Graph transformer neural network for chemical reactivity prediction

David F. Nippa¹², Kenneth Atz¹³†, Alex T. Müller¹, Jens Wolfard¹, Clemens Isert³, Martin Binder¹, Oliver Scheidegger¹, David B. Konrad²†, Uwe Grether¹*, Rainer E. Martin¹† & Gisbert Schneider³⁴,*

¹Roche Pharma Research and Early Development (pRED), Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Grenzacherstrasse 124, 4070 Basel, Switzerland.
²Department of Pharmacy, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, 81377 Munich, Germany.
³ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
⁴ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore, Singapore.
† These authors contributed equally to this work.
* To whom correspondence should be addressed.
E-mail: david.konrad@cup.lmu.de, uwe.grether@roche.com, rainer_e.martin@roche.com, gisbert@ethz.ch

Abstract

Optimizing the properties of advanced drug candidates can be facilitated by directly introducing certain chemical groups without having to synthesize the molecules from scratch. However, their chemical complexity often renders reactivity predictions and synthesis planning challenging. Herein, we introduce a graph transformer neural network (GTNN) approach for computational reaction screening and identification of substrates suitable for late-stage functionalization, taking compound alkylation via Minisci-type chemistry as an example. GTNNs were trained on experimentally generated reactions obtained from miniaturized high-throughput experimentation and literature data. Trained models were prospectively applied to predicting the reactivity of 3180 advanced heterocyclic molecules, identifying potential substrates for Minisci-type alkylation. All predicted substrates were experimentally confirmed. Multiple chemical transformations were identified for each of these compounds. Selected hits were scaled up, isolated, and characterized, delivering 30 novel, suitably functionalized molecules for medicinal chemistry. These results positively advocate GTNN models for reactivity prediction in drug discovery.

1 Introduction

The synthesis of new compounds is the rate-limiting and most time-consuming step for many small molecule drug discovery projects. [1] Late-stage functionalization (LSF) introduces additional functional groups to drug molecules, thereby avoiding de novo synthesis or the need for functional handles. [2] These minor structural modifications facilitate the elucidation of underlying structure-activity relationships (SARs) and enable the optimization of pharmacokinetic properties (absorption, distribution, metabolism, and excretion (ADME)) of lead compounds and drug candidates at reduced synthetic costs. [3] However, not all molecules are susceptible to desired functionalizations, which renders LSF experimentally challenging. Herein, we propose a computational deep learning framework for reactivity prediction of drug molecules that could provide a more rational access to LSF and limit the time and experimental costs typically required.

An increasing number of experimental LSF methods have recently been published that allow medicinal chemists to fluorinate, aminate, arylate, methylate, trifluoromethylate, borylate, acylate, or oxidize structurally intricate molecules. [4, 5] Alkylation reactions have gained interest as they allow the introduction of small cyclic and acyclic alkyl groups through carbon-carbon, carbon-oxygen, or carbon-nitrogen bond formation. [6] In particular, Minisci-type alkylations [7, 8] are considered a valuable LSF methodology for incorporating alkyl building blocks into heterocyclic systems, which often form the core of drug molecules. [9] Originally described in the mid-20th century, Minisci reactions have become a versatile tool in medicinal chemistry for the formation of C-C bonds. [11] Using ammonium persulfate as the oxidant and silver nitrate as the catalyst, alkyl radicals are generated from the corresponding carboxylic acids at elevated temperatures. Upon radical addition to the heteroarene, the reaction product is formed through aromaticity-driven oxidation of the radical intermediate. [11] The scope of both, electron-deficient heteroarenes and alkyl-donating coupling partners, has steadily been expanded. [12, 13] Different radical sources have been reported, such as alkyl carboxylic acids capable of transferring alkyl groups, boronic acids suitable for the incorporation of aryl groups, or sulfinates that were used to transfer trifluoromethyl or tert-butyl fragments. [14, 15] The use of widely available and inexpensive carboxylic acids without the need for prefunctionalization considerably expands the scope of this transformation for drug discovery. [16] The increasing push towards incorporating sp3-rich building blocks into drugs, [17] and the availability of numerous stable cyclic alkyl carboxylic acids make this approach highly attractive for hit-to-lead expansion and drug optimization by LSF.

It has become evident that, by reducing the number of aromatic rings in a drug candidate, the likelihood of clinical success can be increased. [18] Greater fractions of sp3 centers enabled access to new chemical space, potentially enhancing drug selectivity, [19] with a positive impact on physicochemical properties, namely solubility and metabolic stability. [20–22] Despite the
Figure 1: Overview of the research study. Screening plate design: Minisci literature data containing metal-free reactions were extracted and analyzed to determine suitable reaction conditions. For parallel reaction screening, 23 sp³-rich carboxylic acids with relevance for drug discovery were included. Reaction data generation: Using the reaction plate design, physical experiments in high-throughput experimentation (HTE) fashion were conducted with marketed drugs and fragments from an informer library (184 reactions [10]) covering relevant chemical space. In addition, 16 distinctly non-reactive substrates were screened for in silico decoy data generation (368 reactions). Geometric deep learning: The obtained reaction data (SURF, Simple User-friendly Reaction Format) [10] were subjected to geometric deep learning, incorporating 3D structural information of the chemicals. The trained model was applied to 3000 building blocks from the Roche library, with a particular focus on electron-deficient heterocycles. This in silico screening predicted the reactivity of the compounds for substrate ranking and clustering. Validation and application: The prediction models were experimentally validated for a diverse set of 18 building blocks. Selected scale-up reactions led to fully characterized compounds.

availability of guidelines for reactivity prediction for Minisci-type transformations, the limited scope of functional groups tolerated, as well as the varying types of C-H bonds and electronic effects in a complex molecule makes the prediction of alkylation reactions difficult. [3, 23]. Performing individual reactions on a typical medicinal chemistry scale (milligram scale) to augment the reaction database with relevant transformation examples would be a time- and resource-intensive endeavour, with a rather limited value-to-effort ratio.

High-throughput experimentation (HTE) has emerged as a valuable tool to experimentally explore and optimize new chemical transformations in a semi-automated fashion. [24, 25] In order to successfully achieve reaction miniaturization at nanomolar scale, the system needs to be engineered to precisely handle very small amounts of material and guarantee continuous and equal mixing of the reaction components. [26] Another important aspect of HTE concerns the curation of all obtained reaction data, in particular failed transformations, following FAIR principles (findable, accessible, interoperable, and reusable) [27] to obtain high-quality data sets for machine learning. [10, 28–30] In parallel to the advances in LSF and HTE, geometric deep learning, in particular graph neural networks (GNNs) that enable efficient learning on three-dimensional (3D) molecular models, has found various applications in drug discovery and development. [31–33] Next to applications addressing quantum chemistry-related questions [34–37], GNN methods were developed for forward reaction prediction from small substrates to complex drug molecules. [38–40] GNNs have recently been applied to LSF to address reaction yield, binary reaction outcome, and regioselectivity prediction for borylation reactions. [10] A similar approach was introduced for the prediction of late-stage alkylation, mainly focusing on the Baran-type diversinate chemistry using alkyl sodium sulfinate salts. [41] A recent study has shown that hybrid machine learning models augmented with quantum chemical information of transition states enable accurate regioselectivity prediction for iridium-catalyzed borylation reactions in very low data regimes. [42]

Herein, we introduce the combination of GNN-based in silico reaction screening with a miniaturized HTE setup for identification of LSF alkylations by Minisci reaction, addressing various aspects relevant to drug discovery (Figure 1). We demonstrate how GTNNs trained
on a limited amount of reaction data can be utilized for deep learning-based in silico reaction screening and combined with automated reaction screening. The new approach enabled the identification of 276 alkylation opportunities with high accuracy. A diverse array of novel compounds with increased sp³ fraction could be identified, scaled up and characterized.

2 Results

2.1 HTE reaction screening

The Minisci-type reactions described by Sutherland et al. [16] were successfully miniaturized (factor 300) from micromolar (150 µmol) to nanomolar scale (500 nmol) in a parallel set-up on a 24-well plate. During the optimization process, it became evident that the reaction provides considerably higher yields when conducted in a glovebox. Running the reaction with 23 different carboxylic acids a-w (Figure 2) at different temperatures provided the highest conversions at 40 °C, as heating to higher temperatures primarily delivered di-alkylation products. Doubling the equivalents of alkyl carboxylic acids (20 instead of 10 equivalents) and oxidant (6 instead of 3 equivalents) were used to achieve higher conversions (factor 1.2-1.5) on average. A reference reaction (Quinoline 1 with carboxylic acid e in position B4, Figure 2) was included to monitor potential performance issues and ensure the reproducibility of the screening results. In the final set-up, the introduction of 23 different alkyl groups, most of them being small sp³ ring systems, to electron-deficient heterocycles was assessed. Binary reaction outcomes were considered "successful" if the reaction condition with the chosen substrate yielded a mono- or di-alkylation product that could be confirmed by liquid chromatography-mass spectrometry (LCMS) with a threshold of 5%, or "unsuccessful" if the desired transformation was not traceable with LCMS. Four fragments (1-4, Section SI4, Figure S1) and five drug molecules (5-9, Section SI4, Figure S1) from a chemically diverse LSF informer library [10], and 18 fragments (26-43, Section SI4, Figures S3 and S4) from the Roche compound library were screened under these reaction conditions, generating a balanced experimental data set of 691 reactions (379 successful, 312 failed).

2.2 Deep learning-based in silico reaction screening

GTNN models (Figure 3) were trained on a preliminary data set of 621 Minisci reactions (368 originating from decoy generation, 45 from literature, and 208 from the LSF informer library). These models enabled in silico reaction screening of a Roche in-house library of 3180 advanced heterocyclic building blocks. Each substrate was labelled with the ensemble score of six independent models, (i.e., three models for binary reaction outcome prediction and three models for reaction yield prediction). Subsequently, the molecules were grouped into
eight clusters using agglomerative compound clustering (Section 4.7 and Figure 4). Two compound clusters were neglected due to the abundance of unsuitable structures (i.e., heterocycles without free C-H bonds) for the investigated reaction. Three molecules from the six remaining clusters were selected based on their computed reactivity score resulting in 18 N-hetero arenes.

The selected 18 N-hetero arenes were subjected to automated HTE screening, generating an experimental data set of 414 reaction points. For each of the selected substrates, Minisci-type alkylation products could be identified resulting in a total number of 276 successful reactions (100% correct predictions, Figure 5A). Ten of the screened N-hetero arenes enabled 17 to 23 successful transformations among the selected carboxylic acids (Figure 5B). Seven N-hetero arenes allowed ten to 17 successful transformations. For meta-substituted pyridine 42 (Section SI4, Figure S4), fewer than ten successful reactions (Figure 5B) were observed (Section SI9, Figure S9). There were three five-membered N-heterocyclic ring systems (2, 4, 9) in the LSF informer library, for which very low reaction yields (≤4%, averaged over 23 carboxylic acids) were observed.

Based on these initial studies, GTNN models were trained on the full experimental data set of 691 Minisci reactions generated by HTE (277 from the LSF informer library and 414 from the in silico reaction screening). Reaction yields were predicted with a mean absolute error (MAE) of 18.7 (±0.2)% and a Pearson correlation coefficient (PCC) of 0.687 (±0.006) (Figure 5C). Reaction yields were categorized into four ranges: no reaction (<1% yield), poor (>1-11%), medium (>11-35%), and high reaction yield (>35-100%). The model predicted the correct category in 55.7 (±0.7)% of the cases. Binary reaction outcomes were predicted with an absolute accuracy of 81 (±1), and an F-score of 82.7 (±0.6)% (Figure 5D). For additional neural network analyses and descriptions of the chosen metric, see Section SI2.

2.3 Scale-up

Selected screening conditions were used for upscaling to the milligram range. LSF alkylation was carried out for the drug molecules Loratadine (7) and Nevirapine (8), and structurally complex molecular fragments. In total, 30 novel molecules were synthesized, isolated, and characterized by nuclear magnetic resonance (NMR) spectroscopy and high-resolution mass spectrometry (HRMS) (Figure 6).

For Loratadine (7), a molecule from the LSF informer library, several analogues with different cyclic (7b1, 7b2, 7b3, 7j1, 7j2, 7e1, 7e2) and heterocyclic (7s, 7q1, 7q2) substituents were generated. Structurally complex scaffolds with high relevance for medicinal chemistry projects, which could serve as starting points for the development of SAR studies, also provided a variety of interesting alkylation products. Different alkyl groups, covering alkyl chains (e.g., 40h, 33h, 28h), cyclic alkyls (e.g., 26e, 41e, 38e) and cyclic ethers (e.g., 39u, 35m) could be introduced. In general, the observed regiochemistry was consistent with Minisci guidelines, with the alkyl groups being introduced in either the ortho- or para-position on the pyridine core. [23] For molecule 38, different reactivity was observed with the cyclohexyl radical reacting exclusively with the thiocarbonyl functionality affording thiether 38e. No reaction at the pyridine core was observed.
2.4 Reactivity trends

Analysis of the generated data delivered a heterogeneous distribution of the observed reaction yields for the carboxylic acids as well as for the N-hetero arenes. Cyclic ethers (e.g., u, s, a) and alkanes (e.g., b, e, g) were reliably converted to the desired alkylation product, whereas cyclic boc-protected amines (e.g., o, p, q, r) and amides (d) resulted in low yields of the respective desired reaction products (Figure 7). Similarly, substituted pyridines (e.g., 30, 31, 36, 39; see Section SI4, Figures S3-S4) had lower yields compared to compounds lacking a meta-substituent (e.g., 26, 32, 38, 41; see Section SI4, Figures S3-S4). Electron-rich meta-substituted pyridines, such as 3 and 27, had a comparably low average reaction yield compared to their less electron-rich analogues. Overall, compared to their six-membered N-hetero analogues, five-membered N-heterocyclic ring systems (e.g., 2, 4, 9; see Section SI4, Figure S1) did not show significant conversion to the desired alkylation product.

3 Discussion

The Minisci reaction conditions employing ammonium persulfate ((NH₄)₂S₂O₈) as the oxidation reagent and solvent dimethylsulfoxide (DMSO) at 40 °C were successfully miniaturized and converted into a parallel screening format that enables the reaction with a broad variety of alkyl carboxylic acids in an efficient and resourceful manner. [16] The optimized reaction protocol enables fast, metal-free and resource-saving evaluation of reaction conditions in an HTE format to guide further synthesis decisions. It also avoids the time-consuming execution of individual reactions on a milligram scale. Still, this setup comes with some limitations that should be addressed in future research: (i) By design, the current plate only tests one set of reaction conditions, aiming to reduce complexity. However, the evaluation of additional oxidants or solvents, and adjustments of the reaction component equivalents could lead to further improved reaction yields. Furthermore, Minisci-type reactions are usually catalyzed by metals, such as silver or iron. [9] Systematic HTE assessment of different metal salts could deliver further enhanced conditions. (ii) Rather than utilizing solely carboxylic acids as the alkyl donating group, other forms of radical precursors, such as boronic acids or sulfinates, may be explored. [13] This analysis would potentially enable to further broaden the scope of alkyl groups for medicinal chemistry. (iii) Numerous photochemical Minisci-type transformations have been disclosed. [13] These reactions offer alternative radical generation mechanisms that could further expand the scope for LSF.

The user-friendly reaction data format (SURF) [10] proved to be vital for the whole study. It supported reaction data collection from the literature and allowed for standardized reporting of the HTE results and in silico reaction screening data. Thus, the initial reaction data originating from the three different sources (45 literature, 208 experimental and 368 decoy reactions) were readily available for deep learning without the need for manual data curation. SURF guaranteed fast turnaround times between literature evaluation, experiments and model training. As both the experimental and especially the literature data predominantly contain positive results, including decoy data of unsuccessful transformations was crucial for building a reliable prediction model.
A detailed look at the experimental data revealed that cyclic Boc-protected amines (o, p, q, r, v), as well as amides (e.g., d) mainly afforded low yields (5-20%) of the desired reaction products (Section SI9, Figure S8). This observation reflects the half-lives of the generated radical intermediates, [44] e.g., with tertiary carbon radicals (e.g., h) having higher stability than primary carbon radicals (e.g., k) and the latter thus resulting in lower product yields. Another experimental trend relates to the substitution pattern of N-heteroarenes. Meta-unsubstituted pyridines (e.g., 26, 32, 41) consistently provided higher yields than substituted analogues, (e.g., 35, 36, 37) as residues on the meta-position sterically hinder the reaction in ortho- and para-positions to the pyridine (Section SI9, Figure S9). Finally, electron-rich meta-substituted pyridines, such as 3 and 27, had a very low (5-10%) average reaction yield on the screening plate when compared to their less electron-rich analogues (Section SI9, Figure S8). This low reactivity is owed to the electron-rich amine- and methoxy-substituents, respectively. [23]

Compared to a previous study [10], in which GTNNs allowed for a single graph input, the GTNN model described here accept two different molecular inputs, corresponding to the two reactants (N-hetero arenes and carboxylic acids). The network architecture was tailored to the Minisci-type alkylation transformation in such a way that trained GTNNs can be applied to novel N-hetero arenes as well as carboxylic acids. Therefore, the model can be used for in silico molecular library screening for both types of reaction inputs. It could be shown that in silico reaction screening using GTNN models trained on a comparably small preliminary data set consisting of 576 Minisci reactions (i.e., 368 from decoy generation, 45 from literature, and 208 LSF from an informer library) led to the identification of 17 substrates (i.e., 94% of the 18 selected molecules). All newly identified substrates were successfully alkylated with a broad range of at least ten different carboxylic acids. Furthermore, in total 276 successful reactions (i.e., producing alkylation products with a median yield of 26.3%) were identified. The low reaction yields observed for three five-membered N-heterocyclic ring systems (2, 4, 9) indicate that the GTNN models learned to de-prioritize five-membered N-hetero arenes during in silico reaction screening. It was shown how a clustering approach can be combined with in silico reaction screening to assess structural diversity as well as reactivity. The incorporation of electronic features into the molecular graph did not yield improved model accuracy for the investigated tasks (Section SI2). This particular observation suggests that coarse-grained reactivity prediction is possible without the strict need for computationally costly quantum calculations.

With the overall goal of synthesizing novel scaffolds that are relevant to medicinal chemistry, the visualized screening data served to identify appropriate reaction conditions for upscaling to the milligram scale. Again, the SURF data format was instrumental for the laboratory chemist to set up experiments efficiently by providing the CAS number, SMILES string, equivalents, and overall reaction conditions in a comprehensive and easily accessible format. The reaction conditions were reproducible at a higher scale, underscoring the applicability of this approach to drug discovery. With the exception of compound 38e, all reactions yielded C-C coupling products. In general, the observed regioselectivity was in agreement with the expected reaction products according to the rules reported in the literature. [23]
Figure 6: Selected examples of characterized Minisci reaction products from the LSF drug informer library (left panel) and the fragment screening (right panel). The added alkyl groups are highlighted in blue. Late-stage drug alkylation examples include derivates of the drugs Loratadine (7s, 7b1, 7b2, 7b3, 7q1, 7q2, 7j1, 7j2, 7t1, 7t2, 7e1, 7e2) and Nevirapine (8s). Fragment screening highlights the diverse range of introduced substituents, covering cyclohexanes (26e, 41e, 38e), cyclobutanes in different positions (29b, 34b1, 34b2, 37b1, 37b2), heterocyclic alkanes (39u, 35m) and tert-butyl (40h, 33h, 28h). Abbreviations: Boc: tert-Butyloxycarbony, Ph: Phenyl.

However, when moving to more densely functionalized pyridines, these reported literature guidelines do not appear to apply. While the reaction of 34b and 37b primarily generated the expected ortho-substituted reaction products 34b1 and 37b1, also meta-substituted reaction products 34b2 and 37b2 were obtained, albeit in lower amounts (Figure 6). In the literature, amides are described as ortho–para directing groups due to their electron-withdrawing effect, and aryl ethers as ortho-activating moieties due to their electron-donating nature. [23] The formation of regioisomer 34b2 might have been sterically hindered by the amidyl side chain, favouring the meta- over the para position. For 37b2, an explanation of the formation could lie in the several different functional groups that are attached to the pyridine ring, which only leave the meta position available for substitution, despite this position being sterically hindered by the proximity of the aryl sulfide and the CF₃ group. Lastly, 38e showed significantly different reactivity despite bearing a pyridine moiety. This observed reaction product can be rationalized by the greater reactivity of the lone pairs of the sulfur as compared to the C-H bonds of the pyridine side-chain. These results of the scale-up reactions underscore the importance of generating high-quality, single-batch LSF reaction data.

Taken together, the results of this study demonstrate the applicability and benefits of combining laboratory automation, parallel miniaturized screening, and deep learning to enable faster, more efficient, and more economic synthesis in drug discovery. This integrative approach is already in productive use at Roche. The predictive power of the computational model will be continuously improved by feeding the algorithm with an increasing number of newly generated LSF reaction data points covering the relevant medicinal chemical space. For further method development, (i) additional Minisci-type reaction conditions, e.g., variation of oxidation reagents and solvents, including photoredox catalysis and electrochemistry could be explored. [45] (ii) The nature of the alkyl radical precursor could be varied and thus the corresponding alkyl group scope expanded. (iii) The substrate scope could be expanded towards other electron-deficient heterocyclic systems, particularly five-membered heterocycles which are reoccurring motifs in drug-like molecules.
4 Methods

4.1 Literature analysis

A systematic analysis of chemical transformations was carried out to determine the most feasible conditions for reaction miniaturization and parallel screening. Initially, 45 publications covering different Minisci-type alkylation reactions were selected. Most of these methods rely on photo- or electrochemistry. Although it has been demonstrated that these approaches are amenable to HTE [46, 47], carrying out these reaction processes requires specialized equipment that is not readily available in every laboratory. Thus, aiming at broad utilization in medicinal chemistry, publications were screened for a fast, robust and easily adaptable procedure. Sutherland et al. [16] reported a Minisci methodology that fulfilled those criteria. In addition to working without the need for added metals and catalysts, the transformation can be carried out with various alkyl carboxylic acids that do not require any pre-functionalization, allowing for customized templates depending on project needs. Consequently, the reaction data were manually curated and standardized in a simple user-friendly reaction format (SURF, for details, refer to Section SI8). These SURF data were used as literature data set herein. All details of the literature analysis (Section SB3) and the resulting data set in SURF are available as supplementary information.

4.2 Screening plate design and testing

The screening plate was designed around the literature data obtained from Sutherland et al. [16], which showed good yields on average (60%) for a variety of carboxylic acid coupling partners. Aiming at assessing the reactivity of a substrate with a variety of different alkyl groups (rings and chains), a screening plate with 24 different alkyl carboxylic acids was assembled. The carboxylic acids scope from the original publication [16] covering n-alkyl (e.g., h, k, depicted in Figure 2), cyclic alkanes (e.g., e, g) and O-heterocyclic fragments (e.g., m, u) was complemented by sp3-rich N-heterocyclic carboxylic acids with relevance to drug discovery projects (o, p, q, r). The reactions were miniaturized to 0.5 µmol scale, downsizing by a factor of 300 compared to the literature procedure. [16] To achieve this small reaction scale, stock solutions of all components in the reaction solvent (DMSO) were produced. Consequently, the designed screening plate only requires 4.2 mg of starting material (molar mass: 350 Da) to assess 24 different transformations. In comparison, single reactions in [16] were carried out with 52.5 mg of starting material. Using a reference substrate from [16] (Molecule 1, structure depicted in Figure S1 in Section SI4), different oxidant and carboxylic acid ratios (3:10, 6:10, 3:20, 6:20) were tested to identify the more favourable screening condition (higher conversion). Further, the influence of the atmosphere (under air, under nitrogen in a glovebox), and the concentration of the reaction (2, 16 mmol/l) were assessed. Upon determining the highest-yielding reaction parameters, the best-performing condition on the plate (B4, 1 with e) was used as the reference reaction to monitor reproducibility across different plates. The plate layout including all reaction parameters is shown in Figure 2. Further information on the plate testing is provided as supporting information (Section SI5).

4.3 LSF informer library

For the generation of the experimental reaction dataset, the previously published informer library was used as a starting point (see [10] for details). From this collection, three fragments (2-4, Figure S1 in Section SI4 for structures) and five drug molecules (5-9, Figure S1 in Section SI4) were screened. The drug molecule library in [10] was assembled based on clustering of 1174 approved small molecule drugs into eight structurally diverse subsets. As three clusters did not contain any reactive functional groups required for Minisci-type reactions (e.g., electron-deficient heterocycle), only five drug molecules (5-9) were subjected to HTE alkylation screening (see Section 2.4 for details). The screening of the drugs was extended by three fragments (2-4) from [10]. Furthermore, a decoy data set containing 368 unsuccessful reaction examples was generated. The chemical structures of the eight N-hetero-arene substrates (2-9, Figure S1) as well as the 16 decoy substrates (10-25, Figure S2) used to train the machine learning are provided as supporting information (Section SI4).
To assess the performance, i.e., the prediction accuracy, of the developed machine learning model on relevant fragments for applications in medicinal chemistry, a substructure search for heteroaromatic ring systems containing at least one nitrogen atom was carried out in the Roche corporate compound collection. The resulting compounds were retained if (i) there was at least 1 g of powder stock available, and (ii) the structures were not used in any internal project or subject to legal restrictions. This pool of candidates was then clustered using sphere exclusion clustering [48] on ECFP4 fingerprints [43] with a Tanimoto cutoff [49] of 0.6. Based on the clustering results, we manually selected 18 structurally diverse fragments (26-43, Section SI4, Figures S3-S4).

4.4 HTE alkylation screening

Using the 24-well plate design (Figure 2), selected drug molecules and fragments from the LSF informer library (2-9, Section SI4, Figure S1) and a set of relevant building blocks (26-43, Section SI4, Figures S3-S4) were screened. The reaction setup (stock solution, liquid handling) and execution (heating, stirring) in glass vials on a parallel screening plate were conducted in a glovebox under nitrogen. Upon completion of the reactions, the residues were diluted in MeCN/H$_2$O to a defined concentration suitable for LCMS analysis, using a liquid handler. The resulting mixtures were analyzed by LCMS, and the results were subjected to automated reaction data analysis (Section SI6) for the determination of the molecular components. Standardized data output (Section SI7) allowed for direct visualization of the information in TIBCO Spotfire (Somerville, USA). The general screening procedure, including detailed information on the hardware and software utilized, is provided as Supporting Information (Section SI5).

4.5 Scale up reactions

Analysis of the screening results revealed that the drugs Loratadine (7), Nevirapine (8), and 11 fragments (26, 28, 29, 33-35, 37-41) were alkylated with different types of alkyl fragments. From this subset, conditions showing reasonable conversion (>40%, based on UV trace) were subjected to upscaling. Reactions were conducted under nitrogen in a glovebox, in glass reaction vessels with pressure release caps and standard stirring bars. Purification was performed by flash chromatography or reversed-phase high-pressure liquid chromatography (RP-HPLC). Structural elucidation was performed with nuclear magnetic resonance (NMR) spectroscopy and high-resolution mass spectrometry (HRMS). The full analytical results and spectra for all compounds are provided as Supporting Information (Section SI10-SI11).

4.6 Graph neural network architecture

To enable the prediction of reaction outcomes, a GTNN architecture was designed that allows for two distinct and variable molecular graphs (i.e., N-hetero arenes and carboxylic acids) in its input. For both molecular graphs their 3D conformers were calculated using the universal force field (UFF) method [50], and the graph was constructed using nodes represented by atoms and edges defined by all neighbouring atoms within a radius of 4 Å of each atom. Atoms were featureized using embeddings of four atom-level features: (i) twelve atom types (H, C, N, O, F, P, S, Cl, Br, I, Si, Se); (ii) two ring types (True, False); (iii) two aromaticity types (True, False); (iv) four hybridization types (sp$^3$, sp$^2$, sp, s). First, the individual atomic embedding was concatenated and transformed into an initial atomic representation $h^0_i$ via a multi-layer perceptron (MLP). Atomic representations $h^0_i$ were subsequently transformed via three message-passing layers. In each message-passing layer, the atomic representations were transformed via Eq. (1)

$$h^l_{i}+1 = \phi \left( h^l_i, \sum_{j \in N(i)} \psi \left( h^l_j, h^l_j, r_{i,j} \right) \right), \quad (1)$$

where $h^l_i$ is the atomic representation of the $i$-th atom at the $l$-th layer; $j \in N(i)$ is the set of neighboring nodes connected via edges; $r_{i,j}$ the interatomic distance represented in terms of Fourier features, using a sine- and cosine-based encoding; $\psi$ is an MLP transforming node features into message features $m_{ij}$; $m_{ij} = \psi(h^l_i, h^l_j, r_{i,j})$ for 3D graphs, and $m_{ij} = \psi(h^l_i, h^l_j)$ for 2D graphs; $\sum$ denotes the permutation-invariant pooling Operator (i.e., sum) transforming $m_{ij}$ into $m_i$: $m_i = \sum_{j \in N(i)} m_{ij}$; and $\phi$ is an MLP transforming $h^l_i$ and $m_i$ into $h^{l+1}_i$. The resulting atomic features from all layers $[h^{1}_{i}, h^{2}_{i}, h^{3}_{i}]$ were concatenated and transformed via an MLP, resulting in final atomic features. Atomic features were then pooled via a graph multiset transformer (GMT) [51] with four attention heads yielding an overall molecular feature vector.

This procedure was conducted for both input molecular graphs, where no weights were shared between the two GTNN modules except for the initial embedding layers of atom-level representations. The pooled molecular representations were then concatenated to a learned representation of the reaction conditions. This subsequent reaction representation was lastly transformed to the desired output via a final MLP. To assess if incorporation of quantum chemical information can be helpful for model performance, electronic features were obtained from the DelF$^3$Ta software [52] trained on a large set of reference calculations. [53] Further training details can be found in Section SI1.
4.7 Substrate selection

Selection of a diverse and reactive set of N-hetero arenes was based on a Roche-internal library of 3180 advanced heterocyclic building blocks with a molecular weight between 200 and 1000 g/mol. Aiming to check these compounds for potential reactivity in the alkylation reaction, this library was virtually screened with preliminarily trained GTNN models. Each of the \( N = 3180 \) molecules was assigned with an average score value calculated with six independent GTNNs (Section 2.2 for details). Subsequently, agglomerative compound clustering was performed. The molecules were encoded as an \( N \times N \) similarity matrix containing pairwise Jacard similarity values based on ECFP4 molecular fingerprint descriptors. Clustering resulted in eight clusters from which six were used for substrate selection. Three top-scoring compounds were selected for HTE reaction screening for each of the six clusters. This clustering approach was chosen to allow for the selection of chemically diverse reactive substrates.

5 Data and code availability

The SURF-formatted literature, experimental and decay data sets containing 45, 691 and 368 reactions, respectively, and a reference implementation of the geometric deep learning platform based on PyTorch [55] and PyTorch Geometric [56] will be available at https://github.com/atzkenneth/lsfml and https://github.com/ETHmodlab/lsfml after acceptance of the manuscript for publication.

6 Acknowledgements

This research was funded by the Swiss National Science Foundation (SNSF, grant no. 205321_182176). D.B.K. acknowledges funding from the Fonds der Chemischen Industrie (FCI) through a Liebig Fellowship and Roche Basel for funding the PhD position of D.F.N. C.I. acknowledges support from the Scholarship Fund of the Swiss Chemical Industry. We thank Dr. Nicolas Zeidan for helpful discussions.

7 Conflict of interest statement

G.S. declares a potential financial conflict of interest as co-founder of inSili.com LLC, Zurich, and in his role as a scientific consultant to the pharmaceutical industry. D.F.N., A.T.M., C.B., J.W., M.B., U.G. and R.E.M. are full employees of F. Hoffmann-La Roche Ltd.

References


