Application of DFTB and machine learning to evaluate the stability of biomass intermediates on the Rh(111) surface

Chaoyi Chang and Andrew J. Medford*

School of Chemical & Biomolecular Engineering, Georgia Institute of Technology

E-mail: ajm@gatech.edu

Abstract

Biomass compounds adsorbed on surfaces are challenging to study due to the large number of possible species and adsorption geometries. In this work, possible intermediates of erythrose, glyceraldehyde, glycerol and propionic acid are studied on the Rh(111) surface. The intermediates and elementary reactions are generated from first 2 recursions of a recursive bond-breaking algorithm. These structures are used as the input of an unsupervised Mol2Vec algorithm to generate vector descriptors. A data-driven scheme to classify the reactions is developed and adsorption energies are predicted. The lowest mean absolute error (MAE) of our prediction on adsorption energies is 0.39 eV, and the relative ordering of different surface adsorption geometries is relatively accurate. We show that combining geometries from density functional tight-binding (DFTB) calculations with energies from machine-learning predictions provides a novel workflow for rapidly assessing the stability of various molecular geometries on the Rh(111) surface.

Biomass compounds are playing a key role in biorenewable products and are the basis of a sustainable economy. For example, the US Department of Energy has listed the 10

top biorefinery products with the highest potential impact.[1] Selective conversion between those products and other small molecule products are important if biomass compounds are to form the basis of a sustainable economy. Rh is a commonly-used transition metal catalyst for conversion of biomass compounds and has been the subject of numerous experimental studies. For example, Rh catalysts have been used for succinic acid conversion to fumaric acid,[2] hydrogenolysis of furfural to 1,2-pentanediol[3] and production of C1 compounds from ethanol.[4–6] However, there have been relatively few systematic computational studies of biomass intermediates on Rh surfaces.[7] Density functional theory (DFT) is the most common theory used to study the adsorption and reaction of biomass compounds.[8–10] However, complex molecules have multiple binding sites and various geometries, and DFT calculations of larger molecules are expensive and require significant computational effort to converge. Thus, previous studies often utilize empirical and machine learning (ML) methods to relate reaction properties of biomass molecules or intermediates to their structures, physical properties and even experimental conditions.[9,11–13] Most of these ML methods could reach a mean absolute error of adsorption energies/transition state energies within 0.4 eV,[7,13,14] with some examples of predictions as accurate as 0.2 eV using a combination of physical and structural features along with feature selection methods.[12,15]

As mentioned previously, a key challenge with biomass molecules is their complexity, and "model compounds" are often used to simplify systems of interest. A "model compound" refers to a relatively small molecule (typically fewer than 8 heavy atoms) that can be used for studying a larger compounds with similar chemical properties and functional groups. Previous experimental studies showed that the reaction pathway of glyceraldehyde $\rightleftharpoons$ glycerol is similar to linear glucose $\rightleftharpoons$ sorbitol[8,10,16,17] and that propionic acid ketonization is similar to larger carboxylic acid ketonization on Rh surfaces.[9,18] While model compounds are commonly used, they are typically identified by heuristics and intuition. However, a data-driven strategy for systematic identification of model compounds was proposed in a prior study by the authors.[19] It was pointed out that glycerol, glyceraldehyde, propionic acid,

and erythrose are good model compounds to study larger molecules like sorbitol and linear-structured glucose. This makes these four molecules a key starting point for computational and experimental studies seeking a more general understanding of the catalytic conversion of biomass derivatives on solid surfaces.

In this study, we extend the previously-developed embedding models of gas-phase formation energies[19] to adsorbed surface intermediates on the Rh(111) surface. A total of 171 intermediates from the first 2 bond-breaking recursions of erythrose, glyceraldehyde, glycerol and propionic acid are studied. Mol2Vec is used for generating vector descriptors and 83 clusters based on 6 reaction types (C-C, C-O, O-H, C-H, C-M, O-M) are obtained from single-group "radius zero" (R0) Mol2Vec descriptors. Linear discriminant analysis (LDA) and partial least squares (PLS) are used for dimensional reduction of two- and three-group "radius one" (R1) Mol2Vec descriptors, providing low-dimensional vector descriptors for each adsorbate. These vectors are combined with a linear least-squares regression model, yielding a mean absolute errors (MAE) as low as 0.39 eV. Finally, pre-optimization via density-functional tight binding (DFTB) is combined with our embedding models to establish a workflow for rapidly identifying stable adsorption geometries. We show that this workflow identifies 20 new lowest-energy geometries for 171 adsorbates studied, indicating that systematic approaches for identifying the lowest-energy structures of large adsorbed molecules are a necessary addition to the tool set of computational catalysis.

First, we establish an approach to assign detailed classes for each bond type of adsorbed biomass intermediates, similar to the previously-developed approach for gas-phase molecules.[19] Mol2Vec[20] is used for generating vector descriptors for intermediates and reactions with 200 dimensions in R0 and R1.[19] The traditional extended-connectivity fingerprint (ECFP) is modified to include surrounding heavy atom number, surrounding hydrogen atom number, valence, electric negativity and mass as invariants for atoms contained in a structure. Metal atoms are still considered to be heavy atoms but all the other properties are considered as NaN so that metal atoms are a general rather than specific type (e.g. Rh is

indistinguishable from any other metal). An algorithm (see SI) to add metal atoms to unsaturated C, O atoms one-at-a-time is applied to the 171 intermediates generated from the first 2 bond-breaking recursions of erythrose, glyceraldehyde, glycerol and propionic acid, and additional structures are generated by DFTB minima hopping calculations (see below), yielding a total of 2,498 adsorbed structures. The adsorbed structures are combined with 91,098 gas-phase species, resulting in a total of 93,569 structures that are used as the corpus for training the Mol2Vec model (see SI for details). We also utilize 6 basic types of reactions for visualization and checking intuition. These 6 types include 4 types of intra-adsorbate reactions, C-C (C=C and C-C), C-H, C-O (C=O and C-O), OH, and 2 types of elementary reactions with metal atoms, C-M (C-Metal), O-M (O-Metal). The single and double bond breaking reactions are considered as a single class since the distinction between them becomes ambiguous for adsorbed species due to partial bond orders and conjugation. A total of 13,422 gas-phase reactions from our previous work[19] and 1,666 additional elementary surface reaction steps are included for the analysis of reactions. For the convenience of visualization, principal component analysis (PCA) is used for reducing the 200-dimensional reaction vectors to 2 dimensions, and each type of reaction is represented by a different color (yellow for C-C, green for C-H, cyan for C-O, red for O-H, black for C-M, blue for O-M). Fig. 1a and 1b show the visualization of R0 and R1 reaction vectors for the full corpus of intermediates and reactions.

The PCA result in R0 indicates that there are discrete well-defined clusters within the 15,088 reactions. The Euclidean distance between reactions of the original 200-dim R0 vectors are calculated to identify distinct clusters, and the cutoff to separate clusters is set to be 0.05. A total of 83 clusters in R0 are obtained. Each of the 83 clusters contain reactions with different bond-breaking types and different atomic environments of the the atoms within the elementary reaction. The atomic environment refers to the surrounding heavy atoms and hydrogen atoms of the 2 reacting atoms. Table S1 shows all details of the atomic environments for each of these reaction types.

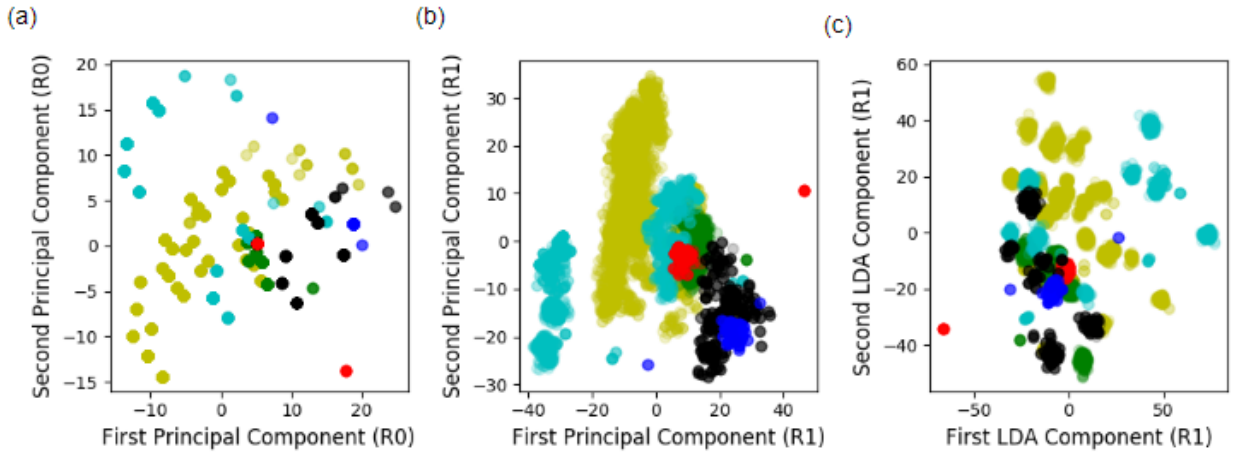

Figure 1: Visualization of reaction vector clusters and 6 reaction types (yellow for C-C, green for C-H, cyan for C-O, red for O-H, black for C-M, blue for O-M): (a) 1st vs. 2nd component of PCA (R0), (b) 1st vs. 2nd component of PCA (R1) and (c) 1st vs. 2nd component of LDA (R1 with class labels from R0 clusters).

The obtained 83 R0 clusters are then used as labels for a supervised classification on R1. Linear discriminant analysis (LDA), is applied on the 200-dimension R1 vector descriptors together with the 83 classes as class labels. Fig. 1c shows the first and the second LDA components of LDA-projected R1 reaction vectors, and the 6 reaction types are represented by 6 different colors for the convenience of visualization. The R1 vectors are reduced to 1-82 dimensions via the LDA projection. The LDA projections represent an unsupervised vector space for analyzing reactions and intermediates, and are used to predict adsorption energies on Rh (111) surface.

Ultimately, the goal is to predict adsorption energies of biomass species. The 171 adsorbates, ranging from C1 - C4 species (see SI for details) included in this study are generated from the first two bond-breaking recursions[19] of erythrose, glyceraldehyde, glycerol and propionic acid. In general, each species can have multiple adsorption energies due to differences in molecular configuration and binding sites. This makes it challenging to directly predict the most stable binding site even with DFT. For this reason, a low-cost pre-optimization tool is needed to identify different stable geometries. Density functional tight-binding (DFTB) provides a rapid physics-based route that provides relatively accurate energies and geometries

of reactive surfaces.[21,22] We use the open-source Hotbit[23] Python package with a previously-developed parameterization for Rh/C/H/O[24] for performing DFTB. The Hotbit calculator is used with constrained minima hopping[25] where Rh slab atoms have fixed positions and adsorbate bonds have fixed lengths[26] to generate local minima geometries for each adsorbate.

The DFTB-based minima hopping process is used to generate up to 50 adsorption geometries for each of the 171 adsorbates, yielding a total of 857 different adsorbate structures. Geometries that are within 1 eV of the lowest Hotbit energies are then calculated by DFT. DFT calculations are performed using Quantum ESPRESSO[27] with the PBE[28] exchange-correlation functional at a planewave cutoff of 450 eV (see SI for details). This process yields 328 unique geometries and associated adsorption energies, which are used as inputs for supervised training of the regression models.

Adsorption energies with both DFT and DFTB are computed for all 328 unique geometries. This data is used to assess the accuracy of the DFTB energies. Since DFT and DFTB do not use a common reference, it is necessary to align the energies based on the stochiometry of each adsorbate:

$$E_{DFT} = E_{DFTB} + \left(\sum_{i \in [C,H,O]} c_i * n_i \right) + \epsilon \tag{1}$$

where E_{DFT} is the DFT adsorption energy, E_{DFTB} is the DFTB adsorption energy, n_i is the number of C, H, O atoms in the adsorbate, c_i are fitted coefficients and ϵ is the residual error. Fig. 2a shows the distribution of the residual DFTB error in energy calculation and Fig. 2b shows the parity plot of DFT and corrected DFTB energies. The mean absolute error (MAE) of DFTB is 1.46 eV, with notable outliers that can have errors of >5 eV. In addition, we use the Spearman's correlation coefficient to evaluate the ability of DFTB to correctly order the energies of different geometries for a given adsorbate. The results, shown in Fig. 3b, reveal that DFTB yields incorrect ordering of adsorbates more than 60% of the time. Despite these large energy errors, the geometries are more accurate with an average position difference of 1.04 Å. This suggests that while Hotbit is a reasonable tool for generating

adsorbate geometries, the energy predictions are not sufficiently accurate to yield chemical insight or even correctly order the stability of various adsorbate geometries.

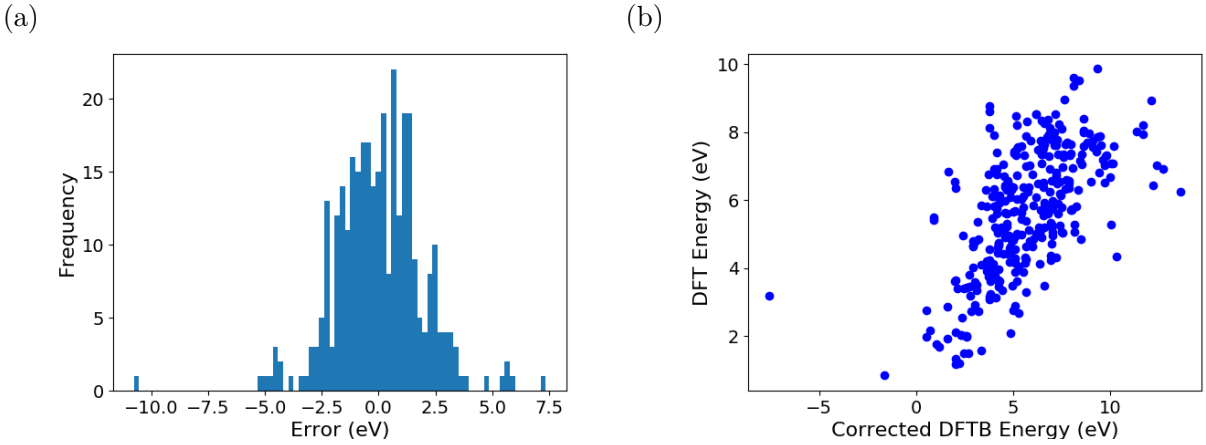

Figure 2: Accuracy of DFTB energies after reference alignment (a) error distribution and (b) parity plot of DFTB vs. DFT energy

To improve the accuracy of energy predictions we turn to a supervised ML approach. The workflow utilizes the Mol2Vec vectors for each adsorbate as inputs, similar to our previously-developed approach for gas-phase energies.[19] We compare three different linear models for predicting adsorption energies from Mol2Vec vectors: The original unsupervised Mol2Vec vectors (OLS), the semi-supervised LDA projection, and the supervised partial least squares (PLS) projection. In each case we utilize a 75/25 train/test split with 4 random repeats with the results from the 328 DFT calculations as the target. The `scikit-learn` package[29] is used for each algorithm (see SI for more details). In each case we evaluate the mean absolute error of the test set as a function of the dimension of the input vector, with a maximum dimension of 82 (the maximum dimension of the LDA vectors).

The results of the ML predictions, shown in Fig. 3a show that the lowest MAE of OLS, PLS and LDA are ∼0.51 eV, ∼0.39 eV and ∼0.43 eV at ∼55-dim, ∼30-dim and ∼65-dim respectively. The MAE of OLS, PLS and LDA decrease at first and reach a plateau or increase slowly after 55-dim, 30-dim and 65-dim. The lowest MAE of the LDA and PLS models are very similar, and we expect that the LDA model will be more transferrable to other adsorbates since the adsorption energies are not used to generate the feature vector.

By contrast, we expect that the PLS performance is more specific to this set of adsorbates since the adsorption energies are used to generate the inputs. Comparing the results of the machine-learning models to DFTB, we see that the prediction errors are much lower (∼0.4 eV for ML vs. ∼1.5 eV for DFTB). While the error of ∼0.4 eV is still somewhat larger than the typical DFT error (∼0.2 eV), the ordering of energies of different geometries with the ML model is relatively good, with a correct ordering at least 59% of the time (Fig. 3c and Fig. 3d). However, the ML models requires geometries as inputs, which must be generated by DFTB. This suggests a synergistic approach between the models is required, where DFTB is used to identify geometries and ML is used to predict energies.

The average error of the best machine-learning models (∼0.4 eV) is relatively large, indicating that DFT will need to be used when accurate energies are required. However, the number of possible geometries and active sites for biomass molecules makes brute force DFT calculations impractical for large numbers of adsorbates. The combination of geometries from DFTB along with predictions from ML can alleviate this issue by identifying the geometries that are most likely to be stable, thus reducing the number of DFT calculations required. Spearman's correlation coefficient is used to quantitatively assess the model's ability to predict the energy order for different geometries of the same adsorbate. We utilize a 75/25 train/test split with 4 random repeats with the adsorbates for PLS and LDA (approximately 1-3 geometries per adsorbate). Fig. 3c and Fig. 3d show the distribution of Spearman's correlation coefficient of PLS and LDA test sets with error bars. PLS and LDA obtain more than 65% and 59% correct energy orders on average with no more than 25% totally inverted (88% of the totally inverted geometries include only 2 geometries). Fig. 3b shows the distribution of Spearman's correlation coefficient of DFTB with only 35% correct, while more than 30% are totally inverted.

This demonstrates that both PLS and LDA have similar performance, but both are better than DFTB in predicting energy orders. However, there are still some mis-ordered energies based on the ML predictions, suggesting that the lowest energy structure may not

be correctly identified in all cases. To increase robustness, we utilize the average standard deviation of the model errors (0.45 eV) as a tolerance factor, meaning any structure within 0.45 eV of the lowest energy is considered as a possible global minimum. Using simple estimates from probability theory this corresponds to 75% confidence that the true global minimum will be included (see SI). This cutoff can be adjusted to improve confidence at the expense of more DFT calculations. Both PLS and LDA models are built based on the 328 DFT calculations and are used for predicting all 857 geometries generated from DFTB and minima hopping. The minimum energy of each adsorbate and the geometries within 0.45 eV of the minimum energy are extracted, and any geometries that have not already been computed are calculated with DFT. Both of the ML models are used, and the geometries are calculated with DFT if the energy predicted by either model is within the threshold.

The ML model identifies multiple possible new global minima for 65 of the adsorbates, corresponding to 154 additional DFT calculations. The MAE between the model predictions and the DFT energies of the new structures are 0.60 eV and 0.62 eV for LDA and PLS models, about 50% higher than the MAE of the test set. The results of these calculations reveal that the energies of many of these structures are lower than the previously-determined lowest energy structure, as shown in Fig. 4a (see SI for details). For 20 of the 65 adsorbates the configuration of the global minima was sufficiently different to lead to a new SMILES representation. For example, for the the adsorbate $[O]CC(O)[CH]$ the ML model identifies a $[CH]$ bidentate-binding geometry with 0.44 eV lower energy than the monodentate geometry that was previously identified as the lowest energy structure, as shown in Fig. 4b. For 13 other adsorbates, the difference in geometry was less drastic so that the SMILES string of the new structure was the same as the old structure, but the energy was slightly different by $<$0.23 eV (see Fig. 4b). The original energy was lower than all newly-computed energies in 32 adsorbates. Overall, the original workflow for identifying global minima failed to identify the global minimum for at least 17% of adsorbates, and had qualitative differences in binding geometries (different SMILES string or energy difference $>$ 0.05 eV) for 14% of adsorbates.

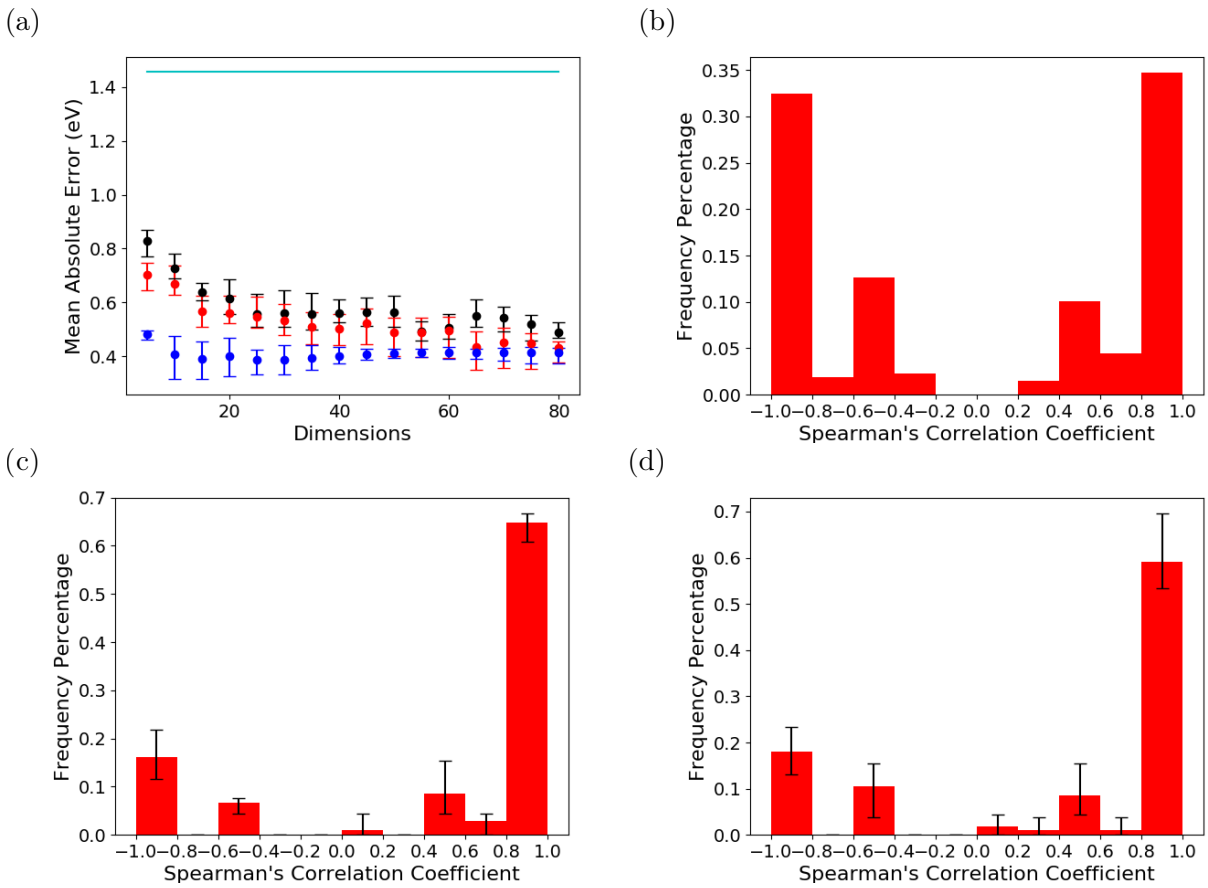

Figure 3: DFTB and ML results of (a) cross validation error of OLS (black), LDA (red), PLS (blue) and DFTB MAE (cyan line), spearman's correlation coefficient of (b) DFTB calculations, (c) 30-dim PLS regression and (d) 65-dim LDA projections with error bars

The results of this work suggest that a combination of physical approximations and ML models is a promising route toward identifying global minima of complex adsorbates. A general workflow involves a first step that uses an approximate physical method (DFTB in this case) to rapidly generate many candidate geometries. The second step involves using DFT to calculate the energies of the most stable structures (structures within 1 eV of the minimum in this case, leading to 328 geometries of 171 adsorbates). Third, these DFT energies are used to train ML models, here based on Mol2Vec and linear regression, and the ML models are then used to predict the energies of all candidate structures. Finally, the predictions of the ML model are used to identify new structures that will be computed with DFT, in this case structures within 0.45 eV (the standard error of the ML models on the

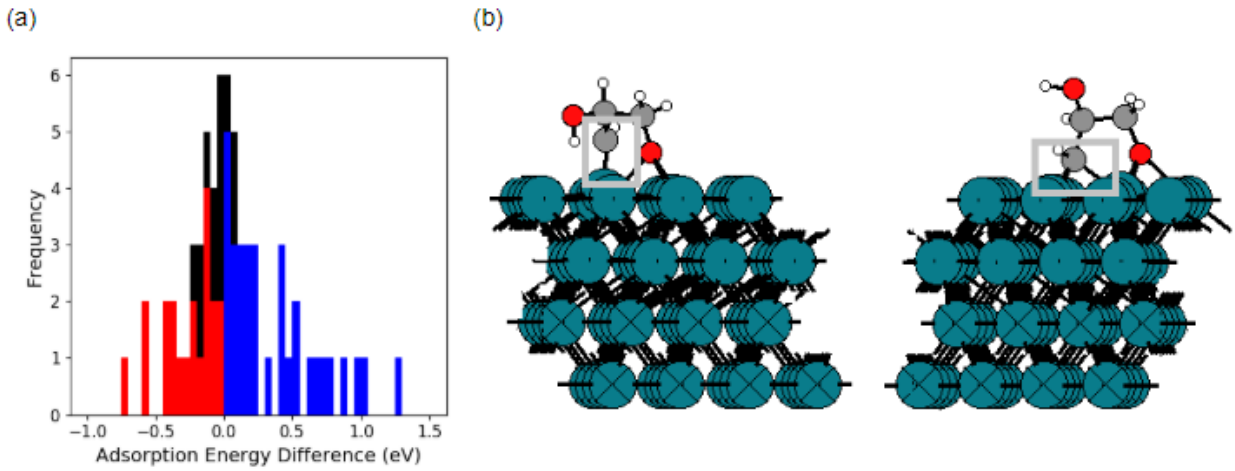

Figure 4: Results of DFT calculations for structures predicted to be low-energy by ML model. (a) Stacked bar plot of adsorption energy difference for structures with the same (black) or different (red/blue) SMILES strings from prior global minimum, with lower-energy structures in red and higher-energy structures in blue. (b) A representative example of an adsorbate ($[O]CC(O)[CH]$) where the new lower-energy geometry binds with a qualitatively different structure, where $[CH]$ binds directly to a single Rh atom (left, gray box) instead of two Rh atoms in the previous structure (right, gray box).

test set) of the predicted minimum. The results of this work show that this process yields 20 global energy minima that would have been incorrect without the use of the ML model. This process can be made more efficient with improved physical approximations and more accurate ML models, which may be necessary to tackle larger and more complex biomass molecules.

The size and complexity of biomass molecules leads to a major challenge in predicting the global minimum adsorption geometry, and this challenge is compounded by the number of possible intermediates that appear in biomass reaction networks. It is clear that new techniques are needed to accelerate the study of these systems since direct calculation with DFT or other quantum chemical techniques is impractical. Here, we show that ML models based on Mol2Vec descriptors can achieve an MAE of 0.39 eV (PLS with 30-dim) and 0.41 eV (LDA with 65-dim) when applied to 171 intermediates derived from erythrose, glyceraldehyde, glycerol and propionic acid. These models provide more accurate estimates of adsorption energies than DFTB (1.46 eV), but the lowest MAE of ML methods are still not

comparable to DFT. We used Spearman's correlation to show that ML methods are much more reliable for assessing the relative stability of different geometries, providing a more robust route to identifying low-energy structures. Finally, the best aspects of ML and DFTB techniques are combined leading to a new workflow of DFTB+minima hopping $\rightarrow$ DFT $\rightarrow$ ML $\rightarrow$ DFT. This approach allows us to discover 20 new global minima for the 171 adsorbates studied here, which would be missed if only DFTB were used to evaluate candidate structures. Nonetheless, the workflow also has the limitation that there is still uncertainty about whether or not the true global minima is found, since an exhaustive search is not feasible for these complex adsorbates. However, the results indicate that combining physical models and ML predictions is a promising path toward solving this challenging problem.

Acknowledgement

CC is supported by a Paper Science and Engineering Fellowship from the Renewable Bioproducts Institute at Georgia Tech. Computational effort was supplied partially by the National Science Foundation under Grant No. MRI-1828187.

References

(1) Werpy, T.; Petersen, G.; Aden, A.; Bozell, J.; Holladay, J.; White, J.; Manheim, A.; Eliot, D.; Lasure, L.; Jones, S. *Top value added chemicals from biomass. Volume 1-Results of screening for potential candidates from sugars and synthesis gas*; 2004.

(2) Lam, E.; Luong, J. H. Carbon materials as catalyst supports and catalysts in the transformation of biomass to fuels and chemicals. *ACS catalysis* **2014**, *4*, 3393–3410.

(3) Pisal, D. S.; Yadav, G. D. Single-step hydrogenolysis of furfural to 1, 2-pentanediol using a bifunctional Rh/OMS-2 catalyst. *ACS omega* **2019**, *4*, 1201–1214.

(4) Zhang, J.; Zhong, Z.; Cao, X.-M.; Hu, P.; Sullivan, M. B.; Chen, L. Ethanol steam reforming on Rh catalysts: theoretical and experimental understanding. *ACS Catalysis* **2014**, *4*, 448–456.

(5) Cavallaro, S. Ethanol steam reforming on Rh/Al2O3 catalysts. *Energy & Fuels* **2000**, *14*, 1195–1199.

(6) Mei, D.; Lebarbier Dagle, V.; Xing, R.; Albrecht, K. O.; Dagle, R. A. Steam reforming of ethylene glycol over MgAl2O4 supported Rh, Ni, and Co catalysts. *ACS Catalysis* **2016**, *6*, 315–325.

(7) Abdelfatah, K.; Yang, W.; Vijay Solomon, R.; Rajbanshi, B.; Chowdhury, A.; Zare, M.; Kundu, S. K.; Yonge, A.; Heyden, A.; Terejanu, G. Prediction of Transition-State Energies of Hydrodeoxygenation Reactions on Transition-Metal Surfaces Based on Machine Learning. *The Journal of Physical Chemistry C* **2019**, *123*, 29804–29810.

(8) Auneau, F.; Michel, C.; Delbecq, F.; Pinel, C.; Sautet, P. Unravelling the mechanism of glycerol hydrogenolysis over rhodium catalyst through combined experimental–theoretical investigations. *Chemistry–A European Journal* **2011**, *17*, 14288–14299.

(9) Yang, W.; Solomon, R. V.; Mamun, O.; Bond, J. Q.; Heyden, A. Investigation of the reaction mechanism of the hydrodeoxygenation of propionic acid over a Rh (1 1 1) surface: A first principles study. *Journal of Catalysis* **2020**, *391*, 98–110.

(10) Wang, S.; Yin, K.; Zhang, Y.; Liu, H. Glycerol hydrogenolysis to propylene glycol and ethylene glycol on zirconia supported noble metal catalysts. *ACS Catalysis* **2013**, *3*, 2112–2121.

(11) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine learning for heterogeneous catalyst design and discovery. **2018**,

(12) Gu, G. H.; Plechac, P.; Vlachos, D. G. Thermochemistry of gas-phase and surface species via LASSO-assisted subgraph selection. *Reaction Chemistry & Engineering* **2018**, *3*, 454–466.

(13) Vorotnikov, V.; Wang, S.; Vlachos, D. G. Group additivity for estimating thermochemical properties of furanic compounds on Pd (111). *Industrial & Engineering Chemistry Research* **2014**, *53*, 11929–11938.

(14) Salciccioli, M.; Edie, S.; Vlachos, D. Adsorption of acid, ester, and ether functional groups on Pt: fast prediction of thermochemical properties of adsorbed oxygenates via DFT-based group additivity methods. *The Journal of Physical Chemistry C* **2012**, *116*, 1873–1886.

(15) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of adsorption energies for chemical species on metal catalyst surfaces using machine learning. *The Journal of Physical Chemistry C* **2018**, *122*, 28142–28150.

(16) Valter, M.; Dos Santos, E. C.; Pettersson, L. G.; Hellman, A. Partial Electrooxidation of Glycerol on Close-Packed Transition Metal Surfaces: Insights from First-Principles Calculations. *The Journal of Physical Chemistry C* **2020**, *124*, 17907–17915.

(17) Kwon, Y.; Koper, M. T. Electrocatalytic hydrogenation and deoxygenation of glucose on solid metal electrodes. *ChemSusChem* **2013**, *6*, 455–462.

(18) Kumar, R.; Enjamuri, N.; Shah, S.; Al-Fatesh, A. S.; Bravo-Suarez, J. J.; Chowdhury, B. Ketonization of oxygenated hydrocarbons on metal oxide based catalysts. *Catalysis Today* **2018**, *302*, 16–49.

(19) Chang, C.; Medford, A. J. Classification of biomass reactions and predictions of reaction energies through machine learning. *The Journal of Chemical Physics* **2020**, *153*, 044126.

(20) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling* **2018**, *58*, 27–35.

(21) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Physical Review B* **1995**, *51*, 12947.

(22) Seifert, G.; Porezag, D.; Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *International journal of quantum chemistry* **1996**, *58*, 185–192.

(23) P. Koskinen, V. M. Density-functional tight-binding for beginners. *Computational Material Science* **2009**, *47*, 237.

(24) Yang, N.; Medford, A. J.; Liu, X.; Studt, F.; Bligaard, T.; Bent, S. F.; Nørskov, J. K. Intrinsic selectivity and structure sensitivity of rhodium catalysts for C2+ oxygenate production. *Journal of the American Chemical Society* **2016**, *138*, 3705–3714.

(25) Peterson, A. A. Global optimization of adsorbate–surface structures while preserving molecular identity. *Topics in Catalysis* **2014**, *57*, 40–53.

(26) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(27) Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I., et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of physics: Condensed matter* **2009**, *21*, 395502.

(28) Perdew, J. P.; Burke, K.; Wang, Y. Generalized gradient approximation for the exchange-correlation hole of a many-electron system. *Physical review B* **1996**, *54*, 16533.

(29) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

Application of DFTB and machine learning to evaluate the stability of biomass intermediates on the Rh(111) surface

Chaoyi Chang and Andrew J. Medford*

School of Chemical & Biomolecular Engineering, Georgia Institute of Technology

E-mail: ajm@gatech.edu

Supporting Information Available

SMILES string generation for adsorbates

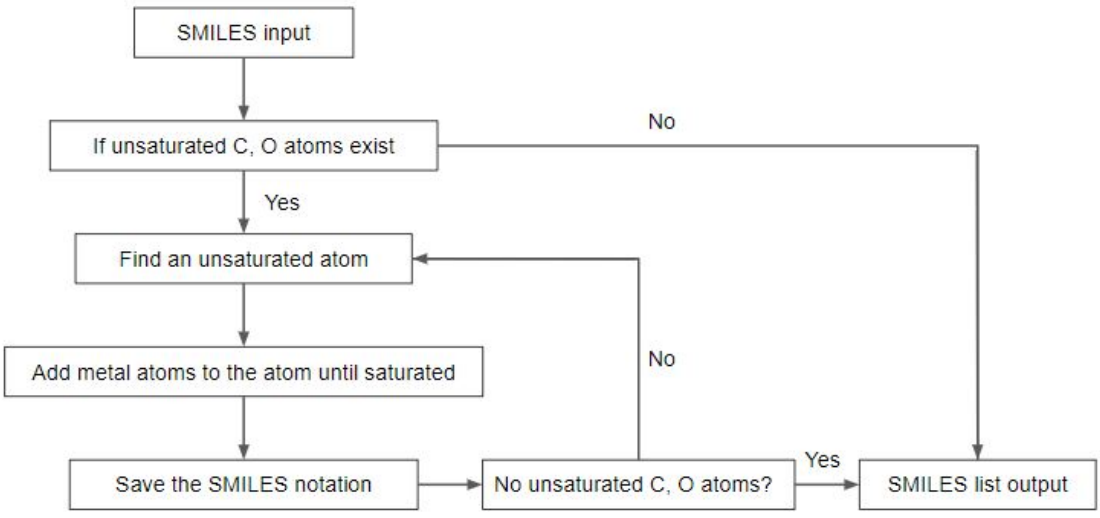

Figure 1: Generating algorithms of SMILES notation of adsorbate with metal atoms

Reaction types identified from vector clustering

Table 1: 83 clusters from R0 vectors with atomic environment (# of hydrogen atom and heavy atom surrounding)

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
1	2	3	1	4	C-C
2	2	3	0	4	C-C
3	0	4	0	3	C-C
4	1	4	1	3	C-C
5	1	3	0	3	C-C
6	0	4	1	3	C-C
7	2	3	1	3	C-C
8	1	4	0	3	C-C
9	0	3	2	3	C-C
10	1	3	1	3	C-C
11	0	3	0	3	C-C
12	0	4	0	4	C-C
13	2	3	2	3	C-C
14	1	4	0	4	C-C
15	1	4	1	4	C-C
16	0	2	1	4	C-C
17	0	2	1	3	C-C
18	0	2	2	3	C-C
19	0	2	0	4	C-C
20	0	2	0	3	C-C
21	0	2	0	2	C-C
22	1	2	2	2	C-C

Continued on next page

Table 1 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
23	2	2	1	4	C-C
24	2	2	0	3	C-C
25	0	4	2	2	C-C
26	2	2	1	3	C-C
27	2	2	2	3	C-C
28	0	2	2	2	C-C
29	2	2	2	2	C-C
30	0	2	1	2	C-C
31	1	2	1	4	C-C
32	1	2	0	4	C-C
33	1	2	1	3	C-C
34	1	2	0	3	C-C
35	1	2	2	3	C-C
36	1	2	1	2	C-C
37	3	2	0	4	C-C
38	3	2	0	3	C-C
39	3	2	2	3	C-C
40	1	3	3	2	C-C
41	3	2	0	2	C-C
42	2	2	3	2	C-C
43	1	2	3	2	C-C
44	1	4	1	2	C-O
45	1	3	1	2	C-O
46	0	4	1	2	C-O

Continued on next page

Table 1 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
47	0	3	1	2	C-O
48	2	3	1	2	C-O
49	2	2	1	2	C-O
50	0	2	1	2	C-O
51	2	2	0	2	C-O
52	1	3	0	2	C-O
53	1	4	0	2	C-O
54	0	3	0	2	C-O
55	0	4	0	2	C-O
56	2	3	0	2	C-O
57	1	2	1	2	C-O
58	1	2	0	2	C-O
59	0	2	0	2	C-O
60	3	1	1	1	C-H
61	3	2	1	1	C-H
62	1	1	1	1	C-H
63	1	4	1	1	C-H
64	1	3	1	1	C-H
65	1	2	1	1	C-H
66	2	2	1	1	C-H
67	2	3	1	1	C-H
68	2	1	1	1	C-H
69	1	2	1	1	O-H
70	1	1	1	1	O-H

Continued on next page

Table 1 – continued from previous page

cluster	hydrogen_0	heavy_atom_0	hydrogen_1	heavy_atom_1	label
71	2	3			C-M
72	1	4			C-M
73	0	4			C-M
74	0	5			C-M
75	1	3			C-M
76	0	3			C-M
77	2	2			C-M
78	3	2			C-M
79	0	2			C-M
80	1	2			C-M
81	0	3			O-M
82	0	2			O-M
83	1	2			O-M

DFT Details

A Monkhorst-Pack k-point sampling[?] of 4×4×1 and a planewave cutoff of 450 eV were used. All surface species were modeled using 3.8034 as lattice constant and vacuum of 10.0 Åwith periodic condition. A BFGS algorithm provided by Atomic Simulation Environment (ASE)[?] was applied to the geometry optimization until the maximum force was no more than 0.05 eV/Å. The adsorption energy is calculated as follow:

$$E_{adsorption} = E_{system} - E_{surface} - E_{adsorbate} \tag{1}$$

where $E_{adsorption}$ is the adsorption energy, E_{system} is the total energy of the adsorbate and the Rh slab, $E_{surface}$ is the energy of Rh slab and $E_{adsorbate}$ is the reference energy of adsorbate relative to CH_4, H_2O and H_2.

Confidence Interval Calculation

The 75% confidence of the 0.45 eV criterion is calculated as following (assuming $\hat{E}_1 - \hat{E}_2 > 0.45$ and $\sigma = 0.45eV$):

$$E - \hat{E} = Z \sim Normal(0, \sigma^2) \tag{2}$$

$$E_1 = \hat{E}_1 + Z \sim Normal(\hat{E}_1, \sigma^2) \tag{3}$$

$$E_2 = \hat{E}_2 + Z \sim Normal(\hat{E}_2, \sigma^2) \tag{4}$$

$$E_1 \perp\!\!\!\perp E_2 \Rightarrow E_1 - E_2 \sim Normal(\hat{E}_1 - \hat{E}_2, 2\sigma^2) \tag{5}$$

$$P(E_1 - E_2 < 0) = \Phi(\frac{\hat{E}_2 - \hat{E}_1}{2\sigma^2}) = 1 - \Phi(\frac{\hat{E}_1 - \hat{E}_2}{2\sigma^2}) \tag{6}$$

$$\frac{\hat{E}_1 - \hat{E}_2}{\sqrt{2}\sigma} > \frac{1}{\sqrt{2}} \Rightarrow P(E_1 - E_2 < 0) < 1 - \Phi(\frac{1}{\sqrt{2}}) = 0.25 \tag{7}$$

$$P(E_1 < E_2 | E_1 \hat{-} \hat{E}_2 > 0.45) < 0.25 \tag{8}$$

$$\Rightarrow P(E_1 > E_2 | E_1 \hat{-} \hat{E}_2 > 0.45) > 0.75 \tag{9}$$

The following files are available free of charge. (https://github.com/cchang373/Rh_paper)

- regression: contains OLS, PLS and LDA regression scripts and the DFT regression data in .json file
- all_adsorbates.smi: contains SMILES notation of all adsorbates used in this study
- model: contains the Mol2Vec training data (groups.smi) and the Mol2Vec corpus data (groups.cp)

- spearman: contains Spearman's correlation coefficient calculation script and the DFT, PLS and LDA regression data in .json file

- dft_diff.json: contains SMILES notations of previous lowest energy geometry, new lowest energy geometry and energy difference for the adsorbates found by ML models