RESEARCH ARTICLE doi.org/10.1002/adsc.202401444 ADVANCED SYNTHESIS & CATALYSIS

Data-Driven Analysis of Ni-Catalyzed Semihydrogenations of Alkynes

Miguel Martinez-Fernandez,[+a] Md Bin Yeamin,[+b, c] David Dalmau,[a] Jorge J. Carbó,[b] Albert Poater,[c,*] and Juan V. Alegre-Requena[a,*]

[a] Departamento de Química Inorgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), CSIC-Universidad de Zaragoza, C/ Pedro Cerbuna 12, 50009 Zaragoza, Spain
E-mail: jv.alegre@csic.es
[b] Departament de Química Física i Inorgànica, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[c] Institut de Química Computacional i Catàlisi, Departament de Química, Universitat de Girona, C/ Mª Aurèlia Capmany 69, 17003 Girona, Catalonia, Spain
E-mail: albert.poater@udg.edu

[+] These authors contributed equally

Manuscript received: November 20, 2024; Revised manuscript received: January 26, 2025;
Version of record online: Februar 11, 2025

Supporting information for this article is available on the WWW under https://doi.org/10.1002/adsc.202401444

Abstract: The semihydrogenation of alkynes to alkenes has historically been an essential technique in organic chemistry. In this context, researchers often employ transition metal complexes to achieve this conversion. Given the pronounced polarization of results, often yielding either very high or very low values, it remains challenging to discern the factors influencing reactivity and selectivity in many cases. In this work, we combine different sub-disciplines of digital chemistry with experimental outcomes to rationalize the results of a model Ni-catalyzed semihydrogenation that leads to *E*-alkenes. First, we analyze the main factors behind successful reactions using a machine learning classification model. The descriptors are computed directly from the SMILES strings of the reacting alkynes using an automated protocol that relies on structural features, molecular mechanics, and semi-empirical techniques. This workflow requires minimal human intervention and provides a fast and effective approach. Next, we couple the same descriptors with activation barriers calculated with density functional theory, generating a regression model that explains reactivity based on the properties of the alkyne substrates. Overall, this study demonstrates the potential of using a combination of digital chemistry techniques to uncover reaction trends in Ni-catalyzed semihydrogenations of alkynes, an area where human intuition proves limited in application.

Keywords: semihydrogenation; Ni catalysis; *E*-alkenes; digital chemistry; machine learning; alkyne

Introduction

The selective semihydrogenation of alkynes to alkenes is a relevant and widely-used reductive transformation.[1,2] In this field, catalytic alkyne semihydrogenation by molecular hydrogen (H_2) is an appealing option since the reactions adhere to the green principle of atom economy.[3] The outcome of this process can vary significantly depending on the specific catalyst and reaction conditions employed, resulting in the formation of either olefins[4–10] or alkanes[11] (Figure 1A).

While numerous catalysts have been developed for *Z*-selective semihydrogenation with H_2 since the introduction of the Lindlar catalyst,[12–15] only a few, primarily homogeneous catalysts can provide *E*-

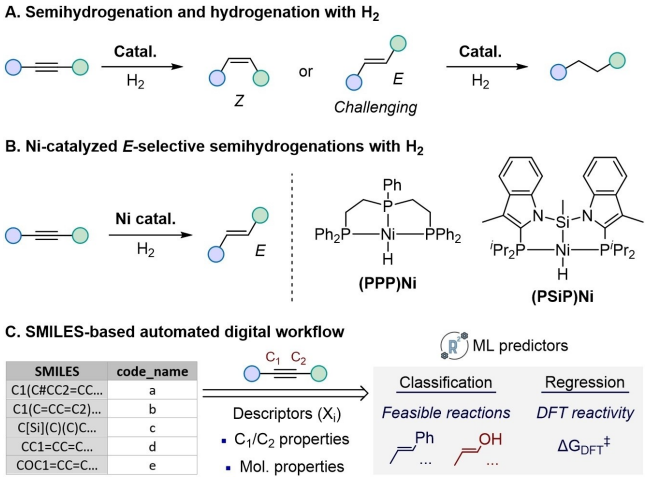

Figure 1. (A) Semihydrogenation and hydrogenation of alkynes. (B) *E*-selective semihydrogenation of alkynes catalyzed with Ni complexes. (C) Summary of the SMILES-based workflows proposed.

alkenes with high selectivity.[16–19] *E*-alkenes are typically generated through the isomerization of *Z*-alkenes, however, they can also be directly formed from alkynes.[20] In this regard, the *E*-selective alkyne semihydrogenations catalyzed by Ni with triphosphine (PPP)[21] and bis(phosphino)silyl (PSiP)[22] ligands represent the first examples of Ni-catalyzed transformations within transition metal complexes (Figure 1B).

In this type of catalysis, it is often difficult to detect trends in reactivity and selectivity using human intuition. For example, the results obtained with (PSiP)Ni[22] show that each product is obtained either in very high or very low yield, with no clear structural patterns to rationalize why substrates fall into either category. Additionally, the polarization of results into two distinct groups makes it difficult to recognize any trends in reactivity based on functional groups, since most molecules that react favorably show yields within a very narrow range.

In the last decade, researchers have increasingly integrated data-driven routines and machine learning (ML) to understand experimental results. This field has proven to be helpful in creating more efficient generations of catalysts,[23–26] discovering drug candidates,[27,28] and designing new functional materials,[29,30] among other applications.[31] These statistical protocols have also been widely used to understand reactivity and selectivity trends, typically employing density functional theory (DFT) descriptors in low-data regimes.[32–34] Nevertheless, and mainly due to their recent rapid evolution, faster alternatives such as xTB,[35] steric descriptors,[36] and RDKit[37] descriptors are gaining popularity. While such cost-efficient techniques are commonly employed in big-data problems, their introduction to regimes with less datapoints is of great interest to the scientific community.

In this work, we employed rapid and automated digital workflows to detect trends in reactivity and selectivity in the (PSiP)Ni-catalyzed semihydrogenation of alkynes.[22] Two ML models are used: one to understand the key parameters behind experimental reaction feasibility and another to analyze reactivity in successful reactions. The models presented rely on cost-effective descriptors derived from SMILES strings and are designed to uncover trends that may not be evident to human intuition.

Results and Discussion

Generation of the descriptor database. First, we compiled a database of descriptors for the initial alkynes (Figure S1)[22] to serve as input for the two target ML models. Choosing meaningful descriptors is a critical step that significantly affects the quality of these models. Starting from a comma-separated values (CSV) file containing SMILES strings, we executed an automated AQME[38]-ROBERT[39] workflow, which included RDKit conformational searches, xTB geometry optimizations, and RDKit-xTB descriptor generation. This workflow only required one command line, minimizing effort and reducing human errors in data manipulation.[40] The resulting descriptors covered a wide variety of molecular properties, including structural, electronic, and steric parameters (Figure 2).

To mitigate the limitations of cost-effective methods, we included atomic descriptors from the two alkynyl carbon atoms, since adding relevant local features has proven to be beneficial in ML studies.[39] These descriptors contain electronic and steric properties calculated using xTB methods, and they are stored as their Boltzmann-averaged values to better represent the molecules in solution and avoid conformer-specific values (Figure 2).[41]

We encountered an identification challenge when assembling the database of descriptors for the two alkynyl carbon atoms. In the workflow, the Boltzmann-weighted descriptors for each atom are initially

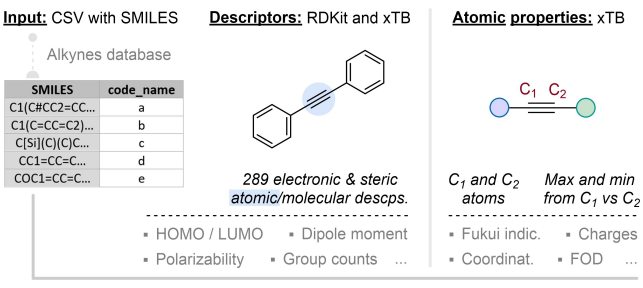

Figure 2. Automated generation of descriptors starting from SMILES strings.

stored using two identifiers, C1_descriptor and C2_descriptor. However, this arbitrary distribution is based on RDKit numbering, making it unclear how to categorize the atoms as C1 and C2. To avoid biases deriving from atom numbering, we followed a more meaningful strategy based on maximum and minimum values. In this approach, the maximum value of a descriptor for C1 and C2 was stored in the database as max_descriptor, while the other was stored as min_descriptor (Figure 2).

Overall, 285 different molecular and atomic descriptors were stored in a CSV database that can be readily used as input to develop ML models. In order to select descriptors rationally from all the possible options, ROBERT executes an automated selection process. The process begins with the elimination of correlated descriptors, after which a model is produced. To provide an additional option, a second model is created using the most influential descriptors identified through permutation feature importance (PFI) from the first model. In this work, of the two alternatives generated, we selected the PFI-filtered models because they showed better ROBERT scores.

Experimental trends: classification ML model. Next, we combine the database and the experimental values of the (PSiP)Ni-catalyzed reaction to create a classification model. Both descriptor generation and ML model screening are part of the same ROBERT workflow starting from SMILES strings.[39] We chose a classification problem because the experimental yields exhibit a bimodal distribution, with the products showing either very high or very low values (Figure 3). In this scenario involving 39 different products, human intuition is of limited use for understanding the results, as there are no clear electronic, steric, or structural patterns that explain the observed reactivity and selectivity trends.[22]

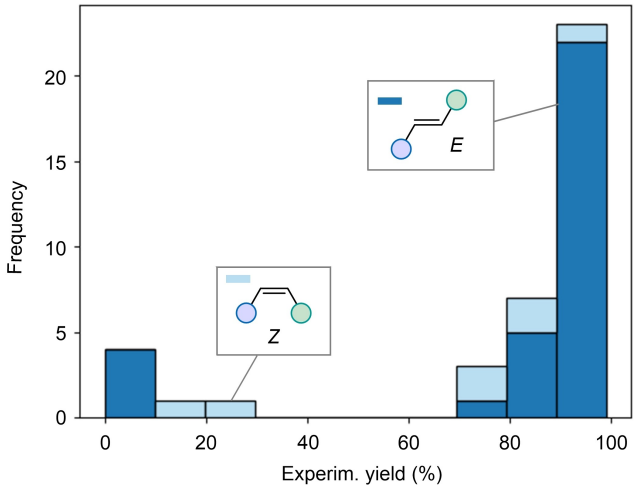

Figure 3. Distribution of experimental yields in the model Ni catalysis.

In the ML classification model, reactions that lead to the *E* isomer with yields over 70% are considered successful (y = 1), while those leading to the *Z* isomer or with yields below 30% are considered unsuccessful (y = 0). One of the substrates (**p**, Figure S1) was removed because it contained two alkyne groups, which were incompatible with the version of AQME used during the creation of the descriptor database. Then, we merged the descriptors and yields of the remaining 38 alkynes and conducted a model screening with ROBERT, considering four types of algorithms and four partition sizes. This workflow was automated, requiring only a single command line to eliminate any human biases when selecting models. The optimal combination included an AdaBoost (ADAB) algorithm with a training-to-validation ratio of 6:4 and two descriptors, leading to significantly good results: an accuracy of 0.94, a balanced F-score (F1 score) of 0.96, and a Matthew's correlation coefficient (MCC) of 0.79 in the 16 datapoints of the validation set (Figure 4A).

No considerable overfitting was observed in the leave-one-out cross-validation test (LOOCV; Figure 4B, left). To ensure the model captures meaningful trends in the data, various "flawed" models were tested, including: (i) y-mean, which assesses accuracy when predicted y values are fixed to the major class; (ii) y-shuffle, which measures accuracy using a model trained on randomly shuffled y values; and (iii) one-hot, which evaluates accuracy when all descriptors are replaced with binary values (0 s and 1 s).[42] In all three cases, the resulting accuracies were significantly lower than those of the original model and the LOOCV (Figure 4B, right).

The predictive ability of the model was further assessed using the ROBERT score (Figure 4C). This score, rated on a scale of ten, is designed to offer users insights into the predictive capabilities of models by considering multiple factors such as the correlation between predictions and measurements, human interpretability, error distribution, sensitivity to features, and avoidance of overfitting and underfitting.[39] The score of 7 achieved is a decent result for this low-data scenario, suggesting that this classification model could be used as a quick check to determine whether new reagents will form the *E* alkene products.

Lastly, we focused on the two descriptors that comprised the model: the Fermi level of the molecules and the Fukui positive index (f_C^+) of the alkynyl carbon atoms. Both features have similar PFI in the model (Figure 4D) and were selected automatically during the data curation and model screening workflows of ROBERT among the initial 285 descriptors. From a human intuition perspective, these two descriptors appear relevant. For example, the Fermi

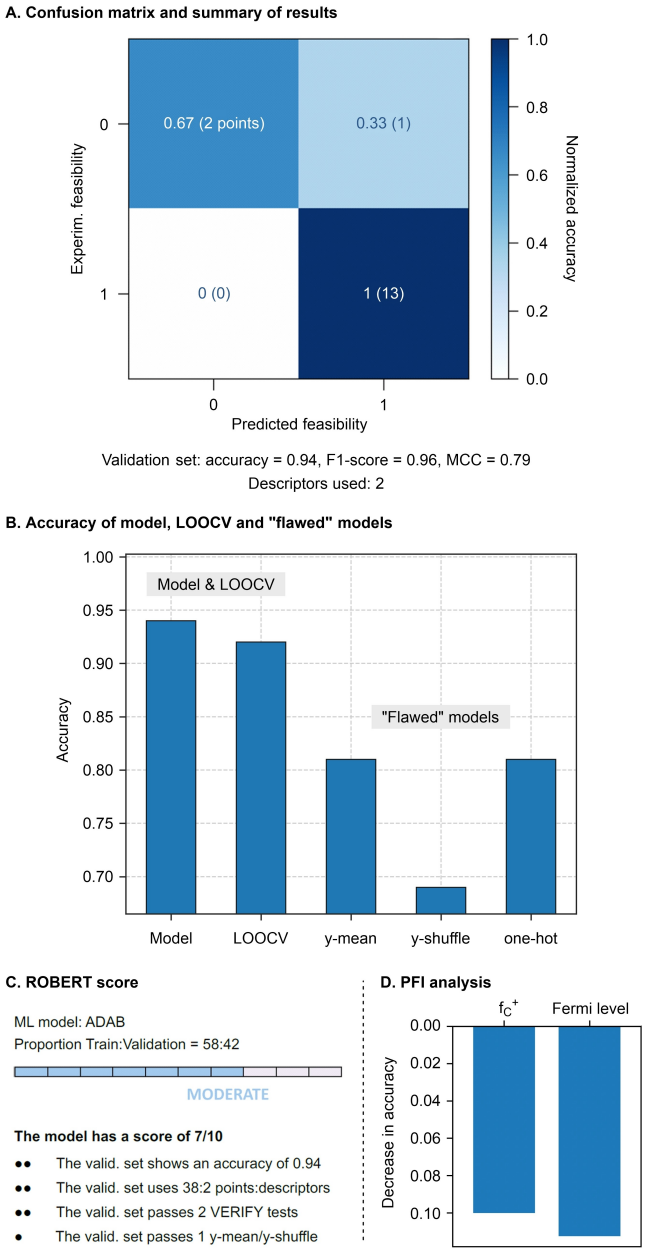

Figure 4. (A) Confusion matrix obtained for the validation set using the optimal classification ML model, including normalized results with the total count of points in parentheses. (B) Accuracy of the model, LOOCV and "flawed" models. (C) Summary of the ROBERT score obtained. (D) PFI analysis of all the variables.

level is related to electronic levels and might influence the C–Ni bond strength and the amount of charge transfer[43] between the Ni center and alkyne occurring at relevant reaction steps.[44] Similarly, the f_C^+ is related to the electrophilicity of carbon atoms,[45] which aligns with the mechanism proposed[44] since H_2 dissociates and one of the hydrogens is transferred to an accepting carbon atom.

Overall, even though reaching a conclusion on the exact origin of the reactivity might be too ambitious, the results suggest that electronic effects prevail over steric factors in this transformation. To further strengthen this conclusion, we developed a new model using a database containing f_C^+ and Fermi levels, as well as multiple steric descriptors calculated with MORFEUS,[46] including buried volumes and pyramidalization of the alkynyl carbon atoms (Figure S2). In principle, if steric parameters were a significant factor influencing reactivity, a notable improvement in the model's performance would be expected. However, the predictions remained identical, supporting the initial findings.

Reactivity trends among successful reactions: regression ML model. The reactivity trends of the model Ni catalysis were also analyzed, focusing on the *E*-alkene isomer. Nevertheless, the distribution shown in Figure 3 suggests that experimental yields may not be effective for developing ML strategies to analyze these trends. For example, a classification model would not be suitable since most of the *E*-alkenes are obtained with very good yields, resulting in an imbalanced model. Similarly, a regression model would not be robust since most reactions fall within a narrow yield range of 78% to 99%, missing important information from the low yield range.

This distribution, biased towards high yields, is likely a consequence of how experimental studies are designed, as researchers often aim to find a set of conditions that work as well as possible with the scope of substrates.[47] As an alternative strategy to create a balanced reactivity scale, we used DFT-calculated activation energies since these values tend to be more varied among different substrates. This digital approach relies on the DFT mechanistic study by Ke and coworkers,[44] which suggested that the rate-determining step (RDS) is the H_2 metathesis that forms the alkene (Figure 5). This transition state occurs favorably from the *Z* isomer, which then undergoes rapid isomer-

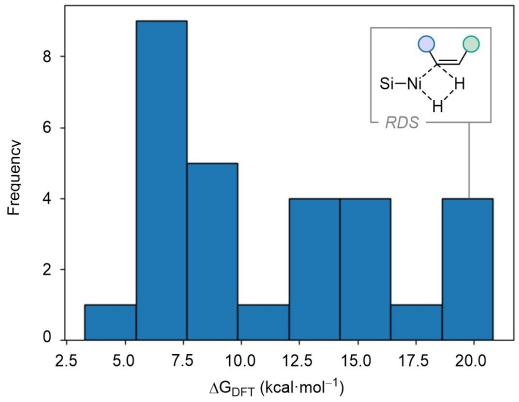

Figure 5. Distribution of calculated $\Delta G^{\ddagger}$ (kcal·mol^{-1}).

ization to reach the final *E* alkene product. Using the individual reagents as the energy reference, the Gibbs free energy activation barriers ($\Delta G^{\ddagger}$) of the RDS were calculated for the 29 substrates that led to the *E* isomer. The resulting population is more evenly distributed than in the previous case and, therefore, we opted to use a regression model to study the trends.

At this point, we combined the descriptors obtained previously with the $\Delta G^{\ddagger}$ values calculated using DFT, and employed ROBERT to screen different ML models. The optimal combination identified was a Random Forest (RF) algorithm with a training-to-validation ratio of 7:3 and one descriptor. This model yielded very good results, with a coefficient of determination (R^2) of 0.95, a mean absolute error (MAE) of 0.91 kcal·mol^{-1}, and a root mean squared error (RMSE) of 1.20 kcal·mol^{-1} in the nine data points of the validation set (Figure 6A). As in the previous case, no considerable overfitting was observed in the LOOCV (Figure 6B, left) and it captured meaningful data trends, as both its RMSE and LOOCV RMSE values were significantly lower than those of the three "flawed" models (Figure 6B, right).[42] The resulting ROBERT score of 9 is robust and further indicates that the model captures efficiently the trends of the data.

The optimal algorithm only kept one descriptor from the initial 285 descriptors, the BalabanJ.[48] This feature is a topological index that assigns a numerical value to the structural complexity of molecules, establishing a structure-property relationship that is highly relevant to the algorithm used to understand the $\Delta G^{\ddagger}$ trends. Indeed, the representation of BalabanJ values vs $\Delta G^{\ddagger}$ already shows a moderate correlation (R^2 of 0.77), with two differentiated groups of substrates that have a reactivity cliff[49] near 12 kcal·mol^{-1} (Figure 7). A closer inspection to the points with higher barriers revealed that the trimethylsilyl group (TMS) leads to higher RDS, which was not possible to determine with experimental results.

The BalabanJ parameter shows a moderate correlation with the maximum buried volume at the alkynyl carbon atoms ($R^2 = 0.73$, Figure S3A). Additionally, the plot of buried volume versus $\Delta G^{\ddagger}$ shows a similar reactivity cliff to the previous case with BalabanJ (Figure S3B, $R^2 = 0.65$). These findings suggest that steric effects significantly influence reaction rates in reactions leading to the *E*-alkene isomer.

Conclusions

We used different techniques of digital chemistry to analyze the results of an *E*-selective (PSiP)Ni-catalyzed semihydrogenation of alkynes. The generation of a comprehensive descriptor database for the initial alkynes was an important step in developing accurate

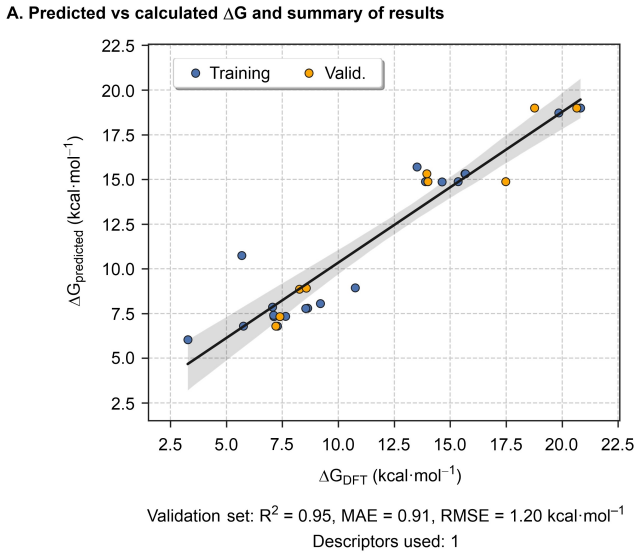

Figure 6. (A) Calculated vs predicted $\Delta G^{\ddagger}$ (kcal·mol^{-1}) obtained with the optimal ML regression model. (B) RMSE of model, LOOCV and "flawed" models.

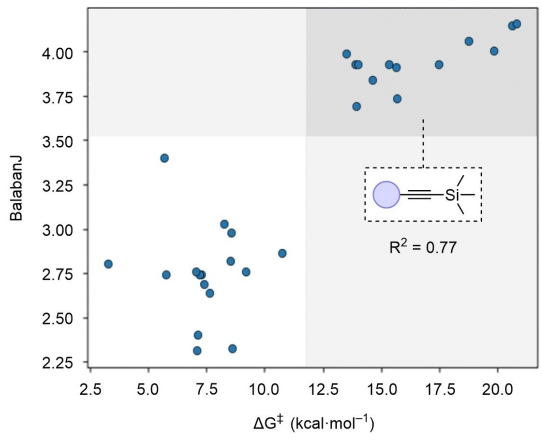

Figure 7. Representation of BalabanJ vs $\Delta G^{\ddagger}$ (kcal·mol^{-1}), along with its R^2.

ML models. By using an automated AQME workflow, we compiled a diverse set of molecular properties from SMILES strings, including structural, electronic, and steric descriptors. This automated approach minimized human error and ensured consistency, resulting in a robust dataset essential for model training. Furthermore, the addition of atomic descriptors from the alkynyl carbon atoms, which are crucial in the catalytic transformation, enhanced the effectivity of the database towards ML modeling. These descriptors were averaged using Boltzmann populations to better represent the molecular properties in solution, avoiding biases from conformer-specific values. We addressed potential issues with atom numbering by categorizing descriptors based on their maximum and minimum values, aiming to improve the chances for obtaining meaningful ML models.

Our classification ML model effectively predicted the feasibility of the (PSiP)Ni-catalyzed reactions, achieving high accuracy (0.94), balanced F1 score (0.96) and MCC (0.79). The model identified the Fermi level and the f_C^+ as key descriptors, suggesting that electronic effects dominate over steric factors in this transformation.

For reactivity trends, we utilized a regression model with DFT-calculated activation energies, which provided a balanced representation of the reaction barriers. The optimal RF algorithm yielded excellent predictive performance (R^2 of 0.95, MAE and RMSE of 0.91 and 1.20 kcal·mol^{-1}, respectively), while requiring only one descriptor, the BalabanJ topological index. The organization of the BalabanJ values suggests that the hydrogen transfer taking place at the RDS might be disfavored by steric effects in the alkynyl carbon atoms.

Computational Details

Computational chemistry calculations. DFT calculations were conducted with Gaussian 16.[50] Geometry optimizations were performed at the B3LYP[51–54]/def2SVP[55,56] level of theory. Vibrational frequency calculations were used to confirm that stationary points were either minima or first-order saddle points on the potential energy surface and to obtain frequencies used to calculate thermochemistry values with the GoodVibes[57] program. Electronic energies were refined using the B3LYP-D3[58]/def2TZVP method. In all cases, the calculations included the integral equation formalism variant of the polarizable continuum model (IEF-PCM)[59,60] solvation model (solvent = benzene) to account for solvent effects.

For simplicity, we calculated the reaction barriers starting from the separated reagents and catalyst. These values allow us to perform a relative comparison of reaction rates while requiring less time than calculating reaction barriers from the resting state intermediate.

GoodVibes thermochemistry analysis. The reported Gibbs free energy activation barriers at the B3LYP-D3/def2TZVP//B3LYP/def2SVP (IEFPCM, benzene) level of theory were calculated with GoodVibes (v3.0.2).[57] We used this program to introduce quasi-harmonic (QHA) corrections to the computed vibrational entropies using a frequency cut-off value of 100.0 cm^{-1}, following the model proposed by Grimme[61] at 298.15 K (temperature from the experimental reaction). Also, a correction for the change in standard state from gas phase at 1 atm to a 1 M solution was introduced (option "-c 1" in GoodVibes).[62] A few of the calculations showed persistent imaginary frequencies lower than 50 cm^{-1}, which were inverted to their respective positive values before the QHA entropic corrections were computed (option "–invertifreq −50"), as seen in previous examples.[63] Entropy corrections due to entropy of symmetry (option "–ssym"), mixing, and multi-structural effects (option "–pes") were also included.

All the thermochemical data including absolute energies, zero-point energies (ZPE) and T·S, among other parameters, at the B3LYP/def2SVP level, as well as the absolute energies, corrected final G and relative G obtained with B3LYP-D3/def2TZVP, were generated in an automated way using GoodVibes and tabulated in a separate file of the electronic supplementary information (ESI, *Thermochemistry.dat* in the *Extra_ESI.zip* file). Molecular coordinates were generated similarly using the "–xyz" option (*Molecular_coordinates.xyz* in the *Extra_ESI.zip* file). The command line used to run GoodVibes was:

" python -m goodvibes -c 1 –spc SP –pes 0_dG_alkynes.yaml –xyz –ssym –imag –invertifreq −50 *.log "

Descriptor generation and ML workflows. The ROBERT program (v1.0.4)[39] was used to generate Boltzmann-averaged descriptors from a CSV file containing SMILES strings of molecules. This automated workflow initially executes AQME (v1.5.2)[38] to perform conformational searches with RDKit, geometry optimization with xTB, and descriptor generation with RDKit and xTB. Both molecular and atomic descriptors were considered, and all the data was stored in a CSV database, which underwent data curation, ML model screening, testing model reliability, and feature importance analysis, among other protocols. The command line used to execute ROBERT was:

" python -m robert –csv_name "FILENAME.csv" –y "TARGET" –aqme –qdescp_keywords "–qdescp_atom [C#C] –qdescp_solvent benzene" "

In this command FILENAME.csv refers to the initial CSV file of SMILES strings containing either classification or regression results and TARGET is the target value to predict ("reaction" for classification and "dG" for regression). The option " –qdescp_keywords "–qdescp_atom [C#C] –qdescp_solvent benzene" " specifies the creation of atomic descriptors for the two alkynyl carbon atoms and accounts for solvent effects during xTB optimization and descriptor generation

After each run, ROBERT generated PDF reports containing comprehensive information about the models, enhancing transparency and providing instructions for reproducibility. These files are available as part of the ESI (in the *Extra_ESI.zip* file).

LOOCV was carried out separately as it was not yet implemented in ROBERT v1.0.4. The results are shown in the ESI document.

Data and Code Availability

All protocols followed in this work are detailed in the *Computational Details* section. For each representation shown in the manuscript, we have included tables with their raw values in the ESI document. The raw thermochemical data from GoodVibes, used for calculating DFT activation barriers, is also available in the ESI, along with the molecular coordinates of all systems (in the *Extra_ESI.zip* file).

Additionally, the descriptor and SMILES databases for each ML model have been uploaded as individual CSV files in the ESI (in the *Extra_ESI.zip* file). The PDF reports from ROBERT, containing comprehensive information about the workflows, are also available in the ESI (in the *Extra_ESI.zip* file).

Acknowledgements

J.V.A.-R., M.M.-F. and D.D. acknowledge Gobierno de Aragón-Fondo Social Europeo (Research Groups E07_23R and E17_23R) and the State Research Agency of Spain (MCIN/AEI/10.13039/501100011033/FEDER, UE) for financial support (PID2022-140159NA-I00). J.V.A.-R. and D.D. acknowledge the computing resources at the Galicia Supercomputing Center, CESGA, including access to the FinisTerrae supercomputer, the Red Española de Supercomputación (grant number QH-2023-2-0003) and the Drago cluster facility of SGAI-CSIC. D.D. thanks Gobierno de Aragón-FSE for a PhD fellowship (2021–2025). A.P. and J.J.C. thank the Spanish Ministerio de Ciencia e Innovación for projects PID2021-127423NB-I00 and PID2021-128128NB-I00, and the Generalitat de Catalunya for projects 2021SGR623 and 2021SGR00110. A.P. is a Serra Húnter Fellow and ICREA Academia Prize 2019. M.B.Y acknowledges the Margarita Salas Grant 2021URV-MS-12 from Rovira I Virgili University (URV) and CSA Trust Grant 2023.

Conflict of Interest

There are no conflicts to declare.

References

[1] K. K. Swamy, A. S. Reddy, K. Sandeep, A. Kalvani, *Tetrahedron Lett.* **2018**, *59*, 419–429.
[2] D. Decker, H. J. Drexler, D. Heller, T. Beweries, *Catal. Sci. Technol.* **2020**, *10*, 6449–6463.
[3] T. M. Saunders, S. B. Shepard, D. J. Hale, K. N. Robertson, L. Turculet, *Chem. Eur. J.* **2023**, *29*, e202301946.
[4] R. Kusy, M. Lindner, J. Wagner, K. Grela, *Chem Catal.* **2022**, *2*, 1346–1361.
[5] R. Kusy, K. Grela, *Green Chem.* **2021**, *23*, 5494–5502.
[6] A. Poater, *Chem Catal.* **2022**, *2*, 1245–1246.
[7] T.-J. Hu, M. Jaber, G. Trans, D. Bouyssi, N. Monteiro, A. Amgoune, *Chem. Eur. J.* **2023**, *29*, e202301636.
[8] Y.-T. Su, X. Wang, Q.-W. Lin, Q. Shen, S.-W. Xu, L.-P. Fang, X. Wen, *Catal. Sci. Technol.* **2023**, *13*, 1718–1724.
[9] D. K. Pandey, E. Khaskin, S. Pal, R. R. Fayzullin, J. R. Khusnutdinova, *ACS Catal.* **2023**, *13*, 375–381.
[10] B. J. Gregori, M.-O. W. S. Schmotz, A. J. von Wangelin, *ChemCatChem* **2022**, *14*, e202200886.
[11] S. E. Sloane, A. Reyes, Z. P. Vang, L. Li, K. T. Behlow, J. R. Clark, *Org. Lett.* **2020**, *22*, 9139–9144.
[12] L. Ling, C.-Y. Hu, L.-H. Long, X. Zhang, L.-X. Zhao, L. L. Liu, H. Chen, M.-M. Luo, X.-M. Zeng, *Nat. Commun.* **2023**, *14*, 990.
[13] A. Kathó, H. H. Horváth, G. Papp, F. Joó, *Catalysts* **2022**, *12*, 518.
[14] M.-Y. Lee, C. Kahl, N. Kaeffer, W. Leitner, *JACS Au* **2022**, *2*, 573–578.
[15] A. Fürstner, C. Mathes, C. W. Lehmann, *Chem. Eur. J.* **2001**, *7*, 5299–5317.
[16] N. O. Thiel, B. Kaewmee, T. Tran Ngoc, J. F. Teichert, *Chem. Eur. J.* **2020**, *26*, 1597–1603.
[17] A. Torres-Calis, J. J. García, *Catal. Sci. Technol.* **2022**, *12*, 30043015.
[18] R. A. Farrar-Tobar, S. Weber, Z. Csendes, A. Ammaturo, S. Fleissner, H. Hoffmann, L. F. Veiros, K. Kirchner, *ACS Catal.* **2022**, *12*, 2253–2260.
[19] D. Srimani, Y. Diskin-Posner, Y. Ben-David, D. Milstein, *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 14131–14134.
[20] Y. Wu, Y. Ao, Z. Li, C. Liu, J. Zhao, W. Gao, X. Li, H. Wang, Y. Liu, Y. Liu, *Nat. Commun.* **2023**, *14*, 1655.
[21] K. Murugesan, C. B. Bheeter, P. R. Linnebank, A. Spannenberg, J. N. H. Reek, R. V. Jagadeesh, M. Beller, *ChemSusChem* **2019**, *12*, 3363–3369.
[22] D. J. Hale, M. J. Ferguson, L. Turculet, *ACS Catal.* **2022**, *12*, 146–155.
[23] Y. Chen, R. Li, H. Suo, C. Liu, *ACS ES&T Eng.* **2021**, *1*, 1246–1257.
[24] Y. Chen, J. Feng, X. Wang, C. Zhang, D. Ke, H. Zhu, S. Wang, H. Suo, C. Liu, *Environ. Sci. Technol.* **2023**, *57*, 18080–18090.
[25] S. Escayola, N. Bahri-Laleh, A. Poater, *Chem. Soc. Rev.* **2024**, *53*, 853–882.
[26] R. Monreal-Corona, A. Pla-Quintana, A. Poater, *Trends Chem.* **2023**, *5*, 935–946.
[27] A. Lavecchia, *Drug Discovery Today* **2015**, *20*, 318–331.
[28] L. Zhang, J. Tan, D. Han, H. Zhu, *Drug Discovery Today* **2017**, *22*, 1680–1685.
[29] A. Merchant, S. Batzner, S. S. Schoenholz, et al., *Nature* **2023**, *624*, 80–85.
[30] S. Axelrod, D. Schwalbe-Koda, S. Mohapatra, J. Damewood, K. P. Greenman, R. Gómez-Bombarelli, *Acc. Mater. Res.* **2022**, *3*, 343–357.
[31] N. Sanosa, D. Dalmau, D. Sampedro, J. V. Alegre-Requena, I. Funes-Ardoiz, *Artif. Intell. Chem.* **2024**, *2*, 100068.
[32] B. C. Haas, N.-K. Lim, J. Jermaks, E. Gaster, M. C. Guo, T. C. Malig, J. Werth, H. Zhang, F. D. Toste, F. Gosselin, S. J. Miller, M. S. Sigman, *J. Am. Chem. Soc.* **2024**, *146*, 8536–8546.

[33] H. D. Clements, A. R. Flynn, B. T. Nicholls, D. Grosheva, S. J. Lefave, M. T. Merriman, T. K. Hyster, M. S. Sigman, *J. Am. Chem. Soc.* **2023**, *145*, 17656–17664.
[34] A. Modak, J. V. Alegre-Requena, L. de Lescure, K. J. Rynders, R. S. Paton, N. J. Race, *J. Am. Chem. Soc.* **2022**, *144*, 86–92.
[35] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *WIREs Comput. Mol. Sci.* **2021**, *11*, e1493.
[36] G. Luchini, T. Patterson, R. S. Paton, *DBSTEP: DFT Based Steric Parameters*. **2022**, DOI: 10.5281/zenodo.4702097.
[37] G. Landrum, *RDKit: Open-Source Cheminformatics* **2010**. https://www.rdkit.org.
[38] J. V. Alegre-Requena, S. V. S. Sowndarya, R. Pérez-Soto, T. M. Alturaifi, R. S. Paton, *WIREs Comput. Mol. Sci.* **2023**, *13*, e1663.
[39] D. Dalmau, J. V. Alegre Requena, *WIREs Comput. Mol. Sci.* **2024**, *14*, e1733.
[40] D. Dalmau, J. V. Alegre-Requena, *Trends Chem.* **2024**, *6*, 459–469.
[41] M. S. Baidun, A. V. Kalikadien, L. Lefort, E. A. Pidko, *J. Phys. Chem. C* **2024**, *128*, 7987–7998.
[42] For a more detailed explanation of the three "flawed" models used, see https://robert.readthedocs.io/en/latest/Modules/verify.html.
[43] A. Stefancu, S. Lee, L. Zhu, M. Liu, R. C. Lucacel, E. Cortés, N. Leopold, *Nano Lett.* **2021**, *21*, 6592–6599.
[44] Y.-W. Li, H.-Y. Liang, Y.-B. Liu, J.-X. Lin, Z.-F. Ke, *ACS Catal.* **2023**, *13*, 13008–13020.
[45] R. Pucci, G. G. N. Angilella, *Foundat. Chem.* **2022**, *24*, 59–71.
[46] K. Jorner, MORFEUS. https://github.com/kjelljorner/morfeus/.
[47] J. M. Cole, *Nat. Chem.* **2022**, *14*, 973–975.
[48] A. T. Balaban, *Chem. Phys. Lett.* **1981**, *89*, 399–404.
[49] S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman, A. G. Doyle, *Science* **2021**, *374*, 301–308.
[50] Gaussian 16, Revision C.01, Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., Nakatsuji, H., Li, X., Caricato, M., Marenich, J., Bloino, A., Janesko, B. G., Gomperts, R., Mennucci, B., Hratchian, H. P., Ortiz, J. V., Izmaylov, A. F., Sonnenberg, J. L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V. G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery, Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Keith, T., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Millam, J. M., Klene, M., Adamo, C., Cammi, R., Ochterski, J. W., Martin, R. L., Morokuma, K., Farkas, O., Foresman, J. B. & Fox, D. J. Gaussian, Inc., Wallingford CT, 2016.
[51] S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* **1980**, *58*, 1200–1211.
[52] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.
[53] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.
[54] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.* **1994**, *98*, 11623–11627.
[55] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
[56] F. Weigend, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065.
[57] G. Luchini, J. V. Alegre-Requena, I. Funes-Ardoiz, R. S. Paton, *F1000Research* **2020**, *9*, 291.
[58] S. Grimme, J. Antony, S. Ehrlich, H. A. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.
[59] E. Cances, B. Mennucci, J. Tomasi, *J. Chem. Phys.* **1997**, *107*, 3032–3041.
[60] J. Tomasi, B. Mennucci, E. Cances, *J. Mol. Struct.* **1999**, *464*, 211–226.
[61] S. Grimme, *Chem. Eur. J.* **2012**, *18*, 9955–9964.
[62] V. S. Bryantsev, M. S. Diallo, W. A. Goddard III, *J. Phys. Chem. B* **2008**, *112*, 9709–9719.
[63] R. Sure, S. Grimme, *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.