Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter

Nadine Schneider,*[,†] Daniel M. Lowe,[§] Roger A. Sayle,[§] Michael A. Tarselli,[‡] and Gregory A. Landrum[†,∥]

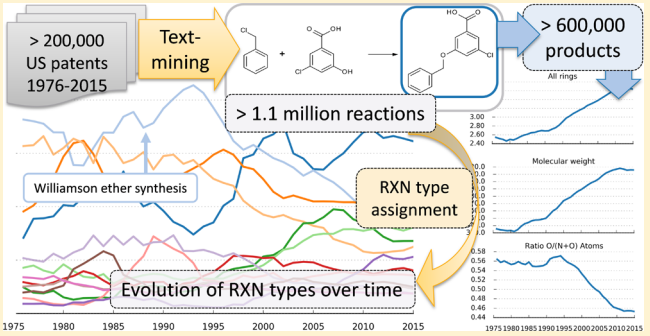
[†]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4002 Basel, Switzerland

[‡]Novartis Institutes for BioMedical Research, 186 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[§]NextMove Software Ltd., Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, U.K.

Supporting Information

ABSTRACT: Multiple recent studies have focused on unraveling the content of the medicinal chemist's toolbox. Here, we present an investigation of chemical reactions and molecules retrieved from U.S. patents over the past 40 years (1976−2015). We used a sophisticated text-mining pipeline to extract 1.15 million unique whole reaction schemes, including reaction roles and yields, from pharmaceutical patents. The reactions were assigned to well-known reaction types such as Wittig olefination or Buchwald−Hartwig amination using an expert system. Analyzing the evolution of reaction types over time, we observe the previously reported bias toward reaction classes like amide bond formations or Suzuki couplings. Our study also shows a steady increase in the number of different reaction types used in pharmaceutical patents but a trend toward lower median yield for some of the reaction classes. Finally, we found that today's typical product molecule is larger, more hydrophobic, and more rigid than 40 years ago.



INTRODUCTION

Which reactions do medicinal chemists use? Here we apply a suite of computational methods to a very large data set to attempt to answer this question. Recent studies to uncover the tools in the medicinal chemist's toolkit have focused on hand-selected and well-curated subsets of publications from relevant journals or on electronic laboratory notebooks (ELNs) from individual companies in the pharmaceutical industry[1−4] (see Table 1). Highlights include a surge in C−C bond formations after the invention of the Suzuki−Miyaura reaction[2,4] and the increasing complexity over time of the compound structures made.[5] Strong toolkit biases exist toward certain reaction types: amide-bond formations, deprotections, or C−C bond formations.[1,2,4] Conventional wisdom suggests that interesting chemistry used in the pharmaceutical industry for building novel structures may either not be published in the scientific literature or will be published with a significant delay;[6−8] influencing factors include a lack of time or motivation to publish and the inevitable legal complexity of publication from within commercial research organizations. Amazingly, a comparison of compound structures extracted from patents and those published in journal articles reveals that only 6% of these molecules actually overlap.[7,9] Cognizant of these challenges, we strive to "complete" the investigation of the medicinal chemist's toolkit, with an analysis of chemical reactions published in granted U.S. patents and U.S. patent applications from 1976 to 2015.

We believe that this rich source of information about the progress of practical medicinal chemistry has been neglected. This is likely due to the difficulty of extracting the data and then accessing it in bulk once it has been extracted. For example, the data in commercial databases like Chemical Abstracts Services (CAS) SciFinder (http://www.cas.org/products/scifinder) or Elsevier's Reaxys (https://www.elsevier.com/solutions/reaxys) have usually been extracted manually, leading to compromises due to the huge amount of published data.[8] These data sources also do not provide the type of bulk access that the motivated data miner needs in order to be able to answer questions "at scale". However, various recent initiatives have started to provide improved public access, for example, the deposition of chemical structures from pre-2000 patents in PubChem[10] (https://pubchem.ncbi.nlm.nih.gov/) in 2011 by IBM[11] or the public launch of the SureChEMBL database[8] (https://www.surechembl.org/search/) in 2014. Advances in chemical text-mining have enabled automatic processing of vast quantities of patent text and chemical structures.[8,12,13] Enhancing these text-mining workflows enables the extraction of entire reaction schemes[12,14] and identification of the reaction roles (reactant, catalysts, reagent, solvent, or product) for each of the molecules in a reaction. With this technology in hand, we compiled a data

Special Issue: Computational Methods for Medicinal Chemistry

Received: January 29, 2016
Published: March 30, 2016

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

Table 1. Overview of Recent Studies Investigating Reactions and Molecules

	This Work	Carey, 2006	Leeson, 2007	Walters, 2011	Roughley, 2011	Brown, 2015
Date Range	1976−2015	2005−2006	1964−2007	1959−2009	2008−2009	1984−2014
Source	Granted U.S. Patents and Applications	In-house synthesis data GSK, AZ, Pfizer	FDA, Prous, GVKBio, Cerep	GVK, ChEMBL, Journals: *J. Med. Chem.*	Beilstein, SciFinder, Journals: *J. Med. Chem.*, *Bioorg. Med. Chem.*, *Bioorg. Med. Chem. Lett.*	Journals: *J. Med. Chem.*, *J. Am. Chem. Soc.*, *Angew. Chem., Int. Ed.*
Documents	126 435	unk	>5000	19 299	139	280
Extraction method	Automated	Manual	Automated	Automated	Automated + Manual	Manual
Sampling over time	Continuous	Snapshot	Continuous	Continuous	Snapshot	Snapshot
Focus	Med-chem patent reactions	Process chem R&D	Oral drugs and development cmpds	Med-chem literature	Med-chem publications from GSK, Pfizer, AZ	Med-chem and natural products publications
Reactions	1.15 million unique	1039			7315	unk
Compounds	628 200	128	>120 000	415 284	3566	unk
Highlights	New frameworks, physicochem. props.	Purchased vs installed chirality, large-scale "practical reactions"	Comparison of various companies' data	Physicochemical properties, complexity, common frameworks	Heterocycle analysis, fluorine	PMI analysis, comparison against natural products literature

set comprising most of the chemical reactions published in U.S. patents over the past 40 years. The automation allowed us to extract 1.3 million unique (defined as the reactants, products, agents, and solvents used) chemical reactions from this enormous data pool. As previously inferred, we could also extract corresponding yields, temperature, time, and preparation and workup steps.[12]

Recently, Brown and Boström[4] manually extracted reactions from 280 different publications in the *Journal of Medicinal Chemistry* in order to be able to systematically classify the reactions using organic chemists' conventions. In this work we have automatically assigned chemical reactions from over 200 000 U.S. patents to over 500 well-known chemical reaction types, such as *amide Schotten−Baumann* or *bromo Suzuki coupling*, using the computational tool NameRxn.[15] The automation enabled us not only to examine a much larger set of reactions compared to other studies[1,2,4] but also to include a more detailed set of reaction types, e.g., different amide bond formations, and to do a continuous sampling over the past 40 years in order to investigate the evolution of reaction types over time. Other computational methods exist that are able to automatically classify reactions based on the reaction center (for an overview see Kraut et al.[16]). The advantage of the method used in this study is that the assigned reaction types are familiar to a medicinal chemist and do not rely on abstract and difficult to interpret patterns. In a recent publication we have provided a predictive model for reaction type assignment that uses the same classification scheme employed here.[17] This model, which covers 50 of the most common reaction types, is freely available for use by others.

As mentioned above, when available, we have also extracted the yields of reaction from the patents. We used this data along with the reaction types to analyze the evolution of the reaction yields over time. To our knowledge, yields have never been investigated at such a scale; nevertheless, it is worth remembering that not all of the stated yields in patents were perfectly optimized. With over 500 000 data points, this analysis provides plausible trends and allows a relative ranking of different reactions by median yield.

Finally, since we have reaction roles assigned to the molecules, we were able to investigate the evolution of the properties of product molecules over time and to relate these findings to the results of our reaction type analysis. We believe that only including the end-points of reactions will reduce the noise in the data introduced by intermediates when compared to other analyses of molecules properties over time.[5,18,19]

In Table 1 we have summarized recent studies examining the evolution of reactions and molecule properties over time. This should help to get an overview of similarities and differences between the studies. Besides these studies concentrating on chemistry designed and executed by humans, Vasilevich and coworkers recently published an analysis of transformations used in natural product chemistry.[20]

A goal of this study is to show how computational tools can help analyze the broad set of chemical reactions found in the patent literature. The study might give some insights into how synthetic medicinal chemistry has evolved over almost a half century. One trend certainly emerges clearly: the enormous increase in the number of patents published in the past 10 years which certainly mirrors the increasing pressure to deliver new chemical entities.

In addition to the comprehensive analysis presented here, an interactive Web page is provided to allow interactive exploration of the data set (see Supporting Information, jm6b00153_si_004.zip).

■ DATA SET AND METHODS

Data Sets. The raw data set used in this study consists of 3 305 795 chemical reactions extracted from over 9 million granted U.S. patents and U.S. patent applications published in the years between 1976 and October 2015 (applications included from 2001 onward). Along with the chemical reactions, solvents, reagents, amounts of the molecules used, yields, temperature, time, and preparation and workup steps were extracted, if available. At least one chemical reaction was found in 122 922 granted U.S. patents and 79 788 U.S. patent applications. We extracted 1 322 045 unique reactions (see Reaction Standardization section for methodology) from these documents, of which 976 472 were found in grants and 345 573 were found in applications. The large number of nonunique reactions is primarily due to the same reaction appearing verbatim in related patent documents from the same inventors. The full set of reactions was extracted from patents from a variety of fields like chemistry, physics, or human necessities. For the analysis presented here we considered only pharmaceutical patents: those having an International Patent

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

Classification (IPC) code of A61K, excluding A61K/8 (cosmetics).[21] This reduces our data set to 126 435 different patents (68 836 grants, 57 599 applications) and 2 967 327 reactions of which 1 150 387 are unique (832 094 from granted patents and 318 293 from applications). About 64% of these unique reactions could be classified to a distinct reaction type using the tool NameRxn[15] (details are described below). In addition, yield values could be extracted for 467 719 of the unique reactions.

A second data set was constructed by extracting all unique reaction products of pharmaceutical patents. Here, we have considered structures that only appear as a product and that were never used as either reactant or agent in other reactions. This resulted in a set of 628 200 unique molecules.

Methods. In the analysis presented here, we have applied a broad set of different computational methods to extract, process, cleanup, classify, and analyze the huge amount of chemical-reaction data. At the beginning we applied text-mining techniques to extract the IUPAC names of chemical structures and make sense of their role in the reaction in order to rebuild the whole reaction scheme.[12,13] The extracted reactions were converted to SMILES format and canonicalized[22] to exclude duplicates from the further analysis. All of these preprocessed reactions were automatically assigned to one of over 500 possible reaction types using a library of predefined and expert-curated chemical transformations (SMIRKS).[15] All the reaction data were organized in a relational database which was used to structure and analyze this huge amount of data. Details of the methods are given in the following section.

Reaction Extraction. The reactions were extracted using an improved version of the text-mining workflow described in ref 12. Experimental paragraphs were identified using a machine learning approach (naïve Bayes classifier[23]) and then grouped with their headings into experimental sections. Within these LeadMine[13] was used to identify chemical entities, physical quantities, and characterization data (e.g., NMR). A version of ChemicalTagger,[24] modified to accept LeadMine's input, then divides the experimental section into phrases, which often can be associated with the reaction action that was performed, e.g., stir, add, yield, dissolve. Within a phrase, quantities such as masses and volumes are associated with the chemical entity that they are describing. Conditions such as durations and temperatures are associated with the phrase.

Chemical entities (IUPAC names, chemical formulas, synonyms) are converted to chemical structures (SMILES[25] and InChI[26−28]) using LeadMine, which employs a combination of manually curated reagent dictionaries, a dictionary derived from ChEMBL,[29] and OPSIN[30] for systematic chemical names. A combination of structural information and textual indications are used to assign compounds the roles of reactant, catalyst, solvent, or product. For example a chemical preceded by the word "in" is likely to be a solvent, a palladium-containing compound is likely to be a catalyst, and a compound in a yield phrase is likely to be the product. At this step more than 24 million potential reactions were identified. The vast majority of these putative reactions (77%) was excluded by a crude sanity check testing for the existence of at least one reactant and product and checking if the product is also contained in the reactants. The Indigo toolkit[31] is then applied to the remaining reactions (5.5 million) to map the reactants' atoms to the appropriate atoms of the product(s). This allows checking that all atoms of the products can be accounted for by the reactants with the assumption being that if this is not the case, then the reaction has been incorrectly text-mined. 35% of the remaining reactions were discarded by that filtering step leaving 3.5 million reactions. Additional heuristics, such as removing reactions with products containing less than 9 heavy atoms, are used to remove other common text-mining mistakes, e.g., the salt of the product being misidentified as the product. Another 6% of the reactions were removed by this step, leaving 3.3 million reactions in our data set prior to filtering to pharmaceutical patents.

Reaction Standardization. The extracted reactions were further processed to identify structural duplicates in the data set using the cheminformatics toolkit RDKit.[32] For the purposes of this analysis, reactions containing exactly the same molecules (reactants, reagents, and products) were identified as structural duplicates. Other characteristics of a reaction, like conditions (temperature, time, etc.) or quantities of the substances, were ignored. First, all nonproduct components, including those labeled as reagents by the text-mining, were considered to be reactants. This was done to remove noise introduced by incorrect reaction role assignment due to ambiguous experimental descriptions in the patent text. Next, canonical SMILES were generated for all components of the reaction.[22] The molecules in SMILES representation were finally sorted lexicographically within the reactant and product sets. Duplicates were removed in the following order: granted patents were preferred over patent application and the earliest publication year was chosen. For the yield analysis, all different yields of a unique reaction were included along with the earliest publication date.

Reaction Classification. Reaction classification was performed using NameRxn (version 2.1.78).[15] NameRxn assigns reactions to one of over 500 distinct reaction types. These are given a position in a derivative of the hierarchy first published by Carey et al.[1] and later refined by Roughley et al.[2] These positions correspond to either named reactions (e.g., Wittig olefination) or where the reaction does not have a trivial name, a description of the reaction (e.g., piperidine synthesis). The hierarchy is formed of three levels: 11 major reaction classes, 80 reaction subclasses, and more than 500 reaction types. For example, *bromo Suzuki coupling* is classified as 3.1.1, where 3.1 is any *Suzuki coupling* and 3 is *C−C bond formation*.

Reaction types are also assigned an identifier in the Royal Society of Chemistry's name reaction ontology (RXNO).[33]

Reactions are classified using a SMIRKS-like pattern syntax[34] that specifies the transformation that must occur for a reaction to be classified. The pattern may also require certain agents to be present, such as a copper catalyst for the Sandmeyer reaction. Additionally when a pattern matches, the correspondence between the atoms of the reactants and products is known, providing the atom-to-atom mapping.

Reaction Database. The reaction data was organized into a relational schema and stored in a PostgreSQL database (PostgreSQL, version 9.4, http://www.postgresql.org). Separate tables were created to store unique reactions, unique molecules, information related to patents like patent number, publication year, or number of preparation steps, and information related to molecules like yield, amounts, or reaction role. The database was used to filter and query the data for the rest of the analyses.

Property Calculation for Products. For the products data set we have calculated several different structural and physicochemical properties using the RDKit toolkit.[32] We included structural properties like number of aromatic rings,

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

number of aliphatic rings, fraction of hydrogen bond donors/acceptors, fraction of sp^3 carbons, or fraction of rotatable bonds and global properties like topological polar surface area (TPSA),[35] calculated log *P* (AlogP),[36] or molecular weight. We investigated the evolution of all these properties over our 40-year time period. The pairwise *T*-test of the SciPy statistics package (ttest_ind)[37] was applied to identify statistically significant differences (p value of <0.05) in the mean of a property from one year to another.

RESULTS AND DISCUSSION

In this section we provide a general overview of the composition of the reaction data set and its evolution over time. An analysis of the number of different reactions types found in the data set is presented along with the evolution of the frequency of major reaction classes and the most common reaction types. The last two subsections cover the development of the yields of the most frequently used reaction types over time and the evolution of the physicochemical properties of the products of these reactions.

General Overview of the Patent Data over Time. The vast majority of the patents and unique reactions in our data set were published after the year 2000 (see Figure 1). We also observed a significant increase in the mean and the median number of unique reactions extracted per patent: for the first 10 years (1976−1985) we found a mean number of 6.2 reactions per patent and a median number of 3 reactions per patent; these numbers increase to 20.1 and 4 in the past 10 years (2006−2015). The trend in numbers of patents and unique reactions is at least partly due to the inclusion of patent applications in our data set after 2001 (Figure 1 green and light green distribution). To demonstrate that the bias in the distribution is not caused by our text-mining approach, we compared our results to a set of 8 505 285 unique molecules from the SureChEMBL database.[8] These molecules were also extracted via a text-mining workflow (a different one from ours), but since only molecules were extracted and not whole reaction schemes, this should be less affected by differences in older patents. The details of the SureChEMBL data set are provided in the Supporting Information (see Figure S1). Comparing the distributions of both sets (the unique reactions and patents of our data set and the unique molecules and patents pulled from SureChEMBL), it is clear that there is indeed a larger fraction of molecules, reactions, and patents published in the past 10 years.

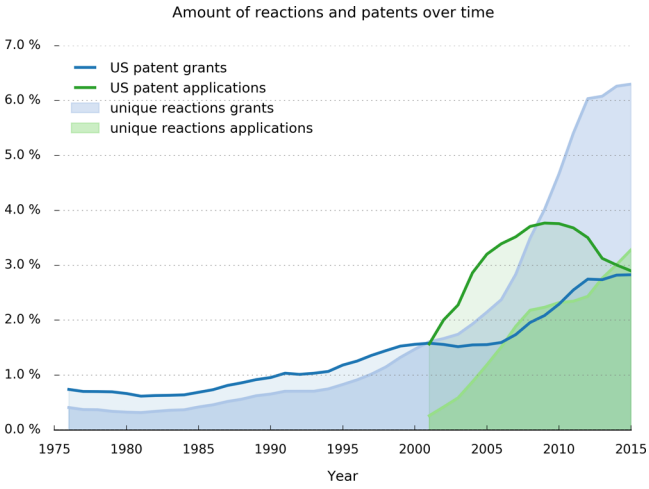


Figure 1. Distribution of U.S. patents and unique reactions extracted from those per year in the time period comprising the years from 1976 until October 2015. Only patents with ICP code A61K[21] (excluding A61K/8) were considered. In light blue and blue, unique reactions of granted U.S. patents and granted U.S. patents are shown, respectively. In light green and green, unique reactions of U.S. patent applications and U.S. patent applications are shown, respectively. Applications only date back to 2001. Note that the values in the plot were smoothed/averaged over 5 years for better visual clarity.

Reaction Type Evolution. In the following we present an analysis of the types of reactions found in our data set. Because the majority of the analysis of this very large data set relies upon an automatic assignment of reaction types, we start with a short overview of the high-level results of that analysis and point out some potential shortcomings. Using NameRxn (version 2.1.78),[15] we were able to assign reaction types to 64% of our data set (738 207 unique reactions); no reaction type could be assigned for the remaining 36% (412 180 unique reactions). Figure 2 top shows the evolution of the fraction classified and unclassified over time. The fraction of unclassified reactions in our data set continuously decreases from 47% in 1976 to 33% in 2008. From 2008 to 2015 the fraction remains rather

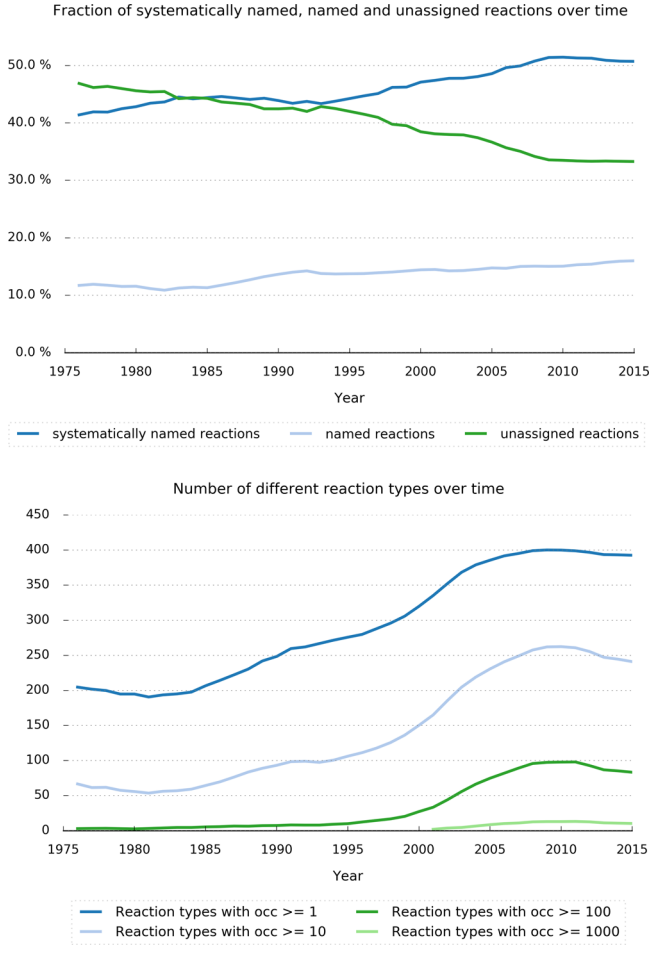


Figure 2. Top: Evolution of classified and unclassified unique reactions over time. Classified reactions are further subdivided in reactions with systematic names and named reactions. Bottom: Evolution of the number of different reaction types showing up in all patents per year. Note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

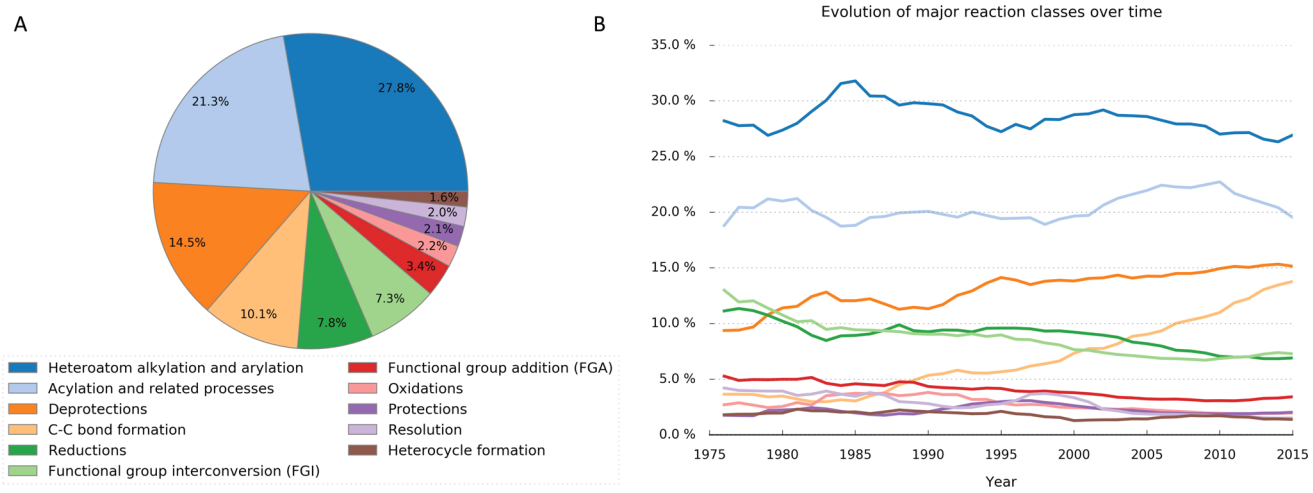


Figure 3. (A) Overall distribution of the major reaction classes of the patent reaction data set. (B) Evolution of the major reaction classes over time. Note that the values were smoothed/averaged over 5 years for better visual clarity.

constant. Possible explanations for the higher fraction of unclassified reactions at the beginning include the higher difficulty of text mining the older patents and a bias in our classification scheme to better coverage of "modern"/recent reaction types. Manual inspection of the unclassified reactions reveals sizable numbers of heterocycle formations and multistep or multisite reactions. The latter reactions include, for example, reactions that form a new bond and simultaneously deprotect another functional group or reactions which introduce two nitro groups at different sites of a reactant. These reactions currently are not classified by NameRxn. The classified reactions were further divided into two subsets, one consisting of systematically named reactions, like *chloro N-arylation*, and the other including "named reactions", like *Wittig olefination*.[38] Here we observed almost four times as many reactions with systematic names as "named reactions". This is reasonably constant over the whole time period (compare Figure 2 top) and could be due to the often rather subtle differences between them, such as the use of special reagents in named reactions.

Moving on to examine the assigned reactions in more detail, a continuing increase in the number of reaction types used can be observed (see Figure 2 bottom). We have applied four different thresholds to confirm this increase in the size of the medicinal chemist's toolbox and the number of tools that are being applied. When considering all reaction types that occur at least once in a year, we observe 217 different types in 1976 and a maximum of 405 types in the 2010 (see Figure 2 bottom blue line). By application of a higher threshold and only considering reaction types that occur at least 10 times in a given year, the same general trend is observed with 78 reaction types in 1976 and a maximum of 274 reaction types in 2010 (see Figure 2 bottom light blue line). Further increasing the threshold to "found at least 100 times within a year" yields six reaction types in 1978 and 110 in 2010. The most restrictive threshold only considers reaction types which show up at least a 1000 times within a year. Here, only a handful of reaction types survived covering a time period from 2001 to 2015. In the year 2010 where a maximum number of 16 different reaction types have been found, the top 5 reaction types are *carboxylic acid + amine reaction*, *chloro N-arylation*, *N-Boc deprotection*, *bromo N-alkylation*, and *amide Schotten–Baumann reaction*. As measured by the number of reaction types used to synthesize the molecules in patents, the size of the medicinal chemist's toolbox has more than doubled over the past 40 years. This increase in the number of reaction types is not completely explained by the increasing number of patents and patent applications published: in 1976, 123 different reaction types accounted for 95% of published classified reactions. This number increases by 27% to 159 in 2010 (see Figure S2 in the Supporting Information). The growth has, however, somewhat stagnated over the past 10 years; we speculate that this can be partially accounted for by the rise of new lead discovery methods (for example, fragment-based drug discovery) which lend themselves to a more limited set of synthetic handles[39] than does traditional medicinal chemistry.

Distribution and Development of the Major Reaction Classes over Time. Going into more detail concerning the distribution of major reaction classes, we have found similar general trends in our analysis of the patent data set as were reported in earlier studies on literature or ELN data.[1−4] The largest class of reactions is made up of *heteroatom alkylations and arylations*, like *chloro N-acetylation* or *Williamson ether synthesis*. This class encompasses 27.8% of the 738 207 unique reactions we have classified (see Figure 3A). The second largest class (21.3% of the data) is *acylation and related processes*, which contains reaction types like *amide Schotten–Baumann* and *carboxylic acid + amine reaction*. The smallest class, with a fraction of only 1.6%, is *heterocycle formation*. This conclusion from our analysis of patents differs from the two studies from Carey[1] and Roughley[2] where heterocycle formation constitutes at least 5% of the reaction data they extracted from in-house process chemistry data and from the medicinal chemistry literature. Brown and Boström[4] found a decrease in the number of applied heterocycle syntheses between 1984 and 2014. They pointed out that purchase of prefunctionalized heterocycles might have superseded their preparation and that recent high throughput chemistry favors amide bond formations over heterocycle synthesis. On the basis of our analysis of the properties of the compounds being synthesized, below, there certainly has *not* been a decrease in the number of heterocycles present in patented molecules. The lack of heterocycle-formation reactions could be a peculiarity of the patent reaction data set but is very likely at least partially explained as an artifact of the SMIRKS-based reaction classification applied by NameRxn. The formation of heterocycles is a reaction class comprising a set of very diverse chemistries that are challenging

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

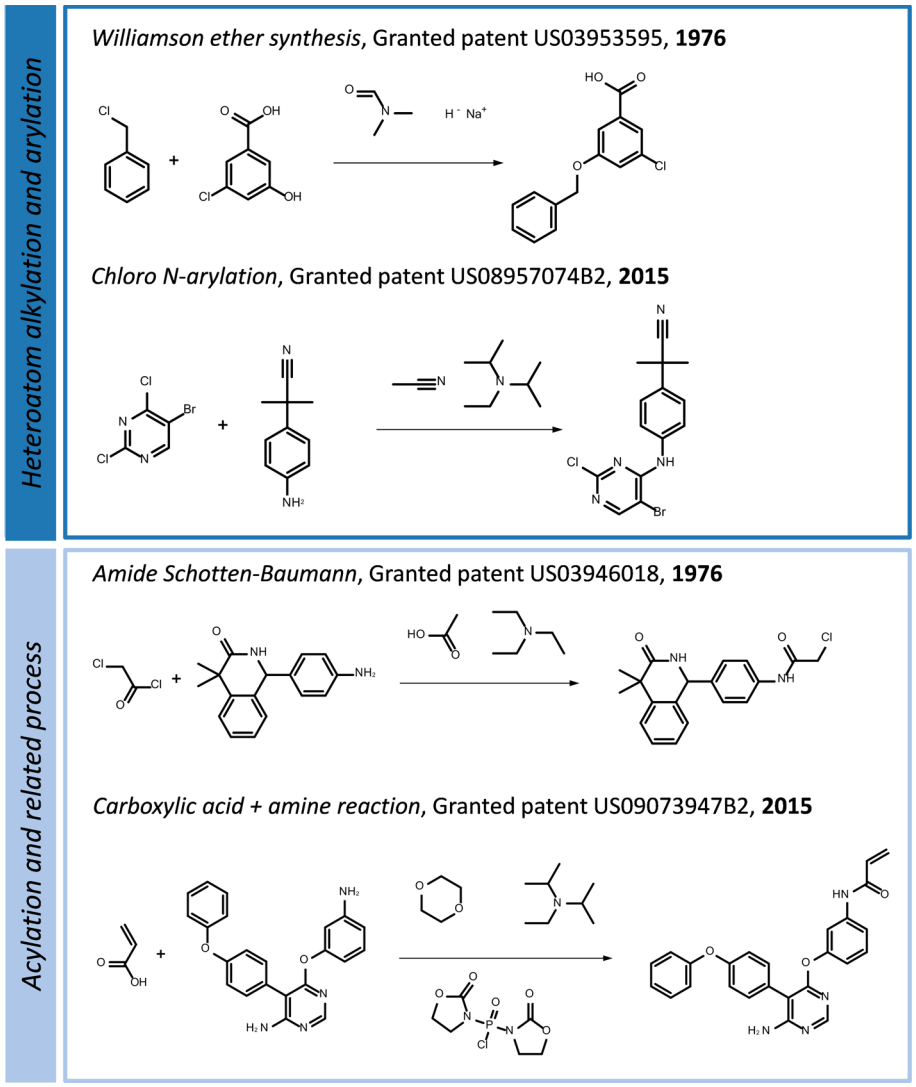


Figure 4. Exemplary reactions for the two largest reaction classes *heteroatom alkylation and arylation* and *acylation and related processes*. For each class the most frequently used reaction types for the years 1976 and 2015 are shown.

to encode as SMIRKS patterns. This is partly confirmed by the fact that we have found a reasonable number of heterocycle formations in our set of unclassified reaction. Currently, the heterocycle formation reaction class is covered with 66 different reaction types in NameRxn and will most likely be extended in future versions. Finally, one has to keep in mind that heterocycle formations themselves could also be assigned to different reaction types, like urea formations, depending on the classification scheme used. This makes it harder to compare absolute numbers for that reaction class across different studies.

We now move on to examine how the use of the major reaction classes has developed over the past 40 years. Figure 3B shows that the two largest classes (*heteroatom alkylations and arylations* and *acylation and related processes*) make up a roughly constant fraction of the data over time and account for a sizable piece of applied chemistry over all 40 years. Exemplary reactions of these two classes from 1976 and 2015 are presented in Figure 4. An especially noteworthy trend in the data is the increase in *C−C bond formations* that started in the mid-80s and that continues until now. While in 1976 this class made up about 4% of the reactions, by 2015 it grew to account for 14% of the reactions reported. Below we will see that this dramatic increase is strongly associated with the invention of the Suzuki coupling. It is however important to note that the large fraction of unclassified reactions in the early years of the data set could also contribute to an undercounting of this reaction class. Another interesting trend is the increase of *deprotections* by about 7%. In 1976 this class was about as common as *reductions* or *functional* group *interconversions*, but by 2015 it was the third most frequent reaction class. A similar result has been found in the analysis of Brown and Boström.[4] Given the transition in industrial medicinal chemistry toward so-called "green chemistry",[40] the use of more protecting groups is somewhat puzzling. We advance two potential explanations: first, the rise in catalytic methods, led by the Suzuki and Buchwald−Hartwig reactions, which paradoxically require that basic heteroatoms that might compete as nucleophiles or poison the catalytic metals must be protected; second, that the sheer increase in available methodology for protecting groups[41] may have influenced the field. It remains to be seen whether the building academic trend toward protecting-group-free synthesis[42] will influence reaction selection going forward.

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

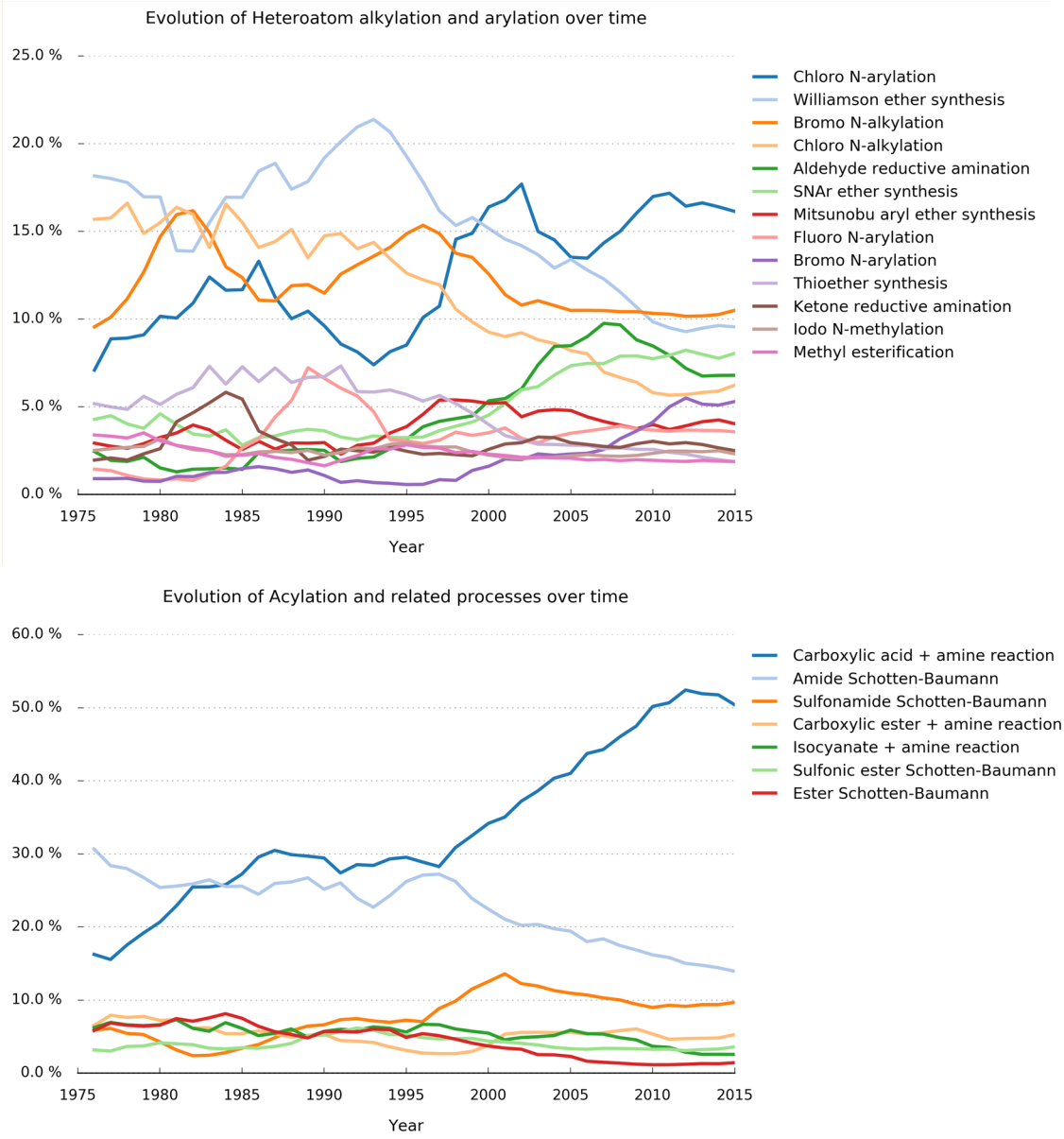


Figure 5. Top: Evolution of largest reaction class *heteroatom alkylations and arylations* over time. Bottom: Evolution of second largest reaction class *acylation and related processes* over time. In both plots reaction types that represent at least 2% of the data were shown. Note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

In the following analysis we examine the compositions of the four largest reaction classes. To make this manageable, we consider only reaction types which cover at least 2% of the reactions within a class. Figure 5 shows the results for the two largest reaction classes. Within the *heteroatom alkylations and arylations* class several interesting trends can be observed (Figure 5 top). While the relative usage of both *Williamson ether synthesis* and *chloro N-alkylation* continuously decreased, dropping by at least 10%, the relative usage of *chloro N-arylation*, *aldehyde reductive amination*, S_NAr *ether synthesis*, and *bromo N-arylation* increased more or less continuously over the past 20 years. What is striking here is the opposing trend over the years of *chloro N-arylation* and *Williamson ether synthesis*. This trend is also mirrored in the decreasing fraction of oxygen atoms compared to nitrogen atoms found in the product molecules (vide infra). Regarding these reactions, while alkylamines are indeed prevalent in pharmaceutically active molecules, their syntheses are complicated by a tendency to overalkylate when reactions such as S_N2 or radical alkylation are used. Reductive amination, really a two-step synthetic process (a condensation with concomitant reduction), compensates for that overreaction and therefore seems to gain growing popularity in recent years.

Much of medicinal chemistry revolves around amide bonds: these linkages occur everywhere in nature: antibiotics, cartilage, peptides, and countless other places. These bonds equally pervade synthetic chemistry: 28 of the top 100 retail drugs contain amide linkages.[43] Brown and Boström's recent review[4] also noted that the most frequently used transformation in medicinal chemistry was amide bond formation, which occurred in 50% of the manuscripts they investigated. After the advent of carbodiimide coupling reagents in the mid-1950s researchers such as Carpino, Li, and El-Faham developed a diverse group of reagents in the 1990s, pyridinium salts,

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

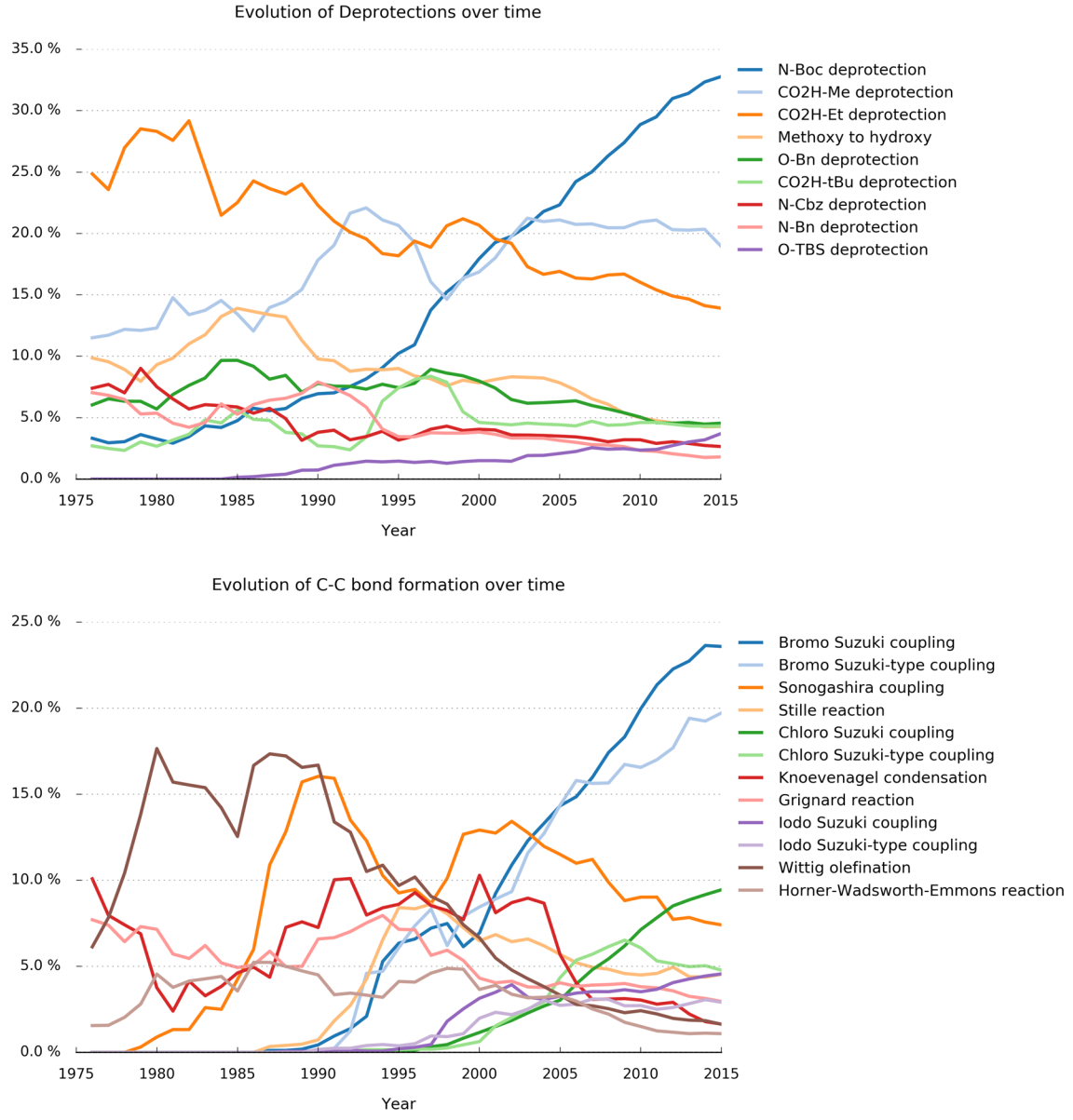


Figure 6. Top: Evolution of third largest reaction class *deprotections* over time. Bottom: Evolution of fourth largest reaction class *C−C bond formations* over time. Note that in *Suzuki-type couplings*, since the palladium catalyst could not always be correctly text-mined from the patent, these subtypes have been introduced. In both plots reaction types that represent at least 2% of the data were shown. Note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

benzotriazoles, uronium salts, among others, transforming amide formation into a reaction undergone primarily between amines and carboxylic acids.[44] The Schotten−Baumann reaction thus lost ground to more modern coupling reagents, an observation confirmed by the heightened prevalence in 2010 of the *carboxylic acid + amine reactions*. We can also observe this shift in the data: the *acylation and related processes* class is dominated by two major reaction types: *amide Schotten−Baumann* and *carboxylic acid + amine reactions*. A strong increase of about 30% of the fraction of the latter is observed (Figure 5 bottom); the *amide Schotten−Baumann* reaction, on the other hand, shows a continuous decrease over the years of about 10%. As noted in other reviews on reaction analysis,[4] the normally harsh reaction conditions of the Schotten−Baumann, combined with the ease of operation of newer coupling reagents, made the *carboxylic acid + amine* essentially a "drop-and-stir." The other reaction types in this class all contribute fractions of below 10%. Only for the *sulfonamide Schotten−Baumann* reaction do we observe a slight increase to a fraction of about 12% at the end of 1990s. This diminished again to less than 10% by the year 2005. Sulfonamide-containing drugs approved during this time period include the PDE5 inhibitors sildenafil (1998) and vardenefil, HIV antiviral darunavir (2006), and COX-2 inhibitor celecoxib (1998).[43]

The two other major reaction classes we have analyzed (*deprotections* and *C−C bond formations*) show distinct trends over the years (see Figure 6). In the *deprotections* class the *N-Boc deprotection* reaction evolves from a rather sparely used reaction type in the 1970s to become the most common deprotection reaction type today (32% in 2015) (Figure 6 top). This rapid growth is likely due to the commercialization of Boc_2O, first prepared at scale in 1977,[45] in the 1980s. *N-Boc deprotection* was also identified as the third most frequently used reaction type in the data set of Brown and Boström.[4] *CO_2H-Me*

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

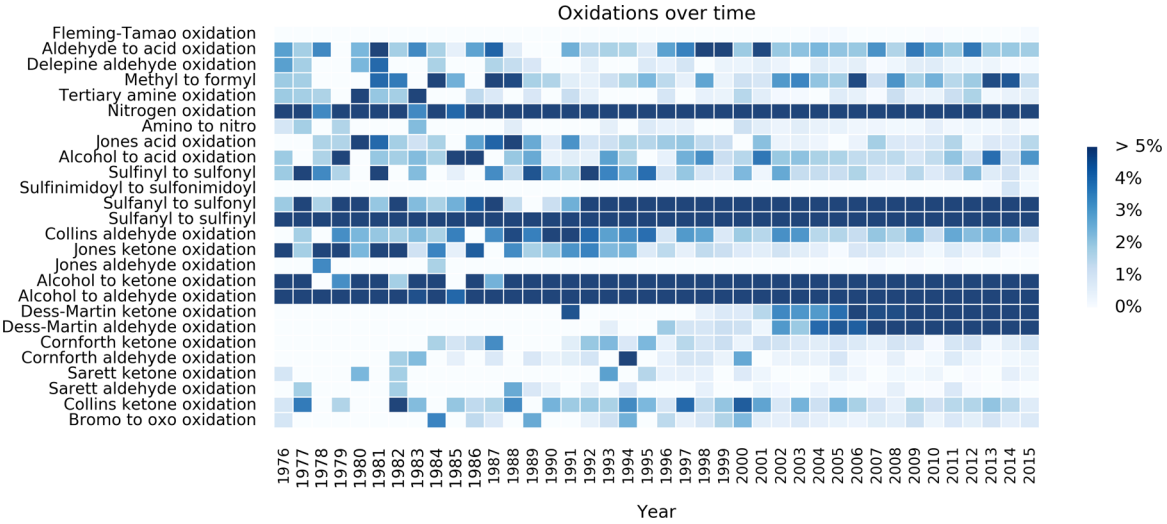


Figure 7. Evolution of *oxidations* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

deprotection and *CO_2H-Et deprotection* are also frequently used deprotection reaction types over the whole time period. These were the predominant types until being displaced by *N-Boc deprotection* around the year 2000. It is worth noting that the strong increase in the use of *N*-Boc deprotections runs in parallel with the remarkable growth of the *carboxylic acid + amine* reactions reaction type discussed above (Figure 5 bottom) and the *bromo Suzuki coupling* discussed below (Figure 6 bottom).

C−C bond formation is probably the most interesting class to observe the development of different reaction types over the years; a number of opposing trends can be seen (Figure 6 bottom). The most noteworthy is the tremendous success of Suzuki couplings since 1989, a few years after their first publication in 1979.[46] From 1989 on there is a steady and steep increase of this reaction type, superseding almost all other *C−C bond formation* reaction types. Others, like *Wittig olefination* or *Sonogashira coupling*, after increasing in popularity until the end of the '80s reverted and decreased with the appearance of the Suzuki couplings. Almost all of the non-Suzuki-coupling reaction types shown at the bottom of Figure 6 have decreased in popularity since the 2000s. Roughley et al. have also discovered in their study that the Suzuki couplings clearly dominate today's C−C bond formation class.[2] While the *bromo Suzuki coupling* reaction types show a steep increase since the beginning of the '90s, the usage of *chloro Suzuki coupling* starts growing about 10 years later. This might be provoked by novel Buchwald-type hindered dialkylarylphosphine ligands, which dramatically increase reactivity of ArCl in Suzuki couplings.[47] Curiously, despite the perceived dominance of Pd-catalyzed couplings, the Stille reaction, which occurs between organotin precursors and aryl halide equivalents, has fallen off since its peak in 2000. Two recent medicinal chemistry trends have likely influenced this change: Trost's atom economy[48] and the more recent focus on green chemistry;[40] organotin reagents, while reactive and selective, presents environmental and toxicity concerns that byproducts of other C−C bond formations do not.

We also observe intriguing trends in other reaction classes (see Figures S3−S8). For example, different heterocycles show crests of popularity in 6-year bursts: pyrroles and epoxides predominate in 1975, yielding to thiazoles in 1983, benzimidazoles in 1988, and tetrazoles in 1994 (Figure S3). This could be due to evolving landscapes in bioisosteres for −COOH groups[49] or due to the evolution of new methodologies; we note that since Sharpless' pivotal "click" paper in 2001,[50] the use of the *azide−alkyne Huisgen cycloaddition* has steadily climbed to about 10% of all heterocycle formations in 2015 (Figure S3). Reductions have mostly held steady, with only *ketone to alcohol reduction* and *nitro to amino* showing major movement over time (Figure S5). Not so with functional group additions; bromination and iodination have become about 10% more frequent relative to other functional group additions since the mid-90s. One might speculate again that this is due to the rising popularity of Suzuki couplings (Figure S8). To allow readers to more flexibly explore the evolution of reaction types over time, we have provided an interactive version of the figures along with this study (see Supporting Information file jm6b00153_si_004.zip).

In order to also track the evolution of rarely used as well as newer and older reaction types, we used a heat-map-like representation (see Figure 7). Here all types within a reaction class that were found at least once are shown. Figure 7 shows that classic chromium oxidants, used in reaction types such as Jones, Cornforth or Collins oxidation, seem to have fallen away and were replaced by an influx of more environmentally friendly and selective periodate oxidations like Dess−Martin oxidation in the early 2000s (Figure 7 and Figure S6). This alternative representation also reveals that the toolkit for C−C bond formations looks much different from 40 years ago: in 1976 the most used reaction types were Wittig olefination, Grignard, Barbier, Friedel−Crafts acylation, Mannich, Horner−Wadsworth−Emmons, aldol and Knoevenagel condensation. All of these depend upon modifying the reactivity of a carbonyl group to form the new C−C bond: most use strong bases (pK_a > 20) or strong Lewis acids ($AlCl_3$, BF_3) as promoters. On the contrary, in 2015 the preferred types are Suzuki, Sonogashira, Stille, and perhaps unexpectedly the Grignard reaction, which has retained its popularity for more than 100 years. The Pd-catalyzed reactions are highly selective, mechanistically well-understood and operationally easy to run (see Figure S11). Figures for the other reaction classes can be found in the Supporting Information (Figures S9−S17).

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

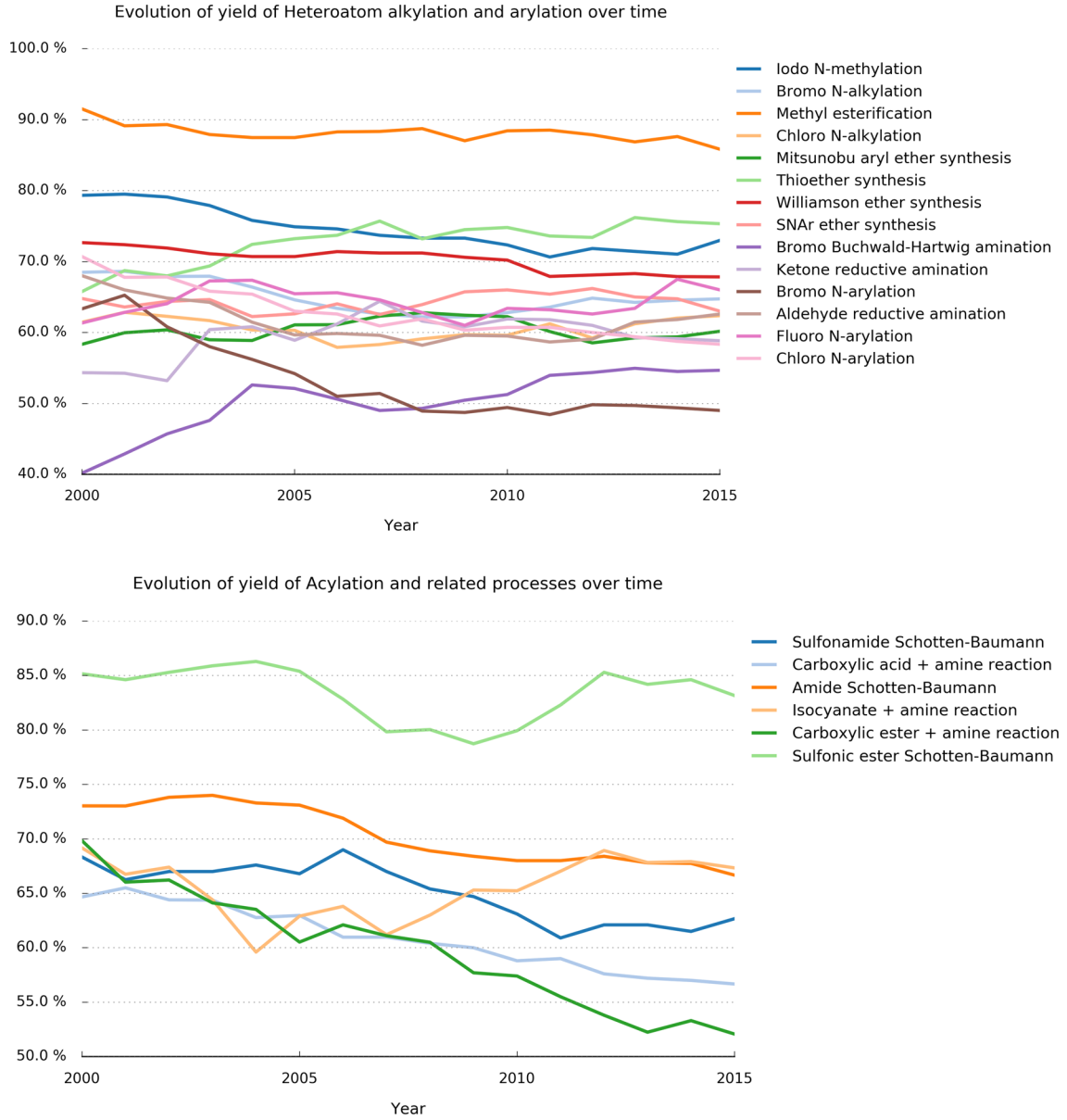


Figure 8. Top: Evolution of the median yield of *heteroatom alkylations and arylations* over time. Bottom: Evolution of the median yield of the *acylation and related processes* over time. In these plots a time period between the year 2000 and the year 2015 (October) was considered. In both plots reaction types that represent at least 2% of the data were shown. Note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

Evolution of Yield over Time. We now consider the evolution of the reaction yield over time in the patent data set. Yield values are available in our data set for over 460 000 unique reactions. In our analysis, when more than one yield value was found for a reaction, all unique instances were kept along with the earliest publication date. We found that although the median yield of 67% in 1976 was slightly larger than the 65% found in 2015, the median yield remains roughly constant over the 40 years examined (Figure S18 top). There is a weak trend in the upper and lower quartiles of the yield: the lower quartile decreases. Given the limitations of the data (the huge bias in the time distribution of the number of reactions/yields and the difficult to assess reliability of yields text-mined from patents), we do not want to overinterpret this. Optimistically, it could be seen as a new culture of including less-than-perfectly optimized results in patents. To overcome the huge bias in the time distribution of the number of reactions and yields, we limit our analysis below to the time period between the year 2000 and the year 2015 (October); this includes 429 692 yield values for 424 596 unique reactions. Over the course of this time period the median yield for all reactions does decrease, going from 72% in 2000 to 65% in 2015 (Figure S18 top). This trend in a general lower median yield might be due to several aspects: First, increasing popularity of automated rapid purification techniques such as flash chromatography[51] or preparative HPLC often results in relatively low recovery even for "pure" samples[52] and enables the isolation of many low-yielding products from mixtures. Second, an emphasis on getting numbers of compounds through screens[53] combined with the miniaturization of the screening techniques themselves means less material is required for a submission, perhaps leading to less effort being dedicated to optimizing reaction conditions.

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

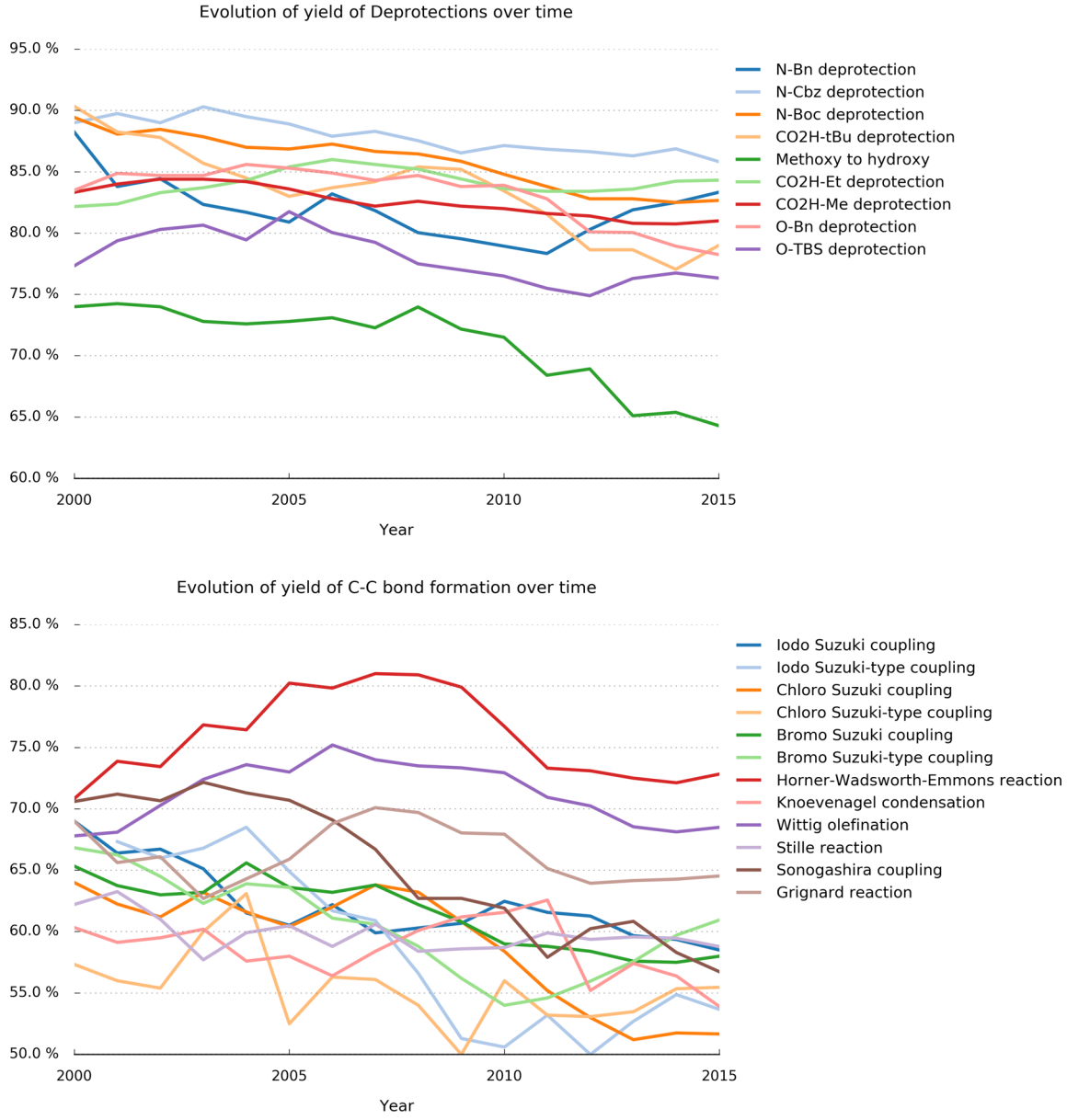


Figure 9. Top: Evolution of the median yield of *deprotections* over time. Bottom: Evolution of the median yield of the *C−C bond formation* over time. Note that in *Suzuki-type couplings* the palladium agent could not correctly be text-mined from the patent; due to this, these subtypes have been introduced. In these plots a time period between the year 2000 and the year 2015 (October) was considered. In both plots reaction types that represent at least 2% of the data were shown. Note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

We also considered the evolution of the median yields of the major reaction classes (see Figure S18 bottom). Although the median yield for most of these classes is more or less constant, the two largest classes (*heteroatom alkylations and arylations* and *acylation and related processes*) along with *C−C bond formation* and *heterocycle formation* do show a significantly lower yield when compared to most of the other classes and those yields go down over the years. Except for *heterocycle formation*, the median yields of these classes have declined by about 7% over the past 16 years.

To shed more light on these results, we once again broke them down to the reaction types of the four largest reaction classes, all of which show a more or less decreasing trend of the median yield in recent years (see Figure S18 bottom). Figure 8 top shows the development of the median yield for the different reactions types in the *heteroatom alkylations and arylation* reaction class. The median yield for most types is constant, with a few showing a slight decrease; only two reaction types (*thioether synthesis* and *bromo Buchwald−Hartwig amination*) have increasing median yields in this time window. As mentioned previously for *bromo Suzuki*, recent developments in ligand architecture, notably by the Hartwig[54] and Buchwald groups, have shown that sterically bulky, electron-rich chelating bisphosphines can drive primary amination and thioethers to completion with as little as 10 ppm of catalyst. These same ligands, such as CyPF-*t*-Bu and "Q-Phos", drive successful reactions with ArCl (vide supra). *Methyl esterification* has the highest median yield in this class, about 88%. This may be unsurprising, given the facility of formation with diazomethane from the corresponding free acid.[55] *Bromo N-arylation* shows both the lowest median yield and the strongest decrease. Note that this is one of the four reaction types in this class which

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

Table 2. Best- and Worst-Performing Reaction Types in Terms of Yield

Major reaction class	Reaction type	Median yield [%]	Occurrence
	Top-10 High Performers Yield		
Functional group interconversion	Carboxylic acid to acid chloride	97	538
Functional group interconversion	Chloro to azido	89	158
Heteroatom arylation and alkylation	Methyl esterification	88	1839
Deprotections	*N*-Cbz deprotection	88	1278
Functional group interconversion	Bromo to azido	88	224
Protections	Aldehyde dioxolane protection	88	143
Protections	O-TIPS protection	87	113
Acylation and related processes	Fischer−Speier esterification	86	1126
Reductions	Nitro to amino	86	9238
Functional group interconversion	Chloro to iodo Finkelstein reaction	86	170
	Top-10 Low Performers Yield		
Heteroatom arylation and alkylation	Bromo Buchwald-Hartwig amination	52	1968
Heteroatom arylation and alkylation	Bromo N-arylation	50	2781
Heteroatom arylation and alkylation	Ullmann condensation	50	243
Heterocycle formation	Quinazolinone synthesis	47	129
Heteroatom arylation and alkylation	Iodo N-arylation	46	744
Heteroatom arylation and alkylation	Chan−Lam ether coupling	46	160
Heteroatom arylation and alkylation	Chloro Buchwald−Hartwig amination	45	1136
Heteroatom arylation and alkylation	Chan−Lam arylamine coupling	45	456
Functional group addition	Fluorination	44	135
Acylation and related processes	Carboxylic acid + sulfonamide reaction	43	840

have shown a significant increase in use in the past 16 years (compare Figure 5 top).

A similar result is observed for the *acylation and related processes* reaction class (Figure 8 bottom). The *carboxylic acid + amine reaction* has become increasingly common since the mid-90s (Figure 5 bottom) while showing a continuous decrease in median yield. Perhaps ironically, these reactions appear to be victims of their own success; as chemists push the reaction into common usage with more challenging substrates, they may be willing to sacrifice gains in yield for operational simplicity. It is far easier to run an array of Suzuki reactions or amide couplings than reoptimize reaction conditions for each substrate. In this class the best performing reaction type is *sulfonic ester Schotten−Baumann* which produces median yields in the range of 80−85%. Unlike other variants of the Schotten−Baumann, this reaction usually produces activated sulfonate esters as synthetic intermediates (less than 0.5% of the product molecules in our data set were made by using that reaction type). Chemists often use the crude sulfonate intermediates, in contrast to the other reaction types that are more amenable to high throughput parallel synthesis and are more likely to be used as late steps and purified by HPLC (see discussion above). Three of the four remaining reaction types analyzed in this class (*amide Schotten−Baumann*, *sulfonamide Schotten−Baumann*, and *carboxylic ester + amine reaction*) all show a noticeable decline in yield in recent years.

We observe a decrease in the median yields for most of the reaction types in the two remaining reaction classes (Figure 9) particularly within the *deprotections* class, where almost all of the reaction types have lower median yields in 2015 compared to 2000 (Figure 9 top). Note, however, that the median yield of this reaction class is in general significantly higher than the three other classes we have analyzed. This seems rational: since a deprotection implies an initial protection step, deprotections must be high-yielding as a reaction class, or chemists would not find their use worthwhile. Finally, in the *C−C bond formation* class it appears that older reaction types like *Horner−Wadsworth−Emmons* reaction, *Wittig olefination*, or *Grignard* reaction outperform the Suzuki couplings in terms of median yield over the whole time period (Figure 9 bottom). While those on average have median yields of 76%, 71%, and 67%, respectively, the Suzuki couplings have median yields between 55% and 62%. This could be explained by the former reactions often being mechanistically significantly simpler and thereby more robust. Suzuki couplings in contrast are influenced by many subtle, and difficult to optimize, factors such as ligand, solvent, counterion, base, electron density of the ring systems, etc., leading to less than optimal yields being reported.

To close the yield analysis, we present the "top-10" high and low performers over the whole 40-year time period in Table 2. For high performers, the median yield is in the range between 86% and 97% and most types were found in the *functional group interconversion* reaction class. Interestingly, the highest-yielding reaction identified was *carboxylic acid to acid chloride* reaction; given this reaction's known propensity to produce water-sensitive, highly reactive intermediates, the corresponding process chemistry and material handling must be improving with time. One aspect of the high yield could be the almost lossless workup of that reaction (evaporate to dryness). In contrast, most of the low performers belong to the *heteroatom arylation and alkylation* reaction class. Here, the median yield ranges between 43% and 52%, substantially lower than the high performers.

Evolution of Reaction Products and Their Properties over Time. With a better understanding of the tools employed, we analyzed reaction products. In order to avoid either reagents or intermediates, we extracted the 628 200 unique compounds from pharmaceutical patents which were only found as products of reactions (i.e., compounds that were never found as either a reactant or reagent in a patent). Since we are interested in analyzing *new* compounds, each product is assigned only to the earliest year in which it appears; grants were preferred over applications. A major difference of this study compared to others[5,18,19] is that we only consider end-points/products of a particular reaction process and do not include intermediates or reagents. We believe that this focus on

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

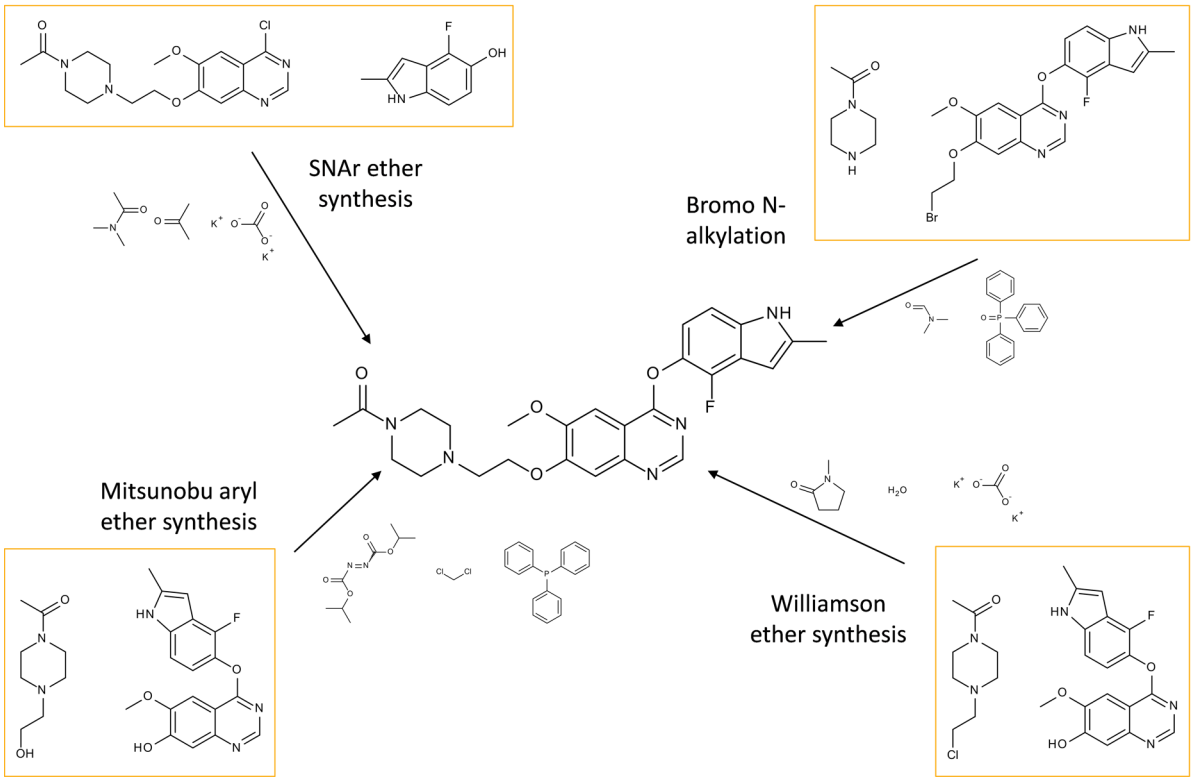


Figure 10. Exemplary product that has been made by four different reaction types. The product molecule and the four different routes stem from U.S. patent US07268230B2.

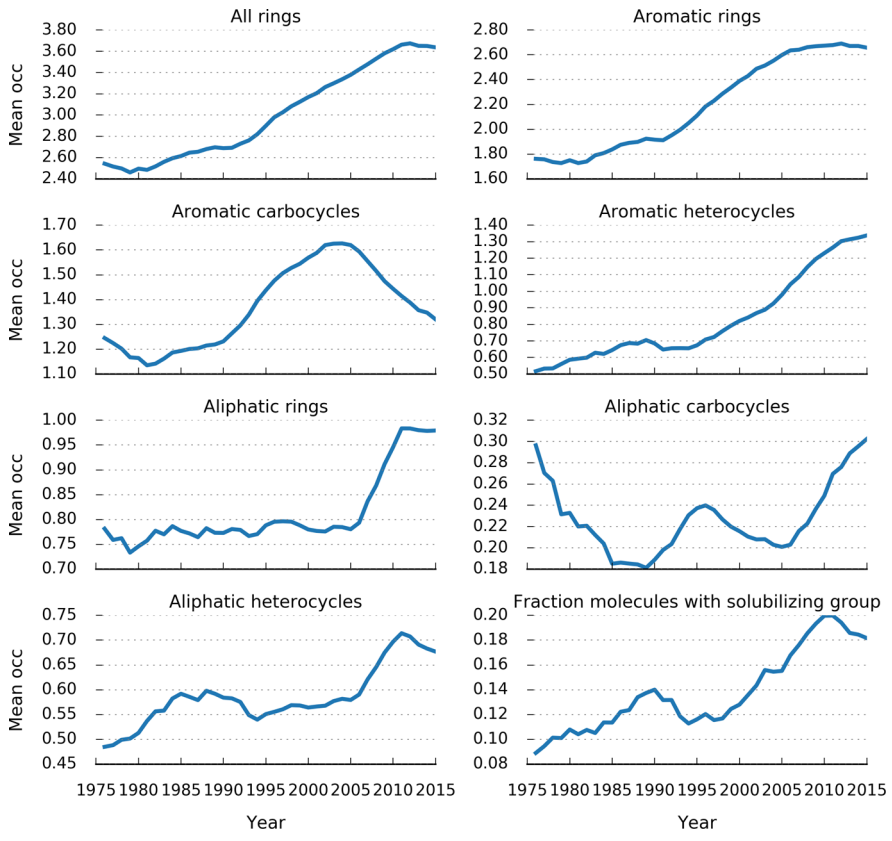


Figure 11. Evolution of the mean number of rings in reaction products over time. In these plots a time period between the year 1976 and the year 2015 (October) was considered. Note that the values in the plots were smoothed/averaged over 5 years for better visual clarity.

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

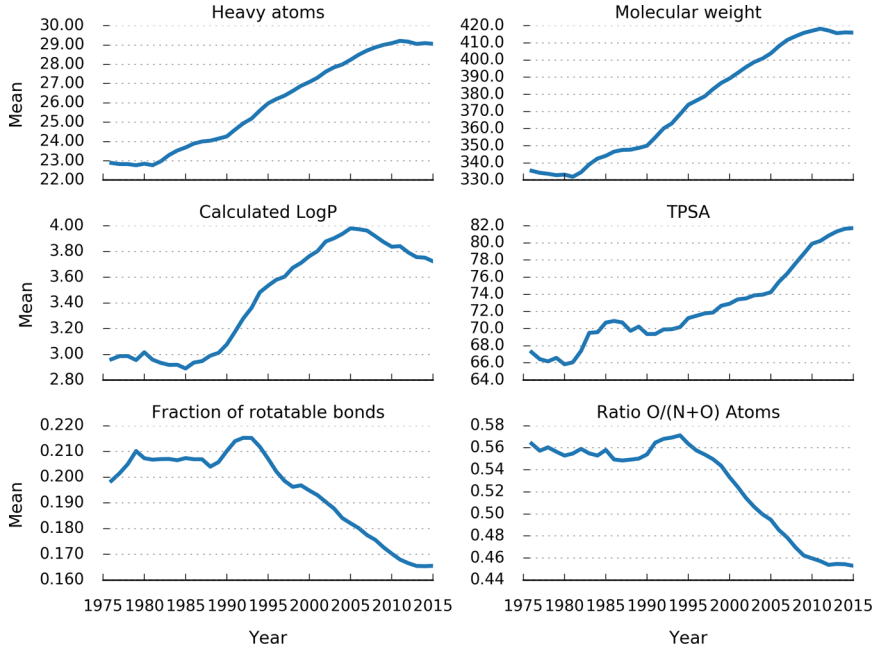


Figure 12. Evolution of the mean number of different structural and global properties of reaction products over time. The *fraction of rotatable bonds* was normalized by the number of bonds. In these plots a time period between the year 1976 and the year 2015 (October) was considered. Note that the values in the plots were smoothed/averaged over 5 years for better visual clarity.

the end result of the reactions, though it ignores exemplified compounds that do not have a producing reaction in the patent, should reduce the noise which is introduced by inclusion of starting materials, intermediates, and protecting groups.

We observe the same trend in the number of unique products over time (Figure S19) as for the unique reactions (Figure 1): most of the products in our set stem from patents published after 2000. After a steep increase starting in 2005 and continuing through 2010 the number of unique products appearing per year in granted patents is stagnant until the end of 2015.

We started with an analysis of the reaction types which were used to produce the unique products in order to determine how many different synthetic approaches to them appeared in patents. After removing the 240 973 products produced by reactions that had not been assigned a type or that were just the product of a resolution reaction, we found that the majority of products (385 428) were made by a single reaction type, 1744 were synthesized using two different reaction types, and only 49 products were produced by three different types of reactions. Continuing, five products were produced by four different reactions (one example shown in Figure 10) and, finally, only one was prepared using five different reaction types. Comparing the statistics of the types of reactions used to synthesize these unique products, results not shown here, we see only minor differences to the overall top 20 reaction types. The types of reactions used as the final step in producing a compound of interest do not differ significantly from the types used in the earlier steps. This differs from the results presented by Brown and Boström;[4] however, that study was done using reactions from the literature.

We next analyzed some properties of the products based on a number of structural descriptors calculated using the RDKit toolkit.[32] First we were interested in the number and types of rings occurring in the products over the years (Figure 11). The overall number of rings increases continuously from a mean of 2.6 in 1976 to 3.5 in 2015. The rate of growth of the number of rings increases in the beginning of the 1990s: in the first 20 years of the plot the mean increases by 0.3, while the second 20 years show an increase of 0.6. The number of aromatic rings shows more or less the same trend: flat growth followed by a steep increase in the early 1990s (Figure 11 top right). On average approximately one ring was added to molecules between 1976 and 2015 and the vast majority of these were aromatic. Several potential explanations could support this trend including the wider adoption and optimization of different reagents for Suzuki-type couplings. The ability to purchase prefunctionalized aromatic synthons from multiple vendors, accompanied by an array of new methods to attach, may also have informed this increase. Further breaking down the aromatic rings into carbocycles and heterocycles (Figure 11 second row), we observe that the initial growth in the number of aromatic rings is dominated by carbocycles while in the 2000s the insertion of heterocycles became more popular. This also parallels the evolution of the Suzuki coupling: since its first report in 1979,[46] this reaction has expanded from relatively limited synthesis of all-carbon biaryls and vinylated species to a wide array of commercially available heteroaryl boronate equivalents.[56] Another contribution would be the surge in leads intended as kinase inhibitors; these molecules tend to be long, polyarylated molecules with a high heteroatom fraction relative to carbon. For aliphatic rings we also observed a smaller but still significant increase around 2005 (Figure 11 third row). Unraveling this further, it becomes clear that most of the increase is due to a growth in the number of aliphatic heterocycles (Figure 11 last row). A simple analysis based on the number of morpholino, piperazine, and pyrrolidine rings in molecules (see Figure 11 last row right) reveals that a large part of this increase is due to incorporation of these common solubilizing groups. Additional trends for saturated rings can be found in the Supporting Information (Figure S20).

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

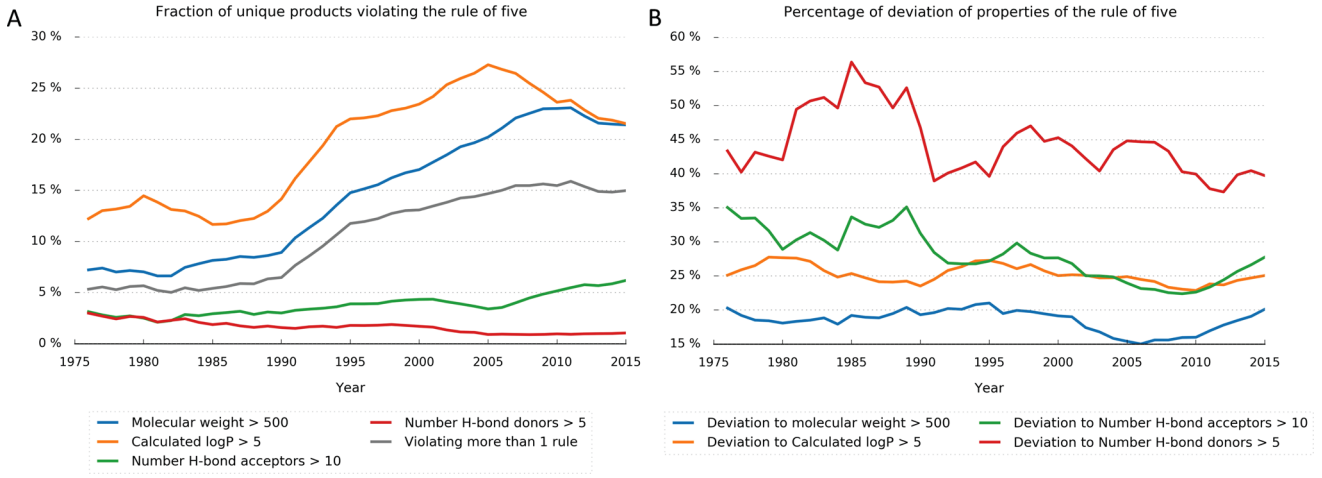


Figure 13. (A) Fraction of reaction products violating Lipinski's rule of five over time. Products that violate more than one rule were counted each time. (B) Percentage of deviation of the four different rule-of-five properties from their thresholds over time. In these plots a time period between the year 1976 and the year 2015 (October) was considered. Note that the values in the plots were smoothed/averaged over 5 years for better visual clarity.

The evolution of other physicochemical properties of the products like the number of heavy atoms, fraction of rotatable bonds, or calculated log *P* is shown in Figure 12. The immediately obvious trend from these plots will not be a surprise to anyone familiar with the literature: the product molecules in pharmaceutical patents have become larger, more lipophilic, and more rigid over the past 40 years. The mean number of heavy atoms has grown from 22.7 to 28.5 (Figure 12 top left), and similar growth rates can be observed for the molecular weight and the calculated log *P* (Figure 12 first and second rows). An average product in 1976 had a molecular weight of about 331 Da and a calculated log *P* of 3.1. In 2015 the average product has MW = 409 Da and calculated log *P* = 3.6, increases of 24% and 16%, respectively. The study of Walters and co-workers also reported an increasing trend for the lipophilicity of druglike molecules over the years, but the growth rate was smaller.[5] The growth in lipophilicity comes despite an increase in polarity: the topological polar surface area (TPSA)[35] of the products has increased by about 15% over the years (Figure 12 second row right). Other properties (fraction of heteroatoms, fraction of sp^3 carbons, fraction of H-bond acceptors, and fraction of H-bond donors) do not show large trends with time (see Supporting Information Figure S21). It is worth noting that decreasing trends in the fraction of rotatable bonds and of sp^3 carbons indicate that more recent products tend to be more rigid (Figure 12 third row, Figure S21). This observation may seem to differ from what is presented in the study of Walters et al.,[5] where the number of rotatable bonds found in molecules was seen to increase. However, these trends are not directly comparable: our value is normalized by the number of bonds to suppress the influence of the molecular size. The decrease in the fraction of oxygen atoms relative to the total number of polar atoms (nitrogen and oxygen) in the product molecules over the years (Figure 12 third row left) could be expected from the trends in the number of reaction types producing amines relative to those producing ethers (compare Figure 5 bottom).

Considering Lipinski's rule of five, published in 1997,[57] our analysis indicates that today's average reaction product falls within these limits. Since the "average product" can easily mask exceptions, we have calculated the percentage of products violating at least one of the rules. We found that 67% of the products obey all rules, 19% violate only one, 12% break two of the rules, 1.6% are above the thresholds of three of the rules, and only 0.07% of the products violate all of Lipinski's rules. Given the amount of time spent discussing them, one could assume that in the years after the publication of Lipinski's rules a trend toward compounds with fewer violations would appear; the patent data, however, tell a different story. Figure 13A shows that the number of products violating the molecular weight threshold of 500 Da increased until 2010 and also the amount of products with more than 10 H-bond acceptors is still growing. For example, 20% of the products published in 2015 violate the 500 Da threshold while in 1976 only 6% had a molecular weight of more than 500 Da. The only parameter that is becoming significantly more "compliant" is the calculated log *P*, which does trend downward over the past 10 years. Interestingly, the log *P* value is probably the most difficult of these properties to directly modulate during the design of a new compound. Since the violation of one of the rules is typically tolerated, we tracked the fraction of molecules breaking more than one rule over time (Figure 13A gray line). Here we also observe an increase of about 10% between 1976 and 2015. This trend does not change after Lipinksi's publication in 1997. Another assumption could be that degree of deviation of the four different properties would shrink, but the percentage of deviation here is also rather constant over the last 20 years (Figure 13B). Finally, it is important to keep in mind that Lipinski's rules were derived from an analysis of orally available drugs and that although our data set was selected from molecular products in pharmaceutical patents, it includes many compounds that have never entered the clinic. For a deeper analysis of clinical candidates that breaks the rule of five, see the publication by Doak and co-workers.[58]

■ CONCLUSIONS

We have used of a set of computational tools to explore the evolution of reaction types over time. Applying those algorithms not only enables the construction and investigation of much larger data sets but also allows a more detailed analysis of reaction types. Furthermore, any bias introduced due to manual selection of data is avoided. These computational

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

approaches have, of course, their own drawbacks and can introduce errors such as incorrect conversion of molecule names to structures, inappropriate reaction-role assignment, or an incomplete set of reaction types. Comparing our results to previous studies which have used manually curated data sets and annotations shows many similar trends in the areas where there is overlap. This helps confirm the reliability of the tools we are using in this study.

Our analysis shows (like others have observed before) that the medicinal chemist's toolkit is quite broad but that the selection of tools actually used is biased toward a manageable set of standard reaction types. This is not particularly surprising given human nature and the increasing automation of processes in the lab. Recent advances in flow chemistry or the increasing interest in methods like DNA-encoded libraries will probably further increase that bias in the future. Still, we have observed a continuously shifting and increasing set of reaction types in use over the past 40 years. Even if not all of them became standard tools, the set of more than 100 different reaction types which have been regularly used in the past years shows the diversity of modern medicinal chemistry. Observing the evolution of the yield over time, we found that once a reaction type has become one of the major types, its yield starts continuously decreasing. This may be due to changes in industrial trends/priorities, more ambitious use of chemistries in places previously not attempted, or increasing use in parallel settings.

The products produced using these tools, an analysis enabled by the computational approaches used here, have become larger, more hydrophobic, and more rigid over the 40-year period. Considering the major reaction types such as Suzuki couplings or the increasing use of reaction types producing amines, we can understand the trends observed in the properties of the product molecules over time. Investigating the druglikeness of these molecules by applying the rule of five indicates that a large number of product molecules appearing in patents violate multiple criteria. While the rule of five seems applicable to oral marketed drugs, our analysis suggests its influence on all patented molecules is lessening with time.

To take a page from the conclusion of Roughley, we might use our data to predict what the medicinal chemists' toolbox will look like over the next decade. Certainly Pd-catalyzed reactions will continue to grow, along with popular amide coupling and reductive amination. "Click" methodology and other heterocycles driven by new [3 + 2] methodology will continue to proliferate. Finally, the old guard seems to have passed: chromium, aluminum, and tin-based chemistries will be mostly abandoned.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.6b00153.

Preparation and analysis of data set; evolution plots (PDF)

Nonaveraged data for all plots and an IPython notebook to generate the plots (ZIP)

All product molecules along with the properties and an IPython notebook to analyze the data (ZIP)

Interactive version of the plots as an html Web page (ZIP)

Molecular formula strings (CSV)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: nadine-1.schneider@novartis.com.

Present Address

[‖]G.A.L: T5 Informatics GmbH, Spalenring 11, Basel 4055, Switzerland.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): R.A.S. and D.M.L. are employees of NextMove Software that markets the NameRxn and the LeadMine tools used in this contribution.

■ ACKNOWLEDGMENTS

N.S. thanks the Novartis Institutes of BioMedical Research Education Office for a Presidential Postdoctoral Fellowship. The authors thank Nikolas Fechner for helpful discussions. The authors thank Minesoft for the use of the PatBase database to determine the set of pharmaceutically relevant patents. We also thank our anonymous reviewers for critical and helpful comments.

■ REFERENCES

(1) Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. Analysis of the Reactions Used for the Preparation of Drug Candidate Molecules. *Org. Biomol. Chem.* **2006**, *4* (12), 2337−2347.

(2) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54* (10), 3451−3479.

(3) Sayle, R. A.; Lowe, D. M.; O'Boyle, N.; Kappler, M.; Pelliccioli, A. P.; Tomkinson, N.; Stoffler, D. Extraction, Analysis, Atom Mapping, Classification and Naming of Reactions from Pharmaceutical ELNs. Presented at Bio-IT World, Conference and Expo, Boston, MA, Apr 9−11, 2013; https://www.nextmovesoftware.com/products/HazELNutPoster.pdf [accessed on December 28, 2015].

(4) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2015**, DOI: 10.1021/acs.jmedchem.5b01409.

(5) Walters, W. P.; Green, J.; Weiss, J. R.; Murcko, M. A. What do Medicinal Chemists Actually Make? A 50-Year Retrospective. *J. Med. Chem.* **2011**, *54* (19), 6405−6416.

(6) Bregonje, M. Patents: a Unique Source for Scientific Technical Information in Chemistry Related Industry? *World Pat. Inf.* **2005**, *27*, 309−315.

(7) Southan, C.; Várkonyi, P.; Boppana, K.; Jagarlapudi, S. A.; Muresan, S. Tracking 20 Years of Compound-to-Target Output from Literature and Patents. *PLoS One* **2013**, *8*, e77142.

(8) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: a Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44*, D1220−D1228.

(9) Southan, C.; Várkonyi, P.; Muresan, S. Quantitative Assessment of the Expanding Complementarity Between Public and Commercial Databases of Bioactive Compounds. *J. Cheminf.* **2009**, *1*, 10.

(10) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(11) IBM Contributes Data to the National Institutes of Health to Speed Drug Discovery and Cancer Research Innovation. http://www.

prnewswire.com/news-releases/ibm-contributes-data-to-the-national-institutes-of-health-to-speed-drug-discovery-and-cancer-research-innovation-135275888.html, 2011 [accessed January 18, 2016].
(12) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Ph.D. Thesis, University of Cambridge, Cambridge, U.K., 2012.
(13) Lowe, D. M.; Sayle, R. A. LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition. *J. Cheminf.* **2015**, *7* (Suppl. 1), S5.
(14) Patent data: http://nextmovesoftware.com/blog/2014/02/27/unleashing-over-a-million-reactions-into-the-wild/ and https://bitbucket.org/dan2097/patent-reaction-extraction/downloads. [accessed December 27, 2015].
(15) *NameRxn*, version 2.1.78; NextMove Software Limited, 2015; https://www.nextmovesoftware.com/namerxn.html [accessed December 23, 2015].
(16) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for Reaction Classification. *J. Chem. Inf. Model.* **2013**, *53* (11), 2884−2895.
(17) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55* (1), 39−53.
(18) Leeson, P. D.; St-Gallay, S. A.; Wenlock, M. C. Impact of Ion Class and Time on Oral Drug Molecular Properties. *MedChemComm* **2011**, *2*, 91−105.
(19) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6* (11), 881−890.
(20) Vasilevich, N. I.; Kombarov, R. V.; Genis, D. V.; Kirpichenok, M. A. Lessons from Natural Products Chemistry Can Offer Novel Approaches for Synthetic Chemistry in Drug Discovery. *J. Med. Chem.* **2012**, *55* (16), 7003−7009.
(21) WIPO: World Intellectual Property Organization. http://www.wipo.int/portal/en/index.html and http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf [accessed December 23, 2015].
(22) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order - An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55* (10), 2111−2120.
(23) Jessop, P. G. Searching for Green Solvents. *Green Chem.* **2011**, *13* (6), 1391−1398.
(24) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *J. Cheminf.* **2011**, *3* (1), 17.
(25) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(26) The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/home/publications/e-resources/inchi.html (accessed May 19, 2015).
(27) Heller, S. R.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. V. InChI - The Worldwide Chemical Structure Identifier Standard. *J. Cheminf.* **2013**, *5*, 7.
(28) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.
(29) Bento, P. A.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: an Update. *Nucleic Acids Res.* **2014**, *42* (D1), D1083−D1090.
(30) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51* (3), 739−753.
(31) Epam Life Sciences, Indigo Toolkit [Online], version 1.2.2beta-r13. http://lifescience.opensource.epam.com/indigo [accessed January 11, 2016].
(32) Landrum, G. A. RDKit: Open-Source Cheminformatics Software [Online], version 2015.09. http://www.rdkit.org, and https://github.com/rdkit/rdkit [accessed December 27, 2015].
(33) RSC's RXNO Ontology. http://www.rsc.org/ontologies/RXNO/index.asp [accessed December 27, 2015].
(34) SMIRKS and SMARTS Language. Daylight Chemical Information Systems. http://daylight.com/dayhtml_tutorials/languages/smirks/index.html and http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html [accessed December 28, 2015].
(35) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714−3717.
(36) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships. I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7* (4), 565−577.
(37) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/, 2001 [accessed December 27, 2015].
(38) Kurti, L.; Czako, B. *Strategic Applications of Named Reactions in Organic Synthesis*; Elsevier Academic: Burlington, MA, 2005
(39) Murray, C. W.; Rees, D. C. Opportunity Knocks: Organic Chemistry for Fragment-Based Drug Discovery. *Angew. Chem., Int. Ed.* **2016**, *55* (2), 488−492.
(40) Anastas, P. T.; Warner, J. C. *Green Chemistry: Theory and Practice*; Oxford University Press: New York, NY, 1998.
(41) Greene, T. W.; Wuts, P. G. M. *Protective Groups in Organic Synthesis*, 1st ed.; Wiley: New York, NY, 1981.
(42) Young, I. S.; Baran, P. S. Protecting-Group Free Synthesis as an Opportunity for Invention. *Nat. Chem.* **2009**, *1*, 193−205.
(43) Top 100 Drugs by Retail Sales. Poster. http://njardarson.lab.arizona.edu/sites/njardarson.lab.arizona.edu/files/Top%20US%20Pharmaceutical%20Products%20of%202013.pdf [accessed January 18, 2016].
(44) Valeur, E.; Bradley, M. Amide Bond Formation: Beyond the Myth of Coupling Reagents. *Chem. Soc. Rev.* **2009**, *38*, 606−631.
(45) Pope, B. M.; Yamamoto, Y.; Tarbell, D. S. Di-*tert*-butyl Dicarbonate. *Org. Synth.* **1977**, *57*, 45.
(46) Miyaura, N.; Suzuki, A. Stereoselective Synthesis of Arylated (E)-alkenes by the Reaction of Alk-1-enylboranes with Aryl Halides in the Presence of Palladium Catalyst. *J. Chem. Soc., Chem. Commun.* **1979**, 866−867.
(47) Martin, R.; Buchwald, S. L. Palladium-Catalyzed Suzuki−Miyaura Cross-Coupling Reactions Employing Dialkylbiaryl Phosphine Ligands. *Acc. Chem. Res.* **2008**, *41* (11), 1461−1473.
(48) Trost, B. M. The Atom Economy- A Search for Synthetic Efficiency. *Science* **1991**, *254*, 1471.
(49) Ballatore, C.; Huryn, D. M.; Smith, A. B., III Carboxylic Acid (Bio)Isosteres in Drug Design. *ChemMedChem* **2013**, *8*, 385.
(50) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem., Int. Ed.* **2001**, *40*, 2004.
(51) Still, W. C.; Kahn, M.; Mitra, A. Rapid Chromatographic Technique for Preparative Separations with Moderate Resolution. *J. Org. Chem.* **1978**, *43*, 2923−2926.
(52) Isbell, J. Changing Requirements of Purification as Drug Discovery Programs Evolve from Hit Discovery. *J. Comb. Chem.* **2008**, *10* (2), 150−157.
(53) Murray, J. B.; Roughley, S. D.; Matassova, N.; Brough, P. A. Off-rate Screening (ORS) by Surface Plasmon Resonance. An Efficient Method to Kinetically Sample Hit to Lead Chemical Space from Unpurified Reaction Products. *J. Med. Chem.* **2014**, *57* (7), 2845−2850.
(54) Hartwig, J. F. Evolution of a Fourth-Generation Catalyst for the Amination and Thioetherification of Aryl Halides. *Acc. Chem. Res.* **2008**, *41* (11), 1534−1544.
(55) Moore, J. A.; Reed, D. E. Diazomethane. *Org. Synth.* **1961**, *41*, 16.

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402

(56) Lennox, A. J. J.; Lloyd-Jones, G. C. Selection of Boron Reagents for Suzuki-Miyaura Coupling. *Chem. Soc. Rev.* **2014**, *43*, 412.
(57) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(58) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space Beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21* (9), 1115−1142.

DOI: 10.1021/acs.jmedchem.6b00153
J. Med. Chem. 2016, 59, 4385−4402