On the Use of Real-World Datasets for Reaction Yield Prediction

Authors: Mandana Saebi,[‡1] Bozhao Nan,[‡2] John Herr,[2] Jessica Wahlers,[2] Zhichun Guo,[1] Andrzej M. Zurański,[3] Thierry Kogej,[4] Per-Ola Norrby,[5,6] Abigail G. Doyle,[3] Olaf Wiest*[2] and Nitesh V. Chawla*[1]

Affiliations:

[1] Department of Computer Science and Engineering and Lucy Family Institute for Data and Society, University of Notre Dame, Notre Dame, IN 46556, USA.

[2] Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN 46556, USA.

[3] Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States.

[4] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

[5] Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

[6] Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden.

‡ These authors contributed equally

*Correspondence to: owiest@nd.edu, nchawla@nd.edu

Abstract

The lack of publicly available, large, and unbiased datasets is a key bottleneck for the application of machine learning (ML) methods in synthetic chemistry. Data from electronic laboratory notebooks (ELNs) could provide less biased, large datasets, but no such datasets have been made publicly available. The first real-world dataset from the ELNs of a large pharmaceutical company is disclosed and its relationship to high-throughput experimentation (HTE) datasets is described. For chemical yield predictions, a key task in chemical synthesis, an attributed graph neural network (AGNN) performs as good or better than the best previous models on two HTE datasets for the Suzuki and Buchwald-Hartwig reactions. However, training of the AGNN on the ELN dataset does not lead to a predictive model. The implications of using ELN data for training ML-based models are discussed in the context of yield predictions.

Introduction

The development of predictive methods is a long-standing goal of computational chemistry. Initially, physics based modeling such as DFT or force field methods were used to understand reaction mechanisms and predict e.g. the stereochemical outcome of reactions[1] or suitable catalysts for their acceleration.[2] More recently, machine leaning (ML) methods[3] were very successful in predicting the likely product of reactions (forward synthesis prediction)[4,5] and promising pathways for the synthesis of organic molecules with a range of complexity.[6-9]

The prediction of yields of chemical reactions is a particularly challenging task because it is not only influenced by the variables of the reaction under study, but also from the influence of all possible side reactions. At the same time, it is an extremely important task due to the significant effort needed to optimize the yield of a reaction by variation of reaction conditions and catalysts. Doyle and coworkers[10] demonstrated that this challenge can be met for the case of the widely used

Buchwald-Hartwig amination by training a ML model on a dataset of 4608 reactions from high-throughput experimentation (HTE). Using a random forest model and computed physics-based features such as NMR shifts or HOMO/LUMO energies, an R^2 of 0.92 was achieved (Fig.1 A). More complex models such as neural networks did not provide higher predictivity.[10] Fu et al.[11] used a dataset of 387 Suzuki-Miyaura reactions[12] and features from DFT calculations to train a deep neural network, resulting in a model with an R^2 of 0.92. Bayesian optimizers[13] and deep reinforcement learning[14] were also successful in the iterative optimization of reaction conditions for a variety of reactions.

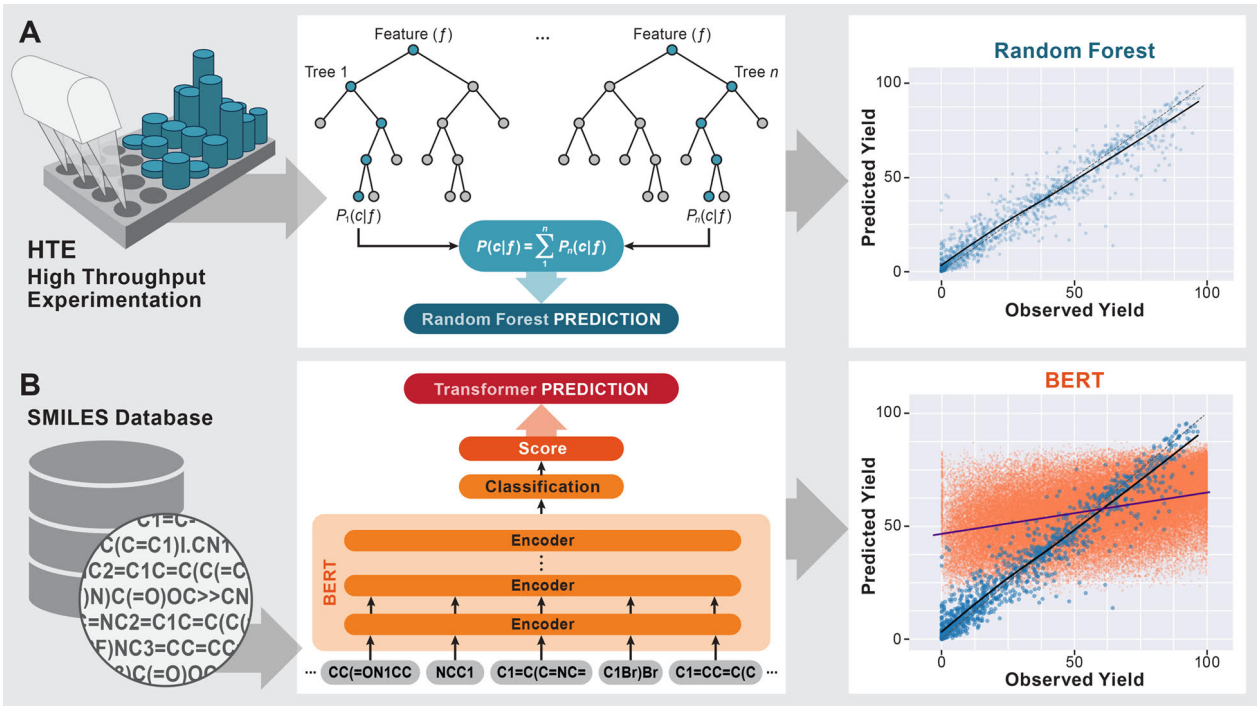

Figure 1: Previous work on yield predictions using ML models: (A) HTE-generated datasets using random forest models (B) HTE (blue) and USPTO derived (red) datasets using the BERT model

In contrast, the use of legacy datasets from published scientific or patent literature for yield prediction has not been successful. The attempt to classify reaction yields as above or below 65% based on a training set of ~10^6 reactions from the Reaxsys database using a large number of descriptors and ML methods gave an accuracy of 65±5%, i.e. a 35% error.[15] The authors of that study attributed this finding to the deficiencies of "currently available chemical descriptors", but it should also be noted that the reaction space represented in their dataset is vast. Schwaller et al.[16] developed a modification of the bidirectional encoder representations from transformers (BERT) model,[17] which uses natural language processing to build a reaction SMILES encoder trained on a large corpus of reactions, followed by a classification or regression layer for a specific task. This approach was very successful for product predictions[5] as well as for reaction yield predictions of the Suzuki-Miyaura (blue in Fig. 1B) and Buchwald-Hartwig reactions.[16] While this approach achieves R^2 values of 0.81 and 0.95, respectively, in line with other ML models when trained on these HTE datasets,[10,18] training on a dataset of Suzuki-Miyaura reactions from the PTSO[19,20] led to a maximum R^2 score of 0.388 (red in Fig. 1B). When the training set was limited to reactions run on a gram scale, the R^2 value dropped further to 0.277, which was attributed to the strong bias of this dataset towards high-yielding reactions.[16] When limiting the dataset to a single reaction, Reymond and coworkers[21] constructed a more qualitative "data-driven cheat-sheet" for the recommendation of conditions for the Buchwald-Hartwig reaction based on a dataset of 62,000 examples from a variety of databases.

Taken together, these previous findings highlight the challenges in using legacy datasets to train ML yield prediction models. As in other areas of ML, there is a lack of suitable datasets to train and validate the models. Although most of the chemical literature is summarized in commercial databases, they are proprietary. The US Patent database, which was converted into a

widely used dataset,[4] is an exception. As a result, studies using commercial databases do not include the data the models were built with.[21,22] Furthermore, databases such as Reaxsys frequently do not contain complete reaction information and reflect the bias of the published literature towards high-yielding reactions and inevitable human error, e.g. in assigning product structures.[23] Finally, the total chemical reaction space is enormous in comparison with even the biggest reaction databases, resulting in a sparse coverage.

As part of our ongoing efforts to explore the potential and limitations of ML methods in synthetic chemistry, we sought to investigate distinct approaches to investigate multiple novel approaches to the use of legacy datasets for reaction yield prediction. Here we introduce a novel dataset extracted from the electronic laboratory notebooks (ELN) of a large pharmaceutical company and an automated procedure for the curation of the dataset using a Jupyter notebook. It has long been hypothesized[8,24,25] that the use of ELNs to train ML models could unlock much larger datasets that are not subject to the publication bias towards high-yielding reactions. While this approach is pursued internally at a number of large organizations,[8] the underlying datasets are proprietary. To the best of our knowledge no such ELN-derived datasets have been made publicly available, and therefore the frequently made assumption that they can be used for training ML models has not been tested. To investigate whether the sparsity, noise, and inherent bias of legacy datasets can be addressed by advanced ML models we developed an attributed graph neural network model and tested it on both HTE and ELN-derived datasets. Finally, we discuss the implications of the findings for the use of legacy data in the prediction of chemical yields.

Results

Model design We hypothesize that the small size and sparsity of typical chemical datasets can be balanced by including the maximum amount of information about the chemical structures involved in the reaction. We propose to combine physically meaningful molecular properties, i.e. chemical features/descriptors, with features capturing the molecular graph structure in an attributed graph neural network (GNN). GNNs have been shown to successfully capture the higher-order interactions between neighboring components of a graph.[26] An overview of the model named YieldGNN is shown in Figure 2. The top module represents the AGNN that learns the structural features while the bottom module captures the features describing the chemical properties.

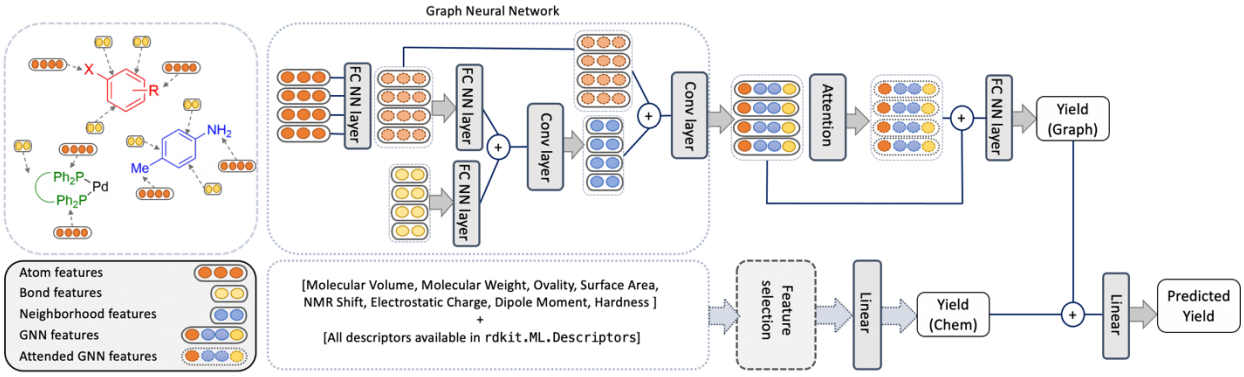

Figure 2: Overview of the YieldGNN model where the structural features are captured by aggregating atom and bond features over the neighborhood (top part) and are combined with the chemical features (lower part) to generate two yield scores Yield (Graph) and Yield (Chem). The two scores are passed through a linear layer to generate the final predicted yield .

For the top module, we use Weisfeiler-Lehman Networks (WLN)[27] to capture the structural features. WLNs are one of the most expressive GNNs studied so far.[28] WLNs learn the structural features by iteratively aggregating features (using convolutional operations) over local node

neighborhoods. This allows WLNs to capture the higher-order neighborhood information in the graph structure.

The bottom module in Figure 2 includes atomic features such as partial charges and NMR shifts as well as molecular features such as orbital energies, electronegativity, dipole moments, molecular volume, and surface area; each of which can be easily generated computationally through programs such as RDKit or Gaussian16. A full list of the features used is provided in the Supporting Information. To minimize the risk of overfitting in the AGNN owing to the large number of chemical features, we trained a random forest (RF) model to select the main chemical features that contribute to the RF model performance. This model serves as a baseline and also helps us reduce the number of the parameters used in our deep learning model. Note that we do not perform feature engineering on the structural features and they are automatically generated by the GNN model.

Training data description The AGNN was trained on three different datasets. For comparison purposes, we used two HTE datasets designed for the Suzuki-Miyaura cross-coupling[18] and the Buchwald-Hartwig amination[10] reactions; both datasets have previously been modelled with ML to make yield predictions.[11,16] As a representative example of a real-world dataset from the pharmaceutical industry, we collected a legacy dataset from electronic laboratory notebooks (ELN) at AstraZeneca. For this purpose, the NextMove software used at AstraZeneca was queried with the term “Buchwald-Hartwig”. The datasets thus obtained were filtered to only include publicly available products, and entries that were recorded prior to August 2016. This resulted in a raw dataset of 1000 entries subsequently saved in UDM format to include the structures of reactants, products, catalysts and bases as well as reaction conditions (e.g., solvents, reaction

temperatures and times) as well as yields. Where available, additional comments from the ELN were also included.

As shown in Fig. 3, the HTE datasets are similar to each other in that they have a dense coverage of a narrow area of the chemical space. If all combinations of variables for the Suzuki-Miyaura are considered 7392 combinations are possible,[29] though the two-stage design of the study decreases this number to 4608. For the Buchwald-Hartwig reaction,[10] a full fractional design was explored, leading to 3960 possible combinations. Both HTE datasets have a broad and relatively uniform yield distribution. The dataset extracted from the AstraZeneca ELNs has, as is typical for ELNs and other legacy datasets, very different characteristics. It covers a much wider chemical space, with 340 aryl halides, 260 amines, 24 ligands, 15 bases and 15 solvents. With 1000 examples to cover ~4.7 x 10^8 possible combinations of reactants, ligands, bases and solvents, the dataset is much sparser. As a result, there are essentially no overlapping conditions for a given substrate combination. In addition, 39.9% of reactions in the ELN dataset did not yield a product for a variety of reasons (see Methods section). Overall, this dataset is much more representative of real-world datasets and the problems associated with them than the datasets from exhaustive HTE. It is therefore an important benchmark if AGNNs cannot only improve the predictive power of ML models for the designed dataset, but provide predictive models based on real-world datasets such as the one derived from the ELNs.

The raw ELNs (in xlm format) were processed to generate a data table suitable for data cleaning. Using a Jupyter notebook, the dataset was converted into a form suitable for ML applications. Molecules were classified as reactants and reagents based on the reaction SMILES strings. As is common in most databases, some of the reaction conditions (e.g., temperature) or reaction components were not listed or had inconsistent structures which required manual curation

for a small subset of reactions, e.g., by correcting based on the product structure. Duplicate and empty entries were removed, reaction conditions were standardized and molecular structures were saved as SMILES files.

As shown in Fig. 3B, a yield of 0% or incomplete reactions were reported for a significant number of entries due to a number of reasons (human error, trial run without yield determination etc.) that were annotated in the comment line of the dataset. These low- or no-yield reactions were classified using an ontology of the reaction description fields using a Jupyter notebook to minimize the need for manual curation and, where possible, adjusted based on duplicate entries. This processing of the ELN entries led to a final dataset of 781 reactions that, in contrast to previous applications of ELN datasets in ML,[8] are made publically available (see Data Availability Statement). Therefore, the ELN dataset for the Buchwald-Hartwig reaction is, to the best of our knowledge, the first publicly available ELN reaction dataset for use in ML applications. Chemical space analysis using Multidimensional Scaling (MDS) as described by Schneider and coworkers.[30] Morgan substructure fingerprints (radius 0-4 bonds, 1024 bit length) were calculated in RDKit and the canonical MDS was calculated using Tanimoto similarity metric. This MDS analysis of the products of the Buchwald-Hartwig reaction (Fig. 3C) shows that the structural diversity of the ELN dataset (shown in gold) is much higher than for the HTE dataset[10].

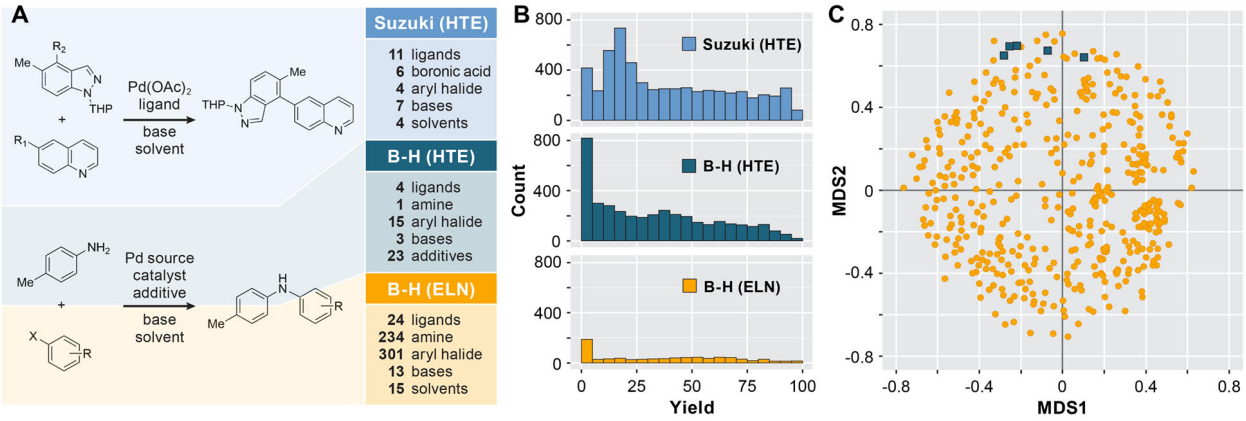

Figure 3 A) Overall reaction and variables for Suzuki-Miyaura (top) and Buchwald-Hartwig (B-H, middle and bottom) datasets. B) yield distributions (middle). C) Chemical space analysis (MDS) of products for HTE (blue) and ELN (gold) datasets (right).

Yield prediction Ten different train/test random 70:30 splits for each dataset were created and used to train and test our model on each respective set. For each dataset, the mean and standard deviation (over the ten test sets) of the R^2 and mean absolute errors (MAE) are reported in Table 1. We tested the YieldGNN without any chemical features (i.e., only the top part of the model shown in Fig. 2, YieldGNN w/o chem. Feat.), followed by tests that included either only the chemical features from the G16 calculations (YieldGNN w/o rdkit feat.) or the complete set of features, including the ones provided by RDKit (YieldGNN with rdkit feat.). Tests of 40 random splits for the case of the Buchwald-Hartwig HTE dataset with the full feature set in YieldGNN did not yield significantly different statistics (See Table S4 in Supporting Information). Further improvements to the YieldGNN were possible by adding the attention layer after the AGNN component, explicit inclusion of solvent and base, and addition of the chemical features into the model. The resulting model that uses the full information as shown in Figure 2 performs as well or better than previously available models, such as random forest or BERT, for the two HTE

datasets studied here. RF–1 and RF–2 are both random forest models with the same hyper-parameters. The only difference between these models is that RF–1 contains all available features, while RF–2 does not contain features derived from RDKit. Thus, the features used in the RF–2 model are very similar to the ones used in a previous study.[10] Below we explain the model performance on HTE datasets and the ELN dataset from AstraZeneca.

Table 1: Results for three reaction datasets. For each data, the mean and standard deviation of R^2 and MAE (in parenthesis) obtained via training each model on 10 random data splits.

	Suzuki-Miyaura [HTE][18]	Buchwald-Hartwig [HTE][10]	Buchwald-Hartwig [ELN]
RF– 1 (with rdkit feat.)	0.824±0.005 (0.083±0.001)	0.899+0.007 (0.059±0.001)	0.259+0.0429 (0.202±0.006)
RF –2 (w/o rdkit feat.)	0.792+0.012 (0.09±0.002)	0.906+0.002 (0.057±0.001)	**0.2619+0.038** (0.204±0.007)
BERT[16]	0.81±0.01	0.951±0.005	---
YieldGNN w/o chem. feat.	0.82+0.028 (0.083±0.001)	0.939+0.025 (0.047±0.008)	0.118+0.05 (0.225±0.014)
YieldGNN w/o rdkit feat.	**0.842±0.013** (0.079±0.003)	0.9±0.031 (0.061±0.01)	- 0.023±0.099 (0.244±0.01)
YieldGNN with rdkit feat.	0.836±0.013 (0.08±0.004)	**0.958** ± **0.003** (0.04±0.001)	--

Performance on HTE datasets. The YieldGNN significantly outperforms the random forest models for the two HTE datasets as indicated by the higher R^2 and lower MAE with the difference being larger in the case of the Buchwald-Hartwig HTE dataset than for the case of the Suzuki-Miyaura reaction. Interestingly, the performance of the BERT and YieldGNN models are within

the standard deviation of each other. Taken together, these results suggest that models that use connectivity data, which in case of the BERT model is encoded in the SMILES files, perform better than the random forest models that are based on chemical features alone. This is in line with the observations that during the training of the YieldGNN model, the weight of the graph features increases and the weight of the chemical features decreases as a function of the training epochs (see Fig.S1 in the Supporting Information). This suggests that the molecular structure provides key information in model training and thus improves the prediction of reaction yield. Although in previous studies, the neural network model performed slightly worse than the random forest model for the Buchwald-Hartwig HTE dataset,[10] the combination of chemical features and structural information shows excellent performance for the focused datasets derived from HTE. This is further supported by "leave-one-group-out" analysis for the Buchwald-Hartwig HTE dataset[10,31] (see Table S5 in the Supporting Information) that shows a modest degradation in the performance as each of the additives is left out of the training set and the YieldGNN is retrained with the remaining 23 additives.

Performance on ELN dataset. Having shown that YieldGNN provides highly predictive models for HTE datasets, we tested whether this information rich, combined approach can treat the more diverse legacy data. The results shown in Table 1 demonstrate that this is not the case and the YieldGNN does not provide meaningful predictions of the yield. Extensive tuning of the hyperparameters of the network or pre-training the model on the HTE dataset for the same Buchwald-Hartwig reaction, followed by fine-tuning the trained model on the target dataset did not improve the performance and led to R^2 values that were negative or close to zero. For this dataset, the random forest models provide better R^2 values, nevertheless these are still too low to provide useful predictions.

Expansion of training data We hypothesize that the reason for the failure of the YieldGNN model to predict the yields for the legacy dataset after being highly successful for the HTE datasets is the much larger diversity and sparsity of the ELN dataset (see Fig. 3C). It should be noted that while the size of the legacy dataset is relatively small compared to the typical data sizes used in deep learning models, it is approximately twice the size of the dataset successfully used by Fu et al.[11] for a yield model of the Suzuki reaction. To improve the model generalizability by increasing the individual dataset sizes, thus decreasing the sparsity of the dataset, a GNN model was pre-trained using the method developed by Hu et al.[32] using attribute masking, context prediction and edge prediction. The resulting model was then fine-tuned separately for the yield prediction task on each of the three datasets. Note that the goal of the pre-training stage is to learn from existing patterns in the data independent of the downstream task. Thus, labels are not necessary at this stage. Two different datasets were used for the pre-training stage. The first dataset contains 2 million molecules sampled from the ZINC15 dataset[33] used previously.[32] For Suzuki-Miyaura reactions, a second dataset contains synthetic Suzuki reactions generated by permutating all commercial available reactants and ligands and generating all possible combinations. This resulted in 440K potential Suzuki reactions that can be used to pre-train the model on a dataset that is more closely related to the target data.

However, none of the above methods resulted in significant improvement on the yield prediction task. The results are shown in Table 2. Note that the GNN model used here is based on the model developed by Hu et al.[32], as a result the R^2 scores after fine-tuning are not similar to our model results. Although we notice a slight improvement in Buchwald-Hartwig reactions from AstraZeneca, the R^2 score of this model is still lower than that of RF–1 and RF–2 baselines. We

conclude that the best result is obtained by training separate models on each dataset. We leave the exploration of other methods to improve model generalizability for future work.

Table 2: Results for three reaction datasets. For each data, the mean and standard deviation of R^2 and MAE (in parenthesis) obtained via training each model on 10 random data splits. For Suzuki-Miyaura data, a second column is added which contains the results of the model pre-trained on our synthetic Suzuki-Miyaura data.

	Suzuki-Miyaura[18]	Suzuki-Miyaura[18] (Pretrain-Synthetic)	Buchwald-Hartwig[10]	Buchwald-Hartwig [ELN]
ContextPred	0.540±6e-4 (0.152±0.0004)	**0.546±3e-4 (0.151±1e-4)**	0.716±6e-4 (0.103±4e-4)	**0.177±0.014 (0.220±0.002)**
EdgePred	**0.540± 6e-4 (0.152±3e-4)**	0.544±3e-4 (0.152±1e-4)	**0.721±0.001 (0.102±1**e-4**)**	0.129±0.011 (0.231±0.002)
AttrMasking	0.535±5e-4 (0.152±0.0004)	0.545±4e-4 (0.152±1e-4)	0.713± 0.001 (0.102±0.004)	0.143±0.008 (0.222±0.002)

Discussion

The key limitation for the application of ML methods in synthesis is the availability of suitable datasets. This is particularly evident in yield predictions which have the potential to greatly accelerate reaction optimization and development but have so far only been demonstrated for specific reactions sets where focused HTE datasets were generated for this purpose. This data challenge is widely acknowledged in the literature and the mining of ELN has been suggested as a possible solution because ELNs are perceived to be less biased towards high-yielding reactions and more information-rich than the primary literature or literature databases.[8,34,35] Although potential problems in the extraction of data from ELNs have been acknowledged,[36] the suggestion was that appropriate tools could overcome the challenges in using ELN data for a variety of applications including yield predictions.[8]

The results presented here, together with the studies in the literature that did not focus on specific reactions,[15,16] suggest that is not the case and the legacy datasets from commercial databases or ELNs, by themselves, might be of limited use for the prediction of yields. The findings that successful models of yield predictions could be built with datasets smaller but more focused than the ELN dataset used here,[11] together with the excellent performance of the YieldGNN for the HTE datasets, suggests that the failure to provide a predictive model owes to the diversity of the chemical space of the training set, visualized in Fig 3C. The suitable balance between dataset diversity and size for yield predictions is not known yet. It should also be mentioned that collection of ELN over an extended period of time by different experimentalists introduces another level of noise that will be hard to control. This has significant consequences for the proposed use of legacy datasets from the literature or ELNs [8,34-36] that will require a more detailed analysis before successful yield prediction models can be build. Developing workflows to curate the ELN-derived datasets and making them available to the scientific community are important steps in this direction.

Methods

Generation and Curation of ELN dataset. The raw dataset was collected from the electronic laboratory notebooks at AstraZeneca using the NextMove software. To curate the raw data, a series of Jupyter notebooks were created, which can be found in the github repository. First, the original data format (.xml) was converted to the internally used library files. The scripts include several steps of data processing for automated curation of the dataset. Examples demonstrating the data format workflows for the generation of the features from the structures are described below. Next, the yield-related information was generated including reactants information, reaction variables

(e.g. temperature, volume, and reaction scale). In many cases, the information was contained in the comment section of the .xml file rather than the appropriate data field. In these cases, the information was transferred (either through scripting or manual curation) to the correct field based on the preparation section which is shown as text form in the original dataset. Cases where no yield was reported were classified to four types of non-yield reactions: (A) no reaction occurred (104 of 173), (B) trace amount of product (41 of 173) and (C) complex mixture of reaction products (28 of 173). Finally, MDL molfiles were generated for each molecule from the compounds database included in the ELN, which were then used to generate SMILES strings. The SMILES files were converted into Cartesian coordinates for the Gaussian calculations using RDKit[37] and OpenBabel.[38]

Feature generation and selection. For each molecule, two sets of chemical features were obtained. The first source is the full set of descriptors available in the RDKit library. The second source are the features from DFT calculations of each of the reactants using Gaussian16[39] with the B3LYP functional and 6-31G* basis set for geometry optimization and 6-311G* basis sets for single point calculations. The remaining features include the surface area generate from pymol, pKa of the base, solvent dielectric constant from the compound database. The following set shows the chemical features used for model training:

Molecular features: molecular volume, surface area, ovality, molecular weight, HOMO/LUMO Energy, electronegativity, hardness, and dipole moment.

Atomic features: Electrostatic charge and NMR shift

Reaction features: Temperature, Reaction scale and volume for some of reactions

To pre-select features, the features from above sources were combined, and a random forest model was trained on ten 70:30 random splits. Then, all features with feature importance of 10^{-4} or greater in any of the 10 random forest models were retained and included in the AGNN. Note that no feature engineering on the structural features is performed, the structural features are automatically generated by the GNN model. The random forest-based pre-selection helps reduce the number of the parameters used in the deep learning YieldGNN model.

Model architecture. The model integrates both the chemical features and the structural features for reaction molecules using two main components. An overview of the model is shown in Figure 2. The top component represents the AGNN which learns the structural features and the bottom module captures the chemical features. In this section the process of structural feature generation is detailed.

For each reaction, attributed graphs containing atom and bond features for each molecule are build first. Each atom feature contains atomic number, formal charge, degree of connectivity, explicit and implicit valence, and aromaticity. The bond features include the bond type, bond order and ring status. Atom features around the atom neighborhood are aggregated in an iterative manner using the Weisfeiler-Lehman Network (WLN)[27] to obtain the local atom and bond features. WLN is a graph kernel based on the Weisfeiler-Lehman test for graph isomorphism. Two graphs are isomorphic if they are topologically equivalent, and the WL test is a necessary condition for graph isomorphism. Thus, the WLN is one of the most expressive GNN methods and is used here. In each iteration, the atom feature representation is updated according to:

$$h_v^l = Relu\left(U_1 h_v^{l-1} + U_2 \sum_{u \in N(v)} Relu\left(V_1 h_v^{l-1} + V_2 h_{uv}\right)\right)$$

where h_v^l is the atom feature representation at iteration l ($1 \leq l \leq L$). U_1, U_2, V_1, V_2 are paramters to be learned, which are shared across L iterations. The final atom feature representation for atom v is obtained at the end of the final iteration using:

$$h_v = \sum_{u \in N(v)} (\theta_1 h_u^L \odot \theta_2 h_{uv}^L \odot \theta_1 h_v^L)$$

where $\odot$ is the convolution operation and θ are the model weights. For the HTE datasets, two iterations are used to capture the 2-hop neighborhoods. Therefore, for these datasets the above operation translates to two iterations to obtain the local representation of atoms.

Next, the local structural features are fed to an attention layer to capture the global structural features. The intuition behind including attention[40] is that different components of the reaction may influence the reaction yield differently. The attention layer is meant to capture the degree to which different atoms influence each other. The global representation of atom v is equivalent to the weighted sum of all other atoms in the reaction:

$$\tilde{h}_v = \sum_z \alpha_{vz} h_v$$

The attention score for a given atom pair (v, z) is calculated using:

$$\alpha_{vz} = \sigma(u^T \times Relu\ (W_1 h_v + W_2 h_z + W_3 b_{vz}))$$

where $\sigma(.)$ is the sigmoid function, b_{vz} is the binary features for atom pair (v, z) and W is the attention weights to be learned by the model. Both global and local structural features are concatenated to generate the final structural features. The YieldGNN model provides two yield scores, one from the structural features (Yield(graph)) and the other from the chemical features (Yield(chem)). The two scores are fed to a linear layer to generate the final reaction yield predictions in analogy to earlier work by Coley et al.,[41] but for prediction of the reaction yield through combining both structural graph-based features as well as chemical properties.

Parameter Selection. A grid-search for each hyper parameter is performed and tuned for each dataset separately. For all datasets, batch size, dropout ratio, and initial learning rate are set to 40, 0.04 and 0.01, respectively. A learning rate decay ratio of 0.5 is used on all datasets if the loss plateaus. A 2-hop neighborhood is used for the HTE datasets and a 3-hop neighborhood for the ELN data. The size of all hidden layers is set to 100 for the ELN data and 200 for the HTE data. The gradient is clipped with a 0.8 ratio on all datasets to avoid the exploding gradient problem. The model is trained for 200 epochs for HTE data and 100 epoch for the ELN data using Adam optimizers[42] with β_1=0.9 and β_2=0.99.

For pre-training, the models developed by Hu et al.[32] we use Graph Isomorphism Network (GIN), which based on the author's finding resulted in the best performance. Following the best parameters suggested by the authors in their original work, we set the batch size, number of layers, and dropout ratio to 256, 5, and 0.15, respectively. We use an embedding dimension of 300 and a learning rate of 0.001, and pre-train the models for 100 epochs. For the fine-tuning module, we set batch size, number of layers, embedding dimensions, and dropout ratio to 32, 5, 0.5, and 300, respectively based on the author's recommended parameters. We set the learning rate to 0.001 and decay it with a 0.9 rate upon loss plateau. We use mean pooling for GNN during both pre-training and fine-tuning. We fine-tune the pre-trained model for 100 epochs as well.

Model Evaluation

Baselines: The model is compared to three main baselines. The first two baselines are random forest (RF) models with 1000 trees each having a maximum depth of 10. Two RFs are trained with different feature subsets. The first model, RF-1, is similar to the previous random forest model[10] and contains the same feature set in that study. The second model, RF-2 includes all features available in RF-1 as well as all descriptors available in `rdkit.ML.Descriptors`. The top

features selected in RF-2 are used in the feature selection step later on. The third baseline is the BERT language model[5,16] that treats reaction smiles as text and fine-tunes the pre-trained BERT language model to predict the reaction yield. For our experimental results, we directly quote the performance of this model[16] on the HTE data.

Evaluation metrics. The performance of each model using coefficient of determination, denoted as R^2, and the mean absolute error, denoted as MAE. 10 models with different random splits of each dataset are run and the mean and standard deviation of the 10 experiments is reported.

Code availability. All models, scripts, Jupyter notebooks and data curation workflows are available free of charge in the Supporting Information and at https://github.com/nsf-c-cas

Data availability The raw ELN dataset derived from the ELNs at AstraZeneca as well as the curated version with associated features used are available free of charge at https://github.com/nsf-c-cas and have been uploaded to the Open Reaction Database https://docs.open-reaction-database.org.

Acknowledgements This work was supported financially by NSF through the Center for Computer Assisted Synthesis C-CAS (CHE-1925607) and AstraZeneca.

Author contributions. M.S., J.H. and A.M.Z. build and refined the models, T.K. and P.-O. N. collected the AstraZeneca ELN dataset, M.S., B.N. and J.W. curated the datasets and generated the features, A.G.D., O.W. and N.V.C. supervised the research. All authors contributed to data analysis and writing of the manuscript.

Competing interests The authors declare no competing interests.

References

1 Rosales, A. R. *et al.* Application of Q2MM to predictions in stereoselective synthesis. *Chem. Comm.* **54**, 8294-8311 (2018).

2 Poree, C. & Schoenebeck, F. A holy grail in chemistry: Computational catalyst design: Feasible or fiction? *Acc. Chem. Res.* **50**, 605-608 (2017).

3 Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **1**, 1-23 (2021).

4 Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281-1289 (2018).

5 Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091-6098 (2018).

6 Molga, K., Szymkuć, S. & Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **54**, 1094-1106 (2021).

7 Bøgevig, A. *et al.* Route design in the 21st century: The IC SYNTH software tool as an idea generator for synthesis prediction. *Org. Proc. Res. Dev.* **19**, 357-368 (2015).

8 Yang, Q. *et al.* Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Comm.* **55**, 12152-12155 (2019).

9 Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Central Science* **3**, 1237-1245 (2017).

10 Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186-190 (2018).

11 Fu, Z. *et al.* Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki–Miyaura cross-coupling reaction. *Org. Chem. Front.* **7**, 2269-2277 (2020).

12 Reizman, B. J., Wang, Y.-M., Buchwald, S. L. & Jensen, K. F. Suzuki–Miyaura cross-coupling optimization enabled by automated feedback. *React. Chem. Engin.* **1**, 658-666 (2016).

13 Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89-96 (2021).

14 Zhou, Z., Li, X. & Zare, R. N. Optimizing chemical reactions with deep reinforcement learning. *ACS central science* **3**, 1337-1344 (2017).

15 Skoraczyński, G. *et al.* Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Scientific reports* **7**, 1-9 (2017).

16 Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* **2**, 015016 (2021).

17 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

18 Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429-434 (2018).

19 Lowe, D. M. *Extraction of chemical structures and reactions from the literature*, University of Cambridge, (2012).

20 Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature Comm* **11**, 1-8 (2020).

21 Fitzner, M. *et al.* What can reaction databases teach us about Buchwald–Hartwig cross-couplings? *Chem. Sci.* **11**, 13085-13093 (2020).

22 Gao, H. *et al.* Using machine learning to predict suitable conditions for organic reactions. *ACS Cent, Sci.* **4**, 1465-1476 (2018).

23 Rosales, A. R. *et al.* Transition state force field for the asymmetric redox-relay Heck reaction. *J. Am. Chem. Soc.* **142**, 9700-9707 (2020).

24 Christ, C. D., Zentgraf, M. & Kriegl, J. M. Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *Journal of chemical information and modeling* **52**, 1745-1756 (2012).

25 Ghiandoni, G. M. *et al.* Development and application of a data-driven reaction classification model: comparison of an electronic lab notebook and medicinal chemistry literature. *Journal of chemical information and modeling* **59**, 4167-4187 (2019).

26 Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. Gnn explainer: A tool for post-hoc explanation of graph neural networks. *arXiv preprint arXiv:1903.03894* (2019).

27 Lei, T., Jin, W., Barzilay, R. & Jaakkola, T. in *International Conference on Machine Learning.* 2024-2033 (PMLR).

28 Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

29 Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429-434 (2018).

30 Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).

31 Żurański, A. M., Martinez Alvarado, J. I., Shields, B. J. & Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **54**, 1856-1865 (2021).

32 Hu, W. *et al.* Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).

33 Sterling, T. & Irwin, J. J. ZINC 15–ligand discovery for everyone. *J. Chem. Inf. Mod.* **55**, 2324-2337 (2015).

34 Moosavi, S. M., Jablonka, K. M. & Smit, B. The role of machine learning in the understanding and design of materials. *J. Am. Chem. Soc.* **142**, 20273-20287 (2020).

35 Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **19**, 353-364 (2020).

36 Engkvist, O. *et al.* Computational prediction of chemical reactions: current status and outlook. *Drug Disc. Today* **23**, 1203-1218 (2018).

37 Landrum, G. (Academic Press, 2013).

38 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminf.* **3**, 1-14 (2011).

39 Gaussian 16 Rev. B.01 (Wallingford, CT, 2016).

40 Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

41 Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370-377 (2019).

42 Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).