Predictive Minisci and P450 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]

Affiliations:

[1] = Department of Physics, Cavendish Laboratory, JJ Thomson Ave, Cambridge, CB3 0HE, United Kingdom

[2] = Development & Medical, Pfizer Worldwide Research, Groton, CT 06340, United States

[3] = Machine Learning Computational Sciences, Pfizer Worldwide Research, Cambridge, MA 02139, United States

[4] = Development & Medical, Pfizer Worldwide Research, Cambridge, MA 02139, United States

[5] = Spectrix Analytical Services, LLC., North Haven, CT 06473, United States

***email:** aal44@cam.ac.uk

Abstract: Structural diversification of lead molecules is a key component of drug discovery to explore close-in 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 message-passing neural network and an ^{13}C NMR-based transfer learning to predict the atom-wise probabilities of functionalization. We validated our model retrospectively and with a series of prospective experiments, showing that it accurately predicts the outcomes of Minisci-type and P450 transformations, outperforming state-of-the-art Fukui-based reactivity indices.

Late-stage functionalization (LSF) is a powerful technique in medicinal chemistry. The "magic methyl effect" is one whereby the addition of a single methyl group, even one distal to the binding motif, can 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 functional handles in

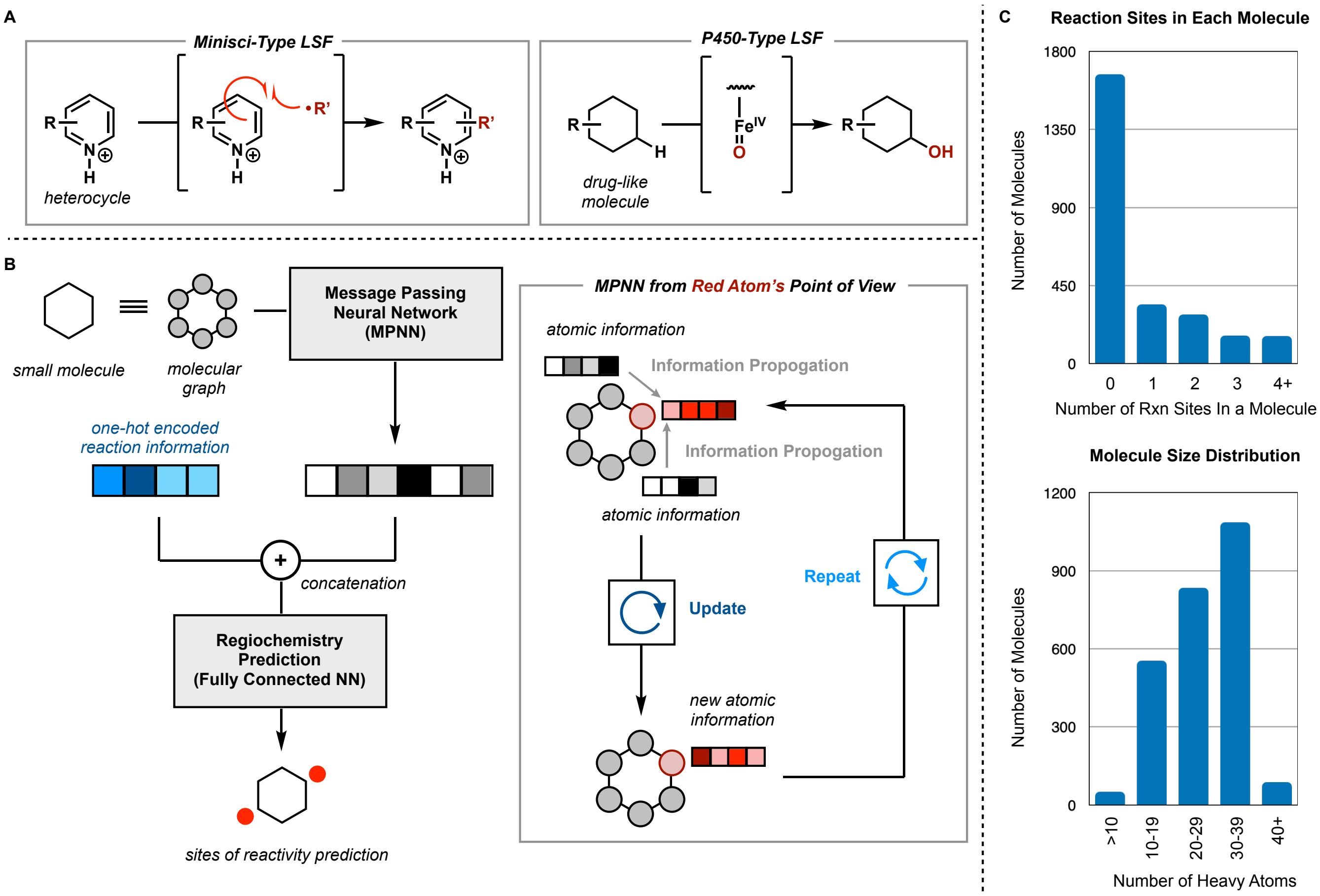

Figure 1. A) Mechanistic differences between the one-electron based transformations of the two major types of reactions in the dataset: Minisci and P450. B) Graphical overview of the basic MPNN model. C) Distribution of reaction sites per molecule and molecule size in the dataset.

generating beneficial motifs. Over the past several decades, numerous methods have been developed to diversify lead compounds and selectively install these biologically privileged motifs directly.[6] One methodology commonly utilized in LSF is the Minisci-type functionalizations, whereby a radical species adds to an electron deficient (hetero)arene (Figure 1A).[7] However, the promiscuity of these single electron methods in conjunction with the inherent structural complexity of LSF molecules make 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. We hypothesized that an ML model can offer improvement upon this state-of-the-art technique (Figure 1B).[9] A consistent and broadly applicable LSF predictive framework would facilitate more rapid and facile access to a diverse array of drug-like compounds, expanding the known possible scope for chemical space 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 real-word data is often noisier and not as readily available as computational DFT data, the utility of ML models built upon experimental data usually outweighs the downsides. Indeed, some of 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 used no pre-computed molecular properties nor any 3D molecular information. As a proof of concept, we highlight our framework's predictive ability on both Minisci and P450 LSFs which outperformed the state-of-the-art Fukui function-based index predictions on retrospective evaluations and a set of prospective experiments.

Data were sourced from Pfizer's internal medicinal chemistry dataset which consisted of 2,613 reactions, 647 unique molecules, and 823 unique LSF conditions which included traditional Minisci functionalizations, Minisci reactions performed with the Baran Diversinates™,[13] P450 mediated oxidations, metalloenzymatic reactions, electrochemical LSF, photoredox alkylations, and even two-electron based LSF such as chlorinations from Palau'chlor.[14] Importantly, our dataset includes reaction condition screening data which contain unsuccessful conditions that led to no significant product formation (zero reactive sites). Given the limited size of the dataset, reactions that yielded oxidative cleavage or hydrolyzed side products were kept in hopes that they would provide the model a deeper understanding of chemical reactivity. When deciding the correct method to split the data in 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.[15] 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

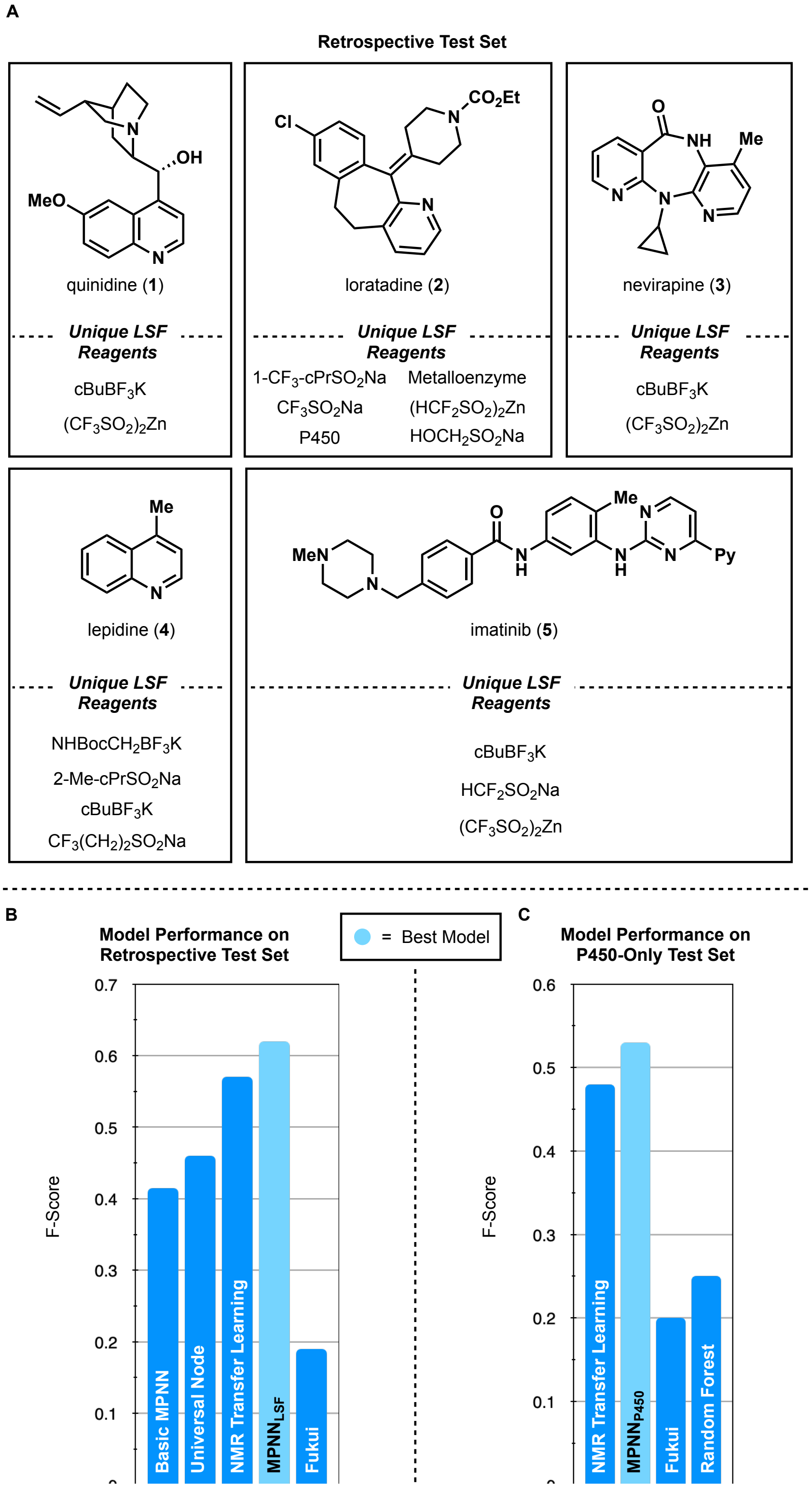

Figure 2. A) The retrospective test set used for optimization of the models. **B**) Top average model performance with each architecture and best individual model on the retrospective test set. Comparisons to Fukui also shown. If not stated, Random Forest baseline is 0. **C**) Top average architecture performance and top model on P450 test set with ^{13}C NMR transfer learning.

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 2A). The first challenge to overcome was to establish automated extraction of reactive sites, which were the labels for the ML task at hand. A notoriously challenging area of research,[16] it was observed that the product structure of successful LSF reactions is nearly identical to the reactant's, different only in the substitution at one or more carbon atoms. Thus, simple automated reaction site extraction could be achieved by treating the product and reactant as graphs and performing an exhaustive subgraph search, identifying the correct site of reaction indices. This was performed by the open-source Glasgow Subgraph Solver.[17] 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.[18] MPNNs are a robust and versatile way to predict macro properties (i.e.

solubility, compound assay activity, IR spectra, energy)[18c, 19] and micro properties (i.e. ^{13}C and ^{1}H NMR shifts, regioselectivity)[11b, 20] 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 atoms' 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).[18c]

We developed an MPNN that sits at ~100 lines of code making it is fast, easy to work with, and highly flexible.[21] This MPNN was only given 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, to predict a given atom's probability of functionalization. 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. For simplicity, the reaction components were one-hot encoded. A random forest regressor was trained on one-hot encoded reaction features and Morgan fingerprints was used as a baseline model. Random forests 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] Comparison of different model variants against any test set was performed with the well-established F-score, which incorporates both precision and accuracy in its calculation.[23] F-scores range from 0 to 1, independent of the number of reactions in a test set, where a score of 1 is a perfectly accurate model. When optimizing model performance, it became apparent that the loss function was critical in model performance. The loss function evaluates how well the model is performing during its training and directs the model to emphasize or deemphasize any given (embedded) molecular features. Exploration of variations on the Binary Cross Entropy Loss revealed several promising avenues (Eq. S1-S4), however the top model's performance was unable to break the 0.5 F-score barrier (Figure 2B). Evaluation of the predictions revealed that the model seemed to be especially 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.[18c] Implementation of a universal node MPNN led to a model with a modest F-score of 0.46 (Figure 2B).

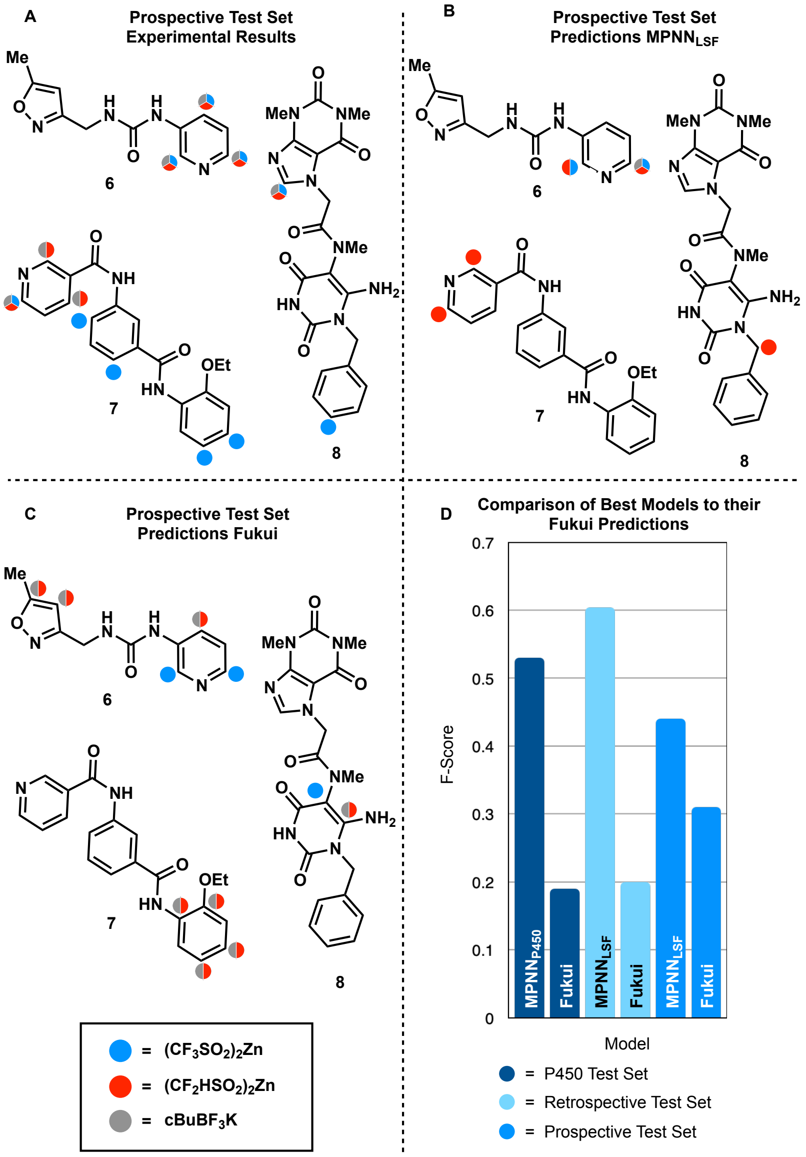

Figure 3. Results on the prospective test set. Color coded by reagent-specific reactivity. Split circles imply more than one reagent functionalized that position. A) Experimental results. B) $MPNN_{LSF}$ predictions on prospective test set. C) Fukui predictions on prospective test set. D) Comparison of $MPNN_{LSF}$ and $MPNN_{P450}$ F-scores against Fukui F-scores.

At this point, we suspected that 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 of each atom, which would hopefully lead to the model gaining a deeper understanding of the local chemical environments. However, given the poor performance of QM-derived atomic descriptors for MPNN regioselectivity prediction in LSF, this was deemed an unsuitable solution.[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. Although these two ML tasks at first glance have little in common with each other, they are two sides of the same coin. Whereas NMR is a measurement of electron density, Minisci LSF (and the state-of-the-art Fukui-derived predictions) are highly predicated on differences of electron density. 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.[20] This step enabled a massive improvement in model performance with the top performing model, MPNN$_{\mathrm{LSF}}$, yielding a top model F-score of 0.62 (4% of reaction sites incorrect) and an average model performance over 5 initializations of 0.57 (Figure 2B). 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 S3). We hypothesize that this is because the negative data allows the model to infer similarities between different one-hot encoded reaction conditions.

To highlight 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.[26] Due to its inherent promiscuity,[27] P450 oxidations are a promising LSF and an excellent test for our framework. Mechanistically distinct from Minisci functionalizations, the crux of a P450 oxidation occurs when an Fe(IV)-oxo complex abstracts a hydrogen atom from a bound substrate, followed by radical rebound to form the newly oxidized product (Figure 1B).[27c] Site of metabolism (SoM) prediction, which deduces the most likely positions for P450 oxidation on a given compound, has seen great strides in the past two decades.[28] We offer this framework as a jumping off point to develop a broadly applicable, isoform-independent SoM methodology. Fukui-based indices have also been shown to be effective at determining the regiochemical outcomes P450 oxidations and thus the indices will continued to be used as the model to beat.[29] Thus, a P450-only test set of 31 reactions and 19 unique molecules (Figure S5), reacting with 18 unique P450s ranging from human to dog to mouse liver extracts was curated. Gratifyingly, our model architecture outperformed the random forest baseline and Fukui predictions, despite only 25% of the training data containing P450 reactions (Figure 2C). The best model, MPNN$_{\mathrm{P450}}$, achieved comparable F-score performance to MPNN$_{\mathrm{LSF}}$ on its P450-only test set (Figure 3D).

A lingering question is 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 utilize QM-derived information including 3D atomic coordinates and find a significant performance improvement. We implemented a MPNN with atomic features built upon molecular dynamics (MD) simulations and atomic density representations with implicit 3D atomic information. We found no significant improvement over the non-MD model (Figure S4). This observation is congruent with Nippa *et al.* which 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.

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.[21] 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 3A) and these experimental results were compared to the Fukui-derived indices and $MPNN_{LSF}$ predictions (Figure 3B & 3C). Gratifyingly, $MPNN_{LSF}$ once again outperformed Fukui predictions (Figure 3D), and the random forest baseline, even with remarkable 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. Specifically, difluoromethylation on compound **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.[30] 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. Fukui-based reactivity index predictions performed moderately well on **6** and **7**, identifying the two correct, non-obvious, trifluoromethylations sites at the 2-ethoxyphenylacetamide moiety. However, 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.

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 on ^{13}C NMR shift prediction which boosts performance well above the current state-of-the-art. Model performance was also highly contingent on the inclusion of negative data in the training set. This paradigm yielded models that outperformed Fukui-based predictions for all three test sets, laying the groundwork for future applications in other LSF regiochemical predictions. 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 further improve the regiochemical predictions.

[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, *ChemRxiv preprint* **2022**, DOI: 10.26434/chemrxiv-2022-5mbcp.

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

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

[7] a) J. M. Smith, J. A. Dixon, J. N. deGruyter, P. S. Baran, *Journal of Medicinal Chemistry* **2019**, *62*, 2256-2264; b) R. S. J. Proctor, R. J. Phipps, *Angewandte Chemie International Edition* **2019**, *58*, 13666-13699; c(M. 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] a) C. A. Kuttruff, M. Haile, J. Kraml, C. S. Tautermann, *ChemMedChem* **2018**, *13*, 983-987; b) Y. Ma, J. Liang, D. Zhao, Y.-L. Chen, J. Shen, B. Xiong, *RSC Advances* **2014**, *4*, 17262-17264.

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

[11] a) C. 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; b) D. F. Nippa, K. Atz, R. Hohler, A. T. Müller, A. Marx, C. Bartelmus, G. Wuitschik, I. Marzuoli, V. Jost, J. Wolfard, **2022**; c) T. J. Struble, C. W. Coley, K. F. Jensen, *Reaction Chemistry & Engineering* **2020**, *5*, 896-902; d) K. Hasegawa, M. Koyama, K. Funatsu, *Molecular Informatics* **2010**, *29*, 243-249; e) N. Ree, A. H. Göller, J. H. Jensen, *Digital Discovery* **2022**, *1*, 108-114; f) E. Caldeweyher, M. Elkin, G. Gheibi, M. Johansson, C. Sköld, P.-O. Norrby, J. Hartwig, **2022**; g) Y. 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] R. A. Rodriguez, C.-M. Pan, Y. Yabe, Y. Kawamata, M. D. Eastgate, P. S. Baran, *Journal of the American Chemical Society* **2014**, *136*, 6908-6911.

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

[16] a) E. 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; b) A. 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; c) W. L. Chen, D. Z. Chen, K. T. Taylor, *WIREs Computational Molecular Science* **2013**, *3*, 560-593.

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

[18] a) D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in neural information processing systems* **2015**, *28*; b) Y. Li, D. Tarlow, M.

Brockschmidt, R. Zemel, *arXiv preprint* **2015**, DOI: arXiv:1511.05493. c) J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, in *International conference on machine learning*, PMLR, **2017**, pp. 1263-1272.

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

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

[21] git@github.com:emmaking-smith/SET_LSF_CODE.git.

[22] a) S. Kapsiani, B. J. Howlin, *Scientific Reports* **2021**, *11*, 13812; b) V. 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; c) B. Kang, C. Seok, J. Lee, *Journal of Chemical Information and Modeling* **2020**, *60*, 5984-5994.

[23] Y. Sasaki, *Teach tutor mater* **2007**, *1*, 1-5.

[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] a) Z. Bibi, *Nutrition & Metabolism* **2008**, *5*, 27; b) G. R. Wilkinson, *New England Journal of Medicine* **2005**, *352*, 2211-2221.

[27] a) N. D. Fessner, *ChemCatChem* **2019**, *11*, 2226-2242; b) C. N. Stout, H. Renata, *Accounts of chemical research* **2021**, *54*, 1143-1156; c) E. King-Smith, C. R. Zwick, III, H. Renata, *Biochemistry* **2018**, *57*, 403-412.

[28] a) A. R. Finkelmann, A. H. Göller, G. Schneider, *ChemMedChem* **2017**, *12*, 606-612; b) A. R. Finkelmann, D. Goldmann, G. Schneider, A. H. Göller, *ChemMedChem* **2018**, *13*, 2281-2289; c) T.-w. Huang, J. Zaretzki, C. Bergeron, K. P. Bennett, C. M. Breneman, *Journal of chemical information and modeling* **2013**, *53*, 3352-3366; d) Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, D. S. Wishart, *Journal of cheminformatics* **2019**, *11*, 1-25; e) S. L. Robinson, M. D. Smith, J. E. Richman, K. G. Aukema, L. P. Wackett, *Synthetic Biology* **2020**, *5*, ysaa004; f) Z. 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.

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

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

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.