PROPHETIC BAD REACTIONS

(https://doi.org/10.6084/m9.figshare.22580266.v2)

Reactions databases contain millions of reactions. Every research group looks for optimal reactions' conditions including, catalysts, solvents, reagents, etc. However, all groups share the common aversion of "bad reactions". This fact is represented by the lack of reactions with bad yields published in the literature and therefore, in reaction databases.

Although this might be considered a quality feature of a reaction database, the lack of "bad reactions" is one of the main problems for machine learning algorithms. Imbalanced data sets are the main problem for data scientists when trying to create models.

The solution to this problem is not trivial since no one wants to publish "bad results". As a result, predictive models for a small set of specific transformations with not many examples published in the literature will always be biased because of the lack of bad reactions reported. The Larock indole synthesis is an example of such a reaction data set (Figure 1) (https://en.wikipedia.org/wiki/Larock_indole_synthesis).

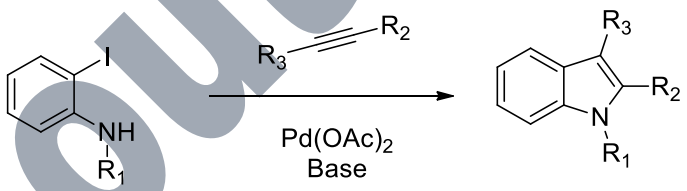

Figure 1. Reaction example published in wikipedia.

The Larock indole synthesis is a heteroannulation reaction that uses palladium as a catalyst to synthesize indoles from an ortho-iodoaniline and a disubstituted alkyne. It is also known as Larock heteroannulation.

One important factor to consider in this reaction is the regioselectivity. On paper, two possible regioisomers are supposed to be formed (Figure 2).

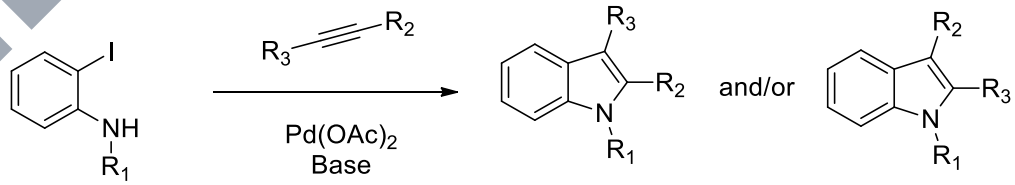

Figure 2. Possible formation of 2 regioisomers during the Larock indole synthesis.

Different studies have shown that the best way to perform this transformation is by using o-iodoaniline together with 2-5 equivalents of an alkyne, a palladium (II) catalyst, an excess of sodium or potassium carbonate as base, PPh_3, and one equivalent of LiCl or n-Bu_4NCl. The role of the lithium or the ammonium chloride salt is to coordinate with the Pd(0) species and favor the oxidative addition of the palladium catalyst to the iodine-carbon bond.

When searching for this type of transformation in a commercial database, approximately 6600 results can be found using the reaction query shown in Figure 3).

Database search template:

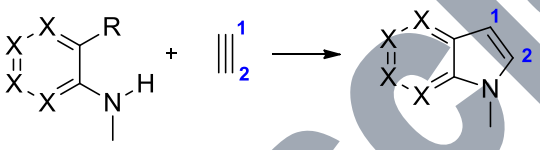

Figure 3. A database search template drawn as generic as possible but with the restriction that the sp-carbons in the alkyne are C2 and C3 in the newly formed indole. R1 represents any possible halogen or pseudo halogen atom, and R2 represents C- or N-atoms.

A visual analysis of the search results shows that not all reactions in the data set can be classified as Larock indole synthesis. The lack of a palladium catalyst, the differences in the electronic properties regarding the haloaniline and/or alkyne, and the so-called domino reactions (transformations occurring in a sequential order in a reaction vessel) indicates that several reaction mechanisms can operate in making the same indole bonds as in the Larock reaction (Figure 4).

Although any reaction fitting the general description shown in Figure 3 will be mentioned as Larock indole synthesis in this manuscript, the fact is that it's only a way to describe the process of making an indole, forming the bonds marked in red (Figure 4).

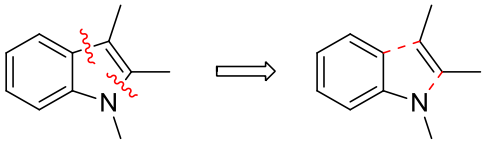

Figure 4. The synthetic approach corresponds with the formation of two bonds in one go to make the indole ring.

The original data set containing over 6,000 reactions was curated by removing duplicates, multistep transformations, and reactions without reported yields. The remaining 3,233 reactions were categorized as good (with a reported yield of at least 60%) or bad (with a reported yield of 40% or less). The specific yield cut-off used is not relevant to the manuscript's goal, and any other distinction between good and bad reactions could have been used with the same result.

After the yield classification, it was observed that the bad reactions category represents approximately 10% of the original data set, resulting in an unbalanced dataset (see Figure 5).

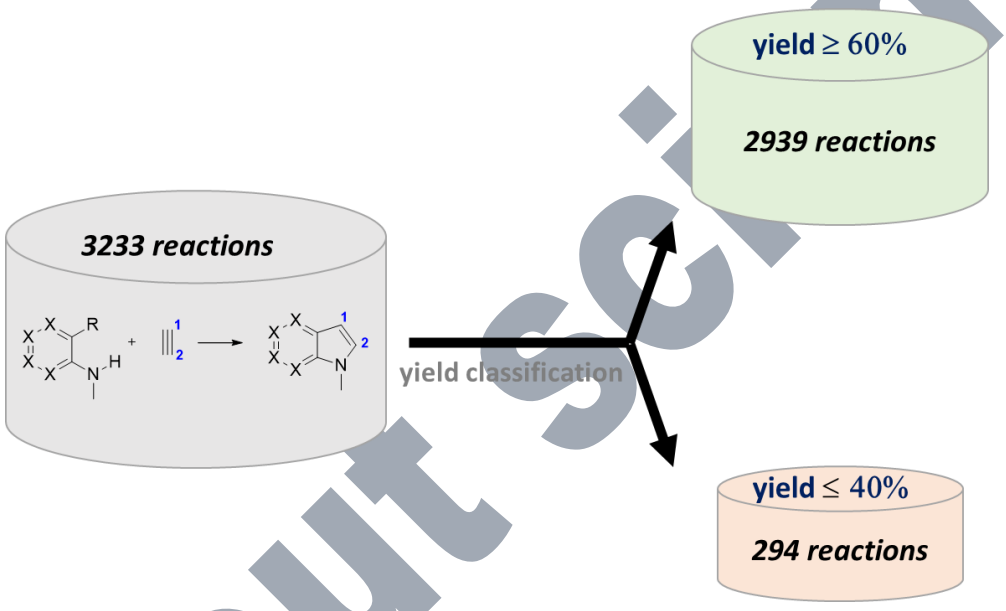

Figure 5. Curated data set divided into good and bad reactions categories.

A dataset of a few thousand reactions may not be large enough for certain algorithms to be applied, and the manuscript does not discuss the corresponding reaction conditions (reagents, catalysts, solvents, etc.) associated with the prophetic bad reactions generated. The process of adding reaction conditions to the prophetic bad reactions will be explained in other manuscripts.

The process of generating prophetic bad reactions begins with the separation of the two primary components in the Larock indole synthesis: the haloaniline (component A) and the alkyne (component B).

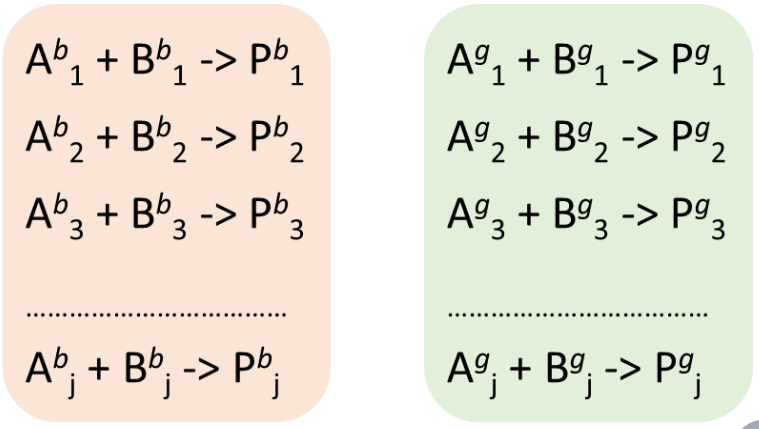

Figure 6. Representation of the "bad reactions" data set (left) and the "good reactions" data set (right)

The components of the bad reactions' ensemble were separated (A^b_j and B^b_j), and all duplicates were removed so that each A^b_j and B^b_j were unique. Theoretically, it would be possible to generate a new combination of components and corresponding products using a combinatorial process with the unique bad components (A^b_j and B^b_j). However, synthetic chemists with experience would understand that this approach might not be correct since, in many cases, it is only one of the components in the reaction that presents a synthetic challenge.

The solution to this experience-based observation may not be very trivial. However, if one could be 100% sure that both bad components (A^b_j and B^b_j) have never been used successfully in the Larock indole synthesis, it might be enough to consider the reaction between them (A^b_j and B^b_j) a bad reaction.

The additional condition described in the above paragraph indicates that each component in the prophetic bad reactions set must be unique within the bad reactions ensemble and must not belong to the good reactions' ensemble (see Figure 7).

$$A^b_j + B^b_j \rightarrow P^b_j$$

$$\text{where } A^b_j \text{ and } B^b_j \notin \{A^g_j + B^g_j\}$$

Figure 7. Prophetic bad reactions hypothesis.

Using Knime version 4.7.0, a set of reactions was downloaded from a database search and cleaned of duplicates, examples with no reported yield, and multi-

step reactions. A structural clean-up removed salts, transition metals, and complex organic structures. A stoichiometry clean-up was also performed, allowing only 1:1 stoichiometry in the final data set.

Additional information such as experimental description, catalysts, and reagents were filtered out, leaving only the reaction field (as smiles) and yield. The reactions were then classified as "good reactions" (yield reported over 59%) and "bad reactions" (yield reported under 41%) and separated into components for each ensemble.

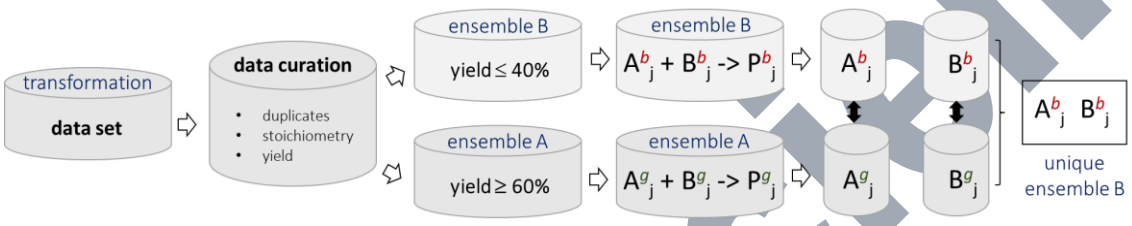

Figure 8. Workflow for the generation of unique reaction components, within the whole data set, with only bad yields reported.

Visual analysis of the components was used to create prophetic bad reactions. In this case study, although the Larock indole synthesis is a specific transformation with specific reagents and/or catalysts, several alternatives can generate the same indole structure, such as via amine addition and conjugated addition. Thus, it is essential to consider all kinds of reactions or sequences of one-pot transformations before applying the procedure to generate prophetic bad reactions.

Each unique component in the bad reaction's ensemble was combined with the corresponding counterpart except for its original, generating a new ensemble of components. Once the new pair of components was generated, the corresponding prophetic bad reactions were created using RDKit Two Component Reaction node in Knime. Figure 9 shows the process for the generation of the new set of reaction components and the formation of the corresponding prophetic bad reactions.

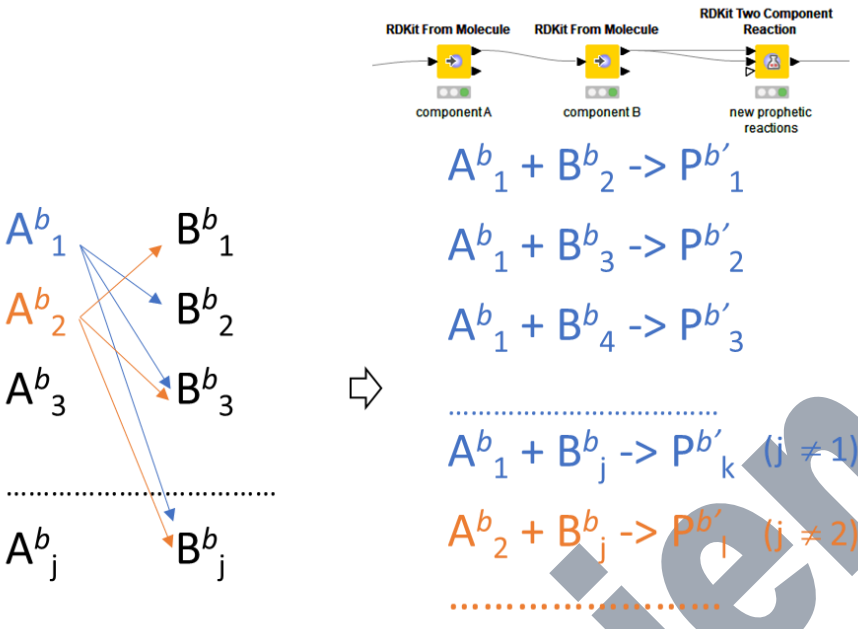

Figure 9. Process for the generation of the new set of reaction components (left) and the formation of the corresponding prophetic bad reactions (right).

The process depicted in Figure 9 generated over 10k prophetic reactions. Considering that the initial data set was slightly over 3k reactions, it was clear that further analysis was needed.

Indeed, the large number of reactions generated has an explanation based on the nature of the reaction itself (Figure 2).

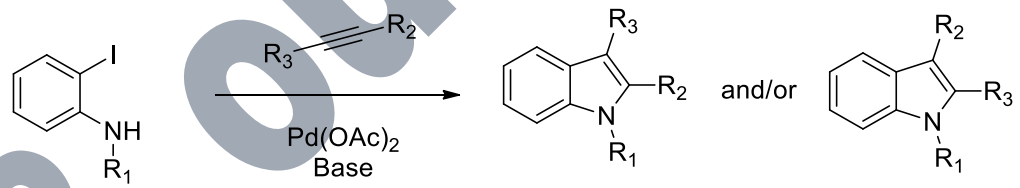

Figure 2. Possible formation of 2 regioisomers during the Larock indole synthesis.

The way the generic reaction smarts were written (for the RDKit Two Component Reaction node), made it possible to obtain two reactions, one for each possible regioisomeric product (with asymmetric alkynes).

There, nevertheless, was an advantage with the large number of reactions generated. Due to the nature of this project, the more diverse representation of the "bad reactions" class the better one can train a model, based on molecular descriptors. Therefore, after the original work to generate (theoretically) "prophetic bad reactions," a diversity selection algorithm was applied.

Spite off the lack of enough data points and the rudimentary of the analysis, the original data set (3233 reactions) was submitted to a classification machine learning algorithm (Random Forest) to predict good reactions (yield > 60%) or bad reactions (yield < 40%). Reactions with yields between 41-59% were disregarded.

Next, the bad reactions class was augmented with a selection of the 100 most diverse prophetic bad reactions. The same descriptors were calculated and the same process for Random Forest classification algorithm was applied.

Comparing the results (besides errors and pitfalls) it was obvious that the prophetic bad reactions generated helped the model to be less biased. Specially in the prediction of “bad reactions” (class 1 in Figure 10).

Despite the limitations of the analysis, this work demonstrates that the generation of "bad reactions" can aid in the development of more reliable machine learning models. Further improvements can be made, but this study provides a promising direction for future research.

original data set

TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy	Cohen's kappa	Normalization	ClassID
467	50	9	0	1	0.9	1	0.15	0.95	?	?	min-max	0
9	0	467	50	0.15	1	0.15	1	0.26	?	?	min-max	1
0	0	526	0	?	?	?	1	?	?	?	min-max	x
?	?	?	?	?	?	?	?	?	0.9	0.24	min-max	Overall

original data set + 100 prophetic bad reactions

TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy	Cohen's kappa	Normalization	ClassID
465	74	5	2	1	0.86	1	0.06	0.92	?	?	min-max	0
5	2	465	74	0.06	0.71	0.06	1	0.12	?	?	min-max	1
0	0	546	0	?	?	?	1	?	?	?	min-max	x
?	?	?	?	?	?	?	?	?	0.86	0.09	min-max	Overall

Figure 10. Summary of the random forest algorithm applied to; a) the original data set (up) and, b) the modified data set with prophetic bad reactions (down).

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