White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/246505/

Version: Accepted Version

Article:
Dobbelaere, Maarten R. and Sneddon, Helen F. (2026) The Stubborn Persistence of Toxic Solvents in Chemical Synthesis. Angewandte Chemie - International Edition. e1677569. ISSN: 1433-7851

https://doi.org/10.1002/anie.1677569

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

RESEARCH ARTICLE

The Stubborn Persistence of Toxic Solvents in Chemical Synthesis

Maarten R. Dobbelaere,[b] and Helen F. Sneddon*[a]

[a] Prof. H. F. Sneddon
Green Chemistry Centre of Excellence, Department of Chemistry,
University of York
Heslington, York, YO10 5DD, UK
E-mail: helen.sneddon@york.ac.uk

[b] Dr. M. R. Dobbelaere
Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering
Ghent University
Technologiepark 125, 9052 Gent, Belgium

Supporting information for this article is given via a link at the end of the document.

Abstract: Solvents dominate the environmental footprint of chemical synthesis, yet their selection remains governed by precedent rather than sustainability. Here, we analyse 1.3 million classified reactions from 40 years of patent data, distinguishing for the first time between reaction, workup, purification, and analytical solvents. Despite decades of awareness of solvent hazards and the publication of influential solvent selection guides, we find no net reduction in the proportion of hazardous solvents used as reaction media. Dichloromethane usage has grown, and rising trifluoroacetic acid consumption intensifies concerns over per- and polyfluoroalkyl substances. Where regulations have removed specific solvents, chemists have defaulted to lateral substitution, replacing one hazardous solvent with another rather than adopting greener alternatives. We identify the reaction classes most dependent on each problematic solvent, providing a data-driven framework for prioritising targeted intervention over blanket regulation. Crucially, these findings expose systemic biases in historical datasets that threaten to turn AI solvent-prediction models into engines for proliferating outdated, hazardous practices rather than advancing sustainable chemistry.

Introduction

The manufacture of active pharmaceutical ingredients (APIs) accounts for more than 1% of global CO_2 emissions.[1] Solvents dominate the carbon footprint of API manufacturing, contributing more than 80% of the process mass intensity (PMI),[2] [3] [4] [5] the mass of materials employed per kg of product. The disproportionate impact of solvents on process sustainability has long been recognised as a major concern,[3] and the fifth principle of green chemistry explicitly encourages the use of safer solvents and auxiliaries.[6] In response, industrial actors and public consortia have introduced solvent guides that classify solvents as "green" (preferred), "amber" (some issues), or "red" (hazardous or problematic), aiming to steer solvent choice towards safer, less environmentally harmful alternatives across the pharmaceutical lifecycle.[7] [8] [9] [10] [11] [12] [13] Solvents are under increasing scrutiny in Europe as part of the assessment framework for 'safe and sustainable by design' chemicals and materials.[14]

So far, reviews have assessed alternatives to chlorinated, dipolar aprotic, and ethereal solvents across synthetic chemistry applications.[15] [16] However, these manually curated analyses are inherently constrained in scope, typically covering no more than a few hundred publications. As a result, a comprehensive picture of general solvent usage over time remains absent.

Recent advances in cheminformatics, fuelled by artificial intelligence (AI), have enabled the automated analysis of very large chemical reaction corpora.[17] These tools can extract reaction components, including solvents, reagents, and catalysts, directly from unstructured patent and literature data, allowing trends to be studied at a scale that could never be achieved by manual curation. Rxn-INSIGHT,[18] [19] an open-source reaction classification algorithm, has classified the entire United States Patent and Trademark Office (USPTO) database of Lowe *et al.*[20] into over 500 named reaction types. This classification provides a unique opportunity to explore solvent use not only across time but also as a function of reaction type, offering a level of granularity that has not been available in previous sustainability assessments. At the same time, the solvent annotations extracted from these large data sources are

WILEY VCH

increasingly used to train predictive models for solvent selection and reaction optimisation,[21] [22] [23] making the accuracy and completeness of the underlying data a concern that extends well beyond historical analysis.

Despite the availability of large-scale reaction databases and automated classification tools, no study has yet leveraged these resources to comprehensively assess how solvent usage in chemical synthesis has evolved over time. Whilst pharmaceutical industry-led solvent guides were first introduced some decades ago,[10] [24] they were compiling known issues with solvents that chemists might have been expected to already be taking into account in their choice of reaction conditions. It remains unknown whether the knowledge of these issues or the introduction of solvent guides has led to a measurable shift away from problematic solvents, or whether historical precedent and practical convenience continued to govern solvent selection. Equally unclear is the extent to which specific reaction types remain locked into hazardous solvents, and where targeted research into alternative reaction media could yield the greatest reduction in environmental impact. This gap has practical consequences, as Beker *et al*. [25] have shown, AI-based condition prediction models largely learn to reproduce historical popularity trends rather than identify optimal choices. If the most frequently used solvents are hazardous, these tools risk perpetuating, rather than correcting, unsustainable solvent selection practices. Addressing these questions requires a systematic, longitudinal analysis that spans reaction media, workup, purification, and analytical solvents, a scope that has not been attempted to date.

Here, we present the first longitudinal study of the nature of industrially relevant solvent use across reaction media, workup, purification, and analysis at scale. From 40 years of USPTO data, we unveil that a small number of – mainly hazardous and problematic – solvents dominate all facets of chemical synthesis. Despite the high citation of various solvent guides, we find no evidence in patents of a shift towards greener alternatives; instead, current regulations primarily drive lateral substitution, where one problematic solvent is replaced by another rather than by a genuinely safer option. Our deep-learning-based reaction classification identifies the types of chemistry most dependent on "red" solvents, pinpointing priority targets for solvent replacement that could substantially improve the sustainability credentials of industrial synthesis routes.

Results and Discussion

A Solvent Plays Many Roles in Chemical Synthesis

Chemical reaction databases generally report either a single solvent or mixture of solvents per reaction entry. Yet, in a typical synthetic chemistry workflow, solvents play many roles. First, reagents and reactants are dissolved in a “reaction solvent”. After the reaction has taken place, quenching by the addition of water or an aqueous solution is typical. Then, an organic solvent is used to separate layers and remove salts from organic components. We refer to this solvent as the “workup solvent”. Further, a “purification solvent” is used for recrystallisation or chromatography, and an “analytical solvent” is used in analysis for HPLC or NMR spectroscopy.

A Handful of Solvents Dominate the Chemistry Laboratory

Fewer than five solvents account for the majority of mentions in each stage of a synthetic procedure. However, the dominant solvents differ sharply between applications (A–D). For reaction media (**Figure 1**A), dimethylformamide (DMF) and tetrahydrofuran (THF) together account for more than a quarter of all solvent mentions, followed by methanol and dichloromethane (CH_2Cl_2). The profile shifts dramatically for workup (**Figure 1**B), where, as one might expect, water alone represents 40.2% of all mentions, followed by ethyl acetate (EtOAc) and CH_2Cl_2 (10.8%). Purification (**Figure 1**C) is dominated by EtOAc and CH_2Cl_2, while analytical (**Figure 1**D), predominantly NMR spectroscopic, characterisation is overwhelmingly carried out in deuterated chloroform ($CDCl_3$) and deuterated dimethyl sulfoxide (DMSO-d_6).

Colour-coding these solvents by the 2016 GSK Solvent Sustainability Guide[9] exposes a clear pattern. Reaction media are dominated by solvents classified as having some or major issues: three of the four most frequently used reaction solvents (DMF, THF, and CH_2Cl_2) carry amber or red classifications. Purification presents a mixed picture, with recommended EtOAc leading, but hazardous CH_2Cl_2 and chloroform ($CHCl_3$) in second and sixth position. Analytical solvents are almost entirely non-preferred solvents, reflecting the limited availability of deuterated alternatives.

Among hazardous reaction solvents, the chlorinated family warrants particular attention. **Figure 1**E shows that CH_2Cl_2 is more than ten times more frequently used than any other chlorinated solvent in 2016, while it was on par with $CHCl_3$ in the 1970s. **Figure 1**F zooms in on other chlorinated solvents, with 1,2-dichloroethane (1,2-DCE), as

the fastest and only climber, due to its growing popularity in reductive amination and acylation reactions (Suppl. Fig. S1). $CHCl_3$ and carbon tetrachloride (CCl_4) saw a strong decline over time, while chlorobenzene (PhCl), 1,2-dichlorobenzene (1,2-DCB), and other chlorinated solvents represent a negligible share. This suggests that, alongside a general growth in this usage, restrictions on one chlorinated solvent may have redirected usage toward another within the same hazard class, rather than driving adoption of structurally different, less hazardous alternatives.

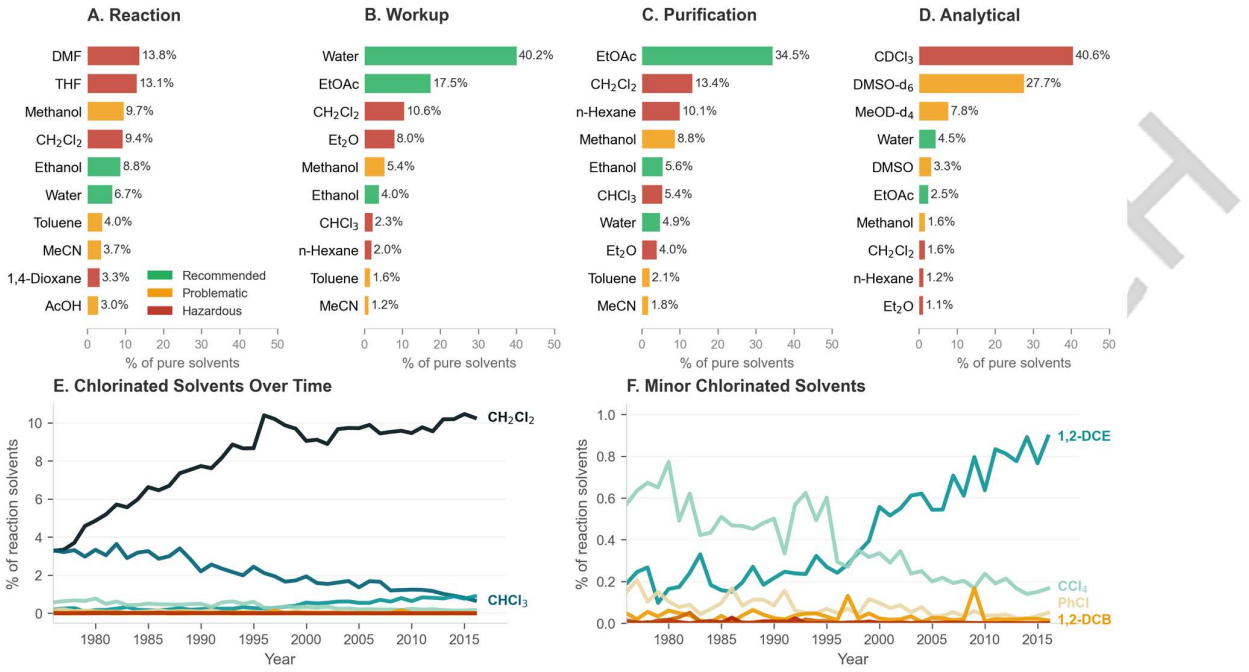


Figure 1. Overview of the most popular solvents. Above are the top 10 solvents for (A) reaction media, (B) workup, (C) purification, (D) analytics, colour-coded using the GSK solvent guide. Deuterated solvents are colour-coded according to their non-deuterated parent compounds. Below are the chlorinated solvents in the period 1976-2016 and their share across reaction solvents. (E) dichloromethane (CH_2Cl_2) and chloroform ($CHCl_3$). (F) Zoom-in on the minor chlorinated solvents 1,2-dichloroethane (1,2-DCE), carbon tetrachloride (CCl_4), chlorobenzene (PhCl), and 1,2-dichlorobenzene (1,2-DCB). Other chlorinated solvents have a negligible share in the database.

Multiple Decades of Solvent Guides Have Not "Greened" Reaction Media

Individual solvent guides differ, however broadly agree on their classification of "green" or "red" solvents.[26] To assess whether the growing emphasis on sustainable solvent selection has translated into measurable change, we tracked the collective usage of solvents over time across all four application stages (**Figure 2**). The results are sobering. Despite the publication of influential solvent selection guides, [7] [8] [9] [10] [11] [12] [13] the proportion of solvents with major issues used as reaction media has remained essentially unchanged over the 40-year study period (**Figure 2**A). The absolute number of solvent mentions has grown in line with the increasing rate of patent submissions,[27] but the relative share of red, amber, and green solvents has barely shifted. Reaction media remain stubbornly dominated by amber- and red-classified solvents.

Workup solvents tell a different story (**Figure 2**B). Here, the green fraction is substantially larger, driven primarily by the use of water in aqueous workup procedures. Purification solvents (**Figure 2**C) show a similar pattern to reaction media, with hazardous solvents maintaining a persistent share, consistent with the extensive usage of dichloromethane, chloroform, and hexane for chromatographic separations. Analytical solvents (**Figure 2**D) are almost entirely red, reflecting the dominance of deuterated chloroform.

The disparity between applications is revealing. It suggests that where a green solvent can serve the same functional role (e.g. ethyl acetate dissolving products in workup) — adoption occurs readily. Where the solvent must meet more stringent requirements, such as dissolving reagents, moderating reactivity, or stabilising intermediates, possibly also involving considerations of boiling point, freezing point, transparency at a given wavelength for reaction monitoring, chemists default to established hazardous solvents regardless of available

guidance. This implies that solvent guides alone, without viable drop-in alternatives validated for specific transformations, are insufficient to drive the transition to greener reaction media.

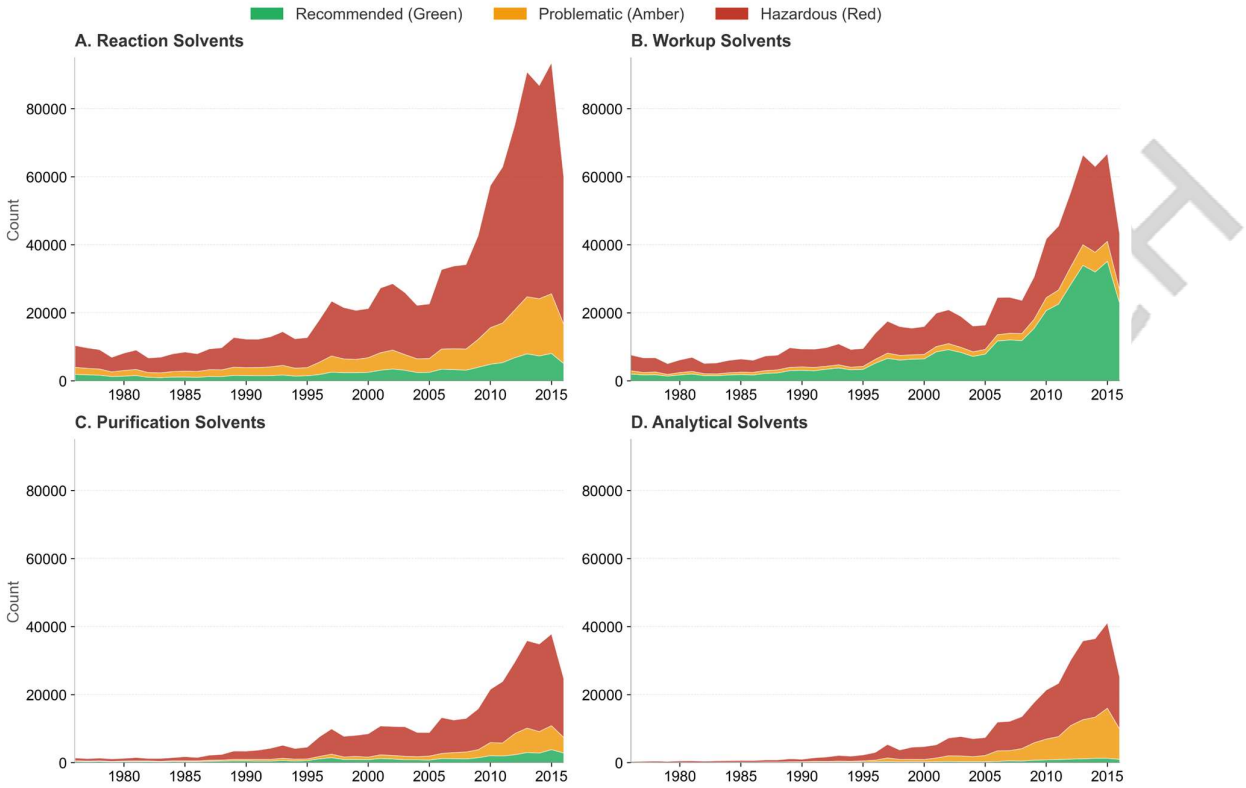


Figure 2: Evolution of absolute solvent usage over time for (A) reaction, (B) workup, (C) purification, and (D) analytical solvents. Solvent mixtures are colour-coded by their least benign constituent using the GSK solvent guide. A total increase in solvents is observed due to an increase in the number of reactions per year. The reporting of purification and analytical solvents is rarer than for reaction and workup solvents.

Differences Can Be Observed Over Time and Geography in Purification Solvent Choice

Mindful that column chromatography uses disproportionately more solvent than reaction or workup solvent, a subset of guides over the years has been dedicated to directing chemists to more sustainable solvent choices in purification.[28] [29] [30] [31] A look at the nature of purification solvents over the years (**Figure 3**) disappointingly shows only evidence of increasing CH_2Cl_2/MeOH use, but does show some evidence that chemists have moved away from hexane, the neurotoxicity of which was first reported back in the 70's.[32] It is converted into the neurotoxic metabolite 2,5-hexanedione by alcohol dehydrogenase[33] which then reacts with lysine residues.[34] Notably, the spacing between the two carbonyl groups is critical for the development of neurotoxicity; diketones lacking this specific spacing do not exhibit neurotoxic effects, and so the move from hexane/EtOAc to heptane/EtOAc is to be welcomed. A geographical analysis of purification solvents (Suppl. Fig. S2-8), confirms the findings elsewhere in this paper that solvent selection is governed by more than the physicochemical requirements of the molecules.

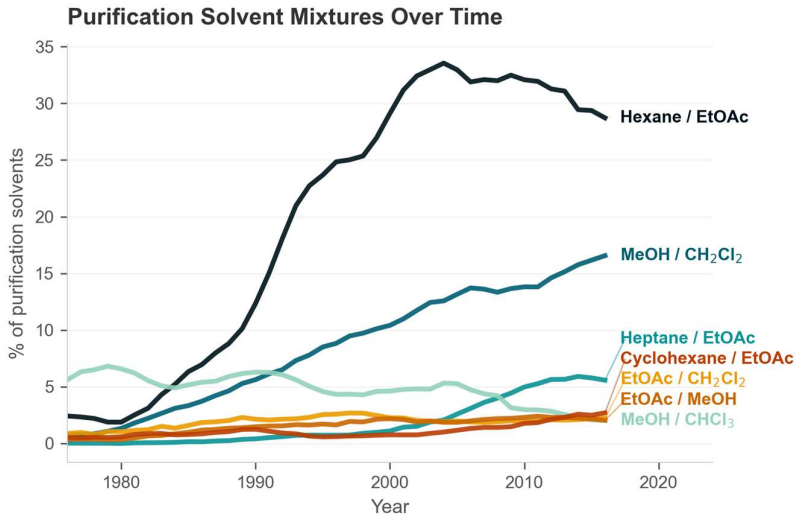


Figure 3: Purification solvent mixtures over time.

Problematic Solvents Do Not Always Concentrate in Identifiable Reaction Classes

We evaluated whether the persistent use of "red" solvents is uniformly distributed across synthetic chemistry by looking at the distribution of each of Rxn-INSIGHT's named reaction classes per solvent. **Figure 4** shows the reaction distribution for the three main chlorinated solvents (CH_2Cl_2, $CHCl_3$, and CCl_4) and three other hazardous solvents (benzene, triflic acid, and diethyl ether). The diversity of reaction types varies strikingly per solvent. CH_2Cl_2 is being used in a broad range of transformations, and no single reaction class accounts for more than 12% of its total usage. This diffuse pattern reflects the reputation of CH_2Cl_2 as a universal solvent and makes it one of the most challenging solvents to replace. There is not a single transformation where a targeted substitution would substantially reduce its overall footprint. Chloroform, while much less frequently used than CH_2Cl_2, also shows a diffuse usage pattern.

CCl_4 shows the opposite extreme pattern in **Figure 4**B with one transformation, Wohl-Ziegler bromination,[35] [36] accounting for 57% of all mentions. The remaining usage is concentrated in a handful of radical and halogenation reactions. This pattern reflects the successful restriction of carbon tetrachloride following the Montreal Protocol,[37] which eliminated its general-purpose use and broadly left only applications where it serves a dual solvent-reagent function, such as the Appel[38] [39] and Reimer-Tiemann[40] reactions.

Benzene (**Figure 4**D) shows intermediate concentration, with aromatic substitutions and Friedel-Crafts reactions prominent — transformations where benzene may function as both solvent and substrate. Trifluoroacetic acid (**Figure 4**E) is dominated by a single application: Boc amine deprotection accounts for a quarter of all mentions. This is significant given the growing concern about environmental accumulation of per- and polyfluoroalkyl substances,[41] [42] and suggests that developing/promoting alternative Boc deprotection protocols would have a disproportionate impact on reducing TFA consumption in synthesis. Diethyl ether (**Figure 4**F) is distributed across diverse reaction types, with Grignard-related additions and reductions featuring prominently.

These treemaps serve a practical purpose: they identify where targeted development of alternative reaction conditions would yield the greatest reduction in hazardous solvent use. For "concentrated" solvents like carbon tetrachloride and trifluoroacetic acid, substitution efforts can focus on a small number of well-defined transformations. For "diffuse" solvents like dichloromethane, a different strategy is needed — one that challenges the default status of CH_2Cl_2 across synthetic chemistry rather than targeting individual reactions.



Figure 4: Most common reaction types per solvent. An overview of the most common named reactions from Rxn-INSIGHT is given for the chlorinated solvents (A) dichloromethane, (B) chloroform, (C) carbon tetrachloride, and for very hazardous solvents (D) benzene, (E) triflic acid, and (F) diethyl ether. The named reactions are colour-coded by their Rxn-INSIGHT superclass.

Banning Hazardous Solvents Drives Lateral Substitution, Not Green Adoption

The treemaps identify which transformations rely most heavily on problematic solvents. We next examined whether solvent usage within these transformations has evolved over time, and specifically whether restrictions or safety reclassifications have driven adoption of greener alternatives or merely redirected usage toward other problematic solvents (**Figure 5**).

The evidence consistently points toward lateral substitution. For the conversion of alcohols to chlorides, benzene usage has disappeared, but dichloromethane usage has increased (**Figure 5**A). For amine protection with tert-butyloxycarbonyl (Boc) the use of trifluoroacetic acid in dichloromethane has increased steadily, pairing a persistent perfluorinated substance with the most prevalent chlorinated reaction solvent. (**Figure 5**B). In urea synthesis (**Figure 5**C), benzene was succeeded by dichloromethane, tetrahydrofuran, 1,4-dioxane, and DMF, all classified as having major issues. In reductive amination (**Figure 5**D), a trend is observed away from water, methanol and ethanol toward CH_2Cl_2 and CH_2Cl_2 containing solvent mixtures. Hazardous solvents continue to predominate across some of the most common transformations across medicinal chemistry,[43] including amide formation (**Figure 5**E). However, the prevalence of alcohols for hydrogenations gives an overall green/amber tone to a consideration of the reduction of nitro groups to amines (**Figure 5**F), where the more flammable, but possibly cheaper, methanol has overtaken ethanol in popularity.

Across all transformations examined, a consistent pattern emerges: when a hazardous solvent is removed, chemists substitute the nearest functional equivalent rather than redesigning the reaction around a greener medium.

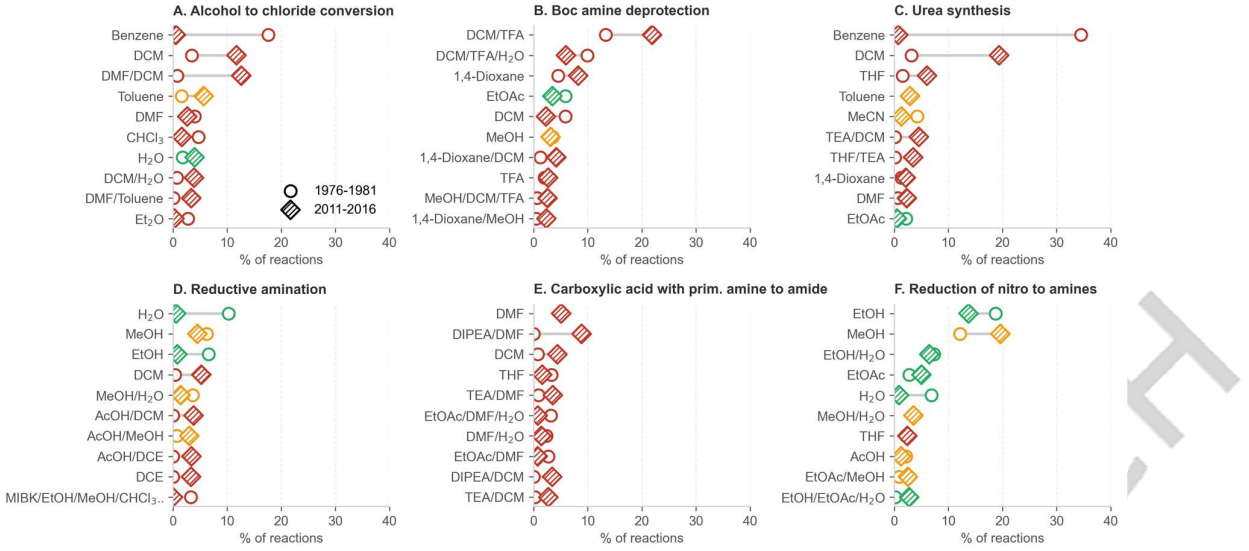


Figure 5: Trends in solvent choice in selected reactions. The popularity of some of the most prevalent solvent choices are shown over time for alcohol to chloride (A), Boc amine deprotection (B), urea synthesis (C), reductive amination (D), amide formation from carboxylic acids and primary amines (E) and the reduction of nitro groups to amines (F).

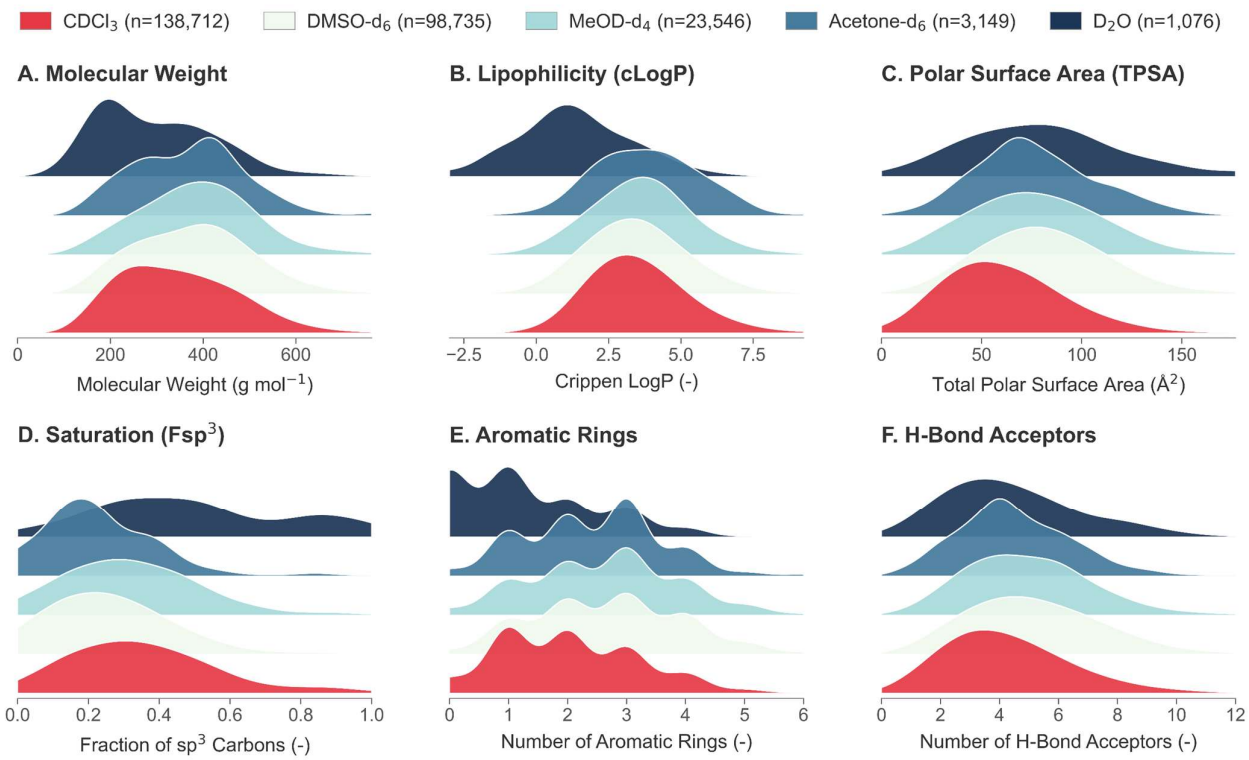


Figure 6: Physicochemical properties of products across NMR solvents. A distribution of six product descriptors is provided for the five most commonly-used NMR solvents.

Precedent, Not Chemistry, Governs Solvent Selection

A natural explanation for the persistence of hazardous solvents would be that they are chemically necessary, that certain substrates or transformations simply require solvents with specific physicochemical properties that greener alternatives cannot provide. Our data do not support this explanation.

For reaction solvents, (Suppl. Fig. S310-311) across six molecular descriptors, molecular weight, lipophilicity (cLogP), polar surface area, saturation (Fsp^3), hydrogen-bond donors, and hydrogen-bond acceptors, the distributions of substrates processed in toluene, THF, ethyl acetate, DCM, DMF, acetonitrile, ethanol, methanol, and water are remarkably similar. If selection were primarily governed by substrate solubility requirements, one would expect distinct substrate profiles for solvents occupying different regions of polarity and hydrogen-bonding space. Instead, whilst acknowledging that solubility is necessary but not sufficient when considering the properties of a reaction solvent, the overlap implies that for most applications, multiple solvents, including greener ones, could fulfil requirements.

This finding is reinforced by the absence of any clear relationship between DCM usage and substrate lipophilicity. Despite DCM's reputation as an effective solvent for hydrophobic substrates, three other solvents, DMF, toluene, and THF, are associated with higher average substrate logP values. The widespread use of DCM, therefore, cannot be rationalised based on a unique solvation advantage.

The same conclusion extends to analytical solvents. NMR solvent selection shows almost no differentiation by substrate properties across $CDCl_3$, DMSO-d_6, MeOD, and deuterated acetone (**Figure 6**). The sole exception is D_2O, whose distinct substrate profile, lower molecular weight, higher polarity, fewer aromatic rings, can be rationalised by the very different properties of water relative to the organic solvents. The more expensive D_2O appears to be reserved for substrates that other NMR solvents genuinely cannot dissolve.

If substrate properties do not explain solvent choice, and if the solvent sustainability profile has not shifted despite decades of guidance, the remaining driver (apart from fundamental compatibility issues with reagents, or reaction conditions such as melting or boiling point, or viscosity or analytical requirements such as an absence of absorbance at a particular wavelength) is literature precedent. Regional variations provide further context to these defaults. The absolute number of reactions in the USPTO database is historically dominated by the United States, Japan, and Europe, with contributions from China and India only representing a substantial share from 2014 onwards (Suppl. Fig. S10-24). Against this demographic baseline, an analysis of the 15 most frequently appearing countries reveals distinct, localized preferences for hazardous reaction solvents, including DMF, NMP, 1,4-dioxane, CCl4, CHCl3, and DCM, that differ noticeably from global averages (Suppl. Fig. S25-30).

Chemical Synthesis Rests on a Fragile Solvent Infrastructure

The preceding analysis reveals a chemical enterprise that is operationally dependent on a small number of hazardous solvents for which *validated* green alternatives often do not exist at scale. This concentration creates systemic vulnerability. Regulatory action on any single dominant solvent would disrupt a disproportionate fraction of synthetic chemistry.

Consider dichloromethane. As the most widely used problematic reaction solvent, it appears across virtually every reaction class in the dataset. Whilst alternatives are known,[15] restrictions on DCM, whether through REACH authorisation requirements or an outright ban, risk leaving the field without a default solvent for a vast number of transformations. Our data show that the historical response to such pressure is lateral substitution, not green adoption, with one possible replacement, chloroform, representing a step backward in hazard profile.

Dimethylformamide presents an equally fragile case. Its classification as a substance of very high concern under REACH has not slowed its adoption. On the contrary, DMF usage has increased steadily, expanding into transformations where water and ethanol previously served (**Figure 4**A–B). The recent discovery that DMF can be a source of nitrosamine contamination in pharmaceutical intermediates, traced to the formation of NDMA from dimethylamine impurities,[44] adds an acute industrial dimension to this regulatory pressure. Yet DMF occupies a specific and difficult-to-replicate functional niche as a high-boiling polar aprotic solvent. The field has not converged on a universally acceptable replacement, and many of the leading candidates each carry their own regulatory or safety concerns.[16]

The vulnerability extends beyond individual solvents to solvent combinations. The growing use of DCM/TFA for Boc deprotection couples the most ubiquitous chlorinated solvent with a persistent perfluorinated substance. Concurrent regulatory pressure is seen on both components, DCM through the 2024 Toxic Substances Control Act and TFA through the proposed ECHA reclassification and broader PFAS restrictions.

This fragility carries tangible consequences. The pharmaceutical industry, which accounts for a substantial share of the patent literature analysed here, operates under increasingly stringent regulatory and environmental scrutiny. It is important to note that our patent-based dataset predominantly captures discovery and early-stage R&D

practices, where synthesis routes are designed for speed and convenience rather than process mass intensity. While hazardous solvents like dichloromethane are actively engineered out of processes during pilot scale-up and manufacturing due to regulatory and safety constraints, their persistent use as defaults in early-stage design creates a substantial bottleneck for process chemists. The treemaps in **Figure 3** and Suppl. Fig S9-S141 provide a map of where this fragility is most acute, and should be read as a call to proactively develop and validate alternative conditions.

Targeted Regulation Can Reshape Solvent Use in Key Transformations

The persistence of solvents with major issues in reaction media does not imply that regulation is ineffective. Rather, our data suggest that the impact of regulation depends critically on how narrowly the affected solvent is used and whether viable alternatives have been identified before the restriction takes effect.

The Montreal Protocol provides the clearest example of successful intervention. Carbon tetrachloride, once employed as a general-purpose solvent, has been effectively eliminated from routine synthesis. Its residual presence in the dataset is confined almost entirely to transformations where it also serves as a reagent. Crucially, alternatives for its former general-purpose applications, including trifluoromethylbenzene[45] and dimethyl carbonate[46] for radical reactions, were already available or emerged rapidly after the restriction. The combination of a clear regulatory mandate with a concentrated usage profile made carbon tetrachloride an amenable target.

The phase-out of benzene tells a more nuanced story. Regulation succeeded in its primary objective: benzene usage dropped sharply following its classification as a human carcinogen in the 1980s. However, the expected compensatory rise in toluene, widely recommended as a safer aromatic alternative, did not materialise. Instead, toluene use as a single solvent declined slightly from the early 1980s to the mid-2010s, and the transformations formerly conducted in benzene migrated predominantly toward dichloromethane, THF, 1,4-dioxane, and DMF (Fig. 4C–D). The ban removed a carcinogen from the laboratory but drove usage of other solvents acknowledged as having issues. This outcome illustrates a general principle: regulation that removes a solvent without simultaneously promoting a green replacement will, in the absence of further intervention, result in lateral substitution.

Taken together, these cases point toward a tiered regulatory strategy informed by the type of data presented in this work. For solvents with concentrated usage profiles, such as carbon tetrachloride and trifluoroacetic acid, targeted restrictions can be effective provided alternatives are developed in parallel. For solvents with diffuse usage, such as dichloromethane and DMF, a different approach is needed. Identifying the specific transformations that contribute most to overall consumption and directing research funding and regulatory attention toward those high-impact targets first. Solvent sustainability will not be achieved by blanket guidance or sweeping bans, but by the systematic, transformation-by-transformation replacement of hazardous defaults with validated greener conditions.

AI-driven Solvent Prediction Models Learn From Mislabelled Data and Potentially Outdated Practice.

A central contribution of this study is the analysis of solvents by their role as reaction, purification, workup, or analytical solvent. This distinction is crucial for advancing both green and AI-driven chemistry research. By coupling reaction solvents to the reaction classes in which they were used, we found that targeted solvent replacements are not straightforward. Various problematic and hazardous solvents, with dichloromethane as the exemplar, are used across diverse reaction classes, necessitating a focus on greener synthesis practices across a broad synthetic toolbox. A growing number of machine learning models aim to predict reaction conditions,[47] including solvents. However, this bias from precedent-based solvent selection will inevitably propagate into model predictions. This phenomenon creates a 'streetlight effect' in chemical synthesis, a pattern already observed in other areas of scientific discovery.[48] Machine learning models optimize for what is highly visible and abundant in historical data, disproportionately recommending well-documented hazardous solvents simply because they are overrepresented in the literature. Consequently, the greener, less-explored chemical space remains in the dark. Since solvent roles are not disambiguated in standard databases, a model trained to predict a reaction solvent may instead have learned to predict the workup or NMR solvent. If restricted solvents, including benzene, remain in training data despite being obsolete in current practice, these tools risk institutionalising outdated practices rather than advancing sustainable chemistry.

This underscores the importance of initiatives such as the Open Reaction Database,[49] which makes recent reaction data available for model development. Our role-disambiguated dataset enables the development of AI models for condition prediction that go beyond reaction solvents to encompass the entire synthetic workflow. Crucially, it highlights the need for both more deliberate experimental solvent selection and more robust data

curation for *in silico* condition recommendations. While this study focuses on patent data, the Rxn-INSIGHT framework is highly adaptable and can be similarly deployed on other standardized databases, such as proprietary electronic lab notebooks (ELNs)[50] or future machine-readable academic journal corpora.

Data and Code Availability Statements

The solvent extraction pipeline and the GSK solvent selection guide are available in the GitHub repository Alkahest at www.github.com/mrodobbe/alkahest and on **10.5281/zenodo.22300099.**

Author Contributions

M.D. and H.F.S. conceived and designed the analysis, analysed the data and wrote the paper.

Competing Interests

The authors declare no competing interests.

Conclusion

Analysis of 1.3 million classified reactions from 40 years of patent data has shown that despite decades of awareness of solvent hazards no net reduction in the proportion of hazardous solvents used as reaction media was seen over the period 1976 to 2016. Dichloromethane usage grew over this time, as did trifluoroacetic acid consumption, particularly relevant given current per- and polyfluoroalkyl substances concerns. Where regulations have removed specific solvents, chemists have defaulted to lateral substitution, replacing one hazardous solvent with another rather than adopting greener alternatives. Addressing this systemic issue requires a fundamental shift in chemical education. We must encourage a critical reading of the literature, emphasizing that historical precedent does not equal chemical necessity. Students should be trained to evaluate solvent selection from first principles, considering factors such as required solvent properties and reagent compatibility, rather than defaulting to the solvents reported in historical protocols. The Rxn-INSIGHT software can provide a data-driven framework for prioritising targeted intervention over blanket regulation.

Supporting Information

The supporting information details methods including details of Reaction Classification, Data Acquisition and Curation, Solvent Extraction and Classification, Solvent Selection Guide, Computational Implementation, Data Analysis and Limitations as well as further analysis of 1,2-Dichloroethane Popularity, Top 20 reaction solvents over time, Geographical Purification Solvent Data, Most Common Reaction Types per Solvent, Trend in Solvent Choice per Reaction Type and Physicochemical Properties of Substrates and Products per Reaction Solvent. The authors have cited additional references within the Supporting Information.[51] [52] [53] [54] [55]

Acknowledgements

Maarten Dobbelaere acknowledges financial support from the Research Foundation – Flanders (FWO Vlaanderen) through postdoctoral fellowship grant 1266226N. This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement n°101057816 (TransPharm). The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO, and the Flemish Government – department WEWIS. Helen Sneddon thanks the EPSRC for funding through Prosperity Partnership EP/X025489/1 undertaken in collaboration with Croda and Prosperity Partnership UKRI605 undertaken in collaboration with Synthomer.

Keywords: Green Chemistry • Prediction • Regulation • Solvent • Sustainable Chemistry

References

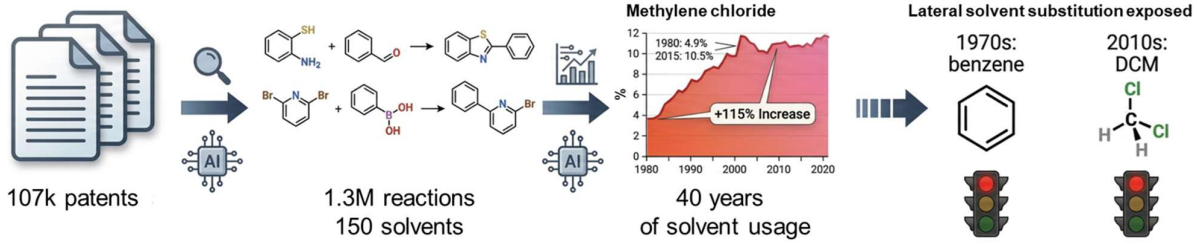
[1] J. S. Karliner, S.; Boyd, R.; Ashby, B.; Steele, K., *Healthcare's climate footprint*, **2019**.

[2] J. Krueger, A. P. Dieskau, J. Hassfeld, J. Gries, O. Block, H. Weinmann, D. Kaufmann, S. Hildbrand, V. Kraft, R. Moeckel, J. R. Dehli, U. Scholz, C. F. Nising, *Chemical Process Development in the Pharmaceutical Industry in Europe—Insights and Perspectives from Industry Scientists Angewandte Chemie International Edition* **2025**, *64*, e202420719.

[3] D. J. C. Constable, C. Jimenez-Gonzalez, R. K. Henderson, *Perspective on Solvent Use in the Pharmaceutical Industry Organic Process Research & Development* **2007**, *11*, 133–137.
[4] C. Jiménez-González, A. D. Curzons, D. J. C. Constable, V. L. Cunningham, *Cradle-to-gate life cycle inventory and assessment of pharmaceutical compounds The International Journal of Life Cycle Assessment* **2004**, *9*, 114–121.
[5] I. Kekessie, K. Wegner, I. Martinez, M. E. Kopach, T. D. White, J. K. Tom, M. N. Kenworthy, F. Gallou, J. Lopez, S. G. Koenig, P. R. Payne, S. Eissler, B. Arumugam, C. Li, S. Mukherjee, A. Isidro-Llobet, O. Ludemann-Hombourger, P. Richardson, J. Kittelmann, D. Sejer Pedersen, L. J. van den Bos, *Process Mass Intensity (PMI): A Holistic Analysis of Current Peptide Manufacturing Processes Informs Sustainability in Peptide Synthesis The Journal of Organic Chemistry* **2024**, *89*, 4261–4282.
[6] P. T. Anastas, J. C. Warner, *Green Chemistry Theory and Practice*, Oxford University Press, **1998**.
[7] D. Prat, A. Wells, J. Hayler, H. Sneddon, C. R. McElroy, S. Abou-Shehada, P. J. Dunn, *CHEM21 selection guide of classical- and less classical-solvents Green Chemistry* **2016**, *18*, 288–296.
[8] D. Prat, A. Wells, J. Hayler, H. Sneddon, C. R. McElroy, S. Abou-Shehada, P. J. Dunn, *Correction: CHEM21 selection guide of classical- and less classical-solvents Green Chemistry* **2015**, *17*, 4848–4848.
[9] C. M. Alder, J. D. Hayler, R. K. Henderson, A. M. Redman, L. Shukla, L. E. Shuster, H. F. Sneddon, *Updating and further expanding GSK's solvent sustainability guide Green Chemistry* **2016**, *18*, 3879–3890.
[10] K. Alfonsi, J. Colberg, P. J. Dunn, T. Fevig, S. Jennings, T. A. Johnson, H. P. Kleine, C. Knight, M. A. Nagy, D. A. Perry, M. Stefaniak, *Green chemistry tools to influence a medicinal chemistry and research chemistry based organisation Green Chemistry* **2008**, *10*, 31–36.
[11] D. Prat, O. Pardigon, H.-W. Flemming, S. Letestu, V. Ducandas, P. Isnard, E. Guntrum, T. Senac, S. Ruisseau, P. Cruciani, P. Hosek, *Sanofi's Solvent Selection Guide: A Step Toward More Sustainable Processes Organic Process Research & Development* **2013**, *17*, 1517–1525.
[12] L. J. Diorazio, D. R. J. Hose, N. K. Adlington, *Toward a More Holistic Framework for Solvent Selection Organic Process Research & Development* **2016**, *20*, 760–773.
[13] P. M. Piccione, J. Baumeister, T. Salvesen, C. Grosjean, Y. Flores, E. Groelly, V. Murudi, A. Shyadligeri, O. Lobanova, C. Lothschütz, *Solvent Selection Methods and Tool Organic Process Research & Development* **2019**, *23*, 998–1016.
[14] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022H2510.
[15] A. Jordan, P. Stoy, H. F. Sneddon, *Chlorinated Solvents: Their Advantages, Disadvantages, and Alternatives in Organic and Medicinal Chemistry Chemical Reviews* **2021**, *121*, 1582–1622.
[16] A. Jordan, C. G. J. Hall, L. R. Thorp, H. F. Sneddon, *Replacement of Less-Preferred Dipolar Aprotic and Ethereal Solvents in Synthetic Organic Chemistry with More Sustainable Alternatives Chemical Reviews* **2022**, *122*, 6749–6794.
[17] P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf, T. Laino, *Machine intelligence for chemical reaction space WIREs Computational Molecular Science* **2022**, *12*, e1604.
[18] M. R. Dobbelaere, I. Lengyel, C. V. Stevens, K. M. Van Geem, *Rxn-INSIGHT: fast chemical reaction analysis using bond-electron matrices Journal of Cheminformatics* **2024**, *16*, 37.
[19] M. R. Dobbelaere, I. Lengyel, C. V. Stevens, K. M. V. Geem, *USPTO Dataset for: Fast Chemical Reaction Condition Suggestion via Rule-Based Classification and Similarity Search [Data set]. Zenodo https://doi.org/10.5281/zenodo.10171745* **2023**.
[20] D. M. Lowe, University of Cambridge (UK) **2012**.
[21] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *Using Machine Learning To Predict Suitable Conditions for Organic Reactions ACS Central Science* **2018**, *4*, 1465–1476.
[22] E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari, P. M. Zimmerman, *Learning To Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst Journal of Chemical Information and Modeling* **2019**, *59*, 3645–3654.
[23] O. Schilter, P. Schwaller, T. Laino, *Balancing computational chemistry's potential with its environmental impact Green Chemistry* **2024**, *26*, 8669–8679.
[24] A. D. Curzons, D. C. Constable, V. L. Cunningham, *Solvent selection guide: a guide to the integration of environmental, health and safety criteria into the selection of solvents Clean Products and Processes* **1999**, *1*, 82–90.
[25] W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, *Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling Journal of the American Chemical Society* **2022**, *144*, 4819–4827.
[26] D. Prat, J. Hayler, A. Wells, *A survey of solvent selection guides Green Chemistry* **2014**, *16*, 4546–4551.
[27] https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm, **U.S. Patent Statistics Chart Calendar Years 1963 – 2020**.
[28] J. P. Taygerly, L. M. Miller, A. Yee, E. A. Peterson, *A convenient guide to help select replacement solvents for dichloromethane in chromatography Green Chemistry* **2012**, *14*, 3020–3025.
[29] D. G. Drueckhammer, S. Q. Gao, X. Liang, J. Liao, *Acetone–Heptane as a Solvent System for Combining Chromatography on Silica Gel with Solvent Recycling ACS Sustainable Chemistry & Engineering* **2013**, *1*, 87–90.
[30] F. M. Chardon, N. Blaquiere, G. M. Castanedo, S. G. Koenig, *Development of a tripartite solvent blend for sustainable chromatography Green Chemistry* **2014**, *16*, 4102–4105.
[31] D. S. MacMillan, J. Murray, H. F. Sneddon, C. Jamieson, A. J. B. Watson, *Replacement of dichloromethane within chromatographic purification: a guide to alternative solvents Green Chemistry* **2012**, *14*, 3016–3019.
[32] A. Herskowitz, N. Ishii, H. Schaumburg, *N-Hexane Neuropathy New England Journal of Medicine* **1971**, *285*, 82–85.
[33] S. J. Crosbie, P. G. Blain, F. M. Williams, *Metabolism of n-hexane by rat liver and extrahepatic tissues and the effect of cytochrome P-450 inducers Hum Exp Toxicol* **1997**, *16*, 131–137.
[34] D. G. Graham, V. Amarnath, W. M. Valentine, S. J. Pyle, D. C. Anthony, *Pathogenetic studies of hexane and carbon disulfide neurotoxicity Crit Rev Toxicol* **1995**, *25*, 91–112.
[35] A. Wohl, *Bromierung ungesättigter Verbindungen mit N-Brom-acetamid, ein Beitrag zur Lehre vom Verlauf chemischer Vorgänge Berichte der deutschen chemischen Gesellschaft (A and B Series)* **1919**, *52*, 51–63.
[36] K. Ziegler, G. Schenck, E. W. Krockow, A. Siebert, A. Wenz, H. Weber, *Die Synthese des Cantharidins Justus Liebigs Annalen der Chemie* **1942**, *551*, 1–79.
[37] P. M. Swamidass, in *Encyclopedia of Production and Manufacturing Management* (Ed.: P. M. Swamidass), Springer US, New York, NY, **2000**, pp. 465–465.
[38] R. Appel, *Tertiary Phosphane/Tetrachloromethane, a Versatile Reagent for Chlorination, Dehydration, and P□N Linkage Angewandte Chemie International Edition in English* **1975**, *14*, 801–811.
[39] I. M. H. Downie, J. B.; Lee, J. B. , *Preparation of alkyl chlorides under mild conditions. Chem. Ind.* **1966**, 900–901.
[40] K. Reimer, F. Tiemann, *Ueber die Einwirkung von Chloroform auf alkalische Phenolate Berichte der deutschen chemischen Gesellschaft* **1876**, *9*, 824–828.

[41] S. Joudan, R. J. Lundgren, *Taking the "F" out of forever chemicals Science* **2022**, *377*, 816–817.
[42] H. P. H. Arp, A. Gredelj, J. Glüge, M. Scheringer, I. T. Cousins, *The Global Threat from the Irreversible Accumulation of Trifluoroacetic Acid (TFA) Environmental Science & Technology* **2024**, *58*, 19925–19935.
[43] S. D. Roughley, A. M. Jordan, *The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates J Med Chem* **2011**, *54*, 3451–3479.
[44] F. Sörgel, M. Kinzig, M. Abdel-Tawab, C. Bidmon, A. Schreiber, S. Ermel, J. Wohlfart, A. Besa, O. Scherf-Clavel, U. Holzgrabe, *The contamination of valsartan and other sartans, part 1: New findings Journal of Pharmaceutical and Biomedical Analysis* **2019**, *172*, 395–405.
[45] D. Suarez, G. Laval, S.-M. Tu, D. Jiang, C. L. Robinson, R. Scott, B. T. Golding, *Benzylic Brominations with N-Bromosuccinimide in (Trifluoromethyl)benzene Synthesis* **2009**, *2009*, 1807–1810.
[46] X. Zhao, W. Y. Ng, K.-C. Lau, A. E. C. Collis, I. T. Horváth, *Generation of (nonafluoro-tert-butoxy)methyl ponytails for enhanced fluorous partition of aromatics and heterocycles Physical Chemistry Chemical Physics* **2012**, *14*, 3909–3914.
[47] M. Ball, D. Horvath, T. Kogej, M. Kabeshov, A. Varnek, *Predicting reaction conditions: a data-driven perspective Chemical Science* **2025**, *16*, 17523–17541.
[48] J. Hoelzemann, G. Manso, A. Nagaraj, M. Tranchero, *The Streetlight Effect in Data-Driven Exploration*, *NBER Working Paper Series*, **2025**, Working paper 32401.
[49] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, C. W. Coley, *The Open Reaction Database Journal of the American Chemical Society* **2021**, *143*, 18820–18826.
[50] C. A. Bauer, T. Kogej, S. Genheden, P.-O. Norrby, *Precedent Finder: Locating Pareto-Optimal Reactions Journal of Chemical Information and Modeling* **2025**, *65*, 9378–9382.
[51] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, *Extraction of organic chemistry grammar from unsupervised learning of chemical reactions Science Advances* **2021**, *7*, eabe4166.
[52] https://www.daylight.com/meetings/summerschool01/course/basics/smirks.html.
[53] D. Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
[54] T. P. Landrum G., Kelley B., Rodriguez R., Cosgrove D., Vianello R., et al. .
[55] A. Vilsmeier, A. Haack, *Über die Einwirkung von Halogenphosphor auf Alkyl-formanilide. Eine neue Methode zur Darstellung sekundärer und tertiärer p-Alkylamino-benzaldehyde Berichte der deutschen chemischen Gesellschaft (A and B Series)* **1927**, *60*, 119–122.

Entry for the Table of Contents



Solvent choice is governed by precedent more than chemistry. Analysis of 1.3M reactions over 40 years reveals trends in solvent use. The various roles of solvents are explored, and findings expose examples of substitution of one problematic solvent by another, and systemic biases in datasets used to train AI solvent-prediction models. Reaction classes dependent on problematic solvents are identified for future focus.