# TRANSFORM DESCRIPTORS

## *(Reactions Data Mining)*

https://doi.org/10.26434/chemrxiv-2023-89v4q

### Introduction

Chemists worldwide encounter new synthetic challenges daily, particularly in contract development and manufacturing organizations (CDMOs) where they often lack background knowledge about target molecules and previously employed synthetic routes. The demand for modified or new routes is also common due to various technical, intellectual property, regulatory, and environmental factors. In such cases, experience, deep knowledge, and intuition play crucial roles in determining the plausibility of finding new routes or improving existing ones.

Computer-assisted synthesis design (CASD) systems have gained popularity in recent years, [1] providing support in generating new synthetic ideas. However, their practical synthetic feasibility is most often limited, even though advancements in applied machine learning techniques have allowed for new perspectives and predictions of possible pathways and conditions.[2] Tools such as ICSynth,[3] ChemPlanner,[4] MIT's ASKCOS,[5] AstraZeneca's AIZynthfinder,[6] and the IBM[7] and ETH Zürich collaboration have emerged, and proved useful, offering open-source or subscription-based access. While these tools significantly advance reaction prediction, their limitations lie in the size and quality of available data sources. Most of these systems tend to prioritize the most obvious disconnections first, among them less interesting functional groups interconversion. This does prove though that these systems do work and certainly give a fresh perspective on viewing results, but they also don't pertain high value for an expert chemist as of yet. For less experienced chemists, or with some adaption, in high-throughput scenarios, these systems already do have interesting impact.

Considering the content and accuracy limitations of existing CASD systems, their current scope is insufficient for process development purposes. Notably, Doyle et al. [8], Johansson et al. [9], and others have recently discussed this topic and various approaches. [2,10] It is furthermore important to acknowledge that not everyone has access to these systems for various reasons.

A brand-new Route Design represents a challenge for all synthetic chemists, not to mention, the difficulties to convince other chemists to invest in a theoretical synthetic route based solely

on knowledge and experience. Here we describe a simple and, easy to calculate descriptors as new parameters to support reactions data mining and hence route design.

## Methods and Discussion

The search for chemical descriptors to help data scientists describing organic reactions with numbers is nothing new.[11] Many different approaches to solve this problem have been considered, where the complexity of the descriptor and additionally their calculations pose a hurdle for most of the experimental chemists that are not familiar with computer science and advanced software applications.

A commonly used approach to perform reaction searches is to employ reaction fingerprints based on the molecular structures of the involved components. This technique enables the comparison of complete reactions based on structure only within the dataset.[12] While evaluating this approach with different (test) reactions, results based on reaction fingerprints in most of our scenarios often did not lead to a satisfactory outcome. This became mainly evident by manual inspection, where for example, some expected reactions were not present among the results and vice versa, expected "incorrect" examples were part of the suggestions, i.e., false positive and false negatives created too much noise for a manual analysis of any practical relevance. Equally disappointing was the use of e.g., RDkit descriptors, or even Mordred, which had previously proven interesting for (binary labeled) yield prediction.[13] This is most likely due to some of the shortcomings of theses descriptors where for example electronic effects of different functional groups are not (sufficiently) accounted for. Additionally, many of the available features, unless hand-picked, are not meaningful for reactions, such as molecular weight or logP, to name but a few.

Another difference in our previous work was using a reaction matrix concept,[14] where each reaction component was described by a of molecular features defining a unique reaction matrix (Figure 2). While it was possible to do aforementioned binary prediction using machine learning methods, it was not suitable for comparing or identifying any given reaction versus similar ones in the literature. Therefore, we considered a different approach to define descriptors that would be more suitable for reaction classifications. In addition, we wanted a template free approach and not work with the common reaction mapping and its pros and cons.[15]

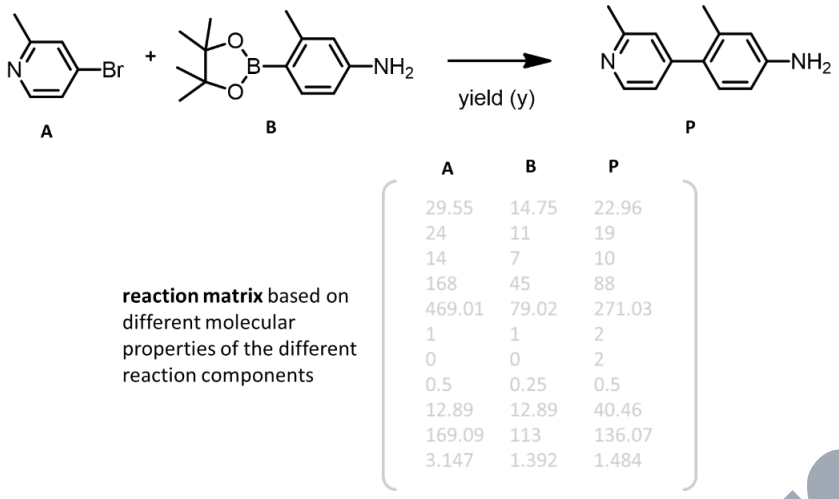



***Figure 2.*** *Reaction matrix concept.*

Such descriptors defining a chemical transformation are based on functional groups, atoms and selected structural changes going from reactants to the final product (Transform Descriptors, TD).[16] These features are carefully selected, a combination of simple out-of-the-box 2d (RDkit) descriptors, combined with additional hand-picked features (vide infra). As shown in Equation 1, a resulting, "transform descriptor" (TD) is then simply calculated by the subtracting descriptor values of the product, from the sum of the reactants' descriptor values.

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

***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.*

For data-handling, calculation, and analysis, either Knime and/or Python coding was applied. Both offer multiple molecular descriptor modules as required, as well as flexibility in terms of speed, etc. Some images shown herein were taken from Knime-based evaluations.

Any dataset analyzed stem most often from external sources and had to be curated and normalized to fit the calculation pipeline. The most common freely accessible one is USPTO dataset.[17] The dataset curation was done to minimize size as much as possible, keeping all relevant information: *reaction_(smiles), yield, an internal ID,* and, if available also *experimental*

*procedure, reagents, catalysts, solvents, citation, other conditions*. The transform descriptors (TD's) were then calculated using the following sequence:

- Structure curation using smiles to Rdkit conversion, salt removal, neutralization.
- Yield normalization, removing missing and erroneous numbers.
- Calculation of selected Rdkit based descriptors (e.g., NumRotableBonds, NumAmideBonds, NumRings, NumAromaticRings, NumAliphaticRings, NumAromaticHeterocycles, NumAliphaticHeterocycles, NumAromaticCarbocycles and NumAliphaticCarbocycles) of the entire molecule (other descriptors, such as Indigo as desired), together with a fragment-based table (Table 1) to capture functional groups. Values are calculated for each component, and as indicated in Equation 1, the sum of reactant descriptor values are subtracted from the product values.
- Elemental analysis (EA), based solely on the string, counting most interesting elements (not RDkit based). The resulting EAs are calculated by summing the reactants' elements together (per element) and then subtracting the product EAs from that resulting value.

*Table 1: Fragment based features and their smarts-representation.*

| SMARTS Representation | Feature Description |
|---|---|
| [#7;R] | NInR |
| [#8;R] | OInR |
| [#16;R] | SInR |
| [#15;R] | PInR |
| [#6]~[#6] | C-C_bonds |
| [#6]~[#7] | C-N_bonds |
| [#6]~[#8] | C-O_bonds |
| [#6]~[#16] | C-S_bonds |
| [#6;X4] | Csp3 |
| C#C | C#C_bonds |
| [#6]~[#14] | C-Si_bonds |
| [#16!H0] | S-H |
| [#16;X1][#16;X1] | -S-S- |
| [#16]=[#8] | S=O |
| [#16][#7] | S-N |
| [#17,#35,#53]-[#6]-1=[#7]-[#6,#7]=[#6,#7]-[#6,#7]=[#6,#7]-1 | Hal-6hetring |
| [#17,#35,#53]-[#6]1[#6]-[#6][#6][!#1!#6]1 | Hal-5het-ring |
| [#6;a]-[#8]S([#6])(=O)=O | aromatic OMs or OTf |
| [#6;a]Br | aromatic bromide |
| [#6;a]Cl | aromatic chloride |
| [#6;a]I | aromatic iodide |
| [#6X2][S][!#6] | -C-S-(noC) |
| [#7;H1] | -NH- |
| [#7X1]=[#6X2] | -C=N- |
| [#7X1]=[#7X1] | -N=N- |
| [#8X1][#8X1] | -O-O- |
| [#8]=[#6X2] | -C=O |
| [$(*-[NX2-]-[NX2+]#[NX1]),$(*-[NX2]=[NX2+]=[NX1-])] | azide group |
| [$([#16X3]=[OX1]),$([#16X3+][OX1-])] | sulfoxide (general) |
| [$([#6]=[N+]=[N-]),$([#6-]-[N+]#[N])] | diazo |
| [$([NX3](=O)=O),$([NX3+](=O)[O-])][!#8] | nitro |
| [$([SX4](=[OX1])(=[OX1])([!O])[NX3]),$([SX4+2]([OX1-])([OX1-])([!O])[NX3])] | sulfonamide |
| [CX1-]#[NX2+] | isonitirile |
| [CX3H1](=O)[#6] | aldehydes |
| [CX3]=[OX1] | any carbonyl |
| [NX1]#[CX2] | nitrile |
| [NX3][$(C=C),$(cc)] | enamine or aniline nitrogen |
| [NX3][CX2]#[NX1] | cyanide |
| [NX3][NX2]=[*] | hydrazone |
| [NX3][NX3] | hydrazine |
| [OX2H][cX3]:[c] | phenol |
| [n] | Aromatic N |
| [o] | Aromatic O |
| [r5] | 5-membered rings |
| [s] | Aromatic S |
| [*;a][#6;A;X3]=[O;X1] | aromatic carboxylic acid/ester |



These transform descriptors allowed then for more automated analysis, offering more promising ways for better clustering of specific reaction similarities. In a somewhat related way, Doyle and colleagues also turned to different ways of featurization since neither

fingerprints nor Mordred descriptors captured correct/sufficient information for their reaction modeling purposes.[8] The main difference in our approach, aside from manually selected features is that we use the difference of product features versus the sum of reactant features to yield what we call transform descriptors (TDs, Equation 1). In contrast, other approaches most often, albeit not explicitly, use e.g., one-hot-encoding, averages, or sums of ***all*** component features, even when used in neural network-based encodings.

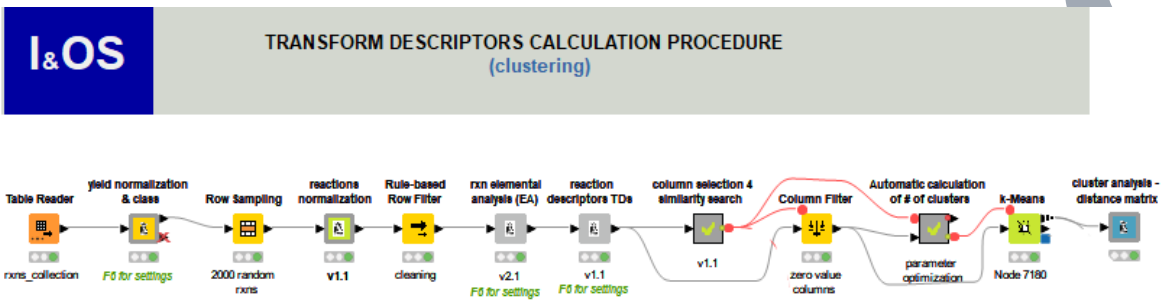



*Figure 1: Overview of a Knime based workflow for calculation of TDs.*

To test our hypothesis, a random (small) set of ca 2000 diverse reactions were collected and submitted to TD calculation to see their potential in clustering reactions of the same type. Any features with only zero values were filtered before clustering. To define the optimal number of clusters an optimization process was carried out using the Silhouette coefficient as

optimization parameter. k-Means algorithm was used for the clustering process and the number of clusters was optimized in this case to 58 clusters (Figure 2).

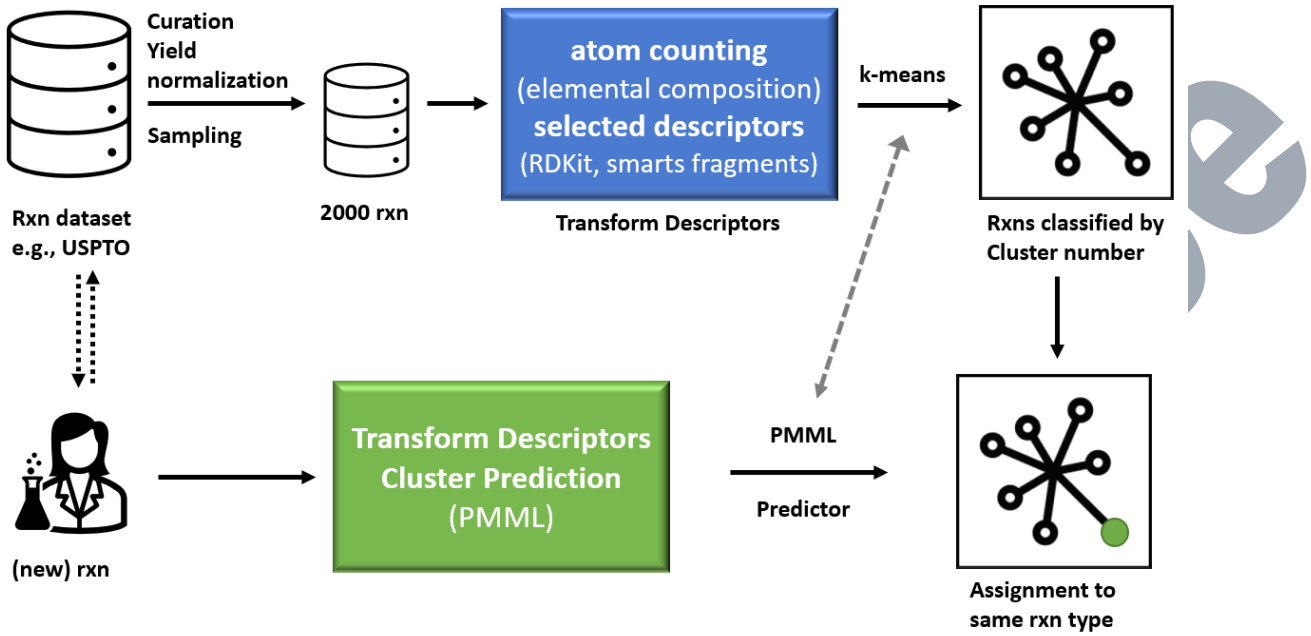



*Figure 2: Overview of workflow concept*

Initial visual analysis of the different clusters gave the impression that the process worked very well to collect similar reaction types into the same clusters, especially considering the diversity of reaction types in the original data set. Nevertheless, it was decided to verify the outcome based on these transform descriptors using reaction fingerprints versus the average Tanimoto distance *within* the different clusters. All of them showed an average Tanimoto distance around 0.8, indicating that the transforms descriptors used are a valid set of parameters to describe reactions (Table 2).

*Table 2. Clusters and their average Tanimoto distance within the cluster. Notice that the calculation is done for each reaction pair within the cluster.*

Clusters & Tanimoto

average distance within clusters

Show 60 entries

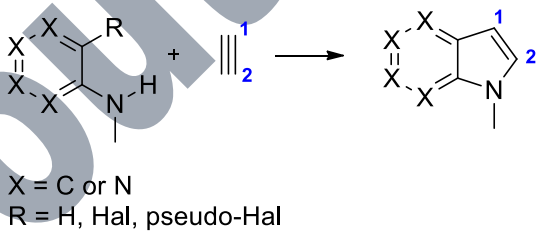
| | | | | | | |
|---|---|---|---|---|---|---|
| **Cluster:** cluster_0<br>**distance_(tanimoto):** 0.6 | **Cluster:** cluster_1<br>**distance_(tanimoto):** 0.6 | **Cluster:** cluster_10<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_11<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_12<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_13<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_14<br>**distance_(tanimoto):** 0.7 |
| **Cluster:** cluster_15<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_16<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_17<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_18<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_19<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_2<br>**distance_(tanimoto):** 0.8 | **Cluster:** cluster_20<br>**distance_(tanimoto):** 0.6 |
| **Cluster:** cluster_21<br>**distance_(tanimoto):** 0.8 | **Cluster:** cluster_22<br>**distance_(tanimoto):** 0.6 | **Cluster:** cluster_23<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_24<br>**distance_(tanimoto):** 0.6 | **Cluster:** cluster_25<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_26<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_27<br>**distance_(tanimoto):** 0.8 |
| **Cluster:** cluster_28<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_29<br>**distance_(tanimoto):** 0.8 | **Cluster:** cluster_3<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_30<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_31<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_32<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_33<br>**distance_(tanimoto):** 0.7 |
| **Cluster:** cluster_34<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_35<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_36<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_37<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_38<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_39<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_4<br>**distance_(tanimoto):** 0.7 |
| **Cluster:** cluster_40<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_41<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_42<br>**distance_(tanimoto):** 0.8 | **Cluster:** cluster_43<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_44<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_45<br>**distance_(tanimoto):** 0.5 | **Cluster:** cluster_46<br>**distance_(tanimoto):** 0.8 |
| **Cluster:** cluster_47<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_48<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_49<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_5<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_50<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_51<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_52<br>**distance_(tanimoto):** 0.7 |
| **Cluster:** cluster_53<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_54<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_55<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_56<br>**distance_(tanimoto):** 0.8 | **Cluster:** cluster_57<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_6<br>**distance_(tanimoto):** ? | **Cluster:** cluster_7<br>**distance_(tanimoto):** 0.6 |
| **Cluster:** cluster_8<br>**distance_(tanimoto):** 0.7 | **Cluster:** cluster_9<br>**distance_(tanimoto):** 0.7 | | | | | |

Showing 1 to 58 of 58 entries

Previous 1 Next

Overall, this method worked well with the available small dataset in categorizing reactions of the same type simply by using easy to calculate descriptors, such as those in Table 2. Of further note is that no reaction mapping or labeling of certain atom types of groups of reaction components were made.

Nota bene, the clusters consisted not only of certain, specific name-reactions, but of reactions generating the same structural motifs regardless the reaction mechanism, meaning, these transform descriptors identify more than one conventional, classical name-reaction. For example, the Larock indole synthesis is a specific transformation where a haloaniline reacts with an alkyne in the presence of a transition metal such as palladium to generate an indole. However, the reality is that there are several reactions that do not follow the same reaction mechanism but that generate the same final structure. Transform descriptors consider the starting materials and final products and therefore the classification is based on the type of structure that is being generated rather than the conventional reaction name. This becomes of importance when using an existing database and most likely search conventionally by keywords and not obtain all possible hits.

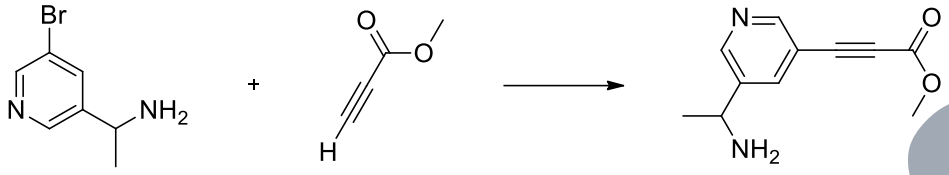



*Scheme 1. All reactions with shown specific structural features would be in the same cluster regardless of their mechanism.*

To evaluate this hypothesis of transform descriptor being applied in this manner, a test-reaction not being present in the original data set, was designed and submitted to TD calculation and evaluated based on the clustering outcome. This would proof highly valuable

since it would be one step closer to the system "learning new chemistry" (within the confines of this limited context).

As test reaction, a Sonogashira type coupling was chosen, see Scheme 2.




*Scheme 2. The designed reaction for the cluster assignment (not present as such in the original data set).*

Transform descriptors and clustering were performed as previous and visual inspection of the resulting cluster assignments showed reasonable reaction classification where our designed reaction was assigned to only one of the original clusters. With all reactions in the predicted cluster being compatible with a Sonogashira-type coupling (Table 3). This demonstrates once again the utility of the transform descriptors here discussed.

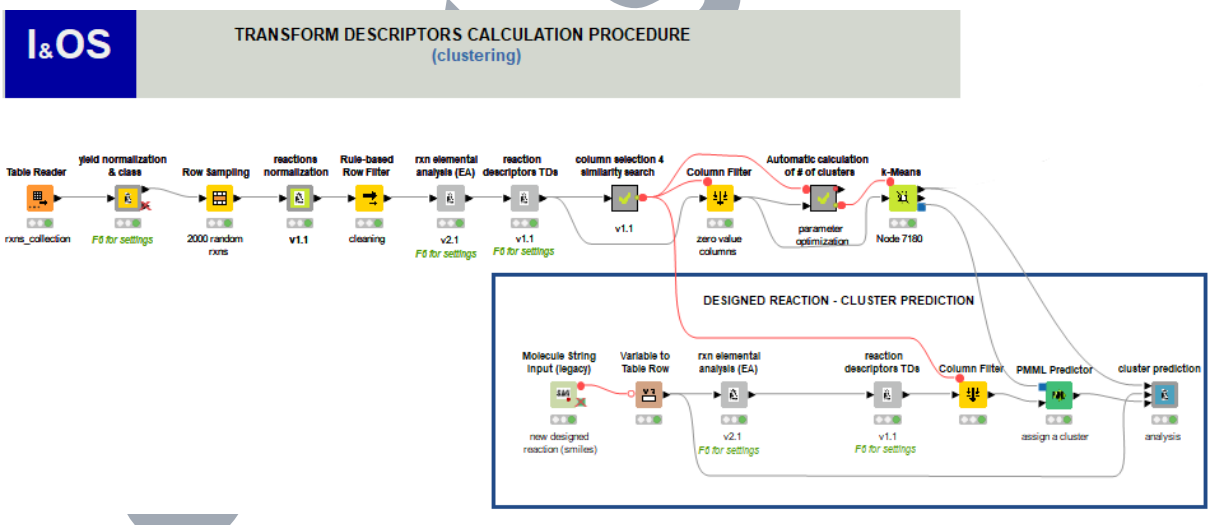
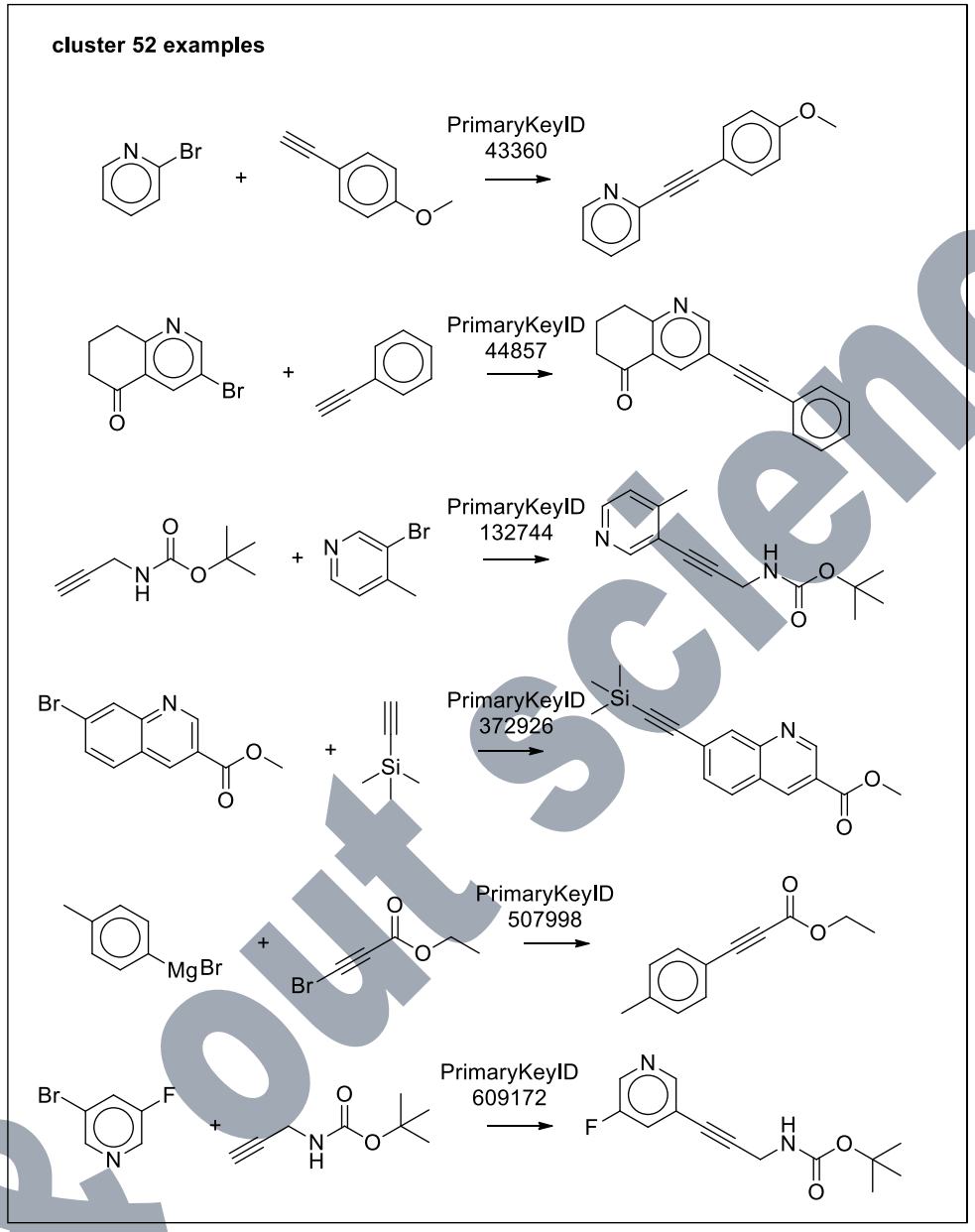



*Figure 3: A Knime workflow overview describing for classification of a reaction type.*

*Table 3: Examples in the predicted cluster for the reaction in Scheme 2. Not all reactions in the cluster operate under a Sonogashira type mechanism but generate the same type of final structure.*




By applying transform descriptors in a similar way, we propose it should even be possible to predict a set of reaction-conditions for a particularly designed reaction. However, this analysis requires much more curation and data preparation for achieving optimal results. As shown by other research groups, such data preparation represents a complex process, (discussed elsewhere).[xviii] Equally, a large dataset, such as the (full) USPTO will be used to challenge the reaction categorization and disclosed elsewhere.

## Conclusions

It has been demonstrated that "easy to calculate" and "easy to understand" descriptors may be applicable to described complex reactions, equaling, and even challenging known methods based on fingerprints or neuronal network type featurization. In addition, we believe these TDs to be conceptually easier to understand by even the average synthetic chemist. This does not exclude the necessity to eventually combine some of these methods, but the performance with the minimalistic input has already proven useful in practical laboratory settings.[xix]

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



[1] Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inform.* **2014**, *33*, 469-476.
[2] Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* **2018**, *23*, 1203-1218.
[3] 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/.
[4] a) Linda Wang. ChemPlanner to integrate with SciFinder$^n$. *C&EN Global Enterprise* **2017** *95*(25), 35-37. b) https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder/synthesis-planning.
[5] For information on ASKOS see: https://askcos.mit.edu/.
[6] Genheden, S., Thakkar, A., Chadimová, V. et al. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform* **12**, 70 (2020).
[7] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, Alpha A. Lee. Molecular Transformer – A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci* **2019**, 5, 9, 1572-1583.
[8] Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and computer-assisted planning for chemical synthesis. *Nat. Rev. Methods Primers* **2021**, *1*, 23.

[9] Johansson, S.; Thakkar, A.; Kogej, T.; Bjerrum, E.; Genheden, S.; Bastys, T.; Kannas, C.; Schliep, A.; Chen, H.; Engkvist, O. AI-assisted synthesis prediction. *Drug Discov. Today Technol.* **2019**, *32-33*, 65-72.

[10] a) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Sci. Rep.* **2017**, *7*, 3582. b) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge. J. Chem. Inf. Model. 2022, 62, 1357-1367.

[11] Skoraczyński, G., DIttwald, P., Miasojedow, B., Szymkuc, S., Gajewska, E. P., Grzybowski, B. A., & Gambin, A. (2017). Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient? Scientific Reports, 7(1), 1–9. https://doi.org/10.1038/s41598-017-02303-0.

[12] Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. J. Chem. Inf. Model. 2015, 55, 39-53.

13 a) G. Landrum, RDKIT: Cheminformatics and machine-learning software in C++ and Python. For the current version, see: 10.5281/zenodo.591637; For the originating project, see: https://www.rdkit.org/. b) Moriwaki H, Tian Y-S, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. Journal of Cheminformatics 10:4: 10.1186/s13321-018-0258-y

14 Fernando Huerta, Samuel Hallinder, Alexander Minidis. Machine Learning to Reduce Reaction Optimization Lead Time – Proof of Concept with Suzuki, Negishi and Buchwald-Hartwig Cross-Coupling Reactions. DOI 10.26434/chemrxiv.12613214.v1.

[15] Schwaller, P., Probst, D., Vaucher, A.C. et al. Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3, 144–152 (2021). https://doi.org/10.1038/s42256-020-00284-w

[16] The origin of the Transform Descriptors here discussed was found on a public KNIME workflow found in the hub https://hub.knime.com/chines/spaces/Reaction%20Navigator/latest/Reaction-Navigator~aTsCiRxVeGzuSzYm.

[17] USPTO curation: a) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, https://doi.org/10.17863/CAM.16293 (2012). b) Minidis, Alexander (2021). Yield curation USPTO rsmi/csv datasets. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14414039.v1

[xviii] Gimediev et al, Reaction Data Curation I: Chemical Structures and Transformations Standardization, *Molecular. Informatics*. **2021**, 40, 2100119. doi.org/10.1002/minf.202100119.

[xix] Undisclosed data.