

The substitution effects on the BDEs were investigated by replacing the R groups (Figure 1) with the corresponding functional groups in the monolignol structures. To further improve the accuracy of the GA-PCA model, weak interactions such as hydrogen bonds and local steric effects were introduced to the group additivity model using the SMARTS[44] description.

A subset calculated by the G4 method was used as a benchmark to enhance the data accuracy via a linear atomistic data fusion approach.[45,46] As shown in eq 6, the property differences (δH_f) of

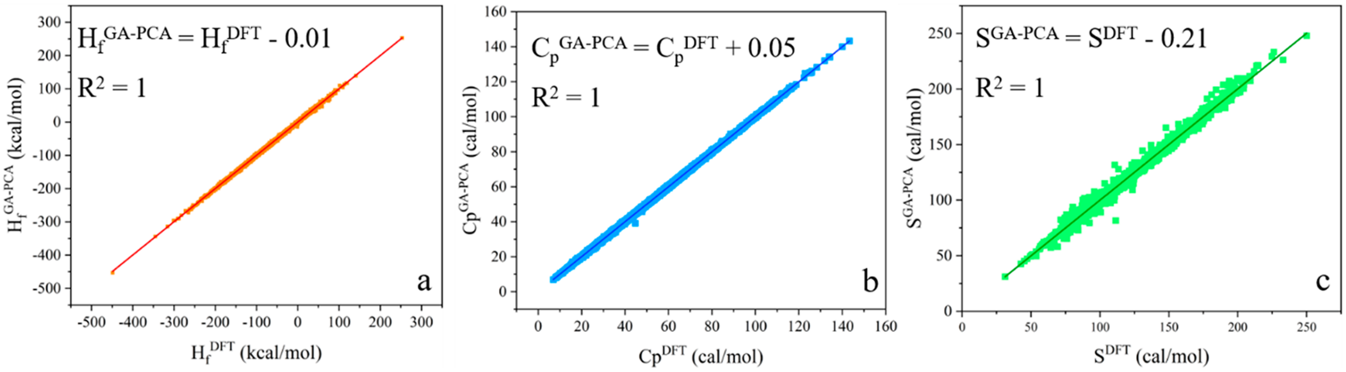


Figure 3. DFT and GA-PCA predicted (a) standard enthalpy of formation ($\Delta_f H^{298.15}$), (b) heat capacity ($C_p^{298.15}$), and (c) entropy ($S^{298.15}$).

M06-2X and G4 were ascribed to errors stemming from the elements without considering larger groups.

$$\delta H_f = H_f^{G4} - H_f^{M06\text{-}2X} = \delta H_f^{C} \times n_C + \delta H_f^{H} \times n_H + \delta H_f^{C} \times n_O \tag{6}$$

H_f^{G4} and $H_f^{M06\text{-}2X}$ are the enthalpies of formation computed using the G4 and M06-2X methods, and n_C, n_H, and n_O are the number of atoms in the species. Here, δH_f^C, δH_f^H, and δH_f^O are the thermal corrections to the M06-2X data for C, H, O, respectively. Details about the thermal correction regressions and the accuracy are described in Section S5.

■ RESULTS AND DISCUSSION

The common linkages in the lignin structures are depicted in Figure 1. The BDEs for the various bonds in these template models are summarized in Figure S2 and Table S1 and follow the sequence of Ar−Ar > (Ar−C$^\alpha$ ∼ Ar−C$^\beta$) ∼ Ar−O (with the O connecting to C$^\alpha$ or C$^\beta$) > (C$^\beta$−C$^\beta$ ∼ C$^\beta$−C$^\gamma$ ∼ Ar−O) > (C$^\beta$−O^4 ∼ C$^\alpha$−C$^\beta$) > C$^\alpha$−O^4. The complete list of substitution functional groups is tabulated in Tables S3−10, as described in Section S3. The substitutions can be at the *ortho*, *meta*, *para*, or the aliphatic tail positions. For α−O−4 and β−O−4, the *ortho* substitutions (R2 and R3) by −OCH$_3$ have the most considerable effect on reducing the BDEs for Ar−O^4 and C$^\alpha$−O^4 bonds by 4.0 and 8.0 kcal/mol, respectively. These values are close to the previously reported values of 6 and 8 kcal/mol by Parthasarathi et al.[29] It should be emphasized that the number of −OCH$_3$ (from 0 to 2) at the R2 and R3 positions represents the lignin source. The −OCH$_3$ substitutions at the *ortho* positions also reduce the Ar−O bond energy in the 4−O−5 linkage by 4.0 kcal/mol per −OCH$_3$ replacement. Substitution effects at R2 and R3 by −OH are also considered since the −OCH$_3$ can be converted to −OH via demethylation. The H-bonds formed via −OH and O^4 reduce the BDE of the C$^\alpha$−O^4 bond by ∼4.0 kcal/mol per H-bond. For the aliphatic tail between two aromatic units, R4 substitutions with −CH$_2$OH and −CHOHCH$_2$OH increase the C$^\alpha$−O^4 bond strength by 3.0−5.6 kcal/mol, while replacing −H with −OH at R5 can reduce its BDE by 2.0 kcal/mol. The −OH effect on lowering the adjacent bond strength is also evident in the β−1(5) and β−β linkages. The 5−5 bond is robust against substitutions at all *meta* positions (R1, R3, R4, and R6) and *ortho* positions (R2 and R5) by −OCH$_3$, −OCH$_2$CH$_3$, and −O−Ph groups. The largest BDE (121.7 kcal/mol) of the 5−5 bond occurs when a H-bond forms between the −OH groups at the R2 and R5 positions. The substitution effects discussed above are then introduced to the following additivity method to improve the prediction accuracy.

The database containing 4100 molecules and radicals was used to generate the configuration matrix through pGrAdd (Excel file in SI zipped file). The matrix is underdetermined with 235 groups and a rank of 216. The PCA method transforms the data into principal components (PCs) to overcome this problem. The additivity values of these principal components (GAVs) for the enthalpy and entropy of formation of species at room temperature were regressed and then applied to predict the thermal properties of species in the database. As shown in Figure S5, in the four tests using data sets with 500, 1000, 1500, and 2000 species, the accuracy of the GA-PCA model kept increasing when the number of principal components, N(PC), is equal to or larger than the rank of the configuration matrix. Clearly, the system is undertermined. We have taken 216 PCs for describing the system with 235 groups (Figure S6). For a new molecule, the groups and their frequencies can be determined, and its thermochemistry can be expressed as a linear combination of of the PCs. Comparison between the DFT and GA-PCA predicted standard enthalpies of formation at room temperature ($\Delta_f H^{298.15}$) shows a mean absolute error (MAE) and a root-mean-square error (RMSE) of 1.68 and 2.49 kcal/mol, respectively.

Despite the low errors of the GA-PCA model, the predicted values of 342 structures deviate from the DFT results by more than 4.0 kcal/mol (the substitution and H-bond effects energy range). The main reason for these deviations could be the lack of weak interactions, such as H-bonds, strain, and other non-nearest neighbor interactions.[18] As discussed in Section S3.1, the stronger H bonds between Ar−O* and adjacent Ar−OH significantly stabilize the radicals by ∼10 kcal/mol. Steric effects from −OCH$_3$ at the 4 and 5 positions also increase the system energies. For instance, when three adjacent groups (−OCH$_3$) are on the same aromatic unit in a sinapyl alcohol type, the −OCH$_3$ group in the middle is forced to be perpendicular to the ring due to crowdedness. The strain from the intramolecular ring structure between two aromatic rings also impacts the system energies. On the basis of these weak interactions' molecular patterns, the number of H bonds, local steric geometry, and strain from nonaromatic rings in model compounds are counted, using the SMARTS descriptions, and added to the configuration matrix (Table S11). The revised GA-PCA model with the corrected configuration matrix for these interactions is more acccurate. For $\Delta_f H^{298.15}$, the number of offset points larger than 4.0 kcal/mol reduces to 147, and the MAE and RMSE decrease to 1.28 and 1.79 kcal/mol,

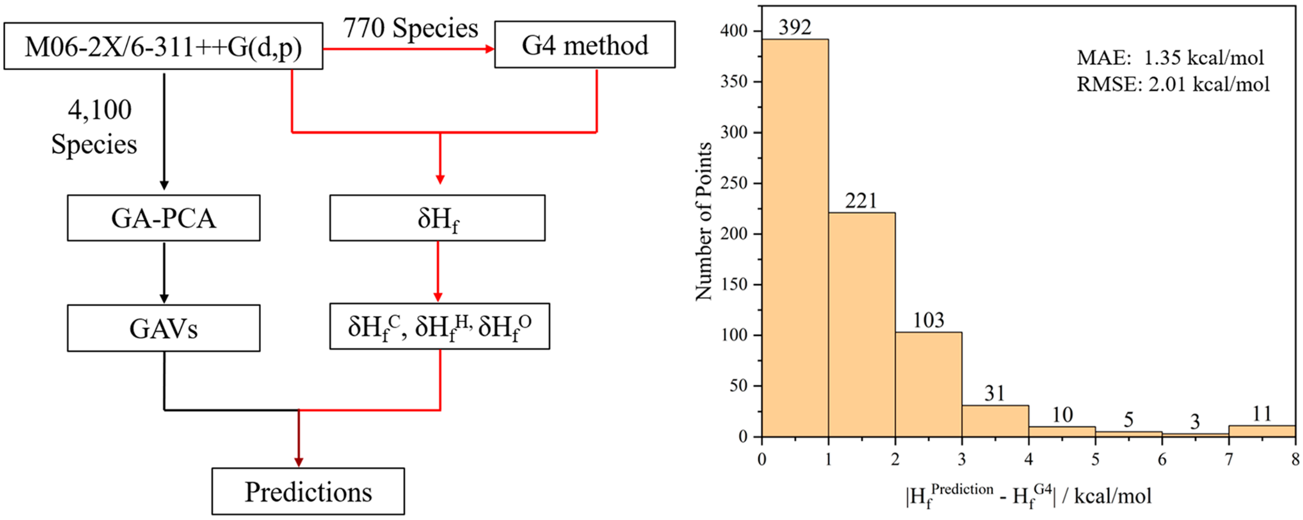


Figure 4. Strategy of correcting the GA-PCA model built at the M06-2X level to the G4 level and evaluation of predictions compared to the G4 subset.

respectively (Figure S7). Considering the large size of the lignin models and the correlation between the predicted GA-PCA and the DFT values for $\Delta_f H$, C_p, and S (Figure 3), the GA-PCA model can successfully reproduce the DFT-based data and enables a cheap way to obtain the BDEs and the effect of substitutions. The predictive ability of the GA-PCA model was also evaluated via typical cross-validation. The RMSEs are 2.05 kcal/mol ($\Delta_f H$), 0.46 cal/mol/K (C_p), and 2.99 cal/mol/K (S) (Tables S13, S14). In this revised scheme, long-range interactions, such as the H-bond formed between distant −OH and O−R groups on aromatic units, are still not accounted for because of the difficulty in determining their SMARTS expressions.

The database properties are computed at the M06-2X/6-311++G(d,p) level. However, intrinsic errors in the thermochemistry associated with the accuracy of the functional and basis set exist.[47] To improve the GA-PCA model's accuracy, a subset containing the smallest 770 species in the database was computed using the G4 method, which is known to be the best composite method for $C_xH_yO_z$ species[48] (Figure 4). Then, the differences in properties at the two levels were assigned to atomic corrections for the C, H, and O atoms, using a new data fusion method. As shown in Table S15, the corrections are nearly constant (−1.098, −0.407, and −0.890 kcal/mol per atom for δH_f^C, δH_f^H, and δH_f^O, respectively) with increasing training data set size, implying that the benchmark data set is sufficiently large, and more extended range corrections are not necessary. The corrected GA-PCA method (eq 6) predicts the formation enthalpies of 770 species with MAE and RMSE of 1.35 and 2.01 kcal/mol, respectively, while 92% of the predictions are between 0.0 and 3.0 kcal/mol.

CONCLUSIONS

The substitution effects at different positions of common structures for various lignin linkages, such as α−O−4, β−O−4, 4−O−5, β−1, β−5, β−β, and 5−5, on bond dissociation energies were investigated. A computational framework was employed to build a database containing 4100 molecules and radicals using an automatically generated library, via the Rule Input Network Generator, along with small $C_xH_yO_z$ structures from the literature. The database was used for generating the configuration matrix and training a group additivity model. To resolve the undetermined matrix problem, PCA was introduced to transform the regular groups in the GA method to new additivity descriptors based on the principal components. The −OCH$_3$ on the aromatic ring and the −OH on aliphatic tails reduce the adjacent bond strength. H-bonds and local steric effects greatly affect the stability of molecules and radicals and their BDEs. We improved the accuracy of the GA-PCA model first by including H-bonds and local steric effects and second via a data fusion approach. In the latter correction, G4 calculations are performed on a relatively small data set, and the atomic corrections are estimated as dominant error contributors and incorporated in the GA scheme. This error reduction method offers the possibility of correcting the GAVs in large databases computed at a low DFT level using only a small set of species computed at a higher level. This work provides a database of complex lignin structures and GAVs for accurate prediction of thermochemical properties of lignin pyrolysis.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssuschemeng.0c08856.

Equations for BDE and Δ_fH calculations, RING rules to generate the structure library, calculated BDEs of all linkage types with different substitution groups and the associated descriptions, SMARTS expressions for different types of H-bonds(PDF)

Brief description of the folder/files in the Supporting Information, coordinates in *xyz* format and DFT calculated thermal properties of species in the database, RING input file for reaction network generation, Python scripts for generating the configuration matrix, determining N(PC), including hydrogen bonds and steric corrections, conducting G4 benchmark and related outputs discussed in the main text and Supporting Information. (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Dionisios G. Vlachos − *Catalysis Center for Energy Innovation and Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States;* orcid.org/0000-0002-6795-8403; Email: vlachos@udel.edu

Authors

Qiang Li − *Catalysis Center for Energy Innovation, University of Delaware, Newark, Delaware 19716, United States;* orcid.org/0000-0001-5568-2334

Gerhard Wittreich − *Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States;* orcid.org/0000-0002-3968-7642

Yifan Wang − *Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States*

Himaghna Bhattacharjee − *Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, Delaware 19716, United States;* orcid.org/0000-0002-6598-3939

Udit Gupta − *Catalysis Center for Energy Innovation, University of Delaware, Newark, Delaware 19716, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssuschemeng.0c08856

Author Contributions

Q.L. performed all DFT calculations. G.W. contributed to the group additivity scheme and statistical assessments of the group additivity scheme. Y.W. contributed to the Python codes and GA discussions. H.B. contributed to the principal component regression and error reduction method. U.G. contributed to the structure generation using RING. D.G.V. conceptualized the problem and oversaw the work. The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work of Q.L. was supported by the Growing Convergence Research (GCR) program at the National Science Foundation (NSF) under Award Number NSF GCR CMMI 1934887. The work of Y.W., G.W., U.G., and H.B. was supported by the U.S. Department of Energy's Office of Energy Efficient and Renewable Energy's Advanced Manufacturing Office under Award Number DE-EE0007888-9.5. The Delaware Energy Institute gratefully acknowledges the support and partnership of the State of Delaware toward the RAPID projects. We also appreciate the kind discussions with Mr. Yue Qiu from Grimwade Centre for Cultural Materials Conservation, University of Melbourne.

■ REFERENCES

(1) Li, C.; Zhao, X.; Wang, A.; Huber, G. W.; Zhang, T. Catalytic Transformation of Lignin for the Production of Chemicals and Fuels. *Chem. Rev.* **2015**, *115*, 11559−11624.

(2) Sun, Z.; Fridrich, B.; de Santi, A.; Elangovan, S.; Barta, K. Bright Side of Lignin Depolymerization: Toward New Platform Chemicals. *Chem. Rev.* **2018**, *118*, 614−678.

(3) Chio, C.; Sain, M.; Qin, W. Lignin utilization: A Review of Lignin Depolymerization from various Aspects. *Renewable Sustainable Energy Rev.* **2019**, *107*, 232−249.

(4) Zakzeski, J.; Bruijnincx, P. C.; Jongerius, A. L.; Weckhuysen, B. M. The Catalytic Valorization of Lignin for the Production of Renewable Chemicals. *Chem. Rev.* **2010**, *110*, 3552−3599.

(5) Fan, L.; Zhang, Y.; Liu, S.; Zhou, N.; Chen, P.; Cheng, Y.; Addy, M.; Lu, Q.; Omar, M. M.; Liu, Y.; Wang, Y.; Dai, L.; Anderson, E.; Peng, P.; Lei, H.; Ruan, R. Bio-oil from Fast Pyrolysis of Lignin: Effects of Process and Upgrading Parameters. *Bioresour. Technol.* **2017**, *241*, 1118−1126.

(6) Kawamoto, H. Lignin Pyrolysis Reactions. *J. Wood Sci.* **2017**, *63*, 117−132.

(7) Kim, S.; Chmely, S. C.; Nimlos, M. R.; Bomble, Y. J.; Foust, T. D.; Paton, R. S.; Beckham, G. T. Computational Study of Bond Dissociation Enthalpies for A Large Range of Native and Modified Lignins. *J. Phys. Chem. Lett.* **2011**, *2*, 2846−2852.

(8) Younker, J. M.; Beste, A.; Buchanan, A. C., III Computational Study of Bond Dissociation Enthalpies for Substituted β-O-4 Lignin Model Compounds. *ChemPhysChem* **2011**, *12*, 3556−3565.

(9) Huang, J. B.; Wu, S. B.; Cheng, H.; Lei, M.; Liang, J. J.; Tong, H. Theoretical Study of Bond Dissociation Energies for Lignin Model Compounds. *J. Fuel Chem. Technol.* **2015**, *43*, 429−436.

(10) Elder, T. Bond Dissociation Enthalpies of A Pinoresinol Lignin Model Compound. *Energy Fuels* **2014**, *28*, 1175−1182.

(11) Elder, T.; Berstis, L.; Beckham, G. T.; Crowley, M. F. Density Functional Theory Study of Spirodienone Stereoisomers in Lignin. *ACS Sustainable Chem. Eng.* **2017**, *5*, 7188−7194.

(12) Berstis, L.; Elder, T.; Crowley, M.; Beckham, G. T. Radical Nature of C-lignin. *ACS Sustainable Chem. Eng.* **2016**, *4*, 5327−5335.

(13) Questell-Santiago, Y. M.; Galkin, M. V.; Barta, K.; Luterbacher, J. S. Stabilization Strategies in Biomass Depolymerization Using Chemical Functionalization. *Nat. Rev. Chem.* **2020**, *4*, 311−330.

(14) Cohen, N.; Benson, S. Estimation of Heats of Formation of Organic Compounds by Additivity Methods. *Chem. Rev.* **1993**, *93*, 2419−2438.

(15) Gu, G. H.; Plechac, P.; Vlachos, D. G. Thermochemistry of Gas-phase and Surface Species via LASSO-assisted Subgraph Selection. *React. Chem. Eng.* **2018**, *3*, 454−466.

(16) Sabbe, M. K.; Saeys, M.; Reyniers, M. F.; Marin, G. B.; Van Speybroeck, V.; Waroquier, M. Group Additive Values for the Gas Phase Standard Enthalpy of Formation of Hydrocarbons and Hydrocarbon Radicals. *J. Phys. Chem. A* **2005**, *109*, 7466−7480.

(17) Paraskevas, P. D.; Sabbe, M. K.; Reyniers, M. F.; Papayannakos, N.; Marin, G. B. Group Additive Values for the Gas-Phase Standard Enthalpy of Formation, Entropy and Heat Capacity of Oxygenates. *Chem. - Eur. J.* **2013**, *19*, 16431−16452.

(18) Ince, A.; Carstensen, H. H.; Reyniers, M. F.; Marin, G. B. First-principles Based Group Additivity Values for Thermochemical Properties of Substituted Aromatic Compounds. *AIChE J.* **2015**, *61*, 3858−3870.

(19) Ince, A.; Carstensen, H. H.; Sabbe, M. K.; Reyniers, M. F.; Marin, G. B. Modeling of Thermodynamics of Substituted Toluene Derivatives and Benzylic Radicals via Group Additivity. *AIChE J.* **2018**, *64*, 3649−3661.

(20) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126*, 084108.

(21) Brönsted, J. Acid and Basic Catalysis. *Chem. Rev.* **1928**, *5*, 231−338.

(22) Evans, M.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11−24.

(23) Gani, T. Z.; Orella, M. J.; Anderson, E. M.; Stone, M. L.; Brushett, F. R.; Beckham, G. T.; Román-Leshkov, Y. Computational Evidence for Kinetically Controlled Radical Coupling During Lignification. *ACS Sustainable Chem. Eng.* **2019**, *7*, 13270−13277.

(24) Orella, M. J.; Gani, T. Z.; Vermaas, J. V.; Stone, M. L.; Anderson, E. M.; Beckham, G. T.; Brushett, F. R.; Román-Leshkov, Y. Lignin-KMC: A Toolkit for Simulating Lignin Biosynthesis. *ACS Sustainable Chem. Eng.* **2019**, *7*, 18313−18322.

(25) Dellon, L. D.; Yanez, A. J.; Li, W.; Mabon, R.; Broadbelt, L. J. Computational Generation of Lignin Libraries from Diverse Biomass Sources. *Energy Fuels* **2017**, *31*, 8263−8274.

(26) Yanez, A. J.; Natarajan, P.; Li, W.; Mabon, R.; Broadbelt, L. J. Coupled Structural and Kinetic Model of Lignin Fast Pyrolysis. *Energy Fuels* **2018**, *32*, 1822−1830.

(27) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, P. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision D.01 ; Gaussian, Inc.: Wallingford, CT, 2009.

(28) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.

(29) Parthasarathi, R.; Romero, R. A.; Redondo, A.; Gnanakaran, S. Theoretical Study of the Remarkably Diverse Linkages in Lignin. *J. Phys. Chem. Lett.* **2011**, *2*, 2660−2666.

(30) Cheng, H.; Wu, S.; Huang, J.; Zhang, X. Direct Evidence from In Situ FTIR Spectroscopy that o-quinonemethide Is A Key Intermediate During the Pyrolysis of Guaiacol. *Anal. Bioanal. Chem.* **2017**, *409*, 2531−2537.

(31) Lym, J.; Wittreich, G. R.; Vlachos, D. G. A Python Multiscale Thermochemistry Toolbox (pMuTT) for Thermochemical and Kinetic Parameter Estimation. *Comput. Phys. Commun.* **2020**, *247*, 106864.

(32) Alecu, I. M.; Zheng, J.; Zhao, Y.; Truhlar, D. G. Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries. *J. Chem. Theory Comput.* **2010**, *6*, 2872−2887.

(33) Rangarajan, S.; Bhan, A.; Daoutidis, P. Rule-based Generation of Thermochemical Routes to Biomass Conversion. *Ind. Eng. Chem. Res.* **2010**, *49*, 10459−10470.

(34) Ghahremanpour, M. M.; van Maaren, P. J.; Ditz, J. C.; Lindh, R.; van der Spoel, D. Large-scale Calculations of Gas Phase Thermochemistry: Enthalpy of Formation, Standard Entropy, and Heat Capacity. *J. Chem. Phys.* **2016**, *145*, 114305.

(35) Ralph, J.; Lapierre, C.; Boerjan, W. Lignin Structure and Its Engineering. *Curr. Opin. Biotechnol.* **2019**, *56*, 240−249.

(36) Weininger, D. SMILES, A Chemical Language and Information System. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(37) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

(38) RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org. (accessed Feb. 9, 2021).

(39) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF Force Filed to the RDKit: Implementation and Validation. *J. Cheminf.* **2014**, *6*, 37.

(40) Rezac, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8*, 141−151.

(41) Stewart, J. MOPAC2016, 2016. http://OpenMOPAC.net (accessed Feb. 9, 2021).

(42) Python Group Additivity. https://github.com/VlachosGroup/PythonGroupAdditivity (accessed Feb. 9, 2021).

(43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(44) SMARTS − A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Feb. 9, 2021).

(45) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087−2096.

(46) Bhattacharjee, H.; Vlachos, D. G. Thermochemical Data Fusion Using Graph Representation Learning. *J. Chem. Inf. Model.* **2020**, *60*, 4673.

(47) Mardirossian, N.; Head-Gordon, M. How Accurate Are the Minnesota Density Functionals for Noncovalent Interactions, Isomerization Energies, Thermochemistry, and Barrier Heights Involving Molecules Composed of Main-group Elements? *J. Chem. Theory Comput.* **2016**, *12*, 4303−4325.

(48) Somers, K. P.; Simmie, J. M. Benchmarking Compound Methods (CBS-QB3, CBS-APNO, G3, G4, W1BD) Against the Active Thermochemical Tables: Formation Enthalpies of Radicals. *J. Phys. Chem. A* **2015**, *119*, 8922−8933.