Predictive Minisci Late Stage Functionalization with Transfer Learning

Emma King-Smith,[1] Felix A. Faber,[1] Usa Reilly,[2] Anton V. Sinitskiy,[3] Qingyi Yang,[4] Bo Liu,[5] Dennis Hyek,[5] and Alpha A. Lee*[1]

[1]Cavendish Laboratory, University of Cambridge, Cambridge, UK

[2]Development & Medical, Pfizer Worldwide Research, Groton, CT, USA

[3]Machine Learning Computational Sciences, Pfizer Worldwide Research, Cambridge, MA, USA

[4]Development & Medical, Pfizer Worldwide Research, Cambridge, MA, USA

[5]Spectrix Analytic Services, LLC., North Haven, CT, USA

Abstract

Structural diversification of lead molecules is a key component of drug discovery to explore chemical space. Late stage functionalizations (LSFs) are versatile methodologies capable of installing functional handles on richly decorated intermediates to deliver numerous diverse products in a single reaction. Predicting the regioselectivity of LSF is still an open challenge in the field. Numerous efforts from chemoinformatics and machine learning (ML) groups have made significant strides in this area. However, it is arduous to isolate and characterize the multitude of LSF products generated, limiting available data and hindering pure ML approaches. We report the development of an approach that combines a message passing neural network and ^{13}C NMR-based transfer learning to predict the atom-wise probabilities of functionalization for Minisci and P450-based functionalizations. We validated our model both retrospectively and with a series of prospective experiments, showing that it accurately predicts the outcomes of Minisci-type and P450 transformations and outperforms the well-established Fukui-based reactivity indices and other machine learning reactivity-based algorithms.

Introduction

Late-stage functionalization (LSF) is a powerful technique in medicinal chemistry. The "magic methyl effect" describes the ability of a single methyl group, even one distal to the binding motif, to dramatically improve (or reduce) potency, solubility, and metabolic stability.[1] However, methyl groups are not the only motif that can radically change pharmacological properties. Fluoro,[2] chloro,[3] trifluoromethyl,[4] and hydroxyl groups[5] are known beneficial motifs and/or temporary functional handles towards other beneficial motifs. Over the past several decades, numerous methods have been developed to diversify lead compounds and selectively install these biologically privileged groups directly.[6] One methodology commonly utilized in LSF is the Minisci-type functionalization, whereby a radical species adds to an electron deficient (hetero)arene (Figure 1A).[7] However, the promiscuity of this single electron method in conjunction with the inherent structural complexity of LSF molecules make

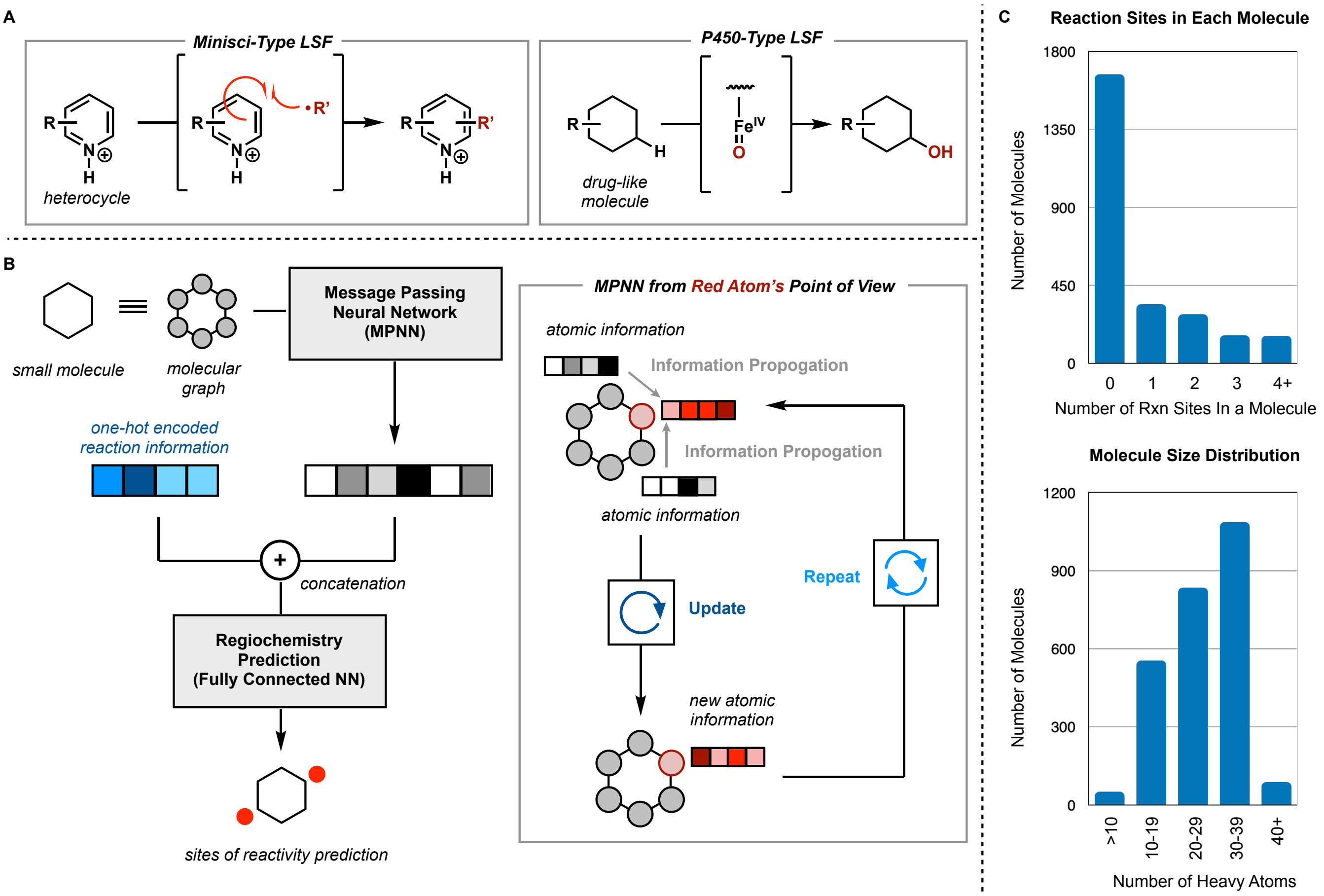


Figure 1: A) Mechanistic differences between the one-electron based transformations of the two major types of reactions in the dataset: Minisci and P450. **B**) Graphical overview of the basic MPNN model. **C**) Distribution of reaction sites per molecule and molecule size in the dataset.

regioselectivity prediction challenging. Regiochemical predictions for Minisci-type reactions were first summarized by O'Hara *et al.* who developed a set of guidelines to determine sites of reactivity based upon the nucleophilicity of the alkyl radical species, pH of the reaction, solvent effects, and electronics of the heteroarene.[8] These observations were later formalized when they were noted to correlate well with the indices from Fukui functions, *i.e.* functions that describe the change in electron density upon the addition or removal on an electron. In the literature, Fukui-based reactivity indices predict the most reactive sites of Minisci functionalization with an average accuracy of 93% (average F-score of 0.77), albeit usually on smaller, minimally functionalized molecules.[9] We hypothesized that an ML model, with its high parameterization, would offer an improvement in accuracy when predicting the regiochemical outcomes of more complex molecules (Figure 1B). The resulting consistent and broadly applicable LSF predictive framework would facilitate more rapid and facile access to a diverse array of drug-like compounds, specifically with respect to structure-activity relationship (SAR)-probing synthesis and expanding the known chemical space available for exploration.

There are two main approaches in the literature for regiochemical predictions: quantum chemical and data-driven. Quantum chemistry-based approaches predict reactivity and regioselectivity by computing energy barriers using techniques such as DFT or machine learning (ML) approximations of DFT-energies.[10] Data-driven approaches to work directly with experimental data, fitting statistical models to correlate known chemical features to real-world observed outcomes in regioselectivity.[11] Whilst computational data is more plentiful and significantly less noisy than real-world data, notable performance can be achieved with carefully curated literature datasets. Some experimentally-based reactivity models can reach human expert performance in their predictions and can, on occasion, surpass them.[11a] However, ML-based regiochemical prediction is still difficult. Due to the challenges of rigorously characterizing the regiochemical outcomes of thousands of reactions, experimental data-based models must often operate in lower data environments, and if gathered from the literature, often with data that contains few negative datapoints, *i.e.* molecules that don't react. In contrast, datasets which include easily extractable yield information often contain ten-fold more data.[12] This makes it more difficult for ML to find relationships between the molecular structure and LSF outcomes. Herein, we report a solution to this problem: the utilization of open-source ^{13}C NMR data in conjunction with LSF data. Our model is a graph-based model which does not require pre-computed molecular properties nor any 3D molecular information for accurate regioselectivity prediction. As a proof of concept, we highlight our framework's predictive ability on both Baran and Molander-type Minisci and P450 LSFs, transformations whose substrate scope is well defined. We show that our model outperforms the Fukui function-based index predictions, and two highly accurate, previously reported, reactivity-based machine learning models: one 1-electron based enzymatic reactivity model and one 2-electron based small-molecule model.

Results and Discussion

The Dataset

Data was sourced from Pfizer's internal medicinal chemistry dataset which consisted of ~2,600 reactions, 647 unique molecules, and 823 unique LSF conditions. The majority of these reaction conditions were Minisci-type functionalizations (1928 reactions), including Minisci reactions utilizing the Baran Diversinates™ (463 reactions).[13] Classic Minisci conditions were included in the training set, however, the majority of the training data consisted of Baran and Molander Minsci reactions (Table S1). Additionally, other single electron based late stage functionalizations were included in the training data such as P450 catalyzed oxidations (642 reactions), electrochemical methylations (12 reactions), and photoredox alkylations (93 reactions) (See Table S2 for a further breakdown of the dataset). Reactions that yielded oxidative cleavage or hydrolyzed side products were kept. A key facet to our dataset was the inclusion of data which contained unsuccessful conditions that led to no significant product formation (zero reactive sites). Despite the significant mechanistic differences between these reaction classes, we hypothesized that additional chemical information relating to the inherent reactivity of both the reagent and the molecule would be advantageous to regiochemical outcome prediction (Figure S1). A mixture of reaction classes has seen use in other reactivity-based predictions with excellent results.[11a, 11g] To implicitly distinguish between the reaction types, each unique reagent, oxidant, solvent, additive, and acid was one-hot encoded to form a specific "reaction vector", unique for each unique reaction condition. Similar to an organic chemist, the selectivity neural network (Figure 1B) would need to interpret the mechanism type from the collection of reagents.

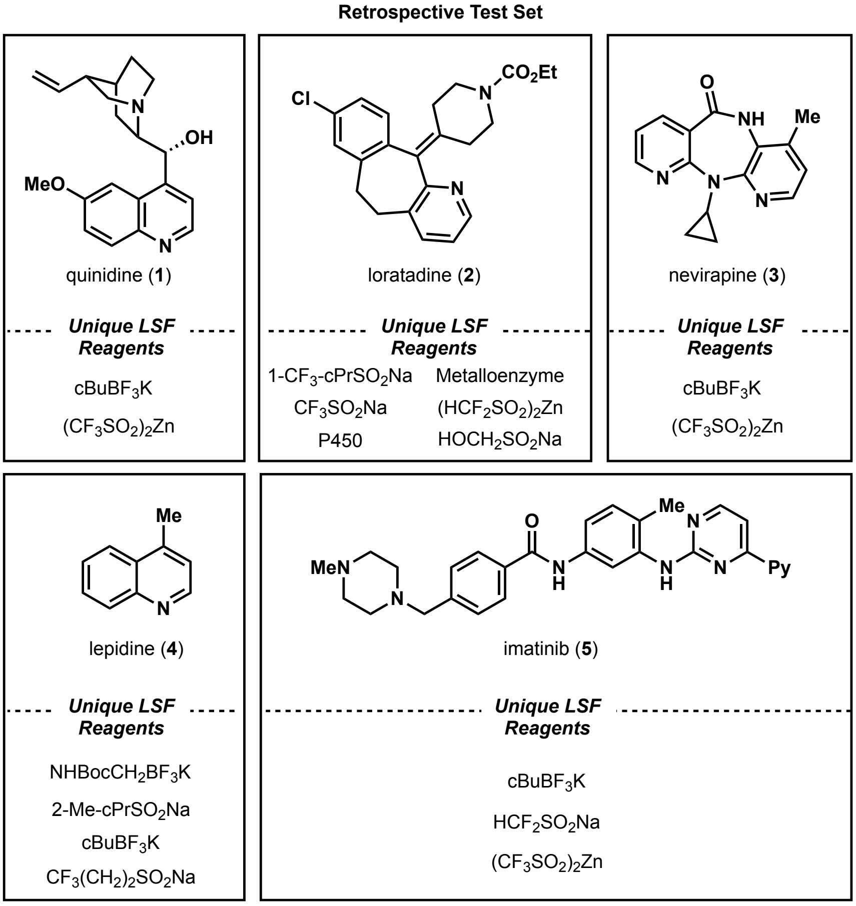


Figure 2: The retrospective test set used for optimization of the models.

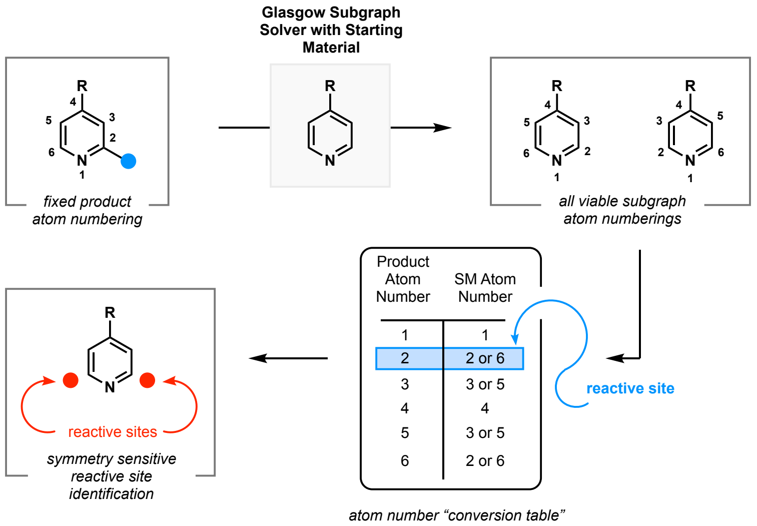
When deciding the correct method to split the data into training and testing sets, we opted for scaffold-based instead of a random split. It has been hypothesized that a random split encourages the model to simply memorize the inherent reactivity of a molecule, instead of applying its learned chemical knowledge to new scaffolds.[14] A scaffold split, where every molecule in the test set is an unseen molecule, provides a more challenging target. The retrospective test set consisted of 25 reactions which was comprised of 5 unique molecules and 17 unique reaction conditions. Of the reaction conditions, 22 were Minisci-type functionalizations with 4 utilizing the Baran Diversinates™, one was a P450 oxidation, and one was a metalloenzyme oxidation (Figure 2).

The Model

One AI architecture that has seen impressive performance has been message passing neural networks (MPNNs), a subset of graph convolutional neural networks (GCNNs), first utilized by Duvenaud *et al.*, Li *et al.*, and Gilmer *et al.* in the mid-2010s.[15] MPNNs are a robust and versatile way to predict macro properties (i.e. solubility, compound assay activity, IR spectra, energy)[15c, 16] and micro properties (i.e. ^{13}C and ^{1}H NMR shifts, regioselectivity)[11b, 17] of molecules by representing molecules as graphs. Graphs, in mathematics, are structures made up by "nodes" and "edges"; nodes are concrete entities (events, people, atoms, etc.) and edges indicate that two things have a connection (these events happened due to the same cause, these people all know each other, these atoms share a bond). Briefly, MPNNs work by transmitting information from one node to another via the edge "highway". Each message pass transmits the atom's information one bond further away, radially, with the intention that after a sufficient number of message passes, each atom will have a comprehensive understanding of its local environment (Figure 1B).[15c]

We developed an MPNN that sits at ~100 lines of code making it is fast, easy to work with, and highly flexible. The implementation of the MPNN and the trained models can be found at: https://github.com/emmaking-smith/SET_LSF_CODE.[18] To our knowledge, this is the first study that discloses predictive LSF models trained on a large scale dataset across drug-like chemical space comprising both positive and negative results. The MPNN was designed to take in basic atomic information (atomic number, atomic symbol, if the atom was a hydrogen acceptor or donor, its hybridization, if the atom was aromatic or not, and the number of explicit hydrogens) and basic structural information (the connectivity of each atom to its neighbors and the type(s) of bonds used in those connections). If the chemist would not know molecular property X by looking at the structure, that information would not be given to the model either. Rather the model must infer relevant chemical and spatial information from the structure. From this information, the MPNN would synthesize an embedded molecule vector which would then be concatenated with the reagent specific one-hot encoding and run through a feed forward neural network to classify each atom within a molecule as "reactive" or not "reactive" (unreactive).

Finding the Reaction Centers


Glasgow Subgraph Solver with Starting Material
R
4
3
5
2
6
N
1
fixed product atom numbering
all viable subgraph atom numberings
reactive site
reactive sites
symmetry sensitive reactive site identification
atom number "conversion table"

Product Atom Number	SM Atom Number
1	1
2	2 or 6
3	3 or 5
4	4
5	3 or 5
6	2 or 6

Figure 3: Diagram of workflow for identification of reactive sites in symmetric molecules.

The first challenge to overcome was to establish automated extraction of reactive sites, the labels for the ML task at hand. Reaction center identification is a notoriously challenging area of research[19] and for our regioselectivity prediction, we required the atom index(es) of the carbon atoms that changed in oxidation state. Visually, this is a trivial task, but due to the arbitrary nature atom indices across chemoinformatics programs, it becomes much more challenging to perform this automatically. One possible solution is to use atom mapped SMILES strings, where every atom in the product has been traced back to its corresponding atom in the starting material.[19c] However, we believed a more user-friendly approach was possible. For our style of LSFs, the core structure of the molecule remained unchanged, with only the "extremities" exchanging a hydrogen atom for a more complex motif. Therefore, the starting materials were mathematically linked: the starting material was a *subgraph* of the product. In mathematical terms, a subgraph is a graph formed by nodes and edges that are only within its parent graph. From the molecular point of view, a subgraph could be a moiety within a molecule or the core of a molecule. The recent development of a fast, accurate, open-source Glasgow Subgraph Solver was the key to automatically find the starting material subgraph within the product structure, facilitating the extraction of reactive sites.[20] Code for the molecule SMILES to reactive site pipeline can be found at: https://github.com/emmaking-smith/SET_LSF_CODE.[18] In addition to automating the task of finding the LSF reaction centers without the need for atom mapping, the workflow is specifically set up to deal with symmetry in molecules. The Glasgow Subgraph Solver was directed to find all possible subgraph solutions for a given starting material and product, elucidating all possible starting material to product atom mappings. Upon identification of the carbon atom indices whose oxidation state had changed, all corresponding starting material atom indices, including the symmetric indices, were identified labeled as reactive (Figure 3). For degradation byproducts, the fragmentation from the resulting oxidation was oftentimes too dramatic for the starting material to remain a subgraph of the product, resulting in 6% of the reactions needing manual elucidation of reaction center.

The Loss Function

With a model architecture and accurately labeled data in place, we turned our attention to the choice of loss function, the system that penalizes the model and directs the learning. Loss functions can be broadly divided into two categories, regression or classification, where regression loss functions are used with regression tasks and vice versa.[21] Our task was to classify each atom in a molecule as a member of the "reactive" class or not a member of the "reactive" class (unreactive) thus classification loss functions were appropriate. The Binary Cross Entropy (BCE) loss, which penalizes the model based on the log-likelihood of correct class prediction, was chosen (Eq. S2). A challenge with reactivity and regioselectivity prediction is that most atoms in a given molecule are unreactive. Our most reactive molecule had only 30% of its structural atoms reacting, leaving 70% of its atoms unreactive and most molecules in our training data had 1 or fewer reactive structural atoms (Figure 1C). Therefore, a model can be technically accurate by simply predicting that all sites are unreactive, though such model would be practically useless. What was required was a loss function that could more heavily penalize incorrect predictions and give less weight to correct unreactive predictions. To this end, a variety of BCE loss weightings, were investigated, whose central theme was that the weight given to correct class predictions was inversely correlated to the frequency that that class was predicted (Eq. S2 - Eq. S4); the value of each correct reactive site prediction was tempered by how often the model predicted any given atom was reactive, and vice versa for unreactive site prediction.

Model Results

Retrospective Test Set

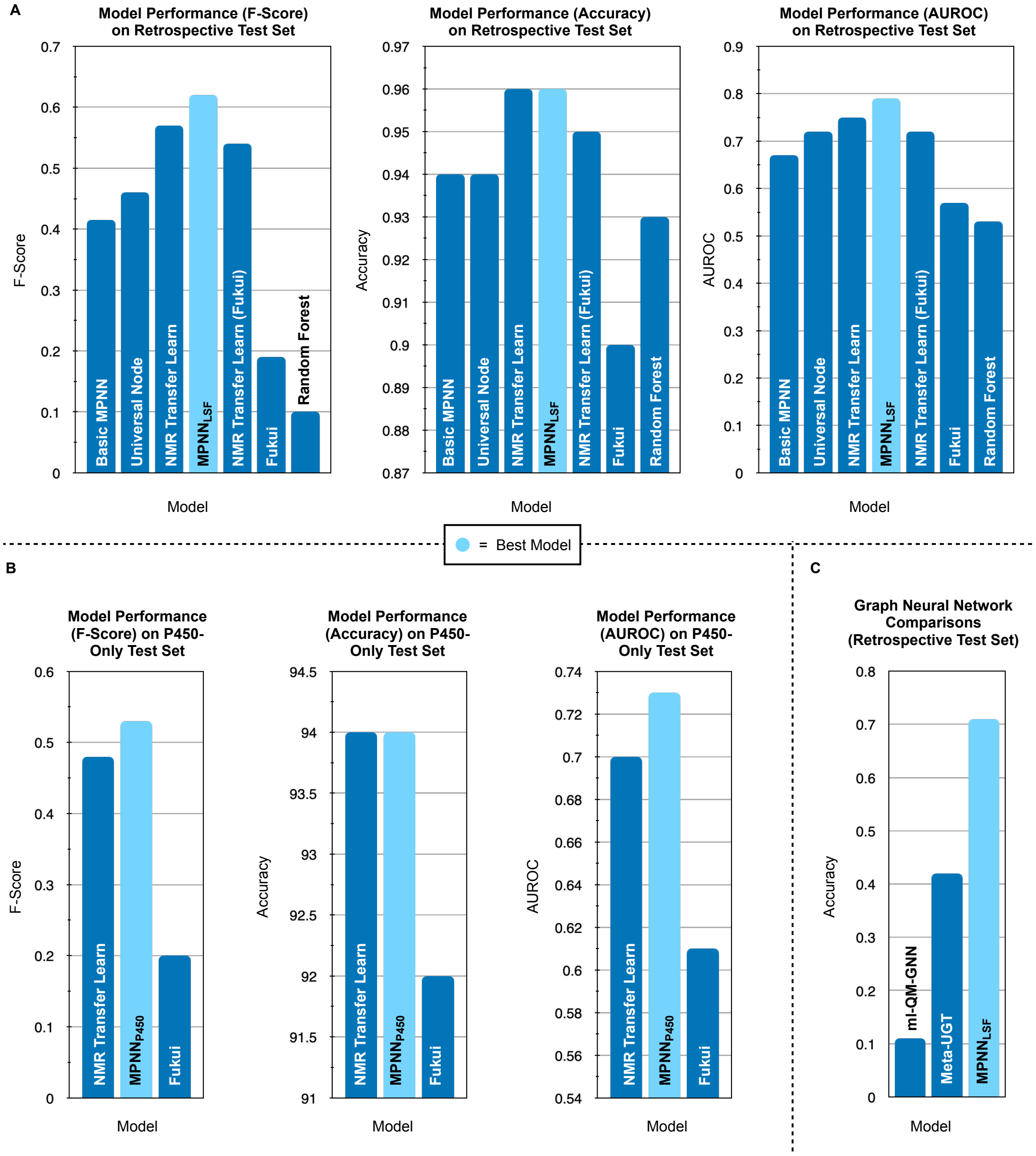


Figure 4: **A**) Average model performance on 5 initializations with each architecture on retrospective test set. "NMR Transfer Learn" is the transfer learned model without Fukui index augmentation. "NMR Transfer Learn (Fukui)" is the transfer learned model with Fukui index augmentation. The best model on the retrospective test set is highlighted. "Fukui" is predictions solely from Fukui indices. **B**) Average model performance on 5 initializations with each architecture on P450 test set with ^{13}C NMR transfer learning. The best model on the test set is highlighted. **C**) Comparison of top-1 accuracy for two graph reactivity models originally developed for 2-electron based (ml-QM-GNN) and 1-electron based (Meta-UGT) transformations.

The baseline model was a random forest, which are known to be excellent predictors of molecular features (*e.g.* compounds increasing the lifespan of *C. elegans*, IC_{50} measurement prediction of drug-like molecules, excitation energies and associated oscillator strengths of fluorophores) especially in low-data environments.[22] Molecules were encoded as their atom-wise Morgan fingerprints. Each row corresponded to the Morgan fingerprint of a specific atom within the molecule. The corresponding one-hot encoded reaction vector was concatenated to the atom-wise Morgan fingerprint and a random forest classifier was then used to predict whether or not each atom in the molecule was reactive or not reactive. We used the well-established classification accuracy metric of the F-score, which balances precision and accuracy to judge model performance. Two other metrics, accuracy (total correct reactive sites predicted / all possible reactive sites) and area under the receiver operating curve (AUROC) are also given for additional interpretability of performance.[23] Initial results on our test set revealed a modest F-score of 0.42 (Accuracy = 94%, AUROC = 0.67), with Fukui-index based predictions yielding a lower F-score of 0.19 (Accuracy = 90%, AUROC = 0.57) (Figure 4A). Fukui indices are predicted only for the molecule, not for the reagent, however, distinctions between different regents are entirely possible. Nucleophilic Fukui indices, $F_i(+)$, correspond to regiochemical outcomes utilizing electrophilic radicals (•CF_3) and radical Fukui indices, $F_i(0)$, correspond to regiochemical outcomes utilizing nucleophilic radicals (•CF_2H, •cBu) (See SI pg. S5 for a mathematical description of each index).[9a] For any radical whose electrophilicity / nucleophilicity reactivity was uncertain, the Fukui indices that best fit the experimental reactivity were used for the calculation of the F-score.

Evaluation of these initial predictions suggested that the model was challenged with extended conjugated systems, such as those present in loratadine (**2**) and imatinib (**5**). We hypothesized that this was due to the difficulty of atoms in one hemisphere of the molecule "seeing" atoms on the other hemisphere in the MPNN. Whilst increasing the number of bonds that every atom's information travels between (the "range" of the atom's message) did not improve performance, the incorporation of a universal node did. This universal node, as described by Gilmer *et al.* (used the term "master node"), is an all-seeing node - information from every atom is given to the universal node, which in turn gives information to every atom about distant atoms.[15c] Implementation of a universal node MPNN led to a model with a modest increase in F-score to 0.46 (Accuracy = 94%, AUROC = 0.72) (Figure 4A).

At this point, we suspected we were running up against the limit of the data. Ideally, this would be solved by the performing additional LSF reactions, however this data is laborious and expensive to generate. Every regioisomer must be isolated and characterized for every new substrate which can be cost and/or time prohibitive. Another obvious solution would be to increase the amount information in each atom's featurization for a deeper understanding of chemical environments. However, given the poor performance of QM-derived atomic descriptors for MPNN regioselectivity prediction in LSF, alternative solutions were sought out first (see the Quantum Chemistry Augmentation Section for a detailed discussion).[11b] Thus, transfer learning was employed. This is a technique whereby a model is trained on off-task data before being trained on the desired-task data to boost performance.[24] It was crucial to choose a transfer learning task that had significantly more data than our current training set which would allow for more complex correlations between structure and reactivity to be inferred. However, it was also imperative that this off-task bore some relationship to atomic reactivity. We hypothesized that ^{13}C NMR shift prediction would be uniquely suited for our goal, which can be abstracted as quantification of local chemical environments. In addition, the inherent symmetry of a molecule is represented in NMR spectra as atoms with identical chemical environments have identical NMR shifts.[25] This would transfer to atoms with identical chemical environments have identical reactivity. Thus, ~27,000 open-source ^{13}C NMR shifts were obtained from Jonas *et al.*'s previous work (originally sourced from NMRShiftDB) and transfer learning from ^{13}C NMR shift to LSF regioselectivity prediction commenced.[17] This step enabled a major improvement in model performance with the top performing model, $MPNN_{LSF}$, yielding an F-score of 0.62 (Accuracy = 96%, AUROC = 0.79) (for every 1 *true positive*, 1.25 incorrect sites are obtained) and an average model performance over 5 initializations of 0.57 (Accuracy = 96%, AUROC = 0.75) (Figure 4A). Interestingly, we observe that negative data is important for model performance.

Removing the entries with zero reactive sites (unproductive reaction conditions) led to a substantial decrease in model performance (Figure S4). We hypothesize that this is because the negative data allows the model to infer similarities between different one-hot encoded reaction conditions.

Comparison to Other Machine Learning Models:

To highlight the difficult nature of predicting Minisci-type transformations without this ^{13}C NMR pretraining protocol, we investigated how another graph-based architecture would perform on our retrospective test set. A recently developed neural network by Jensen *et al.* utilized a joint network approach for 2-electron based regioselectivity prediction. Their first neural network predicted on-the-fly QM properties, which were then given to their second neural network that classified which product was the major product from a user-generated list of possible structures. This approach, dubbed ml-QM-GNN, saw excellent top-1 accuracy performance even in low training data regimes and was validated on a broad range of 2-electron based transformation classes, with a top-1 accuracy of over 85%.[11g] To investigate Minisci-based transformations, we transformed our dataset into the correct format, first elucidating all possible mono-addition C-H functionalizations given our reagent, followed by complete atom mapping of each reaction.[26] Using default parameters, ml-QM-GNN was trained on our training dataset and tested against our retrospective test set. Accuracy was determined using ml-QM-GNN's criteria of top-1 accuracy, where the overall retrospective test set accuracy was the ratio of correctly predicted major products to total number of reactions. As many reactions contained multiple correct possible products, the ml-QM-GNN's classification was deemed correct if its top-1 prediction was any of the valid possible products. Over 5 initializations, the average top-1 accuracy of ml-QM-GNN was 11%, compared to an average top-1 accuracy of 71% for our ^{13}C NMR transfer learning model (Figure 4C).

Finally, we compared our results to a graph-based model specifically developed to predict the outcomes of single-electron-based transformations: Meta-UGT.[27] Meta-UGT was developed to predict the site of metabolism of UDP-glucuronosyltransferases (UGTs). The natural promiscuity of these phase

II metabolic enzymes renders reactivity prediction challenging. The model works in two phases, first predicting if a small molecule is a substrate for the enzyme, followed by the site-specific predictions. When tested upon drug-like molecules, Meta-UGT achieved top-1 site of reactivity prediction accuracy of 89%, making it a suitable candidate to test our model against. Thus, Meta-UGT was trained with default parameters on our training data and tested on the retrospective test set, yielding an average top-1 accuracy of 42% (Figure 4C).

P450-Only Test Set

To investigate this training technique's performance, we devised a different regioselectivity task: P450 oxidation. P450 oxidation plays a central role in drug metabolism, determining the efficacy and duration of a pharmaceutical. Additionally, the interactions of some drugs with human P450s are known to inhibit and/or induce P450 activity leading to drug-drug interactions.[28] Due to its inherent promiscuity,[29] P450 oxidations are a promising LSF and an excellent test for our framework. Mechanistically distinct from Minisci functionalizations, the Fe(IV)-oxo complex acts upon the substrate via radical rebound or through a concerted mechanism, to release the newly oxidized compound (Figure 1B).[29c, 30] Site of metabolism (SoM) prediction, which deduces the most likely positions for human P450 oxidation on a given compound, has seen great strides in the past two decades.[31] We offer this framework as a jumping off point to develop a broadly applicable, isoform-agnostic SoM methodology. Fukui-based indices have also been shown to be effective at determining the regiochemical outcomes P450 oxidations and thus will be used as a baseline measure.[32] Thus, a P450-only test set of 31 reactions and 19 unique molecules (Figure S6), reacting with 18 unique P450s was curated. Employing the aforementioned transfer learning technique to the P450-only test set resulted in an average F-score of 0.48 (Accuracy = 94%, AUROC = 0.70) over 5 initializations. The top performing of these initializations, $\mathrm{MPNN}_{\mathrm{P450}}$, achieved an F-score of 0.52 (Accuracy = 94%, AUROC = 0.73) (Figure 4B). Despite only 25% of the training data containing P450 oxidations, $\mathrm{MPNN}_{\mathrm{P450}}$ outperformed the Fukui-index based reactivity predictions, showcasing the utility of ^{13}C NMR transfer learning.

Quantum Chemistry Augmentation

A lingering question was whether incorporating 3D information and/or quantum mechanical features as input to the graph would help model performance. Conformer generation and quantum chemistry calculations add computational overhead, which would limit this model's applicability in practice. However, many MPNNs that utilize QM-derived information find a significant performance improvement. To this end, a variety of augmentations to the initial atomic features were attempted. However, neither 3D atomic coordinates generated from molecular dynamics (MD) simulations nor electronic information derived from atomic density functions improved overall performance (Figure S5, SI pg. S4-S5). Interestingly, the addition of each atom's electrophilic, nucleophilic, and radical Fukui indices (See SI pg. S5 for a mathematical description of each index) did not see appreciable F-score performance increase in either the prospective or retrospective test sets (Figure 4A & Figure 5E). It is possible that the Fukui indices may not provide any additional information for the MPNN. There have been numerous prior reports which indicate that MPNNs can accurately predict quantum chemical properties from basic atomic information, implying that an MPNN could extract the necessary quantum chemical information from barebones atom featurization, obviating the need for explicit pre-computation of quantum chemical properties.[15c, 16b, 33] This observation is congruent with Nippa *et al.* who independently and concurrently published a MPNN for LSF C-H borylation regiochemical and yield prediction.[11b] They noted that similar augmentation of their atomic information with quantum mechanical features did not lead to noticeable improvement of regioselectivity prediction, and incorporation of 3D atomic coordinates only yielded a modest improvement over 2D molecular representations (scaffold splits). It is possible that the lack of improvement with 3D atomic featurization stems from the difficulty in characterizing properties of the LSF reaction transition state with descriptors that refer to an unperturbed substrate molecule.

Prospective Validation

With the success of our architecture in a variety of LSF regiochemical predictions, we turned our attention to assessing its ability in a completely unbiased setting through prospective prediction. Three maximally structurally different molecules were selected from the Enamine's High Throughput Experimentation catalogue via Butina Clustering.[18] The three compounds were confirmed to not be present within the training or testing data and none had a Tanimoto similarity score over 0.35 with any molecule in the training/testing datasets, indicating low structural similarity between the three prospective compounds and the training/testing data. Each molecule was subjected to CF_2H-, CF_3-, and cBu-functionalization (Figure 5A) and these experimental results were compared to the Fukui-derived indices and $MPNN_{LSF}$ predictions (Figure 5B & 5C). Gratifyingly, $MPNN_{LSF}$ once again outperformed Fukui predictions (Figure 5D), and the random forest baseline, even with respectable performance of Fukui on this prospective test set. All of $MPNN_{LSF}$'s predictions made chemical sense, with predicted functionalizations occurring at known inherently reactive sites or probable sites of oxidation. Fukui predictions often yielded functionalizations at fully oxidized carbons, something that is rarely seen in these LSFs. This is perhaps due to the mechanistically agnostic behavior of Fukui-based predictions, which highlight the site(s) of highest probability for nucleophilic / radical attack, regardless of whether or not those sites lead to productive pathways.

A deeper look at our prospective results sheds light into $MPNN_{LSF}$'s current utility, specifically its highly precise nature. For compound **6**, we see a generally good understanding of inherent pyridine electronics, which is naturally activated the C2, C4, and C6 and positions. However, the effect of the urea motif must be taken into account for a complete picture of regioselectivity. Per the governing heuristics, the π-donating nature of the urea would indicate increased reactivity at the C4 and C6 positions for electrophilic radicals (•CF_3) and reduced reactivity for nucleophilic radicals (•CHF_2, •cBu).[8] Experimentally, it is revealed that the urea motif makes little impact upon the electronics of the pyridine, however, $MPNN_{LSF}$ does not capture this. It instead hedges its bets, correctly finding C2 to be reactive for all three radicals but failing to predict the full chemical reactivity at C4 and C6. This may be in part due to the rarity of the urea motif within our dataset. Out of the ~2,600 training and testing molecules, only 12 contained a urea motif (~0.5% of the data), and of those 12 molecules, functionalization occurred on heterocycles distal to the urea motif. Despite this, MPNN$_{\mathrm{LSF}}$ found 5/9 reactive sites and none of the sites it predicted to be reactive were incorrect.

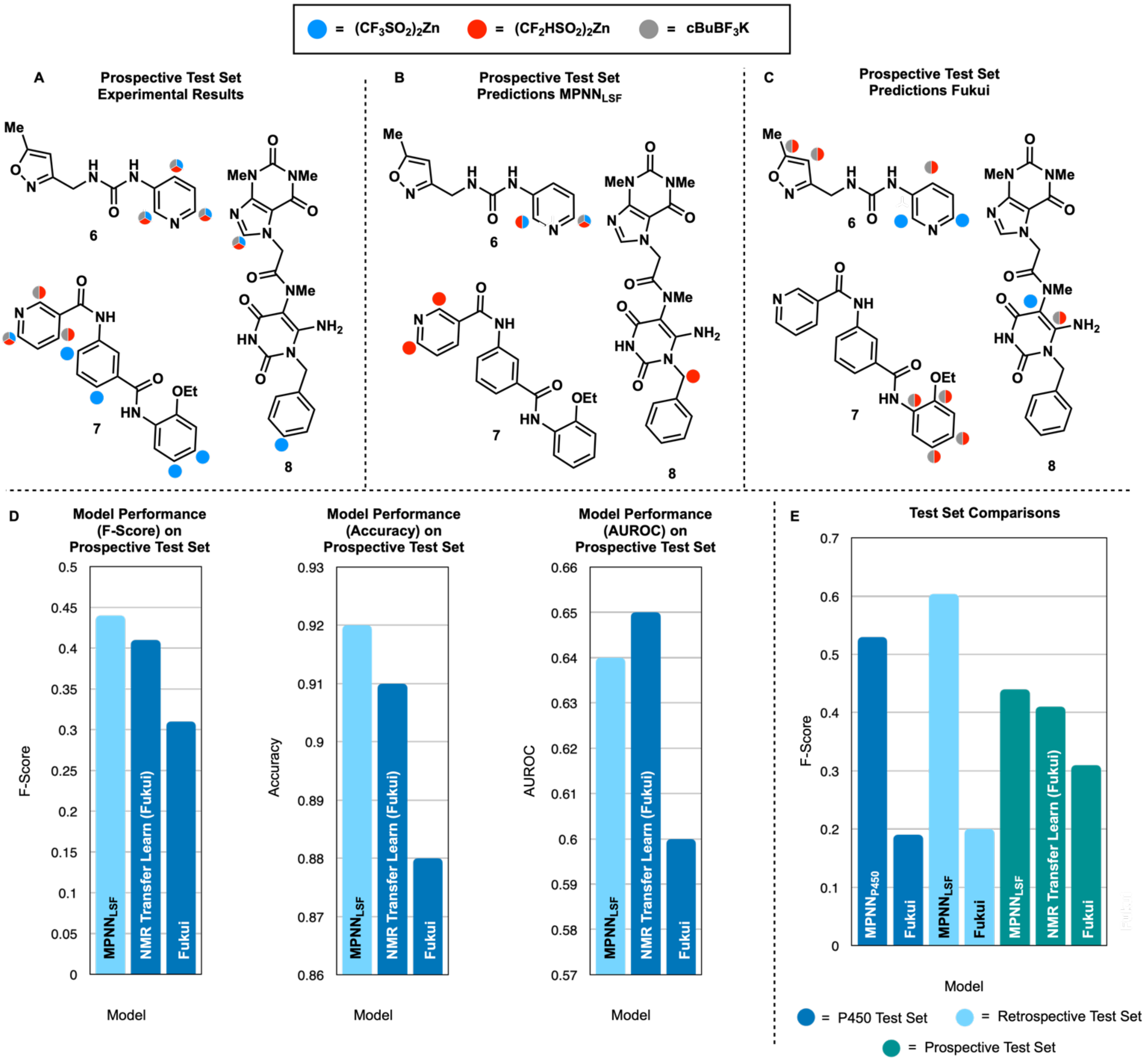
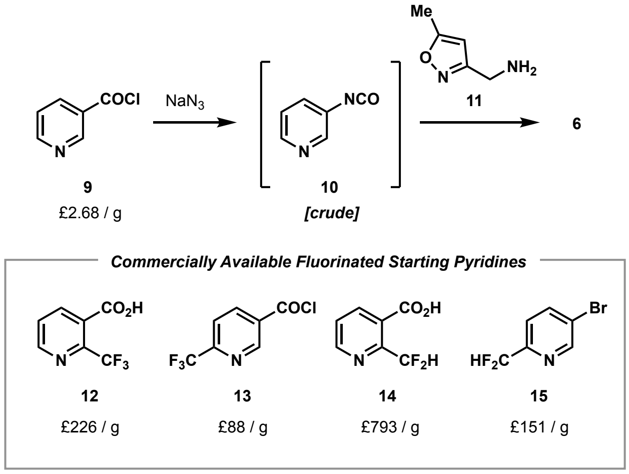


Figure 5: Results on the prospective test set. Color coded by reagent-specific reactivity. Split circles imply more than one reagent functionalized that position. **A**) Experimental results. **B**) MPNN$_{\mathrm{LSF}}$ predictions on prospective test set. **C**) Fukui predictions on prospective test set. **D**) F-Score, Accuracy, and AUROC reported for MPNN$_{\mathrm{LSF}}$, the network that incorporates the Fukui indices as atom features and the baseline Fukui predictions. **E**) Comparison of MPNN$_{\mathrm{LSF}}$, the best Fukui augmented transfer learned model, MPNN$_{\mathrm{P450}}$, and Fukui (no neural network) F-scores on each of the 3 test sets.

For compound **7**, we once again see correct ortho reactivity for •CF_2H however miss the para reactivity for all radicals, perhaps owing to the more sterically congested landscape at that site. However, the clear failure of $MPNN_{LSF}$ was its inability to understand the promiscuous nature of •CF_3 functionalization on **7**. In the majority of Minisci functionalizations, the role of nucleophile is played by the radical, even for electrophilic radicals like •CF_3, and the of role electrophile is played by the heteroarene.[8] Functionalization generally occurs at a (reasonably) electron deficient site. However, compound **7** does not completely follow this trend: all but one of the •CF_3's functionalizations occur on non-heterocyclic, more electron rich arenes, instead of the canonical pyridinyl motif. This atypical substitution pattern plays a large role in the lower performance of $MPNN_{LSF}$ and is even unlikely to be predicted by an expert chemist, highlighting the current limitations of our model: surprising experimental outcomes also surprise $MPNN_{LSF}$.[34]

In compound **8** we finally see a small decrease in $MPNN_{LSF}$'s precision. Instead of identifying the inherently most reactive site on the imidazole, a benzylic oxidation is predicted. The predicted reactivity to difluoromethylation conditions on **8** is likely predicting the major product to be an oxidation byproduct, where the benzylic hydrogen is extracted from the generated alkyl radical and subsequently quenched via TBHP.[35] A prediction of this nature is most likely due to the decision to include byproduct reactions in the training data and lends credence to the hypothesis that the model understands general chemical reactivity trends.

From this analysis, we see that a general trend is the high precision of $MPNN_{LSF}$. This has ramifications in SAR studies, which seek to identify the best decoration of molecular scaffolds for optimal pharmacokinetic properties.[36] In a typical SAR synthesis, one motif is varied and the rest of the molecular structure is held constant. Syntheses of SAR derivates are generally convergent, with the varying motif brought into the synthesis modularly. Despite this workflow's streamlined approach, it still requires each SAR derivative to have its own unique route. A more efficient synthesis would use one reaction to generate multiple desired products. Take compound **6** as an example, with a known route from



Scheme 1: Literature synthesis of compound **6** and the cost of purchasing fluorinated starting materials for a potential SAR campaign.

commercial nicotinoyl chloride (**9**) in an efficient 2-step procedure (Scheme 1).[37] Aryl isocyanate **10** is formed via a Curtius rearrangement, followed by quenching with amine **11** to produce **6**. Current trends in therapeutic molecules have seen the incorporation fluorinated functional groups as substituents on aromatic systems, such as CF_3 and CF_2H, to yield molecules with improved pharmacokinetic properties including lipophilicity, metabolic stability, and cell membrane permeability.[38] Indeed, approximately 20% of all approved pharmaceuticals contain some fluorine-based group.[39] An SAR campaign to investigate the effect of a trifluoromethyl at C2 and C6, would require purchasing the corresponding trifluoromethylated nicotinic acid **12** / nicotinoyl chloride **13**. However, in addition to the added cost of these starting materials (84- and 33-fold more expensive per gram, respectively), the chemist is faced with the challenging task of optimizing and characterizing the outcomes of two small-scale, multi-component, multi-step routes.[40] With MPNN_{LSF}'s high precision, a chemist could be confident that a single route could provide multiple desired derivates in one fell swoop, saving cost of starting material and most importantly, time, both in reaction optimization and in compound characterization. The lower recall isn't as problematic, as any additional "bonus" products can be isolated from the crude reaction mixture concurrently with the correctly predicted functionalizations. The benefit of MPNN_{LSF} becomes more apparent when more exotic functional groups are investigated in SAR. Exploration of difluoromethylation at C2 and C6 by purchasing the necessary difluoromethyl starting pyridines **14** and **15** would be exceptionally expensive: 296- and 56-times more expensive per gram, respectively, of which **15** requires a carbonylation further increasing the time to derivatization.[41] Thus, even without perfect accuracy, MPNN_{LSF} can guide SAR syntheses to produce a multitude of functionalized compounds with minimal time burden.

Conclusion

The regiochemical outcomes of LSF radical-based transformations are governed by many factors: the nucleophilicity of the radical, the BDE of the molecule's atoms, and the steric and electronic landscape to name a few. Interestingly, it has been observed that additional QM-derived or MD-derived data does not yield appreciable improvements in regiochemical outcome prediction. We showcase a transfer learning methodology based upon ^{13}C NMR shift prediction which boosts the performance of zinc sulfinate and BF_3K salt Minisci reaction regiochemical outcome prediction above that of the accurate Fukui-index reactivity scores, and of two reactivity prediction machine learning models, on a narrow yet well-defined slice of chemical space. Promising predictive accuracy was also achieved on P450 enzymatic oxidations, a chemistry with a broader scope than the aforementioned Minisci conditions. Model performance was also highly contingent on the inclusion of negative data in the training set. This paradigm lays the groundwork for future applications in other LSF regiochemical predictions with the current best model showing potential in diversity-oriented SAR synthesis. Our ^{13}C NMR data is open-source and we anticipate that the incorporation of larger proprietary ^{13}C NMR datasets as the first step in this transfer learning methodology will expand this methodology to include in other LSF chemistry.

Data Availability Statement

All code is available under the MIT License at https://github.com/emmaking-smith/SET_LSF_CODE. The repository includes a literature-only dataset of non-proprietary compounds and reactions which is a minimally reproducible example of our workflow.[23, 42] The full dataset is registered with Pfizer as "pfizer_LSF_NatureCommunications_PublicationDate" which can be accessed upon entering a collaboration or legal agreement with Pfizer.

Acknowledgements

Financial support for this work was generously provided by Pfizer and the Royal Society (Newton International Fellowship to EKS and University Research Fellowship to AAL). We wish to thank Rokas Elijošius, William McCorkindale, and Oliver P. King-Smith for their enlightening discussions. We are grateful to Hans Renata and Roger M. Howard for assistance in manuscript preparation. The authors would like to acknowledge several Pfizer colleagues and Spectrix vendor partners who have contributed to this work including: Manjinder Lall, Gregory Walker, R. Scott Obach, and Douglas Spracklin for their leadership and execution of the Lead Diversification Platform (LDP), and Danial Morris for LDP product generation, isolations, bioanalytical support, and anyone else who has contributed to the LDP from the date of its inception.

References

[1] H. Schönherr, T. Cernak, *Angewandte Chemie International Edition* **2013**, *52*, 12256-12267.

[2] H. L. Yale, *Journal of Medicinal and Pharmaceutical Chemistry* **1959**, *1*, 121-133.

[3] E. P. Gillis, K. J. Eastman, M. D. Hill, D. J. Donnelly, N. A. Meanwell, *Journal of Medicinal Chemistry* **2015**, *58*, 8315-8359.

[4] D. Chiodi, Y. Ishihara, **2022**.

[5] S. N. Charlton, M. A. Hayes, *ChemMedChem* **2022**, *17*, e202200115.

[6] aJ. D. Lasso, D. J. Castillo-Pazos, C.-J. Li, *Chemical Society Reviews* **2021**, *50*, 10955-10982; bT. Cernak, K. D. Dykstra, S. Tyagarajan, P. Vachal, S. W. Krska, *Chemical Society Reviews* **2016**, *45*, 546-576; cL. Guillemard, N. Kaplaneris, L. Ackermann, M. J. Johansson, *Nature Reviews Chemistry* **2021**, *5*, 522-545; dM. Moir, J. J. Danon, T. A. Reekie, M. Kassiou, *Expert Opinion on Drug Discovery* **2019**, *14*, 1137-1149.

[7] aJ. M. Smith, J. A. Dixon, J. N. deGruyter, P. S. Baran, *Journal of Medicinal Chemistry* **2019**, *62*, 2256-2264; bR. S. J. Proctor, R. J. Phipps, *Angewandte Chemie International Edition* **2019**, *58*, 13666-13699; cM. S. Lall, A. Bassyouni, J. Bradow, M. Brown, M. Bundesmann, J. Chen, G. Ciszewski, A. E. Hagen, D. Hyek, S. Jenkinson, B. Liu, R. S. Obach, S. Pan, U. Reilly, N. Sach, D. J. Smaltz, D. K. Spracklin, J. Starr, M. Wagenaar, G. S. Walker, *Journal of Medicinal Chemistry* **2020**, *63*, 7268-7292.

[8] F. O'Hara, D. G. Blackmond, P. S. Baran, *Journal of the American Chemical Society* **2013**, *135*, 12122-12134.

[9] aC. A. Kuttruff, M. Haile, J. Kraml, C. S. Tautermann, *ChemMedChem* **2018**, *13*, 983-987; bY. Ma, J. Liang, D. Zhao, Y.-L. Chen, J. Shen, B. Xiong, *RSC Advances* **2014**, *4*, 17262-17264.

[10] aL.-C. Yang, X. Li, S.-Q. Zhang, X. Hong, *Organic Chemistry Frontiers* **2021**, *8*, 6187-6195; bK. Jorner, T. Brinck, P.-O. Norrby, D. Buttar, *Chemical Science* **2021**, *12*, 1163-1175; cX. Li, S.-Q. Zhang, L.-C. Xu, X. Hong, *Angewandte Chemie International Edition* **2020**, *59*, 13253-13259.

[11] aC. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chemical science* **2019**, *10*, 370-377; bD. F. Nippa, K. Atz, R. Hohler, A. T. Müller, A.

Marx, C. Bartelmus, G. Wuitschik, I. Marzuoli, V. Jost, J. Wolfard, **2022**; cT. J. Struble, C. W. Coley, K. F. Jensen, *Reaction Chemistry & Engineering* **2020**, *5*, 896-902; dK. Hasegawa, M. Koyama, K. Funatsu, *Molecular Informatics* **2010**, *29*, 243-249; eN. Ree, A. H. Göller, J. H. Jensen, *Digital Discovery* **2022**, *1*, 108-114; fE. Caldeweyher, M. Elkin, G. Gheibi, M. Johansson, C. Sköld, P.-O. Norrby, J. Hartwig, **2022**; gY. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green, K. F. Jensen, *Chemical Science* **2021**, *12*, 2198-2208.

[12] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, *Chemical Science* **2020**, *11*, 154-168.

[13] Y. Fujiwara, J. A. Dixon, F. O'Hara, E. D. Funder, D. D. Dixon, R. A. Rodriguez, R. D. Baxter, B. Herlé, N. Sach, M. R. Collins, Y. Ishihara, P. S. Baran, *Nature* **2012**, *492*, 95-99.

[14] K. V. Chuang, M. J. Keiser, *Science* **2018**, *362*, eaat8603.

[15] aD. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in neural information processing systems* **2015**, *28*; bY. Li, D. Tarlow, M. Brockschmidt, R. Zemel, *arXiv preprint arXiv:1511.05493* **2015**; cJ. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, in *International conference on machine learning*, PMLR, **2017**, pp. 1263-1272.

[16] aM. Withnall, E. Lindelöf, O. Engkvist, H. Chen, *Journal of cheminformatics* **2020**, *12*, 1-18; bC. McGill, M. Forsuelo, Y. Guan, W. H. Green, *Journal of Chemical Information and Modeling* **2021**, *61*, 2594-2609; cI. Batatia, D. P. Kovács, G. N. Simm, C. Ortner, G. Csányi, *arXiv preprint arXiv:2206.07697* **2022**.

[17] E. Jonas, S. Kuhn, *Journal of Cheminformatics* **2019**, *11*, 50.

[18] https://github.com/emmaking-smith/SET_LSF_CODE.

[19] aE. E. Litsa, M. I. Peña, M. Moll, G. Giannakopoulos, G. N. Bennett, L. E. Kavraki, *Journal of Chemical Information and Modeling* **2019**, *59*, 1121-1135; bA. Lin, N. Dyubankova, T. I. Madzhidov, R. I. Nugmanov, J. Verhoeven, T. R. Gimadiev, V. A. Afonina, Z. Ibragimova, A. Rakhimbekova, P. Sidorov, A. Gedich, R. Suleymanov, R. Mukhametgaleev, J. Wegner, H. Ceulemans, A. Varnek, *Molecular Informatics* **2022**, *41*, 2100138; cW. L. Chen, D. Z. Chen, K. T. Taylor, *WIREs Computational Molecular Science* **2013**, *3*, 560-593.

[20] C. McCreesh, P. Prosser, J. Trimble, in *International Conference on Graph Transformation*, Springer, **2020**, pp. 316-324.

[21] Q. Wang, Y. Ma, K. Zhao, Y. Tian, *Annals of Data Science* **2022**, *9*, 187-212.

[22] aS. Kapsiani, B. J. Howlin, *Scientific Reports* **2021**, *11*, 13812; bV. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947-1958; cB. Kang, C. Seok, J. Lee, *Journal of Chemical Information and Modeling* **2020**, *60*, 5984-5994.

[23] A. Bender, N. Schneider, M. Segler, W. Patrick Walters, O. Engkvist, T. Rodrigues, *Nature Reviews Chemistry* **2022**, *6*, 428-442.

[24] L. Torrey, J. Shavlik, in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, **2010**, pp. 242-264.

[25] M. Kruszyk, M. Jessing, J. L. Kristensen, M. Jørgensen, *The Journal of Organic Chemistry* **2016**, *81*, 5128-5134.

[26] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, *Science Advances*, *7*, eabe4166.

[27] M. Huang, C. Lou, Z. Wu, W. Li, P. W. Lee, Y. Tang, G. Liu, *Journal of Cheminformatics* **2022**, *14*, 46.

[28] aZ. Bibi, *Nutrition & Metabolism* **2008**, *5*, 27; bG. R. Wilkinson, *New England Journal of Medicine* **2005**, *352*, 2211-2221.

[29] aN. D. Fessner, *ChemCatChem* **2019**, *11*, 2226-2242; bC. N. Stout, H. Renata, *Accounts of chemical research* **2021**, *54*, 1143-1156; cE. King-Smith, C. R. Zwick, III, H. Renata, *Biochemistry* **2018**, *57*, 403-412.

[30] B. Meunier, S. P. de Visser, S. Shaik, *Chemical Reviews* **2004**, *104*, 3947-3980.

[31] aA. R. Finkelmann, A. H. Göller, G. Schneider, *ChemMedChem* **2017**, *12*, 606-612; bA. R. Finkelmann, D. Goldmann, G. Schneider, A. H. Göller, *ChemMedChem* **2018**, *13*, 2281-2289; cT.-w. Huang, J. Zaretzki, C. Bergeron, K. P. Bennett, C. M. Breneman, *Journal of chemical information and modeling* **2013**, *53*, 3352-3366; dY. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, D. S. Wishart, *Journal of cheminformatics* **2019**, *11*, 1-25; eS. L. Robinson, M. D. Smith, J. E. Richman, K. G. Aukema, L. P. Wackett, *Synthetic Biology* **2020**, *5*, ysaa004; fZ. Mou, J. Eakes, C. J. Cooper, C. M. Foster, R. F. Standaert, M. Podar, M. J. Doktycz, J. M. Parks, *Proteins: Structure, Function, and Bioinformatics* **2021**, *89*, 336-347.

[32] aM. E. Beck, *Journal of chemical information and modeling* **2005**, *45*, 273-282; bM. M. Fashe, R. O. Juvonen, A. Petsalo, J. Vepsäläinen, M. Pasanen, M. Rahnasto-Rilla, *Chemical Research in Toxicology* **2015**, *28*, 702-710; cP. W. Gingrich, J. B. Siegel, D. J. Tantillo, *Journal of Chemical Information and Modeling* **2022**, *62*, 1979-1987.

[33] J. Zhang, Q. Wang, W. Shen, *Chemical Engineering Science* **2022**, *254*, 117624.

[34] Y. Ji, T. Brueckl, R. D. Baxter, Y. Fujiwara, I. B. Seiple, S. Su, D. G. Blackmond, P. S. Baran, *Proceedings of the National Academy of Sciences* **2011**, *108*, 14411-14415.

[35] J. Tan, T. Zheng, Y. Yu, K. Xu, *RSC Advances* **2017**, *7*, 15176-15180.

[36] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, **2009**.

[37] aE. Rajanarendar, K. Ramu, M. Srinivas, **2004**; bJ. Zhang, W. Tan, Q. Li, F. Dong, Z. Guo, *Marine Drugs* **2020**, *18*, 163.

[38] aW. Zhu, J. Wang, S. Wang, Z. Gu, J. L. Aceña, K. Izawa, H. Liu, V. A. Soloshonok, *Journal of Fluorine Chemistry* **2014**, *167*, 37-54; bY. Zafrani, G. Sod-Moriah, D. Yeffet, A. Berliner, D. Amir, D. Marciano, S. Elias, S. Katalan, N. Ashkenazi, M. Madmon, E. Gershonov, S. Saphier, *Journal of Medicinal Chemistry* **2019**, *62*, 5628-5637; cA. S. Nair, A. K. Singh, A. Kumar, S. Kumar, S. Sukumaran, V. P. Koyiparambath, L. K. Pappachen, T. Rangarajan, H. Kim, B. Mathew, *Processes* **2022**, *10*, 2054.

[39] M. Inoue, Y. Sumii, N. Shibata, *ACS Omega* **2020**, *5*, 10633-10640.

[40] ahttps://www.sigmaaldrich.com/GB/en/product/aldrich/681261, *2-(Trifluoromethyl)pyridine-3-carboxylic acid*; bhttps://www.sigmaaldrich.com/GB/en/product/aldrich/640069, *6-(Trifluoromethyl)pyridine-3-carbonyl chloride*.

[41] ahttps://www.sigmaaldrich.com/GB/en/product/aldrich/741299, *5-Bromo-2-(difluoromethyl)pyridine*; bhttps://www.bldpharm.com/products/P000716069.html, *2-(Difluoromethyl)nicotinic acid*.

[42] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, *Nature Chemistry* **2021**, *13*, 505-508.