Supplementary Information

Using Machine Learning to Predict Chemical Reaction Outcomes in Environmental Applications

Dhiully P. S. Silva,[a] Fabiana F. F. de Godoi,[b] Matheus M. P. da Silva,[c] Isabella A. Guedes,[c] Laurent E. Dardenne,[c] Raphael de A. Gomes[a] and Fernando Schimidt*[,a]

[a]Programa de Mestrado Profissional em Tecnologia, Gestão e Sustentabilidade, Instituto Federal de Educação, Ciência e Tecnologia de Goiás, 74055-110 Goiânia-GO, Brazil

[b]Programa de Doutorado em Inovação Farmacêutica, Universidade Federal de Goiás, 74605-170 Goiânia-GO, Brazil

[c]Grupo de Modelagem Molecular de Sistemas Biológicos, Laboratório Nacional de Computação Científica, 25651-075 Petrópolis-RJ, Brazil

*e-mail: fernando.quimica@ifg.edu.br

*ORCID ID 0000-0003-4094-4762

The equations for the model evaluation metrics, available in the Scikit-Learn library:[1]

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{(S1)}$$

$$\text{Recall} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{\text{TP}}{\text{TP+FN}} \quad \text{(S2)}$$

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positives}} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{(S3)}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}} = \frac{2\text{TP}}{2\text{TP+FP+FN}} \quad \text{(S4)}$$

Where, TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

ROC AUC Score

The ROC is Receiver Operating Characteristic Curve and AUC is Area Under Curve. An ROC curve is constructed, plotted between True Positive Rate (TPR) and False Positive Rate (FPR), i.e., TPR on the y-axis and FPR on the x-axis. AUC is the area under the ROC curve. An excellent classifier has an AUC value close to 1, while a low-performing classifier has an AUC value close to 0. The higher the AUC value, the better the classification model is at predicting classes.

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \text{ (S5)}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \text{ (S6)}$$

In this work, we measured this parameter using the roc_auc_score function available in the Scikit-Learn library.[1]

Chemical reaction patterns (see Table 2) assembled in Marvin Sketch software and its corresponding SMARTS code.

Reactions with •OH radical

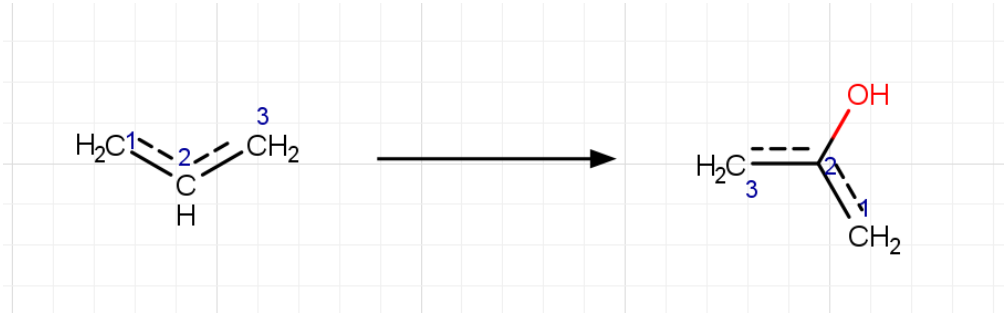

Figure S1. Attack on the double bonds of aromatic chains (reaction #1).

[c:1]:[c:2]:[c:3]>>[#8]-[c:2](:[c:1]):[c:3]

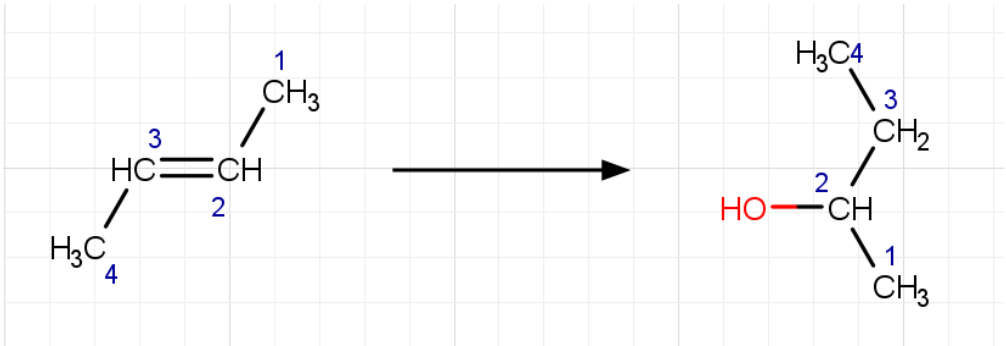

Figure S2. Attack on the double bonds of aliphatic chains (reaction #2).

[#6:1]\[#6:2]=[#6:3]\[#6:4]>>[#6:4]-[#6:3]-[#6:2](-[#6:1])-[#8]

Reactions with ozone

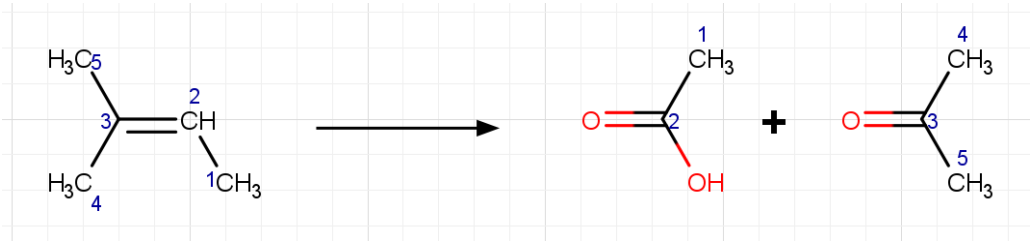

Figure S3. Attack on double bonds of aliphatic chains (reaction #3).

[#6:1]\[#6:2]=[#6:3](/[#6:4])-[#6:5]>>[#6:1]-[#6:2](-[#8])=O.[#6:4]-[#6:3](-[#6:5])=O

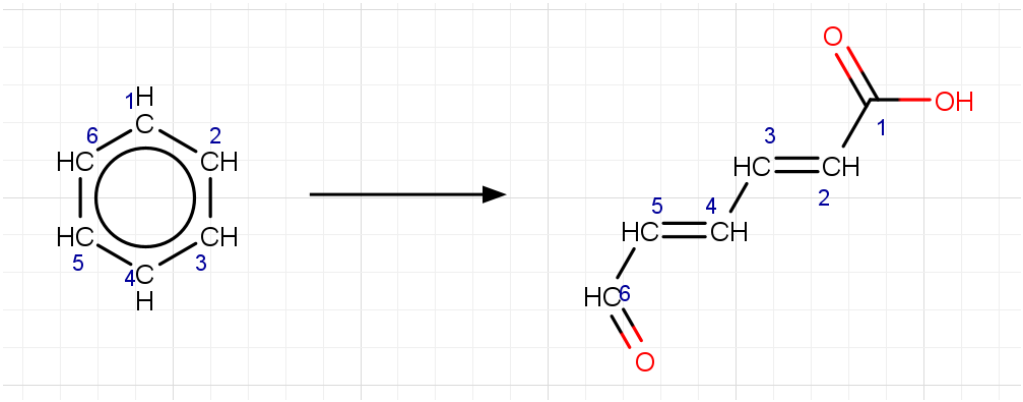

Figure S4. Attacks double bonds in aromatic chains (reaction #4).

[c:1]1[c:2][c:3][c:4][c:5][c:6]1>>[#8]-[#6:1](=O)\[#6:2]=[#6:3]\[#6:4]=[#6:5]\[#6:6]=O

Reactions with atomic oxygen

Reactions with aliphatic chains, three distinct mechanisms and products.

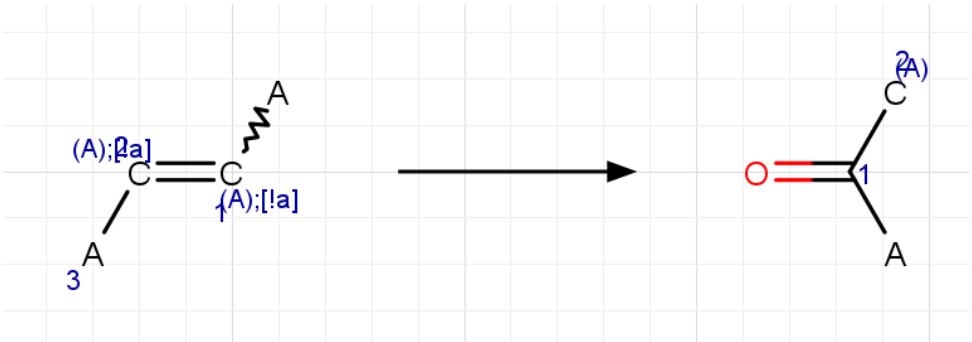

Figure S5. Reactions with aliphatic chains, mechanism A (reaction #5).

[*:0]-[C;!a:1]=[C;!a:2]-[*:3]>>[C:2]-[C:1](-[*:0])=O

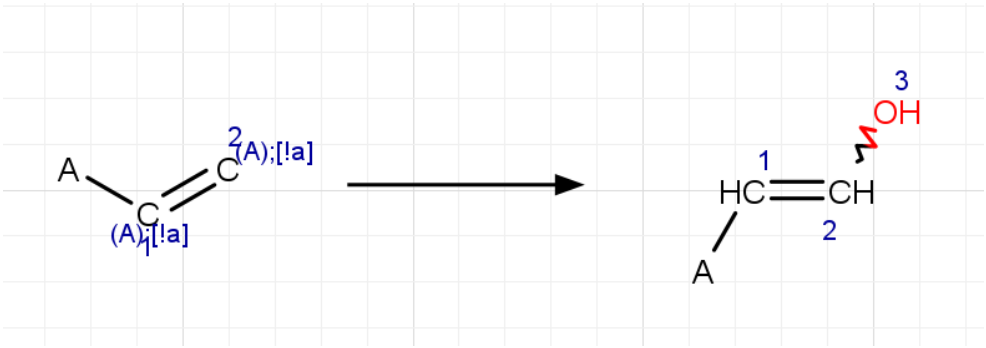

Figure S6. Reactions with aliphatic chains, mechanism B (reaction #6).

[*:0]-[C;!a:1]=[C;!a:2]>>[O:3]-[C:2]=[C:1]-[*:0]

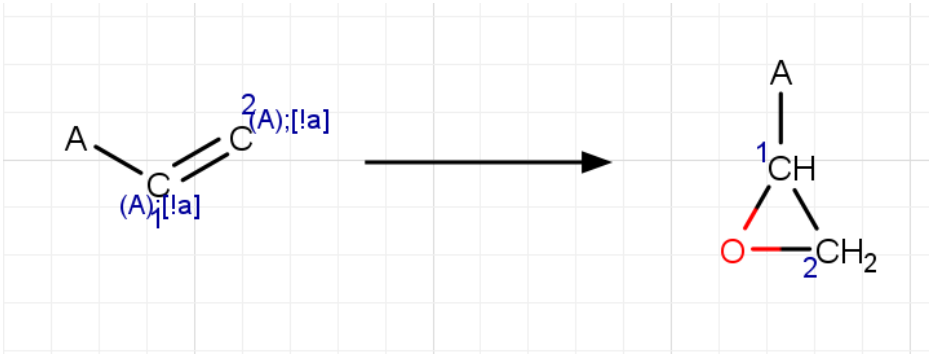

Figure S7. Reactions with aliphatic chains, mechanism C (reaction #7).

[*:0]-[C;!a:1]=[C;!a:2]>>[*:0]-[C:1]-1-[C:2]-[O]-1

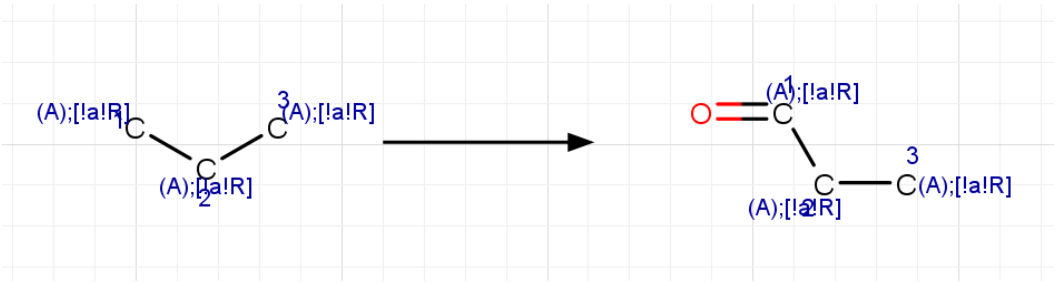

Figure S8. Attack on single bonds in alkanes (reaction #8).

[C;!a;!R:1]-[C;!a;!R:2]-[C;!a;!R:3]>>[C;!a;!R:3]-[C;!a;!R:2]-[C;!a;!R:1]=O

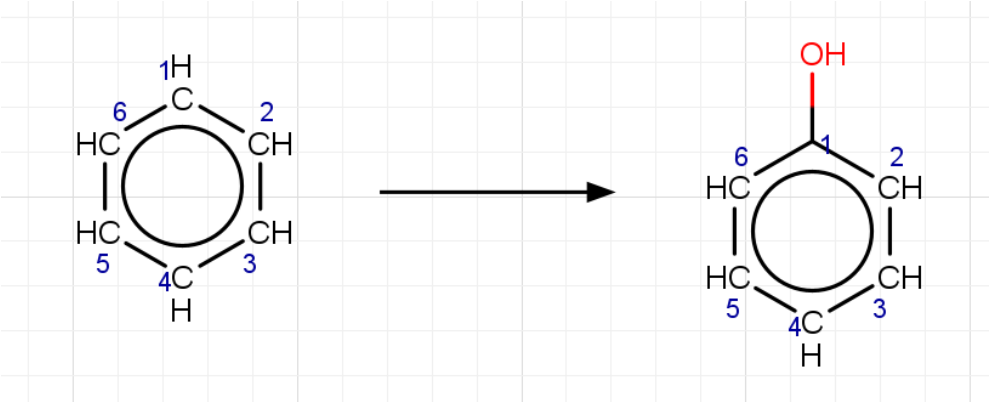

Figure S9. Attack on aromatic hydrocarbons (reaction #9).

[c:1]1[c:2][c:3][c:4][c:5][c:6]1>>[c:1]1([O])[c:2][c:3][c:4][c:5][c:6]1

Table S1. The 70 selected QSAR variables/descriptors (obtained through the RDKit library) and used in the classification models

Variables	
MinEStateIndex	qed
BCUT2D_LOGPLOW	AvgIpc
Kappa2	PEOE_VSA1
SMR_VSA1	SMR_VSA10
SlogP_VSA10	SlogP_VSA12
SlogP_VSA6	SlogP_VSA7
VSA_EState1	VSA_EState10
VSA_EState7	VSA_EState8
NumAromaticHeterocycles	NumAromaticRings
NumHeterocycles	NumRotatableBonds
NumSaturatedRings	NumUnspecifiedAtomStereoCenters
RingCount	fr_Al_COO
fr_C_O	fr_NH0
fr_azide	fr_benzene
fr_halogen	fr_imidazole
fr_nitro_arom_nonortho	fr_nitroso
fr_unbrch_alkane	BCUT2D_LOGPHI
FpDensityMorgan3	HallKierAlpha
BertzCT	PEOE_VSA14
PEOE_VSA13	SMR_VSA5
SMR_VSA4	SlogP_VSA3
SlogP_VSA2	EState_VSA3
SlogP_VSA8	VSA_EState6
VSA_EState4	NumAromaticCarbocycles
FractionCSP3	fr_Ar_N
NumAtomStereoCenters	fr_aniline
NumSaturatedCarbocycles	fr_epoxide
Phi	fr_nitro_arom
fr_ArN	Kappa1
fr_Ndealkylation2	PEOE_VSA2
fr_bicyclic	SMR_VSA7
fr_nitro	SlogP_VSA5
fr_pyridine	EState_VSA9
fr_COO	fr_ester
fr_aryl_methyl	fr_COO2

The descriptors in Table S1 can be organized into functional groups according to the type of molecular information they capture, which facilitates understanding the role of each variable within the predictive model. Table S1 highlights a selection of QSAR descriptors chosen by the Feature Selection function[2] within the Scikit-Learn library, covering everything from global physicochemical properties to detailed structural characteristics.

A first relevant group corresponds to descriptors related to lipophilicity and molecular surface properties, such as those in the SlogP_VSA family (e.g., SlogP_VSA2, SlogP_VSA10, SlogP_VSA12) and SMR_VSA. These descriptors combine solvent-accessible surface area (VSA) information with properties such as logP (partition coefficient) or molar refractivity, being particularly important for modeling processes such as cell permeability, absorption, and interaction with biological targets. In addition, descriptors such as BCUT2D_LOGPLOW and BCUT2D_LOGPHI also capture electronic variations and hydrophobicity distributed throughout the molecule.[3-5]

Another important set is formed by electronic descriptors, especially those derived from the PEOE (Partial Equalization of Orbital Electronegativities) method, such as PEOE_VSA1, PEOE_VSA13, and PEOE_VSA14. These descriptors reflect the partial charge distribution in the molecule, which is crucial for predicting intermolecular interactions, such as hydrogen bonds and electrostatic interactions.[3-5]

The table also includes topological and molecular complexity descriptors, such as BertzCT, Kappa1, Kappa2, and HallKierAlpha. These parameters quantify aspects of molecular connectivity, branching, and overall shape of the molecule, being useful for distinguishing structurally simple compounds from more complex or highly branched compounds. Furthermore, metrics such as *A*vgIpc and Phi contribute to characterizing the distribution of structural information and molecular flexibility.[3-5]

Descriptors related to the counting of structural elements are also widely represented, including variables such as NumAromaticRings, NumHeterocycles, NumRotatableBonds, RingCount, and NumSaturatedRings. These descriptors provide direct information about molecular architecture, such as the number of aromatic rings, the presence of heteroatoms, and the degree of flexibility, factors that directly influence pharmacokinetic and pharmacodynamic properties.[3-5]

Another important group corresponds to functional fragment descriptors (prefix fr_), such as fr_benzene, fr_nitro, fr_halogen, fr_imidazole, and fr_ester. These descriptors indicate the presence or absence of specific functional groups within the molecule, allowing the model to identify structural patterns associated with biological activity or toxicity. This type of information is particularly valuable in QSAR models, as many biological effects are directly related to the presence of certain chemical groups.[3-5]

Descriptors such as FpDensityMorgan3 represent features derived from molecular fingerprints, capturing recurring substructural patterns. The qed (quantitative estimate of drug-likeness) descriptor provides an integrated metric that assesses how "drug-like" a molecule is, considering multiple criteria simultaneously.[3-5]

Finally, descriptors such as FractionCSP3, NumAtomStereoCenters, and NumUnspecifiedAtomStereoCenters provide information on stereochemistry and hybridization, fundamental aspects for specific molecular interaction with biological targets.[3-5]

References

1. Scikit-learn, *User Guide*. [Link] accessed in March 2026

2. Scikit-learn Feature selection. [Link] accessed in March 2026

3. Todeschini, R.; Consonni, V.; *Molecular Descriptors for Chemoinformatics*, vol. I, 2nd ed.; Wiley-VCH Verlag: Weinheim, Germany, 2009.

4. Casteiger J.; *Handbook of Chemoinformatics*, vol. 1; Wiley-VCH Verlag: Weinheim, Germany, 2003.

5. Leach, A. R.; Gillet, V. J.; *An Introduction to Chemoinformatics*; Springer: Dordrecht, The Netherlands, 2007.