Peer Review File

Manuscript Title: Bayesian Reaction Optimization as A Tool for Chemical Synthesis

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee expertise:

Referee #1: cheminformatics / machine learning in synthesis

Referee #2: BO and automation in chemistry

Referees' comments:

Referee #1 (Remarks to the Author):

The manuscript by Adams, Doyle and co-workers describes a suite of methods – Auto-qChem and BO – for the optimization of chemical syntheses and their code is provided. The methods are employed retrospectively in two related, Pd-catalyzed reactions (Suzuki-Miyaura and Buchwald-Hartwig cross-couplings) and one pseudo-prospective example towards the synthesis of BMS-911543 (a single intermediate reaction). Coincidently, that is another Pd-catalyzed reaction (C–H activation/functionalization). The study dwells on a hot topic in both cheminformatics and synthetic chemistry, as innovative tools that statistically motivate a set of experiments over another might reshape how wet lab chemistry is conducted in the future. However, the concept, implementation and conclusions are not new and the study presents numerous major inconsistencies/flaws from a machine learning and organic synthesis point of view. Overall, this study does not reach the high quality standard and soundness expected of a machine learning/organic chemistry manuscript in Nature. Moreover, the study fails to deliver on the identified gaps of knowledge, as it does not reach the state-of-the-art in terms of conceptualization, validation, chemical insights and conclusions. Because these originality and quality criteria are, in my opinion, not met I cannot recommend publication of the study in Nature. Below I provide a series of comments that substantiate my assessment.

Major issues:
1) "However, BO has not yet been investigated as a general method for chemical reaction optimization. In this work we report: (1) the development of a framework for Bayesian reaction optimization, (2) an open-source software tool that allows chemists to easily integrate state-of-the-art optimization algorithms into their every-day laboratory practices, and (3) a systematic study of BO compared to human decision making in reaction optimization." The work is not innovative and is based on incorrect observations. BO/Gaussian Processes have indeed been used to optimize chemical reactions (e.g. ACS Cent Sci 2018, 4, 1134; Chem Sci 2018, 9, 7682) and code is freely available in the Aspuru-Guzik GitHub (for the provided references). Although these examples do not have a GUI, such a software development, in itself, does not constitute ground-breaking innovation and, in my opinion, does not warrant publication in Nature. Moreover, part of the software presented herein (auto-qChem) has been used by the same authors elsewhere or is, at the very least, closely related (Science 2018, 360, 186). This is understandable from the caption to Figure 1B in the Science manuscript, stating that a software tool was built to automate feature generation, while using the same density functional and basis set as here. Extending the software capabilities by adding the LANL2DZ basis set for select atoms only incrementally improves functionality and does not conceptually advance the field. Also, previous studies, in related or similar applications, have performed human comparisons with identical outcomes (e.g. Angew Chem Int Ed, 2017, 56, 10815; ChemRxiv 2018, doi 10.26434/chemrxiv.7291205.v1). Overall,

this study does not provide any further insights into the utility of machine learning over expert intuition and fails at advancing the state-of-the-art at the interface of machine learning/organic chemistry.

2) "Overall, our studies suggest that adopting BO methods into everyday laboratory practices could facilitate faster and more efficient synthesis of functional chemicals by allowing scientists to make more efficient, data-driven decisions about which experiments to run." This is an over-statement as it is not thoroughly supported by data. The method disclosed by the authors served its purpose only in three similar (Pd-catalyzed) examples (two retrospective and one pseudo-prospective). This greatly diminishes the extent to which the machine-learning tool is challenged and validated. For such a broad generalization, and for a manuscript in Nature, the reader expects to find a thorough heuristics validation for every, or the grand majority of use cases (e.g. optimizing catalysts/ligands, solvent, temperature, molar amounts, reaction time or others) and numerous prospective examples of increasing difficulty to process. This is not the case. The study presents only limited validation – taking into account the number of use cases – under strictly controlled and non-challenging conditions: small search space size (3696, 792 and 1728) and optimizable variables (3, 3, and 5). In the first two cases, only categorical features are optimized, whereas in the last example both continuous and categorical features are considered. Multiple, diverse (in terms of mechanism, building blocks, potential conditions and real-world applications) and challenging (multiple variables, beyond what is intelligible to humans, and in HTE-untractable search spaces) prospective examples are warranted and not found in this study.

3) "Optimization of a chemical reaction is a complex, multidimensional challenge which requires experts to evaluate various reaction parameters (substrate, catalyst, reagent, additive, solvent, concentration, temperature, reactor, etc.), many of which have hundreds or thousands of possible settings (Figure 1A). Yet in a typical lab, bench chemists can only evaluate tens or hundreds of conditions during a standard optimization campaign due to time and material limitations." These considerations are partly correct and are not quantifiable. Importantly, there is a misalignment between the perceived needs and what the authors perform herein. There are numerous reaction parameters to optimize in reactions (at least 8 variables as identified here) and each variable can have hundreds or thousands of possibilities. Yet, in the presented use cases, the authors are not nearly as ambitious and only optimize 3, 3 and 5 variables with a limited number of possible configurations. This shows the non-challenging and unrealistic validations provided by the examples and how far off these optimizations are from real world scenarios and needs. It is also debatable the number of reactions a bench chemist can or is willing to perform and evaluate. These subjective considerations should not be present or serve as motivation for the work.

4) "Furthermore, modern advances in high-throughput-experimentation (HTE) have extended experimental capabilities to the collection of a few thousand data points under a limited set of conditions.4 Thus, the chemist's art is differentiating between millions of plausible configurations with a laboratory equipped to run only a tiny fraction of the possibilities." The study fails at answering the research question and address a grand challenge in synthetic chemistry. The challenge of mining millions of possible reaction configurations is real and clearly identified here. The difficulty is even more salient when HTE is not available to screen hundreds to thousands of reaction conditions simultaneously. Active machine learning can effectively help bench chemists at this task by compressing search spaces and prioritizing data acquisition. However, all use cases in this research fail at answering that real-world limitation. No example provided here has a search of space of even 4000 instances. Thus, one cannot conclude that the in silico method provides a general solution to reaction optimizations. Taken together, this constitutes a major limitation and study design flaw, as a reader of Nature expects to find multiple validations addressing unmet needs in real world scenarios. Without such validations, only scepticism (e.g. JACS 2018, 140, 4751) and undesired noise is raised in the expert community, irrespective of any retrospective studies.

5) "For example, design of experiments (DOE) seeks to sample experimental conditions that facilitate modeling of reaction parameters and deconvolution of interactions without bias and with minimum variance (Figure 1B). In conjunction with a response surface model, DOE enables the exploitation of knowledge gained from previous evaluations to guide the selection of future experiments. However, the exploration of reaction space is typically left in the hands of optimal designs, sensitivity analysis, literature precedence, and the operator's intuition." This is highly misleading to the reader. The goal of

DoE is understanding the response surface for a given reaction in order to identify optimal parameters. This goal can only be achieved with a degree of exploration (the authors present indicative data in Figure 3A). Reading the manuscript further it is impossible to distinguish what sets apart this and DoE methods and the conceptual advance relative to the latter. The relevant questions that one would expect to find answered and the authors fail at delivering are the advantages/disadvantages (clearly shown with head to head comparisons in prospective examples) and if the domains of applicability are different. The same questions are relevant when comparing the method reported here to others that operate in the same space (e.g. Phoenics) and various other machine-learning methods.

6) "In contrast to other model-based methods and iterative algorithms, BO is designed to balance exploration of areas of uncertainty and exploitation of available information, leading to high quality configurations in fewer evaluations. Importantly, BO algorithms can be applied to diverse search spaces including arbitrary parameterized reaction domains and enables the selection of multiple experiments in parallel. Accordingly, this approach is well suited to the optimization of chemical processes." The motivation for using BO is not sound and, most importantly, based on incorrect observations and omissions of prior art. BO with balanced exploration/exploitation selection strategies has been used before in the chemical sciences (ACS Cent Sci 2018, 4, 1134; Chem Sci 2018, 9, 7682) and thus it is not a new concept. A balance of exploration and exploitation can also be implemented in random forests for batch selection (e.g. Chem Sci 2016, 7, 3919 for selection of screening compounds), or single experiment selection (e.g. ChemRxiv 2018, doi 10.26434/chemrxiv.7291205.v1 for chemistry design). As exemplified, other algorithms can be employed in different search spaces while using customized, balanced selection policies. A reader would expect to find, a priori, a screen of different methods in a toy example(s) to motivate the use of BO over other heuristics. This is very important because the concept and implementation in this study are neither innovative nor new.

7) "To the best of our knowledge, there are (1) no applications to typical batch chemistry, (2) no general-purpose software platforms which are easily accessible to non experts, and (3) no systematic comparisons to expert chemists' performance.22" This is not true either and further reinforces the lack of novelty of the study, i.e. the narrative only fits the collected data and not the broader machine learning context. I have provided in the comments above multiple references showing several applications with and without BO heuristics and batch selection. Some of them can be found in the manuscript reference list, which makes the authors' statement quite surprising.
At least part of the software released here (descriptor calculations) has been previously used by the authors, as also commented above. There are also other user-friendly tools for the design of experiments (e.g. Minitab).
Furthermore, other colleagues have reported on comparisons between heuristics and humans in different settings and applications (example references above). The present study however does not provide any additional insights into the advantages of ML relative to humans and their biases. This would have been required for a manuscript that, if published, will be highly scrutinized.

8) "(…) open source software that is compatible with automated systems (e.g., flow chemistry) and human-in-the-loop experimentation (e.g., bench scale screening)" The authors over-inflate their findings in the absence of objective scientific evidence. For such claim, one expects to find the integration of the authors' method in a closed flow system loop. This is not the case. Similarly, there is no evidence that the method and humans can co-exist, providing better results than either one alone. Without such evidence, I find the "human-in-loop" expression out of context.

9) "Our approach is designed to integrate with existing synthetic chemistry practices, is applicable to arbitrary search spaces including continuous and categorially encoded reactions and enables the inclusion of physics and domain expertise." Again, this is not new and there is no evidence suggesting that the authors' solution outperforms existing ones at those tasks. While I can relate DTF calculations to physics knowledge, I found nothing in the manuscript showing robust evidence that the method is able to efficiently deal with human expertise (domain knowledge) and their biases (except a ligand list refinement in the last example). This is an unanswered yet relevant question since it has been recently shown that the utility of ML method is greatly affected by human biases (Nature 2019, 573, 251).

10) "First, we highlight the development of our BO platform using experimental reaction data mined from the literature. Next, to test our approach, we undertake a systematic investigation of BO compared to human decision making in the optimization of a new reaction – a key Pd-catalyzed direct arylation from the synthesis of BMS-911543 (Figure 1A)." The authors provide a sub-standard number of examples, with small search space sizes and number of optimizable variables. More specifically, two examples are purely retrospective, with search spaces of 3696 and 792 instances (far away from the authors' real world need of >10^6 – as in main text and Figure 1). The third example (search space of 1728) is also a Pd-catalyzed reaction presented in a retrospective configuration as the authors have performed all possible reactions beforehand through HTE. This is a troublesome option because it defeats the whole purpose of using active learning (compressing search spaces that cannot be fully enumerated experimentally, data availability and acquisition). Besides only probing mechanistically related reactions, the simultaneous use of HTE and small search spaces, renders all the test cases unsuitable to highlight the potential of the in silico method. These are examples in strictly controlled and non-challenging conditions that find little parallel in daily practice. As mentioned in 2), multiple, unrelated/diverse (in terms of mechanism, building blocks, potential conditions and real-world applications) and challenging (multiple variables, beyond what is intelligible to humans, and in HTE-untractable search spaces) prospective examples are warranted for a study published in Nature. The study also fails to answer relevant questions, such as is it possible to optimize reactions towards one regioisomer (if multiple products are possible in one reaction).

11) "(…) and the often-complex relationship between reaction components and yield of the desired product. Importantly, the dimensionality and range of outcomes offer reasonable microsystems to study real-world problems in chemical process optimization." This is too vague and has no place in data-, hypothesis-driven research. How much is the "often-complex relationship" and what are the "reasonable microsystems"? Although this reads well, there is no quantification whatsoever of the size of the problem and it only adds noise to the message.

12) Figure 1 highlights how non-challenging the first two examples are, in particular reaction 1 and 2c-e: small search spaces with a high number (naked eye) of well-performing reactions in low dimensional space (3 variables). I have concerns relative to the behaviour of the algorithm, taking into account the imbalance of reaction yield distribution in 2a,b relative to 2c-e. Admittedly, for 2a,b finding the optimum is not as simple, since random selection in 2c-e would result in >70% yield in ca. 1 out 2 or 3 picks. Probabilistically, it is less likely to achieve the same result in 2a,b. However, scrutinizing Figure 3F-H one finds exactly the same behaviour in all curves, for all compounds/reactions – this is impossible on absolute values. Indeed, the data is masked by a scaled YY axis, which means that a yield of 10% can be set as maximum in one case, whereas in another case a similar value in the YY axis might correspond to 90% yield. Data reporting is not transparent, oversells results and is misleading. In practice, a yield improvement of 1% (insignificant and inside the experimental error) might correspond to a jump of 10–20% in relative terms while only optimizing the reaction to trace amounts. Another aspect that is troublesome and raises a red flag in Figure 3 is that there are no kinks from batches that performed worse than all previous ones (irrespective of the added information). Such a result means that the algorithm always finds better performing reactions in all subsequent batches, independently of the uncertainty and selection policy. This also means that the starting points always correspond to poorly performing reactions, which I hardly find true in 2c-e given that random selection is used for initialization and there is a high percentage of well-performing reactions in those examples.

13) "(…) density functional theory24 (DFT) or cheminformatics descriptors generated using open source libraries." The manuscript is too vague and uninformative. The motivation is unclear and the final size of the descriptor vector is unknown throughout the manuscript and SI files. My understanding, from reading the SI file, is that thousands of descriptors were used to solve an optimization problem in 3 or 5-dimensional space. This disproportion is very questionable as most descriptors only add noise and do not contribute to the algorithm (e.g. what information can molar volume provide?). In addition, the mechanisms for the studied reactions are well known, which weakens the motivations of using all possible computable descriptors. A relevant, unanswered question is which studies did the authors conduct to ascertain the need for an oversized feature space, especially in light of the good performance of OHE and taking into account that no chemical insights were extracted while using DFT.

14) The manuscript is full of uninformative, fuzzy and general narrative, e.g. "(…) facilitate the screening and meta-optimization of the various components critical to achieving good performance with chemical reaction data." What are "various components" since, at best, only 5 variables were optimized; what is a good reaction performance when, in some cases, ca. 30-40% of all reactions have yield >50% (Figure 2). The same applies to the following: "Bayesian reaction optimization begins by collecting initial reaction outcome data via an experimental design (DOE) or by drawing from existing results (…)" and "After training the surrogate model, new experiments in the reaction space are sequentially chosen by optimizing an acquisition function which maximizes the expected utility of candidate experiments for the next evaluation". How many data points? Which selection criteria when drawn from existing results and function? What is the meaning of "expected utility"? Also, "This process continues iteratively, until reaction yield is maximized, resources are depleted, or the space is explored to a degree that finding improved conditions is improbable (e.g. no predicted improvements and estimated variance is low over the entire space)." Which stop criterion for the latter case? Some of the information is available in the manuscript/SI, but is highly scattered and unorganized, making it extremely difficult to follow and immediately grasp what was done. Extensive rewriting and clarifications backed by data would be required in this regard.

15) "(…) we selected surrogate model parameters based on regression performance for reactions 1 and 2a–e. Thus, we elected to take ourselves out of the loop and turn the optimizer on itself (See SI)." This is a nice feature but not new, according to a manuscript in ChemRxiv cited by the authors. The authors' fail to provide experimental evidence and explain how their approach advances the state-of-the-art.

16) "One point of divergence between the models was traced to the reaction representations. For DFT and cheminformatics encodings, which contain hundreds of features for a given reaction (after decorrelation), the automatic identification of important dimensions is critical to model performance. This issue is exacerbated in the low data regime, where the model needs to learn the variation in each dimension from only a few data points." This is an important statement but the study does not answer how the method deals with the curse of dimensionality in low (how many?) data regimes. By starting with 5 random reactions and having a descriptor with hundreds of dimensions, the authors are not able to assure their approach is generally and effectively applicable to all/most chemistry optimizations. Many prospective examples designed to study this limitation are missing. Does it have to do with the granularity of the search space or with the number of starting reactions? Can it be mitigated by batch size in very diverse search spaces, or by constraining search spaces and how (?), or by tuning selection policies? There are too many unanswered questions that would need clarification to help the readership confidently adopting this method. Related to these questions, the authors state that "Thus, we assume that most of the dimensions are irrelevant, impose this belief via a gamma prior centered on longer length scales, and estimate parameters by maximizing the marginal likelihood of the experimental data." This is expected because of the overwhelming amount of descriptors, which include DFT computations. Although I am not arguing about the DFT method here, the authors should at least consider that better options might be available for specific cases and caution the readership that the presented approach is expected to fail, at least occasionally, either because of mis-representations or inappropriate descriptors.

17) Something that is not clear (perhaps because of the scattered information) is for which species are all these descriptors calculated? In the SI, the authors mention "reaction representations based on Density Functional Theory (DFT), Mordred fingerprints" but there is no additional information besides that, in the latter case, >1800 descriptors were calculated for each reaction component. Reading the methods further and inspecting Figure S6 it seems the procedure is identical for DFT descriptors. This constitutes a major design flaw as no rationale can be identified. Given that the examples provided here only optimize 3 or 5 reaction variables and the starting materials are fixed throughout, it is irrelevant to calculate descriptors for those species since there is no variance.

18) "The central tenet of BO is the utilization of both information (exploitation) and uncertainty (exploration) to drive optimization." and "In contrast, a pioneering acquisition function (pure exploration), exemplified by selecting the point of greatest predictive uncertainty for evaluation, will tend to investigate the entire response surface more thoroughly." This oversells the authors' method and

falsely hints at added utility/innovation over previously reported heuristics (some examples provided in comments above). The authors exaggerate on the utility of BO because exploration and exploitation (alone or in combination) are the stepping-stone of active learning (e.g. Drug Discov Today 2015, 20, 458 and references in this review), irrespective of the employed ML algorithms. Indeed, the approach reported herein is not innovative, new nor pioneering. I also argue that exploitation hardly provides added value in terms of information gains, because it focuses on identifying experiments meeting a given target criteria rather than improving the overall model performance (as per Drug Discov Today 2015, 20, 458 and examples in the review).

19) "Over the course of 100 experiments the scores of the explorer and exploiter diverge, with the explorer offering a significantly better fit to the reaction surface. This strategy could be advantageous for users who hope to gain a global understanding of a reaction space." Although this is an interesting observation, there is no quantification in the manuscript of how much the scores diverge and when they start diverging. The manuscript is not objective and clear regarding which example(s) it refers to. Performing such a comparison over 100 iterations is extreme and questionable, depending on the use case. 100 iterations correspond to 3% and 13% of the search space for reaction 1 and 2, respectively, which makes it impossible to generalize findings.

20) The paragraph between lines 188–207 is too cryptic and the general readership will not understand the message. This should be presented with terms just above layman's level. Namely, the EI and TS scores should be more clearly explained. The following paragraph suffers from the same pitfalls. It is too generic and wordy for the small amount of scientific information it conveys. At the end of the paragraph the authors mention: "That is, running fewer experiments at a time results in more efficient use of resources by using information more effectively." The authors fail at identifying the reasons for this rather undesired event. On one hand, it suggests the deficiencies in the selection policy that was implemented. For batch selection, one would expect that all experiments are decorrelated from one another, in the decision function, as to provide the maximum, non-redundant information to the algorithm. Apparently, this is not the case. Other methods (e.g. Chem Sci 2016, 7, 3919) achieve this seamlessly, evidencing that the work developed herein does not reach the expected level of conceptualization. On the other hand, if the selection criteria were correctly implemented, the result indicates the small diversity in the search space. Any of these situations has implications in the conclusions one may draw from the utility of the heuristics and shows that multiple experiments are missing or were incorrectly planned.

21) The model fit score (Figure 3B) is appalling throughout the runs, even for an exploration policy ($r^2 < 0.5$). This is quite unexpected and beyond comprehension unless the exploration policy is not serving its purpose of improving model performance. Seemingly plateauing at experiment 60+ is even more worrisome because the model is not learning anything (or very slowly), despite the poor correlations between predictions and ground truth.

22) Data is presented in a poorly, uninformative and potentially misleading manner. At naked eye, the exploration trace (dotted line) in Figure 3B does not fit any of the plots in the subsequent panels. If the YY axes in panels C–H simply correspond to the max yield within a given batch, one expected to find downward curves, occasionally. For example, one of the first experiments has a yield of almost 100% (Figure 3B exploration). Only after ca. 20 experiments can one find a yield that is higher. This would correspond to a few batches and a downward curve should be visible in the corresponding panel (potentially C?). If the YY axes correspond to the maximum cumulative yield, then the curve would plateau for a few batches. All exploration curves in panels C–H show a steep rise in the first few experiments, which is not supported by the remainder experimental data.

23) The provided examples are not challenging and do not serve as proof-of-concept. Besides the small search space (Figure 3A), which I have debated elsewhere, the unsupervised learning algorithm shows the small diversity of all possible reaction conditions, as evidenced by a small number of mini-clusters (ca. 40–50). These can be further reduced, according to k-means, and as colour-coded by the authors. In practice, the results yield a manifold reduction in difficulty for identifying productive reaction conditions.

24) "(…) we found that both parallel EI and TS gave impressive performance, on average finding optimal configurations in only 50 experiments including 5 random initialization points." and "We were surprised to find, for many of the reactions, pure exploitation gave good results". The authors' observations fail to advance the field and add any relevant information relative to previous reports. Identifying optimized reaction conditions is an achievement, but in this particular case, it fails to at least match what other colleagues have reported (reference 21). What the authors report here is an optimization requiring the pseudo-execution of 1.4% and 6.3% of all possible reaction configurations. This falls short of <0.5% required elsewhere. The present result is also questionable because of the restricted search space, with many similar reactions (Figure 3A), and in light of my concerns regarding the number of optimizable variables. Multiple prospective examples would have been required to confidently ascertain the potential utility of the computational approach.
It is surprising that the authors did not at least consider the possibility of exploitation providing good results. My argument is based on 1) an identical observation in reference 21 – which the authors fail to acknowledge – and 2) the high number of well-performing reactions in the HTE screens. Because the authors did not design any experiment to disconnect the observation from the second hypothesis, it is impossible to conclude that exploitation works as a general approach or is indeed an unexpected result in this particular case. Moreover, speculating that "this could indicate that an objective is relatively convex and thus the optimizer is less likely to get trapped in a local maximum" without designing and executing the appropriate controls for such a potentially important justification is not helpful and falls short of the evidence-based science one expects to find in Nature.

25) "In contrast, chemists typically only implicitly incorporate uncertainty in an optimization campaign, for example by considering a few conditions which do not fit conventional wisdom." This is a major flaw in the design of the research and again, constitutes a narrative that is built to fit the needs of the study. The claim is unsupported by data/references and establishes a ground truth based on a subjective opinion. Motivating work on these grounds is not acceptable.

26) "Reaction optimization truly begins by defining the search space. If the search space is too large, containing many irrelevant or redundant configurations, significant amounts of time and resources can be wasted. If the space is too small, one risks excluding relevant experiments from consideration. We pursued a hybrid approach to selecting candidates, wherein we first consider the larger set of plausible experiments, then quantify similarities between potential reaction conditions via unsupervised learning and select those which ensure a satisfactory distribution across the larger search space." The study runs under highly controlled, curated and non-challenging conditions and does not answer the identified research question. I do understand why the authors have taken this approach – reducing the search space to a number of combinations tractable to their HTE setup. The approach however raises serious questions regarding the study design, contradicts the research question and the potential utility of the algorithm. Therefore, one cannot conclude if the approach is useful for a subset of examples only or can be scaled. The third use case also counter cycles the identified need of mining millions of possible reactions configurations. Given that the search space is condensed to 1728 reactions, HTE is employed as enabling technology. In these settings, active learning is redundant because all search space can be enumerated and outputs obtained. Performing active learning on a dataset for which all outputs in the search space are known (besides being similar to the previous two) does not qualify as a challenging and real-world scenario. The target molecule is of therapeutic interest, but given this irremediable limitation – in the study design and purpose of using active learning – the example does not work as a flagship or convincing application.

27) "Importantly, the final selection of reaction components was done based on valuable knowledge rather than ML". The search space is highly customized and built with domain knowledge. I find no motivation possible for using unsupervised learning and then curating results with potentially skewed domain perception. This has high potential to introduce biases and is detrimental for model building (Nature 2019, 573, 251). More specifically, the use of k-means for ligand curation is interesting but the actual selection of 12 representatives is convoluted. Multiple examples from a same cluster are prioritized, whereas other clusters are neglected (e.g. pink dots in Figure 4). Also, the search space is configured to include unevenly spread temperature values (90, 105 and 120 ºC) and a range of

concentrations (0.057, 0.100 or 0.153 M) that can only be motivated with extensive prior experience. Therefore, no evidence of utility under unknown reaction spaces is provided. This constitutes a major weakness of the study, as the approach is incompatible with the democratization of chemical syntheses, i.e. expert domain knowledge is required. When also considering other arguments presented above (small search space, low number of optimizable variables and ability to perform all reactions in parallel before the actual active learning run) this and all previous examples are deemed unsuitable for a manuscript in Nature.

28) "This was not surprising given the ML optimizers initial choices were made at random." The message conveyed by the authors is confusing and contradicts the remainder of the manuscript. Previously, the manuscript highlighted the utility of BO, stating that it can balance exploration and exploitation strategies in the selection process. Yet, here the authors say that the initial choices were made at random as a reason for the better performance of humans – this is inconsistent. Rather, the result is indicative that a potential benefit of the method can only be found when the reaction budget is high and amenable to some sort of parallelization, which is not always the case in an organic chemistry laboratory. The result is also indicative that the method inefficiently deals with optimization problems in rather low dimensional space (5 optimizable variables) when the descriptor vector has hundreds or thousands of dimensions. For a reasonable performance, the number of descriptor dimensions has to approximate the number of reactions performed (as indirectly cautioned elsewhere in the manuscript). Overall, the number of unanswered questions this manuscript raises shows that multiple experiments are missing and casts doubt on the utility of the algorithm on a wider scale.

29) "Finally, while only around half of human participants achieved within 1% of the optimal yield for reaction 3, the optimizer achieved >99% yield 100% of the time." The manuscript over-interprets data. There is no evidence that 99% or 100% yield are indeed different. No replicates and statistical analyses were performed to ascertain the difference. The comparison is also meaningless without mentioning after how many iterations it was made and without a statistical test. Finding that optimized conditions can be obtained within 50 retrospective experiments (3% of the search space) is interesting, but not ground-breaking in light of the authors' reference 21.

30) The study fails at providing chemical insight or mechanistic interpretations to their findings, falling short of the depth expected in a Nature manuscript. For example, there is no discussion onto the use of CgMe-PPh as a ligand, how this is potentially unexpected and disrupts with established knowledge and how it can advance the state-of-the-art. It is also unknown if the ligand is only applicable to this reaction or is transferrable to similar building blocks – a short substrate scoping study is missing. Overall, using an algorithm as a black box, without attempting to interpret its predictions and augmenting human knowledge is opposite to how the field of machine learning is currently evolving.

31) The authors make comparisons between human intuition and their BO heuristics. I argue their study design is flawed as not all variables are accounted for. Because different initializations are implemented, there is a possibility of having very different starting points for optimizations. This can include reactions that only performed very poorly, very similarly (and thus not informative), or sampling different regions in the search space (ideal scenario). This can affect the optimization efficiency of humans and thus, lead to skewed comparisons and unrealistic statistics. In my opinion, a correct design would have been tiered: First, with the same initialization for all researchers, ask them to optimize the reaction. This could inform on optimization paths, biases and ability to identify a local/global optimum. Secondly, with the same initialization path, allow the heuristics optimize the reaction. This could inform on differences in strategies (if a random state value is set, this could be determined), and ability of the heuristics to out-perform humans in that particular case. Finally, allow the heuristics optimize the reaction with different initializations. The study includes only the final experiment and the preceding questions are unanswered, yet relevant for a thorough comparison.

32) The supplementary information is superfluous in common ground information for colleagues in the machine learning field and simultaneously cryptic for those outside. The SI has no seemingly logical guiding line – lacks detail or is scattered – which made it very difficult for me to follow and find the information I needed. Re-wording, abbreviating and re-organizing are largely required. For example,

there is no irrefutable motivation for using DFT descriptors over others (e.g. OHE). There are no statistics for figure 8 and similar/related analyses throughout the SI file. The authors argue in favour of structurally meaningful encodings but fail at capitalizing on that option as no new knowledge is generated or extracted. There are no comparisons to DoE tools, other BO and machine learning pipelines. "contain hundreds of features" and "hundreds of hyperparameters" – how many? In page 26 the plot suggests OHE works better – is there a meaningful difference to this? Can it be quantified or motivated to go for DFT instead? Figure S26 has nothing to do with OHE as mentioned in page 27, yet a comparison to DFT is made. It is not clear how redundancy in selection is dealt with. To substantiate that the model is learning meaningful patterns it would be useful to show that it can consistently suggest reactions that afford no product or only trace amounts. There are not enough experimental details in the chemistry front – for data transparency, please provide chromatograms for all experiments in an active learning run, at least for one of the examples. Figure S47 shows that the authors' preferred setup (Gaussian processes and DFT descriptors) is never the best option for any given analysis and shows that there is no generalizable solution to what is being studied. This provides preliminary evidence that their algorithm will underperform consistently while in the wild. Gaussian processes with linear regression works better for low data (page 34) and when the number of experiments is <75 it performs equally well as non-linear regression – why not use it throughout and choose the poorer and most convoluted solution? Similarly, if the preferred method works better with higher data regimes, why did the authors start from 5 training examples?

33) I commend the authors for providing code and documentation. I found no issues but in some instances it wasn't intuitive for someone who wants to run in Python. There are too many options, reinforcing that there is no streamlined and general path to using this BO method.

Minor issues:
1) The study is prolific in abbreviations: RF, GP, TS, BO, ARD, DoE, ML, HTE, EI. This makes it very difficult to follow the narrative if not in the machine-learning field. No more than 3–4 abbreviations should be found in the manuscript.

2) The authors fail at citing relevant studies interfacing active learning and drug discovery. This would have been somewhat expected because the active learning concept is not new and there is a reasonable body of evidence that it can be applied to diverse tasks in early drug discovery. While I have no particular issue with the authors citing several books and reviews, these are too uninformative for a reader to follow up. I suggest substituting some of them by original work. For example, reference 29 could be substituted by one of many studies published by the Aspuru-Guzik lab. There are also numerous applications of random forests in chemical sciences (ref 28).

3) There is seemingly no relationship between Figure 1C and the "(...) acquisition function which maximizes the expected utility of candidate experiments for the next evaluation (...)". Captions to Figure 1 do not guide the reader through the overwhelming amount of information in each panel. This is an issue also present in other figures.

4) "To demonstrate this dichotomy, we tracked the exploiter and explorers' decisions (after initialization at the same point) in a two-dimensional representation of 1 (Figure 3A). Indeed, over the first several evaluations, the exploiter remained in a single cluster while the explorer traversed the entire space." The result is quite interesting and corroborates previous findings, namely in ChemRxiv 2018 (reference 21 in this manuscript). Would have been nice to discuss this further in light of independent reports.

5) "Humans made significantly better initial choices than the optimizer, on average discovering conditions which were ~15% yield higher in their first batch of experiments." Please provide statistics.

6) How did the authors make sure the survey participants did not communicate with each other?

7) The 13C chemical shift for the CH3 is missing in the list (page 55 of SI) – it should be 32.8 ppm. The

authors assign peaks to carbon atoms but I find no label (C1, C2, etc) in the structure. Please add those labels to the structure embedded in the 13C NMR spectrum.

Referee #2 (Remarks to the Author):

The manuscript by Doyle and coworkers presents a detailed development of a new tool to apply Bayesian optimization to help guide the trajectory of experimental synthetic chemistry. More specifically, it aims to provide a robust, data-driven approach for catalyst selection for a given experimental landscape.

While this is not the first application of Bayesian algorithms to experimental outcomes (see examples by Häse et al (https://arxiv.org/abs/2003.12127, https://doi.org/10.1021/acscentsci.8b00307 , https://doi.org/10.1371/journal.pone.0229862) this article does a superb job delineating the process of creating the training set, encoding the reactions, and then systematically testing the core "tweaks" needed for successful applications of Bayesian optimization. This includes exploring the impact of highly explorative vs exploitative models, impact of batch sampling, and the necessary rigor of defining the correct search space.

More interestingly, the authors have tested their algorithm against a number of human participants in a simulated game. This latter point is a very unique mode of comparison and serves to highlight experimenter bias. As the authors point out the experts tended to avoid ligands that were not familiar to them, overriding what subtle data trends may have been present. To this reviewer, this application adds a very creative means of both testing the tool, but also serving to educate the community.

Some of the particular observations and conclusions may be very much a product of the reaction data set that was utilized for this study. For example, the authors' observation that highly exploitative models where highly successful is not generically applicable across all catalytic reactions. I agree with the authors' statement that for this case the response surface may be highly convex – but an interesting observation nonetheless.

Overall, this manuscript provides an insightful and interesting addition to the growing community of chemical data scientists. More importantly, it is written from the perspective of educating the community and delineating the thought process that went into design and testing. For this reason in particular – the highly accessible tone in the paper – I would recommend acceptance of the article

Author Rebuttals to Initial Comments:

Major issues:

"1) "However, BO has not yet been investigated as a general method for chemical reaction optimization. In this work we report: (1) the development of a framework for Bayesian reaction optimization, (2) an open-source software tool that allows chemists to easily integrate state-of-the-art optimization algorithms into their every-day laboratory practices, and (3) a systematic study of BO compared to human decision making in reaction optimization." The work is not innovative and is based on incorrect observations. BO/Gaussian Processes have indeed been used to optimize chemical reactions (e.g. ACS Cent Sci 2018, 4, 1134; Chem Sci 2018, 9, 7682) and code is freely available in the Aspuru-Guzik GitHub (for the provided references). Although these examples do not have a GUI, such a software development, in itself, does not constitute ground-breaking innovation and, in my opinion, does not warrant publication in Nature. Moreover, part of the software presented herein (auto-qChem) has been used by the same authors elsewhere or is, atthe very least, closely related (Science 2018, 360, 186). This is understandable from the caption to Figure 1B in the Science manuscript, stating that a software tool was built to automate feature generation, while using the same density functional and basis set as here. Extending the software

capabilities by adding the LANL2DZ basis set for select atoms only incrementally improves functionality and does not conceptually advance the field. Also, previous studies, in related or similar applications, have performed human comparisons with identical outcomes (e.g. Angew Chem Int Ed, 2017, 56, 10815; ChemRxiv 2018, doi 10.26434/chemrxiv.7291205.v1). Overall, this study does not provide any further insights into the utility of machine learning over expert intuition and fails at advancing the state-of-the-art at the interface of machine learning/organic chemistry."

One fundamental difference between our software and that presented in the cited studies is that EDBO is designed and tuned (with chemical reaction data) to be applied in a laboratory setting. While the cited papers focus on developing new BO algorithms, our work seeks to tune established models for reaction optimization, statistically benchmark the methods, and provide a framework for everyday chemists to apply BO at the bench. The related papers did not actually apply BO to chemical reaction optimization. Instead they focus on computational objectives. In turn, the coded implementations for these studies are not directly applicable to the problems tackled in our manuscript. In response to the reviewer's comment we have clarified these distinctions in the manuscript.

Added text: While researchers have begun to explore the application of ML methods to reaction optimization4,6–8,25–27, these efforts have targeted a limited subset of synthetic chemistry including only continuous process parameters.

While the software is distinct from that reported in our previous work, we did not mean to imply that auto-qchem was a focus of this paper, rather just to comment that this is the software that we used to compute DFT encodings.

In response to the reviewer's comment we have included citations for the studies with human comparisons.

"2) "Overall, our studies suggest that adopting BO methods into everyday laboratory practices

could facilitate faster and more efficient synthesis of functional chemicals by allowing scientists to make more efficient, data-driven decisions about which experiments to run." This is an over-statement as it is not thoroughly supported by data. The method disclosed by the authors served its purpose only in three similar (Pd-catalyzed) examples (two retrospective and one pseudo-prospective). This greatly diminishes the extent to which the machine-learning tool is challenged and validated. For such a broad generalization, and for a manuscript in Nature, the reader expects to find a thorough heuristics validation for every, or the grand majority of use cases (e.g. optimizing catalysts/ligands, solvent, temperature, molar amounts, reaction time or others) and numerous prospective examples of increasing difficulty to process. This is not the case. The study presents only limited validation – taking into account the number of use cases – under strictly controlled and non-challenging conditions: small search space size (3696, 792 and 1728) and optimizable variables (3, 3, and 5). In the first two cases, only categorical features are optimized, whereas in the last example both continuous and categorical features are considered. Multiple, diverse (in terms of mechanism, building blocks, potential conditions and real-world applications) and challenging (multiple variables, beyond what is intelligible to humans, and in HTE-untractable search spaces) prospective examples are warranted and not found in this study."

The small search spaces utilized in this study are a practical matter and due to our approach to statistical validation. Collecting all experimental data for larger reaction spaces is not experimentally tractable. Even collecting thousands of controlled experiments (controlled for variance in outcome and translation to larger scale reactions) using HTE (as in our study) requires specialized equipment and significant time and effort. Given this challenge, our approach to validation was to carefully utilize or generate data for a smaller search space to be used as a microsystem so that we can statistically evaluate optimization algorithm performance under different circumstances. With this approach we were able to investigate multiple algorithms in terms of average outcome (i.e., answering the question, what do we expect to see on average?) and robustness (e.g., what is the variance with respect to starting point? What is the worst case?). The result of this analysis is that we can say with high confidence ($p < 0.05$) whether or not BO outperforms other methods (e.g., human intuition, random sampling, ε-greedy, etc.) for a given data set. In turn, these controlled experiments inform our thesis that these methods will be of general use to synthetic chemists tackling routine reaction optimizations.

Given these practical considerations, the alternative approach of running optimizations over larger search spaces but without exhaustively evaluating experimental outcomes (iteratively and with no replicates like old reference 21) would methodologically fall short in terms of accessing generalizability and provide no information about robustness. However, we recognize that prospective real-world test cases were a component missing from our initial submission. Therefore, in response to the reviewer's comments we have included 2 additional examples of BO applied to optimizations over search spaces too large to be evaluated via HTE. We highlight BO applied to the optimization of a Mitsunobu reaction and a deoxyfluorination reaction. In these examples, we show that BO can deliver high quality conditions in more challenging optimization problems which include both more parameters and a larger reaction space (hundreds of thousands of possible configurations).

“3) “Optimization of a chemical reaction is a complex, multidimensional challenge which requires experts to evaluate various reaction parameters (substrate, catalyst, reagent, additive, solvent, concentration, temperature, reactor, etc.), many of which have hundreds or thousands of possible settings (Figure 1A). Yet in a typical lab, bench chemists can only evaluate tens or hundreds of conditions during a standard optimization campaign due to time and material limitations.” These considerations are partly correct and are not quantifiable. Importantly, there is a misalignment between the perceived needs and what the authors perform herein. There are numerous reaction parameters to optimize in reactions (at least 8 variables as identified here) and each variable can have hundreds or thousands of possibilities. Yet, in the presented use cases, the authors are not nearly as ambitious and only optimize 3, 3 and 5 variables with a limited number of possible configurations. This shows the non-challenging and unrealistic validations provided by the examples and how far off these optimizations are from real world scenarios and needs. It is also debatable the number of reactions a bench chemist can or is willing to perform and evaluate. These subjective considerations should not be present or serve as motivation for the work.”

As highlighted in (2), the small search spaces utilized for development in this study are a practical matter and due to our approach to statistical validation. However, when the requirement for statistical validation is lifted, it is feasible to evaluate more challenging optimization problems. In response to the reviewer’s comment we carried out 2 real-world reaction optimizations on distinct reactions (Mitsunobu and deoxyfluorination). Importantly, we included 7 parameters in each optimization and overall search spaces with 180,000 and 312,500 possible experiments. Our findings demonstrate that BO delivers high quality experimental conditions in both test cases.

We do recognize that our quantification of experimental limitations is normative (albeit based in experience). In response to the reviewers comment we have adjusted our discussion – instead we simply state that “chemists can only evaluate a small subset”.

“4) “Furthermore, modern advances in high-throughput-experimentation (HTE) have extended experimental capabilities to the collection of a few thousand data points under a limited set of conditions.4 Thus, the chemist’s art is differentiating between millions of plausible configurations with a laboratory equipped to run only a tiny fraction of the possibilities.” The study fails at answering the research question and address a grand challenge in synthetic chemistry. The challenge of mining millions of possible reaction configurations is real and clearly identified here. The difficulty is even more salient when HTE is not available to screen hundreds to thousands of reaction conditions simultaneously. Active machine learning can effectively help bench chemists at this task by compressing search spaces and prioritizing data acquisition. However, all use cases in this research fail at answering that real-world limitation. No example provided here has a search of space of even 4000 instances. Thus, one cannot conclude that the in silico method provides a general solution to reaction optimizations. Taken together, this constitutes a major limitation and study design flaw, as a reader of Nature expects to find multiple validations addressing unmet needs in real world scenarios. Without such validations, only scepticism (e.g. JACS 2018, 140, 4751) and undesired noise is raised in the expert community, irrespective of any retrospective studies.”

We thank the reviewer for suggesting more challenging examples. As highlighted in (2) and (3), in response to the reviewer's comments we have evaluated more challenging real- world reaction optimization problems.

"5) "For example, design of experiments (DOE) seeks to sample experimental conditions that facilitate modeling of reaction parameters and deconvolution of interactions without bias and with minimum variance (Figure 1B). In conjunction with a response surface model, DOE enables the exploitation of knowledge gained from previous evaluations to guide the selection of future experiments. However, the exploration of reaction space is typically left in the hands of optimal designs, sensitivity analysis, literature precedence, and the operator's intuition." This is highly misleading to the reader. The goal of DoE is understanding the response surface for a given reaction in order to identify optimal parameters. This goal can only be achieved with a degree of exploration (the authors present indicative data in Figure 3A). Reading the manuscript further it is impossible to distinguish what sets apart this and DoE methods and the conceptual advance relative to the latter. The relevant questions that one would expect to find answered and the authors fail at delivering are the advantages/disadvantages (clearly shown with head to head comparisons in prospective examples) and if the domains of applicability are different. The same questions are relevant when comparing the method reported here to others that operate in the same space (e.g. Phoenics) and various other machine-learning methods."

In response to the reviewer's comments we clarified the advantages of our approach over DOE and included head-to-head comparisons with for the 6 development reactions (**1** and **2a–e**). Importantly, we highlight how DOE falls short in applicability to multi-level categorical optimization both conceptually and with statistical evaluation of optimization outcomes. Our findings show that BO is both more straightforward to apply and delivers higher quality conditions that the typical DOE approach.

Added text: We next sought to statistically evaluate the performance of DOE methods compared to BO. In some instances, it may be advantageous to initialize BO using an experimental design. Thus, to highlight the flexibility of our software, we developed a DOE module comparable with the main software methods and utilized EDBO to carry out simulations. When using DOE for reaction optimization, first an experimental design is executed, the results used to build a model, and the model used to predict optimal conditions. However, typical DOE methods for modeling response surfaces (e.g., central composite designs) require a sense of order in each dimension. Thus, DOE is most commonly utilized for the optimization of continuous process parameters in chemistry. We identified two designs which have been employed effectively towards the optimization of chemical processes with categorical variables: generalized subset designs and D- optimal designs.16,45 Generalized subsets, developed for applications in chemistry, are an extension of traditional fractional factorial designs to problems where factors can have more than two levels.45 D-Optimal designs are computer- aided designs which can be utilized when classical designs fail.1
While these designs are compatible with multilevel categorical variables, as the dimensionality of the search space grows, filling either design quickly become intractable experimentally. In addition, one point of experimental variability is that

the ordering of the levels is arbitrary. That is, the experiments selected depend upon the order in which the components in each category are listed. This of course is not a desirable feature as different users may list their reagents in a different order for the same reaction space. Thus, to gauge the variance in outcome to the ordering of the categorical variables we carried out simulations by shuffling the ordering of variables in the data. Then for each design we modeled the experimental data and evaluated predictions using a polynomial regression with quadratic terms using automatic relevance determination to fit the weights of the regression model. For each of the reactions in the development set both generalized subset- and D- Optimal design-based optimizations deviated from BO in both mean outcome ($p < 0.05$), standard deviation (BO $\leq$ 1.9, GSD $\leq$ 6.9, D-Optimal $\leq$ 3.3), and worst-case loss (BO $\leq$ 5, GSD $\leq$ 16, D-Optimal $\leq$ 15). Moreover, when compared to a random design, only the D-Optimal design offered a modest improvement in performance statistics (Table 1). Thus, we conclude that other things equal BO is both more straightforward to employ and provides superior performance in reaction optimization with categorical variables.

In addition, we also investigated the possibility of applying the software Phoenics to the optimization of reactions **1** and **2a–e** as an additional benchmark. However, Phoenics was not readily applicable to these problems. For example, we saw no methods for specifying a categorical parameters or utilizing a reaction encoding – the GitHub site includes the following statement: “Note, that the only currently supported parameter "type" is "float"”. This furthers the argument that such software is not directly applicable to typical synthetic chemistry problems. However, we would like to draw the reviewer’s attention to the fact that we did investigate numerous acquisition functions and models. These all constitute different optimization algorithms screened within our BO framework. For example, our investigation of random forest models and mean maximization are independently analogous to the approach presented in old reference 21 which also uses random forests and top predicted configurations to select experiments. We found that these methods, which differ in either selection criteria or model, offered poorer performance in optimization of the 6 development sets in terms of mean loss, standard deviation, and/or worst-case loss. To draw a direct analogy to old reference 21 we have included statistics for simulations in which we use a RF model and select experiments with the top predicted outcome below. This is essentially a parallel version of the greedy approach presented in old reference 21. In table R1 you can see that this method offers significantly ($p < 0.05$ in 4 out of 6 reactions) worse average performance, larger standard deviation in outcome, and greater worst-case losses.

	1	2a	2b	2c	2d	2e
p-value (H0: RF = BO)	0.001	0.007	0.000	0.020	0.133	0.139
RF mean loss	2	1	5	2	1	2
RF standard deviation	1.8	1.9	6.7	3.0	1.3	2.4
RF worst-case loss	7	8	18	12	5	11
BO mean loss	1	0	0	1	0	2
BO standard deviation	0.9	0.5	0.1	1.1	0.5	1.9
BO worst-case loss	4	1	0	3	2	5

Table R1. Summary of simulation results for reactions **1** and **2a–e** with a RF model and top predicted experiments selected (analogous to old reference 21): p-value is for a hypothesis test for equal average performance (N=30, batch size=5) the selected BO algorithm.

"6) "In contrast to other model-based methods and iterative algorithms, BO is designed to balance exploration of areas of uncertainty and exploitation of available information, leading to high quality configurations in fewer evaluations. Importantly, BO algorithms can be applied to diverse search spaces including arbitrary parameterized reaction domains and enables the selection of multiple experiments in parallel. Accordingly, this approach is well suited to the optimization of chemical processes." The motivation for using BO is not sound and, most importantly, based on incorrect observations and omissions of prior art. BO with balanced exploration/exploitation selection strategies has been used before in the chemical sciences (ACS Cent Sci 2018, 4, 1134; Chem Sci 2018, 9, 7682) and thus it is not a new concept. A balance of exploration and exploitation can also be implemented in random forests for batch selection (e.g. Chem Sci 2016, 7, 3919 for selection of screening compounds), or single experiment selection (e.g. ChemRxiv 2018, doi 10.26434/chemrxiv.7291205.v1 for chemistry design). As exemplified, other algorithms can be employed in different search spaces while using customized, balanced selection policies. A reader would expect to find, a priori, a screen of different methods in a toy example(s) to motivate the use of BO over other heuristics. This is very important because the concept and implementation in this study are neither innovative nor new."

While there are in principal other ML approaches to reaction optimization, our objective was to develop and rigorously test a BO platform tuned for reaction optimization. As highlighted in our response to (1) and (6), previously reported approaches are not readily applied to the problems presented in this study. In fact, these studies do not present any examples of experimental reaction optimization. Rather they focus on simulated objectives. In this respect, many ML papers have presented detailed simulations of BO and other methods on "toy example(s)". This paper is about applying BO to real-world chemical reaction optimization problems. Therefore, we focus on real chemical reaction data. We offer comparison to many different sequential model-based optimization algorithms by screening different surrogate models (RFs, BLMs, GPs) and numerous acquisition functions (PI, EI, UCB, TS, etc.) and selection policies (ε-greedy, MeanMax, etc.).

"7) "To the best of our knowledge, there are (1) no applications to typical batch chemistry, (2) no general-purpose software platforms which are easily accessible to non experts, and (3) no

systematic comparisons to expert chemists' performance.22" This is not true either and further reinforces the lack of novelty of the study, i.e. the narrative only fits the collected data and not the broader machine learning context. I have provided in the comments above multiple references showing several applications with and without BO heuristics and batch selection. Some of them can be found in the manuscript reference list, which makes the authors' statement quite surprising. At least part of the software released here (descriptor calculations) has been previously used by the authors, as also commented above. There are also other user-friendly tools for the design of experiments (e.g. Minitab). Furthermore, other colleagues have reported on comparisons between heuristics and humans in different settings and applications (example references above). The present study however does not provide any additional insights into the advantages of ML relative to humans and their biases. This would have been required for a manuscript that, if published, will be highly scrutinized."

As in comments (1), (5), and (6), we highlight that the peer reviewed references provided by the reviewer do not contain any examples of actual experimental reaction optimization. We address the reviewer's comment about the descriptor computation software (auto- qchem) in (1). In response to the reviewer's comments we have included additional references in which human comparisons were made (new references 25 and 53.

"8) "(…) open source software that is compatible with automated systems (e.g., flow chemistry) and human-in-the-loop experimentation (e.g., bench scale screening)" The authors over-inflate their findings in the absence of objective scientific evidence. For such claim, one expects to find the integration of the authors' method in a closed flow system loop. This is not the case. Similarly, there is no evidence that the method and humans can co-exist, providing better results than either one alone. Without such evidence, I find the "human-in-loop" expression out of context."

In response to the reviewer's comments we have now carried out examples of human-in- the-loop reaction optimization (see above). While we do not provide examples of reaction optimization with automated experimental systems, we do demonstrate the software can automatically optimize any computational objective written as a python function which returns a number (as it was used to optimize its own hyperparameters). We include additional computational optimization examples on the GitHub site. Therefore, we have changed the wording of this sentence to reflect these applications. As noted by the Reviewer, the extension to flow chemistry is out of scope for this work but would simply require a method for communicating between the different systems.

New wording: Herein we report the development of a modular framework for Bayesian reaction optimization and accompanying open-source software that is compatible with automated systems (e.g., computer experiments) and human-in- the-loop experimentation (e.g., bench scale screening).

"9) "Our approach is designed to integrate with existing synthetic chemistry practices, is applicable to arbitrary search spaces including continuous and categorially encoded reactions and enables the inclusion of physics and domain expertise." Again, this is not new and there is no evidence suggesting that the authors' solution outperforms existing ones at those tasks. While I can relate DTF calculations to physics knowledge, I found nothing in the manuscript showing robust

evidence that the method is able to efficiently deal with human expertise (domain knowledge) and their biases (except a ligand list refinement in the last example). This is an unanswered yet relevant question since it has been recently shown that the utility of ML method is greatly affected by human biases (Nature 2019, 573, 251)."

In response to the reviewer's comments we have now clarified the novelty of our approach and have provided a statistically principled evaluation of optimization performance with human expertise and DOE. In the new real-world optimization examples, we have demonstrated additional instances where domain expertise is integrated with EDBO for the refinement of the search space by selecting the components included in each reaction parameter. The reagents included in the reaction spaces were selected by the chemists carrying the experiments based on availability of material and knowledge about the reactions. That being said, the reviewer is right in stating that biases could lead a chemist to exclude a reaction dimension or component critical to reaction success. This would manifest in the optimization stalling at a suboptimal result, which could be a signal that the user needs to expand the search space.

"10) "First, we highlight the development of our BO platform using experimental reaction data mined from the literature. Next, to test our approach, we undertake a systematic investigation of BO compared to human decision making in the optimization of a new reaction – a key Pd-catalyzed direct arylation from the synthesis of BMS-911543 (Figure 1A)." The authors provide a sub-standard number of examples, with small search space sizes and number of optimizable variables. More specifically, two examples are purely retrospective, with search spaces of 3696 and 792 instances (far away from the authors' real world need of >10^6 – as in main text and Figure 1). The third example (search space of 1728) is also a Pd-catalyzed reaction presented in a retrospective configuration as the authors have performed all possible reactions beforehand through HTE. This is a troublesome option because it defeats the whole purpose of using active learning (compressing search spaces that cannot be fully enumerated experimentally, data availability and acquisition). Besides only probing mechanistically related reactions, the simultaneous use of HTE and small search spaces, renders all the test cases unsuitable to highlight the potential of the in silico method. These are examples in strictly controlled and non-challenging conditions that find little parallel in daily practice. As mentioned in 2), multiple, unrelated/diverse (in terms of mechanism, building blocks, potential conditions and real-world applications) and challenging (multiple variables, beyond what is intelligible to humans, and in HTE-untractable search spaces) prospective examples are warranted for a study published in Nature. The study also fails to answer relevant questions, such as is it possible to optimize reactions towards one regioisomer (if multiple products are possible in one reaction)."

Related to our response to points (2), (3), and (4) the small search spaces utilized for development in this study are for rigorous statistical validation. Given experimental limitations and the limited number of combinatorically complete datasets available in the literature this approach necessitated evaluation of smaller-scale optimization problems. However, we believe that this approach pay's dividends by enabling us to tune the optimizer with real chemical reaction data and answer important questions like "what is expected to happen in a typical optimization?" and "how does performance vary with initial starting data?" rather than "what could possibly happen in an optimization?".

In response to the reviewer's comment we carried out 2 additional real-world reaction optimizations for non-Pd-catalyzed reactions. Importantly, we included larger search spaces with 180,000 and 312,500 possible experiments including 7 process parameters each. Indeed, in both examples our findings demonstrate that BO delivers high quality experimental conditions which far exceed controls for standard reaction conditions. For the Mitsonobu reaction, EDBO identified conditions that afford 99% yield where the go-to conditions at BMS provided 60% yield for the specific reaction. Likewise, for the deoxyfluorination reaction, the Doyle lab's published conditions afforded 36% yield whereas EDBO identified multiple sets of conditions that delivered 69% yield. In both reactions BO identified sets of experimental conditions with largely distinct parameter settings from the standard conditions.

"11) "(…) and the often-complex relationship between reaction components and yield of the desired product. Importantly, the dimensionality and range of outcomes offer reasonable microsystems to study real-world problems in chemical process optimization." This is too vague and has no place in data-, hypothesis-driven research. How much is the "often-complex relationship" and what are the "reasonable microsystems"? Although this reads well, there is no quantification whatsoever of the size of the problem and it only adds noise to the message."

In response to the reviewer's comments we have removed these statements from the manuscript.

"12) Figure 1 highlights how non-challenging the first two examples are, in particular reaction 1 and 2c-e: small search spaces with a high number (naked eye) of well-performing reactions in low dimensional space (3 variables). I have concerns relative to the behaviour of the algorithm, taking into account the imbalance of reaction yield distribution in 2a,b relative to 2c-e. Admittedly, for 2a,b finding the optimum is not as simple, since random selection in 2c-e would result in >70% yield in ca. 1 out 2 or 3 picks. Probabilistically, it is less likely to achieve the same result in 2a,b. However, scrutinizing Figure 3F-H one finds exactly the same behaviour in all curves, for all compounds/reactions – this is impossible on absolute values. Indeed, the data is masked by a scaled YY axis, which means that a yield of 10% can be set as maximum in one case, whereas in another case a similar value in the YY axis might correspond to 90% yield. Data reporting is not transparent, oversells results and is misleading. In practice, a yield improvement of 1% (insignificant and inside the experimental error) might correspond to a jump of 10–20% in relative terms while only optimizing the reaction to trace amounts. Another aspect that is troublesome and raises a red flag in Figure 3 is that there are no kinks from batches that performed worse than all previous ones (irrespective of the added information). Such a result means that the algorithm always finds better performing reactions in all subsequent batches, independently of the uncertainty and selection policy. This also means that the starting points always correspond to poorly performing reactions, which I hardly find true in 2c-e given that random selection is used for initialization and there is a high percentage of well-performing reactions in those examples."

In response to the reviewer's probabilistic concerns: We demonstrate in the SI that BO outperforms random selection for each of these reactions. For convenience, we provide an example of the convergence plot for reaction 2b here:

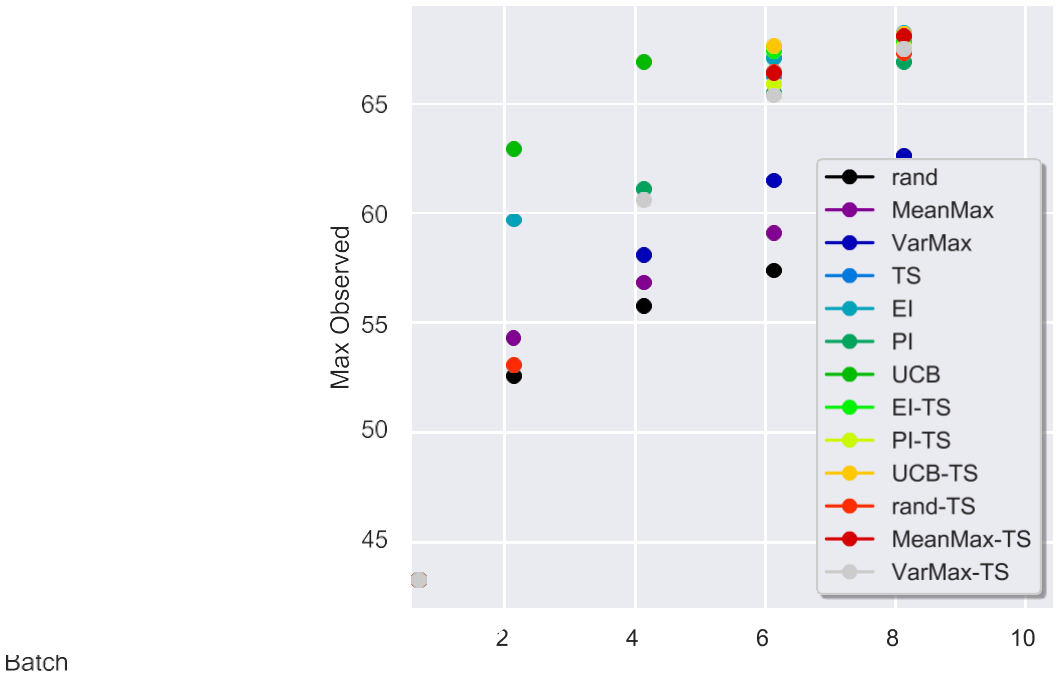

Figure R1. Acquisition function performance on development objective 2b. Average of 30 random initializations with a GP surrogate model, DFT encoding, and batch size = 5.

In addition, to aid in review we provide a summary of statistics supporting this statement for each of the reactions here.

	1	2a	2b	2c	2d	2e
p-value (H0: random = EI)	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
mean loss	4	3	9	3	3	6
worst-case loss	9	8	17	8	10	15
standard deviation	1.7	2.2	5.2	1.7	2.9	3.8

Table R2. Summary of random search simulation results for reactions **1** and **2a–e**: p-value is for a hypothesis test for equal average performance (N=30, batch size=5) the selected BO algorithm.

In response to the reviewer's comments we have included detailed performance statistics for all experiments presented in the manuscript (new Table 1, updated figure 3). We have also included axes ticks to address the reviewer's concerns.

In response to the reviewer's comment about the lack of "kinks": the plots depict the (cumulative) max observed yield for each batch (i.e., after running 2 batches of 5 experiments the plots show the best of all 10 yields). We thank the reviewer for bringing up these points of confusion and have updated figure captions to clarify the presented results.

"13) "(…) density functional theory24 (DFT) or cheminformatics descriptors generated using open source libraries." The manuscript is too vague and uninformative. The motivation is unclear and the final size of the descriptor vector is unknown throughout the manuscript and SI files. My understanding, from reading the SI file, is that thousands of descriptors were used to solve an optimization problem in 3 or 5-dimensional space. This disproportion is very questionable as most descriptors only add noise and do not contribute to the algorithm (e.g. what information can molar volume provide?). In addition, the mechanisms for the studied reactions are well known, which weakens the motivations of using all possible computable descriptors. A relevant, unanswered question is which studies did the authors conduct to ascertain the need for an oversized feature space, especially in light of the good performance of OHE and taking into account that no chemical insights were extracted while using DFT."

In this work we highlight optimization without explicit feature selection for the following reasons: (1) it is challenging to know a priori what features will perform well for a given system when little data is available, (2) it avoids bias, and (3) our approach is more user friendly to chemists with little to no modeling experience. Instead, we focused on tuning the surrogate model(s) to work well in the situation where n-descriptors > m-data (most features are irrelevant) by using automatic relevance determination and calibrated model priors. In response to the reviewer's comments we have added a methods section to the manuscript body which includes details about the encodings (including the size of the descriptor vectors used before and after preprocessing steps) as well as the rational for "implicit" feature selection with automatic relevance determination and priors.

Finally, we have indicated in the manuscript that the ultimate selection of DFT descriptors was made on the basis of performance statistics (new Table 1) relative to OHE and Mordred encoding. Although the differences are small, we found that DFT encoded descriptors gave the most consistent results in terms of worst-case loss and standard deviation. In response to the reviewer's comments we have also added a statement drawing readers attention to the fact that our findings suggest that good performance can be achieved with a number of encodings.

Added text: After independently tuning the optimizer for each data type, we found that the average loss for parallel reaction optimization using each encoding for all 6 reactions was ≤ 3% yield. The average outcome for each encoding was largely indistinguishable ($p > 0.05$, Table 1). However, we found that DFT encoded descriptors gave the most consistent results in terms of worst-case loss (≤ 5% yield for all reactions, versus ≤ 15% and ≤ 8% for Mordred and OHE respectively). Thus, we elected to carry out the remainder of our experiments using DFT descriptors. Importantly, our findings suggest that acceptable performance can be achieved in the wild with a number of reaction encodings.

"14) The manuscript is full of uninformative, fuzzy and general narrative, e.g. "(…) facilitate the screening and meta-optimization of the various components critical to achieving good performance with chemical reaction data." What are "various components" since, at best, only 5 variables were optimized; what is a good reaction performance when, in some cases, ca. 30-40% of all reactions

have yield >50% (Figure 2). The same applies to the following: "Bayesian reaction optimization begins by collecting initial reaction outcome data via an experimental design (DOE) or by drawing from existing results (…)" and "After training the surrogate model, new experiments in the reaction space are sequentially chosen by optimizing an acquisition function which maximizes the expected utility of candidate experiments for the next evaluation". How many data points? Which selection criteria when drawn from existing results and function? What is the meaning of "expected utility"? Also, "This process continues iteratively, until reaction yield is maximized, resources are depleted, or the space is explored to a degree that finding improved conditions is improbable (e.g. no predicted improvements and estimated variance is low over the entire space)." Which stop criterion for the latter case? Some of the information is available in the manuscript/SI, but is highly scattered and unorganized, making it extremely difficult to follow and immediately grasp what was done. Extensive rewriting and clarifications backed by data would be required in this regard."

We thank the reviewer for these comments and questions. We have clarified the text, reorganized the SI, and included a methods section which covers details in response to the above points. Here are the specific responses to the reviewer's comments in this section:

In this section, various components refers to the optimizer configuration. This included the encoding, surrogate model hyperparameters, acquisition function (which to use and what hyperparameters). In response to the reviewer's comments we have removed this statement from the text.

In this work, good reaction performance refers to minimizing loss (i.e., getting consistently closest to the global maximum in yield). In response to the reviewer's comment we have clarified our performance metrics and optimizer selection criteria with statistic tables and discussion (new Table 1).

Bayesian optimization refers to a class of algorithms which can utilize arbitrary amounts of data and encompasses a number of different selection criteria (acquisition functions). Thus, we believe the vagueness of this statement is warranted and well represented in the cited reviews of BO.

Expected utility refers to the process of taking the expectation value of a utility function which, as alluded to in the text, enables the derivation of many acquisition functions. The mathematical derivation of expected utility covered in the cited reviews.

"15) "(…) we selected surrogate model parameters based on regression performance for reactions 1 and 2a–e. Thus, we elected to take ourselves out of the loop and turn the optimizer on itself (See SI)." This is a nice feature but not new, according to a manuscript in ChemRxiv cited by the authors. The authors' fail to provide experimental evidence and explain how their approach advances the state-of-the-art."

Indeed, we used BO to optimize our model hyperparameters which is a common practice. Here we are only trying to convey that this was the method we used and that we utilized our newly developed software to do it. In response to the reviewer's comment we have included a statement about hyperparameter tuning in the newly added methods section.

Added text: The Matérn kernel shape parameter (5/2) and priors were selected via Bayesian optimization using EDBO.

"16) "One point of divergence between the models was traced to the reaction representations. For DFT and cheminformatics encodings, which contain hundreds of features for a given reaction (after decorrelation), the automatic identification of important dimensions is critical to model performance. This issue is exacerbated in the low data regime, where the model needs to learn the variation in each dimension from only a few data points." This is an important statement but the study does not answer how the method deals with the curse of dimensionality in low (how many?) data regimes. By starting with 5 random reactions and having a descriptor with hundreds of dimensions, the authors are not able to assure their approach is generally and effectively applicable to all/most chemistry optimizations. Many prospective examples designed to study this limitation are missing. Does it have to do with the granularity of the search space or with the number of starting reactions? Can it be mitigated by batch size in very diverse search spaces, or by constraining search spaces and how (?), or by tuning selection policies? There are too many unanswered questions that would need clarification to help the readership confidently adopting this method. Related to these questions, the authors state that "Thus, we assume that most of the dimensions are irrelevant, impose this belief via a gamma prior centered on longer length scales, and estimate parameters by maximizing the marginal likelihood of the experimental data." This is expected because of the overwhelming amount of descriptors, which include DFT computations. Although I am not arguing about the DFT method here, the authors should at least consider that better options might be available for specific cases and caution the readership that the presented approach is expected to fail, at least occasionally, either because of mis-representations or inappropriate descriptors."

Low data refers to the start of an optimization where only a few data points may be available. We utilize a combination of tuned priors and automatic relevance determination to facilitate learning and prevent overfitting. In response to the reviewer's comments we have included a more detailed discussion of the selected model in the newly added methods section. This covers our approach to mitigating overfitting in the low data regime. In addition, we have included 2 prospective examples which utilize DFT encodings and randomly selected initial experiments. These examples, along with our statistical analysis with HTE data demonstrate that our approach offers consistent performance in a number of settings (9 reactions for the whole study).

In this work we investigated three distinct methods for encoding chemical reactions and choose the best method for test reactions based on statistical evaluation of performance. The reviewer is right in pointing out that other encodings may exist which give improved performance. Indeed, our finding is that many representations can give good optimization results. In response to the reviewer's comment we have included a statement in the manuscript explicitly highlighting this finding.

Added text: The average outcome for each encoding was largely indistinguishable ($p > 0.05$, Table 1). However, we found that DFT encoded descriptors gave the most consistent results in terms of worst-case loss ($\leq$ 5% yield for all reactions,

versus ≤ 15% and ≤ 8% for Mordred and OHE respectively). Thus, we elected to carry out the remainder of our experiments using DFT descriptors. Importantly, our findings suggest that acceptable performance can be achieved in the wild with a number of reaction encodings.

"17) Something that is not clear (perhaps because of the scattered information) is for which species are all these descriptors calculated? In the SI, the authors mention "reaction representations based on Density Functional Theory (DFT), Mordred fingerprints" but there is no additional information besides that, in the latter case, >1800 descriptors were calculated for each reaction component. Reading the methods further and inspecting Figure S6 it seems the procedure is identical for DFT descriptors. This constitutes a major design flaw as no rationale can be identified. Given that the examples provided here only optimize 3 or 5 reaction variables and the starting materials are fixed throughout, it is irrelevant to calculate descriptors for those species since there is no variance."

Descriptors were only computed for reaction components which were varied. Any singular or highly correlated descriptors were removed in preprocessing. In response to the reviewer's comments we have included a method section discussing featurization and preprocessing of descriptors. In this section, we include an example of the encoded components for reaction **3**.

"18) "The central tenet of BO is the utilization of both information (exploitation) and uncertainty (exploration) to drive optimization." and "In contrast, a pioneering acquisition function (pure exploration), exemplified by selecting the point of greatest predictive uncertainty for evaluation, will tend to investigate the entire response surface more thoroughly." This oversells the authors' method and falsely hints at added utility/innovation over previously reported heuristics (some examples provided in comments above). The authors exaggerate on the utility of BO because exploration and exploitation (alone or in combination) are the stepping-stone of active learning (e.g. Drug Discov Today 2015, 20, 458 and references in this review), irrespective of the employed ML algorithms. Indeed, the approach reported herein is not innovative, new nor pioneering. I also argue that exploitation hardly provides added value in terms of information gains, because it focuses on identifying experiments meeting a given target criteria rather than improving the overall model performance (as per Drug Discov Today 2015, 20, 458 and examples in the review)."

Here we present examples of pure exploration and exploitation as a teaching opportunity to help chemists visualize the differences between algorithms. The reviewer is correct that exploitation/exploration is not unique to BO. In response to the reviewer's comment we have included a comment highlighting that exploration and exploitation are central to active learning in general. In addition, we have updated figure 3 to clearly show how BO balances exploration and exploitation in the Suzuki search space.

Added text: The central tenet of BO (and active learning methods in general) is the utilization of both information and uncertainty to drive optimization.

"19) "Over the course of 100 experiments the scores of the explorer and exploiter diverge, with the explorer offering a significantly better fit to the reaction surface. This strategy could be advantageous for users who hope to gain a global understanding of a reaction space." Although

this is an interesting observation, there is no quantification in the manuscript of how much the scores diverge and when they start diverging. The manuscript is not objective and clear regarding which example(s) it refers to. Performing such a comparison over 100 iterations is extreme and questionable, depending on the use case. 100 iterations correspond to 3% and 13% of the search space for reaction 1 and 2, respectively, which makes it impossible to generalize findings."

In response to the reviewers comments we have rerun the analysis for figure 3 from a different initialization point, including an optimization path for EI (to illustrate how it balances exploration and exploitation), and run the optimizer for only 50 experiments.

"20) The paragraph between lines 188–207 is too cryptic and the general readership will not understand the message. This should be presented with terms just above layman's level. Namely, the EI and TS scores should be more clearly explained. The following paragraph suffers from the same pitfalls. It is too generic and wordy for the small amount of scientific information it conveys. At the end of the paragraph the authors mention: "That is, running fewer experiments at a time results in more efficient use of resources by using information more effectively." The authors fail at identifying the reasons for this rather undesired event. On one hand, it suggests the deficiencies in the selection policy that was implemented. For batch selection, one would expect that all experiments are decorrelated from one another, in the decision function, as to provide the maximum, non-redundant information to the algorithm. Apparently, this is not the case. Other methods (e.g. Chem Sci 2016, 7, 3919) achieve this seamlessly, evidencing that the work developed herein does not reach the expected level of conceptualization. On the other hand, if the selection criteria were correctly implemented, the result indicates the small diversity in the search space. Any of these situations has implications in the conclusions one may draw from the utility of the heuristics and shows that multiple experiments are missing or were incorrectly planned."

In response to the reviewer's comments we have included a more detailed discussion of the EI acquisition function in the newly added methods section. We have also edited the language to focus more on outcome rather than the tradeoff between batch size and performance which is not exclusive to this application. With larger batch sizes the model has less data available to make its decisions. Consider the case in which you are optimizing the same reaction with batch sizes of 1 and 5. Consider further that both optimizers have 5 initial data points. Both the sequential and batched optimizer have 5 data points on which to decide the 6^{th} experiment. However, for the 7^{th} experiment, the sequential optimizer has 6 training data points while the batched optimizer still only has 5.

"21) The model fit score (Figure 3B) is appalling throughout the runs, even for an exploration policy ($r^2 < 0.5$). This is quite unexpected and beyond comprehension unless the exploration policy is not serving its purpose of improving model performance. Seemingly plateauing at experiment 60+ is even more worrisome because the model is not learning anything (or very slowly), despite the poor correlations between predictions and ground truth."

Reaction outcome prediction is a hard problem. Thus, the fit scores presented in figure 3B are consistent with the learning curve (presented in the SI) for reaction **1**. In fact, the results presented in figure 3B suggest that exploration may give an improved learning curve over random selection. For convenience, we present learning curves for all 6 development

reactions below. In these curves, *N* points (x axis) are used to train the model which is then used to predict the outcome of the remaining data points and compute the R^2 value.

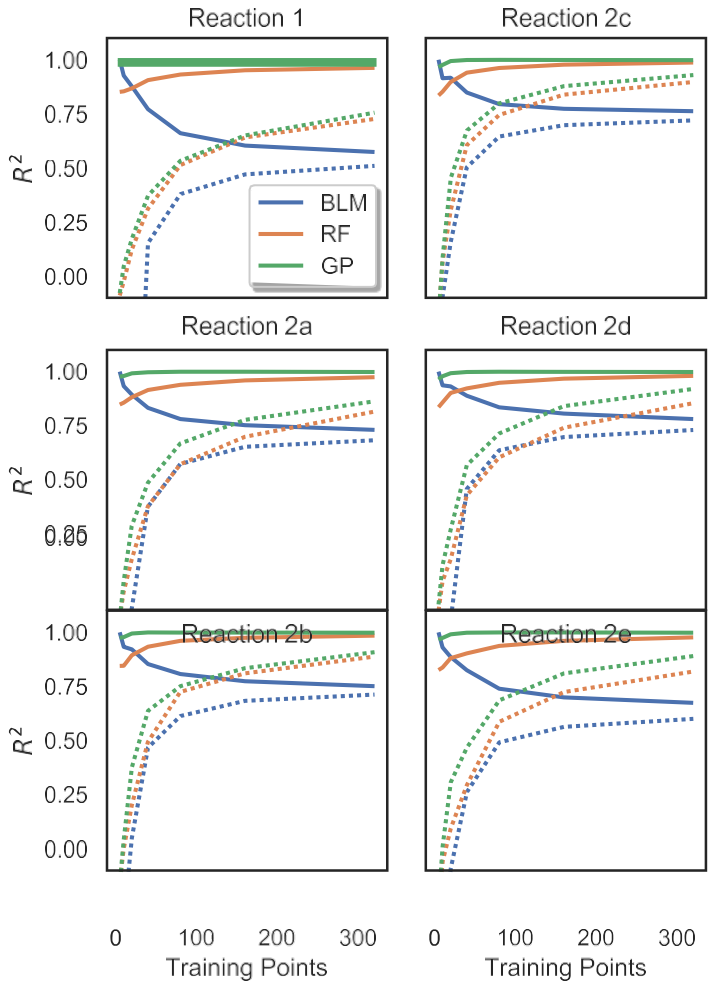

Figure R2. Learning curves for development objectives with Bayesian linear (blue), random forest (orange), and gaussian process (green) models. Training and test scores are solid and dashed lines, respectively. Each curve point represents the average of 25 random splits.

In response to the reviewer's comment we have updated figure 3 to include a clearer example and simulations over only 50 experiments.

"22) Data is presented in a poorly, uninformative and potentially misleading manner. At naked eye, the exploration trace (dotted line) in Figure 3B does not fit any of the plots in the subsequent panels. If the YY axes in panels C–H simply correspond to the max yield within a given batch, one expected to find downward curves, occasionally. For example, one of the first experiments has a yield of almost 100% (Figure 3B exploration). Only after ca. 20 experiments can one find a yield that is higher. This would correspond to a few batches and a downward curve should be visible in the corresponding panel (potentially C?). If the YY axes correspond to the maximum cumulative yield, then the curve would plateau for a few batches. All exploration curves in panels

C–H show a steep rise in the first few experiments, which is not supported by the remainder experimental data."

We apologize for the confusion derived from our presentation of the data in Figure 3. Figure 3B only corresponds to figure 3A. The other optimization curves represent the average cumulative yield over 30 randomly initialized simulations. However, for any given curve a plateau or no change due to random selection of the highest yielding reaction may occur. In response to the reviewer's comments we have clarified the figure caption.

"23) The provided examples are not challenging and do not serve as proof-of-concept. Besides the small search space (Figure 3A), which I have debated elsewhere, the unsupervised learning algorithm shows the small diversity of all possible reaction conditions, as evidenced by a small number of mini-clusters (ca. 40–50). These can be further reduced, according to k-means, and as colour-coded by the authors. In practice, the results yield a manifold reduction in difficulty for identifying productive reaction conditions."

t-SNE is largely a visualization method which attempts to cluster points that are similar in an aesthetically appealing way. Thus, this plot should not be taken as a measure of the complexity of the search space. Instead, we sought to visualize the selection of experiments as a way to teach the exploration/exploitation tradeoff. In response to the reviewer's comments we have included additional prospective examples of higher dimensional optimization over large search spaces.

"24) "(…) we found that both parallel EI and TS gave impressive performance, on average finding optimal configurations in only 50 experiments including 5 random initialization points." and "We were surprised to find, for many of the reactions, pure exploitation gave good results". The authors' observations fail to advance the field and add any relevant information relative to previous reports. Identifying optimized reaction conditions is an achievement, but in this particular case, it fails to at least match what other colleagues have reported (reference 21). What the authors report here is an optimization requiring the pseudo-execution of 1.4% and 6.3% of all possible reaction configurations. This falls short of <0.5% required elsewhere. The present result is also questionable because of the restricted search space, with many similar reactions (Figure 3A), and in light of my concerns regarding the number of optimizable variables. Multiple prospective examples would have been required to confidently ascertain the potential utility of the computational approach. It is surprising that the authors did not at least consider the possibility of exploitation providing good results. My argument is based on 1) an identical observation in reference 21 – which the authors fail to acknowledge – and 2) the high number of well-performing reactions in the HTE screens. Because the authors did not design any experiment to disconnect the observation from the second hypothesis, it is impossible to conclude that exploitation works as a general approach or is indeed an unexpected result in this particular case. Moreover, speculating that "this could indicate that an objective is relatively convex and thus the optimizer is less likely to get trapped in a local maximum" without designing and executing the appropriate controls for such a potentially important justification is not helpful and falls short of the evidence-based science one expects to find in Nature."

Old reference 21 offers a related but different approach to optimization and validation. Most importantly, 21 involves optimization of only continuous variables which are expected to give relatively smooth changes in response (unlike the categorical variables included here), simplifying both modeling and validation. The difference in experimental requirements (<0.5%) for algorithm convergence can be explained by the exclusive optimization of continuous variables alone. Furthermore, given the smooth response to parameters like time, the grid size of the search space can be expanded arbitrarily. In terms of validation, a difference between old reference 21 and this work is statistical relevance. Whereas different starting points in the search space may give very different optimization paths and overall performance, in reference 21 the algorithm is run from a single set of 10 randomly selected experiments. This provides no measure of reproducibility and variance to the initial choice of reaction conditions. In response to the reviewers comments we have included 2 additional prospective examples in which high quality reaction conditions were identified by BO in around 0.02% of the reaction space.

We have included in the SI experiments comparing random search to BO which deconvolutes the high number of well-performing reactions in some of the HTE datasets from the performance of BO. We highlight these findings above in table R2. We found that BO outperforms random search in all cases. In response to the reviewer's comments we have removed speculation about the concavity of the reaction's response surface from the text.

"25) "In contrast, chemists typically only implicitly incorporate uncertainty in an optimization campaign, for example by considering a few conditions which do not fit conventional wisdom." This is a major flaw in the design of the research and again, constitutes a narrative that is built to fit the needs of the study. The claim is unsupported by data/references and establishes a ground truth based on a subjective opinion. Motivating work on these grounds is not acceptable."

In response to the reviewer's comment we have removed this section from the manuscript.

"26) "Reaction optimization truly begins by defining the search space. If the search space is too large, containing many irrelevant or redundant configurations, significant amounts of time and resources can be wasted. If the space is too small, one risks excluding relevant experiments from consideration. We pursued a hybrid approach to selecting candidates, wherein we first consider the larger set of plausible experiments, then quantify similarities between potential reaction conditions via unsupervised learning and select those which ensure a satisfactory distribution across the larger search space." The study runs under highly controlled, curated and non- challenging conditions and does not answer the identified research question. I do understand why the authors have taken this approach – reducing the search space to a number of combinations tractable to their HTE setup. The approach however raises serious questions regarding the study design, contradicts the research question and the potential utility of the algorithm. Therefore, one cannot conclude if the approach is useful for a subset of examples only or can be scaled. The third use case also counter cycles the identified need of mining millions of possible reactions configurations. Given that the search space is condensed to 1728 reactions, HTE is employed as enabling technology. In these settings, active learning is redundant because all search space can be enumerated and outputs obtained. Performing active learning on a dataset for which all outputs

in the search space are known (besides being similar to the previous two) does not qualify as a challenging and real-world scenario. The target molecule is of therapeutic interest, but given this irremediable limitation – in the study design and purpose of using active learning – the example does not work as a flagship or convincing application."

The reduction of the search space was necessary to collect all experimental data required for our statistical analysis. As mentioned before, this enabled us to rigorously evaluate statistical variation and carry out hypothesis testing in an experimentally tractable way. In response to the reviewers comment we have clarified this point in the manuscript. In addition, we have included 2 prospective examples in which exhaustive experimentation is not feasible.

"27) "Importantly, the final selection of reaction components was done based on valuable knowledge rather than ML". The search space is highly customized and built with domain knowledge. I find no motivation possible for using unsupervised learning and then curating results with potentially skewed domain perception. This has high potential to introduce biases and is detrimental for model building (Nature 2019, 573, 251). More specifically, the use of k-means for ligand curation is interesting but the actual selection of 12 representatives is convoluted. Multiple examples from a same cluster are prioritized, whereas other clusters are neglected (e.g. pink dots in Figure 4). Also, the search space is configured to include unevenly spread temperature values (90, 105 and 120 ºC) and a range of concentrations (0.057, 0.100 or 0.153 M) that can only be motivated with extensive prior experience. Therefore, no evidence of utility under unknown reaction spaces is provided. This constitutes a major weakness of the study, as the approach is incompatible with the democratization of chemical syntheses, i.e. expert domain knowledge is required. When also considering other arguments presented above (small search space, low number of optimizable variables and ability to perform all reactions in parallel before the actual active learning run) this and all previous examples are deemed unsuitable for a manuscript in Nature."

As practicing synthetic chemists, we appreciate the value of incorporating domain expertise to solve optimization problems. As an example, one could argue that since technical execution (e.g., order of addition, mixing time of catalyst and ligand, stirring vs shaking) also represents a set of parameters even "fully automated" (chemputer, flow) methods incorporate human expertise. Thus, the optimizer is designed to work with an expert user, not autonomously. Accordingly, we consider the ability to incorporate domain expertise (e.g., in this case the exclusion of bidentate phosphine ligands – pink points - due to the understood mechanism and the cited study which demonstrated the oxidation of a bidentate phosphine, rendering it a mono-phosphine, when employed in a related reaction).

The spread of temperature values is even: 105 – 90 = 15 and 120 – 105 = 15. The concentration range is informed by existing experience, again as a way to incorporate domain expertise.

In response to the reviewer's comments we have carriedout 2 new reaction optimizations. The 2 prospective examples provide evidence that our approach is scalable and could enable application to large search spaces which are not constrained to incorporate domain

knowledge. In addition, as noted in the manuscript, the optimal conditions identified by BO for the Mitsunobu reaction defy conventional wisdom about the impact of temperature and concentration on the reaction outcome.

"28) "This was not surprising given the ML optimizers initial choices were made at random." The message conveyed by the authors is confusing and contradicts the remainder of the manuscript. Previously, the manuscript highlighted the utility of BO, stating that it can balance exploration and exploitation strategies in the selection process. Yet, here the authors say that the initial choices were made at random as a reason for the better performance of humans – this is inconsistent. Rather, the result is indicative that a potential benefit of the method can only be found when the reaction budget is high and amenable to some sort of parallelization, which is not always the case in an organic chemistry laboratory. The result is also indicative that the method inefficiently deals with optimization problems in rather low dimensional space (5 optimizable variables) when the descriptor vector has hundreds or thousands of dimensions. For a reasonable performance, the number of descriptor dimensions has to approximate the number of reactions performed (as indirectly cautioned elsewhere in the manuscript). Overall, the number of unanswered questions this manuscript raises shows that multiple experiments are missing and casts doubt on the utility of the algorithm on a wider scale."

We apologize for the confusion. Our comment was directed at comparing the first batch of experiments selected by chemists versus the BO, where random selection was used to select the first 5 experiments by design (to statistically validate the methods against variations in initial data). Bayesian optimization requires data to make decisions. Thus, before the algorithm can run, we need to collect some initial data. In the case of zero data, exploitation is of course not an option. In subsequent batches, we show that BO can effectively utilize small amounts of data to make better decisions including in the low data regime. In response to the reviewers comment we have adjusted the text to clarify this point.

Added text: Humans made significantly ($p < 0.05$) better initial choices than random selection, on average discovering conditions which were 15% yield higher in their first batch of experiments. However, even with random initialization within 3 batches of five experiments the average performance of the optimizer surpassed that of the humans.

"29) "Finally, while only around half of human participants achieved within 1% of the optimal yield for reaction 3, the optimizer achieved >99% yield 100% of the time." The manuscript over-interprets data. There is no evidence that 99% or 100% yield are indeed different. No replicates and statistical analyses were performed to ascertain the difference. The comparison is also meaningless without mentioning after how many iterations it was made and without a statistical test. Finding that optimized conditions can be obtained within 50 retrospective experiments (3% of the search space) is interesting, but not ground-breaking in light of the authors' reference 21."

We have adjusted the text to reflect the reviewer's comments.

Added text: Notably, in contrast to human participants, BO achieved >99% yield 100% of the time within the experimental budget.

30) The study fails at providing chemical insight or mechanistic interpretations to their findings, falling short of the depth expected in a Nature manuscript. For example, there is no discussion onto the use of CgMe-PPh as a ligand, how this is potentially unexpected and disrupts with established knowledge and how it can advance the state-of-the-art. It is also unknown if the ligand is only applicable to this reaction or is transferrable to similar building blocks – a short substrate scoping study is missing. Overall, using an algorithm as a black box, without attempting to interpret its predictions and augmenting human knowledge is opposite to how the field of machine learning is currently evolving.

In this study, we present BO as a tool for reaction optimization. As such, we do not emphasize the interpretation of ML models. Indeed, both a mechanistic inference and an exploration of the reactivity of CgMe-PPh would be very interesting. However, we feel that these studies fall outside of the scope of this manuscript.

"31) The authors make comparisons between human intuition and their BO heuristics. I argue their study design is flawed as not all variables are accounted for. Because different initializations are implemented, there is a possibility of having very different starting points for optimizations. This can include reactions that only performed very poorly, very similarly (and thus not informative), or sampling different regions in the search space (ideal scenario). This can affect the optimization efficiency of humans and thus, lead to skewed comparisons and unrealistic statistics. In my opinion, a correct design would have been tiered: First, with the same initialization for all researchers, ask them to optimize the reaction. This could inform on optimization paths, biases and ability to identify a local/global optimum. Secondly, with the same initialization path, allow the heuristics optimize the reaction. This could inform on differences in strategies (if a random state value is set, this could be determined), and ability of the heuristics to out-perform humans in that particular case. Finally, allow the heuristics optimize the reaction with different initializations. The study includes only the final experiment and the preceding questions are unanswered, yet relevant for a thorough comparison."

These are all good points. Indeed, we sought to study the variation in optimization outcome to very different starting points by using random selection. Addressing the reviewer's point, in the SI we show that initialization the optimizer using the initial choices of humans provides little change in average outcome.

"32) The supplementary information is superfluous in common ground information for colleagues in the machine learning field and simultaneously cryptic for those outside. The SI has no seemingly logical guiding line – lacks detail or is scattered – which made it very difficult for me to follow and find the information I needed. Re-wording, abbreviating and re-organizing are largely required. For example, there is no irrefutable motivation for using DFT descriptors over others (e.g. OHE). There are no statistics for figure 8 and similar/related analyses throughout the SI file. The authors argue in favour of structurally meaningful encodings but fail at capitalizing on that option as no new knowledge is generated or extracted. There are no comparisons to DoE tools, other BO and machine learning pipelines. "contain hundreds of features" and "hundreds of hyperparameters" – how many? In page 26 the plot suggests OHE works better – is there a meaningful difference to this? Can it be quantified or motivated to go for DFT instead? Figure S26

has nothing to do with OHE as mentioned in page 27, yet a comparison to DFT is made. It is not clear how redundancy in selection is dealt with. To substantiate that the model is learning meaningful patterns it would be useful to show that it can consistently suggest reactions that afford no product or only trace amounts. There are not enough experimental details in the chemistry front – for data transparency, please provide chromatograms for all experiments in an active learning run, at least for one of the examples. Figure S47 shows that the authors' preferred setup (Gaussian processes and DFT descriptors) is never the best option for any given analysis and shows that there is no generalizable solution to what is being studied. This provides preliminary evidence that their algorithm will underperform consistently while in the wild. Gaussian processes with linear regression works better for low data (page 34) and when the number of experiments is <75 it performs equally well as non-linear regression – why not use it throughout and choose the poorer and most convoluted solution? Similarly, if the preferred method works better with higher data regimes, why did the authors start from 5 training examples?"

In response to the reviewer's comments we have included a list of figures and tables at the beginning of the SI. We have also included discussion and detailed performance metrics (Table 1) in the manuscript which clearly motive the use of DFT encoding. In general, we have adjusted the focus of the manuscript to performance by removing discussion about structure-based encodings. We have also included a detailed study comparing the performance of DOE and a method section which gives more detail about the encoding methods.

In response to the reviewer's comment about experimental data we have included example assay data for each of the reactions in the SI.

The plot on page 26 referred to the optimization of hyperparameters for cumulative regression score with a small training set. In this case, OHE indeed achieved a higher score. However, the ultimate selection of encoding was done on the basis of optimization performance. We have presented relevant statistics in the manuscript in table 1.

"33) I commend the authors for providing code and documentation. I found no issues but in some instances it wasn't intuitive for someone who wants to run in Python. There are too many options, reinforcing that there is no streamlined and general path to using this BO method."

Our software contains many arguments which enable customization. However, basic usage requires the user to specify only a search space and the number of experiments they wish to run. In response to the reviewers comment we have included example use cases for the software on GitHub. For example, Jupyter notebooks used for the Mitsunobu and deoxyfluorination reaction optimizations are included.

Minor issues:

"1) The study is prolific in abbreviations: RF, GP, TS, BO, ARD, DoE, ML, HTE, EI. This makes it very difficult to follow the narrative if not in the machine-learning field. No more than 3–4 abbreviations should be found in the manuscript."

Many of these are mentioned so many times that we believe an abbreviation is necessary. However, in response to the reviewer's comment we have removed ARD from the list of abbreviations.

"2) The authors fail at citing relevant studies interfacing active learning and drug discovery. This would have been somewhat expected because the active learning concept is not new and there is a reasonable body of evidence that it can be applied to diverse tasks in early drug discovery. While I have no particular issue with the authors citing several books and reviews, these are too uninformative for a reader to follow up. I suggest substituting some of them by original work. For example, reference 29 could be substituted by one of many studies published by the Aspuru-Guzik lab. There are also numerous applications of random forests in chemical sciences (ref 28)."

In response to the reviewer's comment we have included citation for a review of active learning in drug discovery (39). Old reference 29 is considered the authority on Gaussian processes and thus an excellent reference for anyone interested in these models.

"3) There is seemingly no relationship between Figure 1C and the "(...) acquisition function which maximizes the expected utility of candidate experiments for the next evaluation (...)". Captions to Figure 1 do not guide the reader through the overwhelming amount of information in each panel. This is an issue also present in other figures."

In response to the reviewer's comment we have edited all figures in an attempt to make them clearer. For example, for figure 1C we explicitly label the expected utility function and the resulting next selected experiment.

"4) "To demonstrate this dichotomy, we tracked the exploiter and explorers' decisions (after initialization at the same point) in a two-dimensional representation of 1 (Figure 3A). Indeed, over the first several evaluations, the exploiter remained in a single cluster while the explorer traversed the entire space." The result is quite interesting and corroborates previous findings, namely in ChemRxiv 2018 (reference 21 in this manuscript). Would have been nice to discuss this further in light of independent reports."

In response to the reviewer's comment we have included a citation for old reference 21.

"5) "Humans made significantly better initial choices than the optimizer, on average discovering conditions which were ~15% yield higher in their first batch of experiments." Please provide statistics."

In response to the reviewer's comment we have included a p-value for this statement.

"6) How did the authors make sure the survey participants did not communicate with each other?"

This was explicitly stated in the rules. However, no attempt was made to ensure that participants weren't communicating. In response to the reviewer's comment we have included a comment indicating this in the SI.

Added text: Note: while participants were asked to not share results no attempt was made to ensure that they were not communicating.

"7) The 13C chemical shift for the CH3 is missing in the list (page 55 of SI) – it should be 32.8 ppm. The authors assign peaks to carbon atoms but I find no label (C1, C2, etc) in the structure. Please add those labels to the structure embedded in the 13C NMR spectrum."

We thank the reviewer for identifying the missing data. In response to the reviewer's comment we have included it in the SI. For consistency with the remainder of the data we have removed carbon labels from the tabulated NMR data.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

In the original assessment of the manuscript I had been quite critical regarding some aspects of the data (that was presented in a fuzzy way or was missing) and language, including omissions of prior art or contextualization in light of other relevant work. Although I still have reservations regarding the innovation of the algorithm, the authors have now made it clear in the manuscript that the main goal is answering an utility research question; I think the study now excels at that task as I highlight below.
Overall, I am pleased with the responses that were given to my concerns. Although in some instances I was hoping the authors would have run an extra mile (e.g. interpretation), in other cases I had misinterpreted the manuscript and I am grateful for the clarifications. I am particularly happy with the work that was done regarding point 5), figure quality, rearrangement of the SI and, most importantly, with the new examples. Not having prospective examples incompatible with HTE was a weak point from an active learning utility point of view. In the new figure 5 this major concern is appropriately addressed. The reported method is able to identify optimized conditions that are different from pre-established reaction protocols. Not only that, the ML method is able to identify improved protocols by suggesting a minute amount of experiments relative to the whole search space – this is impressive.
I now resonate with Reviewer 2 in conceding this study does a very good job at educating the broader audience and provides enough examples to convince about the utility of the tool in real-world scenarios.

Minor comments the authors might want to consider:
Figure 1: edit "expected utility" to "expected utility curve".
Figure 5: It would be very helpful to color code the dots in both plots according to the model uncertainty. This can avoid unnecessary criticism suggesting the method is failing by always recommending reactions that underperform. Those are likely the reactions least understood by the model and it is perfectly acceptable (and desirable) to record such "negative" data.

Author Rebuttals to First Revision:

My coauthors and I are grateful to the reviewer for the comments and all efforts in helping us prepare an improved manuscript. Below we provide responses to the reviewer's second round of comments:

(1) "Figure 1: edit "expected utility" to "expected utility

curve”.” In response to the reviewer’s comment we have

edited figure 1.

(2) “Figure 5: It would be very helpful to color code the dots in both plots according to the model uncertainty. This can avoid unnecessary criticism suggesting the method is failing by always recommending reactions that underperform. Those are likely the reactions least understood by the model and it is perfectly acceptable (and desirable) to record such “negative” data.”

We thank the reviewer for the suggestion. BO attempts to balance exploration and exploitation. According to the reviewer’s comment, we prepared a plot that included a color map for the computed expected improvement values. Unfortunately, we found that the plot because harder to digest by including this additional information. Instead, we have provided a method (edbo.bro.BO.acquisition_summary) for users to compute estimated variance for selected experiments and examples of the EI surface for the example reactions on GitHub.

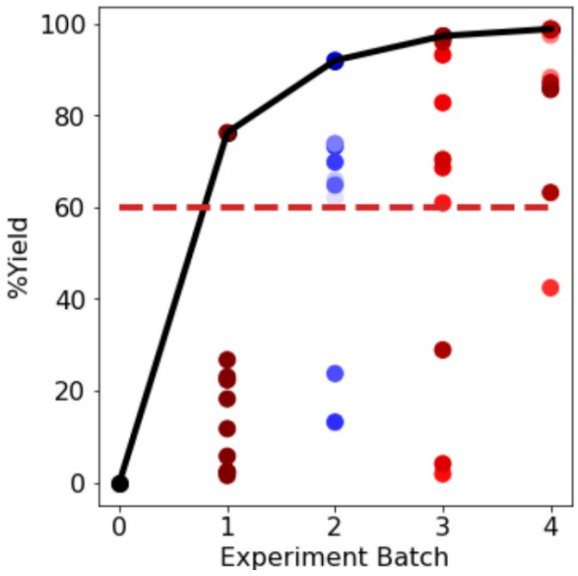

Figure R2-1. Mitsunobu optimization curve with points color coded by expected improvement. Color map: low (red) → high (blue).