

Molecular Descriptors (Transform Descriptors)¹ & Reactivity Prediction:

Is it possible?

(10.6084/m9.figshare.14679576)

Molecular descriptors are mathematical representations of molecules' properties that are generated by algorithms. They are used to describe the features of organic molecules and predict their behavior in chemical reactions. Molecular descriptors can be structural, relating data about the relative position of atoms and types, or calculated data such as electron density using quantum chemical methods. They can be classified by the following representations:

- 0D: Atom types, molecular weight, bond types
- 1D: Indication of presence of C-N, C-S, C=N, or ...
- 2D: Topological indices, connectivity indices, molecular graphs
- 3D: Geometrical descriptors, electronic descriptors

In recent years, prediction of chemical properties by computed tools has become a useful and suitable way to analyze and compare wide libraries of compounds aiming to design and develop new molecules with higher biological activity and/or better and controlled chemical behavior. Machine learning algorithms have gained prominence in predicting the outcomes of organic chemical reactions, with molecular descriptors serving as valuable input features. Such approaches have exhibited promising results and hold the potential to revolutionize the field of synthetic organic chemistry.

However, despite the wealth of literature on various descriptors and prediction methods, one challenge remains for the average synthetic chemist: comprehending the mechanisms behind these predictions. Existing methods rely on fingerprints, smiles and more complex descriptors to represent chemical reactivity. Not to mention the different reaction prediction approaches developed over time using templates and/or chemical rules.

These methods may appear overly complex to the average synthetic chemist. Maybe it was never intended that an average synthetic chemist could understand the mathematics and data science behind the predictions. But can a prediction be trusted without understanding? As scientists, it is not easy.

The question arises: can predictions be simplified to a level where they are understandable and, consequently, trustworthy to the average synthetic chemist? To address this question, it was decided to use descriptors classified as 1D descriptors. When these descriptors are applied to a reaction following the formula product descriptors minus the sum of the reactants descriptors (Equation 1) have been named Transform Descriptors (TD's).²

$$TD = F_P - \sum_{i=1}^n F_{Rn}$$

Equation 1. Transform descriptor (TD) calculation based on product descriptors (i.e., features) (FP) minus sum of reactant descriptors (FR). For simplicity's sake we mainly use descriptors only for the two main components in a reaction, i.e., $n = 2$.

It has been published that using Transform Descriptors and calculating their Euclidean distance followed by k-Means clustering it was possible to cluster a reaction data set into reactions of the same type (reaction classification). Here, the same descriptors will be used to find out if the reactivity of two random reactants (query reactants) could be predicted. Instead of calculating the Transform Descriptors (products minus reactants), the second part of Equation 1 will represent the values to match between the data set and the query reactants (Figure 1).

Initially, the same descriptors used for the reaction classification method will be used here (Table 1).

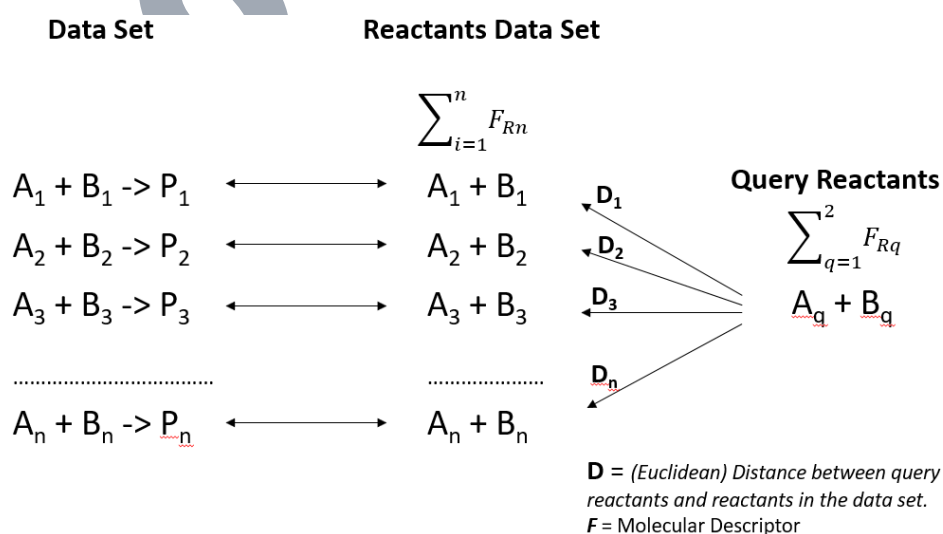


Figure 1. Initial hypothesis; Comparing the descriptors of the query reactants against the reactants in the data set. The closest the distance, the closest the type of reaction possible for the query reactants

The basic principle was to find out whether the collections of descriptors in Table 1 were enough to correlate reactants (or the sum of their values).³ In other words, the idea was to see if the nearest reactants (hence distances) to our query reactants could be used to see what kind of reactivity one could expect and therefore what kind of product could be suggested (Figure 2)

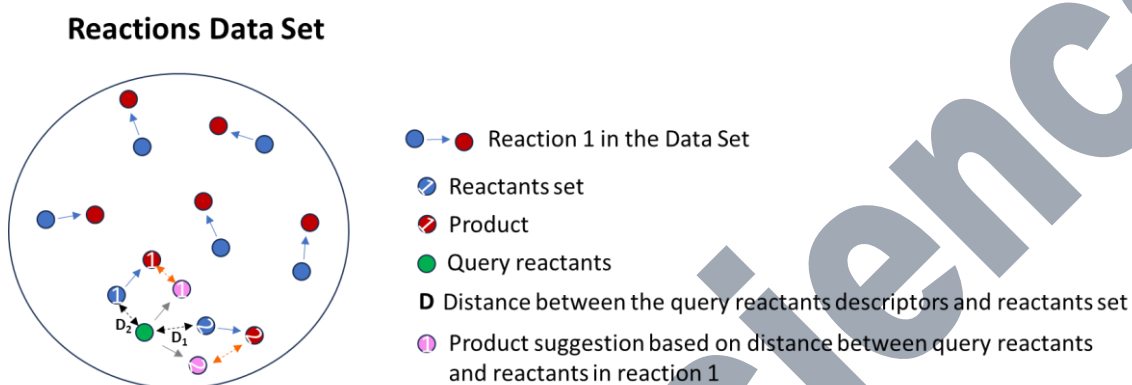


Figure 2. Basic hypothesis, the nearest reactants (from the data set) to our query reactants would imply the nearest distance between the corresponding reactions.

Table 1. List of descriptors (TD's) used for the reaction classification process.

| | |
|-----------------------------|------------------------------------|
| 1. NumRotatableBonds | 31. -C-S-(noC) |
| 2. NumAmideBonds | 32. -NH- |
| 3. NumRings | 33. -C=N- |
| 4. NumAromaticRings | 34. -N=N- |
| 5. NumAliphaticRings | 35. -O-O- |
| 6. NumAromaticHeterocycles | 36. -C=O |
| 7. NumAliphaticHeterocycles | 37. azide group |
| 8. NumAromaticCarbocycles | 38. sulfoxide (general) |
| 9. NumAliphaticCarbocycles | 39. diazo |
| 10. NInR | 40. nitro |
| 11. OInR | 41. |
| 12. SInR | 42. sulfonamide |
| 13. PInR | 43. isonitrile |
| 14. C-C_bonds | 44. aldehydes |
| 15. C-N_bonds | 45. any carbonyl |
| 16. C-O_bonds | 46. nitrile |
| 17. C-S_bonds | 47. enamine or aniline nitrogen |
| 18. Csp3 | 48. cyanide |
| 19. C#C_bonds | 49. hydrazone |
| 20. C-Si_bonds | 50. hydrazine |
| 21. S-H | 51. phenol |
| 22. -S-S- | 52. Aromatic N |
| 23. S=O | 53. Aromatic O |
| 24. S-N | 54. 5-membered rings |
| 25. Hal-6hetring | 55. Aromatic S |
| 26. Hal-5het-ring | 56. aromatic carboxylic acid/ester |
| 27. aromatic OMs or OTf | 57. 7-membered rings |
| 28. aromatic bromide | 58. Number of aliphatic bonds |
| 29. aromatic chloride | 59. Number of cis/trans bonds |
| 30. aromatic iodide | 60. Number of aromatic bonds |

Once the hypothesis was established, a case example was prepared to see if it could be done in practice. As a set of query reactants, the two reactants in Figure 3 were chosen to start our study.

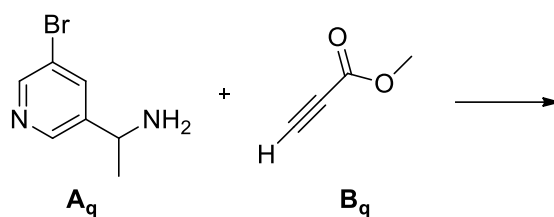


Figure 3. Query reactants.

Upon observing these query reactants, synthetic chemists would typically contemplate the types of reactions that might occur between them. In the context of a Sonogashira-type reaction, our primary concern, aside from the low electron density of the reactants, is the presence of an unprotected primary amine that could potentially interfere with the planned reactivity. Therefore, we anticipate that the analysis should indicate both possibilities.

KNIME 4.7 was used for the analysis and the descriptors calculations were performed as described previously.² The workflow included:

- Reaction data base (e.g. USPTO) curation
- Random selection of 5k reactions
- Descriptors calculation (only for reactants and the complete transform descriptors)
- Descriptors calculation for the query reactants

Once all descriptors were computed, two different distance algorithms (Euclidean and Manhattan) were tested. Initial results were unexpectedly disappointing, with none of the nearest reactions deemed feasible with the query reactants. A closer examination of the descriptors revealed certain values that exerted disproportionate influence on the results. Attempts were made to rectify this issue through various normalization techniques, but regrettably, the outcomes continued to fall short of expectations.

A simple way to enhance the descriptors from the query reaction was to increase their value with a simple mathematical operation. However, the results were still not the ones one could expect (Figure 4).

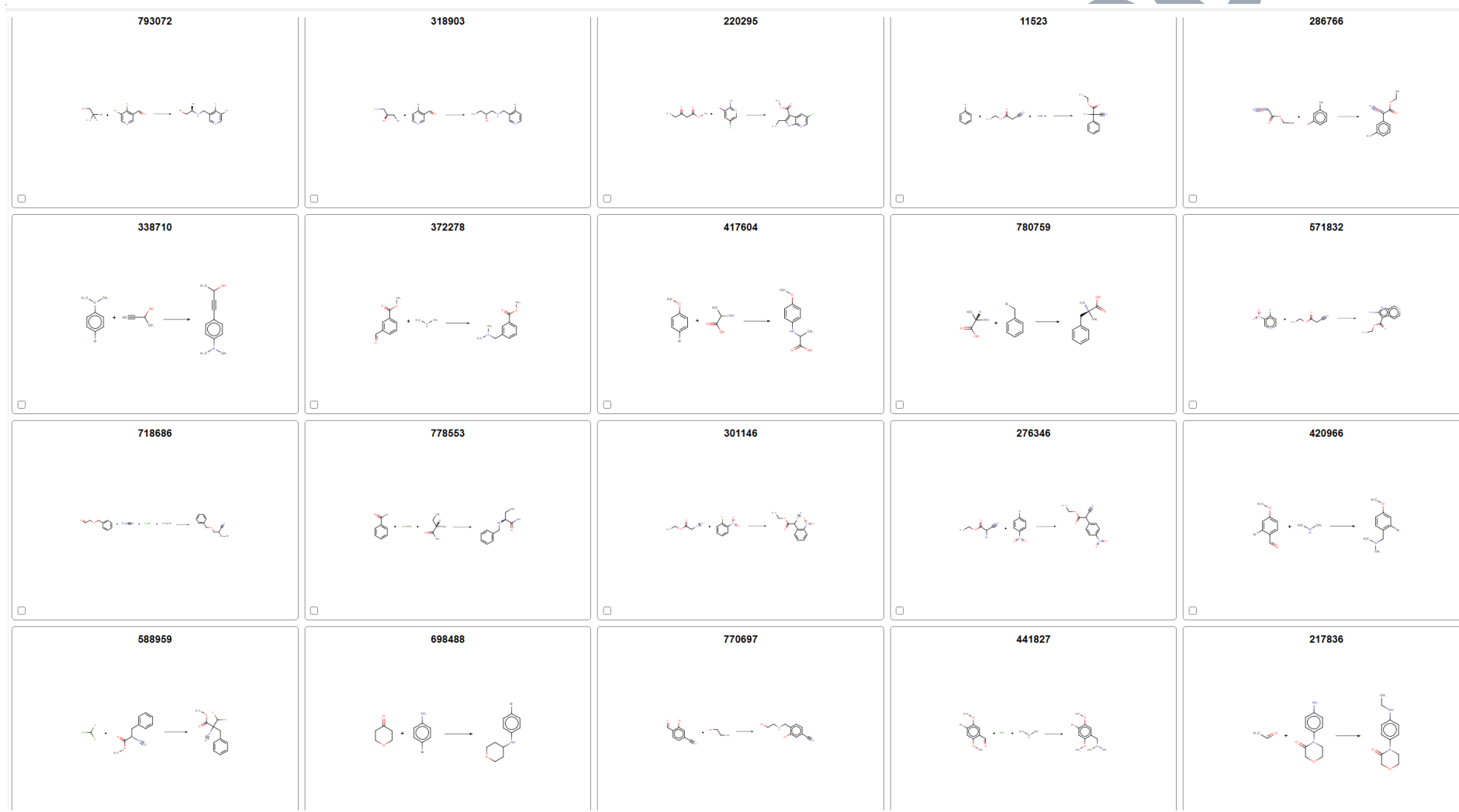


Figure 4. Initial results. The kind of reactivity shown does not correspond with the kind of reactions it could be expected from the two reactants at hand.

To better understand the results obtained, it was decided to investigate the descriptors and their range. In Table 2 it is compared the values for the different descriptors between the reactions data set and the values for the query reaction.

Table 2. Upper and lower values for the different descriptors (for the data set used).

| Descriptors | Reactants (Data Set) | | Query Reactants | |
|---------------------------------|----------------------|-------------|-----------------|-------------|
| | lower bound | upper bound | lower bound | upper bound |
| (R) NumRotatableBonds | 0 | 52 | 1 | 1 |
| (R) NumAmideBonds | 0 | 10 | 0 | 0 |
| (R) NumRings | 0 | 15 | 1 | 1 |
| (R) NumAromaticRings | 0 | 10 | 1 | 1 |
| (R) NumAliphaticRings | 0 | 10 | 0 | 0 |
| (R) NumAromaticHeterocycles | 0 | 7 | 1 | 1 |
| (R) NumAliphaticHeterocycles | 0 | 7 | 0 | 0 |
| (R) NumAromaticCarbocycles | 0 | 10 | 0 | 0 |
| (R) NumAliphaticCarbocycles | 0 | 9 | 0 | 0 |
| (R) Number of aliphatic bonds | 1 | 117 | 10 | 10 |
| (R) Number of cis/trans bonds | 0 | 8 | 0 | 0 |
| (R) Number of aromatic bonds | 0 | 59 | 6 | 6 |
| (R) NinR | 0 | 11 | 1 | 1 |
| (R) DinR | 0 | 8 | 0 | 0 |
| (R) SinR | 0 | 7 | 0 | 0 |
| (R) PinR | 0 | 1 | 0 | 0 |
| (R) C-C_bonds | 2 | 91 | 8 | 8 |
| (R) C-N_bonds | 0 | 28 | 3 | 3 |
| (R) C-O_bonds | 0 | 50 | 3 | 3 |
| (R) C-S_bonds | 0 | 14 | 0 | 0 |
| (R) Csp3 | 0 | 63 | 3 | 3 |
| (R) C=C_bonds | 0 | 9 | 1 | 1 |
| (R) C-Si_bonds | 0 | 13 | 0 | 0 |
| (R) S-H | 0 | 2 | 0 | 0 |
| (R) S-S | 0 | 0 | 0 | 0 |
| (R) S-O | 0 | 4 | 0 | 0 |
| (R) S-N | 0 | 2 | 0 | 0 |
| (R) Hal-6htring | 0 | 0 | 0 | 0 |
| (R) Hal-5htring | 0 | 0 | 0 | 0 |
| (R) aromatic OMs or OTf | 0 | 1 | 0 | 0 |
| (R) aromatic bromide | 0 | 6 | 1 | 1 |
| (R) aromatic chloride | 0 | 9 | 0 | 0 |
| (R) aromatic iodide | 0 | 3 | 0 | 0 |
| (R) C-S (noC) | 0 | 0 | 0 | 0 |
| (R) NH | 0 | 10 | 0 | 0 |
| (R) primary amine | 0 | 3 | 1 | 1 |
| (R) secondary amine | 0 | 4 | 0 | 0 |
| (R) C-N | 0 | 0 | 0 | 0 |
| (R) N-N | 0 | 0 | 0 | 0 |
| (R) O-O | 0 | 0 | 0 | 0 |
| (R) ketone | 0 | 6 | 0 | 0 |
| (R) aldehyde | 0 | 3 | 0 | 0 |
| (R) azide group | 0 | 5 | 0 | 0 |
| (R) sulfoxide (general) | 0 | 1 | 0 | 0 |
| (R) diazo | 0 | 1 | 0 | 0 |
| (R) nitro | 0 | 4 | 0 | 0 |
| (R) sulfonamide | 0 | 2 | 0 | 0 |
| (R) isonitrile | 0 | 1 | 0 | 0 |
| (R) nitrile | 0 | 4 | 0 | 0 |
| (R) enamine or aniline nitrogen | 0 | 10 | 0 | 0 |
| (R) cyanide | 0 | 1 | 0 | 0 |
| (R) hydrazone | 0 | 1 | 0 | 0 |
| (R) hydrazine | 0 | 1 | 0 | 0 |
| (R) phenol | 0 | 8 | 0 | 0 |
| (R) Aromatic N | 0 | 9 | 1 | 1 |
| (R) Aromatic O | 0 | 3 | 0 | 0 |
| (R) 5 membered rings | 0 | 35 | 0 | 0 |
| (R) Aromatic S | 0 | 7 | 0 | 0 |
| (R) ester | 0 | 8 | 1 | 1 |
| (R) carboxylic acid | 0 | 4 | 0 | 0 |
| (R) amide | 0 | 10 | 0 | 0 |
| (R) acyl halide | 0 | 2 | 0 | 0 |
| (R) 7 membered rings | 0 | 7 | 0 | 0 |
| S_count (R) | 0 | 7 | 0 | 0 |
| O_count (R) | 0 | 38 | 2 | 2 |
| N_count (R) | 0 | 16 | 2 | 2 |
| C_count (R) | 5 | 97 | 11 | 11 |
| P_count (R) | 0 | 2 | 0 | 0 |
| Sn_count (R) | 0 | 0 | 0 | 0 |
| Si_count (R) | 0 | 4 | 0 | 0 |
| B_count (R) | 0 | 3 | 0 | 0 |
| F_count (R) | 0 | 25 | 0 | 0 |
| Cl_count (R) | 0 | 9 | 0 | 0 |
| Br_count (R) | 0 | 6 | 1 | 1 |
| I_count (R) | 0 | 4 | 0 | 0 |

reference

Analysis of the values in Table 2 showed that some of the values could have a huge influence on the distance algorithm without (necessarily) being part of reactivity changes. Following this reasoning, it was decided to remove those descriptors not directly involved in potential reactivity changes, such as number of carbons, or number of C-C bonds, etc. Consequently, the definition of some of the descriptors was revised, and it was found that there were some duplicity and gaps regarding functional groups present in the descriptor's table.

Table 3. List of selected descriptors. The (R) indicates that these descriptors were only calculated for the reactants.

| Descriptors List | |
|-------------------------|---------------------------------|
| (R)_NInR | (R)_diazo |
| (R)_OInR | (R)_nitro |
| (R)_SInR | (R)_sulfonamide |
| (R)_PInR | (R)_isonitrile |
| (R)_C-N_bonds | (R)_nitrile |
| (R)_C-O_bonds | (R)_enamine or aniline nitrogen |
| (R)_C-S_bonds | (R)_cyanide |
| (R)_Csp3 | (R)_hydrazone |
| (R)_C#C_bonds | (R)_hydrazine |
| (R)_C-Si_bonds | (R)_phenol |
| (R)_S-H | (R)_Aromatic N |
| (R)_-S-S- | (R)_Aromatic O |
| (R)_S=O | (R)_5-membered rings |
| (R)_S-N | (R)_Aromatic S |
| (R)_Hal-6hetring | (R)_ester |
| (R)_Hal-5het-ring | (R)_carboxylic acid |
| (R)_aromatic OMs or OTf | (R)_amide |
| (R)_aromatic bromide | (R)_acyl halide |
| (R)_aromatic chloride | (R)_7-membered rings |
| (R)_aromatic iodide | S_count_(R) |
| (R)_-C-S-(noC) | O_count_(R) |
| (R)_-NH- | N_count_(R) |
| (R)_primary amine | P_count_(R) |
| (R)_secondary amine | Sn_count_(R) |
| (R)_-C=N- | Si_count_(R) |
| (R)_-N=N- | B_count_(R) |
| (R)_-O-O- | F_count_(R) |
| (R)_ketone | Cl_count_(R) |
| (R)_aldehyde | Br_count_(R) |
| (R)_azide group | I_count_(R) |
| (R)_alpha-H (C=O) | (R)_alpha-H (N=O) |
| (R)_sulfoxide (general) | (R)_alpha-H (S=O) |

Dealing with reactivity, it was not acceptable to define generic C=O bonds, since not all C=O

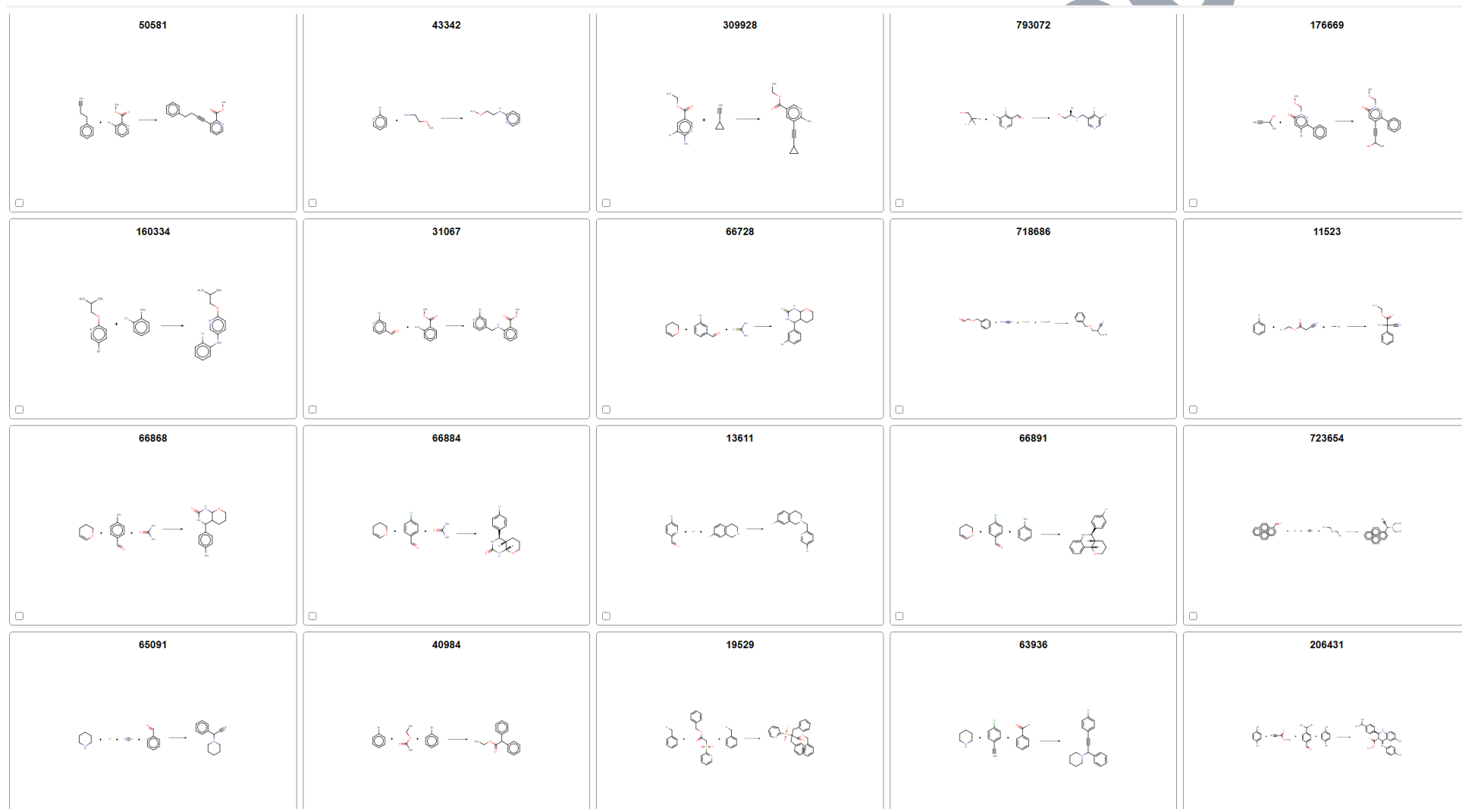


Figure 5. Reactivity patterns prediction using selected descriptors.

bonds have the same type of reactivity, *e.g.*, a C=O in a ketone does not react like the same fragment in an ester. Therefore, a more precise identification of functional groups had to be made to be able to identify reactivity patterns (Table 3).

In conclusion, the descriptors outlined in Table 3, have proven to be invaluable in identifying relevant examples within the dataset that correlate with the expected reactivity of the query reactants, as illustrated in Figure 5.

Employing these carefully selected descriptors, along with Euclidean distance calculations between the reactants in the dataset and the query reactants, we identified twenty nearest examples that align with the anticipated reactivity patterns, as shown in Figure 5.

However, it is important to acknowledge that some examples suggested reactivity patterns that seemed incompatible with the query reactants. For instance, a reductive amination typically requires a ketone or aldehyde and a primary or secondary amine, yet our analysis occasionally proposed this reaction type. This discrepancy can be attributed to the collective effect of numerous descriptors with predominantly low values, as most descriptors in Table 3, have values of 0 or 1. In such cases, the cumulative contribution of various descriptors can outweigh the significance of individual descriptors critical to a particular reactivity pattern. For example, in the case of reductive amination, despite the descriptor (R)_aldehyde having a zero value, the calculations (both Euclidean and Manhattan distances) favored other descriptors, leading to the suggestion of this particular reactivity pattern.

To address this issue, we implemented a filtering step that adhered to a condition derived from the Transform Descriptors formula (Figure 6). According to this formula, Transform Descriptors (TD) for reactions in the dataset should not yield negative values if the corresponding reactants' descriptors (${}^R TD$) have zero values. Consequently, when using the query reactants' descriptors (${}^R TD$), a descriptor with a value of zero in the query reactants would dictate that the TD for suggested reactions must be greater than or equal to zero, ensuring compatibility with the reactivity pattern.

$$TD = F_P - \underbrace{\sum_{i=1}^n F_{Rn}}_{{}^R TD} \quad \text{if } {}^R TD = 0 \Rightarrow TD \geq 0$$

Figure 6. Rule for the Transform Descriptors of the nearest reactions. If the specific descriptor in the reactants set is zero, the Transform Descriptor for the suggested reaction must be ≥ 0 .

As a result, we refined our selection from the twenty nearest examples in the dataset by filtering out those that did not conform to the condition outlined above. The outcomes, as depicted in Figure 7, substantiate the viability of utilizing the descriptors discussed herein (Table 2 and Table 3) for the prediction of reactivity patterns between two query reactants.

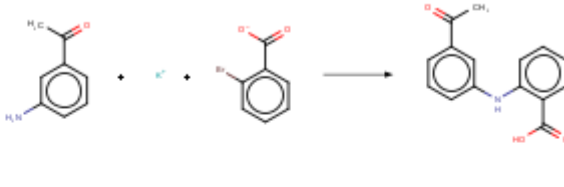
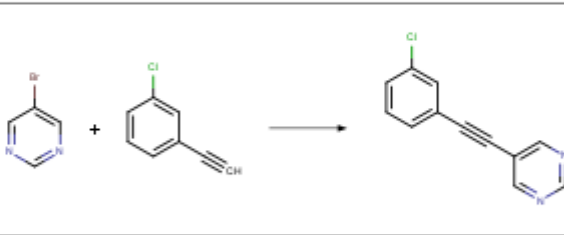
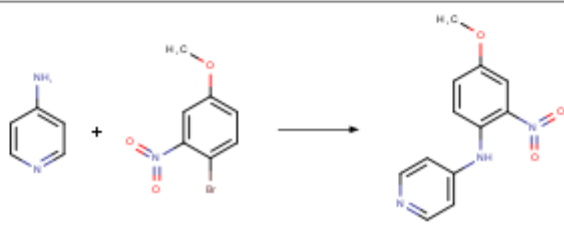
| PrimaryKeyID | SMI reaction_(smiles) |
|--------------|---|
| 108178 |  |
| 14275 |  |
| 643358 |  |

Figure 7. Examples in the data set representing the suggested reactivity among the two query reactants.

Based on these results, it becomes evident that our methodology enables us to anticipate the type of reactivity and consequently the products that may arise from the chosen pair of query reactants (Figure 8).

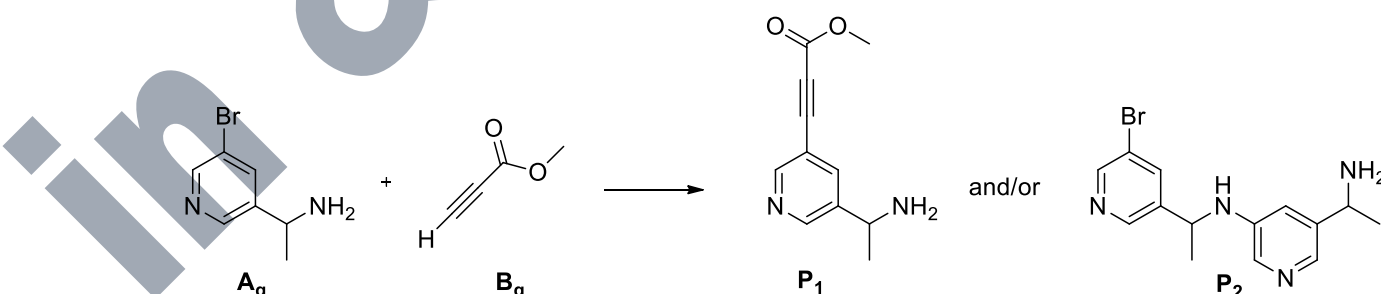


Figure 8. Suggested reaction products based on the examples in Figure 7.

Returning to our initial question regarding the feasibility of predicting reactivity using the descriptors outlined in Table 3, the answer is both affirmative and negative.

The essence of organic synthetic chemistry lies in its diversity. It encompasses a multitude of reaction types, each with its unique characteristics and intricacies. Our methodology, while demonstrating promise, is not a universal solution. Its success hinges on the careful selection and fine-tuning of descriptors to align with the specific reactivity expected.

In the case presented here, our descriptors were meticulously tailored to the desired reactivity, yielding encouraging results. However, the same set of descriptors may not yield accurate predictions for other reaction types, such as Diels-Alder or Beyli-Hillman reactions. These discrepancies stem from the absence of specific descriptors designed to capture the intricacies of these distinct transformations.

In essence, this methodology offers a novel and simplified approach to predicting reactivity patterns, but it is not yet a one-size-fits-all solution.

Some interesting references:

- <https://www.intechopen.com/chapters/58592>
- <https://www.pnas.org/doi/10.1073/pnas.2212711119>
- <https://www.sciencedirect.com/topics/medicine-and-dentistry/molecular-descriptor>
- <https://onlinelibrary.wiley.com/doi/10.1002/jcc.27016?af=R>
- <https://www.nature.com/articles/s41598-023-32347-4>
- Jennifer N. Wei, David Duvenaud, and Alan Aspuru-Guzik, ACS Cent. Sci. 2016, 2, 725–732. DOI: 10.1021/acscentsci.6b00219.
- David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken and Pierre Baldi, Mol. Syst. Des. Eng., 2018, 3, 442.
- Somayyeh Babaei, Mahmood Niad, Polyhedron 188 (2020), 114710.
- For current systems with organic chemistry reaction prediction see (among others): a) Anders Bøgevig, Hans-Jürgen Federsel, Fernando Huerta, Michael G. Hutchings, Hans Kraut, Thomas Langer, Peter Löw, Christoph Oppawsky, Tobias Rein, and Heinz Saller. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. Organic Process Research & Development 2015 19 (2), 357-368. b) <https://www.deepmatter.io/insights/blog/icsynth-40-beta-next-generation-retrosynthetic-planning-software/> c) Linda Wang. ChemPlanner to integrate with SciFindern. C&EN Global Enterprise 2017 95(25), 35-37. d) <https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder/synthesis-planning>. e) ASKOS see: <https://askcos.mit.edu/>.
- Gasteiger, J., Hutchings, M. G., Saller, H., & Löw, P. (1988). Prediction of Chemical Reactivity and Design of Organic Synthesis. Chemical Structures, 343–359. https://doi.org/10.1007/978-3-642-73975-0_36.

*no preaching, no teaching, just a
perspective and an opinion*

i&O S

¹ Minidis ABE, Huerta FF. Transform Descriptors for template free reaction classification. ChemRxiv. Cambridge: Cambridge Open Engage; 2023; This content is a preprint and has not been peer-reviewed. DOI 10.26434/chemrxiv-2023-89v4q