Machine learning to reduce reaction optimization lead time – *proof of concept* with Suzuki, Negishi and Buchwald-Hartwig cross-coupling reactions

Fernando F. Huerta[a*], *Samuel Hallinder*[b], *Alexander B.E. Minidis*[a*]

[a] RISE Chemical Process and Pharmaceutical Development, Forskargatan 20J, SE-151 36, Södertälje, Sweden

[b] Uppsala Biomedicinska Centrum BMC, Husargatan 3, SE-752 35, Uppsala, Sweden.

*** Corresponding authors:**

alexander.minidis@ri.se ORCID – 0000-0003-1100-6064

fernando.huerta@ri.se ORCID – 0000-0002-6686-8895

ABSTRACT

To date, optimizing reactions let alone predicting the outcome (yield) of known reactions requires expert knowledge and can at best be obtained by computationally complex and expensive modelling. The present investigation tests if machine learning represents a viable approach for predicting a model reaction outcome that could be put into daily production. A prerequisite was replacing advanced scripting techniques with a more approachable data science platform such as Knime®. The Palladium catalyzed Suzuki-Miyaura, Negishi and Buchwald-Hartwig reactions were selected for a classification model of high/low yielding outcome combined with a selection of reaction conditions stemming from a commercial database. Here we present preliminary results of a random forest-based classification model using readily calculated standard medicinal chemistry descriptors from substrates and products yielded high ROC AUC of up to 96%. The

descriptors used in the model do not convey anything about the reactivity, only 1D- and 2D-structural information, and performed equal or better than fingerprints, both in terms of prediction and computational requirements. One of the major challenges was the quality of the data and its subsequent curation.

KEYWORDS
machine learning; computer aided synthetic design; palladium cross-coupling; Suzuki-Miyaura; Negishi; Buchwald-Hartwig; Knime.

ABBREVIATIONS
RF, Random Forrest; ML, Machine Learning; FP, Fingerprint; CASD, Computer Assisted Synthetic Design

Introduction

Every chemist knows the difficulty of a supposedly simple search for reaction conditions suitable for ones' specific molecules. Deciding on any testing or final conditions is based on long term gathered experience or searching curated literature databases such as Scifinder® or Reaxys®. This task becomes troublesome and time consuming when the chemist has the luxury of having multiple possible transformational options to synthesize a target molecule. While the selection of a specific transformation is influenced by availability of starting materials, costs, and possibly even environmental factors, non or very little consideration outside of process chemistry development is paid to the optimization process prior to the selection of a transformation. Often, our decisions are based on experience and at best on similarity between reagents and products versus those reported in literature. Since large scale industrial processes must be optimized to the best possible/most suitable reaction conditions, it is necessary to know how much time and effort reaction condition optimization for a specific transformation would require. Both, larger scale industrial applications as well as basic R&D investigations require a solid decision basis. Reducing the lead times of such findings will equal to gain of resources, costs, and a possible competitive edge.

Computer aided synthesis design (CASD) has been known since the 1960s,[1] however, the main concern in all these systems is to find disconnections and the corresponding transformations to make the disconnected bonds. LHASA,[2] IC*SYNTH*[3] and ChemPlanner (now part of Scifinder^n)[4] are clear examples of such software systems. All these systems have in common that the reaction conditions suggested for each transformation found are purely based on similarity with those examples described in the literature.[5] In addition, CASD related publications using machine learning models designed to identify reactivity generally use more complex approaches than the one presented here, and are currently not easily implemented in a synthetic chemist's daily work.[6,7] Considering how optimization work is a vital part of the route selection process, we found that non or very little attention has been paid to help experienced chemists to find the best possible reaction for a given transformation to maximize important outcomes such as yield, or purity. For example, the Negishi and Suzuki-type reactions can both generate the same product from different starting materials and conditions.

The early goal was to develop a model to predict the feasibility in terms of expected yield of a known reaction such as a Suzuki-Miyaura cross coupling.[8] We would then expand the scope of the model to other transition metal catalyzed reactions, e.g. Negishi and Buchwald-Hartwig cross-coupling reactions.[9-11]

Predicting the yield of a given transformation could help us to answer two important questions in process chemistry optimization: How much time would be needed to improve the yield of the transformation, and, did we select the right transformation for our target molecule? In other words, how effective is the transformation in terms of yield.

We thus investigated classification models initially focusing on the afore mentioned Suzuki-Miyaura cross coupling reaction. For practical purposes (see also Results and Discussion) the original hypothesis was to find correlations between the physico-chemical properties of the starting materials and the products. Since a single property cannot be used to describe a chemical transformation, we envisioned that a combination of 1D and 2D based properties would be enough for our first model. It remained to be seen if they suffice to describe the entities in question.[12] While such descriptors do not convey anything about advanced 2D or even 3D structural information, they do render model creation simpler than otherwise commonly used fingerprints (FP).[13] These bit-vector based fingerprints used for structural representation have their own limitations in information content, as well as in calculations requirements that increase exponentially with the

size of the FPs. The latter renders it difficult or even impossible to calculate on a regular PC, depending on the applications used (see below).

Skoraczyñski *et al.* discussed the need of new and better descriptors to represent chemical reactions providing the possibility of developing prediction models more efficiently.[12] While we agree with the affirmation that today we do not have enough general descriptors to represent all interactions present in all chemical transformations, we believe that for a specific transformation one could find such key molecular descriptors that best describe the transformation from starting materials to product. For example, during our investigation we found that the applied molecular descriptors used by the model may differ depending on the reaction.[14]

Some recent approaches attempting to expand the toolbox of predicting or explaining outcomes of reactions are based on machine learning models using initially smaller curated datasets and, more recently, big datasets, i.e. examples with millions of reactions.[15] Sources of data are either tediously curated open-source databases such as USPTO, commercial databases, or in-house generated reactions serving as training sets.

Due to this imperfect availability of amount vs quality, additional curation or data-generation is often required as can be seen in our analysis below and also described by for example Davies.[16] To the best of our knowledge no non-commercial reaction-based database is currently available, aside from USPTO also requiring additional curation in order to obtain a useful dataset as described by Coley *et al*.[17]

These issues can also be seen in previous attempts of predictive approaches. For example, attempting to encompass most parameters of interest for predictions of organic reactions Jensen *et al.* applied neural network algorithms to predict catalyst, solvent, reagents, and temperature from 10 million compounds from Reaxys®.[15] Their overall goal seemed to be to develop one general model, though as in most cases so far, predicting a single parameter alone such as yield, tended to perform better than multiple ones (yield, temperature, solvent, etc.) simultaneously. They were also forced to compromise regarding data curation such as the handling of missing data. The computational infrastructure required to achieve this is also not of practical use for everyone. In another approach, Ahneman *et al.* focused on only one reaction using physicochemical properties instead of fingerprints to describe their model. Using molecular, atomic and vibrational properties calculated using Spartan they achieved up to 92% (R^2 value) in yield prediction for a small set of reactions.[18,19] They also attempted to explain the data obtained with experimental NMR data and

thus to show how ML can be used to give insight into reaction mechanisms. Their data creation required in-house generated experimental data, which, albeit certainly more reliable than external data, is time and resource consuming unless already existent.

Thus, while existing approaches show promising results in the ranges of around 80% accuracy, with few exceptions of over 90% with certain caveats, the outcomes are not practical enough to be deployed for daily use, if only due to computational requirements and accessibility.

Here we present our initial findings on using only molecular descriptors in a binary machine learning model approach. To do so, we choose for starters to remove certain parameters to enable a first proof-of-concept. For example, although certainly an important factor to be optimized, the selection of the catalyst becomes the most troublesome due to the different mechanisms acting in a catalytic cycle. Solvent selection is normally dictated by the polarity and solubility of the starting materials and reagents. In addition, there are various solvent selection models already reported elsewhere.[20] Finally, although the reaction temperature might theoretical be considered a critical parameter for a reaction to take place, in practice and in the actual context of Suzuki-Miyaura transformations, it is usually a transformation carried out under refluxing conditions. Therefore, at this stage and from a practical point of view, we did not consider temperature prediction as a vital parameter for process optimization.

Results and Discussion

The proof-of-concept was limited to analyzing properties of the components involved in a chemical reaction to see if there was a way to predict if said reaction was possible at all. With a potential daily-use implementation in mind the approach should not involve a too complex procedure such as quantum chemical property calculations or mechanism-based calculations, where computational power and user-accessibility would pose a large obstacle. Although an ever-growing library of 3D-based quantum mechanical descriptors are becoming available, they are not yet capable of providing a wide enough range in structural diversity, and neither do they offer an easy way to data-mine on a larger scale, yet.[21,22] Both model-building as well as implementation should be possible on a standard office computer. For instance, a FP based descriptor model in our hands was only possible with bit-vectors up to 1k in size, achieving lower accuracies than the model described below. Although fingerprints represent a structure as graph-based bit-vector and are commonly used for molecular representation in structural similarity analysis,[23,24] they do require

high dimensional representations to perform well. While the actual fingerprint calculations might be fast requiring only modest memory requirements even on a regular PC, the actual model creation is not. Using larger fingerprints such as 4k in size, or Jensen's 16k approach,[15] was therefore not possible on a memory limited computer.

Thus, we focused on using 1- and 2-D based molecular properties that are commonly applied in for example in medicinal chemistry. A somewhat similar approach was described recently by Doyle *et al.*, though they used R programming as modelling tool, for our requirements a less useful approach.[25] On the other hand, Knime,[26] an open source platform that simplifies "programming" allowing for easier adoption by users not equally skilled in advanced computer science, was utilized for data-curation, model building and implementation. A caveat of this system is the memory handling, which is most likely why above-mentioned fingerprint approaches were not possible in our hands, while they would most likely be less of a problem via for example sklearn & rdkit in Python.

As a proof-of concept we focused on the in medicinal chemistry highly popular and very useful Suzuki-Miyaura reaction used for C-C bond formation,[8] as well as similar types of Pd-catalyzed reactions such as Negishi[9] and Buchwald-Hartwig couplings.[10,11] The Suzuki-Miyaura reaction is based on a Boron (**B**) and halide/triflate (**Hal**) containing substrates yielding a product (**P**), see Figure 1. A search in Reaxys® alone reveals over 162k reactions with more than 213k compounds involved (October 2019).[27] Since the Reaxys API for large automated data-extraction is not available to general subscribers, manual export was conducted for the smaller dataset employed in this proof-of-concept investigation.

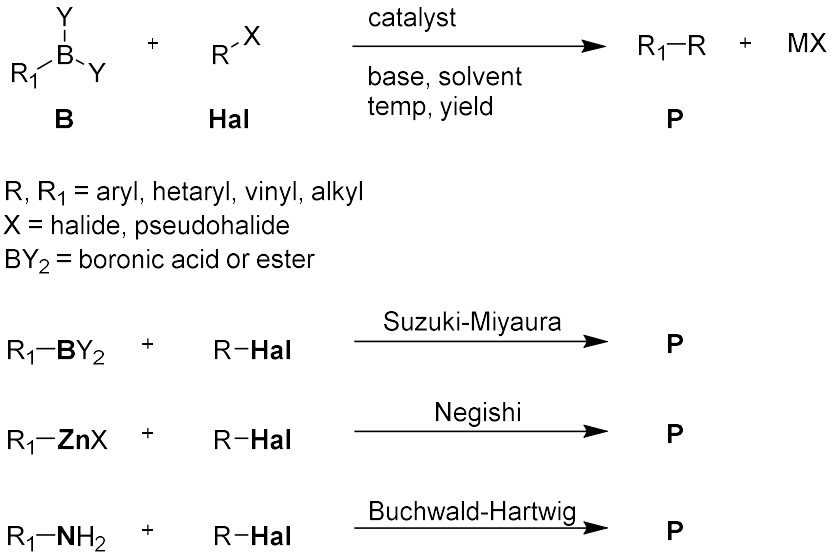

Figure 1. Generalized representation of the Suzuki-Miyaura coupling depicted as **Hal**/**B** components. Negishi is similarly depicted as **Hal**/**Zn** components and Buchwald-Hartwig amination as **Hal**/**N** components.

Despite the high-quality information stemming from a commercial database, further data-curation of the exported dataset was required. The main structures of interest were extracted from the reaction field that contains Smiles based structures. For this, semi-automatic resorting of the input order was applied to obtain "pure" meta-fields for especially **B**-containing structures vs **Hal**-structures. In addition, the stoichiometry of substrates/products had to be normalized, as well as byproducts filtered out. Purely text-based fields are currently not deemed suitable due to none-structured input and conversion difficulties to correct structures, including the catalyst itself. In the Reaxys® dataset one finds it in at least two meta-fields not only as "Catalyst" but also mixed with "Reagent", without an associable ID, and is therefore difficult to curate. Attempts to one-hot encoding of the text-based catalysts did not lead to an enhanced model. Therefore, the catalyst, base and solvent were removed from the dataset and will be part of future investigations. Multi-step reactions and duplicates were removed and finally only one example with the highest yield was selected (see also discussion below). After curation, the dataset contained a total of approximately 95.000 reactions with yields in the range from 5 to 100%.

For the resulting dataset we then calculated molecular properties for the **B**, **Hal,** and **P** structures via the Indigo, CDK and RDKit nodes available in Knime. In case of Negishi this corresponded to

Zn, **Hal,** and **P**, for Buchwald-Hartwig **N**, **Hal,** and **P**. These properties consist of 0D, 1D as well as 2D descriptors, where 0D and 1D descriptors are based on the molecular formula and represent numerical features such as the molecular weight, bond, and atom counts. 2D descriptors are fragment- or graph-based and contain spatial type of information of the molecules such as Topological Surface Area (TPSA), fragment counts, etc. All descriptors are only a portrayal of the components and not of the reactivity since neither 3D-structural information, nor any form of kinetic or thermodynamic information are being considered. Some descriptors though do convey within limits steric or electronic information.[24,28] On the other hand, these descriptors are simple and fast to calculate, as a dataset of nearly 300.000 components is readily computed within minutes (100k of each **B**, **Hal** and **P**). For the final model we limited the descriptors to the RDKit ones since we obtained comparable and even slightly better results using less descriptors.[14] See also Supplementary S2 containing a Knime work-flow performing the clean-up and calculations.

Very early attempts of finding any form of correlation within this dataset based on (multi)-linear regressions or PCA models did not show any form of practical correlations and are therefore not discussed here. Machine learning (ML) algorithms were deemed more promising, certainly in the case of binary classifications. For this, the Random Forest (RF) model offers both predictive accuracy and flexibility. It can be used for both classification and regression tasks. While RF may also overfit a model, it is less prone to do so compared to other models.[29] Furthermore, it performs exceptionally well compared to other learning methods in context of pharma-applications, one of many reasons why the current deep-learning approaches were not considered to offer an advantage, rather the opposite.[30] The Knime platform offers an out of the box solution to implement such ML methods, as well as easy solutions to tweak parameters of such a model.

With a binary model in mind, we chose the yield as a factor of a reaction's success, separating the data into *Good* (>= 60% reported yield) and *Bad* (<= 40% yield). One reason is the difficulty in defining a "good" or "bad" reaction in terms of yield, not including any reasoning of human error that reported results might include. Since we are interested in predicting the success of a reaction in terms of high yields, we accepted to deselect this intermediate range. Further separation is possible, and one may also envision a ternary model that includes the complete range of yields.

The overall model was entailed following steps:

- Curation as described above
- Randomizing the dataset entry order after curation
- Testing with and without equal size sampling
- 10-fold cross validation
- Splitting 80/20 into training/test set.
- Dimensionality reduction to determine most relevant factors for the following model

With the RF model being rather robust, only minor optimization was performed, e.g. the number of trees of 500 were deemed optimal.[14] See also supporting information S1 and S2 or the open access Knime workflow containing above approach in "*Examples > 50_Applications > 21_Model_Selection_and_Managmenet > 01_Model_selection_Sampled*".

While duplicate removal might seem obvious and has been discussed elsewhere,[31] it seems important to highlight that what constitutes as duplicate for reactions may differ for duplicates of single compound comparison used e.g. in biological activity modelling. Two reactions are different as long as they differ in one single variable, such as temperature, additives, stoichiometry, etc. However, for this study, all reactions of the same starting materials and products, independent of conditions, are considered as duplicate.

The highest yield duplicate was kept since it implies the reaction is possible to make at such a yield. Lower yields could be due to a variety of reasons and are therefore dismissed, except in those cases where they are the only example. At this early proof-of-concept stage this is furthermore within our goal of mainly predicting high yielding reaction examples.

This duplicate bias of the data-set was indeed seen in a first approach with a Suzuki dataset was yielding "a too good to be true" ROC of over 98% (Table 1, entry 1a; see also supplementary Table S1).[14] When compared to entries 2a/b one could be led to believe it was due to the larger dataset. Artificially reducing it to the same size (1b vs 2a/2b) the performance impact was noticeable but not substantial. As a result, the AUC was reduced to 89% which is nevertheless a respectable result. The original dataset from entry 1 contained a large number of duplicates stemming from same reactions with different conditions. In terms of imbalance of Good vs Bad, upon comparing entry 2a with 2b no significant difference was observed.

Table 1: Random Forest outcomes for the different data-sets

Entry	Data Set	Size[a]	No. of Descriptors	Size[a,b] Total/Good/Bad	ROC AUC [%]	Sensitivity (True positives)[e]	Specificity (True negatives)[f]
1a	Suzuki[c]	95.4	357 -> 75	75.5 / 37.2 / 38.3	98.2	0.91	0.95
1b	Suzuki[c]	39.3	357 -> 88	39.3 / 19.6 / 19.6	95.6	0.87	0.93
2a	Suzuki[d]	48.9	357 -> 89	39.3 / 21.8 / 17.6	89.1	0.85	0.76
2b	Suzuki[d]	48.9	357 -> 88	35.1 / 17.5 / 17.5	88.7	0.82	0.81
3	Negishi	3.3	357 -> 150	2.9 / 2.7 / 0.2	83.4	1.00	0.25
4	Buchwald Hartwig	26.5	357 -> 88	21.7 / 15.7 / 6.0	90.0	0.94	0.63

a) Numbers given in thousands. b) Data set size after yield split >=60% (good) / <= 40% (bad) and the amount remaining for each binary class. Due to rounding of numbers they might not add up to same as total in Size[a]. c) First tests of data as examined in [14]. d) After additional curation leading to more duplicate removal than initial identified. e) Sensitivity describes proportion of actual positives that are correctly identified. f) Specificity describes the proportion of actual negatives that are correctly identified.

For comparison we also briefly investigated the Negishi and Buchwald-Hartwig cross-coupling. The former dataset is much smaller, only 3.3k reactions after curation. Not counting duplicates, the exported dataset had 7.5k reactions. One of the reasons for this loss is the at times incompatible description of the Zn-components with regards to the descriptor calculations. Nevertheless, a smaller dataset with the same yield separation and RF-model resulted in an AUC ROC value above 80% with an overall high accuracy (Table 1, entry 3). We have not studied if this is possibly due to a high similarity of reactants/products in this smaller data-set. In contrast, the Buchwald-Hartwig amination reaction with an amine component instead of a zinc or boron, was made up of a larger data-set leading to ROC and sensitivity/specificity similar to the Suzuki model (Table 1, entry 4). This might indicate the necessity for larger data-sets greater tens of thousands or more for increased accuracy, although it does not answer how similar/dissimilar the structural information in the data-sets actually is. Many examples in the literature are done on model-reactions with the same substrates (although with small variation in the reaction conditions such as different temperature

or solvents ratio) which in our cased were removed since they were counted as duplicates, one of the reasons for reduced data-set sizes compared to the total entries in the database. The two latter data-sets for Negishi and Buchwald-Hartwig are in addition more imbalanced than the Suzuki set which is somewhat reflected in the lower specificity.

Finally, following up on a comment towards a study by Ahneman and coworkers,[18] published by Keiser *et al.,*[32] we randomized the selection of the molecular descriptors to be used in our model. In the case of the Suzuki set 2a (Table 1), using 5, 10 or 20 randomly chosen descriptors, the maximum ROC AUC value obtained was 77% with sensitivity and specificity in the same range, indicating that a larger number of descriptors is required to obtain higher accuracies, respectively that more work concerning appropriate descriptors is required.[12]

Our results show that basic, easy to calculate molecular descriptors may suffice to create a selection model for the prediction of the chemical yield of a specific transformation. The information thus obtained can be used to estimate whether an optimization process might be worthwhile, or if it were less troublesome to investigate another type of transformation to synthesize the target molecule. (Figure 1).

At present we are still investigating which molecular descriptors are key for each type of transformation, as different descriptors seem to be used in different transformations (Table 1). Thus, currently, a model directly comparing two or more transformations lies outside the scope of this paper.

Class imbalance does have an impact the performance, as is indicated for example by the low numbers of the minor class, i.e. low yields. One approach to counter this could be performed by conformal predictions approach.[33]

Finally, during our investigations we considered experimental validation of the results obtained using the described model. Several experiments were performed to confirm the positive yielding reactions and in all cases the experiments followed the predictions. However, the validation of potential negative results has been troublesome and time consuming and remain therefor part of ongoing investigations.[14]

Conclusion and Outlook

The open source platform Knime® and a "standard laptop" allowed for a simple, still useful and competitive result in order to understand if a transformation is worthy to optimize or not, as in this case, based on the yield as "high versus low".

Model building alone is only one part of approaching this type of reaction prediction, the other is the data and its quality. For common purposes, only non-commercial databases may offer enough data, but their curation remains an issue. Commercial databases are not available to everyone and even they suffer, to some extent, of curation issues. In-house generated data should give the best quality but is restricted to larger efforts from chemists to generate such data and is not possible for everyone.

As discussed, one important factor, curation of catalyst related information requires more extensive work depending on the data-origin. Catalyst and e.g. solvent inclusion should enrich the data-sets in terms of data-set size and information content. Imbalance issues might be able to be countered by conformal prediction.[33]

Another approach for better models will be to use (different) descriptors.[12] What type of descriptors is yet to be seen, especially in terms of complexity. For example, the interesting quantum chemical descriptors approach alone may not suffice, considering Merck's approach, which aside from having a number of caveats, one critical factor certainly being computational power.[34]

In summary, we were able to curate specific reaction sets and create a binary yield predicting model based solely on easily calculated property descriptors. While the predictiveness is yet to come into what we believe would be a practical range of for example >95%, we were able to demonstrate equal or enhanced accuracies versus more complex ones described in the literature. Careful curation of the datasets versus complex calculations poses in our opinion a larger contribution to obtain better models. Nevertheless, we are currently successfully applying the described model and ongoing development ones on in-house cases with the purpose of enhancing the predictiveness.

During the review process some related work in the field of machine learning and synthesis was suggested. Since we have cited other related work, the authors deemed these initially not relevant

due to their arguably (dis)similar approach in methodology or purpose.[35-37] Others were published during manuscript preparation or after submission.[38-40]

ACKNOWLEDGMENT
We thank Prof. Dr. Ulf Norinder for helpful discussion regarding model building and cluster analysis. Furthermore, Prof. Dr. Norinder, Dr. Tomasz Janosik and Dr. Ulf Tedebark for feedback on the manuscript.

Supporting Information

The following files are available free of charge.

- S1_Supporting_Information.pdf
  Contains: Descriptions/Details on Data Origin & Data Export from Reaxys®, Descriptors, Data-Analysis; mentioning of previous review-process.

- S2_Knime_workflows_with_readme.zip
  Contains: A Readme.txt file with brief overview of the enclosed workflows. Knime workflows used for data clean-up and preparation, as well as the Random Forest model.

- S3_Reaxys_ReactionIDs.zip
  Contains: Text files with Reaxys® ReactionIDs for all three reaction types. IDs are listed for both, before, and after data curation (including good/bad classification). These may be used as alternative data-origin starting point instead of the described keyword search in S1.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

Not applicable.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Corey, E. J.; Todd Wipke, W. Computer-Assisted Design of Complex Organic Syntheses. *Science (80-.).* **1969**, *166* (3902), 178–192. https://doi.org/10.1126/science.166.3902.178.

(2) J. Corey, E.; Todd Wipke, W.; D. Cramer, R.; Jeffrey Howe, W. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94* (2), 421–430. https://doi.org/10.1021/ja00757a020.

(3) Bogevig, A.; Federsel, H.-J. J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Low, P.; Oppawsky, C.; Rein, T.; Saller, H.; et al. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19* (2), 357–368. https://doi.org/10.1021/op500373e.

(4) Ravitz, O. Data-Driven Computer Aided Synthesis Design. *Drug Discov. Today Technol.* **2013**, *10* (3), e443–e449. https://doi.org/https://doi.org/10.1016/j.ddtec.2013.01.005.

(5) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.

(6) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - A Eur. J.* **2017**, *23* (25), 5966–5971. https://doi.org/10.1002/chem.201605499.

(7) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in

the Laboratory. *Chem* **2018**, *4* (3), 522–532. https://doi.org/10.1016/J.CHEMPR.2018.02.002.

(8) Miyaura, N.; Yamada, K.; Suzuki, A. A New Stereospecific Cross-Coupling by the Palladium-Catalyzed Reaction of 1-Alkenylboranes with 1-Alkenyl or 1-Alkynyl Halides. *Tetrahedron Lett.* **1979**, *20* (36), 3437–3440. https://doi.org/10.1016/S0040-4039(01)95429-2.

(9) Baba, S.; Negishi, E. A Novel Stereospecific Alkenyl-Alkenyl Cross-Coupling by a Palladium- or Nickel-Catalyzed Reaction of Alkenylalanes with Alkenyl Halides. *J. Am. Chem. Soc.* **1976**, *98* (21), 6729–6731. https://doi.org/10.1021/ja00437a067.

(10) Guram, A. S.; Buchwald, S. L. Palladium-Catalyzed Aromatic Aminations with in Situ Generated Aminostannanes. *J. Am. Chem. Soc.* **1994**, *116* (17), 7901–7902. https://doi.org/10.1021/ja00096a059.

(11) Paul, F.; Patt, J.; Hartwig, J. F. Palladium-Catalyzed Formation of Carbon-Nitrogen Bonds. Reaction Intermediates and Catalyst Improvements in the Hetero Cross-Coupling of Aryl Halides and Tin Amides. *J. Am. Chem. Soc.* **1994**, *116* (13), 5969–5970. https://doi.org/10.1021/ja00092a058.

(12) Skoraczyñski, G.; Dittwald, P.; Miasojedow, B.; Szymkuc, 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**. https://doi.org/10.1038/s41598-017-02303-0.

(13) Huerta, F. F. Innovation in Organic Synthetic Chemistry. In *Nordic Process Chemistry Forum*; Stockholm, 2018.

(14) Hallinder, S. Reaction Conditions Data Mining, *Thesis University Uppsala* **2019**, ISSN: 1650-8297, UPTEC K 19021.

(15) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning to Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**. https://doi.org/10.1021/acscentsci.8b00357.

(16) Davies, I. W. The Digitization of Organic Synthesis. *Nature* **2019**, *570* (7760), 175–181. https://doi.org/10.1038/s41586-019-1288-y.

(17) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3* (12). https://doi.org/10.1021/acscentsci.7b00355.

(18) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Prediction Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science (80-.).* **2018**, *360* (6385), 186–190.

(19) Sadowski, P.; Fooshee, D.; Subrahmanya, N.; Baldi, P. Synergies between Quantum Mechanics and Machine Learning in Reaction Prediction. *J. Chem. Inf. Model.* **2016**, *56* (11), 2125–2128. https://doi.org/10.1021/acs.jcim.6b00351.

(20) Walker, E.; Kammeraad, J.; Goetz, J.; Robo, M. T.; Tewari, A.; Zimmerman, P. M. Learning To Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst. *J. Chem. Inf. Model.* **2019**, *59* (9), 3645–3654. https://doi.org/10.1021/acs.jcim.9b00313.

(21) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J.; et al. Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models. **2019**.

(22) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57* (6), 1300–1308. https://doi.org/10.1021/acs.jcim.7b00083.

(23) 6. Fingerprints - Screening and Similarity https://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

(24) Bajusz, D.; Rácz, A.; Héberger, K. *Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching*; 2017; Vol. 3–8. https://doi.org/10.1016/B978-0-12-409547-2.12345-5.

(25) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl

Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004–5008. https://doi.org/10.1021/jacs.8b01523.

(26) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. {KNIME}: The {K}onstanz {I}nformation {M}iner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, 2007.

(27) Reaxys(R) http://www.elsevier.com/online-tools/reaxys.

(28) Landrum, G. Rdkit documentation: List of available descriptors http://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors.

(29) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(30) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068–2076. https://doi.org/10.1021/acs.jcim.7b00146.

(31) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.

(32) Chuang, K. V.; Keiser, M. J. Comment on "Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning." *Science (80-.).* **2018**, *362* (6416), eaat8603. https://doi.org/10.1126/science.aat8603.

(33) Norinder, U.; Boyer, S. Binary Classification of Imbalanced Datasets Using Conformal Prediction. *J. Mol. Graph. Model.* **2017**, *72*, 256–265. https://doi.org/10.1016/J.JMGM.2017.01.008.

(34) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie - International Edition*. 2016. https://doi.org/10.1002/anie.201506101.

(35) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370–377. https://doi.org/10.1039/c8sc04228d.

(36) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **2019**, *5* (6), 970–981. https://doi.org/10.1021/acscentsci.9b00055.

(37) Awale, M.; Sirockin, F.; Stiefl, N.; Reymond, J. Medicinal Chemistry Aware Database GDBMedChem. *Mol. Inform.* **2019**, *38* (8–9), 1900031. https://doi.org/10.1002/minf.201900031.

(38) David, L.; Arús-Pous, J.; Karlsson, J.; Engkvist, O.; Bjerrum, E. J.; Kogej, T.; Kriegl, J. M.; Beck, B.; Chen, H. Applications of Deep-Learning in Exploiting Large-Scale and Heterogeneous Compound Data in Industrial Pharmaceutical Research. *Frontiers in Pharmacology*. Frontiers Media S.A. November 5, 2019, p 1303. https://doi.org/10.3389/fphar.2019.01303.

(39) Bai, R.; Zhang, C.; Wang, L.; Yao, C.; Ge, J.; Duan, H. Transfer Learning: Making Retrosynthetic Predictions Based on a Small Chemical Reaction Dataset Scale to a New Level. *Molecules* **2020**. https://doi.org/10.3390/molecules25102357.

(40) Eyke, N.; Green, W. H.; Jensen, K. F. Iterative Experimental Design Based on Active Machine Learning Reduces the Experimental Burden Associated with Reaction Screening. ChemArxiv June 18, **2020**. https://doi.org/10.26434/chemrxiv.12465299.v1.