Probing the Chemical "Reactome" with High Throughput Experimentation Data

Authors: Emma King-Smith[1], Simon Berritt[2], Louise Bernier[3], Xinjun Hou[4], Jacquelyn L. Klug-McLeod[2], Jason Mustakis[2], Neal W. Sach[3], Joseph W. Tucker[2], Qingyi Yang[4], Roger M. Howard[2]*, Alpha A. Lee[1]*

Affiliations:

[1]Cavendish Laboratory, University of Cambridge; Cambridge, CB3 0HE, UK

[2]Pfizer Research and Development; Groton, CT 06340, USA

[3]Pfizer Research and Development; La Jolla, CA 92121, USA

[4]Pfizer Research and Development; Cambridge, MA 02139, USA

*Corresponding author for dataset queries. Email: roger.howard@pfizer.com

*Corresponding author. Email: aal44@cam.ac.uk

Abstract: High-throughput experimentation (HTE) has the potential to improve our understanding of organic chemistry by systematically interrogating reactivity across diverse chemical spaces. Notable bottlenecks include few publicly available large-scale datasets and the need for facile interpretation of these data's hidden chemical insights. Herein we report the development of a High Throughput Experimentation Analyzer (HiTEA), a robust and statistically rigorous framework which is applicable to any HTE dataset regardless of size, scope, or target reaction outcome. We improve the HTE data landscape with the disclosure of 39,000+ previously proprietary HTE reactions. HiTEA is validated on this dataset, showcasing the elucidation of hidden relationships between reaction components and outcomes as well as highlighting reaction space that necessitates further investigation.

Main Text: Data-driven chemistry has seen immense strides in recent years, especially in yield and enantioselectivity prediction [1-4]. One major contributing factor to this is the adoption of high throughput experimentation (HTE) data in chemical synthesis [5-8]. Collections of "real-world" HTE data have several beneficial features. They likely have sampled the reaction space that is of direct interest to the field and cover a broad range of substrates and reaction types, ensuring that data-driven findings are relevant [9]. Valuable negative data is also present[10]. Additionally, the data will likely have been gathered in a manner that enables future HTE-guided synthesis, aiding the translatability of the findings. This approach is, however, not without its challenges. Yield calculations are often derived from the uncalibrated ratio of UV absorbances, which assumes that the species have similar UV extinction coefficients and makes this measurement more qualitative than quantitative NMR or isolated yield determinations. The presence or absence of byproducts may also be somewhat obscured (See the SI for a full discussion on yield determination HTE datasets). Moreover, datasets may be subject to biases in reactant and reaction condition selection and have regions of significant data sparsity.

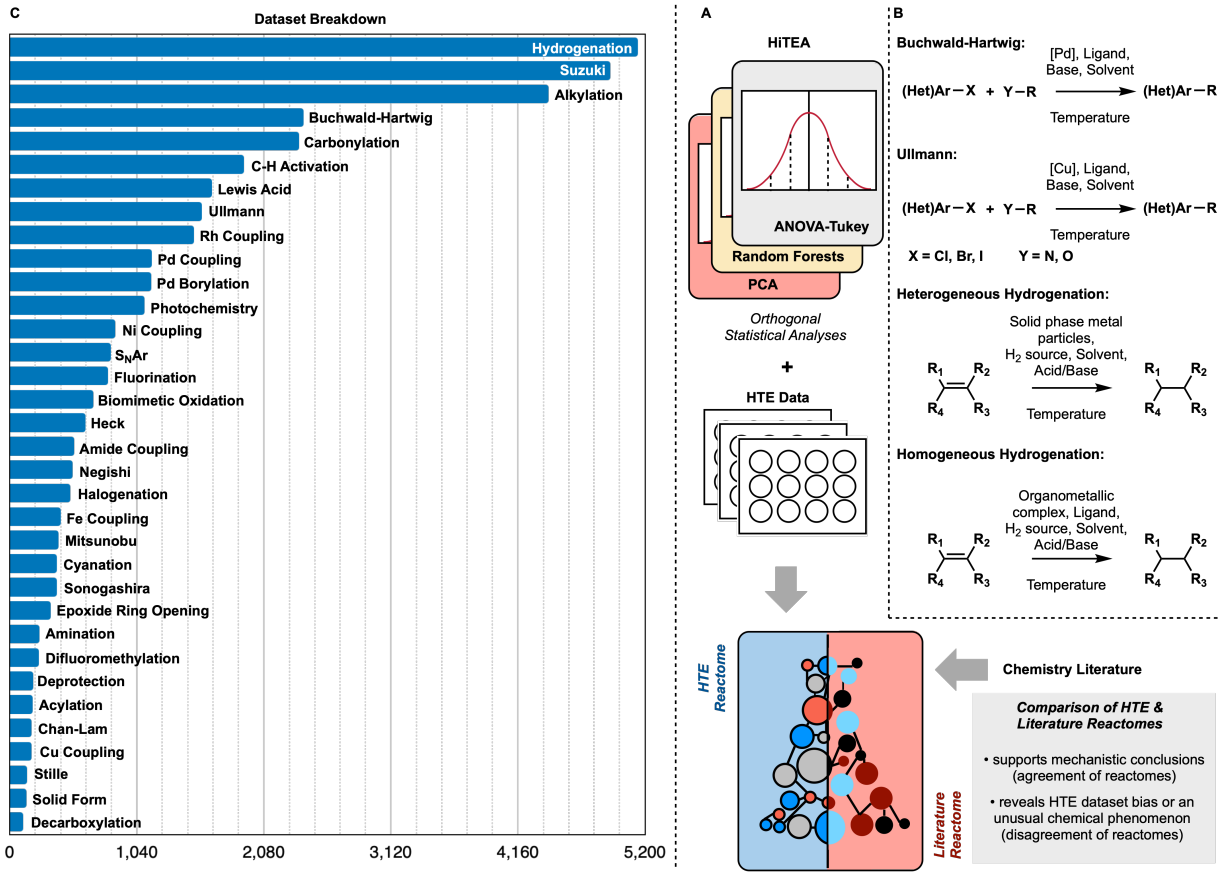

Fig. 1: Overview of the HTE dataset and framework. (**A**) Overview of HiTEA and its analysis. (**B**) Abstracted representations of the four reaction classes analyzed by HiTEA in this publication. (**C**) Breakdown of the HTE Dataset by reaction class.

Despite these known challenges with HTE data, little work has been done to investigate the inherent structure and biases of these datasets [11]. A statistically robust methodology that can be

applied to any HTE dataset to draw out hidden chemical insights is fundamental to driving forward data-driven chemistry. It is important to note that that this statistical framework was not envisioned to predict or generalize any specific reaction property (yield, selectivity, optimal conditions, etc.), but to provide a far more fundamental analysis: what are the chemical insights within a dataset? From these conclusions, we can begin to understand (a) what are statistically important factors that drive good or bad outcomes and (b) what this data will teach an AI model. Finally, comparison of the chemical insights embedded within the HTE data, what we dub the "HTE reactome", to the chemical insights drawn from the literature, the "literature's reactome", may (a) provide further evidence to support the mechanistic hypotheses (agreement of the reactomes), (b) reveal bias within the dataset which limits its usefulness or (c) reveal subtle correlations that may lead to refinement of our chemical understanding (disagreement of the reactomes) (Fig. 1A). For the purposes of this paper, "literature" is defined as information from open-source chemistry databases and published literature in peer-reviewed journals.

To create such a methodology, a **Hi**gh **T**hroughput **E**xperimentation **A**nalyzer (HiTEA) was developed which can deduce the reactome of any HTE dataset. Whilst common chemistry datasets such as the CAS,[12] Reaxys,[13] USPTO, Pistachio, or the Open Reaction Database have impressive coverage, it was a concern that the high level of overlap between their reactions and literature data would shape these datasets' reactomes to be indistinguishable to literature reactomes [14,15]. This would make it difficult to explore the discrepancies between the data and the literature reactomes, something that is likely to be possible utilizing HTE datasets and a fundamental feature of HiTEA that we wished to investigate. Thus, a HiTEA analysis was performed on a ground-breaking release of 10 years of historical medicinal chemistry HTE data. It is an unprecedently large dataset, acquired over 10+ years and spans a wide range of reaction classes (Fig. 1C). Within it are over 39,000 reactions conditions spanning over 350 target products. The reactions are split across numerous classes, ranging from thousands of reactions to tens of reactions, whose reactants and reagents may be over-represented or under-represented. These challenges highlight the necessity of statistical analyses, which can understand the data even in these skewed environments. HiTEA's analysis of several classic reaction types reveal some notable biases as well as some unexpected findings which warrant further investigation.

HiTEA: Statistical Analysis Framework for High Throughput Chemistry

The HiTEA methodology is centered around three orthogonal statistical analysis frameworks, random forests, Z-score-ANOVA-Tukey (**AN**alysis **O**f **VA**riance), and PCA (**P**rincipal **C**omponent **A**nalysis). Each framework answers one of the following questions: Which variable(s) are important? (random forest) [16], what are the statistically significant best-in-class/worst-in-class reagents (Z-score-ANOVA-Tukey) [17-20], and how do those best-in-class/worst-in-class reagents populate the chemical space (PCA) [21,22]. Notably, this combination of statical analysis makes no assumption about the underlying data structure. For example, relationships can be non-linear or even discontinuous, the data does not need to be the full combinatorial cross of all reagents with all reactants, an important feature when considering the sparse nature of chemistry datasets, and smaller datasets are just as feasible as larger datasets. The synergy between these three branches of HiTEA paint a comprehensible understanding of a dataset's reactome, allowing for facile identification of hidden chemical insights. To highlight the flexibility and versatility of HiTEA, we analyze datasets that span upwards of 3,000 reactions across a broad range of substrates to datasets that are just over 1,000 reactions with a narrower substrate scope.

Which Variables are Most Important?

Intuitively, some reactions are more sensitive to certain variables than others. Cross-couplings are highly sensitive to the metal and its ligand, but generally less sensitive to the identity of the solvent [23,24]. The relative variable importance is critical to understanding the chemistry insights that are present in the reactome. Note that importance can be positively correlated or negatively correlated with reaction outcome.

When investigating variable importance, two techniques come to mind as versatile and broadly applicable: random forests and multi-linear regressions. Both have yielded impressive results in chemistry and other fields, however, for HiTEA we chose to utilize random forests [4,16,25,26]. Unlike multi-linear regression, random forests do not stipulate that one's data must be linear, and thus obviate the need for linearization (and ideally normalization). Given the non-linearity of the data, we hypothesized that random forests would yield more accurate variable importances. In general, moderate to good out of bag accuracy of reaction outcome from a random forest with standard hyperparameters was observed (Table S1), with some noted exceptions (see Taking Dataset to HiTEA sections), correlating with poorer mechanistic insights of the reaction class overall. To assess the confidence of the variable importance, ANOVA was performed on each dataset subclass with statistical significance of the variables set at $p = 0.05$.

What are the Best- and Worst-In-Class Reagents?

It is known that there are privileged reagents that perform well across the board for a multitude of reactions, and there are those who's utility is narrow. Identifying the best- and worst-in-class reagents is therefore key to understanding a reactome. However, detangling the impact of a reagent from the inherent reactivity of the reactant(s) is challenging. We chose to compare relative yields which had been normalized through Z-scores, a technique that has shown promise in analysis of HTE data [17,27]. Notably, this framework allows for other target reaction outcomes to be used such as diastereoselectivity / enatioselectivity. ANOVA on the normalized target reaction outcome reveals the broad variables (solvent, base, catalyst system, temperature, etc.) that are statistically relevant for that reaction outcome [17-19]. Tukey's HSD test is then used to identify the outliers in each statistically significant variable, which are then ranked by average z-score to provide the best- and worst-in-class reagents [20].

How do the Best- and Worst-In-Class Reagents Populate the Chemical Space?

A visualization of the best- and worst-in-class reagents is useful to contextualize the scope of the dataset and therefore the extent of the reactome. The selection bias of reagents and clustering of high and low performing reagents can be easily interpreted. Whilst numerous techniques for dimensionality reduction and visualization of high-dimensional space are known, we chose to use PCA as its utility has been widely documented and numerous reliable, user-friendly implementations exist [28,29]. Additionally, PCA is more interpretable than UMAP (**U**niform **M**anifold **A**pproximation and **P**rojection) or tSNE (**t**-distributed **S**tochastic **N**eighbor **E**mbedding) whose non-linearity necessitate warping the high-dimensional shape of the data during projection; the xy-axes of projection lose the easy interpretability of highest variance (x-axis) / second highest variance (y-axis) that is fundamental to PCA [30,31].

Taking the Dataset to HiTEA

To test HiTEA, four distinct reactomes were chosen to be explored. These reactomes were widely used reaction classes: Buchwald-Hartwig couplings, Ullmann couplings, heterogeneous hydrogenations, and homogeneous hydrogenations (Fig. 1B). From the generated reactomes carefully analysis of HiTEA's variable importances, statistically relevant best / worst-in-class bases and catalysts, and ligand distribution was performed, concluding with tailored recommendations for further exploration. This analysis was also performed on temporally segregated data and data with their 0% yielding reactions removed, to mimic a dataset that would be more likely found in literature sources. Generally, temporal analysis appeared to be better correlated with the series of individual substrates screened over time than the evolving screen designs themselves. The removal of 0% yielding reactions lead to a far poorer understanding of the reaction class overall (Fig. S1 - Fig. S4, Fig. S5 - Fig. S8). The disappearance of the worst-in-class reagents and catalysts was expected, however, best-in-class conditions also disappeared. This result highlighted the value of 0% and lower yielding data in the disclosure of all datasets.

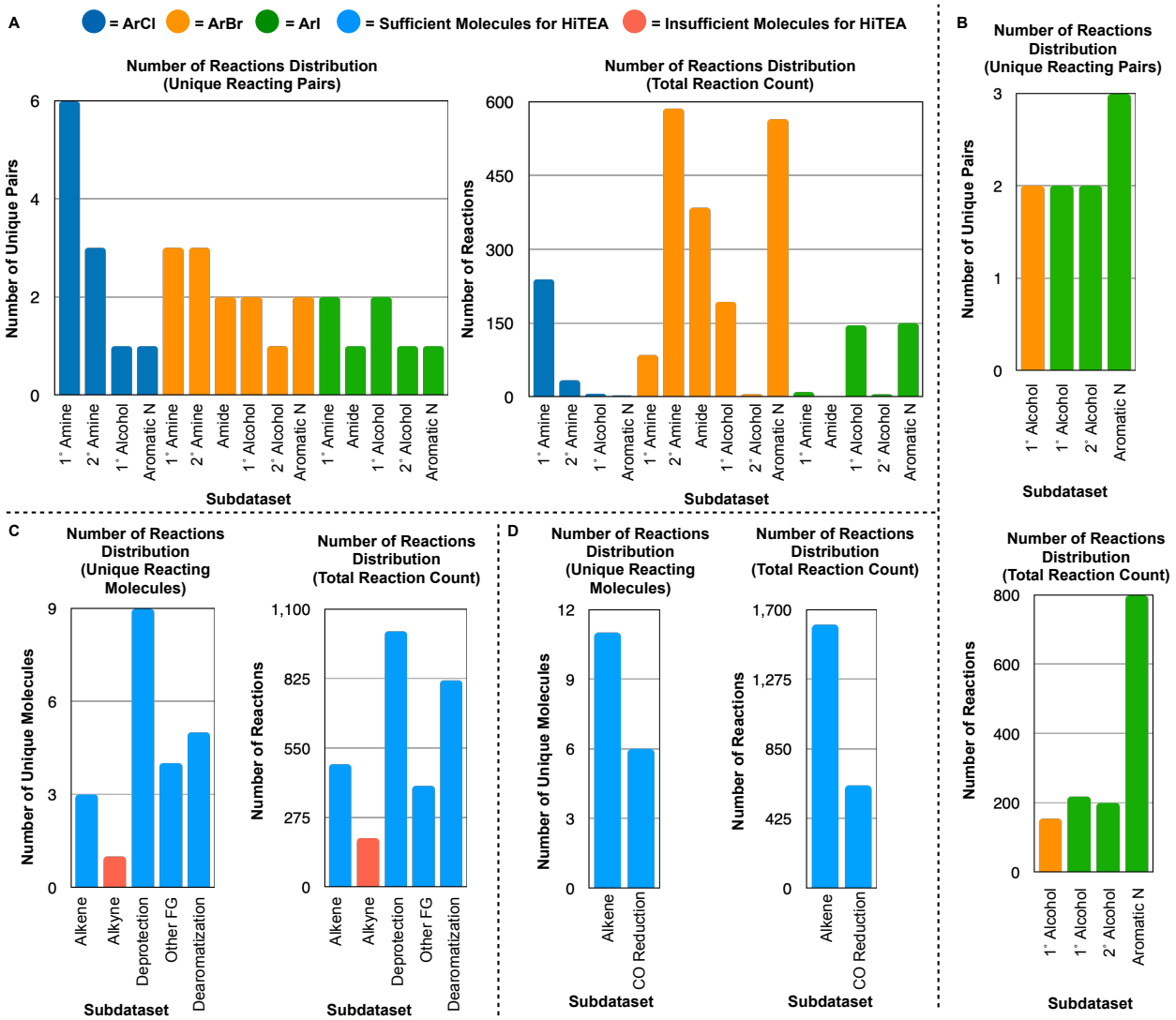

Fig. 2: Unique reacting pairs/molecules for each reaction class. (**A**) Buchwald-Hartwig dataset. (**B**) Ullmann dataset. (**C**) Heterogeneous hydrogenation dataset. (**D**) Homogeneous hydrogenation dataset.

Buchwald-Hartwig Couplings:

Buchwald-Hartwig couplings are a fundamental reaction in medicinal and process chemistry [32]. The dependence of yield upon ligand electronic and sterics is well reflected in this dataset; it is diverse in catalysts and ligands, but less diverse in coupling partners. This was the largest reactome we analyzed consisting of ~3,000 reactions.

Diversity wise, the dataset contained 31 unique halides and 32 unique nucleophiles, encompassing amine, amide, aromatic nitrogen, and alcohol nucleophiles, and 29 unique reacting halide-nucleophile pairs. Interestingly, the nucleophiles were less diverse than the aryl halides, owing to the nature of the ongoing campaigns at the time (Fig. S9). It was also found that aryl bromides made up the majority the reactions, both in number of unique reacting pairs and total number of reactions (Fig. 2A). It was expected that HiTEA on the Buchwald-Hartwig dataset without accommodating for this overrepresentation would reveal an HTE reactome significantly centered around aryl bromide couplings. Indeed, HiTEA credits significant variable importance to

BrettPhos Pd G1 (Fig. S10). This is clearly not in agreement with the literature's reactome in which many ligands show equal or better general performance to BrettPhos [11,23,24,33]. Thus it was hypothesized that a more nuanced analysis would arise if HiTEA was be applied to subdatasets (i.e. ArBr + 1˚ amine, ArCl + 1˚ amine, etc.), to determine their subreactomes. Subdatasets with

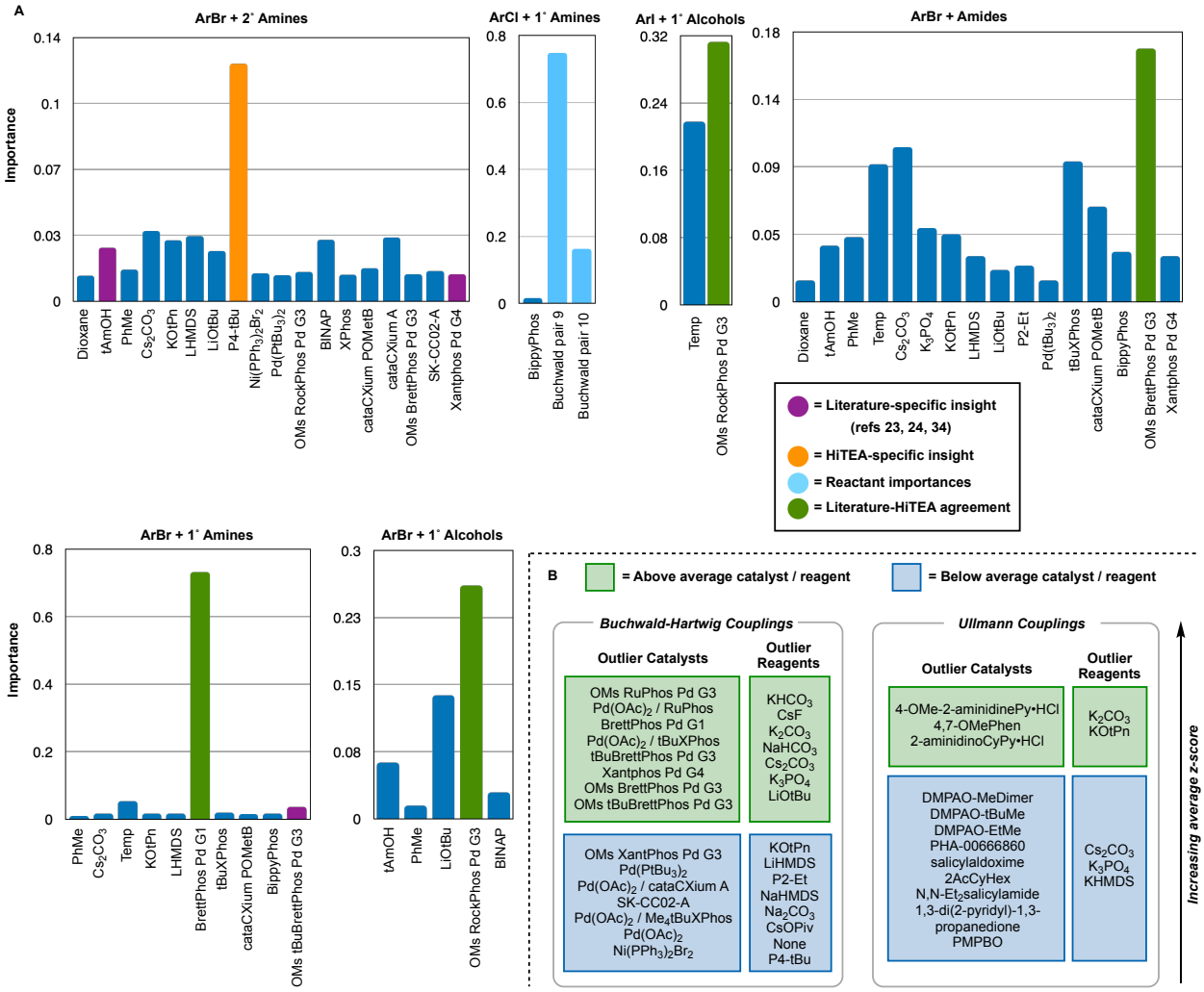

Fig. 3: HiTEA analysis of the Buchwald-Hartwig dataset. HiTEA/Literature-specific variable importances agreement between the literature and HiTEA variable importances highlighted. Acronym structures can be found in in Fig. S11. (**A**) Variable importances. Unless otherwise specified, the metal source for the ligand is $Pd(OAc)_2$. Where appropriate, reactant importances are shown. (**B**) Statistically significant best/worst-in-class catalysts and reagents. Unless otherwise specified, CuI is the copper source for the Ullmann couplings.

more than 80 reactions and two or more unique reacting pairs were analyzed, as these subreactomes were more likely to be differentiated from their literature chemical reactomes.

With ArBr + 1˚ amines (3 unique reacting pairs), the literature precedent suggests a high dependence on bulky biaryl phosphine ligands [23], which can inhibit the unproductive β-hydride elimination pathway and prioritize the reductive elimination, will be observed. BrettPhos ligands were expected to be dominant in the reactome's variable importances [34], and indeed, we see that BrettPhos Pd G1 is by far the most important variable for this subdataset. Surprisingly, the even bulkier *t*-BuBrettPhos was not in contention for the top important variable [23,24]. The ArBr + 2˚ amines (3 unique reacting pairs) show a negative dependence on the presence of P4-*t*-Bu, a

phosphazene base, which despite known utility in cross-couplings [35], is universally bad for this subdataset. The other phosphazene base, P2-Et, is also ranked poorly by HiTEA (Fig. 3B). A recent systematic investigation of optimal standard Buchwald-Hartwig conditions noted that P2-Et underperformed other bases [9]. With ArBr and ArI + 1˚ alcohols (both with 2 unique reacting pairs), and ArBr + amides (2 unique reacting pairs), ligands with rigid backbones and steric bulk which promote easier reductive elimination and prevent the deleterious K^2-amidate complexes [24,36] were expected to dominate, although a lower diversity of catalysts present in the variable importance analysis could be due to the lower random forest out of bag accuracy for these two reaction classes (Fig. SX). For these three subdatasets, the subreactomes are in agreement with the literature's reactome with OMs RockPhos Pd G3 and OMs BrettPhos Pd G3 highlighted in HiTEA's analysis. Finally, we turn to ArCl + 1˚ amine couplings (6 unique reacting pairs). Here, the literature reports electron rich ligands that allow for more facile oxidative addition of the Ar-Cl bond and bulky scaffolds that limit the known β-hydrogen elimination pathway are preferred [37,38]. However, the HTE subreactome had only $Pd(OAc)_2$/BippyPhos as a variable of minor significance (Fig. 3A). Upon closer inspection, a high dependence upon substrate identity was observed, implying that for this subreactome, the most important factor is the reacting halide-nucleophile pair.

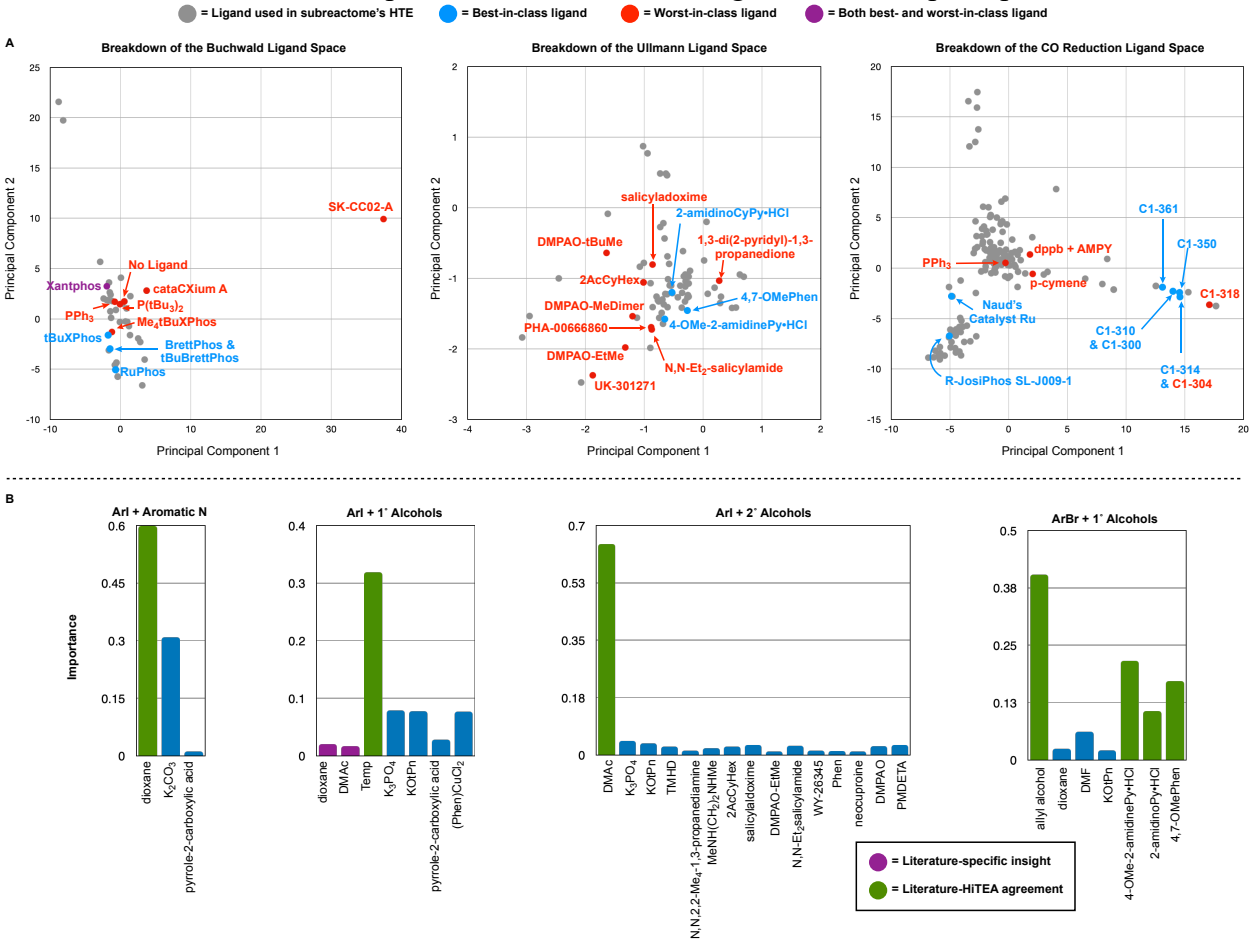

Fig. 4: (**A**) PCA ligand analysis of the Buchwald, Ullmann, and CO reduction ligands. (**B**) HiTEA variable importance analysis of Ullmann dataset. Unless otherwise specified, CuI is the copper source. HiTEA/Literature-specific variable importances agreement between the literature and HiTEA variable importances highlighted. Acronym structures can be found in Fig. S11.

Overall, the best / worst-in-class catalysts fall neatly into chemical intuition for the reasons highlighted above (Fig. 3B), and gratifyingly also cluster neatly in the ligand PCA visualization

(Fig. 4A). A sharp divide between best-in-class and worst-in-class ligand clustering is clear and Xantphos, the single ligand that could be either depending upon the precatalyst employed, resides away from the other ligands. Many of the subreactomes also agree with the literature's reactome, but several areas of interest stick out. First, the ArCl + 1˚ amines reactome differs from the literature's. While ArCl + 1˚ amine yield *are* somewhat dependent upon their reactants' structures, the lack of any significant ligand importance and the dominance of reactant identity suggests to us that this dataset may have some substrate selection bias. A clearer picture of ArCl + 1˚ amines' reactome could be achieved with expansion of the diversity in nucleophiles screened. The second is the little importance placed upon *t*-BuBrettPhos in ArBr + 1˚ amine's reactome. This may be due to the infrequent usage of *t*-BuBrettPhos when compared to the other catalysts in the subdataset. In the instances that *t*-BuBrettPhos was utilized, it was with challenging substrates (hence why it was noted as a best-in-class ligand with z-score-ANOVA-Tukey). In future screens, it could be advantageous to use *t*-BuBrettPhos more frequently to investigate this further.

Ullmann Couplings:

In recent years, palladium free cross couplings such as the Ullmann reaction have gained in popularity due to their cost-effectiveness [39]. Ullmann couplings, in particular are a viable option for aryl bromide / iodide and nucleophile cross couplings. The Ullmann dataset is more modest in scope and scale than its Buchwald-Hartwig counterpart, at about half the size, however even in this smaller space HiTEA is applicable.

Contrary to the Buchwald-Hartwig dataset, which encompassed a "wide but shallow" sampling of the substrate space, the Ullmann reactions are "narrow but deep" with few subdatasets but significant total number of reactions for each (Fig. 2B). The dataset contained 9 unique halide-nucleophile pairs, with good diversity in both the aryl halides and the nucleophiles, albeit a limited number of each (Fig. S12).

HiTEA revealed HTE subreactomes that readily distinguish between subtle differences in solvent. Across the board, a high importance of solvent is observed, with dependencies based on differing reactomes. For example, In the ArI + aromatic nitrogen (3 unique reacting pairs) and ArI + 2˚ alcohol couplings (2 unique reacting pairs), dioxane and DMAc are favored, respectively. For ArBr + 1˚ alcohol's reactome (2 unique reacting pairs), these two solvents are revealed to have less importance. In fact, the solvent of importance, allyl alcohol, is also the nucleophile in these couplings. In this subreactome, ligand identity plays a significant role in yield determination. Finally, for the ArI + 1˚ alcohol's reactome, reaction temperature is a leading factor, a point which is of no surprise [40]. It was expected to also see some importance placed on temperature for the other 3 subreactomes, but due to HTE design, temperature remained nearly constant throughout the entire subreactome, eliminating it as a variable.

For the Ullmann dataset, we believe the HTE and literature reactomes are in broad agreement. Gratifyingly, phenanthroline-based and picolinamide-based ligands are present in the best-in-class ligands, which are well known as privileged scaffolds in Ullmann couplings [41,42]. HiTEA observed that the Ma ligands (DMPAO and PMPBO) were individually less successful than other ligands in the standardized format of this HTE dataset. These ligands are characterized by high yields in the literature which acknowledges that their yields are sensitive to the electronics of the specific ligand-reactant pairing [43]. Thus, it is possible that the true potential of these ligand was masked. Visualization of the ligand space reveals a very narrow scope for ligand choice, perhaps

unsurprisingly given the similarity of privileged scaffolds in Ullmann couplings (Fig. 4A and Fig. S14). A unique observation for the Ullmann's ligand PCA is that the clustering is confined to the best-in-class ligands, supporting the random forest, z-score-ANOVA-Tukey findings that, for the most part, a few select ligands are useful for good yield outcomes. This outcome highlights the sensitivity of HiTEA's best- and worst-in-class catalyst analysis: despite the similarity in structure of the ligands, key differences in performance were identified, leading to a remarkably subtle overall ranking. The selection of specific solvents within the subreactomes was also intriguing. Although all the solvents identified are known to be good solvents within the literature, it is striking how each solvent's importance varies across subreactomes. Solvent effects are known to play a role in the mechanism of Ullmann couplings, but an exact understanding of which solvents are best for SET vs IAT or for C-N vs C-O coupling is not fully characterized, despite observed preferences [40,44,45]. A deeper dive into solvent characteristics is recommended for a more comprehensive understanding of this reactome overall.

Hydrogenations:

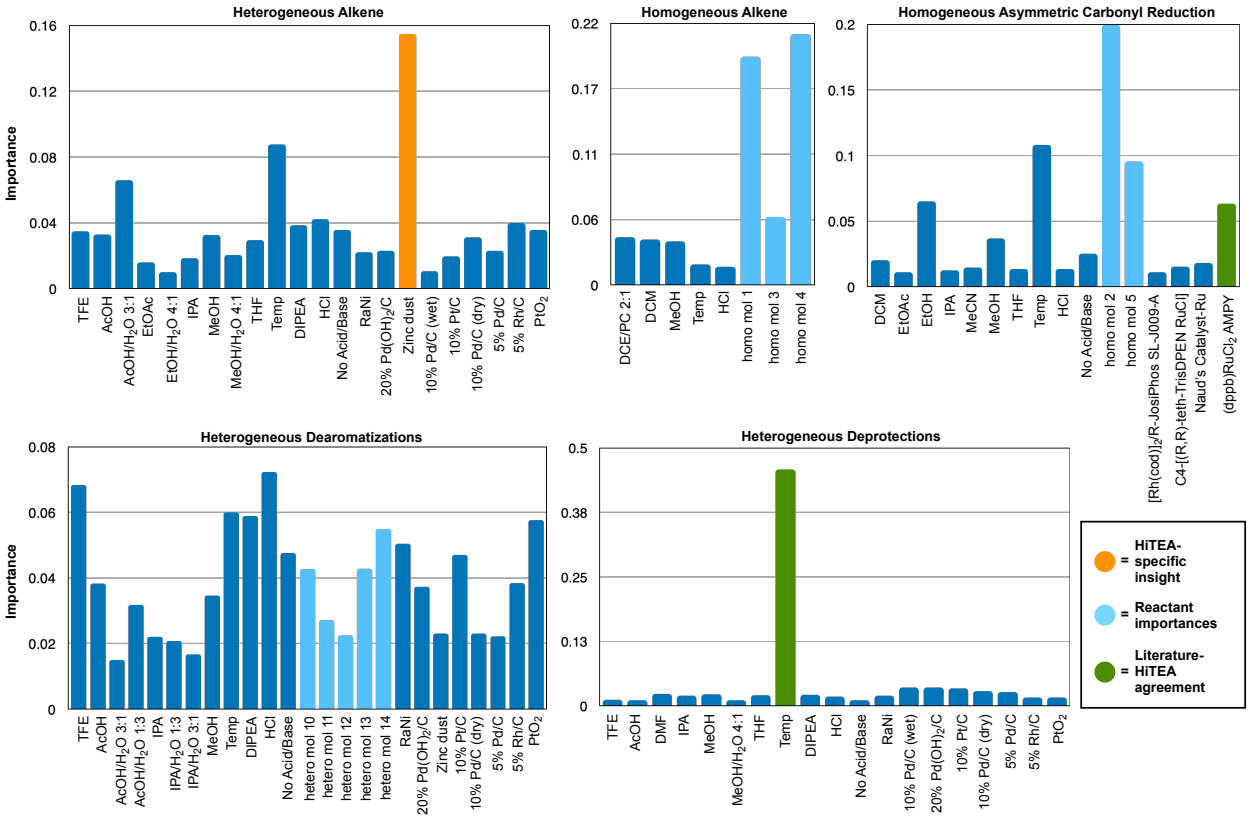

Fig. 5: HiTEA variable importance analysis on heterogeneous and homogeneous hydrogenation datasets. Where appropriate, reactant importances are shown. HiTEA-specific variable importances highlighted, as well as agreement between the literature and HiTEA variable importances. Acronym structures can be found in Fig. S11.

Hydrogenations are a well utilized reaction with a broad range of applications [46,47]. The mechanistic differences between heterogeneous and homogeneous hydrogenations warrant that these datasets be analyzed separately. Similar to the Buchwald-Hartwig dataset, the heterogeneous hydrogenations sample the reaction space in a "wide but shallow" manner whereas the homogenous hydrogenations follow the Ullmann's "narrow but deep" scope. The overall diversity of the molecules for both hetero and homogeneous hydrogenations are broad (Fig. S13). We will

not be delving deeply into the heterogeneous reduction of "Other FGs" (those which contain a mixture of nitro, diazo, and nitrile reductions) nor the heterogeneous alkyne reductions due to this reaction type containing only a single unique molecule undergoing hydrogenation (Fig. 2C & 2D).

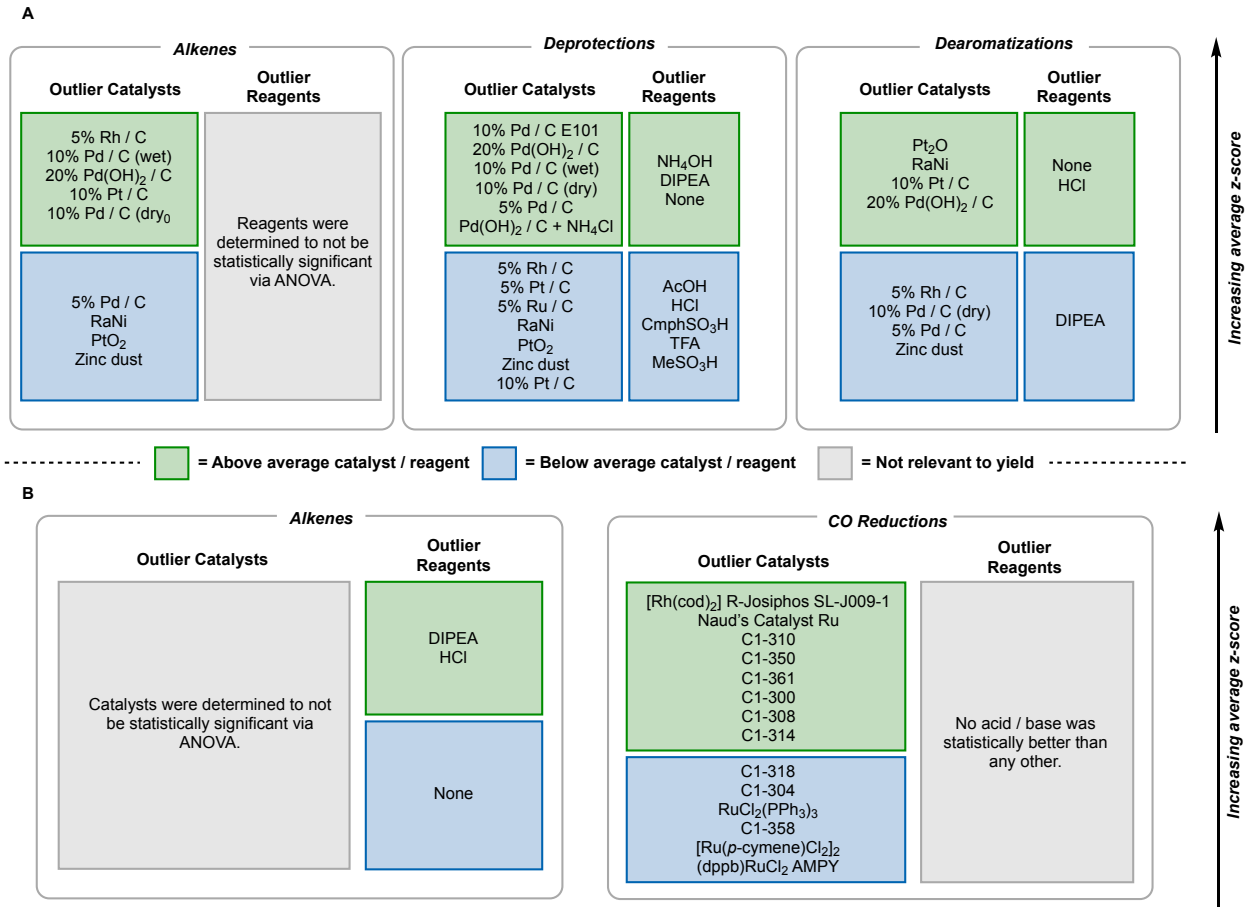

Fig. 6: HiTEA best/worst-in-class analysis of hydrogenation dataset. Acronym structures can be found in Fig. S11. (**A**) Heterogeneous hydrogenation dataset. (**B**) Homogeneous hydrogenation dataset.

HiTEA reveals that the heterogeneous alkene subreactome (3 unique reactants) places high negative importance on zinc dust and for the HTE deprotection subreactome (9 unique reactants) a high positive importance on temperature (Figure 5 & 6). Whilst temperature-correlated deprotections do agree with the literature's reactome [48,49], the negative correlation with zinc dust is a HiTEA-specific insight. This exemplifies the value of the negative results in the dataset which enables HiTEA to confirm negative correlations. The literature's reactome is often unable to confirm such correlations as it lacks publications with the negative data required. Interestingly, the other three subreactomes have no standout variable. In the case of the homogeneous hydrogenations (11 unique reactants), this can be explained by a strong dependence upon the reactants, but dearomatizations (5 unique reactants) show little overall dependence upon any variable, including molecule identity, perhaps due to the diverse and subtle changes that govern the energetically demanding process of dearomatization (Fig. 5) [50]. For these three subdatasets, the HiTEA's best/worst-in-reaction-type reveal more information (Fig. 6).

Overall, higher loadings of Pd/C are better than lower loadings, and Pearlman's catalyst is an all-around good catalyst for heterogeneous hydrogenations, two observations which are mirrored in the literature's reactome [51]. The pH of hydrogenolysis deprotections have been reported to have a marked effect in selectivity of the reaction, although acidic conditions are usually preferential [52]. Dearomatizations, primarily performed on nitrogen-containing heterocycles in this dataset, are partial to acidic conditions [53]. For the yield of asymmetric carbonyl reductions (6 unique reactants), ligand structure was a key factor, with very little importance placed on the base. In the literature, preferential treatment is given to rigid-backboned ligands, as it is hypothesized that flexible backbones deform the chiral pocket, leading to lower stereoselectivities [54,55]. Indeed, even yield in this subreactome is completely dominated by ligand structure, with the top performing ligand featuring a rigid 6-member ruthenium metallocycle of a Josiphos ligand. The middling ligands contain either 7- or 8-member metallocycles. The poorest performer is (dppb)$RuCl_2$AMPY boasts a significantly more flexible backbone from the rotational bonds between the P-P bridge (Fig. 6) [56]. Once again, ligand visualization reveals pockets of best- and worst-in-class ligand scaffolds, with clear distinctions between the best of the best-in-class ((*R*)-Josiphos SL J009-1 and Naud's catalyst Ru - see SI for all structures of acronyms) and the worst of the worst-in-class (*p*-cymene & dppb/AMPY), further supporting our chemical understanding of ligand design in metal-mediated asymmetric carbonyl reductions. Gratifyingly, even amongst the very structurally similar C1-3## family of catalysts from Johnson Matthey, a noticeable delineation between the good and poor performers is visible (Fig. 4A). Finally, homogeneous alkene hydrogenations' best / worst-in-class analysis confirms its variable importances conclusions: across the subreactome as a whole, the choice of catalyst is not statistically relevant in the determination of yield. This is no doubt a case of dataset bias as all of HiTEA's techniques failed to produce reasonable results: the random forest had low out of bag accuracy, the resulting random forest importances not including any catalysts, and the ANOVA deeming the catalysts as not statistically significant. A broader selection of alkene substrates and catalysts, or a subset of this dataset with less noise, would likely improve the utility of this subdataset.

Applications of HiTEA:

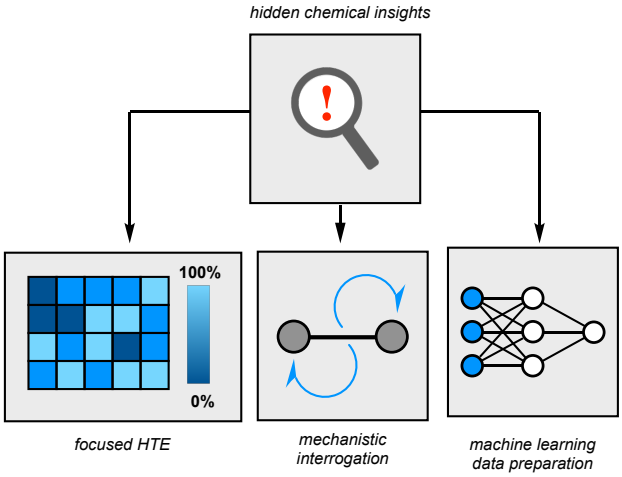

Fig. 7: Possible applications of HiTEA insights to batch scale, high throughput experimentation, and machine learning.

The hidden chemical insights brought to light by HiTEA have a multitude of potential applications. We posit three scenarios for HiTEA application, valuable for synthetic chemists and data scientists (Fig. 7).

Mechanistic Interrogation:

Keen understanding of the underlying reaction mechanism is advantageous for reaction optimization, and oftentimes, a deep understanding of a mechanism can lead to the development of new reactions and catalysts. However, many reaction mechanisms have seen only partial elucidation, especially those which feature organometallic transition states.[57] We imagine that HiTEA could identify hidden correlations between reaction inputs and measured reaction outcome, providing statistically robust evidence for or against mechanistic hypotheses. In the course of our manuscript, we discovered that solvent identity plays a significant role in the yield of Ullmann couplings, however,

unlike their Buchwald counterparts, the effect of solvent polarity on the multitude of potential XAT / SET catalyst intermediates has not been elucidated. As HiTEA has been designed to be applicable in even low data environments, it has conceivable utility in the investigations of other reaction mechanisms with limited screening.

Bias Identification for Machine Learning:

Bias is detrimental to machine learning because it allows the model to "cheat", relying on spurious correlations to get the right answer and leading to a lack of generalizability.[58] Take for example an image classification network recognizing a lion based on the savannah background rather than the animal's own features.[59] Image classifiers now employ a variety of techniques to try to combat bias of this type in addition to using huge image datasets that will have images of their subjects in a variety of backgrounds, poses, and distances.

For chemistry, HTE data has been noted as a valuable source of data for machine learning algorithms, as it is one of the best ways to generate moderate-to-large scale amounts of data in a parallel fashion. However, this data will also some bias: the reagents chosen by the chemist running the screen, the reaction is known to fail with specific motifs thus those motifs are left out of the dataset, or the simple fact that HTE is limited to the set of synthesizable molecules, which can be thought of as a bias, albeit one that we may want the network to learn or to operate in. As observed in the previous sections, HiTEA is adept at finding areas of bias in datasets, which usually take the form of substrate bias. When using these biased datasets for machine learning, one can either a) augment the dataset with further rounds of HTE or additional datasets or b) take a subset of the dataset that is less noisy and less biased; a removal of outliers. Both tasks can be aided by HiTEA through iterative augmentations or reductions followed by HiTEA. Stable and chemically sound HiTEA results (the removal of the surprising insights) indicate a dataset that is relatively robust, and superior for consistent modeling.

HiTEA for Future HTE Screens:

The most straightforward application of HiTEA is for future reaction optimization reactions, either in high throughput or in batch. Whilst HTE can explore swaths of chemical space, the combinatorial cross of all feasible reagents x catalysts x ligands x additives with even a limited set of reactants is unfeasible. HiTEA can give a visualization of the breadth of the scope and rapidly assess the statistically significant best and worst reagents, guiding the chemist to optimal reaction outcome. One could imagine HiTEA being used in conjunction with Shields *et al.*'s Bayesian optimizer for even faster optimization.[60] Additionally, temporal analysis is straightforward to run to visualize trends in poor and excellent conditions over time, adding further versatility to HiTEA's utility in reaction screening.

Conclusions:

Dataset exploration is an overlooked, but critical area of research in data-driven chemistry. The experimentalist is often blind to the chemical insights that have been locked into these datasets, missing key directions towards areas of exploration. With the development of HiTEA, a meaningful step in addressing this challenge has been made. We uncovered several interesting areas of exploration within Ullmann reactomes and identified several reactomes which would most benefit from additional HTE. We hope that this publication serves as a call to arms to the chemical community to collect, publish and analyze additional chemistry HTE data, providing further opportunities to explore the uncharted territories of the chemical reactome.

References:

1 Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186-190 (2018).

2 Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *Journal of the American Chemical Society* **140**, 5004-5008, doi:10.1021/jacs.8b01523 (2018).

3 Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343-348, doi:10.1038/s41586-019-1384-z (2019).

4 Santiago, C. B., Guo, J.-Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chemical Science* **9**, 2398-2412, doi:10.1039/C7SC04679K (2018).

5 Mennen, S. M. *et al.* The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Organic Process Research & Development* **23**, 1213-1242, doi:10.1021/acs.oprd.9b00140 (2019).

6 Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Medicinal Chemistry Letters* **8**, 601-607, doi:10.1021/acsmedchemlett.7b00165 (2017).

7 Krska, S. W., DiRocco, D. A., Dreher, S. D. & Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Accounts of Chemical Research* **50**, 2976-2985, doi:10.1021/acs.accounts.7b00428 (2017).

8 Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429-434, doi:10.1126/science.aap9112 (2018).

9 Xu, J. *et al.* Roadmap to Pharmaceutically Relevant Reactivity Models Leveraging High-Throughput Experimentation. (2022).

10 Strieth-Kalthoff, F. *et al.* Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angewandte Chemie International Edition* **n/a**, e202204647, doi:https://doi.org/10.1002/anie.202204647.

11 Fitzner, M. *et al.* What can reaction databases teach us about Buchwald–Hartwig cross-couplings? *Chemical Science* **11**, 13085-13093, doi:10.1039/D0SC04074F (2020).

12 CAS Content Collection (RXNs, atom mapping for RXNs,
associated RN's, and chemical structures). Available from
CAS, http://www.cas.org.

13 Reaxys, Online. Available:https://www.reaxys.com.

14 Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O. & Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science* **11**, 154-168, doi:10.1039/C9SC04944D (2020).

15 Kearnes, S. M. *et al.* The Open Reaction Database. *Journal of the American Chemical Society* **143**, 18820-18826, doi:10.1021/jacs.1c09820 (2021).

16 Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician* **63**, 308-319, doi:10.1198/tast.2009.08199 (2009).

17 Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J. & Nadon, R. Statistical practice in high-throughput screening data analysis. *Nature Biotechnology* **24**, 167-175, doi:10.1038/nbt1186 (2006).

18 Fisher, R. A. in *Breakthroughs in statistics* 66-70 (Springer, 1992).

19 Bartel, J., Krumsiek, J. & Theis, F. J. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J* **4**, e201301009, doi:10.5936/csbj.201301009 (2013).
20 Tukey, J. W. Comparing Individual Means in the Analysis of Variance. *Biometrics* **5**, 99-114, doi:10.2307/3001913 (1949).
21 Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572, doi:10.1080/14786440109462720 (1901).
22 Kutchukian, P. S. *et al.* Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chemical Science* **7**, 2604-2613, doi:10.1039/C5SC04751J (2016).
23 Surry, D. S. & Buchwald, S. L. Dialkylbiaryl phosphines in Pd-catalyzed amination: a user's guide. *Chemical Science* **2**, 27-50, doi:10.1039/C0SC00331J (2011).
24 Ingoglia, B. T., Wagen, C. C. & Buchwald, S. L. Biaryl monophosphine ligands in palladium-catalyzed C–N coupling: An updated User's guide. *Tetrahedron* **75**, 4199-4211, doi:https://doi.org/10.1016/j.tet.2019.05.003 (2019).
25 Svetnik, V. *et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **43**, 1947-1958, doi:10.1021/ci034160g (2003).
26 Fawagreh, K., Gaber, M. M. & Elyan, E. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* **2**, 602-609, doi:10.1080/21642583.2014.956265 (2014).
27 Qiu, J., Patel, A. & Stevens, J. M. High-Throughput Salt Screening of Synthetic Intermediates: Effects of Solvents, Counterions, and Counterion Solubility. *Organic Process Research & Development* **24**, 1262-1270, doi:10.1021/acs.oprd.0c00132 (2020).
28 Lê, S., Josse, J. & Husson, F. FactoMineR: an R package for multivariate analysis. *Journal of statistical software* **25**, 1-18 (2008).
29 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).
30 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv 2018. *arXiv preprint arXiv:1802.03426* (1802).
31 Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).
32 Caron, S. Honoring 25 Years of the Buchwald–Hartwig Amination. *Organic Process Research & Development* **23**, 1477-1477, doi:10.1021/acs.oprd.9b00177 (2019).
33 Ruiz-Castillo, P. & Buchwald, S. L. Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chemical Reviews* **116**, 12564-12649, doi:10.1021/acs.chemrev.6b00512 (2016).
34 Fors, B. P., Watson, D. A., Biscoe, M. R. & Buchwald, S. L. A Highly Active Catalyst for Pd-Catalyzed Amination Reactions: Cross-Coupling Reactions Using Aryl Mesylates and the Highly Selective Monoarylation of Primary Amines Using Aryl Chlorides. *Journal of the American Chemical Society* **130**, 13552-13554, doi:10.1021/ja8055358 (2008).
35 Buitrago Santanilla, A., Christensen, M., Campeau, L.-C., Davies, I. W. & Dreher, S. D. P2Et Phosphazene: A Mild, Functional Group Tolerant Base for Soluble, Room Temperature Pd-Catalyzed C–N, C–O, and C–C Cross-Coupling Reactions. *Organic Letters* **17**, 3370-3373, doi:10.1021/acs.orglett.5b01648 (2015).
36 Wu, X., Fors, B. P. & Buchwald, S. L. A Single Phosphine Ligand Allows Palladium-Catalyzed Intermolecular C□O Bond Formation with Secondary and Primary Alcohols.

Angewandte Chemie International Edition **50**, 9943-9947, doi:https://doi.org/10.1002/anie.201104361 (2011).

37 Christmann, U. & Vilar, R. Monoligated Palladium Species as Catalysts in Cross-Coupling Reactions. *Angewandte Chemie International Edition* **44**, 366-374, doi:https://doi.org/10.1002/anie.200461189 (2005).

38 Littke, A. F. & Fu, G. C. Palladium-Catalyzed Coupling Reactions of Aryl Chlorides. *Angewandte Chemie International Edition* **41**, 4176-4211, doi:https://doi.org/10.1002/1521-3773(20021115)41:22<4176::AID-ANIE4176>3.0.CO;2-U (2002).

39 Yang, Q., Zhao, Y. & Ma, D. Cu-Mediated Ullmann-Type Cross-Coupling and Industrial Applications in Route Design, Process Development, and Scale-up of Pharmaceutical and Agrochemical Processes. *Organic Process Research & Development* **26**, 1690-1750, doi:10.1021/acs.oprd.2c00050 (2022).

40 Sperotto, E., van Klink, G. P. M., van Koten, G. & de Vries, J. G. The mechanism of the modified Ullmann reaction. *Dalton Transactions* **39**, 10338-10351, doi:10.1039/C0DT00674B (2010).

41 Sambiagio, C., Munday, R. H., Marsden, S. P., Blacker, A. J. & McGowan, P. C. Picolinamides as Effective Ligands for Copper-Catalysed Aryl Ether Formation: Structure–Activity Relationships, Substrate Scope and Mechanistic Investigations. *Chemistry – A European Journal* **20**, 17606-17615, doi:https://doi.org/10.1002/chem.201404275 (2014).

42 Wu, F., Xie, J. & Zhu, Z. 1,10-Phenanthroline: A versatile ligand to promote copper-catalyzed cascade reactions. *Applied Organometallic Chemistry* **34**, e5926, doi:https://doi.org/10.1002/aoc.5926 (2020).

43 Zhou, W., Fan, M., Yin, J., Jiang, Y. & Ma, D. CuI/Oxalic Diamide Catalyzed Coupling Reaction of (Hetero)Aryl Chlorides and Amines. *Journal of the American Chemical Society* **137**, 11942-11945, doi:10.1021/jacs.5b08411 (2015).

44 Zhang, S., Zhu, Z. & Ding, Y. Proposal for halogen atom transfer mechanism for Ullmann O-arylation of phenols with aryl halides. *Dalton Transactions* **41**, 13832-13840, doi:10.1039/C2DT31500A (2012).

45 Chang, J. W. W. *et al.* Copper-catalyzed Ullmann coupling under ligand- and additive- free conditions. Part 1: O-Arylation of phenols with aryl halides. *Tetrahedron Letters* **49**, 2018-2022, doi:https://doi.org/10.1016/j.tetlet.2008.01.062 (2008).

46 Desai, B. & Kappe, C. O. Heterogeneous Hydrogenation Reactions Using a Continuous Flow High Pressure Device. *Journal of Combinatorial Chemistry* **7**, 641-643, doi:10.1021/cc050076x (2005).

47 Palmer, A. M. & Zanotti-Gerosa, A. Homogenous asymmetric hydrogenation: Recent trends and industrial applications. *Curr Opin Drug Discov Devel* **13**, 698-716 (2010).

48 Sinfelt, J. H. & Taylor, W. F. Catalytic hydrogenolysis of ethane. Variation of hydrogen pressure effects with temperature. *Transactions of the Faraday Society* **64**, 3086-3092, doi:10.1039/TF9686403086 (1968).

49 Shu, R. *et al.* Insight into the solvent, temperature and time effects on the hydrogenolysis of hydrolyzed lignin. *Bioresource Technology* **221**, 568-575, doi:https://doi.org/10.1016/j.biortech.2016.09.043 (2016).

50 Wang, D.-S., Chen, Q.-A., Lu, S.-M. & Zhou, Y.-G. Asymmetric Hydrogenation of Heteroarenes and Arenes. *Chemical Reviews* **112**, 2557-2590, doi:10.1021/cr200328h (2012).

51 Nishimura, S. *Handbook of heterogeneous catalytic hydrogenation for organic synthesis*. (Wiley New York, 2001).
52 David, A. & Vannice, M. A. Control of catalytic debenzylation and dehalogenation reactions during liquid-phase reduction by H2. *Journal of Catalysis* **237**, 349-358, doi:https://doi.org/10.1016/j.jcat.2005.11.017 (2006).
53 Glorius, F., Spielkamp, N., Holle, S., Goddard, R. & Lehmann, C. W. Efficient Asymmetric Hydrogenation of Pyridines. *Angewandte Chemie International Edition* **43**, 2850-2852, doi:https://doi.org/10.1002/anie.200453942 (2004).
54 Wang, H., Wen, J. & Zhang, X. Chiral Tridentate Ligands in Transition Metal-Catalyzed Asymmetric Hydrogenation. *Chemical Reviews* **121**, 7530-7567, doi:10.1021/acs.chemrev.1c00075 (2021).
55 Blaser, H. U., Pugin, B. & Spindler, F. Having Fun (and Commercial Success) with Josiphos and Related Chiral Ferrocene Based Ligands. *Helvetica Chimica Acta* **104**, e2000192, doi:https://doi.org/10.1002/hlca.202000192 (2021).
56 Mansell, S. M. Catalytic applications of small bite-angle diphosphorus ligands with single-atom linkers. *Dalton Transactions* **46**, 15157-15174, doi:10.1039/C7DT03395H (2017).
57 Santoro, S., Kalek, M., Huang, G. & Himo, F. Elucidation of mechanisms and selectivities of metal-catalyzed reactions using quantum chemical methodology. *Accounts of Chemical Research* **49**, 1006-1018 (2016).
58 Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **54**, 1-35 (2021).
59 Paullada, A., Raji, I. D., Bender, E. M., Denton, E. & Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* **2**, 100336 (2021).
60 Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89-96, doi:10.1038/s41586-021-03213-y (2021).

Acknowledgments: We would like to thank Rokas Elijošius, Felix Faber, and Oliver P. King-Smith for their insightful discussions. We are grateful to Klaus Dress for his assistance in setting up and maintaining the reaction screening database since its inception in 2011.

Funding: AAL is funded by a Royal Society University Research Fellowship. AAL and EKS acknowledge support from Pfizer Inc. EKS acknowledges support from the Royal Society Newton International Fellowship.

Author contributions: EKS and AAL conceived of the computational work. SB, XH, RMH, JM, JT, and QY conceptualized the HTE dataset. NWS and LB carried out the experimentation for the HTE dataset. RMH, JLK, NWS, JT, and QY curated the HTE dataset. EKS designed and carried out the analysis. EKS wrote the manuscript. SB, LB, XH, RMH, JLK, JM, NWS, JT, QY, and AAL reviewed the manuscript.

Competing interests: AAL is a co-founder and owns equity in PostEra Inc and Byterat Ltd. SB, LB, XH, RMH, JLK, JM, NWS, JT, and QY are employed by Pfizer Inc.

Data and materials availability: Further details of the analysis is available in the supplementary materials. Code and datasets can be found at https://github.com/emmaking-smith/HiTEA. Note that the datasets will only become available for download after the acceptance of the manuscript.

Supplementary Materials

Materials and Methods

Supplementary Text

Table S1

Figs. S1 - S14