Reaction Chemistry & Engineering

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c9re00019d

Making better decisions during synthetic route design: leveraging prediction to achieve greenness-by-design†

Jun Li* and Martin D. Eastgate *

Modern pharmaceuticals are becoming increasingly complex. Incorporating knowledge of a route's holistic sustainability during the route design process could be a critical enabler to minimizing the environmental impact of pharmaceutical manufacturing. The pursuit of the optimal synthesis has historically been characterized by disconnection strategy, or things like step count, however, the optimal synthesis of a molecule may also be assessed through environmentally relevant metrics. The synthesis with the lowest possible cumulative process mass intensity (cPMI) could be considered optimal, a route which may not necessarily be the shortest, but has the best holistic sustainability (for example, considering the synthesis of all reagents and reactants). Previously, we demonstrated the importance of assessing the entire synthetic network by including "above-the-arrow" reagents/reactants into cPMI, to reflect the impact of reagents, such as ligands, on the overall sustainability of the route. Here we present the development of a machine learning approach, using substrate fingerprints, to build a multiclass predictive model to identify which ligands will likely function in a Pd-catalyzed C–N coupling reaction. The resulting predicted multiclass probabilities were then linked to the corresponding ligand cPMIs to yield a probability-weighted predicted holistic PMI for the transformation, integrating the synthesis of the ligand. This proof-of-confidence study may extend our ability to holistically assess different synthetic route options, considering their full impact, to aid decision-making during route ideation. This may lead to greener outcomes in the development of synthetic routes in the pharmaceutical sector and beyond.

Received 13th January 2019,
Accepted 19th February 2019

DOI: 10.1039/c9re00019d

rsc.li/reaction-engineering

Introduction

Since the first synthesis of acetyl salicylic acid (aspirin) in 1853, and its use as a medication in 1897,[1] the synthetic compounds used to treat human disease have been constantly evolving. In the modern era, the pharmaceutical industry is working to gain influence over ever more challenging biological mechanisms, resulting in an inescapable increase in the complexity and diversity of clinical drug candidates. Molecules such as Halaven™ (Eribulin, developed by Eisai),[2] which contains 19 stereogenic centers, 9 ring systems and is prepared commercially by total synthesis, have set a record for the complexity of a small molecule drug being brought to patients through chemical (total) synthesis. Synthetic peptides and oligo nucleotides, such as Spinraza, a molecule which contains chiral thiophosphate linked nucleosides, represent structures of exquisite complexity. These compounds stretch the capability of modern synthetic and analytical chemistry and pose a significant challenge to process research and development organizations – the groups responsible for developing the commercial manufacturing approaches to these compounds. A molecule with the complexity of Halaven makes the challenge of designing a commercial synthesis exceptionally difficult.[3]

A foundational problem in developing a commercial synthesis to a complex compound is simply the number of options and synthetic possibilities available for preparing the system, a natural consequence of combining molecular complexity with the expanding capability of the organic chemistry community. Within this context, the number of options available to potentially prepare a complex molecule represents a problem in decision making. Choosing which potential routes to explore in a laboratory is one of the first and most foundational challenges facing teams when embarking on an effort to design a synthesis of a complex molecule. This is especially true in a time-bound environment where the ability to explore multiple options may be limited. The impact of these

Chemical and Synthetic Development, Bristol-Myers Squibb, 1 Squibb Drive, New Brunswick, NJ, 08903 USA. E-mail: jun.li1@bms.com, martin.eastgate@bms.com

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9re00019d

decisions can be felt for the lifetime of a commercial asset, and define the environmental impact of its commercial manufacture.

We wished to demonstrate, quantitatively, that the complexity of pharmaceutical agents is indeed increasing and that this is not just a perception. Clearly articulating the problem would encourage the exploration of methodologies to mitigate the various impacts of such an increase. With our recently developed complexity index (a method of measuring molecular complexity that reflects both intrinsic molecular features such as the compounds structure, and extrinsic synthetic factors such as the technology available to prepare the compound, which is ultimately expressed through a 1–10 indexed score),[4] we can ensure that the example of Halaven is not simply an easily identified outlier. By analyzing the complexity of marketed drug compounds (from 1988–2018), we are able to detect a clear and rising trend in molecular complexity, exemplified by a study of the evolution of approved antiviral drugs (Fig. 1). This is a fascinating field, which recently delivered curative drug regimens capable of clearing hepatitis C infection to patients – a feat achieved with compounds of exquisite beauty and high molecular complexity.[5]

With a clear trend identified, one that highlights the industry wide challenge that increasing molecular complexity represents, we can explore the impact of this evolution. It is well known that the risks and uncertainties associated with developing pharmaceutical agents are significant (such as the high attrition rates observed during drug development),[6] these challenges compound the impact of synthetic complexity. Additionally, the industry is increasingly competitive, placing a premium on development speed, and regulators, scientists, payers, patients and populations are far more cognizant of the impact of pharmaceutical manufacturing on the environment, making 'green chemistry' an important consideration industry wide. The desire to enhance the efficiency of chemical manufacturing, to promote 'green chemistry' principles, and to develop sustainable processes are highly priority goals for most organizations. This is a truly important pursuit and many development teams are focused on process greenness and sustainability. Cross company collaborative organizations, such as the ACS Green Chemistry Institutes Pharmaceutical Roundtable (GCIPR), work in the pre-competitive space to explore this topic and promote tools and strategies for the adoption of 'green chemistry' principles industry wide.[7]

It is worth addressing the question of exactly why complexity matters. In the context of organic chemistry, the complexity of a molecule is driven by its structure – the way the atoms in the compound are both arranged in space and bound together – it is often expressed through the difficulty of making the substance. As the size (number of atoms) and the complexity of their arrangement in space and connectivity (bonding) increases, the number of potential routes to (or ways to make) a compound increases significantly. While not quantitatively assessed, as molecular size increases, the number of synthetic options seems to expand exponentially. Thus, a high complexity environment results in what we have termed 'the problem of choice'. With a large number of potentially viable ways to make a molecule, it becomes ever more difficult to select an approach that will result in the most efficient synthesis of the molecule at-hand, without human bias entering the decision making process.[8] What decision making tools one has, how one can predict how efficient a route

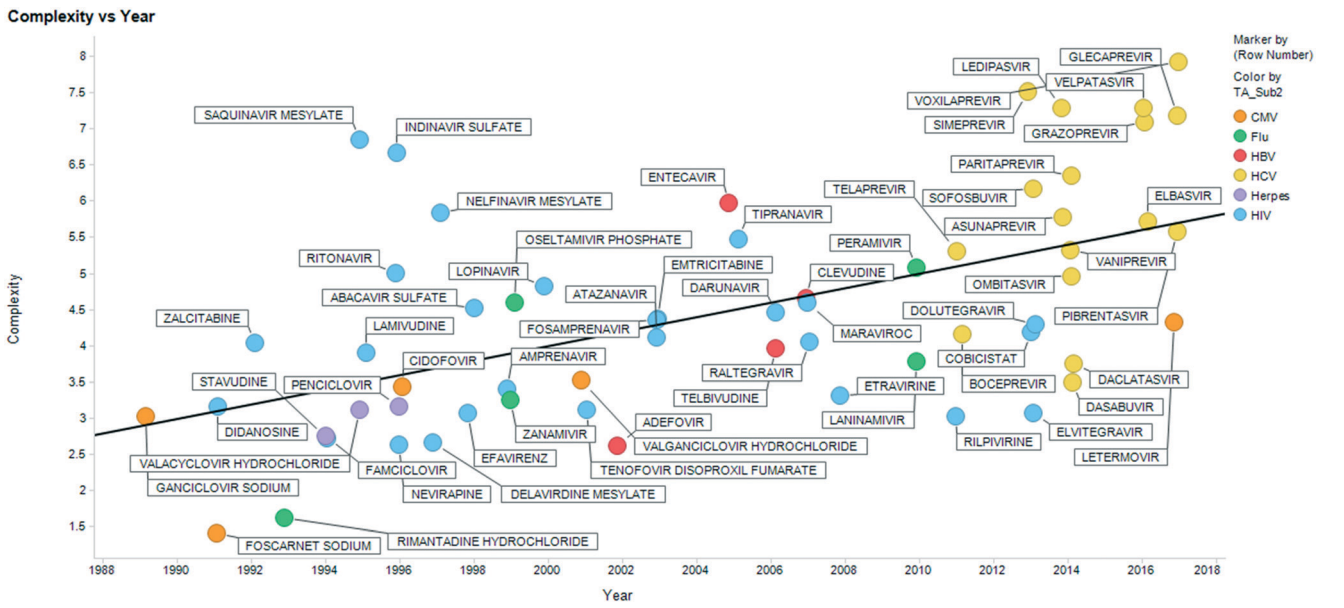


Fig. 1 The complexity of marketed small molecule antiviral drugs from 1988 to 2018. Colors represent the disease target.

will be, how one can estimate development cost and end point (such as yield or efficiency) is a significant problem statement in decision making.

Commercial route selection, the process adjudicating which route of synthesis should be moved forward to regulatory approval and commercial supply, is one of the key decisions of a pharmaceutical development organization, and a defining moment that determines the efficiency and cost of the manufacturing route to the active pharmaceutical ingredient. A rational approach to the assessment of manufacturing routes was developed using the SELECT (safety, environment, legal, economics, control, and throughput) criteria,[9] originally proposed by a consortium which included AstraZeneca, GSK and Pfizer. Based on these and other factors, systematic frameworks to exercise multi-criteria decision analysis, MCDA, on route selection have been reported.[10] In the modern era, the development of a *de novo* synthesis to a drug candidate is often critical to supplying the therapeutic commercially.[11] However, the application of MCDA in the early phase of synthetic route options analysis is limited, due largely to the high degree of uncertainty and risk associated with the synthetic options being considered; an ideal situation would involve analysis and comparison prior to any work being initiated. In such a situation, process details can be estimated by expert opinion, to provide qualitative scores, but without the support of fact-based data. A situation prone to error, inaccuracy, personal bias and unlikely to reliably maximize synthetic efficiency.

Once a proof-of-concept is achieved for a route, it is a relatively simple task to quantitatively compare it to prior known approaches to the same molecule. However, judging whether the 'best' strategy has been selected (one that gives an efficiency greater than any other option, including those not physically investigated) *vs.* the approach only being better than a prior known route, is currently not possible, neither is judging how efficient the route is compared to routes to other molecules (benchmarking the outcome). In our view, this therefore represents a significant gap in our capability and reduces our ability to optimize efficiency in the age of complexity. In the ideal scenario we would be able to explore all the synthetic options and make decisions based on which approaches have the highest probability of delivering the most efficient outcome. Thus, new strategies are needed for comparing options, which can solve the 'problem of choice' by helping innovators understand the probable consequences of the route options being considered – with respect to the SELECT criteria, environmental sustainability and other aspects of efficiency.

In order to address this conundrum, we sought to leverage simple quantitative metrics commonly practiced throughout the industry, and aligned with Green Chemistry principles,[12] to develop a predictive strategy to use in a decision making context. Specifically, we focused on mass-based metrics, such as process mass intensity (PMI),[13] which is one of the most widely used metrics[14] for tracking the efficiency characteristics of a specific chemical process. Based on previous reports[15] mass-based metrics have a clear correlation to both waste generation and manufacturing costs. Thus, we explored the potential to predict the cumulative PMI (cPMI) of an entire proposed synthetic sequence, through analysis of historical BMS data. To test the relevance of this approach, we evaluated PMI outcome data against production costs (excluding raw material costs), where we confirmed that cPMI is positively correlated to production costs (Fig. 2) – demonstrating not only the relevance of PMI to sustainability, but also to improving the costs associated with drug substance manufacturing.

In addition to positively correlating cost and sustainability, we rationalized that PMI can also be utilized to gauge the throughput of the process based on the inverse relationship with process throughput metrics (such as yield per Vmax per week).[16] This provides a multi-purpose synthetic efficiency surrogate, unifying environmental impact (at least in terms of consumption and waste generation), cost, and throughput to metricize a synthetic route during evaluation. This thought process became a foundational rationale for our development of a Monte Carlo based predictive analytics framework,[17] based on cPMI, focused on predicting the efficiency of a proposed synthetic route.

The initial method we developed leverages real-world data to predict probable PMI ranges for potential synthetic approaches being considered, based on the demonstration that certain reaction types have unique PMI signatures,

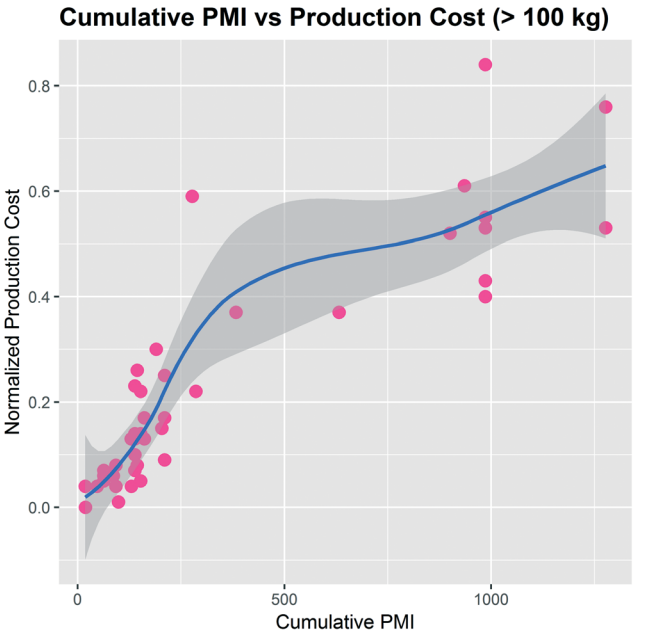


Fig. 2 Normalized production cost *versus* cumulative PMI observed in BMS data.

which span a range of molecules and phases of development. As this approach is based on real-world data, it can also be employed as a benchmarking methodology capable of comparing PMIs across molecules – enabling an assessment of the efficiency of a route *vs.* the aggregate of prior art.[17a] In order to enable broad adoption of this concept, the GCIPR members collaborated to enhance the data within the model and host a web-based application.[18] A simple hypothetical example of this concept (Fig. 3) explores the same total number of nodes for starting materials, intermediates and final API, estimating yields and step PMIs as being equivalent across five different options. The most convergent option 5 provides the lowest cPMI *versus* the linear approach, in line with a chemists' expectation, maximizing the synthetic efficiency *vs.* the convergent route.[19] However, real world examples are not this simple, especially when integrating the impact of reagents, yield variation and different PMI outcomes – not to mention the different reagents and reactants needed to achieve these transformations. Thus, a predictive approach based on ranges observed in similar settings can provide a quantitative assessment to guide decisions on route strategy.

Our initial strategy leveraged reaction PMI – simply the amount (in terms of mass) of reagents, reactants and solvents used in the chemical reaction and isolation *vs.* the amount of product obtained. However, these reagents, reactants and solvents also need to be prepared – and their synthesis is not included in the normal PMI metric. While this information may be included in a full life-cycle analysis, this is a significant undertaking and difficult to use in an early phase of development. However, the ultimate (holistic) impact of a route is the aggregate of the synthesis of all components used in the formation of the desired product. While many of the components used in a reaction will have trivial contributions, some reagents involve very lengthy synthetic approaches – for example, the ligands employed in many modern metal catalyzed reactions may consume a significant amount of resources in their preparation.[17b] With this in mind, selecting a route which employs such reagents has the risk of being unsustainable in the context of other options, which may be preferable when looking at the holistic impact of the strategy. In this context we sought to explore the prediction of high impact 'above and below the arrow' reagents and reactants (Fig. 4), in an attempt to predict the impact of high complexity components on the overall synthetic route. In concept, this meant predicting when a 'high impact' reagent or reactant would be utilized, and including the PMI contribution for the synthesis of the probable compounds needed to effect the transformation in the overall cumulative PMI. The implications of these findings have been discussed recently.[20]

This thought process led us to work towards establishing a predictive strategy, capable of enabling a holistic view of sustainability during the initial stages of the route design process. In order to explore this idea, we first focused on catalytic transformations, which required us to explore methods to predict the likely ligands (amongst a large set of potential options) for certain reactions involving specific (or proposed) substrates.

Taking note of the recent progress in machine aided synthesis planning,[21] we hoped to expand our PMI prediction by leveraging a machine learning (ML) approach to

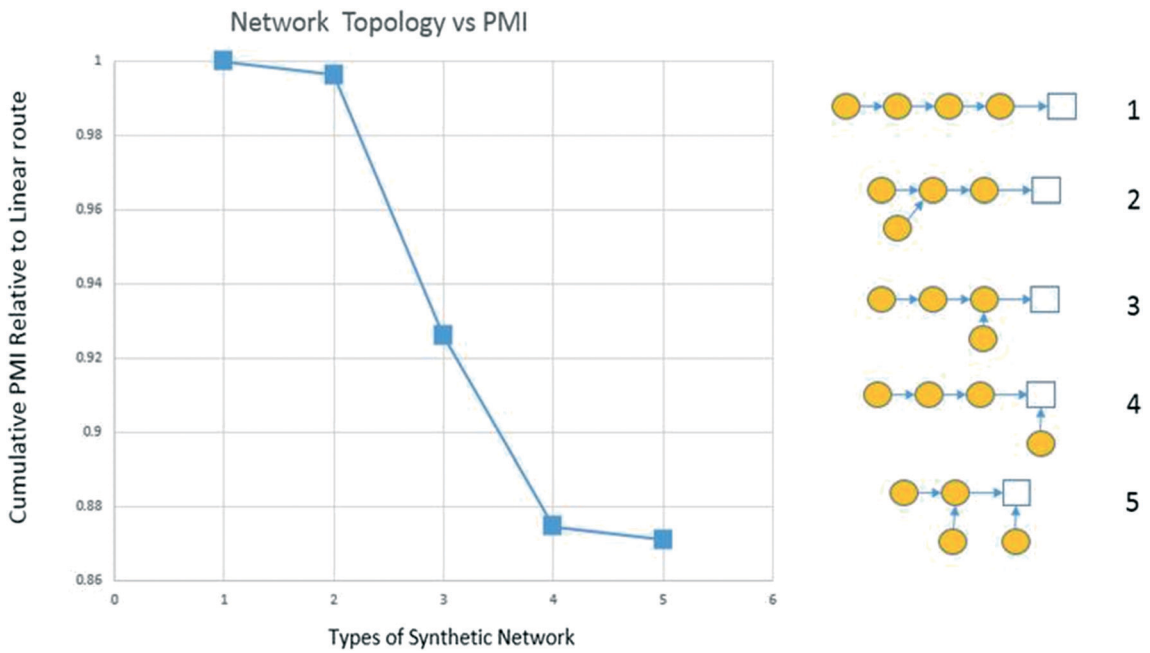


Fig. 3 Synthetic network topology *versus* cumulative PMI.

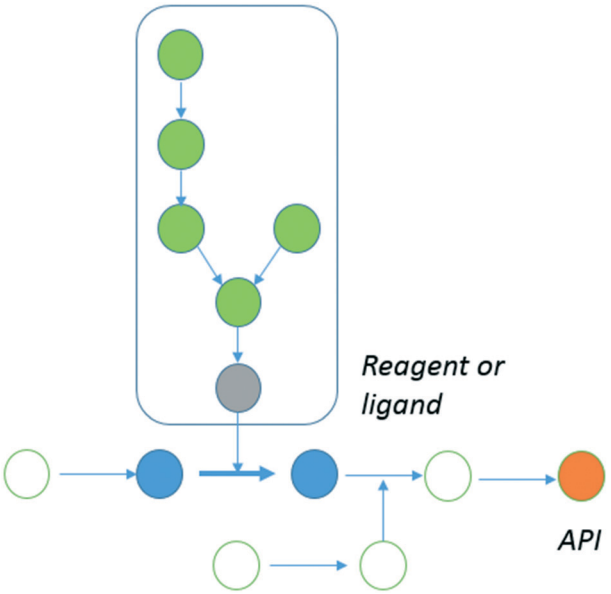


Fig. 4 Consideration of "above-the-arrow" reagents or ligands preparation in cumulative PMI.

ligand identification for proposed catalytic transformations. In the present manuscript we describe a novel ML approach to predicting which ligands will be successful in a given reaction, on a given substrate structure, prior to their experimental exploration. This strategy enables the PMI contribution of the potential ligands to be estimated and included in the context of the synthesis of a target molecule.

Background and approach

Transition metal catalyzed coupling reactions have become an indispensable tool, used routinely for over three decades in both the discovery of new chemical matter and in the large scale synthesis of pharmaceutical intermediates and APIs.[22] To increase selectivity, maximize catalytic efficiency (often defined by the stoichiometry of the mediating metal center – not the overall consumption of materials including ligand preparation), enable new processes and to develop enantioselective processes, different ligand scaffolds have been devised to overcome various challenges in transition-metal mediated reactions. Selecting the correct ligand for optimal reaction performance is sometimes non-intuitive and always involves the screening of multiple options, along with traditional reaction optimization. Several attempts to develop predictive models of various ligand systems and their parameterization, in the hope of understanding the key factor which influence reaction outcome, have been made using multivariate correlation, as well as ML approaches.[23] Atomic, molecular, and vibrational descriptors, amongst others, have been used to improved predictive performance for high throughput screening reactions. Interestingly, a representation system was proposed that did not require prior chemical knowledge, using one-hot encoded vectors for reactants, ligands, solvents, and bases. This matrix of factors was sufficient for predicting reaction performance in a high throughput setting.[24]

As discussed above, we were interested in developing a proof-of-confidence method to predict the probability of a ligand being considered a 'hit,' out of a group of candidates and within a given transition metal mediated coupling. Inspired by recent work leveraging molecular fingerprints to predict reaction classifications, as well as the prediction of reaction outcomes,[25] we explored a ML approach for the prediction of phosphine ligands in a palladium catalyzed C–N coupling reaction – the coupling of an electrophile (–OTf, I, Br, or Cl) and a nitrogen nucleophile (primary, secondary aliphatic or aromatic).[26]

A large array of phosphine ligands have been explored in C–N couplings (Fig. 5), leading to a large amount of reaction data for the preparation of heteroatom containing pharmaceutically relevant compounds.

The concept of ML can be described as 'learning a target function which can best map the input variables to an output variable'. In the example of metal catalyzed processes, the input variables can be represented by topological fingerprints encoding the molecular features from electrophile, nucleophile with or without product. This can be readily actualized by representations such as extended-connectivity fingerprints (ECFP).[27] The output variables are the phosphine ligands observed to have been successful (definition of success based on a used defined criteria) in effecting the desired transformation. Establishing this type of dataset, using substrate structures as diverse as possible, is akin to mimicking an individual's knowledge of the Pd-catalyzed C–N coupling literature. We therefore built our ML approach around this concept, choosing not to include ancillary variables, such as solvent, base, Pd type, ligand-to-metal ratio, temperature, concentration and reaction time, as such conditions are insufficiently reported and remove the ability to compare between different ligands and substrates. By excluding these variables, data aggregation is facilitated across more information sources, allowing us to generate and compare the probability of different ligands working for similar substrates.

Data

We sought to establish an appropriate dataset which captures all the major examples of C–N coupling from our own internal experiments and supporting literature, to generate a dataset with a wide enough coverage of chemical structures to enable our 'proof of concept' exploration. After manually curating ~5000 screening reactions, successful hits were extracted from experiments containing phosphine ligands. To augment the data, additional examples of C–N couplings were added from the literature. One limitation with mining literature information is the absence of broad negative hit

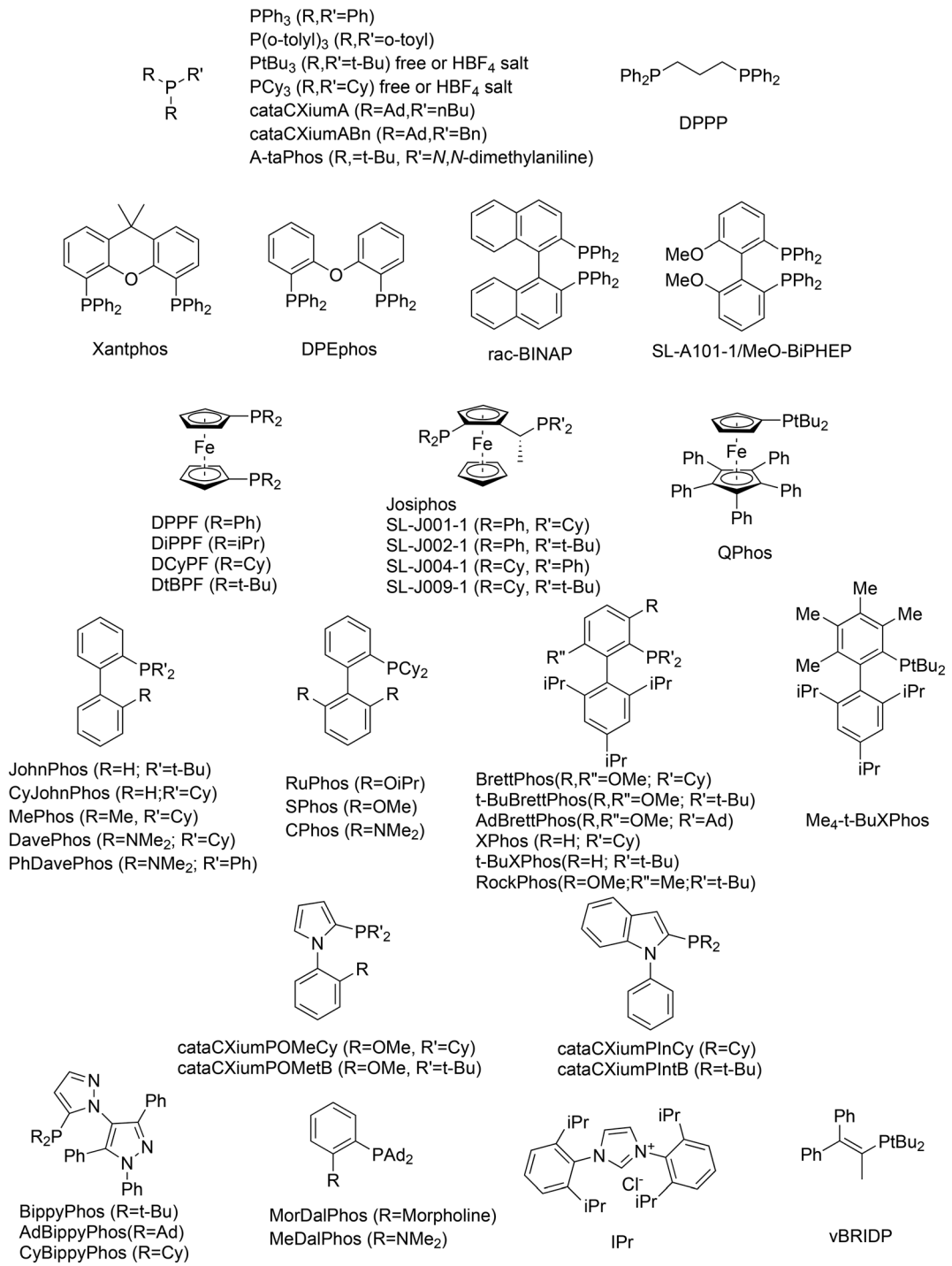


Fig. 5 Some examples of popular phosphine ligands used in C–N coupling processes.

information – limiting us to imply a positive *vs.* negative (absence of positive) criteria (*i.e.* ligands not mentioned = non-functional ligands, in reality this is not true, they may have simply not been tested or just performed worse and therefore not reported). In the area of catalyst screening, multiple ligands can often be found to deliver the desired reactivity, albeit with varying outcomes. A decision then has to be made to narrow the selection for further optimization and eventual scale-up. This poses a second challenge, where unreported ligands can be positive hits while missing from the model, due to a multitude of reasons, including not being tested. One way to assess the classification model using these kind of data is to look at 'Top-N' accuracy metrics (N represents a stratification of the analyte), where identifying the top-1 result is less useful due to the natural inaccuracy of the approach, while the top-N set (a group of probable ligands)

Table 1 Top-N accuracy of different ML methods for testing and validation sets

		Top 1 accuracy		Top 5 accuracy		Top 10 accuracy		Top 20 accuracy	
		Method A	Method B	Method A	Method B	Method A	Method B	Method A	Method B
Testing	Random guess	2%	2%	9.6%	9.3%	19%	19%	38%	37%
	Naïve Bayes	4%	2%	12%	8%	18%	17%	44%	37%
	Random forest	3%	3%	10%	11%	22%	20%	48%	47%
	KNN	2%	4%	9%	13%	15%	19%	53%	54%
	Logistic regression	2%	3%	13%	13%	23%	25%	41%	43%
	SVM	1%	1%	7%	10%	14%	16%	49%	40%
	NeuralNet	19%	18%	42%	39%	63%	58%	80%	79%
Validation	NeuralNet	29%	29%	64%	57%	79%	71%	86%	93%

Note: method A removed some intramolecular reactions while method B kept all the reactions. Testing set is based on average top-N accuracy from *k*-fold cross-validation with *k* is 3 fold.

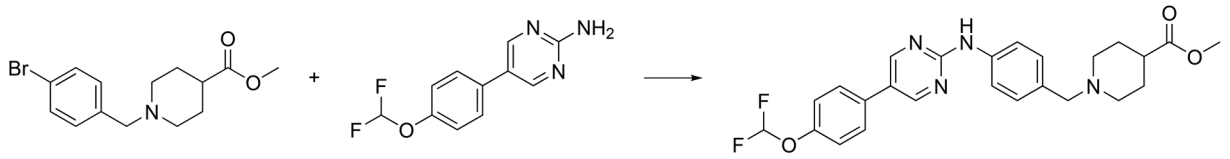
finds the correct class within a number of possibilities. We believed a cluster of top-N ligands with associated probabilities can be reasonably suggested from a ML algorithm, therefore enabling the prediction of a subset of ligands working in a given C–N coupling and likely contribution of these ligands to holistic efficiency.

While we have articulated the problem statement and approach, this strategy represents a challenging machine learning problem. Even though manual data collection[28] can be achieved, there is an enormous design space, compounded by multiple factors: such as the significant substrate diversity which functions in these reactions, the large number of ligand classes that can enable this chemistry and the reality that a diversity of reaction conditions is needed. Thus, in a multiplicative fashion, deconvoluting this complexity is non-trivial. Despite these concerns, we compiled four hundred cases, though manual data collation, to initiate our exploration and inquisition of this strategy.

Machine learning approach

Reaction representations were devised using the following two methods. Method A is to generate molecular fingerprints from both electrophile and nucleophile and have them concatenated as the model input. This flexible framework allows us to include any other parameters such as solvents, bases, or additives in the future. Method B is to adopt the reaction 'difference' fingerprint[25a] without reagents, which helps focus on the location of the transformation rather than using separate reactant and product fingerprints. The general experimental setup for the classification was as follows: the manually curated dataset including literature examples, was split into training and testing subsets. For method A, molecular circular fingerprints such as Morgan radius 3 (similar to ECFP6) were calculated for both electrophile and nucleophile molecule. The python Numpy array of each reactant were concatenated. For method B, a reaction difference fingerprint was calculated for each of the reactions. The resulting fingerprints from both methods were used to train a multi-class machine-learning (ML) model to predict the type of phosphine ligand used. The model was tested on an external validation set collected from BMS internal examples, as well as literature examples not included in the model (a period between June 2016 and 2018). The top-N accuracy metrics were calculated in the cross-validated testing subset and the external validation set. Different machine learning models including naïve Bayes, random forest, support vector machine, logistic regression, KNN, and neural network were used. The approach was implemented using the open-source cheminformatics toolkit RDKit (version 2017.03.01), scikit-learn (version 0.18.1) and Keras (version 2.0.8) (Table 1).

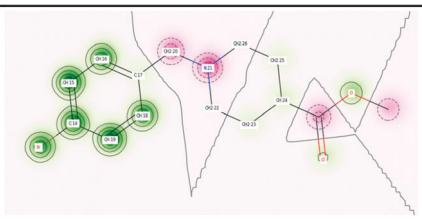
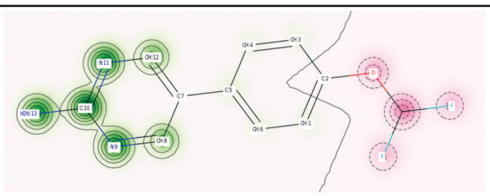
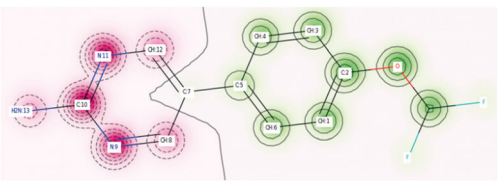
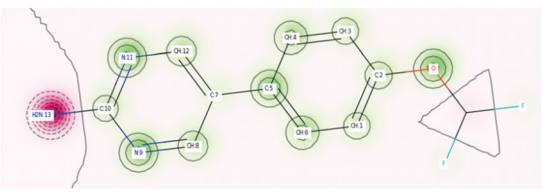
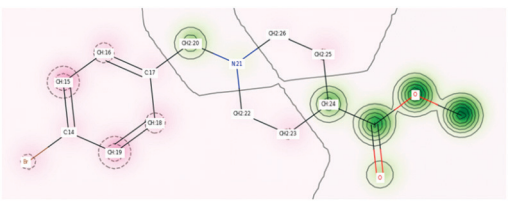
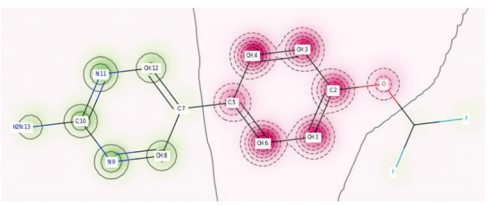
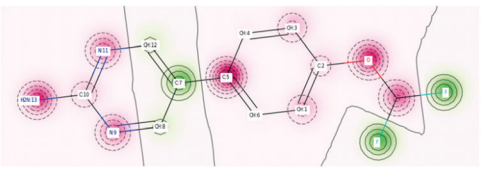
The ML results showed that traditional methods give a very modest improvement over a random guess, except for KNN (the *K*-nearest neighbor algorithm) which showed a minor improvement within the top-20 accuracy category 53–54% *versus* 37–38%. Clearly, none of the traditional ML methods are helpful in predicting the ligand choices. However, the neural network algorithm under Keras library showed significant improvement in prediction performance across the board. Based on the average testing set results from *k*-fold cross-validation, a more than five-fold accuracy increase was observed in the top-5 category, along with three-fold increase in top-10 accuracy category. Most importantly, the validation set provided convincing results to support the neural network model. In the Keras neural network, the objective function being minimized is the categorical cross entropy. The optimizer plays a significant role in tuning the model accuracy. The default parameter setting in Adam



Scheme 1

optimizer provided the highest prediction accuracy in method B (reaction difference fingerprints) while either reducing or increasing the learning rate would deteriorate the model prediction accuracy. We found in method A, the stochastic gradient descent, sgd optimizer gave a better model prediction performance than using Adam optimizer. We believe that the model can be further improved with the addition of new data as research into these coupling processes continues. The framework can also be extended into other transition metal or non-metal catalyzed reaction types to extract more insights from the successful usage of certain catalyst systems associated with reactants.

To further probe if the molecular environment of the electrophile or nucleophile was influencing the choice of a particular phosphine ligand, we explored a fingerprint similarity visualization approach,[29] to see the atomic contributions to the predicted probability of a machine learning model. We used this method to explore method A where concatenated circular fingerprints were used as the model input. In this approach, a "weight" is determined for each atom of either electrophile or nucleophile, by removing the bits set by the atom in the fingerprint of the molecule, recalculating the probability of specific ligand type, and compared to the original unchanged fingerprints to register whether it is increasing (or reducing) the chance of using this particular ligand type. To visualize the influence, the weights are normalized by dividing by the maximum absolute weight value which are then used to calculate bivariate Gaussian distributions centered at the corresponding atom positions. The color scheme is as follows: if the atom is green (meaning that removing the bits associated with this atom reduces the probability of using the specified ligand type), implying the positive influence of this atom toward this ligand type. On the other hand if the atom is pink (meaning that removing the bits associated with this atom increases the probability of the associated ligand type), implying the negative influence of this atom for this ligand type. An example of the validation is shown below (Scheme 1).[30] In our prediction, it is gratifying to see the top-1 candidate Xantphos matched exactly the ligand used in the initial report.

Ligand type	Electrophile	Nucleophile
Xantphos		
BINAP		
t-Bu-Brettphos		
XPhos		
Brettphos		

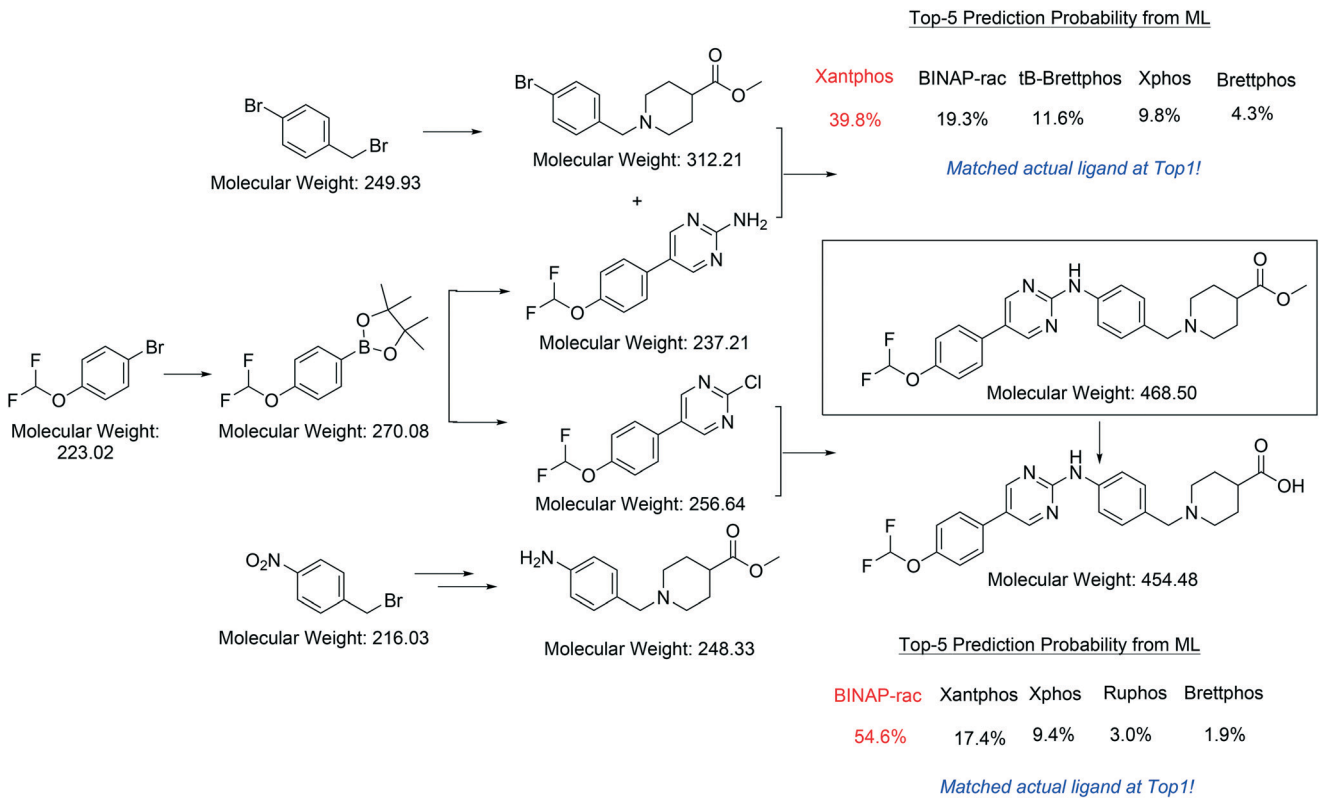
From this visualization we can see that with Xantphos, variation in either component close to the reaction center are favored when using this ligand type (green color around amine and bromide groups). BINAP, for example, showed that the benzyl group in piperidine ring and part of phenol moiety in nucleophile favored this ligand type, while the atoms around the reaction center seem disfavored for this ligand type. With this example, we can showcase the possibility of using machine learning to predict not only the ligand, but to enable us to ask deeper questions of the chemistry itself.

Bridging probability to 'above the arrow' impact in a PMI predictive framework

As noted above, to truly assess a synthetic route's sustainability, one needs to understand the holistic impact of the proposed chemistry. While one can correlate cost of reagents/reactants to impact, being able to estimate potential reagents and quantitate their sustainability is critical to good decision making, thus we decided to integrate our ligand prediction with an estimation of the PMI contribution for the synthesis of those reagents. We did this by using the reported synthesis of the ligand, starting from raw materials with a cost no greater than $100 per mol, as proposed by Roschangar *et al.*[14d] Here we leveraged our recently launched PMI predictor web application,[18] which can readily convert a synthesis to probable cPMI for any number of phosphine ligands incl. Brettphos, *t*-Bu-Brettphos, XPhos, RuPhos, *rac*-BINAP, and Xantphos. Detailed info is provided in the ESI.†

To demonstrate the application of our machine learning approach to the estimation of cPMI, we explored the Novartis synthesis of the c-Kit inhibitor. A key transformation in the synthesis of this compound is a Buchwald–Hartwig coupling. The results of our prediction matched the actual experimental findings, where the top ligand choices of Xantphos and BINAP were employed by the Novartis team and applied in scale-up. In our prediction, the probability for each of the ligands in the top set included, Xantphos (39.8% probability), followed by *rac*-BINAP (19.3%), *tert*-Bu-Brettphos (11.6%), Xphos (9.8%) as well as Brettphos (4.3%) in the top-5 bracket (Scheme 2). The application of this knowledge can be twofold: we can immediately increase the efficiency of the high throughput screening campaign, through better design of the screening (*i.e.* narrowing ligand selection). The 2nd goal is to provide a gateway to explore the potential efficiency of the overall synthetic route during the planning phase, encompassing our previous above-below-arrow concept. Here we want to illustrate how this concept can be realized.

With a knowledge of the individual cPMI for each of the ligands in the predicted top-5, we can easily integrate the efficiency of their synthesis into the main route. As shown above (Fig. 6), between nodes A and B in the network graph represent the borylation step. The arrow between nodes B and C represented a Suzuki reaction. Node H stands for the benzylated piperidine. The arrow between nodes G and F



Scheme 2 Machine learning evaluation of swapping amine and halide in different coupling partners.

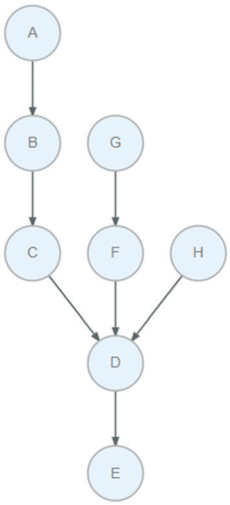


Fig. 6 Synthetic scheme represented as the network figure in PMI web app.

represents the ligand synthesis part where G stands for the ligand starting material on its longest linear sequence and F stands for the phosphine ligand used in making the C–N cross-coupling product D.

One of the most important parameters to introduce is the stoichiometric ratio ranges, required to capture the uncertainty in any predictive approach to PMI. Here we would like to show how each of the ligand loading amounts affected the changes in overall synthetic route cPMI. The following plot (Fig. 7) gave the trend lines of cPMI percentage increase for each of the ligand type loadings. The cPMI percentage increase is based on 2 mol% ligand loading as the baseline condition. Both route 1 and 2 have exactly the same trend lines. As we have pointed out in the previous studies, the step location of the transition metal catalyzed reaction played a significant role on overall cPMI. Currently, both of the C–N coupling reactions are all located in the penultimate step for each route. If the catalytic step is moved further away from the API step, the magnitude of the cPMI increase based on the ligand loading will be magnified. Here the slopes of the trend lines were dictated mainly by the size of the cPMI in each of the ligand synthesis. This concept allows the prediction to inform on the ligand loading needed to achieve a holistically efficient outcome *vs.* other route options which may be competitive.

In the current C–N coupling, due to heterogeneity of the reaction system and reactivity of the amine groups, catalyst deactivation tends to occur. For the purpose of illustration, we will assume a 5 mol% total ligand loading will be needed. We therefore obtain the cPMI for the different ligand types at this specified loading first, later we can explore the expected cumulative PMI contribution from a cluster of ligand types by introducing the normalized probability of success for each ligand type from the top set results. The probability weighted average of cPMI for the entire route is a reasonable representation of this, taking into account a cluster of above-below-arrow ligand contributions for a specific catalytic transformation.

$$\mathrm{cPMI}_{\mathrm{Cluster\ of\ Ligands}} = \sum_{\mathrm{i=ligand}} \mathrm{Probability}_{\mathrm{i}} \times \mathrm{cPMI}_{\mathrm{ligand}}$$

The graph below (Fig. 8) showed the probability-weighted cPMI distributions for route 1 and 2 are not differentiable,

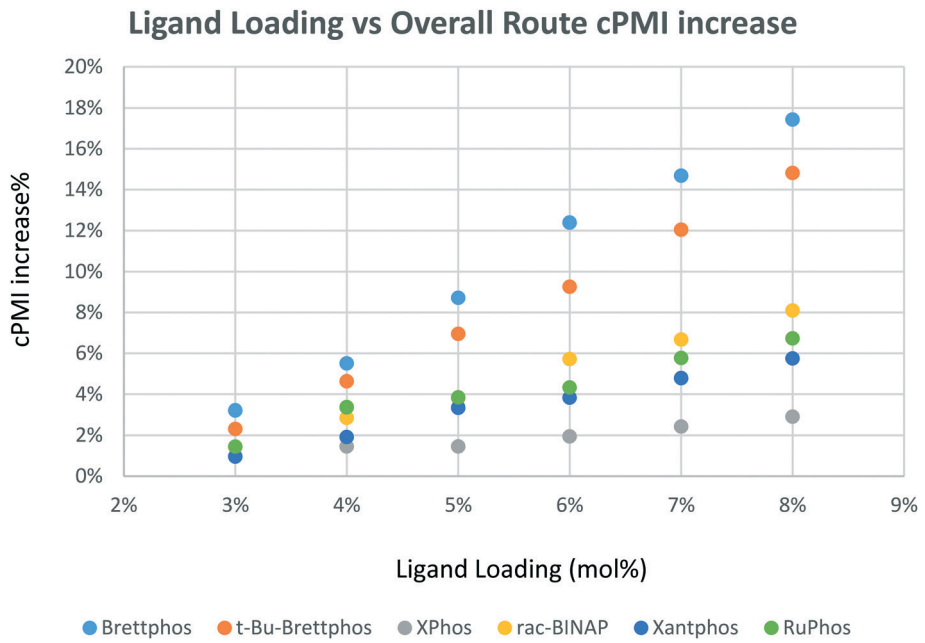


Fig. 7 Ligand loading *versus* overall synthetic route cPMI increase.

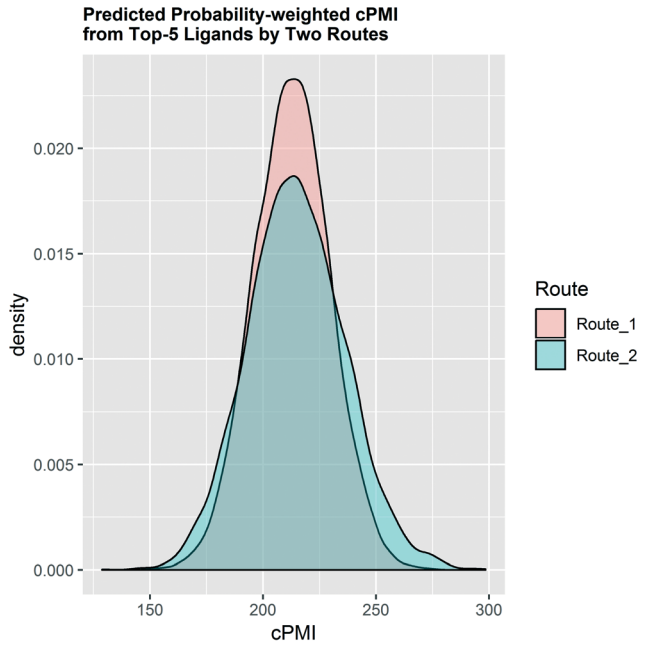


Fig. 8 Predictive distribution of probability-weighted cPMI from two routes with different sets of top-5 predicted ligands.

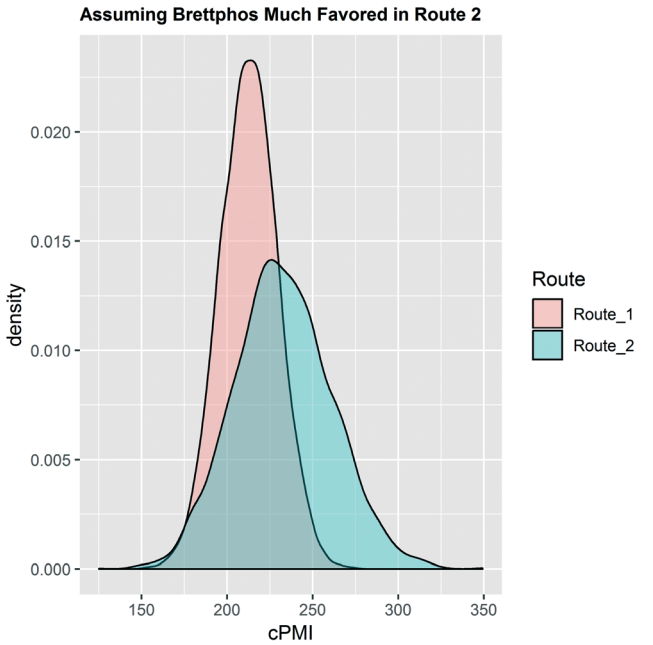


Fig. 9 Predictive distribution of probability-weighted cPMI assuming route 2 favors Brettphos in the top-5 predicted ligand set.

which was expected based on the relatively low cPMIs for the top 2 ligands in Xantphos and *rac*-BINAP.

Now let us assume the normalized probability in route 2 has changed to the following order of sequence favoring Brettphos: Brettphos 90%, RuPhos 7%, and the rest of the ligands are 1% each. In this hypothetical case, we can see the probability-weighted cPMI distribution for route 2 is now shifting toward a higher cPMI for the overall synthetic route (Fig. 9), making route 2 slightly disfavored in terms of the overall synthetic efficiency. If the high probability ligands synthesis has a much larger cPMI, or the C–N coupling step is much further away from the API step, we could see this AI-driven above-the-arrow predictive analytics can guide us toward much greener choices during route design, if no other major advantages were observed from the alternative synthesis plans.

This probability weighted approach helps mitigate the issue of bias from focusing upon a singular ligand contribution, which may be drastically inaccurate in defining the above-the-arrow possibilities. This revised strategy, equipped with AI-driven algorithmic capability, provides a more holistic view of the global sustainability for any designed route in the synthesis plan. With the machine learning derived probabilities of the success for a cluster of ligands in a specific designed reaction in the synthesis plan, we can apply the above formula to obtain the expected cPMI and use them to find potentially credible differentiation in efficiency between alternative synthetic routes.

Conclusion

We have developed a novel machine learning approach to estimate the probability of success for any ligand in a cluster, given specified electrophile and nucleophile combinations, demonstrated in the context of a Pd-catalyzed C–N coupling. The neural network helps improve the predictive performance of the top-N accuracy over other machine learning methods. This demonstration is a proof-of-confidence study to show that predicting the probability of success for a cluster of ligands can be leveraged to integrate their cumulative contribution to the cPMI of an overall synthesis. We believe this AI-driven approach can lead to better and more holistic above-the-arrow assessments, and avoid potential under reporting of the PMI impact of a given strategy and its inherent ligand selections in catalytic transformation, which can ultimately lead to poor decision making at a strategic level.

With this newly enhanced capability to predict holistic PMI, integration of this methodology into the route design process can serve as a key decision aiding tool to bring deeper context to gauging the sustainability impact of synthetic strategy decisions made during the early phases of route exploration. Expansion of these tools will enable richer data analysis in a multi-criteria decision analytics framework, enabling the best manufacturing processes to be selected and developed, thereby reducing the environmental impact of pharmaceutical manufacturing in the age of ever increasing complexity.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thanks the CSD SLT for supporting this work, Dr Jacob Albrecht, Dr Alina Borovika and Dr Eric Simmons for

inspiring discussions. We also thank Victor Hung, Dr Monica Fitzgerald and Dr Elias Mattas for support in accessing data.

References

1 D. Jeffreys, *Aspirin: The Remarkable Story of a Wonder Drug*, Chemical Heritage Foundation, 2008, ISBN 9781596918160.
2 H. Ledford, *Nature*, 2010, **468**, 608.
3 (*a*) G. Pisano, *The Development Factory – Unlocking the Potential of Process Innovation*, Harvard Business School Press, Boston, 1997; (*b*) N. G. Anderson, *Practical Process Research and Development*, Academic Press, San Diego, 2000.
4 J. Li and M. D. Eastgate, *Org. Biomol. Chem.*, 2015, **13**, 7164–7176.
5 (*a*) N. A. Meanwell, *J. Med. Chem.*, 2016, **59**, 7311–7351; (*b*) D. L. Hughes, *Org. Process Res. Dev.*, 2016, **20**, 1404–1415.
6 P. H. Carter, E. R. Berndt, J. A. DiMasi and M. Trusheim, *Nat. Rev. Drug Discovery*, 2016, **15**, 673–674.
7 (*a*) ACS Green Chemistry Institute® Pharmaceutical Roundtable, *Solvent Selection Guide: Version 2.0*, March 21, 2011. https://www.acs.org/content/dam/acsorg/greenchemistry/industriainnovation/roundtable/acs-gci-pr-solvent-selection-guide.pdf; (*b*) L. Diorazio, *Solvent selection tool principles and guidance*, June 2018. https://www.acs.org/content/dam/acsorg/greenchemistry/resources/acs-gci-solvent-selection-tool-principles-and-guidance.pdf.
8 (*a*) H.-J. Federsel, *Acc. Chem. Res.*, 2009, **42**, 671; (*b*) C. M. Cimarusti and D. R. Kronenthal, The Discovery/Development Transition in *Early Drug Development: Bringing a Preclinical Candidate to the Clinic*, Wiley-VCH, 2018, p. 31.
9 M. Butters, D. Catterick, A. Craig, A. Curzons, D. Dale, A. Gillmore, S. P. Green, I. Marziano, J.-P. Sherlock and W. White, *Chem. Rev.*, 2006, **106**, 3002.
10 (*a*) J. D. Moseley, D. Brown, C. R. Firkin, S. L. Jenkin, B. Patel and E. W. Snape, *Org. Process Res. Dev.*, 2008, **12**, 1044–1059; (*b*) R. B. Leng, M. V. M. Emonds, C. T. Hamilton and J. W. Ringer, *Org. Process Res. Dev.*, 2012, **16**, 415–424; (*c*) A. Manipura, E. B. Martin, G. A. Montague, P. N. Sharratt and I. Houson, *Comput. Chem. Eng.*, 2013, **55**, 71–82; (*d*) P.-M. Jacob, P. Yamin, C. Perez-Storey, M. Hopgood and A. A. Lapkin, *Green Chem.*, 2017, **19**, 140–152.
11 M. D. Eastgate, M. Schmidt and K. R. Fandrick, *Nat. Rev. Chem.*, 2017, **1**, 1–16.
12 (*a*) P. T. Anastas and J. C. Warner, *Green Chemistry: Theory and Practice*, Oxford University Press, New York, 1998; (*b*) R. A. Sheldon, *Chem. Ind.*, 1997, 12–15.
13 (*a*) A. D. Curzons, D. J. C. Constable, D. N. Mortimer and V. L. Cunningham, *Green Chem.*, 2001, **3**, 1–6; (*b*) C. Jiménez-González, C. S. Ponder, Q. B. Broxterman and J. B. Manley, *Org. Process Res. Dev.*, 2011, **15**, 912–917; (*c*) D. Hughes, *Pharma and Suppliers: Collaborating on Green Chemistry. Launch of PMI tool*, Feb. 2011, https://www.acs.org/content/dam/acsorg/greenchemistry/industriainnovation/gcipr-informex-2011-pmi-tool.pdf; (*d*) https://www.acs.org/content/dam/acsorg/greenchemistry/industriainnovation/roundtable/process-mass-intensity-calculation-tool.xls.
14 (*a*) D. J. C. Constable, A. D. Curzons and V. L. Cunningham, *Green Chem.*, 2002, **4**, 521–527; (*b*) R. A. Sheldon, *Green Chem.*, 2007, **9**, 1273–1283; (*c*) A. P. Dicks and A. Hent, *Green Chemistry Metrics: A Guide to Determining and Evaluating Process Greenness*, 2015, Springer; (*d*) F. Roschangar, R. A. Sheldon and C. H. Senanayake, *Green Chem.*, 2015, **17**, 752–768; (*e*) F. Roschangar, J. Colberg, P. J. Dunn, F. Gallou, J. D. Hayler, S. G. Koenig, M. E. Kopach, D. K. Leahy, I. Mergelsberg, J. L. Tucker, R. A. Sheldon and C. H. Senanayake, *Green Chem.*, 2017, **19**, 281–285; (*f*) F. Roschangar, Y. Zhou, D. J. C. Constable, J. Colberg, D. P. Dickson, J. P. Dunn, M. D. Eastgate, F. Gallou, J. D. Hayler, S. G. Koenig, M. E. Kopach, D. K. Leahy, I. Mergelsberg, U. Scholz, A. G. Smith, M. Henry, J. Mulder, J. Brandenburg, J. R. Dehli, D. R. Fandrick, K. R. Fandrick, F. Gnad-Badouin, G. Zerban, K. Groll, P. T. Anastas, R. A. Sheldon and C. H. Senanayake, *Green Chem.*, 2018, **20**, 2206–2211; (*g*) F. Roschangar and J. Colberg, *Green Chemistry Metrics, in Green Techniques for Organic Synthesis and Medicinal Chemistry*, ed. W. Zhang and B. W. Cue, John Wiley & Sons, Chichester, UK, 2018; (*h*) R. A. Sheldon, *ACS Sustainable Chem. Eng.*, 2018, **6**, 32–48.
15 J. L. Tucker and M. M. Faul, *Nature*, 2016, **534**, 27–29.
16 D. Kaiser, J. Yang and G. Wuitschik, *Org. Process Res. Dev.*, 2018, **22**, 1222–1235.
17 (*a*) J. Li, E. M. Simmons and M. D. Eastgate, *Green Chem.*, 2017, **19**, 127–139; (*b*) J. Li, J. Albrecht, A. Borovika and M. D. Eastgate, *ACS Sustainable Chem. Eng.*, 2018, **6**, 1121–1132.
18 A. Borovika, J. Albrecht, J. Li, A. S. Wells, C. Briddell, B. R. Dillon, L. J. Diorazio, J. R. Gage, F. Gallou, S. G. Koenig, M. E. Kopach, D. K. Leahy, I. Martinez, M. Olbrich, J. L. Piper, F. Roschangar, E. S. Sherer and M. D. Eastgate, *The PMI Predictor – a Web App Enabling Green-by-Design Chemical Synthesis*, DOI: 10.26434/chemrxiv.7594646.v1.
19 P. Cornwall, L. J. Diorazio and N. Monks, *Bioorg. Med. Chem.*, 2018, **26**, 4336–4347.
20 J. D. Hayler, D. K. Leahy and E. M. Simmons, *Organometallics*, 2019, **38**, 36–46.
21 (*a*) M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604; (*b*) C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289; (*c*) S. Szymkuc, *et al.*, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904–5937; (*d*) J. Law, *et al.*, *J. Chem. Inf. Model.*, 2009, **49**, 593–602; (*e*) A. Bogevig, *et al.*, *Org. Process Res. Dev.*, 2015, **19**, 357–368.
22 J. Magano and J. R. Dunetz, *Chem. Rev.*, 2011, **111**, 2177–2250.
23 (*a*) Z. L. Niemeyer, A. Milo, D. P. Hickey and M. S. Sigman, *Nat. Chem.*, 2016, **8**, 610; (*b*) D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190; (*c*) For comment on this work please see K. V. Chuang and M. J. Kelser, *Science*, 2018, DOI: 10.1126/science.aat8603; (*d*) and response J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher and A. G. Doyle, *Science*, 2018, DOI: 10.1126/science.aat8763.

24 J. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
25 (*a*) N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 39–53; (*b*) J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725–732; (*c*) M. H. Segler and M. Waller, *Chem. – Eur. J.*, 2017, **23**, 5966–5971; (*d*) C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
26 P. Ruiz-Castillo and S. L. Buchwald, *Chem. Rev.*, 2016, **116**, 12564–12649.
27 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
28 S. Lin, S. Dikler, W. D. Blincoe, R. D. Ferguson, R. P. Sheridan, Z. Peng, D. V. Conway, K. Zawatzky, H. Wang, T. Cernak, I. W. Davies, D. A. DiRocco, H. Sheng, C. J. Welch and S. D. Dreher, *Science*, 2018, **361**(6402), DOI: 10.1126/science.aar6236.
29 S. Riniker and G. A. Landrum, *J. Cheminf.*, 2013, **5**, 43.
30 Q. Wu, X. Xiong, Y. Cao, L. He and Z. Fei, *Org. Process Res. Dev.*, 2018, **22**, 557–561.