

MIT Open Access Articles

Opportunities for Machine Learning and Artificial Intelligence to Advance Synthetic Drug Substance Process Development

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Griffin, Daniel J, Coley, Connor W, Frank, Scott A, Hawkins, Joel M and Jensen, Klavs F. 2023. "Opportunities for Machine Learning and Artificial Intelligence to Advance Synthetic Drug Substance Process Development." *Organic Process Research & Development*, 27 (11).

As Published: 10.1021/acs.oprd.3c00229

Publisher: American Chemical Society

Persistent URL: <https://hdl.handle.net/1721.1/158179>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution-NonCommercial-NoDerivatives




Opportunities for Machine Learning and Artificial Intelligence to Advance Synthetic Drug Substance Process Development

Daniel J. Griffin, Connor W. Coley, Scott A. Frank, Joel M. Hawkins, and Klavs F. Jensen*

 Cite This: *Org. Process Res. Dev.* 2023, 27, 1868–1879

 Read Online

ACCESS |

 Metrics & More

 Article Recommendations

ABSTRACT: The goals of this Perspective are threefold: (1) to inform a broad audience, including machine learning (ML) and artificial intelligence (AI) academics and professionals, about synthetic drug substance process development, (2) to break down the general synthetic drug substance process development task into more tractable subtasks, and (3) to highlight areas in which machine learning and artificial intelligence might be beneficially developed and applied. Application of machine learning and artificial intelligence to chemical synthesis of medicinal compounds has long been discussed and has resulted in the development of a number of computer-aided synthesis planning tools by both academic groups and commercial enterprises. The focus of these efforts has primarily centered on the task of retrosynthetic analysis, as seen from the perspective of a medicinal chemist. This has left significant unrealized opportunities in the application of machine learning and artificial intelligence to aid the process chemist or engineer in commercial drug substance process development.

KEYWORDS: *drug substance, accelerated process development, machine learning prediction, computer-aided synthesis planning, route selection and optimization*

1. INTRODUCTION

Bringing a new drug to market is a costly endeavor, with estimates in excess of 2 billion dollars and timelines spanning into decades.¹ This can be attributed to several factors, including the process of discovery, which must go through design–make–test–analyze (DMTA) cycles for large pools of candidates, from which only a few molecules will be successfully moved into clinical trials. As or more significant are the cost and time of conducting clinical trials and the cost of developing and gaining approval for the regulated manufacturing process to supply clinical trials and then patients throughout the world.

In a previous article,² we reviewed the DMTA drug discovery cycle and discussed areas in which artificial intelligence (AI) might be beneficially applied to medicinal chemistry, e.g., with computer-aided synthesis planning (CASP). In this Perspective, we focus on developing a suitable synthetic manufacturing process for clinical and commercial production of drug substance.³

In our experience and our reading of the relevant literature over the last 15 years, we see significant unrealized opportunities for machine learning (ML) and AI to enable more efficient drug substance process development. We suspect that these opportunities have not been in focus because of the complex, multifaceted, and iterative nature of synthetic drug substance process development; such tasks, in their vast formulation, are hard to pin down into confined supervised learning problems that can be tackled with ML or narrow AI. We have three goals in this Perspective: (1) to inform a broader audience, including machine learning professionals, about synthetic drug substance process develop-

ment, (2) to break down the general synthetic drug substance process development task into more tractable subtasks, and (3) to highlight areas in which ML and AI might be beneficially applied.

1.1. Breaking Down the Task of Synthetic Drug Substance Process Development. The overall task of process development is complex.⁴ To gain traction and provide problem formulations that may be more readily tackled with ML, we break down the synthetic drug substance process development task, as shown in [Figure 1](#). This divides process development into two major subtasks: (1) Route Optimization and (2) Process Optimization. Within each of those, we identify sub-subtasks: four for Route Optimization—Route Mapping, Route Narrowing, Route Scouting, and Route Selection—and two for Process Optimization—Problem Formulation and Search. While different pharmaceutical companies approach these subtasks with variations in their workflows, this outline defines the general requisite tasks.

1.2. Outline of This Perspective. The Perspective is structured as we have broken down the process development task: In [section 2](#) we take a deeper look at Route Optimization, and then we turn our attention to Process Optimization in [section 3](#). [Section 4](#) then ends with some overarching conclusions and an outlook for the application of ML and

Received: July 13, 2023

Published: September 25, 2023



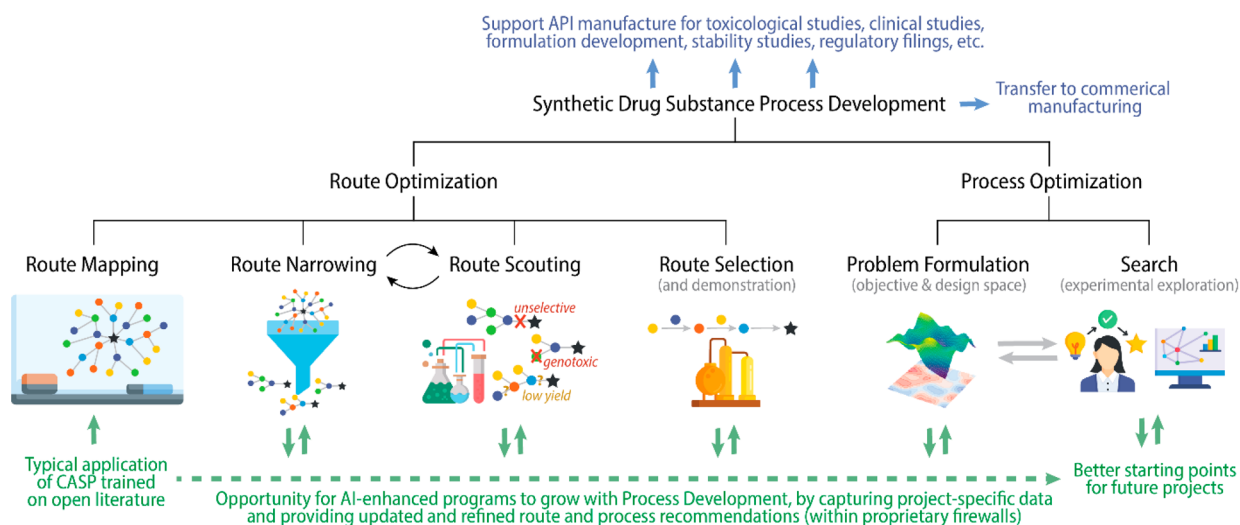


Figure 1. Hierarchical view of the stages and categories of tasks involved in process development. Throughout process development, it is necessary to continue to supply material while also pursuing a more optimal route. This Perspective highlights the many opportunities for AI-enhanced development (green) to support each stage of process development (Icons obtained and/or adapted from Flaticon.com).

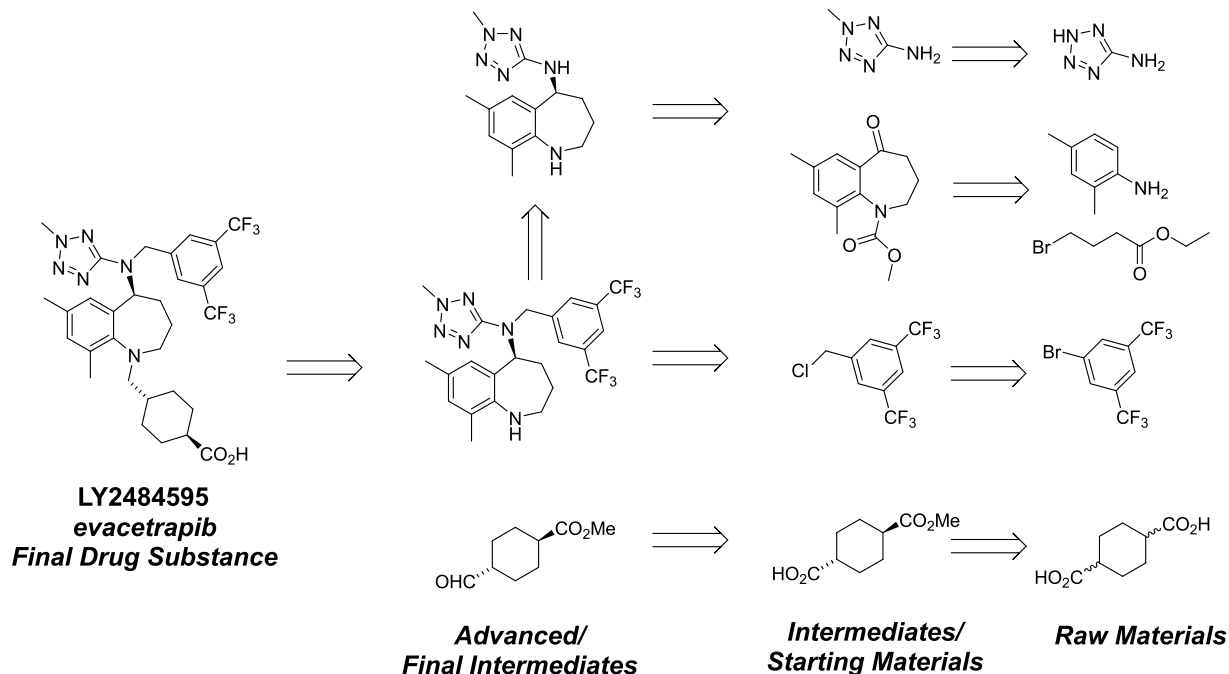


Figure 2. Retrosynthetic approach for the cholesteryl transfer protein inhibitor evacetrapib. Adapted from ref 9. Copyright 2020 American Chemical Society.

AI, including valuable applications that exceed current methodologies and thus invite further advances to the art.

2. ROUTE OPTIMIZATION

In the Route Optimization phase of process development, the goal is to identify the best synthetic route to a given drug substance.⁵ In this goal, a *route* is defined by the selection of the raw starting materials (SMs) and the sequence of drug substance intermediates (DSIs) building from the selected SMs to the final drug substance. A viable route is one that is achievable by applying feasible chemical reactions and isolations to transform SMs into intermediates, intermediates into more advanced intermediates, and the final intermediate(s) into the drug substance.⁶ In their current state, CASP

tools^{7,8} are most useful at the ideation stage (Route Mapping), where precise details of conditions and protection/deprotection chemistry are less relevant. For certain targets, the large number of pathways generated likely needs to be clustered or grouped for manual review along with experimentation to identify viable and efficient routes from those proposed (Route Narrowing).

As an example, Figure 2 presents a proposed convergent route (shown retrosynthetically) to the complex cholesteryl transfer protein inhibitor evacetrapib (shown at the left) from five structurally simple and commercially available raw materials (shown at the far right). As the starting point, access to commodity raw materials is critical and essential to a viable process chemistry route. The SM building blocks shown in

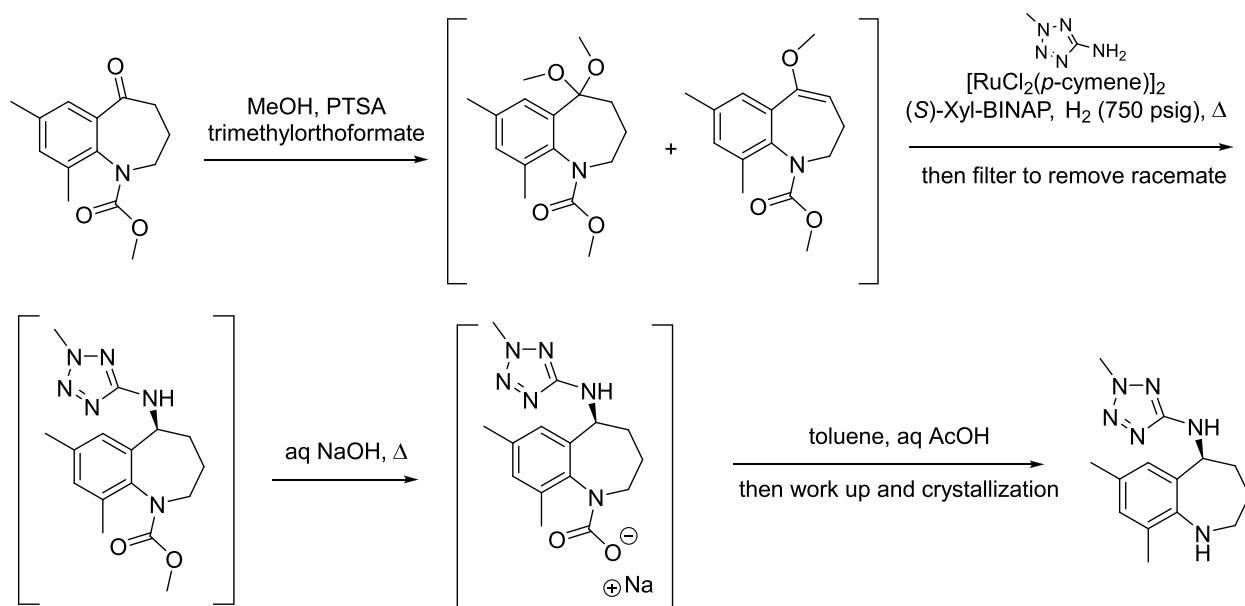


Figure 3. Synthesis of an evacetrapib intermediate. Adapted from ref 9. Copyright 2020 American Chemical Society.

Figure 2 were estimated to be available at a sufficient scale (kilogram availability) and price point (ideally \leq \$100/kg) during ideation.

Additionally, preferred routes are those that can ultimately be achieved with efficient unit operations (simple reaction configurations, minimal workups, high-yielding and highly purifying crystallizations, etc.), be executed safely, and include intermediates with good stability, among many other factors¹⁰ that require additional details of the proposed synthesis process to be filled in. Therefore, to evaluate potential routes, these routes must be further developed into proposed schemes, which include conditions for achieving the transformations, workups, and isolations expressed above and below the arrows. As an example, the scheme to an intermediate in the synthesis route to evacetrapib⁹ is shown in Figure 3. During Route Optimization, proposed routes are scouted with experiments testing key transformations and are further evaluated to eliminate those that are not viable or less desirable. Ultimately, a single route with a proposed baseline scheme is selected, demonstrated on a larger scale, and taken forward into Process Optimization.

The overall task of Route Optimization is multifaceted, and there are at least a few aspects that make this difficult: the virtually infinite number of options (potentially viable routes to select from), the difficulty of experimentally scouting all proposed routes, the multitude of competing objectives or process metrics in defining the *best* route, and the difficulty of evaluating these objectives or process metrics with the incomplete information available when having to make the selection.¹¹

As additional constraints, process development groups are limited in both time and resources in devising a route and must continually make trade-off decisions when selecting routes to ensure that clinical supplies are enabled throughout the process and that the final route is viable for registration with global regulatory bodies (Figure 1, blue). Consequently, due to portfolio prioritization or project acceleration scenarios, business decisions often require the selection of a route at various stages that may not have all of the desired economic

and efficiency attributes, as conceptually depicted in Figure 4. Changes after marketing authorization become even more

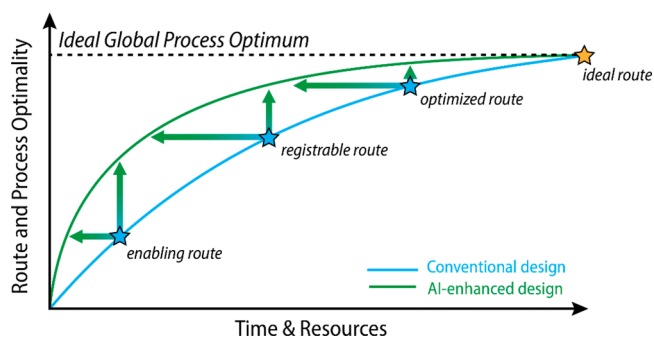


Figure 4. Pharmaceutical Process Optimization conundrum. The path from an enabling route to the ideal route involves several key stages discussed in this Perspective. Criteria for the ideal route may include robustness, cost, and sustainability; safety and drug substance quality are not compromised. Better computational tools should decrease the time/resources required to identify a suitably optimized route and/or an optimized route closer to the ideal route, though there can be diminishing returns for achieving ideality depending on production volumes. The leftward arrows represent the time/resource savings during development that AI-enhanced design can enable, while the upward arrows represent that with a comparable investment a more optimal route can be found that saves resources during manufacturing.

complicated and require a significant return on investment (ROI) justification to undertake. This contrasts significantly with commodity chemical route selection and optimization, in which case route changes and optimizations can be achieved iteratively and over a longer period of time in some examples. We see the potential for the right applications of ML and AI to bring forward much of the evaluation in Route Optimization and enable process development teams to identify enabling, registrable, and optimized routes faster as well as get to more optimal routes in the same finite time; this perspective is also conceptually depicted in Figure 4.

To better manage the complexity of Route Optimization, this is often broken down roughly into four substages, which may be performed concurrently or iteratively:

1. *Route Mapping*, which involves the broad, generative ideation and organization of possible synthetic routes to produce the given target;
2. *Route Narrowing*, where the ideated map of routes is consolidated to top candidate routes and executable work plans are developed for scouting across these (performed iteratively with Route Scouting);
3. *Route Scouting*, where preliminary experiments are conducted to better evaluate the feasibility, productivity, and robustness of proposed routes—with a focus on pivotal transformations;
4. *Route Selection and Demonstration*, where a route is selected and then demonstrated at an appropriate scale going into Process Optimization.

In the following sections, we take a closer look at each of these substages, highlighting challenges and some opportunities for ML/AI.

2.1. Route Mapping. In contrast to synthesis planning for discovery or medicinal chemistry to enable efficient DMTA cycles (as reviewed by Struble et al.²), where the goal is to synthesize many targets in small quantities as efficiently as possible, Route Mapping and synthesis planning for process development have a different starting point and different objectives. Synthesis planning for process development starts with the final synthetic target predefined and has at least one enabling route known at the beginning of the planning exercise; it also has the objective of progressing along the curve shown in Figure 4 as efficiently as possible toward the “ideal” route for clinical supply and ultimately commercial manufacturing.

To begin, process chemists and development teams start with an ideation or Route Mapping exercise. The goal is to identify as many *potentially viable* and *varied* synthetic routes as possible, emphasizing diversity across these to expand the considered route space and improve the odds of including the “ideal” route for commercial manufacturing. As the initiation to Route Optimization, this endeavor is quite important and sets the base for the subsequent stages of Route Narrowing, Route Scouting, and Route Selection.

Ideas for new routes can come from a variety of inputs: experience of the process chemist and institutional knowledge, searches across literature and broadly collected databases using tools like SciFinder or Reaxys, and available CASP tools,⁷ which may employ ML/AI and are becoming more widely used in industry. In particular, CASP methods can stimulate route ideation by more quickly generating a large number of possible routes without being biased by personal preference. Emerging CASP tools incorporating aspects of biocatalysis could further introduce stereo-, regio-, and enantiospecific enzymatic transformations, enabling more efficient synthetic routes and reduced use of organic solvents.¹² However, this comes at the cost of having many options to sort through, and as such, these methods must also have ways to cluster and rank suggestions to avoid overwhelming the user or presenting many near-identical routes (differing in minor features, such as protection/deprotection steps, for example). The ability to collaborate on CASP pathway selections and share annotations and insights would significantly benefit group efforts in Route Mapping.

As part of the mapping exercise for commercial process development, the choice of defined starting materials and the sequence of steps to be registered with regulatory authorities (GMP steps) should also be design criteria. Considering the resource-intensive investigations needed to obtain regulatory approval for the choice of starting materials, there is considerable interest in data-driven models evaluating the risks of candidate regulatory starting materials.¹³ Some routes will have advantages with respect to other approaches in terms of the number of likely registered steps (or GMP steps), the number and complexity of custom starting materials, and synthesis branch points as well as points of entry. For example, a complex late intermediate entering the synthesis route may not be supported by regulatory bodies as a starting material and thus would require an additional GMP branch sequence. In addition, the entry of less complex materials as starting materials into the registered steps may be more easily executed from a supply chain perspective. Accordingly, the mapping exercise should generate routes not only with suitable proposed GMP steps and well-defined, stable custom starting materials but also in a manner where a diverse number of options are devised.

2.2. Route Narrowing. Many possible routes to a given target may be generated in Route Mapping, especially as CASP tools continue to augment human ideation. At the end of Route Mapping, the feasibility of the proposed transformations in any given route may be uncertain, and most of the details of reaction conditions, required workup operations, and isolations have yet to be established. With so much unknown at this stage, it is exceedingly difficult to perform a concrete evaluation of the routes in the map and directly identify the best route for commercial manufacturing from the candidates. Further, it would be cost- and time-prohibitive to perform detailed experimental exploration and optimization for every route in a given map. Instead, a successive approach toward Route Selection and Demonstration is taken. The map is first *narrowed*, potentially down to a single route, but more likely down to two to four routes that can be moved forward to more extensive Route Scouting. Ultimately, one or two routes are taken into Route Selection and Demonstration.

The task of Route Narrowing is one of elimination. The goal is to eliminate those routes that are not feasible and further eliminate those feasible routes that are *dominated* by others on the map. In this usage, dominated means expected to produce a commercial process that is worse on one or more process metrics and not meaningfully better on any process metric. Performing Route Narrowing well requires strategy and the right tools. It could be significantly advanced with newly developed ML and AI tools for evaluating reaction feasibility, especially in combination with high-throughput and automated experimental systems for reaction screening.

The general workflow for Route Narrowing is to first organize or cluster the proposed map, often grouping those routes that are most like one another or that hinge on the same key transformations. Clustering is very useful for making comparisons between routes and highlighting the most informative reaction screening experiments to run. That is, finding those “killer” experiments that assess proposed transformations that have the highest degree of uncertainty with respect to their feasibility is also central to the success of a proposed route or cluster of routes. Additional factors beyond feasibility that might be probed through small numbers of selective experiments include safety inputs (expected reaction

kinetics, high-energy reagents, calorimetry, etc.). Here we can envision AI tools that allow for better identification of the most informative reaction screening experiments to run as well as extrapolate from one set of experimental results to resolve uncertainty on the feasibility of other similar transformations in the route map. The task of experimentally probing the feasibility of proposed transformations through reaction screening is often considered the initial stage of Route Scouting, which happens iteratively with Route Narrowing and is covered in more detail in the following section.

Depending on the size of the initial map, eliminating infeasible routes may not be enough to reduce to the desired two to four routes. To accomplish this, the map must also be evaluated to find and eliminate dominated routes as defined above. However, this can present a challenge, as few details have been filled in for any given route. Moreover, only a small amount of reaction screening plus early medicinal chemistry and first-in-human process development data are available, making it nearly impossible to accurately infer the absolute values of the process metrics for the ultimate commercial process using a given route—like cost of goods manufactured (COGM) or throughput and efficiency. As a consequence, routes cannot be reliably scored according to those metrics and then simply compared to make the selection. Instead, to make this task more tractable, routes can be compared head-to-head with each other using simple criteria that *can* be inferred from the little information on hand but also act as surrogate process metrics, often focusing on comparing routes that are most similar and deviate, for example, only in one or two proposed transformations.

In performing the comparative analysis, it may be possible to use expert judgment and experience to definitively eliminate some routes as being worse than others. It is more often only possible to make probabilistic judgments, eliminating routes as being *very likely* worse than others for commercial manufacturing. In a scenario in which judgment-based comparative elimination does not quickly narrow down to a few routes, additional scouting experiments may be executed with the goal of revealing a better understanding of key steps in those routes being considered. For example, running experiments to better determine if a proposed set of telescoped and complicated transformations, which clearly set up an efficient route, can actually be achieved in practice with reasonably high yield and without quality concerns. The information generated from these further Route Scouting experiments better informs final elimination decisions. Here too, we see significant potential for ML and AI tools. These tools could aid in the organization of routes for comparison, better identify surrogate process metrics that can be used for head-to-head route comparison at this stage, and identify chemical transformations within a given synthesis route essential to the success of the most disruptively innovative routes. The ML/AI methods could also inform experimental planning and comparative analysis across routes and perform predictive modeling to estimate uncertainty about the expected performance of a given route or transformation in a route without having to conduct experiments.

2.3. Route Scouting. Route Scouting is the experimental evaluation and development of proposed transformations and the selection of routes in a given map. This often begins with screening reactions in the map to evaluate the feasibility of uncertain transformations. Beyond elimination of routes that appear infeasible following reaction screening, additional information is generally required to perform head-to-head

comparisons and further narrow the route map by elimination of dominated routes, in particular, information on *how* reactions are to be executed as well as information on the expected performance—e.g., selectivity, yield, robustness, and process throughput. To gain this information requires experiments in which reaction conditions are moderately optimized and possible workup and isolation operations are explored for select steps in the considered routes.

Route Scouting, because of the required experimental campaigns, is likely the most time- and resource-consuming subtask of Route Optimization. The opportunity for AI systems includes developing improved prediction models for feasibility and condition design, as computational evaluation of these factors *a priori* can reduce the experimental burden and potentially reduce the time spent evaluating these routes. At the same time, it is important to note that AI is not solely useful for proposing what to try in the absence of experiments but as a tool for responding to experimental outcomes and proposing what to try *next*. Model-guided experimental design (e.g., active learning) can assist in the identification of the most useful ways to invest experimental resources to eliminate undesired routes quickly and effectively as well as prioritize the most promising routes for further investigation.

It is also useful to recognize that Route Scouting occurs in an iterative fashion with Route Narrowing and evolves through the course of this process. In the early stages of Route Narrowing and Route Scouting, the goal of scouting experiments is generally to probe the feasibility of certain proposed transformations. Ideally this aspect is executed through reaction screening protocols, with initial application of high-throughput experimental (HTE) systems to perform this work efficiently and rapidly. Additional lower-throughput automated reaction screening platforms for preliminary reaction condition definition can also be utilized to provide additional context on specific chemical transformations of interest and to instill further confidence in the robustness of the transformation. However, even with advances in HTE and lab automation,¹⁴ application of reaction screening takes considerable human time and effort as well as advanced quantities of key intermediates. What is more, it is infeasible to examine every reaction under every possible set of conditions. This presents a significant opportunity for the ML and AI tools. We can imagine these tools significantly advancing HTE reaction screening by suggesting the right reactions within a map to explore (as mentioned previously), suggesting the best set of initial conditions for a given reaction to screen across (as well as aqueous workup conditions), automatically processing and interpreting the raw analytical data¹⁵ (for example, with automated peak labeling and structure elucidation of both products and impurities), enabling closed-loop or iterative screening in which the results from one round suggest a next set of experiments,¹⁶ and extrapolating from outcome data collected over a finite set of reactions conditions to say more definitively whether a reaction is feasible or not across *any* conceivable set of reaction conditions. We highlight that in the above, AI may in particular inform the selection of transformations that provide breakthrough innovations in route design but show up as low-yielding or “misses” in initial screens. At the moment, success in pursuing such transformations past screening and accessing these innovative routes relies upon the experience and tenacity of the process chemist. Going forward, we can envision AI tools helping both highlight the potential breakthrough transformations that

deserve extra attention in screening and also help the process chemist more confidently interpret screening results to distinguish between an infeasible transformation and a feasible transformation that requires careful optimization to achieve viable yields.

In the final stages of Route Scouting, it is typical to have two or three moderately differentiated routes still in consideration. The evaluation and comparison of these final routes for Route Selection requires that they be turned into *schemes* such that key process or surrogate process metrics can be calculated, such as yield, atom economy, process mass intensity (PMI), robustness, impurity formation and rejection, and intermediate stability. This, of course, requires an experimental effort to establish moderately optimized reaction conditions, workups, and isolations for each step in the proposed routes. Some of this work may be greatly facilitated by HTE campaigns—for example, running solubility screens to help establish the process solvent design space for reaction, workup, and especially isolations by crystallization. However, the majority of the experimental effort at this stage continues to require process chemists and engineers to manually execute experiments in the lab. Consequently, AI tools that can narrow the design space to explore, extrapolate reliably from small data sets, and/or enable more sophisticated lab automation would offer substantial human-time and cost savings. Note that we use the term “design space” to denote the space in which one designs the commercial route, i.e., the potential operating space, rather than the relationship of material attributes and process parameters to critical quality attributes as defined by regulatory guideline ICH Q8.¹⁷

2.4. Route Selection and Demonstration. For projects to advance to the Route Selection and Demonstration stage, typically one to two likely synthesis routes have been identified. These routes will be carried through to the final target to confirm the initial reaction conditions used in the Route Scouting phase, beginning from a defined starting point. While the precise reaction conditions for all steps used at this stage can be changed in the final commercial version of the process following Process Optimization, this demonstration provides a more in-depth analysis of the overall route. Further examination is required to elucidate parameters regarding reaction performance, yield sensitivity, and telescoping options as well as solvent selection to derisk the route. Many techniques have recently been developed for reaction optimization.¹⁸ Yield prediction remains a challenge for ML/AI, but recent ML models for solvation¹⁹ and solubility²⁰ offer opportunities for predicting suitable green solvents and telescoping opportunities. Process safety metrics, as well as reaction kinetic profiles, can now also be measured in this demonstration. Prediction of adiabatic temperature rise and faster development of reaction kinetic models²¹ represent additional ML/AI opportunities. Exemplification of the route also serves to generate standards for new analytical methodology (e.g., chromatography conditions) and intermediates for future development. Arguably, however, the most important attributes to determine in this final part of route optimization are the purification levers and how formed impurities can be robustly controlled at points in the process, which often do not fully reveal themselves until the route is demonstrated in its entirety. Predicting impurity profiles is closely related to the established tasks of predicting reaction outcomes, but existing ML models do not provide the coverage and quantitative precision needed to design a purification strategy *in silico*.

Failure to demonstrate a viable purification may disqualify the route entirely and necessitate reverting back to earlier stages of Route Scouting to explore alternatives.

In general, the selected route will have many or most attributes of a desired manufacturing process, which may or may not be ideal given the timeline or other resourcing constraints. Figure 5 attempts to depict the multiparameter

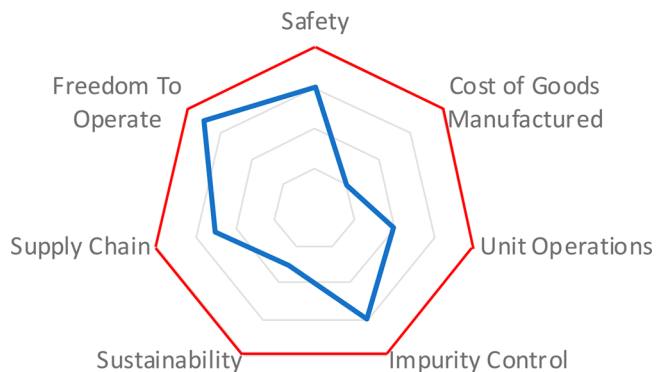


Figure 5. Radar graph depicting a subset of the multiparameter aspects of the Route Selection stage and varying degrees with which key route attributes might be developed prior to Process Optimization.

aspect of Route Selection through a possible scenario where various aspects may not be entirely optimized at the time of selection. These risk elements can be discharged further in the Process Optimization phase, with additional time and resource implications.

Nonetheless, with an exemplified new route successfully preceded, a more complete picture emerges of its attributes, and these can be reassessed against the preceding medicinal chemistry route to determine an ROI. This successful demonstration also establishes the baseline for the nominal process over which future Process Optimization efforts will begin, and while labor- and time-intensive, this detailed in-depth analysis is necessary to justify the major investment required for Process Optimization. At this point, a new synthetic route is selected, and the development effort shifts to the more intensive Process Optimization phase.

3. PROCESS OPTIMIZATION

In the Process Optimization phase, the selected route is developed into a full-fledged manufacturing process in preparation for process characterization, technology transfer, clinical supply, validation, regulatory filing, and ultimately commercial manufacturing. In principle, what has been decided up to the point of Process Optimization is the synthetic route for manufacturing the drug substance, written in terms of the starting materials and the sequence of transformations that form drug substance intermediates to the final drug substance. The conditions for each reaction step and details of the workup and isolation operations have been partially explored in Route Scouting, and a nominal design has been selected and demonstrated, but these aspects have not yet been *locked*. It is these yet-to-be-locked aspects of the process that make up the design space to explore across in Process Optimization. Process Optimization can be broken down conceptually into two substages: *Problem Formulation*, where objectives are established and the design space is identified,

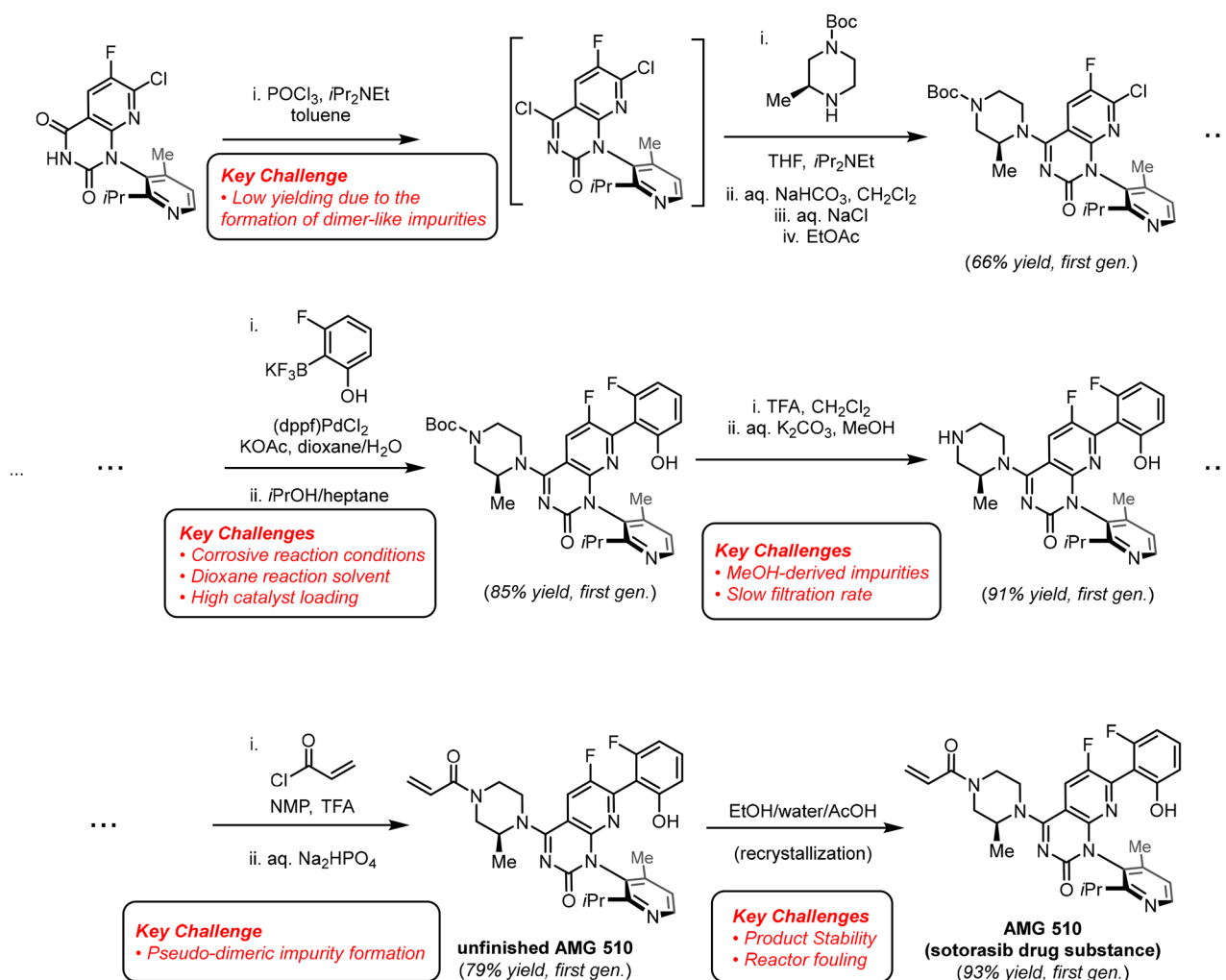


Figure 6. First-generation sotorasib manufacturing process with key process challenges highlighted for each step. Adapted from ref 22. Copyright 2022 American Chemical Society.

and Search, where the design space is explored through multiple stages of experimentation.

How best to formulate and explore the design space and thereby complete Process Optimization as effectively as possible is an outstanding question that process development groups actively grapple with. In the following sections, we take a closer look at the task, provide a description of the current expert-driven approach to tackling Process Optimization using a recent example, and suggest where ML and AI might be beneficially applied (as expanded in [Opportunities and Outlook](#)).

3.1. Problem Formulation. To conduct Process Optimization, development teams must first spend time *formulating* the problem. A set of objectives must be clearly established, and the design space must be identified (and ideally narrowed). This seems straightforward only at a glance; in practice, it requires significant expertise and judgment. What is more, the success and efficiency of the optimization campaign can often be attributed to its formulation.

Development teams commonly have broad objectives in conducting Process Optimization: to identify a safe-to-execute process with the lowest COGM, maximum process throughput, and maximum process robustness. Process throughput, also often termed process intensity, measures how much material can be produced by the specified process in a given

amount of time with a fixed scale of equipment. Process robustness refers to the reliability of the overall process to produce drug substance that meets the quality specifications under normal operating variations; that is, process robustness is a measure of the insensitivity of the drug substance quality output with conceivable variation in the execution of the process.

There are a couple of complications with this set of broad global objectives from the standpoint of conducting Process Optimization. The objectives may be competing—increasing process robustness may result in higher COGM, for example. Even more practically important, these are difficult to directly calculate, making it a challenge for development teams to systematically evaluate the explored variations in the process according to the effects on these global objectives. Take COGM, the simplest of the global objective measures. To directly calculate the COGM requires knowledge of at least the following: how much of each material will be used per amount of drug substance produced, requiring yields at scale through each operational step; the cost of that material, some of which is specialized with bulk cost negotiated over time, with supply chain optimization playing a significant role; and the manufacturing cost per operation, which depends on the required time and complexity of the manufacturing operations as well as where the manufacturing takes place, on the

equipment required for the operations, and often on business relationships and negotiations with contract manufacturing organizations.

To handle the complexity of the optimization objectives, the prevailing strategy is an expert-driven approach: development teams review the current route or first-generation manufacturing process out of Route Selection and Demonstration (section 2.4) and identify what appear to be the biggest challenges or shortcomings of each step in the synthesis with respect to a judgment of how they will influence the overall global objectives. That is, through problem identification at the step level, broad global objectives for Process Optimization are translated to tractable local objectives, and local optimization campaigns can then be formulated around those. As an example, consider the development of sotorasib (LUMAK-RAS) recounted in a recent publication.²² Figure 6 shows the first-generation sotorasib manufacturing process and highlights what were identified by the process development team as key process challenges for each step. Those challenges then informed the local objectives for Process Optimization.

Once the Process Optimization objectives are articulated, the next step is to identify the design space to explore to optimize these objectives. This too can be a challenge. Looking at the yet-to-be-locked aspects from route optimization (everything above and below the arrows in the synthetic route) reveals a very large and hard-to-define design space. Consider even a single reaction—the reaction in the first step of the sotorasib synthesis shown in Figure 6, for example. In this two-part reaction, the process solvents and reagents must be selected along with the solvent volumes, equivalents of all reactants and reagents, and temperatures during the different stages of the reaction. And that is just the start—the order of addition, addition rates/profiles, agitation, and many other parameters may be considered in optimizing the reaction. This leads to a large design space with both categorical and continuous design variables that have clear cross-interactions. The design space expands exponentially when considering the workup and isolation operations in each step and when considering multiple steps in a synthesis simultaneously.

It is clear that the full design space cannot be mapped and fully explored experimentally in a finite amount of time, even with advanced high-throughput experimental setups that utilize automation and parallelization. To deal with this challenge, the prevailing approach, like that in defining the local objectives, is expert-driven and based on current process knowledge. Sticking with the first step in the sotorasib synthesis in Figure 6 as a revealing example, the key deficiency identified going into Process Optimization was that this was a low-yielding sequence, at just 66% step yield; this gave the straightforward local optimization objective of increasing the step yield by adjusting the design parameters.²² What is more, the proximate cause of the low yield was also identified: as executed going with the early-phase process, up to 15 HPLC area % (LCAP) of a mixture of dimer-like species was observed.²² Having identified the key deficiency and an understanding of the proximate cause, experienced chemists on the process development team could hypothesize on the potential mechanism and narrow in on the design variables that should offer the biggest levers. In this case, those could be adjusted to inhibit the formation of the dimer-like species, thereby increasing the step yield. These were the order of addition, addition time, equivalents of POCl₃, and temperature—ultimately providing a much smaller design space that could

be explored experimentally to optimize yield with good success.²²

The described expert-driven approach to Problem Formulation, which recasts general global objectives into local ones with significantly narrowed design spaces based on expert understanding, can be quite effective. However, it also may lead to a local optimum. To manage more ambitious problem formulations that consider full steps or multiple steps at a time and bring us closer to finding globally optimal processes, we will need assistance from ML- and AI-enabled computational tools to accurately capture the global objectives and allow a significantly expanded design space to be explored efficiently.

3.2. Search. Once the optimization problem has been formulated, process development teams consider how to efficiently explore the design space through a series of experiments. It is convenient to break down the search task into three components, operating in a cycle: (1) strategy, (2) experimental execution (including measurements), and (3) interpretation. In this breakdown, strategy refers to the approach taken to determine which experiments to execute over time and the criteria under which the search should be stopped (how to move through the design space and when to stop). Experimental execution is how to run those experiments as well as to collect raw data that can be used to evaluate the outcome, usually over time. Finally, interpretation includes the processing of raw analytical data to relevant outcome measures, the mathematical connection of design variables to the outcome measures, and the evaluation to identify whether there is a point in the already-explored space that meets process optimization stopping criteria and, if not, to suggest the best direction(s) to explore with a next set of experiments (often based on the mathematical connection established between design variables and the outcome measures).

There are a variety of approaches to each component of Search. For local process objectives and sufficiently narrow design spaces, process development teams can use a small Design of Experiments (DOE) that is executed and interpreted mostly manually and expect to get to a local optimum quickly. From a zoomed-out perspective, the prevailing approach in the industry is along these lines: (1) select the points in the design space—usually not too far away from a proven set of conditions—that, based on chemical knowledge and past experience, are expected to produce a different and perhaps more optimal outcome according to a local objective; (2) execute the experiments in the lab, taking advantage of automated lab reactors but requiring manual processing; and (3) process the raw data and interpret the results.

The prevailing strategy is very effective with sufficiently narrow objectives and design spaces. However, with broader objectives and larger design spaces, it becomes more and more difficult, as well as time-consuming, to comprehensively explore the design space through human-driven experiments and interpretation. This is especially true when considering a full step with multiple unit operations and multiple steps in sequence. In those cases, it is not uncommon to conceive of hundreds or thousands of design variables, including categorical and continuous variables and the time dependence of the outcome. To search larger design spaces experimentally, there is a concerted push to formalize the strategy, apply more advanced DOE and optimization concepts, and bring to bear lab automation/robotics systems. Even so, smart experimental design and lab automation alone are unlikely to meet the challenge. To explore truly large and complicated design spaces

will require systems that can explore these complex spaces intelligently and fully autonomously without interruption; this will require new ML/AI tools and new ways of coupling those tools with more flexible robotic systems that are integrated with the right analytical tools.^{23,24}

4. OPPORTUNITIES AND OUTLOOK

The full workflow of synthetic drug substance process development is multifaceted and requires balancing many objectives when specifying (or optimizing) the full details of the sequence of reaction, workup, and purification steps. There is a practical limit to what can be achieved through screening, and hence, new technologies that are able to (a) make use of existing data and information and (b) propose informative experiments to run will help guide this complex design process. While we have touched on several opportunities and use cases already, we will elaborate on some of them here.

Route Mapping is the best connection between process development and the progress in computer-aided synthesis planning that the field has seen largely applied to discovery or medicinal chemistry. CASP tools⁷ are routinely used in industry nowadays to generate large numbers of ideas, some of which may be viable as is, some of which may serve as inspiration, and some of which may not make chemical sense. Assessing the feasibility of the proposed reactions and routes is a challenge shared with discovery chemistry. What is unique about the application to process development is the challenge of narrowing down routes to the most ideal for clinical and commercial production, which has to be done on the basis of tailored objectives and in consideration of complex and hard-to-calculate process performance metrics. For example, the notion of an “efficient” route can be crudely captured by its step count or atom economy, which is easily calculated given a proposed synthetic route; however, the more pertinent metric would be process mass intensity (PMI), which more comprehensively captures material usage and requires quantitative details of conditions and yields for both reaction and purification steps. Even this metric does not directly relay ultimate performance metrics, such as COGM, process robustness, or process safety. Aiming to calculate more complicated but more pertinent metrics raises the bar in terms of the level of precision we expect from our CASP tools, which have tended to focus on qualitative predictions (e.g., whether the reaction is likely to “work”) rather than quantitative ones (e.g., concentrations, reaction times, and yields). Also, unique to process development is the need to meet regulatory constraints, whether in the selection of starting materials or in elucidating and managing impurities. There are no rigid rules as to what acceptable Regulatory Starting Materials are, but efforts have been made to learn guidelines given past FDA approvals.^{13,25} What is more, unlike in applications such as commodity manufacturing, where there is time and financial incentive to improve COGM further and further after proposing a given commercial process, there is a high barrier to making postlaunch changes to API manufacturing routes given regulatory considerations and commitment to safety. The routes that we use do not need to be the mathematical optima, but we would like our computational methods to get us closer to the “ideal route”.

The desire to shift toward CASP tools that include more quantitative understanding of synthetic processes introduces new challenges in molecular property prediction. Property prediction is an overarching goal that supports process

modeling in many different ways. For example, to design a purification following a reaction step requires predicting not just the major product(s) in the reaction but also relevant related impurities and their amounts as well as the solvent system physical parameters and execution process parameters that will influence the operation dynamics and, crucially, the rejection/retention of solvents and key impurities in different phases. This may include pH-dependent distribution coefficients for liquid–liquid extractions with or without pH swings, solubilities in various solvents and solvent mixtures for crystallizations, or even retention times and behavior on chromatographic columns as a function of solvent, temperature, and solid phase. Process intensification strategies (particularly when considering maximizing concentrations in continuous flow chemistry processes) may require estimation of solubilities at elevated temperatures in nonaqueous solvents, for which little data exist. Of these many properties, progress has been made on predicting solvation¹⁹ and solubility.²⁰ Xiouras et al. recently reviewed opportunities for AI to accelerate crystallization processes.²⁶

The use of property prediction models, which resemble the quintessential tasks of quantitative structure–property relationship modeling but may also include *interactions* or *mixtures*, is best suited as components of a strategy to predict outcomes of unit operations across design variation and relevant process metrics at the unit operation or step level, which can then be combined to calculate route-level metrics. We do not envision using a surrogate ML model to predict the PMI of a route end to end, for example, but rather using surrogate models to estimate the values at each step that are needed to calculate the overall PMI. These tools will be most valuable in the early stages of Route Scouting and Route Narrowing to get a sense of pathway viability where some uncertainty and inaccuracy is tolerable. At later stages when intermediates are locked in and there is a better sense of strategy, it is more feasible to obtain experimental measurements (including through the use of automation) instead of relying on uncertain predictions. Uncertainty quantification is itself a major challenge for the field and an important aspect of property prediction.²⁷ In many ways, the step of Route Narrowing is about risk management and anticipating the suitability of candidate routes before committing to a full exploration and optimization of their parameters.

We would like these various predictive tools to lead to a “digital partner” for managing the process development lifecycle: a platform that helps guide us to the optimal route, perform risk assessment, capture ML predictions, design new experiments, learn from new experimental data, predict physical properties, anticipate impurities, and overall guide process development decisions (dotted green arrow in Figure 1). This hypothetical tool could tie together physics-based modeling with empirical data-driven models in the pursuit of a digital twin. Importantly, such a digital partner within company firewalls could learn from project-specific data for the reactivity and properties of proprietary intermediates with the structural motifs of the target API, thus improving the quality of model predictions and overall route guidance throughout the development process. Furthermore, within pharma, heterogeneity of data (both within and across companies) in terms of both inputs and structure further complicates model training and development. A significant potential for AI and a digital twin exists to integrate disparate data streams, which will improve the process development lifecycle and AI-based model

performance. While the complexity of a digital twin for the whole process will be hard to tackle (e.g., a pharmaceutical analogue of Aspen²⁸), progress in addressing individual components of the process can be made, e.g., (1) solvation and distribution prediction to design liquid–liquid equilibrium or crystallization steps, (2) yield prediction as a function of catalyst/ligand and substrate for a particular reaction family, and (3) structural elucidation and spectral prediction to accelerate the analysis of impurities observed experimentally.

Experiments are a key part of the process, from Route Scouting to Search in Process Optimization, and they serve as an important validation step. We should not assume that predictive models will ever fully replace experimental testing, although the confidence in our predictions may become much higher as these tools improve. The role of models trained on previous data may be to predefine a narrower design space and act as a *prior* to inform which condition settings might be most appropriate to explore during Process Optimization. Iterative experimental design techniques (e.g., Bayesian optimization and active learning) also can play a large role in Process Optimization by identifying the most informative or discriminating experiments. Such techniques have been applied successfully to catalyst and condition selection^{18,29} as well as kinetic model discrimination.³⁰ The most well-developed part of this is iterative reaction optimization, but there are many opportunities to facilitate problem formulation and especially design space selection as well as extend these approaches to larger design spaces (more variables), including process steps other than single-step reactions.^{24,31} The automated execution of experiments themselves can also be made more flexible/robust, allowing for the execution of more unit operations and execution of full steps in synthetic sequences (rather than just reaction screening) by integrating feedback loops powered by ML/AI tools, for example, integrating machine vision.³² Additionally, ML/AI tools can be brought to bear to better process raw analytical data and do this automatically; consider automatic structure elucidation from LC-MS data, for example.³³ Finally, ML strategies can almost certainly be used to improve the development of surrogate models that connect design variables to the relevant process outcome metrics (perhaps over time) from data collected during the iterative experimental execution in Process Optimization.

Alongside the opportunities for AI/ML in process development, there are challenges the field must address. Data sparsity complicates model training,³⁴ so it makes sense to consider the use of proprietary company data to improve process models. In addition, improved ability to extract detailed data from written records would be of great benefit for obtaining well-curated data sets.³⁵ There are opportunities for federated learning (collaborative ML without centralized training data hosted by a trusted party), as have been explored by the MELLODDY Consortium³⁶ for certain property prediction tasks. Even with the willingness to share information in this way, it may still prove difficult to build generalizable models that can achieve the tasks we have outlined above without the assistance of specific, proprietary data collected throughout process development. Hybrid models and the use of computational chemistry to augment experimental data sets can help,³⁵ but the total number of molecules that have gone through rigorous process development is in the thousands, compared to the tens of millions of molecules that have been made in the discovery setting. Identifying appropriate data sets and finding creative

ways to augment those data sets to enable the application of ML with sparse data remains an outstanding challenge to meet.

Finally, we should address the wave of excitement around ChatGPT³⁷ and comparable tools. In our view, generative AI tools and in particular large language model (LLM)-based tools do not fundamentally change process development opportunities but may provide far improved interaction with those tools. We expect these advances to have the largest impact in how users interact with written process descriptions and also how users interact with various AI/ML technologies. Conversational, natural language interfaces could provide a more intuitive way for process chemists and engineers to access data and integrate model predictions (first-principles or ML-based) into their workflows. The opportunity to fine-tune LLMs for document-based question and answer also offers new means of accessing the heterogeneous data sources associated with process development, from internal reports to regulatory filings. To the extent that information is already formatted in natural language, LLMs can provide a more convenient and interactive way to learn about the content without reading the full text. As knowledge-based and retrieval-augmented systems mature and the level of trust that we have increases, we may be able to rely on AI to prepare and verify these documents automatically.

As we continue to progress in the field, it is our perspective that effective synthetic drug substance process development will not and cannot be done without domain expertise, but it also should not be done without data science expertise and the use of rapidly advancing ML and AI technology. While the advancement of ML/AI does not change the goal of process development, we do expect these advances to improve our *speed*, *efficiency*, and *effectiveness* in conducting process development and, in doing so, enable us to get material into the clinic and ultimately to patients worldwide faster at reduced costs.

■ AUTHOR INFORMATION

Corresponding Author

Klavs F. Jensen – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-7192-580X; Email: kfjensen@mit.edu

Authors

Daniel J. Griffin – Pivotal & Commercial Synthetics, Process Development, Amgen, Inc., Cambridge, Massachusetts 02142, United States; orcid.org/0000-0002-5051-2069

Connor W. Coley – Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-8271-8723

Scott A. Frank – Synthetic Molecule Design and Development, Eli Lilly and Co., Indianapolis, Indiana 46285, United States; orcid.org/0000-0001-6680-9695

Joel M. Hawkins – Pfizer Worldwide Research and Development, Groton, Connecticut 06340, United States

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.oprd.3c00229>

Notes

The authors declare no competing financial interest.

■ GLOSSARY

| | |
|-------------------|---|
| API | active pharmaceutical ingredient |
| CASP | computer-aided synthesis planning—the task of designing a synthetic route; the level of detail associated with a proposed route varies |
| COGM | cost of goods manufactured—the total cost including processing costs, labor, and equipment used, in addition to the purchase price of consumed raw materials in the synthesis |
| DS | drug substance (often used synonymously with API) |
| DP | drug product—the final formulation (e.g., a pill to be taken orally) |
| DMTA | design—make—test—analyze—the typical iterative process in early-stage drug discovery |
| enabling route | a synthetic route used to manufacture a drug substance for early clinical studies which may not be fully optimized |
| FTO | freedom to operate |
| GMP | good manufacturing practices—the requirements to which manufacture of the final DS must adhere according to regulators (e.g., the U.S. Food and Drug Administration) |
| LCAP | liquid chromatography area percentage—a metric measuring how much product was formed that does not require calibration |
| registrable route | a synthetic route used to manufacture drug substance which is deemed by the innovator company to have desired commercial attributes |
| ROI | return on investment |
| SM | starting material—a commercially available compound used to begin a designated synthetic route |

■ REFERENCES

- (1) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J. Health Econ.* **2016**, *47*, 20–33.
- (2) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; DesJarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63* (16), 8667–8682.
- (3) Cole, K. P. Large-Scale Synthesis. In *Burger's Medicinal Chemistry and Drug Discovery*; Abraham, D. J., Myers, M., Eds.; John Wiley & Sons, Hoboken, NJ, USA, 2021; Vol. 3, pp 1–39.
- (4) (a) Anderson, N. G. *Practical Process Research and Development*; Academic Press: Waltham, MA, USA, 2012. (b) *Complete Accounts of Integrated Drug Discovery and Development: Recent Examples from the Pharmaceutical Industry Volume 1*; Abdel-Magid, A. F., Pesti, J. A., Vaidyanathan, R., Eds.; ACS Symposium Series 1307; American Chemical Society, 2018. (c) *Complete Accounts of Integrated Drug Discovery and Development: Recent Examples from the Pharmaceutical Industry Volume 2*; Pesti, J. A., Abdel-Magid, A. F., Vaidyanathan, R., Eds.; ACS Symposium Series 1332; American Chemical Society, 2019. (d) *Complete Accounts of Integrated Drug Discovery and Development: Recent Examples from the Pharmaceutical Industry Volume 3*; Abdel-Magid, A. F., Pesti, J. A., Vaidyanathan, R., Eds.; ACS Symposium Series 1369; American Chemical Society, 2020. (e) *Complete Accounts of Integrated Drug Discovery and Development: Recent Examples from the Pharmaceutical Industry Volume 4*; Abdel-Magid, A. F., Pesti, J. A., Vaidyanathan, R., Eds.; ACS Symposium Series 1423; American Chemical Society, 2022.
- (5) Campos, K. R.; Coleman, P. J.; Alvarez, J. C.; Dreher, S. D.; Garbaccio, R. M.; Terrett, N. K.; Tillyer, R. D.; Truppo, M. D.; Parmee, E. R. The importance of synthetic chemistry in the pharmaceutical industry. *Science* **2019**, *363* (6424), No. eaat0805.
- (6) Leng, R. B.; Emonds, M. V. M.; Hamilton, C. T.; Ringer, J. W. Holistic Route Selection. *Org. Process Res. Dev.* **2012**, *16* (3), 415–424. Parker, J. S.; Moseley, J. D. Kepner-Tregoe Decision Analysis as a Tool to Aid Route Selection. Part 1. *Org. Process Res. Dev.* **2008**, *12* (6), 1041–1043.
- (7) (a) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365* (6453), No. eaax1566. (b) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynth-Finder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminf.* **2020**, *12*, 70. (c) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W. W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational planning of the synthesis of complex natural products. *Nature* **2020**, *588* (7836), 83–88. (d) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **2020**, *11* (12), 3316–3325.
- (8) Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. Artificial intelligence and automation in computer aided synthesis planning. *React. Chem. Eng.* **2021**, *6* (1), 27–51.
- (9) Mantlo, N. B.; Frank, S. A. Discovery and Commercial Development of Cholesteryl Transfer Protein Inhibitor Evacetrapib. *ACS Symp. Ser.* **2020**, *1369*, 339–372.
- (10) (a) Rossen, K. Greening Organic Chemistry with Process Chemistry. *J. Org. Chem.* **2019**, *84* (8), 4580–4582. (b) Zhang, T. Y. Process Chemistry: The Science, Business, Logic, and Logistics. *Chem. Rev.* **2006**, *106* (7), 2583–2595.
- (11) Eastgate, M. D.; Schmidt, M. A.; Fandrick, K. R. On the design of complex drug candidate syntheses in the pharmaceutical industry. *Nat. Rev. Chem.* **2017**, *1* (2), 0016.
- (12) (a) Sankaranarayanan, K.; Jensen, K. F. Computer-assisted multistep chemoenzymatic retrosynthesis using a chemical synthesis planner. *Chem. Sci.* **2023**, *14* (23), 6467–6475. (b) Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W. Merging enzymatic and synthetic chemistry with computational synthesis planning. *Nat. Commun.* **2022**, *13* (1), 7747.
- (13) Reizman, B. J.; Burt, J. L.; Frank, S. A.; Argentine, M. D.; Garcia-Muñoz, S. Data-Driven Prediction of Risk in Drug Substance Starting Materials. *Org. Process Res. Dev.* **2019**, *23* (7), 1429–1441.
- (14) Eyke, N. S.; Koscher, B. A.; Jensen, K. F. Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends Chem.* **2021**, *3* (2), 120–132.
- (15) Goldman, S.; Li, J.; Coley, C. W. Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks. *arXiv (Quantitative Biology. Quantitative Methods)*, April 25, 2023, 2304.13136, ver. 1. <https://arxiv.org/abs/2304.13136> (accessed 2023-07-10).
- (16) Koscher, B.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Haoyang Wu; Vermeire, F. H.; Jin, B.; Hart, T.; Kulesza, T.; Li, S.-C.; Jaakkola, T. S.; Barzilay, R.; Gómez-Bombarelli, R.; Green, W. H.; Jensen, K. F. Autonomous, multi-property-driven molecular discovery: from predictions to measurements and back. *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-r7b01.

- (17) ICH Quality Guidelines. <https://www.ich.org/page/quality-guidelines> (accessed 2023-07-10).
- (18) Breen, C. P.; Nambiar, A. M. K.; Jamison, T. F.; Jensen, K. F. Ready, Set, Flow! Automated Continuous Synthesis and Optimization. *Trends Chem.* **2021**, *3* (5), 373–386.
- (19) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, 129307.
- (20) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022**, *144* (24), 10785–10797.
- (21) Burés, J.; Larrosa, I. Organic reaction mechanism classification using machine learning. *Nature* **2023**, *613* (7945), 689–695.
- (22) Zhang, L.; Griffin, D. J.; Beaver, M. G.; Blue, L. E.; Borths, C. J.; Brown, D. B.; Caille, S.; Chen, Y.; Cherney, A. H.; Cochran, B. M.; et al. Development of a Commercial Manufacturing Process for Sotorasib, a First-in-Class KRASG12C Inhibitor. *Org. Process Res. Dev.* **2022**, *26* (11), 3115–3125.
- (23) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144* (43), 19999–20007.
- (24) Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. Bayesian Optimization of Computer-Proposed Multistep Synthetic Routes on an Automated Robotic Flow Platform. *ACS Cent. Sci.* **2022**, *8* (6), 825–836.
- (25) Faul, M. M.; Argentine, M. D.; Egan, M.; Eriksson, M. C.; Ge, Z.; Hicks, F.; Kiesman, W. F.; Mergelsberg, I.; Orr, J. D.; Smulkowski, M.; Wächter, G. A. Part 3: Designation and Justification of API Starting Materials: Proposed Framework for Alignment from an Industry Perspective. *Org. Process Res. Dev.* **2015**, *19* (8), 915–924.
- (26) Xiouras, C.; Cameli, F.; Quilló, G. L.; Kavousanakis, M. E.; Vlachos, D. G.; Stefanidis, G. D. Applications of Artificial Intelligence and Machine Learning Algorithms to Crystallization. *Chem. Rev.* **2022**, *122* (15), 13006–13042.
- (27) (a) Hirschfeld, L.; Swanson, K.; Yang, K.; Barzilay, R.; Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60* (8), 3770–3780. (b) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60* (6), 2697–2717.
- (28) *Aspen Plus*. AspenTech, 2023. <https://www.aspentech.com/en/products/engineering/aspen-plus> (accessed 2023-04-22).
- (29) Baumgartner, L. M.; Dennis, J. M.; White, N. A.; Buchwald, S. L.; Jensen, K. F. Use of a Droplet Platform to Optimize Pd-Catalyzed C–N Coupling Reactions Promoted by Organic Bases. *Org. Process Res. Dev.* **2019**, *23* (8), 1594–1601.
- (30) Taylor, C. J.; Seki, H.; Dannheim, F. M.; Willis, M. J.; Clemens, G.; Taylor, B. A.; Chamberlain, T. W.; Bourne, R. A. An automated computational approach to kinetic model discrimination and parameter estimation. *React. Chem. Eng.* **2021**, *6* (8), 1404–1411.
- (31) Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nat. Commun.* **2023**, *14* (1), 1403.
- (32) Eppel, S.; Xu, H.; Bismuth, M.; Aspuru-Guzik, A. Computer Vision for Recognition of Materials and Vessels in Chemistry Lab Settings and the Vector-LabPics Data Set. *ACS Cent. Sci.* **2020**, *6* (10), 1743–1752.
- (33) Tian, Z.; Liu, F.; Li, D.; Fernie, A. R.; Chen, W. Strategies for structure elucidation of small molecules based on LC-MS/MS data from complex biological samples. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 5085–5097.
- (34) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *J. Org. Chem.* **2023**, *88* (9), 5239–5241.
- (35) (a) Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; Barzilay, R. MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. *J. Chem. Inf. Model.* **2023**, *63* (7), 1925–1934. (b) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **2020**, *11* (1), 3601.
- (36) *MELLODDY*, 2023. <https://www.melloddy.eu/> (accessed 2023-08-17).
- (37) (a) *Introducing ChatGPT*. OpenAI, 2023. <https://openai.com/blog/chatgpt> (accessed 2023-04-22). (b) Bran, A. M.; Cox, S.; White, A. D.; Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv (Physics, Chemical Physics)*, April 12, 2023, 2304.05376, ver. 2. <https://arxiv.org/abs/2304.05376> (accessed 2023-04-22). (c) Mahjour, B.; Hoffstadt, J.; Cernak, T. Designing Chemical Reaction Arrays using phactor and ChatGPT. *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-2tfvd.